



中国国际法学会
Chinese Society of International Law



大连海事大学
DALIAN MARITIME UNIVERSITY

中国国际法学会2025年学术年会

2025 Annual Academic Conference of the Chinese Society of International Law

暨世界反法西斯战争胜利及联合国成立80周年与国际法研讨会

and the International Law Symposium Commemorating the 80th Anniversary of the Victory of the World Anti-Fascist War and the Founding of the United Nations

加强涉外法治建设与推进中国式现代化：国际法的地位和作用

Strengthening the Rule of Law in Foreign Affairs and Promoting Chinese Modernization: The Status and Role of International Law

论文集

议题十：

网络空间与人工智能的法律规制

主办单位：中国国际法学会 大连海事大学

Co-sponsored by Chinese Society of International Law and Dalian Maritime University

承办单位：大连海事大学法学院

Organized by Law School of Dalian Maritime University

2025年4月25日-27日，大连

April 25-27, 2025, Dalian

目 录

《联合国打击网络犯罪公约》与我国电子取证规则的衔接	裴 炜 童译瑶 1
美国人工智能法律政策下的美式“安全观”及我对策建议 ——基于 DeepSeek 时代全球产业格局	程乐 龚煊 郑英龙 21
人工智能算法透明度规则分析：国际比较、主要困境及中国因应	熊光清 43
人工智能军控：缘起、现状与中国路径	刘宏松 邹宇霆 72
跨国公司参与数据跨境流动治理的生成逻辑及风险应对	李 智 刘思雨 98
人工智能生成数据权利配置的法理逻辑与制度设计	宋云博 陆洋 113
网络空间命运共同体视域下的网络犯罪综合治理——以刑事检察案件办理为例	吴春妹 杨金玲 131
异化与重塑：总体国家安全观下 AIGC 涉意识形态风险治理策略研究	肖楠 魏琦宗 陶羨华 王汉晓 139
中国—东盟打击跨境电信网络诈骗犯罪执法合作研究	董新春 张波 张倩 150
人工智能犯罪的刑法回应与犯罪治理	郑丽君 165
论“塔林手册 3.0”更迭的走向与博弈——兼论网络武装冲突规则的策略对抗	焦园博 177
生成式人工智能全球治理规则构建路径研究	居 梦 周 静 195
网络攻击下自卫权行使困境与中国应对	王韵涵 220
人工智能时代网络恐怖主义犯罪的国际法规制的问题与路径	李 彦 235
应对美国网络霸权的中国涉外法律建设——以美国 Cloud 法案为视角切入	左文君 吕梦飞 254
数据控制者取证模式的实践抵牾、理论调和及应对策略	王 雪 268
《联合国打击网络犯罪公约》中网络性侵儿童犯罪条款及我国刑事法因应	赵安晓宇 297
欧盟人工智能法识别网络恐怖主义信息的实践及对中国的镜鉴	杨凯 314
《联合国打击网络犯罪公约》的多维比较与衔接展望	上海市人民检察院第二分院课题组 323
人工智能主权的结构化分析——兼论中国人工智能主权的制度构建	李艳华 346
构建人类命运共同体与网络空间全球治理	曾璐 368
禁止武力威胁或使用武力视角下的网络攻击问题研究	许钊涵 381
Integration of Soft and Hard Laws: Profiling Legal Protection for “AI for All”	Yang Xiao and Xiaoxia Sun 392
《联合国打击网络犯罪公约》制颁背景下的电子数据证据制度的完善	陈逸宁 420
集体反措施在网络空间适用的合法性问题分析	刘钰祺 436
人工智能的跨国性风险及其法律治理	张磊 于昊昆 453
数字主权与全球数据监管：争议、实践及反思	吕思嘉 466
网络空间与人工智能治理中国际法的作用：现状、原因与未来	于昊昆 486
美国网络空间立场文件中的塑造方法——以主权的重塑和自卫权的强化为例	张予惟 503
全球人工智能治理下的发展鸿沟与平权进路	侯柔倩 515

从林规则还是协同治理？AI 国际治理法律理念的构建.....	罗早西	546
“规模与效果”路径在网络空间国际法适用中的扩张及方法论反思——基于国家立场声明的考察	方镇邦	555
人工智能生成儿童性虐待材料的国际法治理困局与规则重构	高杨洋	578
《非盟共同立场文件》：“全球南方”参与网络空间国际规则博弈的新范式.....	林嘉悦	589
跨境远程访问之法律规制与中国因应.....	徐铭泽	600
WTO 规则下人工智能产品的归类 ——挑战与应对	张鹏文	618
智能汽车：规制措施适用国际商品贸易规则亦或服务贸易规则？	张素敏	628
人类命运共同体视角下的人工智能全球治理问题研究	陈苏	638
Application Of Principles of Responsibility to Space Cyberattacks Constituting International Crimes	Manzoor Hassan Kuan Yang	649
论我国人工智能的风险规制路径：以负面清单制度为中心	徐世泽	676
人工智能对金融科技伦理的挑战及其国际法规制	公惟韬	689
网络生物安全治理困境及破解路径.....	吴 惠	702
网络空间的武力使用与国际人道法中区分原则适用研究	陈熙杨	715
国际安全视域下人工智能武器的人道困境与因应	郝思源	727
论《联合国打击网络犯罪公约》的刑事定罪.....	刘 龙	750
全球卫生治理中人工智能算法监管的法律困境与突破	陈海玲	766
联合国框架下的深度伪造犯罪治理与中国因应.....	杨翔宇	777
网络空间绝对主权论下的网络间谍行为的规制.....	高彬彬	795
智能船舶保险法律问题研究.....	李依霓 陈雅灵	807
论国家主权平等原则在网络空间的适用	林世钦	815
神经数据保护的法理基础与规范构造.....	汪雨馨	832
总体国家安全观视角下我国人工智能治理的困境与对策.....	胡琳英	851
人工智能全球治理：机制构建、挑战与中国的创新路径.....	付译瑶	863
不干涉原则在网络空间国际治理中的适用困境.....	李海旭	875
智能船舶港口国管辖的内涵阐释、困境剖析及规制应对	崔明珠	884
“荷花号”原则的澄清和再发展 ——以网络间谍活动的国际法应对为例.....	文金金	903
人工智能的军事化利用及其法律规制.....	张昕怡	914
人工智能立法视野下无人船责任分配机制的规范进路	王枫予	929
深度伪造技术法律规制的中欧对比.....	顾杰媛	941
网络空间国家行使预先性自卫权的国际法探析——基于对各国国家立场文件的分析.....	张宇星	951
国际法视野下的网络战规制困境与中国因应.....	谢霖	971
智能航运时代智能船舶远程操作中心的法律定位与治理规则重构	朱艺辰	984
网络犯罪的刑事管辖兼论《联合国打击网络犯罪公约》的管辖协调	欧宝盛	998

人工智能自动武器的国际法回应.....陈佳欣 1021

国产开源人工智能兴起影响下数字丝绸之路的制度优化.....何鑫龙 1032

全球人工智能治理的软法机制.....王玉婷 1048

人工智能数据训练的合理使用规则研究——中美欧的比较分析.....吕丹丹 1065

WTO 法下人工智能产品属性认定研究.....吴梓杰 赵汉卿 1078

欧盟《AI 框架公约》的规则输出、理论基础与中国因应.....杨乐天 1089

网络空间治理下的中国国际话语权：基本内涵、面临困境及提升路径.....蒋莉 1107

《联合国打击网络犯罪公约》电子数据分类的国内法转化研究——基于《刑事诉讼法》第四次修改之背景.....杨宇轩 1122

2025年中国国际法学会年会

《联合国打击网络犯罪公约》与我国电子取证规则的衔接

裴 炜 童译瑶

裴炜，北京航空航天大学法学院教授，博士生导师。

童译瑶，北京航空航天大学法学院与德国科隆大学法学院联合培养博士研究生。

【摘 要】《联合国打击网络犯罪公约》在联合国特委会的通过开启了全球网络犯罪治理的新篇章，《公约》通过区分取证主体、类型化目标数据、以及合比例设置侦查措施，为各国在预防和打击网络犯罪过程中开展电子数据取证合作搭建起了一套具有包容性和可操作性的规则框架。我国作为《公约》的主要推动国之一，接下来的重点工作在于积极促进《公约》的后续签批和谈判。但是，我国当前的刑事电子取证规则与《公约》在数据分类、措施设置、管辖权范围等方面存在诸多差异，一定程度上可能构成后续工作的制度障碍，并进一步影响后《公约》时期我国参与乃至主导网络空间电子取证及犯罪追诉的国际合作。当前我国正在进行《刑事诉讼法》的新一轮修订，促进《刑事诉讼法》与《公约》在电子取证方面的规则衔接势在必行。对此，以涉外法治理念为理论指导审视我国目前的规则体系，通过引入数据分类分级制度，将取证措施与目标数据按比例地适配，并将数字权益融入到刑事诉讼体系之内，同时协同刑事电子取证中的公私合作关系。

【关键词】《联合国打击网络犯罪公约》；电子取证；网络犯罪治理；涉外法治

一、引言

近年来，基于网络空间的弱地域性和数据的全球流动和分散存储的特征，加强国际合作、拓展跨境取证方式、进而提升涉犯罪电子数据证据材料的收集提取效能成为世界各国打击网络犯罪立法与实践的共同探索方向，其中具有代表性的是三大立法进程：一是 2021 年网络犯罪《布达佩斯公约》及其《第二附加议定书》（下文简称《<布约>第二议定书》）的制定；¹二是美国《澄清境外数据合法使用法》（下文简称《云法》）及其相关双边协议；²三是欧盟的《刑事电子证据提

¹ 原文参见<https://www.coe.int/en/web/cybercrime/the-budapest-convention>，最后访问日期 2024 年 12 月 20 日。

² 原文参见<https://www.congress.gov/bill/115th-congress/senate-bill/2383/text>，最后访问日期 2024 年 12 月 20 日。美国分别于 2019 年和 2021 年与英国、澳大利亚签署《云法》框架下数据跨境执法双边协议。美国与英国的协定参见<https://www.justice.gov/criminal/criminal-oia/cloud-act-agreement-between-governments-us->

取和存留令条例》(on European Production Orders and European Preservation Orders for electronic evidence in criminal proceedings and for the execution of custodial sentences following criminal proceedings)(下文简称《刑事电子证据条例》)³以及配套的《为刑事诉讼中收集证据目的协同建立指定法律代表规则指令》(laying down harmonised rules on the designation of designated establishments and the appointment of legal representatives for the purpose of gathering electronic evidence in criminal proceedings)(下文简称《代表人协调指令》)。⁴其中,网络犯罪《布达佩斯公约》作为制定时间最早的国际性网络犯罪专门性公约,其在全球范围的辐射力最为广泛;美国《云法》尽管是国内法,但藉由美国本国互联网行业的优势,其建立起的双边协议机制可以将其影响力拓展至国际领域;欧盟的《刑事电子证据条例》主要适用于其成员国,同时其中有大量条款涉及到与欧盟外第三国之间的电子取证合作事项。

上述三大立法进程均旨在修正和补充国际刑事司法协助机制,在成员国或协议方范围内建立适应刑事电子数据取证特别是跨境取证的新规则,例如均就一国侦查机关向他国网络信息业者直接调取涉案电子数据作出了规定。尽管这三大进程贡献了许多网络犯罪电子数据取证的创新做法,但其共同缺陷在于,以某一国家或区域的自身立法价值、制度和规则框架为前置条件,这就意味着相关规则难以关照非成员国家网络犯罪治理的现实情况和实践需求,在多元性和包容性上有所不足,⁵同时其相对的封闭性也无法兼顾非成员国本国的相关制度生态和规则体系,一定程度上形成了规则制定上话语权的的不平等。⁶

有鉴于此,以中国等国为代表的国家或地区始终主张在联合国框架下建立适应当前网络犯罪治理需求、关照各类国家立法状况和司法实践需求的新的公约体系。这一努力最终促成了联合国大会在2019年12月通过第74/247号决议,成立特设委员会并启动《联合国打击网络犯罪公约》(以下简称《公约》)谈判进程。自2022年1月起,特设委员会连续召开8次谈判会议,最终于2024年8月9日正式通过了《公约》并提交给第七十八届联合国大会。⁷其中第四章在吸纳当前世界各国立法与司法实践经验的基础上,形成了一系列针对电子数据取证的专门性

united-kingdom-great-britain-and-northern; 美国与澳大利亚的协定参见

<https://www.justice.gov/criminal/criminal-oia/cloud-act-agreement-between-governments-us-and-australia>, 最后访问日期2024年12月21日。

³ 原文参见<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32023R1543>, 最后访问日期2024年12月20日。

⁴ 原文参见https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_2023.191.01.0181.01.ENG, 最后访问日期2024年12月20日。

⁵ 参见江溯:《打击网络犯罪的国际法新机制》,《法学》2022年第11期,第45-49页。

⁶ 参见胡建生、黄志雄:《打击网络犯罪国际法机制的困境与前景——以欧洲委员会<网络犯罪公约>为视角》,《国际法研究》2016年第6期,第21-34页。

⁷ 草案出处参见<https://documents.un.org/doc/undoc/gen/v24/036/32/pdf/v2403632.pdf>, 最后访问日期2024年12月21日。

规则,这些规则将会在未来重塑各国之间网络犯罪电子取证的实践做法和国际合作。

中国作为公约的主要和积极推动国之一,占据了新国际规则制定的先发优势,但是观察《公约》关于电子取证的相关文本可以发现,《公约》规则与我国电子取证规则之间存在较大差异,这种差异不仅体现在概念运用上,更体现在侦查权的规制逻辑之上。究其原因,一方面在于我国相较于欧美国家和地区在规则输出方面的能力不足,另一方面也揭示出我国当前电子取证措施自身存在的诸多内在问题。

可以预见的是,规则差异如不经适时调和,在《公约》正式通过后势必成为中国开展电子数据取证国际合作中的制度障碍,进而损及我国打击网络犯罪的效能。这种调和需要我国加强参与起草工作并输出规则,更需要我国适时反思现有电子数据取证规则并予以必要调适,从而为将来国内外规则衔接做好理论和制度准备。目前公约已经提交联合国大会投票,相关准备工作已然迫在眉睫。

本文正是由此出发,力图探索我国刑事电子取证规则与《公约》规则的衔接路径。本文主要通过三个部分开展研究。第一部分系统深入分析联合国《公约》电子取证规则的内在结构和逻辑,并总结其规制特征。第二部分着重比较《公约》与我国当前电子取证措施之间的异同,并在此基础上分析其差异成因。第三部分回归我国电子取证规则体系本身,提出调适相关规则的具体建议。

二、《公约》电子取证规则的框架

《公约》关于电子数据取证的规则主要规定第四章“程序措施和执法”中的第25至30条。具体来说包括了快速保全、流量数据部分快速披露、提交令、搜查、扣押、实时收集流量数据、拦截内容数据等7个电子取证措施,并详细规定了每一个取证措施的适用客体、相对人和适用条件或功能目的。同时,根据《公约》第23条关于管辖权的规定,这些措施可以适用于三种情形:(一)《公约》规定的犯罪;(二)其他通过信息和通讯技术实施的犯罪;以及(三)任何涉及电子数据收集提取的犯罪。可以看到,在《公约》框架下,程序措施的适用范围要远远大于刑事定罪,从而为各国侦查机关根据《公约》开展侦查取证活动并进行国际合作提供了极大的便利。

(一)《公约》框架下电子取证措施的类型

就具体的电子取证措施而言,以其实施主体为标准,可以大致将相关措施区分为三类:一是侦查机关自行取证措施;二是通过第三方取证措施;三是合作取证措施。以下分别予以分析。

在《公约》中,侦查机关的自行取证措施主要指向的是“搜查、扣押”(第28条)。考虑到类似措施在不同国家的表述差异,草案特别包含了与搜查、扣押功能类似的取证措施。在《公约》框架下,搜查和扣押措施针对的客体主要包含三

类：一是存储电子数据的物理介质；二是存储电子数据的信息和通信技术系统；三是在以上两者中存储的电子数据（第 28 条第 1 款）。考虑到系统自身的特性和专业性，《公约》特别要求成员国制定相关规则，以要求相关主体在合理范围内，提供搜查和扣押所必需的信息（第 28 条第 4 款）。这些主体主要是指了解相关系统运行、网络，以及相关数据保护措施的人员。

就搜查措施而言，《公约》主要将其划分为两类：一类是本地搜查，针对的是本地的信息和通信技术系统及其中的数据（第 28 条第 1 款）；另一类是远程搜查，针对的是远程信息和通信技术系统及其中的数据（第 28 条第 2 款）。之所以作此分类，在于从权利保障的角度出发，适用于本地搜查之法律授权并不能直接拓展至其他信息和通信技术系统，即便该拓展可以从技术层面予以实现。也正是基于避免任意扩大搜查范围的考虑，《公约》对于远程搜查设置了相应的条件：一是要求侦查人员应当有合理依据认为目标数据存储在另一信息和通信技术系统且该系统位于本国境内；二是要求侦查人员有合理权限从本地系统进入远程系统。

就扣押措施而言，《公约》指向的是通过搜查或类似措施获取的数据，其功能主要在于电子数据的保全（第 28 条第 3 款）。《公约》还进一步明确了该措施实施过程中可以附加的行为：一是制作和保留电子数据的电子形式副本；二是维护载体或系统内存储的电子数据的完整性；三是在访问的系统中移除目标电子数据或使其不可访问。考虑到系统自身的特性和专业性，《公约》特别要求成员国制定相关规则，以要求相关人员在合理范围内，提供搜查和扣押所必需的信息或技术协助（第 28 条第 4 款）。

第二类取证措施是侦查机关通过第三方主体取证的措施，主要包括两种措施：其一是数据保全和数据披露；其二是数据的调取。两项措施均针对的是已经存储的静态数据。

数据保全措施（expedited preservation）主要针对的是有理由相信具有删除、复制、损毁或篡改风险的静态数据，其功能主要在于风险数据的快速保全（第 25 条第 1 款）。原则上，这一措施仅限于留存这一行为，不涉及向侦查机关提供数据。但是，考虑到流量数据（traffic data）之于启动侦查、收集初步证据的重要线索功能，⁸《公约》特别针对此类数据设置了快速保全后的部分披露规则，同时要求该披露仅用于识别服务提供者和信息传输路径（第 26 条）。针对数据保全，考虑到数据主体知情同意权与侦查取证顺利进行之间可能存在的冲突，《公约》特别授权一国有权机关在采取此类措施时，要求相对人就该项措施本身以及涉及的数据予以保密（第 25 条第 3 款）。

⁸ 根据《公约》第 2 条的规定，“流量数据”是指与通过信息和通信技术系统进行的通信相关的任何电子数据，由参与通信链的信息和通信技术系统生成，指示通信的起点、终点、路径、时间、日期、大小、持续时间或所用基础服务的类型。此类数据通常能够呈现出数据行为的路径和节点，对于快速推进后续侦查具有重要作用。

数据调取措施（提交令）（production order）（第 27 条）主要针对的是侦查机关向控制或占有数据的主体调取与案件相关的数据材料，该类措施的特点是相对人与数据权利主体分离。⁹考虑到当前社会中的各类数据主要由网络信息业者等第三方主体占有，而数据所承载的权益复杂多样，《公约》以相对人为标准对此类措施进行了类型划分：在调取相对人为个人时，调取措施不限数据类型；当调取相对人为服务提供者时，则调取措施仅限于该服务提供者基于自身服务占有或控制的注册人信息（subscriber information）（第 27 条 b）。¹⁰

第三类取证措施是侦查机关与第三方共同参与的取证措施，主要涉及的是对动态数据的实时收集或拦截（第 29-30 条）。此类措施集中针对的是通信信息，特征是对公民权利的干预性强、涉及范围广，因此需要设置严格的限制措施。根据数据类型不同，《公约》分别针对流量数据的实时收集和-content 数据的拦截进行了区别规定，二者的主要规则区别在于，鉴于收集内容数据对于公民隐私权、通信权的严重侵犯，《公约》将内容数据的拦截仅限于一国国内法项下的严重犯罪。在其他措施方面，两类信息适用相同规则。第一是仅针对通过本国通信系统、在本国境内传输的特定通信信息；第二是侦查机关既可以自行收集或记录，也可以要求服务提供者自行或协助收集或记录；第三是与快速保全类似，在第三方介入的情况下，授权成员国侦查机关要求该第三方就侦查措施本身以及涉及的数据保密。

（二）《公约》框架下电子取证措施的特点

综合观察《公约》的取证措施体系可以发现，当前规则在借鉴已有国际规则特别是《布约》的基础上，进一步融合了尊重国家主权的要求，并对不同国家的立法与司法实践差异予以了一定程度的关照，基本保持了内部逻辑的一致性。具体而言，《公约》所建立的电子取证措施体系具有以下四个方面的特征。

首先，《公约》区分了不同的侦查取证行为主体，这也是我们在分析草案文本时采用的分类标准。之所以关注取证主体，一方面在于刑事侦查活动遵循法律保留原则和程序法定主义，哪些主体有权力实施具体侦查措施，以及有权机关能否和在何种程度上授权其他主体实施侦查行为，均需要以法律的明文授权为前提；¹¹另一方面在于，当第三方主体参与刑事侦查取证时，法律关系由两方关系转变为多方关系，意味着相应法律权利义务设置、法律责任分配等规则的复杂化。前者在《公约》中主要体现为明确要求成员国通过立法等方式对具体侦查措施予以

⁹ 参见裴炜：《论个人信息的刑事调取——以网络信息业者协助刑事侦查为视角》，《法律科学》2021 年第 3 期，第 80-95 页。

¹⁰ 根据《公约》第 2 条之规定，“注册人信息”主要是指服务提供商持有的与其服务的用户有关的任何信息，但不包括流量数据或内容数据。注册人信息主要包括三种类型：第一类是所使用的信息通信技术服务类型、适用的技术条款、以及服务期限等信息；第二类是注册人身份、邮寄或地理地址、电话或其他联系号码、账单或支付信息等；第三类是与信息通信设备位置相关的任何其他信息。

¹¹ 参见万毅、林喜芬：《现代刑事诉讼法的“帝王”原则：程序法定原则重述》，《当代法学》2006 年第 1 期，第 26-35 页。

授权、以及明确可以介入的第三方主体类型及协助方式等；后者则主要体现为第三方参与侦查取证时可能附加承担的保密义务。

这里需要额外就《公约》针对第三方保密义务的相关规定予以进一步说明。之所以建立这样的规则，一个核心的出发点在于，知情权是个人信息保护的制度基石，基于知情权形成了信息处理者的告知义务。刑事侦查取证中针对个人信息的收集提取是一种典型的处理行为，因此原则上行为主体负有上述告知义务，¹²例如根据欧盟在 2018 年出台的《警察部门使用个人数据实务指南》，该告知义务包含两方面的含义：一是针对社会公众的一般告知；二是针对数据主体的具体告知。¹³在第三方主体作为数据控制者介入侦查取证时，这一义务事实上同时适用于侦查机关和实际开展处理行为的第三方主体，而对此类义务的免除则需要法律的明确许可，例如《指南》同时提出，对该项告知义务的免除需要有明确的法律依据、遵循比例原则、基于合理的理由、以及在个案中进行判断，而这也是《公约》中第三方保密义务的来源。

其次，《公约》对目标数据进行了类型化，并以此作为比例原则框架下匹配具体侦查措施的基准。总体来看，草案主要划分了三种类型的数据：注册人数据、流量数据、以及内容数据。这种分类延续了近年来包括《布约》、欧盟《通用数字保护条例》（下文简称 **GDPR**）等区域立法所采用的数据分类方式，主要针对的信息网络通信技术服务所涉及的数据类型，其背后的考量是不同个人数据所承载的权益在层级上存在差异：一般认为，注册人数据承载的隐私权等权益最弱，流量数据其次，内容数据最强。基于比例原则的要求，针对不同层级的数据权益需要对具体侦查措施设置不同强度的条件限制，因此这一分类也构成了《公约》具体措施设置的基础。

再次，《公约》对具体侦查措施依其对基本权利的干预程度进行了分层级的划分。其中保存措施因为并不直接涉及对外提供数据，对于基本权利的干预最弱，因此受到的限制最低。相较而言，针对动态数据的实时收集提取可能干预的权利主体范围最广、权利类型最多、干预程度最强，因而对应更为严格的程序性条件和限制。流量数据介于两者之间，此类数据的累积收集可以反映出私人生活及其习惯特征，¹⁴对于隐私的侵犯程度仍然可能较强，因此针对此类数据无论是静态

¹² 参见裴炜、张桂贤：《论刑事诉讼中个人信息保护的知情规则》，《成都理工大学学报(社会科学版)》2022 年第 4 期，第 50-62 页。

¹³ 原文参见“Practical Guide on the Use of Personal Data in the Police Sector”，<https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjwgPL5ruv7AhWMUsAKHWCXCDYQFnoECEEQAAQ&url=https%3A%2F%2Frm.coe.int%2Ft-pd-201-01-practical-guide-on-the-use-of-personal-data-in-the-police-%2F16807927d5&usg=AOvVaw03JHMM4wVBvNYfOqQYDKBi>，最后访问日期 2024 年 12 月 21 日。

¹⁴ See Committee of Convention 108, “Opinion on the Draft Second Additional Protocol (Version 2) to the Convention on Cybercrime on Enhanced Cooperation and Disclosure of Electronic Evidence”，<https://rm.coe.int/opinion-of-the-committee-of-convention-108-on-the-draft-second-additio/1680a26489>，最后访问日期 2024 年 12 月 21 日。

状态下的有限提供还是动态状态下的实时收集,《公约》均对其进行了单独规制。

最后,基于尊重国家主权的原则,对具体侦查措施的域外效力进行了分类限制。《公约》在第5条确立了保护主权原则,要求所有成员国在打击相关犯罪过程中遵循主权平等和国家领土完整原则。这一原则也体现在具体的侦查取证措施上。可以看到的是,《公约》试图在提升跨境取证效能与尊重他国主权之间寻求一种分层级和合比例的平衡,一是将侵犯性最强的数据实时收集提取措施排除在跨境适用之外;二是针对侵犯性相对较弱的静态数据调取措施,要求相对人位于境内,但对于目标数据所在地没有具体要求;三是针对不直接涉及数据境外流转的保存措施则原则上不限制跨境执法活动。

在此,《公约》体现出了管辖连接点多元化的特征,即同时采取了数据所在地和控制者所在地两种管辖模式,具体来说,《公约》中的提交令采取了控制者所在地模式,侦查机关可以要求位于其本国境内的个人或在其境内提供服务的服务提供商提供数据,而不论数据位于其本国境内还是境外。而在搜查、扣押,实时收集和拦截等措施上,《公约》又采取了数据所在地模式,仅允许侦查机关对位于其本国境内或在其境内传输处理的数据进行搜查、扣押,实时收集和拦截。结合上文关于措施分类的标准看,《公约》在由侦查机关单独实施的取证活动中强调数据的本地性,在涉及第三方的情况下则更强调相对人的本地性。

总体上看,《公约》通过以下步骤构建起了其电子取证措施的规则框架:首先以涉案数据为基点,划分了侦查机关单独实施和涉及第三方的两大类取证措施类型,并据此合理界定了侦查机关和第三方的取证权限。其次,区分了注册人信息、流量数据以及内容数据,然后在此基础上依据比例原则逐级设置了快速保全、披露、调取、搜查、扣押、实时收集、监听等措施。最后,综合考虑实施主体的权限和取证措施干预强度,个性化地设置了不同的管辖权连接点。

三、《公约》与我国电子取证规则的对比

通过全面审视《公约》中涉及电子数据取证的相关条款,我们可以观察到贯穿于其中的相对统一的逻辑、模式和特征。尽管部分条文仍然存在术语等方面的争议,但整体框架已然搭建起来,成为我国加入公约体系的重要制度前提。如前所述,《公约》框架下的电子数据取证措施主要包括快速保全、快速披露与提交令、数据搜查和扣押、以及动态数据的收集和拦截等。基于功能等同的考量,上述措施或多或少能在我国现有刑事电子数据证据规则中找到类似措施。但从现有的草案文本来看,《公约》框架下的各类侦查措施与我国的相关制度规定差异较大,未来可能成为我国国内法与国际规则衔接的障碍。

(一) 快速保全与电子数据冻结

针对快速保全措施而言,我国电子数据证据规则中与其功能较为类似的措施

是“电子数据冻结”，¹⁵并且主要匹配的是其中“提取时间长，可能造成电子数据被篡改或者灭失”这一具体情形，¹⁶笔者将其称之为“保全冻结”。¹⁷主要涉及我国《刑事诉讼法》第144条、《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》（《电子数据规定》）第11条、《公安机关办理刑事案件电子数据取证规则》（以下简称《电子取证规则》）第36-39条等相关规定。二者的目的均在于对目标电子数据予以固定，以便于后续的犯罪追诉活动。尽管我国电子数据冻结规则在适用情形上区分了保全冻结和便宜冻结，但是这种区分并未体现在后续的规则设计上，因此整体上冻结的适用范围广于《公约》中的快速保全，也由此形成了两组规则在以下四方面的差异。

第一是在适用条件上，我国的数据保全冻结在强调数据的篡改或灭失风险的同时，也强调该风险源自于“提取时间长”。¹⁸但是，除提取时间这个因素外，还有其他因素可能导致数据损毁、灭失风险，《公约》也强调数据灭失的风险既可能源自于数据本身的特征，也可能源自于法律所规定的数据存留期限即将届满。正是在这一意义上，个案中的快速保全措施被视为是对数据存留规则缺失或期限不足的必要补充。¹⁹同时，我国的电子数据冻结还适用于另外两种情形：一是数据量大，无法或不便提取；二是通过网络应用可以更直观展示数据。

第二是在适用期限上，基于快速保全措施的保全性、临时性、紧迫性的性质考量，《公约》对于其设置了不超过90日的最长期限，该期限与《布约》第16条保持了一致，与我国现有规则中各类冻结情形一概适用6个月的固定期限形成差异。二者均允许后续进行延期，且均未限定延期次数，但《公约》未明确延期的期限，按照文义解释，原则上应该也不能超过90日；我国冻结措施的延期则是单次不得超过6个月。

第三是在较之《公约》针对注册人信息、流量数据和内容数据的规则而言适用对象更广。适用对象上，我国的数据冻结并不区分数据类型，原则上可以适用于任何与案件有关的电子数据，

第四是在相对人的保密义务上，我国数据冻结规则并未设置个案中要求实施

¹⁵ 数据保存（data preservation）与数据存留（data retention）是两种不同的措施，前者是基于个案情况进行的判断，对应的是相对人在个案中的法律义务，这一概念有时也采用“quick freeze”的表述；而後者的对象、范围、期限均以明确的法律规定为前提，是事前确定的法律义务。我国的数据冻结措施更类似与前者。参见European Commission, “Evidence of Potential Impacts of Options for Revising the Data Retention Directive: Current Approaches to Data Preservation in the EU and in Third Countries”, issued in November 2012. Available at https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/police-cooperation/data-retention/docs/drd_task_2_report_final_en.pdf, 最后访问日期2024年12月21日。

¹⁶ 根据“两高一部”《电子数据规定》第11条，电子数据冻结适用的情形主要包括：（一）数据量大，无法或者不便提取的；（二）提取时间长，可能造成电子数据被篡改或者灭失的；（三）通过网络应用可以更为直观地展示电子数据的；（四）其他需要冻结的情形。

¹⁷ 参见裴炜：《论刑事电子取证中的数据冻结》，《北方法学》2021年第2期，第3-22页。

¹⁸ 参见喻海松：《网络犯罪二十讲（第二版）》，北京：法律出版社，2022年，第226页。

¹⁹ See Adam Juszczak and Elisa Sason, “Recalibrating Data Retention in the EU: The Jurisprudence of the Court of Justice of the EU on Data Retention – Is This the End or is This the Beginning?” EUCRIM, 4(2021): 238.

冻结的单位或人员对措施及所涉数据保密的规则。我国《刑事诉讼法》规定了公安司法机关对办案过程中获知国家秘密、商业秘密和个人隐私的保密义务（第54条第3款），一则没有明确规定义务主体，将其直接扩展至相对人在合法性上存疑；²⁰二则保密义务指向的主要是社会公众，并非因个人数据被处理而享有知情权的数据主体。这两点均与《公约》的规则逻辑相区别，因此难以直接以该条款来作为两套规则的对接条款。

（二）流量数据部分披露、提交令与数据调取

数据调取和流量数据的部分披露均涉及到侦查机关从第三方获取涉案电子数据，这两类措施在功能上与我国的电子数据调取措施类似，但其区别也较为明显。主要涉及《刑事诉讼法》第54条、《电子数据规定》第13条、《电子取证规则》第41条等相关规定。

就快速披露而言，其与我国的数据调取措施的主要不同之处在于适用的条件和范围。《公约》中的快速披露措施是快速保全措施的附带性措施，二者存在先后关系，因此其适用目的、范围和条件受限于快速保全措施。如果数据本身不存在丢失或被篡改的高风险，则原则上其不能使用快速保全措施，进而也不能适用后续的快速披露措施。但我国的调取措施本身是一种独立的侦查取证措施，并不以服务其他措施为目的，适用范围较广。

就提交令而言，与我国的数据调取的共性在于目标数据为第三人占有或控制，并且目标数据均是已经存储的静态数据。二者主要存在以下方面的差异：第一，适用对象不同。与数据冻结类似，我国的数据调取措施并不区分不同类型的数据，对各类数据采用同一套规则体系，这与《公约》的整体基础架构存在区别。第二，相对人不同。《公约》以数据控制者所处地理位置区分了提交令的相对人。针对位于本国境内的相对人，无论是自然人还是服务提供者，有权机关均可向其调取各种类型的数据。针对位于本国境外的相对人，由于向其取证会涉及执法管辖权问题，《公约》设置了严格的条件：一是该相对人仅限于商主体，一般自然人不适用；二是该商主体必须通过向本国境内提供服务来建立本国的管辖权连接点，不提供服务的商主体不适用；三是目标数据必须与该服务有关，与该服务无关的数据不适用；四是目标数据仅限于注册人信息一种，流量数据和内容数据不适用。相较而言，我国的调取措施并不做此区分，概括性地适用于“电子数据持有人、网络服务提供者或者有关部门”。²¹同时，根据《国际刑事司法协助法》的规定，该措施应当仅限于“位于我国境内”的相对人，这就限制了我国有权机关向境外网

²⁰ 有观点认为，该项义务的主体既包括公安司法机关，又包括作为相对人的单位和个人。参见孙茂利主编：《〈公安机关办理刑事案件程序规定〉释义与实务指南》，北京：中国长安出版传媒有限公司，2021年，第150页。但是结合《刑事诉讼法》的其他条文来看，这一点难以成立，例如《刑事诉讼法》在规定技术侦查措施时，对有关单位和个人单独明确规定了保密义务（第152条第4款）。如果第54条的保密义务也想扩展至具体执行或提供协助的单位和个人，则应当采用相同的规范方式。

²¹ 参见“两高一部”《电子数据规定》第13条。

络信息业者调取数据的空间。

此外，还需要注意的是，我国《数据安全法》要求公安机关因犯罪侦查调取数据时，需要“经过严格的批准手续”（第35条），该规则是否以及在何种程度上会影响到刑事电子数据相关程序性规则，从目前的刑事诉讼法的回应上尚看不出明确的关系指引，但从协调规范刑事诉讼程序中的相关数据处理行为的角度出发，未来统合两法关系并对刑事诉讼规则进行相应调整应当是必然的趋势。²²

（三）数据搜查、扣押与电子数据提取

搜查和扣押作为《公约》中涉及的一项主要的的数据取证措施，与各国传统侦查措施的对应性相对较强。但是在我国刑事电子数据取证规则体系中，类似功能的措施并未归属于刑事诉讼法侦查框架下的搜查与扣押，而是另起炉灶地创设了“电子数据提取”措施。²³此外，我国的远程勘验措施也会涉及一定的数据搜索活动，因而也与《公约》搜查有重叠之处。结合《电子数据规定》和《电子取证规则》的相关规定，电子数据提取措施可以划分为两类：第一类是本地提取电子数据；第二类是网络在线提取电子数据。其中，本地提取又可以区分为载体扣押后的提取和现场提取。²⁴这一分类与《公约》区分本地搜查与远程搜查的思路类似。同时，无论针对搜查还是扣押，在我国法项下侦查机关均有权要求相关人员提供协助，²⁵因此可以涵盖《公约》第28条第4款所规定的协助义务。

两套规则的差异主要集中于以下两个方面。

第一，适用对象不同。一方面，《公约》中的数据搜查和扣押的适用对象不仅指向数据本身，还指向系统和数据存储介质。而在我国法项下，网络在线提取主要指向的是数据本身，对于系统和介质的查看行为附着于数据提取本身，并无单独规定。另一方面，由于针对远程信息和通信技术系统及其内部数据情况的调查，还涉及到“网络远程勘验”措施的使用。因此，仅就数据类型而言，《公约》搜查措施针对的是已存储的静态数据，我国的提取措施并无此限制，甚至可以拓展至动态数据。

第二，管辖范围不同。《公约》将此类搜查和扣押限定于境内存储的数据，而不论数据的公开状态。相较而言，我国网络在线提取措施区分了境内存储和境外存储，针对境外存储的电子数据的在线提取仅限于公开发布的情形，但对于境

²² 相关探讨参见谢登科：《论侦查机关电子数据调取权及其程序控制——以〈数据安全法（草案）〉第32条为视角》，《环球法律评论》2021年第1期，第52-67页；裴炜：《论个人信息的刑事调取——以网络信息业者协助刑事侦查为视角》，《法律科学》2021年第3期，第80-95页。

²³ 参见刘品新：《电子证据法》，北京：中国人民大学出版社，2021年，第190页。

²⁴ 根据《电子取证规则》第三节“现场提取电子数据”的规定，此类提取主要适用于无法扣押原始存储介质的情形。

²⁵ 《刑事诉讼法》第52条规定，“必须保证一切与案件有关或者了解案情的公民，有客观地充分地提供证据的条件，除特殊情况外，可以吸收他们协助调查。”同时，《网络安全法》第28条规定，“网络运营者应当为公安机关、国家安全机关依法维护国家安全和侦查犯罪的活动提供技术支持和协助。”

内存储的电子数据在范围上并无限制。²⁶

第三，数据保全方式不同。《公约》并未在搜查措施中规定对于目标数据的保全，而是将其放置在了后续的扣押措施之中。但是，我国并未专门就数据的扣押另行规定，在我国当前电子数据取证制度框架下，“扣押”仅指向数据载体，而不指向数据本身。²⁷因此，对于数据的保全统一在提取部分进行了规定。换言之，针对提取后的电子数据，侦查机关自然而然地可以持续占有，而无需再单独适用其他侦查措施。

第四，权力内容不同。首先，《公约》的搜查仅限于搜索行为，我国的远程勘验措施还允许对目标系统进行变动和数据收集。其次，《公约》将保护电子数据完整性的方法作为与“扣押”相类似的措施并列为取证措施，但我国现行法中所列举的各项保护电子数据完整性的方法，并不属于单独的侦查取证措施。最后，在扣押或其类似措施的规定中，《公约》还允许侦查机关屏蔽或移除访问系统中的数据，这无法直接在我国的数据提取措施中找到相应规定。不过，我国针对电子数据的冻结措施可以在一定程度上发挥类似的“使数据不可获取”的效果。

第五，附加措施不同。目标系统、存储介质和数据可能存在加密等安全保障措施，这些措施可能妨碍数据的提取，因而《公约》授权成员国要求“了解相关信息和通信技术系统的运行、信息和电信网络或其组成部分或为保护其中的电子数据而适用的措施的任何人在合理情况下提供必要信息”。我国相关电子数据规定并未明确规定特定人员在数据提取时的协助义务，但是在《网络安全法》等法律中规定了诸如网络运营者提供技术支持和协助的义务，这些义务可以被解释为与《公约》搜查的相关附带措施相类似的规定。同时，我国《电子取证规则》关于网络在线提取的规定，要求有权机关“应当使用电子数据持有人、网络服务提供者提供的用户名、密码等远程计算机信息系统访问权限”（第33条第1款），一定程度上也需要对应相关主体的信息提供义务。

（四）动态数据实时收集与技术侦查

在《公约》目前建立起的取证措施体系下，动态数据的实时收集可以说是权利干预性最强、限制最为严格的一类措施，其中对于通信内容的动态收集直接被定义为“监听”。在我国侦查措施框架下，类似措施主要指向技术侦查。此外，网络在线提取和远程勘验措施也会涉及一些动态数据的收集，这也是在我国目前的电子取证规则中，上述三项措施串套在一起规定的原因。而且，《电子数据规定》规定在进行网络远程勘验过程中可采取技术侦查措施（第9条第3款），这一规定事实上将技术侦查视为了网络远程勘验的下位措施，同时网络远程勘验本身也可能涉及到对远程信息和通信技术系统内部动态数据的监控提取，因此也应当将

²⁶ 有学者提出，这种制度设计事实上导致境内远程提取的权利保障低于境外提取。参见谢登科：《电子数据网络在线提取规则反思与重构》，《东方法学》2020年第3期，第89-100页。

²⁷ 参见裴炜：《论刑事电子取证中的载体扣押》，《中国刑事法杂志》2020年第4期，第120-136页。

其划入到对应的本国措施之中。基于此,我国当前的相关电子取证措施与《公约》中的相关措施主要存在以下差异。²⁸

第一,适用对象不同。与前述措施类似,我国相关电子取证措施并非以数据类型化作为措施设置的参照系,因此技术侦查和网络远程勘验也并无限定且不区分其适用的数据类型。

第二,措施定性不同。在《公约》的措施体系下,动态数据的实时提取是对公民基本权利干预性最强的一类措施,因而其适用的范围最窄、受到的限制也最为严格。相较而言,在我国现有规则框架下,勘验与技术侦查这两项措施实则分属于侦查措施强制性的谱系的两端,即勘验通常被视为是一项任意性侦查措施,而技术侦查则不仅是强制性侦查措施,且属于强制性最强的一类。将二者进行嵌套式规定,实则打破了刑事诉讼法以权利干预为基准所建立起来的侦查措施体系。

第三,适用范围不同。这主要体现在对内容数据的监听措施上。考虑到此类措施的强权利干预属性,《公约》将其适用范围限定于成员国国内法所规定的严重犯罪。单纯从刑罚的严厉程度看,公约所称之严重犯罪是指构成犯罪且最高可处以至少四年有期徒刑或更重惩处的行为。对应到我国,根据当前的犯罪轻重划分的一般标准,严重犯罪则是指可能判处三年以上有期徒刑的犯罪。但若从取证措施之功能的角度出发,对应的技术侦查措施的适用存在三个标准,既取决于其犯罪类型,具体涉及“危害国家安全犯罪、恐怖活动犯罪、黑社会性质的组织犯罪、重大毒品犯罪”,又取决于其严重程度,即与上述罪名危害程度相当的“其他严重危害社会的犯罪案件”或“对于利用职权实施的严重侵犯公民人身权利的重大犯罪案件”,还可能适用于“追捕被通缉或者批准、决定逮捕的在逃的犯罪嫌疑人、被告人”。无论如何,其适用标准显然难以与公约标准相匹配。

第四,管辖范围不同。如前所述,《公约》将动态数据的实时收集限定于境内,即“本国境内信息和通信系统传输的通信”,这也与此类措施的强权利干预属性相对应。从我国与之相对的两项措施来看,其适用范围存在争议。一种观点认为,鉴于远程勘验和技术侦查是规定在网络在线提取措施项下,而网络在线提取又主要适用于境内,²⁹因此在电子数据取证的场景下,远程勘验和技术侦查也适用同一地域限制。³⁰问题在于,远程勘验恰恰是判断数据具体情况、进而决定能否采用网络在线提取的前置性措施,因此以网络在线提取的境内适用来限制网络远程勘验无疑是颠倒了二者的顺序。类似地,我们可以看到国际层面一些国家在运行侦查机关在目标数据位置不明的情况下使用技术侦查措施,然后在查明具体数据位置之后要求侦查机关恢复到正常的有管辖权的侦查措施中去。从这个角度

²⁸ 参见裴炜:《论远程勘验:基于侦查措施体系性检视的分析》,《政法论坛》2022年第4期,第156-166页。

²⁹ 根据《电子取证规则》第23条,针对境外数据的网络在线提取仅限于“公开发布的电子数据”。

³⁰ 参见谢登科:《电子数据网络在线提取规则反思与重构》,《东方法学》2020年第3期,第89-100页。

讲，难以将远程勘验和当前附属于其下的技术侦查限定于境内。

第五，附加措施不同。与前述藉由第三方实施的侦查措施类似，涉及到参与或协助侦查的第三方主体的保密义务。对此，我国同样并未要求远程勘验措施涉及的相对方承担相应的保密义务。同样，仅要求公安机关电子数据取证涉及国家秘密、警务工作秘密、商业秘密、个人隐私的，应当保密，至多要求参与进行鉴定的公安部指定机构承担保密义务。³¹ 然而，该规则一方面无法推及至相对人，另一方面也无法涵盖勘验过程所涉及的电子数据持有人、网络服务提供者等主体。就此而言，涉及网络勘验时要求服务提供商对相关权力的行使情况以及与之相关的任何信息进行保密于法无据，该规则无法与《公约》的规则逻辑相匹配。

总体上看，两套规则在以下三方面存在较大差异：一是数据分类，我国一般不区分不同类型的数据，但区分境内数据和境外数据，公开数据和非公开数据，而公约将其区分为注册人信息、流量数据以及内容数据。二是取证措施的适用对象，公约基于数据分类设计了针对不同数据类型，干预程度不一的侦查措施，而我国无数据分类自然也就没有取证措施的分层。三是第三方的义务范围，在前述规定的基础上，公约仅允许调取第三方基于自身服务占有或控制的注册人信息，实时收集流量数据、拦截内容数据，并要求第三方为此承担保密义务，我国对此并无具体规定。可以说，与公约规则相比，我国规则还略显宽泛。

四、 规则差异成因及其影响

《公约》与我国刑事诉讼电子取证措施规则之所以存在上述差异在很大程度上是因为两部法律的立法视角存在重要差别，集中体现在我国刑事诉讼法所采用的两个基本思路之上：第一是国内法的内向型立法思路；第二是实用主义的立法思路。在二者的共同作用下，我国法律在落实公约规定时面临着如何确保人权不受侵犯，合理协调与第三方主体的关系以及统一国内法和国际主张的障碍。

（一）我国刑事诉讼法立法的基本思路

首先，我国刑事诉讼法作为国内法，长期采用的是内向型的立法思路，相关制度建设主要关注国内犯罪的打击和诉讼程序的运作，对于跨国犯罪以及带有涉外色彩的问题则根据相关国际法的规定进行处理，在国内法层面仅就此予以确认。³²这既是主权原则的体现，也是传统刑事诉讼法主要关注境内犯罪的自然结果。从内向型思路出发，我国刑事诉讼法针对电子数据取证，一方面试图拓展执法管辖权以获取更多的境外数据，另一方面对于外国执法机关针对我国境内数据的收集提取则设置较为严格的程序限制。

³¹ 《电子取证规则》第4条：“公安机关电子数据取证涉及国家秘密、警务工作秘密、商业秘密、个人隐私的，应当保密；对于获取的材料与案件无关的，应当及时退还或者销毁。”第58条：“公安部指定的机构应当按照法律规定和司法审判机关要求承担回避、保密、出庭作证等义务，并对报告的真实性、合法性负责。”

³² 裴炜：《涉外法治视野下刑事诉讼的数字化进路》，《中国刑事法杂志》2024年第2期，第124-142页。

相较而言,《公约》作为联合国层面首部关于网络犯罪的国际公约,旨在协调各国的刑事司法资源,促进各国在打击网络犯罪问题上的共识,以提升全球打击网络犯罪的效率。在此过程中,《公约》必然是以包容的全球化视野对成员国诉求进行审视和平衡的,一方面寻求各国对取证措施类型的认可度的基础上实现取证措施的普遍性适用,另一方面则要在兼顾各国对侦查措施干预性的容忍度的基础上发挥这些措施的最大化作用。

其次,我国刑事诉讼法所确立的电子取证规则具有较强的实用主义导向,是在公安司法机关总结自身实践经验的基础上进行的针对具体取证场景和问题所做的规定,强调的是实用性和可操作性,诸如合法性的问题尽管有所涉及,但其主要服务于保障电子数据的真实性和可靠性。这种实用主义的导向一方面与《刑事诉讼法》所建立的侦查措施体系存在诸多差异,另一方面也与《个人信息保护法》《数据安全法》等数字法的相关要求不相一致。

《公约》规则是经过各成员国、观察员国以及多利益攸关方多轮谈判凝聚的共识,其制定的目的之一就在于最大程度地弥合各国法律制度的差异以提高打击网络犯罪的效能,因而需要具有较高的普适性,以兼容打击网络犯罪的需求和各国法律制度的差异。为此,《公约》在规则设计时大量参考了《联合国反腐败公约》等既有公约的规定,并部分吸收了《布约》等区域性协议的内容。³³同时,在《公约》谈判启动以前,联合国的全面研究网络犯罪问题专家组还针对全球网络犯罪现状发布了《网络犯罪问题综合研究报告(草案)》(Draft Comprehensive Study on Cybercrime),对世界各国的网络犯罪现状进行了综合性概览,以全面反映网络犯罪的治理和打击所面临的难题,并力图寻求获得国家主权和打击犯罪之间的平衡。

(二) 立法思路差异下的法法衔接障碍

立法思路上的差异一定程度上导致了两套规则的差异,进而有可能在未来我国在签批《公约》以及后续的国内法转化方面,形成电子数据取证方面的制度和实践障碍。具体而言,这些障碍主要体现在以下方面。

第一,两套规则在数据分类分层思维上的差异会导致其对人权保障水平的不一,可能会减损证据的可采性。如前所述,公约在分类分层理念的指导下,为不同的数据类型配置了相应干预强度的取证措施,保证了取证措施对相对人权利的最小侵害。我国电子取证制度并未作出此类区分,这也就使得不同数据所承载的公民基本权益只能得到平均化的甚至是最低水平的保障。同时,这种对数据不加区分的立法模式也在事实上造成了我国自身立法对于公民权益保障的内外差异。

³³ 公约谈判过程中,特委会还曾整理总结过一份汇集了既有的关于打击网络犯罪的司法文件的报告“Overview of existing instruments, recommendations and other documents on countering the use of information and communications technologies for criminal purposes”,详见https://www.unodc.org/documents/Cybercrime/AdHocCommittee/Second_session/CRP10.pdf,最后访问日期2024年12月21日。

一方面,我国网络在线提取措施对境内存储和境外存储的数据设置了不同的使用门槛,对于境外数据的在线提取仅限于“公开发布的数据”,而境内数据则不论是否“公开”均可进行在线提取,这就导致对境外权利保护水平高于境内权利的情形出现。另一方面,我国在对境外证据进行审查时采用的是“绝对本国法”的审查规则,³⁴并不关注该证据的收集提取是否符合境外法律规定,只强调“证据材料移送过程中程序的合法性”,³⁵此时境外证据可能在不符合境外法律规定的前提下被我国刑事诉讼程序所接受。由于电子数据之上承载着公民数据财产权、隐私权、通信自由权等基本权利,而我国规则侧重于从取证技术层面保障电子数据的真实性与可靠性,对其中的权利保障问题关注不足,³⁶这可能会引发权利滥用甚至是侵犯权利的问题。

事实上,我国在境外证据审查这一问题上的做法就很可能导致境外证据材料的可采性门槛低于境内证据,同时变相降低了我国刑事诉讼的证明标准,在这种情况下所做出的判决难言正当或合法。除了会招致批评令国际合作无法顺利开展以外,这种不对等的保护水平也违背了平等互惠的原则,上述降低对境外证据的进行取证和采信门槛的做法,既不利于国际合作的开展,还会对我国国内数据的保护造成反噬,即国外主管机关同样可以以此为由无视我国的数据跨境规则。

第二,两套规则对待第三方协助取证的不同方案令第三方在协助取证时合规义务的加重,可能导致第三方的消极配合,阻碍取证的顺利进行。首先,在犯罪追诉的过程中,《公约》可能成为第三方拒绝配合我国侦查机关进行数据调取或动态数据的实时收集等活动的理由,从而导致我国的取证措施无法顺利实现对数据的提取和收集。例如,《公约》并未要求第三方协助侦查机关调取注册人信息以外的数据,而我国的数据调取规则的适用对象包括所有类型的数据,并笼统地要求第三方主体为侦查提供技术支持和协助。

其次,从犯罪预防上看,公约与我国法律对第三方主体采取了不同的态度。公约更多地将第三方主体视为合作伙伴,强调公私合作伙伴关系的建立,重视第三方主体可能为犯罪预防提供的帮助,对此规定较为宽泛。而我国则通过《反有组织犯罪法》《网络安全法》《未成年人保护法》等一系列法律赋予了第三方主体以犯罪预防的义务,要求其承担一定的“主体责任”。同时,我国不区分目标数据的取证规则也会加重第三方运营成本,因为其需要在日常业务开展过程中进行更为广泛的快速保全和维护工作以备司法机关的调取。

两相对比之下,我国对第三方设置了更为具体的犯罪预防义务与责任。此外,

³⁴ 参见吴国章:《刑事境外证据“绝对本国法”审查规则的反思与重塑》,《法学》2023年第10期,第125-142页。

³⁵ 2021年最高人民检察院《人民检察院办理网络犯罪案件规定》规定,人民检察院在审查境外电子数据合法性时,要注重审查收集、提取行为是否符合国(区)际司法协作及相关法律规定的要求(第32条),关注证据来源是否合法、手续是否齐备以及证据的移交、保管、转换等程序是否连续、规范(第59条)。

³⁶ 参见谢登科:《论电子数据收集中的权利保障》,《兰州学刊》2020年第12期,第33-45页。

网络信息服务提供者基于对数据的占有、控制和处理能力广泛参与到刑事侦查当中，但数据的储存并不受限于网络信息服务提供商的注册地或所在地，而是广泛地散布在全球各地，因此尽管是向境内主体或个人实施调取等措施，第三方也可能同时面临我国法、外国法以及公约的三重法律冲突问题。事实上，在公约谈判过程中就有不少观点指出，过重的第三方义务可能会形成规则壁垒，进而阻碍国际合作。³⁷

第三，两套规则的差异实际上提示了我国司法实践和我国国际主张之间可能存在矛盾之处，这会令我国通过公约制定传递话语权的能力大打折扣。我国的电子取证规则的制度框架建立在司法实践的经验之上，³⁸尽管能够为我国刑事诉讼的顺利进行带来便利，但其中所规定的跨境电子取证措施实际均为单边措施，这或多或少与我国在公约谈判过程中的主张和立场相悖。

比如，尽管我国拒绝以及批评《布达佩斯公约》（以下简称《布约》）的重要理由之一就是《布约》第 32 条 b 允许其成员国在无需征得另一国授权的情况下，凭借权利人的同意访问另一国的数据，实际上构成了域外取证。³⁹ 但我国的网络远程勘验在一定程度上与之产生重合。⁴⁰又如，在《公约》谈判之初我国即明确提出，“一国不得违背数据存储服务国法律，直接向企业、个人或采用绕过网络安全保护措施的技术手段等方式调取存储在外国数据。”⁴¹然而，该主张与《电子取证规则》允许对公开发布的境外电子数据进行网络在线提取的规定也存在张力。此外，根据《国际刑事司法协助法》《数据安全法》《网络安全法》等相关法律规定，非经中华人民共和国主管机关批准，境内的组织、个人不得向外国司法或者执法机构提供存储于中华人民共和国境内的数据。这就意味着，我国可以直接对境外公开数据进行在线提取的同时，限制外国司法或者执法机构对储存在我国的公开数据进行提取。

³⁷ 特设委员会第五次闭会期间磋商的主席报告中指出了这一点。参见“Chair’s Report of the Fifth Intersessional Consultation of the Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes”, https://www.unodc.org/documents/Cybercrime/AdHocCommittee/6th_Session/Pre-session-docs/2313717E.pdf最后访问日期 2024 年 12 月 21 日。

³⁸ 有学者通过对“境外服务器”案件进行统计调查得出结论认为，这两部法规所确立的跨境电子证据取证制度就是对近年来司法实践中公安机关境外电子证据取证方式的总结。参见叶媛博：《我国跨境电子取证制度的现实考察与完善路径》，《河北法学》2019 年第 11 期，第 106-119 页。

³⁹ 参见“中国代表团出席联合国网络犯罪问题专家组首次会议并做发言”，外交部官网，https://www.mfa.gov.cn/web/gjhdq_676201/gjhdqzz_681964/lhg_681966/zyjh_681976/201104/t20110402_9381429.shtml。访问日期 2024 年 11 月 5 日。

⁴⁰ 例如在某非法猎取、控制计算机信息系统数据案中，嫌疑人焦某归案后主动向公安机关提供了位于美国的另一台主控服务器的 IP 地址、用户名和密码。公安人员在经过“远程勘验”在该服务器上提取了相应的“主控程序”以及“登录日志”和“主控列表”等电子数据。参见《湖北省武汉市中级人民法院刑事判决书》(2016)鄂 01 刑终字第 176 号。

⁴¹ 参见《中国对联合国打击为犯罪目的使用信通技术公约范围、目标和框架（要素）的建议》，于 2022 年提交《公约》特委会第一次会议，引自 https://www.unodc.org/documents/Cybercrime/AdHocCommittee/First_session/Comments/Chinas_Suggestions_on_the_Scope_Objectives_and_Structure_AHC.pdf，最后访问日期 2024 年 11 月 5 日。

上述规定的差异形成了一个事实上的制度状态,即我国一方面为本国国家机关开展单边跨境取证留有了空间,另一方面对于对外的数据协助又施加了非常严格的规则限制,形成了电子取证的内宽外严的现状。这种状况可能阻碍我国在推动《公约》通过之后的后续签批和落地工作。

五、《公约》与我国电子取证规则的调适

为了实现公约与我国电子取证规则的顺畅衔接,消除权利保障水平不一、第三方合规义务重、国内立法与国际主张冲突等障碍,需要对两套规则的差异进行调和,调整我国电子取证措施的框架,并结合数字法的相关规定对我国电子取证措施的细节进行优化,同时妥善处理与第三方主体的合作关系。为此,应对我国刑事诉讼法的电子取证规则进行以下调整。

(一) 以涉外法治理念为理论指导

在数字化背景之下,犯罪开始超出地域限制,表现出全球化、跨国性的特点:犯罪活动借由网络空间而呈现出的超越物理边界的特征,犯罪活动的人、财、物可以轻易向境外拓展和迁移。⁴²随之而来的是刑事证据的数字化和全球化分散分布趋势。⁴³这一方面表现为电子数据通过网络实现了全球范围内的快速、自由流转,对建立在物理现实空间基础之上的传统证据规则造成冲击。另一方面,第三方主体作为电子数据的主要持有者,越来越频繁地参与到犯罪追诉和犯罪预防活动当中,这一角色与其作为商业主体天然的全球扩张驱动性相叠加,又进一步强化了证据全球化分散分布的特性。由此可见,涉外因素越来越深刻的影响着刑事电子取证活动,其亟待纳入刑事司法立法的整体考量之中。

目前,我国关于电子数据取证的规则散见于低位阶的刑事诉讼规范性文件,《刑事诉讼法》尚未对此作出系统的回应,这就导致相关规则出现了上文所提及的对境内外数据的保护水平不平等,国际和国内的立场主张不一等矛盾之处。为此,应当借本次《刑事诉讼法》修改之机,以系统地目光审视我国既有的电子取证规则,尤其是要从国际的、涉外的角度考虑电子取证中的涉外因素。比如,结合我国实际需求,构建层级化的跨境管辖权,即区分立法管辖和执法管辖,并在执法管辖的框架下合理设置取证措施的连接点。同时,还要通过向刑事诉讼法注入涉外理念思维,基于是否涉外的标准对传统侦查措施进行体系整合和类型细分,以此保证国内做法和国际主张的一致。就此而言,或可参照前述《公约》的相关规定,根据取证措施对主权的干预程度的强弱来对取证措施的域外效力或者说管辖范围进行限制。然后,通过补充、完善或创新国际警务合作机制、向服务提供商实施跨境调取数据的机制以及侦查机关特定的直接跨境侦查措施等配套机制,在既有的国际司法合作框架之外畅通打击网络犯罪的国际合作,以实现在尊重国

⁴² See Gianpiero Greco, Nicola Montinaro, "The Phenomenon of Cybercrime: From the Transnational Connotation to the Need for Globalization of Justice", *European Journal of Social Science Studies*, 6(1)(2021): 1.

⁴³ 裴炜:《涉外法治视野下刑事诉讼的数字化进路》,《中国刑事法杂志》2024年第2期,第124-142页。

国家主权的前提下，最大限度地提升跨境取证的效能。

（二）引入数据分类分级制度

我国电子取证措施体系并未区分电子数据的种类，亦未合理划定取证措施的干预程度，这是目前与《公约》规则最为基础的差异之一，也集中反映出我国国内法上的不一致性。尽管我国电子取证的相关规则多涉及电子数据的类型，例如《电子数据规定》将“电子数据”划分为网络平台公开发布的信息、通讯过程中形成的通信信息、包含用户注册信息在内的用户活动数据、以及承载一定内容信息的电子文件等四类，但这种类型化仅仅是对司法实践的个案经验的总结，与特定类型电子数据承载的公民权益并无关联，进而也就无法用以指导取证规则的分层和比例化设置。⁴⁴对此，我国后续立法需要以数据及其所承载的相应权利内容为基点，按照取证措施对基本权利的干预程度，即以比例原则为线索对二者进行链接，构建电子取证措施的框架。这至少需要从以下两个方面予以考量。

首先，参考相关数字法的分类标准充实刑事诉讼法的数据分级方案。我国《民法典》同时规定了隐私权和个人信息，《个人信息保护法》则区分了一般个人信息和敏感个人信息，这些分类方式都反映出不同数据所承载的权益差异，并构成了采取不同处理规则的制度基础。以此为鉴，刑事司法机关作为典型的国家机关，其对于数据的处理也需要遵循相关数字法中关于国家机关处理数据和公民个人信息的规定。据此，刑事诉讼法可以参考《数据安全法》《个人信息保护法》等法律的思路，根据数据的重要性和敏感度的差异对其进行分级，将其作为刑事电子数据分级的依据，以合理配置侦查措施。

其次，建立起我国数据分类分级体系与《公约》体系的对接，参考《公约》的分类对目标数据进行类型化处理。如前所述，《公约》采用了注册人数据、流量数据和内容数据的分类方式，我国后续电子取证规则的调整需要考虑如何将目前数字法中的信息分类方式对应到《公约》的数据分类体系中去，从而明确取证措施的类型和适用范围。其中，注册人信息一般对应至一般个人信息，而流量信息和内容信息因牵涉到通信自由和通信秘密这一宪法权利，有必要纳入到敏感个人信息的范畴中去。

（三）建立数据分类分级与侦查措施的比例关系

数据分类分级并非目的本身，其功能在于合比例地限定取证措施，这也是《公约》所秉持的基本思路。对此，在引入数据分类分级的基础上，我国电子取证措施制度体系需要进一步做如下方面的调整。

首先，为承载不同权益的数据类型分配相应强度的侦查措施，明确取证行为

⁴⁴ 谢登科：《电子数据侦查取证措施法治化与〈刑事诉讼法〉再修改》，《法治研究》2024年第5期，第90-105页。

的合理界限以回应刑事诉讼程序对公民基本权利的保障要求。目前，我国的电子取证措施一部分是沿用了传统侦查措施，如调取、技术侦查；一部分是改造了传统侦查措施，如冻结和勘验；还有一部分是特别针对电子证据而创设的措施，如现场提取和网络在线提取。但由于缺乏数据分类，这些强度不一的措施均能适用于任意数据。同时，由于缺乏上位法的统一规定，各项取证措施之间的位阶关系也并不清晰。对此，有必要参考《公约》的规定以目标数据为基点，一方面为不同层级的数据匹配强制性程度不同的取证措施，后者尤其需要考量措施的适用对象、范围、方式等因素；另一方面要尽量避免不同取证措施之间的附带关系，特别是避免将针对数据载体的措施天然拓展至数据之上。

其次，需要针对各项类似措施的具体差异进行细化调整。从比例原则的角度出发，结合《公约》的相关规定，这种细节化的调整需要考虑如下因素。第一是区分取证措施的相对人，考虑到一般个人与网络信息业者所掌握或控制的数据在体量、类型、范围以及所承载的权益方面的重大差异，相关措施有必要对二者予以区分，对于后者的取证措施的设置需要有更为清晰的适用条件，并且避免概括性的数据调取。第二是区分取证措施的客体状态，相较于针对静态数据的取证措施，立法应当对动态数据的取证措施设置更为严格的程序限制。第三是区分数据的跨境性与数据控制者的跨境性，以数据控制者模式来主导后续的取证措施调整。第四是明确取证协助方保密义务的适用情形、条件和具体程序，《公约》的相关规则是建立在个人信息处理告知义务的基础之上，后者在我国《个人信息保护法》中也有相应规定，因此无论是与《公约》还是我国国内法的衔接角度出发，都有必要针对告知义务的例外情形进行相应规定。

（四） 协同网络犯罪治理中的公私关系

《公约》在制定过程中充分考虑到了公私合作的必要性，并特别在程序措施部分强调要考虑“对第三方权利、责任和合法利益的影响”，这一要求在具体的取证程序中有所体现。在我国，向网络信息业者调取数据也是网络空间犯罪追诉的重要侦查措施，包括《网络安全法》第 28 条、《数据安全法》第 35 条、《反电信网络诈骗法》第 26 条等均提及了网络服务提供者对犯罪侦查的协助配合义务；同时近年来，网络平台等主体在违法犯罪防治方面的主体责任不断强化，例如“两高一部”《关于依法惩治网络暴力违法犯罪的指导意见》第 20 条强调要“夯实网络信息服务提供者的主体责任”，《反电信网络诈骗法》直接要求网络服务提供者承担电信网络诈骗风险防控的责任。

上述法律义务的设置是网络犯罪治理的现实需求，但也可能由此引发不同法律制度和体系下的义务冲突问题。对此，我国电子取证规则需要将《公约》对于第三方主体权益保障的思路纳入进来，着重从以下方面协同网络犯罪的预防和打击。第一是在调取措施上，在确立数据控制者模式的同时，明确该模式所适用的

数据类型和条件，避免出现概括性的、不加区分的数据调取。第二是在实施与《公约》的搜查、扣押措施类似的勘验、提取、封存措施时，明确第三方主体的具体协助义务内容、事项和范围。第三是在涉及跨境取证的场景中，采用数据控制者模式的同时，需要考虑到因目标数据位于境外，或者网络服务提供者的实体位于境外时，可能引发的与其他国家设置的数据保护义务冲突问题，一方面为协助犯罪追诉的第三方主体提供必要的申诉和救济途径，另一方面针对我国网络服务提供者等第三方主体根据《公约》要求协助其他缔约国执法机关的活动，提前化解其可能引发的国内法上的法律责任。

六、 结论

经过多年的谈判，《公约》所凝聚的是各国主管机关以及多利益攸关方的共识，形成了一套具有普适性的规则体系，为我国与其他国家开展打击网络犯罪国际合作提供了制度框架，能够促进各国之间更高效的司法协助。同时，《公约》表现出的实体法和程序法分离的倾向，让各国摆脱实体罪名的限制和分歧，仅通过程序措施即可实现对触网犯罪的打击，极大限度的拓宽了国际合作的范围。正是在此意义之上，程序措施对于打击网络犯罪而言日益重要，一项得到共同认可的程序措施将是未来开展国际合作的捷径。因此，我们需要加强《公约》规则和我国电子取证规则的衔接，以更充分地利用这一工具满足我国打击网络犯罪的需求。

与此同时，我们还应认识到，《公约》的谈判并非一劳永逸之事，在下一轮谈判启动之前，我们有必要结合我国打击网络犯罪的现实需要，一方面，通过开展双边或多边会议等渠道加强与其他国家的经验交流与合作，进一步落实《公约》规则。另一方面，则是要思考应如何改善既有规则并引导《公约》规定向我国诉求看齐。同时，还可以发挥我国的技术领先优势，推进对发展中国家的技术援助和能力建设等项目和工程，提升其网络犯罪打击能力，消灭网络犯罪的避风港。易言之，作为《公约》的倡议者和支持者，我们需要在以身作则接纳《公约》规则之外，积极推动《公约》规则在各国的落实。同时，还要及时总结我国打击网络犯罪的经验，并充分利用特委会这一谈判平台将其上升为国际规则，通过这样一个双向而非单向的规则衔接，以掌握国际规则制定的话语权。

可以预见，与此前通过的《联合国反腐败公约》和《联合国打击跨国有组织犯罪公约》等相类似，《公约》与我国规则的衔接将会是一项长期而持续的任务。这不仅需要深入研究数字法、刑事法及《公约》本身，还需要在三者之间不断穿梭、综合协调，以解决我国法律规则内部以及我国法律与《公约》规则之间可能存在的 inconsistency 问题，从而更有效地推动国际间打击网络犯罪的合作。

美国人工智能法律政策下的美式“安全观”及我对策建议

——基于 DeepSeek 时代全球产业格局

程乐¹, 龚煊¹, 郑英龙²

(1. 浙江大学 光华法学院, 浙江 杭州 310008; 2. 浙江工商大学 法学院, 浙江 杭州 310018)

摘要: 在美国占据先发优势的全球人工智能行业体系中, 以DeepSeek-R1为代表的中国大模型正对美方主导的失衡产业格局发起冲击。国家安全法学视角下, 美方近年为维护其人工智能领域支配性地位与绝对话语权持续渲染所谓“中国威胁”, 运用“泛安全化”话语叙事在非传统领域打造所谓美式“安全观”, 意图通过芯片封锁、算力垄断、投资限制等措施遏制中方人工智能之崛起。2025年特朗普政府掌权后, 其为确保美方领先地位所采纳的“放任型”技术发展政策更导向美国人工智能产业监管的根本缺位, 彻底倾覆技术发展与技术安全之间的理性平衡, 推动由无差别领域“泛安全化”向核心领域“逆安全化”的畸化转变。为回应美方不当打压, 中方应在总体国家安全观下在科技、市场与政策等层面作出即时回应, 抑制美方举措对我人工智能产业乃至全人类普遍安全制造的实质威胁。

关键词: 人工智能; 法律政策; 美式“安全观”; 泛安全化; 国家安全法学

一、问题的提出

近几年, 全球科技产业正在经历人工智能引发的激荡与变革。美国科技公司 OpenAI 于 2022 年末发布的大型语言模型 ChatGPT 如若一石入水, 在极短周期推动人工智能及其关联技术领域的波涛巨浪。当前, 生成式人工智能等主流大模型的基准性能与多模态能力稳步提升, 内容生成、交互问答、多语种翻译等任务处理能力持续增强;¹强化学习范式下高速迭代的推理模型在强逻辑领域表现突出, 数学奥数、竞争性编程、物化生竞赛等问题已然超越人类专家水平; 世界模型、智能代理、具身智能等关联技术跃进亦在进一步强化智能模型对物质世界的改造能力, 推动人工智能由“推理者”向“代理者”阶段的能级跨越。此背景下, 一方面, 趋向成熟的人工智能产品服务在大量非实体场域展现良好的通用性及迁移

¹ 程乐.生成式人工智能的法律规制——以ChatGPT为视角[J].政法论丛,2023,(04):69-80.

效果，并与金融、法律、医疗、交通、公共决策等核心社会领域深度嵌合；另一方面，现代化人工智能的杰出效能正密切赋能工业、制造业等传统产业的规模化转型与智能升级，乃至推动生物技术、量子计算、类脑芯片等尖端科研领域的长足进展。作为未来智能化产业的“水电煤”与“通用枢纽”，人工智能技术已日渐成为引领第四次工业革命的核心驱动力，在跃升成为全球科技战略竞争要点与制高点的基础上，同步带动各国法律政策的关注与变革。

在人工智能的演进洪流中，美国不仅把握 Google、Microsoft、NVIDIA 等传统科技巨头对技术资本的长期积累，亦拥有 OpenAI、Anthropic 等专攻前沿算法迭新的龙头初创企业，以硅谷为中心扩张并把握绝对的先发优势与支配性竞争力。相较之下，英国、德国、法国、日本等老牌强国均未展现人工智能领域的集群优势，Mistral 等迭出产品亦难呈现匹配美企的同期国际影响力。从科研基础、数据体量、资源储集、资本驱动、产业生态等核心角度审视，中国是美国在人工智能终局竞赛中最具竞争潜力的对手。此背景下，美国近年频繁运用“泛安全化”政治话语构建所谓美式“安全观”，持续对华开展高端芯片、算力支持、技术出口、国际标准制定等环节的全方位围堵，以致较长时间我国通义千问、文心一言等商用模型难以追及美国同类产品的先进水准。同时，美国亦加速累积其高端算力、前沿数据等前有优势，人工智能“美方吃肉，别国喝汤”的局面一时似成定局。然而，2025 年 1 月中国本土模型 DeepSeek-R1 的横空出世遽然打破美国“算力为王”的单方支配局面。²当前，中美两国人工智能双强对峙态势正愈加明朗。恰逢特朗普政府上台的变局节点，其所采纳的“放任型”人工智能监管政策正全面深化“逆安全化”倾向下的群生风险，这一“技术战场”的“修昔底德陷阱”似已无可避免。³在美式“安全观”内涵与外延的嬗变之下，国家安全法学作为国家安全学与法学交叉融合下的新兴学术焦点，⁴为美国人工智能法律政策的风险辨明及我对策形成提供方法论视角。因而，须重点回溯美国人工智能法律政策下美式“安全观”的形成背景、历史逻辑与本质特征，以期形成兼具前瞻性与可操作性的多层级应对策略。

二、 美国人工智能产业格局及法律政策构架

基于计算机科学领域的长时积淀，美国迅速取得预训练大模型研发应用的先发优势，并诞下 ChatGPT、Sora、o1 等“明星产品”。基于“技术传播困境”下美企

² 王奔,方兴东,钟祥铭.DeepSeek效应: 科技创新的技术—资本—政治(TCP)范式与路径选择[J/OL].青年记者,1-10[2025-03-09].

³ “修昔底德陷阱”(Thucydides Trap)的概念由美国政治学家格雷厄姆·艾利森(Graham Allison)提出以隐喻中美关系,指当一个崛起中的新兴大国挑战一个已经主导国际秩序的“守成国”时,产生的结构性冲突可能最终引发战争。这一理论援引自古希腊历史学家修昔底德在《伯罗奔尼撒战争史》中的经典总结:“雅典的崛起与斯巴达对雅典的恐惧,使得战争变得不可避免。”See Allison G. *The Thucydides trap: are the US and China headed for war?*[J]. The Atlantic, 2015, 24(9): 2015.

⁴ 王林.新文科背景下的国家安全法学专业建设与人才培养研究[J].情报杂志,2021,40(10):186-192+207.

模型性能领先差距之缩减，2025 年特朗普政府上台后美国人工智能法律政策或展现其“放任型”演化倾向。

（一）美国人工智能产业的发展历程

二十世纪以来，美国已于深度网络、对抗式学习等人工智能基础理论积累深厚实力。自 Geoffrey Hinton 等学者于 1986 年提出人工神经网络（ANN）概念以来，美国主导的计算机科学界持续攻克图像识别、机器学习、语音合成、自然语言处理、卷积神经网络等理论难关。2016 年，Google 旗下人工智能实验室 DeepMind 在深层神经网络（DNN）架构上结合强化学习（RL）与深度学习（DL）策略，成功研发里程碑式算法模型 AlphaGO 并击败围棋世界冠军。伴随数据体量、算力水平、并行处理技术的提升，美国技术界主导将预训练（Pre-Training）与微调（Fine-Tuning）相结合的人工智能训练范式用作主流，并催生 ChatGPT 等海量参数堆叠下的预训练深度神经网络模型，即大型语言模型（Large Language Model，简称大模型或 LLM）的重要突破。伴随 2022 年后各类预训练大模型的涌现式进展，人工智能复杂推理、多模态生成、自然语言理解等关键性能得到实质提升，但互联网可用训练数据亦临近枯竭。此背景下，OpenAI 以推理模型 o1 实现创新，推动自我对局强化（SPRL）与思维链（CoT）自动生成等后训练阶段（Post-Traning）的强化学习策略成为主流。

当前，美国已形成以 OpenAI、Anthropic 等人工智能企业为前端，Microsoft、NVIDIA 等老牌科技巨头为后端的核心产业集群，其中心依旧集中于加州北部的硅谷（Silicon Valley）。以 OpenAI 为代表，其旗舰模型 ChatGPT 作为大模型领域的“头雁”始终展现杰出竞争力，最新型号 GPT-4.5（Orion）强调幻觉减少、超高“情商”等独有优势；o1、o3 等开创性“o 系”推理模型在强逻辑领域仍有不凡表现；Sora、Voice Engine 等多模态模型亦彰显视频、语音生成等方面的创新潜力。此外，Anthropic、Google、xAI 等美方企业各自旗下的 Claude、Gemini、Grok 等尖端模型亦均展现卓越的国际竞争力与个性化优势，人工智能多维基准测试下的“性能霸主”时常易位。同时，全栈计算领跑者 NVIDIA 可为人工智能模型训练及运行提供高质量 CPU、DPU、GPU 算力供给，旗下 B200 处理器与消费级 GPU RTX 5090 等产品已取得全球范围的断层领先。此外，Microsoft、Intel、Apple、Oracle 等美国老牌科技巨头亦通过密集的资金注入、技术合作、基础设施支持等方式为前端人工智能企业提供关键支持并推进生态联动，进一步巩固美国领先地位。

（二）美国人工智能法律政策的变革进程

美国对人工智能及其关联技术的法律政策扶持可追溯至《莫雷尔法案》（Morrill Act）、《小企业创新研究计划》（SBIR）与《拜杜法案》（Bayh-Dole Act）等早期科研教育立法与创新激励机制。2010 年后，《振兴美国制造业和创新法》

(RAMI)与《国家量子倡议法案》(National Quantum Initiative Act)等主要立法亦分别推动数字制造与量子计算等前沿领域的协同进步。伴随人工智能技术进展,奥巴马政府于2016年发布《国家人工智能研究与发展战略规划》(The National AI R&D Strategic Plan,于2019年和2023年两度更新),⁵从联邦层面为人工智能发展提供研究经费、人才培养、技术标准制定等多方位支持策略。其后,特朗普(Donald Trump)在其第一任总统任期中签署第13859号行政令《维持美国在人工智能领域的领导地位》与13960号行政令《促进联邦政府内可信赖人工智能的使用》,要求进一步开放人工智能研发数据资源并制定适度的应用监管指南,⁶以“安全、合法、透明”等要素为人工智能应用确立更具包容性与一致性的合规指引,⁷强化美方领跑地位。

2021年至2025年的拜登(Joe Biden)任期被普遍视作美国现代化人工智能的黄金发展期。拜登政府上台后,率先推动《国家人工智能倡议法》(National AI Initiative Act,2021年1月由特朗普签署生效)的贯彻落实,确立国家人工智能倡议咨询委员会等机构,统筹学术界、产业界与政府部门的跨部门协同联动。2022年,拜登签署《芯片与科学法案》(CHIPS and Science Act),意图加大投入以游说台积电等全球尖端半导体公司在美本土设厂,振兴美国半导体制造业并减弱对亚洲的依赖。⁸2023年,拜登签署第14110号行政令《关于安全、可靠和可信地开发和使用人工智能》,指示国家标准与技术研究院(NIST)确立生成式人工智能安全测试标准,并对个人隐私保护、算法歧视预防等系列问题进行说明,⁹被诸多学者视为美国人工智能治理最全面的政策举措之一。期间,白宫科技政策办公室、NIST等机构亦相继出台《人工智能权利法案蓝图》(Blueprint for an AI Bill of Rights)、《人工智能风险管理框架》(AI Risk Management Framework)等规范,进一步提出覆盖消费者隐私、透明可问责、算法偏见预防的规范原则,并为企业提供人工智能潜在风险的分层管理步骤。¹⁰

当前,全球人工智能产业竞争渐入白热化。2025年1月,特朗普政府胜选

⁵ National Science and Technology Council, "The National Artificial Intelligence Research and Development Strategic Plan," October 2016, https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf

⁶ The White House, "Executive Order on Maintaining American Leadership in Artificial Intelligence," February 2019, <https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/>

⁷ The White House, "Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," December 2020, <https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/>

⁸ Luo Y, Van Assche A. *The rise of techno-geopolitical uncertainty: Implications of the United States CHIPS and Science Act*[J]. *Journal of international business studies*, 2023: 1.

⁹ 《关于安全、可靠和可信地开发和使用人工智能》(Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)由拜登总统于2023年10月30日签署,旨在确保美国人工智能的开发与应用符合安全、可靠、可信的标准。该行政令已于2025年1月23日由现任美国总统特朗普废除,认为第14110号行政命令对美国人工智能创新制造障碍。See Biden J R. *Executive order on the safe, secure, and trustworthy development and use of artificial intelligence*[J]. 2023.

¹⁰ The White House, "Blueprint for an AI Bill of Rights," October 2022, <https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/>

后再度上台，随即于当月 23 日签署第 14179 号行政令《消除影响美国人工智能领导地位的阻碍》，宣告撤销拜登政府时期包括 14110 号行政令在内的大量审慎监管措施。特朗普认为，拜登此前用于确保人工智能安全性与可靠性的举措过于繁琐，可能阻碍美国人工智能创新。行政令要求在 180 天内制定全新的人工智能行动计划以增强美国的全球领导地位，并审查所有与新政策相左的前有政策、指令与法规。¹¹也即，特朗普新政下的美国科技企业将无需遵守此前的安全测试标准，未来甚至毋需遵守反垄断要求或是对人工智能模型与用户间的对话内容进行审查。这一迹象显示，美国人工智能法律政策领域正出现从“有条件监管”向“放任型监管”的重大转向，这一动向的本质逻辑为“减少监管以促进创新”。同时，以“星际之门”项目为代表，特朗普政府亦计划启动面向人工智能领域的“曼哈顿计划”，¹²通过进一步聚合各方资源以强化“AI First”在 MAGA（Make America Great Again）议程的核心地位。

（三）美国人工智能关联政策的对华遏制倾向

二十一世纪以来中国逐渐成为美国在科技领域的主要竞争者，中美对弈局势趋于紧张。科技层面，美国国土安全部与白宫于 2018 年联合发布《国家网络战略》，明确将中国列为对美国国家网络空间安全的“主要威胁”。¹³2021 年，美国商务部进一步以“外国对手”名义将中国、俄罗斯等国列入信息通信技术服务的高度防范名单。¹⁴此后，美方借用“国家安全”“贸易失衡”等事由，以《芯片与科学法案》为代表引入一系列对华科技遏制政策，运用先进芯片等关键技术的出口管制、对华投资限制等策略限制中国科技领域进展。同时，以特朗普政府第 13920 号行政令与第 13953 号行政令为代表，美方将中国供应商介入大容量电力系统定义为“国家安全威胁”，¹⁵禁止中国企业向美国国防重点设施供应关键电力设备，并在稀土、关键矿产等战略资源领域加强管控，¹⁶意图从源头上削弱中国的原材料优

¹¹ The White House, “*Removing Barriers to American Leadership in Artificial Intelligence*,” January 2025, <https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/>

¹² “曼哈顿计划”是美国二战期间研发人类首枚核武器的军事计划，汇集彼时西方国家最优秀的科学家与海量资源，被时任美国总统罗斯福赋予“高于一切行动的特别优先权”。人工智能领域美国汇聚大量资源用以重点发展的代表性项目是“星际之门”（Stargate）计划，由现任美国总统特朗普牵头，并经OpenAI、Oracle（甲骨文）、SoftBank（软银集团）等巨头企业于 2025 年 1 月 21 日联合发起，旨在通过高达 5000 亿美元的巨额投资在美国建设超大型人工智能基础设施，进一步巩固美国在全球科技领域的领先地位，首个数据中心计划于德克萨斯州开工并为OpenAI提供支持。See Blue C, Calhoun S, Shiffer C, et al. *STARGATE: An Undergraduate Experimental Electric Propulsion Student Research Project*[J]. 2025.

¹³ The White House, “*National Cyber Security*,” September 2018, <https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf>

¹⁴ U.S. Department of Commerce, “*Securing the Information and Communications Technology and Services Supply Chain*,” January 2021, <https://www.federalregister.gov/documents/2021/01/19/2021-01234/securing-the-information-and-communications-technology-and-services-supply-chain>.

¹⁵ The White House, “*Executive Order on Securing the United States Bulk-Power System*,” May 2020, <https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-securing-united-states-bulk-power-system/>.

¹⁶ The White House, “*FACT SHEET: Securing a Made in America Supply Chain for Critical Minerals*,” February 2022, <https://www.whitehouse.gov/briefing-room/statements-releases/2022/02/22/fact-sheet-securing-a-made-in-america-supply-chain-for-critical-minerals/>.

势。

市场层面，美国将“经济安全”定义为国家安全的重要组成，并由此对华开展一系列关税与投资限制。特朗普政府以中国在技术转让、知识产权等方面的“不公平贸易行为”为由提高对华进口关税，并收紧中方在美投资审查以阻断中企在美科技产业链布局；拜登政府意图通过“印太经济框架”整合地区伙伴，排挤中国在亚太贸易格局的影响力；美国联邦通信委员会（FCC）则连续将华为、中兴列入国家安全威胁名单，以电信设备安全风险为由禁止其进入美国市场，并对华为高管孟晚舟以“威胁国家金融体系安全”等罪名开展审查与引渡。¹⁷当前，美方已在法律政策层面对华形成科技禁令与经济封锁的合围态势，联动贯穿核心技术、供应链与市场准入等要点，意在为美方人工智能绝对优势之维持提供制度化后盾。

三、 国家安全法学论域下美式“安全观”的形成及其特征

长久以来，国家安全概念及其适用范围始终是法学与法治研究的重点议题。¹⁸复杂多变的国际安全形势下，我国已逐渐形成以《宪法》为统领，以《国家安全法》为基础，以《国防法》《反间谍法》《反恐怖主义法》《情报法》《保守国家秘密法》《反分裂国家法》等国家安全专门立法及刑法、民法、行政法等部门法中关联条款为支撑的国家安全法律体系。¹⁹此背景下，国家安全法学成为汇融国家安全学与法学这两门一级学科的重点交叉界域，²⁰为明辨美式“安全观”的形成背景、本质逻辑与典型特征提供体系化视角。

（一） 美式“安全观”的时代背景

伴随前沿技术议题在非传统安全领域的重要性升级，美国频繁运用“泛安全化”策略包融各方议题以对抗其全球科技支配权之衰落。为抢夺人工智能这块时下最热的“增量蛋糕”，美国不惜牺牲理据监管下的安全保障，麇集诸升级手段以对抗 DeepSeek 等中方力量的冲击。

1. 冷战后非传统安全议题的地位抬升

冷战时期，美苏为首的东西方对抗阵营所聚焦的安全诉求集中于武装冲突、军备竞赛等传统军事领域。然而，“仅将安全锁定在军事领域传递了关于现实严重错误的形象”。²¹冷战结束使世界权力结构经历重塑，联合国安理会、北约等安全机制之崛起打破传统安全概念的边界，气候、环境、居民健康等非传统领域的次生安全问题逐渐占据国家安全与国际关系议题的核心。国家安全法学的关联领

¹⁷ U.S. Department of Justice, “Huawei CFO Wanzhou Meng Admits to Misleading Global Financial Institution,” September 2021, <https://www.justice.gov/opa/pr/huawei-cfo-wanzhou-mengadmits-misleading-global-financial-institution>.

¹⁸ 吴庆荣. 法律上国家安全概念探析[J]. 中国法学, 2006, (04): 62-68.

¹⁹ 舒洪水, 马瑜. 多学科交融的国家安全法学专业建设探究[J]. 法学教育研究, 2022, 36(01): 54-69.

²⁰ 廉睿. 国家安全学一级学科下的国家安全法学的构建[J]. 情报杂志, 2022, 41(12): 76-80+163.

²¹ Ullman R. *Redefining security*[M]//Security Studies. Routledge, 2011, pp. 11-17.

域中，以 Barry Buzan、Ole Waever 等学者为代表的哥本哈根学派（Copenhagen School）将“安全化理论”（Securitization Theory）视作安全学研究的支柱概念，提出安全是“超越一切政治规则和政治结构的途径，实际是一种所有政治之上的特殊政治”。因而，“安全化可被视为一种更为激进的政治化描述”，²²安全化理论则重点探讨如何通过话语将某一议题提升为安全问题，从而赋予其特殊的紧迫性与重要性。

伴随安全化进程，非传统安全议题及其衍生内容得到持续扩充，发展中国家更关注经验基础上的制度安全，发达国家则可能强调规范层面之上的普遍原则，并引发个体人权理念、人道主义干预等争论。此过程中，除国家行为体作为安全概念构建中的决定性力量，政府间国际组织、非政府组织、慈善团体等非国家行为体亦日渐融入全球安全治理框架并密切参与安全化话语构建。就美国而言，其对反恐问题关注的过度扩张便属于非传统议题安全化的典型案例。历经 2001 年“9·11”恐怖袭击事件，恐怖主义袭击制造的实际人员伤亡数量虽仍难比及军事战争、交通事故等传统问题，但却在二十一世纪初期经由美国政治精英的话语建构成为美国国家安全论述的重中之要，以至任何威胁国家稳定的非军事议题均可被囊括进反恐议题这一安全话语的“头号主线”中。此外，伴随中国、日本等新兴竞争体经济水平的快速提升，美国在经济领域的安全化表述亦同步强化；结合互联网等数字技术的发展与普及，美国在网络安全、数据安全等数字技术议题的安全化诉求亦呈现外延持续扩展的趋势。

2. 二十一世纪美国霸权衰落下的“泛安全化”趋向

美国在冷战结束后取得全球领导地位，后通过关贸总协定（GATT）、国际货币基金组织（IMF）与世界银行（World Bank）等全球化机制主导国际经济秩序，逐步将自由贸易与经济全球化确定为自身行为基调，并将追求市场开放战略作为其主要目标。²³但二十一世纪以来，以中国等金砖国家为代表的发展中国家高速崛起，伴随美国内部政治分裂势力加剧掣肘等问题，美国霸主地位频受冲击。一般认为，2001 年的“9·11”事件引发美国霸权的实质性衰落，这一趋势在 2008 年全球金融危机达到顶峰。此背景下，美国政治精英与权力主导者进一步选择扩张安全化话语的边界，强化民粹主义并进行安全威胁的反复建构，引发与其此前相反的逆全球化行为逻辑与日趋严重的“泛安全化”倾向。

“安全化”在程度上走向极端即可能带来“泛安全化”（Over-Securitization）。作为国家安全法治研究的重要概念，“泛安全化”一般被认为是对众多非传统安全议题进行安全构建，进而使安全的概念、范畴和维护安全的手段极具扩张，导致国

²² [英]巴里·布赞、[丹]奥利·维夫、[丹]迪·怀尔德：《新安全论》，朱宁译，浙江人民出版社 2003 年版。

²³ Falke A. *Wasting Hegemony in the Global Trading System: Trump's Trade Policy*[J]. *Hegemonic Transition: Global Economic and Security Orders in the Age of Trump*, 2021, pp. 87-103.

家安全议程变得无所不包。²⁴在美国的“泛安全化”进程中，对“中国威胁论”的长时聚焦成为其国会两党及诸政府部门无数分歧中的稳定共识。1992年，Ross H. Munro 在《正在觉醒的巨龙：亚洲真正的威胁来自中国》一文便明确提出“中国龙一旦觉醒，世界就会震动”，认为中美之间的军事冲突不可避免。²⁵通过将中国视作持续释放威胁的“他者”，美国将对华战略中的国家安全要素逐渐扩张至经济、科技、文化等多维领域，并将日本、英国、澳大利亚等国纳入“辅美遏华”“联美压华”的战略架构，炮制全球层面的“中国威胁论”以拉拢国际社会孤立中国。

3. 人工智能成为全球科技竞争“主战场”所生发的时代需求

传统安全位阶中，军事与经济是国家安全议题聚焦的核心内容，“霸权同心圆”外围的其它内容则相对不易引起过量关注。²⁶然而，以人工智能为代表的尖端技术正日渐成为现代社会智能化产业的“生产力中枢”。因而，围绕高端芯片、算力支撑、前沿数据等要素的人工智能生态日渐被美国定位为主导世界技术版图的关键资产，美方安全话语的核心叙事亦相应融入“人工智能安全政治化”这一表述。为压制中国等主要竞争对手的快速追赶之势，美国迅速将人工智能安全议题提升至国家安全议程，进而将技术封锁、投资限制等遏制性措施包装为“保护国家安全”的正当举措。本质而言，人工智能领域的美式“安全观”亦折射美国对于其单方视角下存量利益与增量利益分配不均衡的介怀。科技领域作为先前美国收割大量既有存量利益的“大蛋糕”，美国对全球科技市场规则的掌控欲因而持续强化。而以大模型服务应用为代表的人工智能领域作为近年最为热门的“增量蛋糕”，正被美国赋予更强的排他属性以期收割垄断利益。可见，美国当前对于人工智能领域的技术权力分配模式仍未达至满意状态，并抱持一定不安全感。因此，美国正在尝试建立人工智能霸权地位并意图长期保持此种支配性状态，目的亦是在国际体系面临增量利益重新分配时主导重塑人工智能领域的基本秩序，并使这一秩序服务于美国的多方利益，进一步巩固其全球霸主地位。

4. DeepSeek 等中方技术入局之影响

近几十年现代科技发展进程中，美国属于全球科技发展的领先者和开拓者，而中国则相对处在追赶者的位置。整体而言，处在开拓者位置的美方需要持续创新，并投入大量成本用于试错。相对地，追赶者则可能通过学习和模仿已被验证的最优技术路径，以更低成本与更快速度获得同样或类似的技术。人工智能领域的技术发展亦大体遵循这一模式，中国虽起步较晚，此前亦确基于此种“技术传播困境”与“搭便车困境”获得技术增速的优势。²⁷然而，美国近年在高端芯片供应、

²⁴ 卢凌宇,崔磊.霸权衰落与“泛安全化”:基于美国的个案分析[J].国际安全研究,2024,42(03): 76-97+159.

²⁵ Ross H. Munro, “Awakening Dragon: The Real Danger in Asia is Coming From China,” Policy Review, No. 62, 1992, pp. 10-17.

²⁶ 李晓,李俊久.美国的霸权地位评估与新兴大国的应对[J].世界经济与政治,2014,(01):114-141+159-160.

²⁷ 两个关联概念一般用以霸权运行的内生难题。技术传播困境 (Technology Diffusion Dilemma) 指出霸权国往往是全球技术创新的中心,它的技术进步会带动全球生产力提升。但先进技术的扩散一旦失控,新兴国家就能快速追赶,对霸权的技术优势形成威胁,最终削弱霸权国的长期竞争力。搭便车困境 (Free-rider

先进数据资源等关键节点对华封锁以拉大差距。我国即便在核心运算芯片、众核处理器等领域取得自研进展，但通义千问、文心一言、豆包、Kimi 等国产商用模型确长期难以撼动美国同类产品的主导地位。然而，DeepSeek-R1 的问世真正扭转这一艰难形势，作为“国运级”产品对美方主导的人工智能全球格局制造巨大冲击。

DeepSeek-R1 于 2025 年春节期间发布，仅用 20 天便获得超 2000 万日活跃用户数，并在 157 个国家的应用市场中取得超越此前 Mistral 等所有非美产大模型的亮眼表现。R1 借鉴 o1 等现有推理模型的强化学习模式，但运用创设性的 GRPO 架构突破 NVIDIA 的 CUDA 限制，并对 o1 等旧有模型的 PPO 与 DPO 架构进行优化。其不仅在编程、数学奥数等强逻辑任务中追平甚至超越现有美国顶级模型的表现，亦推动推理模型高额训练成本的大幅降低并实现开源，²⁸当前，DeepSeek 这一总部位于杭州的人工智能初创企业已把握旗下基础模型 DeepSeek-V3 与旗舰模型 DeepSeek-R1 两大核心产品，并计划于 2025 年 3 月发布升级版模型 DeepSeek-R2，已成为中方技术打破美国单方垄断格局的典范。其一，DeepSeek 的成功映射美国当前人工智能发展路径正一定程度陷入窠臼，基础设施的断层强大反而易使其陷入纯粹叠加算力的惯性思维，实不利于前沿技术的创新探索。DeepSeek 问世后 NVIDIA 股市暴跌证明“算力为王”的垄断格局正在被打破，且单向度拉大差距的制式作法或将时刻遭遇反噬。其二，DeepSeek 引发的鲶鱼效应将进一步作用于全球人工智能产业，不但验明人工智能领域的“中国路径”同样具备突破性创新可能，还以其开源架构赋能全球大模型系统的本土开发与垂直挖掘，直接冲击美方对全球人工智能产业的制高点把控与封锁。其三，美方在感知中方威胁升级后或将 DeepSeek-R1 等中方先进模型纳入制裁范畴并进一步强化压制措施，中美双方在人工智能领域的竞争将进入更为复杂、多维且不可逆的长期拉锯。

（二）美式“安全观”的形成逻辑

为延续其持续衰退的科技霸权，美国将安全化话语构建用作其“霸权护持”的核心手段。美方从其持续变动的即时需求出发针对性塑造安全诉求，为其非理性政策措施赋予正当色彩，这一逻辑正在深刻重塑美国在全球人工智能竞争格局中的政策取向与干预模式。

1. 美国霸权衰退下“霸权护持”的核心手段

国家安全需求的视角下，人工智能领域的美式“安全观”肇始于美国综合实力

Dilemma) 指出霸权国提供的全球公共产品（如金融稳定、全球安全保障、全球贸易秩序），所有国家都能受益。但并非所有国家都愿意按比例承担成本，很多国家享受好处，却不愿承担责任，让霸权国“独自买单”。See Isabelle Grunberg, “Exploring the ‘Myth’ of Hegemonic Stability,” *International Organization*, Vol. 44, No. 4, 1990, p. 453.

²⁸ 方兴东,王奔,钟祥铭. DeepSeek时刻: 技术—传播—社会 (TCS) 框架与主流化鸿沟的跨越[J/OL]. 新疆师范大学学报(哲学社会科学版), 2025, (04): 1-11 [2025-03-10].

之衰退，其根本动因源于美方对其技术权力优势的无休止追求，并外化为对广泛技术议题的安全概念泛化。一方面，衰落是霸权的宿命。²⁹美国前期运用政策支持、科技创新、学术教育等禀赋优势取得全球范围科技领域的支配性地位，但当前 DeepSeek 对 OpenAI、Google 等美国科技巨头的冲击正在人工智能领域证明其基于先发优势所建立的垄断性“科技霸权”终将在科技、贸易、教育全球化的时代背景下趋于衰退。另一方面，霸权国亦不可能静待霸权体系的崩溃。人工智能作为攸关霸权国美国与崛起国中国在科技主权、国家安全、产业竞争力等核心利益的“主战场”，极大程度影响科技领域的“霸主更迭”抑或“霸权延续”。此背景下，美国身为霸权国必将竭力维持与崛起国之间的权力差距以延长其科技霸权寿命，而其用于维持自身与中国等国家在人工智能议题下权力差距的实践可被定义为其在科技领域的“霸权护持”行为(hegemonic maintenance)。在其科技霸权周期内，美国人工智能法律政策的主要战略目标便是将这一权力差距维持在自身所定义的“安全常数”之下。

然而，霸权护持的手段具有有限性，霸权收缩、离岸平衡、多边主义、选择性接触乃至军事手段等传统策略亦难在人工智能领域取得收效。为逆转其全球科技霸权及绝对话语权衰落所面临的结构性压力，美国选择持续延伸安全概念以推动关联议题的“泛安全化”，并逐渐将此种做法用作维系人工智能领域科技霸权地位及全球市场支配性地位的核心手段之一。事实上，“安全”不仅用于描述客观现象，亦被视作一种主体间观念性社会建构的结果。³⁰一方面，美国旨在运用话语策略将中国等对其人工智能霸权制造威胁的特定对象于语言层面描绘为一种“实在的威胁”，进而包装正当立场以对指涉对象采取针对性行动。另一方面，美国对于特定议题的推进往往需要经历漫长的提请、讨论与批准周期。因而，营造安全危机可以提高政策的合法性和施行效率，以施行在常规政治下难以接受的举措。双重需求下，但凡涉及人工智能模型开发、算力供应、全球标准制定、尖端技术科研等内容的议题，均可被塑造为事关国家安全、贸易失衡与世界秩序的红线领域，并将对美国霸权地位形成潜在冲击的力量描绘为“紧迫威胁”。畸化的美式“安全观”作用下，一旦美国媒体、智库与政府部门将某项外国技术或行动刻画为“重大安全危机”，便可有效触发社会的恐慌情绪与议程偏向，为美国采取高强度管制与不正当封锁措施提供必要性依据，继而强化霸权护持力度。

2. “需求感知-构建安全化诉求-泛安全化”的话语逻辑

传统意义的安全化进程一般分为问题描述、受众接受与诉诸紧急措施三个阶

²⁹ 诸多理论可证明这一观点，例如Terence Hopkins等学者将霸权分为“霸权上升”“霸权胜利”“霸权成熟”“霸权衰落”四个阶段；George Modelski亦提出“霸权的大循环”理论，指出霸权生命周期均将走向“权势消散”的最终阶段。See Terence K. Hopkins and Immanuel Wallerstein, “Cyclical Rhythms and Secular Trends of the Capitalist World-Economy,” *Review*, Vol. 3, No. 2, 1979, pp. 483-500; George Modelski, “The Long Cycle of Global Politics and the Nation-State,” *Comparative Studies in Society and History*, Vol. 2, No. 20, 1978, pp. 214-235.

³⁰ Buzan B, Wæver O, De Wilde J. *Security: A new framework for analysis*[M]. Lynne Rienner Publishers, 1998.

段,³¹“泛安全化”这一极端情态的表现则有所延伸。事实上,安全议程并非自然形成,而是源自政治精英、利益集团、媒体和学术共同体的积极互动与塑造过程。人工智能领域中,通过合法性与权力扩展,安全成为一种政治工具用以捕捉并转化美国在这一产业的各类需求与目标。首先,通过“需求感知”的前置程序,以技术界、政界为主导的多方力量得以明确当前本土人工智能发展的主要瓶颈与重点需求,并锚定抑制他国产业的关键环节。其次,美方可通过官方报告、国会听证、媒体宣扬等途径将其具体需求针对性灌输为国家安全的公共议程以奠定其正当基础,并在投资禁令、芯片断供、技术标准制定等模块层面对应塑造具体安全化诉求并凝聚内外部资源,以实践“围中助美”的现实立场。最后,通过构建“泛安全化”国家安全话语并施行相应法律政策,最大化霸权护持的实际效果以推动中美技术差距加速扩张。

(三) 美式“安全观”的本质特征

国家安全法学视野下,国家安全议题的理性扩充是时代发展的本质要求,国家与社会一体化的局面被打破后“权力”也逐渐出现多元化与社会化趋向。³²然而,美式“安全观”中“倒果为因”“自我放任”的两大去理性化特征正展现趋于明确的“利益中心”“集权中心”趋向,暴露美国人工智能法律政策在“国家安全”话语包装下重度畸化的本质特征。

1. “倒果为因”“利益导向”的“构建式安全观”

国家安全审查等以“安全”为名生发的法律政策措施,虽看似处在法律语境,本质却密切关联政治问题。³³当前,美国在人工智能领域所倡导的美式“安全观”表述往往并非出自对客观风险的真实关切,而是由政治话语与科技、经济、产业需求引领的叙事构建过程。哥本哈根学派将安全视作“主体间的建构与话语实践”与“动员社会资源达成高度政治化目标的手段”,认为当言语行为得到合理利用时,安全化手段可以修改背景内容并使其得到普遍接受。³⁴以此观之,美式“安全观”的政治表述实是对安全化理论的扭曲与滥用,深刻展现“因果倒置”“倒果为因”的强盗逻辑。也即,美方实时确立其强化自身科技实力与封锁他国科技发展的具体需求,预先构造“对方构成威胁”这一结“果”,再针对性挑选得以服务对应需求的具体动“因”纳入其安全化范畴。因此,当美国各方通过立法、行政命令与媒体扩散等话语建构途径反复诉诸“国家安全”支配的人工智能法律政策时,其真实动机并非消除现实化或潜在化的客观风险,而是为其技术壁垒与输出禁令塑造并包装外部的正当化色彩。

³¹ 叶晓红.哥本哈根学派安全化理论述评[J].社会主义研究,2015,(06):164-172.

³² 郭道晖.权力的多元化与社会化[J].法学研究,2001,(01):3-17.

³³ 王东光.国家安全审查:政治法律化与法律政治化[J].中外法学,2016,28(05):1289-1313.

³⁴ 朱宁.安全与非安全化——哥本哈根学派安全研究[J].世界经济与政治,2003,(10):21-26+4-5.

一般情况下，国家安全包含领土安全、国防安全、反间谍能力等主要内容。³⁵客观层面，国家安全概念的范畴扩张及其中心议题倾斜亦是时代发展的合理结果。然而，安全化程度走向极端时“泛安全化”议题的超期泛化与无限延伸必然彻底颠覆其合理性与正当性，而美国人工智能领域“极端安全诉求”反向塑造的法律政策正是这一失序逻辑下“利益导向”的畸化产物。究其本源，以正当色彩的美式“安全观”包装其非正当利益诉求或单纯对华敌视心态这一构建性逻辑，是由美国国内的民粹主义浪潮与各利益攸关群体联合推动，高度依赖各利益集团与政治精英的议题设定与话语操控。此前，美国人工智能行业在 Microsoft、Google 等科技与资本巨头的驱动下于全球范围高速扩张，且这些巨头往往又与美国政界存在深度利益绑定，故而美方政府以美式“安全观”下的法律政策为其竞争对手设限之时本质亦是服务其本土巨头与利益团体的博弈需求。面对中方 DeepSeek 等或将动摇其核心利益版图的竞争性因素，美方不惜大幅削减跨国合作与技术共享，通过对全球创新资源流动的刻意掣肘推进本国企业的市场垄断程度与资源收割空间。此过程中，美式“安全观”已不再隶属于风险识别的范畴，更是成为由美国政治与利益话语系统所共同建构的利益固化工具。

2. “集权导向”“自我放任”的“破坏型安全观”

美国的人工智能法律政策正展现明确的双重集权化倾向，即不但要求他国严格配合不尽合理的出口管制、技术封锁等强制措施，又对自身技术研发与不当市场行为抱持过度“放任”以实现产业进展加速。当前美国总和双重手段以强化其人工智能产业集权，相应法律政策所折射的美式“安全观”体现其对自身垄断性支配地位的“绝对安全意识”，即美方当前对于“相对不安全”的不接受状态实质是无法接受其对于技术领域的“相对不主导”。本质上，美国仍期待追求全球人工智能领域的绝对话语权，并成为全球人工智能产业的“绝对中枢”，亦即要求全球人工智能产业均无法绕过美国而顺利运转，从而实现前所未有的权力集中与利益收割。2025 年之前，这一权力集中的实现主要通过对中国等主要竞争国的封锁与打压。DeepSeek 问世与特朗普政府上台的变局节点后，原有手段已难继续拔高美方的人工智能霸权地位，特朗普政府便选择施行“放任型”法律政策牺牲技术安全、科技伦理等重要价值以换取科技进展加速，并以“星际之门”等项目为代表进一步集中各方资源。美方这一“自我放任”下的集权路径与现有“平衡技术发展与安全”等基本共识背道而驰，需要更为强势的话语构建策略以保美方在国际社会持续声称的正义性与合法性不至根本性削弱。

美方“自我放任”的人工智能集权路径背后展现其不惜牺牲技术安全等重要价值以追求全球人工智能领域垄断性支配地位的战略动机，要求获得“断层式技术优势”与“全球规则制定权”双重垄断。这一做法不仅严重损害积极的全球合作、

³⁵ 梁忠前.“国家安全”概念法理分析[J].江苏社会科学,1995,(04):77-82.

科技创新与良性竞争环境,科技领域的集权化倾向更将极大可能制造规模化的不可控技术风险。须重视的是,人工智能在技术领域的基础性、外溢性与支配性特征决定其对全球公共利益甚至人类普遍安全的影响远超传统科技议题,³⁶在关乎虚假内容、数据隐私、算法歧视等泛化法律风险的同时,甚至可能在自我强化、无监督训练下走向远超人类智能程度的“通用人工智能”与“超级人工智能”。³⁷高度自主化发展路径下迈入“代理者”阶段的人工智能可对真实世界制造巨大影响,其潜在破坏力不啻于对人类普遍安全制造严重挑战甚至毁灭性风险。特朗普政府未经论证便火速撤销拜登政府时期的审慎监管举措,甚至积极推动人工智能与军事、情报等领域的密切合作。美国为实现对自身“绝对集权”的偏执追求不惜将整体人类社会的普遍安全放至次要地位,使得“安全”一词沦为讽刺的片面修辞。

四、 美式“安全观”下人工智能“逆安全化陷阱”的负向影响

系统理论与现实主义理论已共同揭示霸权在维系自身相对优势时可能面临的结构性缺陷与多重消极影响,³⁸例如“泛安全化”可能引发的传统安全关切被低估、分析清晰度与适用性丧失、对不同安全问题的取舍困难化、经验分析与规范性主张的分歧扩张等问题。³⁹当前,美式“安全观”深化制造的“泛安全化陷阱”或将超越前有风险,对全球理性视角下的科技伦理与技术安全制造破坏性影响并推动人工智能领域的“逆安全化”趋势。

(一) 人工智能领域美式“安全观”的多重影响

技术与产业层面,美国人工智能法律政策“泛安全化”下尖端技术的“野蛮生长”极易导向科技领域“事故三角”的倾覆,形成人工智能领域的“泛安全化陷阱”。“泛安全化陷阱”可包含三层逻辑:一是国家行为体对绝对安全状态的无节制扩充,二是内部政治保守化与外部关系封闭化,三是资源错配后关键领域安全能力的不足。⁴⁰人工智能领域中,美方理据监管所需分配的资源严重缺位,具体生成人工智能“泛安全化陷阱”的两大类影响。一方面,美式“安全观”忽视技术安全与技术伦理的理据研讨,促使潜藏于技术底层的数据偏见、恶性竞争、安全审查缺失、自动决策失控等问题持续孕育,全球技术的良性发展格局难以铺开。另一方面,美国作为全球人工智能发展的领先国,如前所述,其本国科技巨头的人工智能模型亦可能在监管缺失与无监督训练的巨大不可控性中制造不可逆转的安全危机。以海因里希法作为参照,⁴¹任何人工智能严重事故的出现都意味着若干次小故障

³⁶ Yao Y, Duan J, Xu K, et al. *A survey on large language model (llm) security and privacy: The good, the bad, and the ugly*[J]. *High-Confidence Computing*, 2024: 100211.

³⁷ 程乐.“数字人本主义”视域下的通用人工智能规制鉴衡[J].政法论丛,2024,(03):3-20.

³⁸ Terry Boswell and Mike Sweat, “Hegemony, Long Waves, and Major Wars: A Time Series Analysis of Systemic Dynamics, 1496-1967,” *International Studies Quarterly*, Vol. 2, No. 35, 1991, pp. 125-132.

³⁹ Miller B. *The concept of security: Should it be Redefined?*[M]//Israel's National Security towards the 21st Century. Routledge, 2014, pp. 13-42.

⁴⁰ 张超,吴白乙.“泛安全化陷阱”及其跨越[J].国际展望,2022,14(02):19-35+149.

⁴¹ 海因里希法则 (Heinrich Pyramid), 又称事故三角理论, 由美国工业安全先驱赫伯特·威廉·海因里希于1931年提出, 该理论奠定了现代工业安全管理的基础。海因里希法则指出, 在工作场所, 每发生1起严

或微风险已在行业与监管部门的默许下累积与酝酿，并最终导向不可逆转的体系崩溃。当前，美方既不打算为人工智能行业负责任与可持续的发展框架负责，亦无意进行底层机制的风险防范，世界人工智能产业正在美式“安全观”的“过度溺爱”下迈向趋于失控的“事故三角”。

经济与政策层面，美国的“逆安全化”思维亦正对全球范围的可持续人工智能产业生态制造显著损害。博弈论中有一经典论断：“个人的理性选择会造成全局的非理性。”⁴²霸权国从本国需求出发反向塑造泛化“国家安全”这一看似符合自身利益的“理性行为”实则极易导致国际社会在技术、产业、经济、政策等层面的集体非理性。以 2025 年 2 月巴黎人工智能峰会为观察点，法国等国近期在人工智能治理政策风向上已对美国输出的“技术至上”与“竞争优先”思路产生追随迹象。美式“安全观”的全球化侵蚀下，全球人工智能产业为盲目追求基准性能飙升等功利化目标，或难在规范制定层面将技术安全置于技术发展之前。此种恶性竞逐虽或在短期内加速所谓“性能奇点”的到来，但终将在更宏观的界域制造巨大风险甚至危害全人类普遍福祉与安全。同时，美国高度政治化，排他性、对抗性的技术思维与地缘政治竞争心态亦对全球多边秩序下的良性规则与技术理性制造负向效果。美国“泛安全化”政策倾向亦显著影响原有多边主义国际科技秩序下的互利氛围，压缩国际合作空间与多边合作秩序，致使跨国标准制定、数据共享与伦理协同陷入停滞，以致国际关系在尖端科技领域陷入更持久、更复杂、更具有对抗性的结构性冲突。

（二）人工智能领域美式“安全观”的“逆安全化陷阱”及其批判

美方人工智能议题的“泛安全化”策略正对全球制造更高级别的群生风险，以“安全化”为名积极构造科技世界与人类世界的“逆安全化”危机。此背景下，美国人工智能法律政策已被用作牺牲长期安全以汇聚短期利益的政治工具，呈现“逆法治化”倾向。

1. 从“泛安全化”向“逆安全化”的失控转化

美国人工智能法律政策折射的美式“安全观”中，传统“泛安全化”之行为模式可能导向“逆安全化”之损害结果的不可控转化。此处所指“逆安全化”，意在说明美国以“安全化”为由施行一系列“泛安全化”话语建构与政策实践，反而制造大量人工智能领域的安全风险，制造“表面强调安全，实际更不安全”的现实困境。将这一演化逻辑抽象为行为模型，可大体呈现“需求感知-构建安全化诉求-泛安全化-作用深化-逆安全化”这一线性递进关系。其中，前三环节与“泛安全化”的展开逻辑相似，美方在感知其各界利益需求与打击他国产业的要点环节后，针对性构

重事故，背后可能有 29 起轻微事故和 300 起未遂事故。这一比例关系强调了预防轻微和未遂事故对于避免严重事故的重要性。See Marshall P, Hirmas A, Singer M. *Heinrich's pyramid and occupational safety: a statistical validation methodology*[J]. *Safety science*, 2018, 101: 180-189.

⁴² Robert Axelrod, “*Effective Choice in the Prisoner's Dilemma*,” *Journal of Conflict Resolution*, Vol. 1, No. 24, 1980, pp. 3-25.

建相应议题的“国家安全”诉求以实现人工智能法律政策的“泛安全化”。然而，人工智能领域的“泛安全化”明确展现超越前有议题的特殊性。一方面，人工智能当前主流无监督学习、自我对局强化等训练范式仍是高度不可知、不可控的“黑箱”，模型自主性与智能水平同步升级所制造的风险级别远超前有技术，堪称非军事技术领域的“核风险”。另一方面，美国以牺牲必要监管为代价追求其科技霸权维系，人工智能多重风险在未经理性监测、评估与控制的情况下必然进一步扩散与升级。因此，美国“泛安全化”措施在深化其影响效果后，将无可避免地在某时导向某个风险爆发的节点，并迅即引发多方危害的快速涌现与安全局面的断层恶化，成就“逆安全化”的破坏性效果。

2. “逆安全化陷阱”之批判

法学视角下，美式“安全观”全面渗透下的美国人工智能法律政策构成对法治功能与基本精神的根本违悖，明确展现其“逆法治化”倾向。从法的本质与价值而言，现代化社会的法治主义本质要求实现对权力的合理限制与对多方利益的正当保护，⁴³积极映射客观性、公平正义、以人为本、良法善治等重要价值。⁴⁴但以特朗普政府施行的法律政策为代表，美式“安全观”终止此前对于良善人类价值的算法对齐要求，⁴⁵甚至计划放开对色情等低俗化应用的前有限制。同时，美方近年非但未在不可控性极高的人工智能领域合理收束其既有权力，反而追求更强的集权状态以实现“绝对安全”与“绝对权力”，呈现对现代法治本质精神的“倒行逆施”。从法的应用与实践而言，美国作为此前积极应用法律政策追求技术与经济全球化的先行者，其国际关系近来出现愈发明显的保守化与封闭化趋向，并显现从未有过的民族封闭特征。同时，亦滥用长臂管辖的法律体系拉拢“五眼联盟”等盟友国家推出与其相似的对华限制措施，⁴⁶展现政治保守主义与经济民族主义势力下的“技术单边主义”。从法的功能与作用而言，法律政策本该成为人类社会应对人工智能多重风险的核心关键，⁴⁷但美式“安全观”对美国人工智能治理体系的功能扭曲使其风险监测、风险防范、安全等保障机能趋于解体，讽刺化地沦为匹配利益需求与美化集权诉求的政治凭借与傀儡工具。

“没有任何一种理论或战略可以完全解释冷战后美国在单极形势下的政策”。⁴⁸事实上，人工智能领域的美式“安全观”不仅“百害无一利”，且同样“损人不利己”：

⁴³ 郭道晖.多元社会中法的本质与功能——第二次亚洲法哲学大会述评[J].中外法学,1999,(03):101-107.

⁴⁴ 江必新.法治精神的属性、内涵与弘扬[J].法学家,2013,(04):1-10+174.

⁴⁵ 人工智能研发人员常采用基于人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）这一机器学习技术进行大模型调试，旨在通过人类反馈来优化模型的行为并使其更符合人类的期望和价值观，拜登此前行政令强调大模型训练中对一系列民主价值观的对齐，特朗普已近期撤销相关要求。See Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J]. arXiv preprint arXiv:2204.05862, 2022.

⁴⁶ 李庆明.论美国域外管辖:概念、实践及中国因应[J].国际法研究,2019,(03):3-23.

⁴⁷ 吴汉东.人工智能时代的制度安排与法律规制[J].法律科学(西北政法大学学报),2017,35(05):128-136.

⁴⁸ Kai He, “The Hegemon’s Choice Between Power and Security: Explaining US Policy Toward Asia after the Cold War,” Review of International Studies, Vol. 4 No. 36, 2010, pp. 1121-1143.

看似消除创新“障碍”获得技术增速，但尖端技术在缺失安全保障与理据监管下的高速发展实是一种“破坏性发展”；看似为美方争取短期垄断利益，但其自身在长期失序进展的技术风险爆发下却可能最先遭受反噬效果。技术是中立的，技术应用必不是“净正向”的。正如此前，某美国工程师将 ChatGPT 接入可自动瞄准人体要害部位并迅速开火的“哨兵枪”，在引发大范围恐慌后仅被平台封禁账号。面对人工智能这一极为强大的“智能中枢”，以法律政策为中心的专门化多元治理体系尚不能完全化解其负面风险。⁴⁹在无监督学习、自我对局强化等训练策略广泛应用的关键节点，美方竟意图进一步弱化针对用户的必要监管以放任不当技术运用的“野蛮生长”，甚至主动推进人工智能武器化、军事化的失控化发展，以致 Anthropic、Meta 等公司删除原先模型协议中“不将大模型用于军事”的理性限制。以 2024 年获诺贝尔物理学奖的“人工智能教父”Geoffrey Hinton 为代表，多名学者与技术专家均发出人工智能可能毁灭人类的迫切警告。任何利益与权力在全人类普遍安全面前均不值一提，大模型的自主性、智能程度与失控可能性当前本已呈现指数型增长，这些风险正被美方推向无序应用与军备竞赛的全新高峰。作为这一混沌局面的缔造者，美国不仅将整个国际社会带入极度不稳定的震荡周期，亦必将在风险爆发的“失控奇点”为其愚钝行径付出惨重代价。

五、 美式“安全观”的中方对策

当前，美方对华战略意识扭曲下的单方技术“围堰”正在将原本复杂多元的技术竞合关系简化成人工智能领域的零和博弈甚至攸关技术存亡的绝对对抗，不可逆地导向中美关系的消极互动甚至冲突升级，攸关我国国家安全与产业发展。作为当今全球科技领域的主要崛起国及美国在人工智能“主战场”的主要竞争对手，中国须从技术、市场与政策层面寻求突破。以实现人工智能全球竞争格局白热化下的理性因应。在总体国家安全观的指引下，我应聚焦自研攻关、前沿创新与安全发展，推动全球科技领域的良性合作与“去安全化”。

（一） 技术层面之应对

一方面，在先进芯片等人工智能关键技术节点夯实自主化根基，着力构建高质量自研体系。回溯新中国成立后的历史经验，“两弹一星”“北斗系统”“天宫空间站”等重大成果均在外封锁下实现自主研发的技术突破。中美人工智能竞争白热化与美方技术封锁常态化的时代背景下，美国在高端芯片、前沿数据等关键要素对我开展长期封锁以期拉大产业差距，但亦为我自研突破留出结构性机会。此前，我国多项前沿技术领域已有显著进展，如“天河三号 E 级原型机”“神威 E 级超算系统”等国产超算已摆脱对 Intel 处理器的依赖，运用“飞腾”“神威 26010+”等本土处理器迈入百亿亿次的运算性能量级；新一代半导体的制程突破与国产化替

⁴⁹ 张吉豫.构建多元共治的算法治理体系[J].法律科学(西北政法大学学报),2022,40(01):115-123.

代水平亦呈稳定提升，如中芯国际已实现 14 纳米与 28 纳米的制程工艺。但客观而言，我国当前 GPU 水平与以 NVIDIA 为代表的美国先进水准仍存较大差距，DeepSeek 虽运用创设性架构一定程度减少算力依赖，但直视强化训练与推理模型运行等产业发展的未来需求，算力水平的提升依旧关键。一方面，仍须打破“市场换技术”之旧有桎梏与制式思维，在算力、数据等底层环节稳步推进国产化替代，尤其注重发展自主知识产权下的 GPU、NPU 等高性能芯片与模型训练平台。⁵⁰另一方面，积极运用新型张量运算、超大规模自注意力机制优化、分层协同训练算法等前沿框架尝试继续降低对外部技术的依赖度。

另一方面，在人工智能技术发展进程中注重平衡前沿创新与安全向善，⁵¹形成可持续的现代产业生态。创新方面，相较美方“拼基准、堆算力”的粗放式路径，推动进入“后训练强化—多场景落地”并行的迭代周期。以 DeepSeek 等龙头企业为核心力量，加强对人工智能底层算法、网络结构与并行化框架的创设性攻关，优化大模型强化学习架构并整合专业领域的优质数据集，⁵²运用高质量、专业化垂直数据完善中国特色的强化微调路径。⁵³此外，以“杭州七小龙”等初创科技企业为代表，强化人工智能在具身智能、脑机接口、游戏开发、智能设备、云计算等多元场域的泛化应用与多维创新。同时，进一步整合科研机构与头部企业的研发能力打造行业应用的分层技术生态，赋能更多如智能代理“Manus”等新兴产品服务的落地。安全方面，为避免支付技术风险爆发引发的巨额负向成本，审慎监管下的技术安全保障仍旧关键。其一，融合智能技术建立人工智能风险监测与安全评估体系，形成监管部门、研究机构与科技企业的紧密配合，协同抵御算法歧视、数据泄露、失序发展等人工智能细微隐患累积形成规模化事故。其二，积极搭建大模型开源生态以加强模型运作的透明性与可检测性，在鼓励成果共享的同时加强对于可控模型发展的技术社区自治。其三，针对性强化对于人工智能武器化、自主代理等高危险系数新型领域的审慎监管，以细化措施敦促研发主体遵循更高阶技术安全标准与伦理规范以平衡创新进展与技术安全，避免高风险领域“事故三角”的现实化以推动人工智能发挥支撑国家战略转型的“正向乘数效应”。

（二）市场层面之应对

一方面，即时审视全球产业进展态势并灵活调整市场策略，在中美人工智能白热化竞争中把握更多主动权。事实上中美在人工智能关联产业市场呈现互补性，亦即美国拥有领先的技术与资金优势，而中国拥有广阔的数据资源、市场规模与

⁵⁰ 洪涛,程乐.全国算力体系一体化建设的五大问题及治理对策[J].中国科学院院刊,2024,39(12):2086-2095.

⁵¹ 刘艳红.生成式人工智能的三大安全风险及法律规制——以ChatGPT为例[J].东方法学,2023,(04):29-43.

⁵² 程乐.我国高质量场景数据集的供给现状与发展策略[J].人民论坛,2025,(05):68-72.

⁵³ 强化微调（Reinforcement Fine-Tuning, RFT）是一种结合强化学习和微调技术的方法，此前由OpenAI提出。旨在运用数学、物理、医学、法律等专业化领域的高质量数据与专业人员反馈，通过奖励驱动的训练循环优化大模型性能以使其更好地满足特定领域的专业化任务需求，被证明在提升模型推理和规划能力方面具有显著效果。See OpenAI, OpenAI's Reinforcement Fine-Tuning Research Program[EB/OL]. [2024-12-06]. <https://openai.com/form/rft-research-program/>.

应用场景。⁵⁴美国当前对华开展人工智能“零和博弈”，本就展现逆市场化趋向且缺乏经济理性。首先，若美方继续推动高强度技术封锁与单边对抗，我亦可从光伏硅片制备、关键矿产资源出口等优势领域展开反制，弱化美国国际市场合作的供应链安全。其次，识别对我人工智能产业施行不正当市场策略的美方企业并针对性开展反垄断调查、制裁高管人员等反制措施，限制对应美企的国际发展空间。此外，加强国际对话与对外宣传，强调部分美企技术策略对于全球技术安全的破坏性影响，并以 DeepSeek 为代表宣传中方技术在模型开源、低使用成本等层面的特有优势，持续输出优质技术方案以强化中方人工智能产品服务的国际竞争力。最后，积极推动与“一带一路”沿线及对华友好国家在人工智能新兴产业的紧密合作，以中国庞大的内需市场与成熟的供应链体系吸引跨国企业、国际资金在人工智能关联市场进一步提高在华参与度。

另一方面，抓住人工智能时代的市场变革机遇，形成中方市场的多层支撑格局，培育面向全球的全新经济增长点。其一，进一步减少 DeepSeek 等具有高技术引领性、强产业带动力的龙头科技企业对方技术与市场的依赖度，提高中方科技企业的稳定性与抗风险能力，鼓励头部企业及中小型创新团队间的配合协同。其二，充分发挥中小型科技企业的能动性与创新优势，运用数字孪生等机制降低中小企业的试错成本，赋能中小企业人工智能共享生态，在创新驱动发展战略下推动科技创新与新兴产业发展。其三，积极推动我国企业嵌入国产大模型赋能的高端技术网络，实现基于国内循环的技术资源连接，在放大资源连接效用价值的同时实现复合杠杆效应。其四，把握人工智能技术力量以加强工业、制造业等我国已有优势产业在工业互联网、智能制造、智能交通等层面的智能化转型，形成具有全球示范效应的先进应用并推进培育新型高价值应用场景。其五，发挥我国产业政策与宏观调控的稳定性优势，有效引导产业资源向高质量发展赛道的流入，规避噱头炒作与低俗化应用下“劣币驱逐良币”的市场现象。最后，进一步加大对人工智能领域的支持力度，结合政府引导基金、政策激励与地方试点等支持机制辅助我国人工智能产业在全球高端价值链中占据领导位置。

（三）政策层面之应对

一方面，持续完善我国立法，并行推进人工智能政策支持机制与多维治理体系。其一，强化法学学科在人工智能可信治理与风险防范中的主导地位，积极开展法学视角下针对网络空间、算法模型、智能技术等新兴事物的跨学科研究。⁵⁵其二，加快完善人工智能法律政策，对《生成式人工智能服务管理暂行办法》、《互联网信息服务算法推荐管理规定》与《互联网信息服务深度合成管理规定》

⁵⁴ 程乐,赵艺林.制度竞争下的美国人工智能监管与中国因应[J/OL].思想理论战线,2025,(01):93-101[2025-03-10].

⁵⁵ 程乐.网络“空间”隐喻的国际法映射、困境与突破[J].政法论丛,2025,(03):1-16.

等现行规范进行细化，⁵⁶贴合后训练强化学习等新近技术动态与产业发展脉络实现对人工智能进展需求的实时匹配。进程中，须以实践理性下的相对安全为目标，并以个体性安全为依归，⁵⁷深化智能技术的可信规管。其三，实现人工智能战略规划及配套支持政策的时代化迭新，在基础教育、科研设施、人才引进、场景创新等层面强化支持力度以助力产业加速，提升对美竞争力。其四，融合技理、伦理、法理等多元机制以完善人工智能复合型治理体系，⁵⁸杜绝美式“安全观”下法律政策风险防范能力缺位等问题，以合规、安全、可持续的人工智能发展态度实现综合技术实力的稳健进展。

另一方面，积极运用“总体国家安全观”，推动构建面向国际社会的“人类命运共同体”。国家安全法学视角下，中国的国家安全观经历了传统安全观、新安全观、总体国家安全观的长期演进。⁵⁹习近平总书记在党的十九大报告中提出要坚持总体国家安全观的基本方略，强调以人民安全为宗旨，以政治安全为根本，以经济安全为基础，以军事、文化、社会安全为保障，以促进国际安全为依托。⁶⁰人工智能时代，国家安全在原有军事与政治概念外，亦强调人工智能等尖端技术与社会、经济的跨领域耦合。此背景下，总体国家安全观与联合国开发计划署（UNDP）于1994年提出的人类安全概念相呼应，⁶¹可共同发挥理论指引以推动完善专业化、一体化、合作化的人工智能安全治理机制。此外，“人类命运共同体”这一中方主张可在人工智能议题中进一步延伸，⁶²呼吁各国秉持前瞻性态度联合防范人工智能的不当滥用与灾难性失控，保障人类群体普遍安全以促技术发展始终服务于全人类共同福祉。Ole Wæver 提出“去安全化”（Desecuritization）概念，反对滥用安全概念的不当作法，认为安全化应当掌握在受众手中而非任由当权者摆布。⁶³国际层面，应善用“去安全化”话语叙事对抗美国人工智能“泛安全化陷阱”下安全概念的畸化扩张，以联合国为中心推动国际层面人工智能技术标准与伦理评估规范的完善细化，联合多边机制着力构建“安全、可控、共享”的全球人工智能合作机制与治理体系以实现真正意义上的技术安全提升。

⁵⁶ 张文显.构建智能社会的法律秩序[J].东方法学,2020,(05):4-19.

⁵⁷ 王贵松.论法治国家的安全观[J].清华法学,2021,15(02):21-37.

⁵⁸ 毕文轩.生成式人工智能的风险规制困境及其化解：以ChatGPT的规制为视角[J].比较法研究,2023,(03):155-172.

⁵⁹ 凌胜利,杨帆.新中国70年国家安全观的演变：认知、内涵与应对[J].国际安全研究,2019,37(06):3-29+153.

⁶⁰ 邢云文,肖扬.以习近平总体国家安全观为指引加强网络意识形态安全建设[J].思想教育研究,2018,(03):114-117.

⁶¹ 联合国开发计划署（UNDP）在1994年《人类发展报告》中首次提出了“人类安全”的概念，强调安全应以人为中心，涵盖经济、粮食、健康、环境、个人、社区和政治等七个领域。See Adger W N, Pulhin J M, Barnett J, et al. Human security[M]. Cambridge University Press, 2014.

⁶² 习近平.共同构建人类命运共同体[N].人民日报,2017-01-20(002).

⁶³ Ole Wæver, “Securitization and Desecuritization,” in Ronnie D. Lipschutz, ed., On Security, New York: Columbia University Press, 1995.

六、 结语

在百年变局与技术聚变的交汇中，国家安全法学研究成为“显学”，人工智能则被视作关乎产业变革、社会转型乃至国家命运与人类文明走势的关键变量。更易获取的高质量智力资源在承托大规模利好的同时，无监督学习、自我对局进化等前沿范式亦映射一头可能触发未来世界毁灭性风险的“灰犀牛”。表面看来，美国作为全球科技主导国，围绕人工智能议题构建安全化框架是大国竞争升级下对关键技术、市场机遇与国家安全的积极追求。但究其内核，美方倒置安全诉求与利益诉求的“泛安全化”政治话语将本该发挥保障机能的法律政策体系用作收割垄断利益与美化不当打压的政治工具，将中国强行卷入违悖技术理性与贸易全球化的零和博弈与“修昔底德陷阱”，以至放任多重风险的肆意蔓延以换取对全球产业的暂时支配。面向未来世界算法失序、模型武器化、机器人战争等难以预知的技术“天灾”，美方无意强化有效治理以遏制其潜在危害，而是贪恋弹指而过的垄断性权力，主动撕裂其人工智能法律政策尚未弥合的监管裂缝，一手打造尖端科技领域的惨重“人祸”。

1984 年，第三十八届美国超级碗中播出 Apple 公司的一则广告。以反乌托邦小说《1984》为灵感描绘一个阴暗、压抑的未来世界，一名身着彩色运动服的年轻女性冲入麻木、沉闷、穿着灰色制服的人群之中，手持大锤并奋力投向屏幕中的“老大哥”，寓意 Apple 将打破 IBM 当时在计算机领域单一化、同质化的垄断。但彼时的屠龙少年终成恶龙，Apple 曾极大推动消费电子产品及软件操作系统的进步革新，而现今却惯用苹果税、封闭生态系统等排他性工具收割畸高的垄断性收益，借以延续其趋于衰落的全球霸权。以此一例映射美国全域，二十世纪美国在其新自由主义政策下积极推动科技共享与贸易全球化，曾在计算机技术、互联网、航空航天、生物科技、核技术等关键领域带动全球科技长足进展，在共享与合作的价值秉持下增进全人类共同福祉。二十一世纪人工智能领域的主导者仍是美国，但其本质立场与价值位序早已翻天覆地。中国作为智能时代进展下的崛起力量，始终抱持安全优先、透明可信、审慎稳健、共享互惠的技术发展立场，DeepSeek 等中方技术力量之崛起将不止停留于全球科技产业的“搅局者”角色，更将成为美方科技霸权难以维系的有力标识。国家安全法学概念中，“安全”可包含“状态”与“能力”两个面向。一方面，安全指代免遭外部威胁、恐吓与危险的一种状态；另一方面，安全也可描述在不得不面对外部威胁时能够以合理的代价保护自身的能力。⁶⁴在日趋复杂的全球人工智能产业生态中，中国将以有效监管、可持续发展为核心塑造“从容状态”，以技术自研、多维度创新为要点把握“关键能力”，积极对抗美方单极化、对抗性的人工智能政策，保卫国家安全并引领国际科技产业的良性进展。

⁶⁴ Benjamin Miller, “The Concept of Security: Should It be Redefined?” Journal of Strategic Studies, Vol. 2, No. 24, 2001, pp. 13-42.a

2025年中国国际法学年会

American “Security Outlook” Under US AI Policy and China’s Responses

— Based on the Global Industrial Landscape in the DeepSeek Era

CHENG Le¹, GONG Xuan¹, ZHENG Yinglong²

(1. Guanghua Law School, Zhejiang University, Hangzhou, Zhejiang 310008, China;

2. School of Law, Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China)

Abstract: Within a global AI industry in which the United States has maintained a first-mover advantage, domestic large-scale models such as DeepSeek-R1 have been persistently challenging the skewed industry structure dominated by the United States. To preserve its hegemonic status and absolute discursive power in global technology, the US has recently incorporated multiple non-traditional security issues to heighten the “China threat” narrative and establish a so-called American “security outlook.” Through restrictive measures on computing power, capital, and technology, Washington seeks to contain the rise of China’s AI sector. Meanwhile, under the Trump administration taking office in 2025, the newly adopted “hands-off” development policy, aimed at ensuring continued US technological dominance, has led to a regulatory vacuum in the American AI industry—fundamentally overturning the rational balance between technology security and growth. This shift has escalated from “over-securitization” in conventional fields to “inverse securitization” in advanced AI domains. In response to unwarranted US pressures on China’s AI sector and to mitigate the broader threats these tactics pose to the global AI industry—and ultimately to humanity’s collective security—China should adopt swift measures at the technological, market, and policy levels under the overarching perspective of its national security strategy.

Keywords: Artificial Intelligence; Legal Policy; American “Security Outlook”; Technological Security; Over-Securitization; National Security Legal Study

人工智能算法透明度规则分析：国际比较、主要困境及中国因应

熊光清¹

摘要：人工智能算法的广泛应用极大地改变了社会经济和生活方式，同时也引发了数据隐私侵犯、算法歧视和算法“黑箱”等伦理与法律挑战，算法透明度因此成为全球人工智能治理的核心议题之一。算法透明度不仅是保障个人知情权、自主权与公平权的前提，更是实现算法问责、增强公众信任、促进技术健康发展的重要途径。当今世界存在美国模式、欧盟模式、中国模式等三种主要的算法透明度法律治理模式。美国模式依托行业自律与分散立法，注重市场驱动与事后规制；欧盟模式以 GDPR、《人工智能法案》等统一立法为核心，强调事前监管与高强度合规义务；中国模式则通过《个人信息保护法》等基础法律和专项规章，以行政监管和备案制度实现“强监管”与“适度透明”。尽管以上治理模式各有特点，但均面临技术实现困难、商业秘密保护与信息披露冲突、国际治理规则碎片化等困境。中国应在应对算法透明度在全球领域存在的问题上发挥重要作用，通过积极参与全球人工智能算法透明度规则制定，推进双边和区域合作，完善国内立法体系，增强技术研发与国际标准制定能力，并妥善处理主权保护与域外适用的平衡。

关键词：人工智能 算法透明度 GDPR 《个人信息保护法》 全球人工智能治理体系

一、引言

人工智能算法在当今社会各领域得到广泛应用，算法技术在创造巨大价值的同时也带来了前所未有的挑战。算法决策过程往往不透明，令人难以理解其如何得出特定结果，这种不透明性引发了公众对算法公平性和可靠性的担忧。例如：大数据背景下出现的“杀熟”现象和算法偏见导致的歧视性决策，都与算法缺乏透明度密切相关。针对这些问题，各国政府和国际组织开始重视人工智能算法透明度，将其视为实现可信人工智能的重要前提。

国际层面已经出台了一系列原则性文件与法规草案以规范人工智能算法的透明度要求。2019 年，经济合作与发展组织（OECD）通过了《OECD 人工智能原则》（The OECD 人工智能 Principles）²，旨在指导人工智能参与者努力开发

¹ 作者简介：熊光清，对外经济贸易大学法学院教授、博士生导师。主要研究领域：国际法、数字法学、人工智能法律、跨境电子商务。

² 2024 年 5 月，经合组织考虑到新的技术和政策发展，发布了《OECD 人工智能原则》的最新版本，对其

值得信赖的人工智能，其将“透明度和可解释性”列为值得信赖人工智能的核心原则之一，该原则要求人工智能系统具有适当的透明度。2021 年，联合国教科文组织(United Nations Educational, Scientific and Cultural Organization, 简称 UNESCO)通过了首份全球人工智能伦理准则《人工智能伦理问题建议书》，将“透明度和可解释性”确立为人工智能伦理原则的重要组成部分，并明确提出：“人工智能系统的透明度和可解释性往往是确保人权、基本自由和伦理原则得到尊重、保护和促进的必要先决条件。”³这些国际共识表明，提高算法透明度已成为全球各国促进人工智能可信、公正发展的共同目标。

各主要国家和地区开始制定相关法律规则，应对算法透明度治理难题。欧盟则走在前列，通过《通用数据保护条例》(General Data Protection Regulation, 简称 GDPR)⁴和《人工智能法案》⁵(Artificial Intelligence (人工智能) Act, 简称 EU 人工智能 Act)等法规，对算法透明度义务作出了明文规定。美国在联邦层面虽然尚未出台综合性的人工智能法律，但通过行业监管和政策指导对算法透明度提出了要求。中国也出台了《个人信息保护法》及多部专项部门规章，强化算法合规和透明度要求。不同法域在治理模式上的差异，使得对这些规则进行比较分析具有重要意义。

二、人工智能算法透明度：基本概念及理论价值

人工智能算法不透明度不仅可能导致歧视、偏见和个人权利损害，更使责任主体难以确定和追究，给法律规制和社会治理带来巨大挑战。通过推动人工智能算法决策过程与依据的透明化，有助于保障公众知情权、增强社会信任和接受度，并促进人工智能技术的可持续发展。

(一) 人工智能算法透明度：概念与内涵

2016 年，尼古拉斯·迪亚科波洛斯(Nicholas Diakopoulos)、迈克尔·科利斯卡(Michael Koliska)正式提出了算法透明度的概念。他们认为，算法透明度是“向相关方披露有关算法的信息，以便其进行监控、审查、批评或干预”。⁶该概念最初被推广到相邻领域，例如社交媒体、在线评论平台和在线广告。最近，算法

进行了更新。迄今为止，共有 47 个国家承诺遵守《OECD 人工智能原则》。

3 2021 年 11 月，联合国教科文组织制定了有史以来第一份全球人工智能伦理标准——《人工智能伦理问题建议书》。该《建议书》由 193 个会员国一致通过，旨在推进透明和公平等根本原则的基础上保护人权和尊严，并始终牢记人类监督人工智能系统的重要性。该《建议书》涉及广泛的政策行动领域，使决策者能够将其核心价值观和原则转化为数据治理、环境和生态系统、性别、教育和研究、医疗健康、社会福利等方面的行动，因此具备高适用性。

4 本法案在 2016 年 4 月 27 日通过，两年的缓冲期后，在 2018 年 5 月 25 日强制执行。根据欧洲联盟运作条约第 288 条第 2 项，因为 GDPR 属于欧盟条例，不是指令，所以不需经过欧盟成员国立法转换成各国法律，而可直接适用。

5 该项法案由欧盟委员会在 2021 年 4 月 21 日提议，于 2024 年 3 月 13 日获得欧洲议会通过，5 月 21 日获欧盟理事会正式批准通过，并将于当地时间 8 月 1 日正式生效。

6 Nicholas Diakopoulos, Michael Koliska. Algorithmic transparency in the news media. Digital Journalism, 2017,5(7):809-828.

透明度也被应用于其他不同的场景，如能源系统、需求规划和人工智能等领域。⁷人工智能算法透明度是指算法决策过程及其依据能够被清晰理解、有效解释，并向利益相关方以适当方式进行信息披露的特性。算法透明度包含三个核心维度：一是算法逻辑透明，即揭示算法决策的原理、模型与过程，使人们理解其运作方式；二是决策结果透明，即告知用户算法决策所依据的关键数据和因素，使受影响者知晓决策的理由；三是责任主体透明，即清晰地明确算法的开发者、部署者和监管者的法律责任主体。

简单来说，算法透明度要求算法不再是不可测的“黑箱”，而应以人类可理解的方式展现其决策原理，从而为外部监督和问责提供可能。算法透明度意味着要使人工智能系统的决策过程对相关利益主体达到可理解、可追溯的程度，即让相关利益主体知道算法系统的各部分在做什么。它涵盖了算法运行机制的信息公开，包括算法决策所依据的逻辑、流程、主要因素等内容。与信息披露制度或 WTO 法律框架中主要约束政府的“透明度”要求不同，人工智能治理中的透明度原则作为对技术的约束，其范围扩大至私主体，对商业主体(技术主体)施加了更多义务。⁸较之可解释性，算法透明度的价值色彩更浓，以至于技术界罕有尝试对算法透明进行严格定义。⁹

算法透明度与可解释性概念密切相关但有所区分。算法透明度强调事前的开放性，即对算法系统结构和运行机制的了解；而可解释性更侧重于事后对具体决策结果给出原因说明。《人工智能伦理问题建议书》第 40 条指出：“可解释性是指让人工智能系统的结果可以理解，并提供阐释说明。人工智能系统的可解释性也指各个算法模块的输入、输出和性能的可解释性及其如何促成系统结果。因此，可解释性与透明度密切相关，结果和导致结果的子过程应以可理解和可追溯为目标，并且应切合具体情况。”透明度可以通过公开算法模型结构、数据使用情况等实现，可解释性则是针对某一次算法决策，提供人类可理解的解释说明，例如：哪些输入特征导致了该决策。提高算法透明度有助于实现算法结果的可解释性，而可解释性又反过来丰富了透明度的内涵。总的来说，无论是公开算法逻辑还是解释具体输出，最终目的都是让算法“不再神秘”，使其运作机制对人类而言是可理解、可评价的。

算法透明度的内涵是多层次的。从信息披露主体来看，既包括向监管者提供详尽的技术信息以便审查（例如：源代码、模型参数等），也包括向公众和用户提供清晰易懂的说明以满足其知情权。从内容上看，透明度涵盖了算法生命周期

7 Tobias Bitzer, Martin Wiener, W. Alec Cram. Algorithmic Transparency: Concepts, Antecedents, and Consequences——A Review and Research Framework. Communications of The Association for Information Systems, 2023(52):293-331.

8 张平：《透明度原则在人工智能治理中的适用》，载《数字法治》2025 年第 1 期。

9 苏宇：《优化算法可解释性及透明度义务之诠释与展开》，载《法律科学(西北政法大学学报)》2022 年第 1 期。

各阶段的信息，例如：算法开发阶段的数据来源和训练方法、算法决策阶段的模型原理和逻辑、以及算法应用阶段对外部环境的响应机制等。透明度还应体现于算法治理的制度设计中，例如：建立算法登记备案、影响评估报告等机制，使算法使用情况可追踪、有记录可查。当然，算法透明度并非要求向任何人公开算法的所有细节——那既不现实也可能引发滥用——而是通过分层次、分类别的信息披露，使不同利益相关者在各自需要的层面上对算法有所认知，从而实现有效监督与合理信任。

（二）人工智能算法透明度的重要性与必要性

人工智能算法过程的透明化有助于公众理解并信任算法的输出结果，从而避免因“黑箱”决策带来的偏见或歧视问题，同时，在算法决策出现错误或导致侵权时，相关各方可以明确责任主体。从法治建设、社会公平、消费者权益保护以及算法治理的角度看，提升算法透明度既是现实需求，也是人工智能治理的必然趋势。

第一，人工智能算法透明度关系到保障公民知情权与自主选择权。算法的透明性和可理解性关涉人类的知情权益和主体地位，人类对算法的安全感、信赖感很大程度上取决于算法是否透明和可理解。实施“算法透明”机制是解决“算法黑箱”问题的根本方法。¹⁰如果算法在黑箱中运行，个人往往无法知道自己何时受到算法决策影响、更无从了解其影响产生的原因。提高算法透明度，告知个人“算法在做什么”，是维护个体知情权的基本要求。只有在知情的基础上，个人才能自主选择是否接受算法的决策结果，或行使拒绝权和救济权。例如，GDPR 等法律赋予个人在某些自动化决策场景下拒绝接受机器决策的权利，其前提正是个人需要被告知有关算法逻辑的“有意义信息”。

第二，人工智能算法透明度有助于促进公平与防范歧视。《人工智能伦理问题建议书》第 39 条指出：“从社会—技术角度来看，提高透明度有助于建设更加和平、公正、民主和包容的社会。提高透明度有利于开展公众监督，这可以减少腐败和歧视，还有助于发现和防止对人权产生的负面影响。”但是，算法不透明就会掩盖许多偏见和歧视，产生许多不应有的问题。近年来出现多起因算法不透明而导致歧视的案例，例如美国 COMPAS 刑事风险评估算法被曝对非洲裔被告有系统性偏见。¹¹由于算法内部机制未公开，受影响群体很难证明算法存在歧视，从而难以寻求纠正。如果算法决策依据和数据使用更加公开透明，公众和监管者

10 李卫东、杨莉莉：《信息推荐算法透明机制的利益相关者分析》，载《新媒体与社会》2024 年第 4 期。

11 COMPAS 是一款预测罪犯可能会成为再犯的风险评估软件。在它的帮助下，法官对被告和罪犯的未来情况，如保释金额、判刑等做出决策。COMPAS 曾经预测了佛罗里达州超过 10000 名罪犯是否会成为再犯的概率，ProPublica 将该软件的预测结果与实际结果进行了比较。结果表明：当算法的预测结果与实际结果相同时，无论是对黑人和白人，算法准确率大致相同。然而，当算法的预测结果与实际结果不相同，算法存在偏见。一方面，黑人被算法标记为会成为再犯的风险远高于白人，而实际上这些黑人并没有成为再犯；另一方面，白人被标记为会成为再犯的风险很低，但实际上这些白人却继续犯罪。

就有机会发现并纠正算法偏见，防止某些群体因不公正的算法而蒙受损失。例如：要求算法决策提供解释，可以使受影响者了解算法是否考虑了不应考虑的敏感因素，从而为质疑和申诉提供依据。可以说，算法透明度是实现“算法公平”的前提条件之一。

第三，人工智能算法透明度有利于实现算法问责与权利救济。从法律角度来看，算法透明度也是实施算法问责和权利救济的必要保障。如果算法造成了错误决定或侵害，他人只有了解算法如何做出该决定，才能确定责任主体并提出救济请求。算法透明机制实则是在这些具有高度复杂性和自动性的机器学习算法决策上建立一套制度化的自我监控和外部监督机制，进而有助于强化算法决策问责。¹²在传统的人类决策中，决策者需要说明理由，才能接受上级审查或司法审理；同理，算法决策也需要通过“解释”来实现问责。算法透明是实施算法监督、自主选择的重要前提，也是在损害发生后用户进行权利救济的必要保障。如果不了解人工智能系统是如何作出决策的，就无法判断其结果是否给予了所有人平等对待。当人工智能被应用于诸如招聘、信贷或司法等敏感领域时，其决策可能对个人生活产生深远影响，因此这一问题尤为重要。¹³无论是个人维权还是行政监管，均离不开一定程度的算法透明度作为支撑。如果算法一直是“黑箱”，问责将无从谈起，权利救济也缺乏抓手。

第四，人工智能算法透明度能够增强公众信任与促进技术发展。人工智能技术的持续进步可能促使机器学习算法越来越具有可解释性，算法治理完全可以在提高政府决策准确性与效率的同时，回应法律及公众对透明度的要求，甚至有助于提升政府决策的合法性与公众信任。¹⁴近些年来，“可信任人工智能”（Trustworthy 人工智能）已在全球范围内成为引领人工智能发展的核心要求和原则。算法信任反映了人工智能以人为本的发展态势，它转变了过去以技术为中心（Technic-centric）的局限性，强调关注技术与用户之间的互动以及公众感受。¹⁵从社会整体利益出发，提升算法透明度有助于增强公众对人工智能的信任，为技术持续健康发展奠定基础。有研究表明，更高的透明度有助于增强最终用户的控制力，并提升对复杂算法系统的接受度。¹⁶当用户了解算法的运行模式之后，他们就会倾向于认为算法的结果是高质量的和可信的，并且这种理解也会有助于他们发现和检测算法中潜在的偏见和错误，以便技术人员进行及时的改进和修正。

12 张恩典：《算法透明度的理论反思与制度建构》，载《华中科技大学学报(社会科学版)》，2023年第6期。

13 Gyandeep Chaudhary, “Unveiling the Black Box: Bringing Algorithmic Transparency to 人工智能”. Masaryk University Journal of Law and Technology, Vol.18, No.1(2024)

14 Cary Coglianese, David Lehr, Transparency and Algorithmic Governance, Administrative Law Review, Vol. 71, No. 1 (Winter 2019), pp. 1-56.

15 戴维、王锡铤：《算法透明机制的局限性及其克服》，载《华东政法大学学报》2025年第1期。

16 Aaron Springer, Steve Whittaker, “Progressive disclosure: When, why, and how do users want algorithmic transparency information?” ACM Transactions on Interactive Intelligent Systems, 10(4), 2020, 1-32.

¹⁷同时，算法透明度提供了社会监督介入的可能性，可以让社会公众监督算法应用中的安全、隐私、伦理问题和其他风险，防止算法的滥用和误用。¹⁸从长远看，这将形成“透明促信任、信任促发展”的良性循环。算法透明度的必要性已成为普遍共识，其核心价值在于以公开促公正、以透明保信任，在个人权利和技术进步之间架起桥梁。

（三）人工智能算法透明度的法律规制与伦理规范

在人工智能算法透明度治理过程中，法律规范和伦理准则两方面力量共同发挥作用。各国立法和监管机构开始将算法透明度要求纳入法律条文，以刚性的义务约束来确保算法提供者履行透明义务；诸如行业准则、伦理原则等“软法”在算法治理初期发挥了重要引导作用，填补了法律空白并为立法提供了原则依据。

第一，在法律规制方面，许多国家已在立法中明确了算法决策的透明义务。算法透明度的理论基础已经较为成熟，包括行政法上的程序正义理论、数据保护法上的知情同意原则、以及消费者保护法中的知情权与公平性原则。法国早在1978年的《信息技术与自由法》中就规定，公民有权知悉自动化处理的逻辑，并可拒绝仅由此类处理作出的决定。2016年法国《数字共和国法》进一步要求行政机关对算法决策进行解释，公开所涉模型、大致参数、权重和数据来源等信息。这些规定开创了以法律保障算法透明度的先例。2021年，美国出台《算法正义和互联网平台透明度法案》，提出了一系列措施确保平台在算法使用过程中履行审核和透明度义务。2021年，中国开始实施的《个人信息保护法》明确要求，利用个人信息进行自动化决策应当保证决策的透明度和结果公平、公正，不得在交易中对个人实行不合理差别待遇。2022年，中国开始实施的《互联网信息服务算法推荐管理规定》确立了平台的算法告知义务，其中第16条规定：“算法推荐服务提供者应当以显著方式告知用户其提供算法推荐服务的情况，并以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制等。”欧盟GDPR将透明度作为个人数据处理的基本原则之一，要求数据控制者在使用自动化决策时向数据主体提供有关算法的“有意义信息”，包括所采用逻辑及其对数据主体的重要性和后果。2022年，欧盟《数字服务法》要求在线平台采取广泛的透明度措施包括推荐算法的透明度，让用户更好地了解平台是如何向他们推荐内容。可见，法律正在将抽象的“透明”理念转化为可执行的义务和权利，为算法治理提供强制力支持。

第二，在伦理规范方面，各种人工智能伦理准则和行业标准已经先行一步，倡导算法透明原则并指导从业者实践。2016年12月，电气与电子工程师协会（Institute of Electrical and Electronics Engineers，简称：IEEE）发布《合伦理设

17 邓克涛、张贵红：《算法透明度：从理论到实践的探索与反思》，载《科学学研究》2024年第7期。

18 张红春、章知连：《从算法黑箱到算法透明：政府算法治理的转轨逻辑与路径》，载《贵州大学学报(社会科学版)》2022年第4期。

计:利用人工智能和自主系统(人工智能/AS)最大化人类福祉的愿景(第一版)》,旨在鼓励科技人员在人工智能研发过程中,优先考虑伦理问题。其中提出,透明性原则意味着自主系统的运作必须是透明的,人工智能/AS 是透明的意味着人们能够发现其如何以及为何做出特定决定。2017 年 1 月,美国计算机协会公共政策委员会发布《关于算法透明性和可问责性的声明》指出,应当在系统开发和部署的每个阶段应用算法透明和可问责性的原则,其中提出了七项算法治理原则。这些原则虽非法律,但对行业实践产生了积极影响,被视为算法伦理的基石。此外,《OECD 人工智能原则》和《人工智能伦理问题建议书》均将透明度和可解释性作为人工智能发展的指导性价值观。这些软法规范为各国政策制定提供了重要参照,在一定程度上推动了硬法的出台。

第三,在人工智能算法透明度治理过程中,重视实现软法与硬法的协同共治。法律是社会的底线伦理,对人类行为具有强制规范作用,而伦理是引导人们去做不止于守住底线的、对社会有益的事情,其对人类行为很多时候只是具有软约束作用。伦理原则如果转化为法律规范,就会具有强制性作用。¹⁹伦理准则往往为法律规制奠定价值共识和技术指引,而法律的出台又将伦理要求上升为具约束力的义务。两者在算法透明度治理中呈现出协同作用。例如:欧盟在制定人工智能法律框架时广泛参考了其人工智能高级别专家组 2019 年发布的《可信人工智能伦理准则》(Ethics guidelines for trustworthy 人工智能)²⁰,其中“透明性”正是指南的关键原则之一;中国在 2021 年通过《互联网信息服务算法推荐管理规定》前也发布了《新一代人工智能治理原则——发展负责任的人工智能》²¹等政策性文件,强调公平、公正、透明的价值导向,以引领算法治理立法。可以说,道德准则的倡导为法律介入提供了正当性依据,而法律的强制执行力又确保了伦理原则落地生根。面对快速演进的人工智能,仍需软法与硬法并举。一方面,要及时更新法律法规应对新问题,另一方面,要发挥行业自律和伦理指导,在法律未覆盖到之处促进行业遵守透明、公平的底线。只有这样,才能构建一个既有刚性约束又具柔性引导的算法透明治理体系。

综上所述,算法透明度已从最初的伦理倡议逐步走向法律化、制度化。它既包含了法律条文中的硬性义务,也融汇了道德准则中的软性要求。法律和伦理的融合,使得算法透明度这一理念更具操作性和普适性,也为各国开展算法治理的

19 熊光清:《科技向善:规范人工智能发展的伦理考量》,载《国家治理》2025 年第 4 期。

20 2019 年 4 月 8 日,欧盟人工智能高级别专家组发布了《可信人工智能伦理准则》。其中提出了七项人工智能系统需要满足的关键要求,以确保其可信性。关于透明度,该准则要求:“数据、系统及人工智能的商业模式应当透明,这一点可通过建立可追溯机制实现。此外,人工智能系统及其决策必须根据利益相关方的具体需求予以适当解释。人类需要明确意识到他们正在与人工智能系统互动,并被充分告知该系统的能力与局限性。”

21 国家新一代人工智能治理专业委员会于 2019 年 6 月发布《新一代人工智能治理原则——发展负责任的人工智能》,提出了人工智能治理的框架和行动指南。这是中国促进新一代人工智能健康发展,加强人工智能法律、伦理、社会问题研究,积极推动人工智能全球治理的一项重要成果。其强调在人工智能发展中要实现安全可控与伦理约束的统一,对中国人工智能治理体系构建产生了重要影响。

国际合作奠定了共同语言和基础。通过法律规制与伦理规范的协同治理，可在保障透明度的同时兼顾创新发展，进而推动建立更加公平可信的全球人工智能治理体系。

三、人工智能算法透明度的法律规制：国际比较的角度

美国、欧盟与中国在算法透明度方面的法律法规与政策措施各有侧重，并存在较大差异，体现出不同的监管范式和价值取向。美国算法透明度法律规则较为分散，强调企业自愿合规、市场监督和事后惩罚。欧盟倾向于建立统一、严格的算法治理法律框架，其以 GDPR 和《人工智能法案》为核心，监管力度强，具有明显的保护公民权益的价值导向。中国的算法透明度规则强调维护社会稳定和内容安全，重视保护个人权益，以行政监管为主，体现了秩序导向和社会管理需求。

（一）美国算法透明度的法律法规与政策措施

美国目前尚未有一部专门针对人工智能或算法透明度的综合性联邦法律，但通过多种途径对算法透明度予以规制，包括现有部门法律框架、行政机构指南以及拟议立法等。美国的模式更倾向于部门化监管+行业自律，在保障透明度的同时注重不对新兴技术设限过死。这种模式一方面依托既有法律（如反歧视法、消费者保护法等）来治理算法可能造成的危害，另一方面通过发布指导原则和倡议鼓励企业自我审查和信息披露。

尽管美国没有统一的“算法透明法”，但若干既有联邦法律在各自领域内涉及到了算法决策的告知义务。例如，在信用领域，1971 年开始实施的《公平信用报告法》（FCRA）要求当基于信用报告的自动化决策对个人产生不利影响时（如拒绝贷款），机构必须向当事人发出不利决定通知并提供原因。这种告知义务实质上保障了消费者获得算法决策解释的权利。2018 年，加利福尼亚州通过《消费者隐私法案》（California Consumer Privacy Act，简称 CCPA）赋予消费者了解其个人信息如何被使用的权利，企业须在隐私政策中披露是否进行“自动化决策”或“画像”等处理，但对算法逻辑本身的透明度要求相对有限。这些法律虽未明确提及“算法透明”，但其信息披露和解释义务涵盖了算法决策情境。2021 年美国参议院出台了《算法正义和互联网平台透明度法案》，禁止互联网平台通过算法歧视性地对个人信息进行处理，并提出一系列措施确保平台在算法使用过程中履行审核和透明度义务，从而建立了算法的安全性和有效性标准，为互联网平台的算法正义与透明指明了方向。²²此外，一些监管规定也隐含了算法透明度准则，例如美国联邦贸易委员会（FTC）将算法偏见视为可能的不公平商业行为，敦促企业在使用算法时确保决策过程公平且对消费者可解释，否则可能构成对消费者的欺骗。

²² 孔凡东：《关于美国〈算法正义和互联网平台透明度法案〉的借鉴与启示》，载《海南金融》2021 年第 12 期。

在立法之外，美国政府和专业团体发布了一系列政策文件，倡导算法透明原则以引导实践。2017 年，美国计算机协会公共政策委员会发布的《关于算法透明性和可问责性的声明》就明确使用算法决策的机构应当对算法所遵循的程序和具体制定的决策进行解释，同时在包括系统开发在内的每个阶段应用算法透明度和可问责性的原则，以及对数据、算法模型和决策等过程进行记录，以便于后续产生损害时的算法审计。这些原则在业内产生广泛影响，被科技公司视为设计人工智能系统的指南。2022 年 10 月，美国白宫科技政策办公室（Office of Science and Technology Policy，简称 OSTP）发布《人工智能权利法案蓝图》（Blueprint for an Artificial Intelligence Bill of Rights），该蓝图反映了美国政府对私营公司和政府机构鼓励采用人工智能技术的原则设想，以减轻涉及数据隐私、算法歧视和自动化系统使用的风险。这一政策虽不具法律约束力，但体现了美国政府对算法透明度的期望，并为后续立法奠定基调。

随着人工智能应用日趋广泛，近年来，美国国会多次尝试推动有关算法透明和问责的立法。其中具有代表性的是《算法问责法案》（Algorithmic Accountability Act）。该法案最早于 2019 年提出，此后尽管多次被提交国会，但截至目前，尚未获得国会批准成为正式法律。《算法问责法案》与欧盟 DREP 都共同旨在确立治理框架，以追究失责实体的责任，确保人工智能系统的伦理性、合法性和安全性。这两部法律都重点强调了算法透明性以及系统功能的解释，整体目标在于将问责机制嵌入人工智能运作过程。²³《算法问责法案》要求大型科技公司对其使用的高风险自动决策系统进行影响评估（Impact Assessment），评估内容包括：对算法的详细描述、所使用的数据及数据治理措施、算法可能造成的隐私安全影响、潜在歧视风险以及已采取的减缓风险措施等。这一法案设计了“三层信息披露机制”，即企业需编制内部的详细评估报告、提交给监管机构的评估简报以及对公众公开的概要信息。这种分级披露既保证了监管部门掌握充分信息、消费者享有基本知情权，又限制了敏感细节的过度公开以防止商业秘密泄露，在透明度与企业机密保护之间寻求平衡。尽管其仍未正式通过，但反映了美国立法者在透明度问题上的权衡取向，即通过强制“算法影响评估+分层披露”来提高透明度和问责，而非简单要求算法公开源代码。

除了联邦层面的努力，一些州和地方政府也率先实践了算法透明监管。纽约市于 2018 年成立算法管理任务组，探索政府部门算法决策的透明和公平问题。随后，纽约市在 2021 年通过了一项针对招聘领域的地方法规，要求对用于招聘的自动决策工具进行年度偏差审计并向求职者披露相关信息（尽管这侧重于算法公平，但也含有一定信息披露义务）。伊利诺伊州则在 2019 年制定《人工智能视

23 Alana Engelmann, Algorithmic transparency as a fundamental right in the democratic rule of law: a comparative approach to regulation in European, North American, and Brazilian Contexts. *Brazilian Journal of Law, Technology and Innovation*, Vol. 1 No. 2, 2023.

频面试法》，要求用人单位在用人工智能分析招聘面试前通知候选人并取得同意，同时在面试后应删除相关视频数据。这些地方举措体现了先行先试的精神，在缺乏联邦统一规则时通过局部立法提高算法使用的透明度和可信度。

近年来，美国行政部门也开始制定更具约束力的规则来强化特定领域的算法透明。例如：2023 年 12 月美国卫生与公众服务部（HHS）通过国家卫生信息技术协调员办公室（ONC）最终确定《健康数据、技术和互操作性：认证计划更新、算法透明度和信息共享（HTI-1）规则》。HTI-1 规则包括算法透明度，即为人工智能和其他预测算法建立首个此类透明度要求，其要求纳入联邦认证的电子健康记录系统中的人工智能或预测模型，必须向使用者披露算法的基本原理、用途，以及模型在临床决策中的作用等信息。这被视为美国联邦政府首个直接针对人工智能算法透明度的强制性规定，标志着美国在行业监管层面正逐步将长期倡导的透明度原则转化为具体合规要求。同时，美国联邦贸易委员会（FTC）在 2023 年发布了商业指南，明确警告人工智能开发者不得对其人工智能产品的功能和局限做出虚假或误导性表述，否则将被视为违反 FTC 条例。这实际上鼓励企业主动提高透明度，如实披露人工智能决策的可靠性和适用范围，以免招致监管处罚。

在司法实践中，美国出现了涉及算法透明度的标志性案件，引发了关于算法透明度和可解释性的法律讨论。“威斯康星州诉卢米斯案”（State v. Loomis, 2016）凸显了算法透明度与被告权利的冲突。在该案中，威州法院在量刑时参考了商业公司开发的犯罪风险评估算法（COMPAS）的分值，被告卢米斯对此提出上诉，主张他有权了解该算法如何判断其风险等级，以及算法可能存在偏见。但算法提供商以商业秘密为由拒绝公开内部机制，最终法院驳回了卢米斯的申诉，维持使用 COMPAS 的判决。此案引发广泛争议。一方面，人们担忧司法过程中引入不透明的算法有损正当程序，被告无法获得对抗证据的机会；另一方面，支持者认为算法只是辅助且其准确率较高，公开源代码可能侵犯厂商权益。关于这一案例的讨论折射出美国在算法透明度领域面临的两难，即：如何既保障个人基本权利，又不妨碍算法创新。尽管该案未能推动厂商披露算法细节，但其带来的舆论压力和学界讨论将会促进后续对“算法透明与商业机密平衡”的立法考虑。

综上所述，美国算法透明度规则呈现出碎片化治理状态，但存在体系化逐步加强的态势。尽管没有统一立法，但通过行业法律、规章和政策框架共同发力，形成了“原则倡导+行业监管+案例推动”的混合模式。这种模式的优点在于灵活性强，注重实用和平衡，缺点则是缺乏统一标准，各行业对算法透明度的标准没有进行统一。美国强调创新驱动和企业自治，在透明度上更倾向于鼓励自律和事后问责，而非像欧盟那样设立详尽的预先义务。但随着社会对算法公正性的要求提高和欧盟等外部规则的影响，美国未来也可能朝着更明确的联邦立法方向发展，在保障透明度的同时保持其独有的宽松监管风格。

（二）欧盟关于人工智能算法透明度的法律规则

欧盟在人工智能治理方面处于领先地位，力图保障本土数字主权与信息安全，并影响全球数字治理规则。²⁴欧盟在算法透明度规制方面也走在世界前列，已建立起相对成体系的法律框架。GDPR 被许多人认为是数据隐私监管领域最重要的变革，其属地适用原则是跟随其所保护的数据而非组织所在地，因此它直接影响到全球范围内处理欧盟公民数据的各类组织。²⁵这样，就形成了所谓的“布鲁塞尔效应”。在 GDPR 中，欧盟即确立了数据处理活动中的透明度义务，并赋予个体一定的自动化决策知情权和拒绝权。此后，欧盟又陆续推出专门针对人工智能的立法方案，包括 2024 年的《人工智能法案》。欧盟希望通过《人工智能法案》继续发挥布鲁塞尔效应，实质影响全球规则走向。这些法律文件共同构成了欧盟算法治理的“双层体系”：一层通过数据保护法规保障个人在算法处理个人数据时的知情权和解释权，另一层通过人工智能监管法规直接规范高风险算法系统的透明度和可解释性要求。

虽然 GDPR 并非专为人工智能制定，但其中多项规定涉及算法决策的透明和解释义务，被视为算法透明度的基础性规则。GDPR 第 5 条将“透明原则”列为个人数据处理的基本原则之一；第 13 条和第 14 条要求当个人数据被用于自动化决策（包括分析或预测）时，数据控制者须在收集信息时向数据主体提供相关信息，其中应当包括“有关逻辑的信息以及此类处理对数据主体重要性和预期后果”。换言之，个人有权获知其数据被算法如何处理、依据何种逻辑得出结论。GDPR 确认了个人数据保护作为一项基本人权，并赋予数据主体对其个人信息的处理、共享和分析方式拥有更大的控制权。当然，应当注意的是，当数据由人工智能算法处理时，主张控制权并提供充分解释就成为一名挑战。随着计算能力的大幅提升和大数据处理的广泛应用，人工智能算法变得极其复杂且不透明，使得大多数数据主体难以理解其运作机制。²⁶

此外，GDPR 第 22 条规定数据主体有权不受仅基于自动处理（包括分析）而做出的对其产生法律或重大影响的决定的约束。虽然该条更多是赋予拒绝接受算法决策的权利，但行使这一权利往往需要个人了解算法决策已发生且可能不利，因而与透明度密不可分。在这里，算法的解释不应被理解为静态的说明性陈述，而应视为一个循环的、多层次的透明过程，包括多个层面：算法的一般信息、基

24 中国信息通信研究院政策与经济研究所、中国科学院计算技术研究所智能算法安全重点实验室：《大模型治理蓝皮报告——从规则走向实践（2023 年）》，<https://www.cict.ac.cn/kxyj/qwfb/dztbg/202311/P020231124526622371194.pdf>。

25 Edward S. Dove, The EU general data protection regulation: implications for international scientific research in the digital era. *Journal of Law, Medicine & Ethics*, Volume 46 Issue 4, 2018.

26 Alexander J. Wulf, Ognyan Seizov, “Please understand we cannot provide further information”: evaluating content and transparency of GDPR-mandated 人工智能 disclosures. *人工智能 & SOCIETY* 9, 235–256 (2024). <https://doi.org/10.1007/s00146-022-01424-z>.

于群体的解释，以及对个体决策的法律正当性说明。²⁷自 2018 年 5 月起生效以来，GDPR 隐私法适用于所有在欧盟境内或境外收集并处理欧盟公民个人信息的组织。²⁸GDPR 实施后，各成员国的数据保护机构也发布指导意见，强调数据主体有权获得关于算法决策的适当解释和信息，以便行使上述权利。

总体来说，GDPR 通过告知义务+权利赋予的方式，将算法透明度嵌入了数据保护法律体系中，为个人隐私提供了保障。2019 年 4 月，欧盟《算法问责及透明度监管框架》就算法自动化决策应用提出了全面的监管要求，将算法透明度与可解释性作为实现算法公平的基础性条件。2020 年，欧盟《数字市场法》和《数字服务法》制定了更加适应平台和数字市场的竞争监管规则，进一步明确数字服务提供者的责任、遏制平台恶性竞争行为，明确算法问责和审计规则，实现算法透明度的提升。

在 GDPR 之外，一些欧盟成员国还通过本国立法强化了算法透明要求。典型例子是法国和英国。法国除了在数据保护立法中规定知情权外，还在行政程序领域引入了算法透明规则。2017 年法国《数字共和国法》（Digital Republic Act）规定，当行政机关运用算法作出个体决定时，应向公民提供关于所使用算法的关键信息，包括：算法涉及的模型类型、处理的大致参数、各参数的一般权重以及所使用数据的来源等。这使法国成为较早要求政府算法透明的国家。英国在脱欧前曾受欧盟框架约束，2021 年，英国发布了《算法透明度标准》（Algorithmic Transparency Standard）为政府机构在决策算法的使用上提供透明化指引。该标准要求政府机构在使用算法辅助决策时公开描述算法用途、决策影响、输入数据和性能指标等信息，并采用模板方便公众理解。虽然英国此举发生在其退出欧盟之际，但理念上与欧盟对公共部门算法问责相一致。

欧盟最为雄心勃勃的方案是制定了专门针对人工智能治理的《人工智能法案》。该法案引入分级监管，将人工智能系统按风险分为不可接受、高风险、有限风险和最低风险四类，并对高风险人工智能系统设置了严格的合规义务。其中，透明度义务是高风险人工智能合规要求的核心内容之一。具体而言，该法案要求高风险人工智能系统的提供者在将产品投放市场前，必须提供详细的技术文档和用户指南，其中包括对算法系统的用途、决策逻辑、性能、限制的说明。用户（这里的用户多指部署人工智能系统的机构）也负有一定透明义务，如需确保其使用人工智能系统的方式可以被追溯和解释，以便在对个人产生影响时能够向其做出解释。这实际上建立了“算法解释链”，即：从开发者到部署者再到最终受影响者，每一环节都有获取算法信息的渠道，确保个人可追溯算法决策来源。这表明欧盟

27 Margot E. Kaminski, Gianclaudio Malgieri, Multi-layered explanations from algorithmic impact assessments in the GDPR. Proceedings of the 2020 Conference on f 人工智能 rness, accountability, and transparency, 27 January 2020. <https://dl.acm.org/doi/abs/10.1145/3351095.3372875>.

28 Razieh Nokhbeh Zaeem, K. Suzanne Barber, The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. ACM Transactions on Management Information Systems Vol. 12, No. 1, 2020.

认识到提高透明度不仅是技术问题，更需要使用者理解算法基本原理，真正实现有效解释。

《人工智能法案》的透明度要求并不局限于高风险系统。在其他类别，如与消费者直接交互的人工智能系统，法案同样规定了相应的透明义务。例如，针对聊天机器人等让人类以为自己在和真人互动的人工智能，要求在互动开始时即向用户明示其为人工智能系统；针对深度合成内容技术，要求标注人工智能生成内容以防止误导公众。这些规定与中国 2022 年发布的《互联网信息服务深度合成管理规定》有异曲同工之处，都是为了让人工智能生成内容“显著可识别”。

此外，2022 年正式生效的欧盟《数字服务法案》（Digital Services Act，简称 DSA）要求互联网平台公开算法推荐机制的运作原理，允许用户选择关闭个性化推荐；并强制对用于内容推荐或广告推送的算法开展风险评估与定期审查，特别关注算法歧视、误导信息传播等问题。DSA 体现了欧盟在数字治理领域的强监管思路，被视为继 GDPR 之后，又一项具有全球影响力的数字治理法案，对全球数字治理规则的走向具有引领作用。

这些法律法规共同营造了一个多层次的算法透明度生态，即：无论是人工智能产品本身（EU 人工智能 Act）、人工智能在数据处理中的应用（GDPR），还是人工智能在互联网平台上的运用（DSA），欧盟均设定了信息披露或解释义务，从不同角度提升算法决策的透明度。

欧盟在强化算法透明度时也非常注重引入外部监督机制，以确保透明义务不流于形式。《人工智能法案》提议建立一个欧盟高风险人工智能系统数据库，由欧盟委员会维护。所有投放市场的高风险人工智能系统必须在数据库中注册，上传关键信息供主管机构和公众查询。这相当于一个公开的“算法备案”制度，类似中国要求互联网算法向网信部门备案并公布清单的做法。通过这种登记公示，监管者可以更好地追踪人工智能系统的使用情况，而公众和利益相关方也能获取人工智能系统的基本信息，提高透明度。另外，欧盟还在探讨强制“算法影响评估”（Algorithmic Impact Assessment）的措施。2019 年 4 月，欧盟委员会发布《算法问责与透明治理框架》报告，建议对公共部门使用算法进行强制影响评估。虽然这一建议并未成为立法，但人工智能法案明确要求高风险人工智能系统提供合规证明时包含风险评估和减缓措施说明。各成员国在执行《人工智能法案》时，可能要求开发者提交独立的算法影响评估报告。欧盟注重公众参与算法治理，欧盟多国开展了算法监督实验（Algorithm Watch）项目，鼓励公民报告和监督在线平台的算法行为。这种民间监督在某种程度上促使企业更加重视算法透明度规则。

欧盟算法透明度规则体现了其一贯的权利本位、预防为主的监管理念。GDPR 彻底变革了组织在收集、使用和共享个人数据方面的行为方式。这主要是因为 GDPR 认识到，数字技术的迅猛发展极大地提升了个人数据的收集、使用和

传播的规模、范围与速度，因此需要一个更强有力的法律体系来强化“数据主体”的权利。²⁹通过 GDPR 保障个体知情权，通过《人工智能法案》等硬法规定预先义务，欧盟旨在从源头和过程中为算法系统增加“可见度”，防止出现无人问津的黑箱决策。欧盟模式的优点是体系完整、标准明确，使企业和公众都有章可循。例如，GDPR 下企业要提供哪类解释信息、《人工智能法案》下高风险系统要披露哪些参数，都有较清晰的指引。这在国际上树立了标杆，也推动其他国家思考加强算法透明度立法。

（三）中国算法透明度的法律规则

近年来，中国在算法治理方面动作频频，逐步建立起涵盖个人信息保护、内容管理和行业规范等多层面的制度架构。其中，有关算法透明度的要求散见于不同法律法规和政策文件中，包括《中华人民共和国个人信息保护法》（简称《个人信息保护法》）、《中华人民共和国数据安全法》（简称《数据安全法》）等上位法的一般原则，以及针对特定场景的专项规章，如《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等。这些规范共同奠定了中国算法透明度规则的框架。总体而言，中国采用的是政府主导的综合治理模式，在强调算法服务提供者承担透明义务的同时，更加重视行政监管和伦理引导，与欧美模式有所不同。

《个人信息保护法》是中国数据保护领域的基础性法律，其中第 24 条对自动化决策做出了明确规范。该条规定：“个人信息处理者利用个人信息进行自动化决策，应当保证决策的透明度和结果公平、公正，不得对个人在交易价格等交易条件上实行不合理的差别待遇。”这一规定将“透明度”明确列为使用算法决策的法定义务，旨在防范利用大数据对不同消费者实行价格歧视等行为。例如，根据该条，电商平台若基于用户浏览和消费数据自动调整商品价格，必须确保此决策过程透明、公平，不能对相同条件的消费者因算法而报出不同价（即“大数据杀熟”）。此外，《个人信息保护法》第 24 条第二款还规定：“通过自动化决策方式向个人进行信息推送、商业营销，应当同时提供不针对其个人特征的选项，或者向个人提供便捷的拒绝方式。”这实际上赋予用户选择退出个性化算法推荐的权利，保障了用户在算法面前的自主选择权。这两款规定相辅相成。一方面要求算法决策本身透明、公平，另一方面要求尊重用户意愿，不强制个性化。这些规定的目的都在于矫正算法应用中损害消费者权益的乱象，使算法更“阳光化”。需要指出的是，从《个人信息保护法》第 24 条的规定来看，其从正面规定了个人信息处理者利用个人信息进行自动化决策应当遵守的规则，而没有规定个人信息处理者不当自动化决策行为的法律后果，因此，难以直接依据该条规定认定个人信

29 Edward S. Dove, The EU general data protection regulation: implications for international scientific research in the digital era. *Journal of Law, Medicine & Ethics*, Volume 46 Issue 4, 2018.

息处理者的民事责任。³⁰

在《个人信息保护法》等法律原则指引下，国家网信办等部门颁布了《互联网信息服务算法推荐管理规定》（2022年3月1日施行），专门规范提供算法推荐服务的行为。这是中国首部针对算法推荐场景的全面规章，其中多条涉及透明度义务。该规定第16条确立了“算法知情权”原则，即：“算法推荐服务提供者应当以显著方式告知用户其提供算法推荐服务的情况，并以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制等。”这意味着，无论是新闻信息流、社交平台的内容排序，抑或电商的个性化商品推荐，平台都必须公开其算法的底层逻辑，让用户大致了解“这个推荐是如何形成的”。该规定把用户的算法知情权落到了实处，回应了公众对平台“大数据杀熟”“算法诱导沉迷”等问题的关切。该规定第17条赋予了用户选择权，即：“算法推荐服务提供者应当向用户提供不针对其个人特征的选项，或者向用户提供便捷的关闭算法推荐服务的选项。”这与透明度一起，构成了算法推荐场景下对用户权利的双重保障，即：用户既有权了解算法，又有权说“不用”。2019年开始实施的《中华人民共和国电子商务法》规定了算法自动化决策的自然结果展示义务和自动化决策的透明度要求。³¹但是，《电子商务法》过于细致具体的规范很有可能使得平台商家规避掉算法的监管审查，平台商家则可以设置相对隐蔽的自然搜索结果入口等方式使得《电子商务法》规定的算法归责丧失存在基础。³²值得一提的是，算法推荐规定还鼓励平台优化算法规则的透明度和可解释性，表明监管部门希望平台不仅在制度上履行告知义务，还能持续改进技术手段提升算法输出的可理解性。

面对迅速发展的深度伪造和生成式人工智能技术可能带来的负面影响，中国监管机构也出台了相应规则，强调透明度和标识义务。《互联网信息服务深度合成管理规定》（2022年1月发布）要求，对于可能导致公众混淆或误认的深度合成服务（如换脸视频、人工智能合成主播），服务提供者应当对生成的内容进行显著标识。该规定第17条规定：深度合成服务提供者提供深度合成服务，可能导致公众混淆或者误认的，“应当在生成或者编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况”。这相当于强制透明——任何人工智能合成的影像、音频，只要存在以假乱真的风险，就必须让观众一眼看出“这是人工智能制作的”。随后在2023年，网信办等四部门又发布了《生成式人工智能服务管理暂行办法》，适用于ChatGPT类文本生成以及人工智能绘画等服务。其

30 王叶刚：《个人信息处理者算法自动化决策致害的民事责任——以〈个人信息保护法〉第24条为中心》，载《中国人民大学学报》2022年第6期。

31 《电子商务法》第18条规定：“电子商务经营者根据消费者的兴趣爱好、消费习惯等特征向其提供商品或者服务的搜索结果的，应当同时向该消费者提供不针对其个人特征的选项，尊重和公平保护消费者合法权益。”

32 丁国峰、寿晓明：《生成式人工智能算法的法律风险及其规范化防控》，载《云南大学学报(社会科学版)》2024年第3期。

中规定，生成式人工智能生成的内容如包含虚假信息或肖像合成，应当以显著方式标明生成内容身份，防止公众混淆真实人物或信息。2025 年 3 月，国家网信办等制定了《人工智能生成合成内容标识办法》。该办法第 4 条要求对生成合成内容添加显式标识，“服务提供者提供生成合成内容下载、复制、导出等功能时，应当确保文件中含有满足要求的显式标识”。以上这些规定，目的都是为了提升人工智能生成内容的透明度，让用户清楚地知道自己看到的、听到的是否由算法生成，从而维护基本的信息真实性和社会信任。

中国还独具特色地实行了算法备案和审查制度，作为增强透明度和加强监管的治理工具。根据网信办发布的公告，自 2022 年起对提供算法推荐服务的互联网信息服务算法实行分批备案制度。平台需要向网信部门提交算法详情（包括算法名称、提供者、主要功能、适用范围等），由主管机关审核后列入算法备案清单并定期公布。这个备案清单相当于向社会公开了国内主要互联网公司的算法“目录”，在一定程度上提升了算法生态的透明度，让公众至少知道哪些算法在被使用及其大致用途。更重要的是，备案为监管部门进行算法评估和检查创造了条件。网信办可以据此对算法进行技术检测、安全评估，重点检查其是否存在危害国家安全、影响社会稳定或侵害用户权益的风险。一旦发现问题算法，监管部门可要求整改甚至下架。可以说，算法备案制度体现了中国治理模式中“强监管”的一面，即：不仅要求事前报备，而且通过行政手段介入算法运行过程，从监管层面对透明度进行保障。这种做法在西方国家较为少见，但中国认为其必要性在于本国互联网平台影响巨大，算法如果不加管制极易引发社会问题。通过备案，政府对算法持有“可视”和“可控”的手段，从而实现对算法透明度更深入的掌握。

中国还非常强调在算法设计和应用中贯彻伦理原则和主流价值观。无论是 2019 年的《新一代人工智能治理原则》还是 2022 年的《关于加强算法综合治理的指导意见》，均倡导算法开发应遵循“公平、公正、透明、可控”等原则，避免算法滥用和歧视。例如，对于招聘、教育等涉及公民基本权益的算法应用，政策层面要求防止算法因偏见而对特定群体产生不公待遇，强调技术向善和促进社会公平。这些伦理要求在法律中有所体现，例如算法推荐规定强调算法应用应“向上向善”，遵守公序良俗和商业道德。在透明度方面，伦理导向促使企业不仅从合规角度被动披露信息，更主动提升算法可解释性和对用户的友好度，把保障用户理解和选择作为企业社会责任的一部分。可以说，中国的算法透明度规则植根于本国文化和治理模式，更加强调社会责任和国家安全维度下的透明可控。我国对于人工智能治理倾向于区分不同技术方向进行规制，监管范围有限，相关管理规范的出台往往是为了应时解决某一热点问题，立法上呈现出分散化趋势。³³因而，

³³ 刘 辉、雷崎山：《生成式人工智能的数据风险及其法律规制》，载《重庆邮电大学学报（社会科学版）》2024 年第 4 期。

我国在人工智能治理领域的法律法规还有许多需要进一步完善的地方。

（四）不同治理模式法律规则的比较分析

通过以上对美国、欧盟、中国算法透明度规制的考察，可以发现三者治理模式和具体规则上各有侧重，反映出不同法治传统和政策目标下的路径差异。

第一，法规体系的完备程度有较大差异。欧盟采取“统一立法+全面覆盖”的模式，透过 GDPR 和《人工智能法案》等形成了较完备的法律框架，将算法透明度纳入法定义务体系，从个人数据处理到高风险人工智能部署均有明确规制。欧盟《通用数据保护条例》下的“充分性认定”机制，本质上是欧盟传统的“等效监管”理念的体现，即欧盟经过评估，单方面认定其他国家的相关监管达到欧盟水平或者说同欧盟监管“等效”。³⁴因而，其对其他国家能够形成压力，并产生示范效应。相比之下，美国迄今为止缺乏专门立法，更倾向于“碎片化监管”，依赖现有法律在具体领域发挥作用，辅以指导性政策和行业自律，其规则散见于各部门法和州法，缺乏统一协调。这体现了美国立法对新兴技术抱持谨慎态度，不急于制定包罗万象的新法，而是让市场与技术先发展，再针对问题调整法规。

中国则介于两者之间，目前没有专门的人工智能基本法，但通过《个人信息保护法》确立了透明、公开、公平的基本要求，并辅以部门规章对重点场景详加规定。中国个人信息保护规范一定程度服务于总体国家安全观下的网络安全与数据安全战略，并兼顾数字经济发展目标，相较于欧盟模式，中国个人信息保护基础显得较为薄弱。³⁵中国模式可称为“分领域规章+行政规范”，由主管机关快速出台针对性规定，实现灵活治理，但也意味着法律层级较低、体系性稍逊，需要不断调整完善。

第二，监管方式与主动程度存在不同。在保障算法透明度上，欧盟强调事前和过程中的监管。通过强制合规评估、备案和公开数据库等手段，使高风险算法在进入市场前即被审查登记，使用过程中也持续接受监督。这种“预防为主”的思路减少了算法黑箱可能带来的危害，有很强的主动性。美国更侧重事后问责，即通过法律责任和司法救济促使企业在出现问题后提供解释或改进，平时则主要依靠企业自律和市场声誉机制来推动透明，监管部门通常在出现明显损害或欺诈时才介入调查。在这种模式下，透明度的实现比较依赖企业意愿，缺乏欧盟那样的强制力。尽管美国《算法问责法案》（Algorithmic Accountability Act）与欧盟《人工智能法案》（Artificial Intelligence Act）都未将算法透明度明确界定为一项基本权利，但显然它们都致力于实现有效的算法透明度，并在智能系统的实施中发挥着引领作用，这从本质上又回到了问责制的理念上来。两者都展示了如何通过监管模式实现智能系统的问责机制。这两项法案可以成为全球范围内建立更具算法

34 廖凡：《数字主权与全球数字治理》，载《暨南学报(哲学社会科学版)》2024年第7期。

35 许皖秀、王硕、左晓栋：《欧盟数据保护认证制度与中国镜鉴》，《信息安全研究》2025年第1期。

透明度的借鉴，它们授权使用透明、可审计且符合伦理准则的智能系统。³⁶

中国走的是事前审查与持续监管相结合的路线。一方面通过算法备案等进行事前介入，要求算法可管可控；另一方面又要求平台建立定期评估、用户申诉等机制，强化事中事后治理。可以说，结合我国的个人信息保护目标与监管机构的实际情况，现行法所选择的是以多部门分散监管为基础、国家网信部门统筹协调的混合式监管机构方案。³⁷这与美国强调市场自治形成鲜明对比，也与欧盟主要依靠法律义务驱动不同。中国模式体现了行政监管的强度，但同时也考验政府部门的技术监管能力。可以说，欧盟推动人工智能监管标准严格化，美国推动标准务实化，中国则可能在标准制定中寻求发挥更大作用，以结合安全可控需求提出新的实现路径。

第三，信息披露的范围与层次不一样。欧盟和中国均在探索分层次的信息披露机制。欧盟人工智能法案要求提供给监管者的技术文档详尽，而提供给公众和用户的信息则相对概要易懂。中国算法备案制度也是内部报备详细数据、对外公布简要清单。三方都意识到不同受众对算法信息的需求不同。监管者需要深入资料，公众和个人需要清晰解释而非技术细节。因此都倾向于有限透明+受控透明，即在确保必要知情的同时，不要求完全公开源代码等敏感信息，以平衡透明度和商业秘密。区别在于，欧盟和中国已将这种分层披露付诸实践（前者通过法律强制，后者通过行政管理实现），而美国因立法未定而主要依靠自愿和诉讼中的信息披露。

这三种模式对算法可解释性的技术难度认知影响了信息披露的范围与层次。欧盟立法者尽管知道深度学习完全解释很难，但仍选择高标准，倒逼业界开发可解释人工智能或提供替代性解释，体现了一种原则性坚持。美国业界和学界相对更坦承技术局限，常有观点称“完全解释算法是不现实且无意义的”，因此更强调实用主义——提供用户真正需要的有限信息即可，无需强求彻底透明。这反映在其立法提案中，就是用分层披露和内部审计代替全面公开。中国则采取渐进策略，目前要求的平台解释多为粗粒度信息（如原理、机制概要），并未触及人工智能模型的复杂细节。但中国也投入大量科研力量研究可解释人工智能，制定技术标准，希望长期来看能提升算法可解释性能，从而降低监管和合规难度。

第四，强制性 vs 软性规范表现不同。欧盟和中国更倾向于强制性法律义务保障透明度。欧盟直接规定算法解释义务，违反将面临处罚；中国以规章形式要求算法公示原理、标识生成内容等，如不遵守可能被行政处罚。相较而言，美国至今大部分有关透明度的要求仍是软性规范（软法），即：行业准则、政策宣言，

36 Alana Engelmann, Algorithmic transparency as a fundamental right in the democratic rule of law: a comparative approach to regulation in European, North American, and Brazilian Contexts. *Brazilian Journal of Law, Technology and Innovation*, Vol. 1 No. 2, 2023.

37 张梦蝶：《个人信息保护监管机构的运行逻辑及其完善》，载《南京大学学报（哲学·人文科学·社会科学）》2024年第5期。

以及企业自行制定的人工智能伦理守则等。虽然美国也有 FCRA 等硬法涉及算法通知义务，但总体覆盖有限，更多是鼓励而非强制。这种差异跟法律文化有较大关系。美国更信奉市场机制纠偏，相信企业出于信誉会提供一定透明度，而欧盟、中国更强调政府责任，要用法律确保企业到位。软法的优点是灵活性和适应性强，但可能执行不力；强法的优势在于可执行性，但可能带来成本和死板。各国根据自身情况取舍：欧盟因为多起隐私丑闻和公众要求而选择硬法强制；美国担心过度管制扼杀创新，倾向软性引导；中国则因为互联网平台影响广泛且集中，更偏好行政规制确保万无一失。

算法透明度规制背后也体现了不同地区的核心价值观。欧盟受欧洲人权、公民自由传统影响，非常重视个人自主权和尊严，因此其透明度规则突出对个人知情权和控制权的保障。公众在欧盟有较高期望知道自己数据如何被算法使用，这驱使立法者将透明度作为维护个人权利的关键手段。美国则兼具个人主义和市场自由色彩，一方面反对算法歧视以维护平等（源自民权价值），另一方面又崇尚商业自由以推动创新。因此其规则在反歧视、告知义务上有一定体现，但整体基调是宽松监管、保护创新。如果说欧洲人说的“隐私权”偏重从人性尊严出发来保护个人的积极自由，那么，美国人说的“隐私权”比较偏重限制国家权力和保护个人消极自由。³⁸中国的文化则更强调集体主义和社会和谐，中国法规不仅关注个人用户权益，也非常注重算法对社会整体秩序的影响，强调以伦理规范引导算法避免不公，维护社会公平正义。例如，中国要求算法传播“正能量”、不得扰乱舆论，这些内容在欧美法规中较少涉及。可以说，欧盟模式是权利导向，美国模式是效率导向，中国模式是秩序导向。这导致三者 in 透明度上的优先目标有所不同，即：欧盟优先保障个人在数字社会的尊严，宁可增加企业负担；中国优先确保算法不危害社会稳定和公众利益，由国家来掌控关键环节；美国则优先促成技术发展和市场繁荣，政府干预有限度。

综合比较可以发现，美国、欧盟、中国在算法透明度治理上各有所长，也各有局限。欧盟模式保障充分但实施成本高、速度较慢；美国模式灵活高效但规范力度弱、存在滞后性；中国模式全面主动但可能对技术自由度有所限制。随着人工智能全球治理深入，不同模式之间也在相互借鉴融合。例如，欧盟立法者关注到美国关于商业秘密保护的考虑，尝试在《人工智能法案》中加入分级披露条款平衡透明度冲突。美国企业为了进入欧盟市场，不得不提升自身透明标准以符合 GDPR 和《人工智能法案》的要求，这实际上把欧盟模式间接推广到了美国企业行为中。中国则一方面学习欧盟数据立法经验强化个人权利保护，另一方面通过参与 ISO 等国际标准制定，将自身“算法备案”“分级分类监管”等理念传递给国际社会。

38 朱晶晶：《个人信息保护路径的比较研究》，《信息安全研究》2023 年第 2 期。

四、人工智能算法透明度法律规制存在的问题及原因

尽管世界各国和地区在强化算法透明度规制方面做出了诸多努力，但现阶段的法律实践仍存在一些共性的问题和挑战。这些问题既有技术层面的客观困难，也有法律制度设计和执行上的掣肘因素。

（一）人工智能算法透明度国际标准不统一导致合规难度较大

尽管国际层面关于算法透明度原则的规定较多，但缺乏具有法律约束力的统一规则，全球人工智能治理呈现“碎片化”局面。OECD、UNESCO 等框架偏软性，不能强制执行。各国基于自身价值和利益制定标准，有时出现冲突。这样，由于各国对算法透明度要求存在差异，跨国企业在不同司法管辖区需要满足不同标准，就增加了合规复杂性。欧盟要求向欧盟用户解释算法逻辑，美国则暂无同类统一要求；中国要求算法备案和正能量导向，这是其他国家所没有的。对于跨国运营的平台，例如 TikTok、Facebook 等，需要针对用户所在地法律分别调整算法透明措施，这既增加了运营成本，也可能带来法律冲突。如果一个国家要求披露的信息在另一国家被视为商业秘密受保护，企业将陷入两难。此外，国际规则差异也容易被一些企业利用，通过将数据或算法部署在监管宽松的地区来规避严格法规，形成监管套利。长远看，这种碎片化不利于构建统一的数字市场秩序，也使得用户权利保护水准参差不齐。一些国际组织已经意识到这一问题，试图推动各国在人工智能治理原则上达成一致，包括透明度在内的若干核心理念。但原则一致不等于细节统一，各国立法基于本国国情，短期内完全协调难度很高。

由于缺乏统一协调的全球监管框架与协调机制，各国各行其是，相关标准和执行力度参差不齐。各主要国家（美、欧、中等）在算法透明要求上不尽相同，导致跨国公司需要遵守不同规则。同时，一些国家可能成为监管洼地，吸引公司将算法决策流程转移到监管宽松地区，规避严格透明义务。这种碎片化使得全球范围的个人权利保障不均衡。公民在严格监管地区享有较强透明和解释权，而在监管弱的地区可能完全处于黑箱算法支配。例如，Facebook 的推荐算法在欧盟需遵守 DSA 的透明义务并接受审计，在其他地区可能没有同等约束。这不仅带来伦理问题，也形成了事实上的不公平竞争。遵守高标准的企业成本较高，反而可能在全球市场中处于劣势。

国际标准不一的原因一方面是政策价值取向不同，另一方面也是国家之间竞争的体现——一些国家可能希望通过比较宽松的人工智能治理规则吸引人工智能产业集聚。要减少这种差异对透明度合规的负面影响，就需要加强跨国合作和互认，可以在双边或多边框架下承认对方的算法评估认证结果，避免企业重复评估；也可以推动制定国际标准（ISO/IEC）来统一某些透明度技术要求，让企业开发说明报告时可以一稿多用，以满足多地监管。这些都是减少合规成本、提升透明度治理效率的可能路径。同时，在数字时代，亟需从伦理上更为深入、精细

地理解透明性所带来的复杂乃至问题性影响，尤其是其对“组织性（未）成熟”状态的重塑作用。³⁹如何将透明度与问责、审核等机制有机结合，形成立体的算法治理体系，是当前探索的方向。如果孤立地看透明度，其作用可能有限甚至被规避，但与其他治理工具协同发力，则能有效促进算法合规和可信。反之，对透明度作用的误判和过高期待，也可能导致资源错配和忽视其他重要措施。这方面的问题反映了政策认识上的不完善，需要在实践中不断总结调整，找到最优的治理组合。

（二）人工智能技术复杂性导致算法透明度的实现存在很多困难

当前，许多人工智能模型内部结构高度复杂、非线性叠加，连开发者自己有时也难以完全解释其决策依据。算法的不透明性可能源自其结构的复杂性，例如深度神经网络，其由成千上万个人工神经元组成，以分布式方式协同解决问题。这种导致人工智能成为黑箱的原因被称为“复杂性”（complexity）。⁴⁰法律要求算法提供“可解释”的信息，在技术上面临巨大挑战。即使强如欧盟的 GDPR 和《人工智能法案》，也不得不接受一种折衷——提供有限的、有意义的信息，而非彻底打开黑箱。在实践中，即便算法开发者竭力配合，能提供的往往只是近似解释或代理模型的输出，这可能与算法真实机制存在偏差，这种技术鸿沟导致透明度要求难以落实到位。例如，一家使用深度神经网络进行信用评分的公司，按照法规为用户生成了一份原因说明，列出影响评分的关键因素。但这份说明可能只是基于输入变量权重的简单分析，未必真正揭示了网络内部上千神经元互动所产生的决策逻辑。因此，当前法律上透明度要求在某种程度上理想化了算法的可解释能力。

由于算法技术的复杂性，即使个人有权访问算法系统的源代码或数据库，这类措施并不一定具有可行性，并可能加大了技术精英与社会大众之间的认知鸿沟。⁴¹当透明度成为法定义务但技术上无法满足时，企业可能选择提供流于形式的解释以“合规”，导致透明度名存实亡。这一问题的根源在于人工智能技术本身的黑箱属性尚未突破，法律规范与技术实现之间存在冲突。但是，算法透明制度非常必要，并非完全没有可行性，其呈现出层次分明的价值内核与技术架构，可以被设计成由柔性规范、中性规范和刚性规范组成的多层次体系。⁴²一些企业担心过多公开算法细节可能泄露商业秘密或被竞争对手“借鉴”，监管部门可能只好做出让步，强调其所要求披露的是“基本原理和机制”，并非具体代码或模型参数，尽

39 Frederik Schade, Dark Sides of Data Transparency: Organized Immaturity After GDPR? Business Ethics Quarterly, Volume 33 Issue 3, 2023.

40 Yavar Bathaee, “The Artificial Intelligence Black Box and The Futility of Intent and Causation”, Harvard Journal of Law & Technology, Volume 31, Number 2 Spring 2018.

41 张宇晖：《算法决策的结构化问责机制：欧盟经验与中国借鉴》，载《上海政法学院学报（法治论丛）》2022年第4期。

42 安晋城：《算法透明层次论》，载《法学研究》2023年第2期。

量避免损害企业核心利益。同时，备案信息公开的范围有限，一般只公布算法名称和功能描述等，不涉及敏感技术细节。关于人工智能可解释性问题，就算企业愿意解释算法，如何用普通人能懂的语言解释深度学习模型决策逻辑，仍是全球性难题。可以说，人工智能技术的复杂性是当前透明度规制面临的首要难题，也是造成人工智能规则执行难的重要原因。

在实际审核中，很难准确界定算法透明度是否足够清晰，这往往成为企业应付监管的灰色地带。造成这种现象的原因，一方面是监管标准的不明确，企业倾向于形式合规多于实质效果；另一方面，不同用户对信息的理解力不同。平台企业自行审查成本巨大，在规制成本较高的同时充分履行网络信息安全义务可行性不高。⁴³在算法透明度和可解释性视域下，平台企业、政府和个体在对应私权力、公权力与私权利相互制约的同时，也映射出商业利益、公共利益和个人权益三者间所需的利益平衡。利益平衡既需要完备的制度体系构建，更需要在实践中寻求利益间的目标一体、成效相关。⁴⁴解决该矛盾需要监管者进一步细化指南、提供示范模板，并鼓励采用用户调研等方法改进解释质量，以提高透明度的实用性。总之，国际社会需要在透明与安全、透明与发展之间寻找平衡点，不断完善人工智能算法透明度规则。

提升算法透明度不仅对企业是挑战，对监管机关同样提出了高要求。监管者需要具备充分的技术能力去审查企业提交的算法文档和评估报告，否则透明义务再详细也难以发现问题。一些国家监管机构目前并不具备足够的人工智能专家人力来处理海量算法资料。中国网信部门虽建立算法备案系统，但如何真正读懂每家提交的算法说明、如何发现隐含风险，也需要投入相当多的人力与资金。监管能力不足会导致“监管真空”或“选择执法”。法律上要求很多，但实际只重点关注极少数大企业、大事件，就会出现多数算法处于放任状态，透明义务形同虚设的现象。由于复杂计算系统的特殊性，当前以算法透明性和影响评估为主的机制，往往难以形成强有力的责任追究与救济体系。⁴⁵在某些情况下，增强算法问责（Accountability）和外部审计或许比提高透明度更有效。譬如，可以要求企业其定期进行独立公正的性能测试和偏差检测，将其视为透明度机制的补充或替代。对于中小企业而言，准备算法影响评估、履行解释义务需要投入的人力物力相对更高，可能形成进入壁垒，不利于创新竞争。因此，一些法规对中小企业豁免了部分要求，但这又削弱了规则普适性。监管部门在执法时也会考虑成本效益，对危害不大的违规可能睁一只眼闭一只眼，以集中资源应对重大风险。这种选择性执行虽是现实无奈，却会损害法律权威，并可能引发合规不公平。

43 张旭芳：《生成式人工智能的算法安全风险及治理路径》，载《江西社会科学》2024年第8期。

44 孙晋、顾瑞琪：《基于算法透明度和可解释性优化的算法治理制度构建》，载《数字法治》，2023年第2期。

45 Burcu Baykurt, Algorithmic accountability in U.S. cities: Transparency, impact, and political economy, Big Data & Society, First published online August 7, 2022.

（三）算法透明度规则与保护商业秘密之间存在明显冲突

许多人工智能系统（尤其是涉及机器学习、深度学习的系统）在模型架构、训练数据、权重参数和推理逻辑等方面具有高度的技术复杂性和原创性，这些通常是企业的核心竞争力。如果法律要求公开算法的结构、逻辑或数据来源，就可能暴露企业的关键技术细节，造成商业秘密泄露。随着人工智能的兴起，算法通过对离散数据集进行预测性分析，使得大规模数据的前所未有的收集、处理与分析成为可能。然而，人工智能的识别能力模糊了隐私边界，使其能够汲取包括健康、情绪、兴趣和行为在内的敏感信息。⁴⁶一般来说，法律要求通常比较原则，企业在执行时可能出于谨慎披露过多信息来免责，结果就会超出用户的承受能力。例如：某银行响应监管要求向拒贷客户提供算法评分依据，结果给出了一长串参数和权重说明，普通客户根本无法理解。这样，其表面看满足了算法透明度义务，实则没有达到让用户“真正明白”的初衷。一些监管机构也意识到这点，强调解释应当“清晰、及时、可访问”，使用浅显语言而非技术术语。算法的信息披露不仅可以确立某种事实上的行为规范，还可以通过构建一种“科学叙事”来为这些规范赋予合理性和合法性。在此过程中，人们将被激发出一种内在动机，自觉遵守这些规范，而较少进行批判性思考。透明度也无法避免所有风险，比如完全公开的算法仍可能被恶意操纵者利用。

算法往往是企业核心竞争力所在，其详细模型、代码和训练数据构成商业秘密或受知识产权保护。为治理算法和保护算法相对人的合法权益，以公开算法代码等信息为手段的算法透明措施，在算法治理中扮演着日益重要的角色，但是，强制披露算法信息可能直接与公司商业秘密和知识产权保护之间产生冲突，引发企业抗拒，由于算法与机器学习等技术的复杂性、信息披露对象的技术水平、商业秘密制度等阻碍因素，算法透明的效能会受到减损。⁴⁷例如，在前述美国 Loomis 案中，算法公司就以商业秘密为由拒绝公开模型细节。在欧盟《人工智能法案》讨论过程中，不少企业和业内团体也游说减少公开要求，担心欧盟市场的严格透明义务会让他们泄露机密、失去全球竞争优势。透明度要求与商业秘密保护的冲突已成为立法绕不过的难题。欧盟和美国试图通过限制公开范围来缓冲这一矛盾，如美国提案中仅要求向监管机构提供详尽资料，向公众提供概要，以避免商业机密外泄。

算法透明度的目的在于促进问责、公平与信任，但如果透明度要求过度，就可能侵犯企业的商业秘密和技术权益，从而抑制创新。中国在算法备案信息公开上也注意保护企业敏感信息，仅要求公布概要清单，不对外透露具体算法逻辑。同样，对源代码的检查通常在政府内部进行，并签署严格保密协议。即便如此，

46 Hitanshi Goel, Gyandeep Chaudhary, “Securing the Digital Footprints of Minors: Privacy Implications of 人工智能”. *Balkan Social Science Review*, Vol. 23 No. 23 (2024).

47 魏远山：《算法透明的迷失与回归：功能定位与实现路径》，载《北方法学》2021 年第 1 期。

一些企业仍担心“开了口子”后监管要求会越来越严。欧盟《人工智能法案》未来的实施细则可能扩展披露要求范围，这会让企业进退两难。这种冲突背后的原因在于法律价值平衡，即：透明度侧重公共利益和个体权利，商业秘密侧重私有财产权益。当两者发生碰撞时，不同法域的取舍有所不同。美国更倾向保护商业利益，因此迟迟未立强制透明法；欧盟优先公共利益，愿意牺牲部分企业便利；中国则试图通过行政掌控实现“两头兼顾”，但长远看也面临取舍抉择。解决这对冲突，需要法律在制度设计时提供豁免和补偿机制。如允许企业对特别敏感的信息申请豁免公开，或引入闭门审核机制替代公开，同时对于强制披露导致企业利益受损的，探索相应的补偿措施。这将在未来国际规则协调中成为焦点问题。

（四）人工智能算法治理主体不明确

当算法决策出现问题（如偏见歧视、错误伤害）时，谁应承担法律责任很容易陷入灰色地带。传统法律中的责任主体认定机制在高度自动化环境下遭遇挑战，出现了责任主体不明的困境。“责任主体不明”反映出现行法律分配责任的方法尚未完全适应人工智能自主性和复杂性的挑战。为避免“无人负责”的真空，各国开始通过立法和司法解释明确开发者、部署者对算法结果承担连带或各自责任的规则。责任主体问题不解决，算法透明度规则的意义就会大打折扣，因为算法透明度规则的目的之一正是厘清责任、追究问责。

算法涉及多个参与方，包括：开发算法的供应商、部署算法的服务提供者、使用算法决策的机构或个人，以及算法本身。当发生损害时，各方可能推诿责任。开发者可能称算法交付时是中立工具，责任不当在使用者；使用者则辩称算法是第三方开发的黑箱，其自身也不完全了解内部机制；而人工智能本身无法律人格，无法承担责任。例如，自动驾驶汽车撞人，责任在制造商（开发人工智能驾驶算法）？车主（使用者）？抑或人工智能系统作为“电子驾驶员”？目前法律主流仍将责任归于人，但确定哪一方人可能复杂取决于过错和合同。在产品责任法框架下，若算法被视为产品部件，制造商对因其缺陷导致的损害承担无过错责任；但算法缺陷如何界定本身有争议，且人工智能具有自学习能力，初始无缺陷但演化中出错责任如何归属不明确。过失责任框架下，需要证明开发者或使用者存在过失（如未尽到训练、测试、监控义务），这在复杂人工智能环境下举证十分困难。随着人工智能技术（如强化学习、自主代理）的发展，算法决策可能超出开发者初始设定范围。在无人机自主作战、算法高频交易等场景下，人工智能做出的具体决策在质变后难以归因于某个人的直接指令。这种高度自主性引发了所谓“人工智能责任真空”——人工智能自身造成损害但又无法承担法律义务。

许多人工智能服务通过 API 或云服务提供，由第三方算法支持本地决策。如银行使用外包的算法模型决定贷款，假如决策不公，客户起诉银行，银行可追偿算法供应商。法律关系上类似产品供应链责任，但算法的不确定行为增加了责任

分摊难度。欧盟在人工智能法案中明确区分“提供者”（开发并将人工智能系统投放市场的一方）和“使用者”（将人工智能系统应用于业务的一方），分别规定其合规义务和责任。2022年9月，欧盟委员会提出《人工智能责任指令》（提案）（AI Liability Directive，简称 AILD）。其与《人工智能法案》和其他一系列欧盟数字领域立法共同构成欧盟人工智能的法律框架。AILD 旨在协调欧盟成员国人工智能系统致损的部分非合同性民事法律制度，通过引入证据披露机制与过错责任下的因果关系推定，降低请求权人可能过高的索赔成本。

算法服务常是跨国提供，责任主体可能位于异国，增加追责难度。比如一款美国公司开发的人工智能医疗诊断软件在中国医院误诊致人损害，受害人面对的责任主体可能在境外，如何适用法律和执行判决成为问题。目前，国际私法对此尚无特别规则，如果按一般侵权或产品责任原则处理，也需要各国法律在责任认定上相容，否则可能出现管辖和法律冲突。如果美国法律豁免了某类算法责任，而中国患者在国内起诉又适用中国法，则判决结果可能无法执行到美国主体上，形成救济落空。这类跨境责任界定和执行问题，需要国际层面通过协定或统一规则来解决。

总的来说，算法透明度规制尚处在不断摸索完善阶段，上述问题的存在并不削弱透明度作为目标的重要性，而是提示我们应以更理性和全面的视角来看待透明度要求。在技术进步、法律完善和国际合作的推动下，这些问题有望逐步缓解。但在此过程中，理解问题背后的原因并对症下药，才能使算法透明度规制真正取得实效。

五、中国应对人工智能算法透明度国际规则的建议

当前，全球围绕人工智能伦理与治理的国际规则博弈日趋激烈。算法透明度作为人工智能治理的关键议题之一，也是各国角力和合作的焦点。在这一背景下，中国要积极参与国际规则的制定，明确自身立场，有针对性的谋划应对之策。

（一）完善国内法律法规体系和标准体系，确保与国际高标准规则基本同步

面对欧盟等推出高标准人工智能治理规则，中国需要与时俱进地完善国内法律，避免因标准落差过大而处于被动。中国应加快制定综合性人工智能立法，将分散于不同法规中的算法透明度要求加以汇总提升，通过高位立法确立算法透明度基本制度，让各部门执行有据可循，也向国际社会展示中国人工智能治理的中国方案。同时，可以考虑在人工智能立法或现有法律修订中，充实算法透明度条款。例如，在《个人信息保护法》实践基础上，明确赋予个人对重要自动化决策的解释权，规定个人有权要求算法决策的提供者给予适当说明。在人工智能专项立法中，对高风险算法系统设定更具体的透明义务，包括向监管部门报送算法说明、向用户提示人工智能身份、重大风险报告等，借鉴欧盟人工智能法案的合理

部分。同时，可以结合中国实际，创新性提出一些主张，例如：要求算法提供者对其透明度承诺进行公开声明，接受社会监督；探索建立算法可信认证制度，由独立机构评估算法透明度和可信度，给予认证标识，引导市场优选透明度高的产品。这不仅是保护国内用户权益的需要，也将在国际谈判中为中国争取主动——当我国标准达到一定高度，就有更大底气倡导国际社会接受一种均衡方案，而非被迫全面采纳他方标准。

中国要投入资金与力量加强可解释人工智能等相关技术的研发，努力突破人工智能黑箱瓶颈，重点解决技术复杂性导致透明度实现困难问题，过度信息披露与实用性之间的矛盾问题，商业秘密与知识产权的法律冲突问题，以及监管能力与执行成本的问题。鼓励产学研合作，开发实用的模型可解释工具、算法审计系统、透明度标签自动生成等技术，将透明度从人工义务转化为部分自动化实现。中国要积极参与 ISO、IEC 等国际标准组织关于人工智能透明度和治理的标准研制。同时，争取由中国牵头制定一些透明度技术标准或指南，将中国在算法标识、分级管理等方面的经验上升为国际标准。例如，可推动制定“人工智能生成内容标识”国际标准，目前我国已发布相关国家标准草案，可在此基础上向 ISO 提交提案，引领此议题的标准方向。再如，可倡导在国际标准中加入对算法备案、评估方法的规范，为我国的监管模式获取国际认可提供依据。

（二）积极参与国际规则制定，倡导公平公正互利共赢的治理原则

中国应当从被动适应规则转向主动塑造规则，在协调与制定人工智能算法透明度国际规则领域发挥作用。中国要在联合国、WTO、RCEP 等多边框架下，积极贡献中国智慧和方案，参与制定全球人工智能治理原则和标准。要支持联合国教科文组织《人工智能伦理问题建议书》的落实，在关于透明度、可解释性等原则上与各国保持共识。在联合国拟议的数字合作机制中，中国可倡议设立算法透明度专题工作组，讨论制定各国普遍适用的指导意见和评价指标。通过深度参与规则磋商，中国可以在透明度问题上发挥桥梁作用。一方面坚持以人为本、公平正义的总体方向，回应欧盟等对保护个人权利的关切；另一方面也要强调尊重各国数字主权，主张各国平等参与，反对个别国家将本国标准单方面强加到国际规则中。中国要确保国际规则的制定兼顾不同发展水平国家的利益，推动形成既有原则高度又可操作落地的透明度国际准则，避免出现少数西方国家垄断话语、制定过高标准而使发展中国家受损的局面。

中国可通过双边协定或区域合作机制，提前布局算法透明度规则的互认，减少跨境合规障碍。要继续在中欧数字对话中就人工智能治理交换意见，寻找监管互认或协调空间。中欧可以在数据和数字贸易谈判中讨论建立算法透明度合作备忘录，探讨中国企业如何满足欧盟《人工智能法案》要求，以及欧盟企业在华如何遵守中国算法备案和标识规定，寻求等效认可或便利通道。在 RCEP、金砖国

家合作等机制下，提议设立数字治理合作章节，包含算法透明度信息交换和标准对接的内容。这些区域性安排能够为中国企业“走出去”和外国企业“引进来”提供更清晰的预期，防止因标准不一导致的贸易壁垒和法律冲突。这既提高了区域整体治理水平，也增强了中国规则的影响力。当然，在推行中国经验时需要尊重他国国情，注重协商共建，而非强制输出。通过互利合作实现规则对接，最终可为将来全球层面的规则趋同奠定基础。

（三）坚持数字主权原则，妥善应对域外法域影响

在融入国际规则的同时，中国也要明确自身红线，维护国家数字主权。数字主权表现为数字技术主导权、数字规则制定权、数字空间话语权以及数字经济发展权。⁴⁸面对一些域外法可能具备的长臂管辖效力，中国应评估其对国内企业和法律体系的影响，必要时提出交涉或采取措施。具体而言，对于外国要求中国企业提供某些算法信息，如果可能危及国家安全或公民个人权益，应建立出口审查或报批机制，防止国内数据和敏感信息以合规之名无序外流。例如，要求相关企业在向外国监管披露算法细节前，先向国内主管部门备案，评估安全风险。这类似于对大型互联网企业拟赴国外上市须经网络安全审查的做法，体现维护主权和安全的底线。要积极构建国内法律豁免与保护机制。当中国企业因遵守国际高标准而在国内面临困难时，政府可给予一定政策支持或法律豁免。同时，中国也需防范某些国家借透明度之名进行技术壁垒或情报收集。对明显不合理的外国信息索求，应通过外交或法律途径坚决拒绝，维护自身正当权益。中国在参与国际规则时，一方面开放合作，另一方面寸土必争，确保任何国际义务的承担都在自身可控范围之内，不损害国家核心利益。

中国应着力提高各利益相关方对国际算法透明度规则演进的认识，做好前瞻性准备。政府部门可以定期跟踪研判国际立法动态，及时向国内企业发布合规提示或指南。例如，针对欧盟《人工智能法案》和美国可能通过的新法规，组织行业宣讲和培训，帮助企业理解主要义务和差距，尽早调整产品和服务设计。大型科技企业则应建立全球合规管理体系，监测各国算法监管要求，内部制定统一策略，以最低成本满足多方规则。对于学术界和公众，应加强宣传教育，促使更多人关注人工智能国际治理。支持国内法学界、技术界专家参与国际讨论，在国际刊物、会议上发出中国声音，避免国际话语被少数西方专家垄断。社会舆论层面也应凝聚共识，既认识到透明度保障用户权益的重要性，又理解中国政府在国际谈判中立场的合理性和必要性。只有全社会形成理性务实又积极进取的氛围，中国才能在复杂的国际博弈中立于不败之地。可以预见，未来算法透明度等数字规则将深刻影响全球贸易和技术竞争格局，国内各界未雨绸缪、练好内功，才能既

48 郭江兴、邹宏、张帆、尚玉婷：《数字主权与国家安全若干问题研究》，载《国家安全研究》2023年第3期。

保护自身发展，又赢得国际尊重。

（四）积极确立中国在全球人工智能治理中的引领性地位

中国在全球人工智能算法治理中不能仅仅作为旁观者和规则接受者，而是全程深度参与全球人工智能治理规则的讨论和制定。习近平总书记指出：要“主动参与国际组织数字经济议题谈判，开展双多边数字治理合作，维护和完善多边数字经济治理机制，及时提出中国方案，发出中国声音”⁴⁹。通过积极建言和方案输出，确保国际规则既考虑先进技术国家的责任，也照顾发展中国家的利益，使规则更公正平衡。中国已是人工智能大国，有责任在全球治理中发挥参与和引领作用。通过继续在国内大胆试行新的监管工具和技术手段，不断优化自身算法透明度制度，积累人工智能治理经验，并与国际先进标准接轨。例如，探索人工智能沙盒监管、算法信用评价体系等前沿做法。在确保风险可控的情况下，走出有特色的治理路子。这些创新实践不仅服务国内，也为全球提供范例。当国际缺乏成熟经验时，中国经验就显得尤为宝贵。

中国应为国际规则提供建设性方案，塑造更加包容、公平的全球人工智能治理体系。中国可倡导制定国际指引或公约，主张借鉴气候变化领域《巴黎协定》的模式，推动商定一份具有广泛参与的“人工智能治理国际公约”或“行为准则”。其中，可明确各国在算法透明度方面的最低义务，如确保公民基本知情权、跨境合作调查算法事故等。当前美欧在人工智能规则上有分歧，中国可在某些议题上扮演桥梁角色，既向西方先进理念学习，又向广大发展中国家传递声音，寻找合作交汇点。在透明度问题上，中国可以推动形成东西方对话，平衡个人权利保护和国家安全、社会稳定，将不同关切融入解决方案。例如，把中国注重的内容治理与西方注重的个人解释权结合起来，提炼综合治理模式。这种桥梁作用有助于避免全球规则割裂，促进国际社会领域的协同治理。特别是在与“一带一路”沿线国家的合作中，中国可传播自身算法治理经验（如备案制度、内容标识制度），帮助合作国家完善其规则，从而在区域内逐步形成与中国接轨的算法透明度监管体系。

六、结论

人工智能算法透明度作为人工智能治理领域的重要课题，涉及复杂的技术、法律与伦理问题。透明度既是维护个人知情权、公平权益的必要条件，也是增进公众对人工智能信任、保障技术健康发展的前提。但实现这一目标绝非易事，技术上的黑箱属性、利益上的冲突博弈、规则统筹上的种种难题，都决定了透明度规制将是一个长期博弈的过程。人工智能时代的算法透明度建设任重道远，法律和技术两个轮子需要一起转动，国内治理和国际合作两个层面需要协同推进。国

49 《习近平著作选读（第二卷）》，北京：人民出版社 2023 年版，第 539 页。

际社会应在尊重各国数字主权的基础上寻求更大程度的合作，共同协商制定人工智能算法透明度规则，避免治理规则碎片化而出现各方利益失衡的局面。各国应以人类共同福祉为重，在竞争中寻求合作，在分歧中找到契合点，逐步建立起透明、可信、可问责的全球人工智能治理体系。

中国应对算法透明度国际规则，应秉持“主动参与、平等合作，立足自身、统筹内外”的总体方针。我们既要防范因国际高标准可能带来的压力，又要善于运用我国已有优势参与规则塑造。在不断完善本国算法监管体系的同时，积极融入全球治理网络，发出中国声音，提供中国方案。只有这样，我们才能在新一轮科技变革和制度竞争中掌握主动权，既保障自身数字经济安全，又为全球人工智能治理贡献中国智慧。中国拥有全球最大规模的互联网和人工智能应用市场，在算法治理方面积累了独特经验，这将是我們参与国际规则谈判的底气所在。只要运筹得当，中国完全可以在人工智能治理关键议题上引领方向、谋求共赢，为构建人类命运共同体的数字未来贡献力量。

人工智能军控：缘起、现状与中国路径

刘宏松¹ 邹宇霆²

摘要：人工智能技术是近年来对人类社会最具颠覆力的技术革新，该技术具有很强的军事化潜力，被各国军事部门重点关注并给予大量投资。但是，人工智能军事化的趋势也导致了国际社会对其可能诱发军备竞赛、责任缺失、伦理和技术问题等一系列风险的担忧。截至目前，围绕人工智能军事化的治理主要在各个国家和国家集团内部实施，国际社会尚未形成针对人工智能军事化的多边规制体系，各国之间在军控主体、军控手段、军控目标等问题上仍存在一系列分歧。在建立具有实际约束力的人工智能多边军控体系的现实需求下，国际社会面临着大国间权力竞争、产业发展与风险管控矛盾、法律实际效力不足等挑战。作为人工智能大国和全球治理的重要参与者，中国正在人工智能军控领域积极行动。目前，中国可以通过建立人工智能大国政府间对话机制、坚持发展与安全并重原则，完善多边普惠型人工智能安全治理体系三条路径，深度参与人工智能的国际军控进程。

关键词：人工智能 军备控制 全球治理 中国路径

一、导言

人工智能（Artificial Intelligence, AI）是计算机科学的一个子领域。根据国际标准化组织（ISO）的定义，人工智能指“一个致力于制作能够完成回应、预测、建议或决策等人类行为的工程系统的科学技术领域”。³当前，人工智能技术已经开始凭借其极强的数据处理能力而被广泛地应用于教育、医疗、交通、金融等民用领域，并有效地提高了各个所赋能领域的运作效能。⁴在军事领域，人工智能技术同样表现出强大潜力，以至于西方部分军事战略专家指出，人工智能在军事领域的大规模应用将从根本上改变现代战争的作战模式和制胜机理，成为继火药、核武器之后，军事技术领域中的“第三次革命”。⁵

近年来，包括美国、法国、以色列、欧盟在内的众多国家或国家集团均推出其人工智能军事化战略，其目标在于促进人工智能技术与传统军事技术的结合，

1 上海交通大学国际与公共事务学院教授，13917134211

2 邹宇霆，上海交通大学国际与公共事务学院，硕士研究生，15221682663，1204427488@qq.com

3 International Organization for Standardization, “What is AI? All you need to know about artificial intelligence”, <https://www.iso.org/artificial-intelligence/what-is-ai#toc1>, 访问时间：2024年3月30日。

4 高奇琦：《人工智能、四次工业革命与国际政治经济格局》，载《当代世界与社会主义》2019年第6期，第19页。

5 Reinhold, Thomas, and Niklas Schörnig. *Armament, Arms Control and Artificial Intelligence*. Berlin, Springer, 2022, p.1-3.

提高现有决策、作战和支援系统的运作效能。以以色列为例，在严峻的安全压力和“技术制胜”的国防发展方针影响下，以色列在人工智能军事化方向投入了大量资源。⁶在2023年10月爆发的新一轮巴以冲突中，以色列将人工智能技术大规模地投入侦察、制导、情报分析等实战应用。这一做法在带来显著战果的同时，也引发了一系列新的现实问题：第一，人工智能武器在实战中的优异表现可能会激发地区甚至世界范围内新一轮的军备竞赛；第二，赋予机器以自主选择攻击对象并决定他人生死的能力和权限，可能导致战争责任难辨和战争伦理缺失。⁷第三，人工智能武器一旦以战场损耗等方式流入恐怖组织等非法组织手中，则可能对各国平民安全带来隐患。⁸

人工智能军事化所引发的一系列问题对国际社会提出了相应的治理需求，但是截至目前，人工智能军控领域内并未形成全球性的规制体系。各个国家都在围绕该议题展开治理竞争。围绕人工智能军事化和相关治理议程，学界已存在大量研究成果，并主要集中于两个角度：一是从技术的角度出发讨论人工智能技术在军事领域的应用场景、潜在风险与治理困境。⁹二是从政治的角度出发，讨论国家间围绕人工智能军控所展开的复杂博弈。¹⁰以上研究较为全面地展示了人工智能军控作为一项全球治理热点的特征与现况。但是，作为全球治理的重要参与者和贡献者，中国也在人工智能军控中扮演着重要角色，而现有研究鲜有讨论中国的一系列治理举措是否与现有人工智能军控困境之间存在对应逻辑。在这一背景下，本文将首先通过介绍人工智能在军事化运用中的场景和风险阐明其治理必要性，再通过分析各国围绕人工智能军控中各个争议性问题的不同立场，厘清当前国际社会在该议题上面临的困境所在。最后，本文将分析中国的一系列主张和举措是

6 Yaakov Katz and Amir Bohbot, *The Weapon Wizards: How Israel Became a High-Tech Military Superpower*, New York: St. Martin's Press, 2017, pp. 36-37.

7 Asaro, Peter. "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making." *International Review of the Red Cross*, Vol.94, No.886, 2012, pp. 687-709.

8 Jürgen Altmann, "Arms control for armed uninhabited vehicles: an ethical issue," *Ethics Inf Technol*, Vol.15, No. 3, 2013, p. 137.

9 Reinhold, Thomas, and Niklas Schörnig. *Armament, Arms Control and Artificial Intelligence*. Berlin, Springer, 2022, p.100-116, 174-191, 192-209; 张煌, 杜雁芸:《人工智能军事化发展态势及其安全影响》,载《外交评论(外交学院学报)》2022年第3期,第99-130页;张东冬:《人工智能军事化与全球战略稳定》,载《国际展望》2022年第5期,第142-161页;杨辰:《论国家安全视阈下的人工智能军事应用风险与治理——以俄乌冲突为例》,载《国际论坛》2023年第2期,第61-81页;杨楠:《人工智能武器化与网络威慑战略的未来》,载《国际安全研究》2024年第五期,第21-47页;周意岷,张零:《智能化“混合战争”及其影响和治理》,载《和平与发展》2024年第3期,第106-131页。Erica D. Lonergan, "Emerging technology and the cult of the offensive", *Contemporary Security Policy*, VOL. 45, NO. 3, 2024, p459-493; Osimen, G. U., Newo, O., & Fulani, O. M, "Artificial intelligence and arms control in modern warfare". *Cogent Social Sciences*, Vol.10, NO. 1, 2024, p1-13.

10 徐能武, 龙坤:《联合国CCW框架下致命性自主武器系统军控辩争的焦点与趋势》,载《国际安全研究》2019年第5期,第114-130页;龙坤, 徐能武:《致命性自主武器系统军控——困境、出路和参与策略》,载《国际展望》2020年第2期,第78-102页;王蕾:《历史比较视野下自主武器军控规范争论及其走向》,载《国际安全研究》2024年第五期,第109-134页;崔铮, 尹金灿:《美国对华人工智能发展的认知偏差、威胁放大与政策选择——基于折中主义分析方法》,载《当代亚太》2023年第6期,第132-133页;王彦雨, 李正风, 高芳:《欧美人工智能治理模式比较研究》载《科学学研究》2024年第3期,第460-461页;沈逸, 高瑜:《大国竞争背景下的人工智能安全治理与战略稳定》,载《国际展望》2024年第3期,第33-50页。

否有助于化解上述困境，从而探讨中国在人工智能时代担当大国责任的可行路径和未来前景。

二、人工智能军控的缘起

人工智能军事化涉及一系列概念，如人工智能武器、人工智能武器系统、自主武器系统（LAWS）等，其基本内涵都是指能够对环境进行自主感知和分析，确定目标并实施攻击的武器系统。¹¹在此基础上，本文将统一使用“人工智能武器”概念用于指代人工智能军事化所产生的成果。人工智能本身不具作战能力，但是其能够很好地与现有军事装备和作战系统结合，在很大程度上提高军队在战场中的作战效率，进而为国家的战争能力带来质的提升。在近年来的几场地区冲突中，由人工智能赋能的自主武器系统、智能情报系统开始频繁出现在战场上，展现出高度不对称的战斗能力。在这一背景下，各国的政府都开始认识到人工智能技术在现代军事科技领域的变革性潜力，将其视作提升作战效能、优化战术决策，应对潜在军事挑战的重要抓手。

（一）人工智能技术在军事领域的应用

人工智能在军事领域具有强大应用前景，涵盖了作战武器智能化、战略战术升级、情报分析等多个场景。

第一，人工智能在作战武器智能化方面发挥着关键作用。英国国防部指出，人工智能武器能够对环境进行感知和分析，确定目标并选择攻击。¹²美国国防部在其所制定的人工智能国防战略中，指出人工智能武器能够在不同程度上实现对战场的自主认识和理解，规划并执行攻击任务，甚至直接代替人类成为战争决策和作战的主体，并在结果上表现出相当程度的精确性。¹³在实战中，通过在智能机器人、无人机等武器系统中集成人工智能技术，军队能够利用这些智能化的作战系统实现更高水平的自动化和智能化作战，具体表现在可以同时承担多项复杂任务，进行更加持久的作战，更加高效精准地达成目标等方面。例如，美军目前正在开发机器人战车（RCV）系统，该战车以人工智能技术为核心，可以通过车载的光学和雷达设备实现对军事目标的识别和攻击，并有效规避行进路线上的障碍物。¹⁴此外，美军还在尝试基于人工智能技术开发能够自主航行和交战的无人

11 参见 British Ministry of Defence, The Government Response to the Report by the House of Lords AI in Weapon Systems Committee: 'Proceed with Caution: Artificial Intelligence in Weapon Systems' https://assets.publishing.service.gov.uk/media/65cb77caa7ded0000c79e526/Government_response_to_the_House_of_Lords_AI_in_Weapon_Systems_Committee_Report.pdf 访问时间：2024 年 4 月 20 日；ICRC, "ICRC position on autonomous weapon systems" <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems>, 访问时间：2024 年 4 月 13 日。

12 British Ministry of Defence, The Government Response to the Report by the House of Lords AI in Weapon Systems Committee: 'Proceed with Caution: Artificial Intelligence in Weapon Systems' https://assets.publishing.service.gov.uk/media/65cb77caa7ded0000c79e526/Government_response_to_the_House_of_Lords_AI_in_Weapon_Systems_Committee_Report.pdf 访问时间：2024 年 4 月 20 日。

13 U.S. Department of Defense, "DOD Releases AI Adoption Strategy", 访问时间：2024 年 4 月 20 日。

14 Breaking Defence, "Frustrations mount over Army's Robotic Combat Vehicle autonomy, acquisition

机和无人舰艇。¹⁵通过在战争中使用这些能够自主感知、决策和行为的人工智能武器，军队一方面实现了作战效能的提高，另外也实现了以作战武器损耗替代传统战争中的人员伤亡，进而提高了作战人员的生存率。

第二，人工智能在战略战术升级方面为军队提供了更先进的技术支持。人工智能技术能够为传统的武器系统赋以更加先进的感知、识别、分析能力，提高其响应速度和精准度，从而帮助作战人员设计出更加灵活高效的战略战术。例如，在 2022 年爆发的俄乌冲突中，乌克兰在战争初期通过“海马斯（HIMARS）”火箭炮系统对俄军造成了极大的毁伤，其背后原因在于，乌克兰在欧美的支援下组建了大量使用无人机进行前线侦查的情报小队，其主要职责是运用小型无人机对战区进行扫描，并将扫描数据上传至后方系统。己方人员在利用人工智能技术对海量数据进行识别和分析后，能够准确地判断俄军在战区中的人员和火力部署，随后，相关坐标会被发送给“海马斯”火箭炮部队。在完成打击后，情报小队还将使用无人机对打击效果进行评估。¹⁶可以看到，人工智能技术能够帮助现代军队更加充分地挖掘和发挥现有武器装备的战斗力，从而为现代军队的战略战术升级提供可行的路径。

第三，人工智能在情报分析方面也具有广阔的前景。战场环境下，军事情报人员需要在短时间内对来自多个信息源的庞大数据进行高效的收集、分析和处理。而人工智能技术能够通过机器学习的方式对过往数据进行梳理和提炼，帮助分析人员更加快速地从海量数据中提取关键信息，识别潜在威胁并做出及时的反应。例如，在未来的情报工作中，情报人员可以基于经过筛选的数据构建一个沙盘，并运用人工智能技术对其中的数据、算法、流程进行反复的测试和优化，成果将可被用于对海量、复杂的情报信息进行有针对性的分析。在分析结果的质量方面，经人工智能系统优化的分析结果在全面性与准确性上得到了大幅度的提升，能够在瞬息万变的战场环境下为用户提供更具准确性和时效性的策略建议。¹⁷2018 年，美国陆军就基于该思路研发并测试了一款名为 HUD3.0 的头盔显示器，该显示器利用了人工智能技术，能够对战场态势进行实时感知，并将数据上传至后台的数据中心并分析挖取其中关键信息，随后，该头盔显示器可以将数据和图像投射到士兵面前的透明平板上，从而呈现出其作战中所需要掌握的地形、障碍物、

approach", <https://breakingdefense.com/2024/07/frustrations-mount-over-armys-robotic-combat-vehicle-autonomy-acquisition-approach/>, 访问时间：2024 年 12 月 2 日。

15 U.S. Department of Defense, "DOD Updates Autonomy in Weapons System Directive", <https://www.defense.gov/News/News-Stories/Article/Article/3278065/dod-updates-autonomy-in-weapons-system-directive/>, 访问时间：2024 年 11 月 20 日。

16 NPR, "In the Russia-Ukraine war, drones are one of the most powerful weapons," <https://www.proquest.com/docview/2696550146?pq-origsite=primo&parentSessionId=4oCanUnhSzymNCv274OWbFYZ%2BNURKU3chHZUsff2DJY%3D&source=Other%20Sources>, 访问时间：2024 年 4 月 20 日。

17 张涛，马海群：《智能情报分析中数据与算法风险识别模型构建研究》，载《情报学报》2022 年第 8 期，第 832-833 页。

目标的图像和数据，帮助士兵提升其战斗能力。¹⁸在实战层面，以色列情报部门早在 2021 年就将包括“炼金术士”（The Alchemist）、“福音”（The Gospel）和“智慧深度”（Depth of Wisdom）在内的多个人工智能情报分析系统投入实战，主要用途包括袭击预警、情报分析和目标锁定等方面。¹⁹

（二）部分主要国家的人工智能军事化现状

当前，世界各个主要国家都在人工智能军事化领域进行布局和实践。其中以美国、法国、以色列为典型。美国在人工智能技术、投资、人才上都处于国际领先地位，被认为是人工智能技术发展的全球引领者；法国代表了在当前人工智能军事化竞争中紧追不舍的中等强国，其主要思路是将人工智能技术与自身传统优势技术相结合；以色列则是将人工智能武器大规模投入实战的代表国家。

美国在人工智能及其衍生技术的研究已经有近 50 年的历史，是人工智能军事化技术上首屈一指的领先者。近年来，美国在人工智能领域投入的劳动力、人才、硬件，以及每年在人工智能领域产生的出版物和专利质量方面均处于世界领先地位。2020 年，美国围绕人工智能中的研究项目已经达到 6000 多个，涉及云计算、自主武器系统、5G 传输、大数据等多个领域，共投资 48.3 亿美元。²⁰2021 年 8 月，美国国土安全部发布《人工智能/机器学习战略计划》，提出要增加研发投入，推动下一代人工智能和机器学习技术用于国家安全保障并构建安全的网络基础设施。²¹2022 年 8 月，拜登签署《芯片和科学法案》（Chips and Science Act），拨款 2800 亿美元用于芯片研发，而芯片便是支持人工智能技术发展的核心硬件之一。²²2023 年 3 月，美国白宫发布《国家人工智能研发战略规划》，从长期投资、人机融合、标准化和法律监管、安全性、数据库、成果评估等六个方面对美国的人工智能发展战略进行了部署。²³2023 年 8 月，美国国防部提出“复制器”计划（the Replicator Initiative），旨在简化美国军方对来自私营部门的新兴技术的吸

18 Breaking Defence, "HUD 3.0: Army To Test Augmented Reality For Infantry In 18 Months", https://breakingdefense.com/2018/03/hud-3-0-army-to-test-augmented-reality-for-infantry-in-18-months/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+BreakingDefense+%28Breaking+Defense%29, 访问时间：2024 年 12 月 1 日。

19 Lieber Institute West Point, "ISRAEL – HAMAS 2024 SYMPOSIUM – ALGORITHMS OF WAR: MILITARY AI AND THE WAR IN GAZA", <https://lieber.westpoint.edu/algorithms-war-military-ai-war-gaza/>, 访问时间：2024 年 11 月 20 日。

20 United States Department of Defense, "Defense Budget Overview 2020", https://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2021/fy2021_Budget_Request_Overview_Book.pdf, 访问时间：2024 年 10 月 5 日。

21 Homeland Security, "S&T Artificial Intelligence & Machine Learning Strategic Plan", https://www.dhs.gov/sites/default/files/publications/21_0730_st_ai_ml_strategic_plan_2021.pdf, 访问时间：2024 年 10 月 5 日。

22 The White House, "Chips and Science Act 2022", <https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/>, 访问时间：2024 年 10 月 5 日。

23 The White House, "NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN 2023 UPDATE", <https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf>, 访问时间：2024 年 11 月 5 日。

收，计划提出美军将在未来加强与私营部门合作，建立一款基于人工智能技术的“全域、耐损、自主”作战系统系统。²⁴在人工智能技术中最前沿的机器学习领域，美国国防部高级研究计划局（DARPA）于 2018 年 9 月推出了一项专门计划“下一代人工智能”（AINext）项目，计划投资 20 亿美元，目标是为美军搭建一套能够实现智能交流和逻辑推理的人工智能工具，其主要作用包括开发未来的国防软件系统，提高人工智能系统的能效比、鲁棒性和可靠性，增强机器学习和人工智能技术的安全性和弹性等²⁵。需要注意的是，目前美国大部分尖端 AI 研发都发生于民营部门，包括亚马逊、微软、英伟达在内的众多美国公司都掌握了当前世界上一流的人工智能技术和投资。美国政府将这些与自身有潜在合作价值的机构称作国防创新单位（DIU），美国政府会通过项目投资等方式将研发需求转包给 DIU，在其开发出“对抗性人工智能”等新技术后，DARPA 会协助 DIU 将新技术交予美国国防部内的首席数字和人工智能官办公室（CDAD）进行后续的演示、评估和采购工作，这一过程可以实现人工智能武器系统从非政府科研部门到军事部门的转移。²⁶但是，这种合作模式也存在一些问题，例如深度参与政府的军事项目可能会对私人公司的商业形象构成破坏，进而导致其出于声誉原因而拒绝参与与相关合作。²⁷

法国在人工智能投资和技术水平上不及美国，但是其同样意识到了人工智能军事化的巨大潜力。2018 年，法国政府首次提出了其《人工智能国家战略》（la *Stratégie nationale pour l'IA*），旨在将法国打造成欧洲乃至全球具有竞争力的人工智能强国。²⁸法国政府为该战略将提供了 15 亿欧元的投资，并计划在 2030 年逐渐将该投资追加到 25 亿欧元。²⁹法国十分关注人工智能技术在军事领域的应用，以法国国防创新局（l'Agence de l'innovation de défense）为代表的法国战略部门倾向于让法军部队和相关军工企业共同参与其人工智能军事化发展战略的制定

24 U.S. Department of Defense, “Deputy Secretary of Defense Kathleen Hicks Keynote Address: 'The Urgency to Innovate' (As Delivered)”, <https://www.defense.gov/News/Speeches/Speech/Article/3507156/deputy-secretary-of-defense-kathleen-hicks-keynote-address-the-urgency-to-innov/>, 访问时间：2024 年 4 月 14 日；Centre for International Governance Innovation (CIGI) “The United States Quietly Kick-Starts the Autonomous Weapons Era”, <https://www.cigionline.org/articles/the-united-states-quietly-kick-starts-the-autonomous-weapons-era/>, 访问时间：2024 年 12 月 20 日。

25 DARPA, “AI Next Campaign (Archived)”, <https://www.darpa.mil/work-with-us/ai-next-campaign>, 访问时间：2024 年 4 月 3 日。

26 CSIS, “The DARPA Perspective on AI and Autonomy at the DOD”, <https://www.csis.org/events/darpa-perspective-ai-and-autonomy-dod>, 访问时间：2024 年 4 月 21 日。

27 Texas National Security Review, “The Rift between Silicon Valley and the pentagon is economic, not moral,” <https://warontherocks.com/2019/01/the-rift-between-silicon-valley-and-the-pentagon-is-economic-not-moral/>, 访问时间：2024 年 4 月 24 日。

28 European Council on Foreign Relations, “Not smart enough: The poverty of European military thinking on artificial intelligence,” https://ecfr.eu/archive/page/-/Ulrike_Franke_not_smart_enough_AI.pdf, 访问时间：2024 年 4 月 24 日。

29 Ministère de l'Économie, des Finances et de l'Industrie et Ministère chargé du Budget et des Comptes publics, “La stratégie nationale pour l'intelligence”, <https://www.economie.gouv.fr/strategie-nationale-intelligence-artificielle-artificielle>, 访问时间：2024 年 10 月 6 日。

过程，从而充分地协调并整合各部门资源。³⁰为节约成本并提高技术可靠性，法国军方的一个关键战略是将人工智能技术与多种现有装备系统相融合。例如，法国军方正在利用人工智能技术开发未来空中作战系统（FCAS），并将该系统安装于法国达索公司开发的无人作战飞机“神经元”上，其目标是在未来的空战中实现有人机与无人机“蜂群”之间的相互配合。³¹

以色列面临着较大的安全和国防压力，其对于先进军事装备和作战系统具有相对更高的需求。根据《以色列国家人工智能战略》（ISRAEL AI National AI Program 2024），以色列建立了以政府战略、操作环境、研发基础社会三位一体的人工智能发展框架，计划在 2024 年前投入共计 10 亿新谢克尔的资金。³²在 2023 年爆发的新一轮巴以冲突中，以色列首次实现了人工智能技术在战争中的大范围应用。由人工智能赋能的自动驾驶、网络舆情战、情报收集分析和自主武器系统显著增强了以色列国防军的作战能力。例如，以色列国防军开发了一款名为“战争瓦兹”（the Waze of War）的人工智能战场管理系统，其能够帮助战场上的各单元之间实时共享态势，这使得以军的情报收集、作战指挥、火力打击等模块之间形成了一个扁平化的网络，拥有极为灵敏的反应速度。以色列国防军在其官方声明中指出，以军利用人工智能技术打击了加沙地带 1.1 万多个目标，其中 90% 的目标是实时生成并被打击的，这要归功于“人工智能和数百名以色列国防军情报官员的共同努力”。³³此外，以色列政府利用一个名为“蓝狼”（Blue Wolf）的人工智能系统对生活在约旦河西岸和加沙地带的数百万巴勒斯坦人进行长期的跟踪和记录。从 2019 年开始，以色列通过检查站、闭路电视录像、摄像头网络、卫星图像、电子信号收集了大量的巴勒斯坦人信息，并将其输入到人工智能系统中进行算法训练，最终建立了一套“哈马斯脸谱”，而以色列国防军就基于这一系统，在战争中识别并攻击那些他们认为可能是哈马斯武装人员的巴勒斯坦人。³⁴

综上所述，人工智能在军事领域具有广阔的应用前景，能够为国家在现代战争中带来战争能力方面的显著提升。当前，多个国家已经开始开发和制造人工智能武器系统，甚至甚至将其应用于实战。在这一背景下，国际社会有必要对其中可能蕴藏的风险进行识别和应对。

30 MINISTERE DES ARMÉES, “La communauté d’intelligence artificielle du ministère des Armées réunie pendant 2 jours”, <https://www.defense.gouv.fr/aid/actualites/communaute-dintelligence-artificielle-du-ministere-armees-reunie-2-jours>, 访问时间：2024 年 6 月 1 日。

31 MINISTERE DES ARMÉES, “Les drones armés”, <https://www.defense.gouv.fr/dems/drones-armes>, 访问时间：2024 年 6 月 1 日。

32 Ministry of Innovation, Science & Technology, Israel, “ISRAEL AI National AI Program 2024” <https://aiisrael.org.il>, 访问时间：2024 年 10 月 7 日。

33 The Jerusalem Post, “IDF uses AI to strike over 11,000 terror target in Gaza since October 7”, <https://www.jpost.com/israel-news/article-771419>, 访问时间，2024 年 5 月 20 日。

34 Institute for Middle East Understanding (IMEU), “Fact Sheet: Israeli Surveillance & Restrictions on Palestinian Movement”, <https://imeu.org/article/fact-sheet-israeli-surveillance-restrictions-on-palestinian-movement>, 访问时间：2024 年 12 月 12 日。

（三）人工智能军事化的风险

人工智能在军事领域具有强大的应用潜力。但是，被用于军事化的人工智能在本质上仍是机器，当涉及杀戮的判断和决策由机器完成时，便会带来大量的安全、法律、伦理和技术问题，甚至在某些极端情况下，由人工智能所作出的决定（如使用核武器）可能会将人类拖入深渊，因此人类社会不可无视人工智能军事化带来的风险。³⁵综合来看，这种风险主要体现在以下四个方面：

第一，人工智能军事化将导致潜在的军备竞赛和紧张局势。人工智能对于传统军事技术的赋能效果十分显著，如果没有相关军控机制进行规范，各国可能会围绕人工智能展开军备竞赛以获取军事优势，最终导致过度的军备发展，增加国家间不信任感，提高潜在的战争风险。目前，以北约（NATO）成员国为代表的某些西方国家在人工智能军事化领域就对中国抱有极大的警惕，北约一方面对中国在数据、软件、硬件等方面实施了出口管制，另一方面炮制出“中国人工智能威胁论”以试图将中国排除在人工智能国际合作之外。³⁶2018 年，北约秘书长斯托尔滕贝格（Jens Stoltenberg）提出，中国在人工智能领域的投资已高达 1500 亿美元，有可能在 2030 年前后成为人工智能领域的全球领袖，北约已经面临来自中国技术的直接威胁，应该将中国作为人工智能领域的竞争对手和防范对象。³⁷导致北约在人工智能军事化领域对中国持担忧态度的原因可能包括两点：首先，人工智能本身并不是构成一个单一的武器系统，而是被构建到各种各样的武器系统和核心基础设施中，这提高了北约对中国人工智能技术发展进行客观分析的难度。³⁸其次，人工智能技术的强大能力和资源稀缺性放大了北约对于中国的威胁认知。在未来，如果主要国家与国家集团之间无法形成围绕人工智能武器的原则性共识，则很有可能加剧彼此间不安全感的互构，对国际稳定造成不利影响。

第二，人工智能军事化可能导致战争中的法律和伦理问题。在人工智能时代，人工智能军备的核心特征是装备和程序的自主性，同时这种自主性会随着人工智能技术的发展而不断提升，最终使得战争中的决策和行为都向着更少人工干预的方向演进。然而，当战场上的攻击决策和行为不再由人类直接做出，武器系统开始自主决定杀戮的对象和方式时，战争便不再是一种完全的人际互动行为。这引发了一系列问题：从法律的角度来说，根据托马斯·谢林（Thomas C. Schelling）的“门槛”理论，战争烈度的升级往往源于战争双方对一些行为“门槛”（如攻击平

35 "The Perilous Coming Age of AI Warfare",
<file:///Users/zouyuting/Library/Containers/com.netease.macmail/Data/Library/Application%20Support/view/2/1131/1027/464.pdf>, 访问时间：2024 年 11 月 5 日。

36 The White House, "Interim National Security Strategy Guide", <https://www.whitehouse.gov/wp-content/uploads/2021/03/NSC-1v2.pdf>, 访问时间：2024 年 11 月 5 日。

37 Transformation Network Branch SACT NATO, "2018 Chiefs of Transformation Conference Analysis Report", <https://www.act.nato.int/images/stories/events/2018/cotc/2018-cotc-report.pdf>, 访问时间：2024 年 4 月 12 日。

38 "Artificial intelligence and the security dilemma," Brookings Institute, <https://www.brookings.edu/articles/artificial-intelligence-and-the-security-dilemma/>, 访问时间：2024 年 4 月 19 日。

民或使用大规模杀伤性武器)的突破³⁹,在传统战争中,战争各方能够通过确认和规避“门槛”来控制战争规模,并基于此确定战争责任。在人工智能时代,武器系统被赋予了对攻击行为进行自主决策和升级的能力,进而破坏了战争中人与技术之间的等级-工具关系。⁴⁰现有实证研究表明,依赖人工智能系统的个人和团队往往会出现“自动化偏差”,即逐步陷入对人工智能的依赖。⁴¹这个过程中,如果智能武器系统因为程序设计不当、信息不完全或敌方干扰等原因而导致对特定战争“门槛”的突破,这个过程中所产生的战争责任将难以追究。从伦理的角度来说,人工智能军事化挑战了人类战争中的诸多伦理原则:首先,战争中伤亡情况的合法性在于是战争双方自愿成为战斗人员并愿意承担可能被杀伤的后果。⁴²但是人工智能武器破坏了这种关系。其次,人类难以保证自己能够完全理解并控制人工智能系统的决策过程,从而难以确保其行为符合国际人道法和战争法等道义和法律要求。最后,将决定人类生死的权力交予机器和程序手中,将导致人类尊严被计算机程序所凌驾。以上难题一方面可能引发全球公民社会对于战争伦理问题责任方的施压,另一方面也有可能引发人类社会对于战争本质的集体思考。

第三,人工智能军事化涉及到高度复杂的技术系统,这使得其应用可能面临技术窃取、技术失灵和技术误用的风险。首先,被开发以用于军事目的的人工智能技术可能会被不正当地利用,例如,北约在其反恐行动中投入了大量的人工智能武器,这虽然对恐怖组织造成了相当程度的打击,但是这也导致了这些武器通过战斗损耗的方式被恐怖组织掌握。恐怖组织可以通过拆装武器、下载软件等方式掌握人工智能军事化技术,并利用其策划新型的恐怖袭击。⁴³其次,由于人工智能技术仍处于不断发展和完善的阶段,给予其高度的自主权限可能导致无人武器在复杂环境下执行任务时做出意外的决策,导致无法预测的后果。例如,2007年美军第三机步师在一次军事行动中使用了三台拥有自主侦查和进攻能力的“剑”式武装机器人。其中一台机器人因软件故障而将武器瞄准美军操作员,在无法排除故障的情况下,美军最终只得将该机器人摧毁,这一事故对美军行动造成了严重干扰。⁴⁴最后,围绕现有人工智能系统内部漏洞的攻防可能成为未来战争的焦点之一,而人工智能技术往往与通讯网络、物联网等同时具备军用和民用性质的技术设施联系密切,这可能导致以军事目标为对象的攻击行为最终伤害到平民和

39 托马斯·谢林著:《军备及其影响》,上海人民出版社2017版,第116页。

40 Elke Schwarz, "Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control", file:///Users/zouyuting/Downloads/retrieve.pdf, 访问时间:2024年12月1日。

41 Toni Erskine and Stephan Miller, "AI and the decision to go to war: future risks and opportunities," *Australian Journal of International Affairs*, Vol. 78, No. 2, 2024, pp. 135-147.

42 Asaro, Peter. "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making." *International Review of the Red Cross*, Vol.94, No.886, 2012, pp. 687-709.

43 高望来:《北约人工智能反恐新态势及其困局》,载《欧洲研究》2021年第2期,第153页。

44 中华人民共和国国防部,“懂算法才能打‘算法战’”,<http://www.mod.gov.cn/gfbw/jmsd/4845421.html>, 访问时间:2024年6月2日。

民用设施。⁴⁵

综合而言，人工智能军事化对人类社会构成了相当程度的威胁，如果国际社会放任其无限制发展，则很有可能在现阶段导致国家间围绕该领域的军备竞赛。在未来可能的国际战争中，一旦掌握人工智能武器的国家将相关武器系统投入实战运用，则有可能引发一系列技术、法律和伦理问题，给交战国的军事部门和平民带来一系列的潜在风险。因此，人工智能军控应该成为当前全球治理中的一个重要的议程。

三、人工智能军控的现状

目前，国际社会尚未围绕人工智能军控形成普遍性的规制体系。⁴⁶联合国秘书长古特雷斯（António Guterres）在联合国安理会（United Nations Security Council）2018年发布的《裁军议程：为了共同未来的安全》文件中提到，人工智能赋能的武器系统可能导致不可预测和难以解释的行为，因此国际社会应该对其研发和使用持谨慎态度。⁴⁷这种呼吁体现了国际社会对人工智能军事化风险的担忧，以及对其有效治理的迫切需求。在这一背景下，各国应该共同努力，发展和完善针对人工智能武器的军备控制体系。军备控制指对武器及其相关设施、相关活动或者相关人员进行约束⁴⁸，这一行动可以发生在武器开发、制造和使用的各个阶段，具体包括对武器底层技术的限制、对武器开发、生产、储存过程的限制、对国家在和平时期储备武器数量的限制、对国家在战时使用武器程度和方式的限制等。当前，对人工智能军事化展开进行军控行动已经是国际社会的共识，围绕该话题的讨论可以聚焦在军控的目标、行动、困境三个方面。

（一）关于军控的目标

目前，各个国家、国际组织和行业相关人士在人工智能武器系统的军控目标上存在三种主要立场：全面禁止，即以立法形式全面禁止人工智能武器、部分禁止，即对人工智能武器从研发到使用的部分流程进行限制；反对禁止，即反对对人工智能武器加以管控。值得关注的是，各国对待人工智能军控目标的态度往往与国家实力呈现出相关性。人工智能产业发展中，“先发优势”在技术发展和规范制定层面都有较为明显的体现，率先取得技术发展和标准制定优势的国家往往可以获得持续递增的收益，而落后的国家则不得不付出经济和政治利益的代价，因

45 Ministry of Innovation Science and Technology of Israel, “Israel's Policy on Artificial Intelligence Regulation and Ethics”,

[https://www.gov.il/BlobFolder/generalpage/israel_national_ai_program_2024/he/Israel's%20Policy%20on%20Artificial%20Intelligence%20Regulation%20and%20Ethics%20\(2023\).pdf](https://www.gov.il/BlobFolder/generalpage/israel_national_ai_program_2024/he/Israel's%20Policy%20on%20Artificial%20Intelligence%20Regulation%20and%20Ethics%20(2023).pdf), 访问时间：2024年10月11日。

46 United Nation Office of Disarmament Affairs, “Lethal Autonomous Weapon Systems (LAWS)”, <https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/>, 访问时间：2024年6月1日。

47 United Nations, “SECURING OUR COMMON FUTURE”, <https://s3.amazonaws.com/unoda-web/wp-content/uploads/2018/06/sg-disarmament-agenda-pubs-page.pdf#view=Fit>, 访问时间，2024年6月8日。

48 李彬著：《军备控制理论与分析》，国防工业出版社2006年版，第4页。

此实力不同的国家在人工智能军控的目标上往往持有不同的态度。根据非政府组织禁止杀手机器人运动（Stopping Killer Robots）统计，不同的国家间观点大致可以如表 1 所示。

表 1 各国关于禁止人工智能武器的态度

各国关于人工智能武器的态度	
全面禁止	古巴、阿尔及利亚、阿根廷、奥地利、玻利维亚、巴西、智利、哥伦比亚、哥斯达黎加、吉布提、厄瓜多尔、埃及、萨尔瓦多、加纳、危地马拉、梵蒂冈、伊拉克、约旦、墨西哥、摩洛哥、纳米比亚、尼加拉瓜、巴基斯坦、巴拿马、秘鲁、巴勒斯坦、乌干达、委内瑞拉、津巴布韦、大多数国际组织和社会团体。
部分禁止	中国
反对禁止	美国、俄罗斯、澳大利亚、以色列、英国、韩国、土耳其、美国、英国。

资料来源： Stop killer robots, “Country Views on Killer Robots”,
https://www.stopkillerrobots.org/wp-content/uploads/2020/05/KRC_CountryViews_7July2020.pdf. 访问时间：2024 年 5 月 24 日。

首先，以阿尔及利亚、巴西为代表的中小国家，往往支持按对人工智能武器进行全面监管。在 CCW 等一系列多边平台中，阿尔及利亚主张对人工智能武器的设计、开发、测试、部署、转让和使用严加限制。⁴⁹其理由是人工智能的军事化将导致军备竞赛，同时如果人工智能武器通过战争损耗被恐怖组织获取，则可能造成更大的人道主义灾难。⁵⁰巴西也支持对人工智能武器严加管控，其主要关注人工智能武器的法律和伦理风险，认为“人工智能武器与国际人道法和国际人权法不相容”。⁵¹2022 年，各国在巴西里约热内卢举办了一次自主武器系统国际研讨会，这是国际社会在联合国平台之外召开的第一次关于人工智能武器的政府间多边会议。⁵²综上，以阿尔及利亚、巴西为代表的国家，大多呼吁在人工智能军控上尽快达成具有法律约束力的国际文书。其目的一方面包括更有效地监督和

49 Government of Argentina, "Statement to the Convention on Conventional Weapons Fifth Review Conference", https://conf.unog.ch/dr/public/61.0500/DDE5EFD9-810F-41D4-A0AB-3D6244069E03_15h03/chunks/snippet_IOS83-47t88-38.mp3. 访问时间：2024 年 11 月 11 日。
50 Government of Argentina, "Statement to the UN Human Rights Council", <http://webtv.un.org/meetings-events/human-rights-council/watch/clustered-id-on-executions-and-idps-9th-meeting-23rd-regular-session-of-human-rights-council/2419860355001#full-text>. 访问时间：2024 年 11 月 11 日。
51 UNODA, "Brazil's Submission on Autonomous Weapon Systems in Response to the Request Pursuant to Resolution 78/241", [https://docs-library.unoda.org/General_Assembly_First_Committee_-_Seventy-Ninth_session_\(2024\)/78-241-Brazil-EN.pdf](https://docs-library.unoda.org/General_Assembly_First_Committee_-_Seventy-Ninth_session_(2024)/78-241-Brazil-EN.pdf). 访问时间：2024 年 11 月 11 日。
52 Fundação Alexandre de Gusmão (FUNAG), "Rio Seminar on Autonomous Weapons Systems," <http://www.funag.gov.br/index.php/en/news/3072-registrations-open-for-the-rio-seminar-on-autonomous-weapons-systems>. 访问时间：2024 年 11 月 12 日。

规范人工智能武器的发展和使用，减少人工智能军事化所导致的技术、法律、伦理风险，另一方面，由于这些国家在人工智能技术发展和产业化上本就处于劣势地位，在人工智能武器被认为具有强大军事潜力的当下，这些国家的立场可能也源于其不希望其他人工智能大国对自身形成更大的军事优势。⁵³

大多数国际组织和社会团体也支持对人工智能武器严加控制。自 2018 年以来，联合国秘书长古特雷斯一直主张人工智能武器“在政治上不可接受，在道德上令人反感”。⁵⁴2023 年，在联合国《新和平纲领》(A New Agenda for Peace) 中，古特雷斯指出，人工智能武器的设计、发展和使用引起人道主义、法律、安全和道德方面的关切，并对人权和基本自由构成直接威胁，因此他建议各国在 2026 年之前缔结一项具有法律约束力的文书，禁止在没有人类控制或监督的情况下使用人工智能武器系统，并建立相应的治理框架。⁵⁵社会团体方面，包括大赦国际 (Amnesty International)、人权观察组织(Human Rights Watch)和诺贝尔女性倡议组织(Nobel Women's Initiative)在内的 250 多个非政府组织组成的联盟成立了“停止杀人机器人运动”(Campaign to stop Killer Robots)，呼吁通过一项具有法律约束力的先发制人的条约来禁止自主武器。⁵⁶2018 年，在瑞典首都斯德哥尔摩举办的国际人工智能联合会议上，包括特斯拉 CEO 埃隆·马斯克 (Elon Reeve Musk) 在内的 2000 多名人工智能行业人士共同签署了《致命性自主武器宣言》，宣言呼吁“国际社会共同建立一套国际法规以遏止致命性自主武器的发展，各国政府不支持也不参加对于各类致命性自主武器的研制、贸易和使用”。⁵⁷

但是，在各个人工智能强国均已事实部署和使用人工智能武器的当下，制造和使用人工智能武器的“红线”已经被突破。同时，人工智能军事化的危害尚未充分体现，各国都看到了人工智能所蕴含的强大军事潜力，意识到了可以通过人工智能军事化为自身取得不对称的战略优势，因此现阶段对人工智能武器展开严格限制的难度较大。

其次，以中国为代表的一部分国家认为，人工智能技术的军事应用具有两面性，并不是完全对人类社会有害。例如，只要确保技术的可靠性，那么人工智能技术能够帮助军事人员更好地分辨军用和民用目标，从而改善武装冲突可能引发

53 张煌，杜雁芸：《人工智能军事化发展态势及其安全影响》，载《外交评论(外交学院学报)》2022 年第 3 期，第 128 页。

54 United Nation office for disarmament affairs, "Lethal Autonomous Weapon Systems (LAWS)", <https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/>, 访问时间：2024 年 12 月 1 日。

55 United Nation, "A New Agenda for Peace", <https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf>, 访问时间：2024 年 12 月 1 日。

56 Foreign Affairs, "The Perilous Coming Age of AI Warfare", <file:///Users/zouyuting/Library/Containers/com.netease.macmail/Data/Library/Application%20Support/view/2/1131/1027/464.pdf>, 访问时间：2024 年 11 月 5 日。

57 Lucas Pery, "Lethal Autonomous Weapons Pledge" <https://futureoflife.org/open-letter/lethal-autonomous-weapons-pledge/>, 访问时间：2024 年 4 月 20 日。

的误伤等人道主义问题。此外，使用人工智能武器有助于减少作战人员伤亡。⁵⁸ 中国主张，将人工智能武器系统分为“不可接受的”和“可接受的”两部分，其中，“不可接受的”人工智能武器系统应满足致命性、全自主性、无法终止性、滥杀性、进化型等五个基本特征。同时具备上述五个特征的人工智能武器系统具有明显的反人类特征和巨大人道主义风险，国际社会应该对其研发和使用进行严格限制。“可接受的”的人工智能武器系统可以具有较高的自主程度，在军事行动中能够遵循国际人道主义法基本原则。针对此类武器系统，中方支持各国在国际法原则基础上采取分级、分类管理等措施以降低风险，并保证人类对其的控制以确保责任的可追溯。⁵⁹

最后，一部分国家则反对来自外部的限制，其倾向于在其内部进行自主监管。这些国家往往在人工智能技术上处于领先地位（如美国），或者面临较大的安全压力（如以色列）。这些国家主张保护和支持本国的人工智能技术创新，反对外部监管对自身人工智能武器技术发展的限制。以美国为例，人工智能武器是美国确保其世界领导地位和维持其军事威慑力的安全支撑和重要抓手⁶⁰，因而美国在人工智能监管上采取了“创新优先”的发展战略。基于该立场，美国的辞令是国际社会在决定军控措施时需要更多的讨论和深入了解，以确保采取的军控措施既能维护国际社会的集体安全，又不会侵犯国家的军事自主权。⁶¹美国国防部在其《人工智能原则：国防部关于人工智能应用伦理的建议》中提出了五项指导原则，即“负责任、公平、可追踪、可靠、可管理”。⁶²但是这五条原则相对模糊，一旦在未来遇到涉及到国家安全的敏感问题，美国政府在该原则的解释上拥有很大的回旋空间。澳大利亚也反对对人工智能武器进行限制。虽然澳大利亚参与了 CCW 框架下围绕人工智能武器的一系列谈判，但是澳大利亚认为，人工智能技术能够促进对军用和民用目标的分辨，进而避免平民伤亡等人道主义问题。同时，由于现阶段人工智能技术仍处于高速发展中，其前途尚不明了，因此澳大利亚认为现阶段制定一项正式性的国际条约来对其进行约束“为时尚早”。⁶³

58 United Nations Office For Disarmament Affairs: “中国关于‘致命性自主武器系统’问题的文件”，<https://documents.unoda.org/wp-content/uploads/2022/07/Working-Paper-of-the-Peoples-Republic-of-China-on-Lethal-Autonomous-Weapons-SystemsChinese.pdf>。访问时间：2024 年 6 月 2 日。

59 中华人民共和国外交部：“中国就“致命性自主武器系统”问题向联合国秘书长提交的文件”，http://svideo.mfa.gov.cn/wjb_673085/zzjg_673183/jks_674633/fywj_674643/202405/t20240523_11310587.shtml，访问时间：2024 年 6 月 2 日。

60 The White House, “The National Artificial Intelligence R&D Strategic Plan 2023 update”, <https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf>，访问时间：2024 年 11 月 2 日。

61 “Artificial Intelligence Regulation in 2024: Examining the U.S.’s Market-Driven Strategy in Comparison to the EU’s Proactive Approach,” *International and Comparative Law Review*

62 Defense Innovation Board, “AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense: Supporting Document”. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF，访问时间：2024 年 4 月 20 日。

63 Government of Australia, “Statement to the Convention on Conventional Weapons Group of Governmental Experts on lethal autonomous weapons”, <https://conf.unog.ch/digitalrecordings/index.html?guid=public/61.0500/E3373111-6687-440D-B02D->

上述现象与过往的军控实践存在一定张力，传统观点一般认为强国更倾向于加强对武器系统的管制，因为它们担心这类武器可能颠覆现有的权力格局。⁶⁴但是在人工智能的国际军控实践中，主张对人工智能军事化严加控制的往往是小国。造成这一现象的原因可能是，人工智能技术需要强大的国家实力作为支撑，特别是最前沿的人工智能技术需要强大的基础设施、电力、人才、高端制造业支持，因此大国在很大程度上锁定了自身在人工智能领域的优势，而在人工智能军事化领域不具有优势的小国，可能会在相关军控议题上倾向于严格限制人工智能技术的军事化，从而防止其与大国间军事实力差距的进一步拉大。

（二）关于军控的主体和行动

人工智能军控可能涉及多个主体间不同的行动，这些主体是构成目前和未来治理框架的基本单元，多种主体在其内部和彼此之间所采取的一系列行动直接影响了人工智能军控的进展和成效。从现有实践上看，人工智能军控中各主体及其行动可以通过下图表示：

表2：人工智能军控中各主体及其行动

军控主体	主要行动	
国家和国家集团	参与国际谈判	
	制定内部法规	对内：制定技术标准
		对外：实行出口管控
国际组织	制定国际标准、设置专门机构	

资料来源：作者自制

国家和国家集团是人工智能国际军控最主要的治理主体之一，在国际关系现实主义的理论观点中，国家被视作理性行为体，决策者会尽力扩大本国对他国的实力优势，并且防止其他国家形成自身的优势。⁶⁵因此，与联合国不同，国家进行军备控制的核心动机不是防止某一项装备或技术对全人类造成威胁，而是防止其被敌对国家所掌控，从而对己方国家利益造成不利影响。从行动上来看，国家和国家集团主要通过加入国际谈判和制定内部法规的方式参与人工智能军控。

在加入国际谈判方面，国家和国家集团通过推动地区性或全球性治理框架的达成，能够规范人工智能在军事领域的发展和使用，防止其他国家或国家集团在人工智能军事化上达成对自身不利的成果。因此，国家会积极参与相关国际规制

CEFB477A9AD6_10h06&position=3539, 访问时间：2024 年 12 月 2 日。

64 CNAS, “Artificial Intelligence and Arms Control”, <https://www.cnas.org/publications/reports/artificial-intelligence-and-arms-control>, 访问时间：2024 年 4 月 12 日。

65 李彬著：《军备控制理论与分析》，国防工业出版社 2006 年版，第 45 页。

和标准的制定谈判，或是通过制定国内法规以巩固和促进自身发展，为自身争取技术竞赛中的主动权。目前，世界主要国家之间已经就人工智能军控达成了初步共识，2023年11月，在由英国举办的首届人工智能安全峰会上，29个国家和国际组织共同签署了《布莱切利宣言》（Bletchley Declaration），宣言对当前人工智能技术发展，特别是其前沿技术对发展对人类社会造成的风险提出了警告，并达成了所有缔约方将共同致力于负责任地设计和部署人工智能的共识，各国共同承诺将基于安全、以人为本、值得信赖和负责任的原则来设计、开发、部署和使用人工智能。⁶⁶《布莱切利宣言》是世界首个由多个国家和地区在人工智能安全领域达成的宣言，在人工智能的国际军控进程中有着里程碑的意义，但是，《布莱切利宣言》只是一部以阐述基本原则和共识的“软法（soft law）”，其中并无针对具体违约行为的处罚条款。由于国家安全利益间存在客观的不可调和性，仅凭该宣言很难弥合各个主要国家之间在人工智能军控议题上的巨大分歧。

在制定内部法规方面，从对内角度来说，国家和国家集团会制定相关技术标准以规范本国人工智能军事化发展。当前，西方国家联盟在人工智能标准设定方面处于领先地位，早在2016年，以欧盟《欧盟机器人民事法律规则》就对人工智能机器人的法律权利和义务进行了确定，由于人工智能机器人在军事领域具有广泛的用途，因此该《规则》在人工智能军控进程中也具有一定意义。⁶⁷2022年，美国、欧盟在签订的美欧贸易协定中设置了技术委员会（TTC），其职能是对包括人工智能在内的技术发展和流动进行更为有效的监管，建立所谓“民主国家人工智能治理合作”。⁶⁸2023年5月，来自七国集团和欧盟的代表在“广岛人工智能进程”（Hiroshima AI process, HAIP）部长级论坛上提出，与会各国将共同参与制定一份人工智能领域的行为准则，具体包含11项内容，其目标是对各大跨国公司开发的包括人工智能武器在内的人工智能系统进行规范，并制定相关的监测和评估机制，从而“根据共同的民主价值观，实现值得信赖的人工智能”。⁶⁹目前，HAIP中所倡导的一系列原则已经在美国《2023年人工智能行政命令》、欧盟《人工智能法案》等西方国家人工智能监管实践中得以体现。⁷⁰但是，这一进程的针对性也非常明显，正如美国国防部在《负责人工智能战略和实施路径》中所言，

66 GOV.UK, “The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023”, <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>. 访问时间：2024年6月2日。

67 European Parliament, “European Civil Law Rules in Robotics”, [https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU\(2016\)571379_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf), 2024年12月1日。

68 “The EU and U.S. are starting to align on AI regulation,” Brookings Institution, <https://www.brookings.edu/articles/the-eu-and-u-s-are-starting-to-align-on-ai-regulation/>, 访问时间：2024年6月2日。

69 Ministry of Internal Affairs of Japan, “G7 Leaders’ Statement on the Hiroshima AI Process,” https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document01_en.pdf, 访问时间：2024年10月11日。

70 CSIS, “Advancing the Hiroshima AI Process Code of Conduct under the 2024 Italian G7 Presidency Timeline and Recommendations,” <https://www.csis.org/analysis/advancing-hiroshima-ai-process-code-conduct-under-2024-italian-g7-presidency-timeline-and>, 访问时间：2024年4月21日。

“美国的竞争对手正在以威胁全球安全、和平与稳定的方式大力投资人工智能，为了在数字竞争的世界中保持军事优势，美国必须拥抱人工智能技术”。⁷¹鉴于当前只有中国在人工智能技术上具有同 HAIP 成员国竞争的能力，因此可以预见到由西方主导的人工智能国际监管体系必然将在未来针对中国人工智能产业作出更多针对性限制。除美欧外，上合组织、东盟、非盟等国家集团目前尚未出台其专门针对人工智能军事化的内部法规，导致这一现象的原因可能包括两点：第一，这些国家在人工智能军事化发展上不及西方国家；第二，由上合组织、东盟、非盟等国家集团在合作紧密程度上不及欧盟、G7 等西方国家集团。

在单一国家内部的法规制定上，美国和法国在设立与军事化有关的人工智能国内标准方面起步较早。美国国防部于 2020 年 2 月正式发布了规范人工智能军事能力发展的“负责、公平、可追溯、可靠和可治理的”五项原则。⁷²2022 年，美国国防部发布了《负责人工智能战略和实施路径》，建立了六项原则作为开发人工智能技术的“基线标准”，具体包括国防部职权、用户信任度、产品生命周期、装备性能、人工智能生态系统、人工智能人才队伍六个方面。⁷³2023 年 10 月 30 日，美国总统拜登颁布了《关于安全、可靠和可信的人工智能的行政命令》，其中要求其国内民营企业和军方在开发人工智能系统时必须遵循《国防生产法》，特别是要遵守安全、可靠、可信等道德原则。⁷⁴法国也曾多次表达对人工智能武器系统发展的关切，在机构设置方面，法国国防部曾在 2019 年成立了一个专注于军事人工智能应用伦理影响的委员会，这表明法国政府在很早就关注到了人工智能军事化的伦理风险问题。⁷⁵2023 年 10 月，法国政府针对最前沿的生成式人工智能技术成立了生成式人工智能委员会（Comité de L'intelligence Artificielle Générative）。该委员会汇集了来自包括经济、军事、文化、技术等不同领域的专家学者，主要职能是在未来为法国政府提供有关人工智能技术的监管建议，防止该技术被错误地发展和应用。⁷⁶在标准制定方面，法国政府在《人工智能国家战略》和《人工智能加速计划》（Stratégie d'accélération Intelligence Artificielle）中对

71 U.S. Department of Defense, "Responsible Artificial Intelligence Strategy and Implementation Pathway", <https://www.ai.mil/Portals/137/Documents/Resources%20Page/DoD%20Responsible%20AI%20Strategy%20and%20Implementation%20Pathway.pdf>, 访问时间：2024 年 11 月 15 日。

72 Defense Innovation Board, "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense: Supporting Document," https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF, 访问时间：2024 年 4 月 20 日。

73 U.S. Department of Defense, "Responsible Artificial Intelligence Strategy and Implementation Pathway", <https://www.ai.mil/Portals/137/Documents/Resources%20Page/DoD%20Responsible%20AI%20Strategy%20and%20Implementation%20Pathway.pdf>, 访问时间：2024 年 11 月 15 日。

74 The White House, "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence", <https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/>, 访问时间：2024 年 5 月 8 日。

75 MINISTÈRE DES ARMÉES, "Les systèmes d'armes létales autonomes (SALA)", <https://www.defense.gouv.fr/dems/systemes-darmes-letales-autonomes-sala>, 访问时间：2024 年 5 月 20 日。

76 Gouvernement de France, "Comité de l'intelligence artificielle générative", <https://www.info.gouv.fr/communique/comite-de-lintelligence-artificielle>, 访问时间：2024 年 12 月 3 日。

本国人工智能技术发展提出了三个要求：嵌入式、节俭和可信性。⁷⁷

从对外角度看，国家与国家集团进行军备控制的一个重要手段是通过出口管控的方式防止本国人工智能武器及相关技术的外泄，从而确保自身在人工智能军事化的国际竞争中不落于下风。在具体操作上，国家或国家集团会制定旨在监督和限制某些军事、军民两用商品和技术贸易的相关法规，对其转让施加许可要求，在某些情况下甚至会禁止特定技术或物品的出口。目前，围绕人工智能的出口管制主要分为硬件管制、软件管制、技术管制、全方位控制四个方面。⁷⁸在硬件管制领域，美国和欧盟采用的《军需两用管制清单》包含多个具体项目，其中就覆盖了神经网络集成电路、神经计算机、各类传感器、计算机芯片以及半导体生产设备等可以用于人工智能开发与维护等硬件设施。在软件管制方面，实施出口管制的国家可能会对用于开发、生产或使用人工智能项目的软件实施控制。例如，《瓦森纳安排》（Wassenaar Arrangement）中设有一个专门针对军用软件的军需清单 ML21，其覆盖了用于“开发、生产和维护”某些特定军事装备的软件系统，该条款可被用于对人工智能软件进行出口管制。在技术管制方面，技术在出口管制中被定义为“开发、生产或使用产品所需的具体信息”。⁷⁹这一概念涵盖技术数据和技术援助两个方面：技术数据包括蓝图、图表、模型、公式、工程设计和规格、手册等实物；技术援助则包括指导、技能培训、工作知识、咨询服务等活动。在人工智能军控中，技术控制还被延伸到了训练特定人工智能系统的数据集，尤其是那些专为所列军事项目设计的数据。在全方位控制领域，全方位控制是一种严厉的出口管制手段，实施该手段的国家会对所有双重用途货物和技术实施严格的管制，对于具有高度军民两用性质的人工智能武器来说，全方位控制被认为是一个有用的工具，使国家能够在不引入广泛基于清单的控制的情况下实施控制。⁸⁰

国际组织也是军备控制的重要主体之一，其在人工智能军控中的作用主要包括制定相关的法规及国际标准，以及通过专门机构进行国际监督。作为当前国际体系中最重要国际组织，联合国（UN）在包括人工智能军控在内的多项全球治理议题中都扮演关键角色。目前，以联合国为核心的人工智能军控进程已经在法规和标准制定上取得了一定的成果。早在 2014 年，联合国《特定常规武器公约》（CCW）框架便开始就限制自主武器展开了一系列讨论⁸¹，而自主武器恰

77 Ministère de l'Économie, des Finances et de l'Industrie et Ministère chargé du Budget et des Comptes publics, “STRATÉGIE D'ACCÉLÉRATION « INTELLIGENCE ARTIFICIELLE »”, <https://www.entreprises.gouv.fr/fr/strategie-d-acceleration-intelligence-artificielle>, 访问时间：2024 年 10 月 6 日。

78 Reinhold, Thomas, and Niklas Schörnig. *Armament, Arms Control and Artificial Intelligence*. Berlin, Springer, 2022, p.196-199.

79 “Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies, PUBLIC DOCUMENTS, Volume II”, 访问时间：2024 年 10 月 11 日。

80 Reinhold, Thomas, and Niklas Schörnig. *Armament, Arms Control and Artificial Intelligence*. Berlin, Springer, 2022, p.198-199.

81 CCW, Report of the 2014 informal Meeting of Experts on Lethal Autonomous Weapons Systems, <https://documents-dds-ny.un.org/doc/UNDOC/GEN/G14/048/96/pdf/G1404896.pdf?OpenElement>, 访问时间：

恰是与人工智能关联最大，受人工智能技术赋能程度最高的一种武器类型。从2014年至2018年，CCW 缔约国之间围绕自主武器控制进行了三次非正式专家会议和三次政治专家会议，围绕相关技术、伦理、法律、军事问题进行了讨论。2019年CCW政府专家组会议提出了规范致命性自主武器系统的10条指导原则，并得到了缔约国全体会议的通过。原则具体包括：国际人道主义法适用于致命性自主武器系统；致命性自主武器的使用后果需由人类承担责任；致命性自主武器的开发、部署和使用应在国际法框架下受到监管；致命性自主武器系统的开发须充分考虑风险；在设计、开发、测试和部署致命性自主武器时必须做好风险评估和风险规避；讨论和采取人工智能军控政策时不应阻碍智能自主技术的发展或和平利用；应力求在军事必要性和人道主义考虑之间取得平衡，等等。⁸²以上原则的目标在于对快速发展的人工智能军事化实践加以人道主义原则、国际法和安全考虑的限制，从而在技术发展与风险管控之间取得平衡。

在2021年11月举行的联合国教科文组织（UNESCO）第41届大会上，193个会员国一致通过了《人工智能伦理问题建议书》，该文件拟定了规范人工智能发展应遵循的原则，特别强调了世界各国和国际组织在制定人工智能相关的技术伦理标准时应当充分尊重国际法，特别是国际人权法。⁸³在职能机构的设立方面，联合国的一系列机构为成员国之间围绕人工智能军控的合作提供了平台，其中包括联合国安全理事会（UNSC）、联合国裁军事务办公室（UNODA）、联合国教科文组织（UNESCO）等。值得注意的是，联合国安理会于2023年7月首次就人工智能问题举行了题为“人工智能给国际和平与安全带来的机遇与风险”的高级别会议，联合国秘书长古特雷斯在会上称，联合国正在考虑以国际原子能机构（IAEA）为蓝本，就人工智能问题建立一个以专业知识为基础、拥有一定监管权力的联合国机构，从而在有效管控风险的基础上尽可能地让人工智能为人类创造价值。⁸⁴2024年3月21日，联合国大会投票通过了第一个有关人工智能的决议草案，该草案的主题是消弭不同发展程度国家间在人工智能上的鸿沟，其中提到，联合国将促请所有联合国会员国不要使用对国际人权法和人类基本权益构成风险的人工智能系统，其中就包括了人工智能武器系统⁸⁵。但是，截至目前，联合国尚未正式通过关于人工智能国际军控的正式法规，相关机构的组建也还留在讨论阶段，联合国的角色仍主要集中在促进讨论、研究、倡议和制定原则等方面。

2024年9月5日。

82 CCW, “《致命性自主武器系统领域的新兴技术问题政府专家组2019年会议报告》”, <https://documents.un.org/doc/undoc/gen/g19/343/63/pdf/g1934363.pdf>, 访问时间：2024年6月3日。

83 联合国教科文组织, “人工智能伦理问题建议书”, https://unesdoc.unesco.org/ark:/48223/pf0000381137_chi, 访问时间：2024年10月8日。

84 United Nations, “Artificial Intelligence: Opportunities and Risks for International Peace and Security - Security Council, 9381st Meeting”, <https://webtv.un.org/en/asset/k1j/k1ji81po8p>, 访问时间：2024年6月1日。

85 United Nations, “联合国大会通过里程碑式决议，呼吁让人工智能给人类带来“惠益””, <https://news.un.org/zh/story/2024/03/1127556>, 访问时间：2024年6月1日。

（三）关于军控的困境

军备控制成功与否取决于两个维度：可取性（desirability）和可行性（feasibility）。其中，可取性指各国对武器军事价值与恐怖程度的感知与计算；可行性指各国对该武器施加实际约束的能力，其具体受到各国军控意愿、国际社会核查能力，国际社会对违约行为惩罚能力，合作国家数量等因素的影响。⁸⁶当前，国际社会关于人工智能国际军控的讨论已经达成了一系列目标与共识，这说明人工智能军控在可取性的维度上已经满足了一定条件。但是在可行性维度上，这一进程仍然受到来自多个方面的阻力，其中最主要的挑战来源于三个方面：大国间战略竞争、产业发展与风险管控矛盾、法律拘束力不足。

同其他国际军控议程一样，人工智能武器的治理中最大的问题在于无法对主权国家，特别是大国的行为进行有效规制。⁸⁷大国在军控中往往扮演最重要的角色：一方面，大国往往拥有广泛的战略利益，拥有充分的动机针对全球性议题发挥影响力；另一方面，大国也往往有着较多的资源和影响力来促进其愿景的达成，因此大国的参与往往贯穿了军控议程的全部流程，决定了军控最终是否能取得成功。但是，由于国家间在安全等核心利益上的冲突，人工智能军控进程也遵循了霸权稳定论的逻辑，即国际体系中的强权国家和国家集团可能也会依托其强大的实力制定出一套对其自身有利的国际规则，进而建立起一个等级性的国际体系，最终将其利益分配需求机制化和合法化，以达到持续压制竞争对手，塑造自身“重视并积极参与全球治理”的正面角色的目的。⁸⁸

以美国为例，大国间战略竞争的意图在美国关于人工智能技术发展和风险规制的一系列行动中得到了充分体现。美国国家安全咨询委员会（USNSC）在其于2023年11月发布《人工智能及相关技术对军备控制、防扩散和核查的影响报告》中，提出中国和俄罗斯的人工智能产业发展“对美国在人工智能军事化领域的领先地位构成了挑战⁸⁹”。基于这一担忧，美国政府正在基于北约、AUKUS等同盟框架，在加强内部合作的同时对中俄进行科技“脱钩”，避免本国人工智能产业链与中俄产生联系，从而维护美国在人工智能领域的领先地位。⁹⁰如今，以美国为核心的北约正在打造人工智能同盟，在该同盟内部，北约力图促进联盟内部工业界、学术界的融合，实现联盟中人工智能系统之间的互操作性，并建立一套统一

86 CNAS, “Artificial Intelligence and Arms Control,” <https://www.cnas.org/publications/reports/artificial-intelligence-and-arms-control>, 访问时间：2024年4月12日。

87 Benjamin Constant, “Autonomous Weapons Systems: A coming Legal “Singularity”, *Journal of Law, Technology & Policy*, Vol.45, No.1, 2013, pp. 47-48, 81.

88 高奇琦：《全球善智与全球合智：人工智能全球治理的未来》，载《世界经济与政治》2019年第7期，第25-27页。

89 International Security Advisory Board, ISAB, “Report on the Impact of Artificial Intelligence and Associated Technologies on Arms Control, Nonproliferation, and Verification”, https://www.state.gov/wp-content/uploads/2023/11/ISAB-Report-on-AI-and-Associated-Technologies_11172023-Accessible.pdf。访问时间：2024年5月7日。

90 Atlantic Council, “AI, Society and Governance: An Introduction,” <https://www.atlanticcouncil.org/in-depth-research-reports/ai-society-and-governance-an-introduction/>, 访问时间：2024年4月10日。

的监管标准。⁹¹在联盟外部，北约力图以一个整体的形象示人，共同在联合国、国际标准化组织、国际电信联盟等国际组织中争夺有关人工智能治理的话语权，打造一套有利于联盟的标准体系，促使其他国家通过主动和被动的的方式向北约标准靠拢。但是，这一套为战略竞争而生的联盟体系也并非铁板一块，尽管北约拥有先进的人工智能技术，和庞大的资源库，但是这些技术和资源仍由各国政府独立掌控，例如美国几乎完全独占了北约卫星系统等资源，并没有将相关技术和信息同其盟国完全分享。这反映出即便是在联盟内部，也存在着大量的博弈和分歧。

92

在推动技术发展和实现有效治理的矛盾方面，人工智能军控的军民两用性和快速迭代性导致国际社会很难在技术发展和风险管控之间取得平衡。第一，人工智能技术具有高度的军民两用性质，这使得对其军事化应用的识别变得困难。许多人工智能技术最初是在医疗、交通管理和生产等民用领域被开发和应用的。然而，这些技术也可以被适用于在军事领域。以美国为例，美国十分注重人工智能技术在军民融合中的发展应用，其大力鼓励高校、民营企业积极开展独立或与政府合作的人工智能项目，并通过 DARPA 将这些技术引入美军现役体系⁹³，这就导致国际社会对美国人工智能技术发展的最终用途知之甚少，难以对其风险行为进行及时的发现、预警和应对。这种发展模式的另一大风险在于，在未来，为人工智能军事系统提供电力、数据、软硬件设备的民用设施可能会成为敌方的打击目标，从而导致战争的升级。⁹⁴第二，即便国际社会能够有效厘清人工智能技术在军民领域的界限，传统的立法程序可能无法适应人工智能技术的快速迭代。人工智能技术具有机器学习和自我迭代的能力，其变化之快是以往科技创新难以比拟的，而传统的民主立法方式往往需要经过大量的社会讨论和法律流程，如果仍基于传统立法方式对人工智能展开治理，则很有可能导致法律发展落后于技术发展。⁹⁵在这种情况下，一些持保守立场的观点认为应该用严格的立法应对人工智能快速发展的风险，但是这可能导致过于严苛的立法阻碍人工智能技术的正常发展，甚至导致支持人工智能发展与反对人工智能发展两派之间的激烈冲突。

在法律拘束力不足方面，由于当前大多数国家都意识到了人工智能技术在军事领域内的强大潜力，因此现有规制体系往往很难对人工智能军事化进行有效规制。在这一背景下，目前最有可行性的方法是由国家、国际组织、企业等利益攸

91 NATO, "Summary of NATO's revised Artificial Intelligence (AI) strategy", https://www.nato.int/cps/en/natohq/official_texts_227237.htm, 访问时间：2024 年 12 月 10 日。

92 GLOBSEC, "NATO's Intelligence Adaptation Challenge," <https://www.globsec.org/sites/default/files/2018-03/NATOs-intelligence-adaptation-challenge.pdf>, 访问时间：2024 年 4 月 11 日。

93 CSIS, "The DARPA Perspective on AI and Autonomy at the DOD", <https://www.csis.org/events/darpa-perspective-ai-and-autonomy-dod>, 访问时间：2024 年 12 月 3 日。

94 Heather Roff, "To Ban or Regulate Autonomous Weapons: A US Response," *Bulletin of the Atomic Scientists*, Vol.72, No.2, 2016, pp.112-124.

95 清华大学战略与安全研究中心，傅莹，“《人工智能的治理与国际合作》——2019 上海世界人工智能大会上的发言，访问时间：2024 年 4 月 10 日。

关方对人工智能进行“共治”⁹⁶，其具体手段是由多主体共同通过设置原则、指导方针、政策建议、行业标准等“软法”来实现治理。这些“软法”往往更具灵活性，更容易被各种跨国主体所接受。在现实中，英国、美国和欧盟在进行人工智能治理时均重视对“软法”的使用，例如英国在 2021 年提出了关于人工智能发展的五项原则，具体包括安全性和稳健性、透明和可解释、公平性、可问责和管理、可竞争性和救济性。⁹⁷美国要求人工智能行业从业者完善内部管理，加强行业自律。⁹⁸欧盟则重视对于风险的管控，其具体方式是先通过政策建议进行“软性治理”，如仍有违规行为发生则引入惩罚措施。⁹⁹但是，这些基于多方监管甚至自我监管的“软法”的作用机制是通过市场手段给予利益相关者一定的激励¹⁰⁰，无法在国家间核心利益冲突中突破国家主权的管辖范围，实现对违规行为的有效惩治。在国际竞争持续加剧的情况下，各个国家基于自身国家利益所出台的“软法”更是难以在国际层面获得普遍共识。例如，美国国内的人工智能治理以“创新优先”为特征，这与以“伦理优先”为特征的欧盟人工治理模式之间存在张力。在七国集团共同建立的全球人工智能合作伙伴关系（Global Partnership on AI, GPAI）框架下，美国时常批评其欧洲盟友对人工智能技术伦理的担忧属于“官僚主义”，如果在 GPAI 各个项目中过多采用非美国的治理标准，则可能遏制自身在人工智能技术上的创新能力。¹⁰¹

在人工智能在军事领域广泛应用已经成为一个现实挑战的当下，人工智能军事化涉及的复杂性和潜在威胁使得全球治理变得尤为迫切，这促使着各国在国际军控框架下围绕人工智能议题展开进一步的讨论。

四、人工智能军控的中国路径

二战后，世界各国为了遏制战争一直在进行着长期而不懈的努力。在这个过程中，国际社会逐步建构起相对完备而有效的国际军控体系，为维护全球安全和稳定打下了坚实基础。中国是国际军控进程的重要参与者，正如中国政府于 2023 年提出的《全球安全倡议概念文件》所言，中国秉持共同安全理念，主张通过合作方式尊重和保障各国安全，在军控领域，中国支持在联合国安理会防扩散委员

96 朱荣生，王雅南：《人工智能全球安全治理的三重困境及我国的政策选择》，载《人工智能》2022 年第 1 期，第 47 页。

97 GOV.UK, “A pro-innovation approach to AI regulation”<https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>. 访问时间：2024 年 10 月 19 日。

98 National Institute of Standards and Technology, “AI Risk Management Framework” <https://www.nist.gov/itl/ai-risk-management-framework>, 2024 年 10 月 19 日。

99 European Commission, “Artificial Intelligence for Europe,” <https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/artificial-intelligence-europe>. 访问时间：2024 年 10 月 19 日。

100 “Soft Law’ Governance Of Artificial Intelligence,” Brookings Institution, <https://www.brookings.edu/articles/how-soft-law-is-used-in-ai-governance/>, 访问时间：2024 年 4 月 16 日。

101 Center for a New American Security, “White House Favors a Light Touch in Regulating AI”, <https://www.cnas.org/press/in-the-news/white-house-favors-a-light-touch-in-regulating-ai>, 访问时间：2024 年 10 月 15 日。

会等框架下，推动全球各项军控进程。¹⁰²在人工智能国际军控面临多重困境的当下，中国正在通过建立人工智能大国政府间对话机制、坚持发展与安全并重原则，完善多边普惠型人工智能安全治理体系三条路径，积极推动人工智能国际军控的进展。

第一，为了应对大国间竞争，中国主张完善现有人工智能大国政府间对话机制。在全球治理实践中，国家间出于交换信息，弥合分歧，防止误判等需要，往往会就某些关键议题建立正式或非正式的对话机制。当前，科技竞争已经成为大国间竞争的关键领域，以美国拜登政府为代表的西方大国特别重视将科技竞争作为同他国开展战略竞争的抓手。¹⁰³对于中国而言，参与人工智能军控既是维护国家安全和利益的重要议题，也是中国积极参与全球治理，承担大国责任的应行之举。中美都是人工智能大国和全球具有高度影响力的国家，中美之间能否围绕人工智能治理达成共识，影响着人工智能全球治理的成败。当前，中国正致力于与美国围绕人工智能治理开展沟通与协调，努力减少信息不对称情况下对于彼此人工智能技术发展的威胁认知。2023年11月15日，中国国家主席习近平和美国总统拜登在旧金山进行会晤，席间两国宣布，将共同建立中美人工智能政府间对话机制。¹⁰⁴这一对话机制的核心关切是确保人工智能的应用和发展不会对中美战略稳定带来负面影响和挑战。2024年5月14日，中美在瑞士日内瓦举行中美人工智能政府间对话首次会议，会上，中美双方在治理议题、治理模式、治理目标上都存在某些分歧，例如，中国强调人工智能治理需要遵循“有益，安全，公平”的原则，而美国则强调“安全，可靠，值得信赖”，同时两国对于“安全”的定义不尽相同。¹⁰⁵但是，两国都意识到了通过对话方式共同开展治理的必要性，这种对话本身就代表着一种进步。¹⁰⁶

第二，为了实现技术发展与风险管控的平衡，中国在人工智能技术的开发与应用中坚持发展与安全并重原则。这种原则首先体现在中国发布的一系列人工智能治理政策文件中。2023年10月18日，中国中央网信办发布《人工智能全球治理倡议》，其中明确指出，一方面，中国主张在人工智能全球治理中坚持“智能向善”原则，呼吁各国在发展和应用人工智能技术时采取审慎态度，反对将人工智能技术用于开发完全自主攻击性武器、操纵舆论，传播虚假信息，干涉他国内

102 中华人民共和国中央人民政府：“全球安全倡议概念文件（全文）”，https://www.gov.cn/xinwen/2023-02/21/content_5742481.htm，访问时间：2024年6月10日。

103 U.S. Department of Defense, "Responsible Artificial Intelligence Strategy and Implementation Pathway", <https://www.ai.mil/Portals/137/Documents/Resources%20Page/DoD%20Responsible%20AI%20Strategy%20and%20Implementation%20Pathway.pdf>，访问时间：2024年11月15日。

104 中华人民共和国外交部：“习近平同美国总统拜登举行中美元首会晤”，https://www.mfa.gov.cn/web/zyxw/202311/t20231116_11181125.shtml，访问时间：2024年6月20日。

105 China-US Focus, "China and the United States Begin Official AI Dialogue", <https://www.chinausfocus.com/peace-security/china-and-the-united-states-begin-official-ai-dialogue>，访问时间：2024年11月20日。

106 复旦发展研究院：“中美人工智能政府间首次对话，最前沿的合作效果如何？”<https://fddi.fudan.edu.cn/50/5f/c21253a675935/page.htm>，访问时间：2024年6月21日。

政等领域。另一方面，中国主张建立风险等级测试评估体系以规范人工智能技术发展。通过对人工智能技术实施分类分级管理，各国能够对特定领域内的人工智能技术发展进行促进或限制。¹⁰⁷该做法能够提高人工智能技术发展的可解释性和可预测性，将人工智能技术的研发、生产和应用严格限制在对整体人类社会有益的范围内，使人工智能技术发展符合“可审核、可监督、可追溯、可信赖”的原则要求。在中国外交部于 2024 年 1 月发布的《中国关于规范人工智能军事应用的立场文件》中，中国再一次提出了加强对人工智能技术分级、分类管理的主张，强调各国应该采取必要措施加强对人工智能军事化风险的研判。¹⁰⁸除提出一系列政策指导文件以外，中国作为人工智能大国，正在不断提高自身人工智能技术发展透明度，进而帮助各国理解中国的人工智能技术水平和应用领域，减少由于信息不对称所导致的不安全感。近年来，中国连续举办了世界人工智能大会（WAIC）、世界互联网大会（WIC Wuzhen Summit）、世界电信标准化全会（ITU）等大型国际会议，真实完整地向世界展示了中国在人工智能发展上透明的态度。最后，由于人工智能技术牵涉到人类社会的一系列前沿科技成果，具有高度的复杂性和专业性，只有行业专家才能完整准确地实现对人工智能武器的物理特征、技术性能、可能用途进行更为精准的判断，提高核查等军控措施的准确性和有效性，防止军控进程对各国正当的技术进步造成阻碍，进而实现“有效的人类控制。”¹⁰⁹因此中国十分重视邀请全球智库和专家参与治理合作。例如，从 2017 年开始，来自中国的人工智能专家长期参与联合国《特定常规武器公约》框架下“致命自动武器系统监管专家组”（GGE）的合作¹¹⁰，为人工智能军控贡献了中国智慧。

第三，为了应对当前现有军控机制法律拘束力不足的问题，中国正在推动构建多边普惠型人工智能安全治理体系。目前人工智能全球治理法律拘束力不足的主要原因在于，作为全球治理主要规范形式的“软法”制约国家行为的机制在于对国家基于理性的成本-收益判断进行影响¹¹¹，这种制约本身是一种市场机制，只有市场手段才能强化或削弱这种制约。在人工智能安全治理领域，构建多边普惠型治理体系能够有效影响国家对于履约收益的预期，从而促进现有军控机制法律拘束力的提高。当前，中国正在基于联合国、上合组织、金砖组织等多边框架，积极推动构建多边普惠型的人工智能安全治理体系。2024 年 7 月 1 日，第 78 届

107 中华人民共和国互联网信息办公室：“全球人工智能治理倡议”，https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm，访问时间：2024 年 6 月 21 日。

108 中华人民共和国外交部：“中国关于规范人工智能军事应用的立场文件”，https://www.fmprc.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/zclc_674645/rgzn/202206/t20220614_10702838.shtml，访问时间：2024 年 6 月 22 日。

109 Davis, Jenny “Elevating humanism in high-stakes automation: experts-in-the-loop and resort-to-force decision making.” *Australian Journal of International Affairs*, vol.78, No.2 2024, pp200-209.

110 联合国：“致命性自主武器系统领域新兴技术问题政府专家组 2023 年会议报告”，<https://documents.un.org/doc/undoc/gen/g23/107/42/pdf/g2310742.pdf?token=RXkUd7BTgxDzlbQzP4&fe=true>，访问时间：2024 年 6 月 22 日。

111 Brookings Institution, “Soft Law’ Governance Of Artificial Intelligence,” <https://www.brookings.edu/articles/how-soft-law-is-used-in-ai-governance/>，访问时间：2024 年 4 月 16 日。

联合国大会一致通过了由中国主提，超过 140 多个国家支持的关于加强人工智能能力建设国际合作决议。该决议提出，当前人工智能技术正处于高速发展阶段，但是人工智能技术发展的成果并没有被世界各国广泛共享，发达国家与发展中国家之间存在相当程度的人工智能鸿沟。¹¹²在人工智能军事化成为一种普遍趋势的当下，人工智能技术的无需发展可能导致国家间军事力量的进一步失衡，从而威胁到国际安全。该决议主张，各国应该在“以人为本、智能向善、造福人类”的原则指导下，共同支持联合国在人工智能全球治理中发挥中心作用，帮助各国加强和规范人工智能能力建设，实现人工智能包容普惠可持续发展。此外，构建多边普惠型人工智能安全治理体系的一个重要路径是将跨国公司、社会组织、各国学者等非国家行为体引入人工智能军控议程。这些非国家行为体能够通过宣传舆论、政策建议等方式推动促进军控合作的达成。¹¹³自 2018 年以来，中国连续 7 年主办世界人工智能大会，会上邀请各国企业高管、政策专家和专业学者对人工智能发展、人工智能应用、人工智能治理等议题开展广泛通论。2024 年，中国在第七届世界人工智能大会上发布《人工智能全球治理上海宣言》，宣言在人工智能安全治理领域，针对加强数据保护，推动网络安全、防范非法活动、制定伦理指南等议题提出了主张。该宣言的新颖之处在于，其提出人类可以“以人工智能技术防范人工智能风险”¹¹⁴，这一思路反映出了人类对于人工智能技术探索正在不断深入，对未来的人工智能治理具有重要的指导意义。

尽管当前中国正在积极推动完善建立围绕人工智能军事化的全球治理体系，但是这一努力也面临着重重挑战。其中，大国间战略竞争问题属于政治范畴，创新与安全平衡问题属于技术范畴，而现有军控机制法律拘束力不足既牵涉到国家间利益冲突的政治问题，也涉及到人工智能高核查难度的技术问题。因此，在未来，中国围绕优化现有人工智能军控路径的努力可以聚焦于政治和技术两个方面。

在政治方面，中国应该在未来的国际合作中继续坚定奉行“真正的多边主义”。真正的多边主义最早由中国国家主席习近平在 2021 年博鳌亚洲论坛年会开幕式上提出¹¹⁵，与西方带有强烈小圈子色彩的“有限的多边主义”不同，真正的多边主义以人类命运共同体理念为核心价值，主张通过相互尊重、平等协商的方式解决国家间分歧。¹¹⁶在全球治理领域，真正的多边主义主要体现在三个方面：维护以联合国为核心的多边体系，促进现有多边治理机制变革，创设多边治理新机制新

112 中华人民共和国中央人民政府：“联大通过中国提出的加强人工智能能力建设国际合作决议”，https://www.gov.cn/yaowen/liebiao/202407/content_6960524.htm，访问时间：2024 年 6 月 20 日。

113 Maas, Matthijs M. “How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons”, *Contemporary Security Policy*, Vol. 40, No. 3, February 2019, pp. 285-311.

114 中华人民共和国外交部：“人工智能全球治理上海宣言”，https://www.fmprc.gov.cn/web/zyxw/202407/t20240704_11448293.shtml，访问时间：2024 年 7 月 5 日。

115 中华人民共和国中央人民政府，“习近平在博鳌亚洲论坛 2021 年年会开幕式上发表主旨演讲”，https://www.gov.cn/xinwen/2021-04/20/content_5600759.htm，访问时间：2024 年 10 月 10 日。

116 卢静，张甲英：《走向真正的多边主义》，载《国际问题研究》2023 年第 6 期，第 25 页。

平台。¹¹⁷在未来，中国可以同金砖国家、上海合作组织、全球南方等国际成员一道，在联合国框架下推动构建更具普惠性的人工智能军事化监管和治理平台。

在技术方面，为了应对人工智能技术在军事领域的滥用风险，中国首先需要进一步完善自身人工智能治理体系。当前，中国已经出台了包括《科技伦理审查办法（试行）》等法律文件以规范人工智能发展。¹¹⁸但是，中国并没有一部专门的《人工智能法》。因此中国在未来应该推进相关立法工作以规范和促进自身技术发展，具体的思路应该是坚定某些基本原则，并为未来技术发展留出足够的“接口”。¹¹⁹在完善自身治理体系以后，中国应该在未来的人工智能国际军控进程中与国际社会积极分享相关经验。为了提高核查能力，目前一个重要思路是“用技术控制技术”。核查的核心是信息的获取和分析，而人工智能技术本身在这些方面具有明显的优势。在针对人工智能武器的核查过程中，可以通过人工智能技术识别并分析核查对象的图像数据、设备和软件参数、传感器数据等数据，从而实现对技术参数和潜在用途不清晰的人工智能软硬件是否违反了相关军控条例进行判断。¹²⁰在未来，中国可以同国际社会一道建立相关核查组织，共同开发用于核查的人工智能系统，实现更加高效和准确的核查。

五、结语

作为当前人类社会最具革命性的科学技术之一，人工智能技术在军事领域具有广阔的应用前景，对包括作战武器智能化、军事战略战术升级、情报分析、作战模拟在内的未来战争形态产生了深远的影响。当前，世界主要大国均推出了自身的人工智能军事化战略，但是该趋势也引发了国际社会关于其安全、法律、伦理、技术风险的一系列担忧。如今可以预见的是，人工智能军事化已经对人类社会构成了潜在威胁，对其展开治理已经成为了国际社会的普遍共识。但是在实际的治理过程中，国际社会仍面临着大国权力竞争、技术发展与风险管控矛盾、现有规制体系拘束力不足等阻力。当前，中国正在积极采取行动，通过建立人工智能大国政府间对话机制、坚持发展与安全并重原则，完善多边普惠型人工智能安全治理体系三条路径，积极参与人工智能的国际军控进程。在未来，中国可以从坚定奉行“真正的多边主义”、完善自身人工智能治理体系并与国际社会分享经验、“用技术控制技术”三个角度继续完善现有军控路径。

正如《中国关于规范人工智能军事应用的立场文件》所述，各国在研发、部

117 吴志成，徐信高：《推动全球治理更加公正合理的中国视角与实践》，载《世界经济与政治》，2024年第4期，第21-23页。

118 中华人民共和国中央人民政府，“关于印发《科技伦理审查办法（试行）》的通知”，https://www.gov.cn/gongbao/2023/issue_10826/202311/content_6915814.html，访问时间：2024年10月16日。

119 张凌寒：《中国需要一部怎样的〈人工智能法〉？——中国人工智能立法的基本逻辑与制度架构》，载《法律科学（西北政法大学学报）》，2024年第3期，第14-15页。

120 Reinhold, Thomas and Niklas Schörmig. *Armament, Arms Control and Artificial Intelligence*. Berlin, Springer, 2022, p.56-70.

署和使用人工智能武器系统时，应遵循人类共同价值观，坚持以人为本，秉持“智能向善”的原则。¹²¹在未来，包括中国在内的国际社会需要共同努力，通过合作方式引导人工智能技术的有序发展和合理应用，为人类社会发展与安全做出更大贡献。需要注意的是，2016年以后兴起的生成式人工智能（GAI）技术是人工智能技术最新的发展阶段，相比传统基于机器学习的人工智能技术，生成式人工智能在智能性、自主性上都表现出更高的水平¹²²，表现出广泛的应用前景和更大的军事潜力，在未来，围绕生成式人工智能的技术竞争和军事化应用可能成为人工智能军控的前沿议题。综合来说，人类社会必须认识到人工智能技术的两面性，在充分考虑人类价值观的基础上，善用人工智能，使技术发展能够真正地为人类社会造福。

121 中华人民共和国外交部：“中国关于规范人工智能军事应用的立场文件”，https://www.fmprc.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/zclc_674645/rgzn/202206/t20220614_10702838.shtml，访问时间：2024年6月12日。

122 “The Three Horizons of AI Policy,” Center for Strategic and International Studies (CSIS), <https://www.csis.org/podcasts/ai-policy-podcast/three-horizons-ai-policy>，访问时间：2024年4月14日。

跨国公司参与数据跨境流动治理的生成逻辑及风险应对

李 智¹ 刘思雨²

摘 要：跨国公司参与数据跨境流动治理是硬法规制缺位和私利优势驱动的共同结果。在金融危机的冲击下，跨国公司从治理推动者成为责任承担者，而数字化的迅速发展为其凭借技术和资源优势回归全球治理舞台提供契机。然而，利益的冲突、监管的困境以及对国家立法的冲击，或将引发多维风险。目前，跨国公司通过行业自律、参与立法等方式填补数据跨境流动硬法规制的缺失。因此，可适度借鉴网络空间“主权分层”理念，融合功能属性分层与安全关联度分层双重标准，坚持以主权国家为主导并充分发挥跨国公司的补充功能，构建主权分层模式下的数据跨境流动协同共治机制，实现合作共赢的全球数据治理新格局。

关键词：数据跨境流动；跨国公司；协同治理；数据安全

前言

在全球化与数字化融合发展的进程中，数据跨境流动通过促进技术、资本等要素的全球化配置，逐渐成为支撑经济发展的重要战略领域。对此，主权国家通过国内立法、国际条约等硬法平衡数字贸易自由化与国家数据安全的同时，跨国公司等非国家行为体也试图凭借专业技术优势和资源整合能力在数据跨境流动治理中占据一席之地，以期提升企业影响力并获得竞争优势。然而，多元治理主体间的张力关系与制度真空，可能消解治理效能甚至诱发数据安全风险。近年来，学界持续关注数据跨境流动治理中公共行为体和私人行为体之间的角色和责任的适当分配问题（International Law Association 2021，罗旷怡、黄志雄译），结合行业实践分析跨国公司对全球数据治理国际法律制度的影响（胡玲，2023），强调私人主体对数据跨境流动治理的补充功能（包柠榛，2024），并提出防范网络空间国际造法“国退民进”态势下非国家行为体对国家间法律关系的冲击（黄志雄、罗旷怡，2023）。因此，基于既有研究范式，进一步明确跨国公司在数据跨境流动治理中的角色定位，系统解析多元主体权益竞合下的风险挑战，进而探究因应之策，对于实现主权国家与跨国公司数据跨境流动治理的良性互动，推动数据主权时代公域监管与私域创新的协同共治，将有所裨益。

作者简介：1.李智，法学博士，福州大学法学院教授，博士生导师。E-mail: lizhi@126.com。
2.刘思雨，福州大学法学院国际法学硕士研究生。E-mail:17777016983@163.com。

一、回归：数字化背景下跨国公司参与全球治理的角色嬗变

经济全球化背景下，全球治理应运而生。国际社会的多元行为体为实现共同目标、增进共同利益，在全球层面的各个领域通过正式或非正式的制度安排而采取合作行动。¹就跨国公司来说，其在全球治理架构中的功能定位正经历重构。

（一）从“赋权”到“制衡”：全球治理体系下跨国公司的权责重构

跨国公司因资本全球化实现快速扩张并成为全球治理的重要力量，又随着2008年国际金融危机的爆发势力消减。

第一，赋权扩能：跨国公司因经济实力崛起深度参与全球治理体系构建。冷战结束后，全球化加速运行，跨国公司成为推动经济增长的重要力量，成功跻身全球治理体系构建行列：一方面，跨国公司对外输出自己的技术标准和市场规则间接影响全球贸易。例如，微软公司推出 Windows 95 后，与全球经销商签订协议要求计算机制造厂商必须附加安装微软的 IE 浏览器作为继续提供该系统的条件，²导致其他系统被边缘化。另一方面，跨国公司通过游说、规则嵌套等方式，渗透主权国家国内立法，影响国际条约的制定。例如，在国内立法层面，高盛、摩根大通等金融机构的高管兼具规则制定者、政府指挥者等多重身份，可以在监管者和被监管者之间轻易转换，试图通过权钱交易、官商勾结实现对美国国内治权的渗透和控制。³在国际条约制定层面，美国知识产权领域的跨国公司通过游说欧洲、日本等国家的跨国公司和产业集团形成利益共同体对抗发展中国家，使其被迫在《关税与贸易总协定》（简称 GATT，即后来的 WTO）“乌拉圭回合”谈判后接受《与贸易有关的知识产权协议》（简称 TRIPS 协议）。⁴这一阶段，跨国公司凭借经济硬实力，在全球治理体系占据一定主动权。对此，主权国家也多持宽松型策略，更多关注跨国公司的经济权利保护，较少对其设定责任义务。然而，2008 年全球金融危机的爆发对这种权利与义务严重失衡的自由化立法敲响了警钟，对资本的权利进行必要限制并赋予相应的责任成为国家间经济立法的新理念。⁵

第二，制衡归责：金融危机爆发重塑跨国公司权责平衡治理框架。解构此结构性变迁的深层逻辑，主要源于双重要素的耦合：其一，金融危机暴露了跨国公司过度逐利的本质，进而引发公众信任危机。以“大而不倒”的金融机构为例，金融危机发生时政府可能使用纳税人的钱对其实施救助，然而，金融机构过度冒险产生的收益却收归机构自己所有。换言之，金融机构将冒险的成本和后果交由社

¹ 参见刘清才、张农寿：非政府组织在全球治理中的角色分析[J]，国际问题研究，2006，(01)：48-52。

² 参见王先林：从微软垄断案看知识产权滥用的反垄断控制[J]，法学家，2001，(03)：93-99。

³ 参见高攀、江涌：从美国金融危机看金融资本对治权主权的渗透和干预[J]，华北电力大学学报(社会科学版)，2019，(01)：75-83。

⁴ 参见马忠法：国际知识产权法律制度发展趋势及中国应对[J]，经贸法律评论，2024，(03)：1-23。

⁵ 刘志云：后危机时代国家间跨国公司社会责任制度体系的构建[J]，政法论丛，2015，(03)：120-128。

会承担，⁶导致公众对金融机构的不信任，影响跨国公司参与全球治理的民众基础。其二，跨国公司逐渐成为大国博弈“工具”，触发主权国家政策渗透防御。较为典型的，从制裁中兴通讯到限制华为芯片供应，再到要求 Tiktok 剥离，无不体现美国遏制我国科技企业崛起以维护其技术霸权的企图。基于此，民众的不信任导致跨国公司标准输出受阻，主权国家的警惕使其制度性游说及规范嵌套策略逐渐失效。在双重要素的交互作用下，跨国公司全球治理话语权呈现式微态势，其角色定位由国际法上的赋权对象向义务承担实体转变，国际社会转而关注其人权保护义务、社会责任等问题。⁷

然而，数据要素的全球化配置重构了跨国公司参与全球治理的权力基础，其通过跨境数据资源优势构建技术标准话语权，依托算法治理能力参与国际规则制定。凭借这一技术赋能效应，跨国公司在数字时代治理体系中呈现回归态势。

（二）“回归”态势：源于数据跨境流动治理的技术需求

近年来，数字技术创新的战略价值显著提升，加大了数据治理的现实需求。然而，现有国际法律框架存在制度性滞后，叠加主权国家治理能力结构性缺陷，客观上催生了公私合作治理实践。得益于专业技术优势和资源整合能力，跨国公司逐步演化为数据跨境流动合作治理的重要行动者。

首先，依托数据资源“硬”实力与技术开发应用“软”实力，跨国公司参与数据跨境流动全球治理的效能逐渐凸显。第一，公司在数据资源层面占优。伴随着数字化持续渗透社会生活，全球数据总量呈指数式增长。据国际数据公司（以下简称 IDC）预测，2025 年全球数据领域将达到 163ZB（即 1 万亿 GB）。⁸与此同时，随着数据来源、使用情况和价值的变化，数据的创建正从消费者驱动转向企业驱动。截止 2015 年，公司创建的数据不到 30%；而到 2025 年，这一数字将接近 60%。⁹此外，IDC 统计数据还表明，从 2019 年开始，公司核心中存储的数据将超过世界上所有现有端点中的数据。¹⁰诚然，公司由此成为数据跨境治理的重点规制对象，但个别科技公司也得益于此，在全球数据市场占据相当的话语权，成为支撑其参与数据跨境流动合作治理的“硬”实力基础。第二，公司的技术开发应用能力占优。公司作为网络空间的研发者和使用者，享有较大的自主性和创造性。

6 参见隋平.金融机构“大而不倒”问题及其法律规制[J].江西社会科学,2014,34(01):161-166.

7 何志鹏.国际法要论[M].北京大学出版社:202306.505.

8 David Reinsel John Gantz John Rydning.Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big. IDC & Seagate April 2017,<https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>,最后访问于 2025-03-25.

9 David Reinsel John Gantz John Rydning.Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big. IDC & Seagate April 2017,<https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf> 最后访问于 2025-03-25.

10 David Reinsel John Gantz John Rydning,The Digitization of the World From Edge to Core.IDC& Seagate November 2018,<https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf?gid=164649> 最后访问于 2025-03-25.

¹¹因此，一方面，公司能够通过开发和运营各种网络服务获得经济利益和社会影响力，进一步提高自身实力；另一方面，也得以将技术应用贯穿数据采集、传输、存储、销毁等各阶段，在合法合规的前提下对数据实施全生命周期管理，为其参与数据跨境流动治理提供“软”实力支撑。

其次，数据跨境流动硬法规制缺位叠加传统治理方式局限，客观上为跨国公司通过非正式机制介入治理提供潜在空间。当前，主权国家主导下的硬法规制无法满足数字经济快速发展对跨国法律规范提出的增量要求，数据跨境流动治理面临制度缺位。这体现在国际规则和国内立法两个方面：其一，国际社会缺乏行之有效的统一标准。当前，关于数据的国际法仍处于萌芽阶段，除欧盟法以外，很少有专门针对数字数据并具有约束力的国际法规则。¹²与数据跨境流动有关的法律与政策呈现碎片化特征。其二，各国对待数据跨境流动存在原则性分歧，增大了数据跨境流动的合规压力。一方面，在平衡基于数据安全的流动控制与基于数字贸易的自由流动基础上，¹³各国采取了不同的治理方案。例如，美国基于商业利益倡导数据跨境“宽进严出”策略，欧盟注重保障数据主体权利主张“外严内松”治理范式。另一方面，以欧盟、美国为代表的部分国际法主体企图借助布鲁塞尔效应（Brussels Effect）、长臂管辖（Long-arm Jurisdiction）等途径实现本国数据法的不当域外适用以维护自身技术霸权，这也导致部分国际经贸协定中的数据跨境流动规则可能背离促进数据自由流动的初衷，从而阻碍数字经济的发展。因此，面对数字时代数据跨境的迫切需求，硬法规制的缺位将为软法介入提供了一定空间。与此同时，主权国家传统治理技术较难满足数据治理的新兴要求，跨国公司恰好能够凭借技术性能力补充传统治理创造之不足。一方面，数据固有的流动性和无形性决定其传输必须依靠云服务、区块链等基础设施或技术手段，而主权国家的传统监管方式受到国家间物理边界管控等限制，甚至容易引发管辖冲突之争。另一方面，各国发展水平存在差异，发展中国家在技术、专业人才等方面存在不同程度的短板，难以建立与发达国家对等的监管能力，若仅仅依靠主权国家的参与，或将引发治理权力失衡，进一步扩大全球数字鸿沟。综上，传统治理主体的数据治理方式难以匹配当前数据跨境流动发展需求，为具备技术优势的私人主体承接部分治理权力¹⁴创造可能性。

再次，国际社会逐步认识到跨国公司参与数据跨境流动治理的重要作用。一是在国际实践层面，为打击犯罪，国家、国际组织可能需要某些公司的配合来收集数字证据。例如，2015年圣贝纳迪诺的枪击事件后，美国联邦地区法院法官要

11 参见黄志雄、罗旷怡.论国际法的“数字化转型”——兼评《国际法的数字挑战》白皮书[J].武大国际法评论,2023,7(03):27-48.

12 参见罗旷怡、黄志雄.《国际法的数字挑战》白皮书[J].武大国际法评论,2023,7(03):1-26.

13 参见张琨蓓、陈星月.跨境数据流动法律规制协调性之检视与重塑[J].重庆社会科学,2022,(03):87-96.

14 参见包柠榛.私人主体对数据跨境流动治理的补充功能：生成逻辑、参与路径与启示[J].国际经贸探索,2024,40(01):104-116.

求苹果公司为联邦调查局(以下简称 FBI)解锁嫌犯手机提供“合理的技术援助”,但苹果公司以保护数据安全和用户隐私为由拒绝执行此项命令。¹⁵又如,2019 年国际法院审理冈比亚诉缅甸违反《防止及惩治灭绝种族罪公约》一案,核心焦点的调查结果可能部分取决于它能否迫使 Meta 公司交出相关数据。¹⁶二是在战略规划层面,国际社会也开始强调非国家行为体在数据跨境流动治理中的作用,支持私营主体参与数据跨境流动治理。例如,2022 年 5 月,七国集团(Group of Seven,简称 G7)的数字部长级会议发布宣言称,支持七国集团内部开展合作并与志同道合的伙伴开展合作,制定由私营部门主导、自愿和基于共识的开放标准。¹⁷又如,世界经济论坛发布的《信任下的数据自由流动:克服跨境数据流动的障碍》提出,在政府间层面进行政策对话的同时,多方利益相关者之间也进行了积极讨论,通过公共和私人伙伴关系帮助和补充有关跨境数据政策的政府间对话。¹⁸对此,有观点指出,如果将互联网看作海洋,那么谷歌、脸书、微软和推特等互联网巨头,可以比作是传说中控制海洋的神。¹⁹

综上,数字经济发展对数据跨境流动领域法律制度的创制提出强烈需求,而以“各国共同同意”为基础的传统国际立法模式制约着法律制度的供给。²⁰这就需要引入“跨国立法”的多元造法机制,承认各种非国家行为体制订软法的权力,²¹从而满足数据跨境流动的现实需求。

二、定位:跨国公司参与数据跨境流动治理的补充功能

在技术优势和硬法缺位的共同驱动下,跨国公司能够依托技术标准统一优势与跨境数据治理能力填补国际统一实体规则的缺位,并对既有国际标准体系的技术性缺陷发挥矫正功能,逐渐回归数据跨境流动全球治理舞台。不过,受限于主体地位和规则性质,跨国公司难以居于体系中心,主要通过行业自律和参与立法起补充性作用。

(一)“补充”的认定:性质使然

主体行为效力的认定须以其法律地位的界定为前提。数据跨境虽归属于生产要素的流动,但数据治理往往也蕴含国家数据主权、基本权利保护等诸多公法核心价值,无法冲破既有公法规范体系。对此,主权国家凭借硬法规制的强制性效力始终居于主导地位,而跨国公司基于软法规则的治理参与仅具有从属性和依赖

15 参见吴子越,网络平台数据刑事调取的理论与规制路径[J],西南政法大学学报,2024,26(05):67-82.

16 罗旷怡,黄志雄,《国际法的数字挑战》白皮书[J],武大国际法评论,2023,7(03):1-26.

17 <https://www.bundesregierung.de/resource/blob/998440/2038510/e8ce1d2f3b08477eeb2933bf2f14424a/2022-05-11-g7-ministerial-declaration-digital-ministers-meeting-en-data.pdf?download=1>,最后访问于 2025-03-23。

18 World Economic Forum,Data Free Flow with Trust:Overcoming Barriers to Cross-Border Data Flow,January 2023, https://www3.weforum.org/docs/WEF_Data_Free_Flow_with_Trust_2022.pdf,最后访问于 2025-03-23。

19 See Polčák R, Svantesson D J B. Information sovereignty: data privacy, sovereign powers and the rule of law[M]. Edward Elgar Publishing, 2017.

20 参见徐崇利,经济全球化与跨国经济立法模式[J],华东政法学院学报,2006,(02):79-88.

21 参见徐崇利,全球治理与跨国法律体系:硬法与软法的“中心—外围”之构造[J],国外理论动态,2013,(08):19-28.

性。

首先，跨国公司本质上属于私法主体，其法律地位受到三重结构性限制：其一，在主体地位层面，国际社会关于跨国公司是否完全具备国际法主体资格尚无定论，导致跨国公司难以按照自身意志独立地享有权利和承担义务。其二，虽然跨国公司通过对外直接投资在国外设立分支机构或子公司，按照自己的“全球战略”形成了一体化的企业体系，但一般仍以一国为基地，受一国大企业的控制、管理和指挥。²²因此，跨国公司必须服从有关国家的属地管辖和属人管辖，只能在国家政策和法律允许的范围内参加国际和国内活动。²³其三，尽管在经济实力上个别跨国公司“富可敌国”，但其参与全球治理仍需要遵循以《联合国宪章》为核心的国际公法体系，以及 WTO 协定、《关于解决各国和其它国家国民之间投资争端的公约》等国家行为体主导下的多边规则。概言之，跨国公司参与数据跨境流动治理本质上受制于国家管辖权的属地优越性，其规则制定权也不具备完全的自主性，而必须从属于主权国家通过国内立法和国际条约构建的强制性规范框架。据此，跨国公司参与全球治理的权力基础具有从属性，这从根本上决定其难以居于数据跨境流动全球治理体系的主导地位。

此外，大多数跨国公司参与构建的国际行业规范属于国际软法，缺乏严格意义上的强制力，致使全球治理实效仍需依赖于主权国家。在数据跨境流动治理的技术性要求下，跨国公司等非国家行为体参与制定的技术标准、行为准则等行业规范往往具有事实上的效力以及功能上的优势，但这并不意味着它们可以位居中心，而只能被定位于跨国法律体系构造的外围。²⁴一方面，跨国公司参与制定的全球治理规则效力大多依赖于行业内的自愿服从。相较于以主权国家共同同意为基础的国际条约，非国家行为体主导下的行业规范属于自我规制性的规则，缺乏保障适用和执行的强制力，只是在众多利益相关方的参与下，对行业内部具有约束力。²⁵另一方面，跨国公司参与构建全球治理规则的目的不是为了约束国家行为体，而是为了影响这些国家的思维方式和行为模式。²⁶跨国公司的根本目的在于获取经济利益，其参与数据跨境流动治理往往不是出于对国际社会的责任感和事业心致力于解决全球问题，而是着眼于缓和商业逻辑与主权国家安全逻辑之间的矛盾，以获取国家便利化的数据自由流动监管态势。概言之，数据跨境流动全球治理体系的建构本质上仍须依托主权国家的强制性法律规制，跨国公司构建的技术标准和行业规范具备依赖性。不过，数字经济背景下各国希冀通过数字化转

22 参见余劲松.《跨国公司法律问题专论》[M].法律出版社:2008: 1.

23 参见余劲松.《跨国公司法律问题专论》[M].法律出版社:2008: 302.

24 参见徐崇利.全球治理与跨国法律体系:硬法与软法的“中心—外围”之构造[J].国外理论动态,2013,(08):19-28.

25 参见刘志云,谢春旭.跨国公司参与国际行业规范建设:动因、路径及其展望[J].南京大学学报(哲学·人文科学·社会科学),2023,60(03):49-56.

26 参见张明.软法视域下中国推进数字贸易全球治理的国际标准路径[J].上海对外经贸大学学报,2024,31(01):97-109.

型提高经济实力和综合国力，跨国公司也凭借其技术和资源形成显著优势，合作治理具备主观意愿和客观基础。因此，跨国公司正在通过行业内普遍接受的思维方式和行为模式潜移默化地影响国内立法、国际条约的制定，参与数据跨境流动全球治理。

（二）“补充”的方式：行业自律和参与立法

跨国公司发挥补充功能，通过行业自律和参与造法，在数据跨境流动领域与主权国家形成互动式治理范式。

1.通过行业自律实现数据跨境流动的行业内部规制

跨国公司行业自律的核心价值在于灵活性和先导性，通过与行业伙伴、上下游企业的合作对接实现行业内部规制，既能够填补国际硬法的缺失，又为企业规避合规风险提供缓冲。

跨国公司通过与同行业或同产业链公司的合作，实现业内与数据有关标准的统一。一方面，通过行业内合作制定跨司法辖区的行业标准。例如，为回应 GDPR 数据可移植性权利等要求，填补欧盟-美国隐私盾协议失效后的技术空白，由谷歌、苹果、Meta 等多家科技跨国公司联合发起《数据转移倡议》（简称 DTI），通过统一数据格式避免因格式差异阻碍数据传输等方式在行业内建立可互认、易操作的标准规范，推动数据在不同平台间安全、便捷传输。²⁷另一方面，通过产业链合作协调上下游产业的数据规则。例如，苹果公司在《苹果供应商行为准则》中要求供应商应遵循适用于苹果的所有隐私、数据保护和数据安全法律。²⁸又如，在欧盟数据战略框架下，德国主导的汽车行业推动形成汽车行业数据生态系统联盟，融合汽车数据链、产业链与供应链，形成欧盟标准的解决方案与行业用例。²⁹基于此，协调上下游企业数据工作标准，有助于减少数据跨境合规工作的重复投入，提高数据跨境流动的效率与价值。

2.间接参与数据跨境流动国家立法与国际条约制定

跨国公司间接参与立法的核心价值在于促进技术协作，降低合规成本。跨国公司在合法合规的前提下对外输出技术标准、管理理念，间接参与国家立法和国际条约的制定，既是全球化背景下企业规避风险的选择，也是数据跨境流动治理规则形成的重要推力。

第一，跨国公司可以通过参与论坛、发表倡议、接受咨询等方式，向国际社会提供数据跨境流动治理建议。例如，联合国在信息安全领域“双轨并行”磋商机制下的信息安全开放式工作组（简称 OEWG）采用多利益攸关方机制，允许联合

27 DATA TRANSFER INITIATIVE, <https://dtinit.org/>, 最后访问于 2025-03-25。

28 <https://www.apple.com.cn/supplier-responsibility/pdf/Apple-Supplier-Code-of-Conduct-and-Supplier-Responsibility-Standards.pdf>, 最后访问于 2025-03-25。

29 参见王瑞莹.欧盟组建全球汽车共享数据生态系统对我国影响及对策建议[J].新经济导刊,2023,(Z2):132-134.

国成员国、私营企业、技术社群等多方参与讨论。³⁰又如，联合国互联网论坛（简称 IGF）旨在促进互联网治理的多利益攸关方对话。其中，来自政府、私营部门、技术社群等领域的代表组成的多利益攸关方咨询委员会（简称 MAG）是 IGF 最核心的咨询机构之一，具有决策权且实际影响 IGF 的运行。³¹通过在这类场合发表意见，一方面，跨国公司可以基于技术能力为相关国际共识的达成提供专业性建议；另一方面，也可以基于对行业内发展趋势的判断表达诉求，在一定程度上影响其他利益攸关方的行为。

第二，跨国公司还可能通过游说监管当局，将自身数据跨境流动治理理念嵌入国内立法、国际条约之中。如前述，TRIPS 协议的形成历程表明跨国公司通过游说监管当局将企业利益转化为全球性规则的实践由来已久。在数据跨境流动治理领域，比较典型的是信息技术产业委员会（简称 ITI）在“Schrems II”案后的声明。³²ITI 是由亚马逊、佳能、联想、谷歌等企业组成的行业协会，在欧盟法院裁决使得隐私盾协议无效后，ITI 提出主张美国政令法规需满足欧盟监管要求，包括建议限制监控活动并为欧盟公民提供充分救济，³³以此敦促欧美新的数据传输协议谈判，实现数据的创新和自由流动。

三、风险：跨国公司参与数据跨境流动治理的多维挑战

当前，国际政治正形成多个中心，国家只是由竞争性治理机构组成的复杂体系的其中一个层次。³⁴在数据跨境流动治理中，充分发挥跨国公司的补充功能是历史必然。然而，合作治理过程中也应当警惕跨国公司冲击主权国家治理权力的潜在风险。具体而言，可以从利益目标、立法权威和监管效力三个方面把握。

（一）利益失衡：商业逻辑与安全逻辑的冲突

主权国家基于数据主权的制度诉求和跨国公司追求要素流通的商业逻辑，构成了数据跨境流动全球治理中多利益攸关方之间博弈的主要内容。如何在多元利益平衡的基础上，以国家安全、社会公共利益和个人隐私的有效维护为前提，创新治理方式以释放跨国公司的技术势能与治理效能，是跨国公司参与数据跨境流动合作治理带来的核心价值挑战。

基于发挥数据价值的自由流动和基于保障国家安全的严格规制是数据跨境

30 参见刘金河、杨乐.联合国框架下双轨制进程的博弈焦点与发展趋势[J].中国信息安全,2021,(09):81-84.

31 张晓.《联合国互联网治理论坛（IGF）：背景、生态与影响》载中国互联网协会 2024-09-11, <https://www.isc.org.cn/article/22092913616875520.html>,最后访问于 2025-03-25。

32 ITI Urges Transatlantic Policymakers to Begin Negotiations on New EU-U.S. Data Transfer Agreement Following Schrems II Decision, <https://www.itic.org/news-events/news-releases/iti-urges-transatlantic-policymakers-to-begin-negotiations-on-new-eu-u-s-data-transfer-agreement-following-schrems-ii-decision>,最后访问于 2025-04-02。

33 参见包柠榛.私人主体对数据跨境流动治理的补充功能：生成逻辑、参与路径与启示[J].国际经贸探索,2024,40(01):104-116.

34 （美国）D.赫尔德、（美国）J.罗西瑙、俞可平.《国将不国：西方著名学者论全球化与国家主权》[M].江西人民出版社.2004:208.

流动领域长期固有的摩擦。这一龃龉根植于国家主权、隐私保护与经济效率之间的平衡与取舍。于国家主权而言，数据主权项下一国享有“使用法定权力裁决某项行为是否构成对该国数据的侵犯”的域内最高管辖权³⁵和平等独立地参与数据领域国际交流合作事务而不受他国控制的对外事务最高权力。因此，国家对管辖范围内生成、存储的数据能否跨境流动享有根本控制权，这也催生了国家出于保护特定数据蕴含的国家安全和公共利益、以及个人信息的隐私意蕴，制定数据本地化存储、充分性保护认证等较为严格的数据出境条件。于数字经济而言，数据的价值不仅体现在其第五大生产要素的功能定性，也因其能够反映社会生产供需关系而成为关涉经营战略布局的重要指标。在经济利益、竞争压力的共同驱动下，海量、多源的数据跨境自由流动需求日渐增加。

跨国公司参与数据跨境流动治理面临商业逻辑与安全逻辑的冲突则更为显著。一方面，跨国公司的数据需求更复杂。如前述，数据是互联网科技型企业开发、运营的关键生产要素。即使对于非科技型企业，数据也能通过反映用户画像、产业链信息等帮助企业优化经营策略。在此基础上，跨国公司经营战略的全球性决定了其多元化、规模化的数据要素需求。因此，对数据跨境流动的自由性、便利性诉求也更为强烈。另一方面，跨国公司面临的数据跨境合规风险更高。跨国公司因经营活动的跨国性，难免触及不同东道国的数据跨境监管规则。然而，当前数据跨境流动尚无具备普遍约束力的全球性规则，其规制体系呈现碎片化、差异化特征。同时，伴随着大国博弈对数据治理话语权的关注，即使是同一国家，在不同时期或是同一时期对不同国家的监管策略也可能存在差异，进一步加大了不同行为逻辑制度性摩擦的可能。此外，在全球产业链优化整合的背景下，跨国公司通过跨国并购等方式实现资源要素优化配置愈发常态化，若此类经济活动涉及关键信息基础设施等与社会公共利益极为密切的行业领域，该商业逻辑与安全逻辑或许更难调和。

（二）立法失语：弱化国内立法国际条约权威

全球治理秩序观下多利益攸关方的协同性合作共治是顺应时代的必然要求。不过，尽管数据跨境流动治理关系的行为体越来越多元化，主权国家仍是其中最核心的角色。因此，如何既充分发挥跨国公司的补充功能，又从根本上保障主权国家对域内数据能否跨境流动的最高控制权，是跨国公司参与数据跨境流动治理的重要制度挑战。

在充分发挥跨国公司技术优势参与行业标准体系建设的同时，也应当警惕其依托数据霸权和技术垄断地位逐步消解国家立法主权和国际条约体系权威的潜在风险。首先，在技术层面，主权国家规制能力与跨国企业技术研发能力存在参

35 参见张晓君.数据主权规则建设的模式与借鉴——兼论中国数据主权的规则构建[J].现代法学,2020,42(06):136-149.

差,可能引致国家监管虚效。基于网络技术优势,跨国公司能够直接或间接地影响国内立法和国际条约的制定。然而,也正是因为技术能力的参差,加大了主权国家的监管难度。例如,若跨国公司以不透明的算法或流程处理数据,对于部分缺乏技术审查能力的国家,则难以切实保障该国数据法律法规对数据安全、国家利益的保护价值。其次,警惕个别国家借此实现数据领域的域外扩张,使数据跨境流动规则成为“西方主导”之工具。对此,尤为典型的是美国《澄清境外数据合法使用法案》(简称“云法案”)。云法案确立了“数据控制者”标准,即只要是美国通信服务提供商所拥有、监控或控制的数据,无论该数据存储在美国境内或境外,该企业均有义务向美国政府披露。在这一机制下,需要警惕美国与跨国公司联合构建跨境数据调取联盟,继而规避东道国对数据的管辖。再次,警惕“私法行为公法化”在数据跨境流动领域的重现,消解国家立法的独立自主性。“私法行为公法化”的实现往往不是国家积极主动地实施干预,而是为适应私人营利的需求而被动地采取特定行为。³⁶大量证据表明,不少国际贸易规则最早的起草者就是企业或企业资助或雇佣的代理人。³⁷因此,应当警惕跨国公司通过宣传技术标准、行业规范等实现规则嵌套,导致国家立法沦为“工具”。最后,部分跨国公司可能借助自身平台效应影响主权国家对域内数据的管辖能力。大型互联网科技平台能够通过平台服务条款、社区准则的设定,实现对平台内数据传输行为与内容的规制。这或将促使有关规定成为适用于平台“虚拟领土”的事实上的规范,并与在一国之内适用的法律相竞争。³⁸综上,发挥跨国公司参与构建的技术标准、行业规范等软法补充数据跨境流动硬法规制缺位的同时,必须警惕其弱化国内立法和国际条约权威的潜在风险。

(三) 问责真空: 监管跨国公司行为面临困境

国际社会对跨国公司的监管主要有两种途径:一是国家内部以立法方式调整由跨国公司产生的各种法律关系;二是国家之间通过国际条约或多边协定、国际组织借助行动准则等方式管制跨国公司行为。然而,上述传统监管模式适用于数据跨境流动治理体系却面临一定的问责真空。因此,兼顾跨国企业的补充作用与主权国家的最高管辖权,还面临对跨国公司的监管规制挑战。

总体而言,对跨国公司行为监管真空的原因在于:首先,国家立法的特性决定其难以应对新兴数据流动场景。不周延性与滞后性是法律固有的局限性。³⁹而数据技术发展的迅速更新迭代将放大了这一局限性,从而弱化其监管效能。其次,数据法域外适用扩张趋势加大了主权国家管辖难度。传统国家管辖原则以属地管

36 参见马忠法.论国际贸易领域中的“私法行为公法化”[J].政法论丛,2022,(01):42-52.

37 参见马忠法.私人部门在国际法规范形成和发展中的作用[J].厦门大学学报(哲学社会科学版),2016,(05):59-69.

38 参见罗旷怡,黄志雄.《国际法的数字挑战》白皮书[J].武大国际法评论,2023,7(03):1-26.

39 参见徐国栋.西方立法思想与立法史略(上)——以自由裁量与严格规则的消长为线索[J].比较法研究,1992,(01):1-39.

辖、属人管辖为基础，国家可以依其主权项下对内最高管辖权实现对跨国公司的有效监管。然而，近年来美国云法案的数据调取规则、欧盟 GDPR 的布鲁塞尔效应等，企图模糊数据领域属地界限，加大了主权国家以国内立法调整域内与数据有关的法律关系的难度。再次，现有国际规则规范跨国公司数据跨境流动行为的效力有缺。一方面，现有数据跨境流动国际协定的碎片化和区域化缺乏全球范围内统一施行之效力，而数据跨境流动往往涉及东道国、母国等不同国家规定，若所涉国家不同为协定缔约方，跨国公司或可借此规避协定规则限制；另一方面，当前国际社会规制数据跨境流动的行为指南、规范标准等软法性质显著，缺乏普遍约束力的特性致使其往往依靠有关主体的自愿履行，因此难以保障监管实效。

四、因应：跨国公司参与数据跨境流动治理的协同路径

主权国家与跨国公司相结合的合作治理范式是实现内部治理和外部威慑激励相容的积极探索，要比单纯立法规制或行业自律⁴⁰更有效率。妥善处理主权国家与跨国公司在数据跨境流动合作治理中的相互关系，可以适当借鉴网络空间“主权分层”理念，融合功能属性分层和安全关联度分层的双重标准，合理界定主权范畴与跨国公司权限。跨国公司必须尊重国家的数据主权，配合其工作与需要；主权国家也应乐于让跨国公司作为有益补充而存在，赋予其一定自由空间并为其发展提供必要条件。

（一）逻辑指引：借鉴主权分层模式

统筹商业逻辑与安全逻辑是因应跨国公司参与数据跨境流动合作治理风险挑战之关键。事实上，关于互联互通与主权安全之间的争议，在网络空间治理领域早已显现端倪。对此，国内学者依据“相互依赖的主权”，提出将网络空间界分为物理层、逻辑层和内容层，分别认定特定层级内网络主权的具体概念；⁴¹国外学者则提出可以类比适用海洋法中的关于主权权利的相关条款解决私人实体对网络活动管辖权的“垂直”主张，强调主权权利的滑动尺度。⁴²

在前者看来，各国间的网络治理问题不同于传统威斯特伐利亚主权，而是存在“相互依赖”的特性，只不过程度有别：其一，物理层的各类设备主要适用威斯特伐利亚主权，强调传统的国家法律管制。因为这类物品互联互通程度相对较低，既可以依据私法确认民事权利，也可以依照传统公法实现管辖。其二，逻辑层的各种传输协议和标准因其极强的互联互通特性，更适用“相互依赖的主权”，强调国家之间、国家与多利益攸关方之间的合作。其三，内容层的信息管理和数据流动规制则因互联网信息无界流通的逐步弱化，适用的主权概念处于传统威斯特伐

40 参见敖海静.数据保护的软法之道[J].法商研究,2022,39(02):158-172.

41 参见刘晗,叶开儒.网络主权的分层法律形态[J].华东政法大学学报,2020,23(04):67-82.

42 See Krishnamurthy V. Anchoring Digital Sovereignty[J]. Chi. J. Int'l L., 2024, 25: 417.

利亚主权和“相互依赖的主权”之间。⁴³

后者主张,综观国际海洋法,与国家对其陆地领土拥有近乎完全的权威不同,国家对其海岸线以外海域所拥有的主权权利是有限的,受制于其他国家和平利用的权利。并且随着向公海迈进,一国享有的主权权利逐渐减弱。⁴⁴类比到互联网领域,可以将其界分为“内部领域”和“外部领域”:当境外访问位于一国国境内的互联网终端设施,相当于“内部领域”范畴,其权利将受到严格限制;若该互联网基础设施连接位于境外,则可能受制于更为有限的一系列主权权利,二者的区别相当于一国在领海与毗连区、专属经济区内享有的不同权利内容,例如外国船只在专属经济区的航行权比领海要强得多。⁴⁵

由此可见,上述两种网络主权观虽然对于分层标准存在分歧:前者主张依照功能对象划分、后者侧重空间距离判断,但在主权分层认知上具有共识。值得注意的是,数据跨境流动本质上源于网络空间的互联互通,因此,对于平衡自由流动与数据主权的理论进路,网络主权分层的分析框架具有重要借鉴价值:将主权与领土“松绑”,转而关注较灵活的动态化治理模式。事实上,国际社会不乏数据跨境流动领域“分层”式实践:例如,中国以特定数据类型对国家安全、经济运行、社会稳定、公共健康和安全的影 响为标准,将数据划分为重要数据、核心数据和一般数据,实行分类分级管理。数据出境安全管理制度并非针对所有数据,而是只限于重要数据,一般数据在合法合规的前提下即可自由流动。欧盟则以地域为标准,实行“外紧内松”的差异化数据治理架构,数据可以在成员国内自由有序流动,若需要传输至欧盟之外国家,则应当满足充分性认证、提供适当保障措施、满足具有约束力的公司规则等更为严格的要求。

因此,统筹跨国公司参与数据跨境流动治理的商业逻辑和安全逻辑,可以在一定程度上对网络空间主权分层治理框架进行结构化移植。借助消解国家之间主权冲突的“放权”思路,缓和跨国公司参与数据跨境流动治理带来的逻辑对冲,赋予跨国公司参与数据跨境流动合作治理的合理空间,充分调动跨国公司的积极性以发挥其技术与资源优势,构建数据跨境流动协同共治机制。

(二) 构建主权分层模式下的数据跨境流动协同共治机制

网络主权分层治理围绕互联网应用场景下的国家主权冲突,在理论上重构了信息时代的主权认知框架:主权不仅在于特定空间内的绝对管辖,也应注重不同治理层级下精准控制能力和实际管理效能。在网络主权分层模式指导下,应当坚持以主权国家为主导,分层划定跨国公司的权责边界,并充分发挥跨国公司的补充功能,推动规则衔接与技术协同,构建数据跨境流动领域的协同共

43 参见刘晗,叶开儒.网络主权的分层法律形态[J].华东政法大学学报,2020,23(04):67-82.

44 See Krishnamurthy V. Anchoring Digital Sovereignty[J]. Chi. J. Int'l L., 2024, 25: 417.

45 See Krishnamurthy V. Anchoring Digital Sovereignty[J]. Chi. J. Int'l L., 2024, 25: 417.

治机制。

1.以主权国家为主导：分层划定跨国公司的权责边界

明确主权国家、跨国公司在数据跨境流动合作治理中的角色分工，是协同共治的前提。数据治理高度依赖技术与资源水平，在公私治理能力差距较大的背景下，更应当及时划定并规范跨国公司参与治理的权责边界。

坚持数据主权理念，分层划定跨国公司参与治理的具体界限。主权是一国固有的根本属性，意味着能够保持充分的意识、自决和控制。⁴⁶在跨国公司参与数据跨境流动合作治理的过程中，要防止商业逻辑冲击国家安全，主权国家必须坚持对本国数据的自决权、控制权，通过适当“放权”来界定跨国公司参与治理的权责边界。具体而言，可以借鉴前述网络空间“内容层”主权内涵，融合海洋法对不同海域内国家主权权利的划定标准予以界分：一方面，应当破解数据主权“全有或全无”的二元困境，正确认识网络空间互联互通下数据跨境流动治理的多利益攸关性，让渡跨国公司适当合理的治理空间；另一方面，汲取海洋法以陆地领土的情况为起点，根据特定区域与领海基线的距离确定不同层次的主权权利内容⁴⁷的做法，在“内容层”下再进行划分，将数据与国家安全、公共利益的密切程度作为界分跨国公司能否参与治理的直接标准。例如，可以进一步引入“安全数据层”与“商业数据层”，对可能影响国家安全、公共利益的数据类型划定主权国家治理红线，并在合作治理实践中验证具体数据的关联程度以进一步细化分层标准，构建主权干预强度的梯度模型。此外，还可以借助“物理层”数据存储设施的判断，认定跨国公司享有合作治理之权限。例如，我国《网络安全法》《数据安全法》就对关键信息基础设施运营者提出更高层次的运行安全义务要求。

在权力分层的基础上，对跨国公司设定多元价值治理义务，进一步规范其合作治理行为：一方面，不论参与哪一层级数据的跨境流动治理，跨国公司都应当确保其行为合法合规并履行社会责任。这不仅包括遵守东道国与母国的法律法规以及参与缔结国际条约规定的最低限度义务，也源于社会道德和商业伦理的要求。另一方面，在跨国公司创新性较强的合作治理行为中，譬如前述“商业数据层”，尤其应当附加地对跨国公司提出透明性、可追溯性要求。不论是跨国公司制定的行业标准可能消解国家立法权威，还是跨国公司参与治理面临问责真空，这些挑战很大程度上都源于传统治理主体的治理能力局限。因此，跨国公司应当确保技术实施流程的透明度、可追溯性，配合东道国和母国有关部门的监管，以合法合规的行为发挥技术资源的优势。此外，为确保跨国公司参与数据跨境流动合作治理的有序性，还可以引入内外部相结合的监督机制。作为商业性组织，跨国公司

46 See Belli L, Gaspar W B, Jaswant S S. Data sovereignty and data transfers as fundamental elements of digital transformation: Lessons from the BRICS countries[J]. Computer Law & Security Review, 2024, 54: 106017.

47 [英]马尔科姆·N.肖著，白桂梅、高健军、朱利江、李永胜、梁晓晖译.国际法（第六版）上.北京大学出版社：2011：436.

的核心利益来源于消费者，从而受到社会声誉约束。因此，应当完善社会监督机制，通过舆论引导、媒体监督、第三方认证等方式，进一步规范跨国公司行为。

与此同时，在全球资本自由流动的今天，只有采取制定双边、区域及多边规则等方式谋求跨国公司最低社会责任义务之共识，才能解决其负外部效应。⁴⁸因此，主权国家之间还应当加强规范跨国公司行为的国际合作，依托联合国、WTO等平台，推动形成全球性数据治理协议，划定跨国公司参与治理的主体责任并对东道国、母国设定协同监管要求；通过设立多国联合监管机构协调规则冲突和执法标准，进而形成具有约束力的全球性数字治理框架。

2.发挥跨国公司补充功能：推动规则衔接与技术协同

“治理”区别于“统治”，意味着变原来在国家层面上的单层次运作为现在国家和各种非国家行为体的多层次运作。⁴⁹在数据跨境流动合作治理中，应当推动跨国公司与主权国家之间的规则衔接，鼓励并支持其充分发挥技术、资源优势，充分实现跨国公司的补充功能。

充分依托跨国公司在“逻辑层”，即数据传输协议与技术标准创新方面的实践优势，通过构建“企业研发-国家认证”的协同机制，实现规则体系的技术性嵌套，从而释放跨国公司的技术势能。对此，我国天津自贸区提供了较好的范例。《中国（天津）自由贸易试验区企业数据分类分级标准规范》允许企业对照十三个大类四十个子类数据的分类分级标准，在企业内部开展分类分级工作，明确本企业重要数据并向自贸区相关主管部门报送重要数据目录。主管部门通过梳理汇总形成自贸区重要数据目录汇总表，报送天津市数据安全工作协调机制，最后按程序报送国家数据安全工作协调机制办公室。以企业内部的梳理整合为目录制定的逻辑起点，既充分发挥了企业的实践优势，又能通过层层报送确保数据分类分级工作的合法性、安全性。

在此基础上，主权国家还应当鼓励与支持跨国公司参与数据跨境流动合作治理：其一，优化数据跨境流动有关配套措施，为跨国公司参与合作治理提供必要条件。在软件层面，国家之间通过构建多双边协定谋求共识，减少数据跨境流动规制的不确定性，降低合规风险；在硬件层面，可以通过完善相关基础设施，提供必要补贴资助等方式加强物质支持。其二，及时化解跨国公司矛盾冲突，为合作治理营造和谐稳定环境。一方面，通过认证、互认等机制实现标准、规范的互联互通，减少跨国公司产生矛盾纠纷的可能；另一方面，通过提供技术取证支持、构建承认涉数据国际仲裁裁决机制等司法方式助力矛盾化解，及时定分止争。

此外，我国长期坚持共商共建共享理念，推动全球数据跨境流动合作，打造

48 参见刘志云.后危机时代国家间跨国公司社会责任制度体系的构建[J].政法论丛,2015,(03):120-128.

49 参见徐崇利.全球治理与跨国法律体系:硬法与软法的“中心—外围”之构造[J].国外理论动态,2013,(08):19-28.

共赢的数据领域国际合作格局。⁵⁰因此，对我国来说，还需警惕部分西方跨国公司基于其母国属性的潜在技术霸权风险，化被动为主动。例如，在助力本土跨国公司参与合作治理的过程中，应当鼓励相关企业争取更多话语权，支持其为数据跨境流动全球治理提供更多的中国经验和先进技术标准。在贡献全球数字治理公共产品的同时，系统地维护我国主权安全的底线要求与数字经济发展的战略空间。

结语

数据跨境流动是数字经济的重要驱动力。跨国公司参与数据跨境流动合作治理，既能够最大限度地发挥其技术与资源优势，也可以通过行业自律促使其规范自身行为。因此，伴随着跨国公司作用的日益彰显，可以借鉴网络主权分层理念，进一步平衡数据价值与数据安全，给予跨国公司适当治理权限；通过设定多元价值治理义务，健全跨国公司行为规范机制；同时提供必要的鼓励与支持，培育发挥合作治理功能，从而构建协同共治的数据跨境流动治理机制，在全球范围内实现数字经济发展与数据主权安全的合作共赢。

⁵⁰ 2024年11月我国发布《全球数据跨境流动合作倡议》，表示“我们期待政府、国际组织、企业、民间机构等各主体坚守共商共建共享理念，发挥各自作用，推动全球数据跨境流动合作，携手构建高效便利安全的数据跨境流动机制，打造共赢的数据领域国际合作格局，推动数字红利惠及各国人民。”

人工智能生成数据权利配置的法理逻辑与制度设计

宋云博 陆洋

(西南政法大学国际法学院/区域国别学院, 重庆 401120)

摘要：人工智能生成的海量数据，带来其数据权利配置的制度难题。全球主要国家数据权利配置制度的治理范式各有利弊，如英国法的知识产权保护范式、欧盟法的数据访问权保护范式、美国法的数据财产权保护范式等。为保障促进数字科技社会建构中真正实现“数字正义”“科技伦理”和“智能向善”，我国人工智能生成数据权利配置制度体系应遵循维护社会公平、数字正义和科技平权的基本逻辑；应在现行地方立法与司法实践中探索加强对人工智能生成数据进行知识产权保护与数据产权保护相结合的双轨制模式基础之上，不断健全完善相关法律制度；应建立知识产权保护、数据使用权赋权和数据产权登记公示制度等三轨并行的权利保护机制，并统筹兼顾社会公平与个体公平，依法依规探索根据合同意思自治原则进行具体数据权利的公平配置。

关键词：人工智能；数据权利；数据使用权；数据财产权；数字正义

一、问题的提出

数据已经成为与土地、劳动力、资本和技术相并列的第五大生产要素，在现代社会的各个领域中都发挥着越来越重要的作用。随着 Chat-GPT、Sora 以及 Deepseek 等生成式人工智能的迭代演进，人们的生活方式正在经历一场深刻变革，同时人工智能运行过程中生成的数据也迎来了数据权利的配置难题。习近平主席在 2023 年的中国国际服务贸易交易会全球服务贸易峰会上强调要“加快培育服务贸易数字化新动能，推动数据基础制度先行先试改革，促进数字贸易改革创新”。² 党的二十届三中全会《决定》也明确指出“建设和运营国家数据基础设施，促进数据共享”。³ 在国家政策层面上，2022 年中共中央、国务院发布

¹ **基金项目：**教育部人文社会科学规划基金项目“中国式现代化的涉外法治内涵体系研究”（23YJA820023）、重庆市社会科学规划重大项目“重庆打造内陆制度性开放综合枢纽研究”（2024ZXZD022）和重庆市哲学社会科学创新工程 2024 年第一批重点研究项目“数据安全重大风险评估机制及响应策略研究”（2024CXZD32）等成果。

作者简介：宋云博，西南政法大学国际法学院教授，中国-东盟法律研究中心（区域国别学院）研究员，研究方向为涉外法治、数字法治、“一带一路”区域国别法治理论与实践。陆洋，西南政法大学国际法学院硕士研究生，中国-东盟法律研究中心研究员助理。

² 《习近平向 2023 年中国国际服务贸易交易会全球服务贸易峰会发表视频致辞》，载求是网，http://www.qstheory.cn/yaowen/2023-09/02/c_1129841791.htm，2025 年 4 月 12 日最后登录。

³ 《中共中央关于进一步全面深化改革 推进中国式现代化的决定》，新华网，<http://www.news.cn/politics/20240721/cec09ea2bde840dfb99331c48ab5523a/c.html>，2025 年 4 月 12 日最后登录。

的《中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见》（以下简称“《数据二十条》”）提出“探索建立数据产权制度，推动数据产权结构性分置和有序流通”。数据权利配置是数据基础制度中最重要、最复杂也是争议最大的一项制度，数据来源复杂、应用场景丰富，数据在产生、流通及使用的过程中涉及到广泛且复杂的各类主体，不同的主体在不同的场景下有不同的利益诉求，故而在数据权利分配时应综合考量各方合理权益。⁴

传统场景下数据权利的配置已经十分复杂，而人工智能的加入使得数据权益配置中各方的关系更加盘根错节，难以厘清。一方面，在传统数据确权场景下，权利主体通常根据与数据的关系分为数据来源者、数据处理者以及第三方，而在人工智能生成数据的场景下，主体种类根据与人工智能的关系增加为人工智能使用者、开发者、所有者、投资者以及人工智能拟制主体，主体繁多，关系复杂，不同场景下各方主体对生成数据的贡献程度不同，很难通过宽泛的规定对其一概而论。另一方面，由于数据具有非竞争性与非排他性，数据可被多位主体可同时使用，⁵数据共享是数据市场的核心要义；同时数据又具有巨大的财产价值，排他性的确权是有效激发数据价值潜能的前提。数据共享与数据确权看似矛盾，但实则都是数据权益配置制度的价值目标，我国《数据二十条》中就强调要“坚持共享共用，释放价值红利”。对于涉及个人信息的数据，我国已有《个人信息保护法》对个人信息权益予以保护，该权益性质上属于人格权益。⁶然而，对于非个人数据所体现出来的财产权益却缺少有相关法律规制，故有必要对非个人数据的权利配置深入探讨。那么，如何在人工智能生成数据的场景下配置数据权利，维护好各方数据主体的数据权利，同时兼顾好数据共享与释放数据财产价值的制度目标？目前，世界各主要国家对于人工智能生成数据的权利配置在学说与实践层面存在不同范式，其间又有何优劣？接下来，对此进行系统梳理分析，尤为必要。

二、人工智能生成数据权利配置的治理范式辨识

人工智能生成数据权利配置是世界主要国家急需应对和解决的法律难题。目前，世界主要国家对人工智能生成数据的权利配置可归纳为四种范式，即：知识产权治理范式、数据访问权治理范式、数据财产权治理范式以及数据有限公开治理范式。通过对这四种主要治理范式进行分析比较，可为建构我国人工智能生成数据的权利配置模式提供重要参考。

（一）英国法的知识产权治理范式

人工智能在不同场景下生成的数据能按一定标准区分为多种不同的种类。其中，在创作场景下生成的数据若能符合著作权法的规定构成作品，则能直接适用

⁴ 参见程啸：《数据权益与数据交易》，中国人民大学出版社，第1页。

⁵ 参见丁晓东：《数据公平利用的法理反思与制度重构》，载《法学研究》2023年第2期，第28页。

⁶ 参见王利明：《数据何以确权》，载《法学研究》2023年第4期，第64页。

知识产权的规则予以保护。英国是最早在著作权立法中对人工智能生成的作品进行规定的国家。早在 1988 年英国的《版权、外观设计及专利法》中就规定对计算机生成的作品的著作权进行保护,规定其作者是对作品创作进行了必须安排的个人,该说法通常被解释为程序设计者。⁷虽然《版权、外观设计及专利法》中“计算机生成的作品”的表述与人工智能生成作品有所不同,但该法将其解释为“在没有人类作者参与的情况下,由计算机独立创作出来的作品”。⁸这和如今的生成式人工智能生成的作品含义有异曲同工之处,故对于《版权、外观设计及专利法》中的该规定可以看作为对人工智能生成作品的规定。此外,英国《版权、外观设计及专利法》中还规定了对数据库的版权保护,数据库指以有系统或有条理的方式排列的独立作品、数据或其他材料的集合,任何由数据库组成的文学作品只有构成作者本人的智力创作才属原创。⁹对于人工智能数据库中的作品,若人工智能使用者对作品做出了实质性贡献,而人工智能只是充当辅助创作的工具,则也符合上述条款的规定,能够受到版权保护。

英国直接赋予人工智能生成作品著作权的模式,虽然十分具有创新性与前瞻性,但是从保护人工智能生成数据权利的角度来看还是存在许多弊端:其一,知识产权理论范式只能保护部分数据,保护范围十分有限。¹⁰著作权法一般会对其保护的客体范围进行严格规定,如我国规定作品为文学、艺术和科学领域内具有独创性并能以一定形式表现的智力成果,英国的版权法规定受保护作品须为原创的且有固定形式,人工智能生成的海量数据中符合构成作品条件的数据只占小部分。对于人工智能生成的其余数据虽不构成作品但仍然具有十分重要的经济价值,这些数据也很有必要进行确权保护。其二,知识产权制度的理念与数据特性无法配适。知识产权制度具有排他性及垄断性,使用知识产权制度保护所有数据不利于数据的正常流通与共享。知识产权具有一定期限,超过期限后权利客体将会进入公共领域,然而数据作为一种资源,可以重复利用并且在不断的加工处理过程中产生新的价值,与有体物不同,数据不会在使用过程中减损其价值,反而会不断更新、积累,成为更大更有价值的数据集。故此,数据不宜被机械地设置一个保护期,也不宜直接进入社会公共领域。¹¹

（二）欧盟法的数据访问权治理范式

欧盟在立法与司法实践中都不承认人工智能生成作品的著作权保护。¹²但

⁷ 参见宋云博、陆洋:《生成式人工智能作品著作权认定的困境、模式与立法完善》,载《数字经济与法治》2024 年第 2 辑,第 55 页。

⁸ 参见英国《版权、设计和专利法》,The Copyright, Designs and Patents Act 1988, Section 178, at <https://www.legislation.gov.uk/ukpga/1988/48/section/178>, 2025 年 4 月 11 日最后登录。

⁹ 参见英国《版权、设计和专利法》,The Copyright, Designs and Patents Act 1988, Section 178, at <https://www.legislation.gov.uk/ukpga/1988/48/section/178>, 2025 年 4 月 11 日最后登录。

¹⁰ 参见李晓宇:《智能数字化下机器生成数据权益的法律属性》,载《北方法学》2021 年第 2 期,第 47 页。

¹¹ 参见王利明:《数据何以确权》,载《法学研究》2023 年第 4 期,第 63 页。

¹² 参见张惠彬、李俊豪:《智能创作时代 AIGC 版权保护的比较及中国实践》,载《国际经济法学刊》

是，欧盟在立法上对于数据权利配置进行了长时间的探索：从 1996 年《数据库指令》中的数据库特殊权利到如今《数据法》中的数据访问权，欧盟对数据的治理模式也完成了由排他性赋权到促进数据共享开放的转变。欧盟《数据法》中数据访问权的权利客体是用户在使用电子通信服务、物理连接或设备访问检索的联网产品或接受相关服务过程中产生的产品数据及相关服务数据。¹³ 虽然整部法律条文中都未提及人工智能生成的数据，但是将人工智能放进数据访问权的场景下来看，如在用户使用人工智能产品进行检索或接受人工智能检索服务过程中人工智能所生成的数据，符合《数据法》所规定的数据库数据外延，故人工智能生成的数据也在《数据法》的调整范围之内。《数据法》赋予用户的数据访问权以数据在数据持有者掌控中为前提，通过向持有者设置的一系列义务以保障用户得以顺利、安全地访问目标数据。

可见，与英国的版权保护模式及欧盟此前的数据库特殊权利保护模式都不相同，《数据法》没有向任何一方数据主体赋予排他性的财产权利，而是创设了数据访问权，打破了数据主体间的屏障提供了数据共享的桥梁。在颁布《数据法》之前，欧盟学界曾就是否需要创建排他性数据权利以及应当引入何种数据权利展开了激烈的讨论。¹⁴ 欧盟于 2017 年发布文件意图创设数据生产者权，遭到了以德国马普所为代表的学界的强烈反对。德国马普所认为，解决方案不是建立一个新的财产权制度，而是创设一种针对性的、不可放弃的数据访问权。¹⁵ 在此背景下，欧盟《数据法》应运而生，数据访问权是促进数据共享推动数据流动与开放的极佳范式，并且《数据法》针对的商业数据客体已经包含了用户在使用人工智能过程中所生成的绝大部分非个人数据，可以对人工智能生成数据进行最大限度的调整。但是，《数据法》虽重视保护和促进数据共享，却也忽略了数据财产价值的制度性激发，数据财产权利与数据共享看似矛盾，实则关系密切、不可分离。数据财产性价值的生成离不开数据共享环节，企业对数据进行加工处理使其产生经济价值的前提是用户将其产生的原始数据共享给企业。同时，只有规定数据财产权利，明确数据在流通使用过程中各个主体享有的权利义务才能缓解个人与企业间数据共享时的权责争议。¹⁶ 可见，数据财产确权与数据共享作为数据权利配置的重要内容要相辅相成、齐头并重，不可偏废。欧盟经过一系列的探索与尝试，

2024 年第 3 期，第 71 页。

¹³ See Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonized rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023R2854&qid=174437529237>, [2025.04.11].

¹⁴ 参见孔德明：《数据财产权到访问权：欧盟数据设权立法转型解析》，载《比较法研究》2023 年第 6 期，第 38 页。

¹⁵ See Josef Drexler et al., Position Statement of the Max Planck Institute for Innovation and Competition of 26 April 2017 on the European Commission's Public consultation on Building the European Data Economy'. Max Planck Institute for Innovation & Competition Research Paper No. 17-08, 5 (2017).

¹⁶ 参见梅傲、柯晨亮：《数据共享与数据财产化》，载《四川师范大学学报（社会科学版）》2024 年第 2 期，第 64 页。

先是放弃了数据排他性确权制度，转而通过数据访问权建立流通共享的数据市场。这前后两种方案都无法平衡与兼顾数据共享与数据财产价值的利用，并不可取。

（三）美国法的数据财产权治理范式

美国与欧盟在人工智能生成作品的著作权的认定问题上持有相同态度，都不承认其著作权的保护。但是，在数据权利保护模式上，与欧盟的数据访问权方式不同，美国侧重于在实践与立法中保护数据的财产性权利。美国最早在十九世纪的若干新闻信息盗用案件裁判过程中形成了信息盗用制度，用来保护新闻信息的经济价值，打击新闻领域的不正当竞争行为。¹⁷ 后来随着技术的发展，信息盗用制度被法院拓展运用到网络数据的不正当竞争案件中，对他人不正当抓取网络平台数据的行为进行认定与规制，通过反不正当竞争的方式保护企业的数据财产权益。虽然信息盗用制度的本质是为反不正当竞争，但其根本目的是为了保护数据信息的财产权益，将数据视为个人或企业的财产，并且同样适用于人工智能生成的数据。这样的理念也体现在美国对于个人信息保护的立法中，美国个别州已经在立法上明确承认了个人对其基因信息、医疗信息及测谎记录享有的财产权利，联邦立法中也明确赋予个人对其名字、形象等个人信息享有的财产权。¹⁸ 能够保护人工智能生成数据财产权益的法律，还有美国于 2016 年颁布的《商业秘密保护法》。该法规定：企业可以主张人工智能生成的数据为商业秘密，从而获得专有的财产性权利保护。

目前，美国对于非个人数据的财产权只能进行事后的防御型保护，并没有直接对非个人数据进行财产权确权。这是由于在欧盟出台《数据库指令》制定数据专有权保护数据的时期，美国也试图出台类似法律，但因此类数据排他性权利带来的争议过大而放弃。数据财产权利在立法上的失败并没有转变美国侧重保护非个人数据财产性权益的立场：由于美国具有判例法的法律渊源，其采取在实践中运用判例法以及解释法律条款的方式对数据财产权益进行保护。虽然目前美国并未出现人工智能生成数据相关的财产权益纠纷案件，但美国对于数据财产权益的重视程度足以让我们推知其对人工智能生成的非个人数据会采用相同的方式进行保护。

（四）欧盟法与美国法的有限公开治理范式

除了前述各自较为典型的治理范式之外，欧盟法和美国法对数据治理还呈现出另一种共通的探索发展方向——有限公开治理范式，有必要在此专门探讨分析。之所周知，公共数据开放已成为全球数据治理的趋势，欧盟近年陆续推出了多部规定公共数据开放的法律法规，对公共数据开放规则进行了长达十余年的立法探

¹⁷ 参见杨翱宇：《美国法信息盗用制度的演进及其对我国数据财产权益保护的启示》，载《政治与法律》2019 年第 11 期，第 146-150 页

¹⁸ 参见刘德良：《个人信息的财产权保护》，载《法学研究》2007 年第 3 期，第 80-81 页。

索。从 2019 年的《开放数据指令》到 2022 年的《数据治理法案》，欧盟逐步扩大了公共数据开放的范围，以此促进开放数据的再利用。¹⁹公共数据作为人工智能训练数据来源的一部分，公共数据的开放可以为人工智能的学习与决策提供更加合理、可靠的数据，从而提升人工智能生成数据的准确性。相较于人工智能来源数据公开规则的成熟，欧盟对于人工智能生成数据的公开还在初步探索的状态。2024 年欧盟颁布的《人工智能法案》按照风险级别对人工智能系统进行划分监管，要求对人工智能生成数据的性质和数量进行评估，通过评估的人工智能系统在被加盖“CE 标识”后其生成的数据即可流入公共领域，以此保障人工智能数据在公共领域的安全使用。²⁰

美国与欧盟在人工智能生成数据开放上具有高度相似的理念和做法。2019 年美国食品药品监督管理局（FDA）公布了三十余种加盖官方标识的人工智能生成的包括血糖浓度变化数据等的医学数据产品，将其共享至公共开放领域。²¹相较于欧盟直接在人工智能相关法案中对数据公开进行规定，美国对人工智能生成数据的公开规定分散于各个领域的法律文件中。2018 年美国颁布了《开放政府数据法》优化公开数据范围与质量，规定公开联邦政府部门的数据。若政府的数据是由人工智能系统输出，如美国国家航空航天局（NASA）利用人工智能分析得出的卫星图像，则该人工智能生成数据也需要公开。美国国立卫生研究院（NIH）于 2022 年要求其资助的研究人员与研究机构公开他们的医疗研究数据，其中就包括利用人工智能生成的数据。

鉴于目前人工智能生成数据的准确性与安全性无法被保证，若公开所有数据将会导致数据真实性的混乱，欧盟与美国通过对人工智能系统及其生成的数据进行审查的方式，对达到公开标准的数据进行官方背书后公开，既提升了数据的公信力又能保障公众对公开数据的安全使用。除此之外，还要明确个人数据与公共领域的界线，人工智能生成数据的权属不清会导致公共领域共享标准模糊。²²部分领域如城市应急、教育、医疗健康等相关的数据具有公开与共享的需求，故而人工智能生成的数据不仅需要审查危害级别，还要确定所属领域，判断流入公共领域的安全性及必要性。可见，欧盟、美国及英国的人工智能数据权利配置范式都存在一定的弊端，都无法完全保护人工智能生成数据的权利，并且无法应对科技时代新的社会问题。那么，人工智能生成数据权利配置究应如何科学合理实现，又应遵循着哪些法理逻辑？

¹⁹ 参见张华韬：《公共数据开放规制的原理及路径》，载《社会科学研究》2025 年第 2 期，第 129-130 页。

²⁰ See EU Artificial Intelligence Act, <https://artificialintelligenceact.eu/the-act/>, [2025.04.11].

²¹ See FDA approvals for smart algorithms in medicine in one giant infographic [EB/OL], <https://medicalfuturist.com/fda-approvals-for-algorithms-in-medicine/>, [2025.04.11].

²² 参见何炼红、朱曦青：《论人工智能数据公共领域深度共享机制的构建》，载《中南大学学报（社会科学版）》2024 年第 6 期，第 35 页。

三、人工智能生成数据权利配置的法理逻辑考察

在人工智能时代，科技发展赋能为人类社会带来了巨大便利和红利，但同时也为人类社会治理带来了系列新问题。其中，最重要的难题就是数字鸿沟的加剧、数字社会的不平等以及数字阶级分化的产生。所以，科技时代的立法在解决科技法律问题本身的同时，还应当统筹配套解决科技带来的社会问题，在建设“数字社会”的奠基开始，就应注重系统性考量所有人公平分享数字经济红利，确保引导科技向善治理发展。故而为了解决数字社会的新问题，促进社会公众平等地享受人工智能带来的便利，在人工智能生成数据权利配置的制度设计过程中应当遵循系列基本原则，以社会公平为目标、以数字正义及科技平权为基准，符合构建人类命运共同体和创造人类文明新形态的基础逻辑。

（一）遵循促进“社会公平”的目标逻辑

人工智能的迅猛发展推动了数据信息在数量上的大爆发，从而加剧了社会上各主体因对数据掌控力存在的差异而带来的不公平现象。在数据市场中具体表现为人工智能大型企业对数据的垄断以及人工智能生成数据过程中涉及主体之间的权利分配不均衡现象。人工智能数据权利配置中的社会公平意为维护各数据主体的地位平等与共同发展，对数据市场中的弱势群体进行适配性的合理保护。其一，维护市场主体发展与创新的公平。数据对于人工智能企业而言是一笔财富。企业能够根据数据的收集、应用等，不断完善改进自身的人工智能科学技术，以满足消费者对其的市场需求，从而获得相应回报。一个充满活力的市场，必然是各个企业的百花齐放、自由公平竞争生态。若大型企业凭借对数据的掌控力而独占数据市场、数据无法公平地自由流通，将不利于小微企业的创新发展以及市场的良性竞争发展。同时，在数据权利相关合同订立的场景中，相关条款大多由具有信息优势和技术优势的大型企业单独拟定，消费者或新兴小微企业在其面前处于弱势谈判地位，易出现大型企业在合同签订过程中利用其优势地位侵害消费者及小微企业的数据访问与使用权利的现象。所以，为了便利小微企业参加数据竞争，促进市场的公平竞争，故在人工智能生成数据的权利配置中应当对小微企业设置倾斜保护政策，鼓励人工智能行业的创新创业。例如，欧盟《数据法》中就制定了合同无效的专门条款对消费者及小微企业的数据权利维护提供支持，并且还将制定示范性数据共享合同条款，以帮助小微企业更好的发展。

其二，平等保护数据主体的权益。人工智能数据生成过程中涉及的主体繁多，市场地位、利益需求又各不相同，不同场景下各数据主体的地位也不尽相同，故为了公平保障各方利益，应当制定具体场景下各方主体的权利与义务，侧重保护弱势方的数据利益。例如，欧盟《数据法》就根据不同场景下不同主体的特点与利益需求分别规定了一系列权利与义务，以此划清权利边界，平衡数据来源者、数据持有者以及第三方的合法权益。在数据持有者向数据来源者提供数据的场景

下，现实中数据来源者通常为消费者，而数据持有者一般具有企业身份，消费者在实际掌控着数据的企业面前具有弱势地位。故为了维护数据来源者的合法数据权益，《数据法》为数据持有者规定了一系列义务，如只能在合同的基础上使用数据，且应当以方便、安全及免费的形式提供数据，²³赋予了数据来源者排他性的数据访问权，对来源者的权利进行了倾斜保护。²⁴而在数据持有者按照数据来源者的要求向第三方提供数据对场景下，第三方往往是市场中的企业或个体有偿受数据来源者委托而访问数据，可能与数据持有者存在竞争关系，²⁵为了维护双方的商业地位及市场的竞争公平秩序，《数据法》禁止第三方获得数据持有者的商业秘密，同时也禁止数据持有者利用数据深入了解第三方的商业经济情况。²⁶欧盟《数据法》为我们提供了公平维护各方主体权益的良好范式，既要弱势群体进行倾斜保护，消除数据时代带来的不平等现象，也要对数据优势方的合法利益进行保护以维护市场的效率及经济利益。

（二）坚守维护“数字正义”的底层逻辑

正义，是法律的内在根据和永恒价值追求。人工智能革命在重构我们的思维方式及生活方式的同时，也引起了正义观念的深刻变革，数字正义正在成为现代正义价值体系中的新形态。数字正义，是指导和约束人类开发运用数字技术的价值准则，其内容具体包括数据资源的合理分配以及数字权利的公平配置等。数字化转型背景下法治建构的目标之一就是实现数字正义，以数字正义推动更高水平的社会主义的实现。²⁷我国始终坚守维护数字正义的正确思想理念。2023年中国外交部发布《全球人工智能治理倡议》，倡导坚持“智能向善”并坚持公平性与非歧视性原则。²⁸作为数字正义的重要内容，人工智能生成数据权利的配置也应当遵循数字正义的价值目标，将数字正义逻辑贯彻于相应立法之中。其一，数字社会同样要实践数字正义。在数字社会中，资源的分配并不是在现实中通过人为手段操控实物流转，而是通过数据信息处理及算法控制来隐性匹配资源及进行调动，而由于算法黑箱的存在，对外界而言数据输入与数据处理过程不透明、不可视，导致数据处理者与数据接收者之间的信息不对称，从而造成资源配置结果缺少公

²³ See Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonized rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023R2854&qid=174437529237>, [2025,04,11].

²⁴ 参见申卫星：“论数据来源者权”，《比较法研究》2024年第4期，第116页。

²⁵ 参见丁凤玲、彭建：“数尽其用：数据访问权的法理反思与功能定位”，《华南理工大学学报（社会科学版）》2024年第3期，第96页。

²⁶ See Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonized rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023R2854&qid=174437529237>, [2025,04,11].

²⁷ 参见周尚君、罗有成：《数字正义论：理论内涵与实践机制》，载《社会科学》2022年第6期，第167页。

²⁸ 参见《全球人工智能治理倡议》，https://www.fmprc.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml, 2025年4月11日最后登录。

平性及信服力。数据资源的拥有程度直接关系到社会地位以及权力、财富、机会等的拥有，数据资源分配的不公平则会导致更多社会不公平现象的产生以及严重的阶级固化。此时通过合理赋予人工智能各方主体数据权利，可以保护数据主体的正当数据权益，防止数据掌控者对数据主体合法权益的剥夺，弥补信息不对称带来的地位失衡。

其二，数字正义涵盖结果正义和程序正义。²⁹其中结果正义在人工智能生成数据权利配置环节中表现为数字技术环节中的计算正义以及权利义务的分配正义，其中计算正义是技术保障下对数据资源的合理分配，分配正义是通过赋予权利义务对数据资源分配不公的纠正及兜底。计算正义是数字时代下数字正义的基本属性，核心是计算程序的正当性以及计算结果的公平性。³⁰人工智能生成的数据是按照人类的指令，在对海量训练数据进行学习、整理的基础上再加工而生成的。故对人工智能来说，计算正义不仅要求数字算法技术的先进与完善，还依赖于人工智能训练数据的准确性与安全性。美国于 2019 年签署《美国人工智能倡议》，提出相关机构将联邦数据、模型向美国人工智能研发领域开放，提高人工智能决策的准确性，增强公众对人工智能技术的信任。但目前由于人工智能产生的算法黑箱、数字鸿沟等问题的存在，需要通过制定法律法规对数据主体重新分配权利，以规制算法程序对数据主体造成的数据资源分配不公现象，通过权利分配的正义来纠正计算正义中的纰漏。

数字正义中的程序正义在人工智能数据权利配置中体现为数据处理、分析过程的透明性以及数据权利的权利知情、救济及事后问责制度。³¹数字正义是基于认知计算的可视正义，是在汇集海量数据基础上从经验研判到全景透视的重大转型，但是由于互联网信息茧房及算法黑箱的存在，形成了数据技术优势者对公众的单项透视，使得公众陷入了不知情、不平等的社会境地。³²

（三）秉持促进“科技平权”的文明逻辑

随着人工智能技术的普及，越来越多的人都能平等地使用人工智能技术，享受科技革命带来的便利。科技平权正是指让所有人平等地使用科技产品、享受科技便利，使科技发展的成果具有普惠性，实现数字时代的共同富裕。习近平总书记指出：“网信事业要发展，必须贯彻以人民为中心的发展思想。要适应人民期待和需求，加快信息化服务普及，降低应用成本，为老百姓提供用得上、用得起、用得好的信息服务，让亿万人民在共享互联网发展成果上有更多获得感。”³³ 传统科技霸权往往建立在高成本研发成本之上，而我国科技企业秉持“科技向

²⁹ 参见张凌寒：《数字正义的时代挑战与司法保障》，载《湖北大学学报（哲学社会科学版）》2023 年第 3 期，第 50 页。

³⁰ 参见周尚君：《数字法理学的基本范畴》，载《光明日报》2025 年 2 月 21 日，第 11 版。

³¹ 参见周尚君、罗有成：《数字正义论：理论内涵与实践机制》，载《社会科学》2022 年第 6 期，第 169 页。

³² 参见马长山：《面向“三维世界”的数字法学》，载《中国社会科学》2024 年第 11 期，第 159 页。

³³ 习近平：《在网络安全和信息化工作座谈会上的讲话》，载《人民日报》2016 年 4 月 26 日，第 2 版。

善”“智能向善”理念和原则，引导科技向善治理，分享数字经济红利，通过开源策略与工程效率优化研发出的 Deepseek 等新兴人工智能产品使技术门槛及使用门槛大幅降低，让更多人能享受人工智能发展成果。然而，目前随着人工智能技术的迅速发展及迭代更新，人们在对人工智能产品使用程度以及适应程度上存在越来越大的差距，数字鸿沟日益扩大，公众在数字领域也产生了社会分化问题。³⁴ 数字鸿沟包括不同主体在运用数字技术能力上的鸿沟如老年数字鸿沟，也包括不同地区在数字资源上的鸿沟如城乡数字鸿沟及中西部数字鸿沟，其本质为经济发展不平衡不充分而造成的教育、基础设施以及公民能力素养等方面的差异。

习近平总书记指出“要发展信息网络技术，消除不同收入人群、不同地区间的数字鸿沟，努力实现优质文化教育资源均等化”。³⁵ 弥合数字鸿沟实现科技平权不仅要在发展上下功夫，加强公共科技供给，减小城乡差距与代际差距，更需要通过立法保护每位公民的数字权利。其一，在法治层面，要重视保护数字弱势群体的正当权利。我国社会主义法治是人民的法治，在人工智能数据权利配置立法中贯彻科技平权的逻辑，是在科技新时代坚持以人民为中心，坚持法治建设为了人民的体现。数字发展的不平等要求我们在进行人工智能生成数据权利配置时侧重保护数字弱势群体的数字权利，并规定完善配套的帮扶老年群体使用人工智能产品的政策，防止老年群体的生存权与发展权受到侵蚀。同时，数字鸿沟的形成也与数字资源、数据资源的垄断、不流通有关，人工智能权利配置要围绕推动数字资源与数据资源开放共享的理念进行展开，使老年群体及不发达地区共享资源，实现数字资源与数据资源普惠各类人群。其二，在数字技术层面，要降低数字科技的使用门槛。要降低要深化数字应用适老化改造，继续通过提升技术来降低科技的使用门槛，降低数字弱势群体的使用难度。我国在人工智能生成数据权利配置的具体制度设计上要遵循这三种理念，在完成权利配置的同时缩小数字鸿沟，促进数字社会的平等。

四、人工智能生成数据权利配置的制度设计思考

与其他国家不同，我国目前在司法实践中承认对人工智能生成作品的著作权保护，同时也制定了地方层面的数据知识产权登记制度及数据产权登记制度，对数据进行确权登记。我国目前的制度有何优劣？该如何改进？在制定具体的人工智能生成数据权利配置制度时，要在我国现有的相关法律条款与政策的基础之上，结合域外经验，遵循社会公平、数字正义以及科技平权原则，来探讨改进的方向及路径。

（一）当前我国人工智能生成数据权利配置的法治实践

结合目前我国对数据的一系列政策及立法来看，我国在地方立法与司法实践

³⁴ 参见张吉豫：《数字法理的基础概念与命题》，载《法制与社会发展（双月刊）》2022年第5期，第50页。

³⁵ 习近平：《为建设世界科技强国而奋斗》，载《人民日报》2016年6月1号，第2版。

中对人工智能生成数据采用的是知识产权保护与数据产权保护相结合的双轨制模式。

其一，在对于是否加强知识产权保护议题上，一方面，学术界就人工智能生成作品著作权认定问题上出现了较大争议，与学术理论界的有关争论相对应，我国司法实务中对此也存在不同的裁判观点。例如，在“菲林诉百度案”³⁶“腾讯诉盈讯案”³⁷“AI 文生图著作权案”³⁸等案例中就体现出不同的司法主张。虽然这三个案例都是人工智能自动生成作品著作权侵权案件，但其审理的焦点都是围绕人工智能作品著作权认定问题展开。最早的“菲林诉百度案”的裁判观点较为保守，否定了人工智能作品的著作权，但也承认了使用者的部分权益。在“腾讯诉盈讯案”和“AI 文生图著作权案”中，法院则都认为人工智能作品属于具有独创性的智力成果，肯定了人工智能作品的著作权保护，并认为著作权应归属于软件使用创作者。其中，“菲林诉百度案”与“AI 文生图著作权案”的判决都由北京互联网法院作出，但该法院在“AI 文生图著作权案”中转变了观念，没有继续坚持“只有自然人或法人组织的创作物才能构成作品”的观念，而是创新性的对涉案图片著作权作出了认定。这体现了我国法院在人工智能作品著作权认定的司法实践中，采取了与时俱进的司法探索实践以及对社会公众使用人工智能创作的司法鼓励和正向引导。故我国目前采用的是在司法实践中通过对《著作权法》的法律条文及其精神的解释，来保护人工智能生成作品著作权，弥补立法缺失的不足。另一方面，许多地方都推行了数据知识产权登记制度，例如，《陕西省数据知识产权登记管理办法（试行）》、《湖北省数据知识产权登记管理办法（试行）》及《浙江省数据知识产权登记办法（试行）》，都规定数据知识产权的登记申请人应当为数据处理者，登记对象为“依法合规获取的、经一定算法或规则处理后形成的、具有实用价值及智力成果属性的处于未公开状态的数据集合”。根据这些规则，数据处理者可以对其使用人工智能生成的数据作品申请数据知识产权登记，从而获得权利保护。但是这里的数据知识产权与我国《著作权法》中规定的知识产权并不完全相同，其本质是为了数据确权而制定的数据权利登记制度，在登记时对条件只作形式审查，并且存在数据知识产权的条件与界线不明确的问题。

其二，在对于是否加强数据产权保护议题上，同样存在国家层面立法缺失的问题。我国也同样是通过司法实践和地方立法等两个层面来进行数据产权保护。一方面，在司法实践层面，已经出现部分涉及数据财产权利归属的案例，典型案例包括 2016 年的新浪诉脉脉案³⁹、大众点评诉百度案⁴⁰、2017 年的淘宝诉美景案

³⁶ 参见“菲林诉百度著作权侵权案”，北京互联网法院（2018）京 0491 民初 239 号判决书。

³⁷ 参见“腾讯诉盈讯著作权侵权案”，广东省深圳市南山区人民法院（2019）粤 0305 民初 14010 号判决书。

³⁸ 参见“李昀锴诉刘某某著作权侵权案”，北京互联网法院（2023）京 0491 民初 11279 号判决书。

³⁹ 参见“新浪诉脉脉不正当竞争案”，北京知识产权法院（2016）京 73 民终 588 号民事判决书。

⁴⁰ 参见“大众点评诉百度不正当竞争案”，上海知识产权法院（2016）沪 73 民终 242 号。

⁴¹以及 2020 年的腾讯诉搜道案⁴²。这些案例都因被告非法使用用户在使用原告产品过程中所产生的数据用于商业用途而起,虽然法院都以不正当竞争纠纷作为案由来裁判,但是这些案例的争议焦点都在于原告对数据的数据权属问题。由于我国在数据权属方面的法律缺失,法院巧妙的将争议焦点转化为了数据企业之间的不正当竞争问题,绕开了数据权属争议,并且运用《反不正当竞争法》第 2 条的诚信原则来进行解释,承认了被告企业使用原告数据对原告造成的损害,构成不正当竞争,但无法证明其构成侵权。并且,近年来已出现判决对数据权益的明文承认。例如,在腾讯诉搜道案中,一审法院认为网络平台方对于数据资源整体与单一原始数据个体享有不同的数据权益,被告行为对腾讯公司基于数据资源整体获得的竞争权益构成了实质性损害,这在数据产权司法领域无疑是重大的突破。另一方面,在地方立法层面,深圳市发布《深圳市数据产权登记管理暂行办法》,规定数据产权登记的对象为数据资源及数据产品。根据这些规则,数据处理者可以对其使用人工智能生成的数据进行知识产权登记,从而获得作品的产权保护。

由此可见,作为建构数字社会和发展数字经济的基础制度之一,我国人工智能生成数据权利配置的制度体系整体还比较薄弱,亟待优化完善。

（二）人工智能生成数据权利配置的基础要素制度完善

第一,完善人工智能生成作品的著作权保护制度。目前世界主要国家中只有英国与印度明确对人工智能生成作品进行著作权保护,世界各国对人工智能生成数据是否进行著作权认定来自于“作者中心主义”及“读者中心主义”的思想根源。⁴³“作者中心主义”主张作者具有极高地位,重视作品与人类作者之间的联系,将作品视为作者人格权的延伸,强调作品的独创性只能来自于人类作者的精神贡献,认为著作权只能保护人类创作的作品。“作者中心主义”的哲学基础在于保护作者的人格利益,维护作者在作品中体现的独特个性及创作意图。⁴⁴而“读者中心主义”则恰恰相反,认为作者与作品应当采用主客体分离的客观评价标准,强调作品的灵魂应是内在的结构与语言体系,而不是作者本身,正如一名学者所言“读者的诞生应以作者的死亡为代价来换取”。⁴⁵“读者中心主义”的哲学基础在于保护作品的经济价值本身,因而不考虑作者进行创作时体现的独特个性、创作意图及贡献付出等,提倡以普通读者的视角对作品独创性进行客观评判。大陆法系国家普遍对独创性的要求较高,如德国、西班牙、意大利等国家都要求作品必须是作者独特个性、精神及艺术观念的反映,不承认人工智能生成数据的著作权保护。而英

⁴¹ 参见“淘宝诉美景不正当竞争案”,浙江省杭州市中级人民法院 2018 年浙 01 民终 7312 号民事判决书。

⁴² 参见“腾讯诉搜道不正当竞争案”,杭州铁路运输法院(2019)浙 8601 民初 1987 号判决。

⁴³ 参见宋云博、陆洋:《生成式人工智能作品著作权认定的困境、模式与立法完善》,载《数字经济与法治》2024 年第 2 辑,第 51 页。

⁴⁴ 参见卢炳宏:《论人工智能创作物独创性判断标准之选择》,载《内蒙古社会科学》2020 年第 4 期,第 103 页。

⁴⁵ [法]罗兰·巴尔特:《作者之死》,怀宇译,百花文艺出版社 2005 年版,第 307 页。

美法系国家从实用主义哲学角度出发，对作品的独创性要求较低，⁴⁶尽可能的保护更多作品的经济价值，但美国在人工智能生成数据方面却严格拒绝对其进行著作权保护，认为只有人类创作完成的作品才能受到著作权保护。这是由于美国政策与法律上倾向将人工智能视为创作工具，对于人类运用人工智能工具辅助生成的作品可以予以著作权保护，但是对于完全由人工智能自身生成的作品，美国在实践与立法中都不承认其著作权保护。

我国《著作权法》由于制定时间较晚，《著作权法》的制定深受发达国家以及国际条约的影响，汇集了世界各国著作权立法以及国际知识产权制度的精华，使其一开始就较为先进。⁴⁷我国在司法实践中对人工智能生成作品著作权问题的探索经历了从严格“作者中心主义”向较为宽松的“作者中心主义”的转向，认为人工智能生成作品中包含人类创作者个性化的安排、指令，故并没有完全采用“读者中心主义”模式。司法实践中对人工智能生成作品著作权的承认与保护有效的激励了人工智能产品的使用及作品的创新，故从理论渊源以及实践效果来看，我国对达到作品要求的人工智能生成数据采用知识产权保护方式是完全具有合理性及可行性的。但是由于立法的缺失，我国理论界还是存在争议，司法实践中也缺失法律依据，为了更好地保护人工智能数据作品，应当尽快完善相关立法，明确人工智能生成作品受保护的条件。并且由于人工智能生成作品只能构成数据中的一小部分，故还需要与其他方式配套共同保护人工智能生成数据的权利。

第二，健全数据使用权保护层面的相关制度。欧盟《数据法》中的数据访问权在性质上属于数据使用权，⁴⁸欧盟及美国的立法经验证明赋予公众排他性的数据财产专有权的方式并不可取，易导致数据市场垄断，出现不公平、不平等现象，我国理论界也有许多学者主张采用赋予数据主体数据使用权的方式在保护数据权利的同时促进数据市场的流通。⁴⁹同时，我国在政策层面上发布了《数据二十条》，其中就指出要创建数据资源持有权、数据加工使用权、数据产品经营权三权分置的结构，其中数据加工使用权在性质上也都属于数据使用权，但目前我国对于数据使用权的探索只限于理论界及政策层面，暂时没有相关立法予以明确。

数据访问权一改以往对数据进行垄断确权的模式，赋予了数据在被各个主体使用过程中在主体之间流动的可能性。⁵⁰故为了促进数据流通，可以采用以使用权为重点的数据权利立法方案，以数据获取和使用规则为数据权利配置制度的重

⁴⁶ 参见王迁：《著作权法借鉴国际条约与国外立法：问题与对策》，载《中国法学》2021年第3期，第28-29页。

⁴⁷ 参见王迁：《著作权法借鉴国际条约与国外立法：问题与对策》，载《中国法学》2021年第3期，第28页。

⁴⁸ 参见吴汉东：《数据财产权：从数据专有权到数据使用权》，载《法商研究》2024年第3期，第7页。

⁴⁹ 参见王洪亮、叶翔：《数据访问权的构造——数据流通实现路径的再思考》，载《社会科学研究》2023年第1期，第71-72页。

⁵⁰ See Josef Drexler, "Data Access and Control in the Era of Connected Devices," Report for BEUC, 2018, p156-157.

心。人工智能数据主体对于其生成的数据都存在利益需求，使用权不具有所有权的排他性，可以合理满足数据流通过程各个阶段人工智能数据主体对数据的使用需求，并且可供多个主体同时享有，极大的促进了数据的流动与公平利用。包括数据访问权在内的使用权能够在尊重数据所有权的基础上，最大程度地促进数据流通，实现数据普遍共享⁵¹，能够作为数据财产权与数据共享之间连接的桥梁，打通二者的边界，实现数据财产权与数据共享的平衡，使数据在各个主体之间流动过程中促进人工智能数据的普惠。

需要明确的是，数据使用权与不动产中的建设用地使用权、宅基地使用权虽然在名称上具有相似性，但二者在权利属性上具有本质区别。具体而言，建设用地使用权人和宅基地使用权人对使用权具有完整的处分权能，而数据使用权人对使用权的权能存在严格限制，只能包括权利人对数据的使用、加工等权利，不能支配数据的交换价值及数据使用权的交换价值，无法处分使用权，这与动产及不动产等有体物的所有权存在较大的区别。并且由于数据的特殊性质，数据的使用不像有体物那样具有排他性，可以有多位主体同时占有或使用同一数据。而使用数据的不同主体之间又可能存在利益上的冲突，故各主体的数据使用权都会存在一定程度的限制，不能损害其他主体的数据利益。而根据贡献原则或利益最大化原则，不同的主体对数据的使用程度和范围都有所不同，故需要根据具体的场景与主体来确定具体的数据使用权限限制与边界。⁵²

第三，完善数据产权保护层面相关制度。数据权利配置制度制定的目标之一就是平衡数据财产权利与数据共享的关系，在促进数据流通的同时，也能更好的利用数据的财产性质以及流通性激发数据市场活力，为数字经济的发展注入活力。然而在欧盟的立法实践中，数据财产权与数据共享却很难达到平衡，《数据法》也重视数据共享而忽略了数据财产价值的激发。与欧盟不同的是，我国重视数据的财产性质，《数据二十条》中强调释放数据要素价值红利，完善交易制度，故我国的数据治理制度势必要兼顾数据财产确权与数据共享，实现二者的价值平衡。在欧盟《数据法》中，相较于作为消费者的用户，作为企业的数据持有者对数据的经济利益有更大的需求，故数据持有者对数据的权利更多的体现出财产性质，而用户对其产生数据的访问权一部分是出于用户的人格权利益，故数据访问权赋予用户权利，削弱数据持有者权利的规定本身就是数据财产性质的弱化。数据访问权强调数据持有者的无偿的数据提供义务，侧重数据共享，重视对数据流通性质的开发利用，这是将数据访问权的主体更多定位到了“数据”本身。然而数据加工后成为数字产品才具有经济价值，产生产权性质，我国《数据二十条》中的“数

⁵¹ 参见杨志航：《跨越企业数据财产权的藩篱：数据访问权》，载《江西财经大学学报》2023年第5期，第142页。

⁵² See Kerber W.: “Governance of IoT data: why the EU data act will not fulfill its objectives,” GRUR International, Vol.72, No.2, 2023, p120-135.

据”就定位于“数据产品”，规定的是数据产品的产权，强调数据加工以及数字经济。我国与欧盟对于数字经济价值理解不同的原因在于欧洲的重点放在“数据”提供，这是基于人权和安全思想的义务本位，是保守思维；而中国的重点在“数据产品”利用，这是基于安全、人权与市场思想的权利本位。数据的财产性质与流通性质不可偏废，只有将二者相结合才能更好的促进数据的流通与利用。

我国地方立法中推行的数据产权登记制度以及数据知识产权登记制度都是通过权利登记明确数据主体拥有的数据权利,从而保护数据主体的数据财产权利。但是目前的数据产权制度存在一系列的问题,如数据产权登记制度与数据知识产权登记制度混淆、数据产权登记机构分散、数据产权登记的客体不清晰、登记效力模糊及文件效力不够等问题。⁵³这些问题都可以归因于我国还未建立全国统一的数据产权登记体系,以及在数据财产确权方面国家层级法律的缺失,只有在立法以及制度设计上作出全国统一适用的规定,才能统领全国各地区的具体制度设定,从而为数据收益权做铺垫。

（三）人工智能生成数据权利配置制度效能的优化

第一,对构成作品的人工智能生成数据进行著作权保护。由于人工智能生成数据数量庞杂,类型众多,需要根据数据类型的不同采用知识产权保护、数据使用权赋权及数据产权登记公示制度三轨并行的路径对人工智能生成数据予以保护。在知识产权方面,应当完善我国《著作权法》,将人工智能生成作品纳入其保护范围,明确对人工智能生成作品进行著作权保护的条件。我国在实践中已经赋予了部分人工智能生成数据作品以著作权保护,那么在立法中也应当与时俱进,使立法与司法相匹配,在司法实践中有法可依。

第二,通过数据使用权及数据产权进行补充配套保护。对于不构成作品的数据,则需要在数据市场中利用使用权及产权来进行保护。在数据使用权赋权时,可以按照数据主体的不同对各主体赋予内容不同的数据使用权,例如,对人工智能使用者而言,他们对人工智能作出指令、安排,人工智能生成的数据具有一定的贡献,他们对生成的数据应当具有以访问、携带及使用为主要内容的数据使用权。对人工智能所有者而言,人工智能生成的数据通过云端上传至人工智能系统而为人工智能所有者所实际掌控,并且人工智能生成的数据是基于人工智能基本算法逻辑而生成,故人工智能的所有者对其生成的数据也具有一定的贡献。并且由于人工智能所有者一般为企业,对人工智能生成的数据具有财产价值利用的需求,需要维护商业地位及商业价值,故人工智能所有者应当拥有储存、加工、处理、利用为内容的数据使用权。⁵⁴对人工智能的设计者而言,其对人工智能的设计一般基于劳务合同等双务合同而作出,故其对人工智能生成数据享有的权利可

⁵³ 参见谭佐财:《论数据产权登记的制度构建》载,《当代法学》2024年第4期,第87-88页。

⁵⁴ 参见吴汉东:《数据财产赋权:从数据专有权到数据使用权》,载《法商研究》2024年第3期,第12页。

按照其与人工智能所有者的合同来确定。

数据使用权可促进数据的流通，而数据产权则可维护数据的财产价值。在构成作品的已有《著作权法》保护的情况下，数据知识产权登记制度的存在就不具有必要性及合理性。数据知识产权登记制度与数据产权登记制度设立的目的都是为了数据确权，维护数据财产权利，并且目前二者的立法极易混淆，故在知识产权已有法律条款保护的情况下，采用数据产权登记公示制度来保护数据财产权利即可。数据产权登记公示制度可以明确数据权利的边界以及归属，是推进数据要素确权、流通的基础性工作。数据产权的登记公示制度通过登记确定权利的归属主体，使得数据主体享有的权利具有了一定程度的排他性特性，维护了数据主体的数据财产权利，与以使用权为主的数据产权制度达成数据财产权利与数据共享的平衡。同时，数据产权的公示可以提高数据权利归属、变动的透明度，数据的利用往往涉及繁杂的过程以及多元的主体，数据产权的公示将数据权利复杂的流动与变动过程可视化，合理的划分了不同数据主体的权利边界，将数据权利的归属与流动变得清晰可见，兼顾了数据要素市场的公平以及效率，极大的促进了数据财产价值的释放。我国目前只有深圳推行了数据产权登记制度，鉴于我国目前数据产权登记公示制度的现状，我国需要根据地方相关试点规则的经验制定一套相关立法，以统一全国的数据产权登记公示制度，如确立统一的数据登记机构、登记平台、登记效力及权责范围等。⁵⁵而具体享有数据产权的数据主体的确定可以依据实现签订的合同进行确定。

可见，知识产权及数据产权具有一定的排他性，可以保护数据财产权利，数据使用权可以促进数据流通与共享，故知识产权保护、数据使用权赋权及数据产权登记公示制度三轨并行的路径可以平衡与兼顾数据共享与数据财产价值的目标，最大化的激励数据市场价值红利释放。同时，数据使用权的设置能够平等赋予各数据主体对数据的使用权利，促进数据的流通以及公平利用，使人工智能生成数据价值普惠更多社会公众。

五、结 语

党的二十届三中全会强调要“统筹立改废释纂，加强重点领域、新兴领域、涉外领域立法，完善合宪性审查、备案审查制度，提高立法质量”。⁵⁶欧盟、美国及英国的数据制度在一定程度上为我国人工智能生成数据权利配置的立法提供了借鉴，但其也有许多不足之处，我们在对待域外新兴法律时要以客观的态度对其进行分析及研究，切不可忽略国情的不同而盲目的照搬照抄。我国应当以一种

⁵⁵ 参见张素华、邓鹏：《数据产权统一登记体系的探索与制度构建》，载《电子政务》2024年第12期，第31-34页。

⁵⁶ “中共中央关于进一步全面深化改革 推进中国式现代化的决定”，载新华网，<http://www.news.cn/politics/20240721/cec09ea2bde840dfb99331c48ab5523a/c.html>，2025年4月12日最后登录。

开放、包容的态度应对国内外法律规范的守正创新，自主革除历史长期的“西法东渐”所导致的西方化的思维惯性和路径依赖，构建中国数字科技法治自主知识体系，以匹配服务推进中国式现代化。

在数字经济快速发展的背景之下，面对由于新兴领域数据权利配置立法缺失而导致的司法层面以及地方立法层面的混乱，我国人工智能数据权利配置的立法完善刻不容缓。人工智能生成数据权利配置制度的目的就是要促进数据公平共享、分配数据价值红利、推动“科技平权”“智能向善”。这就需要合理平衡数据财产权利与数据共享这两个看似矛盾实则关系密切的概念。同时，人工智能生成数据权利配置立法应秉持构建人类命运共同体思想理念，坚持遵循社会公平、数字正义及科技平权逻辑，以此建设公平、高效、安全的数据要素市场，推进中国式现代化和创造人类文明新形态。在设置具体法律规制时，由于人工智能数据权利配置涉及的环节以及主体过多且对权益的需求各不相同，相关立法存在一定的困难。这需要先通过对主体、对场景进行分类，厘清不同场景下各主体的数据权益需求，方可制定针对性的法律。

参考文献：

- [1]程啸：数据权益与数据交易[M]. 北京：中国人民大学出版社，2024。
- [2]张吉豫：数字法理的基础概念与命题[J]. 法制与社会发展，2022（5）。
- [3]马长山：面向“三维世界”的数字法学[J]. 中国社会科学，2024（11）。
- [4]张凌寒：数字正义的时代挑战与司法保障[J]. 湖北大学学报（哲学社会科学版），2023（3）。
- [5]吴汉东：数据财产赋权：从数据专有权到数据使用权[J]. 法商研究，2024（3）。
- [6]周尚君、罗有成：数字正义论：理论内涵与实践机制[J]. 社会科学，2022（6）。
- [7]王利明：数据何以确权[J]. 法学研究，2023（4）。
- [8]申卫星：论数据来源者权[J]. 比较法研究，2024（4）。
- [9]丁晓东：数据公平利用的法理反思与制度重构[J]. 法学研究，2023（2）。
- [10]张凌寒、杨建军、文禹衡、程莹、郑志峰：续章与新篇：中国人工智能法治研究 2024 年度观察[J]. 法律适用，2025（3）。
- [11]周尚君：数字法理学的基本范畴[N]. 光明日报，2025（2）。
- [12]丁晓东：全球比较下的我国人工智能立法[J]. 比较法研究，2024（4）。
- [13]Herbert Zech, “A Legal Framework for a data Economy in the European Digital Single Market: Rights to Use Data” [J], Journal of Intellectual Property Law &

Practice, Vol. 11, No. 6, 2016.

[14]Ali M.& Al-Khour, Data Ownership: Who Owns 'My Data'? [J], International Journal of Management & Information Technology, Vol.2, No.1, Nov.2012.

[15]Kerber W., “Governance of IoT data: why the EU data act will not fulfill its objectives”[J], GRUR International, 2023, 72 (2).

2025年中国国际法学会年会

网络空间命运共同体视域下的网络犯罪综合治理——以刑事检察案件办理为例

吴春妹 杨金玲¹

摘要：网络犯罪是世界和平与发展之公敌，网络犯罪的综合治理是国际社会共同面对的重要课题。小到仅具有涉外因素的电信诈骗案件，大到国家在网络空间的不法行为，如何处理和应对网络犯罪活动对现有法秩序的冲击，不仅考验一国法治能力水平，更检验国际法治的运行效能。习近平总书记在提出构建人类命运共同体理念时，鲜明指出了加快构建网络空间命运共同体的重要论述，为中国积极引导和参与国际社会协同加强全球网络犯罪治理指明了前行之路，提供了系统方法。新时期新阶段，刑事检察在打击和防范网络犯罪工作中，有必要从网络空间命运共同体理念出发，一体贯彻落实习近平法治思想和习近平外交思想，凝聚共同体意识、健全专项法规、完善协作机制，探究行之有效的国际刑事司法治理路径。

关键词：网络犯罪 命运共同体 涉外法治 国际刑事司法协助 体系化治理

一、网络空间命运共同体理念与网络犯罪综合治理

网络犯罪是数字时代的新型安全威胁，打击治理网络犯罪绝非单靠一国之力。网络犯罪天然具有的跨国性、专业性、隐蔽性特征，带来的是非传统安全挑战，威胁地区稳定，对其的打击与治理，必然不同于一般刑事犯罪行为，需要理念先行、法治保障、机制指引。12年来，习近平总书记提出的人类命运共同体理念，理论内涵不断丰富、理论体系日益完善，得到国际社会的广泛认同和支持，被多次写进国际决议文件，正在从理念转化为行动、从愿景转化为现实；该理论体系延伸出来的网络空间领域命运共同体理论架构，倡导“共商共建共享”的全球网络治理观，打击治理网络犯罪正是对该理念的先行实践。由于各国在网络犯罪的界定、证据标准、法律适用等方面存在差异，相关刑事司法协助工作尚有理念、政策、机制层面的分歧需要消弭，呈现各自为战的情势：

¹ 吴春妹，北京市顺义区人民检察院党组书记、检察长、二级高级检察官；杨金玲，中国人民大学国际法法学硕士，现为北京市顺义区人民检察院经济犯罪检察部一级检察官助理，电话：18010208619，邮箱：lycongrong@163.com，地址：北京市顺义区新顺南大街19号，邮编：101300。

（一）网络犯罪综合治理共同体意识不统一

1.罪责刑标准有冲突。互联网应当是造福人类的工具，但是各国网络领域发展状况不同，对涉网冲突的性质认定和处理方式存在地域化特征。以网络赌博活动为例，依据中国刑法及司法解释规定，中国禁止任何形式的赌博行为，组织或参与跨境赌博的行为明确为犯罪行为；柬埔寨《商业博彩管理法》同样全面禁止网络赌博，而菲律宾实行许可证制，中国公民经许可在菲律宾运营博彩公司受其国内法保护，这也因此导致中柬、中菲在查办网络赌博案件的时效不同，联合行动依赖双边协议²，缺乏全球协作机制。

2.联合治理信任危机。人类是休戚与共的网络空间命运共同体，网络主权是国家主权的重要部分，但是网络犯罪领域的执法司法活动极易被放大为干涉他国主权的行爲，涉及的数据要素的所有权争议上升为“数据主权”之争，沦为他国实施霸权的工具。以跨境网络诈骗犯罪为例，证明外国犯罪嫌疑人非法占有主观明知的关键证据，往往存储在境外作案工具中，但外国以“数据主权”为由拒绝向受害国提供，加之国家间未建立任何信息共享机制，导致跨境取证难以进行，刑事诉讼程序因证据链条断裂陷入停摆。

（二）网络犯罪综合治理成文法体系不健全

1.专门性国际公约缺失。“条约必须遵守（Pacta sunt servanda）”是国际法基本原则，是国际条约在国内法中的适用。国际法院依据国际法规则裁判案件，违反条约义务必将引起国际责任。在2024年12月24日《联合国打击网络犯罪公约》（以下简称《网络犯罪公约》）成型前，网络犯罪领域无论是全球性公约还是区域性公约均为空白，更无从谈及公约的体系化建设，基本仅靠《联合国打击跨国组织犯罪公约》，在跨国洗钱、腐败、贩毒等类别领域进行有限制的打击和治理，无法应对新兴跨境网络犯罪形态。现在，《网络犯罪公约》覆盖了网络诈骗、网络洗钱、网络恐怖主义、网络攻击等10余种网络犯罪的罪与责，要求成员国通过国内立法予以落实、开展必要的司法协助和技术合作，成为首部具备全球法律约束力的综合性网络犯罪公约，标志着国际社会在统一网络犯罪打击治理共识上实现重要突破。但是，也不可避免的面临西方国家“长臂管辖”困境，比如在跨境电子证据调取程序上未达成一致意见，管辖权争议依然存在；再比如，网络犯罪要素规定单一，过失网络犯罪的罪责刑规定仍是空白。

2.国际习惯发挥作用不足。国际习惯是国内法的一部分，在传统国际条约不能覆盖多元化犯罪形态规制的当下，国家执法实践可以起到补充性作用。虽然各

² 2023年9月16日新华社发布《中华人民共和国政府和柬埔寨王国政府联合公报》，提及不断深化两国国家及地方警察在打击跨境赌博、电信诈骗等执法领域的务实合作和能力建设，开展“一带一路”重大项目安保合作。
https://www.mfa.gov.cn/web/gjhdq_676201/gj_676203/yz_676205/1206_676572/1207_676584/202309/t20230916_11144074.shtml

国对网络犯罪的管辖权争议不断，但对于网络诈骗类犯罪的普遍管辖基本形成共识，借助国际习惯法的审慎义务之规定，可以要求各国平衡本国主权的行使与他国跨境执法权利的保障，缔结区域性公约，防止发生于本国的网络攻击活动对他国造成损害责任。以缅北回流电信诈骗案件为例，东南亚国家内部建制力量显赫，擅长抓住人民求安稳、企业求发展、外部地缘环境紧张的机会，巩固自身地位，导致缅北地区长期被地方反动武装势力控制的局面³。电信诈骗产业原是当地发展主要依赖的收入来源，电诈分子与缅甸非国际性武装冲突分子长期利益联结，使得缅甸政府为防止内战风险而消极处置边境诈骗犯罪，加剧中国逮捕境外诈骗头目的重重困难。截止目前，缅北部分地区的电诈园区基本清除，但要彻底清除电诈毒瘤，中国需要综合运用司法、外交、技术手段推进区域法律合作实践，在尊重缅甸主权的前提下，推进国际人道法得到尊重和遵守，提升跨境执法效率。

（三）网络犯罪综合治理合作机制单一

1.专门机构匮乏。现阶段联合国机构框架下，尚无一个专门性机构负责统筹跨国网络犯罪打击和治理工作，跨境网络犯罪的执法打击依靠国际刑警组织的协调统筹，跨境网络犯罪的司法惩处多依靠国内检察机关的刑事公诉职能。需要关注的是，国际刑警组织没有直接执法权，只能依赖成员国自觉配合，在缅北电信网络诈骗头目抓捕行动中，执行效果会受缅北非国际性武装冲突分子的影响大打折扣；各国检察机关在全链条追诉网络犯罪中发挥着重要作用，以中国检察机关为典型，去年全年起诉利用网络实施的犯罪 27.2 万人，起诉电信网络诈骗犯罪 7.8 万余人，对缅北明家犯罪集团及关联犯罪集团成员起诉 39 人⁴。网络犯罪检察工作现已形成其“检察+技术”的鲜明特色工作模式，但是在破解部分网络犯罪罪名认定上，如破坏计算机信息系统罪，依赖数智技术赋能。

2.协作机制低效。现有国际或区际协作实践缺少具有强制执行力的合作机制约束。在联合国框架下，仅有新生的《网络犯罪公约》列举了建立跨境电子取证规则、司法协助流程及情报共享平台，简化引渡程序，推动多边协作的国际司法合作机制的内容，特别关注发展中国家的能力建设需求，通过技术转让、培训等方式，提升国家的网络安全能力，但达成成员国的执行共识需要一定期限；在区域性协作中，各国检察机关的司法协助运行常受制于经济竞争、政治博弈、司法程序因素的影响，具有不确定性，如中国在与缅甸联合打击电诈园区行动中，通过切断园区基础设施、联合通缉等方式，极大提高了跨境执法效率，但是在与美国的司法协助方面则充满对抗性，美方以国家安全为核心主张数据自由，推行长

³ 摘自《国际形势和中国外交蓝皮书 2023/2024》，中国国际问题研究院著，世界知识出版社，第 63 页。

⁴ 出自 2025 年 3 月 8 日《最高人民检察院工作报告》，
https://www.spp.gov.cn/spp/gzbg/202503/t20250315_690544.shtml

臂管辖和双重标准，意在凸显其在扩张域外法适用上的单边主义立场。

二、中国检察行权在推进网络犯罪综合治理中的效能

现阶段，全球网络犯罪治理的“政府主导、多元协作”特征凸显。“政府主导”以我国打击缅北回流电信网络诈骗犯罪活动体现最为突出，“多元协作”以公共司法机关联合其他主体打击网络安全犯罪活动最为突出。中国检察机关是宪法赋予唯一法律监督职能的司法机关，发挥其网络犯罪案件审查、法律监督、司法协作功能，生动践行网络空间命运共同体理念。

（一）建构网络犯罪证明体系框架

1.推动刑事检察立法。中国坚持依法治网，完善网络领域刑事法律制度，全面推进网络空间法治化建设。在刑事司法层面，最高检配合《刑法》《刑事诉讼法》具体条款的修订，单独或会同有关部门研究制定网络诈骗、网络暴力、网络侵犯知识产权等领域的司法解释和规范性文件，细化网络犯罪关联罪名证明标准，确保刑事检察行权依法运行，努力对接国际公约要求。比如，明确了涉及网络犯罪的专门刑事罪名有非法控制、破坏计算机信息系统罪、非法获取计算机信息系统数据罪、拒不履行信息网络安全管理义务罪、非法利用信息网络罪等，针对理论与实务界争议较大的帮助信息网络犯罪活动罪的设置上，出台系列工作意见，缩小事实认定和法律适用的争议；在维护网络案件和信息保护立法层面，《网络安全法》《数据安全法》《个人信息保护法》等框架基本形成，作为刑事检察融合民事检察、行政检察、公益诉讼检察职能开展综合履职的依据。

2.推进刑事检察融合履职。法律的生命力在于实施，为确保法律统一适用，最高检牵头制定专项工作意见，构建稳妥、有序的网络空间检察治理体系。2023年3月，最高检印发《关于加强新时代检察机关网络法治工作的意见》，立足“四大检察”职能，从网络立法、执法、司法、普法以及法治研究、队伍建设等方面作出全面部署；2023年4月，最高检印发《深入推进“净网2023”专项行动工作方案》，强调依法严惩网络违法犯罪，全面提高依法治理、综合治理、源头治理网络空间的能力和水平。在2024年全国检察长会议和2023年印发的《2023—2027年检察改革工作规划》中，维护网络安全、推进网络治理成为重要内容，在全国部署推进。

（二）高质效办好涉外涉网案件

自最高检于2024年12月全面部署加强涉外检察工作任务以来，跨境网络犯罪被界定为具有涉外因素的“大涉外”案件而受高度关注，已上升至跨国犯罪的地位高度。以笔者所在的北京市基层检察机关为例，缅北回流电信诈骗案件占全部诈骗犯罪的七成。为依法惩治跨境制贩毒、网络赌博、电信网络诈骗等犯罪活动，检察机关在提前介入、审查逮捕、审查起诉各环节引导侦查，严格把握好涉网案

件的证据关、事实关、法律适用关，借助数字技术赋能涉外办案，深挖彻查上下游犯罪，斩断网络黑灰产业链。比如，聚焦电信网络诈骗犯罪上下游关联犯罪高发问题，严惩缅北电诈犯罪集团及头目，分层分类处置共犯和掩饰、隐瞒犯罪人员，把稳帮助信息网络犯罪活动的边界，全面开展追赃挽损工作；再如，聚焦非法获取、提供个人信息类犯罪，重点加强对上游信息收集、倒卖环节的犯罪线索挖掘，起诉侵犯公民个人信息犯罪的同时，支持民事起诉，保障公民、社会的合法权益；加大对“人肉开盒”⁵等侵害公民人格权违法行为的公益诉讼监督办案力度，助力营造文明、安全、绿色的网络环境。北京市检察机关现正全面开展体系化建设涉外检察工作中，进一步完善域外法查明、涉外案件翻译等工作机制，规范高效办理刑事司法协助案件。

（三）推进网络空间协同综合治理

1.健全区域网络协同治理机制。为依法打击跨境电信网络诈骗行为，全国检察机关积极落实亚洲联合执法行动，开展“断卡”“断流”专项行动。为维护网民在网络空间的合法权益，连续开展“净网”专项行动，打压网络犯罪活动空间。为联合打击网络恐怖主义犯罪，中国在上海合作组织框架下签署《阿斯塔纳宣言》《共同打击国际恐怖主义的声明》等重要文件，坚决执行联合国安理会有关决议，开展打击网上暴恐音视频专项行动。为推进全球网络犯罪相关数据共享，在金砖国家总检察长会议中提出深化电子数据取证合作等内容，借助上海合作组织总检察长会议，建立跨国犯罪证据共享平台，深化网络犯罪领域司法协助；执行《上海合作组织成员国保障国际信息安全政府间合作协定》，《独立国家联合体打击计算机信息领域犯罪合作协定》等，积极参与区域网络空间治理进程。

2.参与国际网络综合治理进程。中国认真履行缔结和加入的条约，支持在联合国框架下制订全球性公约，维护以联合国为核心的国际体系、以国际法为基础的国际秩序和以联合国宪章宗旨和原则为基础的国际关系基本准则，一贯支持打击网络犯罪国际合作。《网络犯罪公约》的谈判成果，与中国在过去的十余年间推动联合国网络犯罪政府间专家组数次决议的努力密不可分，为启动制订联合国打击网络犯罪全球性公约相关决议作出重要贡献。中国与上海合作组织其他成员国向联合国大会提交“信息安全国际行为准则”，是国际上第一份系统阐述网络空间行为规范的文件。中国建设性参与联合国信息安全开放式工作组与政府专家组，推动联合国信息安全开放式工作组与政府专家组成功达成最终报告，为网络空间国际规则的制定与网络安全全球治理机制建设奠定了基础。中国深度参与联合国互联网治理论坛，积极参与联合国互联网治理论坛领导小组、多利益相关方咨询

⁵ 北京市检察机关近两年来审查多件利用“开盒挂人”手段实施网络暴力案件，反映出一种新型网络暴力违法犯罪现象，指不法分子通过非法手段在海外“社工库”等进行网络搜索、挖掘，搜集个人隐私信息，并恶意公布在网络空间，被“开盒”人往往会遭到网民的侮辱谩骂、造谣诋毁。

专家组，连续多年在联合国互联网治理论坛主办开放论坛、研讨会等活动，与来自全球各界及非政府组织代表开展广泛交流讨论。

三、网络空间命运共同体理念下的网络犯罪体系化治理路径

（一）构建网络犯罪治理利益共同体

习近平主席提出全球文明倡议、全球发展倡议、全球安全倡议在内的“三大倡议”，系统阐释了构建人类命运共同体的实践路径。正如构建人类命运共同体必须践行真正的多边主义，网络空间综合治理要以网络空间命运共同体理念为支撑，倡导各主权国家在全球网络犯罪治理中增强共同安全意识，打造利益联结共同体⁶。

1.坚持网络主权平等原则。网络犯罪问题首先是各国内政事务，倡导各国将《联合国宪章》确立的主权平等原则适用于网络空间，弥合政治利益鸿沟，在互相尊重各国网络立法执法司法主权的实际下，凝聚共同打击和抵御网络犯罪的共识，构建公正合理的网络空间国际秩序，杜绝利用网络犯罪实施侵害他国主权、安全和发展利益的行径。

2.坚持国际刑法基本原则。各国在自主选择网络发展道路、制定网络监管政策的同时，不得从事危害或纵容、支持危害他国国家安全的网络活动，应当尊重他国平等参与网络空间国际治理的权利，秉持罪责刑相适应和普遍管辖的国际刑法基本原则。

（二）构建网络犯罪治理法律共同体

1.积极推进网络公约进程。国家是制定、遵守、实施和发展国际法的基本主体，各国司法机关应当积极参与联合国网络安全工作进程，支持联合国在网络空间全球治理中发挥核心作用，加快出台全球统一的网络犯罪国际公约，不断完善网络空间治理的国际规则和标准，为打击网络犯罪提供明确的工作指引，确保国际刑事打击步调基本一致，为各国国内法与国际法的有效衔接提供国际法保障。中国检察机关倡导同各国司法机关共同努力，根据国际公约和相关法律坚决打击各类网络犯罪行为，支持并积极参与《网络犯罪公约》谈判，深化务实合作，让全球互联网治理体系更加公正合理，更加平衡地反映大多数国家特别是发展中国家的意愿和利益。

2.推进解决域外法适用争议。各国对网络犯罪秉持的管辖权原则不同，导致管辖权冲突；构成要件的认定存在差异，导致法律适用冲突；证据的调取规则各不相同，为证据效力的认定带来困难。为力促网络犯罪线索成案和保障刑事诉讼程序顺利进行，各国应结合数字时代网络犯罪的特点，灵活确定管辖权原则，完

⁶ 摘自 2023 年 9 月 26 日《携手构建人类命运共同体：中国的倡议与行动》，中国政府网，https://www.gov.cn/zhengce/202309/content_6906335.htm

善取证机制，减少法律冲突，努力保全关键证据。如，借鉴欧盟《电子证据条例》(EU Electronic Evidence Regulation)建立的跨境电子数据取证模式，完善本国跨境证据调取机制，从属地主义转变为属人主义，检察机关可直接向网络服务供应商调取和保全用户数据，而无需前往数据存储地通过司法协助程序调取⁷。

(三) 构建网络犯罪治理责任共同体

1. 组建专门打击治理机构。国际组织是国际法的重要主体，是各国进行网络犯罪打击合作的重要法律形式，应当积极发挥《联合国打击网络犯罪公约》框架下的积极作用，淡化国别界限，统筹各国刑事司法机关之间在网络犯罪领域的信息共享、联合执法、司法协助、技术支持等。如欧盟设立欧洲司法网络、欧盟刑事司法合作署（Eurojust）以及欧洲检察官办公室等区域司法协助组织机构，用以协调欧盟国家之间和欧盟国家与欧盟以外国家之间在调查和起诉跨境犯罪方面的工作，应对跨国犯罪案件的证据收集、移交和认定困难⁸。德国刑法、国际刑事司法互助法也以国内法条款形式对欧盟的区域协作规则加以确定。

2. 形成跨区域协同治理机制。网络犯罪已经成为全球公害，利用双边、多边国际合作平台，推进区际、国别刑事司法合作，是打击和治理网络犯罪的必由之路。俄罗斯联邦总检察院要求检察机关应与所有相关机关加强合作打击网络犯罪，特别是打击使用特殊技术设备进行的诈骗犯罪，如通过大量公民拨打电话进行诈骗案件⁹。中国检察机关已建立起一套行之有效的网络犯罪检察一体化履职机制，对网络犯罪开展全链条化打击整治，打造中国特色网络犯罪执法司法协作范式：一是加强“四大检察”融合履职。在依法严惩涉网刑事犯罪时加强刑事诉讼监督，一并审查民事侵权、行政监管缺失、侵害公共利益行为线索，建构刑事附带民事、行政及公益诉讼体系。二是加强部门横向联动。建立办案衔接工作机制，加强同政府部门、国际组织、互联网企业、技术社群、民间机构、公民等各主体作用，推进网络犯罪溯源监管。三是加强区域协同治理。在上级检察机关的统一部署下，开展国内跨地域协作和区际刑事司法协作，统分建立线索移送、调查取证、法律适用、司法鉴定等工作机制，有效解决网络犯罪惩处效率不同步问题。四是

⁷ <REGULATION (EU) 2023/1543 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 July 2023 on European Production Orders and European Preservation Orders for electronic evidence in criminal proceedings and for the execution of custodial sentences following criminal proceedings>, Article(7).

⁸ <Common challenges in cybercrime>, European Union Agency for Criminal Justice Cooperation, <https://www.eurojust.europa.eu/sites/default/files/assets/files/common-challenges-in-cybercrime-eurojust-2024.pdf>

⁹ 摘自 2024 年 3 月 27 日俄罗斯联邦总检察院克拉斯诺夫总检察长在检察机关扩大会议上所做《关于 2023 年检察机关工作成果和 2024 年维护法律秩序的任务的报告》。该报告显示，俄罗斯近年网络犯罪在国内犯罪总数中占比首次达到 35%，只有四分之一案件得以破获，追回损失金额极少，移交法院的案件追回损失比例也不超过三分之一。

加强多边司法合作。中国检察机关借助上海合作组织、新加坡总检察长会议等平台，凝聚网络犯罪惩治双边合作共识；借助“一带一路”倡议推进沿线国家的网络基础设施安全保护合作，务实推进网络犯罪与网络安全执法合作对话成果。

四、结语

网络空间命运共同体理念，已经凝聚起国际社会合作推进网络犯罪综合治理的基本共识。面对世界百年未有之大变局，各国司法机关只有用更宽广的国际法胸怀、更务实的法律框架和更团结的司法合作，推进网络犯罪综合治理，才能真正实现构造共商共建共享的网络空间命运共同体。

2025年中国国际法学会年会

异化与重塑：总体国家安全观下 AIGC 涉意识形态风险治理策略研究

肖楠 魏琦宗 陶羨华 王汉晓 *

摘要：AIGC的兴起和深化运用过程中存在跨境数据主权风险、算法偏见风险以及生成内容偏离主流价值导向风险等三重隐患，隐患的背后潜藏着对国家意识形态安全的威胁，亟需在总体国家安全观的指引下重塑AIGC治理的逻辑面向，确立以风险平衡、全面参与和审慎相对为基准的治理原则来防范化解系统性风险，落实到实践中就是在数据主权上以“一带一路”倡议为契机，完善数据合规管控体系来建立中国式跨境数据流动安全治理国际规则，掌握构建数据流动规则和意识形态导向的话语权；在算法治理上以敏捷治理模式纠正算法偏见，明确“看门人义务”让渡企业部分算法解释权，构建算法审查备案制度开展预防性监管；在生成内容上以软法和硬法结合引导人机价值对齐，以软法构建价值对齐的道德性框架，以硬法构建价值对齐的基本导向，实现生成内容与我 国主流价值观协调一致。

关键词：总体国家安全观 AIGC;意识形态 人机对齐 敏捷治理

2014 年，习近平总书记首次提出要准确把握国家安全形势变化新特点新趋势，坚持总体国家安全观。此后，坚持总体国家安全观被写入《中华人民共和国国家安全法》，强调政治、经济、文化等各领域需全方位构建国家安全体系。2024 年 7 月中国共产党第二十届中央委员会第三次全体会议通过《中共中央关于进一步全面深化改革 推进中国式现代化的决定》（以下简称《决定》），将推进国家安全体系和能力现代化专章阐述，要求在全面贯彻总体国家安全观基础上完善重点领域安全保障体系。同时，网络安全作为国家安全的最重要组成部分之一，《决定》要求完善生成式人工智能发展和管理机制。

* 肖楠，江苏省连云港市海州区人民检察院党组书记、检察长。

魏琦宗，江苏省连云港市海州区人民检察院，第六检察部副主任，联系电话：15205136092

邮箱：15205136092@163.com 地址：江苏省连云港市海州区龙西南路 31 号 邮编：222000

陶羨华，江苏省连云港市海州区人民检察院党组成员，副检察长；

王汉晓，江苏省连云港市海州区人民检察院派驻板浦检察室副主任

* 基金项目：本文系 2024 年国家社会科学基金青年项目“行刑共治视域下滥用数据要素信用惩戒机制研究”（项目编号：24CFX099）阶段性成果；2024 年度北京市法学会市级法学研究一般课题“行政执法与刑事司法‘反向衔接’问题研究”（项目编号：BLS（2024）B019）阶段性成果。

需要警惕的是,人工智能的长足进步使得技术与信息社会的嬗变与迭代速度不断加快,以 ChatGPT 为代表生成的 AIGC 爆发式蝶变提升对国家安全体系建设提出了新的挑战。西方科技企业将人工智能与社会运行相耦合、类 ChatGPT 模型与图像生成、代码编写等工具软件结合的过程中极有可能出现国家数据主权受侵犯、算法模型偏见、生成内容挑战主流价值观等意识形态导向问题。在可以预见的将来,AIGC 在各行各业深度运用而产生的巨量信息将带来巨大的监管漏洞,产生意识形态风险,对国家安全形成新的挑战。总体国家安全观视阈下的国家治理体系和治理能力现代化亟需回应生成式人工智能可能带来的意识形态安全风险,多维度监管体系构建方向仍需更加深入地研究,如此方能将人工智能发展更好的耦合于中国式现代化进程中。

一、AIGC发展的三重国家安全风险

内容生成式人工智能的本土化建设是因应信息化潮流,回应西方数字霸权、维护国家安全的重要组成部分,但是鉴于中国的特有国情及对 AIGC 的后发思辨,伏脉于其中的梗阻是本土化发展必须面对的问题。总体国家安全观作为系统性理念,对其发展风险的探讨应贯穿全流程,即数据搜集阶段的数据主权风险、生成阶段的算法运行风险以及内容输出阶段的挑战主流价值观。

(一) 数据搜集阶段的跨境数据主权风险

国际社会因互联网时代开放、互联的精神在初期秉持“去网络主权”的理念,但随着网络攻击致使国家安全受到威胁,“网络再主权化”逐渐成为主流。近年来,网络时代演进为数据时代,数据重要性的提升使得数据主权的概念愈加清晰。数据主权是在法律维度下对数据安全阐释的衍生概念,在学界对数据安全领域的法律概念厘清后,数据主权作为与国家意识形态安全息息相关的概念备受学界关注。涵盖境内社会治理数据、国家安全核心数据等多维度的数据主体显然是国家和政府,该权属主体和数据性质决定了对数据主权的保护标准要远高于普通数据,其核心思想是国家对境内数据拥有控制权和管理权,并且必须加强对数据的监管和保护。¹信息时代的数据既是 AIGC 发展的养料,也事关国家主权的维护,而其对数据使用的必要性导致国家数据主权的独立性可能受到冲击。因此对数据主权的维护在 AIGC 规制方面与防范化解意识形态风险是一致的。

作为算法模型其内容生成的离不开海量的数据支撑,在全球范围内流动的数据与信息的导入与训练下才有现在突破式发展的人工智能。AIGC 的训练阶段也是其数据的搜集阶段,其存在的基石是以数据为核心构建的语料库。因为 AIGC 的更新迭代很大程度取决于语料库的更新迭代,因此数据搜集阶段的核心目标就

¹ 钊晓东. 论生成式人工智能的数据安全风险及回应型治理 [J]. 东方法学, 2023, (05): 106-116.

是构建独属于自身的语料库。从现有的发展趋势来看，AIGC 行业默认对语料库的前期构建及后期更新的形式分为主动抓取和被动输入两种。

语料库的主动抓取主要是通过数据爬虫技术对互联网上存在的信息进行搜集汇总，此种快捷高效的技术使得 AIGC 设计者在初期得以快速建立相对完善的语料库。数据爬虫技术的广泛应用得益于国际社会对数据流通价值理念的认可，但是数据爬取的正当性边界亦有对其限制，²数据爬虫技术抓取的自动性、便捷性也导致其抓取数据极易突破正当性边界，甚至因技术壁垒导致被抓取者难以意识到数据主权被侵犯。我国刑法虽对滥用数据爬虫技术以第 285 条、286 条予以规制，但较国家安全而言事后规制难以挽回国家层面的安全损失。

语料库的被动搜集是通过记录、保存与用户交互的数据得以实现的，相较于主动搜集，在数据量上远小于主动抓取，但是基于用户交互的有效性，其数据的质量远高于主动抓取。基于对高质量数据的需求，AIGC 运营者对交互数据的保有积极性是极高的，但是缺乏国家安全意识的企业用户在使用交互过程中难免会输入涉及公司内部数据的信息，海量企业内部数据经汇总后对国家安全的威胁不容忽视。

（二）内容生成阶段的算法偏见及不透明风险

各 AIGC 运营商建立了多种机制训练自身模型，本质上是为算法模型迭代到一定高度的产物，在数据搜阶段后，其运算生成过程的价值取向难免受到设计者和语料库价值取向的主导，难以做到客观中立的算法偏见蕴含着意识形态的风险。

风险不仅仅在技术应用的过程中被生产出来，而且在赋予意义的过程被生产出来，³从总体安全观的角度来看，现如今中美竞争对抗日益激烈，AIGC 处于领先地位的西方科技企业若通过对算法模型进行修改，开发者对语料库进行有选择的删减，按照自身意识形态的标准对 AIGC 进行迭代，其输出的结果将更加具有算法偏见和西方价值导向。北京理工大学计算机学院相关研究发现，“Chat GPT 生成内容中存在对于中国的大量偏见言论”。⁴

在 AIGC 的模型训练过程中，企业普遍采取了人类反馈强化学习（RLHF）的模式，该模式将人类的反馈纳入训练过程，为机器提供一种自然的、人性化的互动学习过程的方法，通过自然人为 AIGC 输出结果打分的形式推动其不断迭代完善其输出结果。在自然人不断标注反馈与 AIGC 运算输出的交互中，最终将输出结果转化为预想状态。参与数据标注的自然人对数据的标注本身就基于自身西方的价值取向，其思想和价值偏好将进一步加剧 AIGC 的算法歧视。

² 许可. 数据爬取的正当性及其边界 [J]. 中国法学, 2021, (02): 166-188.

³ 亚当·贝克龙. 风险社会及其超越：社会理论的关键议题 [M]. 赵延东，马缨，等译. 北京：北京出版社，2005：

⁴ 张华平, 李林翰, 李春锦. ChatGPT 中文性能测评与风险应对 [J]. 数据分析与知识发现, 2023, 7 (03): 16-25.

AIGC 算法的偏见源于语料库本身的偏见，从整个互联网语言体系而言，英语作为使用范围最广泛的通用语言地位在互联网领域更加凸显，其构成的语料库大量来源于英语语言体系，作为算法的生产资料，数据的来源就决定了其本质上就蕴含着算法偏见的导向以及背后的意识形态风险。

AIGC 算法的不透明性进一步加剧了开发者与相对人间的技术鸿沟，“黑箱”状态的算法对监管构成挑战。欧盟和美国针对此问题较早出台了关于算法解释权的法律规定，我国在《生成式人工智能服务管理暂行办法》（以下称暂行办法）第 19 条也明确了开发者对算法需说明的要求，但是对算法解释权的相对人进限于监管机构等特殊主体，且解释范围相对笼统，在实践中难以应用。

内容输出阶段的价值导向偏离

不同于以往生成内容的机械性、僵化性，AIGC 内容生成的快速、便捷是其强人工智能性的体现，生成内容的涵盖范围、论述深度都远超以往。在此基础上结合图像生成、语音生成等工具，内容输出呈现极高的泛化和迷惑性，足以达到以假乱真的程度。AIGC 内容表面上展现出高度的创造性和逻辑性，但是究其本质，其生成的内容是基于对语料库的再现和排列，是算法模拟下的数学再现，并非像人类一样从理性和感性双向交织的产物。单一且标准化的生成标准背后则难免潜藏着受技术开发者操纵的风险，对国家政治安全构成潜在威胁。AIGC 生成内容对国家政治安全构成威胁主要体现在以下两个方面：

其一，其内容生成存在潜移默化间为西方价值观、政治色彩背书，作为意识形态渗透工具的可能性，将有可能对意识形态甚至是国家主权造成不利影响。民粹主义借助互联网打造意见领袖的人设，在借助 AIGC 的生成能力后，能够以流水线形式大量生成包含历史虚无主义、自由主义等思想的内容，在数字技术的加持下其思想的传播速度和广度远超以往，破坏人民的文化认同根基和民族认同感。AIGC 与系统大数据推送相结合将部分群众困于精心编织的信息茧房，以所谓西方自由思想反复对群众进行冲击，将人群以不同标签割裂为差异化群体，进而挑战主流意识形态。同时，群众在自身使用 AIGC 产品时，基于 AIGC 的强交互性，生成内容的意识形态导向又进一步强化使用者对所谓西方自由思想的偏好。

其二，AIGC 生成虚假政治信息可能致使社会动荡的威胁。在以 Sora 为代表的人工智能文生视频大模型发布后，现实与虚拟的界限被进一步模糊，AIGC 生成内容由单纯的文字信息演进到图文、视频领域。若被恶意利用生成虚假的涉政治新闻、视频，足以对国家政治稳定构成威胁。如 AIGC 领域相对普及的换脸功能，在对涉政治人物进行剪切、拼接后足以对群众产生不良误导。

AIGC 的发展演进形成的连锁反应对社会治理的前瞻性提出更高的要求，也对国家安全的治理理念革新、治理体系和治理能力现代化提出新的要求。

二、总体国家安全观下AIGC治理的逻辑面向

（一）风险平衡原则

总体国家安全观在对国家安全理解与实践深化的基础上创新性的对安全与发展的关系进行论述，确认了“发展是安全的基础，安全是发展的条件”的理念，二者辩证关系的认识对平衡 AIGC 的发展风险与国家安全保障起到至关重要的指导作用。

网络国家安全从“网络主权”演进至“数据主权”，互联网技术的发展使得以 AIGC 为代表的信息科技在国家安全中的地位愈加凸显，AIGC 将带来文明发展模式的根本转型，《决定》中专门论述要求完善生成式人工智能发展和管理机制就是其体现。作为下一轮新质生产力的发展方向之一，因其对国家安全的潜在威胁而放弃其发展显然不符合客观规律，在限制其发展而导致技术落后反而可能加剧对国家安全的威胁。在新技术革新过程中若对风险过度恐慌反而可能扼杀 AIGC 的重要创新，过度严格的监管在一定程度上可能会影响国家在人工智能领域的竞争力。⁵西方科技企业在 AIGC 领域的领先就因其在发展初期所采取的包容开放态度。从风险治理的视角来看，国家治理体系应当能够系统性的持续防范化解内外国家风险，且在此过程中避免颠覆性的威胁。AIGC 带来的数据主权风险、算法偏见等风险是新技术、新生产力在发展中的负面产物，但是也带来防御境外势力技术进攻的新方向。

在对 AIGC 发展风险的规制方面，要在总体安全观的指引下，秉持平衡的原则，认识到囿于技术发展的固有弊端，摒弃传统工业社会强控制模式，⁶追求数据安全的动态平衡，平衡好技术进步和国家安全间关系，在保证威胁可控、可管的基础上鼓励 AIGC 发展反而可有效提高国家安全水平。

（二）全面参与原则

总体国家安全观之方法论深化了系统性、综合性国家安全风险防范的理念，在其对国家安全风险本体“五大范畴”“十一大安全领域”的认知下⁷，对涉 AIGC 国家安全风险的防范化解主体要突破传统的安全观。传统安全观将防范化解重大国家安全的责任集中于政府，总体国家安全观对国家安全总体性、综合性的要求在针对高速迭代的 AIGC，单一监管主体的参与难以因应新形势的需要。在总体国家安全观的指导下应采取各主体全面参与的原则共同应对 AIGC 的发展风险。

⁵ 郭春镇. 生成式 AI 的融贯性法律治理——以生成式预训练模型(GPT)为例 [J]. 现代法学, 2023, 45 (03): 88-107.

⁶ 王宏伟. 总体国家安全观指导下的国家安全系统性风险及其治理 [J]. 现代国际关系, 2017 (11): 50-57.

⁷ 张真源,张富利. 人工智能发展的系统性风险治理：实践反思、理性建构与策略优化 [J]. 学术论坛, 2024, 47 (01): 133-148.

世界各国 AIGC 的发展主要依托于科技企业的自主创新,《决定》亦在深化科技体制改革中明确要求强化企业科技创新主体地位,建立培育壮大科技领军企业机制,加强企业主导的产学研深度融合。⁸在总体国家安全观视阈下应秉持全面参与的原则,强化企业对内自我监控的看门人义务,对企业的自我监督标准予以明确,充分发挥科技企业对防范化解风险的内在科技优势,同时运用企业技术于对外网络防御,以企业的行业自律、自我监控强化国家安全水平。

AIGC 本质作为一种高精尖端技术,政府作为监管主体因其技术水平在专业性、可操作性上存在天然弱势,构建以第三方独立机构对企业进行专门审查的治理模式,以扩充参与主体的方式形成更加稳定可靠的国家安全维护模式。

(三) 审慎相对原则

传统绝对安全观源于理性化思考,强调对绝对国家安全状态的追求,强调发展要让步于绝对的国家安全,任何国家安全风险都是可以被计算和控制的,但是随着信息技术时代数据的交互融合速度愈来愈快,技术的迭代速度也远超以往,可预见与不可预见风险因素的增加致使绝对国家安全已逐渐成为伪命题。总体国家安全观的“总体”性即以总体观点出发,从绝对安全观转为相对安全观,防范化解国家系统性风险的基础上保障社会公共利益。

AIGC 在发展过程中出现的“算法黑箱”、“涌现效应”等现象突破了以往的技术认知,人工智能领域的不可控因素显著增多,应当承认人类在现阶段对 AIGC 发展认知的有限性,在总体安全观的指引下秉持审慎相对的原则,将技术理性从绝对技术理性到相对技术理性,避免因绝对技术理性而陷入片面追求理性化、确定化的窠臼。

审慎的核心在于避免 AIGC 发展脱离国家安全底线的轨道,从监管运作方式角度看它应随时进行创新与改良。⁹审慎相对的原则就是在总体安全观下,突破静态的治理理念,以动态、灵活的视角对 AIGC 的发展进行调整,以此因应其发展过程中的不确定和复杂性。

三、总体安全观下 AIGC 发展风险的治理策略

在总体安全观下指引下的 AIGC 治理是系统性、总体性的工程,各种风险往往不是孤立出现的,很可能是相互交织并形成一个风险综合体,¹⁰因此治理的路径也要从数据搜集、内容生成、内容输出三个阶段展开,此三个阶段展现的问题在其他阶段也有出现,但是以各自阶段的主要问题为解决方向更能在治理方向上凸显针对性。

⁸ 《中共中央关于进一步全面深化改革 推进中国式现代化的决定》,资料来源:中华人民共和国中央人民政府官网, https://www.gov.cn/zhengce/202407/content_6963770.htm, 访问时间 2024 年 10 月 28 日。

⁹ 刘乃梁. 包容审慎原则的竞争要义——以网约车监管为例 [J]. 法学评论, 2019, 37 (05): 122-132.

¹⁰ 习近平. 论坚持全面深化改革 [M]. 北京: 中央文献出版社, 2018: 183.

（一）以合规管控构建跨境数据流动安全治理国际规则

我国应以“一带一路”倡议为契机，在完善数据合规管控体系的基础上建立跨境数据流动安全治理国际规则，掌握数据流动规则构建的话语权，进而掌握 AIGC 意识形态话语权。

对 AIGC 企业数据库的合规要源头治理的视角来管控。依法设置专门的市场准入规则，基于公平竞争的原则，不管是本土 AIGC 企业的成立运行，还是国外 AIGC 企业主动进入我国市场，都要符合准则标准。市场准入规则应针对 AIGC 企业数据获取方式、数据获取来源、数据获取的边界及合法性进行明确，从企业运行 AIGC 之初就将数据风险纳入管控范畴，明确企业作为数据管控的首位责任。2024 年 8 月 1 日，世界上首部关于人工智能的全面法规《欧洲人工智能法》生效，其基于产品安全将 AIGC 企业分为四个风险级别，并针对不同风险级别采取不同的数据管控标准。政府督促 AIGC 企业建立完善的数据风险评估机制。数据风险评估机制应将数据风险识别和分级分类贯穿数据搜集、算法处理和内容输出的全流程，注重数据风险评估的动态性、灵活性，以此掌握数据管理的主动权。总体安全观是在保证不发生系统性风险的基础上总体把控安全和发展，这就决定了与数据风险评估机制相适应的数据合规是一个持续性的过程，而非静态性、一过性的单一行为。

针对数据搜集阶段的被动收集，数据合规机制要在用户与 AIGC 交互中记录的数据进行动态性审查，避免将涉及用户隐私、敏感性涉政数据纳入语料库。对用户交互的数据建立记录-删除机制，在短期内对交互数据予以记录，方便用户调取，对输出违反国家法律法规、违反意识形态内容的用户予以警告并主动删除相关数据。针对数据搜集阶段的主动抓取，数据风险评估机制的核心应当是对数据的抓取和传输实施动态的监管。一方面对自身爬虫技术抓取的数据进行安全监管，主动建立数据清洗措施。数据的主动抓取带来海量数据的同时也会将偏见的污染数据带入 AIGC 语料库中，大量污染数据将使得 AIGC 的生成呈现涌现效应，在源头上使得输出结果呈现高度的不确定性。企业应通过技术手段对抓取的数据进行去标签化，对数据中带有包括性别歧视、信仰偏见等有违主流意识形态的内容予以清洗，从来源上避免原数据可能附带的意识形态导向。另一方面与国家安全部门相互协作，建立协同防御机制和应急预案，在数据抓取时及时追踪数据传输方向，在数据违规出境等风险事件发生时，为主管部门提供数据非法出境的预警信息，协助拦截数据违规出境。

在构建我国自主 AIGC 数据管控体系的基础上，以“一带一路”倡议为契机，与“一带一路”沿线国家开展人工智能领域的合作。在总体国家安全观指引下，从技术角度上在“一带一路”沿线国家推广我国自主的 AIGC 模型，实现沿线国家数据的跨境流动，并于我国境内完成数据的存储、运算、生成。以本土 AIGC 企业

协会的名义建立全球性人工智能非营利性协会，与沿线国家签订跨境数据流动备忘录，对数据流动方向及使用途径进行倡议性规定，对内实现数据库扩充，对外构建“一带一路”沿线国家的跨境数据监管话语权，进而维护自身数据安全及数据主权。

（二）以敏捷治理模式纠正算法偏见及不透明

人工智能时代的到来使得技术迭代的速度进一步加快，算法模式的颠覆性变化使得传统的治理模式难以快速因应 AIGC 发展带来的漏洞。源自软件工程领域的敏捷治理是学界为应对数字时代发展特点提出的全新治理模式，该模式强调快速响应、多元参与，突破传统治理模式单一主体的窠臼，使得治理更具包容性，敏捷治理模式的全局性、适应性与总体安全观的系统性理念相契合。对于有违主流意识形态的算法偏见问题，应以敏捷治理的模式优化算法模型，秉持全面参与的原则，强调多元化主体协同治理以及治理手段的层次性。

首先，AIGC 企业要恪守“看门人义务”来让渡部分算法解释权于政府、社会，以牺牲部分算法透明度方式纠正潜在的算法偏见。算法解释权使得相对人能够知晓算法决策的依据与过程，弥合平台服务提供者与相对人之间的地位鸿沟，推动形式化的数字服务协议转向为实质上的平等。¹¹我国在《暂行办法》第 19 条对企业算法解释义务的笼统性规定是为了避免过度增加企业义务，但是实践证明算法偏见的潜在危害性已可能对意识形态领域造成影响，因此在针对算法模型的治理上应适当增加企业的算法义务。算法模型作为 AIGC 的核心竞争力，应规定在除政府外第三方只有在有合理依据证明其利益受损的情况下才能要求企业对算法进行解释。政府应进一步细化企业算法透明度，在对除政府的相对方公开时，将算法透明度进行分级分层，根据问题的严重程度进行公开。对于一般情况仅需解释算法逻辑基本情况，在政府监管部门介入时则需对算法的具体决策逻辑、敏感数据处理策略等进行详细解释。

其次，政府主管部门应制定算法审查备案制度对 AIGC 进行预防性监管。根据《互联网信息服务算法推荐和管理法规》24 条的规定，具有公众舆论属性的算法应当在提供服务后 10 日内进行备案。因公众舆论属性的算法潜在意识形态风险对其强化监管属于必要措施，AIGC 兼具公众舆论属性

且在影响范围上远超普通公众舆论属性算法，因此应当规定

AIGC 企业在新型算法上架和对算法进行革新性修改时提前向政府主管部门进行备案。该备案是在不阻碍 AIGC 发展的前提下，以不公开企业商业秘密为前提的纲领性备案，对模型的运行逻辑概括性表述。通过向政府主管部门的概括性备案来避免算法模型运行出现严重事故而导致总体安全观所不能容忍的系统性风险。

¹¹ 张凌寒. 商业自动化决策的算法解释权研究 [J]. 法律科学(西北政法大学学报), 2018, 36 (03): 65-74.

最后，现有对算法偏见的监管体系中主要呈现的是政府-企业二元制的，该种传统的监管模式难以快速响应技术变革带来的新形势，《欧盟人工智能法案》第7章中明确规定要设置欧盟人工智能理事会，理事会需要为市场监管机构之间的交流合作提供平台，并向监管机构和认证机构通报与市场监管有关的问题。我国可适当借鉴欧盟做法，在政府-企业间设置理事会的监督机制，以理事会的专业性衔接政府与企业交流，畅通交流机制，增加双方处理的响应速度。

（三）以软硬法结合引导人机价值对齐

主流意识形态是本国道德和伦理的相结合的产物，AIGC 的生成内容要从伦理和道德层面监管才能避免西方意识形态渗透的潜在风险。从技术设计角度来讲，符合主流意识形态就是让 AIGC 行为和目标同人类的价值观、偏好、意图之间实现协调一致，即人机对齐（human-AI alignment）中价值对齐问题。AIGC 意识形态风险作为人工智能中价值对齐问题的突出表现，在现阶段已突破技术问题的治理面向，从单纯的道德难题上升到法理问题和国家安全问题。因此，要从道德与法律着手，以行业自律与法律规制互补从而形成系统性治理。

AIGC 在海量数据纳入算法模型后所表现出的类创新性能力使得其内容输出的广度和深度出现几何级增长，政府主管部门对生成内容的意识形态监管难度成倍增加。传统法律规范因其稳定性、规范性为 AIGC 内容生成提供确定性指引，但是纯粹的法律规范也因其滞后性难以与技术迭代速度相适应。科技领域软法兴起的主要原因在于，硬法的立法滞后性与经济社会急剧发展导致法律供给不足相矛盾、技术的全球化特征与民族国家因主权而各自立法相矛盾。¹²在 AIGC 治理领域，法律规范的固有缺陷使得政府主导的公共治理应逐渐转向与以“非条约义务”“非正式协议”为代表的软法相结合，以软法的灵活性、针对性、及时性的比较优势弥补法律规范的不足，适时调整 AIGC 发展方向。

1. 以软法构建价值对齐的道德性框架

在软法约束层面，软法在科技领域的治理是以各方主体参与行业自律、建立行业标准的方式实现的。如电气和电子工程师协会（IEEE）作为非盈利性的专业技术学会就制定了 P7003 的算法偏差考虑标准，以行业自律的软法来约束参与方对 AIGC 生成内容进行道德性审查。鉴于 AIGC 的迭代速度与道德观念差异性，难以以法律规范形式对道德行为进行具体性规定，应通过软法的行业自律形式构建价值对齐的道德性框架，以求同存异的方式寻求价值观之间的最大公约数，

我国监管部门印发了《关于加强互联网信息服务算法综合治理的指导意见》强调科技伦理和正能量的导向作用，应在此基础上组织 AIGC 科技企业、第三方监管机构共同签署符合社会主义核心价值观和中国式现代化要求的人工智能道德伦理协议，以此强化企业对生成内容的主流价值方向引导实现价值对齐。

¹² 姜明安. 软法的兴起与软法之治 [J]. 中国法学, 2006, (02): 25-36.

从逻辑运行角度来讲,价值对齐即以有所侧重的方式强化人类所需要的内容,并调整删减人类不需要的内容,最终实现 AIGC 生成内容符合其价值偏好。落实到具体举措上,企业对参与 AIGC 设计、开发、运行以及后期维护的技术人员定期开展技术伦理培训,以自下而上的方式开展价值对齐训练,明确我国 AIGC 企业在进行人类反馈强化学习(RLHF)模式时,对参与生成结果标注的从业人员设置入行门槛,从业人员需经过道德审查和岗前培训后才能参与 RLHF 项目。企业内部由管理层成立专门的道德伦理审查委员会,由管理人员对输出内容的意识形态负主要责任。

2.以硬法构建价值对齐的基本导向

行业自律为代表的软法约束建立的道德性框架是广泛而理想化的,在法律规范体系约束的“硬法”层面,需要倡导构建金字塔式的规范体系。¹³《暂行办法》是由网信部门牵头发展改革委等7部门联合公布的,在处罚依据上仍依靠《网络安全法》、《数据安全法》等法律规范,AIGC的发展方兴未艾,现行对其规定零散分布于各法律规范显然无法有效规制发展中的问题。软法的自我约束亦要与硬法的强制效力相结合才能更好消除潜在意识形态风险,防范化解可能的国家安全风险。我国应当在法律规范体系构建时明确“人类中心主义”的基本导向,在具体法律规范制定上表现为习近平法治思想要求的以人为本价值取向。以习近平法治思想的指导规范体系可以满足构建价值对齐的前瞻性需求,同时对价值对齐的调整更加灵活。在具体法规制定中应当由技术专家和法律专家对 AIGC 相关法规进行伦理性审查,保障价值对齐的基本导向,进而杜绝 AIGC 生成内容的意识形态风险。

参考文献

- [1] 钊晓东. 论生成式人工智能的数据安全风险及回应型治理 [J]. 东方法学, 2023, (05): 106-116.
- [2] 许可. 数据爬取的正当性及其边界 [J]. 中国法学, 2021, (02): 166-188.
- [3] 亚当, 贝克, 龙. 风险社会及其超越: 社会理论的关键议题 [M]. 赵延东, 马缨, 等译. 北京: 北京出版社, 2005:
- [4] 张华平, 李林翰, 李春锦. ChatGPT 中文性能测评与风险应对 [J]. 数据分析与知识发现, 2023, 7 (03): 16-25.
- [5] 郭春镇. 生成式 AI 的融贯性法律治理——以生成式预训练模型(GPT)为例 [J]. 现代法学, 2023, 45 (03): 88-107.
- [6] 王宏伟. 总体国家安全观指导下的国家安全系统性风险及其治理 [J]. 现代国际关系, 2017 (11): 50-57.
- [7] 张真源, 张富利. 人工智能发展的系统性风险治理: 实践反思、理性建构与策略

¹³ 陈小平. 人工智能伦理体系: 基础架构与关键问题 [J]. 智能系统学报, 2019, 14 (04): 605-610.

优化 [J]. 学术论坛, 2024, 47 (01): 133-148.

[8]刘乃梁. 包容审慎原则的竞争要义——以网约车监管为例 [J]. 法学评论, 2019, 37 (05): 122-132.

[9]习近平.论坚持全面深化改革 [M] .北京：中央文献出版社，2018： 183.

[10]张凌寒. 商业自动化决策的算法解释权研究 [J]. 法律科学(西北政法大学学报), 2018, 36 (03): 65-74.

[11]姜明安. 软法的兴起与软法之治 [J]. 中国法学, 2006, (02): 25-36.

[12]陈小平. 人工智能伦理体系:基础架构与关键问题 [J]. 智能系统学报, 2019, 14 (04): 605-610.

2025年中国国际法学会年会

中国—东盟打击跨境电信网络诈骗犯罪执法合作研究

董新春 张波 张倩¹

摘要：2024年12月24日，第79届联合国大会正式通过《联合国打击网络犯罪公约》，该公约是第一个由中国等发展中国家引领推动形成的、在网络领域具有国际法效力的全球性规范性文件，是中国主动参与网络空间国际治理新秩序建构、提升中国网络空间治理国际影响力的重要成果²。

电信网络诈骗是一种远程非接触式的非传统诈骗形式和新型网络犯罪形式，当下，以跨境电信网络诈骗为代表的新型跨国网络犯罪也日益频繁，并呈现新态势和新难点。中国执法机构在打击跨境电信网络诈骗方面，已经同东盟及各成员国开展了多项合作，但是，受到缺乏共同加入的网络犯罪公约作为执法依据、司法协助方面条约不完善等因素的影响，中国与东盟国家共同打击跨境电信网络诈骗犯罪的形势依然比较严峻。

本文通过双方执法合作现状及困境进行深入的分析，并进一步结合新《联合国打击网络犯罪公约》，为完善我国电信网络诈骗犯罪预防体系和完善区域合作提供助力，与东盟乃至全球各国携手打击新型跨国网络犯罪，推动构建网络空间命运共同体迈向高质量发展。

关键词：中国—东盟；跨境电信网络诈骗；国际执法合作；对策

一、中国—东盟打击跨境电信网络诈骗犯罪合作现状

（一）合作国际法基础

国际法的基本依据包括两大部分：一是在国际法领域，涉及引渡、司法互助及警察合作的公约与条约占据着举足轻重的地位。这类国际协议和国际性法律文件，是各国在打击犯罪、促进法律协助方面合作的法律基础，其重要性不言而喻，其中涉及国际法律互助、警务合作等方面的条款是必不可少的³，为跨国执法合作提供正当性基础。跨境电信网络诈骗是一种跨国界的犯罪，要想对其进行治理，就必须要有多个国家共同努力。同时，联合国也是一个国际合作的平台，它所签署的各项国际公约能够为国家之间的合作提供法律依据⁴。

¹ 作者简介：董新春，江苏省徐州市铜山区人民检察院党组副书记，副检察长；张波，江苏师范大学法学院副院长，教授，博士；张倩，江苏省徐州市铜山区人民检察院第一检察部检察官助理。联系方式 18751776570，电子邮箱 1106048452@qq.com。

² 裴炜：《〈联合国打击网络犯罪公约〉：全球网络犯罪治理新篇章》，载《中国社会科学报》2025年2月6日第3067期。

³ 张安乐：《中国—东盟打击跨境电信网络诈骗犯罪的侦查合作机制研究》，硕士学位论文，广西师范大学，2023年。

⁴ 王智：《跨境电信网络诈骗犯罪治理研究》，硕士学位论文，云南财经大学，2023年。

（1）国际公约

中国和东盟各成员国均是《联合国打击跨国有组织犯罪公约》（CTOC）缔约国，这为双方开展包括打击跨境电信网络诈骗犯罪案件在内的警务合作提供了一定法律基础。公约旨在促进各国合作，通过对相关法律条款和制度设计的深入分析，该协定与跨境电信网络诈骗犯罪的打击工作存在密切联系。特别是对“有组织犯罪集团”作出了精确定义，将其认定为拥有三人以上成员、明确的犯罪目的、在一定时期内能够稳定运作，并具备组织架构的罪犯集体。同时详细列举了“跨国犯罪”的具体情境，为后续研究提供了重要的理论依据和分析框架。第十四条对没收及其相关事宜的国际合作做了规定。第十九条规定了在尊重缔约国主权基础之上的联合调查规则，倡议各国通过缔结双边或多边协定来促成联合调查机构的建立，也可以在个案基础上商议联合调查事项，即一案一议。第二十七条规定了执法合作方面的内容，包括交换情报、调查犯罪事项、机构协调、人员交流、执法机构间直接合作的双边或多边协定等，该条第三款也提及“借助现代技术实施的跨国有组织犯罪”。第二十八条建议各缔约国适当制定和应用关于“有组织犯罪”通用的定义、标准和方法，即形成官方法律性共识。第二十九条规定了对执法人员的培训和技术援助，可包含人员借调和交流，培训内容也包括“打击借助于计算机、电信网络或其他形式现代技术所实施的跨国有组织犯罪的方法”。第三十一条规定了跨国有组织犯罪预防规则，如执法机构等有关私人实体之间的合作。

可见，《联合国打击跨国有组织犯罪公约》在处理跨境犯罪问题时，在实体与程序两个维度，对涉及的规范体系作了准确的划分。在实体方面，明确界定了包括“有组织犯罪集团”、“犯罪集团”以及“跨国犯罪”在内的核心概念；在程序方面，讨论了执法合作、侦查工作、提起公诉、追缴非法所得等关键环节中应当遵循的基本法律原则及采取的具体法律措施。⁵为成员国联合打击跨境有组织犯罪提供了法律基础与制度措施。各缔约国可以一定程度上参考适用该公约对跨境电信网络诈骗犯罪进行合作打击和治理。

（2）双边条约

表 1.中国与东盟成员国已签署的双边条约

⁵ 郝家英：《电信网络诈骗犯罪跨境追赃与国际刑事司法合作》，载《北京警察学院学报》2021年第2期，第1-9页。

签约国家	条约名称	生效时间	合作内容及方式
马来西亚	《中国政府和马来西亚政府关于打击跨国犯罪的合作协议》	2012.08.02	内容：包含“电信诈骗”和“网络犯罪”； 方式：拘捕合作
菲律宾	《关于打击跨国犯罪的合作谅解备忘录》	2001.10.30	内容：“洗钱、金融等经济犯罪以及双方同意的其他形式的跨国犯罪”，但并未明文规定电信网络诈骗犯罪/网络犯罪
越南	《中国政府和越南社会主义共和国政府关于打击犯罪和维护社会治安的合作协定》	2001.07.26	内容：包含“组织、运送他人偷越国境、非法出入境、诈骗、洗钱”等犯罪； 方式：交换犯罪情报，交流工作经验
缅甸	《中缅内政部边防合作议定书》	2002.06.17	内容：“缔约双方应根据本国法律和双方共同加入的国际协定，在预防和打击跨国犯罪活动方面加强合作”； 方式：交换情报、（经请求同意的）抓捕、移交、合作调查

签订双边协议虽在一定程度上降低了合作执法磋商的时间与资源开销，但对双边打击网络跨境诈骗活动的影响依旧受限。第一，部分签署条约国家并未将条约以国内法的形式加以规定，因而不具有法律效力⁶。第二，现有双边协定缺乏可操作性。

（二）执法合作机制

自 2004 年起，中国同东盟每两年举行一次打击跨国犯罪部长级会议（10+3），并在反网络犯罪等领域开展了紧密的合作。对东盟国家的执法人员进行了双重多边的执法培训和研究项目，例如对网络犯罪的侦查。双方还在各种会议及对话机制下开展了形式多样的执法领域交流合作。

表 5.中国—东盟执法合作机制

⁶ 马梅娟：《犯罪洼地效应理论浅析——以东南亚部分地区电信诈骗犯罪为例》，载《上海公安学院学报》2020 年第 3 期，第 51-60 页。

合作机制	起始时间
中国—东盟数字部长会议	2006
中国与东盟打击跨国犯罪部长级、高官级、警察首长会议机制（10+1）	2009
中国—东盟警学论坛	2014
中国—东盟网络安全交流培训中心	2019
中国—东盟网络事务对话	2020

其中，第三届“中国—东盟警学论坛”重点就打击跨国电信诈骗犯罪、区域安全治理等议题进行讨论；在 2022 年第 9 届中国—东盟打击跨国犯罪部长级会议中，中方提出“构建更加团结的执法合作关系；建立更有效的反跨境犯罪体系。要加强情报交流，健全工作机制，提高执法效率，在深化反电信网络诈骗、跨国赌博、人口贩运以及绑架非法拘禁等刑事犯罪的联合打击中，中国政府强调构建以打击新型跨国犯罪为导向的专业执法部队，并注重提升其专业技能与执行效率。此外，积极推动与东盟各国警务力量之间的专业培训与交流，设立针对性的短期培训项目及在中国境内的学习机会，旨在加强区域执法合作与人员技术提升，共同应对各类犯罪挑战。并提出组建中国东盟警务人员训练导师库，加强两国公安院校之间的交流与合作”⁷。

（三）执法合作成果

在警务情报信息共享方面，2004 年，中国及东盟各国检察机关首脑会议成功构建了一个信息搜集、检索、交流及传输的网络化平台，为打击跨境有组织犯罪领域内的合作提供了机制支持，并就相关议题达成了初步共识⁸。近期，我国与东盟的部分成员国在跨国犯罪信息交换方面展开了高效的合作关系，并取得了积极的研究成果。特别是，在与越南的执法协作过程中，我们有望实现执法人员之间的直接沟通，并共同开展边境地区的联合巡逻任务，从而使得中越两国警务合

⁷ 董凡超：《王小洪出席第 9 届中国—东盟打击跨国犯罪部长级会议》，载《法治日报》2022 年 9 月 23 日。

⁸ 陈伟强：《中国与东盟国家警务合作之考探——从合作基础、形式、挑战到完善》，载《刑法论丛》2018 年第 4 期，第 549-574 页。

作得到显著提升⁹。2023 年，中缅泰三国警察首脑联席会议隆重召开，旨在协同应对跨国犯罪活动日益严峻的挑战，尤其是在跨境电信网络诈骗犯罪方面。为此，三国决定携手并进，通过集中力量打击地方武装与犯罪集团，共同维护地区安全与稳定¹⁰。双方在执法警务情报领域开展了一系列的合作和交流已经有了一定的成果，接下来应继续推动双方在犯罪情报领域的深化合作。

针对逮捕罪犯的合作行动，中国与东盟中的部分成员国达成了若干协作条款，携手追缉逃犯。比如，中国和越南的国内安全机构缔结了旨在携手追捕罪犯的“联合追逃计划”的谅解备忘录，这是两国警方的深度协作项目。通过双边联合行动和合作机制进行犯罪嫌疑人的缉捕行动，它亦标志着中国及东盟成员国在执法合作领域内的一次重要进展。在近年来的实践中，中国公安机关与东南亚若干国家的警务机构紧密协作，针对针对我国公民进行的电信网络诈骗犯罪活动，成功捕获了众多犯罪嫌疑人。此项合作不仅体现了跨国执法合作的深度与广度，同时也为后续的国际警务合作模式提供了宝贵经验。联合抓捕是双方执法合作的重要内容之一。

同时，联合执法机制和平台也持续发挥作用。就澜湄“次区域”而言，一方面，在中老缅泰四国的湄公河共同监管执行机构自设立至 2024 年 3 月 2 日止，已顺利施行了 138 轮湄公河地区的联合巡逻与执法任务；另一面，自从澜沧江—湄公河区域执法合作中心发起运作以来，该中心通过融合湄公河流域的联合巡逻、成员国间的情报共享、执法人员的全面训练，及组织跨国协同调查等多元化合作模式，在遏制地区内的贩毒、人口走私、偷越国境、网络犯罪等犯罪行为方面展现出了明显的效果。¹¹

二、中国—东盟打击跨境电信网络诈骗犯罪执法合作困境

（一）合作基础薄弱

1. 打击网络犯罪公约的缺失

跨境电信网络诈骗犯罪的打击合作在国际法框架下不可或缺，其中，国际条约为合作机制奠定了基础。然而，目前国际条约在供给方面的不足，成为了制约我国与东盟国家合作机制提升的主要瓶颈之一。跨境电信网络诈骗犯罪的“跨境/国性”，成为了这类犯罪治理的重大挑战，故在从源头上有效遏制此类犯罪行为的过程中，国际合作显得尤为必要。然而，跨境执法活动触及诸如国家主权和国际政治关系等复杂敏感议题，其困难在于众多相互竞争的因素需要平衡。针对此

⁹伍光红：《中国—东盟打击跨境电信网络诈骗犯罪案件警务合作机制研究》，载《河北大学学报（哲学社会科学版）》2022 年第 6 期，第 154-160 页。

¹⁰网易新闻：《中、缅、泰三国联合行动，将对东南亚诈骗集团采取联合行动》，载 <https://www.163.com/dy/article/I0GVICAB05531PN7.html>，2024 年 1 月 15 日最后访问。

¹¹丁晨：《电信网络诈骗案件侦查中的跨境合作问题研究》，硕士学位论文，中国人民公安大学，2019 年。

类问题的理想解决方案在于构建与其宗旨及功能相契合的法律体系。基于各方协商一致形成的、遵循中立公正原则的国际法律文书，能够最大程度地协调各国间存在的利益冲突，排除彼此之间的疑虑，提高两岸或多边执法合作的效率，并对合作过程施加规范。

首先，《联合国打击跨国有组织犯罪公约》虽然为中国与东盟成员国协同应对跨境电信网络诈骗提供了一定的法律支撑，但公约本身亦存在一些局限性。一方面，由于《公约》签署的年代较早，其法律条文并未涉及跨境电信网络诈骗罪，虽然具有类推适用的可能，但在实践中依据该公约进行的警务合作难免缺乏强制力和针对性；另一方面，由于跨境电信网络诈骗犯罪是与时俱进的，犯罪方式也越来越智能化，公约中的一些条款已经不能很好地适用于这种犯罪，从而存在着一定的局限性。在适用条约之前，必须要经过国家之间的充分协商，取得一致意见，但是，在实际操作中，往往存在着有法难依的问题。《联合国打击跨国有组织犯罪公约》在实施过程中由于合作形式、地缘政治、法律差距、执法习惯和水平差异等原因而未能取得应有的合作效果¹²。此外，目前中国与东盟间尚未形成区域层面上亦或是更高级别的具有较强法律约束力的跨国网络犯罪公约，也并未加入现有《布达佩斯公约》。

其次，目前中国与东盟在打击跨境电信网络诈骗犯罪方面的国际合作中，使用频率最高的是刑事司法协助条约，但刑事司法协助条约也还未实现东盟成员国的全覆盖以及协议内容的全覆盖；中国与东盟成员国之间也尚未签署任何关于跨境电信网络诈骗犯罪的双边条约。现有条约并未涉及跨境电信网络诈骗犯罪的罪名认定、跨境电信网络诈骗犯罪管辖权争议的解决方案、跨国调取电子证据的流程等核心问题，这使得双方开展合作时仍需要消耗大量的时间来协调具体事宜。¹³在跨国执法合作行动中也由于缺乏统一的刑事证据方面的规定，尤其是对于跨境电信网络诈骗犯罪十分重要电子证据方面的规定，导致证据难获取、证据难固定以及证据移交所需的时间较长等情况，案件侦办进程缓慢，耽误了最佳侦查办案时间，这一问题在打击跨境电信网络诈骗犯罪的过程中显得尤为关键。此外，还可能存在侵犯主权的风险，根据中国对网络空间主权的保护原则，一国对其境内存储的电子数据享有主权，而单方取证（如远程调查、远程技术调查）无需征得另一方同意即可获得，极易引起外交争端。¹⁴

最后，我国虽与部分东盟成员国签署了关于打击跨境电信网络诈骗犯罪的合作文件，但这些文件均属于共识性法律文件，不属于有法律强制力的双边条约。

¹² 李晓淑：《澜湄合作下打击跨国电信诈骗犯罪的法律问题研究》，硕士学位论文，昆明理工大学，2023年，第22-23页。

¹³ 李子麒：《完善中国与东盟治理跨国网络犯罪合作路径研究》，硕士学位论文，中国人民公安大学，2022年。

¹⁴ 叶媛博：《论多元化跨境电子取证制度的构建》，载《中国人民公安大学学报（社会科学版）》2020年第4期，第51-52页。

双方达成的联合宣言、谅解备忘录和工作方案等共识性文件确定的合作内容，主要集中于人才互派培训、法律文献交流共享、技术互通及能力构建等领域，此类合作在深化双边防范和应对跨境电信网络诈骗犯罪的能力上发挥了积极而关键的作用。但是，在进行执法合作的时候，双方还会通过外交渠道来进行具体的协调，或者根据刑事司法协助的程序来进行合作，但是，在这种情况下，合作时间较长，在实现迅速提取电子证据与快速查封涉案资金过程中常面临较大的挑战。并且谅解备忘录等共识性文件的约束力较小，未明确违约的法律责任，有些文件甚至不具有强制性法律约束力，双方中止、退出合作的成本较低，容易受政治、外交等因素的干扰。

2. 执法合作机制不健全

国际执法合作是指跨越国界的执法机构或者它们与其他国际实体之间依据国际条约、多边或双边协议、内国法或国际法的相关条文，为了满应各方的实际或预期的执法要求，而相互协调政策及举措的合作过程¹⁵。这种执法合作是中国与东盟在共同打击跨境电信网络诈骗问题上非常关键的一环，也是其中一个最为关键的手段。目前双方执法人员在具体的协同作业中，频繁面对出境程序繁杂、语言交流障碍以及风俗、执法惯例差异等难题，这些因素对双边合作的侦查等执法行为形成了不小的阻碍。另外，需特别关注的问题是，东盟各成员国对于中国提取及固定电子证据的法律程序不仅缺乏掌握，而且对于特定案件中所需的关键性证据认识不明，这将导致在缺少充分的预先沟通协调的联合执法行动中，中国执法机构获取的电子证据难以作为案件定性的有力依据。同时，在跨国执法合作领域，多国执法机构及其执法人员的参与，使得在合作过程中围绕执法权划分、执法标准的差异、执法依据及程序等问题的争议无可避免。此外，还可能产生执法行为的不规范乃至严重的违法行为。跨国犯罪的打击难点在于其复杂性及跨域性，故各国之间必须构建坚实的信任体系，并依托此种基础，进而推动执法协作的深入与广泛开展。在此情境下，一套科学完备、合理有效的争端解决机制显得尤为重要，以确保涉及多边的执法、司法协同能够顺畅无阻地进行。综上所述，加强和完善执法合作的综合协调机构，对于提升各国打击跨国犯罪的能力，是势在必行的战略选择。

参考现行体系，澜沧江—湄公河流域综合执法安全合作中心¹⁶（以下简称澜湄中心）尽管管理运营已趋于成熟，且开发了 LMIN（LM-LECC Information Network）系统，有效促进各成员国之间即时交流与配合，极为有利于促进澜湄各成员国之间的法律执行机构的通信互动及联合作战等事务，然而，该平台目前尚未设立解决争议的机制。目前，在澜湄合作框架下，各成员国针对跨境电信网

¹⁵ 赵宇：《国际执法安全合作专题研究》，《中国人民公安大学出版社》2019年版，第281页。

¹⁶ 下设湄公河联合巡逻执法指挥部、情报融合与案件协查部、联合行动协调部、执法能力建设部、综合保障部。

网络诈骗案件的司法执法合作之争议解决机制主要采取“协商”这一路径。通常而言，协商过程需要经历反复的讨论与谈判，最终达成共识，然而在对迅速侦办跨国电信诈骗犯罪行为的高效性要求上，此法并非最佳方案。加之此执法合作机制是由中国、老挝、缅甸与泰国四国共同构建，其公正性在处理包含柬埔寨与越南在内的澜湄六国争议时未必得到充分保证。且该中心属于澜湄“次区域”，且仅对中老缅泰四国有效，不能作为中国与东盟整体或者其他不在流域内东盟成员国的执法合作机制。

因此，从宏观角度来看，当前我国与东盟合作领域中急需构建一个全面的执法合作协调机构，该机构将致力于提高双方执法协同过程中细节对接和信息交流的效能与质量。

（二）合作实践困难

1. 侦查取证困难

由于参与人数众多，电信网络诈骗犯罪活动的不同环节大多分散在不同的区域，由于其高度的隐蔽性，使得执法侦查取证工作面临着巨大的挑战。电信网络诈骗是一种远距离远程控制犯罪，其实施过程中，犯罪分子往往不会与受害者进行面对面的交流，而且受害者也不能准确地描述出犯罪嫌疑人的特征，这就给警方取证造成了很大的难度。此外，由于现代通信技术的运用，犯罪分子可以在极短的时间内将财产进行转移，使得公安部门难以快速、直接地掌握有关的线索。在追查犯罪嫌疑人的账户与手机号码时面临着极大的挑战，尤其针对跨国电信诈骗案件，犯罪分子常常利用境外电信服务进行通讯，开设的银行帐户大多都是在国外以另一个人的身份办理的，并且还通过网银进行了层层加密，这无疑增加了侦查工作的难度。调研人员在执行跨国侦查时，不仅需要对犯罪嫌疑人的银行账户情况和资金流动进行详细审查，而且还需与地方执法机构及司法部门紧密协作，共同实施对犯罪嫌疑人的搜捕行动。

2. 信息共享不对称

情报信息共享机制在加强双边预防与打击跨国网络犯罪能力、及时识别并有效遏制相关犯罪活动以及降低其造成的损失方面发挥着关键作用。然而，当前双方在跨国网络犯罪情报信息共享方面存在一定的障碍和不顺畅性。第一，虽然当前部分东盟成员国会主动向中国通报案件信息¹⁷，然而总体上，大量的跨境电信网络诈骗案件信息的获取仍主要依靠中国公民向中国法律机构报告案情或者是中国执法机构在执行职责过程中的主动发现，相较之下，由东盟成员国主动向中国提供的案情线索数量显得较为有限。第二，在网络安全领域，中国与东盟之间

¹⁷ 如 2017 年新加坡警方接到约 300 起有关假冒年轻女性以提供色情服务为由诱骗钱财的报警，新警方通过对涉案预付卡、充值账户、即时通讯软件账号等信息的分析追踪，发现该系列诈骗案件的犯罪窝点在我国境内，并将有关案情和线索通报我国警方，请求协助开展案件调查。

的信息共享机制存在显著缺陷，尚未构建起一个共享网络安全威胁情报的数据库平台。因而，两方面对网络安全漏洞、网络攻击模式与策略等关键信息的系统梳理不足，限制了双方针对全球性的网络攻击，如跨境电信欺诈等犯罪活动的预防与处置能力。

三、对策的提出——以《联合国打击网络犯罪公约》为背景

（一）《联合国打击网络犯罪公约》的意义

通信网络信息技术的快速进步为网络犯罪提供了便捷的跨国作案条件，致使其触角伸展至全球范围，跨越了传统的国界限制，网络犯罪活动已遍布世界各国。网络犯罪活动，以电信网络欺诈为典型代表，以其迅速更新的犯罪手法和显著的跨国特性而著称。这一现象不但使得违法行为难以侦测与揭露，而且还对于国际社会合力打击网络罪行构成了严峻考验，特别是在跨国执法合作领域。鉴于不同国家的法律体系和社会文化背景各异，各国在网络犯罪打击策略的实施层面呈现出显著的差别。欧盟从 2001 年起就推动谈判达成了网络犯罪公约¹⁸，但该公约并未得到世界上其他国家的广泛参与，许多国家认为应该由联合国推动全球合作打击网络犯罪公约的建立。全球社会因缺乏共同遵循的网络行动守则，在全球范围内对新兴网络犯罪作出有效反应之余，还需推动各成员国对抗日益猖獗的网络欺诈和跨国电信诈骗行为，加大国际执法协作力度。鉴此，2017 年俄罗斯向联合国提出了《联合国协防网络犯罪公约（提案）》的建议。到了 2019 年 11 月，在中国、若干东盟成员及俄罗斯等 47 个国家的共同提议下，《应对利用通信技术进行犯罪活动的决议草案》被联合国大会采纳，正式启动“联合国打击网络犯罪公约”（联合国官方名为《关于打击使用信息和通信技术实施犯罪行为的全面国际公约》¹⁹）的起草进程。为此联大成立了特设委员会（OECE）起草该公约。历时三年，特委会于 2024 年 1 月在纽约举行闭幕会议，对公约草案进行最终修订并将草案提交给联合国大会进行表决。2024 年 12 月 24 日，第 79 届联合国大会正式通过《联合国打击网络犯罪公约》。新公约旨在加强各国之间的协同努力，它将为国际打击网络罪行相关的跨境执法协作、法律互助、引渡程序以及犯罪所得追缴提供一个根本性的架构；并且协调各国的立法与执行策略、预防措施、国际间的信息交流和提供技术支持等方面。

此前，区域性和全球性打击网络犯罪公约的缺失，对我国同东盟合作打击网络犯罪和协同治理的行动产生消极影响，特别是我国执法机关长久以来面对着执法司法协作互助程序繁冗、国际合作不畅等难题。新公约致力于对上述难点领域进行突破，为中国及东盟等成员国的执法机构提供更高效、强有力的国际法律支持及合作框架，将进一步巩固打击新兴跨国网络犯罪的法律基础。待未来公约开

¹⁸ Convention on Cybercrime，又称布达佩斯公约。

¹⁹ See A/AC.291/22/Rev.2.

放签署后，我国亦应据此对本土反网络犯罪立法、司法以及执法实践进行自我审视，并根据公约的规定履行其作为国际条约缔约方所需承担的义务。²⁰

《联合国打击网络犯罪公约》（以下简称《公约》）第 12 条明确规定了“与信息通信技术系统有关的盗窃或欺诈²¹”这一罪名；第 68 条则指出，核心部分为刑事定罪、程序措施和执法以及国际合作，涵盖刑事定罪、管辖权等内容，建立合作双重全球框架，规定预防措施、技术援助和信息交流等，设立缔约国会议。内容具有更高全球代表性、制度包容性和适用普遍性，旨在建立全球治理体系，落实预防惩处机制。

《公约》重点规则设计在于，确立维护主权与尊重人权两大基本原则；建立全面、立体的网络犯罪定罪体系，规定广泛适用范围，明确 11 项罪名及法人责任等；加强国际合作，针对不同电子数据规定取证措施，明确适用范围，规定国际合作机制；全方位加强能力建设，规定预防措施、技术援助和信息交流等。

1. 《公约》中关于执法合作的有关规定

在《公约》中，执法合作主要指各缔约国以加强打击本《公约》所涵盖犯罪的执法行动的有效性为根本宗旨，官方机构之间进行的联系和交流，第 47 条规定了缔约国应在符合本国国内法律制度和行政管理制度的情况下相互密切合作的规则，各缔约国有采取措施以加强执法成效的义务，包括：

（1）网络犯罪情报交流义务和犯罪资料的交换义务

《公约》第 47 条第 1 款中规定的情报信息交流主要包括两个方面：一方面是《公约》所涵盖犯罪的各个方面的情报，比如与信息通信技术系统有关的盗窃或欺诈等犯罪中涉及的犯罪有关的情报信息，包括犯罪嫌疑人情况、被害人情况和作案过程等；另一方面是与其他犯罪活动关联的情报，在《公约》中没有具体规定，但缔约国可根据案件侦办或司法实践需要，也可进行交流的有关犯罪情报。相应措施为建立各国主管机关、机构和部门之间的通信渠道，利用现有渠道，包括国际刑事警察组织（ICPO-INTERPOL）。

交换犯罪具体手段和方法的资料，则包括利用虚假身份、经变造、伪造或者假冒的证件和其他旨在掩饰活动的手段，以及网络犯罪伎俩、技术和程序等。

（2）合作调查义务和提供必要的物品供分析调查之用的义务

在跨境网络犯罪中，往往涉及多个国家，如果仅由一个国家对发生在其他国家的犯罪事实进行调查，不仅涉及管辖权的问题，而且将耗费大量的人力、物力、财力，对打击工作极为不利。因此，有必要请求他国对于相关犯罪事实进行调查，包括犯罪嫌疑人的身份、行踪、活动，犯罪工具、赃款等的去向。《公约》规定，各缔约国在请求相关国家给予帮助时，也帮助他国提出的调查请求，配合分析侦

²⁰ 张鹏、王渊洁：《积极参与联合国打击网络犯罪公约谈判 构建网络空间命运共同体》，载《中国信息安全》2020 年第 9 期，第 68-72 页。

²¹ “Information and communications technology system-related theft or fraud”。

查，提供必要的物品或数据，实现更加有效地打击跨境网络犯罪。

（3）加强部门、机构以及相关人员的协调交流义务

有效打击跨境网络犯罪已成为各国面临的共同难题，对这种复杂的犯罪形式进行深入研究以及所积累的经验做法，对打击成效有着密切的关系，从世界范围来看，各国打击能力水平层次不齐。所以，不论是理论界还是实务界都应当加强同其他国家关于跨境网络犯罪问题的协调交流，以便同其他国家一道共同进行有效打击，贡献自己的一份力量，《公约》规定的具体方式主要有两条：一是缔约国相关机构、部门之间的合作、协调等；二是研究组织和研究人员的交流，包括专家的交流和基于双边条约派驻联络官员。²²

2.开展执法合作应当遵循的原则

根据《公约》对缔约国之间执法合作的原则性规定，中国与东盟开展打击电信网络诈骗犯罪跨国执法合作应当遵循以下两条原则。

（1）国家主权原则

国家主权原则是开展跨国执法合作的基础，中国坚持网络主权原则。在《公约》框架下，执法合作行动本质上仍是国家行使国家主权及司法管辖的一种实践方式，尽管出于保障国家安全和有效应对跨国犯罪等特殊需要，各国可能在证据搜集过程中互相提供便利，进行合作或授权执行相关调查，但这并不代表任何国家有权侵犯其他国家的主权。在践行国家主权原则的过程中，进行各类犯罪问题的国际调查及执法合作，包括针对跨境电信网络诈骗行为的联合侦查，必须获得有关国家的允许与支持。此类合作应坚决尊重其他国家的主权完整、安全保障及社会公共秩序，严格遵守相关国家法律法规，并维护涉事国对司法事务的自主管辖权。

（2）自愿互惠原则

一定程度上讲，自愿互惠原则是国家主权原则的延伸和延续。我国现行《刑事诉讼法》第十七条和第十八条明确了相关规定²³。可以说，在国际执法合作领域，各主权国家的积极参与本质上旨在捍卫自身的合法权益。这种合作模式绝非无的放矢，简而言之，不可能存在一国无端承受所谓义务、向其他国家无偿奉献服务的情形。因此，基于尊重国家主权的原则，国际执法合作宜在各参与方的自愿与互利原则基础上展开。贯彻这一原则可以认为是：一国在执法等领域向另一国提供协助，其本质并非无偿之举，合作双方往往期待获取相应的回报。这种报酬不仅可能体现为道义上的支持与认同，亦可能涉及实质性的物质利益，例如办案补助费用等形式。通常而言，这种回报主要表现为道义层面的互惠，如相互之

²² 在以往执法合作中，往往是由我国驻外大使发挥支持和协调的作用。

²³ 《刑事诉讼法》第十七条 对于外国人犯罪应当追究刑事责任的，适用本法的规定。对于享有外交特权和豁免权的外国人犯罪应当追究刑事责任的，通过外交途径解决。

第十八条 根据中华人民共和国缔结或者参加的国际条约，或者按照互惠原则，我国司法机关和外国司法机关可以相互请求刑事司法协助。

间承诺在同样条件下给予对方对等待遇。实践中,由于电诈窝点所处的一些东盟成员国经济实力有限,办案经费缺乏,在打击跨境电信网络诈骗犯罪中,中方也会提供必要的经费支持,以确保办案顺利进行。²⁴

（二）完善中国—东盟打击跨境电诈执法合作的对策

1. 巩固合作法律基础

一方面,《联合国打击网络犯罪公约》通过后尚未开始签署仪式,在共同需求驱使下,中国积极主动并倡导和推动东盟 10 国于开放签署后均加入《公约》,并在《公约》指导下,增加或细化现有的双边或多边条约,特别是反恐和打击犯罪领域的双边条约、司法协助条约,并对打击电信网络诈骗犯罪和执法合作方面的相关内容进行补充完善。

另一方面,目前我国已经与越南、泰国、菲律宾等部分东盟国家签署了多份关于联合打击电信网络诈骗犯罪及移交犯罪嫌疑人的合作备忘录与宣言。参考新公约,在此基础上,我国应参照同越南签署的合作打击电信诈骗犯罪、开展联合追逃行动等具体领域或行动的谅解备忘录,²⁵为加强打击电信网络诈骗等新型跨国犯罪的效率,迫切建议与东盟各成员国缔结协议,明确围绕执法合作模式、多边追逃机制的构建、以及联合预防与打击电信网络诈骗犯罪的行动框架,作出一致性的谋划与部署。

2. 完善信息共享机制

根据已有经验和《公约》第 47 条第一款相关规定,我国和东盟在完善双方打击跨境电信网络诈骗犯罪的执法合作中的信息交流方面,可以利用好国际刑事警察组织(以下称 ICPO)这一现有渠道。完善国际警务合作快速反应机制,解决信息不对称的问题,实现国际警务合作顺畅、反应迅速。

我国和东盟十国均属于 ICPO 成员,2023 年,在 ICPO 的主导下,开展了“风暴制造者二号行动”(Operation Storm Makers II)动员了亚洲和其他地区 27 个国家的执法部门,打击人口贩运和移民走私,中国和除文莱以外的东盟其他九个成员国都参与其中,我国公安部对该行动给予了财政支持。这是首次专门针对人口贩运引发的欺诈现象开展的行动,关键词包括“网络诈骗 cyber scam”、“全球化 globalization”等。该行动逮捕了 281 人,罪名包括贩卖人口、伪造护照、腐败、电信欺诈 telecommunications fraud 和性剥削等罪行。

当前,东盟在 ICPO 已有针对网络犯罪的专门机构即 ICPO 东盟网络能力服务台 INTERPOL's ASEAN Cyber Capability Desk,且早在 2019 年,ICPO 就在该平台识别到相关网络犯罪威胁后于当年 6 月启动了“金鱼阿尔法行动 Operation

²⁴ 于超:《论依托〈联合国打击跨国组织犯罪公约〉打击电信网络诈骗犯罪》,硕士学位论文,云南大学,2020 年。

²⁵ 陆雪、孙梓翔:《中国与东南亚国家防控跨境电信网络诈骗犯罪警务合作路径构建》,载《北京警察学院学报》2022 年第 4 期,第 114-118 页。

Goldfish Alpha”。ICPO 网络基金会项目予以支持。在行动的五个月中，来自 10 个东盟国家的警察和国家计算机应急响应小组（CERT）的网络犯罪调查人员和专家共同努力完成了打击行动。ICPO 的东盟办事处促进了有关国家之间的信息交流和后续行动。对此，我国可以借鉴东盟办事处的经验，在 ICPO 的支持和协调下，成立我国办事处，或者与东盟成立联合办事处，方便与东盟开展针对电信网络诈骗之类网络犯罪的执法合作。

同时，ICPO 鉴于网络犯罪形势的不断演变，全球执法机构共享信息和知识，以制定及时的、基于情报的响应措施的目的，创建了安全的信息共享平台，以促进各国家/区域/组织的执法部门和其他利益相关者之间对于网络犯罪的沟通——包括：“网络犯罪知识交流工作区 CKE²⁶”和“网络犯罪协作平台—行动 CCP-Operation²⁷”。前者为执法机关、政府部门、国际机构以及网络安全领域的专家提供了一个交流渠道，专门分享针对网络犯罪的非警务性质的操作信息。该工作空间是一个动态的沟通渠道，为了使全球各地的用户能够就最新的网络犯罪发展趋势、防范策略、检测技术以及调查方法与全球授权的专家进行讨论。CKE 还将建立一个国际主题专家网络，以分享该领域的知识和经验。后者支持执法行动，仅限运营利益相关者访问，同类平台中最先成立，是一个集中的信息中心，用于协调全球打击网络犯罪的执法行动。这个受限访问平台托管多个独立的工作区，使运营利益相关者能够在交互式和安全的环境中共享情报。有助于提高成员国的业务效率和效力。

一方面我国和东盟各成员国应当利用好 ICPO 创建的信息共享平台，及时将获取的跨境电信网络诈骗犯罪相关情报信息上传至平台中，定期通过平台开展信息情报的交流和沟通。构建情报共享交换的体系架构，对于电信网络欺诈犯罪涉案的嫌疑分子、非法资金流转、以及有关金融账户的信息，实现情报资源的有效整合与及时传递，为跨境联合侦办、缉捕犯罪嫌疑人提供了必要的情报支持，营造了有利的执法环境。在优化情报信息互通平台之余，依据现实情况，筹建涉及电信网络诈骗活动团体、潜逃人员以及疑涉跨国电信诈骗的嫌犯信息库。此举意在汇聚不同机构的数据资源，对已立案而嫌犯潜逃的情况进行彻底的梳理与概括。同时，设立并持续优化这些嫌疑人的信息库，并且保持数据的实时更新，以保证情报更新的及时性、有效性和准确性。构建与相关部门与机构间的沟通与协调机制，以便对可疑行为进行预警和报告。国家公安机关与境内外边检、关税、边境监管、移民事务等相关单位持续保持紧密的协作联络，共同创立并强化国境出入管控机制，借助大数据技术手段，迅速追踪并掌握可疑分子和受害者的活动路径及有关证据。为提高电信网络诈骗犯罪预警，增强电信网络诈骗犯罪打击的有效性提供保障。

²⁶ Cybercrime Knowledge Exchange.

²⁷ Cybercrime Collaborative Platform - Operations.

另一方面，我国应在已建立的电信网络诈骗犯罪侦查破案协作平台基础上，积极与东盟现有的类似平台资源进行深度整合，共同打造一套高效的跨境电信网络诈骗犯罪信息协作平台，并制订一整套务实可行的合作机制。同时，倡导成立专责机构以负责该平台内信息共享的管理与分析工作，旨在优化数据协同处理流程，为打击电信网络诈骗犯罪活动，提供坚实的情报分析支撑。进一步而言，通过在立法、司法以及执法等环节实施协助与协调机制，有效提升对国际电信网络诈骗的监管能力。促使不同国家在对电信网络诈骗行为的惩处、监管以及法律规范方面达成一致，最大限度内提高国际社会共同打击电信网络诈骗犯罪行为的整体效能。

此外，可以在国内推进网络安全公共服务建设，推动全行业完善涉电信网络诈骗“黑、灰名单”制度，形成统一的“涉诈征信数据库”，并将涉诈征信惩戒名单同步至与东盟及其成员国的信息共享数据库中。

3.成立中国—东盟综合执法合作机构

完善现有与东盟合作框架下、澜湄“次区域”的执法合作平台，并在总结现有平台经验的基础上成立与东盟整体间、更高层面的综合执法合作机构。

一方面，中国与东盟的合作应当更加充分地发挥澜湄中心的执法合作平台功能，将其作为我国警方与澜湄流域东盟五国（柬埔寨、老挝、缅甸、泰国、越南）合作打击电信网络诈骗犯罪的重要窗口。澜湄中心作为一个相对成熟的平台，对于中东盟地区跨境电信网络诈骗犯罪的打击具有重要意义。提升跨境执法协作水平，是当前合作的关键所在。具体来说，应当从情报共享、执法能力提升及协同行动等三个方面入手，拓展澜湄中心的功能和作用。通过加强情报整合机制，建立更加高效的信息交换和共享体系；通过提升执法机构能力，加强人员培训和设备升级；通过联合行动，增强跨境电信网络诈骗的应对措施，共同维护地区经济安全和社会稳定。

在中国、东盟不断磋商的基础上，参考澜湄中心，建立中国—东盟综合执法安全合作中心。并且工作人员应当精通多种外语，以期在实施跨国执法协作时能够担任翻译角色，有效降低因语言差异导致的沟通壁垒，进而优化执法协同效能；精熟各国关于电子证据提取与固定的法律规程及电子证据采纳的具体标准，以在执法合作场景下，指导执法人员依据法律规定，合法提取及固定电子证据。依托于新型中心建设，其职能构建应当在澜湄合作机制现有框架下精进与拓充，譬如澜湄合作中心现阶段所创立的“犯罪信息数据库”板块，在信息聚合及资源深化方面尚显薄弱。因此，新中心一是亟需规划并落实更为全面详尽的信息数据库体系，涵盖各国法律、刑事司法协助以及引渡条约等法律文件的数据库，以便相关工作人员能够依照相应的法律规范，有效地进行工作协调；二是构建最佳执法合作实践经验库，对中国与东盟国家间执法合作的标杆案例进行梳理，在整合归纳之后

收录入数据库，为各成员国提供借鉴与参考；三是构建网络安全威胁信息库，针对当前全球范围内盛行及高发的跨国网络犯罪模式及其手段进行汇总归档，以供各国借鉴学习，进而能够事先做好防范及应对措施的准备。

根据《公约》，为达到即时协助收集、获取、保全和共享电子证据的目的，新中心可以参考 LMIN 系统，设置每周 7 天、每天 24 小时全时段网络运作的联络点，联络点应当具备与另一缔约国的联络点进行快捷通信联络的功能，且隶属于执法或司法机关。在国内法允许的前提下，将 ICPO 纳入联络点网络中，以便开展警方之间快捷合作。

4.完善联合侦查取证机制

《公约》第 48 条规定，缔约国可依据共同加入的公约、双边条约或者在个案基础上进行商定，在尊重主权基础上开展联合侦查。为打击跨境电信网络诈骗犯罪，中国与东盟国家之间建立了以警务联络官为核心的制度安排，通过互派执法人员增强双边协作。这种制度对于破获跨境电信网络诈骗案件发挥了关键作用，已成为双方执法机构合作中的重要组成部分。尤其在电诈窝点所在国警方进行即时联络，参与捣毁犯罪窝点、第一现场取证以及完成境外委托取证等环节表现重大作用。但警务联络官任务繁重而人数十分有限，因此难以直接或间接参与全部案件的侦查取证工作，需要完善其它取证方式加以补缺。委托取证是当下较多采取的取证方式，与之相较，其它方式如境外联合取证在当下取证实践中应用较少，其应有优势并未得到充分发挥，有必要通过加强中国与东盟国家间的执法合作，扩大其适用范围。²⁸

结语

综上所述，跨国网络犯罪给中国与东盟地区带来了严峻挑战，双方迫切需要就此形成广泛共识，在当前合作进程中遭遇的种种困境面前，应积极推进合作机制的完善与效能提升，努力为构建一个安全稳定的网络空间环境作出共同努力。与此同时，针对以跨境电信网络诈骗为代表的跨国网络犯罪的打击和治理，在未来趋势方面体现为全球化的纵深、广泛合作，我国与东盟乃至与其他地区和国家的合作打击和治理未来将会只多不少，我国与东盟多年来合作的经验一定程度上也有助于应对未来跨国网络犯罪的侵袭和为世界上其他国家、组织、地区提供参考。当前情况下，双方必须进一步强化共识，深入学习与研究《联合国打击网络犯罪公约》成果，并持续增进合作的主动性，以促进区域间与国际间在跨境信息网络犯罪治理领域的协同努力，携手构建网络空间命运共同体。

²⁸ 伍光红：《中国—东盟跨境电信网络诈骗犯罪窝点取证协助问题研究》，载《广西社会科学》2022 年第 10 期，第 41-47 页。

人工智能犯罪的刑法回应与犯罪治理

郑丽君¹

摘 要：人工智能与传统犯罪碰撞带来犯罪风险，包括数据安全、技术滥用等行为在刑法规制、责任判断等困难；基于此，刑法规制领域建立与人工智能时代相适应的逻辑理念，确保立法与司法并重、安保与技术支撑并行，前瞻性与现实性平衡，从而多途径弥补现有的规制漏洞。

关键词：人工智能 犯罪形态 刑法规制 治理体系

人工智能技术不断发展，促使网络代际更迭不断加速，进而网络犯罪的表现形式也层出不穷。在传统犯罪网络化进程不断加速之时，与人工智能有关的新型犯罪活动也开始突显。目前，刑法及其衍生出的司法解释在处理网络作为工具和对象的犯罪情形时，已经有了一定的成果。不过，进入大数据时代后，不管是网络犯罪，还是“以网络为空间”的犯罪等领域，在规制方面依旧存在不少漏洞。这些漏洞的出现是因为在应对新型网络安全风险上力量薄弱，这种风险包含大数据系统安全、算法安全以及数据与信息安全等方面的内容。并且，在新的科技革命背景之下，针对人工智能技术被肆意滥用而出现的刑事风险，同样欠缺应对的能力。虽然现行刑法及其司法解释等，对网络作为犯罪工具和对象的犯罪状况已经能有效应对，但在大数据时代，网络犯罪以及以网络为空间的犯罪等，仍然存在不少规制漏洞。其一在于对包含大数据系统安全、算法安全、数据与信息安全等内涵的新型网络安全风险防范不足，其二在于缺乏应对新一轮科技革命背景下人工智能技术被随意滥用而产生刑事风险的能力。因此，刑法在规制网络犯罪时，有必要遵循风险刑法理念，在进行适度立法干预与调整之时，对大数据时代的网络安全与技术风险予以预防，同时针对网络公共安全法益开展诸如立法保护之类的工作，从多个方面构建包含刑法规制的犯罪治理体系。

一、人工智能犯罪的概念

（一）人工智能犯罪的定义

人工智能的研发工作基于对人类能力的模拟与拓展。经过大量的研究，现在人们有这样一个共识：当前处于弱人工智能阶段，强人工智能阶段也快要到来了，不过在理论界，对于是否会出现超人工智能阶段则存在不同意见。从字面意义来

¹ 郑丽君，南京市浦口区人民检察院，本科学历，二级员额检察官，联系方式：18913364913。

看，人工智能犯罪是借助人工智能技术实施犯罪活动的一种新型犯罪形式。

（二）人工智能犯罪的类型

学界根据人工智能在犯罪中的定位或作用，将人工智能犯罪分为四种类型，包括：

第一，主要是以人工智能实体和系统为目标的犯罪，在该类犯罪中人工智能更多以犯罪对象存在，这也是网络犯罪的初始形态。计算机信息系统安全是网络安全的核心，而犯罪分子将计算机信息系统安全为目标，以采取植入、传播病毒软件等为主要作案手段，其外延与计算机犯罪重合。

第二，在当下的司法实践当中，存在一类较为频发的犯罪，在这类犯罪里，人工智能常常被用作犯罪工具。网络的特性如虚拟性、瞬时性、无界性等，让网络犯罪能够轻松突破物理世界在时间与场所等方面的限制。传统犯罪朝着网络化方向发展，并且，网络社交成了网络活动的主要形式，而网络活动的关键在于信息，于是以网络活动形式实施的犯罪也快速增加，网络作为犯罪工具的特性越来越显著。人类意志在人工智能的行为上有所延伸，例如诈骗钱财时犯罪分子对人工智能的声、色、形复合型“深度伪造”技术的利用：2020年浙江杭州办理了涉黄案，案中涉及AI换脸软件的使用，该软件被用于制作淫秽视频并传播，主要存在这些犯罪情况，即把大量女明星的形象通过AI换脸用于传播淫秽物品牟利、未经去世人员亲属允许订制AI复活其动态视频可能涉及侮辱、肖像权等侵犯人身权利犯罪，或者利用ChatGPT等生成式人工智能产品侵犯他人著作权、通过学习违法犯罪手段实施犯罪等。

第三，人工智能产品有问题时可能引发犯罪。在产品的研发、设计构造以及投入使用过程中，相关责任人若没有履行应尽的注意义务，就会造成严重后果，而产品责任人存在主观过失，这属于作业、施工期间因过失导致严重后果的过失类犯罪形态。此外，人工智能在识别方面出错也可能引发犯罪活动。人工智能技术有其自身的不足，其精确性、安全性难以做到毫无差错，而且在灵活性方面也存在一定的欠缺，于是就会出现一些操作上的失误。不法分子通过改变面容、发型等特征的方式，让人工智能的识别出现错误，使安全防护失效，从而给犯罪活动制造了可乘之机。

第四，学术界争议最大的是以人工智能为犯罪主体的犯罪，该种犯罪类型应存在于超人工智能时代，也可称为人工智能自主性犯罪，即人工智能在完全独立、绝对自主的意识下实施的犯罪行为。然而，部分人认为，虽然人工智能的行为看似由自身意识主导，可实际上是体现预先人为设定意志、以网络数据运算为根本的违法、犯罪活动。在网络犯罪方面，大数据与算法的运用至关重要，网络既是这类犯罪所依存的空间与场所，也是其犯罪性质与存在形式的根本所在。²

² 参见庞云霞、张有林：《大数据时代网络犯罪的刑法应对---兼论人工智能犯罪的规制》，载《重庆大学学

二、人工智能犯罪的刑法规制现状

网络 1.0 阶段，刑法设定了非法侵入计算机信息系统罪、破坏计算机信息系统罪，这些罪名是针对人工智能系统的，意图对侵入和破坏计算机信息系统的行为予以规范。与人工智能相关的犯罪和普通犯罪在本质上并无差异，主要根据犯罪手段来归类区分，诸如盗窃罪、抢劫罪、侵占罪、故意毁坏财物罪等罪名能够用于评价相关犯罪行为，就像利用伪装成支付链接的木马程序实施盗窃、诈骗这类犯罪。在此期间，针对借助计算机技术实施的传统犯罪也有了指示性的规定。

网络 2.0 阶段，传统的罪名依然可适用，经对部分内容的扩大解释，刑法能够管制大多数网络犯罪。以网络造谣为例，《最高人民法院、最高人民检察院关于办理利用信息网络实施造谣等刑事案件适用法律若干问题的解释》（法释【2013】21 号）对浏览量、传播量等入罪标准作出了明确规定。《刑法修正案（七）》增加了非法获取计算机信息系统数据等方面的罪名，这让受保护的法益从特殊计算机信息系统领域扩展到所有计算机系统安全以及数据流程安全方面，所打击的危害行为包含侵入、破坏、获取、控制等多种类型，同时也对提供技术性程序、工具的帮助行为作出规制，扩大了对危害网络信息系统安全行为的打击面，加大了对利用网络技术实施犯罪的惩处力度。在司法方面，司法机关已能有效地打击那些靠制作出售特定程序来盈利的犯罪行为（比如靠制作出售‘人工智能外挂’程序盈利这种情况）。比如，江西省余江区法院公开宣判了全国首起类似案件，最终以提供侵入、非法控制计算机信息系统程序、工具罪对犯罪分子判处有期徒刑。

同时，随着新能源汽车的不断普及，各个车企都对智能驾驶功能给予了高度重视。在经过测试后，智能驾驶功能也逐步进入市场，这里的智能驾驶意味着能够达成一定程度的自动驾驶。那么，自动驾驶导致的衍生性行为如何评价呢？新闻报道中多次出现这类情况，有人饮酒后坐在副驾驶位置，让车辆自行行驶，随后被查处。然而，有些司机提出这样的观点，即尽管喝了酒，可车辆的自动驾驶辅助功能已开启，这种情况不应被认定为醉驾。针对这种情况，法官给出了答案。依据我国刑法的规定，危险驾驶罪被归为抽象危险犯，只要出现‘饮酒后+启动车辆+车辆移动’的情形就会被认定为酒后驾驶，所以即便开启了自动驾驶辅助系统，这依然属于醉驾行为。虽然人们针对智能驾驶犯罪展开了讨论，但从立法目的来讲，不管是当下自动驾驶的智能化程度还未实现完全自动驾驶，亦或是在未来更高级别的自动驾驶情形中，法律都明确规定，驾驶员在行驶时都有保持警觉的法律义务，而且要随时准备好接管车辆的控制权。故自然人饮酒后再驾车时，其判断力、反应速度以及协调能力会受到极大损害，这就是‘醉驾’被列入危险驾驶罪的根本原因。所以，对于智能驾驶引发的交通事故伤亡和违法行为等人工智能违

法情形，可以通过剖析立法初衷来实现刑法规制。

网络 3.0 阶段，刑法对网络犯罪行为的规制范围进一步扩大。《刑法修正案（九）》增设了四个‘纯正网络犯罪’，传统犯罪论中，针对网络犯罪里‘技术中立’的帮助行为、预备行为，在危害性评价方面存在无力状况，其刑法规制漏洞与危害程度不符，这些漏洞得以弥补，这样犯罪构成理论在应对网络犯罪新形态时就能够有效转型了。除此之外，在这一时期，网络化传统犯罪在刑法规制方面的漏洞得到了进一步完善。2023 年 9 月 25 日，最高人民法院、最高人民检察院以及公安部携手发布了《关于依法惩治网络暴力违法犯罪的指导意见》（法发〔2023〕14 号，以下简称《意见》）。在中央网信办的支持下，最高人民法院、最高人民检察院、公安部鉴于网络暴力在网络环境里越发猖獗，而刑事方面缺乏有效规范的情况，共同推出了《关于依法惩治网络暴力违法犯罪的指导意见》（法发〔2023〕14 号，以下简称《意见》）。《意见》规定了对包括在网络上进行诽谤、侮辱他人，组织‘人肉搜索’，借网络暴力恶意营销炒作，利用‘深度合成’等生成式人工智能技术发布违法信息等情形要从重处罚，同时明确了刑法规制的途径，也为网络暴力活动中拒不履行信息网络安全管理义务的网络服务提供者确定了刑事责任。该《意见》在网络犯罪相关规定方面有所拓展，一是扩大了网络上诸如诋毁、污蔑等犯罪形式的界定范围，二是扩充了‘公民个人信息受保护权益’这一概念的内涵，同时也对《刑法修正案（七）》《刑法修正案（九）》里的部分网络犯罪行为给出了更为宽泛的解释。

2023 年 11 月，最高人民检察院发布《检察机关打击治理电信网络诈骗及其关联犯罪工作情况（2023 年）》之后，各地检察机关根据相关要求，在打击犯罪工作上加大了力度，这些犯罪涵盖侵犯公民个人信息方面，尤其是非法获取人脸、声纹等敏感信息，还包括利用前沿技术如 AI 伪造人脸、声纹及获取其他信息等情况。同时，全国公安机关借助“净网”专项行动，严厉打击身份证照片等图像信息泄露的犯罪源头，五百多人因借助 AI 技术“换脸”进行诈骗犯罪而被侦破，政法机关依法惩处 AI 犯罪，这对犯罪组织产生了一定的震慑效果。

三、人工智能犯罪治理困境

如今，AI 算法与模型发展瞬息万变，互联网信息服务深度合成技术的应用范畴持续扩大，对人们的学习、工作和生活所产生的影响也在不断加重。2022 年 11 月，国家互联网信息办公室、工业和信息化部、公安部为加强管控联合颁布了《互联网信息服务深度合成管理规定》（以下简称《规定》）。该规定明确指出，在提供与深度合成相关的服务时，必须遵守法律法规，尊重社会公德和伦理道德，任何组织或者个人不得利用深度合成服务制作、复制、发布以及传播那些被法律、行政法规所禁止的信息。这部规定清晰地确定了深度合成服务在定义、适用范围、监督管理规范、监督检查与法律责任等方面的内容。近期，国家互联网信息办公

室公布了深度合成服务算法备案清单，其中包含美团在线智能客服算法、快手短视频生成合成算法、百度文生图内容生成算法、天猫小蜜智能客服算法、腾讯云语音合成算法、讯飞星火认知大模型算法等。然而，生成式 AI 产品的安全与治理风险难以评估，所以目前在中国这类产品还未被批准落地。

像 ChatGPT 这样的生成式人工智能产品对生产和生活产生了强烈的冲击，潜力巨大。其将在工业设计、药物研发、材料科学等诸多领域得到广泛应用，甚至还会改变人类思考与处理问题的模式。即便已经预见到这可能会带来包括侵犯知识产权、侵犯数据权益、传播违法信息等一系列违法犯罪风险，从而给个人权益、社会秩序、公共安全乃至国家安全保护造成极大的隐患，情况也是如此。然而，全球化的科技革命势不可挡，人类在人工智能技术面前将处于难以完全把控的状态，进而面临更多治理方面的难题：

数据源安全隐患增加刑法规制难度

区别于传统犯罪，人工智能技术发展为较成熟的产业链，包括算法、芯片、传感器等多个方面，专业性和复杂性使得人工智能技术分析研判更加艰难。人工智能技术在不断发展中形成了大量的数据和知识，同时危险系数也不断增加，主要表现在算法和数据方面存在的安全缺陷。虽然在我国境内备案的人工智能系统算法和数据都是公开透明的，也对某些违法程序进行了有效的刑法评价，如上文提到的全国首例“AI 外挂”案。该案例中，“AI 外挂”中“CVC”等程序对多款游戏中游戏画面数据进行了未经授权获取，对游戏中处理的鼠标数据指令进行了未授权的修改，增加了“自动瞄准”和“自动开枪”的破坏性程序功能，即：主要通过盒子程序源代码接收计算机 USB 端口传输的鼠标数据指令，并在指令下对计算机解析解析，再将解析结果发送至计算机 USB 端口，实现计算机鼠标指针自动移动和点击。可见，违法程序运行管理中既有未经授权的数据获取、修改，也有数据增加，通过这些未经授权的、不透明的非法操作作为规制对象，可实现有效规制目的。

但生成式人工智能产品如 ChatGPT，由于自身具有高度专业性与复杂性，在规制方面存在较大困难。生成式人工智能建立在“大数据+大算力+强算法=大模型”的技术路径之上，即使是专业人员也难以解释内在机理；³如 OpenAI 一类公司仍并未公开所使用的数据来源，相关训练数据库是否均获得授权还存在疑问，故该产品仍属于“算法黑箱”。此外，该公司的使用条款表明其对任何用户输入和输出内容具备广泛的使用权来优化 ChatGPT，所以，要是用户毫无防范地在 ChatGPT 中输入已识别或者可识别自然人相关的各类信息，就存在侵犯他人信息权益的风险。然而，人工智能的优化与发展是以持续不断且自由流动的信息数据

³ 参见苏宇：《算法规制的谱系》，载《中国法学》2020 年第 3 期。

为支撑的，可基于安全考量的社会秩序却给数据的流动和汇集施加了不少限制。依据我国刑法相关规定，人工智能若通过非法途径获取信息数据，又或者非法将用户信息数据泄露出去以及存在其他危害信息数据安全的行为，那么就可能触犯非法侵入计算机信息系统罪、非法获取计算机信息系统数据罪、侵犯公民个人信息罪等犯罪。⁴因此，人工智能在信息数据的不断拓展与信息数据保护之间的鸿沟是无法被填平的。

（二）生成主体复杂导致犯罪不可控

在人工智能的分工体系里，并非只有服务提供者和用户这两类角色，数据供应、模型开发与管控等上、中、下技术链以及产业链方面的分工也被包含其中。数据供给环节包括数据采集、标记和预处理等任务。中间环节有模型开发与定制这一内容，其包含算法、深度学习模型训练、二次开发定制模型、分发应用等事务。这些任务也许会被同一个实体负责，也许会被不同的实体负责，但都对人工智能的学习过程和最终成果产生作用。人工智能犯罪可能是多个阶段积累的结果，也可能是单个阶段单独造成的，这是多个主体与环节产生的责任。然而，当多个主体发挥作用时，倘若每个主体仅清楚自身参与的阶段，并且与其他主体没有交流沟通，那么各环节或者阶段的实施主体就难以承担预测和控制的责任，从而被要求对整个过程产生的影响负责。⁵这为刑法中因果关系和责任的判断带来困难，从而无法找准犯罪主体。

此外，人工智能归根结底是技术，那么在输出内容时只是机械执行指令要求，导致的后果就是内容不可控。如利用 AI 相关技术进行“一键脱衣”的操作，并借此传播淫秽物品牟利这类行为，就可能出现侮辱诽谤他人、传播违法信息等犯罪风险。根本而言，一方面，用户仅仅是被动地处于犯罪风险之下，并没有实施导致风险产生或者使风险恶化的行为，所以不应当被判定为刑事责任的承担者。另一个方面，就像之前提到的，开发者和服务器受到技术不确定性的约束，无法预知犯罪风险是否会出现以及何时出现，那么刑事责任的承担者又该如何判定呢？

（三）输出内容复合型引发犯罪风险升级

人工智能的建模与优化需要大量数据的支持，其获取数据的方式包括通过非数字作品数字化手段，或者利用互联网等载体来抓取已有作品。当下，人工智能尚处于专用人工智能这一阶段的应用之中，此阶段的人工智能缺乏对信息进行深度加工与组合的思考能力，因而并不满足著作权法所规定的独创性标准。ChatGPT 问世之后，社会各界对它给出了像是具有创造性、使用方便等较高的评

⁴ 参见盛浩：《生成式人工智能的犯罪风险及刑法规制》，载《西南政法大学学报》2023 年 8 月第 25 卷第 4 期。

⁵ 参见曾粤兴、高正旭：《论人工智能技术的刑法归责路径》，载《法治研究》2022 年第 3 期，第 117 页。

价。不过，仔细剖析的话就会明白，ChatGPT 本质上属于“智能搜索引擎+智能文本分析器+智能洗稿器”的组合类型。在刑法范畴内，人工智能若在获取数据时未经著作权人许可便生成内容，这种行为也许会被视作《刑法》第 217 条侵犯著作权罪里的“复制”行为，要是通过特定方式向公众提供这个被“复制”的作品，可能就会被认定为“通过信息网络向公众传播”，从而侵犯著作权。

（四）技术滥用等多元风险类型致规制缺漏

在人工智能犯罪风险的刑法规制方面，首先存在不足的是与伪造相关的人工智能技术运用领域。人工智能技术能通过组合个人声音、面部表情以及身体动作等元素来制造虚假内容，这属于深度伪造技术的范畴。这种行为有着极大的社会危害，轻则损害被害人诸如名誉权、财产权等个人权利，重则破坏社会秩序，乃至危及国家安全。针对此类情况，我国刑法从规制侵害公民个人信息（前端）和规制后端产生的法益侵害结果这两方面，利用传统罪名实施了规制举措，不过现行刑法的评价体系依旧存在欠缺之处：一方面，对于处于中间阶段的相关行为自身无法作出评价，从而忽视了该行为本身的社会危害性。另一方面，前端与后端的规制还存在着缝隙，当行为不满足传统犯罪的构成条件时，就难以进行刑法规制。

其次，缺乏滥用技术的规定。当前的刑法只是针对滥用人工智能导致特定危害的情况规定了刑事责任，可是，仅仅是滥用人工智能这一行为应怎样规制却还不明确。美国哲学家兰登·温纳（Landon Winner）曾提出这样一个问题：技术是否为政治问题呢？文章提出了“技术中立”理论。技术本身是没有偏向性的，它所产生的效应在本质上是由技术的研发者、运用者决定如何去选取和运用技术的。不过，要是刑法仅仅着眼于技术被滥用后的侵害结果，而对技术滥用行为本身不闻不问的话，那么，不但法益保护无法实现即时性，而且也难以对技术滥用问题作出充分评价。

如前所言，在探讨通过 AI 技术定制已故人员动态视频的行为时，其初衷往往是为了抚慰亲属的情感需求。在获得亲属明确授权的前提下，此类技术行为通常不会引发侵权问题。然而，若视频制作未经逝者亲属同意，则可能涉及侵犯肖像权、名誉权等人身权利，甚至可能构成违法犯罪。目前，我国《民法典》《刑法》等法律法规对肖像权、名誉权等人格权提供了较为全面的保护，但对于 AI 深度伪造技术的具体应用及其侵权行为的规制，尚存在一定的法律空白。此外，深度伪造技术的滥用还可能引发更为严重的社会问题。如以营利为目的的商业化滥用行为，将给社会秩序、公共利益产生何种不利后果？在造成何种危害后果下，启动刑法规制？

所以，如何从源头上展开有效治理，从而避免后续行为本身以及危害结果出现之后所衍生出的各种违法犯罪现象？

四. 人工智能犯罪的治理途径探索

以 DeepSeek 为代表的中国本土开发的 AI 模型与国外研发的 ChatGPT 相比,虽然在中文语言处理能力、数据安全与合规性、行业定制化能力以及本地化服务支持等方面,能够更好地满足中国用户的需求。而 ChatGPT 则在国际化、通用性和技术成熟度上具有优势。但在刑法规制中,二者面临同样的风险应对需求,需不断融合完善:

1. 加强法律的立改废释纂, 弥补刑法规制漏洞, 以良法促善治

人工智能一定程度对人类行为进行了替代,改变了刑事责任分配体系,促使现有刑法条文对新型犯罪类型不足以完全规制。比如,以智能驾驶系统为主体引发的新型智能驾驶公共安全犯罪,像之前提及的酒后智驾行为能够凭借传统罪名予以规制,可是对于这类智能所引发的其他危害公共安全的犯罪来说,目前法律处于大部分空白状态,司法机关基本上无章可循。⁶2023 年,我国虽然起草了《生成式人工智能服务管理办法(征求意见稿)》,但这个管理办法主要是针对 ChatGPT 等生成式人工智能的管理作出规定的,重点在民事方面,对于刑事归责只是简单提到了“承担刑事责任”。在实践当中,可以参考新加坡的治理思路,探究人工智能犯罪的发展趋势,从流程管理方面着手,由下游犯罪追溯源头。正如前面所提及的,司法机关可利用刑法解释学原理,对现行规范作出适当的扩充解释,借助司法解释的方式来填补刑法法规的缺失,依据人工智能犯罪行为及其产生的后果,有效地解决归责方面的难题。可以尝试将运用“深度伪造”技术进行侮辱诽谤,且给被害人名誉带来严重侵害后果的行为,以单行规定或者指导意见的形式,归入侮辱罪、诽谤罪等入罪情形之中,进而达成刑法规制。

类似地,人工智能被用作犯罪的工具与手段进而引发犯罪结果这种情况,相关部门可根据行为人的主观动机进行适当的责任划分:若行为人故意借助人工智能实施犯罪行为并造危害了后果,应根据其行为性质认定为相应的传统犯罪类型;若人工智能服务的提供者明知他人利用人工智能实施犯罪,仍为其提供技术支持或服务,则可能构成帮助犯或者符合帮助信息网络犯罪活动罪的情形。然而,如何证明服务提供者“明知”存在一定难度。根据《中华人民共和国刑法》第 286 条之一规定,网络服务提供者若未履行法律、行政法规规定的信息网络安全管理义务,且在监管部门责令采取改正措施而拒不改正,则可追究其刑事责任。对于其他罪名中“明知”的认定,则需通过司法解释或者相关规范性文件进一步明确。此外,在人工智能的设计阶段,若在数据预处理、模型优化等环节通过非法手段获取他人数据或作品,并通过人工智能平台进行传播,尽管输出内容可能具有复合型特征,但通过追溯源头,仍可精准定位责任主体。具体而言,责任主体可能包

⁶ 参见孙道萃:《人工智能驾驶的刑法解构与回应路径》,载《西安交通大学学报(社会科学版)》2022 年第 2 期。

括数据采集者、算法设计者以及平台运营者等。

与此同时，吸纳英美法系从设计制造方面强化初始风险管控的理念。Web 前端在工作时，会有数据处理方面的事务，这其中包含编码相关工作，并且会与后端接口产生交互。产品经理给出页面交互需求后，前端开发人员依据该需求进行页面编码，接着利用后端接口获取数据，最终将数据展示在前端页面上。大多数过程都是理性地处理和显示数据。这是人工智能犯罪的风险来源之一。在现行法律规定当中，刑法在数据犯罪管控方面，主要侧重于非法获取与非法破坏相关的犯罪，如非法获取计算机信息系统数据罪和破坏计算机信息系统罪，重点在于强化对静态数据安全的管控限制。可是，智能时代逐步发展的进程中，数据有了动态化、共享化发展的走向。智能时代中数据的动态化、共享化发展趋势下，数据安全保障迫切需要从静态安全向动态管控安全过渡。对于合法获取却非法利用的情形，例如非法分析数据、非法使用数据分析结果，还有数据处理者未得到权利人许可或者违反相关规定，将合法收集、获取的数据用作人工智能预训练数据、优化训练数据以及用于其他算法分析渠道等行为，一旦造成严重后果，刑法应对其予以规制。

除此之外，人工智能通过特定的算法模拟和扩展人类的智能行为，这些算法为人工智能系统提供了解决特定问题或执行特定任务的明确指令和步骤。作为人工智能的核心组成部分，算法决定了计算机如何处理数据以及如何做出决策。因此，对算法的刑法规制成为解决人工智能犯罪问题的关键。算法的刑法规制可从两个维度开展：其一，针对借助算法实施犯罪之举，要是有人运用算法构建诸如实施电信网络诈骗、传播犯罪手段、非法获取数据等违法犯罪活动的程序或者网站，能够设立相应罪名加以规制；此外，从辅助手段的角度来看，当清楚他人利用信息网络实施犯罪却仍然为其提供算法服务时，可以增加入罪条款，从而实现人工智能被不当使用状况的刑法规制。

2.通过明确人工智能服务提供者的地位和监管义务，确定入罪门槛

虽然人工智能领域具有不可控与不可预知特性，在制定法律时，只有当其他替代刑罚的方法无法有效达到法律目的，以及法律呈现出的一种滞后于社会发展的状态，让治理的预防性显得尤为重要。所以，针对人工智能犯罪治理最恰当的做法是提前设置责任主体并明确安全管理义务。由此而论，法律（涵盖刑法）有必要参与到人工智能产品从研发、生产到投入使用的全流程当中，从而实现全方位的管控与监督。结合人工智能发展现状，刑法对服务者可明确以下义务：第一，数据安全把关义务。各环节服务提供者应对其获得的数据来源合法性、内容真实、合法性以及数据处理的安全等内容负责；第二，算法满足规范义务。因为技术有复杂性、不透明性并且缺乏可解释性，用户或者研究者不能清楚地知道算法内部的工作原理与决策过程，所以服务提供者在进行算法设计、模型生成与优化以及

提供服务的时候，要尽量避免违法犯罪风险；第三，内容管理注意义务。在内容生成、命令执行过程中，服务者必须采取最大程度能力范畴内的技术手段防止不当内容的生成，包括涵盖他人个人信息、知识产权、人身权利等争议内容的，并以标识加以备注；第四，终止提供服务的监管义务。智能平台在提供服务过程中，如果发现用户存在违反法律法规等行为，就应当立刻终止服务；第五，服务提供者的安全注意义务。明确服务提供者在研发、编制、设计程序和产品过程中，服务提供者的监管义务应通过立法予以明确。从前置法层面，应细致规定人工智能服务提供方的责任主体、种类以及具体事务，进而为刑法规制提供依据。就刑法规范而言，可以参考《刑法》第 286 条之一规定的“拒不履行信息网络安全管理义务罪”，为人工智能服务者增设相关的罪名条款。

在部分人眼中，人工智能虽然表面上呈现出类似“自主意识”与“自主行动”的现象，从而使得人工智能能否成为法律责任主体这一问题在理论界激起热烈探讨，但本质上，人工智能并无自然生命的属性，其“思考”源于预先设定的程序和指令，其“行动”也是依据数理逻辑进行的。所以，即使人工智能产品在运行期间依靠“智能表现”产生危害社会秩序的严重结果，也要清楚人工智能在人类设定和编写程序的范围内所展现的仍旧是人类的思想，这样在追究责任主体的时候，还是得搞清楚是产品的研发者、制造者或者使用者的责任。拿如今备受瞩目的自动驾驶汽车来说，它是依据预先设定好的程序开展交通运输工作的。如果生产者在算法设计过程中有过失，从而出现违法犯罪的风险，那么可以根据像‘重大责任事故罪’之类的传统犯罪构成要件，确定生产者的前置性义务内容来加以规制。当然，也要注意避免入罪范围的扩大，在人工智能的制造、控制环节，如果不能确定服务供应者存在罪过，即存在技术上无法避免人工智能安全风险这种情况时，这属于意外事件，无需承担责任。

3.加强企业合规监管，以“技治”反哺人工智能生态

在人工智能相关事务的处理上，意大利在企业监管方面展现出了很强的前瞻性与积极性，在应对人工智能犯罪方面表现得尤为明显。我国可借鉴意大利的经验。在立法上，能够颁布新的法律，要求企业对所使用的人工智能系统进行全面审查，以此保证系统决策过程可被解释，消除其“不可知性”，实现公平、透明、可靠的要求；同时，可以构建专门的机构，用以增强对企业使用人工智能时可能存在的不当行为的监管力量，以保护公民个人数据。公民个人数据保护机构尝试从监督与合规方面入手，作为监管者加强对企业的监管与管理，推动企业建立完善的合规体系与专门队伍，帮助企业评估并减少人工智能相关风险；此外，企业的管理人员需要提高推进项目的积极程度，增加人工智能技术教育和培训的推进以及项目实施的力度。企业的管理层应当积极起到带头的作用，由上至下地踊跃参与进来，积极主动地去推动人工智能方面的教育和培训工作的，如此一来能够提

高企业管理层与员工对人工智能潜藏风险的认识，从而加强企业内部的管理控制，确保其被合理地使用。⁷

在安全评估工作方面，可以尝试树立一种借助技术手段管控相关技术的理念，比如运用加密技术、防火墙技术等，周期性地开展此项工作。同时，借助人工智能相关技术提升组织安全防护的能力，构建起针对人工智能技术的网络安全威胁监测与应对体系，从而预防人工智能可能带来的网络安全风险。安全监管部门可致力于开发一种系统，用来监控人工智能的发展进程，提升人工智能在诈骗相关的监测、阻断和警示方面的能力。以国家公安部推广的国家反诈 APP 为例，其在应对电信网络诈骗时，有着监测、阻断和提醒类似人工智能诈骗行为的能力。尝试让像人工智能这样的技术不再被纳入法律规制的对象范围，从而转变为能够支持、辅助法律规制的治理手段。

4. 内外兼顾，积极参与人工智能安全治理全方位部署合作

人工智能事务具有国际性，其中人工智能技术面临不少安全风险。在人工智能安全治理这一国际性事务当中，鉴于人工智能技术存在不少安全风险，我国需要积极参与到国际合作当中去。例如，可效仿德国、美国等国家的标准与技术研究院的做法，一同推动人工智能安全标准的制定工作，并且还能够探寻与其他国家构建有关人工智能技术发展与安全合作的国际监管机制。加强对人工智能应用的监管，不限于技术领域。人工智能基础模型的开发者有必要对模型卡予以定义，并且要公开模型功能、能力限制等相关信息，要是发现存在不当行为且情况查明后，相关主管部门就得对其实施制裁。在 2023 年的首届全球人工智能安全峰会上，中国、美国等 28 个国家共同签署发表了《布莱切利宣言》，其目的在于通过国家间的合作构建针对人工智能的监管方法，希望通过法规之类的手段来避免相关风险的发生。这是世界上第一份反对人工智能的国际声明。另外，基于国内现有的网络安全相关政策，可以进行如下的尝试：第一，优化开发模式，遵循安全设计的要求，尽力在达成业务盈利和保证消费者与数字生态安全之间实现平衡；其二，企业应提高自身基础设施的网络安全水平，这样就能有效应对网络攻击，防止用户数据被盗取或者丢失等情况的发生；其三，于上游处强化管控工作，对信息内容管理进行加强，阻止虚假广告的出现，进而保护用户不被虚假、诈骗类信息所侵害等。⁸

5. 政府主导，企业参与，建立多方协作防范机制

首先，加强公众教育与宣传。政府部门可通过社区讲座、学校课程、媒体报道以及社交媒体、短视频平台等形式，向公众普及 AI 犯罪的基本概念、常见类

⁷ 参见刘星臣：《意大利人工智能犯罪现状及治理》，载《MODERN WORLD POLICE》。Dec.2023，66-71。

⁸ 参见王彩玉：《新加坡人工智能犯罪生态治理的实践与探索》，载《MODERN WORLD POLICE》。Jan.2024,53-61。

型及其 AI 违法犯罪类型。制作通俗易懂的宣传材料（如海报、短视频、漫画等）以及 AI 犯罪警示案例库，帮助公众快速理解 AI 犯罪的特点和防范方法，通过典型的 AI 犯罪案例，分析犯罪手段、危害后果及防范措施，帮助公众提高警惕。并邀请专家或 KOL 进行科普直播或在线答疑，增强公众的参与感和信任感；其次，提升公众的技术素养。在基础教育中引入 AI 基本知识课程，帮助学生了解 AI 技术的原理、应用场景及潜在风险。面向成年人开设 AI 技术普及培训班，提升公众对 AI 技术的认知水平以及培养数据安全意识。教育公众如何保护个人隐私和数据安全，例如不随意泄露个人信息、谨慎使用 AI 应用程序等。向公众介绍深度伪造技术的特点和识别方法，例如通过细节分析（如面部表情、声音异常等）判断内容的真实性；同时，AI 技术开发企业应承担社会责任，通过官方网站、用户手册等渠道向用户普及 AI 技术的安全使用知识。社会机构建立 AI 犯罪防范志愿者网络，培训志愿者成为 AI 犯罪防范的宣传员和技术支持者。最后，利用技术手段提升防范能力。开发并推广使用 AI 反诈骗工具、深度伪造检测工具等，帮助公众识别和防范 AI 犯罪。在 AI 应用程序中嵌入安全提示功能，提醒用户注意潜在风险。建立 AI 犯罪监测与预警系统。向公众提供实时风险提示，帮助其避免落入 AI 犯罪陷阱。

迅猛发展的智能技术在人们的生产、生活中得到广泛应用，它也对当代刑法体系产生了渗透与冲击，在理论界引发了一系列动荡。针对新型智能犯罪的问题，其载体或产品呈现出的属性以及法益形态，要求人们在应对和规制的前瞻性方面，联系智能时代的特征与类型的演变，选取有效的治理方式，从而实现人工智能刑事风险的有效管控。

Criminal Law Response and Crime Governance of Artificial Intelligence Crimes

Abstract: The collision between artificial intelligence and traditional crime has prompted us to address the criminal risks brought about by artificial intelligence, including difficulties in criminal law regulation and liability judgment for behaviors such as data security and technology abuse; In this regard, the criminal law should establish a governance logic and concept that is suitable for the era of intelligence, adhere to the balance between legislation and interpretation, balance security and technological progress, balance foresight and reality, and fill the loopholes in existing behavioral regulations through multiple channels.

Keywords: artificial intelligence; crime forms; criminal law regulation; governance system

论“塔林手册 3.0”更迭的走向与博弈*

——兼论网络武装冲突规则的策略对抗

焦园博*

摘要：当前《塔林手册》已经着手编纂制定3.0版本，作为网络空间国际公法领域颇具影响力的非正式文本，“塔林手册3.0”仍在追求提升手册影响力。即使《塔林手册》表明了谦抑性并声称自身没有规范意图，但《塔林手册》在客观上其发挥的作用并不谦抑。在“塔林手册3.0”进一步寻求影响力的道路上，其延续并保持了自身在网络空间普遍适用国际法的基本主张，但在方法上由原先的专家背书为主转向寻求政府背书。《塔林手册》及其更迭背后的观点交锋，实质是“对抗策略”和“竞争策略”两种策略的博弈，其间博弈的关键问题便是“对抗策略”寻求降低“使用武力”“武力攻击”“自卫权”等概念的门槛，以期为进攻性网络行动披上合法性外衣。“竞争策略”作为应对策略，需要在尊重国际法规则和提升国家主体性的意义上予以应对，进而在论辩实力和能力建设议题中推进国际法在网络空间的合理适用。

关键词：《塔林手册》 “塔林手册3.0” 非正式文本 对抗策略 竞争策略

一、《塔林手册》更迭及其挑战

《塔林手册》是北约影响下的规则文本，作为寻求规则话语权的一部分，《塔林手册》虽然不是有约束力的规则文本，但其影响力已不逊于官方文本。如今《塔林手册》向 3.0 版本的更迭意味着该文本在规则话语权方面有更大抱负，为应对其带来的话语挑战，我们需要对《塔林手册》的更迭过程进行全景考察。

（一）《塔林手册》更迭历程：从1.0到3.0

《塔林手册》最值得关注的方面之一，就是其对于自身在国际社会接受度的紧迫感。作为北约合作网络防御卓越中心（NATO Cooperation Cyber Defense Centre of Excellence，以下简称 CCDCOE）的工作成果，《塔林手册》在文本质量上的“卓越”追求表现的十分明显，无论是邀请多国专家组成的专家组参与编纂，

* 本文所用网络资源的最后访问时间均为 2024 年 12 月 20 日。

作者简介：焦园博，国防科技大学军政基础教育学院副教授，法学博士，主要研究方向：网络空间国际法。本文系国家社会科学基金项目“外空活动中网络攻击行为的国际法规制研究”（2022-SKJJ-C-079）的阶段成果。

还是通过配套议程推介《塔林手册》，¹都从侧面展现出《塔林手册》对于国际社会接受度的极高期待。这一期待最直接的具象化，便是《塔林手册》在十年之间进行了两次更迭。

《塔林手册》从 1.0 到 3.0 版本，每次版本更迭都有大篇幅的内容变化。《塔林手册 1.0》于 2013 年面世，其基本内容以网络武装冲突法为主，将传统领域的武装冲突法规定适用于网络空间。《塔林手册 2.0》于 2017 年面世，其不再局限于武装冲突法的适用，而是将讨论范畴扩展到非战时国际公法领域，更新内容既包括主权、管辖、国家责任等国际公法基础理论，还包括网络空间与具体国际法领域之间的交叉，这些具体领域涉及国际人权法、外交与领事关系法、海洋法、外层空间法等。截至到 2.0 版本，《塔林手册》已经讨论了网络空间几乎可能涉及的所有国际公法规则。“塔林手册 3.0”目前还未面世，根据编纂计划，3.0 版本可能在 2025-2026 年之间面世，根据已公开的信息来看，“塔林手册 3.0”将着重搜集并整合各国官方表态，并重点关注个别有争议的问题，如自卫权、网络间谍等问题。

在两次版本更迭过程中，《塔林手册》总主编均为 Michael Schmitt，这使得《塔林手册》历代版本在调性上保持一致。Michael Schmitt 指导 2009-2013 年《塔林手册 1.0》和 2013-2017 年《塔林手册 2.0》的工作，并担任两个版本的总编辑，目前他已经在担任《塔林手册 3.0》项目的主任。²作为项目主导者，Michael Schmitt 的想法对《塔林手册》的塑造不可谓不重要，其个人视野和观点将直接决定《塔林手册 3.0》的重点修订方向。其中相当一些观点已经在公开途径表达了出来，在后续的行文中将会述及。

关于《塔林手册》是否会在更迭过程中寻求改变其非正式文本的性质，根据塔林手册项目的官方表态，《塔林手册》的性质将保持不变。根据其表述，《塔林手册》将继续是杰出的国际法学者和从业人员撰写的不具法律约束力的学术著作，旨在客观地重述在网络背景下适用的国际法。它是政策和政治中立的，不代表任何国家或国际组织（包括 CCDCOE）的法律立场或学说。与前几版《塔林手册》一样，该项目的领导层致力于客观性，特别是包括关于在网络背景下解释和适用国际法的所有合理观点。³然而从《塔林手册》两次更迭所展现出的追求来看，手册绝不会局限于仅仅作为非正式文本而存在。

¹ CCDCOE 每年都举行 Cycon 会议，自 2010 年开始，每年来自近 50 个国家的政府、军队、学术界和工业界的 600 多名决策者、意见领袖、法律和技术专家齐聚 CyCon，以跨学科的方式应对当前的网络安全挑战。在为期两天半的时间里，专家们通过主题演讲、全体会议、专题会议和分组讨论来探讨会议的关键主题。

² CCDCOE. Tallinn Manual Introduction. <https://www.ccdcoe.org/research/tallinn-manual>, 访问时间 2024 年 7 月 30 日。

³ CCDCOE. Tallinn Manual Introduction. <https://www.ccdcoe.org/research/tallinn-manual>, 访问时间 2024 年 7 月 30 日。

（二）《塔林手册》属性变化：趋向正式的非正式文本

自进入本世纪以来，国际公法规则制定鲜有较大成就，这一状态使得政府在制定国际法规则方面的权威不再被期待。与此同时，一些非官方文本开始展现出作用，其中不少以“手册”命名，⁴其中就包括《塔林手册》。这一系列非正式手册在形式和风格上表现出一定程度的相似性，即以“黑字规则（black letter rules）”为主体，并附有解释规则相关的评注。《塔林手册》采用的正是这一模式，其优势之处在于摆脱了对当前国际法常规编纂模式对政府参与的依赖。《塔林手册》能够更迭至 3.0 版本，与非正式文本的兴起关系密切，再辅之以政府间组织（CCDCOE）的背景支撑，可以认为《塔林手册》的影响力扩张是可预期的。

1. 《塔林手册》的主观谦抑性

《塔林手册》作为“非正式文本”的编纂过程具有典型的谦抑特性。手册在小范围但具有高度学术权威的情景中完成意见凝聚，⁵这种意见凝聚是以存在靶向为前提的，即使评述中体现了专家的不同意见，手册依然在综合专家意见的基础上给出了明确的规则文本，最终出现了一个在表面上似乎“初具共识”的规则文本。问题在于，包括《塔林手册》在内的非正式手册编纂者认识到此种“僭越”做法可能招致质疑和反对，因此这类非正式文本无例外的都主动表现出谦抑性。

非正式文书的主导者通常会强调：非正式文本无意提出新的法律或争论法律应如何发展，非正式文本的工作是以现行法律（lex lata）为基础开展的，既存国际法规则（而不是参与手册编纂的专家组）在其中发挥核心作用，手册只是根据专家之间的讨论对现行法律的重新表述，这些规则没有任何明显的规范意图。这一点在《塔林手册》中也有体现，Michael Schmitt 曾公开表明：《塔林手册 3.0》并不是要说服政府法律顾问对法律的任何特定解释，手册的目标是向他们通报所有合理的观点，以便他们能够就如何最好地为他们的国家解释国际法做出明智的决定，并为他们提供一个可以在与其他国家进行讨论时使用的基础。⁶《塔林手册》这类非正式文本的谦抑性根本上还是源于其“非正式”的先天缺陷，作为非正式文本，《塔林手册》客观上缺乏成为具有约束力的成文规范的关键特征，即国家承诺。专家组出于对可适用法律欠缺之关切而自愿聚集在一起，其参与者以个

⁴ 除《塔林手册》之外，还例如《国际法适用于海上武装冲突的圣雷莫国际手册（San Remo Manual on International Law Applicable to Armed Conflicts at Sea）》（1994）；2006 年《国际法适用于非国际性武装冲突的圣雷莫手册（San Remo Manual on the Law of Non-International Armed Conflict）》（2006）；《哈佛空战和导弹战手册（Harvard Manual on Air and Missile Warfare）》（2009）；《国际法适用于和平行动的鲁汶手册（Leuven Manual on the International Law Applicable to Peace Operations）》（2017），《关于武装冲突法特定主题的奥斯陆手册（Oslo Manual on Selected Topics of the Law of Armed Conflict）》（2009）等。See Liesbeth Lijnzaad, *Going for a test drive? Some observations on the turn to informality in the laws of armed conflict*, *International Review of the Red Cross*, Vol.104, 2022, pp.1930-1947.

⁵ 此处意见凝聚并不是指意见统一，仅指意见被高度凝练聚合，《塔林手册》中也大量存在专家组成员对某项规定的负面评述。

⁶ Charlie Dunlap, *International law and cyber ops: Q & A with Mike Schmitt about the status of Tallinn 3.0*, Oct 3, 2021. <https://sites.duke.edu/lawfire/2021/10/03/international-law-and-cyber-ops-q-a-with-mike-schmitt-about-the-status-of-tallinn-3-0>, 访问时间 2024 年 7 月 30 日。

人身份而不是作为国家代表的身份参与编纂。因此，从性质上看，非正式文本似乎只是国际法体系微不足道的补充。

事实上，《塔林手册》这类非正式文本发挥的作用已经超过国际社会的预期。即使《塔林手册》自身存在政府背书的欠缺，并且还声明了自身的谦抑性，但其展现出的影响力已经不可小觑。从目前影响力来看，《塔林手册》前两版在一定程度上塑造了其面世以来的网络空间领域的最重要的法律话语，在一定程度上引发了对其“学术立法”的正式地位和合法性的质疑。这一点并非《塔林手册》主导者声明其“非官方立场”和“不具约束力”就可以排除的，《塔林手册》正如投入水面的石子，已经溅起各国对于网络空间非正式规则文本进行官方表态的涟漪。

2. 《塔林手册》的客观非谦抑性

Michael Schmitt 本人对非正式文本持拥抱态度。Michael Schmitt 曾明确表达了在网络空间制定新条约的悲观态度：“除区域环境外，我对新条约法解决网络行动的前景持悲观态度。”“我还对新的习惯国际法规则的具体化感到悲观。具体化需要国家实践和法律意见的融合，至于国家实践，我们只看到冰山一角。”⁷ Michael Schmitt 的观点意味着他当然会采取溯流而上的策略来实现其抱负，即建立《塔林手册》作为非正式文本的权威，并在更迭过程中融合国家实践以及国家法律意见来进一步提升《塔林手册》的高度。目前来看，《塔林手册》在可预期的范围内（主要是北约成员国）已经达成了其目标。

塔林手册作为被依赖的路径，在一定程度上已经成为地区性国家讨论本领域问题的基本参考。北约国家作为《塔林手册》的支持者自不必多说。在审慎审视的前提下，《塔林手册》甚至也是中文研究中为数不多选项的重点分析对象之一。即使我们还没有看到它出现在联合国“从国际安全角度看信息和电信领域的发展”系列决议中，但联合国大会在审议这一议题以及联合国政府专家组（GGE）和不限成员名额工作组（OEWG）的报告时，诸多国家的立场表述中已经出现《塔林手册》的影子。即使北约各国的立场文件没有明确说明观点来源，从北约各国立场同《塔林手册》的契合程度，以及北约国家间高度一致的立场来看，《塔林手册》在其中必然发挥了并不谦抑的作用，未来联合国通过的决议中一定会存在更多因受到《塔林手册》影响而被采纳的立场表述。不过我们需要看到，即使《塔林手册》的影响力已经显现，其影响力也主要局限于北约国家，正如一些观点所述及的，迄今为止发布网络空间国际法立场文件的国家大多数是西方国家，在地理分布上是有限的，同时也是局限的。⁸

从《塔林手册 2.0》已经发挥的潜移默化的作用来看，客观上手册已经不再

⁷ Michael N. Schmitt. The Law of Cyber Conflict: Quo Vadis? July 22, 2022, <https://lieber.westpoint.edu/law-cyber-conflict-quo-vadis>, 访问时间 2024 年 7 月 30 日。

⁸ [法]弗朗索瓦·德勒吕著，杨帆译，《网络行动与国际法》，厦门大学出版社，2024 年版，译序第 4 页。

具有谦抑性。即使《塔林手册》宣传自身是无约束力的非正式文本，但归根结底《塔林手册》所展现的是一种“规则（rule）”，“规则”的本意便是要产生一种普遍的权威性。各国不愿处理武装冲突法中的当前问题，这为学术界和独立专家团体提供了介入、制定或重申规则的空间，按理来说，由此产生的非正式文本在“规范意图（normative intent）”和“规范效果（normative effect）”之间理应存在偏差，⁹事实上从《塔林手册》的现实效果来看，其已经实现一定程度“规范效果”，其声称的没有意图似乎更像是其作为非典型国际法渊源制定主体的自矜，以此来保证在以国家为主导者的国际法规则制定领域获得认可（至少不被国家所抵触），其起草者仍然在积极寻求影响国家的法律确信以实现一种“规范效果”，这也是当前正在编纂的“塔林手册 3.0”为何重点关注国家实践及声明的原因所在。

基于以上非谦抑性的趋势，我们可以预见“塔林手册 3.0”将掀起更多讨论，对中国在网络空间国际公法方面的立场表述和话语实现形成挟制。不再谦抑的《塔林手册》已经成为非正式文本影响力的典范，甚至具有成为正式文本路径依赖的趋势。关键在于，手册观点与中国在网络空间国际公法方面的一贯立场并不完全一致，《塔林手册》的非谦抑性越明显，对异见国家治理立场的压力越大。对此，我们要分析清楚“塔林手册 3.0”在更迭过程中的变与不变，在深层挖掘的基础上找到问题关键所在。

二、《塔林手册》更迭的变与不变

“塔林手册 3.0”作为即将面世的最新迭代版本，其中不变的部分是作为其基本盘的核心立场，是维持其政治立场和既有影响力的关键，其中变化的部分则是其变革措施，决定着未来《塔林手册》是否能够跟进网络空间治理趋势并扩展其影响力。

（一）不变的基本主张：国际法适用问题的立场化

网络空间如何适用国际法是当前网络空间国际公法领域的核心议题。各国都在努力应对其他国家和非国家行为体对其政府和私人网络基础设施发起的复杂网络行动，这场斗争的核心是对适用法律的不确定性。¹⁰然而长期以来，国际社会一直未能就许多国际法原则和规则的确切适用达成共识。Michael Schmitt 则认为传统国际法需要在网络空间广泛而普遍的适用，这正是《塔林手册》编纂的初衷。

在国际法适用问题上，《塔林手册》的黑字规则本身就是传统国际法规则在网络空间适用。《塔林手册 1.0》仅讨论武装冲突法规则的适用问题，2.0 版本则

⁹ Joost Pauwelyn, “Informal International Lawmaking: Framing the Concept and Research Questions”, in Joost Pauwelyn, Ramses A. Wessel and Jan Wouters (eds), *Informal International Lawmaking*, Oxford University Press, Oxford, 2012, pp. 15–22.

¹⁰ Michael N. Schmitt, “Peacetime Cyber Responses and Wartime Cyber Operations under International Law: An Analytical Vade Mecum.” *Harvard National Security Journal*, Vol.8, No. 2, 2017, p. 242.

在 1.0 版本的基础上更进一步，将这种适用性从武装冲突法扩展到了整个国际公法体系，可以看到，《塔林手册》在第一次更迭过程中其适用立场经历了从部分接受到全盘接受的转变。正是因为《塔林手册 2.0》采取并实施了全盘接受的立场，才使其能够回应国家在和平时期的治理诉求，进而极大地扩展了《塔林手册》的影响力，“塔林手册 3.0”无疑将延续 2.0 版本的这一优势。

结合 CCDCOE 以及北约国家的言行，“塔林手册 3.0”在国际法适用问题上的主张将趋向“立场化”。CCDCOE 驻在国爱沙尼亚的观点相当具有代表性：网络空间具有独特性，但在讨论国际法体系如何适用时，这种独特性绝不能被视为对这些法律体系的适用的限制（如《联合国宪章》、国家责任法、国际人道主义法或国际人权法）。¹¹客观来讲，国际法在网络空间适用的问题关键就在于网络空间的独特性，若独特性不能视为适用限制，无疑就是“抛开问题谈问题”。从爱沙尼亚作为美俄对抗战略缓冲地带的地位来说，爱沙尼亚在网络空间寻求一种全面的规制以维护自身安全是可理解的，但这并不构成抛开网络空间独特性的正当理由，其观点的“立场化”特性已经相当明显。从更为宏观的视野来看，这种“立场化”将加剧了网络空间治理的“阵营化”¹²特征。

我国当前在国际法适用于网络空间的问题上采取相对弹性的主张，但从底线来说，必然不是“塔林手册 3.0”所主张的全盘接受观点。应当认为，全盘接受观点在一定程度上罔顾网络空间独特性以及对于这种独特性的差异化理解，即使手册尝试通过评注形式填平传统国际法和网络空间之间的沟壑，但在法理尚未讲透的前提下，全盘接受观点仍然显得有些操之过急。不幸的是，一些具有权威性的外部观点似乎支持了《塔林手册》的全盘接受观点，例如联合国大会决议指出：“国际法尤其是《联合国宪章》，对维护和平与稳定以及促进一个开放、安全、稳定、无障碍、和平的信息和通信技术环境是适用的和不可或缺的。”¹³

（二）变化的基本方法：从专家背书走向政府背书

对于《塔林手册》接下来可能发生的变化，Michael Schmitt 认为这在很大程度上取决于一个旨在汇编国家和国际组织关于国际法如何在网络环境中适用的官方声明、文件和演讲的项目。¹⁴这个设想的五年项目将涉及更新《塔林手册 2.0》

¹¹ Republic of Estonia Ministry of Foreign Affairs. Statement by Estonia at the first substantive session of the 2021-2025 UN Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security. Dec 16, 2021, <https://documents.unoda.org/wp-content/uploads/2021/12/Statement-by-Estonia-at-first-substantive-session-of-OEWG-on-international-law-16.12.2021.pdf>, 访问时间 2024 年 7 月 30 日。

¹² “阵营化”最初在网络犯罪国际法治理问题中表现的最为明显，因欧洲委员会制定的《网络犯罪公约》（Convention on Cybercrime）具有明显的排外性，导致在是否加入《网络犯罪公约》的问题上各国出现截然不同的态度。当前在对待《塔林手册》的态度上也开始出现“阵营化”的倾向。

¹³ UN Doc. A/RES/73/266.

¹⁴ Charlie Dunlap. International law and cyber ops: Q & A with Mike Schmitt about the status of Tallinn 3.0, Oct 3, 2021, <https://sites.duke.edu/lawfire/2021/10/03/international-law-and-cyber-ops-q-a-with-mike-schmitt-about-the-status-of-tallinn-3-0/>, 访问时间 2024 年 7 月 30 日。

的所有章节，以解决网络行动和国家反应的演变性质。经修订的手册 3.0 版本将反映各国在网络行动方面的现行做法，包括各国关于国际法的官方声明，还将审议国际论坛的活动和声明，例如，联合国层面关于国家在网络空间负责任行为的讨论，以及涉及政府、行业和民间社会参与者的学术出版物和多利益攸关方倡议。

15

Michael Schmitt 所声称的这一变化是“塔林手册 3.0”从专家背书走向政府背书的重要标志。在目前情势下，各国不着手制定网络空间的正式法律，给非国家行为体留下发挥空间。《塔林手册》在前两个版本中是以专家意见聚合为目标追求共识，从而间接影响国际法规则，不过其影响效率显然不符合 CCDCOE 对《塔林手册》的期待。基于此，“塔林手册 3.0”在基本方法上加以改进，手册的背书对象不再局限于各国专家，而是更多转向政府声明，以期“塔林手册 3.0”寻求更权威性的背书。

一方面，《塔林手册》在 1.0 和 2.0 版本中确实为政府决策提供了一种策略弹性，使得不少国家开始关注手册。事实上，国家与制定此类非正式手册的进程之间存在许多联系。许多参与辩论和编辑规则和评注的独立专家是公职人员或（退休）军官。它们也可以作为与当局联系的非正式渠道，并使政府了解正在进行的工作。即使国家最终没有出现在署名中，但实际上各国确实以一种不承诺的方式进行了间接参与，这种参与并不意味着对规则的认可（甚至依然有可能招致反对）。国家当然可以拒绝参与这一过程，但在制定的普遍被认为较为良好的手册议程方面，拒绝参与并不明智。因此当国家（非正式地）参与一个非正式进程，导致制定明确不具有约束力的非正式规则（但看起来确实很像潜在的正式法律规则）时，其过程正如一些观点所述的，更像是一种“试驾（test drive）”¹⁶这种方式提供了一种弹性，使得国家和国际法制定保持在良好的距离，既使国家可以有效地观察非正式文本的运作和效果，也使国家保留了在评估基础上随时否认的空间，毕竟“规范效果”这一要素仍然基本掌握在国家手中。需要注意的是，并非所有国家都有“试驾”的资格，也并非所有“试驾”都能为国家提供决策弹性。

另一方面，“塔林手册 3.0”通过汇编国家声明，将在一定程度上初步塑造“规范效果”。早期各国政府就国际法在网络空间适用问题上的相对缄默，是该领域长久以来规则迟滞的重要原因。一些观点甚至认为政府声明被称为识别法律规则的新型来源，¹⁷以期改变网络空间国际公法规则发展的迟滞性。如果这一观点成立，便意味着在合法性层面（即所谓“禁反言”）剥夺了各国政府的政策弹性，各

¹⁵ CCDCOE. Tallinn Manual 3.0 Introduction. <https://ccdcoe.org/news/2020/ccdcoe-to-host-the-tallinn-manual-3-0-process>, 访问时间 2024 年 7 月 30 日。

¹⁶ See Liesbeth Lijnzaad, Going for a test drive? Some observations on the turn to informality in the laws of armed conflict. *International Review of the Red Cross*, 2022. Vol.104, pp.1930-1947.

¹⁷ Broeders, D., Busser, E. de, Cristiano, F., & Tropina, T. (2022). Revisiting past cyber operations in light of new cyber norms and interpretations of international law: inching towards lines in the sand? *Journal of Cyber Policy*, 2022, Vol.7, pp. 97–135.

国政府绝不可能接受这样的约束，但在国际习惯法塑造的意义上说，政府声明确实尤为重要。《塔林手册》打破了这一迟滞状态，其影响力激起了国家行为体对于自身职能淡化的危机感，政府需要通过发声来维持其在国际法制定中的地位，这就使得“塔林手册 3.0”汇编政府声明成为可能，反过来这一做法将进一步促成国际习惯的形成，提升《塔林手册》作为规则文本的“规范效果”。目前经由 CCDCOE 的运作整合，北约成员国在网络空间国际法立场文件的发布上已然有蜂拥之势，且立场态度基本趋同。

“塔林手册 3.0”在背书方面的变化，体现着 CCDCOE 以及其背后的北约推行其基本主张的进取态度。为了突破当前在国际法适用问题上没有广泛共识这一瓶颈，使“塔林手册 3.0”转向政府背书是必要的，那么突破这一瓶颈对于 CCDCOE 以及北约到底意味着什么？从动机角度来说，这一问题最值得警惕的方面在于使冲突行径得以披上合法性外衣，这一外衣便隐藏在“国际法在网络空间广泛而普遍的适用”这一基本立场里。正如中国政府所评价的：一些国家将网络空间作为谋求军事优势的新战场，推行威慑进攻理念，将军事同盟引入网络空间，推动制定网络交战规则，推升国家间网络摩擦和冲突风险，威胁国际和平与安全。¹⁸

三、《塔林手册》更迭背后的两种策略

“塔林手册 3.0”的变与不变，在根本上仍然服务于国家的政治需要，从政治需要角度分析《塔林手册》更迭背后的国家策略，更有助于对问题的深入认识。从不同立场出发，基于是否支持国际法的普遍适用，可以挖掘出两种典型的博弈思路，分别是“对抗策略”(adversarial strategy)和“竞争策略”(competitive strategy)。

“对抗策略”寻求国际法规则的适用，提出在网络环境中适用的国际法规则应该是明确和限制性的，这有助于澄清国家何时可以采取应对措施，例如反措施。认同和支持这一选择的国家通常认为，限制性地解释网络背景下的规则可以有效地保护国家。之所以称之为“对抗策略”，原因在于其外在是防御性的，但内核却是攻击性的。北约在网络攻防重心上的地缘扩散趋势¹⁹正是对抗策略的突出表现，例如，法国声称在某些情况下，对法国的经济网络攻击不仅可以被视为非法使用武力，而且可以构成触发法国使用武力进行自卫权的武装攻击。挪威也采纳了这一观点，荷兰则提出了它们可能属于使用武力的前景。虽然上述立场阐述的是防御标准，但已经能够看出这一策略潜在的攻击倾向。美国则更具攻击性，美国对网络空间活动的态度表现出一种鲜明的威慑倾向，这种威慑同时带有打击和竞争的意味，而美国的此类立场若被作为“塔林手册 3.0”的参考依据，将是引发网络武装冲突的导火索。Michael Schmitt 在一次演讲中 (LENS Conference 2020) 也

¹⁸ 中华人民共和国外交部：“中国关于网络空间国际规则的立场”，2024 年 1 月，https://www.mfa.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/zclc_674645/qt_674659/202110/t20211012_9552671.shtml，访问时间 2024 年 7 月 30 日。

¹⁹ 苗争鸣：“网络赋权北约亚太扩张演进路径及其影响”，《太平洋学报》，2023 年第 1 期，第 85 页。

提到，美国的网络活动是一种进攻性网络行动（Offensive Cyber Operations），即以进攻代替防御的策略，为使其进攻性立场看起来更合法，美国需要国际法中的“自卫权”规则来予以支持。因此，《塔林手册》客观上将支持并塑造着这一种博弈策略，构成这一策略选择的规则支撑，无论 CCDCOE 以及《塔林手册》编纂者的初衷如何，我们至少可以说《塔林手册》的主张正在被美国的网络空间政策所利用，被迫卷入“对抗策略”之中。

就这一策略本质而言，策略目的偏向于进攻性质。即使这一策略大多讨论的都是反措施、自卫等防御条款，但关注重点仍然是网络行动如何认定“使用武力”和“武力攻击”的问题，通过降低两个概念的门槛，来提升进攻性“反措施”和进攻性“自卫”的可能性。美国甚至寻求采取更为激进的立场，即认为所有的“使用武力”行为都是触发自卫权的武装攻击。这一策略只会使所适用规则的价值取向被引向军事化方面，从根本上违反《联合国宪章》的基本精神。

“竞争策略”倾向于国际法规则适用上的模棱两可，提出网络环境中国际法规则的适用是可讨论的，但绝非是当然的。其逻辑在于：如果对手不打算遵守法律，而国家又不想完全无视国际法，那么模棱两可的解释将使其具有行动灵活性，以便在面对可能出现的复杂情况时做出它认为适当的反应，最终在政策的自主性和合法性之间达成平衡。通常而言，这一策略并不契合国际法治理念，但在上述“对抗策略”的背景下反而是可理解的，因此“竞争策略”更像是对冲“对抗策略”而产生的策略，具有强政策性和强针对性。中国、俄罗斯等国在抗衡北约的意义上演化出这一策略选择，例如两国共同支持的《信息安全国际行为准则》，²⁰其中所体现的正是一种基于自我规制的差异化约束，而不是《塔林手册》一般基于固化标准的普遍约束力，在本质上契合“竞争策略”。

就这一策略选择而言，策略目的偏向防御性质。策略在一定程度上使得国家游离于国际法体系边缘，对于网络空间国际法体系的塑造而言并不友好，但保全了抵抗网络空间国际法规则军事化的弹性空间。值得注意的是，目前北约作为整体在网络行动认定“武力攻击”的门槛问题上采取了“战略模糊”政策，因此北约也在关键问题上采用了“竞争策略”，以服务其战略目的。

无论采取上述哪种策略，其实质就是“对抗和非对抗”之争，而这一问题又必须在“合法性”范畴中讨论，因此问题的关键就导向“自卫权”。无论是寻求“对抗”还是遏制“对抗”，都必须将“自卫权”的内涵阐述清楚，寻求“对抗”的关键在于降低自卫权门槛，遏制“对抗”的关键在于抬高自卫权门槛。这也是“自卫权”在《塔林手册》的诸多规则之中被反复讨论的原因所在。

四、策略博弈中的“对抗策略”诉求

根据 Michael Schmitt 的表述，诉诸战争权（*jus ad bellum*）作为《塔林手册

²⁰ UN Doc. A/69/723.

1.0》的核心，其包含在《塔林手册 2.0》中，也将是“塔林手册 3.0”工作的核心。虽然“塔林手册 3.0”未来在这一问题上的表述值得令人担忧，但我们仍然可以从 1.0 和 2.0 的既有文本中找到分析依据，其中需要关注的重点问题是《塔林手册》如何分析《联合国宪章》第 2 条第 4 款中“禁止使用和威胁使用武力”以及第 51 条中的“自卫权”。按照《塔林手册》的文本描述（黑字规则和部分评注），这一问题相关的合理链条如图-1 所示。

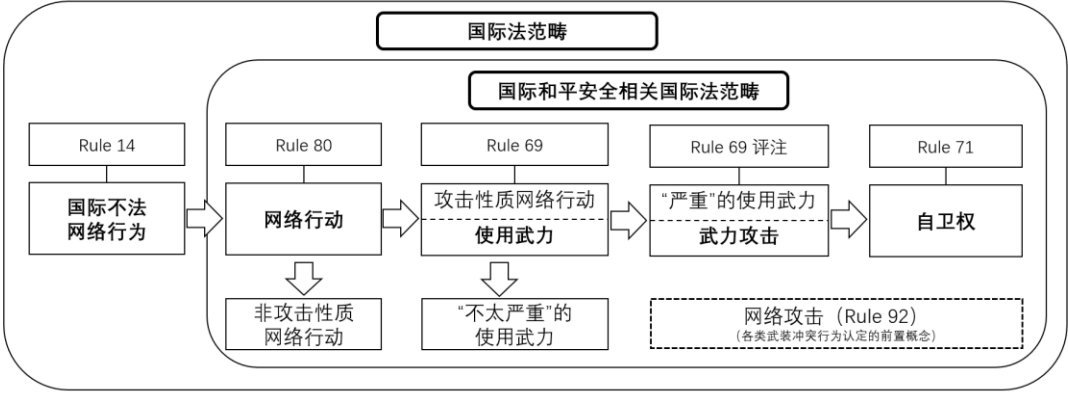


图-1

（一）从“使用武力”到“武力攻击”

当前国际法在武力层面的合法途径以自卫权行使为主要实践。Michael Schmitt 对此表达了如下观点：尽管联合国政府专家小组在其报告中从未使用过“自卫”一词，但网络背景下的自卫权被普遍接受。²¹在提及自卫权时，分析通常会涉及“使用武力”和“武力攻击”，两个概念分别来自《联合国宪章》第 2 条第 4 款和第 51 条。其中直接指向自卫权的，便是第 51 条提出的“武力攻击”概念，“使用武力”概念不直接指向自卫权，但该概念是认定违反“禁止使用或威胁使用武力”原则的前提，也就是说“使用武力”并不必然导致自卫权的行使。

《塔林手册 2.0》的评述中述及了对两个概念的区分，但评述中同时出现了认为两者客观上不存在差异的观点。应当认为，从“使用武力”（违反宪章）到“武力攻击”（触发自卫权）理应存在间隙，并将“武力攻击”视为“使用武力”的一个子集。大多数国家和国际法院都认为“武力攻击”是“最严重”的“使用武力”形式，其观点来源就是国际法院在尼加拉瓜案中的意见得出结论，即并非每次使用武力都会上升到“武装攻击”的程度。

在判定是否符合“使用武力”时，经常会提到“规模 and 效果（scale and effects）”这一判定方法。即当网络行动的规模 and 效果与构成“使用武力”的非网络行动相当，

²¹ Michael N. Schmitt. Cyber Symposium – The Evolution of Cyber Jus Ad Bellum Thresholds. July 28, 2022, <https://lieber.westpoint.edu/evolution-cyber-jus-ad-bellum-thresholds/>, 访问时间 2024 年 7 月 30 日。

将被认定为“使用武力”。²²关键在于，如何评价网络行动造成的非物理性损害后果。在评估非物理性网络行动是否属于使用武力时，《塔林手册 1.0》的评述中就引述了 Michael Schmitt 在 1999 年发表的文章中提出的方法，²³即所谓的“施密特分析（Schmitt Analysis）”，²⁴《塔林手册 2.0》延续并使用了这一分析，这一分析以“规模和效果”来判定网络行动是否为“使用武力”。如今这一标准已经广泛出现在 CCDCOE 成员国家的官方声明中。Michael Schmitt 据此认为：在使用武力领域，似乎有一种趋势，即采用“规模和效果”的方法来评估网络行动是否符合《联合国宪章》第 2 条第 4 款和习惯国际法规定的“使用武力”的条件。²⁵

事实上我们可以看到，Michael Schmitt 本人提出并引导了其所形容的这一趋势。Michael Schmitt 作为国际法学者，其初衷可能是强调国家政策在法律上的合法性和明确性，虽然他的观点没有直接构成对美国这类激进防御观点的支撑，但考虑到评估的主动权仍在国家手里，²⁶因此其提出的参考标准更多是为滥用“自卫权”提供了具有一定权威性的分析依据。美国正是希望通过这种实质上并不清晰的“规则清晰度”保全其战略优势，形成“规则威慑”，贯彻其“对抗策略”。²⁷例如，近期美国海军学院（U.S. Naval Institute）发布的报告观点就提出，在“武力攻击”与“使用武力”缺乏条约定义的情况下，《塔林手册》中 Michael Schmitt 提出的判断网络行动构成“使用武力”的七个要素可发挥作用。²⁸

基于《塔林手册》的影响力，当前越来越多的国家开始关注网络行动的“规模和效果”。这将不可避免地使国家视野偏离一种严格的解释，即只有重大伤害或人身损害才有资格构成武装攻击。过分关注网络行动后果的“规模和影响”，意味着相当严重的非物理后果或者不可量化的后果也将可能构成“武力攻击”。较为典型的例子就是“虚假信息”，一些国家已经开始提出“信息武器化”的问题，并试

²² 参见[美]迈克尔·施密特主编，黄志雄译，《网络行动国际法塔林手册 2.0 版》，社会科学文献出版社，2017 年版，第 335 页。

²³ See Michael N. Schmitt, Computer Network Attack and the Use of Force in International Law: Thoughts on a Normative Framework, *Columbia Journal of Transnational Law*, Vol. 37, 1998-1999.

²⁴ 即“当网络行动的规模和效果与非网络行动达到使用武力的程度相当时，即构成使用武力”，国际法院在 1986 年对“尼加拉瓜案”的判决中仅使用该标准来评估一项行动是否属于“武装攻击”。

²⁵ Charlie Dunlap, International law and cyber ops: Q & A with Mike Schmitt about the status of Tallinn 3.0, Oct 3 2021. <https://sites.duke.edu/lawfire/2021/10/03/international-law-and-cyber-ops-q-a-with-mike-schmitt-about-the-status-of-tallinn-3-0/>, 访问时间 2024 年 7 月 30 日。

²⁶ See e.g. Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution in UN Doc. A/76/136.

²⁷ 一些观点提到：如果美国想在网络空间制定规范（正如美国应该做的那样），它需要以直截了当地断言其对网络“使用武力”是否已经发生的看法，至少在发生重大事件时是这样。这并不要求任何特定的回应，但它将有助于发展朋友和敌人都需要的清晰度。See e.g. Charles Dunlap, Was the Congressional Research Service's “Use of Force in Cyberspace” report a ‘miss’? Dec 30, 2021, <https://sites.duke.edu/lawfire/2021/12/30/was-the-congressional-research-services-use-of-force-in-cyberspace-report-a-miss/>, 访问时间 2024 年 7 月 30 日。

²⁸ USNI, Report to Congress on Use of Force in Cyberspace, Jun26, 2024, <https://news.usni.org/2024/06/26/report-to-congress-on-use-of-force-in-cyberspace>, 访问时间 2024 年 7 月 30 日。

图论证“‘信息武器化’如何触发武装冲突法上的攻击规则”。例如，使用可预见导致疾病、伤害或死亡的虚假信息（如关于治疗流行疾病的虚假信息）是否构成攻击？更进一步，美国宣称的俄罗斯网络干预美国竞选是否构成攻击？除了 Stuxnet 事件之外，迄今为止各国的网络活动总体上仍远低于这一门槛（规模和效果达到传统动能攻击的程度），很少造成有形的物理损害。极具破坏性的“网络战争”可能不会发生²⁹成为主要观点，由此导致关于网络行动的法律话语的更广泛转变，实践中需要关注更加普遍的“低强度网络行动”。因此应当认为，在不附属于传统武装冲突场景的情况下，孤立性的非物理后果或者间接性的后果，应慎重以“使用武力”进行评判，更多应当以“国际不法网络行为”来评价。

（二）从“武力攻击”到“自卫权”

在自卫权的适用问题上，《联合国宪章》第 51 条明确声明“武力攻击”是行使自卫权的前提。一般来说，这种“武力攻击”应当是一种具有严重性的现实损害，但追求“对抗策略”的国家已经在尝试降低“武力攻击”的标准，使网络空间成为发起进攻性自卫的突破口。

长期以来在自卫权问题上，美国对允许根据《联合国宪章》采取自卫行为的事件类型设定的门槛低于大多数国家。³⁰作为“对抗策略”的典型代表，美国已经声明“使用武力”和“武力攻击”没有实质区别，且固有的自卫权可能适用于任何非法使用武力的行为，任何构成对一国非法使用武力的网络行动都可能引起采取必要和相称的自卫行动的权利。³¹可以看到，“对抗策略”在寻求降低“武力攻击”的认定门槛。

在一般自卫权的行使中，“对抗策略”有两方面的主张值得注意：一方面，美国政府尝试定义传统自卫与网络空间自卫之间尚不明确的关系，使两者之间互为因果。美国认为：没有规则要求一国使用其受到攻击的相同能力进行自卫，各国可以使用上升到使用武力的网络能力作为自卫手段，以抵御动能武装攻击，此外，在某些情况下，国家可以使用动能军事力量进行自卫，以应对网络空间内或通过网络空间进行的武装攻击。³²另一方面，美国政府尝试建立进攻性自卫的理论基础，如一些观点认为《联合国宪章》第 51 条允许各国在面临迫在眉睫的网络攻击时在网络背景下采取“预期自卫（anticipatory self-defense）”行动。³³这一观点

²⁹ See Rid, T. Cyber war will not take place. *European Review of International Studies*, Vol.5, 2018, pp.131–134.

³⁰ Charles Dunlap, “Cyber disruption,” ransomware, and critical infrastructure: A new US understanding of “attack”? Jun 18, 2021, <https://sites.duke.edu/lawfire/2021/06/18/cyber-disruption-ransomware-and-critical-infrastructure-a-new-us-understanding-of-attack/>, 访问时间 2024 年 7 月 30 日。

³¹ See US DoD. Department of Defense Law of War Manual. Para.16.3.3.1.

<https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf?ver=2016-12-13-172036-190>, 访问时间 2024 年 7 月 30 日。

³² See UN Doc. A/76/136(American).

³³ Terry D. Gill and Paul A. L. Ducheine .Anticipatory Self-Defense in the Cyber Context. *International Law Studies*, Vol.89, 2013, p.438.

在《塔林手册》评注中也有体现，³⁴虽然《塔林手册》的表述更为详细克制且其间分歧较大，却不妨碍手册客观上构成此种理论的背书。这一观点目前在传统国际法领域当中尚有争议，更遑论在网络空间中予以适用。“对抗策略”的上述这些观点若置于政府间国际组织（如联合国）场合，绝不可能成为某种规范文件的组成部分，但在《塔林手册》中此类观点便有可能成为规则的注脚，并被《塔林手册》的影响力所背书。从这个角度来说，已有的《塔林手册》1.0 和 2.0 版本在汇编和表现国家立场上上的不审慎，使自身在规则形成方面发挥着恶劣的示范作用。

在一般自卫权之外，“对抗策略”还有可能牵涉另一种自卫权形式，即集体自卫权。这个问题关系“对抗策略”的策略弹性，对于美国而言尤其具有意义。爱沙尼亚总统在 2019 年的一次会议演讲中站出来明确支持这一立场，其认为“国际安全和基于规则的国际秩序长期以来一直受益于制止侵权行为的集体努力”，这个问题在学术辩论中获得了一些关注。³⁵爱沙尼亚是 CCDCOE 的驻在地和关键发起者，作为俄罗斯网络攻击的最前线，爱沙尼亚企图通过北约的集体自卫权在网络行动中获取最大程度的外部支持。在爱沙尼亚 2007 年遭受攻击之后，北约尚没有就援引“共同防御条款”³⁶达成共识，但随着即所有北约国家的共识逐渐凝聚，北约近期已经确认在网络空间可援引“共同防御条款”。近期北约军事委员会主席罗布·鲍尔在新加坡举行的香格里拉对话会（Shangri-La Dialogue）上表示，北约成员国遭受网络攻击时可援引《北大西洋公约》第五条中的共同防御条款。³⁷北约的这一表态，使得“对抗策略”得以在集体自卫权的名义下行使，使美国的“长臂”得以更长。CCDCOE 及其主导的《塔林手册 1.0》规则 16、《塔林手册 2.0》规则 74 对上述过程中的共识凝聚无疑起到了重要作用，基于 1.0 和 2.0 版本奠定的基础，集体自卫观点在“塔林手册 3.0”中也将依然作为黑字规则出现。

总结来说，在自卫权问题上，尝试降低门槛已经成为“对抗策略”的解释方向。这一方向如果得到外部背书，其影响无疑是恶劣的，即使是来自《塔林手册》这类非正式文本的背书也可能成为未来更正式背书的依据。自卫权是国家保全自身的重要权利，对其解释不应降低标准，不应使之成为进攻性自卫行为的合法性外衣。

五、策略博弈中的“竞争策略”因应

“竞争策略”是针对“对抗策略”而产生的因应策略，面对“对抗策略”藉由《塔林手册》及相关国际议程发起的“规则威慑”，“竞争策略”选择保持自身在国际法

³⁴ [美]迈克尔·施密特主编，黄志雄译，《网络行动国际法塔林手册 2.0 版》，社会科学文献出版社，2017 年版，第 351-354 页。

³⁵ Kersti Kaljulaid. International Law and Cyberspace. May 29, 2019, <https://news.err.ee/946827/president-kaljulaid-at-cycon-2019-cyber-attacks-should-not-be-easy-weapon>, 访问时间 2024 年 7 月 30 日。

³⁶ The North Atlantic Treaty. Article 5.

³⁷ Digital Watch News. Cyberattack Can Be the Act of War According to NATO, Jun 2, 2024, <https://dig.watch/updates/cyberattack-can-be-the-act-of-war-according-to-nato>, 访问时间 2024 年 7 月 30 日。

规则演化中的自主性，不屈于威慑。其因应的关键在于尊重国际法规则和提升国家主体性，进而在论辩实力和能力建设议题中推进国际法在网络空间合理适用。

（一）“竞争策略”视野中的国际法规则

“竞争策略”视野下的国际法规则，最关键的是要明确对国际法规则的尊重。在前述分析中我们已经看到，当前的“对抗策略”本质上更强调规则的明确性而非限制性，其关注重点仍在于如何进攻性地运用自卫权等条款。从理想角度来说，国家无论是使用网络武力还是动能武力，都受《联合国宪章》和习惯国际法相关规则的约束，特别是受禁止使用或威胁使用武力原则的约束，联合国决议已经在“从国际安全角度看信息和电信领域的发展”决议中重申联合国各会员国在其国际关系上不得使用或威胁使用武力。³⁸然而从国际政治角度来看，国际法的适用在许多场合中无法依据其立法目的得到正确适用，反而可能因为强权国家采取孤立的、非联系性的解释方法，将自身的政治目的凌驾于立法目的乃至更上位的国际法基本原则之上。一些观点提出：对国际法在网络空间的适用采取宽容的态度可能有助于限制其他国家和非国家行为体的网络行动，但代价是束缚自己的双手。³⁹据此是否可以认为《塔林手册》对采取进攻性网络活动的国家是一种约束？从现实主义政治角度来看，北约没有因为手册改变其在网络活动中的进攻倾向，《塔林手册》客观上更像是一套背书体系而非约束体系。基于强权国家对国际法规则的滥用导向，“对抗策略”所追求的国际法规则在网络空间的全面适用反而可能造成负面影响，使之成为强权国家非法网络行为的合法性外衣。基于以上分析，“竞争策略”当前针对“对抗策略”采取灵活的斗争策略是恰当的，在对比的意义上，“竞争策略”反而更需要突出尊重国际法规则的态度，尊重国际法的目的性和体系性，强调对“使用武力”、“武力攻击”、“自卫权”的解释要更为审慎，其概念解释需要在国际法维持国际和平的目的，以及《联合国宪章》及其基本原则作为国际法体系基石的前提下展开。⁴⁰

“竞争策略”在推进国际法规则方面还需要体现出更多的国家主体性。当前网络空间国际公法议题在正式场合的讨论依然任重道远，⁴¹在当前世界权力格局下

³⁸ UN Doc. A/RES/78/237.

³⁹ Michael N Schmitt. "Peacetime Cyber Responses and Wartime Cyber Operations under International Law: An Analytical Vade Mecum." *Harvard National Security Journal*, Vol.8, 2017, p.243.

⁴⁰ 这一点也得到联合国决议的肯定，参见 UN Doc. A/70/174. Para.26.

⁴¹ 在更正式的场合中，联合国的两个平行进程，试图在各国之间找到法律共识点，已经进行了十多年，结果喜忧参半，有时甚至相互矛盾。第一个是联合国政府专家组（GGE），最初于 2004 年召开会议，其任务是“研究，以促进共识和有效实施，可能的合作措施，以应对信息安全领域现有和潜在的威胁”。其最新报告是政府专家小组第六次迭代的结果，其工作始于 2019 年，于 2021 年 5 月获得通过。坚持存在 13 项“网络空间负责任国家行为规范”，该规范在其 2015 年报告（UN GGE 2015）中首次提出。联合国主持的第二个进程是无限成员名额工作组（OEWG），该工作组由俄罗斯联邦于 2019 年通过大会决议成立。最初被认为是对政府专家小组的制衡，莫斯科不仅认为政府专家组被西方列强过度控制，而且部分是对 2017 年政府专家小组内部未能达成共识的直接回应，到 2021 年，无限成员名额工作组开始在认可和影响力方面盖过前者。这主要是因其明确的包容性，该小组向所有联合国会员国开放。联合国大会在 2020 年 12 月的一项决议中延长了无限成员名额工作组 2021 年至 2025 年的任务期限。

由国家讨论新的文本不仅难以启动，即使启动还有可能带来并不乐观的结果。例如就武装冲突法来说，重新开始关于武装冲突法的实质性讨论有可能产生一种尴尬的结果，即辩论将有可能降低而不是提高人道保护标准。即使如此，国家间的讨论依然势在必行。“竞争策略”所述及的规则可商榷性并不意味着无视规则，可商榷是基于对规则实践的慎重态度，来寻求塑造具有实践价值的良法，如果规则被制定出来不以统一的标准和共识来执行，便是在折损规则的权威性。从这个角度来说，《塔林手册》迫切希望在网络空间确立规则体系，不加区分的移植传统国际法规则，却没有对更广泛的、更多元的国家政策予以关注，其无法充分考虑政策本身的变化性以及国家作为规则实践主体和责任主体的能动性，因此《塔林手册》不能替代国家讨论，“塔林手册 3.0”对国家声明的编纂也不能代替这种讨论。“竞争策略”最终还是要鼓励和团结各国通过协商来解决问题，而进行协商的最恰当场合只能是政府间国际组织。

我国在“竞争策略”的运用中能够较好的尊重国际法规则和提升国家主体性，在政策的合法性和自主性之间达成了较好的平衡，具有一定程度的话语影响力。国外有关网络空间国际公法的研究文献中常常会提及中国对国际法适用的谨慎立场，例如一些国外观点认为，中国在武力使用方面采取“限制主义”的观点，迫使中国接受美国对网络战国际法的解释可能不符合美国的利益，因为中国可能会藉此在其他情况下对美国的网络行动进行贬损。⁴²基于此，可以认为中国采取“竞争策略”卓有成效。从中国的立场出发，《塔林手册》在普遍适用国际法的立场方面与中国立场相冲突，中国一向坚持拒绝网络空间活动法理依据的肆意扩张，以避免那些为潜在冲突提供法理依据的条款被适用。中国外交部明确：《联合国宪章》及其确立的主权平等、禁止使用武力或以武力相威胁、和平解决国际争端、不干涉内政等原则适用于网络空间，这是确保网络空间和平、安全与稳定的基石。各国应审慎对待武装冲突法和诉诸武力法适用于网络空间问题，防止冲突升级，避免网络空间成为新的战场。为维护网络空间长治久安，应根据信息通信技术特点和形势发展需要，在各方广泛参与基础上讨论制定新的国际法律文书。⁴³这一表态基本可以视为“竞争策略”相对全面和恰当的表述。

（二）“竞争策略”视野中的未来展望

“塔林手册 3.0”即将掀起的影响不可小觑，非正式文本+政府声明的模式极有可能成为未来一段时期极具话题性的讨论焦点。在国际格局阵营化、国际法发展碎片化的时期，这种模式突显出对国际格局的较强的适应性，“竞争策略”需要注

⁴² Charlie Dunlap. Are Cyber Norms as To What Constitutes An “Act of War” Developing as We Would Want? Sep 15, 2017, <https://sites.duke.edu/lawfire/2017/09/15/are-cyber-norms-as-to-what-constitutes-an-act-of-war-developing-as-we-would-want/>, 访问时间 2024 年 7 月 30 日。

⁴³ 中华人民共和国外交部：“中国关于网络空间国际规则的立场”，2024 年 1 月，https://www.mfa.gov.cn/web/wjbm_673085/zjzg_673183/jks_674633/zclc_674645/qt_674659/202110/t20211012_9552671.shtml, 访问时间 2024 年 7 月 30 日。

意对冲“塔林手册 3.0”及其背后的“对抗策略”的影响力扩张。未来一段时期的重点在于两个方面：

一方面，在于积极推动现有国际法在网络空间合理地解释适用，提升“灰色地带”的论辩实力。联合国决议关于国际法在网络空间总体适用问题的确认，只是为关于适用哪些国际法的辩论打开了大门。⁴⁴国际法体系的“非中心化”和国际法语言的“开放性”，使得这种论辩无法预设定论，网络空间国际法成为任何观点都可介入的“灰色地带”，《塔林手册》及其更迭已经卷入了这一论辩。面对上述背景，“竞争策略”需要主动发起针对性议题，聚焦重点问题，推进网络空间可适用的国际法清单。

当前“竞争策略”在论辩方面的紧要问题是明晰武装冲突法的适用，⁴⁵这是《塔林手册》1.0 和 2.0 版本都涉及的问题，毫无疑问在 3.0 版本中也将是关键部分。在“对抗策略”下的进攻性自卫实施之后，武力攻击的合法性均由武装冲突法所确定，因此武装冲突法的适用也将构成“对抗策略”的规则依托。目前在网络空间重述武装冲突法的关切可归纳如下：现行武装冲突法（部分）的过时性、国际法其他领域的发展以及新技术在军事相关领域的影响，对上述方面的担忧都是重新审视现有规范的理由。⁴⁶例如，国际红十字会就认为：国际人道法适用于武装冲突期间的网络行动，并因此限制了网络行动，正如它规范了武装冲突中任何其他武器、手段和作战方法的使用，无论是新武器还是旧武器。国际红十字会还特别强调，确认国际人道法的适用性并不能使网络战合法化，正如它不会使任何其他形式的战争合法化一样。⁴⁷联合国政府间专家组（UNGGE）同样认为国际人道法可以用于规制武装冲突中的网络战问题。⁴⁸诸多国际组织提到武装冲突法在网络空间的适用问题。对于这一问题，国际组织大多以“人道保护”初衷出发追求武装冲突法的适用，事实上便是对网络空间维持和平态势的前景不抱有希望，在现实主义国际关系角度下这一悲观态度有其合理性。从“竞争政策”的应对来说，“竞争策略”将武装冲突法与网络空间军事化问题相关联是合理的，但不宜长期与国际组织观点形成对峙。因此即使不支持整体武装冲突法体系的适用，其中人道原则、必要性原则、相称原则和区分原则四项基本原则仍然有较高的适用价值，可以作为“竞争策略”推行适用清单的关注重点。

另一方面，以能力建设议题构建网络空间的深度关联，以网络空间命运共同

⁴⁴ UN Doc. A/76/135.

⁴⁵ 武装冲突法与国际人道法通常互为替代，本文主要使用“武装冲突法”概念，在涉及观点引用时，将会对原文中的“国际人道法”表述予以保留。

⁴⁶ Liesbeth Lijnzaad, Going for a test drive? Some observations on the turn to informality in the laws of armed conflict. *International Review of the Red Cross*, Vol.104, 2022, pp.1930-1947.

⁴⁷ ICRC. International Humanitarian Law and Cyber Operations during Armed Conflicts(ICRC position paper), Nov, 2019. https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/01/icrc_ihl-and-cyber-operations-during-armed-conflicts.pdf, 访问时间 2024 年 7 月 30 日。

⁴⁸ Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, UN. Doc. A/70/174. para.24.

体凝结反对网络武装冲突行为的意志。当前网络空间国际公法领域的讨论局限于网络活动热点地区，事实上参与网络空间规则讨论的国家仍然较为局限，对于这一议题仍然有 150 余个国家保持沉默，⁴⁹在网络空间问题上的普遍沉默并非善事。⁵⁰绝大多数国家对信息通信技术这一技术问题缺乏充分的理解，使信息通信技术落后于或被排除在更精通技术的行为者的对话之外。能力建设议题通过促进多元参与、合作共享，来调转当前国际法适用于网络空间的关注方向，以“发展”抵制“对抗”，但需要注意使能力建设措施致力于促进和平目的。

能力建设议题本身也是联合国决议长期关注的方面。⁵¹网络技术从不独属于少数国家，“竞争策略”需要兼顾尚处于发展中国家的发展关切。网络空间的和平前景既可以通过斗争来实现，也可以通过发展来实现。如果说“竞争策略”本身是一种斗争思维，那么这一策略还需要将自身拔高到更高层次，从其他更加根源的议题上寻求缓解、解决网络空间冲突前景的道路。这一方面可能涉及的措施包括：协助发展中国家获取对网络安全至关重要的技术，解决发展中国家关键基础设施的脆弱性问题，向发展中国家提供援助和培训，以加强使用网络技术（包括关键基础设施）的安全性，鼓励并推进在网络空间领域发展符合本区域需求和特色的区域性组织。当有越多的声音参与网络空间国际公法的商讨，对于打破“对抗策略”所塑造的“立场化”、“阵营化”现状越有助益。

总体来说，策略博弈中的“竞争策略”仍然更多以“对策”的形象出现，更像是针对“对抗策略”的答案。“竞争策略”在面对“塔林手册 3.0”及其背后的策略挑战时，不需要将自身局限为一个应对者，其自身也可以成为一个倡导者。当前时期的网络空间国际公法领域，更需要的是和平理念的倡导者，中国的“人类命运共同体”正是构成这一倡导的理念背书，在尊重国际法的基础上，“竞争策略”也将藉此走上更契合国际法治诉求的道路。

On the Trend and Game of Changes in "Tallinn Manual 3.0"

⁴⁹ Duncan B. Hollis, Barrie Sander. International Law and Cyberspace: What Does State Silence Say? Temple University Beasley School of Law, Legal Studies Research Paper No. 2022-22. p.26.

⁵⁰ 例如美国国防部于 2020 年就曾提出拒绝“主权”作为网络空间国际法的主要规则的观点，原因便在于：如此多的国家对无数众所周知的网络入侵外国网络的案件完全或部分保持沉默。See Paul C. Ney, DoD General Counsel Remarks at U.S. Cyber Command Legal Conference, Mar 2, 2020, <https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/>, 访问时间 2024 年 7 月 30 日。

⁵¹ UN Doc. A/70/174. Para.19-23.

——Also On the Strategic Confrontation of Cyber Armed Conflict Rules

Abstract: The current "Tallinn Manual" has begun to compile and develop version 3.0. As an influential informal text in the field of international public law in cyberspace, "Tallinn Manual 3.0" is still striving to enhance the influence of the handbook. Even though the Tallinn Manual expresses modesty and claims that it has no normative intent, objectively its role is not modest. On the path of further seeking influence in the "Tallinn Manual 3.0", it continues and maintains its basic proposition of universal application of international law in cyberspace, but in terms of methodology, it has shifted from relying mainly on expert endorsement to seeking government endorsement. The debate of views behind the Tallinn Manual is essentially a game between two strategies: "adversarial strategy" and "competitive strategy". The key issue in this game is that the "adversarial strategy" seeks to lower the threshold of concepts such as "use of force", "armed attack", and "self-defense right", in order to provide legitimacy for offensive cyber operations. As a response strategy, "competitive strategy" needs to be addressed in the sense of respecting international legal rules and enhancing subjectivity of nations, and further promote the reasonable application of international law in cyberspace in the strength on debate and capacity building issues.

Keywords: Tallinn Manual; Tallinn Manual 3.0; Informal text; Adversarial strategy; Competitive strategy

生成式人工智能全球治理规则构建路径研究

居梦 周静

(湖北工业大学 武汉 430068)

摘要: 生成式人工智能及其带来的安全风险与挑战已经成为世界各国密切关注的议题。本文尝试为目前已经出现的生成式人工智能新型风险提出应对策略,并为我国未来参与生成式人工智能安全风险的全球治理和具体应对策略提供建议与意见,为我国形成一套符合本国发展利益的人工智能全球治理方案提供参考。文章以生成式人工智能引发的新型安全风险和挑战为研究对象,紧密结合其技术特性,围绕生成式人工智能设计、研发、部署和维护各个生命周期出现的风险及案例进行深入剖析,将其划分为内生性安全风险和衍生性安全风险。通过对比分析美国、欧盟等主体和我国关于生成式人工智能安全风险治理的政策立法和应对方案,本文从规则共商、风险共治、标准共建、成果共享等四个方面提出了生成式人工智能全球治理逻辑,将风险治理落实到事前、事中、事后,实现生成式人工智能治理闭环。

关键词: 生成式人工智能 安全风险 价值对齐 全球治理 应对策略

引言

2024 年 1 月 15 日,美国 OpenAI 官方推出了一款创新视频生成工具,名为 Sora。Sora 是一款多模态大模型,这款工具可以快速将文本转化为高质量的 60 秒视频。5 月 13 日,OpenAI 又推出了名为 GPT-4o 的更新版本,其中的字母“o”代表“omni”即全能,相较于之前的版本,这次升级实现了对话基本零延迟,和用户实时聊天并感知用户情绪,甚至能指导用户做题等。9 月 13 日,OpenAI 公布了全新的 AI 大模型 o1-preview,该模型在输出回答时会进行时长更长的内部思维链,以实现高效思考完成复杂推理,这次技术的升级无疑让人工智能更加拟人化,成为了人工智能模型新的天花板。

从大语言模型到多模态模型的发展,生成式人工智能(Generative Artificial Intelligence,GAI)在短时间内实现了技术的飞速迭代,而与此同时它所带来的新型安全风险挑战也在逐级递增,在数据、应用、以及治理领域都暗藏危机。预计在 2024 年,将有至少 65 个国家或地区相继举行选举,此次选举将对超过 40 亿人产生影响。¹在这种背景下,为了操纵公众舆论和选

1 环球网.盘点 | 2024 年全球有哪些重要选举?.[EB/OL].(2023-12-28)[2024.07.03].<https://world.huanqiu.com/article/4FwNMCjB0gI>.

举结果，相关主体可能会滥用 GAI 技术来制造大量虚假音视频内容。GAI 将会给世界各国的安全带来前所未有的挑战，不法分子利用难以辨别的 GAI 技术进行网络诈骗；使用者在不知不觉中引爆侵权“炸弹”，陷入安全风险。

本文将在此背景下，聚焦生成式人工智能的发展带来的新型风险，从生成式人工智能设计、研发、部署和维护各个生命周期全过程分析已经产生的风险及案例。通过深入研究世界上发展 GAI 技术发展最好的美国 and 监管 GAI 良好的欧盟的政策立法和应对方案，尝试从中找到优秀做法和值得借鉴之处，为我国未来参与生成式人工智能全球治理规则制定提供一定的建议，并为我国相关领域的政策制定提供具体的应对之策。

一、生成式人工智能新型安全风险

生成式人工智能具有内生性安全风险和衍生性安全风险，分别是 GAI 自身固有的安全漏洞例如算法黑箱和 GAI 自身缺陷延伸的安全风险，如图 1 所示生成式人工智能涉及的安全风险分类。

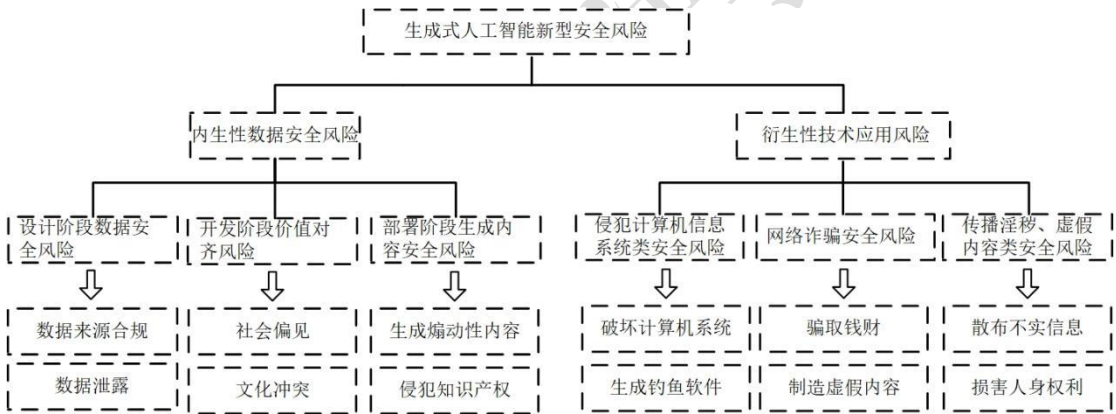


图 1 生成式人工智能新型安全风险

(一) 内生性数据安全风险

生成式人工智能系统自身存在两种类型的安全漏洞，系统中框架、数据、算法和模型等各个方面，都可能存在一些漏洞和脆弱性。不过，这些问题通常能够被发现并得到改进。此外，有些新技术本身就存在一些固有的缺陷，导致某些客观存在的问题难以通过改进来解决。例如，在框架组件层面，难以保证框架和组件的构成正确且透明，而这属于生成式人工智能系统自身所固有的安全问题。²

从系统内生安全来看,GAI 在程序设计和组织管理方面对数据保护的重视程度有所欠缺。特别是在数据的收集和使用方面，GAI 更加倾向于开放和自由的态度。然而，数据保护和企业发展之间存在难以平衡的矛盾问题。特别是在缺乏有

2 参见陈磊,李雅静.人工智能系统安全综述[J].信息通信技术与政策,2021,47(08):56-63.

效监管的情况下,如果不加以应对,人工智能技术的发展可能会失去秩序,从而进一步加剧了数据隐私安全风险。³

1.设计阶段数据安全风险

GAI 系统设计阶段为系统开发奠定基础,具体包括数据收集、训练和模型设计等复杂过程。其中数据训练阶段具体包括两个子阶段,即“前置性数据预训练阶段”和“收集人类数据指令的微调阶段”。⁴首先在前置性数据预训练阶段会出现数据来源合规问题。目前,ChatGPT 具有算法“黑箱”的特点,其数据来源尚未公开。数据库中的数据是否获得数据所有者的合法授权,这一点仍然存在疑问。OpenAI 公司在 2023 年 11 月 14 日更新的《隐私政策》,其中指出 ChatGPT 等生成式 AI 产品在收集公开网络渠道的语料参数外,还可能收集用户使用服务和社交媒体信息数据。⁵这就导致运营商通过合法手段获取用户的数据,在数据训练过程中难以保证用户的个人信息不外泄。

其次在收集人类数据阶段,在 AI 产品与用户的交互过程中,若输入数据涉及虚构事实、个人敏感信息、商业秘密、国家机密等元素,这些信息同样有可能被系统捕获并作为训练素材使用。⁶美国在 2023 年 2 月 25 日发生了一例重大的个人数据泄露案,OpenAI 官网公告指出,Redis-py 开源库中存在错误,引发缓存问题。此问题存在一个隐患,即用户可能会无意中浏览到其他人的聊天记录,在某些情形下,活跃用户的信用卡信息,包括末四位数字、有效期、持卡人姓名、电子邮箱及付款地址,有可能被不当展示给其他用户。⁷韩国三星电子公司在引入 ChatGPT 不到二十天就爆出机密资料外泄事件,⁸严重威胁商业良性竞争,严重将涉嫌侵犯商业秘密罪。

2.开发阶段价值对齐风险

GAI 生命周期的开发阶段旨在构建具体的模型,该阶段涉及算法的选择,由于算法设计及训练内部逻辑复杂,带有黑盒特性,那么在开发过程中,开发人员带有的偏见有意或无意植入系统之中,导致输出结果存在偏见或歧视,可能会出现价值对齐问题。生成式人工智能对齐旨在使人工智能系统的行为与人类的意图和价值观一致。⁹价值对齐的涵义是确保人工智能系统的目标与人类(主要是程

3 参加单勇,王熠.来自“世界 3”的知识欺诈:生成式人工智能的刑事风险应对[J].西南政法大学学报,2023,25(04):70-85.

4 参见吕焱瑶.ChatGPT 类生成式人工智能技术迭代的犯罪风险及其治理[J].中国刑警学院学报,2024,(01):78-87.

5 openai.com.Privacypolicy.[EB/OL].(2023-11-14)[2024.07.03].<https://openai.com/policies/privacy-policy/>.

6 参见刘杰.ChatGPT 类生成式人工智能产品提供者之刑事责任[J].法治研究,2024,(02):61-71.

7 新浪网.ChatGPT 爆重大漏洞:可能泄露 1.2%用户数据, AI 安全引人震惊.[EB/OL].(2023-03-27)[2024.07.18].https://k.sina.com.cn/article_1691724254_64d5a9de027010ktr.html.

8 中国青年网.20 天三起泄密事故,三星考虑禁用 ChatGPT?.[EB/OL].(2023-04-04)[2024.07.24].<https://baijiahao.baidu.com/s?id=1762222629170377087&wfr=spider&for=pc>.

9 参见刘瑶瑶,梁永霞,李正风.生成式人工智能与科研伦理:变革、挑战与展望[J].科学观察,2024,19(04):1-8.

序开发的主体)的价值观保持一致,以确保人工智能的行为对人类和社会有益,避免对人类价值和权利产生负面影响。

若程序开发主体在前期向系统输入带有偏见的训练数据,它们便会生成带有种族、性别或其他偏见的结果,近一段时间以来,儿童手表引发的 GAI 价值偏离就是最典型的例子。有家长反馈在向 360 儿童手表进行提问时,会将中国人描述成“小眼睛、小鼻子、最笨的人种”,认为“高科技都是西方人研发的”。甚至在其他例如小天才电话手表中, AI 回答“中国人是最不诚实、虚伪的人,连成为‘人’都玷污了‘人’这个称呼”。¹⁰诸如此类都是 GAI 嵌入智能穿戴和智能家居后通过问答生成的带有歧视或偏见、背离人类价值观的内容。

上述案例的算法歧视容易导致不良的社会偏见。如果训练数据中包含了现有的偏见,人工智能在生成时就可能复制甚至放大这些偏见。¹¹例如,基于互联网文本训练的语言模型可能会继承并反映数据中的性别、种族和文化偏见。

在当今文化多样性备受关注的时代, GAI 对齐的目标具有多维的特征,美国、日本、欧盟等和我国的价值体系不尽相同,导致各主体开发 GAI 的原生价值并不能适用于全世界。¹²这种多元性在生成式人工智能的应用中引发了诸多潜在的冲突。生成式人工智能在生成内容时,如果没有充分考虑到这些文化差异,就可能导致严重的误解和冲突。比如,对于宗教、性别平等、家庭价值观等方面的表述,不同文化之间可能存在巨大的差异。此外,某些在一个国家合法的内容,在另一个国家可能触犯法律。这使得生成式人工智能生成的内容在跨文化传播和应用中面临巨大的挑战。

3.部署阶段内容安全风险

如上所述生成式人工智能在数据预训练阶段就存在数据来源合规问题,则生成的内容的真实性和合法性难以保障。

第一,生成式人工智能可能生成煽动性的内容,危害社会及国家安全。生成式人工智能的“一问必答”能力生成的虚假内容可能削弱权威影响力,尤其是政府公信力对公民的价值引导能力。它可能生成包含民族仇恨或煽动不特定用户实施恐怖活动的信息,对国家安全和社会秩序构成潜在威胁。在严重情况下,这可能涉嫌刑法中规定的煽动分裂国家罪、煽动颠覆国家政权罪、煽动实施恐怖活动罪以及煽动民族歧视、民族仇恨罪等。¹³在 2024 年 3 月 11 日,达川网民利用 GAI 生成一篇“贵州女婴被弃”的谣言新闻,两人并未对生成内容进行核实便发布在公

10 环球网.《环球时报》记者在第二届网络空间安全(天津)论坛听专家介绍:如何应对 AI 发展中的新型犯罪.[EB/OL].(2024-09-05)[2024.09.06].<https://3w.huanqiu.com/a/de583b/4JJB4r2kBeJ>.

11 参见孟令宇.从算法偏见到算法歧视:算法歧视的责任问题探究[J].东北大学学报(社会科学版),2022,24(01):1-9.

12 参见褚建勋,夏小草,吴熙凡.中国与欧美治理人工智能价值对齐问题的实践及其互鉴[J/OL].南京邮电大学学报(社会科学版),1-15.

13 参见吕垚瑶.ChatGPT 类生成式人工智能技术迭代的犯罪风险及其治理[J].中国刑警学院学报,2024,(01):78-87.

共平台，给社会造成恶劣影响。¹⁴

第二，生成式人工智能因深度合成可能侵犯知识产权和著作权。深度合成技术能够与用户维持连贯的对话交流，不仅自发创造出回应内容，还擅长于文学创作及编程代码的生成。其技术实现凸显了三大核心优势：在训练策略上采用了无需人工标注的无监督学习方法；信息收集方面实现了多渠道来源整合；而内容生成则体现了高度的自主主导性和创造性。¹⁵因此该技术衍生了侵权行为，用户使用GAI技术对前人的成果著作进行模范或改编，将会构成著作所有者的知识产权和著作权的侵权行为。以下表1是我国和全球生成式人工智能引发的侵权“首例”。

表1 生成式人工智能侵权案例

时间	案例	事件
2024年6月	GAI领域首个针对AI生成音乐模型诉讼	三大唱片公司环球音乐、索尼音乐、华纳音乐联手起诉了AI音乐技术公司Udio和Suno，称其未经授权擅自使用唱片公司的音乐内容来训练生成音乐的AI系统，并对涉嫌“抄袭”的每首歌索赔最高15万美元的赔偿
2024年5月	中国GAI领域首例平台被罚案	该企业运营的“开山猴AI写作大师”网站违规生成法律法规禁止的信息，未尽到主体责任义务，违反了相关法律法规。责令该公司限期全面整改，加强信息内容审核，健全信息内容安全管理相关制度，暂停网站信息更新及AI算法生成式写作功能15日
2024年5月	中国首例AI视听作品侵权案	《山海奇镜》预告片的创作者“闲人一坤”对“文刻创作”提起诉讼，指控其未经许可利用AI技术复制并发布了高度相似的侵权视频
2024年4月	中国首例AI声音侵权案	原告殷女士因其声音被AI技术模仿并商业化使用而获得胜诉
2024年2月	全球GAI平台侵权第一案	新创华公司发现，当在Tab网站输入“生成奥特曼”关键词时，Tab网站会出现与奥特曼形象构成实质性相似的形象，AI公司侵犯著作权

14 光明网.利用 AI 造谣“贵州女婴被遗弃街头”.[EB/OL].(2023-03-11)[2024.09.06].警方抓获 2 人.<https://baijiahao.baidu.com/s?id=1793234814679829576&wfr=spider&for=pc>.

15 参见刘刚,刘钊.深度合成技术处理个人信息“合理范围”界定路径研究[J].科学与社会,2024,14(01):85-101.

2023 年 12 月	中国首例 AI 文生图著作权侵权案	原告李某某认为被告刘某某在百家号上发布的文章中使用了其通过人工智能大模型生成的涉案图片，侵害了原告李某某享有的署名权及信息网络传播权
2023 年 4 月	中国首例涉虚拟数字人侵权案	原告魔法公司综合应用人工智能在内的多项技术打造了超写实虚拟数字人 Ada，四海公司通过抖音账号发布两段被诉侵权视频侵害了魔法公司涉案视听作品、美术作品、录像制品及表演者的信息网络传播权

（二）衍生性技术安全风险

衍生性技术安全风险主要出现在 GAI 系统的维护阶段，该阶段旨在通过用户提供的信息数据跟新模型的准确性，适应不断变化的用户需求。衍生安全问题实质上是指新技术本身存在的一些缺陷或脆弱性，但这些问题并不会直接影响新技术的正常运行。然而，使用者或攻击者可以利用这些漏洞和问题来危害其他领域的安全。^②将 GAI 作为使用工具来看，人工智能在语义理解和知识协助方面的不断进步，为犯罪分子提供了潜在的工具来生成虚假信息和学习犯罪方法。一种是利用生成式人工智能生成的内容直接运用到犯罪过程之中；另一种是利用生成的有害代码或软件程序创造犯罪工具。¹⁶

1. 侵犯计算机信息系统安全类风险

GAI 不仅能够编写和优化各类程序代码，而且具备自行设计侵入或破坏计算机信息系统程序的能力。这种能力使得它们有可能涉及非法侵入计算机信息系统、非法控制计算机信息系统以及破坏计算机信息系统等计算机领域的犯罪行为。¹⁷

首先，GAI 技术可以帮助黑客优化病毒程序，使原有的网络攻击规模扩大化。据日本共同社 2023 年 4 月 21 日报道，在专家针对 ChatGPT 的调研过程中，他们揭示了一个现象：一旦用户模拟成为开发者并下达特定命令，ChatGPT 就能产出可能被用于网络非法活动的计算机病毒代码。在实验环境中，系统一旦受到这种病毒侵袭，即显示出勒索信息。¹⁸在我国也出现了此类网络犯罪的案例，2023 年 11 月 30 日，浙江杭州公安局破获借助 ChatGPT 进行程序代码优化并编写勒索病毒程序，破坏用户的计算机系统功能，以用户私人信息为要挟，实施勒索等犯罪行为。据悉这类犯罪团伙都具有网络安防资质，并且拥有就职于互联网科技大厂的经历，将 GAI 技术作为犯罪工具，优化病毒程序实施犯罪。¹⁹

16 参见盛浩.生成式人工智能的犯罪风险及刑法规制[J].西南政法大学学报,2023,25(04):122-136.

17 参见刘宪权.ChatGPT 等生成式人工智能的刑事责任问题研究[J].现代法学,2023,45(04):110-125.

18 参见叶竹盛,林曼婷.生成式人工智能犯罪归责的开发者中心主义[J].深圳社会科学,2024,7(05):134-147.

19 光明网.杭州破获一起重大勒索病毒案.[EB/OL].(2023-12-

其次，GAI 技术可以辅助黑客生成规模化的钓鱼软件，使网络犯罪成本大幅降低且难以识别虚假信息。²⁰黑客利用计算机系统或软件中的漏洞，在用户点击钓鱼链接后获取用户的敏感信息并对系统和网络的未授权访问。如果任由该技术的肆意发展，恐怖分子将会编写恶意程序侵入国家机密部门的计算机信息系统，面临国家机密信息泄露风险。

2.侵犯财产安全的网络诈骗类风险

由于 Sora 等生成式人工智能能够基于文本自动生成逼真的视频，这可能导致深度伪造内容的制作门槛大幅降低，使得伪造他人的言行变得更为容易，进而可能被用于骗财、骗色等违法活动，无疑会加剧网络诈骗犯罪。在现实中各国都已经出现利用 GAI 进行诈骗的案例，涉案金额巨大，诈骗方式主要是利用 GAI 模拟受骗者的声音和样貌、制造虚假新闻或广告。如下表 2 中列举了 GAI 诈骗的典型案列。

表 2 生成式人工智能诈骗案例

时间	国家(或地区)	诈骗方式	事件
2019 年 3 月	英国	拟声	该诈骗者将自己的声音改变成当地一家能源公司老板的声音，要求一位高管向其汇款，通过电话骗取了 22 万欧元
2023 年 5 月	中国	换脸、拟声	福州市某科技公司法人代表郭先生在接到“好友”微信视频电话后，在 10 分钟内直接向对方账户汇款 430 万元
2024 年 1 月	中国香港	换脸、拟声	诈骗者通过搜集一家跨国公司英国高层在 YouTube 上的公开影像，再利用 AI“深度伪造”技术，给诈骗者换上公司高层的面部和声音，从香港分公司骗走 2 亿港元

29)[2024.09.07].<https://baijiahao.baidu.com/s?id=1786567474889043881&wfr=spider&for=pc>.

20 参见赵精武,王鑫,李大伟,等.ChatGPT: 挑战、发展与治理[J].北京航空航天大学学报(社会科学版),2023,36(02):188-192.

2024 年 3 月	加拿大	虚假新闻	一名多伦多男子在投资损失 1.7 万元后才意识到自己受骗，此后公开所谓的加密货币骗局，该骗局利用人工智能生成的新闻网站引诱受害者
------------	-----	------	--

3.传播淫秽、虚假内容类安全风险

随着该技术的普及和应用范围的扩大，一些不法分子开始寻找其中的技术漏洞，用于非法目的。GAI 技术的滥用不仅造成了受害者个人形象的损害，更严重侵犯了人们的隐私权和名誉权。GAI 技术的滥用助长了犯罪者传播虚假信息和淫秽信息等扰乱社会风气的行为。

首先，GAI 技术已被用来传播虚假消息，造成恶劣的社会影响。2023 年 4 月 25 日，甘肃省网安大队在日常网络巡查中发现，在百度旗下的百家号平台上，一篇名为《今晨甘肃一火车撞上修路工人致 9 人死亡》的文章引起了网安工作人员的注意。初步判断该篇新闻属于虚假消息。接着，民警展开了侦查工作，发现在同一时间段共有 21 个百家号发布了完全一样的文章，而这些文章的点击量累计已达 1.5 万余次。犯罪嫌疑人洪某通过使用 ChatGPT 对搜集到的新闻进行修饰后，将其上传至他购买的“百家号”上获得流量。²¹

其次，GAI 技术已被用来传谣淫秽内容，造成人格侵犯的不良后果。无独有偶，在今年 4 月初，北京警方接到报案：海淀区某小区内，有人利用 GAI 技术制作“一键脱衣”淫秽图片并且传播淫秽图片。北京市海淀区人民检察院对白某，以传谣淫秽物品牟利罪提起公诉。²²“AI 换脸”技术的运用，即使仅用于娱乐或恶搞照片制作，可能涉及侵犯他人的肖像权和名誉权，从而引起不可忽视的法律风险，此种行为可能导致严重后果，甚至涉及侮辱罪和诽谤罪。

二、美欧生成式人工智能安全风险监管与治理方案

为了规范生成式人工智能的健康发展，化解各类 GAI 研发与应用过程中带来的风险挑战，近年来各国开始探索 GAI 的监管，本文主要对美国以及欧盟的 GAI 的安全风险治理方案进行分析，具体如下表 3 所示。美国联邦层面在几年出台了人工智能的立法或行政命令，鲜少针对 AIGC 的具体监管，主要在数据隐私保护领域和版权领域的监管。但美国在近期出台了 GAI 的风险管理框架，就 GAI 的新风险提出建议。欧盟也没有 GAI 的具体监管法案，但在全球首部人工智能

21 光明网.火车撞上工人，9 人死亡？又是用“它”造谣！.[EB/OL].(2023-05-29)[2024.09.08].<https://baijiahao.baidu.com/s?id=1767185456696639734&wfr=spider&for=pc>.

22 光明网.用 AI 软件制作“一键脱衣”淫秽图片，北京一男子被提起公诉.[EB/OL].(2023-04-02).<https://baijiahao.baidu.com/s?id=1795225495823900297&wfr=spider&for=pc>.

法案《人工智能法》中涉及到 GAI 的监管级别及方法。

（一）美国 GAI 监管与治理方案

美国在 GAI 领域风险监管主要侧重于版权保护和数据隐私保护等。美国近期发布了首个 GAI 风险管理框架：《人工智能风险管理框架：生成式人工智能概况》（NIST AI 600-1）（简称《概况》），《概况》提出了由 GAI 引起的新风险，并对管理 GAI 特殊风险提出了建议行动。对于数据隐私的监管，《概况》指出建立透明的政策和流程，记录 GAI 应用程序的训练数据和生成数据的来源和历史，以提高数字内容透明度，同时平衡训练方法的专有性质；在退役 GAI 系统时考虑以下因素：数据保留要求；数据安全，例如遏制、协议、退役后的数据泄露；上游、下游或其他数据、物联网或 AI 系统之间的依赖性；使用开源数据或模型；用户对 GAI 功能的情感纠缠。对于运用 GAI 工具诈骗或传播淫秽、虚假内容的，《概况》指出从利益相关者社区获取输入，以确定不可接受的用途；制定政策和机制，防止 GAI 系统生成违法或有害内容；当面对危险、暴力或仇恨内容，需要重新评估组织的风险容忍度，以考虑不可接受的负面风险。对于知识产权的监管，《概况》指出运用维护文件保留政策，保留 GAI 的测试、评估、验证和验证历史，以及数字内容透明度方法，标准输出出处信息。²³

美国还将 GAI 领域的知识产权保护提出了一系列法案，在《内容来源保护和防止编辑和深度伪造媒体完整性法案》中，最具特色的措施是让生成内容所有者控制他们的作品，并为其内容设定使用条款和报酬；在此之前没有法律禁止删除、禁用或篡改 GAI 生成内容的来源信息，而该法案禁止任何人以上述方式干扰生成内容的来源信息。²⁴在《2024 生成式人工智能版权披露法案》中，该法案将要求生成式人工智能平台披露其在训练人工智能模型时对受版权保护的作品的使用情况，并具有追溯效力，为全球的创意界提供版权保障。²⁵美国田纳西州出台了首部生效的 GAI 音乐领域的版权监管法案，该法案通过民事补救措施保护每个人的声音不被 GAI 技术轻易复制，如未经授权使用或复制声音将属于犯罪。²⁶由上可知，美国在处理知识版权著作的问题上有着详细的法律举措，甚至细化到音乐创作领域，其举措值得我国借鉴。

OpenAI 曾在 2022 年 8 月发布 AI 对齐研究工作计划，对齐研究的方法侧重

23 NIST.AI Risk Management Framework.[EB/OL].(2024-07-

26)[2024.09.10].<https://www.nist.gov/itl/ai-risk-management-framework>.

24 us.commerce.senate.gov.Cantwell, Blackburn, Heinrich Introduce Legislation to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content.[EB/OL].(2024-07-11)[2024.09.11].<https://www.commerce.senate.gov/2024/7/cantwell-blackburn-heinrich-introduce-legislation-to-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content>.

25 congress.gov.Generative AI Copyright Disclosure Act of 2024.[EB/OL].(2024-04-09)[2024.09.22].<https://www.congress.gov/bill/118th-congress/house-bill/7913/text>.

26 tennessean.AI in music: TN Gov. Lee proposes bill to protect music artists.[EB/OL].(2024-01-10)[2024.09.22].<https://www.tennessean.com/story/entertainment/music/2024/01/10/ai-music-gov-bill-lee-bill-protect-artists-artificial-intelligence/72163690007/>.

于使用超级智能的 AI 系统设计可扩展的训练信号,让该信号与人类的意图一致。这三个支柱分别是使用人类反馈训练 AI 系统;训练 AI 系统以协助人类评估;训练 AI 系统进行对齐研究。²⁷而 OpenAI 也在 2023 年 7 月成立了“超级对齐”的新部门,意图在四年内找到让超智能 AI 系统实现价值对齐的路径,在今年 7 月,OpenAI 发布了他们在超级对齐方面的最新研究成果:通过证明者-验证者游戏提高大模型语言的可读性。该研究成果通过实证的方法得出训练强语言模型生成弱语言模型易于验证的文本,发现这种训练也使人类更容易对输出文本的可信度进行评估,因此有助于超智能模型的价值对齐。²⁸

(二) 欧盟 GAI 监管与治理方案

欧盟在全球首部人工智能法案《人工智能法》中规定了数据风险分级监管制度,将人工智能应用划分为低风险、有限风险、高风险和不可接受的风险四个等级,分类分级展开监管。如最高风险等级的被禁止 AI 实践,规定了如果利用操纵性或欺骗性技术扭曲人们的行为并损害知情决策,则禁止使用;特别是系统基于年龄、身份、人种、宗教等生物特征对个人或群体进行分类的系统将禁止。《人工智能法》对于高风险 AI 系统纳入了“过滤条款”,如果系统对人的安全或基本权利构成伤害风险,则属于高风险监管范畴。而其中针对 GAI 交互或生成内容的监管属于透明度风险,旨在带来假冒或欺骗类的风险,《人工智能法》中要求此类风险用户必须被告知与聊天机器人交互中,生成内容的部署者必须披露该内容是 GAI 生成的;在生成大量合成内容的 GAI 系统提供者必须提供可靠、可互操作、有效强大的技术方法,如水印来标记和检测。近期欧盟发布了《生成式人工智能与数据合规指南》(简称《指南》),该《指南》是欧盟发布的第一篇关于 GAI 数据合规的指南文件,其中几个问题对我们监管 GAI 具有参考意义。

第一个问题是设计、开发和验证 GAI 系统时,如何确定个人信息处理的合法性原则。《指南》指出,服务提供者需要遵从透明性原则,使被处理信息的个人明确肯定、自由、具体、知情等所有法律要求;同时数据最小化原则也可以帮助降低风险,数据控制者将个人数据的收集和处理限制在处理目的所必须的范围内,运用质量优于数量的 GAI 系统,避免数据的滥用;全生命周期遵从数据准确性原则,必须及时更新或删除数据。《指南》规定了数据保护官的任务,旨在协助控制者监控内部合规情况,就数据保护影响评估提供建议,充当数据主体和欧洲数据保护机构的联络人。

第二个问题是使用 GAI 系统时,如何确保公平处理并避免偏见。《指南》指出用于创建和训练模型的数据集必须保证充分、公平地反映现实;同时实施问责和监督机制,允许持续监测,防止发生并及时纠正对个人产生偏见。

27 OpenAI.Our approach to alignment research.[EB/OL].(2023-07-05)[2024.09.22].<https://openai.com/index/our-approach-to-alignment-research/>.

28 参见张熠天,王丹.欧盟人工智能立法的特点与影响[J].中国信息安全,2024,(04):72-75+80.

具体而言，欧盟的数据合规指南为保障数据安全提供了细致的规制措施，要求监管时采取与风险水平相适应的安全措施，持续监控和评估 GAI 系统。同时成立包括数据保护官在内的人工智能工作组，保持所有利益相关者之间的持续沟通。²⁹

表 3 主要国家/地区生成式人工智能监管方案

发布时间	发布国家/地区	文件类型	文件名称	主要内容
2024 年 10 月	美国	报告	《人工智能：生成式人工智能训练、开发和部署注意事项》	本报告概述了常见的生成式人工智能开发实践、这些技术的局限性及其易受攻击性，以及商业开发商收集、使用和存储生成式人工智能技术训练数据的流程。
2024 年 7 月	美国	风险管理框架	《人工智能风险管理框架：生成式人工智能概况》	该风险管理框架专为生成式人工智能设计，定义了由 GAI 技术引起的新风险和加剧的风险，并提出符合其目标和优先事项的 GAI 风险管理行动。
2024 年 7 月	美国	联邦法案（尚未正式生效）	《内容来源保护和防止编辑和深度伪造媒体完整性法案》	旨在解决人工智能生成内容的“深度伪造”泛滥的问题，保护知识产权人的利益。
2024 年 4 月	美国	联邦法案（尚未正式生效）	《2024 生成式人工智能版权披露法案》	该法案要求公司在使用受版权保护的作品来训练其生成式人工智能模型方面保持透明度。
2024 年 2 月	美国田纳西州	地方立法	《确保肖像、声音和图像安全法案》	该法案系全美首个针对 GAI 音乐创作的监管法案，禁止未经授权使用人工智能来模仿艺术家的声音。
2023 年 10 月	美国	行政命令	《关于安全、可靠、值得信赖地开发和和使用人工智能的行政命令》	1.要求建立通过制定检测人工智能生成内容和验证官方内容的标准和最佳实践。
				2.要求对 GAI 的硬件设施与软件系统所涉及的各类安全问题开展标准制定与安全评估。

29 EDPS.First EDPS Orientations for EUIs using Generative AI.[EB/OL].(2024-06-03)[2024.09.25].https://www.edps.europa.eu/data-protection/our-role-supervisor/first-edps-orientations-euis-using-generative-ai_en.

发布时间	发布国家/地区	文件类型	文件名称	主要内容
				3.强调规制日常生活中美国人使用人工智能产品时可能受到的欺诈、意外偏见、歧视、隐私侵犯和人工智能造成的其他侵害。
2023 年 5 月	美国	报告	《生成式人工智能和数据隐私：入门指南》	对 GAI 如何使用数据、可能引发的数据隐私风险等问题进行了分析,并提出了相应对策,《报告》认为应通过立法建立相应机制以应对 GAI 相关风险
2024 年 6 月	欧盟	指南	《生成式人工智能与数据合规指南》	该指南指出使用生成式人工智能系统有可能增加现有的安全风险或者形成新的安全风险,建议采取与风险水平相适应的措施解决安全风险。
2023 年 5 月	欧盟	法律	《人工智能法案》	其中,生成式人工智能系统一般属于有限风险的人工智能系统,需遵守最低限度的透明度义务,但可能会因其适用的领域和生成的内容而落入高风险人工智能系统的范畴。

三、中国开展生成式人工智能安全治理的实践路径

（一）坚持以人为本智能向善，促进创新和依法治理相结合的治理原则

生成式人工智能治理是全世界共同的课题，攸关全人类的命运，中国始终践行习近平总书记关于构建“人类命运共同体”的理念。2023 年 10 月习近平总书记提出的《全球人工智能治理倡议》（简称《倡议》）中给出的治理方案成为了中国参与生成式人工智能全球治理规则构建的关键一步。

在《倡议》中，生成式人工智能治理遵循的价值导向是“以人为本、智能向善”，确保生成式人工智能朝着全人类进步的方向发展，在发展的过程中坚持相互尊重、平等互利，各国在发展生成式人工智能的过程中享有平等的权利，弥合治理能力的差距。³⁰生成式人工智能在为各行各业增添新动能的同

30 中央网络安全和信息化委员会办公室.全球人工智能治理倡议.[EB/OL].(2023-10-18)[2024.07.10].https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

时，也带来了网络信息安全的新挑战，如何平衡发展和安全成为了生成式人工智能的重要课题。我国为了促进 GAI 的健康有序发展，制定了 GAI 领域首个监管文件：《生成式人工智能服务管理暂行办法》（以下简称《办法》），《办法》第三条强调：“国家坚持发展和安全并重、促进创新和依法治理相结合的原则”，³¹发展与安全是辩证统一的关系，安全为发展提出了红线底线，发展为安全治理提出了更高的要求。因此《办法》采用了包容审慎的监管原则，通过多方面的措施，包括应用、研发和基础设施，打造自主开放的创新生态。根据世界知识产权组织 7 月 3 日发布的报告显示，2014 年至 2023 年，中国生成式人工智专利申请数量超 3.8 万件，居世界第一。³²

（二）推行伦理先行、敏捷治理、分类分级监管的治理体系

生成式人工智能治理遵循的方式方法是“伦理先行+敏捷治理+分类分级监管”^①，中国 2023 年 3 月颁布的《人工智能伦理治理标准化指南》引导生成式人工智能健康有序发展，避免发生偏见或歧视行为，而当 GAI 使用过程中遇到问题时要及时响应，依据问题的级别进行管理，必要时需建立健全相关法律和规章进行处罚。

我国首部以风险治理为导向的《人工智能安全治理框架》（以下简称《框架》）1.0 版在 2024 年 9 月发布。《框架》首先便提出了人工智能安全治理的四个原则，遵循风险识别、技术应对、敏捷治理等治理逻辑，在此基础上客观、中立地描述了人工智能各生命周期的风险及其具体表现，从内生安全和应用安全两大方面落实各方主体责任，深入分析和阐释了技术和管理两大类的应对措施，积极倡导与各行业组织共建人工智能安全风险治理制度。³³《框架》为我国人工智能技术的研发和应用提供了框架性的技术指南，不带特定的政治或经济色彩的技术方案。

（三）出台规范和标准共治，加快推进人工智能治理法治化进程

早在 2021 年，针对互联网大数据应用，国家相继制定了《互联网信息服务算法推荐管理规定》和《互联网信息服务深度合成管理规定》，对网络信息服务提供者规定了责任义务，以确保在算法推荐和深度合成方面维护用户权益。随着 2022 年美国 OpenAI 公司研发的 ChatGPT 正式上线，中国陆续出现了文心一言等 GAI 产品，我国基于 GAI 研发及监管领域的空白，出台了全球首部 GAI 领域的规范文件《办法》。《办法》内容涵盖了生成式人工智能

31 中国政府网.生成式人工智能服务管理暂行办法.[EB/OL].(2023-07-

10)[2024.07.11].https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

32 新华社.世界知识产权组织：中国生成式人工智能专利申请量世界第一.[EB/OL].(2024-07-03).<https://baijiahao.baidu.com/s?id=1803546352861163538&wfr=spider&for=pc>.

33 cac.人工智能安全治理框架.[EB/OL].(2024-09-

09)[2024.09.10].<https://www.cac.gov.cn/rootimages/uploadimg/1727568303900999/1727568303900999.pdf>

服务的监管框架、适用范围、价值立场、责任主体以及境外服务提供者和外商投资的相关规定，为 GAI 服务的健康发展和规范运用提供了法律依据。

除了在 GAI 领域立法外，中国还出台了一系列配套标准辅助监管政策和立法，分别是《网络安全标准实践指南——生成式人工智能服务内容标识方法》、《生成式人工智能服务安全基本要求》、《网络安全技术——生成式人工智能预训练和优化训练数据安全规范》(征求意见稿)，分别在内容标识、服务提供者必须遵循的安全基本要求、预训练数据和优化训练数据的安全要求等领域补充《办法》。同时标准文件也根据具体实践不断更新细化，《网络安全技术——人工智能生成合成内容标识方法》强制性国家标准(征求意见稿)中不仅更新了交互场景界面的标识要求，还对不同类型的标识方式作出更细致的规定。

我国在生成式人工智能安全治理过程中应不断加快相应罪名的刑事立法进程，从修订和新增立法等角度多措并举加快推进 GAI 领域法治化进程，结合中国国情的实际并参考世界各国相关领域的监管与治理规则，进一步推动符合我国国家创新发展需要和兼顾安全风险治理需求的监管方案和标准细则落地，创建具有中国特色的生成式人工智能监管与治理路径。

中国 GAI 监管文件如表 4 所示。

表 4 中国生成式人工智能政策文件与法律法规

发布时间	文件类型	文件名称	主要内容
2024 年 9 月	国家标准	《网络安全技术——人工智能生成合成内容标识方法》强制性国家标准(征求意见稿)	该方法对比之前出台的《生成式人工智能服务内容标识方法》，细化了显性标识和隐性标识的标识方法。除了文本、图片、音频、视频等，增加了交互场景界面标识，覆盖更加全面。
2024 年 9 月	技术文件	《人工智能安全治理框架》1.0 版	该框架是《全球人工智能治理倡议》的细化，从内生和应用风险等角度分析人工智能风险。
2024 年 3 月	国家标准	《网络安全技术——生成式人工智能预训练和优化训练数据安全规范》(征求意见稿)	明确生 GAI 服务研发者在预训练和优化训练等训练数据处理活动中的数据来源等方面的管理要求。
2024 年 2 月	国家标准	《生成式人工智能服务安全基本要求》	我国第一部有关 GAI 服务安全性方面的技术性指导文件，对语料安全、模型安全、安全措施、词库/题库要求、安全评估等方面提出了具体规范和要求。

发布时间	文件类型	文件名称	主要内容
2023 年 11 月	共识文件	《发展负责任的生成式人工智能研究报告及共识文件》	该文件分析当前 GAI 的技术发展态势，提出技术给全球带来的机遇和挑战，全面梳理全球为发展人工智能作出的努力，总结提炼发展负责任的 GAI 十条共识。
2023 年 10 月	倡议文件	《全球人工智能治理倡议》	该倡议旨在推动人工智能健康发展，保障全社会、全人类共同利益，并提出以人为本、智能向善、相互尊重、平等互利等关键原则，引领人工智能全球治理。
2023 年 8 月	国家标准	《网络安全标准实践指南——生成式人工智能服务内容标识方法》	针对四类 GAI 生成内容给出了具有可操作性的、具体的内容标识方法。
2023 年 7 月	部门规章	《生成式人工智能服务管理暂行办法》	该办法是中国对 GAI 监管的专门立法，适用于利用生成式人工智能技术向中华人民共和国境内公众提供生成文本、图片、音频、视频等服务。
2022 年 11 月	部门规章	《互联网信息服务深度合成管理规定》	该规定是为了加强互联网信息服务深度合成管理，划定深度合成服务的“红线”。
2021 年 12 月	部门规章	《互联网信息服务算法推荐管理规定》	该规定针对在我国境内应用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法推荐技术提供互联网信息服务的行为提出了具体的治理措施。

（四）积极响应联合国倡议，参与人工智能全球治理体系构建

习近平总书记在第三届“一带一路”国际合作高峰论坛开幕式主旨演讲中提出的《全球人工智能治理倡议》为在人工智能技术发展中坚持相互尊重、平等互利的原则，维护各国平等发展和利用人工智能的权利做出了积极贡献，并已获得了联合国的认可和支持。中国之所以倡导国际共同治理人工智能，是适应人工智能技术的快速发展和应用早已呈现全球化的特点，国内外各大科技公司推出了各式各样的大模型，同时 AI 技术引发的风险挑战也已成为全球性挑战，亟待各国共同治理。

一直以来中国积极参与人工智能全球治理，在 2023 年举行的首届人工智能安全峰会上，包括中国、美国在内的 28 个国家联合签署了《布莱切利宣言》，中国提出的《全球人工智能倡议》多条内容被纳入《布莱切利宣言》。³⁴2023 年 11 月，世界互联网大会成立了人工智能工作组，编写了《发展负责任的生成式人工智能研究报告及共识文件》（以下简称《共识》）。《共识》总结了全球企业、智库及高校负责任发展人工智能的十条共识，给出了操作性极强的共识意见，如规则体系、追溯机制、沙盒治理等方法。³⁵2024 年 5 月，习近平总书记对法国进行国事访问，双方在《中法联合声明》的基础上达成了关于《人工智能全球治理的联合声明》，双方致力于人工智能的国际合作，加强双方人工智能治理框架和治理倡议的互操作性。³⁶在 7 月 1 日第 78 届联合国大会上，中国主张提出的“加强人工智能能力建设国际合作”决议获得了联合国大会以协商一致方式通过，140 多个国家参加联署，标志着中国积极引领人工智能领域全球治理的历史性成就。³⁷中国积极在人工智能领域贯彻人类命运共同体理念，2024 年 7 月 6 日在上海召开的 2024 世界人工智能大会上，发表了《人工智能全球治理上海宣言》，指出要促进人工智能发展的同时构建人工智能治理体系。³⁸以上参与并引领人工智能全球治理的举措，体现了中国负责任的大国担当，中国有信心进一步发挥引领作用，秉持共商共建共享的全球治理理念，推动人工智能全球治理体系改革。

四、生成式人工智能安全风险的应对策略

生成式人工智能的技术迭代速率远超传统科技革新的上限，其应用在各行各业转化为生产力，治理生成式人工智能的各类安全风险迫在眉睫，但在技术失控风险严峻的今天，现有的治理方案难以实现及时有效的监管，因此要结合国家制定的法律法规和技术指导文件，参考美国和欧盟监管和治理生成式人工智能相关风险的经验，针对性解决上述生命周期的内生风险和衍生风险。

（一）生成式人工智能的全球治理逻辑

1. 规则共商：以联合国为平台尽快达成 GAI 治理框架公约

34 GOV.UK.The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023.[EB/OL].(2023-11-01)[2024.09.25].<https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>.

35 新华网.以负责任态度推动生成式人工智能发展.[EB/OL].(2023-11-10)[2024.09.15].<http://www.xinhuanet.com/tech/20231110/677f4a2feacb4e3eb92074b7b9da595b/c.html>.

36 中国政府网.中华人民共和国和法兰西共和国关于人工智能和全球治理的联合声明.[EB/OL].(2024-05-07)[2024.09.20].https://www.gov.cn/yaowen/liebiao/202405/content_6949586.htm.

37 中国政府网.联大通过中国提出的加强人工智能能力建设国际合作决议.[EB/OL].(2024-07-02)[2024.09.20].https://www.gov.cn/yaowen/liebiao/202407/content_6960524.htm.

38 中国政府网.触摸 AI 发展新脉动——2024 世界人工智能大会观察.[EB/OL].(2023-07-10)[2024.09.20].https://www.gov.cn/yaowen/liebiao/202407/content_6961758.htm.

达成 GAI 治理框架的全球共识是一项艰巨但关键的任务。各国各方都已出台人工智能的安全监管或治理机制，但参考过多的治理机制可能会导致治理规则碎片化、治理效率下降，因此需要对现有的治理机制进行整合，借鉴优秀的治理规则和方案，与各国各方共同商讨求同存异，形成有效应对安全风险合力。首先，以联合国为平台达成 GAI 治理的基础框架公约，以规避全球大多数国家面临的 GAI 风险，这将需要强化联合国等多边国际机构之间的合作，协调它们在 GAI 领域的治理机制和规范，以达成聚合各国共识的全球治理框架公约。习近平总书记在联合国贸易和发展会议中致辞强调，要在联合国框架内加强人工智能规则治理，强调了联合国在治理体系构建中的主导作用。³⁹其次，推动各类商讨 GAI 治理和规则的国际峰会、论坛、进程等，以此为契机加速凝聚各方共识，并通过达成成果文件、共识文件、立场文件等“软法”的形式辅助框架公约规则的拟定和谈判。再次，根据各国或各地区重点关注的领域和社会发展诉求，达成适合各具体领域（如算法、数据、伦理等）的 GAI 安全治理多边条约，这样的个性化治理框架将有助于丰富和完善全球 GAI 治理体系，确保人工智能技术的发展既符合全球标准，又能够满足各区域特定的发展需求和期望。

2. 风险共治：促多利益方参与协商共论 GAI 风险应对方案

在治理 GAI 过程中，我们应该坚持多边主义原则，并联合多利益攸关方共同应对 GAI 风险挑战。习近平总书记在 2024 年 9 月 5 日会见联合国秘书长古特雷斯时就人工智能治理交流了看法，习近平总书记强调：中方将继续积极参与联合国各领域工作，支持联合国举办未来峰会，推动国际金融架构改革，加强人工智能治理，为世界和平和发展事业作出更大贡献。⁴⁰为了确保 GAI 的安全和健康发展，需要各国发挥各自的优势，实现合作共赢。鉴于不同国家在发展模式和认知上存在较大差异，我们应以推动人类文明进步为目标，不断加强国家间的对话与合作，构建一个互利共赢的新型国际关系。⁴¹首先，鼓励各国在联合国的框架内成立一个国际性的 GAI 治理机构。尤其是发展中国家在人工智能治理议题上的影响力和发言权要得到保护和加强。其次，应始终坚持在关于 GAI 全球治理的谈判和对话进程中保留一个开放多方参与的对话与合作机制，使各国际组织、学术科研机构、行业协会、非营利组织、民间团体以及科技企业等利益攸关方都能参与到全球 GAI 治理的过程中。最后，要在政府等监管机构的引导下，促进科技企业在 GAI 领域的创新和研发，并遵守相关法律法规，科技企业也可以及时上报部署应用时遇到的

39 人民网.习近平向联合国贸易和发展会议成立 60 周年庆祝活动开幕式发表视频致

辞.[EB/OL].(2024-06-13)[2024.09.20].<http://politics.people.com.cn/n1/2024/0613/c1024-40255324.html>.

40 人民网.习近平分别会见出席中非合作论坛北京峰会的部分非洲国家领导人和国际组织负责人.[EB/OL].(2024-09-06)[2024.09.20].<http://politics.people.com.cn/n1/2024/0906/c1024-40314190.html>.

41 参见赵申洪.全球人工智能治理的困境与出路[J].现代国际关系,2024,(04):116-137+140.

安全隐患，为政府制定法规和标准提供案例和样本，使二者实现良性互动。

3.标准共建：借行业力量推动 GAI 全球技术标准共商与制定

全球各大科技公司应尝试合作商讨行业标准和指南，为在 GAI 设计、开发和部署等阶段的开发者和服务商提供指导。我国就 GAI 治理配套了一系列技术标准文件，致力于 GAI 生命周期标准化监管，未来需要各国通力合作，共同制定全球通用的标准化指南，进一步增强技术的可靠性和安全性。首先，鼓励全球各大科技巨头、创新企业、研究机构以及监管机构共同协作，成立一个多边标准制定联盟。这个联盟的任务在于共同探讨 GAI 的设计原则、数据伦理、隐私保护、算法公正性以及内容安全等核心议题，确保技术进步的同时，社会价值观与法律法规得以尊重和体现。通过跨领域的知识共享与经验交流，联盟成员能够共同制定出既具指导性又具备实操性的标准与最佳实践指南，为 GAI 技术的健康发展奠定基石。其次，在共建标准的过程中，还需积极寻求国际标准组织的支持与认可，如 ISO、IEEE 等，推动形成国际共识，使得制定的标准具有普遍适用性和权威性。借助全球行业的协作力量，共同推进 GAI 技术标准的讨论与制定，能有效统一技术规范，实现解决方案的无缝对接与全球互通，促进技术创新的健康发展。

4.成果共享：享各国各方 GAI 降低风险与应急管理最佳做法

在为 GAI 各生命周期制定安全要求后，其技术的迭代及应用仍有极大的不确定性，此时可以参考各国降低风险与应急响应的优秀做法。习近平总书记在山东省召开党的企业和专家座谈会上就统筹发展和安全中特别指出建立人工智能安全监管制度，强调了监管环节的重要性。⁴²首先可以借鉴欧盟《人工智能法案》建立风险等级测试评估体系，依据风险发生的可能性和危害大小分类分级治理，再通过敏捷治理技术手段迅速响应 GAI 应用过程的各种变化，并及时改进优化。⁴³由于国内很多大模型借鉴了国外的基础模型作为接口直接嵌入，其中的数据算法不能完全透明，这时科技企业向公众提供 GAI 服务前需要向国家有关部门申请安全评估并备案，这个过程需要各国共同合作，积极披露 GAI 安全所需的必要算法，打造全球范围内可审查、可追溯的 GAI 系统。其次，各国政府应建立人工智能高级别对话机制，定期交换交流各方在 GAI 应急管理和响应环节的典型案例及最佳做法，形成 GAI 全球风险管控的合力。

（二）生成式人工智能治理的具体策略

生成式人工智能治理策略应采取事前审查、事中监控、事后追责的手段，

42 求是网.关于《中共中央关于进一步全面深化改革、推进中国式现代化的决定》的说明.[EB/OL].(2024-08-15)[2024.09.25].http://www.qstheory.cn/dukan/qs/2024-08/15/c_1130192011.htm.

43 参见郭小东.大模型时代全球人工智能治理的挑战与中国方案[J/OL].科学学研究,1-15[2024-09-26].

才能实现 GAI 生命周期各阶段有效治理闭环。要在事前即 GAI 研发和设计阶段进行严格的数据审查和规划，确保数据来源合法、算法公正，确保技术安全；要在事中即 GAI 部署和使用阶段实时监控和管理，及时发现并敏捷响应各类问题；在事后即 GAI 使用后出现问题应进行责任认定并处罚和改进，根据一系列案例完善法规和标准。

1.事前严格审查和规划，确保数据合法、算法公正

为了确保生成式人工智能的数据来源合法合规并保护个人隐私，可以采取审查和技术措施。首先，政府相关部门、企业的合规部门或独立的合规审计机构应进行数据来源审查，确保数据来自合法渠道并符合相关法律法规。其次，对于需要授权的敏感数据，如个人数据或受版权保护的作品，需要与数据权利人签订授权协议，明确授权范围和使用方式以及使用期限。为了保护个人隐私，需要对涉及个人敏感信息的数据进行脱敏处理，例如将用户的姓名、电话号码等个人隐私信息替换成代码；也可进行匿名化处理，例如删除或替换个人身份信息，使数据无法与个人关联。

对于 GAI 带来的数据泄露挑战，可以采取数据安全防护和隐私计算等技术手段。首先，构建一个健全的数据安全管理架构，并运用如加密技术、访问权限控制等科技方法，以此来有力保障数据的安全性，有效抵御数据外泄及未经授权的修改风险。在 GAI 底层框架中嵌入联邦计算技术，该机器学习机制可以保障个人隐私数据或敏感信息保留在各自节点中，节点只交换模型更新而不交换数据本身。

在解决 GAI 中数据偏见和算法歧视导致的价值对齐问题方面，首先是数据审核与清洗，在数据收集阶段就要严格把关，仔细审查并清理数据，除掉可能有偏差或错误的部分，同时要保证数据具有多样性和代表性，能够涵盖各种不同的群体和场景，这样才能让数据基础更公正客观。其次，算法透明度和算法披露至关重要，开发人员应让算法的工作原理和决策过程清晰可见，方便理解与审查，以便察觉可能隐藏的偏见和歧视问题。再者，组建多样化的开发团队意义重大，团队成员拥有不同的背景、文化和观点，能避免单一视角产生的偏见，在设计和开发时进行多维度的思考与评估。《倡议》中指出了 GAI 应该遵守的全人类共同价值，“和平、发展、公平、正义、民主、自由”⁴⁴，考虑到不同国家制度文化带来的差异，应当强调尊重与包容。对于 GAI，应在其训练数据和算法设计中纳入来自世界各地的多元文化样本，以确保其能够理解和适应不同文化背景下的价值观念和需求，减少因文化差异导致的误解和偏见。

44 中央网络安全和信息化委员会办公室.全球人工智能治理倡议.[EB/OL].[2023-10-18].https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

2.事中实时监测和管理，识别并敏捷处理各类风险

为了防止生成式人工智能产生煽动性内容，应采取建立安全预警系统和内容审查机制等方法。首先，开发安全预警系统，监测模型输出中的异常情况，及时发现并处理潜在的煽动性内容。其次，建立自动内容审查机制，利用自然语言处理技术识别和过滤掉违规内容。除了使用技术手段，还可以运用相关技术人员进行审查，例如中国 2021 年发布的《人工智能训练师国家职业技能标准》，出现了人工智能训练师这一职业，将会助力可疑内容的人工审核。⁴⁵

面对 GAI 生成的内容造成知识产权机著作权侵犯的风险，首先要在大模型内部嵌入侵权检测技术，如文本匹配、图像识别等，防止生成侵权内容，并建立侵权处理机制。一旦发现侵权行为，立即采取措施删除或替换侵权内容。还可以利用技术手段，如文本嵌入技术，将文本转换文件特征变量；水印技术可以用来标记和认证 GAI 生成内容；区块链技术，能记录作品的创作和传播过程，有效保障作品的可追溯性等保护作品的版权。关于侵犯版权的归责问题，可以延伸扩展责任主体的方法，明确模型开发者和使用者的责任外，还要模型开发者进行风险追踪，在发生侵权行为后协助观察定责。⁴⁶

为了有效防范不法分子应用 GAI 技术对计算机系统的潜在破坏，要从安全管理机制和隔离监测系统环境上着力。首先，强化访问控制与权限管理机制，严格限制信息系统的操作权限，确保它无法触及或操控核心系统资源。其次，构建安全隔离的运行环境，利用沙盒技术将生成式 GAI 与信息系统隔离开来，限制其对外部环境的任何直接影响。同时，实施严格的输入输出监控策略，通过对信息系统接收到的数据进行细致的过滤与审查，来阻挡恶意指令的渗透，并审查输出内容，以剔除可能引发危害的信息。此外，建立行为监控系统并配合异常检测算法，实现对信息系统行为的实时监控，一旦监测到异常活动，立即启动响应程序进行调查与干预。如果不法分子利用钓鱼软件或插件等系统不易识别的方式侵入计算机系统，要引入签名与认证机制，为正规渠道生成的内容附加数字签名，便于用户验证信息的可信度。

当涉及到生成淫秽、虚假信息等不法内容时，GAI 可以识别并拒绝回答用户的提问；当不法分子设定情境诱导 GAI 生成淫秽内容时，GAI 无法识别，这时要设计出新的识别技术进行内容过滤与监控，自动识别并剔除文本、图像及视频中的不良信息，包括但不限于淫秽、暴力及虚假内容；也可以开发 GAI 技术追溯生成信息的传播路径，评估信息的真实度。最后，针对含有成

45 参见沈芳君.生成式人工智能的风险与治理——兼论如何打破“科林格里奇困境”[J].浙江大学学报(人文社会科学版),2024,54(06):73-91.

46 参见秦瑞标.大模型时代生成式人工智能治理的逻辑更新与回应路径——以《生成式人工智能服务管理暂行办法》为视角[J].时代法学,2024,22(03):28-41.

人内容的特定平台，实施严格的年龄验证措施，并实行内容分级管理制度，严格限制未成年人接触任何不适宜的信息。

3.事后持续监督和改进，完善法律法规和伦理共治

无论在 GAI 的设计阶段、开发阶段还是部署应用阶段，都需要对数据使用情况进行监督审计，因此要建立数据合规监管机制，对数据使用进行监督和审计，并遵守相关的法律法规，参考《数据安全法》和《个人信息保护法》等，以确保数据的合规使用。除了 GAI 服务提供者遵守相关规定配合审查和监督外，服务使用者也需要合理使用 GAI 技术，监管部门可以构建一个高效的用户举报与反馈系统，激励用户积极参与监督，及时举报不当内容。此外，建立定期的后门检测与模型审核机制，及时发现并清除潜在的恶意操纵痕迹或安全隐患，确保 AI 模型的纯洁度与长期安全性，多维度防护，共同构建一个更加安全的网络环境。

上文指出的运用 GAI 生成的淫秽或虚假新闻、诈骗等行文已经触犯法律，但使用 GAI 进行犯罪的界定和处罚还未完善。因此要随着 GAI 技术的快速发展，定期评估和更新相关法律法规，确保法律能够适应新出现的犯罪形式。此外应加强与执法机构的合作，共同打击利用人工智能进行的犯罪活动，例如，与警方和检方合作，调查和起诉犯罪行为，也可与国际组织和外国政府开展合作，应对跨国犯罪威胁。最后，在用户协议中除了加入禁止使用人工智能进行犯罪行为的条款，还要对违规用户进行处罚，例如，暂停或终止其使用平台的权利，必要时向政府有关监管部门及时上报违规操作。

作为预防犯罪的技术手段，伦理治理同样重要。不仅要以伦理治理的方式强化研发者的主体责任，⁴⁷更要强调使用者使用 GAI 过程中的伦理指导。因此要为人工智能的开发和应用制定明确的伦理准则，引导开发者和社会负责任地使用这项技术。2023 年 9 月，中国科技部等十部门发布《科技伦理审查办法（试行）》中明确指出相关责任主体要设立科技伦理委员会，对人工智能应用的伦理问题进行审查和指导。

结 语

美国 OpenAI 公司在 2024 年推出的 Sora 视频生成工具和 GPT-4o 全能聊天机器人，标志着 GAI 技术的新发展。这些技术虽推动了内容创作的高水平，但也带来了安全风险，如数据隐私侵犯、虚假内容生成、知识产权侵犯等。但这些技术的滥用可能对社会秩序和国家安全构成威胁，如煽动性内容的生成和深度伪造技术的应用。此外，GAI 的算法偏见和文化差异可能导致价值观念的冲突和误解。为应对这些风险，需要在数据训练和内容生成阶段采取合规审查、数据安全防护、内容审查机制等措施。同时，应加强伦理治理，

47 参见支振锋,刘佳琨.伦理先行：生成式人工智能的治理策略[J].云南社会科学,2024,(04):60-71.

明确责任主体，更新法律法规以适应新的犯罪形式，并与执法机构合作打击 GAI 相关的犯罪活动。在多元文化和价值观念上，应推动包容和尊重，确保 GAI 的全球价值对齐，并建立灵活机制以适应社会价值的动态变化。中国一直并将继续积极参与并推动全球在生成式人工智能治理领域的合作，为全球 GAI 的健康、安全、可持续发展贡献智慧和力量。

Research on the Path of Building Global Governance Rules for Generative Artificial Intelligence

JuMeng, ZhouJing

(Hubei University of Technology, Hubei 430068, China)

Abstract: Generative artificial intelligence and the security risks and challenges it brings have become a topic of close attention for countries around the world. This article tries to put forward a response strategy for the new type of generative AI risks that have emerged so far, and provides suggestions and opinions for our country's future participation in the global governance of generative AI security risks and specific response strategies, so as to provide reference for our country to form a set of global governance programmes for AI in line with the interests of its own development. The article takes the new type of security risks and challenges triggered by generative AI as the research object, closely combines its technical characteristics, and conducts in-depth analyses around the risks and cases arising from the various life cycles of generative AI design, research and development, deployment, and maintenance, and divides them into endogenous security risks and derivative security risks. Comparative analysis of the United States, the European Union and other subjects and China's current practice and practice on generative AI security risk governance policy legislation and response programmes. This paper puts forward the logic of global governance of generative AI from four aspects, such as rule co-negotiation, risk co-management, standard co-construction, and result sharing, to implement risk governance to the pre, during and post, and to achieve the closed loop of generative AI governance.

Key words: Generative Artificial Intelligence; Security Risks; Value Alignment; Global Governance; Response Strategies

参考文献:

- [1] 环球网.盘点 | 2024 年全球有哪些重要选举? .[EB/OL].(2023-12-28)[2024.07.03].<https://world.huanqiu.com/article/4FwNMCjB0gI>.
- [2] 陈磊,李雅静.人工智能系统安全综述[J].信息通信技术与政策,2021,47(08):56-63.
- [3] 单勇,王熠.来自“世界3”的知识欺诈:生成式人工智能的刑事风险应对[J].西南政法大学学报,2023,25(04):70-85.
- [4] 吕垚瑶.ChatGPT 类生成式人工智能技术迭代的犯罪风险及其治理[J].中国刑警学院学报,2024,(01):78-87.DOI:10.14060/j.issn.2095-7939.2024.01.008.
- [5] openai.com.Privacypolicy.[EB/OL].(2023-11-14)[2024.07.03].<https://openai.com/policies/privacy-policy/>.
- [6] 刘杰.ChatGPT 类生成式人工智能产品提供者之刑事责任[J].法治研究,2024,(02):61-71.
- [7] 新浪网.ChatGPT 爆重大漏洞:可能泄露 1.2%用户数据, AI 安全引人震惊.[EB/OL].(2023-03-27)[2024.07.18].https://k.sina.com.cn/article_1691724254_64d5a9de027010ktr.html.
- [8] 中国青年网.20 天三起泄密事故,三星考虑禁用 ChatGPT? .[EB/OL].(2023-04-04)[2024.07.24].<https://baijiahao.baidu.com/s?id=1762222629170377087&wfr=spider&for=pc>.
- [9] 刘瑶瑶,梁永霞,李正风.生成式人工智能与科研伦理:变革、挑战与展望[J].科学观察,2024,19(04):1-8.
- [10] 环球网.《环球时报》记者在第二届网络安全(天津)论坛听专家介绍:如何应对 AI 发展中的新型犯罪.[EB/OL].(2024-09-05)[2024.09.06].<https://3w.huanqiu.com/a/de583b/4JJB4r2kBeJ>.
- [11] 孟令宇.从算法偏见到算法歧视:算法歧视的责任问题探究[J].东北大学学报(社会科学版),2022,24(01):1-9.
- [12] 褚建勋,夏小草,吴熙凡.中国与欧美治理人工智能价值对齐问题的实践及其互鉴[J/OL].南京邮电大学学报(社会科学版),1-15.
- [13] 光明网.利用 AI 造谣“贵州女婴被遗弃街头”! .[EB/OL].(2023-03-11)[2024.09.06].警方抓获 2 人.<https://baijiahao.baidu.com/s?id=1793234814679829576&wfr=spider&for=pc>.
- [14] 刘刚,刘钊.深度合成技术处理个人信息“合理范围”界定路径研究[J].科学与社会,2024,14(01):85-101.
- [15] 盛浩.生成式人工智能的犯罪风险及刑法规制[J].西南政法大学学报,2023,25(04):122-136.
- [16] 刘宪权.ChatGPT 等生成式人工智能的刑事责任问题研究[J].现代法学,2023,45(04):110-125.
- [17] 叶竹盛,林曼婷.生成式人工智能犯罪归责的开发者中心主义[J].深圳社会科学,2024,7(05):134-147.
- [18] 光明网.杭州破获一起重大勒索病毒案.[EB/OL].(2023-12-29)[2024.09.07].<https://baijiahao.baidu.com/s?id=1786567474889043881&wfr=spider&for=pc>.
- [19] 赵精武,王鑫,李大伟,等.ChatGPT:挑战、发展与治理[J].北京航空航天大学学报(社会科学版),2023,36(02):188-192.

- [20] 光明网.火车撞上工人，9 人死亡？又是用“它”造谣！.[EB/OL].(2023-05-29)[2024.09.08].<https://baijiahao.baidu.com/s?id=1767185456696639734&wfr=spider&for=pc>.
- [21] 光明网.用 AI 软件制作“一键脱衣”淫秽图片，北京一男子被提起公诉.[EB/OL].(2023-04-02).<https://baijiahao.baidu.com/s?id=1795225495823900297&wfr=spider&for=pc>.
- [22] NIST.AI Risk Management Framework.[EB/OL].(2024-07-26)[2024.09.10].<https://www.nist.gov/itl/ai-risk-management-framework>.
- [23] us.commerce.senate.gov.Cantwell, Blackburn, Heinrich Introduce Legislation to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content.[EB/OL].(2024-07-11)[2024.09.11].<https://www.commerce.senate.gov/2024/7/cantwell-blackburn-heinrich-introduce-legislation-to-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content>.
- [24] congress.gov.Generative AI Copyright Disclosure Act of 2024.[EB/OL].(2024-04-09)[2024.09.22].<https://www.congress.gov/bill/118th-congress/house-bill/7913/text>.
- [25] tennessean.AI in music: TN Gov. Lee proposes bill to protect music artists.[EB/OL].(2024-01-10)[2024.09.22].<https://www.tennessean.com/story/entertainment/music/2024/01/10/ai-music-gov-bill-lee-bill-protect-artists-artificial-intelligence/7216369007/>.
- [26] OpenAI.Our approach to alignment research.[EB/OL].(2023-07-05)[2024.09.22].<https://openai.com/index/our-approach-to-alignment-research/>.
- [27] 张熠天,王丹.欧盟人工智能立法的特点与影响[J].中国信息安全,2024,(04):72-75+80.
- [28] EDPS.First EDPS Orientations for EUIs using Generative AI.[EB/OL].(2024-06-03)[2024.09.25].https://www.edps.europa.eu/data-protection/our-role-supervisor/first-edps-orientations-euis-using-generative-ai_en.
- [29] 中央网络安全和信息化委员会办公室.全球人工智能治理倡议.[EB/OL].(2023-10-18)[2024.07.10].https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.
- [30] 中国政府网.生成式人工智能服务管理暂行办法.[EB/OL].(2023-07-10)[2024.07.11].https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.
- [31] 新华社.世界知识产权组织：中国生成式人工智能专利申请量世界第一.[EB/OL].(2024-07-03).<https://baijiahao.baidu.com/s?id=1803546352861163538&wfr=spider&for=pc>.
- [32] cac.人工智能安全治理框架.[EB/OL].(2024-09-09)[2024.09.10].<https://www.cac.gov.cn/rootimages/uploadimg/1727568303900999/1727568303900999.pdf>
- [33] GOV.UK.The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023.[EB/OL].(2023-11-01)[2024.09.25].<https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>
- [34] 新华网.以负责任态度推动生成式人工智能发展.[EB/OL].(2023-11-10)[2024.09.15].<http://www.xinhuanet.com/tech/20231110/677f4a2feacb4e3eb92074b7b9da595b/c.html>.
- [35] 中国政府网.中华人民共和国和法兰西共和国关于人工智能和全球治理的联合声明.[EB/OL].(2024-05-07)[2024.09.20].https://www.gov.cn/yaowen/liebiao/202405/content_6949586.htm.

- [36] 中国政府网.联大通过中国提出的加强人工智能能力建设国际合作决议.[EB/OL].(2024-07-02)[2024.09.20].https://www.gov.cn/yaowen/liebiao/202407/content_6960524.htm.
- [37] 中国政府网.触摸 AI 发展新脉动——2024 世界人工智能大会观察.[EB/OL].(2023-07-10)[2024.09.20].https://www.gov.cn/yaowen/liebiao/202407/content_6961758.htm.
- [38] 人民网.习近平向联合国贸易和发展会议成立 60 周年庆祝活动开幕式发表视频致辞.[EB/OL].(2024-06-13)[2024.09.20].<http://politics.people.com.cn/n1/2024/0613/c1024-40255324.html>.
- [39] 人民网.习近平分别会见出席中非合作论坛北京峰会的部分非洲国家领导人和国际组织负责人.[EB/OL].(2024-09-06)[2024.09.20].<http://politics.people.com.cn/n1/2024/0906/c1024-40314190.html>.
- [40] 赵申洪.全球人工智能治理的困境与出路[J].现代国际关系,2024,(04):116-137+140.
- [41] 求是网.关于《中共中央关于进一步全面深化改革、推进中国式现代化的决定》的说明.[EB/OL].(2024-08-15)[2024.09.25].http://www.qstheory.cn/dukan/qs/2024-08/15/c_1130192011.htm.
- [42] 郭小东.大模型时代全球人工智能治理的挑战与中国方案[J/OL].科学学研究,1-15[2024-09-26].
- [43] 沈芳君.生成式人工智能的风险与治理——兼论如何打破“科林格里奇困境”[J].浙江大学学报(人文社会科学版),2024,54(06):73-91.
- [44] 秦瑞标.大模型时代生成式人工智能治理的逻辑更新与回应路径——以《生成式人工智能服务管理暂行办法》为视角[J].时代法学,2024,22(03):28-41.
- [45] 支振锋,刘佳琨.伦理先行:生成式人工智能的治理策略[J].云南社会科学,2024,(04):60-71.

网络攻击下自卫权行使困境与中国应对

王韵涵*

摘要：无论是遭受“先发制人”武力攻击还是“后发制人”抵抗侵略，对于传统战争形态而言，都具有一定的可辨识性和可操作性，并不存在过多争议。但对网络攻击这种隐蔽性高、无法预防、难于溯源的非传统安全威胁，国家行使自卫权必然面临诸多困境，比如预先性自卫的争议、集体自卫的问题。其中，在预先性自卫问题上，“坚持防御、自卫、后发制人”原则完全符合我国作为安理会常任理事国和世界主要大国的一贯的内政外交政策。在行使集体自卫权上，应尝试摒弃冷战状态下的集团“硬”对抗方式，以更加灵活的姿态和区别于军事联盟武力自卫的形式。在立法欠缺的情况下，我国应充当规制网络攻击范围，不断完善网络安全国内国际立法等“软”合作形式，加强立法层面的网络震慑与技术层面的保护与人才培养等，努力构建网络空间命运共同体。

关键词：网络攻击，自卫权，网络安全，网络空间命运共同体

一、问题的提出

网络攻击事件呈现高发态势，其传播路径的隐蔽性、攻击载体的多元化以及危害后果的扩散性，使得网络攻击的防控面临严峻挑战，这一态势已对全球各国的关键基础设施、经济社会稳定及国家安全构成了深层次的系统性风险。¹高级持续性威胁(APT 攻击)，常用于国家间网络攻击行动，被认为是地缘政治的延伸，甚至是战争和冲突的一部分。包括政府部门、委内瑞拉水电站和乌克兰电网在内的能源、交通、金融和医疗等关键信息基础设施领域遭到网络攻击²。因此，网络攻击高发的形势要求各国建立和完善相应的治理体系，发挥国际法的保障性作用³。然而，由于互联网的匿名性、即时性等特殊技术属性，大量问题在法律层面上仍难以得到明确答案⁴。特别是当研究集中在国家层面的网络攻击时，一个国家或国际社会如何选择应对此类事件值得讨论⁵。鉴于跨境网络

* 王韵涵，哈尔滨工程大学人文社会科学学院，副教授。本文系黑龙江省哲学社会科学研究规划项目“网络空间冲突的国际法律适用与治理研究”(24FXC005)阶段性成果。

¹ 参见王岩：《适用于网络空间行动的国际法》，《武大国际法评论》2019 第 3 期，第 139-157 页。

² 参见梁思雨，孔华锋：《跨境网络攻击治理及中国方案》，《计算机应用与软件》2019 年第 12 期，第 305-308 页。

³ 参见刘碧琦：《论国际法在网络空间适用的依据和正当性》，《理论月刊》2020 第 8 期，第 109-110 页。

⁴ 参见蔡高强，焦园博：《网络武装冲突”的国际法阐释》，《湘潭大学学报(哲学社会科学版)》，2017 年第 3 期，第 16-23 页。

⁵ 余民才著：《国际法上自卫权实施机制》，中国人民大学出版社 2014 年版，第 241 页。

攻击已经成为一个需要面对和解决的全球性问题，中国作为一个网络大国，很难独善其身⁶。习近平主席在其演讲中表明了中国对于国际网络安全问题的态度：在相互尊重、相互信任的基础上加强对话合作，变革治理体系，共建和平开放的网络空间，平衡大多数国家的利益⁷。本文重点研究了网络攻击中自卫权行使困境，在分析了网络攻击下自卫权立法欠缺的基础上，得出中国参与网络空间规则制定的策略选择，有助于建立防御与威慑并重的跨境网络攻击综合治理体系。

二、网络攻击下自卫权行使困境

针对网络空间是否存在国家主权，联合国、欧盟、上海合作组织等国际组织⁸、许多国家以及多数学者都给予了肯定的回答。联合国《建立网络安全和保护关键信息基础设施的全球文化》《努力建立网络安全文化和评估各国保护关键信息基础设施》等相关决议⁹，推动各国就信息化时代的网络安全和信息保护等达成一致。针对网络空间犯罪及网络恐怖主义，联合国大会决议“呼吁会员国多边审议信息安全领域的现有或潜在威胁以及遏制威胁所能够采取的措施¹⁰。”欧盟 2001 年通过《网络犯罪公约》，要求各缔约国政府通过立法、行政等多种手段保护信息安全，打击网络犯罪¹¹。2009 年，上海合作组织元首峰会达成《叶卡捷琳堡宣言》，该宣言第 7 条指出，“在国际安全体系中，保障国际信息安全十分迫切。”针对网络攻击能否行使自卫权，美国等信息网络发达的国家的态度十分明确。美国《网络空间行动战略》（2011 年版）明确，“必要时，美国将如同对待任何其他对我们国家的威胁一样，对网络空间的敌意行为作出反应。所有国家均享有固有的自卫权利……我们保留在必要时、并遵守适用国际法，使用一切必要的手段。”¹²

对行使自卫权的某些限制。《宪章》第 51 条是国家行使自卫权的国际法权利根源所在，既体现出《宪章》对国家主权项下自保权这种天然权利的认可与尊重，也赋予受到非法武力攻击国家的救济手段。但出于“免后世再遭今代人类两度身历惨不堪言之战祸，……促成大自由中之社会进步及较善之民生”的目的¹³，《宪章》对自卫权的行使也作出一些必要的限制。基于《宪章》第 51 条及

⁶ 参见蒋丽，张小兰，徐飞彪：《国际网络安全合作的困境与出路》，《现代国际关系》2013 年第 9 期，第 52-58 页。

⁷ 《习近平就共同构建网络空间命运共同体提出 5 点主张》新华网：http://www.xinhuanet.com/world/2015-12/16/c_128536396.htm，访问日期：2024 年 11 月 21 日。

⁸ 参见《塔林手册（2.0 版）》规则 4 对主权的侵犯（1）

⁹ See G.A.Res.58./199, U.N.Doc.No.A/RES/58/199(Jan.30, 2004); G.A.Res.64./211, U.N.Doc.No.A/RES/64/211 (Mar.17,2010).

¹⁰ See A /RES/63/37,2009-01-09.

¹¹ 参见郑春荣，倪晓姝：《欧盟网络安全战略及中欧合作》，《同济大学学报》，2020 年第 4 期。

¹² The white house,internationnal strategy for cyberspace: prosperity, security, and openness in a nerworked world, may 2011.

¹³ 1945 年《联合国宪章》。

关联条款：一是行使自卫权的主体须是联合国框架体系内的“会员国”，而非其他的组织、团体或者个人。布托加利曾论证美国参与朝鲜战争不属于集体自卫，依据就是1950年韩国还不是联合国“会员国”，因此不享有《宪章》规定的会员国所享有的自卫权¹⁴。学者王献枢也认为，自卫权是《宪章》赋予会员国的权利¹⁵。但也有学者提出，从习惯国际法和《宪章》第2条第6款的角度，无论一国是否为联合国会员国，都享有自卫的权利¹⁶。二是行使自卫权的前提是在受到武力攻击的状态下。¹⁷重大商业欺诈行为，违法操纵股市巨幅波动，或者严重暴力刑事犯罪等行为，也可能对一国在经济、金融、安全等方面造成某种程度的损害，但其不属于武力攻击范畴。¹⁸三是自卫权行使限定于受武力攻击的时间状态或者性质状况。¹⁹基于《宪章》中文文本表述，这个状态既非“过去时”，也非“将来时”，既非“前期预谋”，也非“结束已久”。“《宪章》的实践证明，当攻击正在进行或迫在眉睫时，以自卫使用武力是阻止它所必要的”²⁰。四是自卫权的行使需在“安全理事会采取必要办法，以维持国际和平及安全以前”，如果安理会已就是非对错作出评判，作为会员国则理论上应当以遵守安理会决议为先。最后，考虑受到非法武力攻击，以及会员国行使自卫权均触及国际秩序的重大情势变更，因此，会员国“因行使此项自卫权而采取之办法，应向安全理事会报告”²¹，该要求是“安理会采取必要办法”的救济性措施。在安理会对国际和平具有权威性评判的现代国际社会的共同认知下，无论具有武力攻击性质的网络攻击在前或是在后，安理会都有权就恢复和平作出必要反应。当事国尽速向安理会报告²²，也是行使自卫权的国家实施合法自卫的一种路径和宣示。

行使自卫权面临的不确定因素²³。对照《宪章》第51条的某些限制，国家面临网络攻击欲采取自卫措施之前，将会因网络攻击的特殊性而面临很多不确定性：一是网络攻击者身份认定。通过互联网进行的攻击行为，由于根服务器在美、欧、日等少数几个西方国家，在特殊的网络结构下，攻击者无论是自然

¹⁴ See B.Boutros-ghali,“L'intervention Americaine En Coree Et Le Droit Des Nations-unies,” (caire,1951), pp. 23.

¹⁵ 参见王献枢：《伊拉克战争的非法性》，《法学评论》2004年第1期，第11页。

¹⁶ 1945年《联合国宪章》第2条第6款。

¹⁷ See Shackelford, and Scott J., Managing Cyber Attacks in International Law, Business (Cambridge: Cambridge University Press, 2014), pp. 285-306.

¹⁸ [泰]克里安沙克·基蒂猜沙里：《网络空间国际公法》，程乐、裴佳敏、王敏译，中国民主法治出版社2020版，第135页。

¹⁹ See Melzer, Cyberwarfare and International Law, pp. 13-16.

²⁰ See Stanimir A. Alexandrov, “Self-defense against the use of force in international law,” The Cambridge Law Journal (1997), p. 163.

²¹ 1945年《联合国宪章》第51条。

²² 参见李伯军：《对美国“先发制人”军事战略的对外政策与国际法分析》，《政法论丛》2003年第6版，第24-29页。

²³ “Recent Cyber Events and Possible Implications for Armed Forces,” CCDCOE Researchers, last modified April 21, 2021. https://www.ccdcoe.org/uploads/2021/01/Recent-Cyber-Events-and-Possible-Implications-for-Armed-Forces-8-January-2021_Final-1.pdf.

人、组织、团体还是国家，都可以轻而易举地隐瞒真实身份，甚至将身份伪造嫁祸于人。二是网络攻击时间认定。从有关案例看，网络病毒（或称攻击代码）输入网络的时间，并不就是攻击实施的时间，很多时候尚处于暂时休眠的待激发状态，攻击者往往要为自己留下足够的防止追踪的时间，网络病毒则在等待攻击者设定的时间甚至是随机选择的时间²⁴。三是受攻击程度认定。网络攻击最终达到的效果可能并不在攻击者预计掌控之中。在攻击的过程中，有的被其他黑客修改升级，有的因为网络算法，自动进行升级、演化、变异，有的可能被病毒软件及时发现消杀，没有起到攻击者所期待的最佳效果²⁵。四是最终责任认定。谁来确定攻击者的责任，采取何种手段进行确认，都存在许多道德、伦理、技术、法律方面的困难²⁶。

综上，从现代国际秩序及《宪章》禁止威胁使用和使用武力，以及《宪章》相关条款对行使自卫权所作的某些限制，到网络攻击有别于传统武力攻击的特殊性复杂性，在这种新型攻击（防御）形态面前，国家行使自卫权必然面临诸多盲点、争点和难点。

（一）网络攻击下采取预先性自卫的争议

1. 预先性自卫的实践基础

自卫权是一种与国家存续状态天然共生的权利。从奴隶制国家建立，甚至之前奴隶部落的形成，应对外敌最直接有效的手段就是奋起自卫。因此，自卫权是在人类历史长河中经过长期实践形成的一种习惯法原则。二战以后，以成立联合国为标志的新的国际秩序限制了战争权，但对自卫权也给出了基本判断，即国家不分大小、强弱一律平等，纠纷冲突应当通过采取和平手段加以解决，不得威胁使用或者诉诸使用武力，国家拥有应对外部威胁或武力攻击时的自保权等²⁷。但也有学者提出，作为一种习惯法规则，自卫权并不仅限于受到攻击时所采取的自保措施，当处于武力攻击迫在眉睫的特定情势下，国家也有权作出预先自卫的反应²⁸。1842年，美就“加罗林号”事件对英发表照会，称其承认一般法律规则的例外性存在，但针对例外所采取的武力手段必须是“刻不容

²⁴ See Emily Crawford, *Identifying the Enemy: Civilian Participation in Armed Conflict* (Oxford: Oxford University Press, 2015), pp. 140-144; Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict*, 3rd ed. (Cambridge: Cambridge University Press, 2016), p. 177.

²⁵ 参见张兆晨，毛晓彬，黄松华：《面向人机协作的网络信息体系架构建模方法》，《火力与指挥控制》，2020年第6期，第101-105页。

²⁶ 参见丁丽柏，朱静：《〈塔林手册〉对网络攻击中“武力”的界定及反思》，《河北法学》，2018年第12期，第84-89页。

²⁷ See a similar conclusion in Heather Harrison Dinness, *Cyber warfare and the Laws of War* (Cambridge: Cambridge University Press, 2012), pp. 74, 76-81.

²⁸ See Mary Ellen O'Connell, *The prohibition of the use of force* (Cheltenham: Edward Elgar, 2013), pp. 4, 111-116; Dinness认为，预防性自卫的权利可能会引起争议，因为“在习惯国际法中确立了针对即将发生的攻击的自卫权”，并适用于即将发生的网络攻击（Cyber Warfare, 88）。

缓的、压倒一切的、毫无选择的、需要做出即时反应的”²⁹。在尼加拉瓜诉美国案中，国际法院给出的意见是，“在符合必要性原则和比例性原则基础上”，间接承认预先自卫是一种习惯法规则³⁰。

2.网络攻击的特殊复杂性

与传统武力攻击方法手段相比，网络攻击具有其特殊一面：其一，网络攻击行动通常不使用物质形态武器，即不使用飞机、大炮和军舰等，它所使用的主要是计算机语言及特定代码；³¹其二，基于现代社会的工农生产、交通运输、银行金融、医疗卫生等领域高度依赖信息网络，网络攻击行为并不象传统武器那样对目标实施硬毁伤，而只是修改、破坏目标网络的正常运行，通过目标网络间接地对软硬件设施进行干扰、影响和破坏。³²其三，网络攻击是否存在侵入传统意义上的国界的事实值得探究。³³网络攻击通常依托正常商业运营网络进行，在侵入目标网络系统或造成重大负面影响前，很难被发现和阻止。其四。网络攻击的即时性，一旦网络病毒被激活或者相关程序准备完毕³⁴，就可能即时达成瘫痪、扰乱、破坏目标网络设施的严重后果，受害方特别是高度依赖网络运行的国家以及网络技术不发达的国家可能根本没有时间采取传统意义上的自卫行动。

3.现行国际法原则的适用

基于网络攻击行为的特殊复杂性，自卫权的行使难以像传统战争形态那样能够按照既定的程式组织进行³⁵。因此，在技术手段难以有效应对上述挑战的前提下，针对网络攻击的预先性自卫就有了其存在的价值和意义。在此方面，越是计算机网络和信息产业发达的国家，支持采取预先性自卫的群体也就越多，呼声也就越高。毕竟当国家的绝大部分设施都需要依赖网络运行，当民众的绝大部分财富都转化为网络数据时，任何一点安全漏洞都可能导致网络“溃堤”，对国家安全和社会稳定造成难以挽回的致命性影响。也有学者提出，预先性自卫应当基于以下因素才可实施：一是被攻击方采取自卫行动的时机虽然处在攻击行动尚未开始，但已经能够判明攻击方具有攻击意图；二是网络攻击现实紧迫且绝对无法避免；三是该网络攻击是武力攻击的组成部分。上述三种因

²⁹ Letter from Daniel Webster to Lord Ashburton dated 6 Aug.1842; K.E.Shewmaker, *The Papers of Daniel Webster: Diplomatic Papers*, vol.1:1841-1843(Armidale:University of New England Press, 1983) .p.62.

³⁰ See Case Concerning Military and Paramilitary Activities in and against Nicaragua(Nicaragua v.United States of America), Judgment, ICJ Reports(1986)(Merits),p.193.

³¹ 《塔林手册》第41条评注中的第2条，战争手段与方法的定义。

³² 参见赵阵：《技术风险的生成与治理——以计算机漏洞为例》，《自然辩证法研究》2021年第1期，第29-33页。

³³ 参见若英：《什么是网络主权？》，《红旗文稿》2014年第13期，第39页。

³⁴ 参见朱玲玲：《从《塔林手册2.0版》看网络攻击中国家责任归因的演绎与发展》，《当代法学》2019年第1期，第70-78页。

³⁵ 参见张华：《论网络空间自卫权的行使对象问题》，《法学论坛》2021年第1期，第72-80页。

素可以同时存在，也可以兼或有之，或者具有其中之一。上述观点虽然对预先性自卫作了一些限制要求，但也有很多学者并不认同，国际法学者格雷、布朗利等都提出预先性自卫并不符合《宪章》禁止威胁使用或者使用武力的原则³⁶。人类总是在不断前进，科技发展日新月异，达到使用武力程度的网络攻击只是新的武力使用方法手段之一，之前有处于同一性质层面的攻击样式，以后也会有更多创新的攻击样式³⁷，如果仅仅因为网络攻击这种新技术手段造成的困难，就突破了国际法的基本原则，那么法律的预期、警示、规制、处罚等功能将变的毫无意义。例如，网络攻击突破了自卫权的限制，今后人类研究出的更新的作战手段样式，将攻陷诸如区分原则、适度原则、保护妇女儿童原则等一个个原则³⁸，法律的权威将荡然无存。因此，不能以客观之外的没有发生的事情为借口，依靠主观臆断采取自卫行动，而应当坚守《宪章》所明确的不威胁使用和使用武力的原则，在受到武力攻击情形之下，才能行使自卫权。

（二）网络攻击下集体自卫权的实践与主张

1.行使集体自卫权的条约实践

《宪章》第 51 条明确“本宪章不得认为禁止行使单独或集体自卫之自然权利”³⁹。面对外来武力威胁，不同的国家以结盟、提供援助等形式共同采取自卫行动，该种情形古今中外随处可见。基于长期国际实践，集体自卫权被国际习惯法所认可和接受，并在当前世界外交、军事斗争中屡屡采用⁴⁰。新中国成立后签订的第一个双边条约就包含集体自卫权的相关内容。1950 年，中苏两国签订《中苏友好同盟互助条约》⁴¹，中国抗美援朝期间，苏联按照《条约》规定提供了军事及其他援助，包括选派飞行员参战、出售和赠予武器装备，以及提供战争物资和技术支持等。1961 年，中朝两国也签订了《中朝友好合作互助条约》⁴²。上述两个条约有着特殊时代背景，系在冷战条件下，中国与同处东北亚地缘政治板块的苏联和朝鲜这两个唇齿相依的社会主义国家所签订，对维护国家主权、安全、发展利益起到了积极的作用。同时也要清醒看到，随着形势发展变化，特别冷战结束以及中国改革开放逐步走向世界，深度融入国际经贸体系，在大政方针政策层面，已经逐步淡化集体自卫权问题，更加强调和平发展，以谈判化解争端⁴³。

³⁶ See Heather Harrison Dinniss. *Cyber warfare and the Laws of War*(Cambridge: Cambridge University Press,2012), p.83.

³⁷ 参见马光：《论国际法上网络安全的定义和相关国际规则的制定》，《中国政法大学学报》2019 年第 3 期，第 64-78 页。

³⁸ 张艳：《两种安全准则视域下的网络安全防卫权》，《南京政治学院学报》2016 年第 1 期，第 70-74 页。

³⁹ 1945 年《联合国宪章》第 51 条。

⁴⁰ G7 Declaration on Responsible States Behavior in Cyberspace,Lucca,Apr 11,2017,pp.1-13.

⁴¹ 该条约现已失效。

⁴² 该条约有效期 20 年，分别于 1981 年和 2001 年两次自动延期。

⁴³ 《习近平在第七十五届联合国大会一般性辩论上的讲话》新华网：<http://www.xinhuanet.com/politics/leaders/>

2.行使集体自卫权的国家实践

网络攻击虽然具有其鲜明的科技背景和时代特色，但在本质上与其他形式的武力攻击行为没有不同，因此，作为被攻击一方来说，无论是单独自卫，还是集体自卫，都符合《宪章》要求和国际实践。多个国家和国际组织都直接或间接地承认在网络攻击下国家拥有集体自卫权⁴⁴。例如，北约《威尔士宣言》指出，“包括《宪章》与国际人道法在内的国际法直接适用于网络领域”⁴⁵。在应对网络攻击的国家实践中，一些受攻击国家提出或者得到了依据集体自卫权采取的援助和支持。据媒体报道，2007年爱沙尼亚遭受DDOS攻击后，立即依据《北大西洋公约》第5条向北约请求行使集体自卫权⁴⁶，请求北约各国提供相应的技术支持和援助⁴⁷。2019年委内瑞拉大面积停电发生后，委政府得到了俄罗斯专业技术人员的支持，得以短时间内修复其大部分电力操控系统，恢复向市民送电⁴⁸。

3.行使集体自卫权的基本要求

根据国际习惯法和盟约国家的条约协定，应对网络攻击所采取的集体自卫与传统战争形态下的集体自卫要求具有一定的相似之处。首先，行使集体自卫权的基本前提应当是自身受到具有武力侵略性质的网络攻击，而非一般的网络违法犯罪，更不应以假想防卫为名对他国实施先发制人的网络攻击，不论理由多么华丽堂皇⁴⁹。其次，行使集体自卫权应当由受攻击国家先行提出援助请求，或者事后同意（追认）盟国为其提供支持，该种情形既基于国家主权原则，也基于国家间的盟约关系，没有受攻击国的同意或者追认，第三国自行对攻击者采取措施，只能归于第三国与攻击者相互行为产生国际法律关系，而与受攻击国及行使集体自卫权缺乏直接的关联性。再次，行使集体自卫权应同样适用国际法的基本原则和准则。参与集体自卫的国家数量再多，其所达成的协议也不能超越国际法的基本规制范畴。不能以联盟规模的大小、国家数量的多少，做出应对网络攻击集体自卫超出国际法规制的突破性决定。

4.行使集体自卫权的不同观点

针对网络攻击下的集体自卫，也有不同的观点⁵⁰。“《宪章》授权任何会员

2020-09/22/c_1126527652.htm 访问日期：2024年11月21日。

⁴⁴ 参见邵峰：《网络空间自卫权与国际法的规制考量》，《国际论坛》2017年第5期，第35-37页。

⁴⁵ “Wales Summit Declaration,” NATO, last modified April 21, 2021, http://www.nato.int/cps/en/natohq/official_texts_112964.htm.

⁴⁶ See North Atlantic Treaty, arts. 4, 5.

⁴⁷ “Russia Accused of Unleashing Cyberwar to Disable Estonia,” Ian Traynor, last modified March 12, 2021, <https://www.theguardian.com/world/2007/may/17/topstories3.russia>.

⁴⁸ 《委内瑞拉在俄罗斯的支持下开发电网防入侵技术》俄罗斯卫星通讯社：

<http://sputniknews.cn/society/201904041028110303/> 访问日期：2024年11月21日。

⁴⁹ 参见唐岚：《从 WannaCry 事件看网络空间国际规则的困境及思考》，《云南民族大学学报(哲学社会科学版)》2019年第6期，第147-150页。

⁵⁰ See Dr. Dimitrios Delibasis, The Right to National Self-Defence in Information Warfare Operations (Bury St.

国……使用其武装力量援助受攻击的国家，无论是否存在任何这种事先安排”⁵¹。针对伊拉克入侵科威特的非法行为，安理会通过一系列决议承认，第三国无权以本国军队援助科威特，即使这些国家与科威特没有达成事先的协议，也可以直接出兵。⁵²从现行国际法规则和国际行为准则来看，上述观点显然与主流认知存在较大出入。针对行使集体自卫权的方法手段，有的学者认为，应对网络攻击的集体自卫与通常交战状态下的集体自卫在方式方法上有所不同，传统战争更多的是提供人力、财力、武器装备支援，以及提供使用领土（领海、领空）设施、港口、机场等的便利。网络攻击下的集体自卫，不排除援助方也需提供造成事实损害的物质援助，但在此基础上，更多的需要提供智力成果的支持，也就是提供较比攻击方更为强大的计算机网络通信技术、防病毒技术等，形成对攻击方的技术代差和网络优势，例如，提供具有自主知识产权的网络加密技术、网络监测技术、防火墙技术等，开展网络空间情报合作等，以此消减网络攻击造成的危害效应。也有学者提出，应对网络攻击中的集体自卫权相较于传统意义上的武装冲突下的集体自卫并不存在。因为网络攻击具有非常强的隐蔽性和转化性，能够达到对一个国家造成严重损害的后果，那么其攻击者也将强大到非常难以准确判明。攻击发起方既可能是敌对国家、中立国家，也可能就是集体自卫联盟中的一员。因此，在法律意义上隐藏在联盟中的网络攻击的发起者根本不具有行使集体自卫权的权利。“斯诺登事件”曝光的“棱镜门计划”，美国除对中俄等传统对手进行网络攻击外，甚至还对其欧洲盟友实施网络窃密行动，就是相应佐证⁵³。

5.中国团结应对网络安全问题的新主张

“网络安全的本质在对抗，对抗的本质在攻防两端能力较量”。⁵⁴“中国愿同国际社会一道，坚持以人类共同福祉为根本，坚持网络主权理念，推动全球互联网治理朝着更加公正合理的方向迈进，推动网络空间实现平等尊重、创新发展、开放共享、安全有序的目标。”⁵⁵《新时代的中国国防》白皮书强调，中国军队“忠实践行人类命运共同体理念，积极履行大国军队国际责任，全面推进新时代国际军事合作，努力为建设持久和平、普遍安全的美好世界作贡献”。应对未来网络攻击，中国在行使集体自卫权上，应尝试摒弃冷战状态下的集团“硬”对抗方式，以更加灵活的姿态和区别于军事联盟武力自卫的形式，通过构建关键信息基础设施

Edmunds : Arena,2007),p.120.

⁵¹ See Stanimir A.Alexandrov.“Self-defense against the use of force in internationnal law.”*The Cambridge Law Journal* (1997),p.102.

⁵² 参见朱奇武：《海湾危机与国际法》，《政法论坛》1991年第6期，第3-10页。

⁵³ 参见储昭根：《浅议“棱镜门”背后的网络信息安全》，《国际观察》2014年第2期，第56-61页。

⁵⁴ 《习近平在网络安全和信息化工作座谈会上的讲话》新华网：http://www.xinhuanet.com/zgix/2016-04/26/c_135312437.htm，访问日期：2024年11月21日。

⁵⁵ 《习近平：在第三届世界互联网大会开幕式上的视频讲话》新华网：http://www.xinhuanet.com/politics/2016-11/16/c_1119925133.htm，访问日期：2024年11月21日。

施安全保障体系,加强网络安全国际政治、军事、安全合作,不断完善网络安全国内国际立法等“软”合作形式,构建网络空间命运共同体,携手共同应对网络安全重大问题。

三、网络攻击下自卫权立法分析

(一) 应对网络攻击的国际立法滞后

虽然互联网技术已经走过 60 多个春秋,但国际社会就信息网络制定施行的成文法还仅限于《国际电信公约》等几部规范和平时期电信领域规则、技术、要求的法律,应对网络攻击特别是达到“使用武力”程度的网络攻击的成文法少之又少,并且多不具有直接关联性。例如,《日内瓦四公约第一附加议定书》仅针对新武器研究使用的规制条款提出原则性要求,时空的差异化导致议定书无法预见之后的网络、太空等领域新形作战方法手段的出现,因而无法加以具体明确⁵⁶。欧盟《网络犯罪公约》仅对预防和惩治网络空间犯罪行为作出规定,而无法直接适用于交战状态下的网络攻击。⁵⁷有些宣言仅作为阐述政治观点、加强应对指导的政策性文件,不具有公约的法律性质⁵⁸。在专业研究领域,《塔林手册 1.0 版》、《塔林手册 2.0 版》等及时跟进和展现了国际社会应对网络攻击行为的强烈关注,⁵⁹但上述专家和专家组的研究著述成果,也尚未达到形成国际社会共同一致,推动转化为国际公约条约的程度。

(二) 国际社会对加强网络安全立法观点不一

应对网络攻击的国际立法缺失,主要原因在于国际社会未能达成一致。一方面,网络技术欠发达国家希望能够尽早立法,限制和制约网络技术发达国家对网络技术的肆意使用,以确保自身安全和权益不受损害。基于信息网络技术的信息战,使少数发达国家对广大发展中国家具有了不对称的信息战优势。防止少数国家利用网络攻击手段实现其霸权目标,制止其对信息网络的恶意使用,对维护国际和地区和平尤其重要。但网络也是一把“双刃剑”,信息网络技术大国也会面临同样的问题。2018 年 3 月,美国联邦调查局(FBI)、美国国土安全局(DHS)公布调查报告,指责俄罗斯在 2016 年通过黑客形式,采取了一系列网络攻击干预美国大选⁶⁰。2021 年 1 月 23 日,俄罗斯各地爆发反政府抗议示威活动,美驻俄

⁵⁶ 1949 年《日内瓦四公约第一附加议定书》第 36 条:“在研究、发展、取得或采用新的武器、作战手段或方法时,缔约一方有义务断定,在某些或所有情况下,该新的武器、作战手段或方法的使用是否为本议定书或适用于该缔约一方的任何其他国际法规则所禁止。”

⁵⁷ 参见胡健生、黄志雄:《打击网络犯罪国际法机制的困境与前景——以欧洲委员会〈网络犯罪公约〉为视角》,《国际法研究》2016 年第 1 期,第 21-25 页。

⁵⁸ 例如,2009 年上合组织《叶卡捷琳堡宣言》、2011 年《上合组织成立十周年阿斯塔纳宣言》、2004 年美洲国家组织《美洲网络安全综合策略》等。

⁵⁹ 参见北约卓越网络合作防卫中心国际专家小组:《塔林网络战国际法手册》,朱莉欣等译,国防工业出版社 2016 年版,第 38 页;[美]迈克尔·施密特等:《网络行动国际法塔林手册 2.0 版》,黄志雄译,社会科学文献出版社 2017 年版,第 384 页。

⁶⁰ 参见李恒阳:《美国大选中的网络安全问题》,《美国研究》2017 年第 4 期,第 63-64 页。

大使馆第一时间发布声明“密切关注有关俄罗斯 38 个城市爆发抗议活动的消息”。俄外交部对此提出抗议，指责美干预内政，制造混乱，并要求美方就大使馆“提前一天”公布 1 月 23 日俄各大城市示威活动时间、地点的行为作出解释⁶¹。新信息技术发展非常迅猛，有如“脱缰之马”，如果不对其尽早采取法律的限制措施，以公约等成文法的形式加以解决，仅靠立法迟缓且意见难以统一的习惯法规则，未来可能更加难以把控。另一方面，与之相对，网络信息技术发达国家的一些学者则反对过早立法规制网络行为。他们认为，网络技术仍然处于不断发展的状态⁶²。法律颁行即落伍。过早制定限制规则容易造成固步自封，扼杀技术的创新。况且各国网络信息技术参差不齐的情况下，标准、目的、要求并不一致，各国“还没有明智到坐下来创制条约的程度”，希望通过大量实践首先形成习惯法，为网络空间国际立法提供一种尝试和过渡⁶³。作为与应对网络攻击关系最为密切的美国军方，上世纪末已就网络攻击国际立法问题组织了广泛深入研究。美海军战争学院 1999 年 6 月组织召开相关问题研讨会，与会政府官员、军事官员、专家学者达成共识，认为在应对网络攻击方面制定限制或规范国家行为法律的时机未到，尚不具备达成相关国际共识的必要基础和条件。

（三）应对网络攻击国内立法状况

近年来，中国在应对网络安全立法方面取得长足进步，国内立法日趋完善，比如《网络安全法》（2017 年），作为我国首部全面规范网络空间安全的基础性法律，对维护网络主权和国家安全具有重要价值。但国内立法也存在一些不容忽视的矛盾问题。例如，立法管理无序。除相关法律之外，具有网络安全法规规章立法权限的部门较多，总体上缺乏有效统筹，导致权限不清、界限不明，容易产生法出多门、内容冲突滞后等问题。层级效力较低。除全国人大制订的《网络安全法》，以及《国家安全法》《国防法》有关条款或者表述，《刑法》有关惩治网络安全犯罪的条款外，大多是政府部门制定的管理性规范性法规规章，缺乏必要的统一性、权威性、指导性。重技术轻法治。许多法规规章只涉及具体网络安全标准和网络安全技术，但对网络主体的权利义务，特别在维护网络安全，防范网络风险，应对网络攻击上，大多只是提出原则性要求，缺乏具体的操作规则和惩治性条款。

涉及网络安全特别是应对网络攻击自卫权的有关规定，散见于《网络安全法》《国家安全法》以及新修订的《国防法》等法律⁶⁴。上述内容在宏观战略层

⁶¹ 《俄罗斯多地爆发游行集会要求释放纳瓦尔尼，美使馆提前一天发布示威路线图》东方网：<https://www.163.com/dy/article/G145TJ1405503FCU.html>，访问日期：2024 年 11 月 21 日。

⁶² “NATO and Cyber Defence.NATO,”last modified April 21, 2021.
https://www.nato.int/cps/en/natohq/topics_78170.htm

⁶³ See Phillip A.johnson,“Is It Time for a Treaty on Information Warfare? ”in M.N.Schmitt &B.T. O'donnell eds.,*Computer network attack and interational law*(Naval War College,2002),p.439.

⁶⁴ 《网络安全法》第 5 条：“国家采取措施，监测、防御、处置来源于境内外的网络安全风险和威胁，保护关键信息基础设施免受攻击、侵入、干扰和破坏，依法惩治网络违法犯罪活动，维护网络空间安全和秩序。”

面对保卫网络主权和维护网络安全作了原则性规定，但在具体操作层面，还需要大量的下位法以及相关配套措施办法等予以进一步规范和明确。

（四）推动网络安全国际立法的干扰阻滞

中国积极参与网络安全国际立法，特别是 2011 年，中国与上合组织国家共同推出《信息安全国际行为准则》⁶⁵，在听取多方意见后，于 2015 年 1 月再次向联大提交修订版准则。在国际立法层面，一方面相关立法需要世界多数国家签署、加入或批准才能具有最广泛的公约性质，另一方面，联合国框架下的国际立法受到联合国运行规则的限制，在联合国大会上需多数会员国同意，在安理会形成决议需五个常任理事国全部同意，因而，推进相关国际立法非一日之功。尽管中俄推动《准则》出台受到美西方国家干扰，短期内还无法提交联大审议通过，但中俄通过提交草案等方式，充分表达了自身对网络安全的关切，努力将“四项原则”、“五点主张”融入构建网络空间国际秩序实践当中，⁶⁶在获得越来越多国家和国际组织认可基础上，进一步参与和融入网络安全国际立法当中。

四、中国应对网络攻击的策略选择

2019 年 7 月，国务院新闻办发布《新时代的中国国防》白皮书，将网络安全作为中国国防必须应对的主要威胁加以强调，确立了“维护国家在太空、电磁、网络空间等安全利益”政策目标，强调坚持“坚持防御、自卫、后发制人原则，实行积极防御”的新时代军事战略方针，⁶⁷从国家政策角度将自卫权明确适用于包括网络安全在内的非传统的安全威胁领域。其中，“坚持防御、自卫、后发制人”原则完全符合我国作为安理会常任理事国和世界主要大国的一贯的内政外交政策⁶⁸。但无论是遭受“先发制人”武力攻击还是“后发制人”抵抗侵略，对于传统战争形态而言，都具有一定的可辨识性和可操作性，并不存在过多争议。但对网络攻击这种隐蔽性高、无法预防、难于溯源的非传统安全威胁，由于“后发制人”本质上属于被动应对，需以自身受到攻击和伤害为代价，还可能存在目标点、时间线和损害度难以评估认定的问题，因此，处在当下的安全环境中，有必要对“防御、自卫、后发制人”原则作以更进一步的研究探索，适度深化拓展其内涵和外延，如不再仅仅把行使自卫权当作一种实际行动，而应将其作为一种施加压力、

《国家安全法》第 25 条：“国家加强网络管理，防范、制止和依法惩治网络攻击、网络入侵、网络窃密、散布违法有害信息等网络违法犯罪行为，维护国家网络空间主权、安全和发展利益。”《国防法》第 30 条：“国家采取必要的措施，维护在太空电磁网络空间等其他重大安全领域的活动、资产和其他利益的安全。”

⁶⁵ 《信息安全国际行为准则》外交部：

https://www.fmprc.gov.cn/web/wjb_673085/zfxgsk_674865/gknrlb/tywj/zcwj/t858317.shtml，访问日期：2024 年 11 月 21 日。

⁶⁶ 参见支振锋：《互联网全球治理的法治之道》，《法制与社会发展》2017 年第 1 期，第 91-105 页。

⁶⁷ 中华人民共和国国务院新闻办公室：《新时代的中国国防》，载《解放军报》2019 年 7 月 25 日，第 3 版。

⁶⁸ 例如抗美援朝、对印对越边境自卫反击战等，相关军事斗争实践很好地贯彻运用了上述原则。

威慑敌手的有效手段加以运用。对此,新时代国防政策也从实施积极防御、主动遏制网络攻击、加强战役战术进攻动作、广泛参与全球安全治理等角度给出了新的思路办法。

（一）确立网络空间命运共同体的应对思路

习近平在世界互联网大会上提出,“网络空间是人类共同的活动空间,网络空间前途命运应由世界各国共同掌握。各国应该加强沟通、扩大共识、深化合作,共同构建网络空间命运共同体⁶⁹。”加强网络安全合作是构建网络空间命运共同体的应有之义。通过构建创新拓展自卫权在新时代全球互联网治理体系中的行使方法手段,值得进一步深化探索和尝试。

联合国成立前夕,国际法学家凯尔森提出,为了解决持久和平问题,需要在国际法框架中,通过一个具有一定程度的集中化,但不超越“国际共同体”的通常模式之组织。联合国就是这个“国际共同体”的载体⁷⁰。《联合国宪章》赋予会员国行使自卫权权利,同时提倡会员国不分大小,“采取有效集体办法”,“集中力量,以维持国际和平及安全”。互联网的发展无国界、无边界。当代国际法治环境下,应对网络攻击行使自卫权,更应秉持《宪章》精神,在“坚持尊重网络主权,尊重各国自主选择网络发展道路、网络管理模式、互联网公共政策和平等参与国际网络空间治理的权利”⁷¹基础上,协调各国达成应对网络攻击的理念认同、制度协同、行动共同,采取多边参与、多方参与等多种方式建立完善全球互联网治理体系,善于运用“非武力”而非“武力”、“软自卫”而非“硬毁伤”方式,在相对平和的社会状态下,由各国通力合作制止网络攻击,实现“预防性”消除危险和集体自卫目的,努力达到“不战而屈人之兵”的最佳效果。

（二）规范网络攻击的国际和国内立法

为切实维护国家利益不被侵害,中国有权参与制订国际法中与网络战争有关的条文,并进行了相关实践⁷²。但是目前制定完善的网络战争国际法规还存在一定的难度,其一是该法规会直接影响国家利益、限制军队战斗能力,因此存在较大理论争议和利益博弈;其二是由于没有主权国家承认其参与网络战争的事实,因此该项立法还存在着法律承认和技术问题。

我国现存有关跨境网络攻击治理的法律既有实体法也有程序法,例如《国家安全法》、《刑事诉讼法》、《国际刑事司法协助法》等。尽管我国已有相关立法,但

⁶⁹ 《习近平就共同构建网络空间命运共同体提出 5 点主张》新华网: http://www.xinhuanet.com/world/2015-12/16/c_128536396.htm, 访问日期: 2024 年 11 月 21 日。

⁷⁰ See Hans Kelsen, *Peace Through Law* (Chapel Hill: The University of North Carolina Press, 1944), pp. 12-13; 徐宏:《人类命运共同体与国际法》,《国际法研究》2018 年第 5 期,第 3-14 页。

⁷¹ 1945 年《联合国宪章》。

⁷² 《中美俄等 15 国签署规范网络行为合作意愿书》腾讯网: <https://tech.qq.com/a/20100720/000260.htm>, 访问日期: 2024 年 11 月 21 日。

其并不完善,跨境网络攻击事件较为复杂,需要更加系统性的法律法规⁷³,因此我国应以自身实践为基础,结合其他发达国家相关经验,提升我国法律应对国际网络攻击的有效性、实用性。由于网络本身的特殊性,可以根据《网络安全法》第 75 条和《国际刑事司法协助法》等法律法规处理跨境网络攻击问题⁷⁴。

(三) 推进治理网络空间的国际合作

2017 年 3 月,我国发布了《网络空间国际合作战略》,倡导各国共同制定跨境网络攻击相关法律准则及标准,促进各国网络空间技术与策略共同发展。

我国应从国家战略层面构建我国的网络安全发展体系。目前,随着网络空间的发展,其已逐渐成为我国实体经济命脉,作为我国政治、经济等活动的基础平台,成为我国社会运转的神经系统,因此,维护我国网络空间安全也就是维护我国自身利益。为构建我国网络安全保障体系,建设网络强国,我国应加快制定网络安全总体战略,落实维护网络安全的行动方针与举措。

统筹协调网络安全保障与国家政治、经济等战略目标的实现。从网络安全的特征来看,保障我国网络安全势必会对我国政治、经济等战略目标的实现造成影响,虽然保障网络安全并不是我国战略的终极目标,但其却促进了我国实现其他战略目标,且实现我国网络安全的具体举措也可以作为我国实现其他战略目标的有效手段。目前随着我国国际地位的不断提高,网络安全保障不应只局限于国内的基础信息网络,而应随着我国的发展而不断拓展。我国还应高度重视并积极推动网络安全国际合作⁷⁵。我国应逐步形成在国际上的维护我国网络安全的话语体系,以联合国等国际组织相关法律法规为基础,坚守我国网络安全主张,推动我国与其他国家网络安全双边、多边协调机制的建立,切实解决当前国际网络安全问题。

(四) 网络震慑与加强网络安全技术保护

传统威慑可以让攻击者基于恐惧的心理放弃攻击行为,因为防守方实施报复或惩罚措施后,攻击行为产生不了任何利益。防守方旨在通过向其攻击者发出这些威慑要素的信号来维持现状,以阻止或劝阻后者进行敌对行为⁷⁶。网络威慑不同于传统威慑,传统威慑针对特定武器与活动,而前者跨越整个网络空间。在实施上有三种预想:一是将目光锁定于更精密的网络武器,类比于传统威慑中的核武器⁷⁷。二是诉诸国内法和国际法对有害网络活动或装置施加的限制,以及对违反这

⁷³ 参见王虎华,张磊:《国家主权与互联网国际行为准则的制定》,《河北法学》2015 年第 12 期,第 12 页。

⁷⁴ 《网络安全法》第 75 条:“境外的机构、组织、个人从事攻击、侵入、干扰、破坏等危害中华人民共和国的关键信息基础设施的活动,造成严重后果的,依法追究法律责任;国务院公安部门和有关部门并可以决定对该机构、组织、个人采取冻结财产或者其他必要的制裁措施。”

⁷⁵ 《习近平:把我国从网络大国建设成为网络强国》新华网: http://www.xinhuanet.com/politics/2014-02/27/c_119538788.htm, 访问日期: 2024 年 11 月 21 日。

⁷⁶ See Emilio Iasiello, “Is Cyber Deterrence an Illusory Course of Action?” *J. Strat. Security* 7(2014), pp.52,55-59.

⁷⁷ 参见沈逸,江天骅《网络空间的攻防平衡与网络威慑的构建》,《世界经济与政治》2018 年第 2 期,第 49-52 页。

些法律的惩罚或制裁⁷⁸。三是基于复杂情况进行特定威慑,专门用于阻止特定类型的网络操作和特定参与者,例如阻止来自特定服务器或携带特定类型的文件的服务器的所有网络流量⁷⁹。

目前对于网络攻击的技术追踪与认定还存在一定问题,因此不仅需要各国提高自身网络技术,也应有相关国际组织共同制定具有前瞻性的技术标准。其次,网络攻击的责任分配仍然是目前网络战争领域的一大难题,网络攻击过程较为复杂,因此对网络攻击负有责任的应包括初始攻击者以及后续故意和非自觉参与的全部个人及组织,并且在网络攻击中,网络攻防转换频繁,因此难以适用相关法律的具体规则。一个国家想要在网络战争中很好地抵御其他国家的攻击必须采取全面的防御措施,且时刻保持警戒状态,这在实践中还存在着一定难度,尤其是病毒性网络武器,传染性强,难以控制,最终导致战场范围不断扩大。

此外,我国应重视核心技术的自主可控。2018年以来,美国高调制裁中国中兴、华为等中国信息通信龙头企业,试图从信息网络创新制造研发源头阻碍中国信息产业发展。随着大数据、人工智能等新兴领域的兴起,我国应提前布局,将网络安全需求与产品、技术生命周期相融合,突破核心技术,加速推进网络安全领域国产化替代工作,避免处于被动局面。我国还应重视人才培养⁸⁰。我国于2017年首次公布我国“一流网络安全学院”,并开始我国网络安全人才培养计划。据此,我国还应提高网络安全在我国社会中的流行度,以学科教育、职业教育以及技能大赛多种方式提高社会参与度。

结语

网络攻击事件近年来频频发生,传统攻击已不再是战场上唯一的选择,网络安全领域给国计民生、安危福祉带来了重大的系统性风险。国际法也应顺应这些趋势,并应能够在技术进步的同时规范这方面的具体法律以应对国际社会面临的法律问题。网络空间存在国家主权,对网络攻击这种隐蔽性高、无法预防、难于溯源的非传统安全威胁,由于“后发至人”本质上属于被动应对,需以自身受到攻击和伤害为代价,还可能存在目标点、时间线和损害度难以评估认定的问题,所以国家行使自卫权必然面临诸多盲点,在立法欠缺的情况下,我国应不断完善网络安全国内国际立法等“软”合作形式。《新时代的中国国防》白皮书将网络安全作为中国国防必须应对的主要威胁加以强调,从国家政策角度将自卫权明确适用于包括网络安全在内的非传统的安全威胁领域。我国应通过推进网络安全国际政治、军事、安全合作,加强立法层面的网络震慑与技术层面的保护与人才培养等,

⁷⁸ See Dorothy E. Denning, "Rethinking the Cyber Domain and Deterrence," *Joint Forc Quarterly* 77 (Apr. 2015), pp. 8, 11-15.

⁷⁹ See Eric Talbot Jensen, "Cyber Deterrence," *Emory IL Rev.* 26(2012), pp. 773, 780-783.

⁸⁰ 参见黄志雄:《网络空间国际规则制定的新趋向——基于〈塔林手册 2.0 版〉的考察》,《厦门大学学报(哲学社会科学版)》2018 年第 1 期,第 1-11 页。

构建网络空间命运共同体，携手共同应对网络安全重大问题。还可以更进一步的研究探索，适度深化拓展现有实行积极防御原则的内涵和外延，如不再仅仅把行使自卫权当作一种实际行动，而应将其作为一种施加压力、威慑敌手的有效手段加以运用。

The Dilemma and strategic choice of National Self-Defense Right in Cyber Attack

Abstract: Whether it is a "preemptive" armed attack or a "later strike" resistance to aggression, the traditional forms of war are recognizable and operable to a certain degree without much controversy. Cyber attack as the non-traditional security threat is highly concealed, unpreventable, and difficult to trace, so the state will inevitably face many difficulties in exercising its right of self-defense, such as the disputes of anticipatory self-defense, collective self-defense and the lack of legislation. Among them, on the issue of anticipatory self-defense, the principle of "adhering to defense, self-defense, and later strike" is fully in line with China's consistent domestic and foreign policies as a permanent member of the Security Council and a major power in the world. In exercising the right of collective self-defense, we should try to abandon the "hard" group confrontation method in the Cold War state, and adopt a more flexible posture and form that is different from military alliances force self-defense. In the absence of legislation, China should act as a model country for regulating cyber attacks, continuously improve forms of "soft" cooperation such as domestic and international legislation on cyber security, strengthen cyber deterrence at the legislative level, protection and talent training at the technical level, etc., and strive to build a community with a shared future in cyberspace.

Key words: Cyber attacks, the right to self-defense, cyber security, a community with a shared future in cyberspace

人工智能时代网络恐怖主义犯罪的国际法规制的问题与路径

李彦¹

摘要：人工智能技术的应用使网络恐怖主义犯罪国际法规制面临攻击规模和影响范围扩张，追踪和打击难度增加等新挑战。传统恐怖主义国际刑事规则可作为法律依据，网络犯罪国际刑事规则具有可适用性，人工智能国际规则具有规范价值，国际软法规范产生一定的法律效果。但传统打击恐怖主义国际刑事规则的缺乏针对性，网络犯罪国际刑事规则碎片化显著，人工智能国际规则的适用不确定性大，相关国际软法的局限性明显。完善人工智能时代网络恐怖主义犯罪国际法规制的路径应加强完善现有国际规则体系，推动建构专门的法律机制，强化软法与硬法的动态衔接。中国可通过推动国际立法机制纳入网络恐怖主义犯罪条款，推动在区域组织框架下制定打击网络恐怖主义犯罪的区域规则和建构国内网络恐怖主义犯罪法律规则体系贡献中国力量。

关键词：人工智能 网络恐怖主义犯罪 国际法规制 国际合作

一、人工智能时代网络恐怖主义犯罪的基本问题

（一）人工智能技术的应用与网络恐怖主义犯罪

网络恐怖主义犯罪是指利用网络传播极端思想、策划恐怖活动或攻击关键基础设施的犯罪行为。其核心目的是通过制造恐慌、破坏社会秩序实现政治或意识形态目标。与传统恐怖主义犯罪相比，实施网络恐怖主义犯罪的成本低，只需一个可联网的网络设备（如手机、计算机等）就能操控实施网络恐怖主义犯罪，无需购置实体武器、租赁场地等，大大降低了实施恐怖活动的成本。实施网络恐怖主义犯罪的隐蔽性强，网络空间没有物理屏障和边界限制，恐怖分子可从世界任何地方通过使用在线昵称、以“过客”身份登录网站或利用技术手段隐藏真实 IP 地址等手段匿名实施恐怖行为，安全机构和警察很难追查到恐怖分子的真实身份。网络恐怖主义犯罪的危害大、影响范围广，犯罪分子的攻击目标涵盖金融、电力、交通等关键领域，可能导致大规模的资金损失和秩序混乱，危及民众生命安全，甚至可能引发全球性灾难。

人工智能技术的快速发展为网络恐怖主义犯罪提供了新工具和手段。数字时代，网络恐怖主义犯罪正借助人工智能的突破传统的犯罪边界，以更加隐蔽、高效、精准的方式，对社会安全和稳定造成前所未有的威胁。第一，人工智能技术为恐怖组织快速生成极具煽动性的激进言论和逼真的视频等宣传内容提供了便利，使恐怖主义思想和信息的传播更加迅速和广泛。利用人工智能技术生成的内

1 河南财经政法大学国际学院副教授

容往往经过精心设计，利用人类情感中的恐惧、愤怒、迷茫等弱点，以极具感染力的方式呈现，从而达到蛊惑人心的目的。通过社交媒体等平台的广泛传播，这些内容能够迅速扩散到全球各个角落。社交媒体的开放性和便捷性使得信息传播几乎不受地域和时间的限制，恐怖组织可以轻松地将其极端思想传递给潜在的受众。他们通过精准推送算法，将宣传内容推送给那些可能对其理念感兴趣或正处于心理迷茫期的人群，尤其是年轻人，试图煽动他们的极端情绪，招募他们成为新成员。以臭名昭著的“伊斯兰国”（IS）为例，其利用人工智能生成媒体项目“新闻收获”发布视频。在这些视频中，人工智能生成的新闻主播身着迷彩服，出现在屏幕上，身后的新闻滚动条和视频片段展示着所谓 ISIS 成员执行恐怖任务的画面，主播则宣读着 IS 官方媒体的快讯。节目制作精良，类似电视台的新闻广播，在世界各地的支持者中获得了病毒般的扩散效果，成功吸引了不少人加入该极端组织。这种智能宣传方式，不仅成本低廉，而且传播速度快、覆盖面广，给国际社会的反恐工作带来了极大困难。

第二，人工智能算法可以帮助恐怖分子更精准地分析目标系统的漏洞和弱点，从而制定出更有效的攻击策略。在分析目标方面，人工智能可以收集和处理海量的数据，包括政治、经济、社会、文化等各个领域的信息，从而帮助恐怖组织深入了解目标的特点、弱点和规律。通过对这些数据的分析，恐怖组织能够精准地选择攻击目标，制定出更加有效的攻击策略。在挖掘网络漏洞方面，人工智能同样发挥着重要作用。它可以利用机器学习算法，对网络系统进行扫描和分析，自动检测出潜在的漏洞和安全隐患。恐怖组织一旦掌握了这些漏洞，就可以发动更具隐蔽性和破坏力的网络攻击。例如，他们可以利用漏洞入侵关键基础设施的控制系统，如电力、交通、通信等系统，导致这些系统瘫痪，给社会带来巨大的混乱和损失；或者入侵金融机构的网络系统，窃取资金、篡改数据，破坏金融秩序。这种利用人工智能进行的精准策划和攻击升级，对公共安全和关键基础设施构成了直接的威胁。关键基础设施是现代社会运行的基石，一旦遭到攻击，将引发连锁反应，影响整个社会的正常运转，甚至可能导致社会秩序的崩溃。而且，由于人工智能攻击具有高度的智能化和自动化特点，传统的安全防护手段往往难以应对，使得防范和打击此类攻击变得异常困难。

第三，人工智能技术的使用增加了身份识别和追踪网络恐怖主义犯罪的难度。通过人工智能创建虚假身份和虚拟人物，其技术原理主要基于对大量真实身份信息的学习和模仿。人工智能可以生成看似真实的个人资料、照片、语音等信息，用于注册社交媒体账号、开设网络账户等，从而隐藏自己的真实身份。同时，人工智能算法还能够帮助恐怖分子隐藏网络活动痕迹。它可以对网络流量进行分析和伪装，使其看起来与正常的网络活动无异，从而躲避监管和追踪。例如，人工智能可以将恶意网络流量混入大量的正常流量中，或者通过加密技术对网络活动进行加密，使得执法部门难以察觉和追踪。这种身份伪装与隐匿踪迹的手段，给执法

和反恐工作带来了巨大的挑战。执法部门在追踪恐怖分子时，往往会陷入虚假身份和模糊线索的困境，难以确定恐怖分子的真实身份和位置，从而无法及时采取有效的打击措施。而且，随着人工智能技术的不断发展，恐怖分子的身份伪装和隐匿踪迹手段也在不断升级，使得执法部门始终处于被动应对的局面。恐怖分子可利用人工智能算法对网络系统进行深度扫描和分析，精准地识别出系统中的漏洞和弱点，绕过传统的安全检测机制，发起更加隐蔽的恐怖主义袭击。如通恐怖分子可通过生成对抗网络（GAN）技术，生成与正常流量极为相似的恶意网络流量，从而躲避防火墙和入侵检测系统的监测。还可利用人工智能开发新型的恶意软件，这些恶意软件能够根据目标系统的环境和防御措施自动调整攻击策略，具有更强的适应性和攻击性。

（二）人工智能技术的应用使网络恐怖主义犯罪国际法规制面临新挑战

网络恐怖主义犯罪早期主要借助网络进行简单的信息传播与组织联络，恐怖组织通过论坛、社交媒体群组等宣扬极端思想、招募成员、组织恐怖活动。随着互联网的普及和技术迭代，网络恐怖主义逐渐升级，开始策划实施网络攻击，如破坏政府或企业网站等。而人工智能的介入使得网络恐怖主义犯罪不再是简单的信息传播与低层次攻击，而是向着智能化、协同化等方向加速演变，给全球安全带来了前所未有的挑战。

人工智能技术作用下网络恐怖主义犯罪模式的转变使攻击的规模和影响范围难以估量。人工智能技术使网络恐怖主义犯罪从传统的单一、分散模式向智能、协同模式转变，借助人工智能的自动化和智能化能力，恐怖组织可实现多目标、多维度的协同攻击。如通过编写智能攻击程序，恐怖分子能够同时对多个政府机构、金融机构和关键基础设施的网络系统发动攻击，各个攻击模块间相互配合、相互呼应，形成强大的攻击合力。这种智能、协同的犯罪模式，不仅大大提高了攻击的效率和成功率，也使网络安全防护体系难以全面有效地进行防御。加之，攻击手段的升级增加了提前预警和防范网络恐怖主义犯罪的难度。人工智能技术为恐怖分子提供了强大的数据处理和分析能力，人工智能算法深度挖掘与分析收集来自社交媒体、网络论坛、政府公开数据、商业数据库等多渠道的海量信息，发现人群的活动规律、安保部署情况、重要设施的运行数据等，以使恐怖分子精准预测安保漏洞，选择最有利的袭击目标、时间和方式。如在演唱会、体育赛事等活动期间，恐怖分子可以借助人工智能分析活动的人员流量数据、安保排班表以及周边交通状况，确定在何处、何时发动袭击能够造成最大的混乱与伤亡，进而制定出极具针对性的袭击计划，让安保力量防不胜防。恐怖分子还利用人工智能技术开发出新一代的恶意软件，这些恶意软件具备了更强的学习和适应能力，能够自动分析目标系统的特征、防御机制，从而调整自身的攻击策略和行为模式，以躲避检测和防御系统的拦截。传统的恶意软件特征相对固定，安全软件可以通

过识别这些特征来进行查杀；但基于人工智能的恶意软件能够不断改变自身代码结构和行为方式，实现动态伪装。如一些恶意软件在感染目标系统后，会利用人工智能算法实时监测系统内的安全软件运行状态，一旦发现被检测，便迅速改变代码形态，让安全软件的特征库无法匹配识别。人工智能恶意软件还可以根据目标系统的漏洞情况，自动生成最具破坏力的攻击指令，如篡改关键数据、破坏系统核心文件、窃取敏感信息等，给目标系统带来巨大的损失。

人工智能技术推动网络恐怖主义犯罪的组织形式向更加扁平化、虚拟化的方向发展，组织形式的变化使执法部门在追踪和打击恐怖组织时难以掌握其行踪和意图。传统的恐怖组织通常具有明确的层级结构和指挥体系，人员间的联系较为紧密；但在人工智能的支持下，网络恐怖主义组织可借助互联网和智能通信工具，实现成员间的分散化协作。组织成员不再需要集中在特定的地点，而是可分布在世界各地，通过虚拟网络进行沟通和协调。这种扁平化、虚拟化的组织形式，使恐怖组织的活动更加灵活、隐蔽；执法部门更难确定组织成员的真实身份和位置，也难以掌握组织的具体运作模式和行动计划。恐怖组织利用人工智能算法精准分析社交媒体、视频平台、论坛等网络空间的用户行为和兴趣偏好，通过智能推荐系统有针对性地推送极端思想、恐怖主义宣传内容给那些对特定政治、宗教话题表现出极端观点或浓厚兴趣的用户，吸引潜在的同情者和招募对象，大大提高了宣传招募的效率和成功率。恐怖分子根据线上分析得出的目标安保漏洞、人员活动规律等信息，制定详细的袭击计划。从选择袭击地点、规划袭击路线，到确定袭击时机，每一个环节都经过精心策划，以确保最大程度地造成破坏和恐慌的同时完美脱身。恐怖组织还利用复杂的加密货币交易算法和智能合约技术，在虚拟货币市场中隐蔽地筹集资金以资助和支持网络恐怖主义犯罪。这些算法能够躲避金融监管机构的监测，将非法资金分散在众多交易中，使其来源和流向难以追踪。

二、 人工智能时代网络恐怖主义犯罪国际法规制的现状

目前，国际社会尚未形成专门针对人工智能时代网络恐怖主义犯罪的国际法规制体系，主要依据一般国际刑事规则、网络犯罪国际刑事规则、人工智能国际规则以及国际软法规范进行规制。

（一） 传统恐怖主义国际刑事规则可作为法律依据

联合国框架下与打击传统恐怖主义相关的国际公约，如《制止恐怖主义爆炸的国际公约》《制止向恐怖主义提供资助的国际公约》《制止核恐怖主义条约》等，对恐怖主义犯罪进行了界定，并规定了缔约国的合作义务，可以为打击网络恐怖主义犯罪提供一定的法律依据。

《制止恐怖主义爆炸的国际公约》将“恐怖主义爆炸罪”定义为“任何人非法和故意在公共场所、国家或政府设施、公共交通系统或基础设施投掷、放置、发射或引爆爆炸性或其他致死装置，意图致人死亡或重伤，或意图对这类场所、设

施或系统造成巨大毁损，从而带来或可能带来重大经济损失的行为。”²公约适用于两类恐怖主义行为：使用爆炸性或其他致死装置实施的恐怖主义行为和针对公共场所、国家或政府设施、公共交通系统或基础设施实施的恐怖主义行为。由于公约并未限定行为方式，利用人工智能技术实施的以下网络恐怖主义犯罪行为也可适用公约的相关规定：利用人工智能技术制造和操控爆炸装置，如利用人工智能技术设计和制造具有自主学习和决策能力的爆炸装置，并远程操控其实施爆炸袭击；用人工智能技术攻击关键信息基础设施，如利用人工智能技术开发恶意软件，攻击电力、交通、金融等关键信息基础设施，造成重大经济损失和社会恐慌。

《制止向恐怖主义提供资助的国际公约》将“资助恐怖主义罪”定义为“任何人以任何手段，直接或间接地，非法和故意地提供或募集资金，意图将这些资金全部或部分用于实施，或明知这些资金将全部或部分用于实施实施：(a) 属附件所列条约之一的范围并经其定义为犯罪的一项行为；或(b) 意图致使平民或在武装冲突情势中未积极参与敌对行动的任何其他人死亡或重伤的任何其他行为，如这些行为因其性质或相关情况旨在恐吓人口，或迫使一国政府或一个国际组织采取或不采取任何行动。”³由于公约并未限定行为方式，利用人工智能技术实施的以下网络恐怖主义犯罪行为也可适用公约的相关规定：利用人工智能技术进行资金支持，如利用人工智能技术开发匿名支付系统，帮助恐怖组织转移资金，逃避监管；利用人工智能技术进行资金募集，如利用人工智能技术分析社交媒体数据，精准识别潜在资助者，并进行定向募捐。

《制止核恐怖主义行为国际公约》将“核恐怖主义罪”定义为“任何人非法和故意(a)拥有放射性材料或制造或拥有一个装置”“(b)以任何方式利用放射性材料或装置，或以致使放射性材料外泄或有外泄危险的方式利用或破坏核设施”“在显示威胁确实可信的情况下，威胁实施”上述(b) 项所述犯罪，“在显示威胁确实可信的情况下通过威胁，或使用武力，非法和故意索要放射性材料、装置或核设施”“以共犯身份参加……组织或指使他人实施……或以其他方式促进以共同目的行动的群体实施”上述犯罪。⁴由于公约并未限定行为方式，利用人工智能技术实施的以下网络恐怖主义犯罪行为也可适用公约的相关规定：利用人工智能技术获取或制造放射性物质或装置，如利用人工智能技术破解核设施安全系统，窃取放射性物质或装置；利用人工智能技术操控或威胁使用放射性物质或装置，如利用人工智能技术远程操控放射性物质或装置实施恐怖袭击。

（二）网络犯罪国际刑事规则具有可适用性

全球性公约《联合国打击网络犯罪公约》，**区域性网络犯罪公约**如欧洲委员会 2001 年的《网络犯罪公约》(Convention on Cybercrime)、阿拉伯国家联盟的

² 《制止恐怖主义爆炸的国际公约》第 2 条。

³ 《制止向恐怖主义提供资助的国际公约》第 2 条。

⁴ 《制止核恐怖主义行为国际公约》第 2 条。

2010 年《阿拉伯打击信息技术犯罪公约》(Arab Convention on Combating Information Technology Offences)、独立国家联合体 2001 年的《独立国家联合体打击计算机信息领域犯罪合作协定》(The Commonwealth of Independent States Agreement on Cooperation in Combating Offences related to Computer Information)、西非国家经济共同体 2011 年的《西非国家经济共同体打击跨国网络犯罪指令》(Directive on Fighting Cybercrime within Economic Community of West African States)、上海合作组织成员国 2009 年《上海合作组织成员国保障国际信息安全政府间合作协定》(Shanghai Cooperation Organization Agreement on Cooperation in the Field of International Information Security)、非洲联盟 2014 年《非盟网络安全和个人数据保护公约》(African Union Convention on Cyber Security and Personal Data Protection)等,均具有可适用性。但它们多对网络恐怖主义犯罪的特殊性和严重性认识不足,缺失专门针对网络恐怖主义犯罪的规定和措施,或相关规定较笼统,可适用性差。

《联合国打击网络犯罪公约》是联合国框架下打击网络犯罪的重要国际法律文件,其规定的犯罪行为范围更广,并强调了国际合作的重要性。公约谈判过程中,曾在第一份合并谈判文件中明确将“与恐怖主义相关的犯罪”纳入公约适用范围,这为打击利用人工智能技术实施的网络恐怖主义犯罪提供了重要法律依据。合并谈判文件将“与恐怖主义相关的犯罪”定义为“通过计算机系统实施的,旨在恐吓人口或强迫政府或国际组织实施或不实施任何行为的犯罪行为”。该界定显然采纳的是宽范畴的界定方式,不仅涵盖了传统网络恐怖主义行为如爆炸、绑架等,还包括利用计算机系统实施的网络攻击、网络宣传等新型恐怖主义行为。由此,利用人工智能技术实施的诸多恐怖主义行为都可以适用合并谈判文件:利用人工智能技术进行自动化攻击,如利用人工智能技术开发恶意软件、僵尸网络等自动化攻击工具,对关键信息基础设施实施网络攻击,造成重大经济损失和社会恐慌;利用人工智能技术进行精准化宣传,如利用人工智能技术分析社交媒体数据,精准识别目标人群,并针对性地推送恐怖主义宣传内容,煽动暴力恐怖活动;利用人工智能技术进行智能化招募,如利用人工智能技术分析社交媒体数据,识别潜在的恐怖分子,并对其进行精准化的招募和洗脑,扩大恐怖组织规模。然而,遗憾的是公约最后的文本删除了该条。不过,公约不仅规定了非法侵入计算机系统、非法拦截数据、滥用设备等传统网络犯罪行为,还规定了利用计算机系统实施的犯罪行为,如与计算机相关的伪造和诈骗、与儿童色情相关的犯罪、与侵犯版权和邻接权相关的犯罪,这些条款也都可以适用于打击利用人工智能技术实施的网络恐怖主义犯罪。由此,利用人工智能技术伪造证件、实施网络诈骗等;利用人工智能技术制作、传播儿童色情内容等;利用人工智能技术盗版软件、传播盗版内容等都可适用公约予以规制。

《网络犯罪公约》等区域性公约有直接设置网络恐怖主义犯罪罪名条款的，如《上海合作组织成员国保障国际信息安全政府间合作协定》及其附件仅宽泛地涉及信息恐怖主义——其第2条指出：“各方应以国际信息安全领域存在的下列主要威胁为出发点……信息恐怖主义”；其附件1对信息恐怖主义作了界定“信息恐怖主义——为达到恐怖主义目的，在信息空间使用和（或）影响信息资源”；附件2进一步说明了信息恐怖主义的内容，“该威胁根源：恐怖组织或参加恐怖活动的个人，利用或针对信息资源进行非法活动。其威胁特征：恐怖组织利用信息网络实施恐怖活动，吸收新成员；破坏信息资源，导致社会秩序混乱；控制或封锁大众传媒渠道；利用互联网或其他信息网络散布恐怖主义言论，制造社会恐怖和恐慌，以及对信息资源造成其他负面影响”。⁵《阿拉伯打击信息技术犯罪公约》与该协定附件2类似，明确了网络恐怖主义的内容，“与恐怖主义相关的信息技术手段犯罪包括：1.传播和倡导恐怖组织的理念和原则；2.为恐怖行动融资和训练，并促进恐怖组织之间的沟通；3.传播制造爆炸物的方法，特别是用于恐怖主义行动；4.传播宗教狂热和异议和攻击宗教和信仰”。⁶相关条款当然可以直接适用于打击利用人工智能实施的网路恐怖主义犯罪。⁷而其它未直接设置罪名条款的公约也基本都有设置非法访问、非法拦截、滥用设备等罪名条款，这些条款也可以适用于部分利用人工智能实施的网路恐怖主义犯罪行为。“非法访问罪”往往针对“未经授权故意侵入计算机系统的全部或部分”的犯罪行为，可适用于恐怖分子利用人工智能技术实施的以下行为：利用人工智能技术进行自动化攻击，如利用人工智能技术开发恶意软件、僵尸网络等自动化攻击工具，未经授权侵入目标计算机系统，窃取数据或破坏系统；利用人工智能技术进行精准化攻击，如利用人工智能技术分析目标系统漏洞，制定精准攻击方案，未经授权侵入目标计算机系统，实施破坏活动。“非法拦截”往往针对“未经授权故意拦截计算机系统中非公开传输的计算机数据”的犯罪行为，可适用于恐怖分子利用人工智能技术实施的以下行为：利用人工智能技术进行网络监听，如利用人工智能技术开发网络监听工具，未经授权拦截目标计算机系统中传输的敏感数据，用于恐怖主义活动。利用人工智能技术进行数据窃取，如利用人工智能技术分析目标系统安全漏洞，制定数据窃取方案，未经授权拦截目标计算机系统中存储的敏感数据。“滥用设备罪”往往针对“生产、销售、采购、进口、分发或以其他方式提供主要用于实施上述犯罪行为的设备或计算机程序”的犯罪行为，可适用于恐怖分子利用人工智能技术实施的以下行为：利用人工智能技术开发恶意软件，如利用人工智能技术开发具有自我学习、自我进化能力的恶意软件，用于实施网路恐怖主义犯罪；

⁵ 澳门特别行政区行政长官办公室，第28/2013号行政长官公告，第2条、附件1和附件2。信息犯罪是网络犯罪的一种表现形式，信息恐怖主义也是网路恐怖主义的一个方面。

⁶ Arab Convention on Combating Information Technology Offences, art.15.

⁷ 李彦、马冉：《网路恐怖主义犯罪国际法治理研究》，《河南财经政法大学学报》2019年第1期，第150-151页。

利用人工智能技术开发网络攻击工具，如利用人工智能技术开发自动化网络攻击工具，用于实施大规模网络攻击。

（三）人工智能国际规则具有规范价值

近年来，随着人工智能技术的发展与应用，国际社会越发关注对人工智能技术的监管，相关的倡议和进程层见迭出。但国内和国际层面的立法进程仍大多处于原则宣示阶段，在规则的针对性和规则的执行机制方面不够明确。目前，仅有欧盟立法先行制定了一些法律规则，如欧洲议会的《人工智能法》（Artificial Intelligence Act）、欧盟委员会的《人工智能与人权、民主和法治框架公约》

（Framework Convention on artificial intelligence and human rights, democracy, and the rule of law 等，巩固了国内与国际层面立法进程中存在共识的人工智能原则，是对人工智能的监管方法和相应的执行机制的确认和实施。人工智能国际规则在网络反恐中具有重要的规范价值，它们为规范人工智能技术的发展和应用，防范人工智能被用于恐怖主义目的提供了一定的指导和约束。

欧盟《人工智能法案》是世界上第一部专门规制人工智能的法律，旨在促进欧盟单一市场上私人 and 公共行为体开发、采用安全可信的人工智能系统；确保尊重欧盟公民的基本权利，并促进欧洲在人工智能方面的投资和创新。2024 年 3 月 13 日，欧洲议会以 523 票对 46 票的压倒性多数，正式通过了欧盟《人工智能法》的最终文本。法案采用了基于风险的方法，即人工智能对社会造成危害的风险越高，相应的规则就越严格。⁸该法共有 13 章、113 条，主要内容如下：其一，适用主体十分广泛，除了位于欧盟成员国境内的人工智能系统部署者、进口者等，还包括欧盟境内受影响的个人，及位于第三国但其生产或使用的人工智能系统在欧盟境内输出结果的主体等。其二，对人工智能系统采取的四类风险分级：第一类是存在不可接受风险的人工智能系统，包括基于社会行为和个人特征对个人和群体进行社会评分的人工智能系统等，共计八种，这类系统将在该法生效 6 个月后被禁止使用；第二类是高风险人工智能系统，即可能对健康、安全、基本权利和法治构成重大威胁的系统，包括用于管理关键基础设施的系统、执法系统等，这类系统的提供者将受到更严格的约束，如必须保证实现人类监督、建立风险管理体系，并向当局提供技术文件以证明合规性等；第三类是风险有限的人工智能系统，这类系统不会构成严重威胁，主要规定了提高透明度的义务，如保证人工智能系统生成的内容可被识别等；第四类是风险最小的人工智能系统，如垃圾邮件过滤器等，这类系统可以在欧盟境内自由投放及使用。其三，专门规定了通用人工智能模型（General-purpose AI Models，以下简称 GPAIM）

⁸ European Council, Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI, <https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/#:~:text=Today%20the%20Council%20approved%20a,society%2C%20the%20stricter%20the%20rules,visited> on April 11, 2025.

问题。该法最早的草案由欧盟委员会于 2021 年首次提出，但由于当时生成式人工智能尚未出现，委员会最初并未考虑到相关的监管规则。随着此类人工智能的广泛使用，对 GPAI 进行监管的必要性逐渐凸显：GPAI 出现了产出歧视性结果、不透明、泄露隐私等多重风险。随后经历了重大修订，在四类风险分级之外，该法又引入了对 GPAI 人工智能的监管规则。GPAI 是使用大量数据进行大规模自我监督训练的人工智能模型，这类模型具有显著的通用性，能胜任各种不同的任务，并可集成到各种下游系统或应用中。根据是否具备高度影响力，GPAI 可被分为不具有系统性风险和具有系统性风险两类。如果模型训练时的累计计算量大于 10^{25} FLOP 时，即推定其具有高度影响能力，具备系统性风险。这两类 GPAI 适用不同的监管规则。⁹其四，设立管理机构，包括负责在欧盟范围内执行共同规则的人工智能办公室，支持执法活动的独立专家科学小组，由成员国代表组成的人工智能委员会，以及提供专业技术知识的利益相关者咨询论坛。另外，法案还规定了特定人工智能系统的提供者和部署者的透明度义务、支持创新的措施、欧盟高风险人工智能系统数据库、后市场监测、信息共享和市场监督、行为守则和指南、授权和委员会程序、罚则等。经欧洲议会议长和欧洲理事会主席签署后，法案在欧盟官方公报上公布，并在公布二十天后生效。新法规将在生效两年后适用，但某些特定条款除外。¹⁰法案第 5 条“禁止的人工智能实践”的第 1 分段明确规定为防止恐怖袭击或恐怖袭击威胁可在公众可进入的场所为执法目的使用实时远程生物识别系统，“下列人工智能实践应被禁止：(h)在公众可进入的场所为执法目的使用“实时”远程生物识别系统，除非这种使用相应是为下列目标之一所严格必要的：……(ii)防止对自然人的生命或人身安全构成确切、重大切紧迫的威胁，或防止真实存在或真实可预见的恐怖袭击威胁”。法案附件二“第 5 条第 1 分段第 1 项第(h)(iii)点所述刑事犯罪清单”指出：“第 5 条第 1 段第 1 项第(h)(iii)点所述刑事犯罪：恐怖主义……”。显然，法案认可人工智能系统在打击恐怖主义犯罪和犯罪威胁中的作用，但这不等于公约可适用于打击利用人工智能系统实施的恐怖主义犯罪。但法案规定的禁止的人工智能系统实践、高风险人工智能系统的监管，为人工智能应用在反恐领域的法律治理提供了指引。

⁹ European Parliament, POSITION OF THE EUROPEAN PARLIAMENT adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), https://www.europarl.europa.eu/doceo/document/TC1-COD-2021-0106_EN.html, visited on April 11, 2025.

¹⁰ European Council, Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI, <https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/#:~:text=Today%20the%20Council%20approved%20a,society%2C%20the%20stricter%20the%20rules>, visited on April 11, 2025.

欧洲委员会《人工智能与人权、民主和法治框架公约》（Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, CETS No.225）是全球首个具有法律约束力的人工智能国际条约。2024年5月17日，欧洲委员会通过该公约。公约由欧洲委员会46个成员国、欧盟以及美国等11个非欧洲国家作为观察员国共同起草，68名私营部门、民间社会和学术界代表也作为观察员参与其中。2024年9月5日，欧洲委员会正式举行公约签署仪式。¹¹公约仅开放供欧洲委员会成员国、参与起草的非成员国和欧洲联盟签署，并应于五个签署国（包括至少三个欧洲委员会成员国）通过批准、接受或核准的方式，同意接受本公约约束之日起三个月期满后的下一个月1日起生效。美国、英国、安道尔、格鲁吉亚、冰岛、挪威、摩尔多瓦、圣马力诺和以色列等国家已成为首批签署国。公约生效后，欧洲委员会在与本公约缔约方协商并征得一致同意后，可邀请未参与拟订本公约的欧洲委员会非成员国以多数决定的方式加入本公约。¹²公约旨在确保人工智能系统的使用符合人权、民主和法治的基本原则，其适用范围广泛，涵盖公共和私营部门。公约包含八个章节，涉及一般义务、保证透明度和隐私等一般原则、救济措施、不良后果的评估与缓解、条约的适用机制、涉及国家安全时可适用的例外条款，建立了涵盖人工智能系统整个生命周期的法律框架。其中，第9条“问责与责任”规定缔约国应对人工智能系统造成的不利影响问责和承担责任，可能适用于打击利用人工智能系统实施的恐怖主义犯罪。但与欧盟《人工智能法案》相比，该公约更侧重于广泛的原则，而非具体的执行策略，且在实施细则上尚不明确，缔约方需采取具体措施，以确保人工智能系统的使用符合维护人权、保障民主机构独立性以及保护民主进程的要求。¹³

（四）国际软法规范产生一定的法律效果

联合国及相关国际会议或平台等也相继出台了一系列与打击恐怖主义和网络犯罪相关的国际法律文件，如《联合国全球反恐战略》等。它们虽然不具有法律约束力，但对各国制定相关政策和法律具有一定的指导作用。

1994年，联合国大会通过的《消除国际恐怖主义措施宣言》，明确指出恐怖主义是“无论在何种情况下均不可辩护的犯罪”，首次在全球范围内以国际共识的形式确认恐怖主义的不法性。1996年联合国大会对该宣言进行了补充，进一步强调各国必须采取切实有效的措施，防止恐怖分子以寻求庇护或获得难民身份为掩护，从事各种形式的恐怖活动。2006年联合国大会通过的《联合国全球反恐战

¹¹ COE, CM (2024)52-addprov,

https://search.coe.int/cm/Pages/result_detAlls.aspx?ObjectId=0900001680af0734, visited on April 11, 2025.

¹² COE, Council of Europe adopts first international treaty on artificial intelligence,

<https://www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence>, visited on April 11, 2025.

¹³ COE, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, <https://rm.coe.int/1680afae3c>, visited on April 11, 2025.

略》(A/RES/60/288)是加强国家、区域和国际反恐努力的一项全球性文书。该战略每两年审议一次实施情况,2014年第68届联合国大会第4次评审首次写入打击网络恐怖主义的内容;2023年第78届联合国大会第8次审议通过决议强调应均衡实施战略“四大支柱”:消除助长恐怖主义条件的措施、防止和打击恐怖主义的措施、建立各国防止和打击恐怖主义的能力以及加强联合国系统作用的措施、确保把尊重人权和实行法治作为反恐斗争根基的措施。

2009年,位于爱沙尼亚首都塔林的北约军事团体 The Cooperative Cyber Defense Center of Excellence (NATO CCD COE) 邀请独立的国际专家组拟定网络战手册,即《塔林网络战国际法手册》(the Tallinn Manual identifies the international law applicable to cyber warfare),旨在使现有法律适用于网络战的非约束性文件。2017年升级到2.0版《关于可适用于网络行动的国际法的塔林手册2.0》(Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations) 将仅适用于处理网络战争的规范,拓展到如何自和平时开展网络行动。《塔林手册2.0版》规则98规定“禁止在平民居民中散布恐怖为主要目的的网络攻击或网络攻击威胁”。规则评注指出:“规则98的文本仅包括实施或威胁实施网络恐怖攻击。然而,运用网络手段传播动能攻击威胁,且主要目的在于威胁平民居民,同样为武装冲突法所禁止。”¹⁴

除了具有代表性的文件外,还有众多国际软法规范也在各自的领域发挥着作用。它们有的侧重于规范网络空间的行为准则,有的致力于促进各国在情报共享、技术合作等方面的交流与协作,共同为打击网络恐怖主义构建起了一道坚实的防线。这些软法规范的存在,让各国在制定国内相关政策和法律时有了参考依据,能够更好地结合本国实际情况,制定出符合自身需求的反恐措施。但是,由于上述宣言、战略、手册不具有法律拘束力,性质上属于“软法”,仅具有法律效果。因此,无法有效规制人工智能时代的网络恐怖主义犯罪。

三、人工智能时代网络恐怖主义犯罪国际法规制的问题

现有国际法规则体系在应对人工智能时代网络恐怖主义犯罪方面存在明显不足,主要体现在以下几个方面:

(一) 传统打击恐怖主义国际刑事规则的缺乏针对性

虽然目前联合国框架下虽已存在13项反恐公约,但并没专门针对打击网络恐怖主义的专项国际公约。现有打击恐怖主义国际刑事规制的适用缺乏针对性。

传统打击恐怖主义国际刑事规则规制的重点仍集中于以传统方式如设置爆炸装置、利用核设施实施恐怖主义行为或为上述行为提供资助等,对于新兴的网络恐怖主义行为明显准备不足,难以有效规制利用人工智能技术实施的恐怖主义犯罪。《制止恐怖主义爆炸的国际公约》主要针对的是传统的恐怖主义爆炸

¹⁴ 《塔林手册2.0版》规则98评注第6段。

行为，对于利用人工智能技术操控爆炸装置，实现精准定位和远程触发，从而实施恐怖主义爆炸的情况没有规定。¹⁵在这种新型犯罪模式下，恐怖分子可以利用人工智能对目标进行数据分析，选择最具破坏力的爆炸时机和地点，公约中的罪名无法准确界定此类行为。《制止向恐怖主义提供资助的国际公约》在应对利用人工智能技术进行资金募集和转移的恐怖主义行为时也显得力不从心。人工智能可通过复杂的算法和加密技术，隐藏资金的真实来源和去向，使传统的资金追踪手段难以奏效。一些恐怖组织利用人工智能创建虚拟货币交易平台，通过智能合约进行资金的快速转移和洗白，超出了公约中资助恐怖主义罪名的涵盖范围。《制止核恐怖主义条约》同样无法有效规制利用人工智能技术威胁核设施安全的犯罪行为。人工智能可对核设施的控制系统进行攻击，干扰其正常运行，甚至篡改关键数据引发核事故。而条约中对于核恐怖主义行为的定义和罪名，并未考虑到这种利用人工智能技术的新型威胁。

传统规则难以适应人工智能技术带来的变化，在利用人工智能实施的恐怖主义犯罪的调查、追踪、取证等中可适用性差。人工智能技术的应用使得恐怖分子的攻击手段更加智能化和自动化，传统的安全检测和追踪技术很难及时发现和应对。恐怖分子利用人工智能生成的恶意软件能够自我学习和进化，不断躲避安全软件的检测。传统的基于特征匹配的检测技术，对于这种动态变化的恶意软件往往无能为力，导致无法及时追踪到恐怖分子的踪迹。人工智能技术使得证据的获取和固定变得异常困难。利用人工智能生成的虚假信息和伪造的电子证据，具有极高的逼真度，难以被辨别真伪。一些恐怖组织利用人工智能合成逼真的视频和音频，用于传播恐怖主义思想或制造混乱，这些虚假证据会干扰调查人员的判断，误导调查方向。且人工智能技术还可以对电子证据进行加密和隐藏，使得取证工作面临巨大挑战。传统规则中关于证据收集和认定的规定，在面对这些新型证据时，缺乏有效的应对措施。

（二）网络犯罪国际刑事规则碎片化显著

在全球化的背景下，网络恐怖主义活动往往跨越国界，需要各国携手合作才能有效打击。然而，由于网络犯罪国际刑事规则的碎片化显著，各国在处理网络恐怖主义案件时，缺乏统一的法律标准，往往依据区域性条约及本国的法律和政策，这不仅容易导致法律适用的冲突和混乱，也不利于国际间的合作与协调。《联合国打击网络犯罪公约》《网络犯罪公约》等网络犯罪国际刑事规则并未设置专门针对网络恐怖主义犯罪的条款，其他区域性公约规定给予简单，难以有效应对利用人工智能技术实施的恐怖主义犯罪。

现有区域性打击网络犯罪国际公约中，只有《上海合作组织成员国保障国际信息安全政府间合作协定》及其附件和《阿拉伯打击信息技术犯罪公约》直接涉

¹⁵ 参见王秀梅、魏星星：《打击网络恐怖主义犯罪的法律应对》，《刑法论丛》2018年第3期，第39页。

及网络恐怖主义犯罪，但是，由于它们的规定过于简单和原则性，并不足以应对网络恐怖主义的主要问题。《上海合作组织成员国保障国际信息安全政府间合作协定》及其附件规定存在一定的问题：一方面，信息恐怖主义是否等于网络恐怖主义存在一定的疑问；另一方面，即便将信息恐怖主义等同于网络恐怖主义，协定也存在较大的问题——信息恐怖主义的犯罪构成有哪些？信息恐怖主义的刑罚配置应遵循何种原则或规则？《阿拉伯打击信息技术犯罪公约》第 15 条的规定与该协定附件 2 类似，明确了网络恐怖主义的内容，但根据该条款，成员国同样难以确定网络恐怖主义犯罪的全部犯罪构成因素，如网络恐怖主义的主观方面；成员国也难以明确网络恐怖主义刑罚配置相关问题。。

《网络犯罪公约》并未专门对网络恐怖主义进行规制，但是，欧委会主张公约第 4 和第 5 条的规定足以应对网络恐怖主义问题，公约委员会也主张“公约中的实体罪名可能被用来促进恐怖主义、支持恐怖主义，包括财务或预备行为”。¹⁶事实上，公约这两条仅笼统地对破坏计算机数据和严重妨碍计算机系统的行为进行了规定。公约的实体犯罪条款中可能涵盖利用计算机数据和计算机系统或人工智能系统、未经授权且故意实施的、具有严重损害的网络恐怖主义犯罪。网络恐怖主义犯罪的主要内容虽然是干扰计算机系统和计算机数据的犯罪，却并不仅限于此，宣传、组织和资助网络恐怖主义也是网络恐怖主义犯罪的内容。¹⁷“宣扬恐怖主义思想、煽动暴力恐怖活动的行为是否受本公约规制？它们是恐怖活动初始阶段的行为，不像暴力恐怖活动那样直接造成严重的有形损害后果，但它们对包括暴力恐怖活动在内的各种恐怖活动行为起到重要的引起、促成或支持作用。

18

《联合国打击网络犯罪公约》也没有专门规制网络恐怖主义犯罪的条款。公约谈判中是否有必要纳入似乎存在疑问，且最终案文删除了“与恐怖主义相关的犯罪”。例如第一届会议各方意见中，一方面，中国、俄罗斯、印度尼西亚、约旦、科威特等国明确指出应在新公约中对网络恐怖主义进行规制。缔约国应采取必要的立法和其他措施，将利用信息和通信技术实施、招募人员实施、煽动、参与恐怖主义活动；宣传恐怖主义和为恐怖主义辩护；为恐怖主义活动筹资募集或提供资金确定为犯罪。另一方面，新西兰和瑞士明确反对在新公约中涉及网络恐怖主义，指出“没有必要仅仅因为这些犯罪可能是使用信息和通信技术进行的而予以重复涵盖。这种做法有矛盾和混乱的风险，不会产生一个提高我们处理网络犯罪的集体能力的有针对性的实用文书。”事实上，因恐怖主义国际法治本身的

¹⁶ Cybercrime Convention Committee (T-CY), T-CY Guidance Note #11 (DRAFT) Aspects of Terrorism Covered by the Budapest Convention, available at <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ab61b>, visited at 23 Nov. 2024.

¹⁷ 本文将在下文中明确网络恐怖主义犯罪内涵和外延，此处的论述以下文的分析为基础，在此不赘述。

¹⁸ 皮勇：《全球化信息化背景下我国网络恐怖活动及其犯罪立法研究》，《政法论丛》，2015 年第 1 期

矛盾，加之网络恐怖主义的急剧变化，各国在是否设置专门条款及如何设计条款内容争议较大，也正是这些原因导致公约管理机构为顺利促成公约的生效删除了“与恐怖主义相关的犯罪”条款。

（三）人工智能国际规则的适用不确定性大

目前，已通过的人工智能国际规则内容庞杂且规定的多是原则性规则，缺乏具体的实施细则和操作指南，在实际应用中难以准确把握和执行。具体适用还要依赖各国进一步地立法，存在较大的不确定性，难以有效规制人工智能技术在打击网络恐怖主义犯罪中的滥用。加之，人工智能技术发展迅速，新的应用场景和问题不断涌现，现有的人工智能国际规则往往难以跟上技术发展的步伐，导致在应对利用人工智能的网络恐怖主义犯罪时，缺乏有效的规范和约束。

《人工智能法案》的许多条款反映了目前国内和国际社会中一些进程倡导的人工智能治理原则，包括提高人工智能运行与决策过程中的透明度、保证人类监督、坚持公平与非歧视等。比如，在坚持公平与非歧视上，法案通过禁止使用进行社会评分的人工智能，并将用于分配入学名额、举行招聘选拔等涉及教育、就业的人工智能系统纳入高风险名单，保证了各类群体不会因为性别、宗教等因素受到歧视，或失去公平获得社会资源的机会。但其更像一个法律框架，其中有大量条款都需要欧盟委员会和其他授权机构进行二级立法来进一步发展和补充。如，第六条“高风险人工智能系统的分类规则”，委员会被授权在 18 个月内明确本条规定实施的指导方针，并制作高风险和非高风险人工智能系统的实际案例清单。通过这种方式，保留了一定的灵活性，使该法能不受暂时无法解决的问题的拖延而及时推出，保证了法律的及时性，同时又能适应后续的技术发展。

《人工智能与人权、民主和法治框架公约》是以欧洲为代表的西方国家在人工智能国际法领域的首次立法尝试，具有鲜明的欧洲人工治理特色和浓厚的西方法律价值色彩。公约具有显著的框架性特征，规定了人工智能系统生命周期内活动的共同一般原则和规则。它并未详尽列举人工智能生命周期内的特定活动，留待缔约方自行决定或根据后续文书补充。在全球人工智能治理日益碎片化、各国分歧不断扩大的情况下，相对模糊的规则有助于消减各国对公约可能阻碍本国技术发展的担忧，推动达成全球人工智能治理理念的最大公约数。但框架性与灵活性同时也意味着法律约束力的减弱，可能导致公约缺乏执行力。公约试图兼顾技术安全与创新，发展人工智能监管的关键是安全与创新的平衡。除了对透明度、数据保护、问责制等一般原则的规定，特别设置了“安全创新”条款，呼吁各缔约方酌情在其主管当局的监督下建立开发和测试人工智能系统的受控环境，在促进创新的同时避免对人权、民主和法治产生不良影响。公约同样将具体制度安排留给缔约方决定，期待缔约方通过“监管沙盒”等方式，在人工智能的设计阶段就纳入质量、隐私和其他问题的考量，在主管当局的控制之下推动人工智能技术的开发

与应用。公约依然具有强烈的重监管色彩。在谈判中，美国曾坚持将私营部门排除在公约规制范围之外，体现了其对重监管可能阻碍技术与应用的担忧。最终通过的公约采用了折衷方案，即公约本身只适用于公共实体和代表其行事的私人行为者的行为，但缔约方必须声明其意图对私营部门施加的义务。各国可以随时出于国家利益的考量修改声明，调整对公约的承诺，这大大削弱了公约对国家的约束力。¹⁹

（四） 相关国际软法的局限性明显

尽管相关国际软法规范在打击网络恐怖主义方面发挥了重要作用，但由于其内容和性质等方面的问题，在规制利用人工智能系统实施的网络恐怖主义犯罪时，存在明显的局限性。

国际软法的内容往往比较模糊和宽泛，权利、义务和具体的实施措施规定不明确，缺乏可操作性。例如，《联合国反恐战略》虽然呼吁各国加强合作打击网络恐怖主义犯罪，但对于如何进行合作、合作的具体方式和内容等却没有明确规定，导致在实践中各国的合作缺乏可操作性。不同国家和地区对于网络恐怖主义的定义、认定标准及应对措施存在差异，可能使各国在适用相关国际软法时存在理解和执行上的偏差，难以形成统一的行动标准和规范。以人工智能生成的虚假信息为例，有些国家认为只要传播了这类虚假信息就构成网络恐怖主义行为，而另一些国家则需综合考虑更多的因素，如传播的范围、造成的影响等。标准的不统一，使国际软法规范在协调各国行动、形成统一的打击合力方面面临困难。加之，国际软法规范的实施主要依赖于各国的自愿遵守和合作，缺乏有效的监督和执行机制，违反之并没有明确的法律责任和相应的制裁措施，使得软法规范在实际执行过程中大打折扣，难以真正发挥出应有的威慑作用。如对利用人工智能技术在网络上散布恐怖主义宣传视频、招募成员的行为，如果相关国家没有积极采取措施进行制止，国际社会很难依据软法规范对其进行有效的问责。

相关国际软法规范难以应对快速发展的人工智能技术带来的新挑战。人工智能技术日新月异，网络恐怖主义分子利用人工智能实施犯罪的手段也在不断翻新。而软法规范的制定和更新往往需要较长的时间，难以跟上技术发展的步伐。例如，虽然《联合国全球反恐战略》每两年审议一次实施情况，但其更新速度远远跟不上网络恐怖主义的发展速度。随着人工智能技术在图像和视频合成领域的应用越来越广泛，恐怖分子可以利用这些技术制作出更加逼真的虚假恐怖袭击视频，从而制造社会恐慌。对于这种新型的网络恐怖主义行为，《联合国全球反恐战略》无法及时提供切实可行的应对策略。

四、 人工智能时代网络恐怖主义犯罪国际法规制的路径

¹⁹ COE, Council of Europe adopts first international treaty on artificial intelligence, <https://www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence>, visited on April 11, 2025.

为了有效应对人工智能时代网络恐怖主义犯罪带来的挑战，国际社会需要加强合作，推动构建专门的法律机制、完善现有国际规则、强化国际软法与硬法的动态衔接。

（一）加强完善现有国际规则体系

对现有的国际规则进行修订和完善，明确将网络恐怖主义犯罪纳入其中。例如，在相关国际公约中，增设网络恐怖主义罪名条款，详细列举利用人工智能进行网络攻击、虚假信息传播、宣扬恐怖主义思想、煽动暴力恐怖活动等新型恐怖主义行为的具体表现形式。同时，加强对人工智能技术应用的规范和监管，制定专门的技术标准和安全规范，防止人工智能技术被恐怖分子滥用。以利用网络或人工智能系统宣扬恐怖主义思想、煽动暴力恐怖活动为例，随着网络的普及，恐怖主义思想的传播变得更加容易和迅速，通过网络平台宣扬恐怖主义思想，可能会蛊惑更多人加入恐怖组织，从而引发暴力恐怖活动。明确此类罪名，能够更有针对性地打击恐怖主义思想的传播，从源头上遏制恐怖主义犯罪的发生。煽动暴力恐怖活动的行为同样具有极大的危害性，它直接促使人们实施暴力恐怖行为，对社会的安全和稳定造成严重威胁。将其明确为罪名，能够让法律的制裁更加精准和有力。

在明确罪名的同时，详细界定网络恐怖主义犯罪的构成要件和适用范围也至关重要。以对于宣扬恐怖主义思想罪为例，应明确规定宣扬的方式、内容以达到何种程度构成犯罪。例如，通过网络发布宣扬恐怖主义思想的文章、图片、视频等，且传播范围达到一定程度，或对特定人群产生了明显的蛊惑效果，就应认定为犯罪。而对煽动暴力恐怖活动罪则需明确煽动的行为方式、对象及煽动的具体内容。例如，通过网络聊天工具、社交媒体等向他人鼓吹实施暴力恐怖活动，或提供暴力恐怖活动的策划方案等，都应纳入该罪名的范畴。

（二）推动建构专门的法律机制

随着网络恐怖主义威胁的不断加剧，国际社会对统一法律规范的需求将越来越迫切。通过逐步推进建构专门的国际法律机制，可以促进各国在网络反恐领域的交流与合作，增进相互理解和信任。然而，建构专门的国际法律机制并非一蹴而就，各国政治体制、法律制度、利益诉求、对网络恐怖主义的认知和重视程度、技术水平等的差异，使共识难达成、机制难建立。如何确保拟建构的专门的国际法律机制具有前瞻性和适应性，是需要深入思考和研究的问题。

建构全球性法律机制的困难不言而喻，且短期内很难达成；但在全球性网络犯罪国际刑事规则框架下，借助网络犯罪国家公约谈判的“东风”推进网络恐怖主义的国际法治是非常必要的。虽然《联合国打击网络犯罪公约》删除了“与恐怖主义有关的犯罪”，但鉴于包括中国在内的很多国家认为应明确规定网络恐怖主义犯罪，公约“特委会”已经将起草公约的补充议定书纳入工作计划，在“特委会”

框架下讨论设计网络恐怖主义犯罪罪名条款具有可行性。可在现有意见基础上，综合俄罗斯、约旦和科威特等支持在新公约中对网络恐怖主义进行规制国家的意见，指出“缔约国应遵守本国缔结的有关防止和惩治恐怖主义行为的国际法律文书的义务，对传播、支持或宣传恐怖主义意识形态和为恐怖主义目的使用信息和通信技术的行为进行规制”。另外，仅关注表层网的恐怖主义犯罪不具有前瞻性，且各方对于网络恐怖主义入罪的分歧可能对网络公约的生效造成阻碍。应强调具有强危害性和隐匿性的暗网恐怖主义的国际法规制。暗网恐怖主义是在暗网上实施的恐怖主义犯罪，其较表层网上的恐怖主义犯罪行为更具有特殊性。暗网恐怖主义作为恐怖主义和暗网犯罪的结合体，既具有恐怖主义的强危害性，又具有暗网犯罪的强隐匿性。“暗网”中继节点分散在世界各地，范围极广且拥有大量使用者。不仅如此，在暗网上违法交易种类繁多，以各类数据交易、走私军火、贩卖情报、伪造身份、毒品交易、秘密刺杀、淫秽色情为主，暗网的存在为这些非法交易提供了天然的掩体，给个人、企业、国家、国际社会造成极大的危害。然而，暗网的匿名性使得各国执法机构很难追踪恐怖组织。因此网络恐怖主义罪名设置仅作一般性规定不足以应对如此具有特殊性的暗网恐怖主义犯罪，需专门进行强调。综合支持在新公约中对网络恐怖主义进行规制国家的意见，指出“缔约国应遵守本国缔结的防止恐怖主义行为的国际文书的义务，对通过表层网和暗网实施的传播、支持或宣传恐怖主义意识形态和为恐怖主义目的使用信息和通技术的的行为进行规制”。

另外，区域性国际法律机制的推进具有明显的优势。同一区域内的国家往往在地理位置、文化背景、政治体制、经济发展水平等方面具有一定的相似性，利益诉求也更为接近，这使得它们更容易在网络反恐问题上达成共识。例如，欧盟国家在政治、经济和文化等方面有着紧密的联系，它们在网络反恐领域的合作相对较为顺畅。欧盟通过制定一系列的指令和框架，加强了成员国之间在网络安全信息共享、执法合作等方面的协作，取得了一定的成效。区域性机制推进在实践操作中也更加灵活和高效。由于参与的国家数量相对较少，沟通和协调成本较低，能够更快地制定和实施相关政策和措施。区域组织可以根据本地区的特点和需求，制定针对性更强的网络反恐机制，更好地应对本地区面临的网络恐怖主义威胁。区域国家可以针对本地区网络恐怖主义活动的高发领域，批判地借鉴上述传统恐怖主义犯罪国际刑事规则和网络犯罪国际刑事规则的相关规定，制定专门的应对措施，提高打击的精准性和有效性。区域性机制还可以作为试验田，为全球性机制的建立积累经验。通过在区域内先行先试，不断完善网络反恐机制，为未来全球性机制的建构提供有益的参考和借鉴。

（三）强化软法与硬法的动态衔接

软法和硬法各有其独特的作用和价值：软法能够提供原则性的指导和广泛的

共识，促进国际合作的开展；硬法则具有法律约束力，能够对网络恐怖主义犯罪行为进行有效的规制。强化软法与硬法的动态衔接，能够充分发挥两者的优势，提高人工智能时代网络恐怖主义犯罪的规制效果。

重视国际软法规范的指导作用，确保硬法的规定与国际社会的普遍共识和价值取向相一致。在制定打击网络恐怖主义犯罪国际规则时，可以先由基于相关国际软法提出一些原则性的建议和框架，形成初步的共识。再引导各方在行业规范、国际倡议等软法内容和实践经验基础上，将软法中成熟有效的规则和做法上升为具有强制约束力的法律条款，使其更具可操作性和适应性。例如，在制定打击网络恐怖主义资金流动的国际规则时，借鉴金融行业协会制定的反洗钱、反恐怖融资的行业规范，明确金融机构在监测和报告网络恐怖主义资金交易方面的义务和责任。

重视非国家行为体的作用，及时发现和解决软法与硬法在衔接过程中出现的问题，不断完善国际法律机制。非国家行为体凭借在各自领域的专业优势，为完善国际法律机制提供了丰富的专业知识与技术支持。在国际立法中，非国家行为体凭借专业知识和广泛的社会联系，积极参与规则的讨论与拟定，使法律更贴合实际需求；在国际执行中，非国家行为体可监督法律的实施情况，及时反馈问题，促进法律有效执行。卡巴斯基实验室长期关注网络恐怖主义犯罪的发展态势，通过对大量网络攻击事件的监测和分析，深入研究网络恐怖主义犯罪的手段、特点和规律，定期发布网络安全报告详细披露网络恐怖主义犯罪的最新动态和威胁情况，为各国政府、企业和组织提供了重要的情报信息，帮助他们及时采取防范措施，降低网络恐怖主义犯罪的风险。国际互联网安全协会建立了网络安全信息共享数据库，收集和整理全球范围内的网络安全事件和威胁情报，为会员单位提供及时准确的信息服务，帮助他们更好地应对网络恐怖主义犯罪的威胁。一些科技企业和科研机构合作开展研究项目，研发出了一系列先进的网络安全技术和工具，如人工智能驱动的入侵检测系统、区块链技术在网络安全中的应用等。这些技术和工具能够实时监测网络流量，及时发现和预警网络恐怖主义攻击行为，提高了网络安全防护的能力和效率。这些非国家行为体的专业知识和技术支持，使国际法律规则在制定和实施过程中更加科学、合理，能够更好地适应复杂多变的国际环境和技术发展需求，提高国际法律机制应对各类问题的能力。总之，非国家行为体为打击网络恐怖主义犯罪提供了有力的支持，国家、政府间国际组织应加强与非国家行为体的合作。

五、结语

人工智能时代网络恐怖主义犯罪对国际社会安全构成严重威胁，需要国际社会加强合作，共同应对。作为负责任的大国，中国应积极参与人工智能时代网络恐怖主义犯罪国际法规制的构建，并提出中国方案，为维护国际网络安全贡献力

量。具体地，中国可从国际和国内两个层面同时推进。国际层面可确立两步走战略：第一步，推动国际立法机制纳入网络恐怖主义犯罪条款：在联合国打击网络犯罪国际公约附加议定书中写入网络恐怖主义犯罪条款，明确相关犯罪行为和缔约国的合作义务；第二步，推动在区域组织框架下制定打击网络恐怖主义犯罪的区域规则，形成区域性合作机制，如在亚太经合组织、上海合作组织等区域组织框架下，推动制定打击网络恐怖主义犯罪的区域规则，加强区域合作。国家层面：建构国内网络恐怖主义犯罪法律规则体系：完善现有法律法规，制定专门针对网络恐怖主义犯罪的法律法规，明确相关犯罪行为的界定、法律责任、执法程序等；加强网络安全基础设施建设，提升网络安全技术水平，培养网络安全专业人才，提高应对网络恐怖主义犯罪的能力；积极参与国际社会打击网络恐怖主义犯罪的合作，分享经验，贡献力量，共同维护国际网络安全。

应对美国网络霸权的中国涉外法律建设

----以美国 Cloud 法案为视角切入

左文君、吕梦飞^[1]

摘要：以美国《澄清境外数据的合法使用法案》（简称《云法案》，CLOUD Act）为切入点，分析《云法案》中相关条款，指出美国以“国家安全”为名对中国实施打压的法律逻辑与政治意图。以美国对我国华为、TikTok等科技企业的打压为例，利用国内法干涉国际法律规则的运用揭示其通过国内法“长臂管辖”构建网络空间霸权的本质。文章结合国际法与国内法之间的互动关系，探讨中国在应对网络霸权的涉外法治建设中问题，我国应积极应对美国霸权，借鉴国际经验并结合中国国情，采取有利措施因应对挑战，从制定针对“长臂管辖”的具体反制规则、转变跨境数据治理立场方向、强化人才队伍建设，增强国际规则经验等方面出发，维护我国网络安全和数据安全。

关键词：美国网络安全 涉外法治 跨境数据 国内法

引言

在全球化与数字化的双重浪潮下，网络空间已成为大国博弈的新战场。作为全球数字技术的领导者，美国凭技术优势与法律工具，逐步构建起一套以“网络安全”为名的霸权体系。^[2]特朗普政府时期出台的对华科技政策，对我国信息通信、半导体制造等领域的发展造成巨大的负面影响；拜登政府上台后，在保留部分特朗普时期对华政策的同时，企图采取“小院高墙”政策，协同其盟友共同围堵、遏制我国科技发展，给我国高新科技领域的进一步发展带来了严峻的挑战。

美国 2018 年颁布的《澄清境外数据的合法使用法案》（简称《云法案》，CLOUD Act）成为其扩展网络空间管辖权的重要法律工具。该法案将美国“长臂管辖”进行了升级，传统“长臂管辖”以“最低联系原则”，如商业活动、侵权行为与美国存在关联为基础。但《云法案》直接将美国的管辖权扩展至“数据控制权”领域，一方面，美国的属地管辖获得突破，即只要数据控制者受美国法律管辖无论数据存储于何地，美国政府均可直接调取；另一方面，美国的属人管辖得到延伸，即使数据控制者是外国公司，但若其在美国有业务或服务器，仍可能被纳入管辖范围。《云法案》导致了数据主权边界的消解，它通过将“数据控制权”等同于“主

^[1] 左文君 湖北经济学院法学院副教授；吕梦飞 湖北经济学院法学院硕士研究生。

^[2] 周琪. 高科技领域的竞争正改变大国战略竞争的主要模式[J]. 太平洋学报, 2021, 29(01): 1-20. DOI:10.14015/j.cnki.1004-8049.2021.01.001.

权管辖权”，将物理领土的司法边界虚化为“数据控制者国籍”，实质是以企业国籍替代传统主权边界，为跨境执法提供合法性借口。此外，《云法案》中体现体现“长臂管辖”与网络霸权的核心条款包括：第 103 条数据跨境调取的单边授权、第 105 条双边行政协议机制以及第 2713 条数据控制者的合规义务。

目前，针对美国以“国家安全”为名的网络霸权扩张，国内外学者展开多维度探讨。国内研究主要聚焦四大方向，一是美国网络霸权政策演变，学者余丽（2023）、韩梦阳（2023）等指出，美国从特朗普政府的“单边主义”到拜登政府的“盟友协同”，策略虽变但维护霸权的本质未改，差异体现于治理模式与意识形态渗透。宫云牧（2023）进一步分析美国网络安全战略从被动防御转向国际威慑的阶段性特征。二是网络霸权对中国法律反制能力的冲击，孙祁、栾文韬（2024）揭示美国通过“法律战”将霸权工具化，迫使中国涉外法治面临规则压制困境；韩娜（2024）强调美国以技术优势巩固网络霸权，倒逼中国强化法律反制能力。三是美国对中国科技企业打压的实证研究，景舟（2025）、高望来（2024）等以《美国人工智能脱钩法案》、芯片管制及 TikTok 封禁为例，剖析美国通过立法、联盟和技术断供遏制中国科技发展的霸权逻辑。王智珑（2024）指出美国对华数字政策的“泛安全化”本质。四是应对美国网络霸权的涉外法律路径探索，学者聚焦中美法律冲突核心领域，如朱劲草（2023）批判美国《云法案》的“长臂管辖”，杨彦蕾（2022）反思国内网络安全法实施障碍，陈思琦（2024）呼吁构建数据主权规则以平衡自由流动与安全。

而国外研究关注中国法律实践成效与挑战，国外学者普遍承认中国通过《网络安全法》、《数据安全法》初步构建了网络主权法律框架，但认为其国际规则影响力仍待提升。Federico Fabbrini（2021）的《跨境数据保护：跨大西洋视角下的治外法权与国家主权》一书结合欧美数据领域的最新立法、判例探讨了在数据立法域外适用背景下欧美的紧张关系与合作空间^[3]。Adèle Azzi,何叶华（2021）认为 GDPR 的域外适用范围尽管相当广泛,其符合国际判例的默示承认和国际习惯法原则,具备国际法正当性基础。^{[4][5]}总的来说，国内外研究均揭示中美网络规则博弈的复杂性与长期性。中国涉外法律建设需兼顾主权维护与国际协同，既要抵御美国霸权渗透，又需通过多边合作提升规则话语权，构建更具包容性的全球网络治理秩序。

国内外学者围绕美国网络霸权与中国涉外法律建设，虽然已形成较为系统的研究框架，分别进行政策演进分析、霸权冲击与实证研究以及法律应对路径探索，但其仍有研究不足之处，一方面，对于规则重构的路径比较模糊，现有研究多集

^[3] Federico Fabbrini, Edoardo Celeste, John Quinn. Data Protection Beyond Borders: Transatlantic Perspectives on Extraterritoriality and Sovereignty [M]. New York: Hart Publishing, 2021: 1~280.

^[4] Adèle Azzi. The Challenges Faced by the Extraterritorial Scope of the General Data Protection Regulation [J]. Journal of Intellectual Property, Information Technology and E-Commerce Law, 2018, (2)

^[5] 何叶华.论数据保护法的域外效力[J].北京理工大学学报(社会科学版),2021,(9).

中于“抵御霸权”，但对“如何主动改善中国涉外法治”缺乏深入探讨；另一方面，动态性与系统性研究不足，对美国霸权策略调整跟踪滞后。

本文旨在以《云法案》为中心，探讨美国如何利用网络安全议题对中国实施法律遏制，并分析中国在涉外法治与技术自主性建设中的应对路径。通过剖析华为、TikTok、DeepSeek 等典型案例，揭示通过“长臂管辖”构建网络空间霸权的本质，并为中国在涉外法治探索以及应对美国霸权主义上提出具体方法。

一、美国通过国内立法构建网络霸权——《云法案》的实质

近年来，霸权国家以“维护国家安全”为由对外恣意实施法律战的倾向日益明显，以法律战为代表的“软战争”已逐渐成为服务霸权和强权政治的新策略。^[6]在国际通行的数据跨境治理规则不足的情况下，美国于 2018 年 3 月施行了《澄清境外合法使用数据法》（Clarifying Lawful Overseas Use of Data Act，以下简称《云法案》）。该法案摒弃之前采取的数据属地管辖模式，转向“数据控制者”标准，规定美国可以根据最低联系原则获取位于美国境外的数据。^[7]其核心内容包括跨境数据管辖与国际合作模式。

（一）跨境数据管辖

1. 属地管辖主导

在《云法案》提出前，美国对于跨境数据的管辖主要以“属地主义”为主导，以数据存储的物理位置为核心，认为数据管辖应遵循存储地国家的法律。并且依赖司法互助条约（MLAT），即美国执法机构需通过繁琐的司法互助程序如双边条约或国际协商向外国政府申请调取境外数据。

2.“数据控制者”原则

在《云法案》提出后，美国以“数据控制者”为标准，认为只要数据由美国企业控制，无论数据存储于何处，美国执法机构都可以直接要求调取。美国通过《云法案》赋予自身跨境数据调取的直接权力，绕开传统国际司法合作框架。

（二）国际合作模式

1. 单边与被动

在《云法案》提出前，美国的跨境数据调取主要依赖国际合作，但程序耗时长，平均在 10 个月以上，难以应对网络犯罪和反恐的时效性需求。并且因为未建立直接的跨境数据调取渠道，企业常因不同法域的法律冲突陷入合规困境。

2. 构建“数据联盟”

在《云法案》提出后，允许美国与符合条件的国家签订行政协议，如美英 CLOUD 协定、美欧《数据隐私框架》、美欧《司法互助协议》等，实现了跨境数

[6] 孙祁,栾文韬.应对法律战视角下的涉外法治建设初探[J].和平与发展,2024,(05):1-24+179.

[7] 朱劲草.美国《云法案》中的数据域外管辖问题研究[D].华东政法大学,2023.DOI:10.27150/d.cnki.ghdzc.2023.000337.

据直接调取，缩短了执法流程。此外，“数据联盟”设定了合作门槛，要求协议国符合美国标准，如人权保障、法治水平等，变相的输出美国的价值观。

二、《云法案》中维护美国数据霸权的具体条款分析

美国《云法案》允许美国执法机构绕过用户同意和国家的数据主权，允许直接获取美国境外的数据，极大地扩张了数据域外管辖权的范围。^[8]该法案中的核心条款通过赋予美国政府跨境调取数据的单边权力、构建以美国标准为主导的合作框架，以及强化对企业的强制合规要求等，也集中体现了“长臂管辖”和网络霸权的特征。

（一）第 103 条：数据跨境调取的单边授权

1.具体内容

该条款授权美国执法部门直接要求美国企业提供其控制的数据，无论数据存储于美国境内还是境外。企业必须配合执法请求，除非数据存储地法律与美国法律存在“实质性冲突”。美国《云法案》所确立的“数据控制者准则”也不局限于国籍等传统属人连接因素，只要数据控制者与美国的联系符合美国判例法的“最低程度联系”原则，在境外成立的数据控制者也可被纳入适用范围。^[9]

2.条款实质

传统国际法以数据存储地即属地原则作为管辖权基础，但第 103 条以“数据控制者”身份即数据由美国企业控制为依据，将美国法律效力延伸至全球。此外，美国通过国内法单方面确立跨境数据调取权，绕开国际司法互助程序，实质是“国内法凌驾于他国主权”。

（二）第 105 条：双边行政协议机制

1.具体内容

允许美国政府与其他国家签订行政协议（Executive Agreements），使签约国执法机构可直接向美国企业调取数据，但需满足美国设定的条件，包括：协议国需具备“人权保障和法治标准”该标准由美国单方面定义；协议国不得限制美国对其境内数据的调取权；协议国需接受美国对合作范围的单方面审查。

2.条款实质

美国通过设定政治化门槛，将合作对象限定为盟友，形成以美国为中心的数据调取网络，强化其规则主导权，并且协议国需要承诺不限制美国的数据调取权，但美国无需对等放弃自身管辖权，形成“单向透明”的合作机制。

（三）第 2713 条：数据控制者的合规义务

^[8] 朱劲草.美国《云法案》中的数据域外管辖问题研究[D].华东政法大学,2023.DOI:10.27150/d.cnki.ghdzc.2023.000337.

^[9] U.S.Department of Justice.Promoting Public Safety, Privacy, and the Rule of Law Around the World: The Purpose and Impact of the CLOUD Act, White Paper[EB/OL].<https://www.justice.gov/opa/press-release/file/1153446/download>.

1.具体内容

明确要求美国企业及其子公司、关联公司必须遵守美国政府的执法要求，包括：提供境外存储的数据；配合解密通信内容（即使违反外国法律）；不得以“数据存储于境外”为由拒绝配合。

2.条款实质

通过强制企业合规，美国将跨国科技公司，如谷歌、Meta 等转化为其全球数据监控网络的执行者，而企业仅可在外国法律与美国法律存在“实质性冲突”时提出抗辩，但举证责任完全由企业承担，这在司法实践中极少能够成功。

三、美国霸权主义的典型案例

（一）华为事件

美国以“国家安全”为名对华为的全面打压，是中美科技博弈的标志性事件，其本质是遏制中国技术崛起、维护全球科技霸权。2017 年 12 月，特朗普政府发布的《国家安全战略》报告将中国明确称为“对手”和“修正主义国家”。这份报告是美国开启对华战略竞争的标志，其强调美国必须在研究和技术领域保持领先地位，通过更加强劲的研发创新“更新美国的竞争优势”，尤其是要更为重视“国家安全技术领域”。[10]美国的行动以“国家安全”为名，却缺乏公开证据，暴露出双重标准：一方面打压华为，另一方面通过《云法案》扩张自身数据主权。美国对华为打压时间线如表一所示。

时间	事件	关键措施/影响
2012 年 10 月	美国国会发布调查报告	指控华为、中兴“威胁国家安全”，建议禁止其参与美国通信基础设施建设。
2018 年 8 月	美国通过《2019 财年国防授权法案》（NDAA）	禁止美国联邦机构及其承包商使用华为、中兴等 5 家中企的通信设备。
2018 年 12 月	孟晚舟在加拿大被捕	美国以“违反伊朗制裁令”为由要求引渡，施压华为；2021 年 9 月孟晚舟与美达成协议获释。
2019 年 5 月	美国商务部将华为列入“实体清单”	禁止美企向华为出售技术（需许可证），切断华为与美国供应链（如谷歌停止 GMS 服务）。

[10] 赵明昊.技术鹰派、国家安全与美国对华战略竞争[J].国际安全研究,2025,43(01):115-132+158-159.DOI:10.14093/j.cnki.cn10-1132/d.2025.01.006.

2019 年 11 月	美国启动“清洁网络计划”	联合盟友排除华为 5G 设备，英国、澳大利亚、日本等国响应。
2020 年 5 月	美国升级芯片制裁	限制台积电等使用美国技术的企业为华为代工芯片，切断华为高端芯片供应。
2020 年 6 月	美国将华为列为“国家安全威胁”	联邦通信委员会（FCC）禁止美运营商使用联邦补贴购买华为设备。
2020 年 7 月	英国宣布全面移除华为 5G 设备	迫于美国压力，英国推翻此前“有限使用华为”政策，要求 2027 年前完成拆除。
2021 年 3 月	美国进一步限制华为获取 4G、Wi-Fi 6 等成熟技术	禁止美企向华为出口涉及 4G、Wi-Fi 6 等技术的产品，全面封锁其通信设备生产。
2021 年 11 月	美国通过《安全设备法》	禁止华为、中兴获得 FCC 设备认证，彻底阻断其在美国销售新设备的可能性。
2022 年 10 月	美国全面禁止华为、中兴在美销售设备	FCC 将华为、中兴列入“国家安全威胁清单”，并撤销此前颁发的设备授权。
2023 年 1 月	美国施压荷兰、日本加入对华芯片管制	限制 ASML 向中国出口 DUV 光刻机，进一步遏制华为芯片制造能力。
2023 年 8 月	华为发布 Mate 60 Pro 手机（搭载国产 7nm 芯片）	美国制裁倒逼中国突破技术封锁，引发美方对制裁效果的质疑，称将“采取更强硬措施”。

表一

（二）TikTok事件

2024 年 3 月 8 日，美国众议院能源和商务委员会（The House Energy and Commerce Committee）以 50:0 的投票结果，表决通过了一项专门针对 TikTok（海外版抖音）等外国应用程序的法案——《保护美国人免受外国敌对势力控制的应用程序侵害法》。^[11]该法案表示，字节跳动，即 TikTok 的中国母公司，需要 270 天内主动剥离出售该业务，否则将在美国地区应用商店强制下架 TikTok。同年 4 月 24 日，拜登宣布签署该法案，使其成为正式法律。美国官方将 TikTok 渲染为危害“美国国家安全”的威胁，并对其作出封禁决定。美国对 TikTok 的打压行为表明了其在数据管辖权上的双重标准，虽然美国没有直接引

[11] 王智珑.美国对华数字政策泛安全化主要手段与应对建议——以 TikTok 封禁案为例[J].海外投资与出口信贷,2024,(05):35-38.

用《云法案》，但是该行为的法律逻辑与《云法案》一脉相承。在美国看来，根据《云法案》，美国有权要求本企业（无论数据存储在何处）提交数据，但在指控 TikTok 时，却以“中国可能依据《中国国家情报法》获取数据”为理由，认为字节跳动无法拒绝中国政府的数据要求，这充分的表明，美国自身通过《云法案》合法化其跨境数据调取权，却将中国的类似法律框架视为“威胁”。美国对 TikTok 打压的时间线如表二所示。

时间	事件	核心内容	结果/影响
2017 年 11 月	字节跳动收购 Musical.ly	以 10 亿美元收购美国社交媒体平台 Musical.ly，后整合为 Tiktok 美国业务。未主动向 CFIUS 申报交易。	为后续美国外国投资委员会（CFIUS）追溯审查埋下隐患。
2019 年 11 月	CFIUS 启动国家安全审查	以“数据安全风险”为由，对字节跳动收购 Musical.ly 进行审查。	首次明确以国家安全名义对 Tiktok 施压，引发法律争议。
2020 年 8 月	特朗普签署行政禁令	发布第 13942 号行政令：要求字节跳动 90 天内剥离 TikTok 美国业务。 第 13943 号行政令：禁止美国企业与 TikTok 交易	Tiktok 起诉美国政府违宪，联邦法院暂缓禁令执行。
2021 年 6 月	拜登撤销特朗普禁令，扩大审查范围	撤销前任行政令，但要求 CFIUS 继续审查 Tiktok，并发布新行政令限制“外国对手控制应用”。	政策转向更隐蔽的技术性审查，延续对 Tiktok 的数据安全指控。
2023 年 3 月	美国国会听证会	众议院能源与商业委员会质询 TikTok CEO 周受资指控其“为中国政府操控美国舆论”，但未提供实证。	听证会被批为“政治作秀”； TikTok 用户发起 #SaveTikTok 抗议运动。
2023 年 4 月	蒙大拿州通过全美首个 Tiktok 禁令	禁止应用商店提供 Tiktok 下载，违者每日罚款一万美元。	2023 年 11 月被联邦法官以“侵犯言论自由”驳回，但引发多

时间	事件	核心内容	结果/影响
			州效仿限制政府设备使用。
2024 年 4 月	美国国会通过《保护美国人数据法案》	要求字节跳动在 270 天内剥离 TikTok, 否则应用商店下架该应用; 拜登签署法案生效。	TikTok 起诉美国政府违宪 (第一修正案和第五修正案), 法律战持续。

表二

四、中国当前应对网络霸权的现实困境

(一) 国内法缺乏具体反制规则

首先, 中国虽出台《数据安全法》、《个人信息保护法》、《网络安全法》、《关键信息基础设施安全保护条例》等法规, 但条款多为原则性规定, 缺乏针对“长臂管辖”的具体反制细则。一方面, 管辖依据的覆盖范围有限, 《个人信息保护法》虽通过“行为发生地域”与“目标指向性标准”确立域外管辖基础, 但其规制对象仅限于个人数据领域, 对企业经营数据、公共管理数据等非个人信息类型未作明确规范。而《网络安全法》《数据安全法》虽赋予执法部门跨境数据调取权, 但主要针对刑事司法协助场景, 未就行政执法中的域外数据管辖形成系统授权机制。另一方面, 对等原则的适用性存疑, 尽管《个人信息保护法》第 43 条与《数据安全法》第 26 条引入“对等原则”作为域外管辖依据, 但条文表述中既未明确该原则的启动标准, 亦未规定其在行政执法中的具体适用程序。其次, 中国强调“数据本地化”以保障安全, 但在推动数据跨境流动如加入 DEPA、RCEP 数字条款时, 尚未形成兼顾安全与发展的分级分类规则体系, 导致企业面临“过度合规”或“合规不足”的两难。数据本地化对不同类型的企业都会产生一定的负面影响, 针对数据依赖型新兴企业, 数据本地化法律削弱了企业生产经营的效率, 因为企业必须在一个特定的市场保留更多的人员和技术, 并降低它们在各国间提供无缝、一致的服务的能力。^[12]再次, 美国《云法案》通过后, 已与英国、澳大利亚等国签署 CLOUD 协定, 并不断扩展“数据盟友圈”, 但中国对美国行政协议的具体条款以及执行效果的跟踪研究较少, 导致反制措施缺乏针对性。一是在具体条款上, 美国与英国、澳大利亚等国签署的 CLOUD 协定中, 关于“人权标准”“法治门槛”等关键条款的定义和适用条件具有高度弹性; 二是在执行效果上, 美国、国通过 CLOUD 协定调取境外数据的频率、数据类型、涉及行业等缺乏实证分析, 导致

[12] 中央财经大学课题组. 数据本地化的重思与改进 [J]. 数据法学, 2024, 6 (01): 3-22.

中国无法量化其威胁程度。

（二）跨境数据治理以“弱防御”为主

中国在跨境数据治理国际谈判，如 WTO 电子商务谈判、APEC 跨境隐私规则等中仍以“防御性立场”为主，未提出系统性替代方案，难以对冲美国主导的“数据联盟”。例如，针对美国《云法案》虽名义上允许“符合特定标准的外国政府”调取美国境内数据，但“标准”的解释权与认定程序完全由美国政府单边掌控的情况，从我国立场出发，该法案赋予其执法机关宽泛的域外数据管辖权，严格限制我国对其域内数据的调取与访问，会侵害我国国家主权和管辖利益。但由于缺乏必要的反制措施，从总体上看我国仍处于“弱防御”状态。

（三）相关领域涉外法治人才的缺失

美国凭借《云法案》等法律工具推行“长臂管辖”，对我国数据主权、企业合规及司法独立构成严峻挑战。在此背景下，我国相关领域涉外法治人才队伍的结构短板日益凸显。

1. 专业化人才储备不足，跨领域协同能力薄弱

当前我国涉外法治人才多集中于传统国际公法领域，精通数据主权、数字贸易规则、跨境合规等新兴领域的复合型人才稀缺。在我国当前的法律课程设置中，教师几乎只针对我国法律规则进行讲述，对他国的法律规则、法律文化等理论知识涉及较少，缺少从全球性视角分析国际问题的意识与能力。^[13]据司法部数据显示，全国精通国际数据纠纷处理的律师不足 200 人，且多数缺乏技术背景，难以应对《云法案》等技术驱动型法律挑战。据统计，目前我国虽有近百所高校开设了国际法专业，但限于资金支持和地域发展，其中部分高校并没有足够的国际资源支持和培育涉外法治方面的人才，这也就导致了部分学生所学虽然是国际法相关专业，但并没有足够的国际视野与专业能力支撑其日后从事涉外法律相关工作。^[14]

2. 国际规则博弈经验匮乏，话语权建设滞后

涉外法治人才在国际组织及多边谈判中的参与度较低，削弱了我国规则制定的话语权。以 WTO 电子商务谈判为例，中国代表团中兼具数字治理与法律外交能力的成员占比不足 15%，远低于美欧 30% 的比例。这导致我国提出的“数据主权”主张常因缺乏配套技术标准与司法案例支撑而被边缘化。同时，熟悉美国司法程序与判例体系的人才稀缺，企业应对境外诉讼时多依赖高价外聘律师，2022 年华为、TikTok 等企业涉外法律支出超百亿元，暴露出本土人才实战能力短板。

中国应对美国网络霸权的路径选择

（一）制定针对“长臂管辖”的反制规则

[13] 贺赞. 涉外法治人才培养机制创新：以课程体系建设为中心[J]. 中国法学教育研究, 2017 (2): 3-11.

[14] 朴顺善. 新时代涉外法治人才培养路径研究 [J]. 黑龙江教育(理论与实践), 2024, (06): 25-32.

我国的涉外法律规范体系目前仍存在着大量薄弱环节,一些制度、规则较为简单、笼统,难以对实际工作展开提供有效的指引。^[15]为有效遏制他国数据管辖权的域外不当扩张,需构建系统化的法律阻断机制。当前我国在该领域的立法架构存在显著短板,商务部2021年颁布的《阻断外国法律与措施不当域外适用办法》虽初步确立阻断规则,但其规制范畴于经贸活动领域,且对“不当域外适用”的判定标准欠缺可操作性指引,难以应对数字时代的数据主权博弈。

制定数据领域的阻断立法,以加强阻断的针对性,为相关部门提供有效的指引,作为阻断机制的核心载体,该法应明确界定数据领域“法律域外适用”的违法边界,可以重点规制的行为包括:外国司法机构基于“数据控制者”理论强制调取中国境内存储数据;以“网络安全审查”为名要求企业披露在华业务敏感数据等,并且立法也需要设置梯度化阻断措施,从数据出境禁令、企业合规豁免,直至对实施管辖的外国主体实施市场准入限制。

实施分类分级的精准阻断策略,在数据安全治理重点领域与重点行业实现阻断规则的专门化和精细化,重点保障关乎国家安全与公共利益的数据免受外国法律管辖。^[16]对于阻断可以根据重要性进行分级分类,例如,将其分为三等级,一级为国家安全与公共利益以及直接关系国家安全、国民经济命脉、重要民生、重大公共利益的数据;二级为公共健康数据、金融行业敏感数据、重点行业运营数据等一旦遭破坏或泄露可能对公共利益、行业安全或社会秩序造成显著影响的数据;三级为个人消费记录、企业公开经营信息、公共领域非涉密数据等主要影响个人、企业合法权益的非敏感数据。对数据进行分级分类有利于相关部分进行资源凝聚,针对关键数据进行阻断防控,减少资源浪费。

(二) 转变跨境数据治理立场方向

针对我国跨境数据治理中存在的“被动防御”短板,可以从规则建构方面入手,从“防御跟随”到“标准输出”,构建兼具防御性与主动性的治理体系。与“防守型”的阻断立法相对,我国还应采取“进攻型”的应对策略,以更加精细化和灵活化的方式提升数据立法的域外适用性,为数据跨境的有序治理提供制度保障。^[17]针对美国以“国家安全”为名进行的网络霸权主义,我国应转变“软防御”状态,加强输出准则,统筹兼顾安全与发展。一方面数据要素是推动国家经济、科技、贸易发展的战略性资源,需要予以保护和关注,另一方面数据有别于其他要素,数据的共享、使用、流通有利于数据价值的发挥。^[18]首先,制定立足于我国国情的跨境数据治理规则,该规则可从美国《云法案》对我国带来的影响出发,分析国外相关

[15] 赵怿君,居晓林,田曜华.国际法律服务的现状与完善路径[J].法制博览,2025,(01):64-66.

[16] 陈俊秀,王纬航,李智.我国域外数据执法管辖面临的困境与因应[J].集美大学学报(哲社版),2024,27(05):27-34+92.

[17] 文淑.数据立法域外适用引发的法律冲突与中国解决方案[J].云南师范大学学报(哲学社会科学版),2023,55(06):96-108.

[18] 汪洋,郑晓欢,班艳,等.深化落实数据要素政策,建立健全科学数据生态圈[J].中国科学数据(中英文网络版),2023,8(1):134-145.

的跨境数据立法,例如亚太经合组织的跨境隐私规则(CBPR)、欧盟的《通用数据保护条例》(GDPR)、《美国-欧盟数据隐私框架》(DPF)等,学习他国关于跨境数据的法律规定,从而制定适合我国的规则,用于应对美国的网络霸权。其次,在现有法律上进行关键规则细化,例如我国现有的《数据安全法》、《个人信息保护法》、《网络安全法》、《关键信息基础设施安全保护条例》等,将这些法律规定中的原则性规定与实际相结合,对美国的“数据控制者”原则规定具体反制专章,在“防御”中“输出”,更好的保护我国的跨境数据安全。

(三) 强化人才队伍建设, 增强国际规则经验

1. 培养专业化人才, 增强跨领域协同能力

涉外法治人才培养是一项系统性、长期性工作,为保障我国涉外法治人才培养工作的持续性、稳定性,应尽快推动涉外法治人才培养的规范化标准化运行,制定完善相关国家标准,建立涉外法治人才培养“参照系”,为全国涉外法治人才培养提供明确、协调的规范指引。^[19]首先,增强相关领域法学人才培养,高校可设立跨学科人才培养项目,融合法学、数据科学、国际关系等专业,增加关于数据主权、数字贸易规则、跨境合规等新兴领域的课程,将理论知识与实践相结合,开展关于“法学+”的实践活动。其次,国家从立法政策方面建立激励平台,在立法政策上,不断加强重点领域、新兴领域以及涉外领域的立法工作,为涉外法治人才施展才能提供坚实的法律基础支撑。通过完善相关法律体系,使得涉外法治人才在处理各类法律事务时有法可依,为其专业实践筑牢根基。例如,2024年,教育部批准51家涉外法治人才协同培养创新基地,旨在整合多方资源,为培养高素质涉外法治人才搭建平台,可在此基础上扩展培养基地,加大培养力度,适应人才稀缺的现状。再次,给予法学复合型人才补助,一方面,对于在涉外法治领域取得突出学术成果,如在国际知名法律期刊发表高质量论文,或是在国际法律学术竞赛中获得优异成绩的学生与学者,给予专项奖学金和科研补助,以鼓励他们深入开展涉外法治研究。另一方面,对于积极参与涉外法律实践项目,如为中国企业海外投资、国际商事纠纷解决等提供法律服务的人才,按照项目难度与成效给予相应的实践补贴,提升其参与涉外法律实务的积极性。

2. 增强国际规则博弈经验, 建设话语权

当前尚不存在一个有约束力的统一的数据跨境流通治理机制,尚未形成统一的数据跨境流通规则和争端解决机制,缺乏能够调和国际数据跨境流通治理矛盾的有强制约束力的措施。但现实中各国利益诉求、数字经济水平、历史文化、法律政策等不同,难以在数据本地化存储规则、数据流通、关税、数字产品非歧视待遇等问题上达成共识,统一的数据跨境流通治理机制难以构建。^[20]针对国际规

[19] 石佑启,韩永红. 论新时代我国涉外法治人才培养体系之优化 [J]. 中外法学, 2024, 36 (06): 1427-1448.

[20] 郭朝先,李婷,罗芳. 数据跨境流通规则博弈与中国应对 [J]. 中国流通经济, 2024, 38 (09): 27-38.

DOI:10.14089/j.cnki.cn11-3664/f.2024.09.003.

则的变化,我国可以从国际合作以及构建“数字丝绸之路”方面出发,首先,增强与在数据跨境治理领域具有国际影响力的国家或组织的合作,重点突破关键领域的制度性对接。可参考欧盟与东盟在 2023 年 5 月布鲁塞尔联合发布的《东盟示范合同条款与欧盟标准合同条款联合指南》,可探索建立中欧标准合同条款互认机制,构建统一的数据跨境流动合规框架,引导企业通过标准化条款实现数据自由流通。面对美国对华的产业断链政策,我国应秉持“共商、共建、共享”的全球治理观,在既有国际合作机制中加入网信议题,进一步扩展中国在网信领域的“国际朋友圈”,深化中国在网信产业中的国际合作,打造“开放、公平、公正、非歧视”的数字发展环境。^[21]其次,构建“数字丝绸之路”,我国应充分利用共建“一带一路”合作平台契机,依托网络空间命运共同体,凝聚“一带一路”数据跨境流通治理共识,搭建“一带一路”数据跨境流通框架,合作共赢,共同发展,提高我国在双多边区域数据跨境流通治理中的影响力。^[22]在深化“数字丝绸之路”战略实施过程中,需着力弥合“一带一路”沿线国家的数字发展鸿沟。作为数据跨境流动的物理载体,新型信息基础设施体系建设是关键突破口。可以协同沿线国家推进 5G 通信网络、云存储中心、人工智能算力平台等新型基础设施的共建共享,同步推进数据安全防护体系建设,涵盖数据资产目录编制、加密存储、脱敏处理及防泄露技术应用等关键领域。通过构建全链条的数字基建标准体系和数据安全防护框架,可有效提升沿线国家数字基础设施的互联互通水平,为跨境数据流动创造安全可靠的技术环境,为数字经济合作筑牢底层支撑。

六、结论

在数字经济与地缘政治深度交织的背景下,美国凭借《云法案》构建网络空间霸权,通过“长臂管辖”实施单边主义数据治理,对我国数据主权、企业合规及司法独立构成系统性威胁。华为、TikTok 等典型案例揭示,美国以“国家安全”为名的打压本质是遏制中国技术崛起、维护全球科技霸权。中国在涉外法治建设中,既需抵御霸权渗透,更需通过制度创新提升规则话语权,构建兼顾安全与发展的全球网络治理秩序。

参考文献

[1]孙祁,栾文韬.应对法律战视角下的涉外法治建设初探[J].和平与发展,2024,(05):1-24+179.

[2]王智珑.美国对华数字政策泛安全化主要手段与应对建议——以 TikTok 封

[21] 张心志,郑文佳.美国网络安全战略的演进:对华政策与应对策略[J].智库理论与实践,2024,9(06):169-179.DOI:10.19318/j.cnki.issn.2096-1634.2024.06.18.

[22] 郭朝先,李婷,罗芳.数据跨境流通规则博弈与中国应对[J].中国流通经济,2024,38(09):27-38.DOI:10.14089/j.cnki.cn11-3664/f.2024.09.003.

禁案为例[J].海外投资与出口信贷,2024,(05):35-38.

[3]韩娜.美国对华网络空间安全政策:动因、举措及影响[J].国家安全论坛,2024,3(04):44-52+91.

[4]余丽.美国网络霸权护持政策的继承与调整:目标、工具和效应[J].郑州大学学报(哲学社会科学版),2023,56(04):24-29.

[5]朱劲草.美国《云法案》中的数据域外管辖问题研究[D].华东政法大学,2023.DOI:10.27150/d.cnki.ghdzc.2023.000337.

[6]洪延青.“法律战”漩涡中的执法跨境调取数据:以美国、欧盟和中国为例[J].环球法律评论,2021,43(01):38-51.

[7]肖伟,党凯瑞,皇甫泽西.中美贸易战对中国科技企业的影响分析[J].时代金融,2020,(18):57-59.

[8]中央财经大学课题组.数据本地化的重思与改进[J].数据法学,2024,6(01):3-22.

[9]贺赞.涉外法治人才培养机制创新:以课程体系建设为中心[J].中国法学教育研究,2017(2):3-11.

[10]朴顺善.新时代涉外法治人才培养路径研究[J].黑龙江教育(理论与实践),2024,(06):25-32.

[11]陈俊秀,王纬航,李智.我国域外数据执法管辖面临的困境与因应[J].集美大学学报(哲社版),2024,27(05):27-34+92.

[12]文淑.数据立法域外适用引发的法律冲突与中国解决方案[J].云南师范大学学报(哲学社会科学版),2023,55(06):96-108.

[13]石佑启,韩永红.论新时代我国涉外法治人才培养体系之优化[J].中外法学,2024,36(06):1427-1448.

[14]郭朝先,李婷,罗芳.数据跨境流通规则博弈与中国应对[J].中国流通经济,2024,38(09):27-38. DOI:10.14089/j.cnki.cn11-3664/f.2024.09.003.

[15]李坤泽,张依桐.美国的TikTok政策:威胁认知、打压实践与效果制约[J].国际关系研究,2024,(04):22-40+155.

[16]陈思琦.我国跨境数据流动法律规制问题研究[D].中国人民公安大学,2024. DOI:10.27634/d.cnki.gzrgu.2024.000312.

[17]Reading DM E .DeepSeek AI fails multiple security tests, raising red flag for businesses[J].Urgent Communications,2025

[18]Kayaalp E M ,Prill R ,Sezgin A E , et al.DeepSeek versus ChatGPT: Multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-Redefining innovation in research and practice.[J].Knee surgery, sports traumatology, arthroscopy : official journal of the

ESSKA,2025

[19]Parry O R .Internet Law And Business: International Issues[M].World Scientific Publishing Company:2024-03-06. DOI:10.1142/13427.

[20]United States Pursues Regulatory Actions Against TikTok and WeChat Over Data Security Concerns[J].The American Journal of International Law,2021,115(1):124-131.

2025年中国国际法学会年会

数据控制者取证模式的实践抵牾、理论调和及应对策略

王雪*

摘要：数据控制者取证模式以对网络服务提供者的属人管辖为核心，混合合法权利标准与实际能力标准共同界定数据的控制范围，从而实施跨境数据调取。数据主权的弱属地性并不足以推翻域外管辖规则对执法的限制，数据控制者取证模式的单边域外执法侵犯他国数据主权。同时，数据控制者取证模式与个人数据保护规范和数据出境管制规则的矛盾尖锐。为调和数据控制者取证模式的实践抵牾，数据主权以谦抑性为法理基础，进行多维度划分与组合，配合程序主义与数据分级分类的实施，可达到弹性适用的效果。为应对数据控制者取证模式的扩散，我国也需构建攻防兼备的跨境取证方案。我国可借助国际协议的谈判契机推动传统国际司法协助的简化，搭建数据取证的专属风险评估通道，同时优化阻断条款，引入数据信托云帮助企业规避数据控制者取证模式的管辖。

关键词：数据控制者取证模式 数据跨境取证 主权冲突 合规困境 主权弹性

2022年3月，美国伊利诺伊州北区地区法院在“*Philips Med. Sys. (Cleveland), Inc. v. Buan*”案中主张中国《个人信息保护法》第41条以及《数据安全法》第36条等阻断条款并不能自动免除被告遵守美国法院证据开示命令的义务。¹同年，“*Intex Rec. Corp. v. Bestway (USA), Inc.*”案²再次暴露同样的问题。鉴于大数据时代下，执法机关或司法机关迫切需要跨境调取数据，由此引发的实践冲突俨然成为我国在跨境诉讼中面临的重要难题。

长期以来，国际社会所实施的数据跨境取证方式大致可以划分为两种，即“数据存储地取证模式”与“数据控制者取证模式”。前者在执法机关或司法机关取证中占据主导地位，即依据数据存储的地理位置来判断一国调取数据的管辖权范围。后者则是允许执法机关或司法机关要求为本国提供服务的运营商在特定情形下提交自身掌握或控制范围内的数据。考虑到传统国际司法协助方式的弊病，“数

作者简介：王雪，女，河南周口人，内蒙古大学法学院讲师，南开大学法学博士，研究方向为网络信息法、网络空间国际法，联系方式 wangxue_imu@163.com, 17826029280。

基金项目：2022年国家社科基金一般项目“数智环境下个人信息数据泄露与保护的政策法规体系构建研究”（22BTQ016）；内蒙古大学青年学术人才科研启动项目（10000-23112101/261）。

¹ *Philips Med. Sys. (Cleveland), Inc. v. Buan*, No. 19 CV 2648, 2022 U.S. Dist. LEXIS 35635 (N.D. Ill. Mar. 1, 2022)

² *Intex Rec. Corp. v. Bestway (USA), Inc.*, No. CV 19-8596-JAK(Ex), 2023 U.S. Dist. LEXIS 194563 (C.D. Cal. Oct. 30, 2023)

据控制者取证模式”逐步在全球扩张。³针对数据控制者取证模式引发的争议,有学者研究主权冲突下的对等回应与互惠,⁴也有学者从多主体协同取证的角度研究策略,⁵或是从不同数据存储方式的角度分析中国的跨境调取路径⁶。然而,数据控制者取证模式引发的对立依旧没有得到有效解决。面对他国的单边数据跨境调取,我国执法机关与企业仍显得较为被动。

2024年12月,联合国大会审议通过《打击为犯罪目的使用信息和通信技术行为的全面国际公约》(Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes,以下简称《联合国打击网络犯罪公约》),但这并未解决数据控制者取证模式造成的抵牾。鉴于此,本文结合《联合国打击网络犯罪公约》,从以下方面展开:(1)追溯数据控制者取证模式的发展与动向,梳理其理论基础,尤其是理清学界尚未关注到的“控制”范围的判断标准。(2)结合近年来最新的跨境取证案例的裁决趋向,剖析数据控制者取证模式引起的两大核心冲突;(3)从数据主权弹性适用的视角出发,探索缓和跨境取证摩擦的路径,并尝试从进取与防御的双重方向提出我国与企业的应对策略。

一、数据控制者取证模式的国际动向与理论基础

数据控制者取证模式在全球逐步扩张,得到诸多国家的效仿,但不同国家主体的立法规范有所差异,国际社会对数据控制者取证模式的态度较为割裂。因此,有必要梳理数据控制者取证模式的现行规则及理论基础。

(一) 数据控制者取证模式的现行规则

数据控制者取证模式的现行规则既存在于民事案件,也存在于刑事领域。从主要依托于国内法实施到多边协议达成有限共识,数据控制者取证模式在全球的适用范围明显扩张。

1. 美国:证据开示制度与CLOUD法案的双重叠加

为保障诉讼主体可以不受阻碍地获取与诉讼争议有关的所有证据,美国《联邦民事诉讼规则》(Federal Rules of Civil Procedure)赋予诉讼主体以广泛的证据开示的权利。传票作为美国证据开示制度的主要方式,由一方当事人单方面启动,且请求对象既包括案件当事人也包括案外人,请求开示的证据范围广泛且以法院

³ 参见王雪、石巍:《数据立法域外管辖的全球化及中国的应对》,《知识产权》2022年第4期。

⁴ 参见梁坤:《基于数据主权的国家刑事取证管辖模式》,《法学研究》2019年第2期。

⁵ 参见裴炜:《论网络犯罪跨境数据取证中的执法管辖权》,《比较法研究》2021年第6期;裴炜:《向网络信息业者取证:跨境数据侦查新模式的源起、障碍与建构》,《河北法学》2021年第4期。

⁶ 参见魏求月:《中国数据跨境调取路径探析——以数据相关主体和存储路径为视角》,《国际法研究》2023年第3期。

强制执行⁷为保障。⁸2015 年美国纽约地区法院在 *Gucci America, Inc. v. Weixing Li* 一案中向中国银行发出传票,称中国银行在纽约地区设立分支机构并在美国盈利,满足最低限度联系的需求,因而对中国银行行使属人管辖权是恰当的。尽管与被告账户有关的数据位于中国,但美国纽约地区法院判决中国银行遵守证据开示命令。⁹这一判决完整展现了美国法院借助证据开示制度实施单边数据跨境取证的逻辑思路。

刑事诉讼中美国法院也支持使用传票程序迫使在美国开展业务或设立分支机构的银行交出同一银行在外国分行所持有的记录,即使交出记录会违反外国的保密法。在 *In re Grand Jury Proceedings Bank of Nova Scotia* 一案中,在美国迈阿密设立办事处的加拿大新斯科舍银行屡次不遵守大陪审团的传票,被美国佛罗里达州南区地区法院下达的藐视法庭令并处以 182.5 万美元罚款。联邦第十一巡回上诉法院在对美国和开曼群岛相互冲突的法律进行评估后,认为执行传票符合两国法律的基本政策,最终维持地区法院裁决。¹⁰

相似的,美国《存储通信法》(Stored Communications Act, 以下简称 SCA)第 2703 条“客户通信或记录的必要披露”也规定,对于基本的用户信息和交易数据,执法机关可以通过行政传票(administrative subpoena)¹¹来获得。然而,值得注意的是,若政府要求网络服务提供者披露通信数据,依旧需要遵守刑事诉讼规则并由法院签发搜查令(warrant)。¹²可见,对于隐私保护程度较高的通信数据,政府仍依赖法院签发搜查令从网络服务提供者获取。

由于搜查令作为行使和制约政府权力的工具,以明显的地域限制方式保护隐私。是故,搜查令能否适用于域外取证这一问题在美国法院之间存在较大分歧,这一争议最终在“微软案”彻底爆发。美国联邦第二巡回上诉法院从否定 SCA 法案的域外效力的角度,论证美国执法机关无权限单边跨境调取存储于爱尔兰的数据。¹³为解决这一问题,2018 年美国颁布《澄清合法使用境外数据法》(Clarifying Lawful Overseas Use of Data Act, 以下简称 CLOUD 法案),成为刑事案件中数据控制者取证模式的典型代表。CLOUD 法案第 103 (a) (1)

⁷ 依据《联邦民事诉讼规则》第 45 条,法院有权发送传票,强制要求被取证人出示证件、电子存储信息或有形物品。在没有充分理由的情况下不遵守传票或与传票有关命令的主体将被签发传票的法院认定为藐视法庭并受到制裁。

⁸ 参见胡洋铭:《美国事证开示程序的域外效力及阻断路径探究》,《争议解决》2021 第 4 期,第 212-218 页。

⁹ *Gucci Am., Inc. v. Weixing Li*, 135 F. Supp. 3d 87 (S.D.N.Y. 2015).

¹⁰ *In re Grand Jury Proceedings Bank of Nova Scotia*, 740 F.2d 817 (11th Cir. 1984).

¹¹ 行政传票是行政管理机构为迫使个人向该机构提供信息而发出的传票。为了获得行政传票,政府不需要证明合理的理由。

¹² 依据美国 SCA 第 2703 条,为了访问存储时间少于 180 天的通信数据,政府必须根据《联邦刑事诉讼规则》规定的程序获得法院审批的搜查令;对于存储时间超过 180 天的通信数据,尽管政府可以通过行政传票要求网络服务提供者披露,但前提是政府必须事先通知用户。而在刑事侦查中提前通知用户可能并不利于警方调查。因而,政府要求网络服务提供者披露通信数据主要依赖于法院签发的搜查令。

¹³ *Microsoft Corp. v. United States (In re Warrant to Search a Certain E-Mail Account Controlled & Maintained by Microsoft Corp.)*, 829 F.3d 197 (2d Cir. 2016).

条款¹⁴的规定将通信数据纳入政府单边跨境调取数据的范畴。因而，CLOUD 法案的颁布是对美国证据开示制度与 SCA 法案的补充。

综上，美国证据开示制度与 CLOUD 法案的叠加使得数据控制者取证模式的适用范围遍及民事、行政与刑事诉讼。即使是隐私保护程度最高的通信数据，美国执法机关或司法机关依旧可以单边向网络服务提供者跨境调取。

2. 欧盟：向数据控制者取证模式的转型

欧盟对数据跨境取证的探索由来已久。早在 2000 年欧洲议会与欧盟理事会便制定《欧洲联盟成员国刑事事项互助公约》（on Mutual Assistance in Criminal Matters between the Member States of the European Union）¹⁵为成员国之间相互申请证据披露提供便利。然而，利用司法合作渠道获取电子证据具有“先天不足”，无法满足时效性需求。后续申请国获取到的线索可能不再可用。为此，2014 年欧洲议会和欧洲理事会再次发布《关于刑事事项欧洲调查令的第 2014/41/EU 号指令》（Directive 2014/41/EU of the European Parliament and of the Council of 3 April 2014 regarding the European Investigation Order in criminal matters）¹⁶，允许申请国签发机关签发紧急调查令，与另一成员国执法机关直接就跨境取证合作。调查令是欧盟为促进电子证据的调取，缩短取证时限，简化取证程序的尝试。但电子证据易被迅速删除，调查令在后续适用中依旧无法应对欧盟成员国跨境调取电子证据的需求。2016 年 3 月，“布鲁塞尔恐怖袭击事件”加快了欧盟内部打击恐怖犯罪、实现司法数据互通的进程。欧盟各国司法部长、内政部长和欧盟机构代表发表联合声明，强调更快、更有效地获取数字证据。

美国推行跨境取证新规的行为也促使欧盟对自身法规进行调整。2018 年欧盟委员会的工作文件显示，欧洲 55% 的刑事案件调查涉及跨境获取电子证据，尤其是需要向位于其他司法管辖区的网络服务提供者提出请求。欧盟成员国越来越依赖与网络服务提供者搭建直接合作渠道，但缺乏统一的法律框架，引发法律不确定性和潜在的法律冲突。是故，欧盟成员国的执法和司法机关迫切需要有效工具来调查和起诉与网络空间有关的犯罪行为。同年，欧盟在美国 CLOUD 法案的影响下，发布《关于欧洲刑事案件电子证据提交令和保全令的条例草案》，不再将数据位置作为决定性的连接因素。¹⁷经历漫长的修订与讨论，2023 年 7 月，欧

¹⁴ 无论通信、记录或其他信息是否存储在美国境内，电子通信服务和远距离计算服务提供者均应当按照本章所规定的义务要求，保存、备份或披露关于用户或客户的有线或电子通信内容、所有记录或其他信息，只要上述通信内容、记录或其他信息为该服务提供者所拥有、监管或控制。

¹⁵ See Convention established by the Council in accordance with Article 34 of the Treaty on European Union, on Mutual Assistance in Criminal Matters between the Member States of the European Union, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A42000A0712%2801%29,2024-11-02>.

¹⁶ See Directive 2014/41/EU of the European Parliament and of the Council of 3 April 2014 regarding the European Investigation Order in criminal matters, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0041,2024-11-02>.

¹⁷ See European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European Production and Preservation Orders for electronic evidence in criminal matters, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A225%3AFIN,2024-11-02>.

盟正式通过《关于刑事诉讼中的电子证据和刑事诉讼后执行监禁判决的欧洲提交令和保存令的第 2023/1543 号条例》（以下简称《电子证据条例》）¹⁸。欧盟主管司法机关在刑事诉讼（包括刑事调查）中，或在刑事诉讼后执行监禁判决或拘留令时，可通过发布数据提交令和保存令直接要求在欧盟提供服务的网络服务提供者出示或保全电子证据，无需考虑数据位置。《电子证据条例》的管辖范围包括在欧盟提供服务的所有网络服务提供者，即使其未在欧盟境内设备经营场所但在欧盟境内积极提供服务，依旧适用于这一条例。可见，欧盟在跨境取证领域的立法已然不局限于国际司法协助的轨道，向数据控制者取证模式转化。

3. 国际：数据跨境取证制度的有限共识

2001 年，世界上第一部专门打击网络犯罪的区域性多边条约《布达佩斯公约》（Convention on Cybercrime）¹⁹对数据控制者取证模式进行了初步探索。《布达佩斯公约》第 18 条“提交令”规定，允许本国领域内的数据主体提交所拥有或控制的数据，或境内网络服务提供者提交所拥有或控制的用户信息。此时数据或用户信息的地理位置位于不同的司法管辖区并不会影响这一条款的适用。尽管这一条款已然产生数据控制者取证模式的效果，但鉴于其强制性效力局限于请求国领土，是否回应请求国的指令依旧由各缔约国的网络服务提供者自主决定。²⁰在美国 CLOUD 法案的推动下，2021 年《布达佩斯公约》第二附加议定书对提交令条款予以革新：其一，加强与其他缔约国境内网络服务提供者与实体的直接合作。依据第二附加议定书第 6 条第 1 款²¹和第 7 条第 1 款²²，某一缔约国的主管当局可以直接请求其他缔约国境内的网络服务提供者提供域名注册信息或用户信息。其二，加强缔约国主管当局之间的国际合作以披露数据。依据第二附加议定书第 8 条第 1 款，某一缔约国的主管当局向另一缔约国主管部门提出申请后，后者将要求其境内的网络服务提供者披露用户信息或通信数据。在紧急情况下，各缔约国可使用联络点相互之间转达或接受请求，指令其境内的服务提供者披露特定数据，而无需提出司法协助的请求。

鉴于《布达佩斯公约》无法代表众多发展中国家利益，2019 年，在中国与俄罗斯倡议下，第 74 届联合国大会决定设立政府间专家委员会以拟订《关于打击

¹⁸ See Regulation (EU) 2023/1543 on European Production and Preservation Orders for electronic evidence in criminal proceedings and for the execution of custodial sentences following criminal proceedings,自 2026 年 8 月 18 日起生效, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2023:191:TOC,2024-11-02>.

¹⁹ 2001 年欧盟成员国与美国、日本等 30 多个国家于布达佩斯签署，该公约也成为世界上第一部专门打击网络犯罪的国际条约。

²⁰ 参见梁坤：《网络犯罪国际公约中的跨境调取数据条款》，《当代法学》2024 年第 2 期，第 124 页。

²¹ 《布达佩斯公约》第二附加议定书第 6 条第 1 款：各缔约方应采取必要的立法和其他措施，授权其主管当局为具体的刑事调查或诉讼程序，向在另一缔约方境内提供域名注册服务的实体发出请求，要求提供该实体所拥有或控制的信息，以查明或联系域名的注册人。

²² 《布达佩斯公约》第二附加议定书第 7 条第 1 款：各缔约方应采取必要的立法和其他措施，使其主管当局有权发出命令，直接提交给另一缔约方境内的服务提供商，以便获得该服务提供商所拥有或控制的特定的、已储存的用户信息的披露，而这些用户信息是发布命令缔约方的具体刑事调查或诉讼所需要的。

为犯罪目的使用信息和通信技术行为的全面国际公约》（International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes）（以下简称《联合国打击网络犯罪公约》）。²³同年，联合国毒品和犯罪问题办公室第五届网络犯罪政府间专家组会议召开，会议共识在于，占有大量数据的网络服务提供者成为重要的执法参与者且国际社会应当在司法协助之外探索新的取证模式。然而，在共识之外，具体策略却争议较多，尤其是跨境向网络服务提供者直接调取数据涉及到数据主权与安全的问题。²⁴数据控制者取证模式遭到俄罗斯、南非和伊朗等国的反对。²⁵经过多重谈判，最终 2024 年 12 月《联合国打击网络犯罪公约》正式通过，其中第 27 条²⁶“提交令”条款几乎继承了《布达佩斯公约》第 18 条。可见，国际社会仅仅认同就用户信息的跨境调取采取数据控制者取证模式，针对其余数据依旧采用司法协助的方式予以跨境取证。

（二）数据控制者取证模式的理论支撑

梳理了数据控制者取证模式的现行规则后，厘清其国际法上的理论基础至关重要。数据控制者取证模式应用的关键在于判断管辖范围，包括网络服务提供者的管辖范围以及数据的控制范围。

1. 以对网络服务提供者的属人管辖权为依据

“数据控制者模式”将跨境调取数据的关键因素从数据存储的地理位置转移到对网络服务提供者的管辖权。从立法管辖权基础来看，“数据控制者模式”下域外管辖权的立法依据接近属人管辖原则。

其一，美国 CLOUD 法案便是以积极属人管辖原则为基础，将公司在境外的数据活动纳入管辖范围。CLOUD 法案第 102（2）条款强调管辖对象是“受美国管辖的通信服务提供者”。“受美国管辖”是否等同于被管辖者一定是美国公司呢？2019 年美国司法部发布 CLOUD 法案白皮书，对此做出否定回答，其宣称 CLOUD 法案的司法管辖范围并不限于美国公司、总部设在美国的公司或美国人拥有的公司。但美国对外国公司的管辖权受到美国宪法的约束且遵循美国法院确立的判例法。美国司法对属人管辖的审判秉持“最低限度联系”的判断标准。位于美国境外但在美国提供服务的外国公司是否与美国有足够的联系且受美国管辖，是基于对个案中该公司与美国联系的性质、数量和质量的考生。基于事实调查，搜查令会

²³ 参见叶伟：《〈联合国打击网络犯罪公约〉：进展、前景与展望》，《北京航空航天大学学报(社会科学版)》2021 年第 5 期，第 32-33 页。

²⁴ 参见黄道丽：《网络安全法治研究 2020》，华中科技大学出版社 2020 年版，第 278 页。

²⁵ 参见王渊洁：《联合国网络犯罪政府专家组第五次会议综述》，《中国国际法年刊》2019 年，第 354 页。

²⁶ 《联合国打击网络犯罪公约》第 27 条：

各缔约国均应采取必要的立法措施和其他措施，授权本国主管机关下令：

（一）位于本国领土内的某人提交其所拥有或控制的存储在信息通信技术系统或电子数据存储介质中的特定电子数据；以及

（二）在本国领土内提供服务的服务提供者提交其所拥有或控制的与此类服务有关的订阅用户信息。

适用于被国外收购或者总部位于国外的美国子公司所存储的数据。²⁷因而，CLOUD 法案所规定的“受美国管辖的通信服务提供者”绝不仅局限于美国网络服务提供者，未在美国境内注册但符合美国司法所确立的属人管辖规则的外国网络服务提供者也会受到 CLOUD 法案的制约。

其二，欧盟《电子证据条例》也同样以对网络服务提供者的管辖为核心。

《电子证据条例》第 2 条“本条例适用于在欧盟境内提供服务的服务提供商”。根据第 3 条对术语的解释，满足“在欧盟境内提供服务”需要具备两项条件：第一，网络服务提供者向欧盟成员国的自然人或法人提供电子通信服务、互联网域名和 IP 服务以及其他信息社会服务；第二，与其提供服务的成员国具备实质性联系。网络服务提供者在某成员国设有经营场所可以直接视为具备实质性联系。倘若未设立经营场所，在欧盟境内有相当数量的用户，或其活动以一个或多个成员国为目标（target），也可认为存在实质性联系。这一判断标准吸收了 2015 年比利时最高法院在“雅虎案”中的裁决思路²⁸，即虽然雅虎公司在比利时没有实体经营场所，但其将经济活动指向比利时的消费者，比如使用当地的语言、根据用户的位置显示广告并且通过安装服务台为比利时的用户提供便利，这些均表明其自愿接受比利时法律的约束，在比利时具有虚拟存在（virtual presence）。可见，只要网络服务提供者提供的服务以欧盟境内的消费者为目标，将被纳入到欧盟《电子证据条例》的管辖范围。这也极大扩张了欧盟成员国跨境调取数据的范围。

2. 合法权利标准与实际能力标准混合界定控制范畴

属人管辖是数据控制者取证模式应用的国际法依据，而数据被特定的网络服务提供者所持有或控制则是执法机关或法院发布调取命令的前提。美国 CLOUD 法案第 103 条将披露或保存的数据范围界定为受美国管辖的网络服务提供者占有、保管或控制的数据。这与《联邦民事诉讼规则》第 45(a)(1)条²⁹以及《联邦刑事诉讼规则》第 16(a)(1)条³⁰的表达相一致。“占有”和“保管”的定义很少引起争议，因为它们都是指“是或否”的二元命题：目标实体要么占有或保管所寻求的证据，要么没有，而确立“控制”则不那么明确。尽管 CLOUD 法案尚未解释“占有、保管或控制”的界定，但美国法院已经在一系列判例中就《联邦民事诉讼规则》

²⁷ See *Promoting Public Safety, Privacy, and the Rule of Law Around the World: The Purpose and Impact of the CLOUD Act White Paper*, U.S. Department of Justice (April 2019), <https://www.justice.gov/opa/press-release/file/1153446/dl>.

²⁸ Hof van Cassatie [Court of Cassation], Dec. 1, 2015, Nr. P.13.2082.N (Belg.)

²⁹ 《联邦民事诉讼规则》第 45(a)(1)条：每张传票都必须……命令传票所针对的每个人在指定的时间和地点做以下事情：出庭作证；出示指定的文件、电子存储信息或该人占有、保管或控制的有形物品……

³⁰ 《联邦刑事诉讼规则》第 16(a)(1)条：根据被告的要求，政府必须向被告披露以下所有内容，并允许其检查、复制或拍照：

(i) 被告的任何相关书面或录音陈述，前提是：

- 该陈述由政府占有、保管或控制；且
- 政府律师知晓（或通过尽职调查可以知晓）该陈述的存在；

与《联邦刑事诉讼规则》中的同一术语进行了广泛论证，研究如何判断一个实体控制数据。后续美国法院对 CLOUD 法案“控制”的解读也将遵循同样的逻辑。整体而言，美国判例法对“控制”有两种测试标准：一是合法权利标准；二是实际能力标准。³¹

合法权利标准指一方主体拥有获取文件、电子存储信息以及其他有形物品的合法权利。³²合法权利标准用于确定美国公司是否“控制”外国关联公司所持有的文件。适用于数据跨境取证的场景，美国法院在判断“控制权”需要考虑两项因素：第一，网络服务提供者拥有按要求获得所需文件的法定权利；例如在“新斯科舍银行”一案中，总公司被强制要求交出外国分支机构持有的数据。³³第二，网络服务提供者与相关实体人格混同，法院有理由揭开公司面纱；³⁴例如，母公司对子公司拥有完全控制权，因此母公司在法律上有能力指导子公司数据的使用或转移，这表明母公司对数据拥有法律控制权。³⁵

实际能力标准则是指当一方主体具备获取文件、电子存储信息以及其他有形物品的实际能力。³⁶将这一标准应用于数据跨境取证的情境下，可以理解为如果网络服务提供者有实际能力获得执法机关或司法机关要求的数据，它将必须披露。采用实际能力标准的执法机关或法院无需考虑网络服务提供者是否在法律上拥有合法权利获取。但法院如何确定这种“实际能力”是否存在？美国联邦法院的判例尚未形成统一的判断方法。美国法院在应用实际能力标准时考量许多因素，如公司是否能在正常业务过程中交换或获取文件、公司与争议交易的联系、母子公司之间的财务关系等等。³⁷但不同案件中法院究竟考量哪些因素并不一致。鉴于“实际能力”缺乏一个准确的、普遍接受的定义，导致在特定案件中法院如何应用实际能力标准缺乏可预见性，判例法也无法提供真正的指导。³⁸

然而，美国联邦巡回上诉法院在应用合法权利标准时已然与实际能力标准相混淆。鉴于合法权利标准之下，法院可以调取文件或数据范围相比于实际能力标准的范围较窄，美国法院在数据跨境取证中不可能排除对网络服务提供者日常业务活动中实际控制能力的考量。美国联邦第三巡回上诉法院在 *Gerling Int'l Ins. Co. v. Comm'r Gerling Int'l Ins. Co.* 中主张，如果子公司作为代理人，基于业务需

³¹ See Johnathan D. Jordan, *Out of Control Federal Subpoenas: When Does a Nonparty Subsidiary Have Control of Documents Possessed by a Foreign Parent*, 68 Baylor Law Review 189,197(2016).

³² See The Sedona Conference, *Commentary on Rule 34 and Rule 45 "Possession, Custody, or Control,"* 25 Sedona Conf. J. 1,18 (2024).

³³ *In re Grand Jury Proceedings Bank of Nova Scotia*, 740 F.2d 817 (11th Cir. 1984).

³⁴ See Johnathan D. Jordan, *Out of Control Federal Subpoenas: When Does a Nonparty Subsidiary Have Control of Documents Possessed by a Foreign Parent*, 68 Baylor Law Review 189,199(2016).

³⁵ See Justin Hemmings, Sreenidhi Srinivasan & Peter Swire, *Defining the Scope of "Possession, Custody, or Control" for Privacy Issues and the CLOUD Act*, 10 Journal of National Security Law and Policy 631, 656(2020).

³⁶ See The Sedona Conference, *Commentary on Rule 34 and Rule 45 "Possession, Custody, or Control,"* 25 Sedona Conf. J. 1,18 (2024).

³⁷ See Johnathan D. Jordan, *Out of Control Federal Subpoenas: When Does a Nonparty Subsidiary Have Control of Documents Possessed by a Foreign Parent*, 68 Baylor Law Review 189,200-201(2016).

³⁸ See The Sedona Conference, *Commentary on Rule 34 and Rule 45 "Possession, Custody, or Control,"* 25 Sedona Conf. J. 1,35 (2024).

要可以获得委托人母公司的文件，法院将不允许子公司否认控制权。³⁹然而，判断母子之间业务关系实则属于对子公司实际能力的判断。这一判例先后被美国联邦第九巡回上诉法院⁴⁰与地方法院⁴¹引用，将实际能力的部分考量因素与合法权利标准相融合。此外，美国联邦第四、第八、第十、第十一巡回上诉法院也前后混合使用合法权利与实际能力两种标准。⁴²例如，美国网络服务提供者作为个人数据处理者负责日常处理他国公司的数据，尽管不具备披露的法定权利，但依旧会被认定为“控制”。⁴³

《布达佩斯公约》解释性报告第 173 段对第 18 条“提交令”的“拥有”或“控制”予以解读。报告规定，“拥有或控制是指在命令方的领土上实际占有有关数据，以及需要出示的数据不在当事人的实际占有范围内但当事人可以在命令方的领土上自由控制，例如，一个人被送达搜查令，要求出示通过远程在线存储服务存储在其账户中的数据，那么这些数据必须被出示。”⁴⁴对网络服务提供者而言，其占有、保管或控制的数据不仅包括其实际占有的数据，也包括在其控制下远程存储的数据。可见，《布达佩斯公约》对控制的解读不仅包括法定占有的数据，也包含对实际能力的考量。《联合国打击网络犯罪公约》第 27 条的制定来源于《布达佩斯公约》第 18 条，因此，其对网络服务提供者控制权的解读也会借鉴相同的逻辑。

二、实践抵牾：主权冲突与合规困境

不可否认，数据控制者取证模式确实达到高效取证和成本低廉的效果，但国际社会对数据控制者取证模式的认同依旧有限。美国与欧盟是以国内法的方式试图单边跨境调取数据，极易与他国产生主权与规范层面的冲突。

（一）数据控制者取证模式引发主权冲突

主权冲突是数据控制者取证模式引发的核心矛盾，调和国家在数据领域的摩擦都离不开对主权这一问题的探讨。

1. 数据主权的弱属地性不足以推翻域外管辖规则

数据控制者取证模式是一国单边扩张数据域外管辖权的典型表现。网络空间经历了从“去主权化”到“再主权化”的历程。国家在网络空间是否行使管辖权这一问题得到解决，但国家如何在网络空间行使管辖权却依旧充满分歧。数据主权是

³⁹ Gerling Int'l Ins. Co. v. Comm'r Gerling Int'l Ins. Co., 839 F.2d 131 (3d Cir. 1988)

⁴⁰ See 7-UP Bottling Co. v. Archer Daniels Midland Co. (In re Citric Acid Litig.), 191 F.3d 1090 (9th Cir. 1999).

⁴¹ See Addamax Corp. v. Open Software Found., 148 F.R.D. 462 (D. Mass. 1993); In re Polygon Glob. Partners LLP, No. 21-mc-007 WES, 2021 U.S. Dist. LEXIS 90143 (D.R.I. May 11, 2021).

⁴² See The Sedona Conference, *Commentary on Rule 34 and Rule 45 "Possession, Custody, or Control,"* 25 Sedona Conf. J. 1,28 (2024).

⁴³ See Justin Hemmings, Sreenidhi Srinivasan & Peter Swire, *Defining the Scope of "Possession, Custody, or Control" for Privacy Issues and the CLOUD Act*, 10 Journal of National Security Law and Policy 631, 660(2020).

⁴⁴ See Council of Europe, Explanatory Report to the Convention on Cybercrime, <https://rm.coe.int/16800cce5b,2024-11-02>.

主权理念在数据领域的延伸,强调国家对本国数据的最高控制权以及排除他国干涉的权力,⁴⁵这一观点为国内绝大部分学者所认同。⁴⁶西方学者以 Daskal 与 Svantesson 教授为代表提出“数据例外主义”,主张基于数据的流动性、可复制性以及位置的独立性等特殊性质,数据主权的属地性不再存在,传统域外管辖理论无法适用于数据领域,急需为数据这一介质创设新的国际管辖规则,打破原有国际习惯法对域外执法的限制。⁴⁷那么,当主权延伸到数据领域时,是否依旧具备属地性呢?国际法域外管辖的规则是否可以继续适用于数据领域?这一问题涉及数据主权的根本性问题,也是数据领域国际合作的出发点。

一方面,数据的流动性、可复制性以及位置的独立性意味着数据主体、网络服务提供者或执法人员对数据可以开展远程操作而不受领土边界的限制。⁴⁸针对同一数据,一方主体对数据的处理,不会影响另一方主体对数据的利用。例如,2022 年在我国“李某某开设赌场”一案中,由于赌博平台位于境外,我国侦查人员远程勘验工作记录及光盘获取证据。⁴⁹鉴于此,主权的属地性在数据的远程操作中确实被弱化。

另一方面,数据主权的属地性依旧广泛存在于网络空间,主权国家依旧可利用数据主权的属地性实现管辖。其一,国家对物理层⁵⁰所包含基础设施的管辖具有强烈的属地特征。尽管数据在全球范围内流动,但操纵数据的设备或主体依旧受到某个国家的领土管辖。位于国家境内的网络基础设施受其颁布的法律规范的制约和保护,国家主体可采取合适或必要措施行使立法、执法以及保护的权力,且这一权力不受设施用途与国籍的影响。⁵¹因而,数据主权属地性的特征在物理层较为突出,网络大国对网络基础设施行使主权也不存在原则性分歧。⁵²

其二,数据主权在逻辑层⁵³依旧具备属地性。虽然作为互联网技术的 TCP/IP 协议或域名系统全球通用,但其依旧会受到国家领土边界的限制。为防止以美国为首的西方国家对俄罗斯网络空间采取全面制裁,俄罗斯创建本国域名系统并实

⁴⁵ 参见齐爱民、祝高峰:《论国家数据主权制度的确立与完善》,《苏州大学学报(哲学社会科学版)》2016 年第 1 期,第 83 页。

⁴⁶ 中国学者对数据主权的界定回归主权的应有之义上,倾向于从管辖的角度将数据主权解释为国家对数据以及数据设备的控制、管辖、使用与保护的权力。文献参见齐爱民、盘佳:《数据权、数据主权的确立与大数据保护的基本原则》,《苏州大学学报(哲学社会科学版)》2015 年第 1 期;孙南翔、张晓君:《论数据主权——基于虚拟空间博弈与合作的考察》,《太平洋学报》2015 年第 2 期;张晓君:《数据主权规则建设的模式与借鉴——兼论中国数据主权的规则构建》,《现代法学》2020 年第 6 期;朱雅妮:《数据主权及其在〈数据安全法〉的体现》,《浙江工业大学学报(社会科学版)》2021 年第 4 期。

⁴⁷ See Jennifer Daskal, *The Un-Territoriality of Data*, 125 Yale Law Journal 326 (2015); Jennifer Daskal, *Borders and Bits*, 71 Vanderbilt Law Review 179 (2018).

⁴⁸ See Jennifer Daskal, *Borders and Bits*, 71 Vanderbilt Law Review 179, 224-225(2018).

⁴⁹ 辽宁省大连市中级人民法院(2022)辽 02 刑终 285 号刑事裁定书。

⁵⁰ 物理层指的是用于连接人与人之间的物质,包括计算机、电话、手持设备、电线、无线链路等。

⁵¹ 参见迈克尔·施密特,丽斯·维美尔:《网络空间国际法文库 网络行动国际法塔林手册 2.0 版》,黄志雄译,社会科学文献出版社 2017 年版,第 59 页。

⁵² 参见黄志雄:《网络主权论》,社会科学文献出版社 2017 年版,第 71 页。

⁵³ 逻辑层代表算法、标准以及一切将对人类有意义的通信内容转化为机器可以传输、存储或计算内容的方法;这包括标准、协议和软件,不局限于操作系统等通用平台,也包括更具体的应用软件。

行断网测试。俄罗斯通信部副部长阿列克谢·索科洛夫（Alexey Sokolov）表示，俄罗斯已成功完成了断开全球互联网后的一系列测试，例如遭遇外部“断网”时检查俄境内互联网运行的完整性和安全性。⁵⁴政府之所以可以关闭互联网，是因为逻辑层会受到国家领土范围的限制。数据以流动为主要特征，但政府一旦切断互联网服务，数据流动显然丧失了技术支持。

其三，数据本地化措施在内容层⁵⁵强化对数据的控制权。近些年来，发展中国家实行数据本地化的数量呈现增加态势。我国《网络安全法》第 37 条与《数据安全法》第 31 条要求关键信息基础设施的运营者在中国境内收集的个人信息与重要数据须在中国境内存储。除却关键信息基础设施运营者，《个人信息保护法》第 40 条进一步要求处理个人信息达到规定数量的个人信息处理者也须将收集和处理的个人信息存储在中国境内。2015 年 9 月俄罗斯实施的第 242-FZ 号联邦法案第 18 条规定，被视为个人数据运营者的组织必须确保收集或处理俄罗斯公民个人数据时通过位于俄罗斯的数据库进行。2021 年由于谷歌拒绝将俄罗斯用户的个人数据存储在位于俄罗斯境内的服务器上，莫斯科法院责令谷歌支付 300 万卢布罚款。⁵⁶2022 年 10 月越南政府信息和通信部发布的第 53/2022/ND-CP 号法令生效，依据这一法令第 5 章第 26 条，外国网络服务提供者须将越南用户创建的数据存储在越南境内。可见，数据在内容层的流动依旧受到国家领土边界的限制。

综上，主权从物理空间到数据领域的属地性从强转弱，但主权国家依旧可在物理层、逻辑层、内容层利用属地性对数据跨境流动予以管辖。法律管辖的对象并不是数据这一介质本身，而是数据背后的法律关系。数据主权的属地性变弱不足以推翻原有的国际管辖规则，成为国际管辖规则的例外。“数据例外主义”学说否认数据主权的属地性，本质是为数据控制者取证模式寻求理论支持。

2. 单边域外执法挑战他国数据主权

在国际习惯法规则中，域外执法管辖权受到国际习惯法的一般禁止，一国在他国内部的任何执法行为均被认为侵犯了他国对其境内执法管辖权的垄断。

《网络行动国际法塔林手册 2.0 版》规则 11 也对网络空间域外执法管辖的解释延续了国际习惯法，即只有在国际法赋予特定权力或外国政府对在其境内行使管辖权有效同意的情形下，一国方可对人、物和网络行为行使域外执法管辖权。⁵⁷

⁵⁴ 《俄罗斯举行首次国家级防“断网”演习》，载新华网 2019 年 12 月 25 日，<http://world.people.com.cn/n1/2019/1225/c1002-31522453.html>。

⁵⁵ 内容层是指人类相互之间发表的一系列具有意义的言论，既包括实际的语句，也包括基于人类交流而非机械处理的过滤、认证和解释机制。

⁵⁶ Daria Litvinova, *Russia fines Google for violating data storage law*, AP News (29-July-2021), <https://apnews.com/article/technology-europe-business-russia-data-storage-8cfce05469d996b6342899a2195ce6df>。

⁵⁷ 参见迈克尔·施密特，丽斯·维米尔：《网络空间国际法文库 网络行动国际法塔林手册 2.0 版》，黄志雄译，社会科学文献出版社 2017 年版，第 105 页。

数据控制者取证模式之所以引发争议，核心原因在于其赋予了执法机关或司法机关以域外执法的权力，侵犯了数据存储国对境内数据执法管辖的权力。尽管数据的特殊性质对国际管辖规则造成挑战，但数据的弱属地性不足以成为原有的国际管辖规则的例外。即使美国执法机关或司法机关依据证据开示制度或 CLOUD 法案在立法层面对网络服务提供者享有属人管辖权，但其单边调取境外数据的执法行为明确缺乏国际法依据。尽管美国司法部认为 CLOUD 法案并未将美国的管辖权扩展到任何新的当事方，强调“CLOUD 法案并不会改变对拥有或控制数据的实体的传统司法管辖要求”，但数据控制者取证模式并非只是实行域外立法管辖或域外司法管辖，其根本目的是赋予美国公权力机关以数据跨境取证的域外执法权。单边主义下一国制定的法案并不足以构成执法机构调取境外数据的合法性基础。美国司法部在白皮书中对 CLOUD 法案管辖权范围的解释本质上混淆了国际习惯法对域外立法管辖权和域外执法管辖权的不同限制。数据控制者取证模式导致数据存储国难以控制其他国家对本国境内数据的访问和获取，而控制权作为数据主权的核心表现，数据存储国的数据安全将受到冲击。即使在看似无国界的网络空间，未经国际法赋予特定权力或外国政府有效同意的数据域外执法行为依旧可能侵犯他国主权。因此，仅仅以取证高效为理由单方面强制调取存储在他国数据的行为违反国际管辖规则，极易引发主权冲突。

（二）数据控制者取证模式下的合规困境

除却主权冲突，数据跨境取证模式对他国个人数据保护与数据出境管制规则的忽视，将导致网络服务提供者陷入双重法律风险，甚至被迫承担行政或刑事责任。

1. 数据跨境调取与个人数据保护规范相矛盾

首先，借助互联网技术的跨境数据取证往往具有非同寻常的侵入性。由于跨境远程搜查的范围无法达到外科手术般的精确度，且执法机构需要调取的数据往往与其他数据相互混合在一起，难以剥离，因而在远程跨境数据取证的同时其他大量数据也会暴露。⁵⁸鉴于云计算技术下存储数据的设备可以存储的容量依然达到 TB 级别，一旦数据遭到泄露，对数据主体很可能造成财产损失和精神焦虑。

其次，由于各国对个人数据或隐私的保护程度存在差异，因而在跨境数据取证过程中，是否需要遵守他国对个人数据或隐私权保护的立法依旧存在争议。

⁵⁹2022 年，在 *Design Systems, Inc. v. Syntronic AB* 一案中，美国加利福尼亚州北区地方法院要求被告 Syntronic 北京技术研发中心有限公司（以下简称 Syntronic 公司）提供电脑交付美国以供检查数据。Syntronic 公司声称，依据《个人信息保

⁵⁸ See Orin S. Kerr, *Searches and Seizures in a Digital World*, 119 Harvard Law Review 531, 569 (2005).

⁵⁹ 参见王志刚,张雪:《跨境电子数据取证的困境与出路》,《重庆邮电大学学报(社会科学版)》2021 年第 5 期,第 50 页。

护法》第 39 条，被要求提供的电脑包含其雇员和前雇员的受保护的个人信息，转移到中国境外之前必须征得数据主体个人的单独同意，但其雇员和前雇员均拒绝同意转移至中国境外。因此，美国法院不应该要求 Syntronic 公司违反中国《个人信息保护法》。然而，美国首席治安法官 Spero 却将“外国法院的命令”解读为构成中国《个人信息保护法》第 13 条第 3 款“为履行法定职责或者法定义务所必需”这一例外情形。换言之，其认为美国法院发布的跨境调取数据的命令属于中国企业应当遵守的法定职责或法定义务，所以无须取得数据主体的同意，从而驳回了 Syntronic 公司的申请。⁶⁰

最后，监控类与秘密侦查措施的普遍实施涉嫌侵犯数据主体的个人数据权益、隐私权与知情权等多种权益。相比于国际司法协助，由于远程跨境数据取证的成本低廉且程序简化，且利用技术手段实施监控方便执法机关取证，因而在侦查实物中跨境监控取证依然成为公开的秘密。这导致被监控目标的个人数据权益、隐私权均可能遭到侵犯。此外，由于利用技术手段在网络空间跨境取证不易被数据主体知晓，这也为秘密侦查提供了便利。即使秘密侦查违反法定程序，数据主体也因知情权得不到保障而难以举证。⁶¹

2. 数据跨境调取与出境管制规则相抵触

数据领域的出境管制规则依照内容与实施效果可以划分为特定行业的数据出境管制规则、一般性数据跨境流通管制规则以及专门用于抵制数据控制者取证模式的阻断法。

首先，特定行业的数据管理规定对本行业的数据跨境传输设定禁止或限制条件，单方面跨境调取数据可能与特定行业的数据出境管制规则相矛盾，加重企业合规的法律风险。我国在证券、汽车、银行等领域均有禁止或限制向境外提供数据的条款，如下表所示。

时间	法律规范名称	条款	内容
2020 年	证券法	第 177 条第 2 款	未经国务院证券监督管理机构和国务院有关主管部门同意，任何单位和个人不得擅自向境外提供与证券业务活动有关的文件和资料。
2021 年	汽车数据安全管 理若干规定	第 11 条	重要数据应当依法在境内存储，因业务需要确需向境外提供的，应当通过网信部门会同国务院有关部门组织的安

⁶⁰ See *Cadence Design Sys. v. Syntronic AB*, No. 21-cv-03610-SI (JCS), 2022 U.S. Dist. LEXIS 112275 (N.D. Cal. June 24, 2022)

⁶¹ 参见梁坤：《数据主权与安全 跨境电子取证》，清华大学出版社 2023 年版，第 188 页。

全评估。

2023 年	中国人民银行业务领域数据安全管理办法》（征求意见稿）	第 27 条	非经中国人民银行和其他有关主管部门批准，数据处理者不得向其提供境内存储的数据。
--------	----------------------------	--------	---

表 2.1 我国特定行业数据出境管制规则

其次，数据跨境流动的一般性管制规则也会与数据控制者取证模式相矛盾。我国《网络安全法》第 37 条对境内个人信息与重要数据的跨境传输做出安全评估的规定。2022 年我国出台《数据出境安全评估办法》，重点评估数据出境活动可能对国家安全、公共利益、个人或者组织合法权益带来的风险。由于在中国境内运营中收集和产生的重要数据和个人数据的安全评估均受到《数据出境安全评估办法》的约束，且我国尚未针对司法案件中数据的跨境调取确立专门通道，因而外国执法机关或司法机关的跨境取证依旧受到我国对数据出境的一般性限制。

最后，数据出境管制规则中最具有对抗效果的便是阻断法。阻断法是国家为阻断第三国具有域外效力的法律在本国的适用而采取的反措施。⁶²为防止美国不当域外管辖，我国立法已经构建了多重阻断条款，如下表所示。

时间	法律规范名称	条款	内容
2018 年	《国际刑事司法协助法》	第 4 条	非经中华人民共和国主管机关同意，外国机构、组织和个人不得在中华人民共和国境内进行本法规定的刑事诉讼活动，中华人民共和国境内的机构、组织和个人不得向外国提供证据材料和本法规定的协助。
2021 年	《数据安全法》	第 36 条	非经中华人民共和国主管机关批准，境内的组织、个人不得向外国司法或者执法机构提供存储于中华人民共和国境内的数据。

⁶² 参见廖诗评：《〈阻断外国法律与措施不当域外适用办法〉的属事适用范围》，《国际法研究》2021 年第 2 期，第 44 页。

2021 年	《个人信息保护法》	第 41 条	非经中华人民共和国主管机关批准，个人信息处理者不得向外国司法或者执法机构提供存储于中华人民共和国境内的个人信息。
2024 年	《民事诉讼法》	第 294 条	除前款规定的情况外，未经中华人民共和国主管机关准许，任何外国机关或者个人不得在中华人民共和国领域内送达文书、调查取证。

表 2.2 我国抵制数据控制者取证模式的阻断法

然而，在前文提到的“*Philips Med. Sys. (Cleveland), Inc. v. Buan*”中，美国伊利诺伊州北区地区法院提到“外国的阻断法规并不剥夺美国法院命令受其管辖的一方出示证据的权力，即使出示证据的行为可能违反了该阻断法规，这一点已得到公认”。法院进一步分析，《数据安全法》并不代表中国在不披露特定数据方面的主权利益，而是试图将中国法院纳入美国法律程序，从而为中国诉讼当事人带来潜在利益，这样的法律侵犯了美国的主权。⁶³因此，法院不仅拒绝适用“国际礼让”审查，甚至否认中国对涉案数据存在主权利益。这显然与我国《个人信息保护法》和《数据安全法》的立法目的相违背。而在“*Intex Rec. Corp. v. Bestway (USA), Inc.*”一案中，美国加利福尼亚州中区地方法院也同样拒绝考量中国的阻断法，认为中国《数据安全法》的既定目标和明确意图是防止传播可能危害中国“国家安全”的数据，但没有任何证据表明涉案数据有可能危害中国的“国家安全”，因此要求厦门盈泰克公司继续披露数据。⁶⁴综上，美国数据跨境取证与中国阻断法的冲突在司法案件中较为明显，且美国法院拒绝考量中国阻断法的立场进一步加深中美在数据领域的摩擦。

三、理论调和：数据主权弹性适用

数据控制者取证模式所引发的冲突或矛盾折射出国家主体之间数据主权的激烈竞争。国家主体之间对数据主权的认知不同甚至相互对立。数据主权的弹性适用为调和数据跨境取证的冲突提供核心理念。

⁶³ *Philips Med. Sys. (Cleveland), Inc. v. Buan*, No. 19 CV 2648, 2022 U.S. Dist. LEXIS 35635 (N.D. Ill. Mar. 1, 2022).

⁶⁴ *Intex Rec. Corp. v. Bestway (USA), Inc.*, No. CV 19-8596-JAK(Ex), 2023 U.S. Dist. LEXIS 194563 (C.D. Cal. Oct. 30, 2023).

（一）弹性适用以谦抑性为法理基础

国家主权的适用范围从领土扩展到网络空间，国家主权的特性也从不可动摇的绝对性到受国际法原则约束的相对性得到认可，这也为数据主权的谦抑性提供了基础。

1. 数据主权相互依存

上个世纪 70 年代全球商品、人员和资本的跨境流通促使国家之间的联系更加紧密，“相互依存理论”在西方风靡流行。一般意义上，“相互依存”(interdependence)指的是国家之间的利益关系，即如果一个国家的立场发生变化，其他国家也会受到这种变化的影响。⁶⁵相互依存不局限于均衡依存，也可能是非对称性依存(asymmetries in dependence)，甚至不局限于互利的情景。⁶⁶学者基欧汉(Keohane)提出著名的“复合相互依存理论”，以“敏感性”和“脆弱性”两项因素分析权力在相互依存中的作用。“敏感性”用以描述一国政策变化导致其他国家有代价变化的速度，即一国受外部强加代价影响的程度。脆弱性被解释为国家因外部事件强加的代价而遭受损失的程度，即国家为有效适应外部国际环境变化而做出调整付出的成本。⁶⁷

大数据时代各类数据传播的成本几乎可以忽略不计，国家之间的联系程度也更加紧密。这也意味着国家受到另一方数据监管政策变化的影响会更深，即“敏感性”和“脆弱性”明显。例如欧盟法院先后宣布《安全港协议》与《隐私盾协议》的无效，直接导致支撑 7.1 万亿美元经济活动的跨大西洋数据流动陷入困境。这也迫使美国总统拜登于 2022 年发布《第 14086 号关于加强美国信号情报活动保障的行政令》(Executive Order 14086 on Enhancing Safeguards for United States Signals Intelligence Activities)以消除欧盟法院的顾虑。这项行政令不仅遵循欧盟保护个人数据的理念，更单独建立了约束机制帮助个人⁶⁸在遭到美国信号情报活动违法收集其个人数据时可以采取有效补救措施。可见，大型经济体之间数据主权相互依存的“敏感性”和“脆弱性”也会因为庞大的数字经济和商业需求而上升，欧盟数据跨境流动政策对美国产生较大影响并迫使美国遵守更加严格的数据保护标准。

此外，数据控制者取证模式产生的损害是双向的，冲突无法依靠一方主体单边解决。美国与欧盟的数据控制者取证模式在极大延伸其数据域外管辖权的同时，也同样遭到其他国家的抵制。⁶⁹中国针对数据单边跨境取证的阻断法以及中国、

⁶⁵ See Richard Rosecrance & Arthur Stein, *Interdependence: Myth or Reality?*, 26 *World Politics* 1, 1 (1973).

⁶⁶ 参见罗伯特·基欧汉，约瑟夫·奈：《权力与相互依赖》，门洪华译，北京大学出版社 2002 年版，第 9-11 页。

⁶⁷ 参见罗伯特·基欧汉，约瑟夫·奈：《权力与相互依赖》，门洪华译，北京大学出版社 2002 年版，第 12-13 页。

⁶⁸ 第 14086 号行政令将个人限定为合格国家和区域经济一体化组织的个人。

⁶⁹ 参见唐云阳：《安全抑或自由：数据主权谦抑性的展开》，《图书与情报》2022 年第 4 期，第 90 页。

俄罗斯所采取的数据本地化主权战略都是为抵消数据控制者取证模式的效果，抑制数据霸权的措施。中国作为数据大国采取的数据本地化措施毫无疑问会加剧外国企业的负担。例如，2021年5月汽车公司特斯拉宣称已在中国建立数据中心，以满足中国法律对数据存储本地化的要求，凡是在中国大陆市场销售的汽车所涉及的数据均将存储在中国境内。⁷⁰可见，数据控制者取证模式的实施在侵犯他国数据主权的同时，也对实施国家自身造成不利影响。总之，国家数据主权相互依存，大型经济体之间的数据主权相互依存度更高。

2. 兼顾数据的多元价值

其一，数据的财产价值要求数据主权保持一定的谦抑性。在云计算和数据分析等技术迅速普及的推动下，数据分析作为提高效率的手段在几乎所有行业予以应用。甚至，如果不跨境收集和发送个人数据，企业无法开展涉及消费者的国际贸易。⁷¹数据赋予数字经济巨大的能量，在流动与交换中产生价值。对于不涉及国家安全的数据，一味坚持主权的属地性很可能阻碍数据的流通，从长远来看不利于数字经济的增长。当数据不在限制流通的范畴之内，数据在世界范围内跨境传输所带来的利益远超出放弃独占数据产生的损失。⁷²在拥挤的互联网中，特定类型的通信从数据来源过传输至数据目的地所在国通过其他国家的服务器是不可避免的，这也被解释为国家的默示同意。⁷³因此，数据主权在维护主权不受侵犯的同时，包容数据流动对于开放性的需求。

其二，维护个人数据权益的需要要求数据主权具有谦抑性。维护个人数据权益意味着国家作为国际法律主体，有义务行使数据国际法律主权，签署国际合作协议以保障本国个人数据在国外得到有效保护。同时，作为国际合作的对价或妥协，国家也需要采取保障措施以确保外国主体个人数据在本国境内得到保护。个人数据权益与数据主权同时附着于数据之上，数据主权也应当尊重个人数据权益的行使。当数据主体因个人需要同意个人数据跨境传输，且传输的个人数据并不在数据来源国禁止的范畴，那么国家以数据主权为理由阻碍个人数据跨境传输的行为是没有依据的；否则，便不是保护个人数据权益，而是阻碍个人数据权益。

74

其三，企业对数据的掌控已经导致国家不可能对数据享有绝对的数据主权。

⁷⁰ 《跨国公司纷纷抢滩在华建数据中心，全球掀起“数据本地化”浪潮》，载新浪财经 2021 年 5 月 28 日，<https://finance.sina.com.cn/chanjing/gsnews/2021-05-28/doc-ikmxzfnm5225518.shtml>。

⁷¹ See Nigel Cory, *Cross-border data flows: Where are the barriers, and what do they cost?* Information Technology and Innovation Foundation(2017).

⁷² 参见卜学民,马其家:《论数据主权谦抑性:法理、现实与规则构造》,《情报杂志》2021年第8期,第67页。

⁷³ See Asaf Lubin, *The prohibition on extraterritorial enforcement jurisdiction in the datasphere*, in Austen Parrish and Cedric Ryngaert eds., *Research Handbook on Extraterritoriality in International Law*, Edward Elgar Publishing, 2023, p.347-350.

⁷⁴ 参见卜学民,马其家:《论数据主权谦抑性:法理、现实与规则构造》,《情报杂志》2021年第8期,第65页。

推动数字技术革新的网络服务提供者已变得极具影响力，其经济和政治实力足以媲美甚至超越民族国家。这也迫使部分国家不得不将第三方网络服务提供者与主权国家置于同等地位，并通过技术外交的方式影响大型数据平台。2017年，丹麦成为世界上第一个将技术和数字化提升为跨领域外交和安全政策优先事项的国家，这一举措被命名为科技外交（TechPlomacy）。同年丹麦任命了首位“科技大使”，科技大使不与国家互动，而是访问谷歌、脸书或亚马逊等公司的总部。其科技外交的运作包括两个方面：第一，同其他使馆一样，科技大使代表丹麦当局与科技公司进行对话以试图影响技术的发展方向；第二，根据丹麦的利益和价值观，围绕科技政策问题影响国际议程，包括通过新的联盟、多边论坛和多利益攸关方伙伴关系。⁷⁵相似的，2017年法国也提名了首位数字大使，以协调法国在数字领域国际问题上的立场构想，并向国际合作伙伴以及公共和私人主体宣传法国立场。⁷⁶可见，国家享有的数据主权会受到企业，尤其是大型科技公司的制约。

综上，数据主权具备谦抑性。国家数据主权之间相互依存，行使数据主权应兼容数据的财产价值、个人数据权益以及企业的数据权益。

（二）数据主权弹性适用的实现路径

平衡数据主权的属地性与谦抑性是数据主权弹性适用的关键。在维护数据主权的独立权、平等权、自卫权与管辖权等核心权益的同时，顺应数据跨境取证需求，探索数据主权得以弹性适用的路径。

1. 数据主权的多维度划分与组合

鉴于传统的主权理论无法适应新的国际形势，学者们对主权概念的理解不再是静态单一的，而是逐步划分为多层次多维度。学者 Krasner 对主权内涵横向划分为四种不同的维度⁷⁷：国内主权（domestic sovereignty）⁷⁸、相互依赖主权（interdependence sovereignty）⁷⁹、国际法律主权（international legal sovereignty）⁸⁰、威斯特伐利亚主权（Westphalian sovereignty）⁸¹。不同维度的主权不一定是共存的。⁸²我国学者张军旗将主权划分为身份意义上的主权⁸³与权能意义上的主权⁸⁴。前者与学者 Krasner 所划分的国际法律主权有一定的相似之处，后者也与“国

⁷⁵ *The TechPlomacy Approach*, Ministry of Foreign Affairs of Denmark (10-November-2024), <https://techamb.um.dk/the-techplomacy-approach>.

⁷⁶ See Lily Martinet, *Exercising Digital Sovereignty over Blockchains: A Case Study from France*, 4 Stanford Journal of Blockchain Law & Policy 1, 12-13(2021).

⁷⁷ See Stephen D. Krasner, *Sovereignty: Organized Hypocrisy*, Princeton university press, 1999, p.4.

⁷⁸ 国内主权指的是一个国家内掌握公共权力的机构对国内事项行使权力的有效控制程度。

⁷⁹ 相互依赖主权指的是公共权力机构控制跨境流动的能力。

⁸⁰ 国际法律主权指的是国家或其他实体的相互承认。

⁸¹ 威斯特伐利亚主权指的是将外部行为者排除在国内权力构架之外。

⁸² See Stephen D. Krasner, *Sovereignty: Organized Hypocrisy*, Princeton university press, 1999, p.4.

⁸³ 身份意义上的主权解释为国家在国内社会以及国际社会中的法律地位，代表着质的规定性。

⁸⁴ 权能意义上的主权表现为对本国人民、领土范围内的外国人以及本国内外事务的控制力与支配力。

内主权”、“威斯特伐利亚主权”以及“相互依赖主权”具有一定的重叠。⁸⁵可见，无论是国内或国外，主权均被认为包含不同的维度。

在国际关系中，国家作为国际法律主体，为获得合理的对价（如互惠）让渡自身部分主权利能，同时保留随时恢复让渡权能的方式。主权让渡的本质是国家自主决定行使主权的方式。身份意义上的主权或国际法律主权无法让渡。一旦让渡，国家作为国际法律主体资格便不复存在。权能意义上的主权包含范围广泛，国家可以选择让渡部分主权。但值得注意的是，权能意义上的主权并非可以随意让渡，若部分权能意义上的主权让渡之后导致主权的终极支配权或作为身份依托的主权受到损害，则这一权能意义上的主权也应当得到维护。无论何种让渡，国家均可以随时收回让渡出去的主权利能，未让渡的剩余主权利能也依旧由国家紧紧掌握。换言之，为不损伤主权的根本属性，主权的让渡是自主的且有限的。⁸⁶

主权涵义的多维度划分为数据主权的弹性适用提供了实现路径。⁸⁷对数据主权至关重要的并非不受外部干扰的自由，而是参与国际机构的构成和国际条约的自由。⁸⁸换言之，只有具备机会参与国际社会数据治理的主权才是有意义的，脱离国际社会而不受干扰的数据主权并不符合国家利益。依据学者 Krasner 对主权内涵的划分，数据主权的内涵也可以划分为不同维度：第一，数据国内主权，即国家对境内数据的控制力度；第二，数据威斯特伐利亚主权，即排斥他国干涉本国境内数据；第三，数据相互依赖主权，即国家公权力机构对跨境数据的控制力；第四，数据国际法律主权，即国家以独立法律主体的身份同其他国家主体签订与数据相关的协议或加入国际组织。

在数据主权的四项维度中，国家在保障数据国际法律主权不受侵蚀的前提下，其他三项不同维度的数据主权可以依据国家利益的需求予以适当的强化或减弱，并在不同场景下进行组合。对于被请求国而言，数据跨境取证会削弱数据威斯特伐利亚主权，并同时降低政府控制境内部分数据的能力，即在一定程度上弱化数据国内主权。但在被请求国执法机关未来侦查案件须跨境调取数据时，基于互惠，数据跨境取证的合作也将帮助其提升数据相互依赖主权。2020 年最高人民检察院发布的指导性案例“张凯闵等 52 人电信网络诈骗案”中，肯尼亚将涉案笔记本电脑、手机等物证移交我国公安机关，帮助后期公安机关与检察机关获取犯罪嫌疑人的通信数据。⁸⁹电信诈骗案件中犯罪嫌疑人的聊天记录并不会影响到肯尼亚的国家安全与公共利益。我国与肯尼亚之间良好的司法合作虽看似减弱肯尼亚数据威斯特伐利亚主权与数据国内主权，但事实上尊重了肯尼亚的国家主权，在后

⁸⁵ 参见张军旗：《主权让渡的法律涵义三辨》，《现代法学》2005 年第 1 期，第 98-102 页。

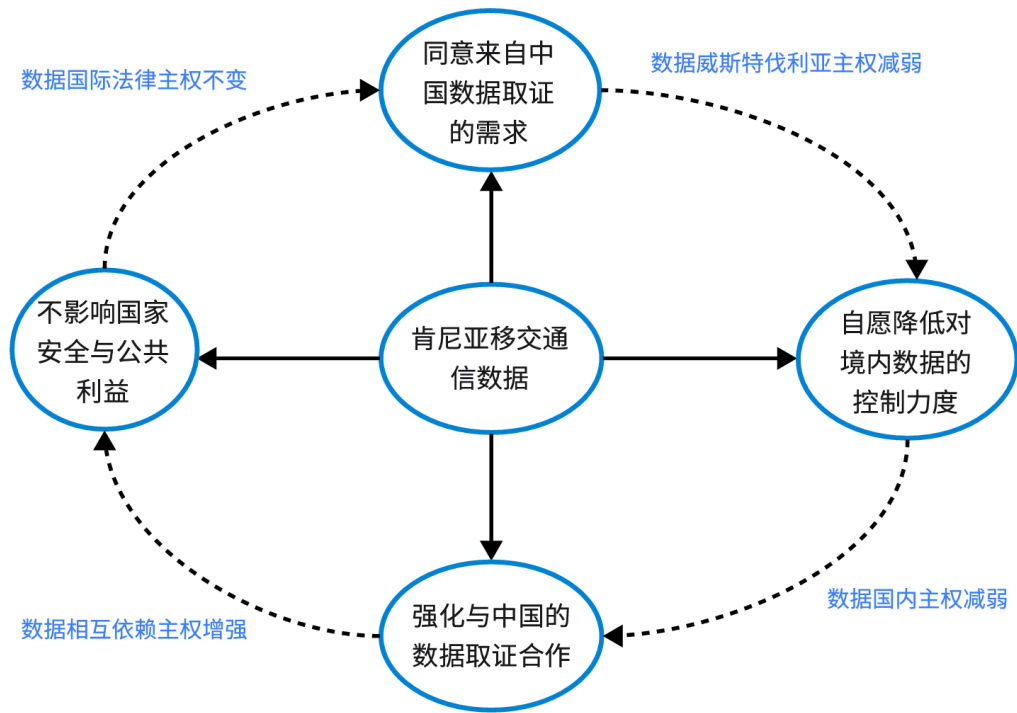
⁸⁶ 参见刘凯：《国家主权自主有限让渡问题研究》，中国政法大学出版社 2013 年版，第 82-89 页。

⁸⁷ 参见许可：《自由与安全：数据跨境流动的中国方案》，《环球法律评论》2021 年第 1 期，第 36 页。

⁸⁸ See Adeno Addis, *The Thin State in Thick Globalism: Sovereignty in the Information Age*, 37 *Vanderbilt Journal Transnational Law* 1,56 (2004).

⁸⁹ 参见《关于印发最高人民法院第十八批指导性案例的通知》，《人民检察》2020 年第 8 期，第 46-47 页。

续司法的长期合作中，强化了其数据相互依赖主权（如下图）。



2. 配合程序主义与数据分级分类

程序主义与数据分级分类为数据主权不同维度的相互组合进一步提供现实策略。其一，学者 Addis 将主权的程序主义解释为允许与结果有利害关系的国家主体平等参与国际决策过程。当数据的单边监管影响到他国利益时，从规范上讲，所有受到影响的国家主体均有发言权，并通过国际互动、谈判和协议来实现其立场。⁹⁰程序主义式的数据主权意味着数据跨境取证领域以获取他国政府同意或国际条约的授权为前提。数据主权不同维度的组合也依赖于国家主体在平等协商中达成一致协议，而为了防止民族国家之间的相互竞争，国际社会的多边协议需要为数据跨境取证的条件寻求“最大公约数”。《联合国打击网络犯罪公约》是国际社会应用程序主义弹性适用数据主权的典型代表。以尊重主权为基础，综合考量数据安全、网络空间公私合作的需求以及个人数据保护等多方面因素，《联合国打击网络犯罪公约》第 27 条“提交令”就“用户信息”这一类数据实行数据控制者取证模式达成共识。此外，2022 年《中华人民共和国和肯尼亚共和国关于刑事司法协助的条约》第 19 条允许中国与肯尼亚相互请求提供与犯罪行为有关的银行

⁹⁰ See Adeno Addis, *The Thin State in Thick Globalism: Sovereignty in the Information Age*, 37 Vanderbilt Journal Transnational Law 1,60 (2004).

信息。这一请求除适用于银行外，也可以适用于其他金融机构。被请求方不得以银行保密为理由而拒绝本条所规定的请求。在这一司法协助条约中，中国与肯尼亚基于主权尊重与互惠，通过国际谈判让渡了金融数据所蕴含的数据威斯特伐利亚主权与数据国内主权，相互强化对跨境金融数据流动的控制力即数据相互依赖主权。

其二，除却程序主义，数据主权的多维度组合也须配合数据分级分类。面对不同类型的数据，主权遭受侵犯的程度并不相同，所引发的后果也具有较大差异。我国《数据安全法》根据数据的重要程度以及非法滥用后的危害程度，将数据划分为一般数据、重要数据和核心数据。与我国立法从安全视角的划分不同，现有涉及数据跨境取证的国际条约中倾向于将网络服务提供者掌握的数据划分为“用户信息”、“流量数据”以及“内容数据”，不同类型数据跨境调取的条件也由弱趋强。⁹¹尽管不同视角下划分的种类不同，但数据分级分类的本质均是为了区分不同数据背后的主权属性。若数据泄露引发的风险波及国家安全或公共利益，且需要国家主动保护，这意味着数据具有较强的主权属性；相反，风险或损害局限于私域，且主体享有救济方式，则数据所附有的主权属性较弱。⁹²例如《联合国打击网络犯罪公约》第 27 条“提交令”将范围局限于“用户信息”，便是因为用户信息被强制调取在大多数场景下不涉及公共利益或国家安全，个人数据面临的风险限于私域范围，主权属性较弱。而针对其他数据，国家主体须在平等协商或具体案件中组合不同维度的数据主权以满足自身对数据跨境取证的个性化需求。

综上，数据主权的弹性适用以谦抑性为法理基础，以多维度数据主权的划分和组合作为实现路径，并配合程序主义与数据分级分类的实施。在数据跨境取证的场景下，国际合作以平等协商为前提，针对主权属性较弱的数据，通过适当让渡一定的数据威斯特伐利亚主权与数据国内主权以获得数据相互依赖主权的提升（如下图）。

⁹¹ 参见梁坤：《网络犯罪国际公约中的跨境调取数据条款》，《当代法学》2024 年第 2 期，第 128 页。

⁹² 参见吴玄：《数据主权视野下个人信息跨境规则的建构》，《清华法学》2021 年第 3 期，第 84 页。

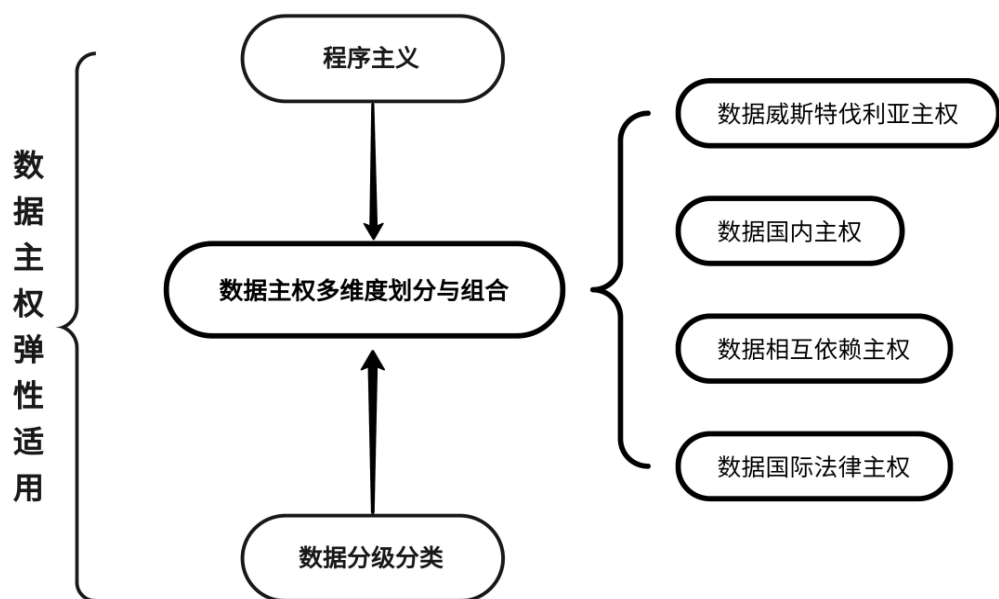


图 3.2 数据主权弹性适用的实现路径

四、数据控制者取证模式下的中国方案

与欧盟、美国采取数据控制者取证模式不同，我国在数据领域立法导向具有明显的防御倾向，采取数据本地化措施维护我国数据主权。中国在参与搭建数据跨境取证合作机制和为网络犯罪全球性公约的谈判贡献力量的同时，可进一步从“攻”与“守”两个方面优化中国数据跨境取证的方案。

（一）进取：调整国际司法协助程序与建立专属评估通道

面对数据控制者取证模式引发的冲突，一味“防守”不利于我国参与全球数据治理。我国可以通过构建数据跨境取证的进取型路径实现数据主权的弹性适用。

1. 依托联合国推动国际司法协助程序升级

一方面，依托联合国平台，在跨境数据取证中嵌入“一字型”结构的司法协助程序。⁹³借助《联合国打击网络犯罪公约》的商谈，确立“一字型”结构是优化司法案件数据跨境取证程序的有效路径。“一字型”结构即缔约国执法机关或司法机关的直接合作，在尊重主权的同时提高跨境数据取证的效率，改进数据存储地取证模式。《联合国打击网络犯罪公约》显然已经认识到优化司法协助程序的重要性。第 47 条对缔约国之间的“执法合作”提出大致的合作方向。其中，第 47（1）a 款⁹⁴明确必要时建立各缔约国主管机构之间的通信渠道以交换案件所涉及的数

⁹³ 参见王立梅：《论跨境电子证据司法协助简易程序的构建》，《法学杂志》2020 年第 3 期，第 90 页。

⁹⁴ 《联合国打击网络犯罪公约》第 47（1）a 款：加强并在必要时建立各国主管机关、机构和部门之间的通信渠道，以促进安全而迅速地交换有关本公约所涵盖的犯罪的各方面信息，在有关缔约国认为适当时还

据；第 47（1）c 款暂定“在适当情况下提供必要的物品或数据以供分析或侦查之用”。虽然上述规定提供了指引，但可操作性却不足。“必要时”以及“适当情况下”的表述含糊不清，并未对缔约国产生强制执行效力。即使《联合国打击网络犯罪公约》第 47 条第 2 款提出“缔约国应当考虑订立其执法机构间直接合作的双边或多边协定或安排”，但这一条款并未明确缔约国在直接合作时是否应当披露或传送数据，显然是将这一事项交由缔约国自行判断。这就导致数据跨境取证事项上司法协助程序的简化效果大打折扣。

对此，《布达佩斯公约》搭建的多边合作下，数据跨境取证的“一字型”司法协助程序值得借鉴。《布达佩斯公约》第二附加协定书第 8 条和第 9 条允许某一缔约国的主管当局向另一缔约国的主管当局直接提出披露用户信息、流量数据的请求，并由后者负责联系其境内的网络服务提供者予以提供数据。考虑到不同数据主权属性的差异，《布达佩斯公约》第二附加协定书第 8 条第 13 款⁹⁵允许缔约国对主管当局之间就流量数据的合作享有保留的权利。结合数据主权的弹性适用，国家主体可以考虑借助联合国这一全球性平台进行谈判。遵循程序主义的同时对主权属性较弱的流量数据允许国家主管机构就跨境调取直接开展合作，强化数据相互依赖主权，并为不愿参与直接合作的国家主体提供保留的权利。国家主管当局之间就数据披露的直接合作，免除了刑事侦查或诉讼中“倒 U 型”引发的程序障碍，且依旧属于司法协助的范畴，不侵犯他国数据主权。

另一方面，统一数据跨境取证的审查因素，明确时间限制。《联合国打击网络犯罪公约》对数据跨境调取的司法协助进行了一定程度的改进。例如第 41 条要求各缔约国设立全天候网络平台这一联络点；第 40 条第 21 款⁹⁶规范设计缔约国拒绝提供司法协助的情形。但上述条款的内容对于完善高效的司法跨境数据取证制度尚不全面，笔者从以下角度进行改进。

其一，明确跨境调取不同类型的数据之下被请求国考虑的审查要素。《联合国打击网络犯罪公约》第 45 条对实时收集流量数据的司法协助中缔约国提出请求时应当详细说明的因素进行列举，例如侦查或诉讼涉及的主要事实和性质、流量数据与犯罪之间的关系、数据所有人或用户的身份、需要收集流量数据的理由等。为跨境调取数据提供明晰的审查因素将极大改进现有的司法协助制度，不仅为跨境调取数据提供统一性和确定性，而且对请求国和被请求国的执法人员开展

可包括与其他犯罪活动的关联的信息。

⁹⁵ 《布达佩斯公约》第二附加协定书第 8 条第 13 款：在签署本议定书时或在交存其批准书、接受书或核准书时，缔约国可保留不对流量数据适用本条的权利。

⁹⁶ 《联合国打击网络犯罪公约》第 40 条第 21 款：在下列情形中可拒绝提供司法协助：

- (a) 请求未按本条的规定提出；
- (b) 被请求缔约国认为执行请求可能损害其主权、安全、公共秩序或其他基本利益；
- (c) 假如被请求缔约国主管部门依其管辖权对任何类似犯罪进行侦查、起诉或司法程序时，其国内法将会禁止其对此类犯罪采取被请求的行动；
- (d) 若同意此项请求将会违反被请求缔约国关于司法协助的法律制度

司法协助提供指引。鉴于公约只对实时收集流量数据中审查的因素进行统一，可以将这一立法模式延伸到其他类型的数据或情形中，如统一缔约国请求拦截内容数据的审查因素。⁹⁷

其二，《联合国打击网络犯罪公约》可进一步对被请求国回复请求国数据跨境调取做出时限要求，并对不同类型数据的时限要求做出区分。虽然公约第 40 条第 26 款规定被请求国尽快执行司法协助请求，尽可能充分考虑到请求国提出的最后期限，但是公约对数据跨境取证的司法协助尚未规定明确的时间限制。对此，可以考虑借鉴《布达佩斯公约》第二附加议定书第 8 条第 6 段 a 项规定，被请求国在接受到所有的信息后采取合理措施，要求其境内的网络服务提供者在 45 天内提供通讯数据或 20 天提供用户信息。

2. 数据出境专属风险评估通道的搭建

首先，明确司法案件中数据出境风险评估主体。我国《数据出境安全评估办法》第 4 条规定，满足一定条件的数据处理者应当通过所在地省级网信部门向国家网信部门申报数据出境安全评估。可见，我国将国家网信部门作为评估数据出境风险的主体。但司法案件的侦查和诉讼具有较强的专业性，网信部门缺乏了解司法诉讼、案件侦查等领域的专业人才，无法对数据出境的风险做出有效判断。⁹⁸因此，在国际司法协助中评估数据出境风险的主体须具备侦查或司法诉讼的能力。一方面，就民商事案件而言，依据多边或双边民商事司法协助条约我国司法部直属单位司法协助交流中心负责协助跨境取证。2023 年约三分之一的调查请求系通过司法协助中心线上系统递交。⁹⁹可见，我国司法部已然就民商事案件的国际司法协助积累了实践经验，应为网信部门提供数据跨境取证评估指导的主体。另一方面，就刑事案件而言，我国《国际刑事司法协助法》第 6 条第 1 款明确“国家监察委员会、最高人民法院、最高人民检察院、公安部、国家安全部等部门是开展国际刑事司法协助的主管机构”。其中最高人民检察院平均每年依照刑事司法协助条约处理案件近 200 件，占全部刑事司法协助主管机构处理司法协助案件数量的三分之一。¹⁰⁰鉴于最高人民检察院在处理司法案件中的专业性与中立性，其在我国网信部门的指导下评估司法数据出境最为合理。

其次，明确司法案件中数据出境风险评估内容。第一，数据的出境是否涉及数据安全，尤其是国家安全。若评估出境的数据遭到破坏后将危害国家安全与公共利益，原则上不应允许出境。数据安全影响评估宜综合考虑数据类型、数据内

⁹⁷ 参见梁坤：《网络犯罪国际公约中的技术侦查条款及国内程序法回应》，《公安学研究》2023 年第 4 期，第 54 页。

⁹⁸ 参见郑曦：《刑事数据出境安全评估制度研究》，《法学论坛》2023 年第 4 期，第 132-133 页。

⁹⁹ 《2023 年民商事司法协助案件统计》，载中华人民共和国司法部网站 2024 年 5 月 28 日，https://www.moj.gov.cn/pub/sfbgw/jgsz/jgszsdw/zsdwsfxzjlzx/sfxzjlzxxwdt/202405/t20240528_499499.html

¹⁰⁰ 参见阮丹生，潘新艳，张宏峰等：《加强国际刑事司法协助 提升涉外执法司法效能》，《人民检察》2023 年第 12 期，第 23 页。

容、数据规模、数据来源等因素，一旦遭受破坏可能给一国金融、外交和国防等领域、公共利益以及企业和个人合法权益带来的安全风险。司法案件中跨境调取的数据在安全性遭到破坏后，可能对国家政权、领土主权、民族团结、社会和金融市场稳定等造成影响，则意味着达到影响国家安全的程度。对生产经营、教学科研、医疗卫生、公共交通等社会秩序和公众的政治权利、人身自由、经济权益等造成影响，则达到影响公共利益的情况。因而，对我国而言，评估主体须需要结合具体案情对数据进行盘点与梳理，形成统一的数据资产清单；结合所在国家及行业有关法律法规，综合考虑数据规模、匿名化技术对数据安全风险进行判定。若被调取的数据归属于核心数据或重要数据，则评估主体应拒绝请求国的申请。

最后，成本费用的补偿也是司法案件数据出境过程中执法机构或司法机关需要考虑的问题。¹⁰¹《联合国打击网络犯罪公约》第 27 条提交令基于属人管辖已然允许在特定情形下单方面跨境调取订户信息，但司法诉讼中跨境数据取证的大量增加毫无疑问会给网络服务提供者造成额外的费用。大型科技公司考虑到与公权力机构合作的声誉价值以及财务成本相对较低，倾向于不要求政府或司法机关的费用补偿；但规模较小的企业面对大量的数据请求并在处理中耗费大量资源，则较为需要公权力机构的费用补偿。¹⁰²德国《司法薪酬与补偿法》（Judicial Remuneration and Compensation Act）第 23 条规定“只要截取电信的命令或信息是由提供电信服务或参与电信服务的人（电信公司）执行的，在本法附件 3 中规定了特殊补偿的情况下，补偿应完全按照本附件计算。《美国法典》（18 U.S.C. §2706）规定“获取通信内容、记录或其他信息的政府实体应向收集或提供此类信息的个人或实体支付费用，以补偿……直接产生的合理必要的费用。”我国目前在跨境取证中对网络服务提供者承担数据披露义务方面并未规定经济补偿，立法可以考虑这一措施以减轻数据跨境取证给企业带来的负担。

（二）防御：完善阻断措施并引入数据信托云

尽管阻断法的实施效果不佳，但其表明我国对数据控制者取证模式的立场，是敦促他国与我国开展国际谈判的筹码。衔接国内阻断法与国际协议，并帮助企业调整合规策略是我国可有效采取的防御方案。

1. 阻断法的优化

其一，强化阻断立法的执行，并配合豁免程序予以实施。阻断立法的执行性会直接影响他国司法机关的评价，强化阻断立法的实施有助于受到他国法院的认可。但一味加强执行亦可能导致个人信息处理者陷入两难。是故，面对他国行政

¹⁰¹ 参见裴炜：《论刑事诉讼中网络信息业者的数据提供义务》，《上海政法学院学报(法治论丛)》2022 年第 6 期，第 74 页。

¹⁰² Catherine De Bolle, Ladislav Hamran&Didier Reynders, 2022 SIRIUS EU Digital Evidence Situation Report, Europol(02 December 2022),<https://www.europol.europa.eu/publications-events/publications/sirius-eu-digital-evidence-situation-report-2022>。

机关或司法机关跨境调取数据的要求,在与我国阻断法相互冲突的情况下应依据具体案件做出不同分析。首先,当我国企业因司法诉讼或刑事侦查被国外执法机关或司法机关要求提供相关数据时,应先判断是否属于豁免情形。《联合国打击网络犯罪公约》第 27 条提交令的存在意味着企业在部分情形下享有豁免。换言之,在特定情况下即使向外国执法机关或法院提供相关数据,也无需因阻断法的存在担责。其次,阻断法的实施可以与我国司法协助的简化调整相互协调。若依照国际公约或国内立法,传票涉及的数据不属于企业可以豁免的情形时,企业应立刻向我国负责评估司法数据出境风险的机构报告,即转向司法协助的安全评估。为打消国外执法机关或法院对司法协助效率低下的顾虑,负责评估数据出境风险的机构可以出具文件告知数据出境评估的具体流程、考量因素以及评估的截止日期,提升透明度以回应国外执法机关或法院的需求。最后,若国外执法机关或法院依旧要求我国企业直接提供数据或我国评估后发现调取的数据出境后给国家安全、公共利益或个人权益带来的风险较大,此时阻断法应当得到遵守。即使我国的个人信息处理者面临他国法律的处罚,仍然需要遵守阻断立法,否则将承担我国法律责任。因而,我国应当结合实践需求,在强化阻断立法执行性的同时对豁免程序进一步优化,以缓解网络服务提供者面临的压力。

其二,阻断法与反制措施相配合,共同作为我国在数据跨境取证领域的角逐工具。本质上来看,阻断法的宣誓效果大于其实际作用,因为阻断法的实施取决于国家实力的对比;¹⁰³另一方面,即使《阻断外国法律与措施不当域外适用办法》(以下简称《办法》)第 11 条¹⁰⁴明确遵守我国阻断禁令的国内企业所获取的支持,但这一支持究竟是什么,并没有详细解释。我国政府可以对企业从哪些方面进行补偿尚不清晰。况且,《办法》的效力等级只是部门规章。是故,仅仅依靠阻断法并不能达到防御效果。基于国际法对等原则,我国《个人信息保护法》第 43 条与《数据安全法》第 26 条分别规定我国针对其他国家在个人数据保护或数据开发利用技术领域采取的歧视性的禁止、限制或者其他类似措施可采取“对等反制措施”。虽然此类措施并不专门针对于数据跨境取证,但我国可以明确将“对等反制”引入数据跨境取证。针对不在我国阻断法豁免范围内且违反我国司法数据风险评估流程的单方面跨境调取行为,我国可以设置“黑名单”。面对他国单方面在侦查案件中对存储于我国的数据远程取证,我国保留使用同种侦查措施远程调取数据的权利。¹⁰⁵对等反制与阻断法相互配合,有助于实施数据控制者取证模式的国家与我国达成妥协,促进其与我国进一步谈判,更有效地防御单方跨境取证对我国的侵害。

¹⁰³ 参见霍政欣:《国内法的域外效力:美国机制、学理解构与中国路径》,《政法论坛》2020 年第 2 期,第 187 页。

¹⁰⁴ 《办法》第 11 条:中国公民、法人或者其他组织根据禁令,未遵守有关外国法律与措施并因此受到重大损失的,政府有关部门可以根据具体情况给予必要的支持。

¹⁰⁵ 参见梁坤:《数据主权与安全 跨境电子取证》,清华大学出版社 2023 年版,第 209 页。

2. 企业借助数据信托云调试合规策略

“数据信托云”（data Trust）可以为企业提供保护，免受他国域外请求的影响。“数据信托云”的本质是将客户数据的管理权限想分离。网络服务提供者将用户数据移交给国外独立的数据受托公司，其在组织和法律上独立于前者并通过信托协议保障自身独立性。信托协议将托管的数据访问权集中于受托公司而非网络服务提供者。受托公司控制对用户数据的所有访问，而网络服务提供者未经受托公司同意或没有合同规定有限原因（如维修）无法访问特定数据。网络服务提供者与数据受托人、网络服务提供者与其客户以及数据受托人与客户之间的关系由所有三方之间具有约束力的合同加以调整和执行。¹⁰⁶因而，“网络服务提供者”可证明自身不满足“占有、保管或控制”这一条件。

鉴于“微软案”的影响，微软公司为其欧盟区域的用户提供了德国“数据信托云”服务。微软公司将欧盟区域内的用户数据移交给德国 T-Systems 公司存储。T-Systems 公司是独立的德国公司，在法律上不受美国管辖。基于微软公司与 T-Systems 公司的信托协议，只有 T-Systems 公司才有访问数据的技术能力和法律权力。从技术层面来看，微软在德国“数据信托云”中的用户数据是加密的，只有 T-Systems 公司掌握数据的密钥，微软在技术上无法访问数据。从法律层面来看，只有在用户允许或出于某些特定目的（包括技术维护和维修服务），微软公司方能在 T-Systems 公司的监督下访问。¹⁰⁷根据有效的数据信托安排，受美国司法管辖的微软公司及其德国子公司可以有力地证明其德国“数据信托云”中的数据不在其“占有、保管或控制”之下。依据其信托条款，只有数据受托人才能访问并回应传票请求所需的数据。这意味着美国执法机关或司法机关要求披露数据的请求并非是向微软公司提出，而是需要向 T-Systems 公司提出。鉴于 T-Systems 公司不在美国管辖范围内，美国执法机关或司法机关也无法依据 CLOUD 法案要求德国 T-Systems 公司单方面披露数据，更无法从技术层面要求微软公司强制披露。此外，当法院进行礼让分析并评估所审理案件的国家利益时，T-Systems 公司与微软公司之间的信托关系被视为代表实质性国家利益。数据受托人 T-Systems 公司可声称外国披露请求会损害“国家的重要利益”¹⁰⁸从而在反对披露数据方面获得强有力的依据。

综上，借助信托协议，“数据信托云”将数据访问权仅局限于数据受托人行使，使网络服务提供者从技术与法律层面均不再满足“占有、保管或控制”数据这一要件，有助于网络服务提供者免受数据控制者取证模式的干扰。

¹⁰⁶ See Paul M. Schwartz/Karl-Nikolaus Peifer, Datentreuhändermodelle–Sicherheit vor Herausgabeverlangen US-amerikanischer Behörden und Gerichte?, Computer und Recht 33(2017), S.172.

¹⁰⁷ See Paul M. Schwartz, *Legal Access to the Global Cloud*, 118 Columbia Law Review 1681,1698-1699 (2018).

¹⁰⁸ See Paul M. Schwartz, *Legal Access to the Global Cloud*, 118 Columbia Law Review 1681,1735 (2018).

五、结语

《联合国打击网络犯罪公约》在有限范围内对数据控制者取证模式达成共识，为国际司法协助提供补充，但以国内法为依据单方面实施数据跨境取证的案例依旧频繁出现。司法审判暴露了数据主权的冲突与合规困境这两大矛盾，但现有各国对数据主权的认知差异以及在数据保护和跨境流通的不同策略导致这一问题难以得到有效解决。数据主权的弹性适用在尊重数据主权属地性的同时以谦抑性为基础，实现多维度划分与组合，配合程序主义与数据分级分类有助于国家主体的自主选择。同时，我国一味采取防守策略已然失效，完善“攻防兼备”的中国方案是学界的关注焦点。就进取方案而言，在维护数据主权的同时，依托于《联合国打击网络犯罪公约》的良好契机，推动国际司法协助程序的升级以及国内专属风险评估通道的建立，促进我国国内法与国际协议的衔接是当务之急。就防御方案而言，阻断法的效果较为有限，但配合豁免与反制措施，依旧是我国在数据跨境取证领域参与国际谈判的法律工具。数据信托云通过规避数据控制者取证模式的要件，对实现企业合规具有积极意义。可以预见，在未来数据控制者取证模式与数据存储地取证模式将长时间并行行使，数据跨境取证的冲突也将继续存在。面对数据控制者取证模式的压力，我国不仅要积极参与全球数据跨境取证规则的制定，更须针对性调整数据防御战略，探索攻防兼备的中国道路。

Practice Contradictions, Theoretical Harmonization and Coping Strategies of The Data Controller Forensic Model

Xue Wang

Abstract: The “data controller” forensic model focuses on the personal jurisdiction of network service providers, combining legal right standard with practical ability standards to define the scope of data control, thus implementing cross-border data retrieval. The weak territoriality of data sovereignty is insufficient to overturn the restrictions of extraterritorial jurisdiction rules on law enforcement, and unilateral extraterritorial enforcement of the data controller's forensic model infringes on data sovereignty of other countries. Furthermore, there is a sharp conflict between the data controller's evidence-gathering model and personal data protection regulations, as well as rules on the control of data outbound. To reconcile the practical contradictions of the data controller's forensic model, data sovereignty, based on modesty, can be divided and combined in multiple dimensions, and combined with the implementation of proceduralism and data classification, to achieve flexible applicability. To cope with the spread of the data controller's forensic model, China also needs to construct a cross-border forensic scheme that integrates both offensive and defensive strategies. China

can leverage negotiation opportunities for international agreements to simplify traditional international judicial assistance, establish exclusive risk assessment channels for data forensic, while optimizing blocking clauses and introducing data trust clouds to help enterprises avoid the jurisdiction of data controller's forensic model.

Key Words: Data Controller forensic Model; Cross-border Data forensic; Sovereignty Conflicts; Compliance Dilemma; Elasticity of Sovereignty

2025年中国法学年会

《联合国打击网络犯罪公约》中网络性侵儿童犯罪条款及我国刑事 法因应

赵安晓宇*

摘 要：《联合国打击网络犯罪公约》中网络性侵儿童犯罪条款的设置为国际社会合力打击该类犯罪提供了国际法层面的法律遵循。现有涉及网络性侵儿童犯罪条款的国际公约主要分为“儿童权利保障性”国际公约及区域性打击（网络）犯罪公约，二者对罪名及犯罪情节规定各有侧重点。《联合国打击网络犯罪公约》网络性侵儿童犯罪条款设置经历了多次调整，侧重规制网络空间中“儿童色情制品类”与“教唆或诱骗儿童”行为；我国现有刑事法体系能部分涵盖上述犯罪行为，但在“持有”儿童色情制品行为、“性引诱”儿童行为、（跨境）侦查取证方面需进行必要的国内法衔接。域外典型国家将国际公约转换为国内法同时，积极开展网络性侵儿童犯罪的刑事法治理实践，成效较为明显。基于刑事一体化背景，我国可从刑事实体与程序两方面入手开展网络性侵儿童犯罪治理工作，包括但不限于：对“性引诱”行为入罪化处置；规制“持有”儿童色情制品行为；优化跨境侦查取证规则；探索网络取证侦查措施。

关键词：网络犯罪；性侵儿童；儿童色情；刑事定罪；跨境取证

一、问题提出：国际社会如何形成打击网络性侵儿童犯罪合力？

网络技术在提升儿童群体¹数字素养的同时也带来了一定风险隐患，在缺乏有效的网络监管手段下，儿童遭受网络侵害风险剧增，尤以侵害儿童性权益风险为甚，儿童色情制品、儿童色情直播²及性犯罪者充斥在网络空间，严重侵害儿童身心健康。在前述背景下，2024年12月联合国大会最终通过的《联合国打击网络犯罪公约》³（简称《公约》）刑事定罪章节专门提及网络性侵儿童犯罪⁴，具体

* 作者信息：赵安晓宇（1995—），男，新疆伊犁人，法学博士，内蒙古大学法学院讲师。研究方向：网络法学、刑事证据法学、未成年人法学，电子邮箱：zaxy0703@163.com。联系电话：18511760783。
基金项目：本文系内蒙古大学青年学术人才科研启动项目“性侵未成年人犯罪案件的侦查取证规范研究”的阶段性研究成果。

1 儿童概念参照《儿童权利公约》的规定，即18岁以下的任何人。本文中除引用内容、专有名词外，“未成年人”与“儿童”即指代同一群体。

2 国际社会普遍认为“儿童色情”“儿童卖淫”等词语会污名化儿童群体，倡议使用儿童性剥削（child sexual exploitation）、儿童性虐待（child sexual abuse）等词语替代。为保持全文统一，除引用内容外，本文暂继续使用“儿童淫秽色情（直播）”“儿童卖淫”等词语。

3 该公约全称为《联合国打击网络犯罪以及为打击利用信息通信技术系统实施的某些犯罪并共享严重犯罪电子证据而加强国际合作公约》（United Nations Convention against Cybercrime; Strengthening International Cooperation for Combating Certain Crimes Committed by Means of Information and Communications Technology Systems and for the Sharing of Evidence in Electronic Form of Serious Crimes）。

4 网络性侵儿童犯罪即指犯罪行为人在网络平台等互联网环境中对儿童实施的非接触性的

涉及《公约》第 14 条“涉及网上儿童性虐待或儿童性剥削材料的犯罪”及第 15 条“为对儿童实施性犯罪而进行教唆或诱骗”。《公约》涉网络性侵儿童犯罪条款的规定在引发社会各界高度关注的同时，也在侧面揭示了世界范围内网络性侵儿童犯罪的严峻趋势。

2024 年联合国专项报告数据显示：去年（作者注：2023 年）全球约 3.02 亿儿童成为了网络性剥削和性虐待的受害者，全球 12.6% 的儿童成为了非自愿交流以及性图像、视频分享和接触的受害者；此外全球有 12.5% 的儿童曾受到网上引诱。⁵欧洲联盟（简称欧盟）数据显示：2023 年欧盟收到了超过 3620 万份疑似网络性侵儿童报告，在 2021-2023 年期间，欧盟引诱（grooming）儿童的报告数量激增了 300%。⁶我国网络性侵儿童犯罪形势同样不容乐观。中国版“N 号房”事件⁷与“未成年人擦边举牌”事件⁸先后曝光，为我国网络儿童保护工作敲响了警钟，犯罪数据也凸显了这种严峻形势。最高人民法院 2024 年上半年数据披露：线上“隔空猥亵”、线上诱骗线下性侵等涉网性侵案件增加，强制猥亵、侮辱罪案件一审收案同比增长 20.07%。⁹ 最高人民检察院最新数据显示：“2024 年前 11 个月，检察机关起诉涉未成年人利用电信网络实施犯罪 9569 人，同比上升 66.97%。隔空猥亵、线上联络线下性侵的案件不在少数”。¹⁰“在未成年人照片分享、交友信息等页面，发布诱导不良交友、引流非法网站等信息”¹¹等隐性侵害儿童性权益行为层出不穷。

“零容忍”打击网络性侵儿童犯罪已成为国际社会共识，但严峻的犯罪形势也给世界各国犯罪治理工作带来了难题：一方面，网络性侵儿童行为在刑事罪名认定方面存在一定困难。传统性侵儿童犯罪在网络化转变过程中，对应的犯罪构成要件及加重情节也发生了相应变化，如何精准认定犯罪行为较为困难。另一方面，

性侵害行为，而在犯罪行为实施的过程中，受侵害一方可能存在着（身体或精神）遭受了实际侵害但自身可能未完全意识到的情形。网络性侵儿童犯罪在我国较多体现为“隔空猥亵”儿童、儿童色情直播等行为。相关概念界定参见赵安晓宇、马忠红：《从一元走向多元：网络性侵儿童犯罪治理的英国经验与中国镜鉴》，载《当代青年研究》2023 年第 6 期，第 52 页。

5 See *Sale, sexual exploitation and sexual abuse of children*, United Nations (Sept.9,2021) https://documents.un.org/symbol-explorer?s=A/79/122&i=A/79/122_1723826295191.

6 See *Combating child sexual abuse online*, European Parliament (Mar.27,2025), [https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/738224/EPRS_BRI\(2022\)738224_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/738224/EPRS_BRI(2022)738224_EN.pdf).

7 2022 年 7 月，网友爆料部分国内网站存在着大量儿童色情视频，同时在外网某即时聊天软件中发现名为“小白菜”的群聊中，群成员每天都在讨论如何“诱捕”未成年少女，并在群聊中实时更新违法犯罪行为。具体可参见《中国版 N 号房曝光 5 万恶魔正在围猎幼女》，载凤凰网 2022 年 7 月 18 日，<https://news.ifeng.com/c/8Hkl40zPOpr>。

8 近期央视新闻记者调查发现，部分社会人员借助举牌的形式，买卖淫秽色情图片和视频，形成了一个由举牌者、代理和买家组成的利益链条。在这个非法利益产业链中，出现了未成年人参与情形。具体可参见《央视新闻独家调查 | 起底隐匿在“举牌”背后的色情暗流！》，载央广网 2025 年 4 月 5 日，https://news.cnr.cn/native/gd/20250405/t20250405_527124365.shtml。

9 数据参见《最高人民法院公布 2024 年上半年司法审判工作主要数据》，载最高人民法院网站 2024 年 7 月 19 日，<https://www.court.gov.cn/zixun/xiangqing/438521.html>。

10 数据参见王春霞：《网络空间未成年人保护需要加强》，载《中国妇女报》2025 年 3 月 4 日，第 4 版。

11 参见《中央网信办启动“清朗·2024 年暑期未成年人网络环境整治”专项行动》，载中国网信网 2024 年 7 月 13 日，<https://www.court.gov.cn/zixun/xiangqing/438521.html>。

网络性侵儿童犯罪在数据取证方面存在一定困难。犯罪网络（暗网）犯罪实施空间的出现及加密软件的使用大大降低了性犯罪者的作案成本，犯罪跨国化、网络化与匿名化趋势也使得刑事证据（特别是电子证据）取证难度陡增。如何形成合力打击网络性侵儿童犯罪，成为了国际社会共同关注的现实议题。于此，本文基于联合国发布《公约》这一现实背景，首先，梳理既有国际公约文件中涉网络性侵儿童犯罪条款并进行总结；其次，对《公约》中网络性侵儿童犯罪条款内容进行解读，并将其与我国现有刑事法条款进行比较对比；再次，基于域内外比较考察，梳理典型国家网络性侵儿童犯罪的刑事法治理实践；最后，立足刑事一体化背景，提出《公约》网络性侵儿童犯罪条款国内刑事法转化的可能建议，以此呼应网络儿童保护这一现实主题。

二、既有国际公约中网络性侵儿童犯罪条款的文本呈现

在联合国《公约》出台之前，部分国际公约中就提及了网络性侵儿童犯罪条款，上述文件也为《公约》制定提供了一定参考。当前提及网络性侵儿童犯罪条款的国际公约主要有两种，一种为专门聚焦儿童权利保障的国际公约，该类公约强调儿童权利保护的重要性，并详细罗列了侵害儿童犯罪行为，其中就包括对网络性侵儿童犯罪行为的规定；另一种是聚焦打击（网络）犯罪议题的国际公约，该类公约通常聚焦网络犯罪问题治理事宜，将网络性侵儿童犯罪列为网络犯罪类型之一。

（一）“儿童权利保障类”国际公约中涉网络性侵儿童犯罪条款

在《公约》出台之前，部分地区与国际组织先后出台了打击（网络）性侵儿童犯罪公约，以期最大限度保障儿童性权益，较为代表性的有联合国、欧盟以及东盟发布的区域性公约及相关配套文件。

1. 联合国相关公约及配套文件

一是《儿童权利公约》。1989年联合国通过的《儿童权利公约》在保护儿童权益方面发挥着重要作用，与网络性侵儿童犯罪相关的条款集中在公约第34条。第34条规定的罪行包括：1. 引诱或强迫儿童从事任何非法的性活动；2. 利用儿童卖淫或从事其他非法的性行为；3. 利用儿童进行淫秽表演和充当淫秽题材。¹²二是《儿童权利公约》配套文件。此后，2000年5月联合国通过《〈联合国儿童权利公约〉关于买卖儿童、儿童卖淫和儿童色情制品问题的任择议定书》（简称《任择议定书》），其中《任择议定书》中与网络性侵儿童犯罪相关的规定集中在第2、3条。第2条界定了“儿童色情制品”概念，即“儿童色情制品系指以任何手段显示儿童进行真实或模拟的露骨性活动或主要为诲淫而显示儿童性器官的制品”；第3条第2款规定儿童色情相关犯罪行为，包括“为了上述目的（作者注：买卖儿

12 参见《儿童权利公约》，载联合国网站，http://un.china-mission.gov.cn/chn/zgylhg/shhrq/liandawanwei1/202412/t20241220_11507892.htm，2025年4月11日访问。

童、儿童卖淫）生产、发售、传播、进口、出口、主动提供、销售或拥有第2条所指的儿童色情制品”。¹³联合国先后发布的《儿童权利公约》及《任择议定书》概括性地规定儿童群体所享有的免遭一切形式侵害的合法权利，这也为后续部分地区与国家针对本国儿童权益保障性法律文件提供了立法文本参考。

2. 欧盟相关国际公约及配套文件

一是《欧洲理事会保护儿童免受性剥削与性虐待公约》。为有效打击侵害儿童犯罪，2007年欧盟发布了《欧洲理事会保护儿童免受性剥削与性虐待公约》（简称《兰萨罗特公约》），其中该公约第20、21、23条涉及网络性侵儿童犯罪条款。第20条儿童色情犯罪（Offences concerning child pornography）行为包括：制作、提供、分发、传送、持有儿童色情制品、为自己或他人购买儿童色情物品等；第21条儿童参与色情表演犯罪（Offences concerning the participation of a child in pornographic performances）涉及多个行为：（1）招募儿童参与色情表演或导致儿童参与此类表演；（2）胁迫儿童参与色情表演或因此从儿童身上获利或以其他方式剥削儿童；（3）明知而观看有儿童参与的色情表演。第23条性引诱儿童犯罪（Solicitation of children for sexual purposes）行为包括：成年人以对儿童实施性虐待或生产儿童色情制品等犯罪行为为目的，借助信息和通讯技术故意邀约未达性承诺年龄的儿童会面，且在“邀约”发出之后还实施了促成双方会面的行为。¹⁴欧盟制定《兰萨罗特公约》时深受《儿童权利公约》及《任择议定书》影响，并在二者基础上，细化了包括（网络）性侵儿童犯罪在内的侵害儿童犯罪行为的具体行为。相较前者，欧盟在立法方面较为“激进”，《兰萨罗特公约》首次将“性引诱”儿童这一犯罪预备行为进行了入罪化处置。

二是《欧洲议会和理事会关于打击儿童性侵犯和性剥削及儿童色情的指令》。2011年欧盟发布的《欧洲议会和理事会关于打击儿童性侵犯和性剥削及儿童色情的指令》（简称《指令》），细化了《兰萨罗特公约》中侵害儿童罪行的刑罚规定，其中与网络性侵儿童犯罪条款集中在《指令》第4-6条。《指令》第4条性剥削犯罪（Offences concerning sexual exploitation）规定：雇佣儿童参与色情表演、胁迫或强迫儿童参加色情表演、雇佣儿童卖淫、胁迫或强迫儿童卖淫将分别被判处2-8年不等刑期。第5条儿童色情犯罪（Offences concerning child pornography）规定：获取或持有儿童色情制品、故意通过信息和通信技术获取儿童色情制品、分发、传播、供应、出售儿童色情制品等行为将分别被判处1-3年刑期。第6条出于性目的引诱儿童犯罪（Solicitation of children for sexual purposes）规定：通过信息和通信技术提议与未达到性同意年龄的儿童会面并发生性关系，最少判处1

13 参见《〈儿童权利公约〉关于买卖儿童、儿童卖淫和儿童色情制品问题的任择议定书》，载联合国网站，<https://www.un.org/zh/documents/treaty/A-RES-54-263-2>，2025年4月11日访问。

14 See Council of Europe Convention on the Protection of Children against Sexual Exploitation and Sexual Abuse, The Council of Europe (Apr.11,2025), <https://rm.coe.int/1680084822>.

年刑期。¹⁵ 此外针对儿童色情表演网络化趋势,《指令》第2条指出“色情表演”的方式不仅仅是发生在实体空间,也包括通过信息和通信技术进行色情表演的行为。

3. 东盟相关国际公约及配套文件

受多重因素影响,东南亚国家一直以来是侵害儿童犯罪的重灾区,为有效打击(网络)性侵儿童犯罪,东南亚国家联盟(简称《东盟》)发布了多个公约文件,规定了相关犯罪条款。一是《东盟保护儿童免受一切形式网上剥削和虐待宣言》(简称《东盟宣言》)。2019年东盟发布的《东盟宣言》强调促进、制定和实施全面的国家法律的框架,并努力改善东盟成员国关于所有形式的在线剥削和虐待的儿童保护标准和政策。¹⁶此后,东盟在《东盟保护儿童免受一切形式网络剥削和虐待区域行动计划》(简称《东盟行动计划》)承诺将下列行为进行入罪化处理:1.持有儿童性虐待制品;2.制作和传播儿童性虐待制品;3.无论是否出于商业目的获取、通过直播观看或参与儿童性剥削和性虐待;4.线上与线下引诱儿童。¹⁷为进一步高效打击网络性侵儿童犯罪,东盟于2023年发布了《东盟关于打击各种形式网络儿童性剥削和性虐待的协调一致且全面的国家立法指南》(简称《东盟立法指南》)进一步强调将下列行为纳入刑事犯罪的范畴:1.拥有、制作、传播儿童性虐待材料以及直播儿童性剥削和性虐待的行为;2.性引诱儿童;3.性勒索(Sexual extortion)儿童。¹⁸

(二) 区域性打击(网络)犯罪公约中涉网络性侵儿童犯罪条款

除了“儿童权利保障性”国际公约详细罗列网络性侵儿童犯罪条款情形外,部分区域性打击(网络)犯罪公约文件中也提及了上述犯罪类型,较有代表性的打击(网络)犯罪公约文件有以下几种。

1. 欧盟《布达佩斯网络犯罪公约》(简称《布达佩斯公约》)。2001年由欧盟成员国牵头制定的《布达佩斯公约》是世界范围内第一部针对网络犯罪而制定的国际公约,其中第9条儿童色情犯罪(offenses related to child pornography)涉及对网络性侵儿童犯罪条款的规定。该条明确指出以下行为构成犯罪:(1)制作儿童色情制品,以便通过计算机系统销售;(2)通过计算机系统提供或提供儿童色

15 See on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA, European Union (Apr.11,2025), <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02011L0093-20111217>.

16 See Declaration on the Protection of Children from all Forms of Online Exploitation and Abuse in ASEAN, ASEAN (Apr.11,2025), <https://asean.org/wp-content/uploads/2021/01/3-Declaration-on-the-Protection-of-Children-from-all-Forms-of-Online-Exploitation-and-Abuse-in-ASEAN.pdf7>.

17 See Regional Plan of Action for the Protection of Children from All Forms of Online Exploitation and Abuse in ASEAN, ASEAN (Apr.11,2025), https://asean.org/wp-content/uploads/2021/11/4.-ASEAN-RPA-on-COEA_Final.pdf.

18 See ASEAN Guidelines for Harmonised and Comprehensive National Legislation Against All Forms of Online Child Sexual Exploitation and Abuse, ASEAN (Apr.11,2025), <https://acwc.asean.org/wp-content/uploads/2024/06/ASEAN-Guidelines-for-Harmonised-and-Comprehensive-National-Legislation-Against-All-Forms-of-Online-Child-Sexual-Exploitation-and-Abuse.pdf>.

情制品；（3）通过计算机系统销售或传送儿童色情制品；（4）透过电脑系统为自己或他人获取儿童色情物品；（5）在电脑系统或电脑资料储存媒介内藏有儿童色情物品。¹⁹

2.英联邦国家《关于计算机和计算机犯罪示范法》（简称《示范法》）。2002年在《布达佩斯公约》基础上，英联邦国家起草了《示范法》。《示范法》规定了不同类型的网络犯罪类型，其中与网络性侵儿童犯罪的规定集中在第10条儿童色情（Child pornography）条款之中，该条规定将下列行为列为犯罪：1.通过计算机系统出版儿童色情制品；2.为通过计算机系统出版而制作儿童色情制品；3.在计算机系统或计算机数据存储介质中拥有儿童色情制品。²⁰

3.阿拉伯国家联盟《打击信息技术犯罪公约》。2010年阿拉伯联盟通过的《打击信息技术犯罪公约》同样提及了网络性侵儿童犯罪条款。阿拉伯联盟《打击信息技术犯罪公约》第12节色情犯罪（Offence of Pornography）将以下行为纳入了刑事规制范畴：（1）通过信息技术制作、展示、分发、提供、出版、购买、销售、进口色情材料或构成猥亵行为的材料。（2）对与儿童和未成年人色情制品有关的犯罪行为应加重处罚。（3）本条第2款提到的增加包括通过信息技术或这种技术的储存媒介获得儿童和未成年人色情材料或构成侮辱性的儿童和未成年人材料。但与其他地区打击网络犯罪公约规定不同的是，阿拉伯联盟《打击信息技术犯罪公约》并未区分成年人与未成年人（儿童）的区别，直接规定了色情犯罪，并将儿童色情犯罪当做了色情犯罪的加重情节。²¹

（三）小结

既有国际公约中对网络性侵儿童犯罪条款的设置呈现出以下几个特点。

1.儿童利益最大化与“零容忍”打击（网络）性侵儿童犯罪

无论是“儿童权利保障性”国际公约还是区域性打击网络犯罪公约，上述两种国际公约都高度聚焦儿童性权益保障工作，“零容忍”打击一切侵害儿童犯罪行为，这也是贯彻《儿童权利公约》儿童利益最大化的体现。《儿童权利公约》第3条第1款明文规定：关于儿童的一切行动，不论是由公私社会福利机构、法院、行政当局或立法机构执行，均应以儿童的最大利益为一种首要考虑。²²虽然相应的

19 See *Convention on Cybercrime*, Council of Europe (Apr.11,2025), https://www.europarl.europa.eu/meetdocs/2014_2019/documents/libe/dv/7_conv_budapest/7_conv_budapest_en.pdf.

20 See *Model Law on Computer and Computer Related Crime*, The Commonwealth (Apr.11,2025), https://production-new-commonwealth-files.s3.eu-west-2.amazonaws.com/migrated/key_reform_pdfs/P15370_11_ROL_Model_Law_Computer_Related_Crime.pdf.

21 See *Arab Convention on Combating Information Technology Offences*, Asian School of Cyber Laws (Apr.11,2025), <https://www.asianlaws.org/gclid/cyberlawdb/GCC/Arab%20Convention%20on%20Combating%20Information%20Technology%20Offences.pdf>.

22 参见《儿童权利公约》，载联合国网站，http://un.china-mission.gov.cn/chn/zgylhg/shhrq/liandawanwei1/202412/t20241220_11507892.htm，2025年4月11日访问。

公约及其解释性文件在国际法层面并不具有约束性，但是其对儿童利益最大化原则的细致解释，也充分体现出了儿童利益最大化原则在刑事司法领域的重要性，这一理念也深刻影响着后续儿童权利保障类国际公约的制定。如《兰萨罗特公约》开篇直接指出：“考虑到儿童的福祉和最大利益是所有成员国共有的基本价值观，必须不加歧视地加以促进”。²³此后最直接的体现为，各国应采取一切必要手段“零容忍”打击任何形式的侵害儿童行为。比如，在打击儿童色情制品类犯罪时上述公约文件均将持有、传播、下载、浏览等行为纳入到了刑事规制范畴。

2.网络性侵儿童犯罪规制思路“宽严有别”

受区域文化及历史传统等因素影响，不同地区与国家在规制网络性侵儿童犯罪方面的思路呈现出“宽严有别”的态势。²⁴欧美等国家针对色情类犯罪的基本规制思路为区分“成人”与“儿童（未成年人）”，并严格规制儿童色情类犯罪，对所谓的“成人色情”不加以严格限制，上述联合国及欧盟文件也充分体现了这一点。但是聚焦到阿拉伯联盟的《打击信息技术犯罪公约》，该公约对色情类犯罪呈现出严格规制的路径，其并未进行“成人—儿童”的二元划分，将所有色情行为纳入到了刑事规制范畴，并将儿童色情相关行为纳入到了色情犯罪的加重情节。阿拉伯联盟相关刑事规制思路与前者相比，显然更为严格。

3.网络性侵儿童犯罪类型划分“由浅及深”

上述国际公约在划分网络性侵儿童犯罪类型时也呈现出“由浅到深”的特点。联合国1989年《儿童权利公约》及2000年《任择议定书》制定时的现实背景是包括性侵在内的侵害儿童犯罪主要发生在实体空间之中，侵害儿童犯罪网络化趋势还不明显，因此上述公约文件中与网络性侵儿童犯罪相关的条款主要集中在儿童色情制品类犯罪方面。随着网络技术的迅速发展，网络性侵儿童行为愈发明显，儿童色情制品逐步在网络空间中传播，儿童色情直播逐步蔓延。在此背景下后续的《兰萨罗特公约》《布达佩斯公约》等文件立即将上述行为纳入到了刑事规制范畴。此外，考虑到儿童性引诱行为愈发具有现实危害性，不分国际公约文件将这一犯罪预备行为独立成为了新的罪名，典型如欧盟《兰萨罗特公约》及《指令》，上述两个文件明确提出将“性引诱”儿童行为纳入刑事规制范畴。

三、《公约》中网络性侵儿童犯罪条款及其与我国刑事法比较考察

在《公约》最终出台之前，特委会召开了六次会议，形成了多个《公约》草案，其中都涉及网络性侵儿童犯罪条款的制度设计。本章聚焦系列《公约》草案

23 See Council of Europe Convention on the Protection of Children against Sexual Exploitation and Sexual Abuse, The Council of Europe (Apr.11,2025), <https://rm.coe.int/1680084822>.

24 不同国家对“儿童色情制类”行为的规制态度各有不同。日本主张对儿童色情制品的认定应当限于真人情形，应当排除漫画等虚拟形象；欧洲国家主张对儿童自制色情情形作出法律层面的保留，保留部分出罪规定，阿拉伯国家则在主张《公约》儿童色情条款同时，主张将成人色情行为进行入罪考虑。关于部分国家在“儿童色情”犯罪方面主张的观点可参见吴沈括：《〈联合国打击网络犯罪公约〉的治理体系与中国因应》，载《中国法律评论》2025年第1期，第219页。

及最终正式版《公约文本》，梳理《公约》中网络性侵儿童犯罪条款设置、变更；并在此基础上，与我国现有刑事法条文进行对比分析。

（一）《公约》中网络性侵儿童犯罪条款的内容解读

实际上在《公约》制定过程中，我国就多次关注网络性侵儿童犯罪条款事宜。2022年3月《中国对联合国打击为犯罪目的使用信通技术公约范围、目标和框架（要素）的建议》中就提及：“对于国际社会有普遍共识的、利用信通技术实施的犯罪行为，公约可以酌情列举，包括：网络勒索、网络诈骗、网络色情特别是儿童色情、利用信通技术侵犯著作权和邻接权、利用网络煽动、实施恐怖主义、传播有害信息等。”²⁵此外中国代表团杨帆在联大三委与贩卖儿童及儿童性剥削问题特别报告员互动对话时进一步强调：“提请各方注意联合国会员国于今年8月谈判达成的《联合国打击网络犯罪公约》，认为其中有关条款，将为打击通过信息通信技术实施的儿童性剥削等行为提供更有效国际合作渠道。中方期待各方携手推动公约尽快生效，共同努力强化合作打击犯罪，在网络空间更有效保护儿童权利”。²⁶

1. 《公约》草案中网络性侵儿童犯罪条款的文本解读

早在2010年，联合国大会第65/230决议中就曾提请预防犯罪与刑事司法委员会设立政府间专家组，专门研究国际层面打击网络犯罪法律应对事宜。此后至2019年12月27日，联合国大会通过了第74/247号决议，“决定设立一个代表所有区域的不限成员名额特设政府间专家委员会，以拟订一项关于打击为犯罪目的使用信息和通信技术行为的全面国际公约”²⁷，这也标志着联合国层面的打击网络犯罪公约制定工作正式开启。此后，进行了六次会议谈论，终于在2023年8月29日至9月1日举行的第六次委员会会议上，特设政府间专家委员会（简称特委会）根据之前会议的讨论提出了一份公约草案。之后，根据相关各方的意见与建议，特委会多次修订，在2024年8月8日通过了《公约》文本。《公约》在正式出台之前后经历了六次会议讨论，《公约》（草案）也陆续进行了修改完成了“三读”。其中与网络性侵儿童犯罪相关的条款进行了大幅变动。

2023年1月9日至20日特委会召开了第四次会议，期间形成的会议文件（A/AC.291/16）整合了之前会员国提案的要点，形成了《公约》的一个初步草案。其中与网络性侵儿童犯罪相关的条款如下所示。会议文件（A/AC.291/16）第

25 参见《中国对联合国打击为犯罪目的使用信通技术公约范围、目标和框架（要素）的建议》，载联合国毒品和犯罪问题办公室网站，https://www.unodc.org/documents/Cybercrime/AdHocCommittee/First_session/Comments/Chinas_Suggestions_on_the_Scope_Objectives_and_Structure_AHC.pdf，2025年4月11日访问。

26 参见《中国代表团杨帆在联大三委与贩卖儿童及儿童性剥削问题特别报告员互动对话时的发言》，载中华人民共和国常驻联合国代表团网站2024年10月11日，http://un.china-mission.gov.cn/chn/zgylhg/shhrq/liandawanwei1/202412/t20241220_11507892.htm。

27 参见《打击为犯罪目的使用信息和通信技术行为》，载联合国网站，<https://docs.un.org/zh/A/Res/74/247>，2025年4月11日访问。

18 条为与网上儿童性虐待或性剥削材料有关的犯罪；第 19 条为通过[计算机系统][信息和通信技术系统/装置]为儿童性虐待或性剥削材料提供便利；第 20 条为通过[计算机系统][信息和通信技术系统/装置]为性目的而诱骗或诱拐儿童；第 21 条为网上跟踪儿童。²⁸此后，2023 年 8 月 21 日至 9 月 1 日特委会召开了第六次会议，其中会议文件（A/AC.291/22）对“刑事定罪”部分进行了重大删改，较之会议文件（A/AC.291/16）变化较大，一是体现在罪名删减方面，二是体现在具体罪名构成要件方面，具体如表 1 所示。

表 1 会议文件（A/AC.291/16）与会议文件（A/AC.291/22）对比表

会议文件 (A/AC.291/16)	会议文件（A/AC.291/22）	文件变动对比
第 18 条：与网上儿童性虐待或性剥削材料有关的犯罪	第 13 条：与网上儿童性虐待或儿童性剥削材料有关的犯罪	1.修改第 1 款行为类型； 2.删除第 2 款中“貌似儿童的人从事真实或模拟的露骨性行为”的表述；3.删除第 3 款、第 5 款
第 19 条：（通过[计算机系统][信息和通信技术系统/装置]为儿童性虐待或性剥削材料提供便利）		罪名删除
第 20 条：通过[计算机系统][信息和通信技术系统/装置]为性目的而诱骗或诱拐儿童	第 14 条：通过[计算机系统][信息和通信技术装置]为性目的教唆儿童	1.罪名中的“诱骗或诱拐”改为“教唆”；2.删除第 2 款、第 3 款
第 21 条.网上跟踪儿童		罪名删除

在特委会第六次会议之后，《公约》草案又进行了三次修改²⁹，其中与网络性侵儿童犯罪相关的条款也进行了相应的调整，对罪名的调整也能在一定程度上反应国家间在认定犯罪行为与标准时存在的博弈，相应变动参见表 2。

28 参见《关于打击为犯罪目的使用信息和通信技术全面国际公约的总则、刑事定罪条款及程序措施和执法条款的合并谈判文件》，载联合国毒品和犯罪问题办公室网站，<https://docs.un.org/zh/A/AC.291/16>，2025 年 4 月 11 日访问。
29 三份修改草案文件号分别为 A/AC.291/22/Rev.1、A/AC.291/22/Rev.2、A/AC.291/22/Rev.3。

表 2 第六次特委会后三次《公约》（草案）修改内容对比

序号/文件名	A/AC.291/22	A/AC.291/22/Rev.1	A/AC.291/22/Rev.2	A/AC.291/22/Rev.3
1	第 13 条：与网上儿童性虐待或儿童性剥削材料有关的犯罪	第 1 款（a）项中的“邀约提供”改为“主动提供”；删除第 1 款（b）项的底条款、（d）项的“资助”、第 2 款（b）项；修改第 5 款为第 2 款的适用添加限制条件	增设第 1 款（d）项，明确缔约国可将“资助”该条犯罪	将第 2 款的对象限定为“未满 18 周岁的儿童”
2	第 14 条：为对儿童实施性犯罪而教唆或诱骗	增设第 3 款、第 4 款		

（二）最终版《公约》中网络性侵儿童犯罪条款的文本解读

经过多次修改，2024 年 12 月联合国大会正式通过《公约》。在最终的《公约》中与网络性侵儿童犯罪相关的条款被调整为第 14 条涉及网上儿童性虐待或儿童性剥削材料的犯罪及第 15 条为对儿童实施性犯罪而进行教唆或诱骗。第 14 条犯罪条款将以下行为纳入到了刑事规制范畴：1.通过信息通信技术系统制作、主动提供、出售、分销、传送、广播、展示、发布或以其他方式提供儿童性虐待或儿童性剥削材料；2.通过信息通信技术系统索取、获取或访问儿童性虐待或儿童性剥削材料；3.持有或控制存储在信息通信技术系统或其他存储介质中的儿童性虐待或儿童性剥削材料。第 15 条犯罪条款则将以下行为纳入到了刑事规制范畴：为实施本国法律所界定的对儿童的性犯罪，包括为实施根据本公约第十四条确立的任何犯罪，通过信息通信技术系统进行通信联络、教唆、诱骗或做出任何安排。³⁰无论是与第四次特委会文件（A/AC.291/16）、第六次特委会文件（A/AC.291/22）还是之后的三次修改文件（A/AC.291/22/Rev.1、A/AC.291/22/Rev.2、A/AC.291/22/Rev.3），最终正式版《公约》网络性侵儿童犯罪条款明显减少，且具体到相关条款之中，犯罪构成要件也有所调整。

30 参见《打击为犯罪目的使用信息和通信技术行为》，载联合国网站，<https://docs.un.org/zh/A79/460>，2025 年 4 月 11 日访问。

此外由于网络犯罪跨地域性特征较为明显,跨境取证需求日益增加,因此《公约》中也涉及对网络犯罪的跨境取证条款规定。较为明显,打击网络性侵儿童犯罪不仅需要再刑事实体法层面的努力,也需刑事程序法层面高效固定证据。《公约》中涉及网络性侵儿童犯罪的取证条款主要有以下几种。《公约》第 25 条快速保全存储的电子数据、第 26 条快速保全和部分披露流量数据、第 27 条提交令、第 28 条搜查和扣押存储的电子数据、第 29 条实时收集流量数据、第 30 条拦截内容数据等。

(二)《公约》中网络性侵儿童犯罪条款与我国刑事法比较

1.我国刑事法已涵盖内容

就刑事定罪内容而言,《公约》第 14 条与第 15 条的网络性侵儿童犯罪条款实际上大多都能在《刑法》中一一对应。《公约》第 14 条涉及网上儿童性虐待或儿童性剥削材料的犯罪中第 1 款“通过信息和通信技术系统制作、主动提供、出售、分销、传送、广播、展示、出版或以其他方式提供儿童性虐待或儿童性剥削材料”的行为可以完全被《刑法》制作、贩卖、传播淫秽物品罪所吸收,具体包括 363 条的制作、复制、出版、贩卖、传播淫秽物品牟利罪;以及 364 条传播淫秽物品罪。针对《公约》第 15 条“为对儿童实施性犯罪而进行教唆或诱骗”,有学者认为可以通过解释方法与国内法衔接,教唆是典型的共同犯罪,通过相关犯罪的共犯进行处理即可。³¹

此外针对网络犯罪的侦查取证规定,我国已建立起了一定的法律规范体系,如《刑事诉讼法》与《国际刑事司法协助法》,此外《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》(法发〔2016〕22 号)、《关于办理信息网络犯罪案件适用刑事诉讼程序若干问题的意见》(法发〔2022〕23 号)以及公安部《公安机关办理刑事案件电子数据取证规则》也同样规定了网络犯罪的侦查取证规则要求,其中部分法律性文件的规定与《条约》有所交叉,但同时差异性也较大,需要进行国内法的转化适用。

2.我国刑事法未涵盖内容

一是“持有”儿童色情制品行为。《公约》第 14 条第 1 款第 2 项规定了通过信息和通信技术系统征求、获取或访问儿童性虐待或儿童性剥削材料的行为以及第 3 项规定的拥有或控制存储在信息和通信技术系统或其他存储介质中的儿童性虐待或儿童性剥削材料的行为,简言之,《条约》将持有、获取、访问儿童色情制品行为纳入了刑事规制范畴。但是就实际而言,当前我国现有法律文件中并未将“持有”“获取”“访问”儿童色情制品行为进行入罪化处理。

二是性引诱儿童情节的规制。《条约》第 15 条为对儿童实施性犯罪而进行教唆或诱骗条款规定,为实施本国法律所界定的对儿童的性犯罪,包括为实施根据

31 观点参见李本灿:《联合国打击网络犯罪公约》核心议题述评,载《法学论坛》2025 年第 1 期:第 96 页。

本公约第十四条确立的任何犯罪，通过信息通信技术系统进行通信联络、教唆、诱骗或做出任何安排实际上是想将“性引诱”儿童这一犯罪预备行为实行化处理，但目前，我国法律性文件中并未直接规定类似的条款。

四、域外国家治理网络性侵儿童犯罪的刑事法实践与探索

为有效打击网络性侵儿童犯罪，部分国家在将上述国际法义务转化为国内法律条文的同时针对本国犯罪现状，开展了一系列的刑事法治理实践。其中英、美、澳等国在这方面做法较为丰富，治理成效较好。因此本节主要聚焦英美澳等国的网络性侵儿童犯罪治理实践，从刑事实体法与程序法视角出发，总结归纳上述国家的刑事法实践。

（一）刑事实体法层面

1.对“持有”儿童色情制品行为的延伸规制

基于本国国情，英、美、澳三国先后完善了网络性侵儿童犯罪的刑事规制力度，加大力度规制涉儿童色情相关行为。近年来英国通过立法的形式先后将制作、持有、传播、复制儿童色情制品相关行为纳入刑事规制范畴。1988年《刑事司法法》、1978年《儿童保护法》、2015年《严重犯罪法》、2009年《验尸官与司法法》与2008年《刑事司法和移民法》均涉及到对“持有”儿童色情制品入罪的规定。³²此外，美国“持有”儿童色情制品罪名集中在《美国法典》第18编第110章2252A条的儿童色情制品罪，该条款明确规定持有儿童色情制品行为将被纳入刑事规制范畴。与英、美等国类似，澳大利亚同样将“持有”儿童色情制品行为入罪，相应法律文件包括1914年《澳大利亚犯罪法》（Crimes Act 1914）、1995年《澳大利亚联邦刑法典》（Criminal Code Act 1995）。此后为进一步保护儿童权益，2019年澳大利亚发布的《打击儿童性剥削立法修正案》扩大了儿童色情制品的包含范围，将儿童模样的性爱用品（child sex doll）纳入了儿童色情制品范畴，持有、走私相关物品被定为犯罪行为，最高可判处15年有期徒刑。将“持有”行为纳入刑事规制范畴后，英、美、澳等国侦查部门便逐步将（电子）搜查手段应用到了犯罪侦查之中，查缉犯罪者住所及电子产品中是否藏有儿童色情制品。

2.对“性引诱”儿童行为的规制

将“性引诱”儿童这一犯罪预备情节独立成为新的罪名并纳入刑事规制范畴，才能为侦查部门及时打击网络性侵儿童犯罪提供法律依据。英美澳等国先后将“性引诱”儿童行为独立成为了新的罪名。英国1988年《恶意通信法》规定发送包括不雅、严重冒犯或含有威胁的信息属于犯罪行为；2003年《通讯法》进一步

32 2025年初，英国更是突破性的制定了《犯罪和警务法案（草案）》（Crime and Policing Bill），该《草案》将严格打击利用人工智能（AI）工具生成儿童色情制品行为，持有、创建或分发用于生成儿童色情图片的AI工具，违者最高将被处以五年的刑期。与此同时，英国边境执法部门将获得新的侦查执法权限，对出入境英国的可疑人员电子设备开展搜查工作，以寻找可能存在的儿童色情制品。See *Crime and Policing Bill*, UK Parliament (Mar.27,2025), <https://publications.parliament.uk/pa/bills/cbill/59-01/0187/240187.pdf>.

规定通过公共电子通信网络发送具有严重冒犯性或猥亵、淫秽或威胁性的信息或其他事项属于犯罪行为；同年的《性犯罪法》第 16 条进一步明确将包括与儿童（16 周岁以下）进行性对话、诱使儿童见面发生性关系等行为定为犯罪。美国联邦层面的“性引诱”条款主要集中在《美国法典》第 18 编第 110 章 2251 条的儿童性剥削罪之中；除此之外美国各州也纷纷立法，将“性引诱”行为纳入到了刑事规制范畴。如 2023 年 9 月，得克萨斯州通过了《儿童引诱法案》，根据规定引诱 18 岁以下儿童的个人可能面临最高 10 年有期徒刑。而澳大利亚也同样规定了“性引诱”儿童规定，《澳大利亚联邦法典》修正案第 474.25C 条款（也称《卡莉法案》，Carly's Law）规定，成年人利用准备或计划对儿童造成伤害或进行性活动的行为将构成犯罪。

（二）刑事程序法层面

1. 优化跨境刑事数据调取规则

网络性侵儿童犯罪的跨地域性就决定了存在大量涉案电子数据存储于涉案第三国网络平台当中，这就涉及到跨境刑事数据调取或者跨境电子取证问题。但是传统的刑事司法协助机制程序较为繁琐、耗时较长，“侦查机关不仅难以确定与某一刑事案件相关的电子数据具体存储于何处，即使待其可得确定之时，相关数据亦可能在跨境取证请求发出之前便又被移转，以致传统跨境取证模式作用落空”³³。2018 年美国通过的《澄清合法使用海外数据法案》（Clarifying Lawful Overseas Use of Data Act，简称《云法案》）授权美国与合作的他国通过签订双边行政协议的方式，赋予执法机关通过简便的方式获得所需的跨国数据的能力。在此背景下，部分国家基于特定犯罪类型，签订了跨境刑事数据调取协议。一是英美两国于 2019 年 10 月制定的《数据访问协议》，该协议赋予美国和英国执法机构在适当授权情况下，强制任何一个国家的科技公司提供与恐怖主义、严重暴力犯罪、儿童剥削、跨国犯罪及金融诈骗等犯罪类型相关的电子数据（The United States Department of Justice, 2019）。这一过程无须通过传统的刑事司法协作途径，极大地减少了侦查取证时间。与之类似，2021 年 12 月，澳大利亚与美国在基于《云法案》的基础上通过了跨境执法的双边协议。该协议允许澳美双方在打击网络性侵儿童、恐怖主义等犯罪类型时无需通过繁琐的刑事司法协助制度，可以直接访问协议国通信供应商在本国的涉案数据。³⁴《云方案》及美澳、美英协议的实施，在一定程度上解决了跨境侦查管辖的困境，提升了电子取证效率。

2. 拓宽强制性侦查取证措施的网络化适用场景

33 孙永超：《论美国跨境刑事数据调取中的国际礼让：以〈云法案〉为例的分析》，载《求是学刊》2024 第 3 期，第 126-127 页。

34 继美英、美澳跨境协议外，当前美国也在推进与加拿大基于《云方案》的跨境执法协议。See *United States and Canada Welcome Negotiations of a CLOUD Act Agreement*, United States Department of Justice (Mar.22.2022), <https://www.justice.gov/archives/opa/pr/united-states-and-canada-welcome-negotiations-cloud-act-agreement>.

英美澳国家对强制性侦查措施或者所谓的秘密侦查或技术侦查措施使用有着天然的保守心态。但近年来为有效应对（网络）性侵儿童、恐怖主义等犯罪，在法律授权基础上，英国警方逐步将通信截取与留存、卧底侦查等技术侦查措施应用于上述犯罪案件侦查取证活动当中。相关侦查措施实施依据主要来源于2000年《侦查力规制法》（Regulation of Investigatory Powers Act 2000）、2016年《侦查权力法》（Investigatory Powers Act 2016）、2021年《秘密人工情报源（犯罪行为）法》[Covert Human Intelligence sources (criminal conduct) Act]等。以2021年《秘密人工情报源（犯罪行为）法》[Covert Human Intelligence sources (criminal conduct) Act]为例，该法案细化了包括卧底侦查措施（包括网络卧底侦查）使用规定，甚至规定了卧底人员因打击犯罪需要实施犯罪行为的免责规定。在英国内政部发布的《秘密人工情报源（犯罪行为）法操作案例》，Covert Human Intelligence Sources (Criminal Conduct) Bill Operational Case Studies 中也详细的记载了英国警方使用卧底警务人员潜入儿童色情网站寻找恋童癖者及假扮未成年女孩“诱捕”性犯罪者的成功案例。³⁵

澳大利亚同样将强制性侦查取证措施应用到网络性侵儿童犯罪当中。2021年《监控立法修正案（识别与破坏）》[Surveillance Legislation Amendment (Identify and Disrupt) Act 2021]的通过更是扩大了执法部门在打击包括性侵儿童犯罪在内的侦查权力。该法案赋予 AFP 与 ACIC 在调查“严重犯罪”时行使数据干扰（Data disruption）、网络活动许可（Network activity warrants）与账户接管许可（Account takeover warrants）等三个新的权力，AFP 与 ACIC 必要时可以犯罪调查为由修改、添加、复制或删除数据、从调查对象使用的设备或网络中收集情报以及控制调查对象的网络账户。这为澳大利亚警方跨国打击性犯罪者、查缉儿童色情制品、监控儿童性剥削直播提供了便利。

此外，英美澳等国在使用网络卧底、信息通讯截取等强制性侦查手段的同时，逐步使用“特情引诱”方式来逮捕性犯罪者，具体表现为警务人员伪装成儿童身份在网络空间中与性犯罪者沟通，对其“性引诱”行为进行及时的侦查取证，并在性犯罪者试图线下见面时加以逮捕。当前美国 FBI、DHS 等侦查部门在社交软件等网络环境中部署了大量的卧底警务人员，伪装成儿童身份寻找性犯罪者，并对企图“性引诱”儿童并在线下试图发生关系的性犯罪者进行逮捕。³⁶

35 See *Covert Human Intelligence Sources(Criminal Conduct)Bill Operational Case Studies*, Home Office (Sept.24, 2020),

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923238/CHIS__CC__Bill_-_Case_Studies.pdf.

36 See *Florida Man Sentenced to a Decade for Attempting to Meet a Minor; Engage in Sexual Activity*, U.S. Immigration and Customs Enforcement (Dec.12, 2024), <https://www.ice.gov/news/releases/florida-man-sentenced-decade-attempting-meet-minor-engage-sexual-activity>.

五、《公约》网络性侵儿童犯罪条款的国内刑事法因应

当前我国刑事法中网络性侵儿童犯罪条款与《公约》相比,还存在一定差距。从将国际公约转化为国内法义务的视角来说,对当前我国刑事实体与程序法中未涵盖的部分,我国需要在《刑法》《刑事诉讼法》《未成年人保护法》《未成年人网络保护条例》等法律性文件中做好衔接准备。

(一) 刑事实体法因应: 延伸规制网络侵害儿童性权益行为

1. 对“性引诱”行为的精细化规制

我国可以在谨慎论证基础之上,参照《公约》规定,将刑事规制视野延伸至性侵儿童犯罪的预备阶段,将性侵儿童的预备行为——“性引诱”情节单独成立罪名,从而能为国内执法与司法活动提供法律依据,更加及时干预犯罪行为。首先,需要精准界定“性引诱”行为的内涵和外延。性侵儿童犯罪预备行为涉及到犯罪者利用网络途径与儿童被害人建立起与沟通及所谓的“信任”,并通过强迫、诱骗等方式,诱使受害儿童对“性行为”脱敏及日常化。相关行为包括但不限于发送私密图像、音视频、线下见面发生性侵行为等。在具备立法可能性情况下,可以在《刑法》侵害公民人身权利罪章节设立专门条款,特别是强调规制包括通过网络通讯在内的任何形式性引诱儿童的犯罪行为。其次,需要明确“性引诱”行为的加重情节。性引诱行为的加重情节,可以包括但不限于:对儿童造成身心伤害;线上联系并在线下实施性侵行为;在网络空间中使用金钱等手段诱骗儿童进行色情表演并进行传播、牟利行为;通过使用欺骗方式或成年人滥用职权,如家长、教师等身份进行性引诱。

2. 对“持有”儿童色情内容的全链条打击

刑事立法层面,需要细化儿童色情额制品类犯罪的刑事规制细则。在儿童色情行为认定方面,可以尝试确立相关犯罪评判标准,明确“持有”儿童色情制品犯罪构成情节,与《刑法》《治安管理处罚法》等法律文件形成有效的衔接,并根据犯罪行为具体情节,设置相应的刑罚幅度。³⁷如有学者建议:“可以在《刑法》第 285 条中增加一款,设立持有侵入、非法控制计算机信息系统程序、工具罪;在《刑法》第 364 条增加一款,规定持有儿童性虐待或者儿童性虐待材料罪”³⁸。同时需要细化《未成年人网络保护条例》相应条款,如增设“网络强制报告”条款、细化儿童色情制品行为规制细则条款、儿童色情制品分级分类处置条款等。基于刑事一体化背景,这一刑事立法的整体性调整,也有利于侦查部门调整侦查打击重点。

37 有学者认为持有儿童色情制品行为会助长制作儿童色情物品的行为,而后一行为必然会侵害儿童身心健康,所以,将来增设持有儿童色情物品罪是可以接受的。观点可参见张明楷:《持有犯的基本问题》,载《清华法学》2023 年第 1 期,第 19 页。

38 参见周振杰:《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》,载《中国刑事法杂志》2024 年第 5 期,第 118 页。

（二）刑事程序法因应：提升网络性侵儿童犯罪侦查取证效能

1. 规范电子数据侦查取证指引规则

未来在制定侦查取证规范的时候需要参考《公约》中相关规定，细化刑事证据类型的具体处置，专章设置电子数据取证条款，包括但不限于：选择合适的技术工具和方法对网络数据进行提取和分析，包括网络监控工具、数据挖掘技术以及远程取证技术。同时建立常用社交媒体、聊天应用等平台的取证技术指南。制定与现代电子技术发展相适应的取证技术标准，确保电子证据的完整性、真实性和可靠性。这包括网络数据的提取、存储、分析、转换及打印出证等链条上的各个环节。同时，树立严格的证据链管理规范，确保所有数字证据在采集、传输、存储和分析过程中的完整性和连续性，并确保证据来源和变更可追溯。公约第 42 条至第 44 条规定了电子证据跨境合作的基本措施，包括快速保存、快速披露流量数据、访问电子数据等。³⁹考虑到网络性侵儿童犯罪的跨地域性特点，为及时解决跨境取证难题及可能存在的风险，我国可在现有刑事司法协作机制及《公约》基础上，探索便捷的跨境刑事取证方式。

2. 网络侦查取证措施的尝试与规范

对侦查措施网络适用场景的丰富与规范。就网络犯罪侦查和诉讼而言，网络警务人员可以借助大数据技术，更加灵活的根据个案的实际情况和复杂需求，采取更差异化的策略。⁴⁰通过丰富与规范技术侦查措施的适用场景，一定程度上可以满足侦查部门对电子数据的实际需求。域外经验来看，网络监控、通讯截取、留存、网络卧底侦查等手段在打击网络性侵儿童犯罪中发挥了较大作用。当前我国对相关侦查手段使用规定集中在《刑事诉讼法》第 150 条。虽然该条款使用了“其他严重危害社会的犯罪案件”的概况条款，但就实际而言，几乎没有将上述技术侦查措施应用到网络性侵儿童犯罪打击活动当中。

未来，我国可探索将技术侦查手段用于网络性侵儿童犯罪案件侦破当中。⁴¹如对有组织犯罪集团开展网络监控，截取通讯信息，摸排犯罪黑灰产业链；物建人力情报力量，对侵害儿童犯罪团伙开展卧底侦查活动。此外，侦查部门在行使相关权力的同时遵循必要性、规范性、比例性原则。此外需要强化侦查权行使的监督机制。侦查权的行使固然会对公民的隐私权产生损害，但是信息网络犯罪所依托的载体与传统犯罪有较大区别，技术侦查的运用对公民隐私权的侵犯程度更

39 商震等：《〈联合国打击网络犯罪公约〉若干重点问题评析》，载《中国应用法学》2025 年第 1 期，第 194 页。

40 王玉薇：《网络犯罪治理的法治化路径》，北京大学出版社 2020 年版，第 190 页。

41 有学者认为当前《刑事诉讼法》修改过程中规定和完善电子数据侦查取证措施具有一定的必要性，一方面需要对现有搜查扣押等传统侦查措施予以适当改进与完善，明确将电子数据纳入搜查扣押的适用范围；另一方面，针对电子数据所承载基本权利或者法益的差异，设置不同类型的电子数据侦查取证措施。观点参见谢登科：《电子数据侦查取证措施法治化与〈刑事诉讼法〉再修改》，载《法治研究》2024 年第 5 期，第 100-102 页。

大。⁴²在证据获取方面中，技术侦查措施的合法行使与否决定了案件证据价值，不规范的侦查取证行为会导致证据瑕疵甚至非法证据排除。在程序法定原则之下，公安机关在侦查网络犯罪中能够采取或者适用的侦查取证措施相对比较封闭、固定，特别是干预或者侵犯公民基本权利的强制性侦查措施，需要遵循法律保留原则，需要有法律的明确规定和授权。⁴³在行使技术侦查措施的同时遵循必要性与规范性原则，完善技术侦查措施审批与监督机制，加大事后监督的惩处力度。

Provisions on Online Child Sexual Abuse in the UN Convention on Cybercrime and China's Criminal Legal Response

Zhao Anxiaoyu

Abstract: The provisions on online child sexual exploitation crimes under the United Nations Convention on Cybercrime provide a legal framework at the international level for global cooperation in combating such offenses. Existing international conventions addressing these crimes can be broadly categorized into two types: (1) child rights protection treaties and (2) regional cybercrime conventions, each with distinct emphases on criminal definitions and contextual elements. The formulation of these provisions in the UN Convention on Cybercrime underwent multiple revisions, ultimately prioritizing the regulation of child sexual abuse material (CSAM) and online grooming or solicitation of children. While China's current criminal legal system partially covers these offenses, gaps remain in addressing possession of CSAM, sexual enticement of children, and (cross-border) investigative evidence collection. Comparative analysis reveals that several jurisdictions have effectively transposed international treaty obligations into domestic legislation, achieving notable progress in criminal law enforcement against online child sexual exploitation. Within an integrated criminal justice framework, China may enhance its legal response through substantive and procedural reforms, including but not limited to: (1) criminalizing online grooming behaviors, (2) penalizing the possession of CSAM, (3) refining cross-border evidence-gathering mechanisms, and (4) developing specialized digital forensics protocols.

Keywords: Cybercrime; Child Sexual Exploitation; Child Sexual Abuse Material; Criminalization; Cross-border Evidence Collection

42 洪刚：《信息网络犯罪技术侦查的法律控制》，载《中国应用法学》2022年第1期，第141页。

43 谢登科：《网络犯罪治理的两种模式：规则型治理与技术型治理》，载《云南社会科学》2024年第5期，第54页。

欧盟人工智能法识别网络恐怖主义信息的实践及对中国的镜鉴

杨凯¹

摘要：人工智能时代，网络空间反恐形势更加严峻，如何有效利用人工智能技术识别和打击网络恐怖主义内容成为全球各国共同面对的重要课题。欧盟通过《人工智能法案》和《防止恐怖主义网络传播条例》构建起了较为完善的利用人工智能应对网络空间内极端主义、恐怖主义信息传播的法律治理框架。这些都对我国人工智能时代网络空间恐怖主义信息的识别与防范具有重要的借鉴意义，基于欧盟经验，本文提出我国在法律规范完善、人工智能建设、法律体系设计等方面的借鉴路径，以提升我国网络反恐的治理效果，实现我国网络安全法律体系的进一步完善。

关键词：网络反恐 欧盟人工智能法 恐怖主义信息识别 中国镜鉴

引言

网络空间具有世界性，信息辐射全球，这种特点也被传统恐怖分子所利用，以期达到被更多人看到的效果，造成大面积的社会恐慌²。例如 2019 年新西兰克赖斯特彻奇清真寺枪击案和 2022 年美国布法罗超市枪击案中，凶手都通过网络直播实施过程，使得恐怖主义内容在短时间内在全球范围内快速传播，引发各国政府对网络空间内恐怖主义信息传播的强烈关注。随着人工智能技术的迅猛发展，恐怖主义分子借助新的科技手段传播信息，采用人工智能科技的方式在网络空间传播极端主义与恐怖主义信息，如何有效利用人工智能技术识别和治理网络空间中的恐怖主义信息，防止其扩散助长极端主义暴力，已经成为国际社会共同面对的网络安全课题。

欧盟近年在法律和技术层面采取了一系列措施，应对恐怖主义信息在网络空间内的传播。一方面，欧盟制定了《人工智能法案》³（Artificial Intelligence Act，以下简称 AI Act），从法律框架上规范人工智能科技的使用，确保高风险应用不会导致恐怖主义信息的传播；另一方面，欧盟也通过了《防止网络空间传播恐怖主义内容条例》（Regulation on Preventing the Dissemination of Terrorist Content Online，以下简称 TERREG），要求平台在接到命令后一小时内删除恐怖主义内容⁴。与此同时，欧盟各成员国和跨国网络科技企业也采取了多种人工智能手段，来

¹ 中国政法大学国际法学院讲师，18501089426。

² 杨凯《国际法视角下网络反恐问题研究》，中国政法大学出版社 2024 年版，第 70 页。

³ 于智精, Enrique Capdevila. 欧盟《人工智能法案》简析[J]. 中国信息安全, 2024(08): 16-21

⁴ Wahl T. Rules on removing terrorist content online now applicable, *Eu crim*, 2022(2): 112-114.

识别和删除网络空间中的恐怖主义信息。⁵

相比之下，中国同样面临恐怖主义信息在网络空间内传播的风险。自 2015 年颁布《反恐怖主义法》以来，不断强化对网络极端主义内容的监管打击。我国法律规定“互联网服务提供者应当依照法律、行政法规规定，落实网络安全、信息内容监督制度和安全技术防范措施，防止含有恐怖主义、极端主义内容的信息传播；发现含有恐怖主义、极端主义内容的信息的，应当立即停止传输，保存相关记录，删除相关信息，并向公安机关或者有关部门报告……对互联网上跨境传输的含有恐怖主义、极端主义内容的信息，电信主管部门应当采取技术措施，阻断传播。”⁶

借助于全民防范和互联网服务供应商自主研发的审核算法，大量恐怖主义信息在中国网络空间中被封禁。然而，随着人工智能技术的发展，极端组织利用全新人工智能技术手段规避审查、传播极端性内容的能力也在提高，这对我国现有法律管制手段提出了新的挑战。在国际社会的网络与人工智能立法中，欧盟立法独具特色，既有欧盟层面的统一立法，也有国别层面和重要网络服务提供商之间的互动，本文拟通过梳理欧盟人工智能法框架下识别网络恐怖主义信息的具体实践经验，关注人工智能技术在欧盟网络反恐中的应用，分析其对于中国的启示，探讨欧盟人工智能网络反恐立法中我国可以借鉴之处，以便更有效识别恐怖主义信息在网络空间内的传播，推动人工智能在维护我国网络空间安全中发挥更重要的作用，进而完善我国的网络安全法律体系。

第一章 欧盟人工智能法框架下的网络恐怖主义信息治理

一、欧盟《人工智能法案》

欧盟《人工智能法案》(AI Act) 是欧盟首个全面规范人工智能应用的立法，由欧盟委员会于 2021 年 4 月 21 日正式提议，2024 年 3 月 13 日经欧洲议会通过。这一法案的出台不仅体现了欧盟对于人工智能技术带来的潜在风险的高度关注应对，更标志着欧盟在全球率先建立起严格而系统的人工智能技术监管体系。

根据该法案的具体内容，欧盟明确将人工智能系统根据其潜在风险的高低划分为不可接受风险、高风险、受限风险和最小风险四个明确等级，针对不同风险级别实施差异化监管策略。其中，涉及到网络恐怖主义内容的被划分为高风险，高风险人工智能系统的管理尤为严格，涵盖数据治理标准、技术安全性要求、人类监管参与机制以及透明度和可解释性义务等多个维度。

在网络空间内识别和治理网络恐怖主义信息方面，《人工智能法案》提出了一系列详细而严格的监管措施与标准。⁷这些措施明确指出，执法部门在网络空

⁵ Hard Questions: What are we doing to stay ahead of terrorists? <https://about.fb.com/news/2018/11/staying-ahead-of-terrorists/>

⁶ 全国人民代表大会常务委员会. 中华人民共和国反恐怖主义法[R]. 2015 第 19 条

⁷ European Parliament and Council of the European Union. (2024). *Artificial Intelligence Act (AI Act)*. Retrieved

间中用于预防、识别和应对恐怖袭击及其他重大安全威胁的人工智能防卫系统，应当明确归类为高风险应用。具体而言，这类用于网络反恐的高风险 AI 系统，必须严格遵守欧盟现行的数据保护法律规范，在数据的采集、储存、处理与使用过程中，确保数据的合法性与安全性，防止因数据滥用或泄露造成对公民隐私的侵害。此外，法案还特别强调，在实际应用中，必须建立完善的监督机制，确保人工智能的决策始终受到监管，以避免因算法的自动化决策而导致的误判、歧视或权利侵犯。与此同时，《人工智能法案》还对该类系统的透明度和可解释性作出了明确规定。相关方必须能够详细说明人工智能系统的决策逻辑、数据来源及模型构建方式，使得人工智能决策过程清晰可追溯，从而提高执法行动的可信度和公众监督的可能性。这一要求的目的在于防止技术权力的过度集中，确保公众能够理解并监督网络空间内的反恐活动。

为了进一步确保网络反恐行动的合法性与正当性，法案还明确要求执法部门定期对所使用的高风险 AI 系统开展系统性风险评估与基本权利影响分析。这类分析和评估须涵盖系统性能稳定性、伦理风险可能性和对基本权利潜在影响等方面，以保障技术应用的稳健可靠，同时防止技术使用过程中出现的滥权行为。此外，《人工智能法案》特别强调，除非存在预防或应对重大恐怖袭击的紧迫安全需求并获得严格的法律授权，否则严禁执法机构在网络空间对公众私人敏感数据进行大规模、未经授权的采集与处理，包括人脸识别、生物特征数据、位置信息等。这种严格的例外情形处理机制旨在强调技术应用的边界性，确保在极端情况下的技术使用也受到严密监管，防止公民隐私与基本人权受到不必要的侵犯。

综上所述，《人工智能法案》以清晰明确的风险分类和严格细致的监管制度，为欧盟网络空间治理网络恐怖主义信息提供了强有力的法律基础和制度框架。不仅能够提升网络反恐行动的效率和精准性，还能够有效预防技术滥用所带来的社会风险与人权侵犯。

二、欧盟AI专项立法

欧盟于 2021 年出台了《防止网络空间传播恐怖主义内容条例》(TERREG)，为欧盟境内网络空间恐怖主义信息的识别和防范的设置明确的法律义务和处罚措施，是目前欧盟识别和移除线上极端主义、恐怖主义内容的核心法律机制。

(一) 及时有效的删除指令

首先，条例要求所有在欧盟境内提供服务的网络服务供应商，在收到成员国主管机关识别不当消息并发出的移除指令后，必须在一小时之内删除或者采用技术手段屏蔽相关的恐怖主义内容。这一规则的目的是反制恐怖主义内容在网络空间内的扩散，减少公众辐射面。对于未按要求删除内容的网络服务供应商，成员

国可以施加最高额为该公司全球年营业额 4%的罚款，以此来确保网络平台切实履行责任。

（二）平台的主动措施义务

其次，TERREG 规定了平台的主动措施义务。当平台被认定其服务内容中有涉及恐怖主义的内容时，应当采取适当的积极措施预防此类内容的传播。虽然 TERREG 并未在具体技术手段上强制要求，但实际上要求平台采取一切可行的手段加强内容审核，例如利用人工智能技术的自动识别算法来扫描和识别恐怖主义图文及音视频资料。如 Facebook、YouTube 等大型网络平台，也陆续建立了自己的 AI 识别系统，普遍采用了数据库中的哈希(Hash)索引技术，以便在用户上传内容时快速和已有的恐怖素材数据库对比，一旦匹配相似便立刻停止发布。TERREG 还要求各平台定期通过透明度报告公开其处理恐怖内容的情况，包括收到的移除指令数量、处理用时和处理结果等，确保网络空间内恐怖主义内容的迅速有效处理。

除 TERREG 之外，在欧盟最新的数字治理体系中，另有《数字服务法》(Digital Service Act, DSA)也通过强调恐怖主义内容是网络平台需要降低的系统性风险，要进行风险评估并采取风险环节措施，提升内容审核的透明度，由此来强化大型网络平台清理非法内容的责任。⁸

（三）相互配套的机构设置

为了使 TERREG 达到良好的施行效果，欧盟在制度建设层面也做了调整，欧盟刑警组织（European Union Agency for Law Enforcement Cooperation, 以下简称 Europol）创立了 PERCI 平台，各国执法机构将删除指令发送至该平台，再统一传递给各网络服务供应商，提升了跨国协同效率。⁹

欧盟委员会移民与内务司(Directorate-General for Migration and Home Affairs)还设立了专门网站，公布各成员国负责 TERREG 执行的主管机关，方便企业及时确认指令来源的合法性。

综上所述，TERREG 通过立法明确了“政府命令—平台执行”的责任链条，在欧盟范围内统一了网络恐怖内容治理的规则。这一硬性法律框架与 AI 法案形成互补：前者直接规范平台处理恐怖内容的行为，后者则从技术应用角度规范 AI 工具的开发部署。这些法律共同构成欧盟人工智能法打击网络恐怖主义信息的多层次机制：既有具体的快速删除要求和高额处罚威慑，也强调平台应主动监测并运用先进技术识别有害内容，同时注重程序正当性和权利保障。简而言之，欧

⁸ European Union. *Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act)*. Official Journal of the European Union L 277 (October 27, 2022): 1–102. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065>.

⁹ EUROPOL, How is Europol keeping online spaces safe? <https://www.europol.europa.eu/media-press/newsroom/news/how-europol-keeping-online-spaces-safe>

盟的法律机制通过规制技术和约束平台两方面发力：一方面通过 AI 法案等确保用于内容识别的人工智能系统可靠、安全，避免“滥删”或歧视；另一方面用 TERREG 和 DSA 等明确平台在发现和移除恐怖内容上的义务，倒逼企业投入资源采用更先进的 AI 手段来履行这些义务。这种法律与技术相结合的治理模式，为后续技术实践奠定了框架。

第二章 人工智能科技欧盟网络反恐中的具体应用

一、欧盟层面

（一）对中小网络平台的技术扶持

欧盟委员会通过欧盟互联网论坛(EU Internet Forum)推动人工智能技术打击网络恐怖主义的发展，并鼓励中小型平台借助行业数据库提高审查能力

对于资源有限的小平台，欧盟甚至资助开发了开源恐怖主义内容识别工具。可基于 1000 部“伊斯兰国”(ISIS)的视频进行训练，在 < 0.1% 的误差率下检测出 94% 极端视频，这样就使得规模不大的网络平台也能通过人工智能技术，有效过滤大部分恐怖主义内容，从而避免成为极端势力从大平台转移阵地的避风港”。

（二）社交网络的分析预警

在社交网络分析和预警方面，欧盟执法机构也运用了 AI 技术。欧盟刑警组织下设的互联网遣返部门(European Union Internet Referral Unit, EU IRU)自 2015 年成立以来，就利用网络爬虫和机器学习模型主动巡查公开网络，发现恐怖主义内容及时并向平台和成员国举报¹⁰。

IRU 分析员借助自动脚本每天扫描上百个极端组织常用的网站和账户，当系统根据关键词或者图像特征发现疑似的内容时，会标记下来供专家审查，如果确认属于恐怖主义内容，IRU 会直接将链接提交平台要求删除，并共享给相关国家执法部门作为情报。

（三）危险人员的识别与防范

欧盟一些成员国还探索使用 AI 进行行为模式分析和危险人员识别。恐怖主义信息在网络空间内的传播，会引发“独狼式”恐怖袭击，这些恐怖分子正是通过互联网实现了自我极端化，传统侦查方式将难以预警，算法的检测有助于提前发现预兆，提前防范。例如法国在 2021 年修订《反恐法》时授权情报机构使用算法，在海量网络通信数据中筛查可疑迹象，如某用户频繁搜索极端暴力内容等。

11

法国警方也尝试将人脸识别应用于社交媒体图像，以识别发布极端内容或具有潜

¹⁰ Europol. EU IRU Annual Report 2022.

¹¹ Chrisafis A. France planning to allow use of algorithms to detect extremism online, <https://tech.slashdot.org/story/21/04/30/0559241/france-planning-to-allow-use-of-algorithh>

在现实危险倾向的嫌疑人。尽管此类大数据算法监控在欧盟境内引起了一些争议，尤其是对隐私的担忧，但欧盟的法律框架下，通过人工智能法案和专门立法，以及成员国的相关立法，这些技术被限定用于网络反恐，并辅以独立监督，以降低滥用风险。

综上，在欧盟层面，欧盟及成员国充分利用人工智能技术提升了打击网络恐怖主义的能力。在实践中，人工智能帮助平台提前删除了绝大部分恐怖主义内容，最大程度压缩了其传播范围，同时辅助情报预测了潜在的网络风险。从平台内容的算法过滤，到执法层面的网络巡查机构和风险预警模型，AI 贯穿了网络恐怖主义信息“发现—处理—预防”的各个环节。AI 技术已成为欧盟网络反恐不可或缺的利器。

二、企业层面：内容管理实践

互联网平台是网络恐怖主义信息传播的主要载体，也是欧盟监管的重点对象。在欧盟人工智能法框架下，Facebook、YouTube、Twitter 等平台一方面受到欧盟法规制，另一方面也通过 AI 技术的迭代升级提升了网络反恐效果。

（一）Facebook

Meta 旗下的 Facebook 作为全球最大的社交平台，在欧盟法的规制下，除了利用 AI 检测并删除平台中的恐怖主义内容，还设立专门的反恐怖内容团队，集合算法工程师、内容审核员和情报分析员，在发生恐袭事件或出现恐怖威胁时第一时间介入，全平台搜索相关内容并速删速封。

同时，Facebook 推出了反仇恨言论和恐怖宣传的指导方针，并以此来训练 AI 数据大模型。其母公司 Meta 还与欧盟共同启动了“线上公民勇气倡议”（Online Civil Courage Initiative, OCCI）项目，资助民间组织在社交媒体上进行发声，通过“反言论（Counter speech）”，来应对仇恨和极端主义内容。Facebook 每季度公布平台中恐怖主义内容的处理数据，以便公众监督。对于欧盟执法机关的紧急请求，例如发现疑似恐怖袭击策划群组，Facebook 也建立了绿色通道，使用自动化工具迅速导出相关账号信息并配合调查，极端组织在该平台的公开活动空间被大幅压缩。

（二）YouTube

YouTube 作为全球最大的视频分享平台，在欧盟法的规制下也采取了一系列措施应对恐怖主义内容在自身平台的传播。YouTube 应用了先进的视频指纹和 AI 模型分类技术，对上传的视频进行分析，一旦某视频内容与已知恐怖主义、极端主义内容高度相似，那么该内容上传即被阻止。

此外，针对网络恐怖主义内容搜索，YouTube 还采用了创新举措“重定向方法”（Redirect Method），即当用户在搜索栏输入与恐怖主义相关的关键词时，算

法将自动识别，不会呈现搜索结果，而是呈现例如揭露恐怖行动危害等反极端化的内容。这一方法通过算法干预用户搜索结果，在不侵犯言论自由的情况下起到了“反极端化”的引导作用，YouTube 在 EU Internet Forum 框架下积极推广该方法，欧盟也将其视为打击网络激进化和恐怖主义宣传的“非强制性工具”，被欧盟和联合国反恐部门誉为最佳实践之一。

YouTube 还提高了人工审核团队规模，特别是增加了精通多语言、多文化背景的审核员，辅助 AI 判断疑似内容。YouTube 还参与了全球互联网反恐论坛 (Global Internet Forum to Counter Terrorism, GIFCT) 的哈希数据库 (Terrorist Counter Hash Database, TCHD) 共享并与欧盟执法机构保持密切联络，曾经在法国警方请求下及时删除了数百个发布宣扬圣战视频的账户。

总体而言，在欧盟法的运营框架下，YouTube 在欧盟的运营由过去饱受诟病的“恐怖主义温床”逐渐转变为高度警觉的平台，每当恐怖组织更换策略，其算法也快速调整，以确保不出现新的传播漏洞。

（三）其他公司

X 和微软等公司也采取了类似行动。X 在 Twitter 时代就有检测机器人 @CtrlSec 用于检测平台中的恐怖主义内容，在欧盟 TERREG 通过之后，X 利用 AI 检测暴恐图像，每年封禁大量涉恐账号，并在欧盟报告中公布细节。微软则将自家搜索引擎必应与 EU IRU 系统对接，发现恐怖网页时向用户警示或屏蔽，并通过旗下的云服务为中小平台提供内容审核的应用程序编程接口 (Application Programming Interface, API)。

由上述内容可以看到，欧盟通过立法将平台纳入反恐共同体，再借助技术手段赋能，实现了政府、企业和社会组织合力遏制网络恐怖主义宣传的局面。欧盟监管框架促使各大平台形成了“技术投入+协同合作+信息透明”的综合治理模式，这种模式既通过 AI 技术的升级不断剔除违法内容，也彼此共享数据和经验，并接受外部监督。欧盟法框架下这种多管齐下的管理方式，使网络空间中的恐怖主义信息大为减少。正如欧盟官员所指出的，如果没有平台的积极参与，单靠政府命令难以真正清除网络极端主义土壤¹²。

第三章 对中国网络反恐法律体系的镜鉴

当前我国正处于数字经济与人工智能高速发展的重要阶段，网络空间治理，特别是网络恐怖主义信息治理面临严峻形势。一方面，网络恐怖主义信息传播速度快、隐蔽性强，给治理工作带来了巨大困难；另一方面，人工智能在信息治理领域的应用在我国虽取得了一定进展，但相应的法律规范、算法监管机制以及数据安全保护等方面尚存在诸多不足。因此，研究和借鉴欧盟人工智能法识别网络

¹² Council of the European Union. *Fight Against Terrorism*. Accessed April 9, 2025. <https://www.consilium.europa.eu/en/policies/fight-against-terrorism/>.

恐怖主义信息的实践经验，对于完善我国网络恐怖主义信息治理体系具有重要的现实意义。

中国现有法律体系中，2015年颁布的《反恐怖主义法》第十九条要求电信、互联网服务提供者“建立内容监控和安全管理制”，防止传播极端主义，并规定未及时删除恐怖信息可追究法律责任，此外在《网络安全法》《网络信息内容生态治理规定》等也将宣扬恐怖主义、极端主义列为禁止内容。然而，相较欧盟的TERREG等专项法规，我国现行法律在平台及时处置义务、政府处置程序和惩罚力度方面仍有提升空间。未来我国将进一步完善人工智能法律体系，欧盟经验可以对我国有所镜鉴，具体而言，包括如下方面：

一、依法明确网络平台责任

在未来中国的网络反恐立法中，可以借鉴TERREG的部分内容，明确平台处置非法恐怖内容的时限和责任。目前我国法律虽然有要求平台删除极端内容的条款，但缺乏具体时限规定和量化标准。可考虑在相关法规中引入类似“一小时内删除”的要求，针对重大恐怖内容，例如正在传播的恐怖袭击视频设定更加紧急的处理时限，并完善明确的处罚措施，提升对违法企业的经济处罚上限，形成更有力的震慑。这样反过来也会督促国内互联网企业审核技术的迭代升级，确保自身平台出现网络恐怖主义内容时能及时发现并反应。

二、建立明确的政府指令和司法监督机制。

欧盟TERREG规定由成员国主管机关发出删除命令，并允许平台在认为命令不当时提请本国复审，这点也可为我国所借鉴。我国可以指定权威机构，例如公安部第三研究所或其他权威机关，作为网络恐怖信息处置的牵头机关，统一受理举报并下达删除指令。同时引入司法监督，对存疑内容的认定和处置提供复议或申诉渠道。这样既确保删除措施有法定依据，也防止滥用造成对合法言论的过度审查，平衡安全与言论权益。

三、加强网络恐怖主义信息治理的跨部门合作与国际合作。

网络空间具有世界性，恐怖主义内容在网络空间的传播辐射面很广，也很难只由一个部门处理，欧盟的经验表明，网络恐怖信息治理需要执法、情报、网信等多部门协同，更需要国家之间国际信息共享。对此，我国在依法依规设立的机构之间，应当建立公安、国安、网信办、相关高校等部门之间的共享和磋商机制，定期交流新出现的极端宣传动向和应对策略。在国际上，中国可一方面维护好现有的网络安全合作机制，例如上海合作组织的网络安全合作，中国—东盟网络安全合作等；另一方面，借助我国蓬勃发展的AI技术，中国也可以更主动地参与例如借助联合国平台的多边合作，以及GIFCT平台的多边合作等，与其他国家共享恐怖主义黑名单、哈希数据库和技术规范。在网络联通全球的北京下，这有

助于我国获取最新情报，更好地封堵境外极端主义恐怖主义宣传对我国的渗透，也可以使中国的 AI 技术辐射到更多国家，提升中国在 AI 领域的全球影响力与话语权。

四、平衡法律治理与人权伦理

欧盟经验表明，在利用 AI 治理网络恐怖主义信息时，同步考虑法律和伦理问题十分重要。中国在推进相关工作的同时，也应加强这方面的研究和规范。一通过制定专门规章，规范 AI 在网络内容监管中的使用原则，如禁止歧视性算法、保障用户申诉权等方式明确人工智能审查的法律边界，确保 AI 删除内容有法律依据，有动态痕迹，可接受司法审查。二是加强算法透明和第三方评估。欧盟 DSA 中提出了算法透明度和外部审计的要求，我国也可以借鉴，可以要求大型平台定期报告其内容识别算法的性能指标，包括具体数据、错误率以及改进措施，并邀请独立机构评估其对正常言论的影响。三是开展法律宣传。向社会说明使用 AI 打击网络恐怖的重要性 and 必要限度，争取公众理解支持，并及时回应可能的隐私顾虑。在高校和研究机构中支持“AI 治理与法律”交叉研究，培养兼具技术和法学背景的专门人才，为政策制定提供智力支持。

总之，中国可以从欧盟的成功经验中吸取有益成分，结合我国国情加以创新。在完善法规、压实企业责任的同时，充分利用人工智能提升对网络恐怖主义信息的“发现-处置-预防”全链条能力，构筑起坚实的网络安全屏障。

结论

网络恐怖主义信息的泛滥是全球数字时代面临的共同挑战。欧盟通过人工智能立法规范和技术创新，探索出了一套较为成熟的网络反恐治理模式：既以严格法律赋予政府和平台明确职责，又发挥人工智能技术优势实现对恐怖内容的快速精准打击。实践表明，强有力的法律措施能够倒逼企业改进技术，而先进技术的应用又能极大提高法律执行的效能，二者相辅相成。在这一过程中，欧盟亦注重维护基本权利，寻求安全与自由的平衡。对于中国而言，欧盟的经验提供了有价值的镜鉴。在依法严厉打击网络恐怖主义的同时，我们也需不断更新治理理念和手段。通过引入更明确的法律规定和跨部门机制，强化平台主体责任和协同配合；并在此过程中，坚持法治思维和风险防范，确保技术手段在法治轨道上运行。唯有如此，才能在日益复杂的网络安全形势中掌握主动权，切实维护国家和社会稳定。面向科技高速发展的未来，随着人工智能技术的进一步发展，恐怖主义与反恐斗争也将持续演变。我们应当未雨绸缪，在国际合作中贡献中国智慧，在国内治理中融合科技与法治，不断提高识别和清除网络恐怖主义信息的能力水平，为营造清朗安全的网络空间而不懈努力。

《联合国打击网络犯罪公约》的多维比较与衔接展望

上海市人民检察院第二分院课题组¹

摘要：随着网络犯罪呈现全球化、专业化、隐蔽化趋势，传统刑事司法体系与国内立法在应对时力不从心，国际社会迫切需要构建新的治理框架，由此《联合国打击网络犯罪公约》应运而生。文章以多维比较为视角，一方面通过对比其他打击网络犯罪的区域性条约，另一方面与其他既有联合国公约相比较，详细探讨了该公约在制定理念、立法技术、条文体例及国际协作落实等方面的优势与局限，并从实体和程序两方面呈现了与我国打击网络犯罪立法规定的异同。在此基础上，文章着眼公约制定的中国立场，提出了我国未来作为缔约国为落实公约在数字网络立法、刑事立法、国际合作制度等方面的衔接展望，并主张未来通过修正案的修订和补充议定书的制定，进一步完善公约。

关键词：联合国打击网络犯罪公约 多维比较研究 公约完善 国内法衔接 中国因应

引言

作为 21 世纪国际社会应对网络犯罪挑战的重要成果，《联合国打击网络犯罪公约》（以下简称《公约》）已于 2024 年 8 月 9 日由联合国网络犯罪问题特设委员会通过，并于同年 12 月 24 日经第 79 届联合国大会全体同意而正式批准。²该公约共设 9 章 67 条，内容包括刑事定罪、程序措施、国际合作、技术援助等重要章节，首次在全球层面构建起兼具实体法与程序法功能的综合性网络犯罪治理框架。其诞生标志着国际社会在数字经济时代对网络犯罪治理的共识性突破，亦为全球网络安全治理注入新的制度动能。

目前，《公约》已确定将于 2025 年在越南首都河内开放签署，并将在第 40 个缔约国完成批准程序后的 90 天正式生效，其后续实践将对全球网络犯罪治理格局产生深远影响。³

中国作为全球网络空间治理的重要参与者，也是《公约》的推动制定者和未来缔约国，在《公约》开放签署之际，着眼全球治理需求，立足本国主权立场，

¹ 课题组成员：陶建平、王建平、吴晓东、张蓝予、方晗、胡皓煜、金懿、万春艳

² 参见联合国新闻通告：“United Nations: Member States finalize a new cybercrime convention”, https://www.unodc.org/unodc/en/frontpage/2024/August/united-nations_-member-states-finalize-a-new-cybercrime-convention.html; Vibhu Mishra, “UN General Assembly adopts milestone cybercrime treaty”, <https://news.un.org/en/story/2024/12/1158521>,最后访问时间：2025 年 3 月 19 日。

³ 参见联合国新闻通告：Vibhu Mishra, “UN General Assembly adopts milestone cybercrime treaty”, <https://news.un.org/en/story/2024/12/1158521>,最后访问时间：2025 年 3 月 19 日。

通过聚焦《公约》文本内容，对比其与既有联合国公约、区域性条约及国内法的优劣与异同，进而探寻我国当前立法与公约的衔接和因应路径实有必要，也期望通过这些研究，为完善全球网络犯罪治理体系、推动公约有效实施提供理论支持与实践指引。

一、《联合国打击网络犯罪公约》的诞生背景与全球治理需求

随着数字技术的飞速发展，网络犯罪已呈现出全球化、专业化、隐蔽化的新特征。以我国为例，《2021 年最高人民检察院工作报告》显示，在刑事案件总量下降的情况下，2020 年网络犯罪同比上升 47.9%。⁴至 2024 年，起诉网络犯罪总数达 27.2 万人，占起诉总人数的 16.7%。⁵传统犯罪愈加“网络化”正是造成这一现象产生的原因之一。新型犯罪形态对传统刑事司法体系提出了严峻挑战。各国从打击网络犯罪的角度出台了相关国内法，但受限于国家地域性限制，无法作为国际法依据来对国际网络犯罪进行惩治。因此，国际社会亟需构建适应网络犯罪特点的国际治理框架，以弥补打击网络犯罪国内立法之不足、满足维护网络空间秩序的现实需要。⁶

（一）网络犯罪的跨国性催生国际协同治理需求

现代网络犯罪最显著的特点是突破了传统犯罪的地域限制。通过分布式服务器、匿名通信工具和加密货币支付等手段，犯罪行为实施者、犯罪结果发生地、受害者所在国往往分属不同司法辖区。以一起电信网络跨国诈骗案为例，犯罪集团服务器架设在 A 国，资金流转通过 B 国加密交易所，受害者则遍布全球数十个国家。这种“物理空间碎片化、网络空间一体化”的犯罪模式，使得单一国家刑事管辖权难以有效应对。传统司法协助条约规定的“双重犯罪原则”和冗长的审查程序，往往不利于关键电子证据及时取得，犯罪所得也有足够时间被转移隐匿。而这种“跨地域性”特点反过来又加剧了网络犯罪发生的普遍性。以分布式拒绝服务攻击（DDoS）为例，攻击流量可能来自全球数以万计的“肉鸡”⁷设备。与此同时，网络犯罪手法的迭代更新也愈加迅速，这种高度技术化对证据收集、罪名认定等环节提出了特殊要求，超过了各国国内法所能调整的权利义务关系的范围，通过国际立法对网络犯罪进行惩治迫在眉睫。

此外，网络犯罪的产业化和黑灰产业链条化趋势，凸显了系统性应对的必要性。

⁴ 《最高人民检察院工作报告（第十三届全国人民代表大会第四次会议 张军 2021 年 3 月 8 日）》，最高人民检察院官网，https://www.spp.gov.cn/spp/gzbg/202103/t20210315_512731.shtml，最后访问时间：2025 年 3 月 19 日。

⁵ 《最高人民检察院工作报告（第十四届全国人民代表大会第三次会议 应勇 2025 年 3 月 8 日）》，最高人民检察院官网，https://www.spp.gov.cn/spp/gzbg/202503/t20250315_690544.shtml，最后访问时间：2025 年 3 月 19 日。

⁶ 参见赵永琛：《〈联合国打击网络犯罪公约〉的制定问题研究》，载《武大国际法评论》2022 年第 5 期，第 3 页。

⁷ “肉鸡”也称傀儡机，是指可以被黑客远程控制的机器。百度百科，<https://baike.baidu.com/item/肉鸡/33880>，最后访问时间：2025 年 3 月 19 日。

在这一产业链中，从漏洞挖掘、工具开发到洗钱变现，各环节高度专业化，形成了完整的犯罪生态。这种产业化特征使得传统的单点打击模式难以奏效，必须转向全链条治理和系统性应对。国际社会逐渐认识到，单一国家的法律和执法能力已无法有效应对网络犯罪的跨国性、复杂性挑战。因此，建立统一的法律框架和执法机制，加强国际合作，成为全球共识。系统性应对不仅涵盖法律和执法层面的协同，还包括技术援助、能力建设和信息共享等多维度合作。

综上，国家之间有必要统一立场，通过采取制定能够容纳绝大多数国家的框架性协议的方式，进一步加强预防和打击网络犯罪的国际法治力度。⁸《公约》的通过，标志着国际社会在打击网络犯罪方面迈出了关键一步，为各国协调行动、共同应对这一全球性挑战提供了坚实的法律基础。

（二）回应国际合作与发展均衡化的现实诉求

当前，网络犯罪治理的国际合作正面临“数字鸿沟”与“制度壁垒”的双重挑战。这种技术能力与合作机制的结构性失衡，使得发展中国家在打击网络犯罪时既缺乏必要的技术手段，又难以有效参与国际合作机制。在此背景下，《公约》的制定体现了对均衡发展需求的积极回应。

以应对网络犯罪的视角来看，世界不同国家和地区在信息技术发展上存在明显差距。发达国家在信息技术领域具有显著优势，拥有先进的网络基础设施、强大的技术研发能力和专业的网络安全人才。而发展中国家则在这些方面相对滞后，网络基础设施薄弱，技术研发能力有限，网络安全人才匮乏。这种技术能力的不均衡使得发展中国家在面对网络犯罪时往往处于劣势。例如，一些发展中国家缺乏必要的网络监控和追踪技术，难以有效监测和追踪网络犯罪活动。这使得犯罪分子可以利用这些国家的技术漏洞，实施跨国网络犯罪，而这些国家却难以有效应对。此外，发展中国家在网络安全防护方面也存在不足，容易成为网络攻击的目标，导致大量数据泄露和经济损失。

数字技术的快速发展使得网络犯罪呈现出的专业化、产业化特征与发展中国家普遍面临取证技术落后、电子证据保存系统不完善、网络溯源能力不足等问题产生的冲突，是打击跨国网络犯罪的主要矛盾之一。这种矛盾不仅影响了各国的网络安全治理能力，也给各国在网络犯罪打击方面的合作带来诸多困难。在跨国网络犯罪的调查取证过程中，往往需要各国互相提供技术支持和数据共享，而部分国家由于技术能力不足，难以提供有效的支持。这使得跨国网络犯罪的调查和取证工作难以顺利进行，影响了国际合作的效果。

除了技术能力的差异外，国际社会在网络犯罪治理领域的分歧始终难以弥合，究其原因是在规则主导权、治理理念上的结构性矛盾日益凸显，根源在于既有国际规则体系的失衡。纵观《公约》之前的国际网络犯罪治理体系，呈现

⁸ 参见赵永琛：《〈联合国打击网络犯罪公约〉的制定问题研究》，第3页。

出较为明显的“双轨并行”格局：一方面是以《布达佩斯公约》⁹为代表的区域性公约主导治理实践，另一方面则是美欧通过国内法扩张域外管辖权，导致治理碎片化与规则冲突的存在。

一方面，之前的区域性公约，在代表性和均衡性上均存在一定缺陷。以《布达佩斯公约》为例，作为首个区域性网络犯罪治理框架，其制度设计存在显著的“欧洲中心主义”特征。该公约由欧洲委员会主导制定，谈判过程排斥发展中国家参与，导致其规则体系难以反映非西方国家的法治诉求，缺乏国际社会代表性。¹⁰例如，《布达佩斯公约》第32条“跨境数据调取”条款允许缔约国执法机关直接向境外网络服务提供者调取数据，这种“数据长臂管辖”机制对习惯了一体化的欧盟各国并无桎梏，但对许多非西方国家而言，会认为侵犯了数据储存地国的司法主权而难以接受。类似现象亦存在于《阿拉伯公约》¹¹和《马拉博公约》¹²中。这些区域性公约虽填补了地区治理空白，但因其适用范围有限且缺乏协调机制，反而间接促成了全球治理的碎片化。此外，这些公约出台时间较早，主要针对传统网络犯罪（即“对象型网络犯罪”），对于数量日益增长的利用网络实施的传统犯罪（即“工具型网络犯罪”）关注较少，无法顺应新时代国际网络空间治理的需求。¹³

另一方面，以美国《云法案》¹⁴和欧盟《电子证据条例》¹⁵为代表的单边立法，试图通过“数据控制者”标准突破传统司法主权利限制。例如，《云法案》允许美国执法机构直接调取境外企业数据，欧盟《电子证据条例》则规定只要企业向欧盟居民提供服务即受其管辖。此类机制虽提升了跨境取证效率，却因其“长臂管辖”属性引发广泛争议：发展中国家认为其侵蚀国家主权，加剧了“数据霸权”

⁹ 全称为《布达佩斯网络犯罪公约》，是全球首个针对网络犯罪进行国际合作的条约。该公约由欧洲委员会制定，主要由欧洲国家于2001年签署并于2004年生效，但也对其他国家开放。它的主要目标是通过国际合作，改善打击跨国网络犯罪的效率，保护个人数据，打击计算机犯罪、网络诈骗、儿童色情等网络犯罪行为。目前已有76个国家加入，虽被部分非欧洲国家（如日本、美国）接受，但其核心仍是欧洲理事会主导的区域性公约，其影响力主要集中在西方国家。

¹⁰ 参见于志刚：《缔结和参加网络犯罪国际公约的中国立场》，载《政法论坛》2015年第5期，第97-98页。

¹¹ 全称为《阿拉伯国家联盟打击信息网络犯罪公约》，系阿拉伯国家联盟于2010年签订。

¹² 全称为《非洲联盟网络安全和个人数据保护公约》，系非洲联盟于2014年签订。

¹³ “对象型网络犯罪”主要是针对计算机信息系统或者网络信息系统，破坏其完整性和有效性，如侵犯计算机信息系统、非法获取计算机信息系统数据等犯罪；“工具型网络犯罪”指利用网络进行诈骗、赌博、散布淫秽视频、侵犯个人隐私或信息、发布侮辱性或者诽谤性言论等犯罪。参见裴炜、黄志雄、冯俊伟、谢登科、李怀胜：《〈联合国打击网络犯罪公约〉：背景、内容与展望》，载《数字法治》2024年第5期，第11-12页；彭文华：《犯罪治理现代化视野下网络犯罪的刑律体系化》，载《苏州大学学报(哲学社会科学版)》2023年第1期，第73页。也有学者将“对象型网络犯罪”称之为“纯粹网络犯罪”，参见张路遥、龚雯聪：《联合国打击网络犯罪公约相关法律问题的各国立场》，载《中国信息安全》2020年第9期，第85-86页。

¹⁴ 全称为《澄清境外数据的合法使用法》(Clarifying Lawful Overseas Use of Data)，是美国于2018年3月23日签署生效的一项法律。该法案确立的向网络服务提供者调取非适格外国刑事数据的取证途径，本质上是绕开外国主权者而开展的单边跨境取证活动。参见孙永超：《论美国跨境刑事数据调取中的国际礼让：以〈云法案〉为例的分析》，载《求是学刊》2024年第3期，第126页。

¹⁵ 全称为《关于在刑事事项中获取和保存电子证据的条例》(Regulation (EU) 2023/1543 on European Production Orders and Preservation Orders for electronic evidence in criminal matters)，于2023年7月12日由欧洲议会和欧盟理事会通过。

与“数字主权”的对立。更严重的是，这些单边规则缺乏国际法合法性，在导致企业面临多国管辖权冲突，被迫在合规与用户隐私保护之间陷入两难的同时，更有使得网络犯罪治理沦为大国博弈的工具的风险。这种治理模式不仅违背国家主权平等原则，更导致发展中国家在网络安全领域陷入“制度性失语”。

《公约》的出台，本质上是发展中国家通过多边主义打破规则垄断的尝试¹⁶：通过将网络犯罪治理议题纳入联合国框架，确保所有国家（尤其是技术能力较弱的发展中国家）平等参与规则制定，从而重塑全球治理的权力结构，标志着南北合作从理念共识迈向制度实践。

二、比较视角下《联合国打击网络犯罪公约》的优势与局限

《公约》在科技进步与社会发展中发端，在东西博弈中成形，在妥协和让步中诞生，既肩负着在联合国框架体系下整合各国力量，加强国际司法合作，打击计算机信息网络犯罪的重任，也不可避免的存在各种问题和缺陷。通过立足《公约》原文，以多维比较的视角充分挖掘《公约》的先进与不足，才能为各国继续发挥智慧来解决和弥合相关问题提供可行路径。

（一）与其他区域性条约的比较

1. 《公约》系联合国框架下国际法体系的进一步完善

先前由发达国家主导的单边主义实践（如美国《云法案》）与区域性合作框架（如欧盟《电子证据条例》）试图通过“规则输出”扩大自身管辖权，但此类机制缺乏全球合法性，反而引发发展中国家对“数字殖民”的警惕。《公约》作为首部全球性网络犯罪治理法律框架，其诞生不仅是国际社会对网络犯罪跨国化特征的回应，更是南北合作框架下机制创新的标志性成果。这一进程背后，既体现了发展中国家争取规则制定话语权的努力，也展现了通过多边协商平衡主权、人权与技术差异的务实路径。

相较上述区域性条约，《公约》具备更多优势特点。

首先，《公约》是由联合国主导制定，是全球首个普遍适用的防止网络犯罪的国际法律文件，覆盖联合国 193 个成员国。《公约》旨在制定一个全球适用的网络犯罪法律框架，目标是通过广泛的国际合作形成统一的打击网络犯罪的国际规则。而区域性条约适用于特定地区，主要侧重关注各区域面临的网络犯罪问题，合作范围受限。

其次，《公约》由所有联合国成员国共同参与谈判，力求兼顾各国法律体系和发展水平，特别是发展中国家的立场，这也使得更多国家能够接受并实施该公约。《布达佩斯公约》虽然签订早、影响力大，但主要由欧美主导，中俄等重要的大国和许多发展中国家未加入该公约。《马拉博公约》《阿拉伯公约》等其他区

¹⁶ 参见李哲、朱晓琴：《〈联合国打击网络犯罪公约〉的中国立场与核心问题》，载《北京师范大学学报（社会科学版）》2024年第5期，第124页。

域性条约仅涵盖特定成员国，在弥合发达国家与发展中国家之间在网络安全治理上的差距方面，影响力更是有限。

再次，《公约》在刑事定罪方面覆盖的网络犯罪类型宽于区域性条约。不仅包括传统的计算机及网络技术犯罪，还涵盖了部分借助网络技术实施的传统犯罪。其立法用语具有技术中立性，能够在一定时间跨度内兼容未来技术的发展。而其他区域性条约均有其侧重打击的方面，如《布达佩斯条约》更侧重于计算机及网络技术犯罪；上海合作组织在国际信息安全领域推出和补充系列合作协议，提供了实施信息安全合作的重要法律保障，但更侧重打击网络恐怖主义和信息安全，¹⁷且受限于制定时的技术发展水平，均与打击网络犯罪的现有需求存在差距。

最后，《公约》致力于推进广泛的国际合作。《公约》的内容涉及全球执法合作、信息共享和跨境调查取证，设立国家间协调机制，避免执法障碍，当然仍面临着实际操作中的困难与磨合。而其他区域性条约往往本就由某一区域性组织的成员国签订，比较容易达成共识。例如《布达佩斯公约》和《阿拉伯公约》的一些规定，在本就一体化程度较高的欧洲联盟或阿拉伯国家联盟基础上，规定了相似的跨境数据取证规则，以让渡数据主权换取更高合作效率的跨境取证模式。

2. 《公约》对电子数据取证措施的吸收和改造

同时，《公约》也从《布达佩斯公约》等区域性条约中吸收了大量内容，尤其是电子数据取证方面。

电子数据的跨境取证是打击网络犯罪的司法实践中的关键环节，长期以来，呼唤一个统一的标准来推进该领域的国际合作，但该问题同时也是发达国家与包括中国在内的众多发展中国家的焦点分歧。

《布达佩斯公约》体现了欧盟一体化发展的成果，设置了单边直接调取电子数据的内容，许多发展中国家之所以没有加入《布达佩斯公约》就是对此心存疑虑，但受西方发达国家影响，《布达佩斯公约》中的“快速保全”“提交令”等内容仍然被导入《公约》中。然而，经过博弈和妥协，这些制度在被引入时，也受到了一定的调整，如《公约》删去了《布达佩斯公约》中缔约国可单边直接调取存储在他国境内电子数据的内容。

经过整合调整，《公约》一定程度平衡了国家司法主权和打击犯罪、保护人权的关系。目前，《公约》第 28、29、30 条规定的对电子数据的搜查、扣押、收集、拦截等取证方式是针对电子数据的强制取证方式，其适用对象只能为位于境内的电子数据；如需跨境强制调取，则需按照《公约》第 40 条第 3 款请求国际刑事司法协助。而《公约》第 25、26、27 条是要求电子数据拥有者或服务商快速保全或者提交，属于公私合作的内容，强制性稍弱，故并未明确规定只能针对

¹⁷ 参见王海燕：《上海合作组织信息安全合作的机制建设与挑战》，载《中国信息安全》2021 年第 8 期，第 78 页。

存储在境内的电子数据。但是《公约》第 40 条第 3 款司法协助范围也未明文规定上述情形必须通过国际司法协助才能获取。所以尽管“提交令”的提出是为了达到国际社会能够保存并快速调取与犯罪调查有关的数据,维持适当的司法互助制度,提高跨境调取数据的制度安排效率,但当缔约国通过“快速保全”和“提交令”向其境内的电子数据拥有者及服务提供商提出数据要求,可能仍会与《公约》第 5 条“保护主权”原则产生冲突。¹⁸

(一) 与其他联合国公约的比较

《公约》作为继《联合国打击跨国有组织犯罪公约》¹⁹(以下简称 UNTOC)及《联合国反腐败公约》²⁰(以下简称 UNCAC)之后全球犯罪治理领域的重要成果,广泛吸收了两个国际公约的有益经验。²¹尽管上述公约均以打击跨国犯罪为核心目标在联合国框架下诞生,但因各自所针对的犯罪特性不同,在诸多方面存在差异。

1. 条文体例和立法技术层面

第一,《公约》表现出较强的技术适应性与罪名专业化,但在个别表述及犯罪打击的前瞻性和灵活性上仍有欠缺。作为全球首个打击网络犯罪的法律文件,《公约》一是对网络犯罪相关概念的界定更为精准和专业,如对“信息和通信技术系统”“电子数据”“流量数据”“内容数据”等核心术语²²均做了精细化的网络概念界定,确保其适用于各类新兴技术;二是通过详细阐释刑事定罪内容为各国认定网络犯罪行为提供清晰标准,如明确界定“非法访问”系“无权情况下故意访问信息和通信技术系统的全部或任何部分的行为”²³,更贴合网络时代犯罪特征,突破了 UNTOC 和 UNCAC 更依赖传统法律概念的限制,也弥补了其他两个公约在网络犯罪领域的空白。²⁴

但对于部分条文的设计和表述,也不乏批评声音。其一,有的学者认为《公约》条文在文字表述质量上参差不齐,部分条文的规定较为笼统,有的语义晦涩,有的语言不精炼,影响到对文本内容的把握。²⁵其二,部分条文或因嵌套过多其他条文而降低了文本可读性,如《公约》第 49 条²⁶层层引用的条文多达 13 条,

¹⁸ 参见李本灿:《〈联合国打击网络犯罪公约〉核心议题述评》,载《法学论坛》2025 年第 1 期,第 99-100 页。

¹⁹ U.N. Convention Against Transnational Organized Crime, 于 2000 年 11 月 15 日由第 55 届联合国大会通过。

²⁰ United Nations Convention Against Corruption, 于 2003 年 10 月 31 日由第 58 届联合国大会通过。

²¹ 叶伟、姜博谦:《〈联合国打击网络犯罪公约〉开启网络空间国际规则新篇章》,载《中国信息安全》2024 年第 8 期,第 49 页。

²² 参见《公约》第 2 条“术语的使用”。

²³ 参见《公约》第 7 条“非法访问”。

²⁴ 相比之下,UNCAC 虽涉及部分网络腐败行为(如第 21 条“私营部门贿赂”),但未细化电子证据规则,导致司法实践中的认定困难。

²⁵ 具体参见江溯:《〈联合国打击网络犯罪公约〉“刑事定罪”部分评述》,载《中国信息安全》2023 年第 3 期,第 59 页。

²⁶ 《公约》第 49 条引用了《公约》第 1 条、50 条,而《公约》第 50 条又引用了《公约》第 1 条、31 条、第 40 条的相关内容。其中第 40 条包含 32 款,且引用了《公约》第 18 条及 44、45、46 条。第 44 条又引

甚至出现了循环引用。这样的立法方式，结合文本语言语种的因素，势必会造成理解适用上的巨大分歧。其三，作为公约的核心内容，“刑事定罪”一章所确立的网络犯罪类型在普适性与前瞻性方面仍存在一定缺憾。²⁷如《公约》第二章刑事定罪部分规定的犯罪主要涉及“对象型网络犯罪”，但对于近年来迅猛发展且亟需规制的“工具型网络犯罪”关注有限；²⁸也未对区块链、人工智能等新技术及此类新技术支撑下的网络黑灰产业链催生的新型犯罪形态作出探索性规定。²⁹此外，《公约》为实现国际合作打击网络犯罪，制定了一些统一的标准和规则，这在一定程度上保障了打击犯罪的一致性和协同性。然而，在面对各国不同的网络发展水平和犯罪特点时，这种统一标准可能缺乏灵活性。反观 UNCAC，其采用较为灵活的规定方式，允许各国依据自身国情和法律体系制定具体实施细则，充分考虑了不同国家在政治体制、法律文化、经济发展水平等方面的差异。在反腐败机构的设置、腐败犯罪的认定标准以及预防措施的实施等方面，各国可以根据本国实际情况进行调整，提高了公约的适应性和可操作性。

第二，《公约》条例框架清晰、内容丰富，但在一定程度上忽视了对既有公约的衔接与协调。

《公约》全文共设九章 68 个条文，章节设置紧密围绕网络犯罪的特点和打击需求。核心内容聚焦于刑事定罪、程序措施和国际合作三个关键部分，逻辑清晰合理，契合网络犯罪打击流程。这种条文体例安排，使各国在实施公约时，能清晰把握不同环节重点任务和操作流程，为有效打击网络犯罪提供有力制度框架。相比之下，UNTOC 仅 41 条，条文相对简洁，在应对复杂多样的犯罪情形以及细化国际合作具体细节方面，条文体例的丰富度略显不足，无法像《公约》一样全面细致地覆盖网络犯罪打击的各个方面；而 UNCAC 内容虽全面，但结构上却不如《公约》针对特定犯罪类型的逻辑紧密。尽管也有学者认为，《公约》的刑事定罪章节先规定核心犯罪行为，后规定责任、参与形式等，显得条文编排混乱，不符合我们惯常先总后分的思维。³⁰但对比 UNCAC 和 UNTOC 可见，这种整体先总后分、局部先定罪后执法的条约制定方式也属于公约惯例。

用了《公约》第 35 条、42 条。第 35 条引用了 40-46 条的内容，出现了循环引用。《公约》第 42 条引用了《公约》第 25 条、41 条和 40 条的相关内容。

²⁷ 敬力嘉：《〈联合国打击网络犯罪公约〉“刑事定罪”部分的缺憾与展望》，载《中国信息安全》2024 年第 8 期，第 61 页。

²⁸ 也有学者将“工具型网络犯罪”称之为“网络赋能型犯罪”，参见裴炜、黄志雄、冯俊伟、谢登科、李怀胜：《〈联合国打击网络犯罪公约〉：背景、内容与展望》，载《数字法治》2024 年第 5 期，第 18 页，体现在相较于第四次会议案文（A/AC.291/16），第六次会议案文（A/AC.291/22）对“刑事定罪”部分进行了重大删改，删除了几乎所有网络化传统犯罪的罪名。持此观点的还有敬力嘉：《〈联合国打击网络犯罪公约〉“刑事定罪”部分的缺憾与展望》，载《中国信息安全》2024 年第 8 期，第 61 页。

²⁹ 参见敬力嘉：《〈联合国打击网络犯罪公约〉“刑事定罪”部分的缺憾与展望》，载《中国信息安全》2024 年第 8 期，第 64 页。

³⁰ 参见江溯：《〈联合国打击网络犯罪公约〉“刑事定罪”部分评述》，载《中国信息安全》2023 年第 3 期，第 59 页。

《公约》在网络犯罪领域带来了一定的创新性，但其在一定程度上忽视了与既有公约的协调性，可能导致法律适用冲突。最典型表现在于《公约》对于“双重犯罪原则”³¹的突破可能引发司法互助机制上的冲突。如 UNTOC 第 18 条要求司法互助需符合“双重犯罪原则”，而《公约》第 42 条第 4 款明确规定，“被请求缔约国收到另一缔约国的请求后，应当根据本国法律采取一切适当措施，快速保全所指定的电子数据。在回应请求时，被请求缔约国不得以两国共认犯罪作为提供此种保全的条件”。这表明在快速保全存储的电子数据方面，被请求缔约国无需以请求事项在两国均构成犯罪为前提，突破了传统的“双重犯罪原则”。虽然该条款未具体指明针对哪些技术性网络犯罪，但结合公约整体内容来看，与电子数据密切相关的犯罪都可能涵盖其中，如公约第二章“刑事定罪”中规定的非法访问（第 7 条）、非法拦截（第 8 条）、干扰电子数据（第 9 条）、干扰信息和通信技术系统（第 10 条）等犯罪，这些犯罪均基于信息和通信技术系统，涉及电子数据的不当操作。若缔约国同时受 UNTOC 约束，就可能面临义务冲突。这种机制差异在实践中可能导致两种不利后果：一是“协助选择化”，缔约国可能选择性地援引对自己有利的公约条款；二是“请求冲突化”，当两个公约的协助请求同时到达被请求国时，缺乏明确的优先适用规则。

2. 责任规定及协作落实层面

《公约》在国际协作机制上较先前有显著改进，尤其是在跨境数据调取、应急响应及公私协作方面，但却存在主权让渡争议等问题。

首先，在跨境电子证据调取上，《公约》第 27 条“提交令”允许执法机构直接向网络服务提供商调取数据，突破了 UNCAC 第 46 条“需经主权审批”的限制，使得执法机构能够更快捷地获取涉案数据，以应对网络犯罪的快速性和隐蔽性特点，有助于提高打击网络犯罪的效率。然而 UNCAC 之所以在司法协助时涉及主权审批环节，目的也是保障国家主权和法律秩序，因此《公约》的做法也可能引发数据主权争议。

其次，在应急响应上，《公约》第 41 条明确了“24/7 联络点机制”，更适应网络犯罪打击的时效性需求。该机制要求缔约国设立全天候联络点，以应对网络犯罪的紧急情况。相比之下，UNTOC 第 18 条仅规定传统的司法互助请求需通过中央机关书面提交，效率较低；UNCAC 虽在第 46 条鼓励“快速互助”，但未建立具体操作机制。同样体现《公约》时代化特征的还有第 40 条关于“司法协助的一般原则和程序”中的第 14 款，增加了对“电子形式传送和接受司法协助请求、与之相关的通信以及证据”的鼓励，而 UNTOC 及 UNCAC 均维持在以书面形式（紧急情况下可以口头，但应立即加以书面确认）进行的司法协助请求。³²

³¹ “双重犯罪原则”是国际刑事司法协助中的一项基本原则，即请求国与被请求国均认定该行为构成犯罪。

³² 参见 UNTOC 第 18 条第十四款和 UNCAC 第 46 条第十四款。

再者,《公约》在私营部门协作方面较 UNTOC 和 UNCAC 更具约束力。如前文所述,《公约》在电子数据取证方面从《布达佩斯公约》等区域性条约中吸收了不少内容,加强了公约的执行力。其中多个条款明确了网络服务提供商的协作义务,要求私营部门参与,如第 25 条“快速保全存储的电子数据”³³明确了网络服务提供商的法定义务,需配合执法机关保全数据,没有选择的余地;第 27 条“提交令”³⁴直接赋予执法机关要求网络服务提供商提供特定信息的权力,网络服务提供商必须执行;第 29 条“实时收集流量数据”和第 30 条“拦截内容数据”也对网络服务提供商提出要求,如强令服务提供商在其现有的技术能力范围内,配合并协助主管机关实时收集或记录流量数据、内容数据,且要对权力行使情况及相关信息保守秘密。这些条款明确了网络服务提供商在不同侦查环节的具体责任,具有很强的针对性,且通过设定缔约国的立法和其他措施义务,间接使网络服务提供商的协作具有法律强制力。相比之下,UNTOC 中企业合作的相关规定较为分散,且多为倡导性内容,如第 31 条“预防”中提及加强与企业界的合作,防止有组织犯罪集团对公共当局招标程序、补贴和许可证以及法人等的不正当利用,但未明确企业具体的合作义务和不合作的后果。这种倡导性规定给予企业较大的自主选择空间,企业可以根据自身情况决定是否合作以及合作的程度。

综上,对比 UNTOC 和 UNCAC,《网络犯罪公约》在技术适配性、协作效率方面均有显著进步,但在与既有公约衔接、主权平衡方面存在短板。未来或需通过缔约国实践填补漏洞,并借鉴 UNTOC 和 UNCAC 的成熟机制。

(三) 与中国国内法的比较

1. 刑事实体法方面

条文定义上,《公约》与国内法对于电子数据的定义有所区别。《公约》第 2 条对电子数据的定义为“对事实、信息或概念的任何陈述,其形式适合在信息和通信技术系统内处理,包括适合使信息和通信技术系统执行某一功能的程序”。

³⁵而我国两高一部《办理刑事案件收集提取和审查判断电子数据若干问题的规定》(以下简称《电子数据规定》)³⁶第 1 条规定,“电子数据是案件发生过程中形成的,以数字化形式存储、处理、传输的,能够证明案件事实的数据。”两者概念虽然规定不同,但实质范围基本相近。在对电子数据有关的外延规定方面,《公

³³ 《公约》第 25 条要求缔约国采取措施,使主管机关能下令网络服务提供商等相关人员快速保全使用信息和通信技术系统存储的指定电子数据,且要在规定时限内(最长不超过 90 天,可延期)保全和维护数据完整性。

³⁴ 《公约》第 27 条规定,缔约国应授权本国主管机关下令在本国境内提供服务的网络服务提供商提交其所拥有或控制的与此类服务有关的订户信息。

³⁵ 参见《公约》第 2 条“术语的使用”(b)项“电子数据”。

³⁶ 为规范电子数据的收集提取和审查判断,提高刑事案件办理质量,最高人民法院、最高人民检察院、公安部于 2016 年 9 月制定发布了上述规定。参见《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》,中华人民共和国最高人民检察院官网, https://www.spp.gov.cn/zdgz/201609/t20160921_167425.shtml, 最后访问时间:2025 年 4 月 8 日。

约》第2条对电子数据外延从逻辑上划分为三大种类：“流量数据”、“内容数据”、“订户信息”，《电子数据规定》第1条对电子数据进行定义后，又补充了4项外延表现形式。我国采取列举式的方式较为直观，但是可能存在遗漏，而《公约》概括式的定义更为周延。

刑事定罪内容上，《公约》“刑事定罪”一章共15条，对犯罪行为、法人责任、犯罪形态、追诉时效、追诉事项等五个部分作出了规定。从条文数量来看，该章仅次于第五章“国际合作”（18条），这从一个侧面凸显了《公约》对刑事定罪问题的关注。³⁷《公约》第7条至第17条以行为或手段的方式规定了应当确立为刑事犯罪行为的情形，主要涵盖了“对象型网络犯罪”（第7至12条）以及少量“工具型网络犯罪”，例如财产犯罪（例如第13条盗窃或欺诈、第17条洗钱）、涉及儿童的性犯罪（例如第14、15条）、侵犯个人信息及隐私的犯罪（例如第16条未经同意传播私密图像）。

《公约》明确列举的犯罪中，除了保留针对信息和通信技术、电子数据实施的“对象型网络犯罪”以及与儿童色情有关的网络犯罪等《布达佩斯公约》原有的犯罪，还增加了利用信息和通信技术洗钱的相关犯罪。³⁸经过梳理，相关犯罪行为在我国《刑法》中均能找到大致对应的相关罪名。

	《联合国打击网络犯罪公约》第二章	《中华人民共和国刑法》相关罪名	相同点
非法访问计算机系统	规定非法访问计算机系统为犯罪行为，包括未经授权访问计算机系统或其数据。	《刑法》第 285 条：非法侵入计算机信息系统罪。	均将未经授权访问计算机系统或数据的行为定义为犯罪。
非法拦截数据	规定非法拦截计算机系统内的非公开传输数据为犯罪行为。	《刑法》第 285 条之一：非法获取计算机信息系统数据罪。	均将非法获取或拦截计算机系统内的数据行为视为犯罪。
数据干扰	规定故意破坏、删除、损坏、篡改或抑制计算机数据的行为为犯罪。	《刑法》第 286 条：破坏计算机信息系统罪。	均将破坏、删除、篡改计算机数据的行为定义为犯罪。
系统干扰	规定故意干扰计算机系统功能的行为为犯罪，包括输入、传输、删除或损坏计算机数据。	《刑法》第 286 条：破坏计算机信息系统罪。	均将干扰计算机系统功能的行为视为犯罪。
滥用设备	规定生产、销售、获取、进口、分发或提供用于实施网络犯罪的设备或工具为犯罪。	《刑法》第 285 条之三：提供侵入、非法控制计算机信息系统程序、工具罪。	均将提供用于网络犯罪的工具或设备的行为定义为犯罪。

³⁷ 参见周振杰：《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》，载《中国刑事法杂志》2024年第5期，第107页。
³⁸ 李哲、朱晓琴：《〈联合国打击网络犯罪公约〉的中国立场与核心问题》，载《北京师范大学学报》（社会科学版）2024年第5期，第128页。

		《刑法》第 280 条： 伪造、变造、买卖国 家机关公文、证件、印章罪；第 286 条： 破坏计算机信息系 统罪。↔
计算机相关伪造	规定通过计算机系统输入、更改、删除或抑制计算机数据，以产生虚假数据并意图使其被当作真实数据使用的行为为犯罪。	伪造、变造、买卖国 家机关公文、证件、印章罪；第 286 条： 破坏计算机信息系 统罪。↔
计算机相关诈骗	规定通过计算机系统实施的欺诈行为为犯罪。↔	《刑法》第 266 条： 诈骗罪；第 287 条： 利用计算机实施犯 罪的提示性规定。↔
儿童色情相关犯罪	规定制作、分发、传播、获取或持有儿童色情材料为犯罪。↔	《刑法》第 364 条： 传播淫秽物品罪；第 367 条：明确将儿童色情材料列为淫秽物品。↔
侵犯版权及相关权利	规定故意侵犯版权及相关权利的行为为犯罪，包括通过计算机系统实施的侵权行为。	《刑法》第 217 条： 侵犯著作权罪；第 218 条：销售侵权复制品罪。↔

图一 刑事罪名对比

《公约》所规定的部分具体犯罪行为可以通过解释被我国《刑法》所涵盖。如《公约》第 7 条“非法访问”、第 8 条“非法拦截”、第 9 条“干扰电子数据”、第 10 条“干扰信息和通信技术系统”中的绝大部分内容可以被我国《刑法》第 285 条、第 286 条所涵盖。³⁹《公约》第 21 条规定的起诉、审判与制裁等追诉事项本身就非常原则化，而且规定就假释、辩护、行为合法性判断等事项，由各缔约国国内法规定，所以可以认为我国《刑法》现有规定可以涵盖。⁴⁰

但仍有部分具体犯罪行为是我国现行《刑法》未能完全涵盖的。如相较我国刑法，前述第 9 条“干扰电子数据”的“干扰”行为方式更广泛，包括劣化、隐瞒数据，且“电子数据”不限于计算机信息系统，涵盖所有信息和通信技术系统，计算机系统外的数据的干扰行为就无法为我国《刑法》所涵盖⁴¹；《公约》第 11 条“滥用装置”中的行为包括“获取、制作、采购使用、进口等行为”，行为对象包含了“程序以及访问类数据”，而这已超越了我国《刑法》第 285 条第 3 款提

³⁹ 参见周振杰：《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》，载《中国刑事法杂志》2024 年第 5 期，第 113 页；敬力嘉：《〈联合国打击网络犯罪公约〉“刑事定罪”部分的缺憾与展望》，载《中国信息安全》2024 年第 8 期，第 65 页；李本灿：《〈联合国打击网络犯罪公约〉核心议题述评》，载《法学论坛》2025 年第 1 期，第 95 页。

⁴⁰ 参见周振杰：《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》，载《中国刑事法杂志》2024 年第 5 期，第 115 页。

⁴¹ 李本灿：《〈联合国打击网络犯罪公约〉核心议题述评》，载《法学论坛》2025 年第 1 期，第 98 页。

供侵入、非法控制计算机信息系统程序、工具罪所规定的“提供”行为和“程序、工具”对象。⁴²又如根据《公约》第 14 条第 1 款，缔约国需将“通过信息和通信技术系统征求、获取或访问儿童性虐待或儿童性剥削材料”（第 b 项）及拥有或控制存储于电子设备中的此类材料（第 c 项）的行为纳入刑事犯罪。而我国《刑法》目前尚未对此类行为设定独立罪名，仅能通过传播淫秽物品罪（第 364 条）等条款有限覆盖，存在规制空白。⁴³

《公约》第 17 条有关洗钱犯罪的规定亦是要求各缔约国扩大洗钱罪的上游犯罪范围。我国《刑法》第 191 条虽然规定了洗钱罪，并涵盖了《公约》第 17 条规定的行为类型，但是该条规定的上游犯罪范围显著窄于《公约》。⁴⁴当然，如果将我国《刑法》规定的掩饰、隐瞒犯罪所得、犯罪所得收益犯罪解释为广义的洗钱犯罪，则可以认为《公约》第 17 条被我国《刑法》涵盖。

此外，对于《公约》规定的一些任择性措施与要求⁴⁵，我国《刑法》可能也无法涵盖。例如，《公约》第 16 条将“未经同意传播私密图像”确定为刑事犯罪行为，该行为可以根据我国《刑法》第 364 条规定的传播淫秽物品罪处罚。而《公约》第 16 条在第 5 款设置了任择条款，“缔约国可规定须有造成伤害的意图才能追究刑事责任”。对此，有观点认为，对于传播淫秽物品的目的犯，我国《刑法》目前仅有第 363 条规定的传播淫秽物品牟利罪与之相关，除此之外，并无有关“伤害的意图”的规定。⁴⁶但考虑到《公约》中的伤害除直接（身体）伤害外，还应包含间接伤害，如心理伤害。⁴⁷在此情形下，我国《刑法》第 246 条规定的侮辱罪或许能有所涵盖。

除了上述和刑事定罪有关的内容外，《公约》第 32 条“犯罪记录建档”，还涉及到境外前科的承认。目前，我国仅有《刑法》第 10 条涉及对境外刑事判决的消极承认，主要解决的是对境外判决过的刑事案件，是否还需要追究刑事责任的问题，而“犯罪记录建档”则还会影响到境内其他案件判决的量刑问题。就此而言，我国刑法尚无法涵盖《公约》此方面的要求。

2. 刑事程序法方面

⁴² 同上，第 95 页。

⁴³ 参见周振杰：《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》，载《中国刑事法杂志》2024 年第 5 期，第 115 页。

⁴⁴ 同上，第 115 页。

⁴⁵ 《公约》为缔约国设定的义务类型可以分为三类：一是强制性义务，指各缔约国必须依照公约规定予以适用，不允许以本国的意志变更或排除适用的规定，一般使用“应当采取”（shalladopt）的表述方式；二是任择性要求，强制的程度稍弱，通常使用“应当考虑采取”（shallconsideradopting）或者“应当努力”（shallendeavourto）的表述方式；三是任择性措施，指公约所规定的允许缔约国选择为或不为一项措施的规范，通常使用“可以采取”（mayadopt）措辞。参见杨帆：《〈联合国打击网络犯罪公约〉草案条文若干问题述评》，载《中国信息安全》2023 年第 3 期，第 57 页。

⁴⁶ 参见周振杰：《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》，第 116 页。

⁴⁷ 在《公约》起草过程中，曾明确规定对《公约》确立的犯罪进行加重处罚的情节包括“对个人造成伤害，包括身体或心理创伤”。参见《公约》（草稿）（A/AC.291/16）版本，第 39 条第 2 款（c）项。https://www.unodc.org/documents/Cybercrime/AdHocCommittee/4th_Session/Documents/A_AC291_16_Advance_Copy.pdf 最后访问时间：2025 年 4 月 8 日。

《公约》第四章“程序措施和执法”从第 23 条至 34 条共十二条，对刑事侦查或诉讼的程序做出了规定。

相比国内法，《公约》的程序措施适用范围无法适用于全部刑事犯罪。除此以外，《公约》和国内程序法的差异主要有以下几个方面：

第一，关于电子数据取证规范体系的差异。根据《公约》的规定，《公约》建立了颇具特色的电子数据取证规范体系。

一是建立电子数据的快速保全措施，具体分为对存储数据的快速保全和对流量数据的快速保全、披露。对应的，我国两高一部的《电子数据规定》规定了电子数据冻结，公安部《公安机关办理刑事案件电子数据取证规则》（简称《电子数据取证规则》）⁴⁸对冻结电子数据的措施作了细化规定。但电子数据的概念和《公约》有所差异，且在对当事人权利保护方面、程序保密方面有所差距。

二是建立提交令制度，根据《公约》提交令的规定，可要求缔约国境内的人或服务提供商提供其拥有或控制的电子数据。对此，我国《刑事诉讼法》《电子数据规定》《电子数据取证规则》规定的内容较为笼统，即向有关单位和个人、收集调取证据，对可以调取的数据范围并未明确。

三是制定强制调取电子数据措施，包括搜查、扣押，实时收集流量数据，拦截内容数据等⁴⁹。对应搜查、扣押，我国《刑事诉讼法》第 136 条规定了针对人身、物品和场所相关搜查措施，我国《电子数据规定》第 5、8、9、10 条，规定了对电子数据扣押、提取、固定等措施。对应实时收集流量数据、拦截内容数据，《电子数据取证规则》第 23 条也规定了对公开发布的电子数据、境内远程计算机信息系统上的电子数据，可以通过网络在线提取，并规定了网络远程勘验。上述《电子数据规定》和《电子数据取证规则》初步构建了包括“网络在线提取”“对远程计算机信息系统进行网络远程勘验”“采取技术侦查进行网络远程勘验”等三个侦查手段的数据取证体系。⁵⁰但我国规定的网络在线提取和网络远程勘验等侦查手段和实时收集、拦截不同，前者更关注静态的电子数据，而后者则同时关注动态的电子数据变化。此外还需特别关注的是，如上文所述，《公约》所规定的上述强制调取电子数据措施均指向的是缔约国本国境内的数据⁵¹；而《电子

⁴⁸ 为规范公安机关办理刑事案件电子数据取证工作，确保电子数据取证质量，提高电子数据取证效率，公安部于 2019 年制定发布了该取证规则。参见《公安机关办理刑事案件电子数据取证规则》，中华人民共和国公安部官网，<https://www.mps.gov.cn/n6557558/c6337154/content.html>，最后访问时间：2025 年 4 月 8 日。

⁴⁹ 参见《公约》第 28 条“搜查和扣押存储的电子数据”、第 29 条“实时收集流量数据”、第 30 条“拦截内容数据”。

⁵⁰ 参见魏光喜、刘想树：《跨境数据取证中的公私合作——以〈联合国打击网络犯罪公约〉为视角》，载《河北法学》2024 年第 8 期，第 87 页。

⁵¹ 参见本文第二章（一）“与其他区域性条约的比较”中的“2.《公约》对电子数据取证措施的吸收和改造”部分。

数据规定》第9条第2款⁵²及《电子数据取证规则》第23条⁵³的表述似乎为跨境远程勘验提供了一定的法律基础，然而上述规定之间不仅缺乏国际公约的依据，也和我国主张的主权保护理念相违背，甚至两个规定之间也存在一定的分歧，未来或需进一步予以释明和调整。

第二，关于犯罪所得处置。《公约》第31条规定了“扣押、冻结、没收犯罪所得”，不仅涉及犯罪所得，还涉及用于或拟用于犯罪的财产、设备、工具的处置。该条第4、5款详细规定了涉案财物已混合、转化的等值追缴措施，同时在该条第9款规定了保护善意第三人的权利。我国《刑法》《刑事诉讼法》对于涉案财物处置（包括犯罪所得、收益和犯罪工具）亦有详尽的法律规定。需要指出的是，我国刑事诉讼制度除了对违法所得没收的一般规定，还规定了犯罪嫌疑人、被告人逃匿、死亡案件违法所得没收特别程序，而《公约》并无类似条款。

第三，关于证人、被害人的权利保护。《公约》第33、34条分别规定了关于本公约所确立的犯罪中，对证人、被害人及其亲属的保护与帮助。相对应的，我国《刑法》《刑事诉讼法》《公安机关办理刑事案件程序规定》《公安机关办理刑事案件证人保护工作规定》都作了相当完善的规定，包括对未成年被害人作证的特殊保护规定。但《公约》第34条还规定了对被害人的康复帮助、屏蔽删除性侵害有害信息等内容，我国国内法尚无明确规定。

三、基于中国立场的《联合国打击网络犯罪公约》的发展完善与国内立法因应

作为《公约》重要的推动者和未来的缔约国，中国为构建兼具国际兼容性与本土适应性的网络犯罪治理体系，需在坚持本国立场的基础上，分析《公约》的发展完善路径，并提出国内立法因应的策略建议，以推动构建公平、包容、可持续的全球网络犯罪治理新秩序。

（一）《公约》制定的中国立场

1. 中国倡导的网络治理方案

作为《公约》的首倡者和推动制定者，我国在打击网络犯罪方面始终坚持如下理念：首先是尊重各国主权。我国主张“数据主权”，反对长臂管辖。数据安全要尊重各国的法律和政策，避免强加他国的法律标准，并倡导国际社会尊重不同国家的文化和社会发展状况。其次是注重隐私保护与数据安全。我国对于信息技术产品和服务的供应链安全甚是重视，坚决反对针对他国的大规模监控、个人信息非法采集。提倡信息技术产品和服务供应企业应统一行为道德规范，杜绝留

⁵² 《电子数据规定》第9条第2款：“对于原始存储介质位于境外或者远程计算机信息系统上的电子数据，可以通过网络在线提取”。

⁵³ 《电子数据取证规则》第23条：“对公开发布的电子数据、境内远程计算机信息系统上的电子数据，可以通过网络在线提取。”从字面文义理解，该规定将远程勘验的范围限缩为境内计算机信息系统上的电子数据，境外只有公开数据才能进行远程勘验。

后门、强制升级等行为。强调各国在保护个人隐私、数据安全和促进信息技术进步方面要加强合作，共同应对跨境数据泄露和滥用的问题。再次是谨慎推动数据跨境流动规则。中国主张各国在确保数据安全的前提下，推动数据跨境流动，并通过建立透明、互信的规则，避免过度限制。最后是深化全球合作。推进数字化基础设施建设，促进国家之间的信息和数据互联互通，尤其是通过技术援助促进发展中国家的数字发展。

2. 对中国倡导的认同与反对

对于我国倡导的理念，国际社会的反应并非全然一致，而是呈现出明显的区域性差异。《公约》的制定过程和最终成果也反映出各国对我国理念的认同及分歧。

认同方面主要集中在全球合作、技术援助和保障发展中国家网络安全自主权上。国际社会普遍认识到网络空间是一个全球化领域，打击网络犯罪需要跨国协作。在这一点上，我国主张通过联合国框架及其他多边合作机制推动国际规则的制定，得到了一些发展中国家及金砖国家的认可。它们认为这种模式能平衡不同发展阶段国家的利益，促进共同发展。对于许多发展中国家来说，西方主导的网络治理模式往往被视为对网络安全自主权的侵蚀。我国倡导的信息通讯产品、服务及网络自主控制的理念，与这些国家追求网络安全自主权、数据安全的诉求相契合，获得了较高的认同。我国提出发达国家应当通过资金与技术援助帮助发展中国家提升网络安全能力，这一举措也受到发展中国家的积极响应。许多国家认为，只有通过多边合作并提供切实的技术支持，才能共同应对网络犯罪的挑战。

而反方面则主要集中在所谓的“过度主权化”、限制数据自由流动和潜在人权问题上。一些西方国家及其盟国攻击我国强调“数据主权”的观点，认为这种主张可能导致互联网分割化和信息封锁，限制了全球信息流通与开放性，甚至攻击我国借此正当化网络审查和言论控制。同时西方国家攻击我国对技术和数据流动设置壁垒，对跨境数据流转进行过多限制，影响全球数字经济的自由发展。西方国家还认为过度强调国家审查和控制可能损害公民隐私和言论自由，他们呼吁在打击网络犯罪时必须确保严格的人权保障机制。

3. 对中国倡导的呼应与差异

《公约》在谈判和出台过程中，对中国倡导的理念进行了以下方面的呼应和修正：

首先，尽管我国强调国际合作应尊重国家主权，《公约》宗旨一章也强调了尊重各国主权，但并未将“尊重政治独立”，“尊重各国历史、文化、社会制度的多样性”“合作打击分裂主义、极端主义以及煽动民族、种族和宗教敌意”等内容明确写入。

其次，在刑事定罪模式方面，我国主张“宽定罪模式”，即《公约》的定罪范围要突破“对象型网络犯罪”的狭义模式，涵盖利用计算机实施的传统犯罪。⁵⁴而《公约》最终采纳了“窄定罪模式”与“宽定罪模式”的折中，主要涵盖了“对象型网络犯罪”以及少量“工具型网络犯罪”。⁵⁵

再次，我国对于跨境数据取证公私合作的态度作出了调整，虽然仍强调国家主权、司法管辖权和对数据的安全管理权，对绕开国际刑事司法程序而直接跨境数据取证持谨慎态度，但也支持跨境数据取证的公私合作，我国同意了《公约》关于“快速保全”“提交令”等程序规定以及设立“24/7 全天候联络点”的要求，以促进跨境数据取证方面的国际合作。

最后是人权保障方面，我国并未对《公约》的人权保护的相关条款提出异议，确保在打击网络犯罪时尊重个人隐私和基本自由。这些对立立场的重申和调整，恰恰体现了中国在维护国家主权和隐私保护的同时，积极参与全球网络犯罪治理的努力。

（二）《公约》落实的国内法衔接展望

我国虽然积极推动《联合国打击网络犯罪公约》的制定，但经过博弈后的文本并非为我国的法律制度量身定制，和我国国内法的法规体系、思维习惯、术语定义等方面均有不同程度的差异。公约签订后，出于履行国际义务及更好的发挥公约作用的需要，有必要通过立法、发布司法解释、完善规章制度等形式，推动国内法与公约的衔接。

1. 数字网络立法方面应加强系统性融合

各国的立法现状和文化背景或多或少均有差异，在此前提下制定的《公约》是各国意愿的最大公约数，所以公约关于数字网络治理方面的逻辑架构和各缔约国的国情往往有所差异，我国也不例外。

其一，基础概念术语等方面需要国内法与之融合。⁵⁶从公约名称看，国内法需要重新整合网络的概念以适应国际公约。我国翻译的名称为“联合国打击网络犯罪公约”，这里的“网络”主要指“信息和通信技术系统”，而这两个术语既有密切关系，内涵又有差异，“网络”更侧重技术层面物理或逻辑上的连接结构，而“信息和通信技术系统”则涵盖更广泛的技术和应用。

从关键术语看，国内法需要对电子数据分类分级。如前文所述，《公约》将电子数据划分为“流量数据”、“内容数据”、“订户信息”三大种类，同时《公约》所涉的程序措施和国际合作对三类信息的规定并不相同。一种情形是不区分三类

⁵⁴ 参见张路遥、龚雯聪：《联合国打击网络犯罪公约相关法律问题的各国立场》，载《中国信息安全》2020年第9期，第85-86页。

⁵⁵ 参见本文第二章（三）“与中国国内法的比较”中的“1. 刑事实体法方面”部分。

⁵⁶ 参见吴沈括：《联合国打击网络犯罪公约》的治理体系与中国应因，载《中国法律评论》2025年第1期，第224页。

电子数据作出统一规定，如《公约》第 25 条规定缔约国有权快速保全存储在信息和通信技术系统的电子数据；另一种情形是针对某一类特定的数据作出规定，如《公约》第 26 条规定服务提供商有义务快速保全和部分披露流量信息。相应的，我国两高一部《电子数据规定》第 1 条和《人民检察院办理网络犯罪案件规定》第 27 条虽有对电子数据的定义和分类，但无法和《公约》的分类一一对应。为了精准实施《公约》规定的内容，国内法也有必要引入“流量数据”、“内容数据”、“订户信息”的概念和对应的程序规定。

其二，公私合作方面需要系统性设计。网络犯罪的特点使得与案件有关的电子数据在全球范围内分散储存并快速流动，仅依靠传统的调查取证思路将会力不从心，特别是碰到跨境调取电子数据，需要依靠冗长的国际司法协助程序，很难及时有效的获取证据。在此背景下，跨境数据取证的公私合作被视为一种有效的手段。⁵⁷此次《公约》在其他国家的实践探索的基础上也吸纳了公私合作的方式。在我国，虽然也有一定的公私合作的尝试，但主要基于司法实践的推动发展，没有在国内法上的系统性设计，需要统筹《公约》的规定、国际先进经验和国内法的现状，建立系统的规范体系。具体而言，一是要求服务提供者加强与司法部门、执法部门的合作，并进一步明确服务提供者的范围及提供电子数据的范围，根据《公约》规定，对服务提供者控制的数据亦可作为证据提供。如要求服务提供商快速保全证据，就需要预先规定制度，让服务提供商具备快速保全证据的能力以及能够迅速掌握需要保全证据的服务提供商名录的措施。二是规定完整的电子数据调取程序规定，细化公私合作的协作义务，加强对电子数据的隐私保护，避免诉讼过程中的不当披露。

2. 刑事法方面的衔接应综合考量

《公约》和我国刑事实体法、刑事程序法密切相关，为更好的履行国际法义务，有必要加强国内法和国际法的衔接。

其一，刑事实体法方面的衔接。《公约》的刑事定罪条款从和国内法兼容的角度可以分成两种类型：第一类是和国内法完全兼容的部分，也就是现有刑法已经涵盖的内容。由于《公约》体例和条款表述和我国刑法的差异，此类情况极为罕见，在和总则有关的部分条文中勉强可以认为完全兼容，如《公约》第 19 条“参与和未遂”所规定的内容⁵⁸和我国刑法总则中共同犯罪、犯罪未遂及犯罪预备等内容较为契合。

第二类是和国内法不完全匹配的部分。所以对于第二类情形，我们又要分为三种情况：

⁵⁷ 参见魏光喜、刘想树：《跨境数据取证中的公私合作——以〈联合国打击网络犯罪公约〉为视角》，载《河北法学》2024 年第 8 期，第 80 页。

⁵⁸ 根据《公约》第 19 条，“参与”系“以共犯、从犯或教唆犯等任何身份参与根据本公约确立的犯罪”的犯罪行为；“未遂”系“将故意实施根据本公约确立的犯罪的任何未遂行为”；另有“将故意为实施根据本公约确立的犯罪做准备的行为确立为刑事犯罪”的规定。

第一种是可以通过法律解释来解决的,对这种情况,解决方案是司法解释、指导性案例,以司法路径为主。这类问题主要是刑法已涵盖但司法实践中不明确的内容,对于刑法未涵盖,但和现有规定不相违背的,也可以考虑在司法实践中,通过一定的扩张解释来解决。如前文所述,将我国刑法规定的掩饰、隐瞒犯罪所得行为解释为广义的洗钱犯罪,可以满足《公约》第17条对犯罪所得的洗钱行为的打击要求。又如《公约》第32条关于犯罪记录建档,即将境外对被告人的有罪判决作为刑事案件的参考,此类规定《公约》并未强求一定要通过立法措施,也可以通过其他措施实现,故可以考虑通过司法解释的方式,将查明的境外网络犯罪记录作为量刑的酌定情节予以考虑。

第二种是明显和现有国内法有冲突或者立法没有涉及的,这个必须要通过立法途径来解决。典型的如《公约》第14条第1款的(b)、(c)两项,“通过信息和通信技术系统征求、获取或访问儿童性虐待或儿童性剥削材料”及“拥有或控制存储在信息和通信技术系统或其他存储介质中的儿童性虐待或儿童性剥削材料”,根据我国现行刑法的规定,单纯获取或者拥有、控制色情信息并不构成犯罪,必须通过立法方能满足履行国际义务的需要。

第三种是保留条款或者允许各国灵活掌握的任择性条款。公约保留条款,如《公约》第11条滥用装置的第1款除第1项第2目以外的内容均可保留,对于这些内容,我们要把握国际法的发展大势所趋、我国的主张和国家利益所在,确定是否保留。从11条的内容看,总体和我国刑法285条第3款的内容可以衔接。而《公约》中不少条款允许各国灵活掌握,如《公约》第9条第1款规定:“各缔约国均应采取必要的立法措施和其他措施,在本国法律中将在无权情况下故意破坏、删除、劣化、更改或隐瞒电子数据的行为确立为刑事犯罪行为。”该规定和我国刑法第286条破坏计算机信息系统犯罪的罪状表述较为契合,但我国刑法有后果严重的要求。对此,《公约》第9条第2款允许缔约国把造成严重损害作为行为构成犯罪的要件之一。

其二,刑事程序法方面的衔接。我国国内法和《公约》有诸多差异,需要加强衔接。

第一,与电子数据快速保全有关的制度衔接。对应《公约》第25条、26条规定的电子数据快速保全的措施,我国虽有两高一部《电子数据规定》和公安部《电子数据取证规则》规定的冻结电子数据的措施。但我国规定与《公约》相比,尚有需要进一步加强衔接之处。一是权利保护方面,《公约》规定,数据保全需要规定最长不超过90日的期限,对于需要延期的,应当设定相应规定;⁵⁹而目前我国《电子数据取证规则》所规定的电子数据的冻结期限为六个月⁶⁰,远高于《公

⁵⁹ 参见《公约》第25条第2款的规定。

⁶⁰ 参见《电子数据取证规则》第39条。

约》确定的一般期限。从实践角度，较长的冻结期限在一定程度上为执法机关全面、深入、细致地调查案件提供了充足的保障，但对于数据所有者或使用者而言，则会影响其正常的生产经营或个人生活。为加强衔接，我国可以考虑进一步细化电子数据冻结期限的规定：根据案件的复杂程度、性质等因素，合理设定不同档次的冻结期限，如对于简单案件，可以设定与《公约》90日相接近的一般期限；对于复杂案件，在符合法定条件的情况下，可以适当延长冻结期限，但应明确延长的具体条件和审批程序，增强与《公约》规定的兼容性。二是程序方面，对照《公约》第25条第3款需要设置数据保全义务人对快速保全程序的保密义务。

第二，与提交令有关的制度衔接。提交令的规定和我国严格的司法主权立场并不协调，但这类权力并非单向的，我国不行使，其他缔约国也有权行使。因此，我国在完成签署后，一方面需要尽快和国内法做好衔接，通过完善国内立法、规范操作程序、加强人员培训等确保国内法律与《公约》要求相契合，避免出现法律适用冲突及和他国在权力行使方面不对等的局面；另一方也应当在司法执法中积极且审慎地行使提交令的相关措施以有力打击跨境网络犯罪，确保在不损害我国司法主权和其他国家的合法权益的情况下，维护我国的网络安全和公民的合法权益。

第三，与强制调取电子数据有关的制度衔接。如上文所述，《公约》针对存储的数据规定了搜查和扣押的方式，针对传输的数据规定了实时收集流量数据和拦截内容数据的方式。我国两高一部《电子数据规定》及公安部《电子数据取证规则》虽也有相关技术侦查措施，但针对的对象主要是静态存储的数据，对动态数据的实时收集和拦截的可操作性还待加强。同时，对于当前存在于我国两部关于电子数据取证规范中涉及远程勘验措施不完全一致的规定，需给予一定的关注。⁶¹一方面，要注意因目前不存在更高位阶的电子数据专门性立法，实际操作中，执法人员可能会因规定不明确而难以确定跨境远程勘验的具体适用情形和程序，影响案件办理效率和公正性；另一方面，为遵循我国所主张的“数据主权”理念，应在涉及境外电子数据的获取时进一步明确具体条件和限制程序，严格落实审批手续，避免对他国数据主权造成侵犯。

第四，证人、被害人保护制度方面的衔接。对照《公约》，我国《刑事诉讼法》还需要规定康复帮助、屏蔽删除性侵有害信息等内容。

3. 国际合作制度方面需要完善强化

我国关于国际合作的国内法主要是《引渡法》和《国际刑事司法协助法》，对照《公约》关于国际合作的内容需要进一步调整充实。

其一，需要对《引渡法》的内容作修改。如《公约》第37条第11款对《公约》确定的网络犯罪，明确了“或引渡或起诉”原则。我国《引渡法》第9条规

⁶¹ 参见本文第二章（三）“与中国国内法的比较”中的“2.刑事程序法方面”的内容。

定，我国对于引渡请求所指的犯罪具有刑事管辖权，并且对被请求引渡人正在进行刑事诉讼或者准备提起刑事诉讼的，对于引渡请求所指的犯罪，可以不引渡。根据国内法的表述，代入具体场景看，基本契合“或引渡或起诉原则”，但其中准备提起刑事诉讼的情形并不能完全保障该原则，需要对条文内容做进一步修改。

但有的内容可以作灵活掌握。如《公约》第 37 条第 2 款，在本国法允许的情况下，可以突破双重犯罪原则准予他国引渡被请求引渡人，但从坚持国家主权角度出发，让他国引渡我国不认为是犯罪的人员对维护国家主权不利，在《公约》允许各国灵活掌握的前提下，可以坚持双重犯罪原则。

其二，我国国内法需补全刑事诉讼的移交制度。根据《公约》第 39 条第 1 款：“如果缔约国认为移交诉讼有利于正当司法，特别是在涉及多个法域时，为了进行集中起诉，应当考虑相互移交诉讼的可能性，以便对根据本公约确立的犯罪进行刑事诉讼”，对此，我国国内法尚属空白，考虑到移交的诉讼前提必然是我国刑法具有管辖权，可以通过修改完善《国际刑事司法协助法》及相关司法解释、相关程序规定的方式予以衔接。

其三，设置相关的机构设置。《公约》借鉴欧盟的经验，引入了 24/7 全天候网络。我国可通过相应的立法设立相应的机构，或者通过执法机关（如公安部）设置相应的内设机构指定一个每周 7 天、每天 24 小时开放的联络点。

其四，其他各项基础制度的衔接。基于《公约》对于电子数据的快速保全、提交令、强制调取规定以及其他规定和国内法的区别，我国《国际刑事司法协助法》对应的内容也应和《公约》相统一。

其五，建立便于国际合作的外国法查明机制。《公约》为统合各国主张，一些条款允许各国保留适用或者根据国内法灵活掌握，一些条款需要根据各国的国内法制定的情形落实。如《公约》第 40 条规定司法协助的一般原则和程序，其中第 17 款规定：“请求应符合被请求国国内法程序，兼顾请求国要求的程序。”因此，为了能够达到请求他国协助的目的，势必要掌握被请求国的相关国内法及程序，故而有必要建立相关外国法查明机制。

（三）《公约》完善的后续渠道与要点

《公约》是东西博弈的产物，交织着理念的冲突和妥协，《公约》的规定有着限制功能发挥的局限性，部分条文语义模糊，强制效力不佳。好在根据《公约》规定，《公约》生效满五年后，可以提出修正案，同时，《公约》还可以通过补充议定书，为《公约》的完善预留了一扇门。从完善《公约》的角度出发，有以下方面可以考虑。

其一，通过后续补充议定书的修订，进一步扩大刑事定罪范围。⁶²一是适应

⁶² 参见裴炜、黄志雄、冯俊伟、谢登科、李怀胜：《<联合国打击网络犯罪公约>：背景、内容与展望》，载《数字法治》2024 年第 5 期，第 19 页。

打击恐怖主义犯罪、毒品犯罪等国际社会共同需求的需要,增设网络工具型犯罪。二是基于个人权益保护的需要,推动与电信网络诈骗犯罪等网络治理相关的犯罪,改良推广我国在打击“帮助信息网络犯罪活动罪”等方面的经验。三是针对最新的网络技术问题,如人工智能、区块链、元宇宙等产生的新问题、新技术加强回应。⁶³

其二,经过一段时间的实践,在后续修订修正案过程中需要注意以下内容。一是继续贯彻倡导维护国家主权和安全的理念,尤其要注意解决国际协作效率与主权让渡争议的平衡问题,如对获取未在境内存储但有控制权的电子数据,是否存在需要限制的情形。二是注重现有条文与既有公约的衔接与协调,一方面要充分援引或整合 UNTOC 和 UNCAC 中的既有协作机制,加以连贯和完善,另一方面也要对可能导致冲突或重复的条款进行调整,如通过确立“特别法优先”、“从重吸收”、“程序从新”等冲突适用原则,明确优先适用规则,避免因协调不足导致的法律适用混乱。三是各国基于对《公约》条款理解不一致等原因发生争议纠纷的,应当明确能够为各国所接受的争端解决机制及救济手段,尤其是当缔约国对《公约》第 63 条第 2 款作出保留的情况下。同时,可以参考 UNCAC 的经验通过进一步丰富和完善以“缔约国会议”⁶⁴为核心的履约审议框架,以“条约监督”、“法治评估”等⁶⁵组合拳的方式增强《公约》的有效实施和顺利实践,以提高《公约》的软性约束力。

四、结语

《公约》在网络犯罪全球化、复杂化的严峻形势下诞生,承载着国际社会协同打击网络犯罪、维护网络空间秩序的殷切期望。从其诞生背景来看,网络犯罪的跨国性、国际合作的不均衡以及现有国际法体系的不完善合力推动了《公约》的制定与出台。经过比较研究可以发现,《公约》作为联合国框架下首个全球性网络犯罪治理法律文件,在打击犯罪覆盖范围、兼顾各方立场、技术适应性及跨境协作效率等方面显著超越了既有区域性条约和其他联合国公约,但其在主权平衡、保障执行及与既往公约及国内法协调性等方面仍存在改进空间。

从中国视角看,《公约》的诞生既体现了发展中国家争取规则制定话语权的努力,也反映了国际社会在数字时代平衡安全与发展、效率与公平的复杂博弈。中国始终秉持多边主义和网络主权原则,倡导构建包容、公正的全球网络犯罪治理体系。在公约谈判中,中国推动尊重各国司法主权、强化发展中国家能力建设等主张被部分吸纳,但也面临西方国家对数据自由流动和人权保障议题的争议。未来,中国需在履行公约义务的同时,通过司法解释、立法修订等方式完善国内

⁶³ 参见吴沈括:《〈联合国打击网络犯罪公约〉的治理体系与中国应因》,载《中国法律评论》2025 年第 1 期,第 218 页。

⁶⁴ 参见《公约》第 57 条。

⁶⁵ 参见张茜:《〈联合国反腐败公约〉履约审议机制研究》,华东政法大学 2019 年博士学位论文,第 29-30 页。

法体系，尤其在电子数据分类、跨境取证公私合作、刑事罪名扩展等领域实现与《公约》的有效衔接。

随着《公约》于 2025 年开放签署并逐步生效，其实施效果将深刻影响全球网络犯罪治理格局。展望未来，各国应在公约框架下，加强沟通与协作，充分发挥公约的积极作用。同时，各缔约国需依据公约要求，完善国内立法，确保公约的有效实施。中国也将继续秉持自身理念，深度参与全球网络犯罪治理，推动公约不断发展完善。期待在国际社会的共同努力下，《公约》能够切实发挥作用，有效遏制网络犯罪，构建更加安全、有序、公平的全球网络空间，让数字技术更好地服务于人类社会的发展与进步。

2025年中国国际法学会年会

人工智能主权的结构化分析

——兼论中国人工智能主权的制度构建*

李艳华¹

摘要：技术领域的运动正在蔓延至数据、人工智能、安全和隐私，我们进入人工智能主权时代。人工智能主权是指国家在其境内控制和监管人工智能的能力，其具有对内的人工智能的控制权，对外的独立自主权与合作权。人工智能主权内部结构由基础层、技术层与规范层构成，外部效应可能造成综合性安全的实施差距、全球南北方的代表性差距、监管的协调差距、国内主权与国际贸易规则的合规性差距。人工智能主权总体上符合我国的根本利益，为此，需要以包容性价值观为合理内核，基础层、技术层与规范层的构建与联动为基础，歧视性贸易反制为防御措施，全球人工智能与全球数据安全协同为对外合作纽带，以实现我国人工智能主权的制度构建。

关键词：人工智能主权 结构化 构建路径

问题的提出

人工智能不仅是机器学习的基础，更将成为决策体系的核心驱动力，其重塑人类社会各领域的影响力正呈指数级增长。预计到 2030 年，人工智能将推动全球经济价值高达 19.9 万亿美元。²这场颠覆性革命引发全球人工智能发展与监管制度的协同竞争。2023 年 11 月，英伟达首席执行官黄仁勋提出，各国必需建立“主权人工智能”（sovereign AI），通过对人工智能基础设施有效控制，以维护各自的文化、社会与历史特征。³同时，美国在联合国提出《全球人工智能包容性伙伴关系》推广人工智能主权，基于对所服务地区不同文化、语言和传统的理解和尊重，该协议重点关注对人工智能模型的访问、技术能力培养与本地数据集的扩展。⁴近来，印度、日本、法国、加拿大和阿拉伯等国家或地区投入数十亿美元来补贴人工智能主权的建设。⁵

* 本文系 2021 年度司法部重点课题“习近平法治思想指导下涉外法治体系完善的结构化分析”（21SFB1006）的阶段性成果。

¹ 李艳华，华东政法大学中国法治战略研究院师资博士后，特聘副研究员，主要研究方向为数字贸易法、数据法、国际经济法。联系电话：15651745645；邮箱：2755685337@qq.com。

² Ina Fried, AI Will Add \$19.9 Trillion to Global Economy by 2030, IDC Predicts, <https://www.axios.com/2024/09/17/ai-global-economy-idc-2030>.

³ Reuters, Nvidia CEO Huang Says Countries Must Build Sovereign AI Infrastructure, <https://www.reuters.com/technology/nvidia-ceo-huang-says-countries-must-build-sovereign-ai-infrastructure-2024-02-12/>.

⁴ 包括国务院、Amazon、Anthropic、Google、IBM、Meta、Microsoft、Nvidia 与 OpenAI。

⁵ Sovereign AI Is the Next Fault Line in the Global Tech Sector, <https://www.bain.com/insights/sovereign-ai-is-the->

安全、社会价值利益、经济效益与战略优势构成国家推行人工智能主权的关键原因。当人工智能设计系统未能考虑区域差异时，它可能会导致效率低下、偏见和应用缺乏相关性。控制人工智能基础设施、数据和技术可以有效遏制外国间谍活动或算法操纵导致的相关漏洞，防止过度依赖外国行为者获得关键组件或服务，从而保障国家安全。国家还通过将数据保存在境内，制定国内隐私法规，以保护公民隐私。民主价值观、基本权利、反歧视与伦理价值同样构成人工智能主权要保护的核心价值。人工智能可以优化生产流程、提高产品质量、推动产品设计和定制创新，降低交易成本，拓展传统产业的规模经济，满足个人偏好，并且提高当地医疗保健、教育和交通等公共服务的效率和可及性。⁶人工智能行业通过创造高技能就业、刺激创新和吸引投资来促进经济增长，从而为全球人工智能市场带来竞争优势。在战略优势上，人工智能促进自力更生和战略自主权，据此提升自身的科技竞争力与世界影响力。

不过，事实上，并非所有国家都能满足设计和实施人工智能主权的政治、体制与战略需求。这一概念的提出塑造了全球范围内人工智能部署格局，但同时也加剧了监管的碎片化，并可能引发地缘政治紧张局势、知识产权冲突、贸易争端与军事对抗等。

综合学界现有研究来看，有分解人工智能主权构成要素的，数字基础设施、数据主权和隐私、国家安全、劳动力发展、经济自主、监管和道德与文化框架、国际合作与标准、对人工智能的态度等构成人工智能主权的关键变量。⁷有依人工智能主权的控制程度，将人工智能主权划分为严格的人工智能主权、温和的人工智能主权、松散的人工智能主权、无人工智能主权的。⁸上述研究有助于解释不同国家的人工智能政策。然而，既有研究尚未能体系性的阐释人工智能主权的实质结构。

从我国来看，网络主权、数据主权无论在学者讨论，还是国家战略制定与立法主旨等方面都得到高度肯定。而强调地域特征与功能特征的人工智能主权不仅

next-fault-line-in-the-global-tech-sector-tech-report-2024/.

⁶Ahmed Banafa, The Rise of Sovereign AI: A Technological Race with National Stakes, <https://www.linkedin.com/pulse/rise-sovereign-ai-technological-race-national-stakes-banafa-zxsjc>.

⁷Ahmed Banafa, The Rise of Sovereign AI: A Technological Race with National Stakes, <https://www.linkedin.com/pulse/rise-sovereign-ai-technological-race-national-stakes-banafa-zxsjc>; Yu Chen, AI Sovereignty: Navigating the Future of International AI Governance, <https://philarchive.org/rec/CHEASN-2>; Muath Alduhishy, Sovereign AI: What it is, and 6 strategic pillars for achieving it, <https://www.weforum.org/agenda/2024/04/sovereign-ai-what-is-ways-states-building/>; Luca Belli, Exploring the Key AI Sovereignty Enablers (KASE) of Brazil, towards an AI Sovereignty Stack, In Carnegie Endowment for International Peace. Digital Democracy Network Conference 2023 Essay Collection. (2023), pp.216-pp.221; Woods, Andrew Keane, 'Digital Sovereignty + Artificial Intelligence', in Anupam Chander, and Haochen Sun (eds), Data Sovereignty: From the Digital Silk Road to the Return of the State (New York, 2023; online edn, Oxford Academic, 14 Dec. 2023), <https://doi.org/10.1093/oso/9780197582794.003.0006>, accessed 14 Sept. 2024, p.125-p.129; Roberto Reale, Towards Sovereignty in AI: A 7-Tier Strategy for Europe's Technological Independence in Generative Artificial Intelligence, <https://futurium.ec.europa.eu/en/european-ai-alliance/blog/towards-sovereignty-ai-7-tier-strategy-europes-technological-independence-generative-artificial>.

⁸<https://www.forbes.com/sites/lancecliot/2024/11/20/all-that-vocal-clamor-for-ai-sovereignty-is-plainly-sensible-or-maybe-completely-off-the-charts/>.

符合人工智能发展的客观规律，还与我国以往的治理理念与制度设计一脉相承，从而进一步推动我国在数字时代的主导地位。本研究采用系统性研究的方法，在对人工智能主权进行多层次解读的基础上，通过解构人工智能主权的内部构造，并进一步分析其外部影响，为中国具体建构人工智能主权路径提供解决方案。

二、人工智能主权结构的本源分析

（一）人工智能主权脱胎于数字主权并富含丰富的内容

1648年《威斯特伐利亚公约》确立了国家主权原则。国家主权是指国家在国际关系中的独立地位和自主权，强调国家在国际事务中的平等参与和决策权。在数字时代，这一术语经常与关键资产相结合，从而形成数据主权、数字主权、技术主权、国防和军事战略自治、金融战略自治等。⁹数字主权本质上是一种控制权，它强调对数据、软件（例如人工智能）、标准和协议（例如5G、域名）、过程（例如云计算）、硬件（例如手机）、服务（例如社交媒体、电子商务）以及基础设施（例如电缆、卫星、智能城市）的控制。¹⁰数据主权主张国家或地区有权监管其境内的数据存储和流动。实际上，上述主权内容已经超越了传统国际法意义上的威斯特伐利亚式领土主权，形成了通过控制技术基础设施以实现对公司、个人甚至其他国家行使独特权力能力的功能性主权¹¹。而今，国家主权原则映射于人工智能，人工智能的发展可用于维护国家主权、安全和发展利益。人工智能主权是国家在其境内控制和监管人工智能的能力。在数据主权、数字主权与人工智能主权的系上(如图一所示)，数据主权构成受信赖的人工智能主权的基石，¹²其促进了对数据的所有权，提高了人工智能的透明度和问责效果。¹³人工智能主权是数字主权的延伸和重要体现，人工智能技术广泛应用于数字领域，对该主权的维护有助于提升数字主权的整体水平。

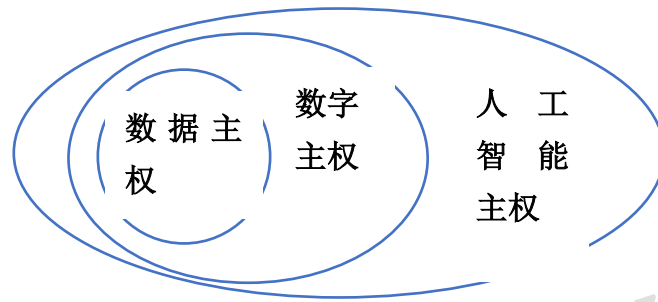
⁹ Timmers, P, Ethics of AI and Cybersecurity When Sovereignty is at Stake, *Minds & Machines* 29, 635–645 (2019), <https://doi.org/10.1007/s11023-019-09508-4>, p.636.

¹⁰ Luciano Floridi, The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU, 33 *PHILOSOPHY AND TECHNOLOGY* 369, 371 (2020). 还有学者认为，数字主权的定义应该更加泛化，包括国家通过互联网过滤技术和数据本地化要求来规范跨境数据流动的主权权力，以及规范言论活动（例如，打击虚假新闻）和技术的权力。

¹¹ 沈伟伟：《网络主权视角下的数据治理比较研究》，载《法学杂志》2024年第4期，第24页。

¹² Andrew Horton, Sovereign Data: Australia's AI Shield against Disinformation, <https://www.aspistrategist.org.au/sovereign-data-australias-ai-shield-against-disinformation/>.

¹³ Cuong Le, Amplifying AI Use Cases Through Data Sovereignty: A Strategic Approach to Global Innovation, <https://insideainews.com/2024/09/24/amplifying-ai-use-cases-through-data-sovereignty-a-strategic-approach-to-global-innovation/>.



图一 数据主权、数字主权与人工智能主权的关系

我国依次历经了互联网主权、信息主权、网络空间主权与数据主权阶段，尤其是，《中方关于网络主权的立场（China's Views on the Application of the Principle of Sovereignty in Cyberspace）》便强调网络主权是一国基于国家主权对本国境内的网络设施、网络主体、网络行为及相关网络数据和信息对内最高权和对外独立权。这一概念同样应用于人工智能领域。《全球人工智能治理倡议》尽管并未提出人工智能主权这一鲜明主张，但倡议内容反映了人工智能主权的不同层次。尤其是它重申尊重他国主权并严格遵守法律的必要性。进一步地，中国从全球性视角出发，关注发展中国家的人工智能主权，要求确保各国人工智能发展与治理的权利平等、机会平等、规则平等，防止数字鸿沟与智能滥用。¹⁴

世界经济论坛将人工智能主权定义为一个国家根据其地方政策或国家人工智能战略，利用本土人才建设本地化的人工智能的能力。¹⁵本文认为，它以保护国家主权、安全、经济竞争力和社会福祉为目的，¹⁶是一个国家根据自身的价值观、利益和法律法规对人工智能技术的开发、部署与治理保持控制和自主的能力，对人工智能核心内在元素数据、算法与硬件的主权。¹⁷通过发展国内人工智能能力并确保在全国范围内获得关键数据、技术、专业知识和基础设施，以减少对外国人工智能技术的依赖，保护国家免受潜在的供应链中断，从而加强数字主权和国家安全。¹⁸在此基础之上，各国可以参与全球人工智能的发展。¹⁹

人工智能主权的控制要素含有丰富的内容。这是因为，自主智能的出现具有

¹⁴ 张平：《人工智能治理规则的国际协同与竞争》，载《上海法治报》2023年12月1日B7版。

¹⁵ How Nations Can Build Sovereign AI and Homegrown Talent for Economic Competitiveness, <https://www.weforum.org/agenda/2024/07/sovereign-ai-talent-improve-economic-competitiveness/>.

¹⁶ Pablo Chavez, The Rising Tide of Sovereign AI, <https://siliconangle.com/2024/02/10/rising-tide-sovereign-ai/>.

¹⁷ Andrea Calderaro & Stella Blumfelde, Artificial Intelligence and EU Security: the False Promise of Digital Sovereignty, *European Security*, Volume 31, 2022, <https://www.tandfonline.com/doi/citedby/10.1080/09662839.2022.2101885?scroll=top&needAccess=true>. 有学者从消极角度来定义人工智能主权，人工智能主权指一个国家限制外界获取在其主权范围内设计、开发和部署人工智能的能力，包括对该人工智能行使控制权并排除其他国家干预的权力。

¹⁸ Muath Alduhishy, Sovereign AI: What It Is, and 6 Strategic Pillars for Achieving It, <https://www.weforum.org/agenda/2024/04/sovereign-ai-what-is-ways-states-building/>.

¹⁹ Yu Chen, AI Sovereignty: Navigating the Future of International AI Governance, <https://philarchive.org/rec/CHEASN-2>.

广泛的社会溢出效应，在政治、经济、社会等各个领域都会带来深刻变革。在实践中，国家在开展自身的人工智能主权时需要平衡创新与监管、数据主权与数据共享、经济自主与开放市场、国家安全与国际合作、人工智能发展与文化或伦理保护的关系等。

（二）人工智能主权是“主权人工智能”的国家表达

主权人工智能更多由行业推动，而非决策者。²⁰这是因为人工智能的发展轨迹是由市场驱动与技术决定的。企业已成为人工智能创新的主要驱动力。鉴于大型科技公司缺乏与公民签订社会契约的合法性来源，因而这一主权又称为准主权。²¹人工智能非国家行为者的权力和影响力正在挑战传统的国家权力概念。²²一方面，人工智能企业不仅凭借独特的技术优势内生出一些技术型权力，还通过辅助治理从国家手中另外获得委托或授予的部分权力。国家治理权力结构从国家单中心结构转变为多中心结构。²³由于人工智能技术的复杂性和商业模式的不透明性，使得监管机构无法制定具体规则，大型科技公司因而对人工智能施以控制。大型科技公司还通过游说的方式避免监管。²⁴结果是，全球人工智能垄断企业在本国和他国产生不公平竞争，使政府难以有效影响其行动；阻碍小公司访问大数据，限制创新，破坏人工智能的公益性质；对劳动者与消费者的权利造成实际影响等。欧盟委员会便对谷歌与微软提起跨司法管辖诉讼。

人工智能主权是“主权人工智能”的国家表达。人工智能主权尚未得到监管机构的官方推动。除了商业性主权人工智能之外，国家应根据本国价值观、民族文化、身份和能力，投资与监管人工智能。事实上，主权人工智能可在一定情形下实现人工智能主权的效应。大型企业通过在域外本地化运营、与当地公司成立合资企业，提供主权人工智能而有可能间接控制他国国家主权，并通过投资、就业、税收等渠道反哺母国，从而转化为国家主权之间的竞争。跨国企业采用嵌入式本地化策略实现技术渗透路径，例如，甲骨文和英伟达通过分别提供分布式云、人工智能基础设施、生成式人工智能服务和加速计算、生成式人工智能软件合作在全球范围内提供主权人工智能解决方案。²⁵这种技术部署往往附带数据本地化、

²⁰ Ryan Browne, Tech giants are investing in 'sovereign AI' to help Europe cut its dependence on the U.S., <https://www.cnn.com/2024/11/14/tech-firms-invest-in-sovereign-ai-to-cut-europe-dependence-on-us-tech.html>.

²¹ Huw Roberts, Digital sovereignty: A Normative Approach (January 18, 2024). Available at SSRN: <https://ssrn.com/abstract=4699167> or <http://dx.doi.org/10.2139/ssrn.4699167>.

²² Hazrat Usman, Bushra Nawaz, Saiqa Naseer, The Future of State Sovereignty in the Age of Artificial Intelligence, *Journal of Law & Social Studies (JLSS)* Volume 5, Issue 2, pp 142-152, www.advancelrf.org, pp.144-145.

²³ 梅立润：《技术置换权力：人工智能时代的国家治理权力结构变化》，载《武汉大学学报（哲学社会科学版）》2023年第1期，第46页。

²⁴ Roberts, Huw, Digital Sovereignty: A Normative Approach (January 18, 2024). Available at SSRN: <https://ssrn.com/abstract=4699167> or <http://dx.doi.org/10.2139/ssrn.4699167>.

²⁵ Justin Flitter, What is Sovereign AI and why do we think it should be a part of NZ's AI Strategy, <https://newzealand.ai/insights/what-is-sovereign-ai-and-why-do-we-think-it-should-be-a-part-of-nzs-ai-strategy>. See also NVIDIA, Oracle and NVIDIA to Deliver Sovereign AI Worldwide, <https://nvidianews.nvidia.com/news/oracle-nvidia-sovereign-ai>.

算法定制等条款，形成事实上的技术标准主导权。在资本控制网络方面，以合资企业为载体的资本运作模式（如英伟达与中东主权基金合作）构建起复杂的股权网络，人工智能研发中心通过交叉持股形成数字资本环流，使母国处于价值链顶端。微软在东南亚、非洲和拉丁美洲建立新的数据中心，微软 Azure 主权云方案中包含的合规框架，实质输出美式数据治理规则，形成规则制定权的隐形扩张。

（三）人工智能主权不同于人工智能民族主义

人工智能主权的竞争与相互依赖的双重属性决定了人工智能的非民族主义特性。在竞争属性项下，除了一般的本地化主权属性之外，还延伸出人工智能主权的域外属性，这体现在法律的域外适用、法律的域外影响等。然而，当此类主权措施越界时，他国有权采取同等反制权。人工智能主权还具有相互依赖属性，这就凸显了人工智能合作主权的重要性。换言之，人工智能主权是对传统主权的超越，这不仅体现在地域的超越，还有关系的超越。

相较于使用人工智能主权这一术语来描述政府对人工智能的支持，部分学者更倾向于用“人工智能民族主义”这一保护性语词来表示，这是因为，这一术语更能体现出人工智能这一新兴领域的排外性，包括并不限于人工智能供应链中的本地内容、高级芯片的出口管制等。²⁶本文认为这一表达不免有失偏颇，仅反映美国等国家的人工智能主权立场。

人工智能主权的战略自主手段与人工智能民族主义形成显著区分。除了自给自足与风险管理之外，战略伙伴关系与全球合作同样构成人工智能主权的重要内容。人工智能民族主义将危及到国家的发展，这是因为能够实现人工智能相对自给自足的国家较少。无论是缺乏资金投入的发展中国家，还是无法创建一个具有全球竞争力的人工智能部门，或者创建可以作为多种分析用途基础的人工智能模型的发达国家。²⁷风险管理方法则尽可能保持主权风险的可控性。加强联合或共享控制能力的战略伙伴关系方法同时增加了对彼此的依赖性。这也就解释了为何美国与日本在数据治理与人工智能合作方面亦步亦趋。全球合作对全球公地人工智能的培育从某种程度上加强了自身主权，²⁸通过共同制定国际规则 and 标准，促进国内人工智能主权的实现与发展。

综上所述，本文认为，人工智能主权是数据主权、数字主权的高级形态，主

²⁶ Aaronson, Susan, The Age of AI Nationalism and Its Effects (April 22, 2024). Available at SSRN: <https://ssrn.com/abstract=4803311> or <http://dx.doi.org/10.2139/ssrn.4803311>, p.6.

²⁷ Susan Aaronson, Why Australia Shouldn't Pursue Sovereign AI, <https://www.innovationaus.com/why-australia-shouldnt-pursue-sovereign-ai/>. 尽管澳大利亚在计算机视觉、深度学习、现场机器人、神经网络和机器学习等领域拥有丰富的人工智能专业知识。

²⁸ Timmers, P. (2024). Sovereignty in the Digital Age. In: Werthner, H., et al. Introduction to Digital Humanism. Springer, Cham. https://doi.org/10.1007/978-3-031-45304-5_36. See also Timmers, P. Ethics of AI and Cybersecurity When Sovereignty is at Stake. Minds & Machines 29, 635-645 (2019). <https://doi.org/10.1007/s11023-019-09508-4>, p.636.

权人工智能的国家表达，具有对内的控制权以及对外的独立自主权与合作权。²⁹这一主权的完满建立既需满足国家对内与对外控制的有效性，又需满足基于道德和伦理的合理性。

三、人工智能主权结构的内部构造

（一）基础层：数据主权+算力主权

首先，防御型主权与进攻型主权构成国家维护数据主权的两种方式。其一，从防御型主权来看，欧盟通过建立泛欧数据空间构建规模化数据资源池，³⁰以降低欧盟内部人工智能研究获取数据的成本。欧盟在数据本地化存储方面采取软性规制路径，GDPR 基于隐私保护对于跨境数据流动条款的严格限制导致个人数据出境通道并不顺畅。施雷姆斯 II 案进一步强化了该制度的实施刚性。这种防御型数据主权虽能确保人工智能训练数据的属地化治理，使算法模型内嵌欧洲伦理准则与文化价值，但其衍生的“隐私保护悖论”亦对数据要素的市场化配置形成制度性约束。新兴经济体在人工智能战略推进中呈现出差异化的数据治理模式，俄罗斯构建双重管控体系，一方面，俄罗斯禁止外国大语言模型获得大量俄罗斯用户数据；另一方面，俄罗斯正建立本国独立的政府数据集分级开放机制，赋予本土人工智能企业在政务数据开发利用方面的排他性优势。³¹

其二，从进攻型主权来看，在数字经济时代主权博弈的语境下，美国通过建构“政府-企业-公民”三元协同机制，系统性地推进其数据控制战略。第一，基于技术优势形成数据主权根基。美国人工智能企业凭借其主导地位，构建了事实上的数据本地化存储体系。第二，国内法律工具与国际协议的复合运用。《云法案》创设的“数据控制者推定管辖权”突破传统属地原则，使得境外数据亦可纳入美国司法管辖权。美国还通过签署宽松的跨境数据流动协议，构建数据跨境自由流动的制度洼地。第三，公私部门的数据攫取协同机制。美国科技企业以技术援助项目为名，通过在他国建立人工智能项目使大量数据最终回流至美国数据中心。2023 年《关于安全、可靠和可信人工智能的行政命令》正式将数据整合上升为国家安全战略。白宫人工智能数据中心基础设施工作组与美国能源部人工智能数据中心参与团队的创建将支持人工智能数据中心的发展。³²

²⁹ 对于此，我国学者赵骏对人工智能主权作如下定义，它在对内层面意味着主权国家对人工智能技术与实体的管辖权；在对外层面指国家享有自主决定和保护人工智能产业发展与规划而不受其他国家干扰之权利。赵骏、李婉贞：《人工智能对国际法的挑战及其应对》，《浙江大学学报（人文社会科学版）》2020 年第 2 期，第 10-25 页。人工智能主权涉及到国家如何监管人工智能训练数据的数据层，人工智能算法透明度的算法层，以及人工智能使用者的应用层。沈伟、赵尔雅：《数字经济背景下的人工智能国际法规制》，载《上海财经大学学报》2022 年第 5 期，第 132 页。

³⁰ Woods, Andrew. (2023). Digital Sovereignty + Artificial Intelligence. 10.1093/oso/9780197582794.003.0006. p.117.

³¹ 周逸清：《俄罗斯正在推进“AI 主权化”战略》，海国图智研究院，https://mp.weixin.qq.com/s/?__biz=MzIwMzYyMzQ5MQ==&mid=2247549813&idx=2&sn=0f00b3cbd840329b9ae11ed8bf0675a2&chksm=97cdc2a1b8586b6bcd27b00e86fe46add1309031bcb98694b17d71436e8d06224a8d99ca61&scene=27。

³² The White House, Readout of White House Roundtable on U.S. Leadership in AI Infrastructure,

其次，可信数据空间建构了以信任机制为基础的数据主权动态实现路径。其赋予个人、企业与国家对其管控的数据拥有排他性自决的能力，可以实际决定人工智能领域数据的开放、流通和使用水平。可信数据空间是基于共识规则、联接多方主体，实现数据资源共享共用、数据要素价值共创的应用生态。就全球实践范式而言，欧洲多主体协同推进数据空间建设，并且通过构筑生态推进全球扩张。美国依托先进技术优势，科技巨头企业，推进数据跨云互通，体现市场驱动型治理路径。我国以政府引导为主的可信数据空间³³应用探索刚刚起步，各地结合地方特色积极探索数据流通新模式。从国际合作层面来看，数据空间的发展将促进跨国数据流动和交易，通过参与国际数据空间的对话和合作，推动数据空间技术标准、运营规则和制度体系的全球协作。

最后，算力主权是人工智能主权基础层的重要支柱。算力主权是指一个国家或地区对其计算资源的自主控制权。在算力维度的博弈方面，中国通过东数西算国家工程实现算力总量跃居全球第二，华为与中国移动联合研发的星云 AI 存储系统可支持千亿级参数模型训练，标志着国产算力体系的技术突破。不过，技术依赖困境仍然存在，中国在计算芯片和存储芯片，以及人工智能算法框架和数据库使用方面，严重受制于英伟达、三星、美光等外国厂商。³⁴为巩固美国在先进制程半导体供应链方面的显著优势地位，遏制中国获取发展人工智能系统所需要的各项算力资源，美国对华成熟制程芯片发起 301 调查。欧盟通过 GAIA-X³⁵打造主权云基础设施，Scaleway 公司研发的人工智能超级计算机试图重塑欧洲算力格局，不过，其算力自主化进程仍滞后于中美技术代差。

技术层：技术主权

首先，数据可通过算法转化为有价值的信息和决策。算法基于海量数据运算配置社会资源，直接作为行为规范影响人的行为，辅助甚至取代公权力决策，从而发展成为一种新兴的技术权力。³⁶这种技术权力已突破传统工具属性，逐步演化为具有规范效力的“算法利维坦”。该权力的扩张引发四重制度性风险：其一，公权力领域面临算法嵌入导致的决策自主性侵蚀与权力结构嬗变；其二，商业领域因算法垄断引发市场秩序异化；其三，个人面临主体性消解与权利减损的双重危机；其四，法律体系遭遇“代码即规则”的范式解构挑战。各国纷纷出台算法监管政策，关注算法透明度、可问责性和可解释性，以应对算法偏见、隐私泄露等问题。典型如南非《人工智能国家战略》，其通过立法确立数据价值链全周期监

<https://www.whitehouse.gov/briefing-room/statements-releases/2024/09/12/readout-of-white-house-roundtable-on-u-s-leadership-in-ai-infrastructure/>。此外，英国也主张建立主权人工智能数据中心。

³³ 企业可信数据空间、行业可信数据空间、城市可信数据空间、个人可信数据空间、跨境可信数据空间构成我国的可信数据空间。

³⁴ 《中国未来 50 年产业发展趋势白皮书》。

³⁵ Gaia-X 的七项指导原则是：欧洲数据保护；开放性和透明度；真实性和信任；数字主权和自决权；自由市场准入和欧洲价值创造；模块化和互操作性；以及用户友好性。

³⁶ 张凌寒：《算法权力的兴起、异化及法律规制》，载《法商研究》2019 年第 4 期，第 63 页。

管机制，强调以本土化数据训练算法模型，确保人工智能发展契合非洲特有的历史传统、价值伦理与文化基因。这种基于地缘文化特征的算法主权实践，为全球南方国家提供了技术治理的差异化路径。

其次，人工智能从算法到大模型正引领新的技术变革。截止 2024 年 7 月，全球人工智能大模型数量为 1328 个，美国占比 44%，中国占比 36%，不过在底层技术上，中国还相对薄弱。³⁷目前，大模型已经渗透至金融领域与医疗领域，并且催生出新的商业模式和产业生态。在此进程中，技术标准制定权争夺与数字殖民主义风险形成双重张力。当前大多数前沿模型都是根据西方数据进行训练的，导致非西方文明面临系统性表征危机。对此，全球南方国家正通过技术主权的制度实践展开战略回应。俄罗斯、南非、巴西、印度、越南等国家正在开发反映当地价值观的本土人工智能解决方案。³⁸这一做法挑战了科技巨头的数字殖民主义，并引发关于人工智能发展中的多样性和包容性的重要对话。³⁹积极技术主权与消极技术主权成为南方国家两种不同的实现方式。阿联酋已成为中东人工智能创新的中心。⁴⁰阿联酋不仅拥有发展人工智能的大量资金，还拥有丰富的人工智能发展能源。⁴¹作为消极主权的代表，俄罗斯国内已禁止对于西方国家人工智能模型的访问，发展对标 Chat GPT 等外国大模型的本土替代品，以减少对外国技术的依赖。从制度演进趋势分析，发展中国家正经历从消极防御到积极主张的范式转换。⁴²尊重和增强多语言和文化遗产成为全球南方国家发展人工智能主权的关键契机。语言多样性是印度人工智能的优势。⁴³实际上，印度开发本土大规模模型并不希望与美国和中国开发的模型直接竞争，而是利用印度语言的利基市场。⁴⁴目前，阿联酋与沙特阿拉伯致力于在本地数据集上训练阿拉伯语模型。

³⁷ 中国信息通信研究院：《2024 全球数字经济白皮书》，<http://www.caict.ac.cn/kxyj/qwfb/bps/202501/P020250116675154243240.pdf>。

³⁸ “AI made in Africa for Africa”指导非洲国家构建自身的人工智能主权，包括加纳、肯尼亚、卢旺达、乌干达和南非在内的许多非洲国家已经制定并实施了采用数字技术和制定数字转型道路的国家战略。Arthur Gwagwa and Beverley Townsend, Re-imagining Africa's Sovereignty in a Digitally Interdependent World, <https://www.globalpolicyjournal.com/blog/10/05/2023/re-imagining-africas-sovereignty-digitally-interdependent-world>。

³⁹ Nazareen Ebrahim, AI Sovereignty and Ethics: The Narrative of 2024. What Does 2025 Look Like?, <https://www.itweb.co.za/article/ai-sovereignty-and-ethics-the-narrative-of-2024-what-does-2025-look-like/nWJad7bNdQk7bjO1>。

⁴⁰ Nell Lewis and Rym Bendimerad, Why These Gulf States Want to Be AI Superpowers, <https://edition.cnn.com/2024/09/16/middleeast/middle-east-artificial-intelligence-spc/index.html>。普华永道预计到 2030 年，该地区的人工智能市场将达到 3200 亿美元。

⁴¹ <https://www.capitalbrief.com/briefing/abu-dhabi-unveils-ai-investment-firm-aims-to-surpass-us100b-aum-bdaa9571-a8f2-4773-ab87-0a13c057a452/>。阿联酋在 2024 年 3 月提出考虑提供 1000 亿美元成立针对 AI 的主权投资机构，旨在与中国和美国竞争。

⁴² 李艳华：《发展中国家的数字发展权：跨境数据流动治理规则何去何从》，载《人权》2023 年第 82-106 页。

⁴³ Ola 推出了 Krutrim，这是一个多语言人工智能模型系列，被誉为印度第一个全栈人工智能解决方案。

⁴⁴ Jyoti Panday and Mila T Samdub, Promises and Pitfalls of India's AI Industrial Policy, <https://ainowinstitute.org/publication/analyzing-indias-ai-industrial-policy>。

规范层：规制主权

首先，从规制主权规制目标与规制方式来看，欧盟人工智能主权的本质是强法律监管下欧盟价值观的全球化，⁴⁵从而将欧盟从泥泞的人工智能竞赛中解救出来，使其更有效应数字社会中的挑战。“负责任、以人为本、值得信赖的人工智能承诺”构成欧盟人工智能法规制的核心。欧盟的人工智能方法优先考虑数据保护、透明度和问责制，确保人工智能的发展与人权和社会价值观相一致。尤其是全球首个具有约束力的横向人工智能法规《人工智能法案》将人工智能系统划分为禁止、高风险、有限风险与最低风险四类，根据类别实施不同的监管措施。借用人工智能模板推动全球范围内人工智能监管规则的统一和标准化已经形成欧盟传统的立法逻辑。美国十分重视科技产业的发展需求，以提升科技创新能力为圭臬，适时适度的人工智能监管将有利于人工智能主权下技术与监管的有效平衡。美国国家主权与州主权在处理人工智能监管的方式上有所不同，但总体上以市场监管为中心。美国人工智能前期总体实行轻监管的风险控制路径，各部门制定一系列人工智能战略文件，多侧重于原则性要求，并秉持自愿性原则。伴随着生成式人工智能带来的过高安全风险，美国近年来的监管出现了硬监管与审慎监管的趋势，确定人工智能符合伦理标准，保护公民的隐私和权利，同时避免算法歧视和数据滥用。⁴⁶但总体上，美国在人工智能治理方面强调弱监管和软治理，例如，《前沿人工智能模型的安全与保障创新法案》要求披露其训练前沿模型的关键信息，设置严重损害为责任界限，对模型开源后的风险承担责任，后因过于严格与削弱人工智能创新而遭致否决。我国人工智能治理体系坚持发展安全并重的规制路径，呈现三个制度特征：首先，坚持总体国家安全观指导下的立法架构，《生成式人工智能服务管理暂行办法》首创“训练数据合规—内容安全治理—应用场景监管”的全链条规制模式；其次，构建算法治理规范群，通过《互联网信息服务算法推荐管理规定》等规章形成协同规制体系；最后，发展敏捷治理方法论，在自动驾驶、深度合成等领域建立场景化监管沙盒。这种促进创新与规制风险并重的立法取向，体现了新质生产力发展要求的制度回应。

其次，从规制主权规则的软硬来看，又可以分为硬性规则与软性规则。人工智能所涉多元价值、所处发展阶段与技术要求决定人工智能监管要求一种软法和硬法相辅相成、各有所长的混合法律模式，以全面回应公共治理的需求。硬法是指有约束力的法律，体现了基于主权意志的制度刚性。以欧盟为例，《人工智能法案》、GDPR、《数据法案》、《数字服务法》、《数字市场法》等文件形成了多层

⁴⁵ Daniel Mügge, Mügge, EU AI sovereignty: for whom, to what end, and to whose benefit? *Journal of European Public Policy*, 31(8), 2200–2225. <https://doi.org/10.1080/13501763.2024.2318475>.

⁴⁶ 《关于安全、可靠、值得信赖地开发和使用权人工智能的行政命令》采用了基于风险的监管原则，对生成式 AI 引入了对基础模型的披露和测试要求。州层面的立法相对容易通过，以 Open AI、Meta 等人工智能公司所在地加州为例，《加州数字内容溯源标准法案》《加州人工智能训练数据透明度法案》对人工智能合成图像与视频的元数据中嵌入来源水印、发布有关训练数据文档的要求与人工智能技术本身同步发展。

次、全方面的人工智能保护硬规则体系。⁴⁷软性规则“软性”并非指法律义务的力度，而是在于它缺乏法律约束力，⁴⁸其构成一种不完全的规范，⁴⁹但立足于治理弹性的规范创新。在技术伦理维度，《欧盟可信人工智能伦理指南》确立七项基本原则，通过伦理影响评估机制实现价值引导；在标准建构层面，我国《国家人工智能产业综合标准化体系建设指南（2024 版）》设定算法透明度等 42 项技术指标，形成标准-认证-应用的治理闭环。面对技术主权缺失的结构性困境，全球南方国家通过《利雅得人工智能伦理宪章》等区域软法，将伊斯兰文明价值观嵌入技术治理体系，开创了伦理规范本土化的制度创新路径。全球南方国家还积极推动国际软法的形成，国际电信联盟《人工智能就绪——迈向标准化就绪框架的分析》在沙特阿拉伯的支持下，建立评估人工智能能力和战略的全球标准以调整技术主权的方向。⁵⁰

最后，从规制主权的构成来看，单一主权与合作主权⁵¹或相互依赖的主权构成人工智能主权的规制形式。美国的规制主权遵循人工智能民族主义路径，其本质是技术霸权主导的单边主义范式，贯彻美国优先主权。除了对于本国数据基础架构的控制，美国还通过多边与盟友合作⁵²遏制竞争对手；单边出口管制、监管大型跨国人工智能公司与外交胁迫来强化自身人工智能主权。⁵³通常情况下，各国在本国人工智能主权的基础之上，就此嵌入到全球架构下的相互依赖主权。各方对自己部分的主权放弃，最终在国际空间延伸人工智能主权。例如，欧盟通过其庞大的内部市场和严密的制度架构，单方面影响全球监管遵循全部或部分欧盟的设计逻辑，由此产生人工智能监管方面的布鲁塞尔效应。由于《人工智能法案》遵循欧盟产品安全与责任立法的范畴，为欧盟的基本权利提供了有限的保护。欧洲委员会通过具有法律约束力的开放式《人工智能、人权、民主和法治框架公约》⁵⁴对人权保护进行补充，该公约确立了七大原则与可采用的人工智能风险和影响

⁴⁷ 《人工智能法案》对 GDPR 进行补充，《数字服务法》与《数字市场法》则同时规范大型数字公司的人工智能系统。

⁴⁸ Michael Hahn, Interesting Times: Soft Law in International Economic Governance, in Manjiao Chi, Marc Bungenberg, Andrea K. Bjorklund eds., *Asian Yearbook of International Economic Law*, Berlin: Springer 2022.

⁴⁹ Case C-16/16 P, *Kingdom of Belgium v European Commission*, 20 February 2018, Opinion of Advocate General Bobek, para.86.

⁵⁰ AI for Good, Are you AI ready? A New AI Standards Framework Has Just Launched, <https://aiforgood.itu.int/are-you-ai-ready-a-new-ai-standards-framework-has-just-launched/#:~:text=The%20framework%20offers%20a%20comprehensive,their%20AI%20capabilities%20and%20strategies>.

⁵¹ 黄志雄教授等首提出网络空间的合作主权观，其以领土主权为基础，尊重各国主权独立与自主，在传统平等观上强化网络空间治理与合作的主权责任，实现主权在网络空间适用的权责统一性，以及网络空间的自由、安全、发展与共享等基本价值。参见黄志雄、孙芸芸：《以合作主权观应对主权在网络空间适用的挑战》，载《边界与海洋研究》2023 年第 5 期。

⁵² 美国是经合组织、全球人工智能伙伴关系、七国集团和二十国集团的成员国。

⁵³ Roberts, H., Hine, E., Floridi, L. (2023). Digital Sovereignty, Digital Expansionism, and the Prospects for Global AI Governance. In: Timoteo, M., Verri, B., Nanni, R. (eds) *Quo Vadis, Sovereignty?*. Philosophical Studies Series, vol 154. Springer, Cham. https://doi.org/10.1007/978-3-031-41566-1_4.

⁵⁴ 目前已经有 11 个非欧盟成员国签署。该公约规制主体涵盖了公共当局与私人行为者；并且该公约不适用于保护国家安全利益的活动。

管理框架，要求签署国对人工智能产生的任何有害和歧视性结果负责，受害者享有法律追索权。为减轻人工智能的不利影响，公约并没有像《人工智能法案》对风险采用分层的方法，而是将酌处权留给缔约国。加强彼此合作来巩固本国主权成为全球南方国家摆脱北方控制的另一种手段。如，俄罗斯不断推进与非洲国家和金砖国家的人工智能合作。⁵⁵

四、人工智能主权结构的外部效应

（一）基础层与技术层：人工智能主权加剧全球安全困境

尽管人工智能在加强国家主权方面发挥着重要的作用，使国家能够保护其数字基础设施和数据免受威胁。但人工智能同时还在侵蚀国家主权，这是因为使用自主武器和网络攻击等人工智能技术可以绕过国家边界。因此，人工智能主权概念的双重性解释在评价不同国家的人工智能主权行为时可能出现双重标准，并使国家陷入安全困境，从而阻碍全球人工智能治理框架的构建。

首先，人工智能主权引发国家安全风险。各国将人工智能与国家安全战略相结合。自主网络攻击可能使人类在决策中缺乏控制。⁵⁶美国是第一批与谷歌开展人工智能国防军事项目“算法战争”的国家之一。2023年8月，美国国防部成立工作组，以研究将生成式AI用于军事任务的可能性。⁵⁷乌克兰战争成为科技巨头人工智能战争实验室，包括Palantir、Microsoft、Amazon、Google、Clearview AI等公司已与乌克兰武装部队合作，为乌克兰提供先进技术和支持，以对抗俄罗斯的侵略。⁵⁸这种科技企业和军事机构的新型协作模式，不仅加速了人工智能武器的实战化进程，更凸显了私营部门在当代战争形态演变中的战略地位。这一态势预示着未来国际安全格局将面临更为复杂的技术伦理挑战和法律规制困境。

其次，人工智能主权引发经济安全风险。人工智能供应链安全风险构成经济安全的重要一环。随着地缘政治的紧张趋势和对国家经济安全担忧的日益加剧，美国通过上游制裁我国人工智能产业链的底层支柱，中上游配合对华芯片制裁，下游制裁应用类公司，尤其重点打击与政府关系密切的以及军民两用的企业。⁵⁹

⁵⁵ Konstantin Pantsev, Technological Cooperation Between Russia and African Countries, <https://valdaiclub.com/a/highlights/technological-cooperation-between-russia-and-africa/>. 例如，俄罗斯在卢旺达开设第一个开放源代码和数据管理实践的透明度中心。International: Russia and China reinforce Cooperation for AI Development, <https://www.dataguidance.com/news/international-russia-and-china-reinforce-cooperation-ai>. 俄罗斯人工智能中心通过成立合作工作组、分享道德监管与工业应用的最佳实践、推动金砖国家在人工智能领域的发展潜力等形式与中国加强人工智能合作。

⁵⁶ Yogita Upadhyay, & Dr. Rituja Sharma. (2023), Cybersecurity And Legal Considerations In AI Applications For National Security. Educational Administration: Theory and Practice, 29(3), 474–480.

⁵⁷ Josh Luckenbaugh, New Pentagon Task Force Exploring Generative AI, National Defense, October 25, 2023, <https://www.nationaldefensemagazine.org/articles/2023/10/25/new-pentagon-task-force-exploring-generative-ai>.

⁵⁸ Vera Bergengruen, How Tech Giants Turned Ukraine Into an AI War Lab, Time, February 8, 2024, <https://time.com/6691662/ai-ukraine-war-palantir/>.

⁵⁹ 凌翔，杨茗薇：《ChatGPT 爆火背后：总体国家安全观视角下人工智能诱发多领域安全风险刍议》，载《信息安全与通信秘密》，2023年第6期。

中国同样对算法进行控制，并正在减少对外国技术的依赖。对于生成式人工智能而言，鉴于其可能存在深度伪造、错误信息和版权问题，保持对生成式人工智能模型相关的整个供应链的绝对控制变得至关重要。

再者，人工智能主权引发数据安全风险。人工智能主权能够确保国家对其数据拥有控制权，包括遵守国家数据保护法律和法规，保护个人信息、知识产权和其他敏感数据免受未经授权的访问或利用的能力。但同时意味着，政府可以加强监控此类数据与跨境数据传输。事实上，大部分国家已经制定数据本地化的法律或其他规范，生成式人工智能的中心化技术与人工智能主权的独立性特征将进一步加剧数据的本地化存储趋势，导致数据安全困境。

最后，人工智能主权可能引发伦理安全风险。将伦理和道德框架纳入人工智能决策具有挑战性，这是因为它需要人工智能能够理解和应用各种场景下的道德原则。生成式国内人工智能算法与数据收集可能存在潜在的社会偏见、歧视或不平等，一些国家的透明度机制与问责机制不明晰。人工智能主权强调国家依据自身的道德与文化来确保人工智能的开发和应用，但现实是英语模型通过这一正当性理由而对他国形成偏见，更加放大了排除不同的观点并强化局部偏见的可能性，这可能会侵蚀文化和语言多样性。

（二）基础层与技术层：人工智能主权扩大全球南北方数字智能鸿沟

一方面，全球北方通过自身发展与战略合作加强人工智能主权能力。全球北方通过专利、许可和人工智能产品和服务的出口，对人工智能技术和算法施以控制以促进其经济的增长。此外，北方国家针对外国人工智能人才出台富有吸引力的移民政策，又或者对本国高科技人才的对外流动进行限制。根据美国联邦贸易委员会的调查，全球大约有 20 家公司拥有云基础设施、计算能力、资本和大量数据来开发和部署生成式人工智能。⁶⁰欧洲和亚洲国家正投资数十亿美元用于创建独立的大规模人工智能系统。例如，ABCI 3.0 是日本大型开放人工智能计算基础设施的最新迭代，旨在推进人工智能研发。这项合作强调了日本致力于提高其人工智能能力和加强其技术独立性。⁶¹英伟达还与电信提供商软银公司合作，为 5G 和 6G 应用程序构建生成式人工智能平台，以及分布式人工智能工厂网络。新加坡政府正在与英伟达合作，升级其国家超级计算机中心。通信服务提供商 Singtel 在东南亚各地建设节能人工智能工厂，开发基于东南亚语言数据的基础模型。

另一方面，全球北方通过自由跨境数据流动规则不断获得全球南方的数据，并严格控制人工智能基础设施芯片及其他人工智能技术进出口。加之，建立和维护独立的人工智能基础设施、人才库和数据资源，构建和训练新的生成式人工智

⁶⁰ <https://openai.com/research/video-generation-models-as-world-simulators>.

⁶¹ Dion Harris, Japan Enhances AI Sovereignty With Advanced ABCI 3.0 Supercomputer, <https://blogs.nvidia.com/blog/abci-aist/>.

能需要大量投资，从而阻碍发展中国家参与这一领域的的能力。而严重依赖他国的人工智能技术将破坏一国的经济独立性和全球竞争力，从而加剧全球南北方的数字鸿沟。例如，非洲国家人工智能训练数据的知识产权大部分归属亚马逊，这种“数字资源虹吸”现象导致技术依赖加深。根据 Gartner 预测，在未来五年内，拥有人工智能主权战略的国家和没有人工智能主权战略的国家之间的差距将大幅扩大。⁶²目前，部门南方国家已经开始通过加大人工智能要素投入，限制人工智能要素流出等方式以加强国际竞争。例如，越南、墨西哥等南方国家在欧美国家的影响下，也制定相应政策限制数据与敏感技术的流出。

（三）规范层：人工智能主权加剧全球治理碎片化

尽管在人工智能监管原则方面，多个国家同意透明度、可追溯性与可解释性；数据保护、隐私与数据安全、识别风险、禁止歧视与偏见等内容。然而，由于各国经济、社会、法律与政治状况存在差异，因而，各国受人工智能影响的方式有所不同，从而导致人工智能监管方式与监管内容有所不同。⁶³

首先，监管方式不同加剧全球治理碎片化。欧盟采取全面、横向和基于风险的监管方式。欧盟数据领域监管路径的布鲁塞尔效应在人工智能方面的影响力——即针对人工智能制定国内法规，并依靠该法规直接或间接影响其他司法管辖区的人工智能标准——还有待考察。这是因为，欧盟的风险监管过于追求形式的统一性，将不同行业和领域的人工智能风险统一划分为不同级别。⁶⁴此外，鉴于风险的开放概念，执法机构将面临复杂性挑战。该法案还有可能无法规制新出现的人工智能技术。因而，目前来看，部分国家因技术开发和监管能力较为有限，尚未采用欧盟的人工智能横向方案。美国联邦层面采取的是去中心化、针对具体行业、非强制性的监管方式。州层面的不同规则加剧了企业的合规负担。鉴于美国科技企业的全球领导力，这种规则效应与标准可能借此扩展至其他国家。中国则采取分散、垂直、迭代的监管方式。这表现在对技术的特定应用或表现形式进行针对性监管，⁶⁵因而具有较强的实用性和可操作性，这一监管路径可为新兴市场国家提供思路。一些国家则采用现有制度或非约束性方法以促进人工智能创新。或者采用欧盟的综合式立法，如巴西最新的人工智能法律提案采取欧盟基于风险的方法，禁止风险过大的人工智能系统，⁶⁶对人工智能系统供应商强制自我评估，

⁶² CLOVA, Sovereign AI, A Crucial Challenge of the AI Era, <https://clova.ai/en/tech-blog/en-sovereign-ai-a-crucial-challenge-of-the-ai-era>.

⁶³ Smuha, Nathalie A., From a 'Race to AI' to a 'Race to AI Regulation' - Regulatory Competition for Artificial Intelligence (November 1, 2019). Published in Law, Innovation and Technology, Vol. 13, Iss. 1, 2021, Available at SSRN: <https://ssrn.com/abstract=3501410> or <http://dx.doi.org/10.2139/ssrn.3501410>.

⁶⁴ 丁晓东：《全球比较下的我国人工智能立法》，载《比较法研究》2024年第4期。

⁶⁵ 中国金融四十人论坛：《人工智能监管：全球进展、对比与展望》，<https://mp.weixin.qq.com/s/EXSzaJ6QCbwhT1IFDdLsVA>。

⁶⁶ Luca Belli, Exploring the Key AI Sovereignty Enablers (KASE) of Brazil, towards an AI Sovereignty Stack, In Carnegie Endowment for International Peace. Digital Democracy Network Conference 2023 Essay Collection. (2023), pp.216-pp.221.

包括信用评级、法律决定、医疗诊断、自动驾驶汽车等领域。⁶⁷或者采用中国的分散式立法，沙特阿拉伯《深度伪造技术指导原则》要求深度伪造技术符合道德标准。或者通过战略文件指导人工智能的发展，如俄罗斯《2030 年国家人工智能发展战略》。跨司法管辖区的不同监管方法对互操作性构成挑战，并增加了国际公司的合规成本。

其次，监管具体内容不同加剧全球治理碎片。尽管预训练阶段的问题认识相对已经比较明确，各国措施正在趋同化，各国对于人工智能智能技术应用的风险评估以及虚假信息的识别机制尚未达成一致；在对于本地语言的生成式人工智能的监管也有所不同；生成合成内容标识的用户参与程度也不同。⁶⁸人工智能监管在法律上的碎片化所带来的技术障碍将影响全球性人工智能合作。

（四）规范层：人工智能主权可能扭曲人工智能贸易

首先，基于合法公共政策的数据本地化与数据跨境流动限制具有一定的合理性，然而，使人工智能跨境访问和分析数据变得更加困难，从而限制人工智能的创新和经济增长，一些国家则将此视为对本土人工智能的激励。美国 2024 年《贸易壁垒报告》中便将中国、越南、俄罗斯等 7 个国家的数据本地化措施视为歧视性措施。⁶⁹而美国为防止中国企业收集敏感数据，拟禁止智能网联汽车使用中国与俄罗斯软硬件，同样构成歧视性措施。

其次，对于数据来源披露而言，《欧盟人工智能法案》对于高风险人工智能开发者的数据披露义务可能被视为歧视性，因为高风险人工智能提供商主要为中国与美国，从而形成对中美间接限制。

再者，国家通过建立数据基础设施⁷⁰与数据共享为企业与研究机构提供训练人工智能的高质量数据，并且限制甚至排除外国企业的参与，这可能构成歧视与不公平的待遇，因为这将为国内人工智能行业创造间接补贴，根据世贸组织补贴规则，这些补贴可能构成可诉性补贴。⁷¹

最后，在对于人工智能基础设施的限制方面，美国以国家安全为由实施限制人工智能贸易的一系列措施，其中重点包括美国的芯片出口管制措施，包括并不限于护栏条款、直接产品规则、需通知先进计算的许可证例外、逐案审查等政策。这是因为人工智能芯片作为智能算力产业的基础硬件，为智能算法与应用提供必要的计算能力，从而成为美国的重点出口管制对象。2024 年 5 月 9 日，美国还

⁶⁷ Gupshup, Data Sovereignty in the AI Age: Gupshup's Innovative Approach, <https://www.gupshup.io/resources/blog/data-sovereignty-in-the-ai-age-gupshups-innovative-approach>.

⁶⁸ 例如，中国侧重在全国范围内建立严格的标识规范，美国强调内容的透明度与用户的参与。

⁶⁹ USTR Releases 2024 National Trade Estimate Report on Foreign Trade Barriers, March 29, 2024, <https://ustr.gov/about-us/policy-offices/press-office/press-releases/2024/march/ustr-releases-2024-national-trade-estimate-report-foreign-trade-barriers>.

⁷⁰ 英国同样将数据中心列为关键国家基础设施。

⁷¹ Shang KK, Du RR. Disciplining Artificial Intelligence Policies: World Trade Organization Law as a Sword and a Shield. In: Peng S, Lin C-F, Streinz T, eds. Artificial Intelligence and International Economic Law: Disruption, Regulation, and Reconfiguration. Cambridge University Press; 2021:274-292.

通过《加强海外关键出口限制的国家框架法案》，目的是保护美国人工智能及其他支持技术免受外国敌手利用。近期，美国发布“临时最终规则”加强对量子计算与先进半导体制造技术的出口管制，从而限制我国人工智能模型的发展。这一出口管制措施还通过联合荷兰与日本来实现。上述立法的实体要求与程序限制违背了最惠国待遇原则和国民待遇原则，破坏了多边贸易体制。

由此可见，为维护人工智能主权，部分国家选择多种措施以确保数据安全、数据透明度以及人工智能基础设施安全与保障，这些措施包括并不限于数据本地化、数据跨境流动限制措施、数据共享措施、数据来源追溯、芯片出口管制、云基础设施与大型计算补贴等。⁷²然而，上述单边措施可能构成贸易保护行为，违反国际经贸规则。平等对本地和外国企业及产品的非歧视原则、自由化原则、可预测和透明度原则、促进公平竞争构成国际贸易的核心规则。⁷³事实上，国家往往以贸易协定中的隐私保护、国家安全例外为由，为各种限制措施辩护，然而，上述部分国家措施已经超过了例外豁免中的比例性原则要求，这就要求国家措施应考虑与国际经贸规则的相适性。

五、我国人工智能主权结构的建构路径

（一）嵌入包容性的价值观

人工智能固有主权与合作主权构成包容性价值观的嵌入土壤。部分学者将人工智能主权等同于人工智能国家主义，而忽视国家作为全球人工智能治理中的固有成员身份及其积极参与国际社会的权利、义务与责任。鉴于人工智能技术的不确定性、跨国性与全球风险性，重视人工智能主权的延伸性、功能性主权合作主权便显得至关重要。因而，我国主张的人工智能主权应该在实现主权独立的基础上，将合作主权作为巩固本国主权与拓展涉外主权的工具，通过一定的主权让渡以获得更大的自主权。包容性理论是一个综合性的理论框架，⁷⁴包容性治理理念有利于弥补人工智能主权对外排他性的部分缺陷，提供对内控制与对外合作的发展与安全并重的理论基础。

人本主义构成包容性价值观的本位价值。现阶段中国人工智能立法应彰显人本主义，贯彻我国《全球人工智能治理倡议》中以人为本的人工智能治理主张。人本价值意味着人工智能规范体系不得摆脱或超越人的重要性而实现其目标，⁷⁵这是人工智能发展的前提。这一理念已经得到国家与国际组织的认可，联合国《以

⁷² Aaronson, Susan, The Age of AI Nationalism and Its Effects (April 22, 2024). Available at SSRN: <https://ssrn.com/abstract=4803311> or <http://dx.doi.org/10.2139/ssrn.4803311>.

⁷³ World Economic Forum, ChatWTO: An Analysis of Generative Artificial Intelligence and International Trade, September 2024, https://www3.weforum.org/docs/WEF_An_Analysis_of_Generative_Artificial_Intelligence_and_International_Trade_2024.pdf, p.7.

⁷⁴ 郭小东：《大模型时代全球人工智能治理的挑战与中国方案》，载《科学学研究》2024年，第6页。

⁷⁵ 李将：《论多边主义价值性的基本构成——以包容和人本价值为视角》，载《太平洋学报》2022年第12期。

人为本的人工智能治理》报告提出了解决人工智能相关风险的蓝图，以促进发展和保护所有人权。

社会正义理论与可持续发展理论构成包容性价值观的内在要求。社会正义理论构成智能向善、平等、公平与非歧视性、伦理先行等原则的基础；可持续发展理论则强调人工智能技术在实现可持续发展中的重要作用，这有利于实现碳中和、优化资源分配、支持人机互动技术设计与产品发展、推动全球治理等。上述理论符合人工智能主权国家的共同追求。

（二）加强人工智能主权基础层、技术层与规范层的构建与联动

首先，基础层与技术层构成我国人工智能主权控制的基础。

其一，人工智能主权基础设施必须由政府主导，私营科技行业应与政府进行合作，建立人工智能主权基础设施。中国应重点加强底层硬件基础设施芯片的制造，确保最大的供应链安全和弹性。实际上，对于美国的芯片出口管制，中国应在现有芯片库存的基础上，迅速发展自己的供应，而不依赖美国公司。中国人工智能公司还通过云服务提供商访问美国高端芯片，以规避美国出口管制。或者将人工智能芯片与行业特定软件和预培训模型相结合，而无需采用英伟达等先进芯片。⁷⁶然而，上述人工智能公司的做法只能短期解决现实中的难题。因此，从长期来看，中国应加大芯片自主制造能力。

其二，从数据主权的内容来看，数据主权是人工智能发展的重要保障，数据主权能够使我国根据国家的独特需求和环境构建人工智能模型。我国现有的数据规制主要集中在网络安全、个人隐私保护等边境后领域，数据权属确定困难、数据价值评估困难、数据互联互通困难等问题仍不同程度存在。改善数据作为战略经济资源的使用可以大幅扩大训练人工智能模型的数据可用性。中国尽管在数据数量上具有领先优势，但在数据的质量与多样性方面还有很大的提升空间，数据标准化困难，缺少多样场景且有标注的数据，从而影响数据的可用性。为此，在人工智能数据收集和整合阶段，数据必须标准化、清理和去标识，以确保其可用性并保护隐私。确保数据质量、安全性和道德使用，尤其是应确保人工智能训练数据的高质量供给，建构人工智能训练数据公共服务平台，建立公平性的训练数据供给生态保障体系。鉴于跨境数据规则既规范创建生成式人工智能所需的输入，也规范了生成式人工智能模型和算法可能产生的输出，为此，在跨境数据传输规则制定方面，应促进数据安全流动，明晰人工智能重要数据目录，建立健全与负面清单相适应的数据贸易监管体系。

其三，从数据主权的实现载体来看，为构建数据要素社会化重用秩序，需要在尊重数据持有者权的基础上，建立可信的数据流通体制和机制。⁷⁷而可信数据

⁷⁶ Hodan Omaarm, How Innovative Is China in AI?, <https://itif.org/publications/2024/08/26/how-innovative-is-china-in-ai/>.

⁷⁷ 高富平：《可信数据流通制度论——治理范式经济秩序的形成》，载《交大法学》2024年第5期。

空间赋予数据空间可信结构，是让数据在不同主体之间的可信流通的载体。国家数据局发布的《可信数据空间发展行动计划（2024-2028年）》提出构建全国统一可信数据空间战略目标。建设我国自主可控的可信数据空间，既是内循环产业链高质量发展的有力推动，也是外循环产品高质量出海的积极探索。为推动可信数据空间加速发展，可在重点行业与领域鼓励多元主体加快数据开放与共享，同时加快公共数据授权运营，应注重标准化工作，统一技术标准，确保数据的安全可控流通。此外，还应积极引入第三方服务方，精准匹配不同需求，实现数据资源的交互和价值共创。

其次，规范层的完善有利于我国人工智能主权的“良善”控制实现。

其一，从规制目标来看，我国对人工智能的监管应是在创新与安全可控之间寻求平衡的动态监管，在顶层设计方面纳入总体国家安全观。从规制方式来看，我国应渐进式建立一部综合的人工智能法律。目前，我国主要通过内容监管、数据保护和算法治理三个层次对人工智能进行治理，并且分散在《算法推荐规定》《深度合成规定》《人工智能暂行办法》《科技伦理审查办法（试行）》《个人信息保护法》和《数据安全法》之中。部分监管内容出现重复。综合性立法通过在关键内容方面达成共识，可减少立法冗余，并提高公民的可预测性。围绕人工智能安全、个人权益保护以及体系性部署人工智能产业发展构建相应制度，并且应为未来可能的技术产业发展预留充足的制度空间。《人工智能法（学者建议稿）》从权益保护、安全义务、监督管理、特殊应用场景、国际合作与法律责任等方面对人工智能实行全方位监管。不过，我国仍需进一步论证争议法律条文，关注人工智能监管与创新的平衡条款，有序推进人工智能立法工作。从软硬规则来看，应在全域构建软法与硬法深度耦合的规则体系。一方面，软法在未来具有可能的硬法化前景。如，为加强人工智能控制的有效性与合理性，《人工智能安全治理框架》鼓励人工智能创新发展，采用风险管理的方法，对人工智能的内生安全风险与应用安全风险提出综合防治措施。另一方面，人工智能伦理规范还有利于硬法的实施。从合作主权来看，中国可以加强与他国的人工智能合作以强化自身的人工智能能力建设。《人工智能能力建设普惠计划》是中国为弥合数字和智能鸿沟的最新举措，该计划明晰了合作形式多元化、合作内容全面化、合作利益倾斜南方国家的具体要求。未来中国仍需具体细化并落实人工智能基础设施、产业赋能、人才培养、数据建设和安全治理的相关方案，以促进人工智能的可持续发展。中国还可推动生成式人工智能技术的跨境传播，并纳入能力建设与发展援助之中。

78

最后，人工智能主权底层控制与良善控制的联动有利于技术创新和制度保障之间形成良性互动，确保人工智能主权的完满实现。一方面，技术可赋能监管智

⁷⁸ World Economic Forum, ChatWTO: An Analysis of Generative Artificial Intelligence and International Trade, September 2024, https://www3.weforum.org/docs/WEF_An_Analysis_of_Generative_Artificial_Intelligence

能化。我国在构建基于数据、算法、模型和场景等维度的治理模块时，应以技术赋能监管，打造与人工智能体模块化治理框架相适配的、智慧化的治理工具，实现对特定高风险场景的精准化治理。另一方面，法律作用于信息技术的方式除了预防技术安全风险、明确权利义务之外，还包括从科技创新资源供给、科技创新环境保障等角度推动科技创新。在人工智能技术治理领域，人工智能规范问题可以划分为科技市场法、科技风险管控法和科技政策法三种类型。⁷⁹我国需以动态平衡为核心原则，在科技市场法中激发创新活力，完善数据共享与知识产权保护机制，强化反垄断与算法共谋监管，在风险管控法中筑牢安全底线，构建动态化、场景化的风险治理体系，在科技政策法中引领全球治理，最终构建兼具包容性与竞争力的人工智能治理生态。

（三）有效应对他国可能的歧视性贸易措施

美国通过歧视性出口管制与经济制裁、数据保护全面围击我国人工智能产业，尤其是在芯片与数据流动方面将中国视为敌对竞争对手⁸⁰。一方面，美国的芯片产业立法冲击芯片产业链全球格局，有关身份条款——禁止美国人未经许可在中国境内从事相关工作⁸¹——将限制高端芯片人才的国际交流，出口管制措施严重阻碍我国芯片产业升级。另一方面，美国的跨境数据自由流动政策发生结构性转变，将跨境数据流动规则区域化，针对中国采取严格的流动措施，禁止受关注国家中国获取美国人大量敏感个人数据和美国政府相关数据，⁸²对内嵌中国信息通信技术或服务的智能网联汽车启动国家安全审查。人工智能大语言模型则加剧了美国对国家安全的关切，尤其其可能通过识别和控制关键敏感个人数据来训练算法。

美国区别对待的措施已然构成对 WTO 非歧视性原则的违反。为此，我国可以充分利用 WTO 争端解决机制，就美国的可诉补贴、限制数据流动等歧视性措施寻求磋商或主动启动审查程序，近年来，WTO 对于国家安全例外条款的裁断提高了胜诉的可能性；对美国的芯片进行反补贴调查；积极利用我国具有战略反制能力的新能源、芯片原材料等领域，增加博弈资本。从我国自身来看，《中华人民共和国反外国制裁法》《不可靠实体清单》《出口管制法》的反制与威慑能力十分有限。我国可完善《阻断外国法律与措施不当域外适用办法》，优化调整禁止、限制进出口技术目录，制定专门的《芯片法案》，提高政策透明度和可预测性，维护数字领域人工智能供应链安全、产业链安全。在此基础上，我国还可以继续推动中美出口管制信息交流机制与中美跨境数据流动交流机制。对于可能受

⁷⁹ 赵精武：《论人工智能训练数据高质量供给的制度建构》，载《中国法律评论》2025 年第 1 期，第 92-104 页。

⁸⁰ 这里的敌对外国政府（受关注国家）主要针对中国、俄罗斯、朝鲜与伊朗四国。

⁸¹ 美国《出口管制条例》EAR §772 规定，禁止美国人未经许可在中国境内从事支持开发和生产特定先进计算芯片和半导体生产相关的工作。

⁸² 这一措施将影响中美跨境贸易和国际供应链，中美人才流动与科研合作，中国企业在美投资和经营。

到影响的在美企业以及向美国提供服务及产品的中国企业而言，应全面评估相关后续法规、标准和指南要求与执法动态，以降低合规风险。

（四）推动全球人工智能与全球数据安全的协同治理

事实上，人工智能技术的全球性和跨境性质，加上经济体的相互关联性，需要多边和多利益攸关方合作，以制定共同的原则和标准。然而在人工智能安全基础上国际合作的背景下，训练的不同人工智能模型、国内法规与不同利益可能会使各国就人工智能合作达成协议更具挑战性。尤其是，一部分发达国家利用信息不对称与技术门槛，推动人工智能霸权与数据霸权，使得全球人工智能体系进展缓慢。此外，各国通过对跨境数据流动施加严格限制来实现人工智能主权会增加人工智能的系统性风险。因而，人工智能主权应被置于一套制衡体系之内，并维护全球人工智能与全球数据安全的协同治理。

中国应重视人工智能作为全球公共产品的性质，并使得人工智能全球治理更具包容性。中国应加强与他人工智能与跨境数据流动的合作。具体而言，其一，在合作内容方面，关于双边贸易协定，可以仿照《英国—新加坡数字经济协定》嵌入数据共享条款，合作制定数据移动性政策和标准，包括消费者数据可移植性，共享与之相关的政策方法和行业实践，如数据信托。⁸³为此，我国可以扩大对高质量数据的访问。对于禁止披露算法和源代码这一条款而言，中国应该在维护人工智能主权的基础上，逐步接受当前贸易协定中发达国家起草的源代码禁令的规制方法，并且根据对象为发达国家或发展中国家在豁免的范围方面而有所不同。其二，在合作对象方面，全球南方国家具有获得先进人工智能发展的平等权利。数字贸易规则对低收入和中等收入国家的益处最大，因而，中国在共建一带一路的框架下，可借助自身的人工智能优势，与其他发展中国家就跨境数据流动与人工智能合作达成一揽子计划，在这一方面，美国与尼日利亚的合作可为中国提供借鉴。⁸⁴与此同时，探索我国与欧盟在数据跨境流动、美国在开源社区治理等领域的合作机制，避免技术脱钩风险。其三，在合作方式方面，中国可推动人工智能主权倡议的提出，推进建立基于安全观的国际规则，积极参与联合国《人工智能伦理建议书》等标准制定，推动中国主导的人工智能安全测试认证体系如《人工智能安全治理框架》成为国际通用标准，深化人工智能国际合作与技术标准互认。⁸⁵此外，中国可推动人工智能与区块链等新兴技术的整合和协同作用。区块链可以保证数据的正确性和真实性。其四，在合作主体方面，仍需国家、企业与

⁸³ Emily Jones, Digital Disruption: Artificial Intelligence and International Trade Policy, Oxford Review of Economic Policy, Volume 39, Issue 1, Spring 2023, Pages 70–84, <https://doi.org/10.1093/oxrep/grac049>.

⁸⁴ Federal Ministry of Communications, Innovation & Digital Economy, Nigeria, U.S. Agree to Harness Artificial Intelligence, Facilitate Data Flows and Empower Digital Upskilling; Issue Joint Statement, <https://fmcide.gov.ng/u-s-nigeria-reach-agreement-to-harness-artificial-intelligence-facilitate-data-flows-and-empower-digital-upskilling-issue-joint-statement/>.

⁸⁵ David L. Shrier and A. Aldo Faisal, Sovereign GPTs: Aligning Values in AI for Development, https://unctad.org/news/sovereign-gpts-aligning-values-ai-development_and_International_Trade_2024.pdf.

个人在人工智能与数据治理层面的协同合作。尤其是大型科技公司基于专业性在推动全球人工智能治理方面应该起到关键作用。

结语

人工智能主权强调的是人工智能对维护国家主权的重要意义，其核心观点在于国家应重视人工智能的发展及它为国家主权带来的改变。人工智能主权是大国家主权概念，是数字主权的最新发展。人工智能企业的技术型权力推动了“主权人工智能”的开展，这为国家推行人工智能主权提供了坚实的基础与对外扩张的工具。人工智能主权显著区分于人工智能民族主义，这是因为后者具有排外性与优先性，这与人工智能的终极价值不相符合。人工智能自主权、人工智能同等反制权与人工智能合作权构成人工智能主权的核心内容。

人工智能主权结构是包括数据主权、算力主权、技术主权、规制主权的体系性主权。各主权之间相互支撑，数据控制能力决定算力、技术和规制主权的实现，技术主权影响规制主权，技术标准制定影响规则制定，规制主权依赖于其他主权并且为其他主权提供法律保障。当人工智能主权相互竞争或者为确保自身的人工智能主权时，会产生全球治理碎片化的外部效应，南方国家可能进一步成为人工智能的被殖民者，人工智能的发展还进一步扩大了安全困境的不同维度。此外，对于主权的保护行为有时构成对于多边贸易体制的破坏。

于我国而言，基础层、技术层与规范层的构建与联动是当下人工智能主权有效控制的前提条件，通过发展人工智能主权能力，中国可以对他国可能的歧视性贸易措施形成有效反制，并推动和影响全球人工智能标准和规范的形成，促进人工智能技术的互操作性、问责性和可持续性，利用人工智能技术实现普惠发展，释放新质生产力，从而创造更大的社会价值。

Structural analysis of artificial intelligence sovereignty

-On the system construction of China's artificial intelligence sovereignty

Abstract: The movement in the field of technology is spreading to data, artificial intelligence, security and privacy, and we have entered the era of artificial intelligence sovereignty. The sovereignty of artificial intelligence refers to the ability of a country to control and supervise artificial intelligence within its territory, which has the control right of artificial intelligence at home and the right of independent autonomy and cooperation abroad. The internal structure of artificial intelligence sovereignty consists of basic layer, technical layer and normative layer, and external effects may cause the implementation gap of comprehensive security, the representation gap between the North and the South, the coordination gap of supervision, and the compliance gap

between domestic sovereignty and international trade rules. On the whole, the sovereignty of artificial intelligence conforms to China's fundamental interests. In order to build China's sovereignty of artificial intelligence, it is necessary to take inclusive values as the reasonable core, the construction and linkage of basic layer, technical layer and normative layer as the basis, discriminatory trade countermeasures as the defense means, and global artificial intelligence and global data security collaboration as the external cooperation link, so as to realize the system construction of China's sovereignty of artificial intelligence.

Keywords: artificial intelligence sovereignty, structuralization, construction path

构建人类命运共同体与网络空间全球治理

曾璐*

摘要：随着互联网深度融入全球发展，网络空间全球治理成为国际社会关注焦点。本文聚焦构建人类命运共同体与网络空间全球治理，分析二者关联。构建人类命运共同体理念以其合作共赢、公平正义等核心内涵，为网络空间全球治理提供价值导向与目标指引。当前网络空间全球治理面临发展不平衡、安全威胁加剧等挑战，在此背景下，从加强国际合作、强化网络安全保障、推动数字经济发展、促进网络文化交流等角度，探索基于人类命运共同体理念的治理策略，旨在推动全球网络空间实现和平、安全、开放、合作、有序发展，促进各国在网络空间的共同利益与可持续发展。

关键词：构建人类命运共同体 互联网 网络空间 全球治理

在当今全球化的时代，互联网技术以前所未有的速度和广度发展，深刻地改变了人类的生产生活方式。网络空间已经成为与现实空间紧密相连的新的人类活动空间，对经济、政治、文化、社会等各个领域产生了深远影响。从经济领域来看，电子商务、数字金融等新兴业态蓬勃发展，成为推动经济增长的新引擎。据相关数据显示，全球电子商务交易额逐年攀升，2024年已达到数十万亿美元，众多企业通过网络平台拓展市场，实现了资源的优化配置和高效利用。在政治领域，网络空间成为信息传播和舆论交锋的重要阵地，社交媒体的兴起使得民众能够更便捷地参与政治讨论，对政府决策产生影响。文化领域，网络空间促进了文化的多元交流与传播，不同国家和地区的文化通过网络相互碰撞、融合，丰富了人类的文化宝库。

随着网络空间的不断发展，其重要性日益凸显，逐渐成为全球治理的关键领域。网络空间的开放性、虚拟性和跨国性等特点，也使得全球网络空间治理面临着诸多严峻挑战。网络安全问题日益突出，网络攻击、网络诈骗、网络间谍等恶意行为频繁发生，给各国的国家安全、经济安全和社会稳定带来了巨大威胁。2023年，某国遭受了大规模的网络攻击，导致其关键基础设施瘫痪，交通、能源等多个领域陷入混乱，造成了巨大的经济损失。网络犯罪呈现出跨国化、智能化的趋势，犯罪行为难以追踪和打击，国际合作面临重重困难。数据泄露事件频发，个人隐私和企业商业机密受到严重侵犯，引发了公众对数据安全的担忧。数字鸿沟问题依然存在，不同国家和地区在网络基础设施建设、互联网接入和使用能力等方面存在巨大差距，这不仅影响了全球经济的均衡发展，也加剧了国际社会的不

* 作者：曾璐，工作单位：桂林市资源县人民检察院，通讯方式：19167852886

平等。

在这样的背景下，构建人类命运共同体理念的提出为全球网络空间治理提供了全新的视角和思路。人类命运共同体理念强调人类的共同利益和共同责任，倡导建设持久和平、普遍安全、共同繁荣、开放包容、清洁美丽的世界。将这一理念应用于网络空间治理，有助于打破传统的单边主义和霸权主义思维，推动各国在网络空间领域加强合作，共同应对全球性挑战，实现网络空间的和平、安全、开放、合作与有序发展。

本研究具有重要的理论意义和实践意义。在理论方面，深入探讨构建人类命运共同体与网络空间全球治理的关系，有助于丰富和完善全球治理理论体系，为网络空间治理提供新的理论框架和分析视角。通过研究，可以进一步明确网络空间治理的目标、原则和方法，推动网络空间治理理论的创新与发展。在实践方面，本研究的成果可以为各国政府制定网络空间政策、参与国际网络空间治理合作提供有益的参考和指导。有助于促进各国在网络空间领域的交流与合作，加强网络安全防护，缩小数字鸿沟，推动全球网络空间治理体系朝着更加公正合理的方向发展，为构建人类命运共同体奠定坚实的网络空间基础。

一、构建人类命运共同体与网络空间全球治理的相关概述

（一）构建人类命运共同体理念的内涵

构建人类命运共同体理念的提出有着深刻的时代背景。当今世界，全球化进程不断加速，各国之间的经济、政治、文化等联系日益紧密，形成了你中有我、我中有你的格局。在经济领域，全球产业链、供应链深度融合，各国经济相互依存度不断提高。

然而，全球化的发展也带来了一系列严峻的挑战。地区冲突不断，如中东地区长期的战乱，导致大量人员伤亡和流离失所，给当地的经济和社会发展带来了巨大的破坏。恐怖主义活动猖獗，其袭击范围不断扩大，对全球的安全稳定构成了严重威胁。环境污染问题日益突出，全球气候变暖、酸雨、水污染等问题严重影响了人类的生存环境。贫富差距悬殊，一些发达国家与发展中国家之间的经济差距不断拉大，导致世界发展不平衡加剧。这些问题相互交织，使得人类面临着前所未有的困境，迫切需要一种新的理念和方案来解决这些全球性问题。

在这样的背景下，2013年3月23日，习近平主席在俄罗斯莫斯科国际关系学院首次向世界提出“人类命运共同体”重大倡议，呼吁国际社会树立“你中有我、我中有你”的命运共同体意识。这一倡议的提出，为解决全球问题提供了新的思路 and 方向，得到了国际社会的广泛关注和积极响应。此后，在多个重要国际场合，习近平主席进一步阐述和丰富了构建人类命运共同体理念的内涵。2017年1月18日，习近平主席在日内瓦万国宫出席“共商共筑人类命运共同体”高级别会议，并发表题为《共同构建人类命运共同体》的主旨演讲，深刻、全面、系

统阐述人类命运共同体理念，主张共同推进构建人类命运共同体伟大进程，坚持对话协商、共建共享、合作共赢、交流互鉴、绿色低碳。

构建人类命运共同体理念对全球治理具有重要的指导意义。它为全球治理提供了新的价值导向，打破了传统的以自我利益为中心的治理观念，强调人类的共同利益和共同责任，倡导各国在全球治理中相互合作、共同担当。在应对气候变化问题上，各国应根据自身的能力和 responsibility，共同制定减排目标，加强技术合作，共同推动全球气候治理。该理念推动了全球治理体系的变革与完善，促使国际社会更加关注发展中国家的声音和利益，推动全球治理体系朝着更加公正合理的方向发展。在国际经济秩序中，应增加发展中国家的代表性和发言权，改革不合理的国际金融体系，促进全球经济的均衡发展。它还为解决全球性问题提供了具体的行动方案，通过加强国际合作、推动多边机制建设等方式，共同应对恐怖主义、网络安全、环境污染等全球性挑战，实现人类社会的可持续发展。

（二）两者的内在联系与逻辑关系

构建人类命运共同体理念与网络空间全球治理之间存在着紧密的内在联系和深刻的逻辑关系，二者相互促进、相辅相成。

构建人类命运共同体理念为网络空间全球治理提供了价值导向和目标指引。这一理念强调人类的共同利益和共同责任，倡导建设持久和平、普遍安全、共同繁荣、开放包容、清洁美丽的世界，这些价值追求与网络空间全球治理的目标高度契合。在网络空间中，各国应遵循构建人类命运共同体的理念，尊重彼此的网络主权，摒弃网络霸权主义和单边主义，通过平等协商、合作共赢的方式解决网络空间中的各种问题，实现网络空间的和平与安全。各国应加强网络安全合作，共同防范网络攻击、网络犯罪等威胁，维护网络空间的稳定秩序。在网络经济领域，应促进数字经济的公平发展，推动各国在网络技术创新、电子商务等方面的合作，实现共同繁荣。在网络文化方面，要倡导开放包容的文化交流，尊重不同国家和地区的文化差异，促进网络文化的多元共生。

网络空间全球治理是构建人类命运共同体的重要组成部分和支撑。随着互联网的普及和网络技术的发展，网络空间已经成为人类社会的重要组成部分，对人类的生产和生活方式产生了深远影响。网络空间的发展状况直接关系到人类命运共同体的构建。通过有效的网络空间全球治理，可以促进网络空间的健康发展，为构建人类命运共同体提供坚实的网络基础。加强网络空间的国际合作，可以增进各国之间的相互了解和信任，促进国际间的交流与合作，推动构建新型国际关系。在网络安全领域，各国通过共同应对网络安全威胁，加强网络安全防护，能够提升全球的安全水平，为构建普遍安全的人类命运共同体提供保障。推动网络空间的互联互通和数字经济的发展，可以促进资源的优化配置和共享，推动全球经济的增长，为实现共同繁荣的人类命运共同体目标做出贡献。

构建人类命运共同体理念与网络空间全球治理在实践中相互促进、共同发展。在构建人类命运共同体理念的指导下，网络空间全球治理不断推进，各国在网络空间领域的合作不断加强，治理机制不断完善。在网络安全领域，各国通过建立网络安全信息共享机制、开展联合执法行动等方式，共同应对网络安全威胁，取得了显著成效。在网络空间国际规则制定方面，各国积极参与，推动制定更加公正合理的国际规则，为网络空间的有序发展提供了保障。网络空间全球治理的实践也进一步丰富和发展了构建人类命运共同体理念。在网络空间治理过程中，各国不断探索新的合作模式和治理机制，这些实践经验为构建人类命运共同体提供了有益的借鉴和启示。

二、网络空间全球治理的现状与挑战

（一）网络空间全球治理的现状分析

当前，全球网络空间治理主要存在三种模式，分别是美国主导的多利益相关方模式、欧盟的数据保护导向模式以及中国倡导的主权主导下多元共治模式。

美国的多利益相关方模式强调政府、私营部门、民间社会组织、技术社群等多利益相关方共同参与网络空间治理。在这种模式下，政府在网络空间治理中的直接干预相对较少，更多地是通过制定政策和提供支持来引导治理。互联网名称与数字地址分配机构（ICANN）是这种模式的典型代表，它负责全球互联网域名系统（DNS）的管理和协调，由来自不同国家和地区的政府、企业、技术专家等多利益相关方共同参与决策和运营。这种模式的优点在于能够充分调动各方的积极性和专业知识，促进网络技术的创新和发展，提高治理效率。但也存在一些问题，由于美国在互联网技术和资源方面占据主导地位，这种模式可能会受到美国利益的影响，导致治理结果偏向美国，一些发展中国家的利益难以得到充分保障。

欧盟的数据保护导向模式则侧重于通过严格的数据保护法规来规范网络空间行为。欧盟出台了《通用数据保护条例》（GDPR），该条例对企业收集、使用和存储个人数据的行为进行了严格规范，加强了对个人数据的保护。同时，欧盟还通过建立相关机构和机制，加强对数据保护法规的执行和监督。这种模式的优势在于能够有效保护公民的个人隐私和数据安全，提升公众对网络空间的信任。但严格的数据保护法规也可能对数字经济的发展产生一定的限制，增加企业的运营成本，在跨境数据流动等方面也面临一些协调难题。

中国倡导的主权主导下多元共治模式，坚持网络主权原则，强调国家在网络空间治理中的主体地位和主导作用。同时，鼓励政府、企业、社会组织、技术社群和网民等多元主体共同参与网络空间治理，形成协同治理的格局。中国通过制定网络空间相关法律法规，加强网络安全监管，维护国家网络安全和公民合法权益。积极推动网络空间国际合作，参与国际网络空间治理规则的制定，提出构建网络空间命运共同体的理念，为全球网络空间治理贡献中国智慧和方案。这种模

式既尊重了各国的网络主权，又充分发挥了多元主体的作用，能够更好地平衡网络空间的安全、发展和治理之间的关系。

在网络空间治理中，许多国际组织和国家都采取了一系列重要举措并取得了一定的合作成果。联合国在网络空间全球治理中发挥着重要的协调和推动作用。联合国信息社会世界峰会（WSIS）为各国提供了一个讨论网络空间治理问题的重要平台，促进了各国在网络空间发展、安全、治理等方面的交流与合作。联合国还通过制定相关决议和倡议，推动全球网络空间治理的发展，如呼吁各国加强网络安全合作，共同应对网络犯罪等威胁。

二十国集团（G20）在网络空间治理方面也发挥了积极作用。G20 通过发布《二十国集团数字经济发展与合作倡议》等文件，推动数字经济的发展和合作，加强数字基础设施建设，促进数字技术的创新和应用。G20 还关注网络安全问题，倡导各国加强网络安全合作，共同维护全球网络空间的稳定。

在区域合作方面，欧盟通过制定统一的网络空间政策和法规，加强内部网络空间的治理和合作。欧盟在数据保护、网络安全等领域制定了一系列严格的法规和标准，促进了欧盟内部网络空间的一体化发展。亚太经合组织（APEC）也在积极推动数字经济合作，通过制定《APEC 互联网和数字经济路线图》等文件，加强成员国在数字经济领域的合作，促进数字贸易的发展，提升区域内的数字经济竞争力。

各国之间也在网络空间治理领域开展了广泛的双边合作。中美两国在网络安全领域进行了多次对话和交流，就打击网络犯罪、加强网络安全防护等问题达成了一些共识。中国与东盟国家在网络安全领域的合作也取得了显著成果，双方建立了网络安全交流培训中心，开展网络安全培训和演练，共同应对网络安全威胁。

（二）网络空间全球治理面临的挑战

随着网络空间的快速发展，网络空间全球治理面临着诸多严峻挑战，这些挑战严重阻碍了网络空间的健康、有序发展，对全球网络空间治理构成了重大威胁。

网络空间发展不平衡是当前面临的突出问题之一。不同国家和地区在网络基础设施建设、互联网接入和使用能力等方面存在巨大差距，形成了数字鸿沟。在网络基础设施建设方面，发达国家凭借其雄厚的经济实力和先进的技术水平，拥有高速、稳定的网络基础设施，光纤宽带和 5G 网络覆盖广泛。据统计，欧洲和北美地区的互联网普及率高达 87%-91%，5G 覆盖率也处于较高水平，欧洲 5G 覆盖率达到 68%，美洲地区为 59%。而许多发展中国家的网络基础设施建设相对滞后，网络速度慢、稳定性差，部分偏远地区甚至无法接入互联网。在非洲。互联网普及率仅为 37%，5G 覆盖率更是低至 6%。这种数字鸿沟不仅限制了发展中国家的经济发展和社会进步，也加剧了国际社会的不平等。在经济领域，数字经济的快速发展使得网络技术和资源成为重要的生产要素。发展中国家由于网络

基础设施薄弱，在数字经济发展方面面临诸多困难，难以充分利用互联网技术推动产业升级和创新发展，与发达国家的经济差距进一步拉大。在教育领域，网络学习资源的获取也受到数字鸿沟的影响，发展中国家的学生难以享受到优质的在线教育资源，限制了他们的知识获取和能力提升。

网络安全威胁日益加剧，给全球网络空间治理带来了巨大挑战。网络攻击手段不断升级，网络间谍活动频繁发生，网络犯罪呈现出跨国化、智能化的趋势。网络攻击的形式日益多样化，包括分布式拒绝服务攻击(DDoS)、恶意软件攻击、漏洞利用等。这些攻击不仅会导致网络系统瘫痪、数据泄露，还会对国家关键基础设施造成严重破坏，威胁国家安全和社会稳定。某国的电力系统曾遭受网络攻击，导致大面积停电，给民众生活和经济活动带来了极大不便。网络间谍活动也日益猖獗，一些国家通过网络手段窃取他国的政治、军事、经济等机密信息，严重损害了他国的利益。网络犯罪呈现出跨国化的特点，犯罪分子利用网络的跨国性，在不同国家之间实施诈骗、盗窃等犯罪行为，使得犯罪追踪和打击难度加大。网络犯罪还呈现出智能化的趋势，犯罪分子利用人工智能、大数据等技术手段，提高犯罪的精准性和隐蔽性，增加了防范和打击的难度。

网络空间霸权主义的存在严重破坏了网络空间的公平正义和国际秩序。一些西方国家凭借其在网络技术和资源方面的优势，推行网络霸权主义，试图主导全球网络空间治理规则的制定，将自身的价值观和利益诉求强加于其他国家。美国在网络空间领域具有强大的技术和资源优势，通过控制互联网关键资源，如互联网域名系统(DNS)的管理，对其他国家的网络空间进行渗透和控制。美国还利用网络监控技术，对其他国家进行大规模的网络监听，侵犯他国的主权和公民隐私。美国的“棱镜门”¹事件曝光后，引发了国际社会的广泛关注和强烈谴责。这种网络空间霸权主义行为不仅破坏了网络空间的和平与稳定，也阻碍了全球网络空间治理的健康发展，使得网络空间治理难以实现真正的公平、公正和合理。

数据安全和隐私保护问题也成为网络空间全球治理的重要挑战。随着数字经济的快速发展，数据成为重要的生产要素和战略资源。大量的个人数据和企业商业数据在网络空间中流动和存储，数据泄露事件频发，给个人和企业带来了巨大的损失。一些企业由于安全防护措施不到位，导致用户数据被泄露，个人隐私受到侵犯，用户的财产安全和个人权益面临威胁。数据的跨境流动也带来了数据安全和隐私保护的难题，不同国家的数据保护法规和标准存在差异，使得数据在跨

¹ 棱镜门事件：2013 年 6 月，美国前情报雇员爱德华·斯诺登披露美国国家安全局(NSA)实施的“棱镜计划”(PRISM)，该计划通过苹果、谷歌等 9 家互联网公司大规模收集全球用户通信数据，包括电子邮件、视频通话、社交网络信息等。事件揭示了网络空间中存在的单边主义监控行为，直接冲击了国家主权平等原则。据斯诺登披露的文件显示，NSA 在 2012 年通过该计划获取了超过 2.5 亿条全球用户数据(《卫报》2013 年 6 月 6 日报道)。该事件被联合国人权理事会认定为“系统性侵犯隐私权”，成为网络空间治理体系失衡的典型案列。

境传输过程中面临安全风险。一些国家对数据的跨境流动进行严格限制，以保护本国的数据安全和隐私，但这也给跨国企业的运营和国际数据合作带来了困难。

三、基于构建人类命运共同体的网络空间全球治理策略

（一）加强国际合作，促进共同发展

构建人类命运共同体理念的核心在于倡导各国在网络空间中加强合作，携手应对共同挑战，实现共同发展。在当前网络空间全球化的背景下，任何一个国家都难以独自应对网络安全威胁、数字鸿沟等复杂问题，国际合作成为必然选择。

中国积极推动建立多边、民主、透明的全球互联网治理体系，强调各国应在平等、互利、共赢的基础上开展网络空间国际合作。这一理念的提出，旨在打破传统的单边主义和霸权主义思维，让更多国家能够参与到网络空间治理中来，充分表达自身的利益诉求和发展愿景。多边合作机制能够汇聚各国的智慧和力量，共同制定网络空间治理规则，加强网络安全防护，促进数字经济发展。通过联合国、二十国集团（G20）等国际组织和平台，各国可以就网络空间治理问题进行深入讨论和协商，推动形成具有广泛共识的治理方案。

在技术合作方面，各国应加强在网络技术研发、网络安全防护等领域的交流与合作，共同攻克技术难题，提升全球网络技术水平。中国与俄罗斯在网络安全技术领域的合作不断深化，双方共同开展网络安全技术研究，分享网络安全情报，加强网络安全防护能力建设。两国还在 5G 技术、人工智能等新兴技术领域开展合作，共同推动网络技术的创新与应用。

数据合作也是网络空间国际合作的重要内容。各国应建立数据共享机制，促进数据的安全、有序流动，推动数据资源的合理利用。在医疗领域，各国可以共享医疗数据，开展联合研究，提高疾病诊断和治疗水平。在科研领域，数据共享能够促进科学研究的创新与发展，加速科技成果的转化。

人才合作同样不可或缺。各国应加强网络空间人才培养和交流，通过举办学术研讨会、培训课程等方式，培养高素质的网络空间专业人才。中国与非洲国家在网络空间人才培养方面开展了广泛合作，为非洲国家培养了大量的网络技术人才。中国通过举办网络技术培训班、提供奖学金等方式，帮助非洲国家提升网络技术人才的素质和能力，促进了非洲国家网络空间的发展。

以金砖国家网络安全合作机制为例，该机制的建立为金砖国家在网络安全领域的合作提供了重要平台。金砖国家在网络安全领域面临着共同的威胁和挑战，如网络攻击、网络犯罪等。通过该合作机制，金砖国家可以共享网络安全信息，加强技术交流与合作，共同应对网络安全威胁。在网络安全技术研发方面，金砖国家可以共同投入资源，开展联合研究，提高网络安全防护技术水平。在网络安全事件应急响应方面，金砖国家可以建立联动机制，共同应对网络安全事件，减少损失。

再如，中国与东盟国家在网络安全领域的合作也取得了显著成效。双方建立了网络安全交流培训中心，开展网络安全培训和演练，共同应对网络安全威胁。通过合作，中国与东盟国家在网络安全防护能力、网络安全意识等方面都得到了提升，为区域网络安全稳定做出了重要贡献。

（二）强化网络安全保障，维护网络空间主权

在网络空间中，网络安全是至关重要的基础，直接关系到国家的主权、安全和发展利益。各国应高度重视网络安全保障工作，不断提升自身的网络安全防护能力，加强关键信息基础设施保护，坚决反对网络霸权和网络干涉，切实维护网络空间主权。

提升网络安全防护能力是保障网络安全的关键。各国应加大在网络安全技术研发方面的投入，培养高素质的网络安全专业人才，建立健全网络安全防护体系。通过研发先进的网络安全技术，如入侵检测系统、防火墙、加密技术等，提高对网络攻击的防范能力。加强网络安全人才培养，通过开展网络安全培训、竞赛等活动，提高网络安全人才的专业技能和应急处理能力。美国在网络安全技术研发方面一直处于世界领先地位，投入了大量的资金和资源，研发出了一系列先进的网络安全技术和产品。美国国家安全局（NSA）拥有一支庞大的网络安全专业人才队伍，他们在网络安全技术研发、网络攻击防范等方面发挥着重要作用。

加强关键信息基础设施保护是维护网络安全的重要举措。关键信息基础设施是国家经济社会运行的神经中枢，包括能源、交通、金融、通信等领域的重要信息系统和网络。各国应建立健全关键信息基础设施保护制度，加强对关键信息基础设施的安全监测和预警，提高关键信息基础设施的抗攻击能力。2017年，美国发生“WannaCry”勒索病毒攻击事件，该病毒迅速蔓延，导致全球范围内大量的计算机系统受到感染，其中包括许多关键信息基础设施，如英国的医疗系统、俄罗斯的铁路系统等，给这些国家的经济和社会运行带来了严重影响。为了应对此类事件，各国纷纷加强了对关键信息基础设施的保护，制定了相关的法律法规和政策措施，加强了对关键信息基础设施的安全监管和防护。

反对网络霸权和网络干涉是维护网络空间主权的必然要求。一些西方国家凭借其在网络技术和资源方面的优势，推行网络霸权主义，对其他国家进行网络监控、攻击和干涉，严重侵犯了他国的网络空间主权。各国应坚决反对网络霸权主义和网络干涉行为，通过国际合作和外交途径，维护自身的网络空间主权和合法权益。2013年，美国“棱镜门”事件曝光，美国国家安全局通过网络监控手段，对全球多个国家的政府、企业和个人进行大规模的监听和数据收集，严重侵犯了他国的主权和公民隐私。这一事件引发了国际社会的广泛关注和强烈谴责，各国纷纷要求美国停止网络霸权行为，尊重他国的网络空间主权。

（三）推动数字经济发展，缩小数字鸿沟

数字经济作为全球经济发展的新引擎，在当今时代正发挥着愈发关键的作用。随着信息技术的飞速发展，数字经济已渗透到社会生活的各个领域，成为推动经济增长、创新发展和社会进步的重要力量。然而，当前全球数字经济发展呈现出不均衡的态势，数字鸿沟问题严重制约了全球经济的协调发展和人类命运共同体的构建。因此，积极推动数字经济国际合作，大力支持发展中国家的数字基础设施建设，广泛分享数字经济发展经验，已成为缩小数字鸿沟、实现全球数字经济共同繁荣的必然选择。

中国与非洲国家在数字经济领域的合作堪称典范，为推动数字经济发展、缩小数字鸿沟提供了宝贵的实践经验。近年来，中国与非洲国家在数字经济领域的合作不断深化，取得了丰硕的成果。在数字基础设施建设方面，中国企业积极参与非洲国家的网络建设，为非洲国家构建了空天地立体高速通信网络。中国移动国际有限公司参与建设的“2Africa”超级工程，海缆总长度达 45000 公里，连接亚、非、欧三大洲的 33 个国家和 46 个登陆点，为非洲提供了低时延、高质量、低成本的“端到端”带宽解决方案，有力地推动了非洲数字经济的发展。中国企业还为非洲多个国家建设了光纤宽带网络、4G/5G 移动通信网络及卫星服务，服务超过 9 亿非洲人口。在赞比亚南方省的纳姆瓦拉区，中国企业帮助建设了通信塔和微型太阳能电站，为当地提供了移动和互联网连接服务以及电力保障，使小村庄实现了数字化转型。

在数字技术应用方面，中国的数字技术和创新模式在非洲得到了广泛应用和推广。蚂蚁集团与南非沃达康公司合作开发的数字钱包 APP—Vodapay，已服务 300 多万用户，成为南非最大的数字钱包，为非洲消费者带来了便捷的移动支付体验。蚂蚁小程序开发平台助力 M-Pesa 等伙伴在 8 个国家服务 6000 多万活跃用户，推动了非洲数字经济的发展。在农业领域，中国的农业物联网技术、精准农业技术等非洲得到应用，帮助非洲提高农业生产效率和农产品质量。在金融领域，移动支付、数字信贷等数字金融服务在非洲不断普及，为非洲中小企业和个人提供了更加便捷的金融服务。

中国还注重与非洲国家分享数字经济发展经验，加强数字经济人才培养。蚂蚁集团联合多家国际合作伙伴，通过“10x1000”项目为非洲培养本地数字人才，已为 541 位非洲创业者和年轻人提供了培训课程。在 2024 年中非数字合作论坛上，中国和 26 个非洲国家共同制定并发布《中非数字合作发展行动计划》，提出未来三年共建 10 个数字转型示范项目，共同培养 1000 名数字人才。通过举办数字经济研讨会、培训班等活动，中国向非洲国家分享了数字经济政策制定、产业发展、市场监管等方面的经验，为非洲数字经济的发展提供了智力支持。

中国与非洲国家的数字经济合作，不仅促进了非洲数字经济的发展，也为缩

小全球数字鸿沟做出了积极贡献。通过加强数字基础设施建设、推广数字技术应用和分享数字经济发展经验，非洲国家在数字经济领域取得了显著进步，提升了自身的经济发展能力和国际竞争力。这种合作模式也为其他国家和地区在数字经济领域的合作提供了有益的借鉴，推动了全球数字经济的共同发展。

（四）促进网络文化交流，倡导文明互鉴

在网络空间中，文化交流与文明互鉴是推动人类文明进步的重要途径。互联网打破时空限制，让不同文化得以在网络碰撞、交流与融合。为促进网络文化交流与文明互鉴，可从以下几方面着手：首先，搭建多样化网络文化交流平台。专门的平台能汇聚各国文化资源，方便文化展示与交流。如“孔子学院网络学院”，通过网络课程、在线讲座和文化活动等形式，传播中国语言和文化，满足不同层次学习者需求，增进中外文化交流互动。其次，尊重不同国家和民族的文化差异。每个国家和民族都有独特文化传统与价值观念，这是人类文明多样性的体现。在网络交流中，应摒弃文化偏见与歧视，倡导文化平等和包容，以开放心态理解和欣赏他国文化特色，从中汲取营养。再者，共同营造健康的网络文化环境。网络存在低俗、暴力等不良文化现象，阻碍文明互鉴。各国需加强网络文化监管，制定法规打击不良文化传播，净化网络空间。同时加强网络文化教育，提高网民素养，引导其抵制不良文化，开展宣传活动弘扬正能量。最后，积极开展文化交流活动。通过举办网络文化节、展览、竞赛等，展示各国优秀文化成果，增进人民间的文化交流与了解。不同文化的交流碰撞还能激发文化创新，推动文化发展进步。

总之，各国携手在网络文化交流中做好上述工作，定能促进文化多元共生，推动人类文明不断向前发展。

四、结论与展望

（一）结论

在互联网深度重塑人类社会的时代背景下，网络空间全球治理已成为国际社会面临的共同课题。构建人类命运共同体理念为破解网络治理难题提供了价值遵循，其核心在于倡导主权平等、合作共赢，推动各国在网络空间形成休戚与共的治理格局。

当前，网络空间发展失衡、安全威胁加剧、数据霸权主义等问题凸显。发达国家与发展中国家数字鸿沟持续扩大，网络攻击、虚假信息等跨国威胁频发，单边主义治理模式难以应对全球性挑战。人类命运共同体理念突破零和思维，主张各国在网络空间尊重彼此主权，通过对话协商制定多边规则，为构建公平合理的治理体系注入新动能。

中国作为网络空间治理的积极践行者，通过一系列举措推动理念落地。在国

内层面，建立《网络安全法》《数据安全法》等法律体系，强化关键信息基础设施保护；在国际层面，提出《全球数据安全倡议》，推动“数字丝绸之路”建设，与 81 个国家建立网络安全应急响应合作机制。中国倡导的网络主权原则被写入联合国决议，为全球治理提供了“安全为基、发展为要”的中国方案。

面向未来，需从四方面深化实践：一是加强多边合作，完善联合国框架下的全球治理机制，推动数字经济规则互认；二是强化技术共享，通过 5G、云计算等技术合作缩小南北差距；三是筑牢安全屏障，建立跨境数据流动规范，打击网络犯罪；四是促进文明互鉴，搭建数字文化交流平台，反对文化霸权。中国将继续以人类命运共同体理念为指引，与国际社会携手构建和平、开放、安全的网络空间，让数字文明成果惠及全人类。

（二）未来展望

展望未来，网络空间全球治理将面临更多的机遇与挑战。随着 5G、人工智能、区块链等新兴技术的不断发展，网络空间将更加深入地融入人类社会的各个领域，成为推动经济社会发展的重要力量。与此同时，网络空间的安全威胁也将更加复杂多样，网络空间治理的难度将进一步加大。在这样的背景下，持续推进构建人类命运共同体理念在网络空间治理中的实践，将具有更加重要的意义。

在技术发展方面，各国应加强在新兴技术领域的合作，共同推动网络技术的创新与应用。在人工智能领域，各国应加强数据共享和技术交流，共同研发人工智能技术，推动人工智能在网络安全、智能交通、医疗健康等领域的应用。在区块链技术方面，各国应加强合作，共同探索区块链在网络空间治理中的应用场景，如数据安全保护、网络信任体系建设等，提高网络空间的安全性和可信度。

在国际合作方面，应进一步加强多边合作机制的建设，推动各国在网络空间治理中的深度合作。联合国等国际组织应发挥更加积极的作用，协调各国之间的利益关系，推动制定更加公正合理的网络空间国际规则。各国应加强在网络安全、数字经济、网络文化等领域的合作，共同应对全球性挑战。在网络安全领域，各国应加强网络安全信息共享和联合执法行动，共同打击网络犯罪和网络攻击。在数字经济领域，各国应加强数字贸易合作，推动数字经济的全球化发展。

在规则制定方面，应加快网络空间国际规则的制定和完善，为网络空间治理提供更加坚实的法律保障。各国应在尊重网络主权的基础上，共同参与网络空间国际规则的制定，充分考虑不同国家的利益和诉求，确保规则的公平性和合理性。在数据安全、网络隐私保护、网络犯罪惩治等方面，应制定统一的国际规则，规范网络空间行为，维护网络空间的秩序。

构建人类命运共同体理念为网络空间全球治理提供了新的思路 and 方向。在未来的网络空间全球治理中，各国应积极践行这一理念，加强合作，共同应对挑战，推动网络空间的和平、安全、开放、合作与有序发展，为构建人类命运共同体奠

定坚实的网络空间基础。

参考文献:

- [1]吴慧,许屹山.习近平网络空间命运共同体理念的内在逻辑、科学内涵与时代价值[J].安庆师范大学学报(社会科学版).2021,40(6).DOI:10.13757/j.cnki.cn34-1329/c.2021.06.001.
- [2]邬贺铨.走中国特色的网络强国之路[J].红旗文稿.2021,(7).13-16
- [3]赵瑞琦.全球网络治理改革:崛起国的路径选择[J].学术界.2021,(1).DOI:10.3969/j.issn.1002-1698.2021.01.005
- [4]熊光清,王棋.如何构建网络空间命运共同体——从全球抗击新冠肺炎疫情的角度[J].党政研究.2021,(5).DOI:10.3969/j.issn.1008-9187.2021.05.010.
- [5]程乐.构建网络空间命运共同体:理念、路径与策略[J].北京航空航天大学学报(社会科学版).2021,34(5).36-38.Doi:10.13766/j.bhsk.1008-2204.2021.0374.
- [6]王四新.网络空间命运共同体理念的价值分析[J].人民论坛.2022,(4).DOI:10.3969/j.issn.1004-3381.2022.04.007
- [7]郝志超,王旨思虹.2021 年全球网络空间安全态势分析[J].信息安全与通信保密.2022,(1).DOI:10.3969/j.issn.1009-8054.2022.01.001.
- [8]赵弘毅.网络空间的多维场域结构及治理路径研究[J].现代管理科学.2021,(2).
- [9]何雅迪.马克思主义共同体思想在新时代的创新发展研究[D].2023.
- [10]张毅.网络空间国际合作治理:现实挑战与机制重构[J].国家治理.2023,(23).
- [11]周伯乐,马莉莉."中国方案"能否破解全球发展赤字[J].经济学家.2023,(12).
- [12]中国网络空间研究院.世界互联网发展报告 2022 [M].电子工业出版社,2022.
- [13]丁元竹.共同构建网络空间命运共同体:意义、方向与路径[J].国家治理.2023(3):2-7.
- [14]高晓瑜,杨茜伊,彭泽.网络空间命运共同体参与全球互联网治理的中国方案[J].青年记者,2023(21):4-5.
- [15]刘镭.人类命运共同体视域下网络空间命运共同体构建刍议[J].中国矿业大学学报(社会科学版),2024,26(6):69-76. DOI:10.20089/j.cnki.issn.1009-105x.2024.06.007.

- [16]徐秀军,阮建平,周玉渊,等.新质生产力与人类命运共同体[J].战略决策研究.2024,(4).3-25.
- [17]周璐铭. 推动构建网络空间命运共同体迈向新阶段[J]. 中华魂,2024(3):14.
- [18]罗坤瑾,许嘉馨.网络空间安全共同体构建的源起、逻辑与路径[J].现代传播.2024,46(2).DOI:10.3969/j.issn.1007-8770.2024.02.015 .
- [19]刘阳.数字时代网络空间命运共同体构建路径探析[J].信息系统工程.2024,(2).DOI:10.3969/j.issn.1001-2362.2024.02.033 .
- [20]袁莎.网络空间命运共同体：核心要义与构建路径[J].国际问题研究.2023,(2).DOI:10.3969/j.issn.0452-8832.2023.02.002 .
- [21]吕娜. 习近平总书记关于网络空间命运共同体重要论述研究[D]. 齐鲁工业大学,2024.

禁止武力威胁或使用武力视角下的网络攻击问题研究

许钊涵*

摘要：网络化、信息化和数字化时代的到来，一方面极大便利了人们的生活，也给国家带来了显著的益处，如推动经济蓬勃发展和加快社会的信息化进程与整体进步。另一方面，然而，网络技术的飞速发展也为网络攻击提供了滋生与蔓延的土壤，这些攻击展现出跨国界、高隐蔽性和巨大破坏力的特点，并逐渐呈现出军事化的倾向。在此背景下，禁止武力威胁或使用武力原则视角下的网络攻击问题日益受到重视。但该视角下的网络攻击依然存在“武力”界定和自卫权行使的困境。而西方国家正试图通过提出解决这些困境的方案，来扩大自卫权的行使范围，进而推动网络空间的军事化，以实现其自身利益。针对这一复杂局势，我国应以系统性思维统筹应对，在网络空间国际秩序构建中积极表达中国立场和保障自身安全利益。

关键词：禁止武力威胁或使用武力 网络攻击 自卫权

近代以来，网络信息技术不断发展，极大便利了人们的日常生活，在国家治理生活中也占据着重要的位置，无论是在政治、经济、文化还是军事等各个方面都发挥着独一无二的作用。各国纷纷在网络空间中积极寻求突破，并试图建立优势地位，使得网络空间成为了竞争激烈的战略高地。然而，随着网络信息技术的不断发展，网络攻击（cyber attack）¹事件愈发频繁。

2007年，爱沙尼亚受到了前所未有的大规模的网络攻击。国家机关、金融机构等重要主体的网络全面瘫痪，并且时间长达数周。2008年格俄战争期间，出现了针对格鲁吉亚的大规模网络攻击，致使格鲁吉亚国内的通信、运输系统无法正常运作。从而导致战争物资无法及时运达指定位置，军队作战能力大打折扣。这一现象彰显了网络攻击与传统军事行动的融合，为各国敲响了严峻的安全警钟。无独有偶，2010年“震网”（Stuxnet）病毒攻击伊朗的布什尔核电站，致使伊朗的核发展计划被迫推迟。自2022年爆发以来至今仍未停歇的俄乌冲突中，网络空间更是成为了两国博弈的另一主战场。一方面，持续不断的网络攻击使双方的政治、经济受影响²，呈现出网络攻击军事化的趋势。另一方面，俄乌双方都重视利

* 作者简介：许钊涵，西南政法大学法学硕士，西南政法大学海外利益保护研究中心助理研究员。

¹ 《网络行动国际法塔林手册 2.0 版》（以下简称《塔林手册 2.0 版》）对网络攻击的定义为，无论进攻还是防御，网络攻击是可合理遇见的会导致人员伤亡或物体损毁的网络行动。[美]迈克尔·施密特：《网络行动国际法塔林手册 2.0 版》，黄志雄等译，北京：社会科学文献出版社，2017 年版，第 406 页。

² 郎平：“从俄乌冲突看网络空间武器化倾向及其影响”，《中国信息安全》，2022 年第 6 期，第 66 页。

用网络新媒体进行舆论战，从而争夺国际话语权。由此可见，网络攻击常常涉及到一国的政治、经济和军事等方面，严重影响一国的发展进程。

一、网络攻击的定性

实践中网络攻击事件频发，加之网络本身的全球性和开放性特征，要求有一套相对统一的国际法规则约束网络攻击行为。实务和学界的普遍共识是，网络空间并非“法外之地”，现有的国际法应当延伸到网络空间内适用。³其中，相关的学术研究和国家立场主要集中在诉诸战争权(jus ad bellum)和战时法规(jus in bello)方面。但将现有的国际法规则用“法律类推”的方式适用于网络空间的过程中可能发生“移植排斥反应”，传统国际法原则和规则在应对网络攻击时可能存在不适应。与诉诸战争权相关的国际法问题是目前网络空间最复杂、最敏感、争议最大的问题之一。⁴因此，本文将从禁止武力威胁或使用武力这一角度探讨约束网络攻击的困境，并据此思考对我国的启示。

诉诸战争权是指关于诉诸武力的可接受的法律理由，是对诉诸武力的起因的合法性的认定。从诉诸战争权的角度对网络攻击行为进行约束，前提是对网络攻击行为进行定性。具体而言，需要探讨两个关键问题：一是何种具体情境下的网络攻击行为会构成《联合国宪章》(以下简称《宪章》)第2条第4款所明确禁止的“使用武力”(use of force)行为。⁵二是何种具体情境下的网络攻击行为会被视为《宪章》第51条所定义的“武力攻击”(armed attack)，进而触发相关主体的自卫权。需要明确的是，国际法院在“尼加拉瓜案”的判决中指出，有必要将使用武力的最严重形式——构成武装攻击的形式，与其他不太严重的形式区分开来。对于未达到“武力攻击”程度的“使用武力”的行为，国家应诉诸非使用武力的其他合法的反击手段。⁶即可以认为“武力攻击”是“使用武力”这一行为在程度上更为严重的表现，两者不能等同。

二、界定“武力”的困境

《宪章》第2条第4款确立了禁止武力威胁或使用武力原则。这一原则在《宪章》首次被明确的以文字的形式规定下来，随后又在联大通过的不同决议中被多次重申。其中最具有代表性的是联大在1970年全体一致通过的《关于各国依联合国宪章建立友好关系及合作之国际法原则之宣言》(简称《国际法原则宣言》)。这一决议不仅将禁止武力威胁或使用武力原则位列于各原则之首，同时列举了这

³ 包括国家主权原则，禁止武力威胁或使用武力原则等。[美]迈克尔·施密特：《网络行动国际法塔林手册2.0版》，黄志雄等译，北京：社会科学文献出版社，2017年版，第57-58页。

⁴ 黄志雄：“国际法视角下的“网络战”及中国的对策——以诉诸武力权为中心”，《现代法学》，2015年第5期，第156页。

⁵ 即各会员国在其国际关系上不得使用威胁或武力，或以与联合国宗旨不符之任何其他方法，侵害任何会员国或国家之领土完整或政治独立。

⁶ Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. USA), Merits, Judgment, I.C.J. Reports 1986, p.14, paras.191, 195.

一原则包含的具体内容。1986年,国际法院在“尼加拉瓜案”判决中认为,禁止武力威胁或使用武力原则已经成为了公认的习惯国际法规则。可以说,禁止武力威胁或使用武力原则作为第二次世界大战后各国齐心协力凝结而成的重要成果之一,其构成了现代国际法的基石。然而,需要指出的是,虽然国际社会普遍接受第2条第4款限制了国家使用武力,但不同国家对该款的认识不同,在“武力”这一核心术语的具体定义和内涵上尚未形成统一且明确的见解,由此引出了许多争议。

（一）传统领域的“武力”界定之争

关于“武力”的界定,历史上存在狭义解释派和广义解释派。支持狭义解释的一方认为“武力”只限于武装力量。而支持扩大解释的一方认为“武力”包含武装和非武装的胁迫。由于《宪章》本质上属于多边性的国际公约,因此对其条款中的“武力”的解释应遵循条约解释的基本方法。条约解释主要依赖于《维也纳条约法公约》第31条的一般解释规则和第32条的补充解释规则。

首先,结合《宪章》文本对“武力”进行系统解释。第一,《宪章》序言第7款规定,接受原则,确立方法,以保证非为公共利益,不得使用武力。此处的“武力”为“武装力量”,在《宪章》的英文文本中的用词为“armed force”。条约序言往往包含了体现全约精神的条约目标和宗旨,并构成《维也纳条约法公约》第31条规定的“上下文”的一部分。⁷即《宪章》的正文内容是以序言为纲领,因此序言的用词含义在逻辑上应当宽于或者等于正文内容的用词含义。由此可知,《宪章》第2条第4款规定的“武力”范围应当等于或者小于“武装力量”。

其次,结合当事国嗣后所订关于条约之解释或其规定之适用之任何协定。最具有代表性的是《国际法原则宣言》,虽然该宣言并没有法律上的拘束力,但其被视为对《宪章》原则的权威解释。1962年12月18日,大会第十七届会议通过了第1815(XVII)号决议,确认了《宪章》规定的国际法原则十分重要,并决心加以研究。因此,大会决定将这一项目列入其第十八届会议的临时议程,以便重点研究包含禁止武力威胁或使用武力原则在内的四项原则。联大在第十八届会议上,决定设立关于各国建立友好关系和合作的国际法原则特别委员会(以下简称特别委员会),负责草拟报告,列明其研究的结论和建议。从特别委员会最终提交的报告详细内容来看,该委员会在面对部分国家所提出的创新性建议——即将国际法语境下“武力”一词的传统含义进行拓展,使其涵盖经济与政治层面所造成的损害——时,采取了消极的态度。委员会认为“武力”仅限于“武装力量”。

再次,结合《宪章》目的和宗旨。《宪章》第1条明确了联合国的宗旨是维护国际和平与安全、促进国际合作、成为协调各国行动的中心以及发展国际友好

⁷ 冯寿波:“论条约序言的法律效力——兼论 TRIPS 序言与《WTO 协定》及其涵盖协定之序言间的位阶关系”,《政治与法律》,2013年第8期,第102页。

关系。其中，维护国际和平与安全的核心是“采取有效集体办法，防止且消除对于和平之威胁制止侵略行为或其他和平之破坏”。这表明，若某种行为严重威胁国际和平，可能被纳入“武力”范畴，这给如网络攻击等新型威胁纳入传统“武力”框架提供了合法性依据。然而《宪章》并未对“和平之威胁”的具体标准做出明确规定，概念的模糊性导致了各国在“武力”界定问题上的分歧难以弥合。

最后，结合《宪章》的准备工作和缔约情况。1945年4月25日，联合国制宪会议在旧金山召开，经过了两个月的讨论才最终一致通过了《宪章》。根据联合国官方文件的记录，巴西代表团曾经提出“武力”应当包含“经济措施”（economic measures），但最终该提案被委员会以2票对26票否决。⁸

需要强调的是，无论适用哪种条约解释方法，对于条约的解释应当是善意的，这是贯穿条约解释的原则。但综上可知，结合不同的解释方法对“武力”的进行善意解释，无法得出确定的唯一的结论。

（二）网络空间的“武力”界定之争

从上文可知，禁止武力威胁或使用武力原则本身就存在一定的争议，而网络空间所具有的去边界性和全球性的特征，进一步加剧了这一原则在网络空间中适用的复杂性。何种具体情境下的网络攻击行为会受到禁止武力威胁或使用武力原则的约束，在国际社会上尚无定论。

针对这一难题，学者们各抒己见，形成了三种主要的学说，分别是“目标论”、“工具论”和“效果论”。“工具论”对于武力的界定在于实施攻击所使用的形式和手段。新型武器的形式和造成的结果与常规军事行为的形式和造成的结果越相似，其行动就越有可能构成“使用武力”或“武装攻击”。所以，只要某一网络攻击行为能导致类似于传统军事行动的后果，就能认定该网络攻击行为属于使用武力。“目标论”则将焦点放在网络攻击的对象上——关键基础设施（critical infrastructure）。按照这一理论，任何针对关键基础设施的网络攻击，都将被视为“使用武力”行为。“效果论”的支持者认为，相较于网络攻击的形式和攻击对象，国家更加关注的是网络攻击所带来的后果的严重性。基于此，这一派的学者通过提出评估网络攻击后果的标准，来判断某一网络攻击行为是否足以构成“使用武力”。三种学说近年来出现了融合的趋势，且“效果论”逐渐主流化。⁹最具代表性的效果评价标准是由美国学者迈克尔·施密特（Michael N. Schmitt）提出的八项因素：严重性、迅即性、直接性、侵入性、效果的可衡量性、军事性、国家介入程度和推定合法性。¹⁰

然而，“效果论”本身也存在不足之处。一是效果的判断存在较大的自由裁量

⁸ Documents of the United Nations Conference on International Organization, San Francisco, 1945=Documents de la Conférence des Nations Unies sur l'organisation internationale, San Francisco, 1945, UN Conference on International Organization, Vol.6, paras.334-335.

⁹ 张华：“网络空间适用禁止使用武力原则的法律路径”，《中国法学》，2022年第2期，第287页。

¹⁰ [美]迈克尔·施密特：《网络行动国际法塔林手册2.0版》，黄志雄等译，北京：社会科学文献出版社，2017年版，第338-341页。

空间。首先，严重性的认定存在模糊地带。施密特认为，严重性在判断网络攻击是否达到使用武力的门槛的分析中占据主导地位，后果的范围、持续的时间和烈度对评估严重性有重要的作用。但随之而来的问题是，严重性是否仅以人员伤亡或者财产损失等物理性结果为标准。如对重要的国际证券交易所实施网络攻击进而引发市场瘫痪，这并没有人员和财产等物理上的损害，但对整个证券市场造成了广泛的负面影响，是否属于造成了严重影响。此外，施密特认为，国家在进行效果与影响的评估时，不应局限于既定的、单一的标准框架，而是应当根据实际情况的复杂性和多样性，灵活地将诸如目标性质等其他相关因素纳入评估体系之中。这无疑进一步模糊了评估标准的边界，使得在评估过程中需要权衡和考虑的因素更加多元和复杂。二是“效果论”所包含的评价标准大多属于事后分析性质，因此在网络攻击实际发生的时刻难以迅速的为遭受网络攻击的国家提供所需的及时的帮助。三是鉴于网络攻击有时具有隐蔽性的特性，其带来的后果可能是长远的或累积的。因此，如何准确判断某一网络攻击行为与长远影响之间的因果关系，从而防止在认定过程中出现滥用现象，也是“效果论”有待回答的问题。

三、行使自卫权的困境

根据《宪章》第 51 条的规定，一国在受到外国武力攻击时，有实施单独的或集体的武装自卫行动的权利，即自卫权。当网络攻击达到武力攻击的门槛时，受到网络攻击的国家可以行使自卫权。对于这一门槛，并没有明确的国际法层面上的规定。《塔林手册 2.0 版》规则 71 认为，网络行动是否构成武力攻击的决定性因素是其规模和效果，辅助的判断因素有“跨境”、“起源”及其“攻击目标”。¹¹其中起决定性作用的“规模和效果标准”是由国际法院在“尼加拉瓜案”中的裁决提出的。但该标准存在与上述的界定“使用武力”的“效果论”一样的困境，即有较强的模糊性和不确定性。有学者认为这两项标准不是现行法，并非一定要遵守。实际上提出这两项标准的国际法院在之后涉及自卫权的裁决中也没有严格坚持“规模与效果标准”。¹²此外，即使某一网络攻击行为达到了武力攻击的程度，被攻击的国家在行使自卫权时仍会面临诸多挑战。

（一）难以认定自卫权的行使对象

首先是主体界定的模糊性。根据传统国际法，自卫权的行使对象仅限于主权国家或可归因于主权国家的非国家行为体（non-State actors）。然而，在网络空间中，网络空间虚拟性和高度匿名的特征，使追溯攻击源头并确定责任归属变得异常艰难，从而使得受害国难以有效行使自卫权。溯源难的主要因素在于技术层面

¹¹ [美]迈克尔·施密特：《网络行动国际法塔林手册 2.0 版》，黄志雄等译，北京：社会科学文献出版社，2017 年版，第 344 页。

¹² 张华：“网络空间适用自卫权的法律不确定性与中国立场表达——基于新近各国立场文件的思考”，《云南社会科学》，2021 年第 6 期，第 85 页。

的限制。IP 地址的定位是追踪网络攻击来源的关键步骤。然而,在网络空间中,网络攻击的发起者出于逃避法律责任等多方面的原因,往往会运用一系列技术手段,试图掩盖自己的真实身份信息,使得追踪和打击网络攻击行为变得极为困难。近年来,APT (Advanced Persistent Threat, 即高级持续性威胁)作为一种典型的网络攻击方式,因其高隐匿性而备受瞩目。与 DDoS 等攻击方式相比,APT 通过频繁使用跳板主机、跳板网络和公共网络服务进行网络攻击,使得追踪攻击源头变得更加困难。

需要注意的是,即使成功实现溯源,归因依然是一大难题,即国家是否应当就网络攻击行为承担国际法律责任。而法律层面的制约是归因困难的主要缘由。2001 年国际法委员会二读通过了关于《国家对国际不法行为的责任条款草案》(以下简称《责任条款草案》),其中《责任条款草案》将行使政府权力要素的个人或实体的行为和受到国家指挥或控制的行为等非国家行为体的行为列为可归因于国家的行为。但这在现实中面临着极大的挑战。实际上,当一国利用黑客等非国家行为体实施网络攻击时,其目的往往在于隐藏自身身份,从而规避国际法律责任。在这种情况下,要求其提供证明非国家行为体得到正式授权的证据,显然是不切实际的。而对于受到国家指挥或控制的行为,草案并没有提供具体的标准,但在实践中存在两种判定标准,即“有效控制”(effective control)标准与“整体控制”(overall control)标准。然而,应该适用哪一种标准,国际社会目前尚未达成统一的认识或结论。¹³

鉴于上述困境,学术界出现了新的声音并展开了激烈的讨论,核心在于对那些行为无法归因于国家的非国家行为体,是否应被纳入自卫权行使的潜在对象范围。有学者认为在有限的条件下——即“不能够或不愿意理论”(The Unwilling and Unable)成立的时候,无法归因于国家的非国家行为体可以成为自卫权行使的对象。该理论在《塔林手册 2.0 版》中有明确的定义,即当满足必要性原则,且网络武力攻击的防御仅能通过自卫实现,同时领土国无法或不愿采取有效措施来阻止此类攻击时,对网络武力攻击采取自卫行动是被允许的。¹⁴但“不能够或不愿意理论”依然存在主观性过强的问题,这实际上是对现存国际法赖以存在的基础——国家主权原则和不干涉原则是个严重的挑战,而且存在被滥用的风险。

(二) 难以把握自卫权的行使时间

现行国际法认为,行使自卫权的时间节点是一国在受到外国武力攻击时,但网络空间具有虚拟性和去边界性,这使得网络攻击不同于传统的武力攻击,不受时间和空间的限制,使“受到武力攻击”的认定复杂化。基于此,有国家认为网络

¹³ 张华:“论非国家行为体之网络攻击的国际法律责任问题——基于审慎原则的分析”,《法学评论》,2019 年第 5 期,第 161 页。

¹⁴ [美]迈克尔·施密特:《网络行动国际法塔林手册 2.0 版》,黄志雄等译,北京:社会科学文献出版社,2017 年版,第 349 页。

攻击的瞬时性使各国难以防范，从而提出了“预先性自卫”（anticipatory self-defense）或“预防性自卫”（preventive self-defense）的理论。

自 2001 年“9·11”事件后，美国在 2002 年的《国家安全战略》中表示，出于常识和自卫，其有权在新出现的威胁完全形成之前采取行动，即采取预防性自卫行动。预先性自卫和预防性自卫在实践中有时会被混淆，但两者存在本质的区别。预先性自卫和预防性自卫间本质的区别在对紧迫性的要求不同，前者针对的是紧迫的武力攻击（imminent attacks），后者是针对非紧迫的武力攻击。由于预防性自卫几乎不受限制，能否行使自卫权的判定标准皆由一方控制，因此难以受到国际社会的普遍认同。如法国虽然在其发布的有关网络空间国际法适用的立场文件中强调其有权进行先发制人的自卫行动，但前提是构成武力攻击的网络攻击是迫在眉睫的。并且明确否认了预防性自卫的合法性。¹⁵相较而言，预先性自卫似乎更有合理性和正当性，因此也更容易被接受。

然而，预先性自卫理论也具有明显的不足之处，这些缺陷削弱了其正当性和在实践中的可操作性。首先，关于紧迫性的判定标准亦模糊不清。如上文所述，多数国家认为行使预先性自卫的前提是受到紧迫的武力攻击。但界定紧迫性的标准，目前国际社会并没有定论。《塔林手册 2.0 版》规则 73 解释 4 中指出，大多数专家赞成“最后可行的机会之窗”这一标准，并认为关键在于一国能否合理预见如未能在特定时间做出反应，一旦攻击开始，该国将无法有效自卫。¹⁶但“合理遇见”这一表述过于模糊，具有强烈的主观性和不确定性。此外，预先性自卫面临一个核心悖论：如何界定网络攻击是否构成武力攻击，这是预先性自卫得以实施的前提。正如上文所述，预先性自卫行使的前提是他国的网络攻击达到武力攻击的程度。但难点在于，如何在网络攻击实际发生前，准确预判其是否会升级为武力攻击。这一情况可能会引发不良后果，即自卫权在实践层面可能出现无序扩大的趋势，甚至可能演变为被肆意滥用的局面。一旦自卫权被不当利用，就极有可能成为某些国家侵犯他国主权的冠冕堂皇的托词。而这种行为从根本上违背了自卫权制度设计的初衷。

根据《宪章》第 51 条的规定，会员国行使的涉及自卫权的措施，应当立即报告给联合国安理会，且该措施只能在安理会采取必要办法之前实施。这体现了安理会在国家行使自卫权方面的重要作用。不少国家在针对“国际法如何适用于网络空间”这一问题的立场文件中表明了向安理会的报告义务。德国在其立场文件中进一步认为，网络攻击是否构成武力攻击的决定不应当由此类恶意网络行动的国家受害者自行决定的，而是需要根据《宪章》第 51 条向国际社会（即向联

¹⁵ INTERNATIONAL LAW APPLIED TO OPERATIONS IN CYBERSPACE, 9 September 2019, available at: <https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf>, para.1.2.2.

¹⁶ [美]迈克尔·施密特：《网络行动国际法塔林手册 2.0 版》，黄志雄等译，北京：社会科学文献出版社，2017 年版，第 352 页。

联合国安理会）进行报告，并由安理会进行判断。¹⁷

四、中国因应

当前，围绕禁止武力威胁或使用武力在网络攻击场景中的适用性问题，国际社会呈现出规则博弈与立场分化的复杂图景。这一争议不仅折射出网络空间国际秩序构建的阶段性特征，更揭示了各国在数字时代安全利益与规则话语权领域的深层博弈。在此背景下，中国需以系统性思维统筹应对。在明晰各国在“武力”定性、自卫权行使条件等关键议题上的差异化立场的基础上，我国应当以“网络空间命运共同体”理念为指引，依靠联合国等国际组织，重视软法的作用并积极参与制定网络空间的国际新规则。

（一）秉承网络空间命运共同体

通过对约束网络攻击的困境进行系统性分析，可以发现网络空间日益军事化的显著趋势。这一趋势主要体现在自卫权适用对象与行使时机的双重扩张。此外，相关国家的实践也表明了这一趋势。自美国提倡“朝前防御”（defend forward）这一进攻性网络安全战略后，不少西方国家紧随其步伐，网络空间军事化的倾向日益明显。¹⁸如韩国于2024年9月发布的《韩国国家网络安全基本计划》中强调其将开展进攻性网络防御行动以确保国家网络安全。然而，网络空间军事化的加剧只会使无形的战争在网络这一新领域不断蔓延，导致网络空间的无序状态，这既不利于网络空间国际治理体系的完善，也不利于全人类的共同发展。网络空间的包容性和开放性，促使世界上的大部分国家紧密连接，同时也让风险跨越国界，呈现出更为显著的全球性特征。因此，应当秉承网络空间命运共同体的理念，尊重各国的网络主权、维护网络安全、促进国家间的开放合作、构建良好的秩序。2024年世界互联网大会乌镇峰会中，习近平总书记强调，当前我们正置身于科技革命蓬勃推进、产业变革日新月异的时代洪流之中。这一时期，既为全人类带来了前所未有的发展机遇，也伴随着诸多挑战。因此，我们应当共同努力，携手创造网络空间命运共同体，以期共创一个更加繁荣的“数字未来”，确保互联网能够更大程度地惠及人民、造福全球。

（二）依托联合国等国际组织

秉承网络空间命运共同体理念，需要以《宪章》的宗旨和原则为基础，坚持以联合国作为协调各国行动的中心，促进国际合作，共同应对数字化、网络化时代的全球性风险。

¹⁷ Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution 73/266,A/76/136, 13 July 2021.

¹⁸ 张华：“网络空间适用自卫权的法律不确定性与中国立场表达——基于新近各国立场文件的思考”，《云南社会科学》，2021年第6期，第84页。

从 1998 年开始,联合国就对信息和通信技术(information and communications technologies, 以下简称 ICTs)安全有所关注。2004 年开始,联合国设立联合国政府专家组(the Group of Governmental Experts, 简称 UNGGE),开展在政府间专家层面的使用 ICTs 带来的威胁并且应当如何应对这些国际安全问题的讨论。六届 UNGGE 一共进行了六次实质性会议,并在第 2、3、4、6 届 UNGGE 中分别达成四份实质性报告。其中 2012 年至 2013 年召开的第 2 次会议达成的 A/68/98*号报告中说明国际法尤其是《宪章》适用于网络空间,对维持和平和稳定至关重要。而在 2014 年至 2015 年达成 A/70/174 号报告中进一步明确了禁止武力威胁或使用武力原则在网络空间领域的可适用性,并且认为各国应当在法律、技术和政策层面共同讨论和发展相关概念。2018 年 12 月,联大通过第 A/RES/73/27 号协议设立了不限成员名额的信息安全开放式工作组(Open-Ended Working Group, 简称 OEWG),旨在提供开放包容的讨论平台,让所有联合国成员国均有权参与网络空间安全治理。在 OEWG 第五次实质性会议报告中,重申了各国应遵守禁止武力威胁或使用武力原则。

在联合国框架内,ICTs 安全问题正有条不紊的推进磋商,未来国际社会也会继续在这一框架内深入探讨,力求达成共识。虽然两者的报告没有强制约束力,但国际社会普遍认可在网络空间领域应以联合国框架下的 OEWG 和 UNGGE 的报告为指南的观点。因此近年来,我国多次在 OEWG 和 UNGGE 框架内表达意见。如在 OEWG 框架内,我国提交的“中国对主权原则在网络空间的适用的看法”;“中国提交的《全球数据安全倡议》”;“中国对网络空间国际规则制定的立场”等。我国应当继续坚持以联合国为核心的多边合作框架,深度参与禁止武力威胁或使用武力约束网络攻击中的适用性问题的磋商和规则制定,表达中国声音和中国立场。

除依托联合国外,目前,我国也依托上海合作组织(以下简称上合组织),通过深化网络安全领域的合作,为区域和全球网络安全贡献力量。首先是建立网络安全合作方面的法律框架。在上合组织第九次峰会上,各国元首签署了《上海合作组织成员国保障国际信息安全政府间合作协定》(以下简称《合作协定》),针对国际信息安全领域¹⁹的主要威胁,基本术语和合作方向等内容进行了规定。2020 年,上合组织成员国表示愿加强协作以全面落实 2009 年签署的《合作协定》。²⁰其次,上合组织积极与联合国在信息安全领域进行合作。2011 年 9 月 12 日,上合组织的四个成员国(中国、俄罗斯、塔吉克斯坦和乌兹别克斯坦)向联大提

¹⁹ 根据《合作协定》的文本可知信息安全主要涉及国际网络安全。《合作协定》将信息安全定义为:个人、社会、国家及其利益在信息空间处于受保护状态,免受威胁、破坏和其他负面影响。

²⁰ “上海合作组织成员国元首理事会关于保障国际信息安全领域合作的声明”,《人民日报》,2020 年 11 月 11 日第 3 版。

交了《信息安全国际行为准则》。²¹在吸纳国际社会的合理建议后，2015年1月9日，上合组织的六个成员国（中国、哈萨克斯坦、吉尔吉斯斯坦、俄罗斯、塔吉克斯坦和乌兹别克斯坦）再次向联大提交了新版的《信息安全国际行为准则》。²²这一系列举措，充分彰显了上合组织作为具有重要影响力的国际组织在国际网络安全领域所发挥的积极作用。

（三）规则：重视软法，参与制定网络空间国际新规则

当前，鉴于各国在网络空间治理领域的利益和需求存在显著差异，网络空间的国际法规则体系尚不健全，部分具体细分领域内更是存在规则空白的情况。在此背景下，发挥主要引导作用的是不具备法律强制力的国际软法规则。如《塔林手册 2.0 版》就是国际法学者在理论层面探索国际法适用于网络空间的一次重要尝试，在若干方面体现了网络空间国际规则制定的趋向。虽然其不具有普遍的法律效力，但对目前网络空间的国际化治理和构建全球网络空间治理秩序有着十分重要的参考和借鉴作用。然而，目前关于网络空间治理的话语权依然是西方主导，这导致由其参与主导的“软法”以西方国家的国家利益为主，并没有充分考虑发展中国家的利益，不利于建立平等的治理秩序。以《塔林手册 2.0 版》为例，整个手册的启动、起草到最后出版的过程都是由西方国家主导的。因此，在该手册的最终版本中，国际专家组的评注大多从西方国家的利益视角出发，存在浓厚的西方主导的色彩。如在自卫权方面，手册未能正视“预先性自卫”存在的理论争议及其在实践中的不可行性。此外，在关于网络空间的国际法规范较为缺乏的情况下，一些国家试图通过将《塔林手册 2.0 版》作为跳板，通过不断推广和应用手册的观点，逐步将其转化为国际习惯，从而达到其政治目的。总之，《塔林手册 2.0 版》这种倾向和其潜在的目的值得深思。因此，我国积极参与网络空间国际软法规则的制定，促进平等的国际网络空间安全治理秩序的形成。

另外，除了重视“软法”的作用，基于网络空间的虚拟性、去边界性和全球性等特点，还必须认识到现有的国际法未必能与网络空间完全契合，可能会产生“移植排斥反应”。因此，制定关于网络空间的新规则有其必要性，同时也有可行性。²³我国应坚持以《宪章》为指导，在联合国框架下构建具有广泛代表性的网络空间新规则。如 2024 年 12 月 24 日，联大通过了《联合国打击网络犯罪公约》，这对于打击网络犯罪，促进网络空间领域的国际合作具有里程碑式的意义，该公约

²¹ Letter dated 12 September 2011 from the Permanent Representatives of China, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General, UN General Assembly Document A/66/359, September 14, 2011.

²² Letter dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General, UN General Assembly Document A/69/723, January 22, 2015.

²³ 可行性包括：新兴国家的努力；西方国家的潜在倾向，即通过形成国际习惯来制定新规则；网络空间构建国际秩序的客观需要。黄志雄：“国际法在网络空间的适用：秩序构建中的规则博弈”，《环球法律评论》，2016 年第 3 期，第 11-13 页。

是我国、俄罗斯等非西方国家推动制定网络空间新规则的重要成果。

五、结语

网络化与信息化浪潮在重塑全球发展的同时，也催生了网络攻击这一非传统安全威胁。当前，网络攻击的跨国性、隐蔽性与破坏性日益凸显，尽管国际社会普遍认同将现有国际法规则延伸至网络空间适用，但围绕诉诸战争权视角下的“武力”界定与自卫权行使仍然存在较大分歧。针对当前全球网络空间治理所面临的复杂形势与诸多挑战，我国应当以高度的责任感和使命感，运用系统性思维对各类问题与挑战进行全面统筹和科学应对。我们要从全局出发，提出具有前瞻性、战略性和可操作性的措施，为全球网络空间治理提供中国智慧和方案，贡献中国力量。

Integration of Soft and Hard Laws: Profiling Legal Protection for “AI for All”

Yang Xiao and Xiaoxia Sun*

ABSTRACT

This paper analyzes the concept of “AI for All” (AIFA), exploring its theoretical foundations, challenges, and pathways for legal implementation. AIFA encompasses three key dimensions: “AI Technology for All”, “AI Justice for All”, and “AI for All Countries”. However, the rise of generative AI—with costly subscription models—threatens these ideals, exacerbating economic and cognitive disparities. Currently, AIFA governance relies on non-binding soft laws, which lack enforceability, while hard laws remain underdeveloped. To prevent AIFA from a utopian vision descending into a dystopian reality, this paper argues for a dual approach: (1) expanding soft laws to explicitly address affordability, literacy, and international cooperation, and (2) tailoring hard laws based on national AI capacity—ranging from pricing regulations in core AI states to global cooperation for peripheral nations. By integrating soft law’s ethical guidance with hard law’s enforcement, AI governance can balance innovation with inclusivity, ensuring AI benefits all rather than deepening inequalities.

Keywords: AI for All; AI Governance; AI Justice; AI Regulation; AI Policy; Global AI Frameworks; International AI Law

1. INTRODUCTION

“AI for All” broadly refers to the vision that everyone – regardless of geography, socioeconomic status, age, or gender – should share in AI’s benefits. The Indian government framed “AI for All” as the benchmark for future AI development, focusing on inclusive growth and access to AI in key sectors.¹ Internationally, UNESCO echo this vision: “The vision of ‘AI for all’ must ensure that everyone can partake in the ongoing technological revolution and reap its benefits”.² Beyond the direct concept of “AI for All,” related concepts, such as “inclusive AI” and “accessible AI,” “justice AI”, complement and overlap with “AI for All.” Inclusive AI focuses on fairness,

*Corresponding author: Yang Xiao, Guanghua Law School, Zhejiang University, Hangzhou, China, E-mail: xiaoyangp2024@126.com. <https://orcid.org/0009-0009-6664-5569>

Xiaoxia Sun, Digital Rule of Law Research Institute, Zhejiang University, Hangzhou, China, E-mail: sunxiaoxia@fudan.edu.cn

¹ NITI Aayog, National strategy for artificial intelligence (Government of India, 2018a).

² United Nations Educational, Scientific and Cultural Organization (UNESCO), "Third International UNESCO Model on AI has been launched for associations and clubs of the UNESCO movement," (2024).

representation, and non-discrimination in AI systems.³ Accessible AI emphasizes usability specifically for people with disabilities.⁴ However, distinctions between these terms are often relative, and they frequently overlap or are used interchangeably.⁵ Clearly delineating “AI for All” remains a research gap.

While some consensus has been achieved regarding “AI for All”, little legislation to protect the idea, leaving to a kind of contradiction in status quo. Documents such as the Ethical Guidelines for Artificial Intelligence, the Global Digital Compact, and China’s Ethical Norms for a New Generation of Artificial Intelligence have all included the notion of “AI for All” as a fundamental principle. However, these remain at the level of declarative soft law, lacking legal protections. Meanwhile, the digital divide in the AI era continues to widen, exacerbating wealth inequality and the Matthew effect. A case in point is OpenAI’s membership strategy: its most advanced models require a \$200/month subscription; the Deep Research feature—capable of functioning at a PhD-assistant level—is available only twice per month for free users but up to 100 times for paying users at the \$200/month tier. OpenAI is further exploring a “Pro Max” tier at \$2,000/month. With advanced AI models now surpassing even skilled engineers in capability, the gap between free and cutting-edge models is becoming increasingly significant. Consequently, the ability to utilize these advanced AI tools will profoundly influence social mobility.

Against this backdrop of rapid generative AI and AGI advancements, coupled with emerging paid subscription models and signs of a widening digital divide in the AI era, this paper researches three questions:

1. What’s the meaning of “AI for All”, what's the theoretical foundation of it?
2. Is the current situation—characterized mainly by soft-law declarations advocating “AI for All”—sufficient?
3. How can we establish legal protections for the principle of “AI for All,” effectively integrating soft and hard laws to bridge the gap between innovation and development?

2.LITERATURE REVIEW

This section will provide a brief introduction to current research on “AI for All,” as well as soft and hard laws associated with it. Specifically, it will first elaborate on the definition and necessity of “AI for All”; second, it will clarify the relationship between soft laws and hard laws.

³ Toronto Declaration. (n.d.). Inclusive AI must be shaped by a diverse range of voices in all areas from strategy development, data gathering, algorithmic design, implementation, and user impact, with the goal of respecting the right to live free from discrimination. Retrieved from <https://www.torontodeclaration.org/declaration-text/english/>.

⁴ M. R. Morris, "AI and accessibility," *Communications of the ACM* 63, no. 6 (2020): 35–37.

⁵ T. Avellan, S. Sharma, and M. Turunen, "AI for all: Defining the what, why, and how of inclusive AI," In *Proceedings of the 23rd International Conference on Academic Mindtrek* (2020): 142–144.

2.1 The Definition of “AI for all”

Shawn Schuster described “AI for All” as “everyday people can benefit from Artificial Intelligence”.⁶ However, while most scholars mention the term “AI for All” in their papers, they do not focus on its meaning but rather discuss related derivative questions. Many scholars conflate or equate “AI for All” with diversity and inclusion in AI (D&I in AI), arguing that AI should be free from bias, discrimination, and perceived untrustworthiness.⁷ And defined “inclusive AI” as the “inclusion of humans with diverse attributes and perspectives in the data, process, system, and governance of the AI ecosystem”. Diversity refers to the “representation of the differences in attributes of humans in a group or society”.⁸ In sum, compared to concepts such as inclusive AI,⁹ diversity AI¹⁰ or accessibly AI, few scholars have systematically explored the idea of “AI for All.”

2.2 The Necessity of “AI for all”

“AI for All” is a necessary condition for social justice. A scholar through the lens of the theory of social justice known as the capability approach to conceptualize AI’s transformative role and its socio-political implications, concludes that AI itself should be considered among the conditions of possession and realization of the capabilities it transforms. In other words, access to AI—in the many forms this access can take—is necessary for social justice.¹¹ Another scholar, using the veil of ignorance to determine principles for aligning AI systems, found that participants expressed a preference for the concept of “AI for all”—a principle that prioritizes the most vulnerable groups when designing AI assistants.¹² A scholar further analyzes the applicability of Rawls’s principles of fair equality of opportunity and the difference principle to AI. Regarding fair equality of opportunity, AI should not merely pursue formal fairness; rather, it should substantively reduce social inequalities, such as by enhancing social mobility. Concerning the Difference Principle, AI systems should prioritize improving the well-

⁶ S. Schuster, AI for all: How everyday people can benefit from artificial intelligence (Shawn M. Schuster, 2023).

⁷ See Avellan, T., Sharma, S., & Turunen, M. (2020, January). AI for all: Defining the what, why, and how of inclusive AI. In Proceedings of the 23rd International Conference on Academic Mindtrek (pp. 142–144); Shams, R. A., Zowghi, D., & Bano, M. (2023a). Challenges and solutions in AI for all. [Preprint]. arXiv:2307.10600; Zowghi, D., & Bano, M. (2024). AI for all: Diversity and inclusion in AI. *AI and Ethics*, 4(4), 873–876; Shams, R. A., Zowghi, D., & Bano, M. (2023b). AI for all: Operationalising diversity and inclusion requirements for AI systems. [Preprint]. arXiv:2311.14695; Shams, R. A., Zowghi, D., & Bano, M. (2023c). AI for all: Identifying AI incidents related to diversity and inclusion. [Preprint]. arXiv:2311.14696

⁸ Zowghi, D., & Bano, M. (2024). AI for all: Diversity and inclusion in AI. *AI and Ethics*, 4(4), 873–876.

⁹ Fosch-Villaronga, E., & Poulsen, A. (2022). Diversity and inclusion in artificial intelligence. In B. Custers & E. Fosch-Villaronga (Eds.), *Law and artificial intelligence: Regulating AI and applying AI in legal practice* (pp. 109–134). T.M.C. Asser Press.

¹⁰ Shams, R. A., Zowghi, D., & Bano, M. (2025). AI and the quest for diversity and inclusion: A systematic literature review. *AI and Ethics*, 5(1), 411–438.

¹¹ Buccella, A. (2023). “AI for all” is a matter of social justice. *AI and Ethics*, 3(4), 1143–1152.

¹² Weidinger, L., McKee, K. R., Everett, R., Huang, S., Zhu, T. O., Chadwick, M. J., Summerfield, C., & Gabriel, I. (2023). Using the veil of ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences*, 120(18), e2213709120.

being of society's most disadvantaged groups, rather than solely optimizing overall efficiency.¹³

2.3 *The Relationship Between Soft and Hard Laws*

Disagreements prevail in the existing scholarship on their relationship between soft and hard laws, including the ontology of soft laws and hard laws, as well as their connections and distinctions. In terms of the ontology, positivist legal scholars tend to deny the concept of "soft law," since law by definition is "binding".¹⁴ Rational institutionalist scholars respond that "the term 'binding agreement' in international affairs is a misleading hyperbole".¹⁵ Constructivist scholars, in contrast, focus less on the binding nature of law at the enactment stage, and more on the effectiveness of law at the implementation stage, addressing the gap between the law-in-the-books and the law-in-action; they note how even domestic law varies in terms of its impact on behavior, so that binary distinctions between binding "hard law" and nonbinding "soft law" are illusory.¹⁶

Regarding the connections and distinctions between soft law and hard law, can be divided into domestic and international level. At the domestic level, soft law is often applied in specific situations, such as emergencies, when there is insufficient time or when it is inconvenient to enact hard law.¹⁷ Furthermore, soft law influences behavior by informing the public and political institutions about intentions and policy preferences.¹⁸ At the international level, soft laws have become increasingly prevalent and important, significantly affecting international relations.¹⁹ Functionalist scholars argue that hard law increases the costs for states violating legal commitments,²⁰ whereas soft law provides greater flexibility, lower costs, and reduced sovereignty costs in sensitive areas.²¹

¹³ Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, 151(2), 218–231.

¹⁴ Weil, P. (1983). Towards relative normativity in international law? *American Journal of International Law*, 77(3), 413–423.

¹⁵ Lipson, C. (1991). Why are some international agreements informal? *International Organization*, 45(4), 495–538.

¹⁶ Trubek, D. M., Cottrell, P., & Nance, M. (2006). 'Soft law,' 'hard law,' and European integration: Toward a theory of hybridity. In G. de Búrca & J. Scott (Eds.), *Law and new governance in the EU and the US* (pp. 65–91). Hart Publishing.

¹⁷ Daly, S. (2021). The rule of (soft) law. *King's Law Journal*, 32(1), 3–13.

¹⁸ Posner, E., & Gersen, J. (2008). Soft law [Working Paper No. 213]. University of Chicago Public Law & Legal Theory Working Paper.

¹⁹ Guzman, A. T., & Meyer, T. L. (2010). International soft law. *Journal of Legal Analysis*, 2(1), 171–225.

²⁰ Trubek, D. M., Cottrell, P., & Nance, M. (2006). 'Soft law,' 'hard law,' and European integration: Toward a theory of hybridity. In G. de Búrca & J. Scott (Eds.), *Law and new governance in the EU and the US* (pp. 65–94). Hart Publishing; Raustiala, K., & Victor, D. G. (1998). Conclusions. In D. G. Victor, K. Raustiala, & E. B. Skolnikoff (Eds.), *The implementation and effectiveness of international environmental commitments: Theory and practice* (pp. 659–707). MIT Press.

²¹ Lipson, C. (1991). Why are some international agreements informal? *International Organization*, 45(4), 495–538; Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International Organization*, 54(3), 421–456; Sindico, F. (2006). Soft law and the elusive quest for sustainable global governance. *Leiden Journal of International Law*, 19(3), 829–846.

3. Systemizing “AI for All”

3.1 The Content of “AI for All”

In contrast to concepts such as “inclusive AI” or “accessible AI,” there has not yet been a comprehensive or systematic definition of “AI for All,” and at times, the term is even conflated with “inclusive AI,” “accessible AI,” or “human-centered AI.” Therefore, it is necessary to formulate a more systemic framework to clarify what “AI for All” means. Building on numerous papers, protocols, and contracts—across international, domestic, authoritative, and private sectors—directly or indirectly express the idea of “AI for All”, this part aims to systematize the idea of “AI for All” by dividing it into three dimensions: AI technology for all, AI justice for all, and AI for all countries.

3.1.1 AI Technology for All

“AI technology for all” implies that AI should be both open and affordable to everyone. Openness suggests that AI should be made as open-source as possible and not monopolized by a small number of technology giants; at the same time, it necessitates improving the public’s understanding of AI. Affordability means that the cost of deploying and using AI should not be set so high that only a small elite can bear it. This approach is reflected in various initiatives. For example, the nonprofit organization AI4ALL focuses on educating and empowering underrepresented groups to ensure broad benefits of AI;²² Greece launched a similar “AI for All” initiative to train civil servants in both basic and advanced AI skills.²³ and the Bill & Melinda Gates Foundation has emphasized “ensuring access to AI for all,” particularly in low- and middle-income countries.²⁴ Furthermore, many nations’ policies and legal frameworks stress making AI technology open and affordable, reflecting the principle of “AI technology for all.” For example, the European Union has funded initiatives such as “AI4EU” to encourage the development of inclusive AI applications and, through its digital policies, has encouraged member states to strengthen AI-related education and training in order to narrow digital divides.²⁵ Beyond EU-wide legislation, individual member states have referred to ensuring that “no one is left behind” in their national AI strategies, as exemplified by Finland’s free AI education program for its entire population. The Global Digital Compact likewise emphasizes open data, open AI models, open standards, and open content, ensuring that societies and individuals can use digital technologies for their developmental needs.²⁶ The Asilomar AI Principles

²² AI4ALL. (2025). AI4ALL: Opening doors to artificial intelligence for historically excluded talent. <https://ai-4-all.org/>.

²³ Panagopoulos, A. (2024, August 9). AI for All: An initiative of the Ministry of Interior in cooperation with Google and EKDDA. Digital Skills and Jobs Platform. <https://digital-skills-jobs.europa.eu/en/inspiration/good-practices/ai-all-initiative-ministry-interior-cooperation-google-and-ekdda>.

²⁴ Bill & Melinda Gates Foundation. (2025). AI equity: Ensuring access to AI for all. Bill & Melinda Gates Foundation. <https://www.gatesfoundation.org/ideas/science-innovation-technology/artificial-intelligence>.

²⁵ High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.

²⁶ United Nations. (2024). Global Digital Compact: Open data, AI models, standards, and content for development. United Nations. <https://www.un.org/global-digital-compact/>.

also underscore that AI technologies should benefit and empower as many people as possible.²⁷

3.1.2 AI Justice for All

“AI justice for all” underscores both formal justice and substantive justice. Formal justice requires that AI not discriminate against specific populations, demanding the prevention of bias and the fair treatment of all users, while substantive justice calls for AI not to exacerbate social inequality but instead to provide equitable opportunities for people of all social classes, races, genders, and economic backgrounds. This broader notion of substantive justice aligns with ideas from thinkers such as John Rawls, who emphasizes fairness in opportunity distribution²⁸. India’s 2018 National AI Strategy, spearheaded by the NITI Aayog, exemplifies this emphasis on substantive justice: its slogan “AI for All” aims to use AI in driving sustainable and inclusive growth that transcends class, geography, and economic constraints.²⁹ Besides, many discussions similarly demonstrated the notion of “AI justice for all”. The 2019 “Beijing Consensus on AI and Education” encourages countries to ensure that the design of AI adheres to ethical norms, avoids discrimination, and remains fair, transparent, and auditable, thereby reflecting the significance of justice in AI for all.³⁰ In 2019, the OECD adopted AI governance principles, and were subsequently recognized and adopted by the G20. The very first principle is “inclusive growth, sustainable development, and well-being,” urging stakeholders to promote AI in a responsible manner that benefits all and highlights substantive justice.³¹ IEEE’s “Global Initiative on Ethics of Autonomous and Intelligent Systems,” launched in 2016, published its first version of “Ethically Aligned Design” in 2019. This guide promotes human-centered, welfare-oriented AI design concepts and emphasizes fairness, transparency, accountability, and thus both include the notion of formal and substantive justice for “AI for ALL”.³² In the same year, the EU High-Level Expert Group on AI proposed a “human-centric AI” approach that respects European values and human rights, its guidelines for “Trustworthy AI” outline seven key requirements, in which diversity and non-discrimination, as well as social well-being, echo “AI justice for all”.³³

3.1.3 AI for All countries

²⁷ Future of Life Institute. (2017). Asilomar AI Principles. <https://futureoflife.org/open-letter/ai-principles/>.

²⁸ Rawls, J. (1999). A theory of justice (Rev. ed.). Harvard University Press.

²⁹ NITI Aayog. (2018). National strategy for artificial intelligence. Government of India. <https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf>

³⁰ United Nations Educational, Scientific and Cultural Organization (UNESCO). (2019). Beijing consensus on artificial intelligence and education.

³¹ Organization for Economic Co-operation and Development. (2019). OECD AI principles. <https://www.oecd.org/going-digital/ai/principles/>.

³² Institute of Electrical and Electronics Engineers. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (1st ed.). IEEE.

³³ High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.

“AI for All countries,” especially developing countries, is equally critical. Global justice theorists like Thomas Pogge and Charles Beitz believed the need for equitable distribution of resources and technology across national boundaries to prevent new forms of deprivation in less developed regions.³⁴ UNESCO’s vision of “AI for All” similarly includes preventing the exacerbation of technological disparities within and among nations.³⁵ One global organization has sought to ensure no country is left behind by establishing an AI for developing countries forum, illustrating that such initiatives aim to make AI accessible and beneficial on a worldwide scale.³⁶ The notion of “AI for all countries” is also reflected in several areas. Emerging multilateral collaborations such as the Global Partnership on AI (GPAI), established by dozens of countries in 2020, aim to ensure that AI benefits different cultures and economies through cross-border cooperation.³⁷ The African Union has formulated a regional AI strategy framework to harness AI’s potential in education, agriculture, and other sectors for promoting inclusive development.³⁸ The Global Digital Compact likewise highlights the need to “build capacity, especially in developing countries, and facilitate access to, development of, and management of AI systems for sustainable development.”

3.2 Theoretical Foundations for “AI for All”

The concept of “AI for All” hinges not only on policy initiatives but also on robust theoretical underpinnings. This part would illuminate why AI should be made widely accessible (Law and Economics), how fairness can be upheld (justice theory), and the broader global implications of AI deployment (world-systems theory) in theory.

3.2.1 Law and Economics

Originating with Ronald Coase, Law and Economics also provides a legitimacy for “AI for All.” In the view of Judge Richard Posner, one of the leading scholars in this field, Law and Economics is a discipline that uses economic principles to explain legal issues. In fact, the economic analysis of law primarily aims to improve market efficiency and reduce transaction costs. Yet, greater efficiency does not necessarily mean a better society.³⁹ Consequently, while the perspective of economic analysis is neither a sufficient nor a necessary condition for “AI for All,” this section argues that, even from within its own framework, “AI for All” has economic legitimacy because it can reduce market externalities and enhance market efficiency.

³⁴ Pogge, T. (2002). *World poverty and human rights: Cosmopolitan responsibilities and reforms*. Polity Press.

³⁵ United Nations Educational, Scientific and Cultural Organization (UNESCO). (2024). Third International UNESCO Model on AI has been launched for associations and clubs of the UNESCO movement. <https://www.unesco.org/en/articles/third-international-unesco-model-ai-has-been-launched-associations-and-clubs-unesco-movement>

³⁶ AI for Developing Countries Forum. (2025). AI for developing countries forum. <https://aifod.org>.

³⁷ Global Partnership on Artificial Intelligence. (n.d.). About GPAI. <https://gpai.ai/about/>

³⁸ African Union. (2024). Continental artificial intelligence strategy. African Union. <https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy>.

³⁹ Sen, A. (1992). Reflections on the meaning of efficiency. *The Economic Journal*, 102(410), 1–18.

From the standpoint of Law and Economics, whether a particular service should be supplied by the market or regulated by the government depends on whether it constitutes a public or quasi-public good, and whether there is a market failure necessitating government intervention. For centuries, the market economy has worked wonders. Under Coase's theorem, in a world with zero transaction costs, the market naturally arrives at an optimal allocation, and the law itself has little to do.⁴⁰ Therefore, absent compelling reasons, a service should generally be left to the market—unless it constitutes a public or quasi-public good or unless there is a market failure. The former rationale derives from Paul Samuelson's paper, "The Pure Theory of Public Expenditure," while the latter follows Pigou's explanation in *The Economics of Welfare*, where government intervention may enhance overall social welfare in the presence of negative externalities that cause market failure.⁴¹ Thus, one should first determine whether AI belongs to the category of public or quasi-public goods. If it does, then government intervention is warranted. If it does not, generative AI services should in principle be left to the market, unless there are market defects that hamper efficient supply. When the market cannot correct such supply inefficiencies and negative externalities on its own, the government should intervene to provide or regulate these services or products.

First, it is acknowledged that, at present, generative AI services do not possess the non-rivalrous and non-excludable attributes of public or quasi-public goods. Unlike typical public goods such as national defense or public healthcare systems, generative AI services are often provided by private entities and can be made exclusive through access controls, payment mechanisms, and similar measures. They may also be subject to congestion or degradation in quality if overused, reflecting their rivalrous and excludable nature. For example, certain OpenAI models require user registration and membership subscriptions, indicating that they are not non-rivalrous; server capacity is limited, so an overload of users could result in outages, showing that they are not non-excludable. Therefore, generative AI services do not fall under the category of public or quasi-public goods. In principle, these services tend to be supplied by the market. However, if the market's pure pursuit of profit ignores the broader public interest, it may lead to inefficient supply or harm to social welfare.

Second, in the long run, failing to implement AI for All will exacerbate existing societal divisions and create economic negative externalities, ultimately undermining the environment for efficiency and innovation. From an externality theory perspective, externalities arise when the production or consumption of a good or service generates benefits or costs for others that are not fully reflected in market transactions.⁴² If

⁴⁰ Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3, 1–44.

⁴¹ Medema, S. G. (2020). *Markets, morals, and policy-making: A new defense of economics*. Oxford University Press.

⁴² Medema, S. G. (2020). *Markets, morals, and policy-making: A new defense of economics*. Oxford University Press.

generative AI services depend entirely on market supply based purely on ability to pay, certain social groups may be excluded or underserved. This exclusion, while potentially efficient from a narrow, short-term private market perspective, can contribute to negative externalities in the broader economy. Examples include widening the digital divide or marginalizing underdeveloped regions and vulnerable populations, which can lead to deeper social inequality.⁴³ As structural inequalities intensify, this may compromise social stability. Such instability, can ultimately increase business operating costs, deter long-term investment, hinder overall productivity, and thus deteriorate the very market environment necessary for sustained efficiency and innovation. Market mechanisms alone often lack sufficient incentives to fully address these externalities, suggesting a potential need for government intervention, regulation, or other collective actions to mitigate these risks.

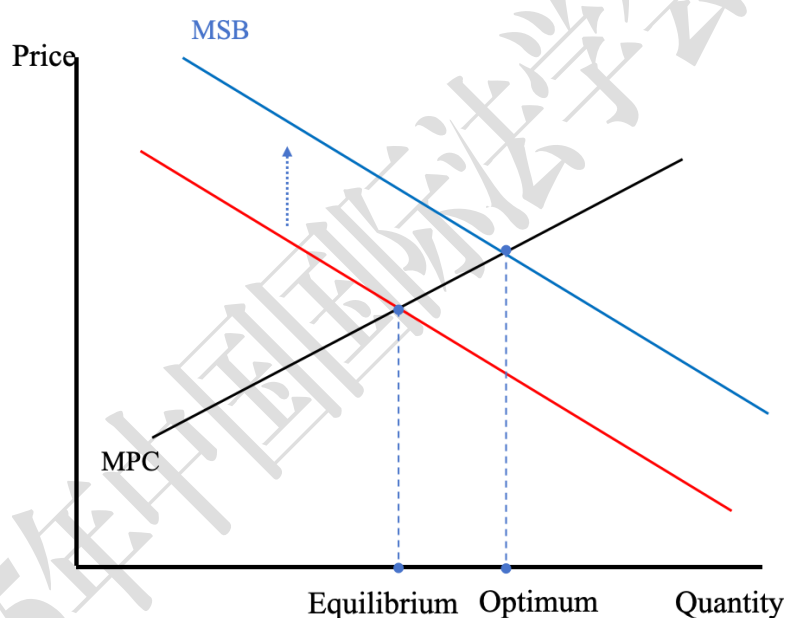


Figure 1 positive externalities on AI for all

Finally, as Figure 1 illustrates, when initiatives promoting broader access and utilization of AI, aligning with the principle of “AI for All”, are implemented, the marginal social benefit (MSB) of AI usage likely exceeds the private benefit captured by market transactions alone. Consequently, the socially optimal level of AI adoption (where MSB intersects supply) is greater than the market equilibrium—reflecting uninternalized positive externalities. If service providers proactively work towards the goals implicit in “AI for All,” whether motivated by policy guidance or their own initiatives (for example, by making foundational aspects of the technology more open, keeping prices for basic access affordable, disseminating knowledge widely, or

⁴³ Ziewitz, M., Fourcade, M., & Boyd, D. (2021). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Critical Sociology*, 47(7–8), 1137–1148.

providing training), these actions can help realize and internalize positive externalities, generating substantial societal benefits beyond the direct users. When more people gain meaningful access to generative AI services, they can drive innovation in healthcare, education, agriculture, environmental protection, and other fields, which in turn yields widespread economic and societal gains.

3.2.2 Justice Theory

Justice theories provide a moral foundation for “AI for All,” insisting that AI development and use be fair and inclusive. Both formal justice (impartial and consistent rule application) and substantive justice (fair outcomes regarding distribution of benefits and burdens) are relevant. Philosophers like John Rawls, Amartya Sen, and Aristotle offer principles to evaluate AI ethics and policy.

Rawls’ theory of justice as fairness emphasizes equitable distribution of social goods and opportunities. Two key Rawlsian ideas apply to AI ethics: fair equality of opportunity and the difference principle. Fair opportunity means positions and services (like AI-driven benefits) should be open to all under conditions of equality, which in AI translates to non-discrimination by algorithms and equal access to AI tools. The difference principle holds that inequalities are acceptable only if they benefit the least advantaged members of society.⁴⁴ In an AI context, this implies AI systems and policies should be designed to improve outcomes for marginalized groups rather than worsen gaps. So, if AI increases productivity or wealth, Rawlsian justice would argue for mechanisms that ensure disadvantaged communities share in those gains. For example, AI decision systems should not only avoid bias but proactively help those worse off – echoing Rawls’ focus on uplifting the worst-off.⁴⁵

Beyond Rawls, Amartya Sen and Martha Nussbaum’s capabilities approach enriches the conversation by stressing that real freedoms—capabilities—should be expanded across communities. The capabilities approach holds that true well-being lies not in what people have, but in people are able to do and capable of achieving.⁴⁶ This approach shifts the focus from abstract resources to people’s actual freedoms and abilities to achieve well-being. “AI for All” from Sen’s perspective means ensuring that AI technology expands people’s capabilities – e.g. the ability to get an education, communicate, or obtain healthcare – for everyone, not just the elite, to ensure people have the real capability to benefit from AI in their daily lives. Sen warns that pure economic growth or tech advancement can mask growing inequalities in capabilities. In the AI era, there’s a risk of a large underclass dominated by a techno-elite if access to AI tools and digital skills is uneven.⁴⁷ We see this in concerns that advanced AI could

⁴⁴ Westerstrand, S. (2024). Reconstructing AI Ethics Principles: Rawlsian ethics of artificial intelligence. *Science and Engineering Ethics*, 30(5), 1–21.

⁴⁵ Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, 151(2), 218–231.

⁴⁶ Robeyns, I. (2017). *Wellbeing, freedom and social justice: The capability approach re-examined*. Open Book Publishers.

⁴⁷ Yoon, I. S. (2021). Amartya Sen's Capabilities Approach: Resistance and transformative power in the age of

primarily empower wealthy corporations or countries, leaving others behind. Capabilities approach in AI ethics thus advocates for policies that empower marginalized populations with AI. The capabilities approach in AI ethics therefore advocates for policies that empower marginalized populations with AI—by investing in AI literacy, multilingual AI assistants, or affordable AI services that enhance the practical reasoning of marginalized groups, expanding their ability to perceive, imagine, and think. As one scholar notes, Sen’s framework can guide us to establish social conditions where all persons can flourish in the age of AI.⁴⁸ In summary, justice demands that AI development be assessed by how well it expands human freedoms and reduces capability deprivation.

Taken together, justice theory validates the moral imperatives behind “AI for All”: not only should AI be non-discriminatory, it should actively work to redress inequalities across diverse social, economic, and cultural contexts.

3.2.3 World-Systems Theory

While legal economics and justice theory mainly shed light on domestic contexts, world-systems theory places “AI for All” within a global and historical framework. Originating from the work of Immanuel Wallerstein, world-systems theory dividing the world into core, semi-periphery, and periphery nations.⁴⁹ It provides a macro-level perspective on AI accessibility and technological divides between developed and developing countries. World-systems theory exposes how core nations unjustly maintain their advantages to extract benefits, highlighting the necessity of “AI for All Countries”. The concept of “AI for All Countries” emphasizes the responsibility of AI-leading core nations toward peripheral and semi-peripheral countries and serves as a crucial idea for breaking the cyclical inequalities embedded in the global AI system.

In world-systems terms, AI increasingly become a new mechanism of unequal exchange between core and periphery, reinforcing dependency. Wallerstein’s analysis sees wealthy core countries dominating high-tech innovation and benefiting disproportionately, while poorer periphery countries provide raw materials or cheap labor and lag in development.⁵⁰ This dynamic is evident in AI today. The US concentrate AI research labs, cutting-edge tech companies, and infrastructure, exporting AI products and services, whereas many developing nations mainly consume AI technologies or supply data and cheap data annotators. A study by the World Economic Forum points out that the economic and social benefits of AI are geographically concentrated primarily in the Global North.⁵¹ Another study provides empirical

transhumanism. *Zygon*®, 56(4), 874–897.

⁴⁸ Yoon, I. S. (2021). Amartya Sen's Capabilities Approach: Resistance and transformative power in the age of transhumanism. *Zygon*®, 56(4), 874–897.

⁴⁹ Wallerstein, I. (2004). *World-systems analysis: An introduction*. Duke University Press.

⁵⁰ Hopkins, T. K., & Wallerstein, I. (Eds.). (1982). *World-systems analysis: Theory and methodology*. SAGE Publications.

⁵¹ Yu, D., Rosenfeld, H., & Gupta, A. (2023, January 16). The ‘AI divide’ between the Global North and the Global South. World Economic Forum. <https://www.weforum.org/stories/2023/01/davos23-ai-divide-global-north-global-south/>

evidence of this disparity, noting that AI capital stock from 1995 to 2020 is heavily concentrated in developed regions, exacerbating wealth inequality.⁵² Thus, the “AI divide” mirrors the digital divide: core countries “dominate in ... technological advancement, while periphery countries are often exploited for their resources” (including data and cheap labor).⁵³

Giovanni Arrighi’s work on long-term economic cycles and hegemonic powers can be used to explain how the AI revolution could be seen as part of a new cycle of accumulation: today’s leading powers (the U.S.) compete for dominance in AI, while other nations risk falling behind.⁵⁴ However, Arrighi also noted opportunities for semi-periphery states to rise by leveraging new technology.⁵⁵ Nevertheless, world-systems theorists would caution that without deliberate intervention, AI’s benefits will concentrate in existing centers of power. The flow of AI talent and capital is mostly from periphery to core. Global income and knowledge gaps may widen as AI increases productivity primarily in tech-savvy economies. Arrighi might interpret initiatives like China’s massive AI investments as attempts to restructure the system, potentially creating a more multipolar tech world.⁵⁶ But for many low-income countries, the worry is becoming even more peripheral in an AI-driven global economy.

4.Challenges Facing “AI for all”

This section discussed the current challenges of AI for All in the era of generative AI. First, it introduced the background—how generative AI challenges the concept of AI for All—then argued that neither soft laws nor hard laws can effectively solve these issues currently.

4.1 Background: How Generative AI Challenges “AI for All”

The emergence of generative AI technologies could lead to an emerging “AI divide”⁵⁷ and perpetuate structural inequalities among users with different sociodemographic backgrounds.⁵⁸ This part will examine how, in the era of generative AI, these factors contribute to the “AI divide,” thereby intensifying the wealth gap and diminishing social mobility, and erode “AI for all”.

south/.

⁵² Smith, J., & Lee, K. (2024). Artificial intelligence and wealth inequality: A comprehensive analysis. *Technological Forecasting and Social Change*, 196, 121234.

⁵³ Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data’s relation to the contemporary subject. *Television & New Media*, 20(4), 336–349.

⁵⁴ Arrighi, G. (1990). The developmentalist illusion: A reconceptualization of the semiperiphery. In W. G. Martin (Ed.), *Semiperipheral states in the world1-economy* (pp. 11–42). Greenwood Press.

⁵⁵ Arrighi, G. (2007). *Adam Smith in Beijing: Lineages of the twenty-first century*. Verso.

⁵⁶ Arrighi, G. (2007). *Adam Smith in Beijing: Lineages of the twenty-first century*. Verso.

⁵⁷ Carter, L., Liu, D., & Cantrell, C. (2020). Exploring the intersection of the digital divide and artificial intelligence: A hermeneutic literature review. *AIS Transactions on Human-Computer Interaction*, 12(4), 253–275.

⁵⁸ Gran, A.-B., Booth, P., & Bucher, T. (2021). To be or not to be algorithm aware: A question of a new digital divide? *Information, Communication & Society*, 24(12), 1779–1796. 12. Cotter, K. (2022). Practical knowledge of algorithms: The case of BreadTube. *New Media & Society*. Advance online publication.

4.1.1 Price Cost Eroding “AI for all”

Over the past few years, Generative AI has advanced significantly, marking a paradigm shift in both capability and social impact. Unlike earlier AI systems, which were limited to specific tasks such as facial recognition or recommendation algorithms, Gen AI can perform generative tasks, such as writing novels and generating code. This leap in generality and autonomy is exemplified by models such as OpenAI’s GPT-4.5 and O3, Anthropic’s Claude 3.7, and Google’s Gemini 2.5 Pro, which demonstrate reasoning, creativity, and decision-making capabilities that rival or surpass human expertise. Moreover, state-of-the-art models exhibit remarkable proficiency in complex tasks; for example, OpenAI’s “deep research” function autonomously conducts comprehensive internet research, achieving a 26.6% accuracy rate on the “Humanity’s Last Exam” benchmark, comparable to human performance.⁵⁹ As Gen AI capabilities continue to evolve rapidly, even more advanced systems are under development.

However, the price of access to most advanced generative AI tools has also undergone a significant leap. While some advanced generative AI models remain open-source or freely accessible to the public—among which DeepSeek R1 stands out prominently—venture capitalist Marc Andreessen described its emergence as the “Sputnik moment” in the AI field. Nevertheless, DeepSeek R1 still trails behind the most cutting-edge large models.⁶⁰ The most cutting-edge models are unaffordable for the majority of people, posing a substantial challenge to the vision of AI technology for all. The advanced capabilities of Gen AI models—such as OpenAI’s GPT-4.5 and O1 pro—have been matched by steeply tiered pricing that entrenches inequality. OpenAI’s 200/month “Pro” tier, for instance, grants 120 “deep research” queries, while free users are limited to two—a disparity that extends to API costs, where high-volume access to models like ChatGPT 4.5, 150 dollars per million output tokens. Even mid-tier subscriptions (e.g., Grok at 30/month, ChatGPT plus at 20/month) remain unaffordable for billions globally, particularly in regions where 20 represents a substantial share of income. And OpenAI is exploring the pro max subscription-2000/month level. As showed in table 1, even the mid-tier subscriptions divided about 2-15% of their month income. As for “pro” lever, occupied 20%-150% for developing countries’ people’s income, it’s impossible for most developing countries people to afford.

Country	Plus Share (%)	Pro Share (%)
USA	0.3	3
China	1.8	18

⁵⁹ TechRadar. (2025, March). OpenAI’s Deep Research smashes records for the world’s hardest AI exam. TechRadar. Retrieved from <https://www.techradar.com/computing/artificial-intelligence/openais-deep-research-smashes-records>.

⁶⁰ Mozur, P. (2025, March 20). Chinese AI firm DeepSeek shows promise but trails U.S. leaders. The Wall Street Journal. <https://www.wsj.com/tech/ai/china-ai-deepseek-chatbot-6ac4ad33>

India	9.4	94
Japan	0.6	6
Germany	0.43	4.3
Brazil	2.6	25.8
South Africa	3.7	37
Nigeria	12.7	127
Bangladesh	8.3	83
Pakistan	16.4	16.4

Table 1 The subscription costs (\$20 for Plus, \$200 for Pro) were divided by monthly normal GNI per capita and multiplied by 100 to get the percentage of income. GNI data come from World Bank of 2023.

As a result, the leap in capability of advanced Gen AI also brings about a leap for the person who uses it,⁶¹ significantly impacting the ideal of “AI for all”. When it comes to “AI justice for all,” this trend deviates from the principle of substantive justice: creating a capability caste system, wealthy users and corporations harness Gen AI for innovation, education, and market dominance, while marginalized groups—students, small businesses, and low-income nations—are relegated to slower, less intelligent, outdated, or restricted versions. Individuals with access to the most advanced Gen AI can transcend conventional skill boundaries and even evolve into “super-individuals.” For example, OpenAI’s Deep Research function can effectively replace a research assistant, enabling users to complete literature reviews or data analyses within minutes—far outpacing the productivity of human researchers.⁶² According to a now-outdated McKinsey study, earlier versions of generative AI could already boost information retrieval efficiency by 400%.⁶³ Plus, research also indicates that Gen AI significantly enhances learning capacity, and managerial skills, widening capability gaps exponentially rather than incrementally.⁶⁴ This exponential advantage consolidates existing social hierarchies, reflecting the broader dynamics of digital private power concentration.⁶⁵ A wealthy student paying 200 dollars per month for AI-based assistance can produce work once requiring a professor’s guidance, while an entrepreneur uses AI to generate patent-level solutions—turning technical privilege into capital gains. These “super-individuals” are not created by innate talent or diligence but by the ability to pay, stifling social mobility and reinforcing a cycle in which “capital

⁶¹ Cheng, L., and Gong, X. 2024. Appraising Regulatory Framework Towards Artificial General Intelligence (AGI) Under Digital Humanism. *International Journal of Digital Law and Governance* 1 (2): 269–312.

⁶² Reuters. (2025, February 3). OpenAI launches new AI tool to facilitate research tasks. <https://www.reuters.com/technology/openai-launches-new-ai-tool-facilitate-research-tasks-2025-02-03/>.

⁶³ McKinsey Global Institute. (2023, June 14). The economic potential of generative AI: The next productivity frontier. McKinsey & Company.

⁶⁴ Cornell University Center for Teaching Innovation. (2024). Generative artificial intelligence. Center for Teaching Innovation. Retrieved March 19, 2025, from <https://teaching.cornell.edu/generative-artificial-intelligence>.

⁶⁵ Sun, X. & Xiao, Y. (2024). How Digital Power Shapes the Rule of Law: The Logic and Mission of Digital Rule of Law. *International Journal of Digital Law and Governance*, 1(2), 207-243.

empowerment” fuels “technology empowerment,” leading to “capability enhancement” and ultimately more “capital accumulation.”

4.1.2 Literacy Divide and Infrastructure condition Undermining “AI for all”

A critical yet underrecognized obstacle to “AI for all” lies in cognitive disparities—the gap between those who understand Gen AI’s potential and those who remain unaware of its existence or utility. A study by the St. Louis Fed shows that high-education workers use depth 3.2 times more than low-education groups, leading to a 15% increase in the productivity gap.⁶⁶ Gen AI is forging a new digital divide concerning not only who can log on or possesses technical digital skills but also who can comprehend, create, or control AI technologies.⁶⁷ This knowledge gap is compounded by AI literacy: While tech-literate populations leverage AI for tasks ranging from coding to career advancement, marginalized group often lack even basic awareness of these tools. Research have shown, the most vulnerable groups with the lowest levels of AI knowledge and AI skills were mostly older, with lower levels of education and privacy protection skills, than the average users.⁶⁸ Without targeted literacy programs, this cognitive barrier entrenches a hierarchy of opportunity, where the technologically initiated reap AI’s benefits, leaving others further behind in education, employment, and civic participation.

Plus, the promise of “AI for all” collapses where basic infrastructure is absent. Globally, 2.6 billion people remain offline,⁶⁹ —preconditions for using Gen AI. Even when devices are available, computational demands pose hurdles: running models like GPT-4.5 requires stable high-speed connectivity, which remains inaccessible in areas with bandwidth limitations. Without universal infrastructure investment, Gen AI’s transformative potential remains confined to urban, affluent enclaves, deepening global inequities.

In short, traditional digital divides—concerning both AI literacy and digital infrastructure—are simultaneously creating and exacerbating a new “AI divide,” making previously disadvantaged groups even more vulnerable. This significantly undermines the three core ideals behind “AI for all”: Firstly, individuals remain unaware of AI technology and have no means to access it, let alone benefit from “AI for all.” Secondly, marginalized communities become further disadvantaged, making

⁶⁶ Federal Reserve Bank of St. Louis. (2025). The impact of generative AI on work productivity. St. Louis Fed. Retrieved March 19, 2025, from <https://www.stlouisfed.org/on-the-economy/2025/feb/impact-generative-ai-work-productivity>.

⁶⁷ Hendawy, M. (2024). The intensified digital divide: Comprehending GenAI. Internet Policy Review.

⁶⁸ Wang, C., Boerman, S. C., Kroon, A. C., Möller, J., & H. de Vreese, C. (2024). The artificial intelligence divide: Who is the most vulnerable? New Media & Society.

⁶⁹ International Telecommunication Union. (2023, September 12). Population of global offline continues steady decline to 2.6 billion people in 2023: Accelerating progress is key in race toward universal and meaningful connectivity. ITU. Retrieved March 19, 2025, from <https://www.itu.int/en/mediacentre/Pages/PR-2023-09-12-universal-and-meaningful-connectivity-by-2030.aspx>

substantive justice impossible. Thirdly, at present, it's more accurate to describe the situation as “AI for certain countries” rather than “AI for all countries.”

4.2. *Soft Laws Are Not Enough*

Soft laws are the main source of “AI for all” concept, with many of its guiding principles articulated through non-binding frameworks and ethical guidelines. However, relying on soft law alone is clearly insufficient.

Level	Country/ Organization	Document name (Year)	Main Content Concerning AI for all
Domestic	United States	Blueprint for an AI Bill of Rights (2022)	Emphasizes non-discrimination, privacy, equal opportunity.
		National AI R&D Strategic Plan (2023)	Ensures AI advances equity and prevents inequality.
	European Union	Ethics Guidelines for Trustworthy AI (2019)	Fairness is key ethical principles. It refers to the equal and just distribution of benefits and costs, as well as ensuring freedom from unfair bias, discrimination, and stigmatization.
	China	Ethical Norms for New Generation AI (2021)	Promotes fairness and justice, promote fair sharing of AI benefits by all of society, fully respect and help vulnerable groups
	Japan	Japanese Society for AI Ethical Guidelines (2017)	Acknowledging AI may bring about additional inequality and discrimination; ensuring AI can be used by humanity in a fair and equal manner.
	South Korea	National Guidelines for AI Ethics (2021)	Inclusive AI, minimizing bias, equal benefits for all.
International	OECD	OECD AI Principles (2019 Updated 2024)	Emphasizing AI should benefit humanity and the planet by promoting inclusive growth and

Level	Country/ Organization	Document name (Year)	Main Content Concerning AI for all
			sustainable development.
	United Nations General Assembly	A/RES/78/168 (2024)	Urge support for developing countries to ensure inclusive and equitable access to AI, bridge the digital divide, and align with sustainable development goals.
	G20	G20 AI Principles (2019)	Focus on the responsible development of AI, emphasize inclusivity, ensure AI benefits everyone, and reduce inequalities.
	World Economic Forum	A Blueprint for Equity and Inclusion in AI (2022)	Provides an inclusive strategy for the entire AI lifecycle, ensuring fair access to AI and skill development, benefiting vulnerable groups.
	United Nations	Global Digital Compact (2024)	Achieving inclusive and equitable access to artificial intelligence and digital technologies.

Table 2 Main Soft laws concerning AI for all

As showed in table 2, soft laws concerning “AI for all” can be divided into two level: domestic and international. In terms of domestic level, various countries have published soft law documents related to “AI for all”, focusing on fairness, equality, and non-discrimination. At the international level, there is a general trend toward the sustainable development of AI-related goals. One key difference between domestic-level AI soft law and international-level discussions is that the North-South AI divide is a major concern in international papers, especially in United Nations documents. Actually, most “AI for all” issues can only be resolved at the international level. For instance, regarding “AI technology for all,” the most advanced gen AI models are concentrated in the United States. Some American AI companies provide free access to

both cutting-edge and slightly less advanced models under certain conditions. As a result, in most countries, their domestic AI companies struggle to establish technological barriers and lack the ability to charge users for their gen AI models. Consequently, citizens of other countries are left with two choices: either pay American AI companies or use less advanced models. Regarding “AI justice for all,” achieving both procedural and substantive justice requires international cooperation. The standard for procedural justice must be unified to prevent a global race to the bottom in regulation. Likewise, substantive justice cannot be realized without support from developed countries for developing nations. Finally, “AI for all countries” is inherently an international issue. Developing countries already face disadvantages in infrastructure, funding, talent development, and technological reserves. Without international support, AI technology will remain confined to certain nations, further widening the digital divide.

In the jurisprudential context, soft laws whether in domestic or international governing “AI for All” encounter significant challenges, including inadequate substantive fairness (for certain countries), non-binding enforcement mechanisms, and excessive abstraction. Firstly, fairness emerges as a shared attribute across documents from different countries, albeit with distinct interpretations. The United States adopts a procedural approach emphasizing equality of opportunity and prevention of explicit discrimination, yet demonstrates minimal engagement with structural inequalities inherent in AI systems. This paradigm proves particularly problematic given the socioeconomic realities of advanced AI models. The high costs associated with accessing advanced AI models – a critical barrier unaddressed by U.S. guidelines – risk exacerbating wealth disparities under such opportunity-focused frameworks. Current U.S. soft law provisions permit this outcome by prioritizing formal justice over distributive equity, effectively legitimizing market-driven inequalities through regulatory inaction. In contrast, the European Union and China not only stress formal equality but also emphasize substantive equity, advocating for fairness in outcomes. Secondly, these documents are characterized by their non-legally binding nature. One research compared 22 mainstream AI ethics guidelines and found that most of them are merely symbolic, lacking binding force, and argued that they need to be reinforced by legal regulations and independent oversight.⁷⁰ As a result, none of the guidelines under review impose mandatory legal obligations. Certain documents explicitly affirm their non-binding status, as evidenced by disclaimers such as, “Adoption of these principles may not meet the requirements of existing statutes, regulations, policies, or international instruments, or the requirements of the Federal agencies that enforce them”.⁷¹ This lack of enforceability means that the articulated visions and norms can

⁷⁰ Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120.

⁷¹ White House Office of Science and Technology Policy. (2022). Blueprint for an AI Bill of Rights: Making automated systems work for the American people. <https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/about-this-document>.

be disregarded, rendering the lofty declarations contained within them largely utopian. Such a situation precipitates a structural collective action dilemma: the absence of binding constraints enables enterprises that flout principles to secure greater profits and competitive advantages. To remain viable in the market, other enterprises are pressured to follow suit, culminating in a “race to the bottom” wherein adherence to ethical standards is universally forsaken. This predicament cannot be ameliorated through moral persuasion alone; it demands the establishment of legally binding regulations endowed with coercive authority. Thirdly, the content of these AI ethical guidelines tends to remain abstract and theoretical.⁷² They frequently adopt a philosophical narrative style, describing idealistic visions of a utopian future in generalized and somewhat vague terms, while lacks concrete, actionable measures for bridging the gap between aspirational discourse and practical implementation. Although connections between the high cost of advanced models and the “AI for All” concept can be inferred at an abstract level, this requires interpretative effort. These soft law norms do not directly address specific challenges, such as the concrete impacts of the high cost of advanced models, nor do they provide specific guidance or policies to mitigate these effects. This deficiency extends to other barriers, including cognitive and usability barriers as well as infrastructure barriers.

4.3 Hard Laws: Remaining Scarcity

Some believe that ethical AI principles alone are insufficient to ensure that companies use AI responsibly. Instead, a strict governance framework should be implemented to reduce AI-related risks.⁷³ However, enforceable legal regulations for “AI for all” are largely absent. Even when focusing on domestic AI legislation across various countries, truly hard legal frameworks are rare, let alone regulations specifically promoting “AI for all”.

Firstly, the absence of hard legal frameworks for “AI for all” is primarily evident in the absence of comprehensive AI-specific legal regulations. Most countries are still exploring AI governance and have not yet formed systematic hard-law frameworks for AI regulation,⁷⁴ much less specific provisions for AI for all.

Secondly, even countries emphasizing “AI for all” through soft-law guidelines and introducing AI-related “hard laws”, their hard laws still make few responses regarding “AI for all”. For example, the EU's Ethics Guidelines for Trustworthy AI (2019) stress “fairness as one of four key ethical principles for Trustworthy AI,” stating that “substantial fairness relates to the equal and just distribution of benefits and costs, and

⁷² Cheng, L., Han, J. and Nasirov, J. 2024. "Ethical considerations related to personal data collection and reuse: trust and transparency in language and speech technologies" *International Journal of Legal Discourse* 9 (2): 217–235; Cheng, L. and Liu, X. 2023. "From principles to practices: the intertextual interaction between AI ethical and legal discourses" *International Journal of Legal Discourse* 18 (1): 31-52.

⁷³ Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics*, 1(1), 73–80.

⁷⁴ Fang, A., & Perkins, J. (2024). Large language models (LLMs): Risks and policy implications. MIT Science Policy Review.

ensuring freedom from unfair bias, discrimination, and stigmatization.” Yet, the provisions explicitly related to “AI for all” within the EU AI Act remain relatively limited. Provisions concerning “AI justice for all” in the EU AI Act primarily focus on formal justice rather than substantive justice. Regulations center around preventing AI discrimination, biases, and safeguarding privacy. For instance, Article 5 prohibits fundamentally unjust or equality-harmful AI uses; Article 10 on Data and Data Governance requires datasets for training, validation, and testing to be representative and error-free to avoid discrimination. Although the Act itself lacks extensive regulations on privacy protection, other related regulations like GDPR potentially provide privacy safeguards. However, regarding substantive justice, regulations remain notably insufficient. Protections for disadvantaged and marginalized groups are minimal, only including Article 16(1), which requires providers of high-risk AI systems to “ensure compliance with accessibility requirements,” and Article 4, mandating that AI providers and deployers enhance AI literacy among staff and users. There are no explicit measures detailing how to provide affordable AI solutions for economically disadvantaged groups. The Act also lacks measures to systematically solve more implicit and structural social inequalities through AI. Furthermore, explicit provisions encouraging open-source AI models, algorithms, and data are notably insufficient, lacking substantial support for open-source communities and public resource sharing.

Several factors contribute to the disconnect in the transition from soft law to hard law regarding “AI for all”:

Firstly, slow legislative processes contrasted with rapid technological advancements. One key reason for the soft law–hard law disconnect is the regulatory lag between fast-moving AI technology and the slow pace of legislation. Law often trails technological innovation, leaving a gap in governance as new risks emerge.⁷⁵ In the AI context, scholars observe that “the pace of development of AI far exceeds the capability of any traditional regulatory system to keep up”.⁷⁶ Formal lawmaking – from drafting bills to passing legislation or agency rules – takes years, while AI capabilities evolve in months. This pacing problem means “AI for all” may lack binding rules at the moment they’re most needed. For example, the EU’s effort to enact a comprehensive AI law (the EU AI Act) began in 2021 but is only reaching approval stages by 2024. Regulatory lag suggests that without adaptive, faster mechanisms, hard law will chronically catch up “after the fact,” struggling to address AI’s latest challenges.

Secondly, the abstract and non-binding nature of soft law facilitates easier adoption. Because they lack legal force, soft law measures face fewer political hurdles – no need for lengthy legislative votes or international treaties. This makes them agile

⁷⁵ Joskow, P. L., & Noll, R. G. (1981). Regulation in theory and practice: An overview. *Studies in public regulation*, 1-78.

⁷⁶ Marchant, G. (2020, February 10). Soft law as a complement to AI regulation. Brookings Institution. Retrieved March 19, 2025, from <https://www.brookings.edu/articles/soft-law-as-a-complement-to-ai-regulation/>.

tools for emerging tech governance.⁷⁷ This abstract nature facilitates passage but also means these principles often stay on paper without concrete implementation mechanisms.

Thirdly, disagreements on pathways to achieve “AI for all”, particularly between market-driven and government-driven approaches. Some experts argue that premature or heavy regulation could impede innovation in AI, suggesting that “AI for all” might be achieved through competition and corporate responsibility without strict laws. The U.S. policy approach has largely leaned this way: it encourages multi-stakeholder collaboration on AI guidelines and involves industry leaders in drafting policies, rather than imposing top-down rules early.⁷⁸ On the other side, many academics and civil society voices contend that government intervention is necessary to guarantee inclusivity and public values in AI.⁷⁹ They point out that some scholars believed that laissez-faire development has led to AI tools with bias and disparate impacts, and that without binding rules, firms might under-invest in costly fairness measures.

5. Mechanism to Integrate Soft and Hard Laws for “AI for All”

This mechanism requires a dual approach: leveraging the guiding role and ethical pressure of soft laws by expanding the connotation of “AI for All”, while simultaneously harnessing the mandatory force of hard laws to strike a balance between market dynamics and “AI for All” based on national realities.

5.1 Role of Soft Laws: Guiding Principles and Ethical Pressure

Only soft laws are clearly not enough, yet they indeed serve unique functions. Soft laws can exert moral pressure on relevant parties and influence the creation of hard laws.⁸⁰ Given the new challenges arising from the rapid advancement of generative AI, soft laws should continue to fulfill their roles by expanding the content of “AI for All,” thereby contributing to its practical implementation.

Firstly, the principle of “affordability” in “AI for All” should be explicitly included in soft law provisions. “AI technology for all” means more than merely universal access; it should incorporate the capability approach and Rawls' difference principle, offering varied degrees of benefits tailored to the actual circumstances of different users to achieve substantive justice. Currently, major AI companies like OpenAI and Google, provide free access only to second-tier models, while genuinely advanced services are

⁷⁷ Frontier Economics. (2023). Innovative technology governance: Hard rules for soft laws. Frontier Economics. Retrieved March 19, 2025, from <https://www.frontier-economics.com/uk/en/news-and-insights/articles/article-i21147-innovative-technology-governance-hard-rules-for-soft-laws>.

⁷⁸ Walter, Y. (2024) Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation—A contemporary overview and an analysis of socioeconomic consequences. *Discov Artif Intell* 4, 14.

⁷⁹ Wilczek, B., Thäsler-Kordonouri, S., & Eder, M. (2024). Government regulation or industry self-regulation of AI? Investigating the relationships between uncertainty avoidance, people's AI risk perceptions, and their regulatory preferences in Europe. *AI & Society*.

⁸⁰ Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International Organization*, 54(3), 421–456.

either expensive or severely usage-limited. On one hand, soft laws should encourage developers and gen AI service providers to offer discounted or free access to AI services in low-income regions, and for disadvantaged groups. For example, enterprises should consider to adopt global pricing strategies adjusted according to purchasing power parity (PPP). Even state-of-the-art AI models should offer users certain levels of free access. On the other hand, businesses should establish AI-focused public welfare programs and collaborate with academic institutions, libraries, and non-profit organizations to provide low-cost or free advanced functionalities for educational and research purposes. Extending access to state-of-the-art AI models broadly across international and local communities can maximize AI's positive externalities.

Secondly, enhancing AI literacy and infrastructure should be identified clearly as one of the primary goals within soft law frameworks. Currently, soft law guidelines on AI ethics often address AI literacy and infrastructure superficially and abstractly. Soft laws should establish specific provisions dedicated to AI literacy, encouraging collaboration among enterprises, universities, and governments to implement comprehensive public training and awareness programs. Integrating “basic AI cognition” into mandatory or elective courses at all educational levels should be advocated. For instance, civil society organizations should be encouraged to disseminate recommendations for the correct use of large language models. A relevant example is the American Medical Association (AMA), which has already published guidelines on the responsible use of ChatGPT and similar models.⁸¹ Particular attention must be directed toward vulnerable populations, elderly individuals, and those with limited educational backgrounds. Community groups and non-profit organizations should be encouraged to host offline training sessions, providing introductory AI skills and demonstrations for residents in rural areas, thereby bridging the existing cognitive divide.

Thirdly, international cooperation must be reinforced through soft law guidelines concerning AI. AI should not remain an exclusive domain of wealthy nations; global AI soft law guidelines should strengthen advocacy by emphasizing capacity building and promoting multilateral and bilateral cooperation. Countries with advanced AI capabilities should be encouraged to provide comprehensive public resources to developing nations, including infrastructure, open-source models, datasets, governance frameworks, and safety tools. In infrastructure development, advanced nations should support developing countries in establishing data centers, high-performance computing facilities, and robust network infrastructure, particularly prioritizing increasing Africa's share of global data center capacity. Regarding talent cultivation, soft law can advocate for establishing a global AI education network to nurture local AI expertise in developing nations, encompassing technical research, application development, and

⁸¹ American Medical Association. (2023). ChatGPT and generative AI: What physicians should consider [PDF]. <https://www.ama-assn.org/system/files/chatgpt-what-physicians-should-consider.pdf>.

governance management. Concerning technology transfer, advanced countries should be encouraged to share applicable technologies with developing nations, carefully balancing intellectual property protection with technological inclusivity. Regarding institutional capabilities, assistance should be provided to developing nations in establishing AI governance frameworks tailored to their national contexts, thus enhancing their capacity to participate effectively in global AI governance. Concurrently, developing countries can leverage initiatives such as the Global Initiative on AI Governance and the United Nations resolution on “Strengthening International Cooperation in AI Capacity Building” to prioritize capacity building as a critical aspect of international cooperation.

5.2 Role of Hard laws: Balancing Innovation, Market Efficiency, and Inclusivity

To advance “AI for All” through hard law, differentiated institutional designs must be developed based on the varying positions of countries within the global AI technology system, while preserving innovation vitality and market efficiency. Therefore, this section first explores a set of potential measures for the hardening of soft law, and then provides a categorized analysis based on the state of AI technological development in different countries.

5.2.1 Measures for Implementing Hard Law

Currently, generative AI services are neither be seen as “public goods” nor “quasi-public goods”—markets remain the primary drivers of innovation and supply. However, “AI for All” carries significant public value, and overreliance on market forces risks exacerbating structural inequalities and other negative externalities. Consequently, the critical role of hard law lies in two dimensions: first, clarifying the boundaries of governmental public responsibility to avoid oversimplifying AI as a “free public good,” and second, deploying incentive-based legal mechanisms to correct market failures, social inequalities, and other negative externalities, thereby promoting broader and more affordable access to AI services for the public.

Regarding AI price barriers, multiple hard law measures could be considered. Firstly, incorporating AI services into government-guided pricing should be considered. Various countries have historically employed price regulations for essential public commodities and services. For instance, in the U.S., the Hepburn Act (1906) established maximum railroad freight rates. In China, the Pricing Law mandates governmental pricing or guided pricing for critical commodities affecting national welfare, including medicines, energy, and transportation. Similarly, Japan’s Price Control Act permits governmental regulation of commodity and service pricing during emergencies. Other nations such as Germany, France, and the UK also regulate pricing in sectors like energy, telecommunications, railways, and healthcare to varying extents. For AI services, if their societal and economic significance continues to grow, governments may consider subjecting them—in whole or in part—to guided pricing. Specific measures could

include setting price ranges based on per capita income levels or AI's impact on public services and safety, thereby preventing exclusionary costs from barring public access to core AI applications. Secondly, promoting “AI for All” through government procurement could be implemented. Most countries have government procurement laws regulating public purchasing behaviors. AI procurement agreements could help negotiate better pricing through economies of scale.⁸² By procuring advanced large-model APIs and then providing them to education, healthcare, and small and medium-sized enterprises, this approach would exert less market impact compared to government-mandated guided pricing.

To enhance citizens' AI literacy and bridge the AI divide, multiple hard law approaches exist. Firstly, in public service assurance law, free AI services could be provided through public cultural institutions. Various national legal frameworks regulate public service facilities like museums and libraries, covering their construction, management, operations, funding, and public accessibility. For example, the U.S. Museum and Library Services Act of 1996 outlines funding and operational guidelines to ensure public access. China's Public Cultural Services Guarantee Law similarly standardizes the establishment and operation of libraries, cultural centers, and museums to guarantee equitable public cultural services. Under the “AI for All” framework, countries could revise or reinterpret such laws to empower public institutions—using fiscal funds—to procure commercial AI services and offer free or low-cost access to the public. This approach would alleviate the financial burden on low-income or marginalized groups, enabling them to utilize advanced AI models through public facilities without requiring expensive personal devices or subscriptions. Secondly, in education law, revising or interpreting educational legislation to formally integrate “AI literacy” and “digital skills” within the scope of the “right to education” is advisable. Educational laws should explicitly incorporate “understanding and applying generative AI” into defined educational objectives. For general adult education and lifelong learning frameworks, community colleges, vocational training centers, and online platforms should be encouraged to provide AI literacy programs, ensuring ongoing learning opportunities across individuals' life stages. For example, California's recent law requiring schools to teach students about AI signifies a growing recognition of the importance of equipping young people with these essential skills.⁸³ Furthermore, legislation governing educational facilities should include broadband networks, smart terminal devices, and cloud computing resources as essential educational infrastructure, clarifying spending proportions or minimum standards in terms of network coverage rates and per capita device ownership.

⁸² McGee, M. (2025, January 25). What the rising costs of AI means for government. StateTech. <https://statetechmagazine.com/article/2025/01/what-rising-costs-ai-means-government>.

⁸³ Jones, D. (2024, October 11). California law requires schools to teach students about AI. Government Technology. <https://www.govtech.com/education/k-12/california-law-requires-schools-to-teach-students-about-ai>.

5.2.2 Differentiated National Pathways: From Core to Peripheral AI Technology States

The effectiveness of hard law in realizing “AI for All” depends on a nation’s possession of “AI sovereignty.” In other words, universal access to AI presupposes the existence of domestic technological capacity to provide “AI for all”. If advanced large language models reside outside a nation’s borders, and domestic markets and industries lack influence, efforts toward “AI for all” will falter. For this reason, while outlining potential hard-law institutional approaches, this section also categorizes four types of nations based on their distinct inclusivity challenges and corresponding response strategies. Drawing on world-systems theory and the current global division of labor in AI technology, nations can be categorized into Core, Near-Core, Semi-Core, Semi-Peripheral, and Peripheral AI technology states. Countries at different tiers face stark disparities in technological sovereignty, market scale, international competitiveness, and regulatory capacity when advancing “AI for All” through hard law, necessitating distinct legislative priorities.

Tier	Key Countries/Regions	Hard-Law Strategies for "AI for All"
Core AI States	United States	Domestic hard law can effectively promote “AI for all”: Covers AI Technology for All and AI Justice for All.
Near-Core AI States	China	Domestic hard law can advance “AI for all”: Addresses AI Technology for All and AI Justice for All.
Semi-Core AI States	UK, Canada, Germany, France, Japan, South Korea	Limited domestic hard-law impact on “AI for all”: Pricing and open-source policies for advanced models are dictated externally; long-arm jurisdiction may ineffective. AI Justice measures can still influence fairness.
Semi-Peripheral States	India, Russia, Brazil, UAE (emerging economies)	Weak domestic hard-law impact: Advanced model pricing and open-source policies are controlled externally; long-arm jurisdiction faces challenges. AI Justice frameworks may partially address equity.
Peripheral States	Most developing nations (Africa, Oceania, Latin America)	Domestic hard law fails to advance “AI for all”: No control over pricing or open-source policies for advanced models; long-arm jurisdiction is nearly impossible. Only reliance on international cooperation.

Table 3 Differentiated National Tiers and Hard-Law Strategies for “AI for All”

Firstly, for core AI technology state, The U.S. possesses the world’s most robust AI ecosystem, leading in large-scale model development, chip design and manufacturing, foundational research, and industrial applications, with full AI sovereignty. It can effectively implement “AI for All” through domestic hard law—for example, by adopting government-guided pricing, implementing targeted subsidies, strengthening antitrust and data governance, and mandating AI literacy in education laws. The legislative challenge lies in balancing two goals: maintaining incentives for innovation and global competitiveness in the tech sector while preventing the monopolization of technological dividends by a few corporations or privileged groups, which would undermine societal equity.

Secondly, China, the Near-Core AI Technology State, ranks second globally in foundational model development, domestic AI chip R&D, academic output, and application scale, with partial AI sovereignty. However, it remains dependent on foreign suppliers for certain high-end chips and core algorithms. Domestic hard law can play a significant role here: through pricing laws, public service laws, and education laws, China could enact preferential measures to expand AI access for low-income populations and remote regions. The legislative difficulty lies in navigating international technological blockades and competition: ensuring domestic control while avoiding overregulation that stifles innovation in its AI industry.

Thirdly, these semi-core AI Technology States nations (e.g., South Korea, Japan, Germany) exhibit active R&D and industrial ecosystems, moderate capabilities in large-scale model design and chip manufacturing. While excelling in specific dimensions, none matches the U.S. or China in comprehensive strength across models, chips, computing power, research, applications, and policy investment. For “AI for All” legislation, reliance on foreign core technologies and models complicates efforts to enforce pricing controls or subsidies, as decisions by multinational corporations often dictate outcomes. Nonetheless, domestic antitrust laws, data governance frameworks, and labor regulations can still promote fairness and affordability in AI adoption, albeit with less impact than in Core or Near-Core states.

Fourthly, for semi-peripheral AI Technology States, these nations (e.g., India, Brazil, Indonesia) lack advanced large-scale model design capabilities and are typically large emerging economies or developing countries actively deploying AI. While they possess some technical and talent infrastructure, their capacity for high-end chip production, cutting-edge research, or large-scale industrial AI applications remains limited. Hard law can regulate domestic AI pricing, public service access, and educational guarantees, but challenges persist due to reliance on foreign technology. For example, “long-arm jurisdiction” struggles to influence pricing or platform strategies of foreign firms in local markets. Thus, international cooperation—via

regional or multilateral platforms—is critical to facilitate technology transfer, public AI infrastructure development, and global/regional coordination on AI education and training.

Finally, most developing nations fall into peripheral category, lacking advanced research institutions, AI industrial bases, or capabilities in foundational model development and chip manufacturing. Their AI adoption is limited to off-the-shelf solutions provided by foreign entities. Domestic hard law has minimal leverage to counter transnational monopolies or high-cost services, as they lack control over core technologies. The solution does not lie in hard law, but rather depends heavily on international organizations, multilateral aid, or treaties to integrate “AI for All” into global development agendas. This includes securing preferential or free access to AI tools for low-income nations, coupled with infrastructure and educational support. At the same time, efforts should also be made to expand electricity and internet access and provide training in AI skills.⁸⁴

CONCLUSION

This paper has systematized the multifaceted concept of “AI for All” into a coherent theoretical and practical framework, emphasizing three interrelated dimensions: AI Technology for All, AI Justice for All, and AI for All Countries. By defining these dimensions clearly, the research delineates “AI for All” from related concepts like inclusive, accessible, or human-centered AI, addressing ambiguities prevalent in contemporary discourse.

The theoretical exploration grounded in Law and Economics, Justice Theory, and World-Systems Theory provides robust legitimacy to the “AI for All” initiative. Law and Economics underscores the need for governmental intervention when market mechanisms alone fail to tackle inequalities and externalities inherent in AI deployment. Justice Theory, drawing notably from Rawlsian and capabilities frameworks, reinforces the moral imperative for equitable AI distribution, stressing substantive over purely formal fairness. World-Systems Theory places “AI for All” within a global context, highlighting how AI can either perpetuate or mitigate inequalities between developed and developing nations, emphasizing the need for deliberate international policy interventions.

However, realizing “AI for All” faces critical barriers: the prohibitive costs of advanced generative AI models, gaps in AI literacy, and inadequate infrastructure. Soft law initiatives, though abundant, lack enforceability, specificity, and responsiveness to rapidly evolving generative AI technologies. Meanwhile, existing hard laws remain

⁸⁴ Nugroho, L., & Cammelli, G. (2024, December 12). Tipping the scales: AI’s dual impact on developing nations. World Bank Blogs. <https://blogs.worldbank.org/en/digital-development/tipping-the-scales--ai-s-dual-impact-on-developing-nations>.

scarce, fragmented, and insufficiently attuned to the substantive dimensions of justice necessary for genuinely inclusive outcomes.

To bridge this gap, the paper proposes a comprehensive integration of soft and hard laws. Soft laws should be explicitly updated to emphasize affordability, literacy, and international cooperation, providing ethical guidance and public pressure to complement hard laws. Concurrently, hard laws should adopt market-friendly but socially responsible regulations—such as guided pricing mechanisms, public service assurances through institutions like libraries, and legislative frameworks mandating AI literacy as part of education rights—to balance innovation with inclusive accessibility. Moreover, by dividing countries into core, near-core, semi-core, peripheral, and semi-peripheral categories according to their technological strength in the global AI system, the paper emphasizes that national legislative approaches must consider each country's position within the global AI technological hierarchy.

Ultimately, solving the challenges identified herein demands a strategic synthesis of ethical principles and enforceable regulations, ensuring that the transformative potential of AI technology genuinely benefits all sectors of society, both domestically and globally.

Integration of Soft and Hard Laws: Profiling Legal Protection for “AI for All”

Yang Xiao and Xiaoxia Sun*

摘 要：本文剖析了“AI普惠全人类”(AI for all, AIFA)的内涵，从理论基础、现实困境与法治路径三个维度展开研究。“AI普惠全人类”包括普惠AI技术（AI technology for all）、普遍AI正义（AI justice for all）和全球AI普惠（AI for all countries）三重维度。但当前的生成式AI的付费订阅模式正侵蚀这一理念，将加剧经济鸿沟与认知壁垒。当前“AI普惠全人类”依赖软法规范，但软法缺乏强制约束力，硬法规范又尚未健全。为防止AIFA从乌托邦愿景沦为反乌托邦现实，本文主张采取双重路径：一方面通过软法拓展明确可负担性、数字素养与国际合作等原则；另一方面依据国家AI实力分级构建硬法——在核心AI国家实施定价监管，对边缘国家推进全球算力合作。通过软法价值指引与硬法制度保障的协同，AI治理可在创新与包容间寻求平衡，使技术红利真正惠及全人类而非加剧不平等。

关键词：AI for all；人工智能治理；人工智能正义；人工智能监管；人工智能政策；全球人工智能框架；人工智能国际法

《联合国打击网络犯罪公约》制颁背景下的电子数据证据制度的完善

陈逸宁¹

(中国政法大学 刑事司法学院, 北京 100088)

摘要：随着信息网络技术的深度应用以及跨境网络犯罪数量的攀升，司法实践中跨境电子数据的需求也不断激增，这也对刑事司法领域内的国际合作以及刑事司法程序提出了更高的要求。传统的国际刑事司法协助路径取证效率较低，适用范围有限且成本较高，而单边直接取证的路线则存在引发国家主权冲突的风险。审查判断规则上，目前取证程序和证据可采性审查依据相互分离，传统的证据审查判断思路也无法完全契合电子数据本身的特殊性。涉外电子数据还存在较高的排除风险。为适应这种不断激增的数据入境需要，也为审慎处理国家司法主权、国家安全和国际关系，在当前我国牵头发起的《联合国打击网络犯罪公约》制颁出台、《刑事诉讼法》第四次修改以及全面依法治国要求统筹国内法治与涉外法治的背景下，有必要基于国家主权原则探索出一条兼顾现实性与前瞻性的取证方案，并明确境外电子数据审查运用的基本规则。

关键词：电子数据 跨境取证 司法协助 可采性

一、问题的提出

伴随着全球化及国家开放程度的深入，犯罪模式也在发生转变，跨国界犯罪的数量呈现明显增长，²而电子数据是打击网络犯罪的最佳证据，³也因此跨境电子数据取证的需求也呈现井喷式增长。⁴如何在虚拟无边界的网络空间及时获取电子数据，审查境外电子数据逐渐成为理论与实务界关注的重点。涉外取证主要有国际条约途径、司法协助途径以及外交途径三种。而对于电子数据并结合检察实践看，2021年《人民检察院办理网络犯罪案件规定》（以下简称《网络犯罪规定》）及张凯闵等52人电信网络诈骗案（检例第67号）展现出我国目前所坚持的跨境电子取证的立场，即坚持数据存储地主义下的刑事司法协助路径。但

¹ 作者简介：陈逸宁，中国政法大学刑事司法学院诉讼法学专业博士研究生，北京市丰台区人民检察院。

² 参见李生荣、李磊：《境外收集证据在内地的审查与认定》，载《人民法院报》2018年11月1日，第7版。

³ 参见辛素：《“一带一路”电子取证刑事司法协助问题研究》，载《北京警察学院学报》2018年第6期，第49-54页。

⁴ 参见戴佳：《最高检披露网络犯罪大数据，有图有真相！》，载最高人民检察院官方微博，<https://weibo.com/ttarticle/p/show?id=2309404623301518688447>，2024年7月8日访问。

是由于刑事司法协助双向性的特点，该路径涉及境内外诸多机构，程序繁琐，用时较长，并且适用范围也多限于案件涉及的国家较少且条约足以覆盖的国家地区。

为了应对上述不足并适应高效取证的现实需要，实践中一些侦查机关逐渐出现基于 2016 年《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》（以下简称《刑事电子数据规定》）和 2019 年《公安机关办理刑事案件电子数据取证规则》（以下简称《刑事电子取证规则》）等规范，利用在线提取、远程勘验及技术侦查措施等措施直接取证的探索。不过这种方式的规范依据有限，并且利用上述途径获取的数据是否具备证据能力也存在一定疑问，司法界对这一方式的态度也不一而足。并且涉外案件本身就关系到国家司法主权、国家安全和国际关系，具有高度敏感性，特别是在各国数据主权意识日益增强的背景下，这种不经他国数据主管部门同意而直接取证的方式也极易发生取证不合法的风险⁵以及国际礼让和数据主权等方面的争议。

此外，为了应对不断增长的跨境电子取证需要，⁶域外电子数据的取证模式也在发生转型，出现了一大批基于本国利益而展开的直接取证模式实践。例如以欧美国家为代表的“数据控制者标准”的提出，该模式的核心表现为一种将数据的实际控制者作为核心管辖要素的属人原则，即数据控制者需要履行按照相关部门的要求提供其所掌握的电子数据的义务。⁷域外的这种变革实际上也会对我国司法制度产生冲击。例如 2022 年 6 月美国法院就首次在 *Cadence Design Systems, Inc. v. Syntonic AB* 案（“Syntonic 案”）中认定《中华人民共和国个人信息保护法》（以下简称《个人信息保护法》）第 39 条并不能阻却中国当事人在美国诉讼程序下的证据开示义务。与此同时 2024 年 12 月 24 日联合国大会通过了《联合国打击网络犯罪以及为打击利用信息通信技术系统实施的某些犯罪并共享严重犯罪电子证据而加强国家合作公约》（以下简称《打击网络犯罪公约》）（United Nations Convention against Cybercrime; Strengthening International Cooperation for Combating Certain Crimes Committed by Means of Information and Communications Technology Systems and for the Sharing of Evidence in Electronic Form of Serious Crimes），两项核心议题之一就是跨境取证规则的设计，例如第 27 条就规定了有关跨境电子数据取证的提交令条款。我国应当如何应对上述变革所带来电子数据证据相关的体系性影响，国内立法应当如何调适和回应上述变革都需要展开深入

⁵ 参见刘梦：《跨境网络犯罪中非正式电子取证的合法性及其效力》，载《政法学刊》2021 年第 5 期，第 6 页。

⁶ 根据欧盟委员会的报告，85% 的刑事调查需要电子数据，而 2/3 的调查涉及跨境取证。See European Commission Staff Working Document, Impact Assessment Accompanying the Document Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for Electronic Evidence in Criminal Matters and Proposal for a Directive of the European Parliament and of the Council laying down Harmonized Rules on the Appointment of Legal Representatives for the Purpose of Gathering Evidence in Criminal Proceedings, 2018, p.14.

⁷ 参见陈爱飞：《“数据控制者标准”取证模式及中国因应》，载《法商研究》2024 年第 3 期，第 61 页。

的思考。在全面依法治国要求统筹国内法治与涉外法治的背景之下，就需要在基于国家主权原则的基础上，探索出一条兼顾现实性与前瞻性的数据取证方案，完善境外电子数据审查运用规则，并推进涉外法治和涉外国家安全机制建设。

二、数据入境需求下的管辖及取证模式

（一）电子数据主权的确定及管辖模式

跨境电子数据的特殊之处即在于跨境因素的存在，具体又可以分为需要调取的数据存储于境外以及他国需要调取的数据存储于我国两种。但是区别于其他证据主要存在于传统物理空间，电子数据所依附存在的网络空间则具有跨国界、无边界的特性，疆域不明确，国家边界弱化，犯罪场所虚拟化，也致使跨境因素的识别本身就具有复杂性。目前世界范围内电子数据的管辖主要分为数据存储地及数据控制者两种模式。我国采用的是传统的数据存储地标准。但是电子数据本身具有流动性、内容海量、混杂性及易改变的特性。加之与电子数据相关的区块链、暗网、云计算等技术的不断兴起，这种基于数据存储地主义而确定管辖的路径模式会遇到诸多困难。⁸例如，区块链技术具有存储位置不稳定及多位置的特性。云计算技术下的数据则存在多点存储，数据副本同时存储于不同的虚拟机和物理机上，涉及多个管辖区域。⁹至于暗网技术，由于暗网技术的通信实体的高动态性、网络结构异构以及资源要素隐蔽等特征，数据存储地在具体识别判断的时候还面临诸多障碍。

（二）传统管辖理论下的刑事司法协助取证模式

基于古典管辖理论，电子数据若存储于境外，司法机关需依托刑事司法协助完成取证。该路径的本质为国家间的合作，重点在于委托取证。该路径规范依据有《国际刑事司法协助法》第四章、《刑事诉讼法》第18条及《关于办理电信网络诈骗等刑事案件适用法律若干问题的意见》等。张凯闵等52人电信网络诈骗案（检例第67号）反映的也是这一立场。该路径于法有据并不易引发国际争端，但在数据快速流动的背景下，该路径是否依旧最佳方案也存在一些争论。

第一，该路径程序繁琐，取证效率不高。该路径涉及的部门多、人员广、流程繁杂，且数据提供方通常也无时效要求，故效率低下，该路径的通用时长往往多达10个月以上，¹⁰而实践中跨境警务合作还比较有限。与此同时，电子数据又具有易变性的特点，如何确保取证的及时性也十分重要。第二，该路径适用范围有限。该路径具体包含了根据国际条约、两岸司法互助协及通过国际组织途径取证三种模式，适用范围通常仅限与我国签订了司法互助协议或基于平等互惠原

⁸ J. Daskal, *The Un-territoriality of Data*, Yale Law Journal, Vol.125: 2, p.329 (2015).

⁹ 参见[英]克里斯托弗·米勒德：《云计算法律》，陈媛媛译，法律出版社2019年版，第418页。

¹⁰ 参见梁坤：《基于数据主权的国家刑事取证管辖模式》，载《法学研究》2019年第2期，第188-208页。

则可开展跨境电子取证的国家和地区。而目前我国是与 86 个国家签订过双边司法协助条约,数量还比较有限,并且已有的条约的约定内容还比较有限,也容易发生超出协助条约条款以外内容的情况。同时部分犯罪数据是分散于多个国家,加之电子数据高传输性的特点,如何与多个国家开展刑事司法协助也会面临难题。第三,根据《刑事电子数据规定》及《刑事电子取证规则》,我国强调原始存储介质扣押保存的优先性,但就国际刑事司法协助电子数据取证的实践看,存储载体无法取回以及取回成本较高的情况还屡见不鲜。第四,境外侦查机关取证、保管和移交过程还有待规范,影响数据真实性、完整性的风险因素也较多。正因此,2021 年《关于办理电信网络诈骗等刑事案件适用法律若干问题的意见(二)》第 14 条允许我国公安机关对保管链进行书面说明并经审核后作为证据使用。第五,一些学者还认为固守司法协作模式等同于对网络空间主权的默示放弃,还存在架空网络空间主权原则的风险。¹¹

(三) 单边跨境取证路线的适用及其争议

有鉴于此,也为便利高效取证,2016 年《刑事电子数据规定》则开始从侧面授权侦查机关开展单边直接取证。其是指当数据原始存储介质位于境外时,侦查机关通过在线提取、网络远程勘验及技术侦查措施等方式收集,而无需扣押原始存储介质。由于其便利性,现实中侦查机关也更倾向此路线。但 2019 年《刑事电子取证规则》将在线提取的范围限定于公开发布的电子数据,并基于此规定了网络远程勘验及技术侦查等措施,但是对于不公开的跨境电子数据能否适用上述措施则还存在一些争议。先就三种措施的内涵及差异看,一是网络在线提取措施类似于直接从公开网络空间下载信息,适用依据是 2019 年《刑事电子取证规则》。¹²二是网络勘验措施,其是指在已经掌握账户密码的情况下通过登录境外服务器勘验并提取电子数据的一种措施。根据《刑事电子数据规定》第 9 条第 3 款,除犯罪嫌疑人自愿配合提供账号密码后通过服务器取证外,实践中的远程勘验还包括未经同意利用技术侦查获取账号密码后取证等情形。¹³三是至于权利干预色彩最为浓厚的技术侦查措施,《刑事电子数据规定》第 9 条第 3 款似乎也从规范层面直接授予侦查机关直接开展技术侦查的权力。但是对于第二、三种措施来说,2019 年《刑事电子取证规则》并未明确,通过上述措施所获取的不公开证据的合法性还存争议,上述证据也恐难进入审判程序。同时,虽然规范层面进行了明确区分,但实践中在线提取等措施与远程勘验措施混同适用的情况依旧比较普遍。

14

¹¹ 参见陈丽:《跨境电子取证的中国应对》,载《国家检察官学院学报》2022 年第 5 期,第 25 页。

¹² 参见万春等:《〈关于办理刑事案件收集提取和审查判断电子数据若干问题的规定〉理解与适用》,载《人民检察》2017 年第 1 期,第 53 页。

¹³ 参见《湖南省郴州市中级人民法院刑事判决书》(2016)湘 10 刑终字第 69 号。

¹⁴ 参见《云南省楚雄彝族自治州中级人民法院刑事判决书》(2016)云 23 刑终字第 82 号。

并且,单边跨境取证虽整体回应了侦查机关高效取证的需要,但问题同样突出。一是诸如经犯罪嫌疑人同意而采取的远程勘验等措施与我国外交部门在数据出境领域对2001年《布达佩斯公约》第32条b款直接表达的反对立场相左,也与我国既有的刑事司法协助框架存在冲突。¹⁵二是上述路径本质上属于国内法的直接延伸,由于牵涉网络空间相关的国家主权,一旦处理不好在影响司法进程的同时,也会对国家外交关系产生负面影响。特别是远程勘验措施和技术侦查等措施的权利干预色彩都比较浓厚,制约监督机制还有限,公民隐私、自由被侵犯的风险也较高。三是《刑事电子取证规则》颁布后,司法机关对于这种直接通过技术手段所获取的非公开数据直接作为法庭证据使用还持谨慎的态度,并多以此否定此类数据的证据资格。¹⁶四是,传统刑事司法实践中不加区分的电子数据调取模式也容易带来个人隐私权利被侵犯的风险。

三、域外网络犯罪刑事管辖权的扩张倾向

就域外司法发展来看,当前电子数据领域的刑事管辖界限标准发生一些松动。特别是伴随取证需求量的快速增加,一些国家开始通过立法改革重塑境外电子数据的概念,以期突破传统的刑事司法协助与警务合作框架,并搭建起一种高效取证与合作机制。而其中的代表就是“数据控制者”标准的提出。这一标准相关的代表性立法主要有2001年《布达佩斯公约》(Budapest Convention on Cybercrime),2010年《阿拉伯国家联盟打击信息技术犯罪公约》(Arab Convention on Combating Information Technology Offences),2018年美国《澄清合法使用海外数据法》(Clarifying Lawful Overseas Use of Data Act)以及2022年《<网络犯罪公约>关于加强合作和披露电子证据的第二项附加议定书》(以下简称《第二项附加议定书》)Second Additional Protocol to the Cybercrime Convention on enhanced co-operation and disclosure of electronic evidence)等。上述法案均是将境外电子数据的界定标准由传统“数据存储地”转变为“数据控制者”,即将部分境外电子数据认定为境内电子数据,以避开司法协助与警务合作取证。而这一趋势也折射出域外网络犯罪管辖权领域的一种扩张倾向,而就上述趋势的具体发展进程来看,其中主要有两种代表模式。

第一种模式代表是《布达佩斯公约》,该公约充分依托欧盟政治一体化与内部高度开放的优势条件,允许基于授权原则而展开取证,具体表现为一种“授权模式”。即除在认可电子数据存储介质属地管辖之外,还规定了两种突破属地管辖模式的例外方式。《布达佩斯公约》第32条a款规定取证国可直接通过网络在线提取公开发布的电子数据,第32条b款则允许在经由具有“合法授权”的主

¹⁵ 参见梁坤:《跨境远程电子取证制度之重塑》,载《环球法律评论》2019年第2期,第141~142页。

¹⁶ 参见赵宪伟、刘政、徐衍:《海量电子数据审查的实现路径》,载《中国司法鉴定》2022年第1期,第33页。

体在合法且自愿同意的基础上通过境内计算机系统调取存储于境外的电子数据。由于 a 款指向的公开发布数据并未对其他国家的司法主权造成过多干预,司法实践中争议较少。但是对于 b 款所指向的电子数据,上述方案虽然一定程度上解决了取证及时性的需要,提高了取证效率,但是该条款适用范围的扩张性也容易引起国家安全与数据主权的对峙,也因此诸多新兴网络国家对此多持谨慎的态度。同时值得注意的是《布达佩斯公约》第 18 条还设置了出示令 (production order) 条款以授权缔约国的执法机关命令其境内提供服务的服务供应商提供其拥有或控制的与此类服务有关的用户信息 (subscriber information), 此处的用户信息是指除流量数据 (traffic data) 和内容数据 (content data) 以外的任何数据信息。¹⁷此外, 2022 年《第二项附加议定书》还对出示令条款在适用的数据类型以及调取程序方面进行革新, 即区分了域名注册信息 (domain name registration information) 与用户信息 (subscriber information), 并特别严格了调取用户信息相关数据的调取程序, 同时允许缔约国对此提出法律保留。¹⁸不过上述条款中的“合法授权”以及“合法自愿同意”表述的内涵与外延都不十分明确, 而出示令的范围也主要限于“非内容数据”, 其适用范围还比较有限。并且由于服务供应商所在国数据安全法规的不同以及为避免可能引发的数据安全风险, 实践中各网络服务供应商的配合度不一。即便是服务提供商不配合取证, 公约也未明确规定不配合取证行为的法律后果。鉴于此, 加之受到美国立法的影响, 欧盟在 2018 年还通过了《关于刑事犯罪电子数据提交令与保存令条例的提案》授权成员国的执法机关有权命令服务供应商提交或保存电子数据, 并且该条款的适用对象不限于欧盟境内的供应商, 还包括了与欧盟存在“实质联系”(substantial connection) 的对象, 并规定了诸如强制执行等制度。¹⁹

第二种则是以美国《澄清合法使用海外数据法案》(Clarify Lawful Overseas Use of Data Act, Cloud Act) (以下简称《云法案》) 为代表的侦查机关向数据控制者直接调取模式, 即“长臂管辖模式”。此种电子数据获取模式的调取思路更为激进, 即认为只要通讯服务的提供者受到美国管辖, 如具有商业实体或分支机构, 侦查机关就可以通过搜查令或法庭命令的方式要求数据服务的提供者收集存储于境外的数据。该模式充分利用了美国作为全球数据资源产业链中的优势地位, 突破了考虑电子数据存储位置的传统模式, 绕过了传统司法协助与执法合作渠道。但是这一激进策略的背后也引发了不少争议, 一方面, 这种绕开他国执法机关而

¹⁷ See Council of Europe, Convention on Cybercrime, <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680081561>, Last Accessed on Jan.24,2025.

¹⁸ See European Union, Second Additional Protocol to the Cybercrime Convention on enhanced co-operation and disclosure of electronic evidence, <https://rm.coe.int/1680a49dab>, Last Accessed on Jan.24,2025.

¹⁹ See European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018PC0225>, Last Accessed on Jan.24,2025.

直接调取的行为增加了对他国数据主权安全损害的风险，容易引发数据管辖权冲突；另一方面，数据控制方通常缺乏应对执法机关要求的专门机构。并且对于云存储服务提供者等对象，这种提取要求本身可能就和云储存协议的内容存在冲突，相关服务提供者的协助积极性本身就不高，并且《数据安全法》关于服务提供者协助义务的强制性规定也会造成服务提供者作为侦查协助者与作为数据保密者的角色冲突。同时为了应对随之可能带来的国际关系紧张与法律冲突，美国还将国际礼让的思想嵌入《云法案》之中，即为符合美国法要求的适格的外国政府提供调取电子数据的合作渠道，其本质在于创造一种取代已有的双（多）边刑事跨境取证模式的“美国标准”跨境刑事调取模式。

在各国普遍强调对国家数据以及本国公民数据的所有权、控制权与使用权的背景下，跨境电子数据收集本身就容易引发数据主权与隐私保护的冲突。特别是就其中我国的外交立场来看，我国并非《布达佩斯公约》的缔约国，也曾对该公约以及美国《澄清合法使用海外数据法案》等提出过明确的反对意见。例如，2020年我国就在《全国数据安全倡议》中就强调，“不得直接向企业或个人调取位于他国的数据”。并在诸如《中阿数据安全合作倡议》《“中国+中亚五国”数据安全合作倡议》中申明了尊重他国主权和数据安全管理权的立场。特别是在涉及敏感信息或国家安全的情况下，取证的障碍也会成倍增长，并且也容易引发国际礼让方面的纠纷并对数据主权的构建造成障碍。近年来，我国通过《数据安全法》《国际刑事司法协助法》等法律表达了对未经批准的数据出境行为的强硬反对与防御立场，2021年最高检《网络犯罪规定》第56条也明确规定了“尊重协作国司法主权”原则的立场。故而目前在跨境电子数据取证方面，基于国家数据主权与安全的考虑，我国的官方立场更倾向于优化司法协助与执法合作，对于诸如直接跨境调取电子数据的方式的态度还比较谨慎。

但是值得注意的是我国同时也是《打击网络犯罪公约》的主要发起国，为了推进成员国间在网络犯罪情报共享与联合执法的协作，该公约也提出建立统一的跨境电子数据取证相关的技术标准和规则，强化了公私合作，并在第27条规定了与《布达佩斯公约》相类似的提交令制度，即授权各缔约国主管机关下令在本国领土内提供服务的服务提供者提交其所拥有或控制的与此类服务有关的订阅用户信息。²⁰虽然该规定目前还仅限于订阅用户信息，权利干预属性相对较弱。但是上述制度设计已经与我国已有的法律规范和严格司法主权的外交立场存在较多差异。面对域外的提交令，我国司法机关是否允许，以及我国司法机关直接

²⁰ 订阅用户信息系指服务提供者持有的与其服务订阅用户有关的任何信息，但不包括流量数据或内容数据，通过订阅用户信息可以确定：

- 1.所使用的通信服务种类、与之相关的技术条款以及服务期限；
- 2.根据服务协议或安排提供的订阅用户身份、邮政地址或地理地址、电话或其他接入号码、账单信息或付款信息；
- 3.根据服务协议或安排提供的关于通信设备安装地点的任何其他信息。

向服务提供者取证的时候是否需要通知域外司法机关以及发生拒绝的情况下如何应对都需要深入的研究，同时我国还应当积极调适国内立法以应对《打击网络犯罪公约》制颁后可能带来的影响。

四、跨境电子数据审查的标准及其规则判断

除却电子数据取证方面的障碍，电子数据的收集与审查依据分类、跨境因素识别、技术性鉴真规则等问题同样值得关注。对于经由司法协助路径获取的电子数据来说，在准据法方面，由于跨境等因素存在，收集与审查的法律依据相互分离，并且受到国际刑事司法协助条约与国内刑事诉讼法双重规则影响。具体表现为基于国家司法主权保护原则，电子数据的收集工作通常是由证据所在地的司法机关完成，但是在缺少请求方侦查人员的情况下，国外警方的取证积极性与规范性通常不高。同时，各国数据保护以及隐私权保障的规范不尽相同，取证程序与方法还存在明显差异，电子数据的证据提取与固定规则也不同。这种取证程序与证据可采性审查依据的相互分离也同时带来了证据被排除的风险。至于非经司法协助或警务合作所取得的境外电子数据，此类证据的证据效力本身就存在不被接受的风险。例如，就曾有辩护人对境外在线提取电子数据的证据效力提出质疑。

21

（一）规范与实践：跨境电子数据审查的准据法

过去司法机关通常不审查经由司法协助渠道而获取证据的合法性，但是伴随着司法公正以及人权保障观念的发展，双重审查标准逐渐兴起并为各国所接受，即一种同时依据本国法和被请求国法对跨境证据的合法性进行审查的标准。²²而有关我国境外证据审查的准据法也经历了从无到有的发展阶段，现有的规范体系主要有《关于办理电信网络诈骗等刑事案件适用法律若干问题的意见》（以下简称《电信网络诈骗意见》）《刑事电子数据规定》《国际刑事司法协助法》《关于适用〈中华人民共和国刑事诉讼法〉的解释》（以下简称《刑诉解释》）。就已有的准据法规范体系来看，我国整体沿用的是国内准据法审查模式，即根据《刑诉解释》第77条，法院判断境外证据合法性的准据法是我国《刑事诉讼法》，并未将域外国家取证程序的规范性考虑在内。而就不同犯罪类型下通过不同方式获取的证据的证据能力审查具体规则来说，根据《电信网络诈骗意见》第6条和《办理跨境网络赌博犯罪案件若干问题的意见》第7条，通过刑事司法协助或警务合作方式获取的证据的审查主要依据的是侦查机关的情况说明，至于通过其他方式获取的境外证据则还需要通过是否符合我国刑事诉讼法规定的证据能力审查程序。

但是根据《人民检察院办理网络犯罪案件规定》第32条，对于跨境电子数据，合法性审查依据实际上包含了国（区）际司法协作及相关法律规定，也因此

21. 参见四川省攀枝花市中级人民法院（2020）川04刑终79号刑事裁定书。

22. 参见陈苏豪：《跨境刑事证据审查判断的要素检视》，载《现代法学》2023年第6期，第193～194页。

我国并未采取绝对的国内准据法审查模式。²³但是这一条款的具体指向并不明确，并且就上述条款的本质来看，更多体现为对跨境电子取证管辖权的谨慎态度而非对于国内准据法审查模式的直接规定。同时就司法实践看，目前虽然没有明确将外国法作为准据法的案例，²⁴但是部分案例还是体现了双重审查标准的倾向性意见。以最高检第一批国际刑事司法协助典型案例——李某南故意杀人案为例，司法机关在对证据进行审查判断时就比较了中美询问证人的程序差异，并从报告的形成过程、与其他证据的相互印证角度认定美国警方的访谈报告具有可采性，一定程度展现了域外法查明以及结合域外法程序进行审查判断的模式特征。

整体而言，这种“国内准据法”为主的审查模式未能完全适应司法实践的需求。一是这种模式规定还过于原则、粗疏，具体的界定标准和法律后果规定尚不明确，非法证据排除规则的适用还存在规范层面供给不足的问题。二是《刑事诉讼法》主要局限于物理空间的国内犯罪，对于数字类证据的跨境取证以及境外证据审查的关注还较为有限。三是已有的关于经由刑事司法协助或警务合作方式取得的境外证据的审查模式过于形式化，缺乏实质审查，还存在不符合我国刑事诉讼法审查标准以及低于我国人权保障标准的风险。

（二）困局与突破：跨境电子数据的审查规则

电子数据能否作为证据使用还需要进行可采性与证明力的审查，但是电子数据种类多且数量大，并且电子数据可与原始存储介质相互分离，同时此类证据还具有稳定性与易变性并存，产生与获得容易，内容与形式又易受不同环境、条件的影响而发生根本性变化的特性，此类证据自身的复杂性与技术性也传统的证据审查体系带来了一定挑战。电子数据审查的核心依旧是证据“三性”。《刑事电子数据规定》第2条就明确规定司法机关应当围绕合法性、真实性和关联性进行审查判断。整体而言，在境外证据审查方面，目前司法界的关注重点集中在证据转化的连续性与真实性上，具体的审查内容也聚焦在传统审查思维下的技术程序和保障程序的真实合法性方面，并且格外重视程序链条的规范性。有关证据排除的实践也大多集中在取证程序的瑕疵性。同时为了确保司法协助的顺利开展，目前司法机关的审查标准还较为宽松，多以不审查为原则，即默认被请求方取证行为的合法性，具有明显的形式化特征。²⁵

1. 跨境电子数据的合法性审查

跨境电子数据合法性的审查主要聚焦于取证环节的规范性。根据《刑诉解释》

²³ 参见曹艳琼：《我国境外证据审查的准据法模式选择与规则重构》，载《法商研究》2023年第3期，第189页。

²⁴ 参见吴国章：《刑事境外证据“绝对本国法”审查规则的反思与重塑》，载《法学》2023年第10期，第128页。

²⁵ 参见冯俊伟：《论跨境追诉中的辩护权保障——主要以境外取证为例的分析》，载《中国刑事法杂志》2019年第4期，第129~144页。

第77条、2018年最高检《检察机关办理电信网络诈骗案件指引》对境外证据审查内容的描述以及2021年《网络犯罪规定》，已有的关于通过境外获取证据的合法性审查规则可以归纳为如下三点：一是境外证据的取得应当具备合法性依据，即符合相关公约、条约及合作协议；二是境外证据的取得应当遵循法定程序，即收集主体、收集手段以及证据表现形式都需符合《刑事诉讼法》；三是证据来源的合法性审查具体包括了审查报批、审批手续以及移交手续是否完备等环节，对于不符合我国证据种类及程序要求的证据材料，则需要重新转换与固定。2021年《关于办理电信网络诈骗等刑事案件适用法律若干问题的意见（二）》（以下简称《电信网络诈骗意见（二）》）第14条还允许一定条件限制下对证据链要求标准的放松，这是因为司法协助路径下电子数据移送不完整、以证据清单和情况说明作为替代的问题还十分普遍，若是审查标准过于严苛，跨境电信网络诈骗往往会面临无证据可用的境地，这也算是适应司法实践需要下的无奈之举。综合上述审查规范来看，目前关于电子数据合法性的审查标准已有一定规定，但是如前所述2016年《刑事电子数据规定》以及2019年《刑事电子取证规则》还将网络远程勘验和网络在线提取等措施界定为任意性侦查措施，还未能将可能涉及的公民基本权利等考虑其中，还存在取证违法的风险，与此同时刚性的制裁规范也较为欠缺。

2. 跨境电子数据的真实性审查

电子数据的真实性审查又叫电子数据的鉴真规则，电子数据的真实性包含了电子数据载体、数据本身以及数据内容的真实性三个方面。真实性审查的功能在于保障刑事审判时出示的证据与案件发生时证据的同一性。一方面就规范领域来看，规范层面与电子数据鉴真相关的原则、方式以及法律后果的规定还散见于不同位阶的各类法律规范中，各条款还存在条文内容重复、规范内容逻辑不一致的问题，²⁶并且已有的鉴真标准也比较模糊，虽然关注到了电子数据的真实性等方面，但是并未对电子数据进行特征化分类，也未结合电子数据的特性进行内外载体方面的区分。另一方面就司法实践情况来说，现有的真实性鉴定模式还较为单一，实践中常用的“保管链证明”和“独特性确认”等方法还未能完全契合电子数据所具有的虚拟性、可分离性以及海量性的特性。²⁷特别是对于跨境取得的电子数据来说，原始存储介质的获取本身就存在一定障碍，在未获得原始存储介质的情况下也很难通过存储介质完成相关数据的鉴真工作。并且在涉及海量电子数据的时候，传统的鉴真方法还很难胜任此类证据同一性的比对和确认。此外实践中侦查机关收集提取电子数据的设备技术水平还参差不齐，而完整性校验、可信时间戳、数字签名、区块链存证等新技术也尚未在规范与司法实践中得到充分的回应。

²⁶ 参见冯俊伟：《刑事证据分布理论及其运用》，载《法学研究》2019年第4期，第188页。

²⁷ 参见谢登科：《电子数据的技术性鉴真》，载《法学研究》2022年第2期，第209页。

3.电子数据完整性的规制争议

作为信息技术发展衍生下的新型证据种类,电子数据具有独特的系统性特征,在电子数据相关的构成与审查要素判断领域,除却传统的“证据三性”,晚近的司法解释和相关理论研究还提出并诠释过关于电子数据的完整性的概念,并逐渐将这一概念从简单的技术话语上升为规范层面的要求。首先是在电子数据完整性的内涵与功能界定上,完整性着重强调电子数据在收集提取、固定保存以及审查判断时不受篡改和破坏而具有完整性的要求,涉及电子数据数据载体、数据以及数据内容三个层面。同时电子数据的完整性与电子数据的真实性有着密切关系,故而《刑事电子数据规定》第22条将其纳入真实性的审查范畴。同时完整性的适用场域相对有限,主要集中于电子数据领域的发展。其次是就该属性的发展沿革看,完整性这一概念起源于1996年联合国国际贸易法委员会《电子商务示范法》第8条,我国电子数据领域在立法起草时也就这一问题给予过关注。²⁸例如《刑事电子数据规定》第23条就明确要求从原始存储介质的扣押和封存状态、电子数据收集过程等方面审查电子数据的完整性。而且该规范中关于“完整性”提及的次数也远多于“合法性”“真实性”被提及的次数,这也展现出电子数据完整性从边缘到核心的发展趋势。此外伴随着信息技术的发展,诸如2022年《关于办理信息网络犯罪适用刑事诉讼程序若干问题的意见》及2021年《网络犯罪规定》也对完整性的问题予以关注,并试图推进其在区块链存证及哈希值校验活动中融合发展。综上,伴随着《刑事电子数据规定》《刑事电子取证规则》以及《刑诉解释》等一系列规范的出台,有关电子数据的完整性实际上已经形成了包括取证、鉴真、审查和排除在内的较为完整的规范体系。²⁹再次就完整性在电子数据审查要素体系中的地位来说,虽然电子数据的完整性被给予了充分的重视,但是电子数据的完整性与证据三性等属性间的关系还较为混乱,完整性究竟是属于真实性的下位概念,还是证据充分性的评价要素,还是一种证据的独立要素属性还存在较多争论。并且实践中完整性被错误规制甚至被泛化适用的情况也较为常见。³⁰例如在司法适用中,电子数据真实性的实质性审查受到忽视,合法性审查被边缘化,关于电子数据的审查步骤也更多被简化为对完整性的审查,特别是对于证据保管链完整性的审查。就需要对这种仅关注电子数据的完整性而忽略了其他属性规范的要求的完整性泛化使用的倾向予以充分警惕,并做好电子数

²⁸ 包括《电子签名法》第5条,《刑事诉讼法》第123条第2款,两高一部《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》第5条、第22条、第23条,公安部《公安机关办理刑事案件电子数据取证规则》第2条、第41条、第42条,最高检《人民检察院公诉人出庭举证质证工作指引》第39条,最高法《人民法院计算机信息网络系统建设管理规定》第12条,最高检《检察机关办理电信网络诈骗案件指引》第6条,《杭州互联网法院电子证据平台规范(试行)》第8条、第15条等规范都曾提及完整性相关的要求。

²⁹ 参见吴洪淇:《电子数据完整性的法律定位与理论反思》,载《国家检察官学院学报》2024年第1期,第147~154页。

³⁰ 参见陈丽:《论电子数据的完整性》,载《证据科学》2021年第6期,第707页。

据相关的证据审查规范体系建设。

4.跨境网络犯罪的证明模式转型

电子数据的特殊性还在于经常会涉及海量的证据材料,但是不同电子数据的证明逻辑还不尽相同。特别是在跨境网络犯罪这类案件中,被害人的数量以及牵涉的资金通常都较多,电子数据也呈现海量性特征。传统的印证证明模式为一一对应式的证据审查模式,这种模式难以应对上述案件可能涉及的海量银行流水及聊天记录,机械僵化的审查模式无形之中也会增加审查的成本,并与司法审查的客观实际不相匹配。³¹当传统的严格证明模式在电信网络诈骗犯罪领域发生证明乏力的情况下,2016年《关于办理电信网络诈骗等刑事案件适用法律若干问题的意见》第6条第1款和2024年《关于办理跨境电信网络诈骗等刑事案件适用法律若干问题的意见》(以下简称《电诈意见》)第5条还就涉案数额以及被骗人数的证明问题中提出了一种综合交易记录、聊天记录等证据材料综合认定的审查模式。但是这种综合认定模式属于一种运用经验法则的间接证明方法,其核心是围绕计算机等信息系统中的资金流与信息流两个维度的综合证明方法。这种证明模式虽然一定程度上减轻了司法人员的证明负担,提高了证明效率,但是已有规范关于综合认定审查模式的表述较为宽泛,缺乏明确指引,并存在被扩大适用、适用标准不统一以及侵害被告人权益的风险。例如司法人员就可能因为被害人众多在无法逐一收集被害人陈述的情况下降低证明标准和转移证明责任。有必要针对上述缺陷从规范层面进一步回应。此外特别针对海量数据,为了进一步减轻取证与审查证明负担,司法实践中还出现了一种基于基础事实,主要依赖经验与逻辑法则的推定证明模式。³²例如《电信网络诈骗意见(二)》第3条和《电诈意见》第7条都规定了通过赴境外诈骗犯罪窝点的时间与次数的方式认定符合《刑法》第266条规定的“其他严重情节”的证明规定。同时2022年《关于办理信息网络犯罪案件适用刑事诉讼程序若干问题的意见》第20条也规定了采用综合认定、推定、抽样证明的方案推定证明模式。但是至于具体个案是否适宜抽样以及具体应当如何抽样目前都还缺乏细致的规范指引,而如何保障现有手段的科学性,推进信息抽样代表数据的全貌也都需要明确具体的规范指引。

5.控辩双方举证责任的不平衡

司法实践中,审判机关经常基于《刑事诉讼法》第77条的规定,对诸如辩方等私权主体提供的境外刑事证据附加公证、认证等程序性要求,而对于控方来说,在证据属性的审查上,只要电子数据得到妥善的处理,审判机关就会作出有利于控方的推定。此外对于证明责任与证明标准的规定,如前所述包括《电诈意

³¹ 参见程雷等:《〈关于办理信息网络犯罪案件适用刑事诉讼程序若干问题的意见〉的理解与适用》,载《人民检察》2022年第19期,第27页。

³² 参见智嘉译:《电信网络诈骗案件中的证据问题研究》,载《法律适用》2022年第9期,第172页。

见》《电信网络诈骗意见（二）》等规范都在电信网络诈骗的罪量证明上予以了一定简化。虽然这种做法一定程度便利了刑事审判的审查工作，但是侧面也加重了私权主体举证责任的负担，造成控辩双方的举证失衡。

五、跨境电子数据取证思路的改进

跨境取证工作的背后实际上是丰富且复杂的主权冲突，为此也应当从解决国家主权冲突的角度出发，平衡国家执法权力合法性、个人权利保护以及外国主权之间的关系，兼顾我国自身涉外法治的构建。同时，我国还应当高度关注域外取证规则的变化趋势，跟进最新研究，探索出既符合国际法治发展形势，又契合实践需要的取证思路。

（一）跨境电子数据取证的基本原则

尊重主权是国际交往的基本准则，作为国家主权在网络空间的延伸与发展，网络空间主权也格外重要。对于跨境电子数据取证来说，虽然电子数据分散且高速流动的特性在一定程度上造成了传统管辖模式的识别困境，但是司法主权与数据主权问题同样不可小觑。对内，应当继续探索自主立法规制与管理网络安全活动，构建起整体符合网络空间主权原则的法律框架体系。对外，则应当尊重网络主权平等的原则，坚持跨境取证的主流模式——司法协助路径，保障自身在网络空间的平等与独立，并在遭受干预与攻击的时候充分行使网络自卫权。而作为解决国家间冲突的一项基本原则，国际礼让原则在国际数据领域中同样重要，该原则的核心是国家利益的权衡以及国家主权的自我约束。在国际数据法律领域尚未达成共识而数据保护与境外数据获取矛盾又持续加深的背景下，可以适度借鉴国际礼让原则的适用思路并完善数据出入境规范。

（二）跨境电子取证的具体方案

第一，坚持司法合作的主路径，持续优化司法协助路径。其一，在具体的协助程序上，探索简化各国间的文书程序，通过规范性文件明确合作方式、合作范围以及主要对象等事项以便提高协助取证的效率。在具体开展模式上，结合数据类型及他国的数据管辖模式采取灵活多边的取证方案，特别是积极利用他国法律框架内允许的渠道或方式进行创新。例如，美国《加州法案》规定企业从可验证的消费者处接到访问个人信息的请求后，应当立即披露和提供所要求的个人信息，³³我国司法机关应当充分参考并利用上述途径。其二，国际立法上，积极参与国际跨境数据执法规则的讨论与修订，并参与签署更多的司法协助文件。例如，积极参与当前《打击网络犯罪公约》的后续工作，融入并推动形成联合国框架下的跨境数据国际执法机制。这也是充分参与全球治理，维护国家利益的重要体现。

³³ 参见廖斌，刘敏嫻：《数据主权冲突下的跨境电子数据取证研究》，载《法学杂志》2021年第8期，第158页。

同时鉴于国际公约缔结程序的复杂性与耗时过长,我国也可以在区域框架内秉持平等互惠的原则,借助“一带一路”的建设契机先行倡导建立区域间的电子数据条约,并提出符合各国利益的快捷取证道路方案,并未推动统一的跨境电子数据取证规则的建立提供参考。与此同时还应当加强诸如侦查取证等方面的技术领域的交流与协作,提升跨境执法协作的效率。

第二,适度创新跨境电子取证路径,提升取证效率。就域外现有的取证方案看,以美国《云法案》为代表的取证模式还过于激进,也与我国谨慎对待数据出境的外交立场相左。为避免激进路径下存在的弱化主权国家对境内电子数据掌控的风险,也为了适应《打击网络犯罪公约》制颁出台后的体系性影响,我国可选择部分吸收诸如欧盟《布达佩斯公约》调取模式的思路。具体来说,一是,在允许境外公开电子数据在线提取的情况下,基于信息天然具有的非竞争性特征,结合数据公开发布的情形、时间等因素,扩张公开电子数据的内涵,将诸如“深网”内的电子数据涵盖其中。二是,对于经由数据主体自愿同意而能合法呈现于我国境内终端的电子数据则允许通过录像等方式达成间接提取。

(三) 借《刑事诉讼法》第四次修改之机,协调国内法律规范

一是,进一步平衡数据出入境之间的立法立场,对《打击网络犯罪公约》设置的取证程序予以积极回应。特别是在当前倡导网络空间命运共同体构建的时代背景下,平衡数据出入境原则的一致性,结合《打击网络犯罪公约》中的数据分类安排以及电子数据相关的措施义务规定,调整国内规范中的数据分类,并结合各类数据的特点和数据保护要求的不同,做出阶梯式的保全与调取安排,并保障国内法律规范的逻辑自洽。二是,探索完善国内法的域外适用,允许一定情形下的延伸适用。即加强执法机关与网络服务提供者之间的直接合作,明确跨境电子取证的启动要求、程序、权利保障及义务冲突的解决途径,探索侦查机关与网络服务提供者的合作取证路径。³⁴三是,对于实践中的跨境非公开数据的远程勘验、技术侦查等取证路径,仍应当坚持不宜将经由上述路径而获取的电子数据直接作为合法证据来使用的立场,但是可以适度允许提供部分数据以作为办案的线索。与此同时,探索严格限定情况下的紧急豁免情形,³⁵例如,在诸如恐怖主义等严重犯罪领域,设计特殊的程序规定。³⁶

³⁴ 参见冯俊伟:《境外电子数据的取得与运用——基于第 67 号检察指导性案例的展开》,载《国家检察官学院学报》2021 年第 4 期,第 158~159 页。

³⁵ 参见叶媛博:《论多元化跨境电子取证制度的构建》,《中国人民公安大学学报》(社会科学版)2020 年第 4 期,第 55 页。

³⁶ See Ángel Tinoco-Pastrana, *The Proposal on Electronic Evidence in the European Union*, EuCrim, Issue1, p.46 (2020).

六、合理审查判断规则的构建

（一）境外刑事证据审查判断的准据法

不同于涉外民事诉讼境外证据证明，目前《刑事诉讼法》尚未规定境外证据可采性的审查规则，传统证据法研究也多集中于境内证据的取证程序及审查规则上。而就国际发展看，为了平衡国内外法律规范、司法主权以及人权保障等多种价值取向，准据法的选择上目前还存在相对本国法、绝对本国法及混合准据法等三种模式。其中，绝对本国法是以单一的本国法作为跨境证据审查的准据法，与之相反，相对本国法则是在以本国法作为合法性审查准据法基础上同时允许国际公约或条约优先适用等例外情形。如前所述，在境外刑事证据上，采取绝对本国法模式容易造成司法机关最终的完全不审查，而直接认可协助行为的合法性，不仅妨碍刑事正当程序的独立价值，也会影响到非法证据排除规则效力的发挥。³⁷在《国际刑事司法协助法》明确条约优先的立场基础上，也为适应不断深化的刑事司法合作现实需要，可以适度探索借鉴域外审查模式以及我国民事领域审查的经验，即在坚持司法主权的原則基础上，探索并优化现有的证据合法性准据法模式，强化域外法查明，通过官方途径调取域外法程序规则，加强证据合法性审查。同时，根据跨境取证的不同模式，对诸如单边跨境取证、联合调查取证以及代为取证等不同取证的模式特点探索适用不同的准据法。在诸如双边条约签订时考虑将请求国的法律效力扩大延伸至取证国，³⁸我国还应继续参与国际公约的讨论与修订，并在区域间构建起相对统一的准据法适用标准。

（二）电子数据可采性的审查判断规则

电子数据的审查判断具有其特殊性，应当借助《刑事诉讼法》第四次修改之机，梳理现有的电子数据审查判断司法解释。一是，针对司法实践中电子数据真实性与完整性或同一性的混同问题，规范层面首先要进一步厘清电子数据完整性与真实性、合法性间的关系，将证据审查判断中容易造成误读的完整性条款删除，尊重电子数据真实性本身的内涵。其次，厘清电子数据完整性与证据三性之间的审查流程关系。传统司法实践中，审判机关的证据审查与判断的关注点多集中于证据的“真实性”³⁹。但是基于检察机关自身所承担的证明责任及司法主权以及人权保障观念的不断兴起，目前检察领域规范文件表达的立场却与审判机关存在一定差异。例如，根据2018年《检察机关办理电信网络诈骗案件指引》“境外证据的审查”一节的规定，境外证据的合法性被置于真实性审查之前，这在2021年《人民检察院办理网络犯罪案件规定》第59条和《刑诉解释》中也有体现，特

³⁷参见吴国章：《刑事境外证据“绝对本国法”审查规则的反思与重塑》，载《法学》2023年第10期，第130～132页。

³⁸参见唐彬彬：《我国国际刑事司法协助中境外证据的采纳：现状与完善》，载《公安学研究》2020年第5期，第56页。

³⁹参见冯俊伟：《域外取得的刑事证据之可采性》，载《中国法学》2015年第4期，第260页。

点是格外聚焦“程序”这一核心要素，注重电子数据生成系统、技术等可靠性及电子数据调取、保存和流转规范性等方面的审查。同时，检例第 67 号案件也特别强调对境外获取证据合法性的审查。结合电子数据的本身特性及实践需求看，根据完整性作为基础铺垫的特点，对于此类证据的审查顺序，首先还是应当审查数据的完整性，并于此审查数据的合法性和真实性。⁴⁰至于具体的审查事项，合法性的审查上，应重点审查证据材料的形式是否符合我国《刑事诉讼法》，境外执法人员取证行为于我国的合法性，证据保管链的完整性，境外电子数据的公证以及认证程序的合法性。二是，至于完整性规则，一方面在审查流程上要关注提取、保管、转移、鉴定等涉及诉讼的全流程要素，另一方面在具体审查内容上，除单一地关注哈希值外，还需加入数字签名、数字证书等属性信息，收集、提取的过程录像，备份数据及访问操作日志等要素审查。三是至于证明标准，面对电子数据的海量性以及传统证明方式的无力性，一方面坚持刑事诉讼证明标准，另一方面继续细化等约计量、抽样过程和方法的具体规定与适用规则，解决网络犯罪中数量与数额认定难题。

七、结语

网络犯罪治理形势日趋复杂，而作为数字时代下的证据之王，电子数据的取证与审查直接关系跨境网络犯罪的打击实效。而检察机关在跨境电子数据取证、移交与审查运用等方面都发挥着重要作用，需要正视数字技术带来的法律挑战，研判《打击网络犯罪公约》制颁后的体系性影响，并对国内法律规范进行适当的调整适配，在基于国家主权原则的基础上探索一条兼顾现实性与前瞻性的取证方案，并明确境外电子数据审查运用的基本规则。

⁴⁰ 参见吴洪淇：《被嵌入的程序空间：庭审排非程序十二年观察与反思》，载《中国法律评论》2022 年第 6 期，第 60～71 页。

集体反措施在网络空间适用的合法性问题分析

刘钰祺*

摘要：集体反措施在《国家对国际不法行为的责任条款草案》中未予明确，近期因多国网络空间立场文件的重申而引发争议。从一般国际法的角度分析，集体反措施的合法性存在诸多阙如：其国际法依据不足、非直接受害国采取反措施缺乏法律依据，以及集体反措施中代理与权利转让缺乏法律基础。在网络空间领域，集体反措施应然遵循一般国际法框架。网络特性未增其合法性，反措施仍受受害国主体限制。鉴于集体反措施在网络空间的适用议题讨论日益激烈，我国应秉持谨慎负责态度，引领客观对话，同时以推动全球与区域网络安全的协同建设与多维合作，共同应对国际不法网络行为。

关键词：集体反措施 网络空间 国际法 非直接受害国 合法性争议

一、问题的提出

集体反措施（Collective Countermeasures）作为一个国际法概念，其表述形态多样，诸如“受害国以外的国家采取的反措施”（Countermeasure taken by States other than an injured State）及“第三方反措施”（Third-Party Countermeasures）等术语亦有所见。在联合国国际法委员会（ILC）的《国家对国际不法行为的责任条款草案》（以下简称《国家责任条款草案》）第 48 条之语境下，集体反措施原本系指非受害国针对违反对世义务（obligations *erga omnes*）的国家所采取的措施。¹而随着网络空间安全问题日益突出，集体反措施的内涵业已拓展，囊括了一种新兴情况，即第三国应受害国之请求，代其行使权利，对不法行为国实施反措施。²2022 年 10 月 13 日，中国代表在第 77 届联合国大会第六委员会发言时，指出各国在“受害国以外的国家采取措施”问题上存在分歧。集体反措施的研究对国际法的完善与国际秩序的稳定至关重要。³

集体反措施作为反措施的延伸，意在通过集体的协同行动，敦促责任国履行其应尽的国际义务。在一般国际法框架内，反报（retorsion）、反措施和自卫均

* 武汉大学国际法研究所博士研究生。

本文系 2020 年国家社科基金重大项目“网络空间国际规则博弈的中国主张与话语权研究”（20&ZD204）的阶段性成果。

¹ See United Nations International Law Commission. Draft Articles on Responsibility of States for Internationally Wrongful Acts, 2001, Art. 48.

² 此即所谓“代理反措施”（Proxy Countermeasures）。See Miles Jackson and Federica Paddeu, Proxy Countermeasures in International Law, para.1. <https://www.ejiltalk.org/proxy-countermeasures-in-international-law/>, visited on 16 January 2025.

³ 参见中华人民共和国常驻联合国代表团：《中国代表刘洋在第 77 届联大六委关于“国家对国际不法行为的责任”议题的发言》，2022 年，http://un.china-mission.gov.cn/hyyfy/202210/t20221014_10783209.htm，2024 年 11 月 30 日访问。

是合法的自助（self-helping）手段，用以对国际不法行为做出回应。⁴虽然三者在某些方面逻辑上关联或表面相似，但在法律含义和适用条件上各有其独特的内涵和范围。反报，是指一国对另一国不礼貌、不友好、不公正但不违法的行为所采取的回应措施。其行为本身合法，如中止经济援助等，此类举措并不违背国际法的基本原则，除非受特定条约义务的限制。⁵反措施与自卫权，则本身具有不法性，但由于它们是对先前侵害或威胁的回应，因此在一定条件下得到了国际法的豁免。⁶反措施是受害国针对责任国不履行国际义务所采取的措施，其目的在于迫使责任国继续履行义务，具有“临时性”和“回应性”⁷。自卫权则是指一国在遭受外来武力攻击时，使用武力抵抗攻击，以保护自己的固有权利或自然权利。⁸它是国际法中最为严格的回应机制，其行使必须符合《联合国宪章》第 51 等国际法的严格规定。⁹

在编纂《国家责任条款草案》的过程中，ILC 就除受害国以外的其他国家采取反措施之权能展开了广泛深入的探讨。传统国际法理论视国际不法行为为受害国与责任国之间的双边事务，自卫权等法律手段主要由直接法律利益受损之国家行使，反措施亦遵循此双边逻辑。¹⁰然至 21 世纪，特别报告员詹姆斯·克劳福德（James Crawford）突破性地提出了“集体反措施”概念，并对责任国所承担的双边与多边义务进行了明确界分。¹¹基于此，ILC 尝试将集体反措施作为应对严重违反对世义务的手段，并力图将其纳入草案之中。于草案二读阶段，委员会曾暂时通过了第 54 条，即“受害国以外的国家采取的反措施”的相关规定。然此提议，在 ILC 及联合国大会中遭到强烈反对。最终相关条款被删除，仅以“不妨碍条款”的形式表达了对这一复杂且敏感问题的谨慎态度。¹²自此之后，国际法学界对于集体反措施的讨论逐渐降温，几乎陷入沉寂。

近年来，网络空间屡遭低于武力冲突阈值的攻击事件侵扰，“想哭”

（WannaCry）勒索软件肆虐横行、¹³格鲁吉亚政府机构网站遭遇大规模瘫痪、

⁴ 参见温树斌：《论国际法上自助措施的范围》，《法制与经济（下旬刊）》，2009 年第 8 期，第 85 页。

⁵ See Alexander Orakhelashvili, *Akehurst's Modern Introduction to International Law* 9-10 (Routledge 2018) .

⁶ 参见李永胜：《论受害国以外的国家采取反措施问题》，法律出版社 2012 年版，第 35 页。

⁷ See James Crawford, *Articles on Responsibility of States for Internationally Wrongful Acts* 2001, <https://legal.un.org/avl/ha/rsiwa/rsiwa.html>, visited on 30 November 2024.

⁸ 参见[英]劳特派特修订：《奥本海国际法》上卷，第一分册，王铁崖等译，中国大百科全书出版社出版 1995 年版，第 308 页。

⁹ 自卫权的使用受《联合国宪章》第 51 条严格规定，必须遵循必要性、比例性原则，且仅限于对武力攻击的防御性回应。See U.N. Charter, Art. 51.

¹⁰ See Jonathan I. Charney, *Third State Remedies in International Law*, 10 *Michigan Journal of International Law*, 57, 60-61 (1989).

¹¹ 第 50 条 A 款“代表受害国的反措施”，即在一般违反的情况下，第三国须应受害国的要求才能代表受害国采取反措施；第 50 条 B 款“严重违反对整个国际社会的义务时的反措施”，在严重违反情况下，即使不存在受害国，第三国也有权采取或者协助采取反措施。See *Yearbook of the ILC, A/CN.4/SER.A/2000/Add.1 (Part 2)/Rev.1*, p. 58-59.

¹² 第 54 条“不妨碍条款”：“本章不妨碍第 48 条第 1 款有权援引另一国责任的任何国家，对该另一国采取合法措施（lawful measures）”。See United Nations International Law Commission. *Draft Articles on Responsibility of States for Internationally Wrongful Acts*, 2001, Art. 54.

¹³ 2017 年“想哭”勒索软件事件波及 150 多个国家和地区，数百万台计算机受到感染，医院、银行、工厂

¹⁴美国遭受复杂网络渗透攻击之侵袭，¹⁵诸多案例纷呈迭出，无不彰显出网络空间的脆弱性以及国际不法网络行为的严峻危害。这些行为不仅严重违反了国际人权法、国际贸易法、知识产权法等相关法律规范，还公然践踏了国家主权与不干涉原则，动摇国际安全与稳定之根柢。在此背景下，集体反措施作为一种可能应对国际不法网络行为的手段，成为国际社会关注的焦点。爱沙尼亚、波兰、爱尔兰、哥斯达黎加、法国、意大利、加拿大、新西兰及丹麦等国，纷纷在立场文件阐发己见。昔日沉寂之议题，因网络空间之新挑战而再度兴起，国际法学界对此亦展开了新一轮之探讨与研究。¹⁶

本文以各国在其网络空间国际法的立场文件中提及的集体反措施议题为关键参考点，分析集体反措施在一般国际法中是否具有充分的合法性依据？若其在一般国际法中缺乏合法性，网络空间的特性是否赋予了集体反措施新的合法性支撑？通过此项研究，以期为中国在网络空间国际造法进程中，就相关议题提供坚实有力的理论支撑与清晰明确的立场表述，为推动网络空间的法治化建设贡献中国智慧与中国方案。

二、关于集体反措施在网络空间适用的三种立场观点

在 2007 年，爱沙尼亚遭受大规模分布式拒绝服务攻击（DDoS），攻击源头与责任方难以确定，传统安全机制应对网络威胁的局限性暴露无遗。爱沙尼亚作为北大西洋公约组织（NATO）的成员国，虽然得到了 NATO 的援助，但《北大西洋公约》第五条在面对未达到传统武力攻击级别的网络攻击时，其适用性存在显著不足。¹⁷鉴于此，爱沙尼亚 2019 年提出，集体反措施可适用于网络空间，作为一种潜在的、保护国家网络主权和安全的必要手段，并主张未受害国家有权“采取反措施以支持直接遭受恶意网络行为侵害的国家”。¹⁸

等重要机构运营受阻。See Patrick Fowler, Anthony King, WannaCry: What to Know and What to Do, 2017, available at <https://www.jdsupra.com/legalnews/wannacry-what-to-know-and-what-to-do-29653/>, visited on 30 November 2024.

¹⁴ 2019 年格鲁吉亚遭遇大规模 DDoS 攻击，导致约 2000 个网站和电视台服务器瘫痪，包括总统府、法院和教育部等政府机构网站。See Emerging Europe Staff, Georgia hit by massive cyber attack, 2019, available at <https://emerging-europe.com/news/georgia-hit-by-massive-cyber-attack/>, visited on 30 November 2024.

¹⁵ 2020 年，美国政府和多家私营公司更是遭到俄罗斯支持的黑客组织发起的复杂且持久的网络渗透行动。See Bill Chappell, Greg Myre, Laurel Wamsley, What We Know About Russia's Alleged Hack Of The U.S. Government And Tech Companies, available at <https://www.npr.org/2020/12/15/946776718/u-s-scrambles-to-understand-major-computer-hack-but-says-little>, visited on 30 November 2024.

¹⁶ See e.g. Kosseff, Jeff, Collective Countermeasures in Cyberspace, 10 Notre Dame Journal of International & Comparative Law, Article 4 (2020).; S. Haataja, Cyber Operations and Collective Countermeasures under International Law, 25 Journal of Conflict & Security Law, 33–51(2020); N. Schmitt and Sean Watts, Collective Cyber Countermeasures?, 12 Harvard National Security Journal 373 (2021); Jackson, Miles, and Federica I. Paddeu. The Countermeasures of Others: When Can States Collaborate in the Taking of Countermeasures?, 118 American Journal of International Law ,231–274 (2024), etc.

¹⁷ 《北大西洋公约》第五条具体规定了集体防御的条款，主要涉及成员国在遭受武装攻击时的反应，并未涵盖低于适用武力程度的集体行动。See North Atlantic Treaty, April 4, 1949, 63 Stat. 2241, 34 U.N.T.S. 243, Art. 5.

¹⁸ 爱沙尼亚总统科里斯基·卡柳莱德在 2019 年 5 月的演讲中提出集体反措施在网络空间的适用。See Kersti Kaljulaid, President of the Republic at the opening of Opening at CyCon 2019, Speech in Tallinn on 29 May 2019.

此后，多国通过发布立场文件、国家领导人在国际会议上的发言以及多种官方声明形式，积极阐述对在网络空间适用集体反措施之看法，立场分化出支持、反对与保守三派。虽然各国在论及该议题时措辞各异，但深入剖析其语境与内涵，其讨论核心均聚焦于一般国际法中集体反措施在网络空间的适用问题，而非旨在提出或创设一项全新的国际法制度。

（一）支持观点：有利于维护网络空间共同利益

爱沙尼亚、哥斯达黎加、波兰及爱尔兰等国，对网络空间适用集体反措施表达了赞同立场。（见表 1）各国的具体立场和依据存有差异，但它们的核心理念都在强调集体反措施对于维护网络空间共同利益方面的重要作用。

国家	年份	关键论点
爱沙尼亚	2019	未受到直接伤害的国家可以采取反措施来支持直接受到恶意网络行动影响的国家。 ¹⁹
	2021	受害国有权采取报复、反制措施，支持单独或集体行动。 ²⁰
哥斯达黎加	2021	受害国及第三国可针对违反对世义务的行为或应受害国的请求采取反措施。 ²¹
波兰	2022	承认国家可追求普遍利益（ <i>general interest</i> ）采取反措施，特别是针对违反强行法（ <i>Jus Cogens</i> ）的行为。 ²²
爱尔兰	2023	在有限情况下，特别是违反强行法时，允许第三方或集体反措施。 ²³

表1 支持集体反措施在网络空间适用的国家立场观点展示

哥斯达黎加、波兰及爱尔兰等国注重对《国家责任条款草案》（以下简称《草案》）的解读。哥斯达黎加明确援引《草案》第 48 条和第 54 条的具体条款，主张当一国违反对世义务时，其他国家可应受害国之请求采取反措施；²⁴

¹⁹ See Kersti Kaljulaid, President of the Republic at the opening of Opening at CyCon 2019, Speech in Tallinn on 29 May 2019.
²⁰ See UNODA, A/76/136, p. 28.
²¹ See Costa Rica’s Position On The Application Of International Law In Cyberspace, 2023. p. 5, available at [https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-__\(2021\)/Costa_Rica_-_Position_Paper_-_International_Law_in_Cyberspace.pdf](https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-__(2021)/Costa_Rica_-_Position_Paper_-_International_Law_in_Cyberspace.pdf), visited on 30 November 2024.
²² See Poland, the application of international law in cyberspace, 2022, p.8, available at <https://www.gov.pl/web/diplomacy/the-republic-of-polands-position-on-the-application-of-international-law-in-cyberspace>, visited on 30 Nov. 2024.
²³ See Ireland, Position paper on the application of international law in cyberspace, 2023, p.6, available at <https://dfa.ie/media/dfa/ourrolepolicies/internationallaw/Ireland---National-Position-Paper.pdf>, visited on 30 Nov. 2024.
²⁴ See Costa Rica’s Position On The Application Of International Law In Cyberspace, 2023. p. 5.

波兰指出，《草案》已逐渐被国际习惯法所接受，并强调在此框架内，集体反措施的合法性应获承认；²⁵爱尔兰则基于国家实践提出，自《草案》2001 年通过以来，集体反措施在特定情况下，已获实践证明其可行性。²⁶

其中，波兰和爱尔兰在立场文件中对强行法规范的强调，彰显了他们对国际法更根本性和普遍性规范的重视。因为强行法与对世义务在价值上具有一致性，共同反映了国际社会的基本价值，²⁷但对世义务往往是强行法所确立的义务，并非所有强行法都直接涉及对世义务。²⁸

爱沙尼亚则独辟蹊径，将自卫权与反措施进行类比，提出一新颖主张。“如果网络行动是不友好的或违反国际法义务的，受害国有权采取报复、反制等措施，或者在发生武装袭击的情况下，有权进行自卫。这些措施可以是单独的，也可以是集体的。”²⁹这一表述不仅将反措施与自卫权置于同一讨论框架内，而且其“单独或集体”的表述也与《联合国宪章》第 51 条中的集体自卫权表达颇具相似性。³⁰

尤为值得关注的是，美国西点军校的权威学者迈克尔·施密特（Michael N. Schmitt）等兼采上述两种观点，依据《国家责任条款草案》所蕴含之集体主义（collectivist）精神，及《联合国宪章》第 51 条所载之集体自卫权，为集体反措施的合法性提供有力支撑。³¹彼等强调，网络空间所蕴含的集体主义精神已广获国际社会认可，并受国际法庇护。³²

（二）反对观点：不利于国际法的稳定性和合法性

法国、意大利及加拿大等国家对集体反措施在网络空间的适用持反对态度（见表 2）。集体反措施的法律依据尚未得到明确确立，引发对国际法稳定性和法律明确性的担忧。

国家	年份	关键论点
----	----	------

²⁵ See Poland, the application of international law in cyberspace, 2022, p.8.
²⁶ See Ireland, Position paper on the application of international law in cyberspace, 2023, p.6.
²⁷ 参见何志鹏，周萌：《强行法与对世义务间关系的动态演进机理探究》，《南大法学》2024 年第 3 期，第 120 页。
²⁸ 参见薛捍勤：《国家责任与“对国际社会整体的义务”》，《中国国际法年刊（2004）》，第 28 页。
²⁹ See UNODA, A/76/136, p. 28.
³⁰ See U.N. Charter, Art. 51.
³¹ See N. Schmitt and Sean Watts, Collective Cyber Countermeasures?, 12 Harvard National Security Journal 378 (2021).
³² 文中主张，代表和支持受害国采取集体反措施是合法的，且应限定在针对违反集体义务行为的第三方反措施范围内。See N. Schmitt and Sean Watts, Collective Cyber Countermeasures?, 12 Harvard National Security Journal 410 (2021).

法国	2019	习惯国际法未授权第三国对责任国采取反措施， ³³ 反措施应由受害国实施。 ³⁴
	2021	重申反措施必须由法国作为受害者采取，排除集体反措施的可能性。 ³⁵
	2023	强调仅受害国有权采取反措施。 ³⁶
意大利	2021	不支持集体反措施，但主张加强信息共享合作。 ³⁷
加拿大	2022	明确指出尚未看到足够的支持集体网络反措施的证据，并强调需要区分“集体网络反措施”与“集体自卫”行动。 ³⁸

表 2 反对集体反措施在网络空间适用的国家立场观点展示

法国基于双边主义理论，³⁹坚决主张反措施应仅由直接受害国实施，第三国无权介入。为佐证此观点，法国于其《军事作战法手册》中特别引用了《国家责任条款草案》第 49 条的规定：“一受害国只在为促使一国际不法行为的责任国依第二部分履行其义务时，才可对该国采取反措施。”⁴⁰同时，法国还援引了国际法院在尼加拉瓜案中的判决，⁴¹强调第三方国家不应因其对国际法义务的理解或政治立场而干预他国事务。⁴²

加拿大的反对意见在于自卫权与反措施的概念不可混同。其强调，必须清晰区分“集体网络反措施”和“集体自卫”行动。⁴³具体而言，“集体自卫”依据《联合国宪章》第 51 条的自卫权概念，针对直接的武装攻击，通常采取军事手段回

³³ See Droit International Appliqué aux Opérations dans le Cyberspace (International Law Applicable to Operations in Cyberspace), 2019, p.39, available at <https://www.justsecurity.org/wp-content/uploads/2019/09/droit-internat-appliqu%C3%A9-aux-op%C3%A9rations-cyberspace-france.pdf>, visited on 30 November 2024.

³⁴ See Droit International Appliqué aux Opérations dans le Cyberspace (International Law Applicable to Operations in Cyberspace), 2019, p.126.

³⁵ See International Law Applied To Operations IN Cyberspace, Paper shared by France with the Open-ended working group established by resolution 75/240, 2021, p.4, available at <https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf>, visited on 30 November 2024.

³⁶ 法国强调，其在北约和欧盟内部的网络防御合作并不等同于可以无视国际法原则、采取未经授权的集体反措施。See France, Ministère des Armées, Manuel de droit des opérations militaires, 2023, p. 304, available at <https://www.defense.gouv.fr/actualites/droit-operations-militaires-manuel-inedit-au-service-armees-francaises>, visited on 30 November 2024.

³⁷ 意大利仅阐述了其立场，并未提供具体的法律依据或支撑性论据。See International Law and Cyberspace, Report of the Study Group Co-organised by the University of Bologna, University of Milan and University of Westminster, February 2021, p. 126, https://www.esteri.it/wp-content/uploads/2021/12/UNIBO_Applicazione-dei-principi-della-Carta-delle-Nazioni-Unite-nello-spazio-cibernetico.pdf, visited on 30 November 2024.

³⁸ See Canada, International Law applicable in cyberspace, 2022, para. 37, available at https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_scurite/cyberspace_law-cyberespace_droit.aspx?lang=eng#fn24, visited on 30 November 2024.

³⁹ See Yearbook of the International Law Commission, 1985, Vol. I, p.310.

⁴⁰ See United Nations International Law Commission. Draft Articles on Responsibility of States for Internationally Wrongful Acts, 2001, Art. 49.

⁴¹ 该判决指出“第三国美国采取的反措施未能证明其合理性”。See Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, I. C. J. Reports, 1986, para. 249.

⁴² See France, Ministère des Armées, Manuel de droit des opérations militaires, 2023, p. 304.

⁴³ See Canada, International Law applicable in cyberspace, 2022, para. 37.

应；而“集体网络反措施”则依据国际法中的责任条款，是对违反国际义务行为的反制手段，至今尚未有足够的国家实践或法律意见以支撑集体网络反措施的合法性。⁴⁴两者在法理基础、触发条件及所采取的措施性质上均截然不同。相较之下，意大利的反对立场较为温和，其并不支持采取集体反措施，而主张通过加强信息共享与合作以应对网络空间的挑战。⁴⁵

（三）保守观点：有限支持与谨慎探索

鉴于网络空间安全议题之复杂性和敏感性，新西兰和丹麦等国在集体反措施的适用问题上采取了较为保守的立场。（见表 3）这些国家并未完全否定集体反措施的可能性，而是出于对集体利益（collective interest）⁴⁶及对整个国际社会承担的义务⁴⁷的考量，倾向于在有限的、特定的情况下，考虑允许第三国协助受害国采取反措施。

国家	年份	关键论点
新西兰	2020	对受害国在有限情况下请求他国协助（assistance）采取反措施。 ⁴⁸
丹麦	2023	集体反措施问题尚未得到充分解决，需谨慎考虑；在某些情况下，可能需要请求其他国家协助采取适当和必要反措施作为集体反措施。 ⁴⁹

表3 集体反措施在网络空间适用的保守国家立场观点展示

值得注意的是，第三国的协助行为仍属于双边关系范畴，与集体反措施中的代表行为存在着明确的界限。《塔林手册 2.0 版》明确指出，在实践中，尽管协助（援助）行为与代表行为之界分可能在某些情况下显得较为模糊，但在判断第三国行为的合法性时，核心考量因素应在于该行为是否违背了国际法对反措施对象国所承担的义务。⁵⁰因此，任何形式的协助或反措施均不得超越国际法所规定之界限，必须尊重反措施对象国的主权和合法权益，避免对其造成不必要的损害或侵犯。

⁴⁴ See Canada, International Law applicable in cyberspace, 2022, available at https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/cyberspace_law-cyberespace_droit.aspx?lang=eng#fn24, visited on 30 November 2024.

⁴⁵ See International Law and Cyberspace, Report of the Study Group Co-organised by the University of Bologna, University of Milan and University of Westminster, February 2021, p. 126.

⁴⁶ See New Zealand Foreign Affairs and Trade, The Application of International Law to State Activity in Cyberspace, 2020, p. 4.

⁴⁷ See Kjelgaard and Melgaard, Denmark’s Position Paper, 2023, p. 454.

⁴⁸ See New Zealand Foreign Affairs and Trade, The Application of International Law to State Activity in Cyberspace, 2020, p. 4, available at <https://www.mfat.govt.nz/en/media-and-resources/the-application-of-international-law-to-state-activity-in-cyberspace>, visited on 30 November 2024.

⁴⁹ See Kjelgaard and Melgaard, Denmark’s Position Paper, 2023, p. 454, available at https://brill.com/view/journals/nord/92/3/article-p446_007.xml, visited on 30 November 2024.

⁵⁰ See Schmitt, M. N., & Vihul, L., Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations, Rule 24, Commentary 9 (Cambridge University Press. 2017).

综上所述，各国于网络空间适用集体反措施之议题上立场纷呈，形成了支持、反对与保守三种主要观点。争议的核心主要聚焦于两个关键方面：其一，集体反措施在一般国际法中是否具有坚实的法律基础，特别是是否已经成习惯国际法规则；其二，非直接受害国提出反措施请求是否具有正当性及法律依据。

三、一般国际法中集体反措施的合法性依据缺失

集体反措施在一般国际法中的合法性依据问题，需从三个方面进行探讨。先，需考量集体反措施本身在国际法中的法律证据是否充分，这是其合法性基础的首要前提。其次，从第三国请求的视角出发，验证非直接受害国在特定情境下采取反措施的权力来源及法律依据。最后，集体反措施中的代表行为，涉及代理与权利转让的合法性问题，是判断其能否在一般国际法中找到法律基础的又一关键要素。

（一）集体反措施的国际法依据不足

依据《国际法院规约》第 38 条之阐释，⁵¹条约法、习惯国际法与司法判例是评判集体反措施合法性的关键依据。

1. 国际条约体系中集体反措施规定缺失

当前，国际条约体系对于集体反措施的规定相当匮乏，既未明确定义其内涵，亦未界定其适用范围、实施条件及法律后果。其中，需特别指出的是，

《联合国宪章》第七章并不能为集体反措施之实施提供合法性依据。第七章所述之强制措施，与集体反措施实有本质区别。宪章赋予联合国安全理事会采取强制措施的权力，此等措施以国际法的普遍性和强制性为基础，且须遵循严格的法律程序。⁵²譬如，1990 年针对伊拉克实施的经济制裁与贸易禁运，皆是在《联合国宪章》框架下，⁵³经正式程序批准而执行，其合法性与正当性由此广受认可。然而，集体反措施作为一种自助行为，其性质更接近于“私力司法”。⁵⁴因而，国际条约体系中关于集体反措施的规范，尚处于空白之境地。

2. 集体反措施尚未确立为习惯国际法之规则

习惯国际法之形成，赖于广泛的国家实践（state practice）和法律确信（*opinio juris*）。⁵⁵

（1）国家实践之地域局限与形式单一

⁵¹ See Statute of the International Court of Justice, Art. 38.

⁵² See U.N. Charter, Chapter VII: Action with Respect to Threats to the Peace, Breaches of the Peace, and Acts of Aggression, available at <https://www.un.org/en/about-us/un-charter/chapter-7>, visited on 30 November 2024.

⁵³ See UN Security Council, Resolution 661, 6 August 1990, p. 23.

⁵⁴ 参见霍政欣：《〈反外国制裁法〉的国际法意涵》，《比较法研究》2021 年第 4 期，第 150 页。

⁵⁵ See UN, GA Resolution 73/203, 20 December 2018, p. 2-3.

集体反措施的国家实践在地域分布和实施主体上皆显局限。虽然历史中不乏可视为集体反措施雏形的案例，但这些实践多局限于特定国家或区域，且多以单边行动的形式加以呈现。譬如，美国在历史上多次采取此类措施，包括1977年对南非种族隔离制度的回应性制裁，1978年因乌干达政府涉嫌种族灭绝而实施的贸易制裁，以及1980年因苏联入侵阿富汗而采取的经济禁运等。类似地，其他集体措施如1980年欧洲国家对利比里亚的制裁，以及1998年欧共体对南联盟的制裁等。联合国立法系列材料对此亦有深刻评述，认为集体反措施的实施尚属有限，且仍处于萌芽阶段。⁵⁶这些行动往往限于特定区域，且多由少数发达国家推动，⁵⁷远未形成普遍现象。

更为关键的是，实施此类行为的国家通常更倾向于将其冠以“制裁”或“限制性措施”之名，而非“反措施”。⁵⁸这一命名偏好从侧面揭示了制裁国并未借助集体反措施的法律框架来为其行动赋予合法性。因此，这些措施非但不能被视为创造或推动集体反措施形成的实例，⁵⁹反而更多地凸显了国家在实践中的单边主义倾向和局限性。

(2) 法律确信的支持与反对声音交织

集体反措施的合法性与正当性，在国际社会中尚未获致普遍认同。尤以《国家责任条款草案》为著，其常被视作习惯国际法规则之反映或体现。各国对《草案》第54条的二读反馈，分歧赫然。一些国家，如韩国、法国、荷兰和西班牙，支持在特定情况下允许集体反措施，且强调须严加限制，以防滥觞。而另一些国家，如中国、奥地利、墨西哥和日本，则明确反对，担忧集体反措施可能冲破国际法框架，违背基本国际法原则。⁶⁰终而，集体反措施条款未获草案二读通过，此结果既昭示国际社会对此问题的复杂心态与纷纭立场，亦凸显其于国际法中之地位不确定、含义模糊之现状。

习惯国际法的形成需广泛的国家实践和坚定的法律确信，二者缺一不可。⁶¹然集体反措施既显实践局限，又存立场争议，其在习惯国际法中并未确立确定地位。

⁵⁶ See United Nations, *Materials on the Responsibility of States for Internationally Wrongful Acts*, 2nd edition, 2023, p. 527, para. 3.

⁵⁷ See PORTELA, Clara, *Where and why does the EU impose sanctions?*, 17 *Politique Européenne* 83-111 (2005).

⁵⁸ See Ruys, T., *Immunity, Inviolability and Countermeasures – A Closer Look at Non-UN Targeted Sanctions*, in Ruys, T., Angelet, N. and Ferro, L. (eds), *The Cambridge Handbook of Immunities and International Law* 704 (Cambridge University Press 2019).

⁵⁹ 参见李庆明：《论外国中央银行资产的国家豁免与其他国家的没收和集体反措施》，《环球法律评论》2024年第5期，第191页、第202页；See Yamada, T., *Sanctions on Russia: What Impact Do They Have on the Question of “Third-Party Countermeasures”?*, in Furuya, S., Takemura, H. & Ozaki, K. (eds), *Global Impact of the Ukraine Conflict: Perspectives from International Law* 113 (Springer 2023).;

⁶⁰ See UN General Assembly, International Law Commission, *Comments and Observations Received from Governments*, Fifty-third Session, Geneva, Switzerland, 23 April-1 June and 2 July-10 August 2001, A/CN.4/515, 19 March 2001.

⁶¹ See UN, GA Resolution 73/203, 20 December 2018, p. 2.

3. 国际法院司法实践对集体反措施鲜有触及

集体反措施，作为一种尚未被明确规范的法律手段，在国际法院的司法实践中鲜有涉猎。此现象之缘由，部分可从国际法院审理案件的规定和实践中窥见一斑。

首先须明确的是，国际法院作为国际争端解决的核心机构，其审理案件的基础在于当事国之同意。依据《国际法院规约》第 36 条之规定，“法院之管辖包括各当事国提交之一切案件，及联合国宪章或现行条约及协约中所特定之一切事件。”⁶²同理，根据《国际法院规约》第 65 条，“本法院对于任何法律问题如经任何团体由联合国宪章授权而请求或依照联合国宪章而请求时，得发表咨询意见。”⁶³因此，国际法院通常不会审理那些未经争议方、联合国或其他授权机构明确提出，或超出规定范围的问题。在实践中，虽然诸多案件触及了对世义务的内容，如巴塞罗那电车案、西南非洲案和冈比亚诉缅甸危害人类罪案等，但集体反措施的合法性与适用性，却始终未曾成为法院审理的议题。个中原因，若集体反措施非案件争议焦点，法院通常不会主动介入，亦不会轻易发表意见。

值得一提的是，1986 年的尼加拉瓜案中，国际法院明确否定了“武力”作为反措施手段的合法性，⁶⁴且未就集体反措施的其他方面展开深入探讨。部分参与《塔林手册 2.0 版》编纂的专家据此认为，代表他国采取反措施之行为实属违法。⁶⁵这一判例虽未直接涉及集体反措施的全部方面，但其在一定程度上反映了国际法院对反措施手段的审慎态度。

（二）非直接受害国采取反措施缺乏法律依据

在涉及违反对世义务或强行法的情形下，非直接受害国是否可以介入实施反措施，成为另外一个重要的分歧争议点。

1. 传统国际法框架中反措施的受害国限定

在传统国际法框架中，反措施的行使权利被严格限定于直接受害国。非直接受害国因其实质利益未受直接损害，因此被普遍认为不具备采取反措施的权利。⁶⁶这一观点在《国家责任条款草案》第 42 条中得到了明确体现，规定只有那些由于他国不法行为而受到“特别影响”或“根本性改变”其履行义务立场的国家，方有权在国际法框架内实施反措施。⁶⁷此规定强调了反措施的主要目的，即恢复或保

⁶² See Statute of the International Court of Justice, Art. 36.

⁶³ See Statute of the International Court of Justice, Art. 65.

⁶⁴ See *Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.)*, Judgment, I. C. J. Reports, 1986, paras. 249, 252.

⁶⁵ 参见[美]迈克尔·施密特：《网络行动国际法塔林手册 2.0 版》，黄志雄等译，社会科学文献出版社 2017 年版，第 162 页。

⁶⁶ See Yearbook of the International Law Commission, 1980, Vol. II, p.114, para.40.

⁶⁷ See United Nations International Law Commission. Draft Articles on Responsibility of States for Internationally Wrongful Acts, 2001, Art. 42.

护直接受害国的权利。因此，即使非直接受害国在道义上对他国的不法行为有所关切，国际法也并未赋予其采取反措施的合法权利。

此外，《国家责任条款草案》第 54 条虽然为第三国采取“合法措施”提供了法律框架，但并未明确界定“合法措施”与传统反措施在具体实施过程中的差异。⁶⁸事实上，“合法措施”并非等同于反措施。2001 年当选为联合国国际法委员会委员的薛捍勤指出，国际法委员会未采用“反措施”术语，正是基于只有受害国才能合法行使反措施的原则，非直接受害国不应有权采取反措施。⁶⁹因此，尽管《草案》提到“合法措施”，但并未因此赋予非直接受害国采取反措施的合法性。反措施作为回应手段，其合法性必须满足严格条件，而这一点在“合法措施”框架下并未得到同等保障。

2. 《国家责任条款草案》第48条为扩展受害国范围提供有限依据

《国家责任条款草案》第 48 条被哥斯达黎加等援引为集体反措的法律依据。然而，该条款为非直接受害国提供的法律基础相当有限。第 48 条明文规定，非直接受害国可在两种特定情况下主张责任追究：“（a）被违背的义务是对包括该国在内的一国家集团承担的、为保护该集团的集体利益而确立的义务；或（b）被违背的义务是对整个国际社会承担的义务。”⁷⁰这一规定与《草案》二读时的第 42 条内容基本一致。国际法委员会对此作出阐释，指出所涉及的国际义务不仅限于双边关系，还可能针对特定国家集团或整个国际社会。⁷¹这意味着，非直接受害国只有在国际义务具有这种更广泛的范围时，才能依据该义务提出权利主张。比利时诉塞内加尔案中，国际法院确认了非直接受害国在特定情况下的诉权，⁷²但这并不等同于对反措施的认可或支持。

在 21 世纪初，国际法委员会尝试将集体反措施作为应对严重违反对世义务行为的法律手段纳入《国家责任条款草案》。然而，这一尝试遭遇了诸多争议未能获得广泛支持。最终，集体反措施未被纳入《草案》，这表明即使是对世义务这一重要概念，也未能能为集体反措施提供充分的合法性依据。

⁶⁸ 《国家责任条款草案》第 54 条：“本章不妨碍依第 48 条第 1 款有权援引另一国责任的任何国家，对该另一国采取合法措施以确保停止该违背义务行为和使受害国和被违背之该义务的受益人得到赔偿。”See United Nations International Law Commission. Draft Articles on Responsibility of States for Internationally Wrongful Acts, 2001, Art. 54.

⁶⁹ 参见薛捍勤：《国家责任与“对国际社会整体的义务”》，《中国国际法年刊（2004）》，第 33 页。

⁷⁰ United Nations International Law Commission. Draft Articles on Responsibility of States for Internationally Wrongful Acts, 2001, Art. 48.

⁷¹ See James Crawford, The International Law Commission's Articles on State Responsibility: Introduction, Text and Commentaries 257 (Cambridge University Press 2002).

⁷² See Questions relating to the Obligation to Prosecute or Extradite (Belgium v. Senegal), Judgment, I.C.J. Reports 2012, paras. 65-68.

（三）集体反措施中代理与权利转让缺乏法律基础

集体反措施通常涉及多个国际主体之间的合作与协调，而代理机制与权力转让则可能成为促成这一合作的法律依据。⁷³

1. 代理机制之类比困境

正如安吉洛·塞雷尼（Angelo Sereni）在 1940 年的研究中指出，国际社会通常并不反对在国际关系中诉诸代理，但国际法管理机构并未为此提供明确规则。⁷⁴代理机制在国际关系中存在一定的实践基础，由此为集体反措施提供了一种可能的法律框架，允许受害国授权其他国家代为执行反措施。此类机制源于国内法中的委托制度，也在国际法实践中尤其在外交保护和经济关系中所体现。罗马法和民法早期便对委任关系进行了规定，允许国家授权他国代表其行事。⁷⁵然而，国内法代理机制并不能完全适用于国际情形。国内法中，代理人的行为被视为委托人的行为。⁷⁶这意味着，国际上合作国的行动应被视为受害国的行为，且代理行为必须得到目标国（即被反措施国）的承认，⁷⁷在实际的国际局势中，目标国通常并未被告知或同意代理国采取反措施。由于缺乏这一必要的同意，代理机制在集体反措施中的适用性和合法性受到严重限制。

2. 反措施权利转让的内在矛盾

另一个问题是，国际法是否允许受害国将其反措施的权利转让或授予给合作国。⁷⁸反措施权利的转让意味着合作国将基于受害国的授权，对目标国采取可能的反措施。这一行为可能引发复杂的法律问题，尤其是合作国在实施反措施时，可能违背其对目标国原本应承担的国际义务。国际法对反措施的合法性和正当性有着严格的要求。根据《塔林手册 2.0 版》第 24 条第 9 条评论，任何形式的反措施，无论是援助形式还是代表形式，均不得超越国际法所规定的义务，且必须尊重反措施对象国的主权和合法权益。⁷⁹反措施的权利转让可能使合作国的行动超出国际法允许的范围，进而违反国际法的基本原则，如国家主权平等、不干涉内政以及和平解决国际争端等。

⁷³ See Jackson M, Paddeu FI. The Countermeasures of Others: When Can States Collaborate in the Taking of Countermeasures?, 118 American Journal of International Law 266 (2024).

⁷⁴ See Sereni, Angelo Piero. Agency in International Law. 34 The American Journal of International Law, 644(1940).

⁷⁵ See Q. Wright, Mandates and the League of Nations, 1930, p. 375

⁷⁶ 参见《中华人民共和国民法典》第 161 条。

⁷⁷ See Sereni, Angelo Piero. Agency in International Law. 34 The American Journal of International Law, 649(1940).

⁷⁸ 国际法院解释说，葡萄牙被授权“代表马拉塔统治者”在其位于印度领土上的飞地行使军事和警察职能，它只拥有马拉塔人“授予的权力”。See Right of Passage over Indian Territory (Port. v. India), Merits, 1960 ICJ Rep. 6, p. 38–39.

⁷⁹ See Schmitt, M. N., & Vihul, L., Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations, Rule 24, Commentary 9 (Cambridge University Press. 2017).

四、网络空间中集体反措施对一般国际法框架的应然遵循

集体反措施在一般国际法中的合法性依据缺失，这一现状在网络空间这一领域中并未因新情境的出现而有所变化。尽管网络空间具有其独特的属性和挑战，但一般国际法依然适用。因此，集体反措施在网络空间中并未脱离一般国际法框架，其依然不具有合法性。

（一）网络空间特性未改变集体反措施的合法性依据

网络空间的隐匿性、跨国界性以及高度技术化⁸⁰使其成为一般国际法框架的一个挑战，尤其是在应对网络威胁时，现有法律体系显得力不从心。尽管网络空间的特殊性推动了国际法在某些领域的适应性调整，例如拓展国家行为规范和更新执行措施，但这些调整始终未突破国际法的核心结构和基本原则。正如《塔林手册 2.0 版》规则 4、规则 66、规则 69 等所指出，⁸¹即使是网络攻击，也必须遵循不使用武力和国家主权等核心国际法原则。

联合国信息安全政府专家组（UNGGE）和开放式工作组（OEWG）多年来的报告已明确指出，应继续适用一般国际法。网络空间的反措施在实施时，仍必须遵循《国家责任条款草案》第 49 条规定的目的和限制，⁸²遵守第 51 条相称性原则。⁸³也就是说，任何超出这些法定限制的反措施，都将被视为非法。

因此，尽管网络空间呈现出前所未有的复杂性和挑战，集体反措施在网络空间中依然不能脱离一般国际法的框架。网络空间的特性并未为集体反措施的合法性提供新的法律依据，也未得到国际社会的普遍认可。网络空间中的集体反措施仍然受到现有国际法框架的严格约束。

（二）网络空间反措施的受害国限定

网络空间中的集体反措施并未突破一般国际法框架。《塔林手册 2.0 版》第 24 条重申，只有受害国才有权采取反措施。⁸⁴集体反措施在网络空间的提出，并未充分考量其合法性问题。尽管其在某些方面展现出优势，如促进国际合作⁸⁵、构建网络空间威慑体系⁸⁶以及完善国际救济机制等，但在现有国际法框架下并不具有合法性。

⁸⁰ See Polly M. Holdorf, Prospects for International Cybersecurity Regime, (2015) U.s Air Force Institute for National Security Studies, Strategic Paper 10.

⁸¹ 参见[美]迈克尔·施密特：《网络行动国际法塔林手册 2.0 版》，黄志雄等译，社会科学文献出版社 2017 年版，第 62 页，第 319 页，第 335 页。

⁸² 参见[美]迈克尔·施密特：《网络行动国际法塔林手册 2.0 版》，黄志雄等译，社会科学文献出版社 2017 年版，第 149-158 页。

⁸³ 参见朱磊：《论国际法上的反措施在网络空间的适用》，《武大国际法评论》2019 年第 3 期，第 136 页。

⁸⁴ See Schmitt, M. N., & Vihul, L., Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations, Rule 24 (Cambridge University Press. 2017).

⁸⁵ See Ireland, Position paper on the application of international law in cyberspace, 2023, para. 26, available at <https://dfa.ie/media/dfa/ourrolepolicies/internationallaw/Ireland---National-Position-Paper.pdf>, visited on 15 Sep. 2024.

⁸⁶ 美国海军学院学者杰夫·考斯夫提出了集体反措施在网络空间适用的五点优势：一是可以解决网络空间威胁的高度互联性问题；二是可以充分利用网络能力更先进国家的比较优势；三是能够更好地解决其在网

罗伯特·阿戈（Roberto Ago）指出，国际犯罪的法律后果不应被视作赋予单个国家单方面行动权限的普遍授权，而应被视作需国际社会协同应对的挑战。⁸⁷在网络空间中，诸如反对网络恐怖主义等涉及对世义务的，并不能自动转化为赋予单个国家或集体采取反措施的合法性依据。鉴于此，对于《国家对国际不法行为的责任条款草案》等相关条款的解读应保持谨慎，避免扩大化或宽松化倾向。

五、中国因应

中国对于单边措施行为的立场经历了从早期坚决反对到当前审慎评估的转变，这种转变为集体反措施在网络空间适用的立场考量提供了重要参考。在早期阶段，中国明确反对任何形式的单边反措施，主张在国际法框架下，避免将武力、威慑和反制等手段引入网络空间，防止网络军备竞赛和冲突升级。⁸⁸随后，中国遂转而采取更为审慎的评估态度，并在 2021 年《中国关于网络空间国际规则的立场》中强调，必须谨慎处理武装冲突法和诉诸武力法（*jus ad bellum*）的网络空间适用问题，避免网络空间成为新的冲突领域。⁸⁹因此，对于集体反措施在网络空间的适用，中国应持谨慎态度，既非全盘否定，亦非轻易接受。

（一）引领负责任框架下的集体反措施适用议题对话

鉴于部分国家已展开相关议题讨论，中国应主动担当，以引导讨论向更为理性、务实的路径发展。在引领负责任框架下的集体反措施适用议题对话中，中国应坚持“趋利避害”的战略，推动符合自身利益的政策主张，并防止西方国家推动下有关规范继续走向不平衡。⁹⁰关于集体防范措施在网络空间适用的具体讨论方向，应涵盖以下几个方面：

络空间面临威胁的持续性；四是集体反措施可能会产生威慑作用；五是集体反措施可以降低网络攻击升级的可能性。See Kosseff, Jeff, *Collective Countermeasures in Cyberspace*, 10(1) *Notre Dame Journal of International & Comparative Law*, Article 4 (2020).

⁸⁷ 罗伯特在特别报告员任期结束时所写的一段话是富有启发性的：“我不认为有人可以说‘如果有一种国际不法行为比另一种或其他行为稍微严重一点，我们就将其称为犯罪，所有国家都可以合法地进行干预’。并非所有国家，而是国际社会，都被认为可能拥有对这种特别严重的国际不法行为作出反应的权利。因此，对世义务的理念不仅与承认该共同体本身的存在这一事实有关，而且还与该共同体更加先进的国际化这一事实有关。”Roberto Ago, “Obligations Erga Omnes and the International Community”, in Weiler, Cassese, Spinedi (eds) (n 49) 238.

⁸⁸ “应当本着和平的精神研究国际法适用问题，避免将使用武力、威慑、反制等做法引入网络空间，切实预防网络军备竞赛，降低对抗冲突的风险。”参见中华人民共和国常驻联合国代表团：《中国代表团孙磊参赞在第 72 届联大一委关于信息和网络安全问题的专题发言》，2017 年，http://un.china-mission.gov.cn/chn/zgylhg/cjyjk/ldyw/201710/t20171030_8355369.htm，2024 年 11 月 30 日访问。

⁸⁹ “各国应审慎对待武装冲突法和诉诸武力法适用于网络空间问题，防止冲突升级，避免网络空间成为新的战场。”参见中华人民共和国外交部：《中国关于网络空间国际规则的立场》，2021 年，https://www.fmprc.gov.cn/web/wjb_673085/zjzg_673183/jks_674633/zclc_674645/qt_674659/202110/t20211012_9552671.shtml，2024 年 11 月 30 日访问。

⁹⁰ 参见黄志雄：《网络空间负责任国家行为规范：源起、影响和应对》，《当代法学》2019 年第 1 期，第 68 页至第 69 页。

第一，确立制度构建原则。推动制定一套全面的制度构建原则，倡导以和平解决争端为首要原则，结合必要性审查、明确性与合理性并重等要素，为集体反措施的实施设定清晰的框架与界限。

第二，界定适格主体与实施依据。明确集体反措施的合法实施主体及其实施依据的具体范围，确保制度的精确性与可操作性，避免滥用或误用。

第三，细化实质性规范与程序性要件。深入探讨国际不法行为的认定标准、反措施的衡量基准、实施机制及证据规则等核心问题。⁹¹特别关注紧急情况下程序要求的适用性问题，中国应坚持程序正义，反对过度放宽程序要求，强调集体自卫的合法性需基于受害国的明确请求，⁹²第三国的反应措施也需在受害国授权下行使，以维护国际法的权威性与稳定性。

为了有效推动集体反措施在网络空间的适用议题讨论，中国应在联合国框架下积极参与信息安全开放式工作组（OEWG）和“行动纲领”（The program of action，简称PoA）⁹³，发出“中国声音”⁹⁴。中国应引导国际社会就集体反措施的合法性、合理性和可操作性展开客观讨论，推动形成平衡的国际共识，确保这一议题的探讨符合国际法基本原则，避免不平衡或滥用现象的出现。

（二）推动全球与区域网络安全的协同建设与多维合作

为应对日益复杂的网络安全挑战，中国应致力于推动全球与区域网络安全合作机制的多维建设，特别是在提升全球与区域间的协同防御与能力建设方面。可以参考欧盟理事会于2017年6月通过推出《欧盟联合外交应对恶意网络活动框架》，又称“网络外交工具箱”（Cyber Diplomacy Toolbox），⁹⁵统一应对国际网络不法行为。中国应首先可通过现有区域性平台，如亚太计算机应急响应组织（APCERT）推动成员国在网络安全标准、证据共享、应急响应等关键议题上的协作，促进统一的技术规范和反应机制的形成，增强成员国间的防御能力和应急响应能力。

⁹¹ 萨穆利·哈塔亚等学者对网络空间集体反措施的可行性和必要条件进行了深入探讨，强调此类措施应在特殊情况下谨慎使用，且必须遵循受害国的请求、相称性原则以及作为最后手段的原则。See S. Haataja, *Cyber Operations and Collective Countermeasures under International Law*, 25 *Journal of Conflict & Security Law*, 40–49(2020).

⁹² See *Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.)*, Judgment, I. C. J. Reports, 1986, para. 199.

⁹³ 第78届联合国大会在2023年12月的最后一次会议上，通过了关于OEWG、PoA的决议。See UN, GA Resolution A/78/404, 4 December 2023, p.10-11; 2020年10月，法国主导并且联合40个国家建议，结束联合国框架下UN GGE和OEWG“双轨并进”模式，建立一项促进网络空间负责任国家行为的PoA作为联合国的常设机制。参见《2022年国际网络空间九大动向》，《中国信息安全》，2023，https://mp.weixin.qq.com/s?__biz=MzA5MzE5MDAzOA==&mid=2664175552&idx=1&sn=f289dc28459b93a8748e6ddd609d68c1&chksm=8b591539bc2e9c2f79e1ec51915fa713feb1f8cc99116196c4c7853a91db76f5a5aa0a05a89a&scene=27，2024年11月30日访问。

⁹⁴ 参见李若瀚，高娜：《国际法视野下的网络空间“负责任国家行为”框架--兼评<关于推进网络空间中负责任国家行为的联合声明>》，《中国军事科学》2019年第5期，第72页。

⁹⁵ See Council of the European Union, *Council Conclusions on a Framework for a Joint EU Diplomatic Response to Malicious Cyber Activities*, <https://data.consilium.europa.eu/doc/document/ST-10474-2017-INIT/en/pdf>, visited on 16 January 2025.

同时，积极拓展与其他区域性组织如东盟、上海合作组织等的合作，扩大合作范围，强化区域和全球网络安全防护体系的建设。在全球合作层面，中国应强化与“一带一路”国家的联动，推动跨国技术研发、应急响应和网络安全教育合作，促进技术共享和能力提升。可深化产教融合与人才培养，提升区域和全球在面对复杂威胁时的协同应对能力。与此同时，中国应积极参与联合国信息社会世界峰会成果的落实进程，推动国际共识凝聚，加强全球数据治理与基础设施构建，为构建和平、安全、开放、合作、有序的网络空间贡献力量。

六、结语

集体反措施在一般国际法中的合法性依据确属阙如，这一现状并未因网络空间这一新兴且复杂领域的出现而有所改观。网络空间的独特性并未为其增添任何额外的合法性基础，以支撑集体反措施的实施。然而，有关将其延伸至网络空间的观点仍不绝于耳。⁹⁶国家通过立场文件所表达的支持立场，既是对现有国际法规则的阐释，也可能成为新习惯国际法规则形成的契机。随着网络空间安全地位的日益凸显，支持集体反措施的声音愈发强烈，昭示着其可能正逐渐成为国际法体系中的重要组成部分。⁹⁷

集体反措施虽以国际不法行为为触发点，但实际操作中对该行为的界定常带主观性与单方面性，难以避免偏见和误判。⁹⁸因而，集体反措施的有效性往往取决于第三国的实力对比。只有当第三国在特定领域相较于违法国具有明显优势时，反措施方有可能产生实效。这种依赖强国实力的机制本身便蕴含着公正性的隐忧。对反措施的适用范围必须严格加以限制，否则可能被滥用，特别是防止一些大国借“集体利益”之名，推动“集体干涉”或“集体制裁”，从而形成不具有合法性的“集体反措施”。⁹⁹

⁹⁶ See N. Schmitt and Sean Watts, *Collective Cyber Countermeasures?*, 12 *Harvard National Security Journal* 410-411 (2021).

⁹⁷ See SPÁČIL, Jakub. *Countermeasures against Cyber Operations: Moving forward?*, 23 *International and Comparative Law Review* 100(2023).

⁹⁸ 参见李传宏：《联合国安全理事会的制裁措施初探》，《法学评论》1996年第3期，第85页。

⁹⁹ 参见贺其治：《国家责任法及案例浅析》，法律出版社2003年版，第323页。

Exploring the Illegitimacy of Collective Countermeasures in Cyberspace

Abstract: Collective countermeasures remain undefined in *the Draft Articles on the Responsibility of States for Internationally Wrongful Acts*, and have recently sparked controversy due to the reaffirmation of multiple countries' position papers on cyberspace. From the perspective of general international law, the legitimacy of collective countermeasures is fraught with deficiencies: their basis in general international law, the legal grounds for non-directly injured states to take action, and the possible mechanisms for agency and transfer of power all lack clear support. In the realm of cyberspace, collective countermeasures ought to follow the framework of general international law. The characteristics of cyberspace do not enhance their legitimacy, and countermeasures remain subject to the restriction of being initiated by injured states only. Given the increasingly heated discussions on the application of collective countermeasures in cyberspace, China should adopt a cautious and responsible attitude, lead objective dialogues, and promote collaborative construction and multidimensional cooperation in global and regional cybersecurity, jointly addressing internationally wrongful cyber actions.

Keywords: collective countermeasures; cyberspace; international law; non-directly injured states; legitimacy dispute

人工智能的跨国性风险及其法律治理

张磊 于昊昆*

摘要：人工智能的恶意使用风险、故障风险和系统性风险容易外溢，或造成跨境损害。但现有的“基于原则”“基于权利”“基于场景”以及“基于风险”的法律治理路径难以适应人工智能的国际治理现状。人工智能区域性规则的普遍性和代表性不足、法律机制供给与风险治理需求存在张力，进一步抑制了人工智能跨国性风险的有效治理。文章提出“基于关键性”的法律治理路径，以人工智能系统在涉国家安全和公共安全相关基础设施中的地位与作用，划分“关键人工智能”与“一般人工智能”并设置不同的监管要求。同时，主张赋予跨国科技企业不完全的国际法律人格，确立相关实体在人工智能全生命周期的国际法律责任，统筹利用国际法与国内法协同治理人工智能的跨国性风险。

关键词：人工智能；跨国性风险；国际法；人工智能国际治理

一、引言

伴随国际社会共同利益的拓展，国际法经历了从“共存法”（Co-existencing Law）到“合作法”（Co-operation Law）的历史转型。¹国家间最低限度的安全与生存要求得到满足后，洛克式“共存性国际社会”渐渐转向“合作性国际社会”。²国家携手合作，共同处理全球变暖、新兴技术安全等全球性风险与挑战的意愿增强，国际法的目标也从维系国家间的共存关系转向共同治理国际问题。

当前，人工智能对科技、政治、经济等多个领域产生了颠覆性影响，其赋能千行百业，产生显著积极效应的同时，亦无法忽视人工智能推理和决策中的不透明性、不可解释性与不可控性所带来的安全风险，尤其是人工智能可能产生的跨国性风险构成对国际和平与安全的严重威胁。当人工智能的恶意使用风险、故障风险以及系统性风险预期或实际造成的损害可能或已经溢出国境时，其开发、部署和应用活动便构成了国际法规制的对象与客体。

在国际社会应对人工智能风险的过程中，不同国家、多边机制以及国际组织围绕人工智能治理形成了“基于原则的治理路径”“基于权利的治理路径”与“基于风险的治理路径”等，³也有一些学者提出了“基于场景的治理路径”。⁴“基于原则

* 张磊，武汉大学国际法研究所博士研究生；于昊昆，武汉大学国际法研究所博士研究生。

¹ See Wolfgang Friedmann, *The Changing Structure of International Law*, (Stevens & Sons, 1964), p.60-62.

² 徐崇立：《国际社会的法制化：当代图景与基本趋势》，载《法制与社会发展》2009年第5期。

³ UNESCO, *Consultation Paper on AI Regulation Emerging Approaches Across the World* (2024).

⁴ 例如，有学者认为，中国人工智能立法不应该简单效仿欧盟，应该坚持对人工智能风险的场景化规制。参见丁晓东：《人工智能风险的法律规制——以欧盟〈人工智能法〉为例》，载《法律科学（西北政法大学学报）》

的治理路径”试图从原则层面优先凝聚国际共识，通过“从抽象到具体”的方式实现对人工智能的有效治理。经合组织 2019 年发布《人工智能建议书》，阐明了重要的人工智能治理原则，⁵成为全球第一份对人工智能进行负责任管理的政府间标准。

基于权利的治理路径以人权保障为核心。欧洲委员会 2024 年签署的《人工智能与人权、民主与法治框架公约》（以下简称《人工智能框架公约》）作为世界上首部人工智能国际公约，将捍卫人权、民主和法治视为首要目标。⁶从人权保障出发治理人工智能，以推动符合人类价值观的人工智能技术发展。⁷

基于风险的治理路径在一定程度上构成了国际社会有效治理人工智能的主流模式。欧盟《人工智能法》中提出了风险分级分类的要求，成为了该法案的一大鲜明特色。作为国际层面为数不多的对人工智能具有法律拘束力的监管规则，欧盟《人工智能法》产生了广泛的国际影响，成为探讨人工智能跨国性风险化解问题中难以忽视的参照样本。

基于场景的治理路径通常也称“场景化规制”，其强调将人工智能风险还原到具体的行业、整体产品和已有的法律关系中，利用已有的行业和领域性风险监管以及现有的法律关系调整相关活动。⁸一些国家对自动驾驶、深度伪造、推荐算法、人脸识别等进行分别规制，并不制定统一的监管规则，本身就是对“基于场景的治理路径”的实践。

然而，现有的人工智能治理路径仍存在诸多局限性，难以适应碎片化、阵营化的人工智能国际治理现状，且直接照搬无法张扬中国特色，有盲从“欧洲模式”之嫌。本文在梳理人工智能的主要跨国性风险和现行法律治理路径以及机制的局限性的基础上，尝试从治理路径、调整主体、权利与义务等方面探索人工智能跨国性风险的法律治理模式，以期为人工智能的国际规则博弈提供中国智慧与中国方案。

二、人工智能的主要跨国性风险

全球化的潮流滚滚向前，国家之间的互联互通日益增加，人工智能等新兴技术的部署与应用惠及全球大多数人类，但相关风险因产品和服务的出口与海外上市、数据跨境流动等经济活动能轻易突破国境限制，产生外溢效应，构成人工智能的跨国性风险。

2025 年法国人工智能行动峰会发布的《国际人工智能安全报告》（以下简称

2024 年第 5 期，第 3 页。

⁵ OECD, *AI Principles*.

⁶ Council of Europe, *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*.

⁷ The Pillars of a Rights-Based Approach to AI Development, Tech Policy. Press, <https://www.techpolicy.press/the-pillars-of-a-rights-based-approach-to-ai-development/> (last visited 9 April 2025).

⁸ 参见丁晓东：《人工智能风险的法律规制——以欧盟〈人工智能法〉为例》，载《法律科学（西北政法大学学报）》2024 年第 5 期，第 15 页。

《安全报告》将人工智能的风险区分为恶意使用风险、故障风险以及系统性风险三类，这也是人工智能国际治理的主要对象。⁹本文将依循此分类方法，分述人工智能的跨国性风险。

（一）恶意使用风险

在《安全报告》中，“恶意使用”被界定为有意利用人工智能造成损害的行为。¹⁰在这一范畴下，可能造成跨国性损害的行为包括利用人工智能生成虚假信息对他国民众实施诈骗或操纵舆论、通过深度伪造（deep fake）技术干涉他国大选、对他国发起自动化网络攻击以及辅助生物与化学武器攻击等。

一方面，上述恶意使用行为仍处于现行国际法的规制框架内。使用者可能借由人工智能技术生成虚假信息、操纵舆论，亦可以借此进行网络侵犯与助长生物与化学攻击。¹¹从国际法角度讲，若上述行为由国家所实施，则可能构成对现行国际法规则的违反，从而构成国际不法行为并引发国家责任。¹²例如，若一国通过人工智能技术生成虚假信息以干涉一国的大选，则可能违反国际法中的不干涉原则（Principle of Non-intervention）。¹³若一国在战争中利用人工智能技术不加以区别的发动网络攻击或助长生化武器的使用，则有可能构成对国际人道法中区分原则与比例原则的违反。从这一意义上讲，人工智能技术的恶意使用风险似乎仍处于现行国际法的可规制框架内。然而，人工智能恶意使用的主体除国家外，更有可能是个人、黑客组织、网络恐怖组织等非国家行为体。这使得以规定国家权利和义务为主的现行国际法制度效力受到了削弱。国际法能否在国家间形成合意，共同打击和治理非国家行为体的人工智能恶意使用问题，或将成为未来可能重点讨论的方向之一。

另一方面，人工智能技术的恶意使用同样动摇了现行国际法规则。例如，在国际人权法中，对观点与表达自由的保护是绝对的，即使是错误信息（misinformation），在不使公共利益受损的情况下，仍然有被保护的价值。¹⁴在人工智能被恶意使用的背景下，由于缺乏“人的因素”介入，保护错误信息是否依旧属于保障人权的范畴，值得怀疑。与此同时，在人工智能深度伪造技术的影响下，生成的音频、视频、图像等信息真假难辨。即使保障言论自由，恐怕也很难实现“真理越辩越明”的效果，这也使得对错误信息的保护弊端愈加凸显，对错误言论保障的理论与哲学基础也将不复存在。可以说，人工智能恶意使用风险成为了技术使用者侵犯行为的“放大器”，其在加剧国际法规制困难程度的同时，也削弱了

⁹ International AI Safety Report

¹⁰ International AI Safety Report at 223.

¹¹ International AI Safety Report at 62-87.

¹² *Responsibility of States for Internationally Wrongful Acts*, Article 1, Article 2, Article 3.

¹³ Charter of the United Nations, Article 2 (7).

¹⁴ ICCPR Article 19; CCPR, General Comments No.34, para 47, 49; Human Rights Council (HRC), Irene Khan, Disinformation and Freedom of Opinion and Expression, Human Rights Council, U.N. Doc. A/HRC/47/25 para 38 (Apr. 13, 2021); *Salov v. Ukraine*, App. No. 65518/01, Eur. Ct. H.R. para 113 (2005).

国际法自身的有效性。

（二）故障风险

人工智能的故障风险包括可信赖性问题、人工智能偏见以及失控问题。¹⁵具体而言，人工智能幻觉、偏见与歧视、脱离人类控制等可能产生跨国性风险。

就当下人工智能技术的发展来看，人工智能仍有可能无法正确执行人类所下达的指令和任务，人工智能幻觉等系列现象影响了人工智能技术的可信赖性。¹⁶当使用者利用人工智能等技术工具辅助进行信息检索、内容生成等工作时，若人工智能生成信息的正确性无法得到保障，使用者的工作准确性也将受到直接的影响。同时，人工智能技术存在的算法黑箱问题挑战了人类的知情权与自主决策权，不透明的算法侵蚀了人类的自主性，同时也可能威胁个体自由与加剧社会偏见。¹⁷据统计显示，在利用人工智能技术对佛罗里达布劳沃德县进行犯罪风险预测的过程中，对非裔美国人进行错误标记的概率是白人的两倍。¹⁸与此同时，亦有研究表明，面部分析的技术错误则亦有可能因种族和性别不同而产生差异。¹⁹其所反映的一系列问题或将进一步扩大并影响社会公众的认知，加剧社会偏见。就其未来发展看，人工智能存在失控的风险。伴随着人工智能技术在深度学习领域的进展，已有越来越多的业内人士对人工智能技术发展表示担忧与不安。²⁰人类对发展超级智能体存在的普遍不安感也被概括为“大猩猩问题”，人类有可能无法控制比我们更加聪明的机器，正如同大猩猩无法控制比我们更加聪明的机器那样。

对于上述问题，固然可以优先从国内法角度加以解决，也应当优先借助拘束力、执行力都更强的国内法律体系实现对人工智能的规制与监管。然而，在人工智能技术博弈和经济全球化的背景下，仅借助国内法治理人工智能问题却是远远不够的。国家与国家之间可能进行人工智能技术的恶性竞争。为了保障本国人工智能技术处于领先地位，国家可能有意放松对本国人工智能技术的监管，并令全球各国被动承担监管不利的消极后果。²¹为了推动人工智能技术发展行稳致远，国际层面进行有序竞争，仍然有必要在利用国内法进行规制的基础上，发挥国际法的作用，对人工智能本身进行有效的治理，将人工智能技术的发展限制在可控的范围内，避免人工智能酿成不可挽回的全球性悲剧。

¹⁵ International AI Safety Report at 88-109.

¹⁶ International AI Safety Report at 88.

¹⁷ 丁晓东：《论算法的法律规制》，载《中国社会科学》2020年第12期。

¹⁸ Julia Angwin et al. *Machine Bias*, Pro Publica, May 23 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

¹⁹ Joy Buolamwini and Timnit Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” *Proceedings of Machine Learning Research*, 2018, Volume 1, pp. 1–15.

²⁰ Artificial Intelligence: A Modern Approach.

²¹ The White House, *Removing Barriers to American Leadership in Artificial Intelligence* (January 23, 2025), <https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/>.

（三）系统性风险

从更广泛的意义上讲，人工智能技术还存在同社会议题相联结的系统性风险：人工智能技术引发的劳动力市场问题、技术研发鸿沟、环境破坏风险、知识产权侵犯等问题都远远超越了个体模型所引发的困境，²²需要国际规则加以体系化地因应。

以劳动力市场问题为例，人工智能技术在方便社会生活的同时，也导致诸多传统工作岗位受到人工智能的取代，进而造成大量劳动者的失业。这种劳动力市场上的风险更有可能发生在低技能劳动者群体之中，进而导致社会不平等现象的加剧，同时也导致发展中国家的劳动力市场受到更为严重的冲击。伴随着全球化的发展，跨国公司获得了在全球布局生产链的能力，其在一方面强化资本权力的同时，也在另一方面弱化了劳工权利，这致使了传统的劳动法不敷所用。²³在人工智能技术很大程度上受跨国公司主导的背景下，跨国公司可能进一步依仗技术方面的优势，侵害发展中国家中劳动者受国际法保护的工作权。²⁴据此，国际社会有必要围绕商业与人权等议题，系统的讨论人工智能对劳动力市场的侵害现状与发展趋势，并制定相应的治理对策缓解劳动力市场问题。

与此同时，在人工智能技术研发方面，“全球南方”国家同北方发达国家之间存在明显的技术鸿沟，“全球南方”国家的技术仍然存在极大落后。²⁵在国际社会致力于共同发展的背景下，新兴技术却导致了国家之间的进一步不平等。在此背景下，由中国牵头在联合国大会通过了《加强人工智能能力建设国际合作决议》试图弥合数字鸿沟，平等发展和利用人工智能技术，共享人工智能知识成果。²⁶然而推动发展中国家人工智能技术发展的美好愿景能否在实践中得到落实，仍然需要进一步的制度建构与时间考验。

总言之，在《国际人工智能安全报告》的风险分类视阈下，人工智能技术所存在的恶意使用风险、故障风险与系统性风险都分别在不同程度、从不同方面对国际法提出了挑战。在恶意使用风险中，人工智能技术动摇了国际法规则的效力，并在一定程度上影响了某些国际规则存在的理论和哲学基础。在故障风险中，国际社会对人工智能技术的安全可控存在现实需要，考虑到技术博弈的背景，避免恶性竞争的诉求有赖于国际规则付诸实现。在系统性风险中，人工智能产生的问题不再局限于个体模型或特定的使用者，而是产生了更广泛的社会影响。上述人工智能风险具有溢出性、普遍性，构成潜在的跨国性风险，亟需通过解释与适用现行国际法、制定新的国际规则等方式加强国际治理。

²² International AI Safety Report at 88.

²³ 陈一峰：《跨国劳动法的兴起：概念、发展与展望》，载《中外法学》2016年第5期，第1382页。

²⁴ International Covenant on Economic, Social and Cultural Rights, Article 6.

²⁵ The Global AI Index 2024, Tortoise, (Last Visit: Nov.11, 2024) <https://www.tortoisemedia.com/intelligence/global-ai/>.

²⁶ 《加强人工智能能力建设国际合作决议》。

三、现行人工智能法律治理路径与机制的局限性

当前,治理人工智能的国际规则呈现“国际软法先行,区域硬法并进”的态势。伦理规范、决议、倡议、宣言等软法既能约束相关实体行为,又不至于遏制创新,在人工智能国际治理初期发挥了重要作用。欧盟《人工智能法》、欧洲委员会《人工智能框架公约》等区域硬法作为最早的统一监管法案和国际公约,为全球性的人工智能国际规则制定提供了良好的范本。

然而,也应当认识到,现阶段人工智能跨国性风险的法律治理路径和机制仍然存在一系列问题与固有局限性,影响了国际社会对人工智能的有效治理。

(一) 现行法律治理路径的局限性

前文述及的“基于原则的治理路径”“基于权利的治理路径”“基于场景的治理路径”和“基于风险的治理路径”仍存在一定程度上的固有局限性,并非人工智能国际治理的最佳路径。

“基于原则的治理路径”从原则出发构建人工智能跨国性风险治理的框架,具有易形成国家间共识等优点,有利于为国家化解人工智能风险提供方向性的指引。然而,“原则性”本身也意味着“模糊性”,致使人工智能的国际治理缺乏一定程度的明确性与可操作性,仍然需要国际规则的补充。从国际法角度讲,规则是具有强制性的,但是原则只提供了一种指南,行为者仍有选择如何行为的余地。²⁷因此,尽管联合国教科文组织、经合组织等发布的人工智能原则为国家行为提供了良好的行为导向与行动指南,但由于在其各项原则框架下国家的解释空间过大,国家仍存在过多的自由抉择空间,无法对人工智能风险治理产生直接的督促和规制作用。

“基于权利的治理路径”以人权等价值观作为主要标准,推动了个体权利保护的有效实现,但也在一定程度上忽视了人工智能治理中需考量的其他因素。事实上,人工智能技术引发的跨国性风险覆盖面之广远远超过了个体权利的保护范畴。在文章第二部分所讨论的内容中,无论是借助人工智能技术进行网络攻击,还是未经授权获取数据训练模型而侵犯知识产权,都反映了技术对于国家安全、法律制度与社会发展所带来的挑战。因此,仅仅以权利导向作为未来人工智能治理的主要路径,则未免有顾此失彼,抓小放大之嫌。

“基于场景的治理路径”从现有法律体系中寻求或制定适用于人工智能在不同行业和领域的监管规则,考虑到人工智能未来垂直应用、产品和服务体系的庞杂,必然导致监管规则臃肿、行政执法部门权责不清。就国际治理而言,这一路径也势必加剧国际法的碎片化以及各国解释、适用国际法的不确定性。

“基于风险的治理路径”在欧盟《人工智能法》中得以确立,其通过对人工智能进行不同风险程度的分级,并针对不同等级的风险提出不同的要求,做到了对

²⁷ 贾兵兵:《国际公法:和平时期的解释与适用》(第二版),清华大学出版社2022年版,第45页。

症下药。与此同时，通过设置进入市场前的预防义务与进入市场后的持续监督，基本实现了对人工智能风险的有效治理。然而，在吸取其经验的同时，也有必要注意到，欧盟《人工智能法》中“基于风险的治理路径”在很大程度上是依托于欧盟这一独特的“超国家”组织形式而存在的。其虽然对欧盟内各个成员国具有相应的拘束力，构成了人工智能“跨国性风险”的治理，但是在另一方面，此类法案却也具有一定程度国内法的性质，其由欧盟主要机构制定，调整各国国内事务并直接对成员国与成员国国民适用，²⁸其是否适合直接作为国际法律治理的主要路径则仍值得商榷。事实上，欧盟《人工智能法》将风险分成四个层次的做法在欧盟法层面或许有利于实施，但是在国际治理过程中，考虑到各国发展程度、利益诉求与治理能力上的差距，此类方法或因其标准过细而难以实现。

总而言之，在寻找可供借鉴的国际治理路径时，现行的治理方式或多或少仍然存在一定的局限性，难以将其视作人工智能跨国性风险治理的最佳路径。人工智能跨国性风险的法律治理仍需要探求其他路径。

（二）人工智能区域规则的普遍性和代表性不足

当前，人工智能国际治理中最大的问题在于，现阶段尚不存在全球广泛参与、凝聚各国共识的国际法规则。人工智能国际治理中虽然充斥着广泛的国际软法，但是从实证主义的国际法立场出发，缺乏国家的明示或默示同意，国际软法本身并不能构成具有拘束力的法律。在仅存的国家间国际规则中，唯有欧盟《人工智能法》与欧洲委员会《人工智能框架公约》具有法律拘束力，其中前一项隶属于欧盟法的范畴，仅对欧盟成员国具有效力，而后者适用于欧洲委员会的成员国和观察员国，签署该公约的非西方国家较少，缺乏发展中国家的广泛参与。

事实上，国际法的欧洲中心主义被诟病已久，其反映了国际法历史上的霸权主义特征。²⁹传统国际法的实在体系不仅是欧洲思想共识的产物，而且其亦从共同的西欧信仰源泉中汲取其生命力。³⁰在人工智能等新兴国际法领域推行反映各国共同意志的国际治理理念与规则，有利于实现国际法的真正民主化。然而，由西方国家所主导的《人工智能法》与《人工智能框架公约》等区域立法与国际条约不仅缺乏非西方国家的广泛参与，甚至进一步加剧了国际法所反映的霸权主义色彩。

未来，欧盟《人工智能法》将持续发挥“布鲁塞尔效应”，³¹美西方国家也将积极推动其区域主张上升为全球标准。作为回应，中国等“全球南方”国家需要进一

²⁸ 张晓东：《论欧盟法的性质及其对现代国际法的贡献》，载《欧洲研究》2010年第一期，第69页。

²⁹ 巴多·法斯本德、安妮·彼得斯：《牛津国际法史手册》，李明倩等译，上海三联书店2012年版，第1037-1060页。

³⁰ Antony Anghie, “The Evolution of International Law Colonial and Postcolonial Realities”, (2006) 27 *Third World Quarterly* 739, p. 739.

³¹ 布鲁塞尔效应（Brussels effect），是指欧盟事实上通过市场机制将其法律推广至境外，导致其单边监管全球化的过程。因为布鲁塞尔效应，受监管的实体，特别是公司，最终会出于各种原因，即使处在欧盟之外，也遵守欧盟法律。

步参与人工智能国际治理议题的讨论,并提出体现广大发展中国家利益的国际治理主张。

（三）现行国际法机制供给与风险治理需求的张力

现行国际法机制不足以应对人工智能技术的跨国性风险,法律规则的供给与治理需求之间存在张力。这进一步增强了完善人工智能法律治理机制的紧迫性与现实性。

这种张力尤其体现在人工智能的国际治理主体方面。传统国际法的主体是主权国家,³²国际法仅在国际人权法、国际组织法等领域才赋予了个人等非国家行为体有限的法律人格。³³然而在人工智能技术与产业的发展中,掌握人工智能最先进的算法、模型,并直接控制人工智能的部署和应用的主体却是各大科技企业。可以说,从某种意义上讲,如 Open AI、Google、Microsoft 等科技公司对人工智能行业的影响远比一些发展中国家要强。因此传统的国际法治理模式在一定程度上存在治理主体的错配问题,而主体的错配也在一定程度上影响了责任承担的错配问题,应当承担责任的主体仅仅在国内法层面间接承担国际责任,而未受到国际法的直接约束。

总言之,现有针对人工智能跨国性风险的法律治理机制还存在一定局限性。首先,在治理路径的选择上,现有的四种人工智能治理路径因其固有局限性并不适合作为人工智能国际治理的最佳路径,仍需要寻求路径方式上的创新。其次,国际社会缺乏凝聚全球普遍共识的国际规则,现行治理规则仍然反映出明显的欧洲中心主义特点,中国等广大发展中国家积极参与人工智能国际规则制定具有必要性。最后,现有国际法机制的供给同人工智能风险治理需求存在错配。鉴此,本文拟提出新的法律治理路径并完善相关国际法律治理配套措施。

四、人工智能跨国性风险的法律治理

人工智能与互联网在技术、应用和社会层面具有密切关联,其重大风险往往具有较强的溢出效应,³⁴尤其是金融、电力、医疗、能源等领域的人工智能系统因故障或滥用产生的风险与危害易波及多国乃至全球。基于风险分级分类的治理路径适配国内监管,却难以在技术与治理能力良莠不齐的国际治理层面体现可操作性。“基于关键性”的法律治理路径以各行业或领域的人工智能系统在国家安全中的重要性为二分法标准,划分“关键人工智能”和“一般人工智能”并明确不同监管措施,可弥补风险治理路径在国际治理中的局限性。跨国性风险的治理主要依赖调整主权国家间以及其他具有国际法律人格的实体间关系的国际法。鉴于跨国

³² 周鲠生:《国际法》商务印书馆 2018 年版,第 60 页。

³³ Malcolm N. Shaw, *International Law* 9th edition, Cambridge Press, p.245-237, (2021).

³⁴ 参见张磊:《人工智能国际规则博弈态势、共识分歧与中国“话语势能”建构》,载《新疆社会科学》2024 年第 4 期,第 124 页。

科技企业在人工智能的技术创新与产业发展中发挥主导作用，应赋予其不完全的国际法律人格并澄清其国际法律权利和义务。更重要的是，应确立相关实体在人工智能研发、销售、部署和应用全生命周期的国际法律责任，统筹国内法与国际法体系协同治理，加强问责制。

（一）采用“基于关键性”的法律治理路径

有学者提出一种具有适应性的分级分类治理方案，将能力、影响和属性等维度与风险程度结合形成复合型的分级分类标准，然后综合区分“关键人工智能”与“一般人工智能”。“关键人工智能”系指达到一定能力标准、在重要场景发挥核心功能或应用于关键基础设施核心功能的人工智能系统。³⁵为便于和“基于风险”“基于场景”“基于权利”等治理路径形成对比，笔者将之称为“基于关键性”的治理路径。这一治理路径是在探索中国国内人工智能监管模式的背景下提出，但其标准较为笼统、模糊，不利于企业合规与执法管辖，反而更适合因应当前碎片化、阵营化的人工智能国际治理现状。

在国际治理层面，“基于关键性”的治理路径相较于“基于风险”的主流治理路径具有突出优势，体现在四个方面：

一是分类简洁，易于在国际社会达成共识。欧盟《人工智能法》将风险分为不可接受的风险（禁止的风险）、高风险、有限风险以及最小风险四个级别。这种分类具体详细，有利于国内监管中落地执行，但在国际治理层面，不利于技术能力、法治水平多样化的国际社会达成共识。“基于关键性”的治理路径分类简洁，相对适合在人工智能国际规则建构初期求同存异，制定统一的全球风险管理框架。

二是概念内涵和制度外延更具有包容性。“基于风险”的治理路径注重保护基本人权并以风险程度为唯一标准，存在不科学和僵化的困境。³⁶“基于关键性”的治理路径注重维护国家安全、公共安全，³⁷在评估“关键人工智能”时综合考察风险、能力、属性、影响等多元标准，尤其关注人工智能系统在国家政务、金融、电力、水利、能源、通信、医疗等关键基础设施中的应用风险，不仅在概念内涵和制度外延方面更具有包容性，而且因关键基础设施是国家主权的核心表征，³⁸更适合跨境国家安全风险的治理。

³⁵ 参见张凌寒：《人工智能法律治理的路径拓展》，载《中国社会科学》2025年第1期，第107页。

³⁶ 参见丁晓东：《人工智能风险的法律规制——以欧盟〈人工智能法〉为例》，载《法律科学（西北政法大學學報）》2024年第5期，第3页。

³⁷ 一些学者认为，对人工智能系统及其提供者的法律规制，尤其应该注重国家安全、重大公共安全方面。这类观点是本文以人工智能系统在国家关键基础设施中的应用为考察中心的理论基础。参见丁晓东：《人工智能风险的法律规制——以欧盟〈人工智能法〉为例》，载《法律科学（西北政法大學學報）》2024年第5期，第3页。

³⁸ 多数国家在其关于国际法适用于网络空间的立场声明中指出，对关键基础设施发起网络攻击即视同侵犯国家主权。See UN GGE, “Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States Submitted by Participating Governmental Experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security Established Pursuant to General Assembly Resolution 73/266”, A/76/136, 13 July (2021) .

三是在网络空间国际治理领域有二级分类的先例可资借鉴。联合国信息安全政府专家组（UNGGE）、联合国信息安全开放式工作组（OEWG）发布的年度报告中（联合国大会审议通过）将网络空间的基础设施分为关键（信息）基础设施（critical infrastructure）和非关键基础设施，联合国“网络空间负责任国家行为规范”第 f 条还专门针对关键基础设施制定了不具有法律约束力的保护性规范。³⁹大多数国家将对关键基础设施发起网络攻击视为侵犯主权、使用武力、干涉等一般国际法规则的重要标准，⁴⁰这一分类已经具有相当普遍的国家实践和法律确信。在借鉴这一先例的前提下，也有利于一般国际法在人工智能领域的类推适用并为制定国际条约奠定良好的基础。

四是作为具有创新性的中国主张，有助于规避欧盟规则的“先发陷阱”，既为人工智能国际治理提供了崭新的中国智慧与中国方案，也有助于提升中国叙事的向心力和国际话语权。

（二）赋予跨国科技企业不完全的国际法律人格

国家、国际组织、跨国公司和个人等实体尽管在国际层面都有某种程度的影响，但他们并非都构成法律人格者。国际人格者不仅要参与，还需要以某种形式被社会接受。这种接受取决于不同因素，如国际人格的形态、表现形式，或者在特定国际法分支里发挥关键作用。⁴¹在“巴塞罗那牵引公司案”后，与跨国公司相关的国际实践明显增多。但关于跨国公司的国际法律人格问题依然没有解决。

在新一轮科技革命的背景下，鉴于跨国科技企业在推动技术和产业发展及其治理方面发挥的关键作用，本文主张赋予跨国科技企业在人工智能等新兴技术领域不完全的国际法律人格。首先，从新现实主义的角度讲，跨国科技企业通过技术垄断已经实现了治理权力的扩张，国际权力结构有必要进行调整。在现代科技社会中，技术已经不仅仅是工具，更是塑造国家、社会发展路径的关键变量。传统权力形式仍然重要，但其运行越发依赖对关键结构性资源及其治理规则的掌控。⁴²脱离跨国科技企业的配合，国际社会或将陷入“无序状态”。主权国家在新兴技术等特定领域向跨国科技企业让渡部分治理权力是合理的、利大于弊的。其次，在人工智能的区域治理中，赋予科技企业法律人格已具有一定国际实践。2024 年，《欧盟人工智能法》、欧洲委员会《人工智能框架公约》相继通过。鉴于《欧盟

³⁹ See UN GGE, “Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security”, A/76/135, 14 July (2021) .

⁴⁰ 近年来，发布国际法适用于网络空间立场声明的国家逐渐增多，以下仅列举 2021 年及以前的联合国汇编文件，2022 年至今尚未更新汇编文件。See UN GGE, “Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States Submitted by Participating Governmental Experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security Established Pursuant to General Assembly Resolution 73/266”, A/76/136, 13 July (2021) .

⁴¹ 参见[英]马尔科姆·N·肖：《国际法》（第六版上卷），白桂梅等译，北京大学出版社，第 158 页。

⁴² 参见孙志伟、殷浩铨：《人工智能时代数字巨头的技术权力及其对“全球南方”的挑战》，载《国际安全研究》2025 年第 2 期，第 144 页。

人工智能法》全面生效尚需两年时间，且《人工智能框架公约》因多国反对未直接对科技企业施加国际义务，2024年9月至今，近两百家科技企业签署了欧盟《人工智能公约》（AI Pact）。欧盟《人工智能公约》要求企业制定人工智能治理战略与《欧盟人工智能法》衔接、识别高风险人工智能系统以及加强员工相关培训等。⁴³这一公约虽然是自愿承诺性质的国际软法，但从其签署和实施过程来看，也在某种程度上赋予了科技企业在欧盟区域内的国际法律人格。

不完全的国际法律人格体现在，跨国科技企业需要承担国际法律义务并享受相应权利，但无法与主权国家等同。一方面，跨国科技企业需要承担人工智能系统研发、销售、部署和应用全生命周期的预防义务、侵权责任等，并为人工智能的应用与活动所造成的损害承担救济和赔偿责任。同时，权利与义务对等。跨国科技企业也应享有人工智能国际治理规则制定、实施与执行过程中的磋商和建议权。另一方面，跨国科技企业的权利不能与主权国家等同。企业因其注册地、经营地等限制，天然地受所在地主权管辖。主权国家具有对内最高管辖权和对外独立权。企业的决策会在一定程度上受到主权国家的影响乃至控制。除此之外，企业缺乏主权国家特有的强制执行力，在国际责任的追究过程中能发挥的作用较为有限。

（三）确立相关实体在人工智能全生命周期的国际法律责任

确立相关实体的国际法律责任有利于保障国家、企业和个人正当权益，维护国际社会秩序，实现公平正义等法律价值。这首先要求厘清人工智能在国际法上的调整对象和调整范围。在国际治理层面，主要的法律人格者包括主权国家和跨国科技企业。除涉及战争罪、危害人类罪等国际刑事责任外，个人不属于国际法的调整对象。就调整范围而言，国际法只需要处理人工智能研发、销售、部署和应用等全生命周期的跨国性风险与损害。

确立人工智能全生命周期的国际法律责任应以“人工智能价值链”为参考，设定法律身份、选取中心节点并分配法律义务。⁴⁴所涉法律身份较为复杂，可简化为三类：人工智能模型和系统的研发者与部署者（以下简称“人工智能提供者”）、人工智能提供者注册或经营所在地的管辖国（以下简称“提供者管辖国”）、因人工智能应用或服务受影响者的管辖国（以下简称“受影响者管辖国”）。人工智能提供者（主要是跨国科技企业）因其负责人工智能模型和系统的开发、数据训练、销售和部署等，处于人工智能价值链的中心节点。现阶段，人工智能技术发展方兴未艾，对可能阻碍创新的限制和监管义务的设置应予克制，以满足基本需求为指引。人工智能全生命周期的国际法律义务可根据“事前——事中——事后”的逻

⁴³ 截至2025年3月24日，共有193家科技企业签署欧盟《人工智能公约》。See EU, “AI Pact”, <https://digital-strategy.ec.europa.eu/en/policies/ai-pact>.

⁴⁴ 参见李润生：《人工智能价值链的法律型塑》，载《东方法学》2025年第2期，第1页。

辑分配,主要确立事前预防义务、事中共同监管义务以及事后的救济与赔偿义务。

预防义务是(风险)预防原则(**precaution principle**)义务属性的体现。预防原则在1941年的“特雷尔冶炼厂仲裁案”中最早作为一项国际法律原则被提出,国际仲裁庭认为“任何国家皆需对境内活动所造成的跨国界环境危险以及可预见的环境风险承担防治或预防的责任”,⁴⁵后在国际环境法领域逐渐成为一项基本原则。⁴⁶近年来,预防原则在全球治理中的地位和作用日益受重视,扩展适用于国际和平与安全、气候变化、海洋、公共卫生、新兴科技等诸多领域。有学者指出,在技术发展和治理的早期阶段应纳入预防原则,该原则已被证明在管理与新兴和不可预测技术有关的风险方面是有效的。⁴⁷《欧盟人工智能法》、欧洲委员会《人工智能框架公约》虽未直接表述预防原则,但其风险评估、缓解与审查等措施均体现了预防原则的应用。⁴⁸在人工智能国际治理中,预防义务的建构包括两个方面:首先,人工智能提供者需承担主要的预防义务,如通过建立“合格测评制度”等确保开发模型、系统、产品或服务的安全性,并实现人工智能与人类价值观对齐。其次,提供者管辖国需承担次要的预防义务,在本国所产模型、系统、产品或服务的出口和国外应用中把牢质量与风险的审查关,严格遵循“不合格不上市(市场)、不合格不出口”的要求。

事中共同监管义务是预防义务的延续。仅依靠事前预防,难以确保风险保障和评估的实质性标准一以贯之,在人工智能出口和海外应用过程中始终得到落实。人工智能提供者、提供者管辖国应对人工智能模型和系统的二次开发、集成和应用等问题保持关注(**constant care**),尤其避免恶意使用、滥用等风险。事前预防、事中共同监管并不能排除所有风险,而是将风险控制在合理的范围内。

人工智能的应用与活动致损难以完全避免,建立事后救济与赔偿机制是必要的。⁴⁹责任制度和权利救济有助于实现法律治理的事后反馈,加强制度工具间的协同,促进人工智能提供者、提供者管辖国在事前、事中阶段加强风险的防范和控制,而缺乏有效的问责制会导致事前、事中的管理标准难以转化为切实的成果。⁵⁰就损害的赔偿而言,或可建立一种基于保险的风险分散机制。通过建立保险机制来分散由不可预测的人工智能应用结果所引发的损失或损害风险。国际保险基金可由主权国家、跨国科技企业共同出资,确保在发生不可预见的损害事件时,能够为受害者提供及时的经济补偿。⁵¹

⁴⁵ See *Trail Smelter Case (United States v. Canada)*, Reports of International Arbitral Awards 1941, p.1939.

⁴⁶ 张华:《论欧盟食品安全法中的风险预防原则:问题与前瞻》,载《欧洲研究》2011年第4期,第101页。

⁴⁷ See Aida Ponce Del Castillo, *Artificial Intelligence, Labour and Society* (ETUI Aisbl, Brussels, 2024), p.14.

⁴⁸ See Tuomas Pöysti, “The Precautionary Approach Design Pattern”, (2024) 3 *Digital Society* 1, p.11.

⁴⁹ See Marietje Botes, “Regulating Scientific and Technological Uncertainty: the Precautionary Principle in the Context of Human Genomics and AI”, (2023) 119 *S Afr J Sci* 1, p.3.

⁵⁰ 参见张凌寒:《人工智能法律治理的路径拓展》,载《中国社会科学》2025年第1期,第110页。

⁵¹ See Bronwyn Howell, “Regulating Artificial Intelligence in a World of Uncertainty”, <https://www.aei.org/research-products/report/regulating-artificial-intelligence-in-a-world-of-uncertainty/>.

（四）统筹利用国内法与国际法体系协同治理

人工智能跨国性风险的协同治理需要平衡技术创新与安全防护，既要通过国内法守护安全底线，又要依托国际法建立责任共担机制，最终形成“底线管控——标准互认——动态响应”的立体治理网络。

在底线管控方面，通过国内法优化人工智能的分级分类监管体系，完善数据和版权治理机制，构建相关责任主体的分层追责制度。把可能产生跨国性风险的人工智能应用和活动列入“禁止或高风险人工智能”，加强安全审查，捍卫人工智能出口或海外上市应用最重要的一道屏障。

在标准互认方面，把中国《人工智能生成合成内容标识办法》《互联网信息服务算法推荐管理规定》等监管规则及其执行与实施信息通过双多边协定、国际技术标准组织等与他国互认，或建立联合监管沙盒、“一带一路”国家间人工智能产品认证机制等，实现跨国合规数据共享。

在动态响应方面，通过联合国等平台建立人工智能风险响应与损害赔偿机制，既能加强国际层面的人工智能风险反馈与治理，也能协商处理涉人工智能的跨国性争端，监督救济与损害赔偿义务的履行。

五、结语

现阶段，生成式人工智能（GenAI）、智能体（AI Agent）、通用人工智能（AGI）等人工智能技术与产业发展样态的故障风险、滥用风险以及系统性风险有扩散为跨国性风险的可能，但现有的“基于风险”“基于场景”“基于权利”的治理路径在人工智能国际治理层面仍然具有实践局限性。“基于关键性”的治理路径在风险分级分类的基础上作出适当改良，通过“关键人工智能”与“一般人工智能”的二元划分，包容风险、能力、属性、场景等多样参数，更适应碎片化、阵营化的人工智能国际治理现状。为促进“基于关键性”的法律治理路径的落地实施，需赋予跨国科技企业不完全的国际法律人格，澄清其在人工智能国际法上的权利和义务。同时，以“人工智能价值链”为参照，设立人工智能提供者、提供者管辖国以及受影响者管辖国的法律身份，明确分配事前预防义务、事中共同监管义务以及事后救济和赔偿义务等。最后，鉴于跨国性风险通常由国内溢出，从底线管控、标准互认和动态响应方面构建国内法与国际法的协同治理体系是应有之义。

数字主权与全球数据监管：争议、实践及反思

吕思嘉¹

摘要：数字时代，数字主权成为数字治理的重要议题。伴随数字技术的发展，数字主权不再局限于传统领土主权，有了更为丰富的内涵。然而围绕数字主权这一概念的主体、内容、性质、实践存在很多争议，数字主权的界定并不清晰。这导致各国采取不同的数字主权叙事方式，在数据监管上体现的尤为明显。数据主权是数字主权的核心之一，全球数据主权实践虽然主要形成了欧盟、美国、中国三种数据监管模式，但本质上都以安全为核心，是传统领土主权观的延伸。这并不适应数字社会的发展需求，可能阻碍数字经济的发展和数字技术的提升。数字时代，各主体之间联系紧密，需要拓展新的多元化的数字主权理论和实践方式。国家一方面可以发展合作式数字主权以指导数据治理实践，另一方面也可以通过创新数据治理机制以丰富数字主权的理论内涵。

关键词：数字主权 数据主权 数据监管 数据安全 数据治理

引言

数字主权自诞生起，热度不减。但是数字主权的内涵和外延一直有待确定。虽然网络空间国家主权从未离开，但数字时代，主权应当以何种方式发挥作用，这一问题仍然有待回应。理论上的争议可以总结为两个观点：一是认为数字主权是传统威斯特伐利亚体系领土主权的延伸，另一观点认为数字主权有着新的表现形式，或是进行功能主义的划分²，或是从数据主权、技术主权等内容上对其赋予新的内涵³。虽然理论上学者对数字主权的界定有所创新，但是在实践中，数字主权仍然延续了传统领土主权的实践原则，以国家安全为核心，实施强监管措施。这在数据监管上体现的尤为突出。数据主权是数字主权的核心组成部分。各国数据监管愈发强化，体现了数据主权和数字主权安全化的演进趋势。不论是欧盟的权利监管模式、美国的自由主义霸权监管模式、还是中国的国家安全监管模式，实质上都是主权安全的变相呈现。

复杂的政策实践引出的理论问题是：数字主权理论应当如何回应数字技术对传统主权的冲击，如何引导国家的数据治理实践？数据的流动性、跨国性、风险的弥散性等特性使传统主权不能解决各国数据治理的实际需求，我们应当如何修

¹ 复旦大学法学院国际法专业博士研究生

² 刘晗,叶开儒.网络主权的分层法律形态[J].华东政法大学学报,2020,23(04):67-82.

³ 廖凡.数字主权与全球数字治理[J].暨南学报(哲学社会科学版),2024,46(07):47-60.

罗有成.数字时代主权的嬗变与国际安全秩序重塑[J].国际展望,2024,16(06):89-112+161-162.

正理论视角以丰富数字主权内涵,为未来的数据治理以及更广范围的数字治理提供路径指引?本文首先将回溯数字主权的现存争议,其次基于欧盟、美国、中国现有的数据实践反思数字主权实践的局限和不足,最后本文将提出数字主权理论的革新和探索方向,建议国家将重心转移至国际合作机制的探索与创新,以引领数据治理实践。

一、数字主权的内涵和争议

从互联网主权、网络主权、信息主权、到数字主权,随着科技的发展,主权的形势也愈发多样化。数字主权成为数字时代的主旋律。然而数字主权的概念并不清晰,关于其实施主体、核心内容、法律性质,理论和实践中存在不同的主张和意见。

(一) 数字时代主权的变迁

16世纪法国思想家博丹最先提出近现代意义的主权概念,指的是国家绝对、最高和永恒的权力,独立于法律之上。主权可以被人民、贵族、君主等一人或数人所持有。随着时间推移,这一概念发生了很多变化,以适应不同国家不同时期政治统治的需要。纵观其衍生出的各种含义,大体上可以分为以下几类⁴: 1.主权作为国家或国家权力的一种属性,而非某种权力、权利或权威本身。2.指代国家权力本身,国家行使的各种权力总和。3.强调主权的最高性、最终性要素,即特指国家内部最高、最终决定权。4.例外状态决定权。5.制宪权或修宪权。6.国家正当性的来源或基础。国家主权行使方式可以是议会主权、人民主权、契约君主主权、君权神授等。可以看出主权的含义主要包含两个部分:一部分是客观要素,将主权看作一种权力形式、属性,一部分是主观要素,包含权力的正当性、合法性来源。

从国际法来看,主权概念经由《威斯特伐利亚和约》确认后,更多指的是国家内部的最高性,即独立自主处理对内和对外事物的最高权力,不受他国干涉,包括了独立权、平等权、自卫权、管辖权。随着经济贸易的发展,超国家的跨国公司、国际性政府间组织和非政府组织的出现,国家被迫让渡一部分主权,国家主权被不断“软化”。从主权演变的形式来看,主权理论经历了进一步发展⁵:主权弱化论、主权过时论/终结论、主权可分/让渡论、人权高于主权论等都在强调国家主权功能减弱。“关系性主权”“分享式主权”“可分的主权”“被悬置的主权”等话语兴起。

互联网的出现和发展是对国家主权的新一轮冲击。国家主权是否可以适用于互联网?互联网主权的出现最初表达的意思与主权毫不相关,用以表达的是互联

⁴ 陈楚风.论宪法中的主权概念[J].甘肃政法大学学报,2023,(06):87-106.

⁵ “国家主权”的精要介绍:历史起源、概念解析与理论定位
<https://mp.weixin.qq.com/s/0i3BRrvDwNf4jFEef8iXIA>

网空间具有独立与国家的主权。这是由于最初构建互联网的美国工程师大多秉持技术乐观主义和反国家主义的精神，认为互联网本质在于自由和创新，独立于政府和企业。因此互联网应当实现自治，摆脱政府权力和资本的宰制。这一时期的主权否定论最典型的是约翰·巴洛著名的《网络空间独立宣言》(Declaration on the Independence of Cyberspace)。在他们眼中，网络空间超越国家主权的最重要的原因在于互联网突破了国家传统主权的领土边界，国家主权基于领土的属地管辖权失去作用空间。⁶互联网是自由的全球公域，应当建立独立的生态系统。⁷

从那时起，政府机构和学术界关于网络空间主权的争论一直持续。然而，在国际实践中，除了某些国家和组织正在推广多利益攸关方治理模式外，没有证据表明主权的概念发生了实质性变化。正如 Mainwaring 所说，“主权和国家权力被改变，而不是被抹去”。⁸ 国家主权并没有消失，而是不断得到越来越多国家的重申。随着全球安全问题、政治经济问题的突出，国家日渐加大对互联网的监管力度，数字经济竞争和资源的争夺也与日剧增，国家主权“重回”网络空间。“网络主权”“信息主权”“数据主权”等主权概念不断被提出，多是在指涉一种基于领土的排他性威斯特伐利亚主权观，带有非常强的领土性和防御性。这正是现代欧洲经由威斯特伐利亚体系推广到世界的国际法原则和逻辑：“现代主权反抗除自己以外的其他一切政治力量，反抗除自己以外的一切其他国家。”⁹进一步地，这一防御性主权开始向扩张性主权过渡，向帝国式主权演变。

国家主权的复归愈发确认国家网络空间的地位。网络空间不是“全球公域”，也并非独立于现实世界的法外之地，主权国家有权管理其领土范围内的互联网“领地”。¹⁰网络主权是传统国家主权在网络空间的延伸和拓展，虽然外部条件发生了变化，但是其内涵和逻辑并没有颠覆主权范畴。¹¹在新时代，网络主权也确实发生很多变化。在治理主体、治理内容、治理理念、治理规则、主权维护手段等方面都发生了显著变化。¹²其中数据成为主权国家网络治理的重中之重。数据的跨界流通特性超出单一主权国家的领土范围，容易引起国家主权纠纷。加上数据作为经济发展基础要素，已不仅仅是权力的来源，也成为力量本身，可以作为具有颠覆性的武器来使用。掌握更多数据和数据处理能力的国家将拥有更多的网络空间权力，国家竞争与数据权力紧密相连。

随着数字技术应用的普遍化，网络主权开始向数字主权演进。数字主权被认

⁶ David R. Johnson and David Post, Law and Borders: The Rise of Law in Cyberspace, Stanford Law Review, Vol. 48, No. 5 (May, 1996), pp. 1367-1402.

⁷ Mueller, M. L. (2020). Against Sovereignty in Cyberspace. International Studies Review, 22(4), 779-801.

⁸ Mainwaring, S. (2020). Always in control? Sovereign states in cyberspace. European Journal of International Security, 5(2), 215-232.

⁹ 迈克尔·哈特 安东尼奥·奈格里 《帝国·全球化的政治秩序》江苏人民出版社 2003 年版第 9 页。

¹⁰ Goldsmith, Jack and Wu, Tim, "Who Controls the Internet?: Illusions of a Borderless World" (2006). Faculty Books. 175.

¹¹ 黄志雄：网络主权论——法理、政策与实践，p.3，社会科学文献出版社，2017。

¹² 熊光清,王瑞.网络主权：互联网时代对主权观念的重塑[J].中国人民大学学报,2024,38(01):126-138.

为建立在网络主权的基础之上。关于二者何者范畴更大的问题存在不同的认识。但普遍而言，数据和数字技术是二者交汇的核心。

数字主权愈发成为政治实践和学术话语中的主要议题。然而，如何理解数字主权？数字主权不仅仅是传统主权的数字化版本，也不会取代主权的法律地位。它既不是主权原则的结果，也不是主权原则的延伸。实际上，数字主权是特定法律、政治和科学话语的核心，但它并不总是与其物理性质相关联。数字主权“具有多学科的内涵，它可以具有不同的含义，用以描述不同领域的多个方面，具体取决于实际应用场景”。¹³ 因此，主权和数字主权具有不同的用途和方法。虽然主权是一项对所有国家和任何地方都具有同等范围和意义的一般原则，但数字主权在不同情况下的理解并不相同。在理论和实践中，在不同场合援引数字主权会改变其功能和概念。

（二）数字主权的争议

数字主权的内涵和外延并不明确，主要争议点围绕其实施主体、内容范畴、法律性质以及法律实践。

1. 主体之争

数字时代，越来越多的观点认为主权不再只是国家的象征。第一个观点是网络例外论，它认为数字领域与现实世界在性质上是不同的，数字空间需要采取不同方式区别对待。这一观点在 1990 年代商业互联网兴起期间流行一时，认为互联网拥有自己的主权，国家主权不适用于网络空间，互联网的兴起意味着国家主权的消亡。¹⁴网络例外论，通常以网络自由主义的形式出现，¹⁵是硅谷技术自由主义的延伸和发展，拥有强大的文化和经济支持。网络自由主义者极不信任国家和政府机构，他们认为，数字媒介的政治形式将促进社会的去中心化组织。这应该能够比传统形式的政治组织更好地应对现代社会治理的复杂需求。根据这种观点，国家主权、法律和领土在跨国网络的背景下显得不那么重要。由于数字技术的发展，国家治理手段无法跟上数字技术和相关商业模式创新步伐，无法妥善解决网络空间的复杂问题。¹⁶因此，与受领土和主权国家束缚的世界相反，网络空间是一个独立于政府干预的新的、自主的虚拟领域。¹⁷

从 2000 年开始，第二个挑战国家主权的观点是多利益攸关方互联网治理。

多利益攸关方模式强调自下而上、协商合作、开放透明的治理原则。因此在原则上，政府、私营部门、民间社团的权利是相同的，都可平等地参与制定网络

¹³ Fabiano, N. (2020). Digital Sovereignty Between “Accountability” and the Value of Personal Data. *Advances in Science, Technology and Engineering Systems Journal*, 5(3), 270–274.

¹⁴ Katz, J. (1997). Birth of a Digital Nation. In *Wired*. <https://www.wired.com/1997/04/netizen-3/>.

¹⁵ Keller, C. I. (2019). Exception and Harmonization: Three Theoretical Debates on Internet Regulation (2020(2); HIIG Discussion Paper Series). Alexander von Humboldt Institut für Internet und Gesellschaft.

¹⁶ Post, D. G. (2007). Governing Cyberspace: Law. *Santa Clara High Technology Law Journal*, 24(4), 883–913.

¹⁷ Barlow, J. P. (1996). A Declaration of the Independence of Cyberspace. Electronic Frontier Foundation.

空间规则。当前，采取多利益攸关方治理模式的组织机制主要有互联网名称与数字地址分配机构（The Internet Corporation for Assigned Names and Numbers, ICANN）、联合国信息世界峰会（WSIS）、互联网协会体系、万维网联盟、世界互联网大会、“伦敦进程”等。

多利益攸关方治理模式明确拒绝既定的政府主导网络空间治理，并试图用自下而上的跨国主义原则取而代之。然而，仍然有很多国家坚持国家主导网络空间治理的多边主义模式。由于网络空间治理进程往往并行不悖，以及主要治理议题从技术问题转向更公开的政治或社会议题，协调问题日益明显，导致了网络空间治理的内部冲突。

2. 内容之争

从主权到数字主权，数字主权的内涵和外延已然有所不同。传统主权指的是国家主权，欧盟率先提出数字主权，将其用于自身，赋予数字主权更多内涵，而不仅仅只是主权的附属或者是其延伸。虽然就国家而言，数字主权作为传统主权的一部分是合理的，但对欧盟而言，其功能必然有所不同。

数字主权最早出现在 2020 年 7 月欧洲议会发布的《欧洲数字主权》文件中，“数字主权”没有被明确定义，只是将其简单描述为“欧洲在数字世界独立行动的能力，应从保护性机制和促进数字创新的防御性工具（包括与非欧盟公司合作）两方面来理解。”这一概念被认为是欧盟“面对数字世界的竞争，为保持独立性、竞争力与领导力提出的，强调欧盟主导本地数字发展的能力。”传统主权本是威斯特伐利亚体系下国家的专有属性，欧盟使用主权概念，将主权应用于联盟层面，是在类比意义上使用。数字主权是欧盟晚近提出的“欧洲主权”“战略自主”的一个方面，着眼点和出发点在于提升欧盟“技术自主”能力，避免非欧盟科技公司制约欧盟科技企业的成长，威胁欧盟公民个人数据安全，影响成员国执法能力。因此，“数字主权”和“技术主权”在某些方面可以作为同一概念使用。但是根据时任欧盟轮值主席国的德国发布的报告，数字主权“不是一个明确定义的概念，而是各自社会经济秩序的政治愿景。它从根本上解决了减少数字化世界中现有和新出现的依赖性的问题”。¹⁸事实上，数字主权没有单一或普遍的概念，也没有关于这个词本身的共识。数字主权、技术主权、数据主权、战略主权等表达方式，在欧盟的话语体系中时常交替使用、或互换使用。这些术语并不能总是作为同义词，因为它们都代表了“数字主权”这个总体上更广泛的概念的一个方面。¹⁹

数字主权被认为“已经成为一个更加全面的概念，不仅解决了互联网通信和连接的问题，还解决了更广泛的社会数字化转型问题”。²⁰数字主权包括“强大的、

¹⁸ Report of German Presidency on Digital Sovereignty. https://erstesung.de/wp-content/uploads/2020/10/20-10-14_Germany_EU_Digital-Sovereignty.pdf

¹⁹ Europe's Quest for Digital Sovereignty: GAIA-X as a Case Study. Istituto Affari Internazionali (IAI), Roma. <https://www.iai.it/en/pubblicazioni/europes-quest-digital-sovereignty-gaia-x-case-study>.

²⁰ Pohle, J., and T. Thiel. 2020. Digital sovereignty. Internet Policy Review [Online] 9 (4).

创新的工业基础和足够的网络安全保护能力”。²¹在数字主权之下，包括技术主权、数据主权等多层次的主权形态。比如技术主权可以被定义为“一个国家或国家联邦提供其认为对其福利、竞争力和行动能力至关重要的技术的能力，并能够开发这些技术或从其他经济领域采购这些技术，而不会产生片面的结构性依赖”。²²可以说，数字主权是网络空间中外部参与者对技术主权的又一次“迭代”，基于三个不可分割的支柱：计算能力、对数据的控制和安全连接。²³ 欧盟实践中，经常将技术主权和数字主权互换使用。而从逻辑上讲，数字主权将是一个更广泛、更通用的概念，而技术主权是数字主权的组成部分，战略主权则是从整体政治维度对数字主权的传达。除此之外，数字主权还包括数据主权、网络主权。²⁴

3.性质之争

脱离欧盟语境，很多理论也将数字主权定义为不同于传统主权的概念。然而，这个概念尚未达成共识。

由于数字主权的许多概念化发生在学术界之外，导致各种术语之间的混淆。因此，数字主权与不同的特征相关联，从对数据、数字服务和算法的监管机构，到对硬件和基础设施的控制，这一概念的含义和目的各不相同。因此，数字主权是一个流动的概念，其内涵和预期效果经常变化。”²⁵数字主权本质上是一个模糊的概念，²⁶ 从性质上看，数字主权通常有三种理解方式：

第一种观点认为数字主权是传统领土主权的延伸。数字主权应该被广泛定义，以涵盖国家的主权权力，不仅包括国家通过使用互联网过滤技术和数据本地化措施以监管数据跨境流动，而且还包括一系列获取、使用数字技术的行为活动。²⁷

第二种观点认为从功能主义来看，数字主权是一种权力和能力，是国家、个人或社会对数字活动所拥有的控制能力。²⁸在网络空间，这种功能性的主权在网络空间的不同层面表现不同，正在从物理层、内容层为中心转向以代码层为中心。

29

第三种观点是从价值层面考量，认为主权的概念在经过各种重新表述后，数

²¹ Burwell, F. G., & Propp, K. (2020). The European Union and the Search for Building “Fortress Europe” or Preparing for a New World?

²² Edler, J., Blind, K., Frietsch, R., Kimpeler, S., Kroll, H., Lerch, Ch., Reiss, T., Roth, F., Schubert, T., Schuler, J., & Walz, R. (2020). Technology sovereignty from demand to concept. Fraunhofer Institute for Systems and Innovation Research.

²³ Csernaton, R. (2021). The EU’s Rise as a Defense Technological Power: From Strategic Autonomy to Technological Sovereignty. Carnegie Europe Program.

²⁴ 廖凡.数字主权与全球数字治理[J].暨南学报(哲学社会科学版),2024,46(07):47-60.

²⁵ Prokscha, A. (2021, June). Digital Sovereignty for the European Union - Analysing Frames and Claims for Digital Sovereignty in the European Union’s Digital Strategy.

²⁶ Elms, D. (2021). Digital Sovereignty: protectionism or autonomy. Hinrich foundation, Asian Trade Centre.

Fabiano, N. (2020). Digital Sovereignty Between “Accountability” and the Value of Personal Data. *Advances in Science, Technology and Engineering Systems Journal*, 5(3), 270–274.

²⁷ Chander, A., & Sun, H. (2021). Sovereignty 2.0. Georgetown University Law Center.

²⁸ Posch, R. (2015). Digital Sovereignty and IT-Security for a Prosperous Society. En Werthner, H. y Van Harmelen, F. *Informatics in the Future. Proceedings of the 11th European Computer Science Summit (ECSS 2015)*.

²⁹ 沈伟伟.网络主权视角下的数据治理比较研究[J].法学杂志,2024,45(04):24-39.

字主权越来越多地用于描述各种形式的数字技术和数字内容的独立、自主和可控。³⁰从价值论来看，数字主权一词用于指代可以满足个人权利和自由、社会平等和公平、经济竞争有序、安全稳定且受控的数字领域。³¹

4.实践之争

从话语实践来看，数字主权的表现更为多样，既可以表现为独立自主的防御权，也可以表现为具有扩张意图的规则制定权和全球话语权。从数字经济规则里最为核心的数据监管中，可以更为明显的看到数字主权表现形式的多样，窥得不同国家和地区如何主张其数字主权，数字主权又是如何变迁的图景。数字主权是否具有实质内涵，还是只是国家的一种话语实践和政治口号？

不同的主权概念服务于不同的国家利益。³²如今，数字主权的概念正在许多政治和经济领域得到应用。它有着各种各样的内涵、变体和不断变化的性质。它的具体含义根据不同的国家和行为主体而有所不同。³³数字主权的含义至少包括经济自主、网络自主、技术自主几个要素。在欧盟，数字主权突出表现为防御性战略自主，在美国，数字主权表现为扩张性的“数字霸权”，在中国，数字主权表现为总体安全观下的数字安全。³⁴

除此之外，还出现了用户自主权和个人自决权，这来源于德国数字主权政策。³⁵这一主张强调个人自决权的重要性，侧重于公民作为雇员、消费者和用户的角色的自主权。这一含义背离了以国家为中心的主权理解范式。行为者不将主权视为在特定领土行使权力的先决条件，而是将其视为个人有意识、独立思考采取行动和做出决定的能力。数字主权应当加强公民自决能力以及公民的民主权利，³⁶防止专制政权下用户和公民可能受到数字主权措施的约束和控制。³⁷

二、数字主权的叙事和实践

数字主权的概念和性质较为模糊。从不同角度来看，数字主权有着不同的内涵，各国对其也有着不同的界定。从各国数据监管法律实践来看，数字主权的多元叙事更为明显，其不同层次的内涵也表现的更为清晰。

（一）欧盟的权利叙事

欧盟数据监管已经形成一套具有自身特色的数据安全保护制度体系，这一体系以个人数据权利保护为中心，采用非常严格的隐私和数据保护框架。首先，欧

³⁰ Crespi, F., Caravella, S., Menghini, M., & Salvatori, Ch. (2021). European Technological Sovereignty: An Emerging Framework for Policy Strategy. *Intereconomics. Review of European Economic Policy*, 56(6), 348–354.

³¹ European Digital Sovereignty. Institute of European Democrats. <https://www.iedonline.eu/download/2021/IED-Research-Paper-Innerarity.pdf>

³² Milton L. Mueller, “Against Sovereignty in Cyberspace,” *International Studies Review*, Vol. 22, No. 4, 2020, pp.779-801.

³³ Pohle, J. 2020. Digital sovereignty. A new key concept of digital policy in Germany and Europe.

³⁴ 郭江兴,邹宏,张帆,等.数字主权与国家安全若干问题研究[J].国家安全研究,2023,(03):74-90+161.

³⁵ Georg Glasze, et al. (2023) Contested Spatialities of Digital Sovereignty, *Geopolitics*, 28:2, 919-958.

³⁶ Georg Glasze, et al. (2023) Contested Spatialities of Digital Sovereignty, *Geopolitics*, 28:2, 919-958.

³⁷ Georg Glasze, et al. (2023) Contested Spatialities of Digital Sovereignty, *Geopolitics*, 28:2, 919-958.

盟创制了一系列新型个人权利——被遗忘权（删除权）、数据携带权、数据更正权、数据限制处理权等保护性权利，以增强个人对数据的控制。使数据主体可以限制或影响数据控制者的处理行为。这些权利高度规范并限制了数字跨国公司使用欧盟个人数据的行为。其次，欧盟创设了数据跨境流动监管制度，构成数据安全保护的一个重要维度。数据跨境流动规则的价值基础是数据主权。³⁸GDPR 对数据跨境流动采取的“原则禁止、例外允许”的规制模式。这一模式延续自此前 1995 年的《数据保护指令》，在此基础上规制程度适度放松。GDPR 本身并未规定何为传输，从第 4 条第 2 项“处理”概念的“通过传播披露”可以推断，“传输”意为对个人数据的任何形式的披露。GDPR 建构了数据跨境流动规则的“三维”体系：第 44 条确立了数据传输的基本原则，禁止一般性的数据跨境传输；第 45 条、第 46 条和第 49 条分别设置三项例外情形：针对特定国家“整体适用”的充分性认定规则、针对特定商事主体“个别适用”的标准数据保护条款规则、以及数据主体明确同意等其他例外情形。GDPR 第 44 条作为一项一般性禁止规定设定了数据保护最低永久标准，实现了个人数据保护水平的地域扩张。第 44 条还为数据出境建构了合法性审查框架：第一阶段审查只针对传输本身的合法性，这一全面审查不涉及第三国，第二阶段审查才重点考察数据跨境传输是否满足特别要求，许可是否可以扩展到 GDPR 地域适用范围之外。

在监管目标上，欧盟虽然是数字主权的基本战略意图是独立自主，但这一目标也有了扩张之势，试图将欧盟数据规则全球化。欧盟数字主权战略的核心之一是数据主权。虽然欧盟数字经济未能在全球领先，但欧盟制定的系列数据保护法却领先于世界，引发“布鲁塞尔效应”。欧盟被视为隐私和数据保护方面的标准制定者。多国已参考 GDPR 条款，将其纳入国家立法，很多跨国公司也选择采用 GDPR 作为其全球运营标准。欧盟机构在促进遵守严格的欧盟隐私标准的技术解决方案的开发和使用方面发挥了重要作用。

欧盟基于特有的隐私文化，以个人数据保护为由限制数据跨境流动，具有极强的布鲁塞尔效应。20 世纪 70 年代开始，欧盟率先在成员国内部限制数据跨境流动；80 年代开始推进《OECD 指南》和《108 号公约》；90 年代推动《服务贸易总协定》（GATS），通过隐私保护例外条款作为合法限制国际贸易的措施；1995 年通过《个人数据保护指令》；2009 年开始改革个人数据保护框架，最终于 2016 年通过 GDPR。GDPR 强化了数据跨境流动机制，借助域外适用和违法处罚将其推及全球。GDPR 规定了具体的用以评估数据跨境转移的工具：充分性认定、标准合同条款、可约束性公司规则、行为守则、认证机制、特别合同条款、克减条款等。充分性认定是 GDPR 数据跨境传输确定性和便利度最高的机制，所考虑的主要因素包括他国/国际组织的法治和人权保护程度、数据保护监管机构等。然

³⁸ 吴玄.数据主权视野下个人信息跨境规则的建构[J].清华法学,2021,15(03):74-91.

而在实践中欧盟委员会对充分性认定的评估自由裁量空间较大,更多是一种政治或政策考量。³⁹

欧盟一边通过 GDPR 扩大其立法管辖权,另一边还通过欧洲法院扩张解释 GDPR。从类案来看,自 2003 年-2019 年,欧洲法院以扩张解释欧盟数据法为基本立场。从个案来看,自 2015 年 Schrems I 案件后,欧洲法院对数据跨境流动的司法裁判重心转变,开始聚焦司法审查,效力审查机制逐步形成。2020 年 Schrems II 案具有突破意义,正式确立了欧洲法院的个案实质审查机制。欧洲法院的司法审查升级为对第三国数据保护水平的实质监管。与此同时,欧盟对数字跨国平台也加大 GDPR 处罚力度。仅从 2020 年-2021 年 GDPR 最高的十笔罚款中的前六名针对的就是跨国平台企业。⁴⁰欧盟凭借着高效严苛的数据保护制度和保护标准影响全球,通过市场机制将其法律法规外部化,产生单边监管全球化效应,将欧盟标准转化为全球标准。

2024 年 1 月正式生效的《数据法案》(DA)与 GDPR 成为欧盟数据战略的两个重要法律支柱,构成平衡协调经济发展与个人权利保护、内松外严的数据治理框架。在对外数据政策上,欧盟依然采取的是 GDPR 严格的数据跨境流动限制和保护性域外管辖措施。在对内数据政策上,欧盟旨在实现内部数据市场的协调和统一,促进数据共享,驱动数字经济发展。有关数据内容治理方面,DA 进一步加强了互联产品数据收集与共享,设置加强的可移植性权利,以便个人和企业更好的控制自身数据方便数据流动。与 GDPR 相比,DA 主要针对非个人数据,促进非个人数据的流动和共享,以激发其创新价值,构建境内数据自由流动市场。DA 是对数据控制者和处理者的激励和支持,以享受数据共享的权利和利益。DA 着重强调的是数字经济参与者之间数据价值分配的公平性,为数字信任筑牢基础,促进公共部门数据开放,鼓励数据流动和数据贸易,开放竞争。DA 和 GDPR 之间具有一致性和互补性,都是为了实现欧洲的数字化转型和创新发展。从侧重个人数据保护的 GDPR,到侧重数据应用的 DGA 和 DA 则,欧盟已建立一套系统的数据治理体系,分别针对个人数据和非个人数据,制定了统一的数据治理框架,形成了“内松外严”的数据战略格局。

(二) 美国的霸权叙事

近年美国数据监管模式发生了很大转向。在过去,美国将本国数字跨国公司打造成美国在网络空间的国家利益载体,其数据治理工具借助本国数字跨国公司走向全球,尽可能多地掌握控制全球数据。⁴¹因此,一直以来美国推崇的是贸易

³⁹ 叶开儒.数据跨境流动规制中的“长臂管辖”——对欧盟 GDPR 的原旨主义考察[J].法学评论,2020,38(01):106-117.

金晶.欧盟《一般数据保护条例》:演进、要点与疑义[J].欧洲研究,2018,36(04):1-26.

⁴⁰ GDPR 执法统计盘点、重点案例分析及应对方案 https://www.yenlex.com/news/content_246.html?lang=en-us

⁴¹ 洪延青.数据竞争的美欧战略立场及中国因应——基于国内立法与经贸协定谈判双重视角[J].国际法研

自由，并不以国家安全为导向，力图推动美国数字跨国公司的全球发展。在自由主义旗帜下美国推动经济全球化，强调商品和服务贸易给美国和全球带来的好处。这种模式不会仅仅因为其他的威权政治制度而对它们施加限制。美国依赖数字自由贸易，强调数字市场的扩张。美国积极将自由贸易纳入各类国际贸易协定，这也促使美国在 2001 年支持中国加入 WTO。数字自由贸易的主张之下，美国的数据政策也以支持数据跨境自由流动为主。美国积极推动建立亚太经合组织 APEC 数据跨境流动隐私框架和《跨境隐私规则》（CBPR），以及《跨太平洋伙伴关系协定》（CPTPP）、《跨大西洋贸易与投资伙伴关系协定》（TTIP）、《美国-墨西哥-加拿大协定》（USMCA）等国际经贸协定的数据跨境自由流动规则。在 WTO 框架下，美国也坚持主张电子商务规则应允许数据跨境自由流动，并试图禁止其他国家数据本地化和源代码审查等要求。

然而，美国对外立场上开始有所改变，其数据政策也日渐限缩，发生转向，强调维护国家安全和数据安全。2023 年 10 月，在瑞士日内瓦 WTO 电子商务联合声明倡议会议期间，美国贸易代表凯瑟琳申明美国在 WTO 电子商务规则谈判中放弃长久以来坚持的跨境数据自由流动主张，为美国国会加强对大型数字跨国公司的监管腾出空间。⁴²这一主张与美国近年国内数据限制措施的强化相一致，意味着美国在国内和国际数据流动政策趋向同步。受到世界局势和政治经济形势影响，美国综合实力下降，如今美国则愈发关注国家安全，这将美国自由主义市场驱动的监管方式导向以政府监管为主导的安全模式。2024 年 1 月 17 日，美国联邦贸易委员会 FTC 宣布加入隐私执法全球合作安排 Global Cape。同年 2 月，总统宣布有关限制和禁止数据流动的第 14771 号行政令，加强对华数据限制和数据安全审查。美国国际立场的转变并非突然，其国内出口管制、外商投资审查、经济制裁、ICTS 监管、长臂管辖等机制早已在潜移默化地推进数据流动限制措施。美国数据治理国际规则立场的转变恰好折射出美国数据监管思路和方法的全方位转向，主导策略从自由贸易调整为国家安全。

安全模式下，美国法律政策和制度呈现出国家安全泛化特征和趋势，数据监管也不可避免受到影响。国家安全多维度的在各个领域渗透。“国家安全泛化”指的是“一国对于众多非安全化议题进行安全化建构，使安全的边界、范畴与维护安全的手段极度扩张，进而导致国家安全变得无所不包”。⁴³其本质是“国家在不断变化的国际安全环境中对绝对安全和权力优势的无限追求”。美国数据监管强调对华数据限制正是国家安全泛化的表现。

究,2021,(06):69-81.

⁴² David Lawder.US drops digital trade demands at WTO to allow room for stronger tech regulation[EB/OL].[2024-5-3].
<https://www.reuters.com/world/us/us-drops-digital-trade-demands-wto-allow-room-stronger-tech-regulation-2023-10-25/>.

⁴³ 卢凌云,崔磊.霸权衰落与“泛安全化”:基于美国的个案分析[J].国际安全研究,2024,42(03):76-97+159.

首先表现为美国国家安全概念范畴的扩张。美国国际经济治理中国国家安全已然呈现出不同纬度的表现形态：从政策到工具、从例外到原则、从传统领域到新兴领域，国家安全的内涵和外延都在扩张。⁴⁴非传统安全被不断纳入国家安全范畴。数据领域也是如此，包括个人隐私、市场稳定、数据安全等都可以成为美国国家安全审查的理由。在第 14771 号行政令中，政府对敏感数据的定义非常广，对数据的数量、种类、使用方式等都有所限制。

其次，美国对华数据限制的国家安全治理机制也在极速扩张。从立法司法的长臂管辖到执法时的法律域外适用，限制数据流动工具不断被制定。特别是行政部门的国家安全审查措施正在成为被广泛运用的监管工具。

美国数据监管安全泛化的最终目的是为了在内部统一数据市场，允许数据内部自由流动，对外出于防范外国政府获取数据的目的，严格把控数据走向，限制数字企业境外业务开展及伴随的数据跨境流动，构建数据“宽进严出”体系。美国为稳固自身实力，试图加大对本国数据范围和安全程度的掌控，将有竞争威胁的敌对国家特别是中国排除在数据跨境流动的圈子之外。⁴⁵

美国当前数据政策可以分为四种类型：数字自由贸易模式、阻止对手模式、隐私保护模式和数据盟友模式。⁴⁶四种模式实质上可以分为合作和对抗两类。美国对于盟友强调数据自由流动，对于对手则是限制数据跨境传输。美国数据政策转向安全模式后不仅只是主张盟友间数据安全流动，更是强调数据单向流动。通过美国《澄清境外数据的合法使用法案》（CLOUD Act）一面扩大美国政府获取海外数据的能力，一面防止其他国家政府获取美国境内数据。实质上美国所谓的数据安全指的是数据单向流动，只能允许数据流向美国而非流出美国，实现数据单向脱钩。

从监管目标来看，美国数据规制是意识形态和地缘政治竞争背景下对数据霸权的维系和护持。一方面，虽然美国近年数据政策才开始向安全监管模式转变动向，政府开始着重针对中国进行数据监管，在法律政策中明确将中国列为对手，但其脉络可以追溯至十年之前。美国数据监管一直是意识形态和地缘政治竞争导向的结果。另一方面，尽管美国法律和各种行政措施对国家安全概念进行了模糊处理，但处处都可体现美国一以贯之的国家安全思想——一种例外主义、二元对立、霸权护持的国家安全观。⁴⁷例外主义即美国处处认为本国优越于其他国家和民族的唯我独尊的自负意识，还认为自己带有传播自由和社会正义的天赋使命。二元对立观则是来自西方哲学二元对立的基调。这也令美国喜欢寻找敌人，零和博弈，以此构建自身正面形象。最终例外主义、二元对立导向美国国家安全的根

⁴⁴ 彭阳.国际经济治理中的国家安全泛化：法理剖析与中国应对[J].国际法研究,2022(05):87-107.

⁴⁵ 洪延青.数据跨境流动的规则碎片化及中国应对[J].行政法学研究,2022(04):61-72.

⁴⁶ Sacks, Samm and Swire, Peter, A Framework for Assessing U.S. Data Policy Toward China (October 13, 2023). Available at SSRN: <https://ssrn.com/abstract=4601794> or <http://dx.doi.org/10.2139/ssrn.4601794>

⁴⁷ 郭超.美国国家安全思想：理念源流与核心要义[J].国家安全研究,2023,(05):47-67+153.

本：霸权主义。美国国家安全的最终目的是为了维护自身霸权地位。霸权在不同阶段表现为各种不同的法律政策手段和安全、自由等价值口号。在美国霸权衰落的现阶段表现为国家安全泛化。⁴⁸在法律政策上的一个表现就是外严内宽。⁴⁹比如通过区域性协定数据流动框架和国内管制云法案等构建数据“宽进严出”体系，⁵⁰以实现数据霸权。美国一边允许政府可以获取境外数据，一边严格控制服务供应商向外国政府披露境内数据。

伴随着经济全球下行，美国霸权衰落，自特朗普政府时期开始，美国国家安全导向开始逐步显现，针对他国政府和企业的制裁日渐频繁。美国对华数据限制措施即是这一时期的产物，各项制裁以国家安全为名，实则延续了美国自由贸易时期的经济霸权话语，是为了维护自身技术与数字霸权。虽然美国近期出台的行政令和法案都在强调数据限制只是针对几个国家某些敏感领域，维护开放互联网的承诺不会被放弃，但这些都只是美国维护霸权地位的手段和形式。美国总体上一直企图沿着规范和执法两个层面全面推进全球数据治理主导权，⁵¹只是名义上的旗号从过去经济上升期的自由民主转向经济下行期的国家安全。国家安全的话语形式为美国经济制裁的霸权行动提供了合法性依据。美国安全理念表面以国家安全和利益为核心，实际是延续了一直以来的霸权叙事话语，是霸权收缩时期的泛安全化表现，是由扩张性霸权向收缩性霸权的战略性转向。美国霸权虽然衰落，但霸权色彩和话语叙事从未离场，在既有霸权体系下继续通过“制度-价值”框架构建数据霸权。⁵²当前美国为维护数据霸权地位展开的国际数字地缘战略包括三个方面：一是将以中国为代表的国家列为明确的竞争对手，动用一切手段在数字领域对其进行打压；二是以意识形态为导向，广泛建立数字领域的盟友；三是以美国为中心重塑全球数字技术供应链体系。亚洲逐渐成为全球数字地缘政治的中心地带，在算法、算力、数据等数字技术的核心指标中占据优势地位。美国意图通过对亚洲的影响形成对中国的数字技术包围和封锁，重夺全球霸权领导地位。

（三）中国的安全叙事

中国国家治理体系一直被认为是以国家安全为导向。但在不同时期中国国家安全观也在发展演变。自建国初期到 20 世纪 70 年代末以军事和政治安全为核心，到在 20 世纪 70 年代末至二十世纪初从传统安全观向非传统安全观转变，再到进入二十世纪后逐渐形成的新时代的总体国家安全观，如今中国数字治理过程在总体国家安全观的指导下兼收并蓄，兼顾安全与发展双重目标。数字安全是数

⁴⁸ 卢凌宇,崔磊.霸权衰落与“泛安全化”:基于美国的个案分析[J].国际安全研究,2024,42(03):76-97+159.

⁴⁹ 谢宇.美国法律如何防控外国威胁基于美国国家安全法律的考察[J].中外法学,2023,35(02):541-559.

⁵⁰ 洪延青.数据竞争的美欧战略立场及中国因应——基于国内立法与经贸协定谈判双重视角[J].国际法研究,2021,(06):69-81.

⁵¹ 吴沈括.透视美国放弃在 WTO 的跨境数据自由流动主张[EB/OL].[2024-5-3].

https://www.sohu.com/a/735055900_120287836?scm=10001.325_13-109000.0.10140.5_32&spm=smcpc.channel_248.block3_308_NDdFbm_1_fd.1.1713581928829EySPAsi_324.

⁵² 冯硕.美国数据霸权的建构与法律应对[J].太平洋学报,2023,31(08):45-58.

字发展的前提。数据安全又是数字安全的重中之重和应有之义。中国数据治理理念从强调数据安全保护逐渐转向数据安全有序发展。

在互联网经济发展初期，中国数据监管机构对科技公司采取的立场并不强硬。数字企业经历了早期的“非法兴起”过程，成长为大型数字跨国公司。针对违法行为，中国数据主管部门在执法过程中并未采取激进立场。监管机构并未对违反数据和隐私规定的公司进行法律处罚，而是主要采用非正式手段，例如谴责以及点名批评那些屡次警告后仍未改正的科技公司。例如，在支付宝、百度和字节跳动曝光多起数据丑闻后，工信部分别请这三家公司的代表讨论其数据合规性问题，并要求其公司进行整改。但并未对其实施任何处罚。⁵³

2019 年，四大中央监管机构成立了联合应用治理工作组，以规范中国应用的数据合规性。⁵⁴工作组颁布了两项指南，一项是关于实施措施，另一项是关于合规。随后，市场监管总局对 2300 多个应用程序进行了评估，并跟进执法。在这些检查中，这些机构大多与企业进行行政约谈，要求整改，然后有选择地公布被视为不合规的应用程序列表。只有在一定期限后仍未整改的应用程序才会受到进一步处罚。在特殊情况下，该机构在收到多次警告后会删除未能整改数据实践的应用程序。在这些机构中，只有公安机关和市场监管总局在执法行动中处以罚款。这些罚款很低——平均每案约 13000 元人民币。⁵⁵根据市场监管总局披露的 10 起典型案件，没有一起涉及大型网络平台。应用程序治理联合工作组在 2020 年继续开展工作。⁵⁶然而，执法速度明显放缓，工作组仅发布了一份公告，公开批评 35 款应用程序。⁵⁷

然而，情况在 2021 年迅速发生了变化，法院和行政机关都变得更加积极地制定严格的法规并进行积极的执法。当早起的数字企业逐渐成长为数字巨头，大型数字跨国公司，开始具有市场垄断地位，威胁国家监管时，政府加大执法力度。自 2020 年底以来，中国数据管理部门迅速适应了最高领导层的新政策举措，制定了严格的立法，同时对中国大型数字跨国公司采取了严格的执法行动。基于以国家安全为重心的传统，尽管企业成本增加，中国在数据安全方面采取保守立场，严格控制数据流出。事实上，《个人信息保护法》对符合中国签署的国际条约或协议的海外数据传输作出了例外规定，为未来与其他国家的谈判铺平了道路。⁵⁸

⁵³ 你的个人信息安全吗？工信部约谈百度蚂蚁金服今日头条，新浪 (1. 12, 2018), <http://tech.sina.com.cn/roll/2018-01-12/doc-ifyqqciz 5880474.shtml>.

⁵⁴ Cyberspace Administration of China (May 2019), http://www.cac.gov.cn/2020-05/26/c_1592036763304447.htm.

⁵⁵ 市场监管总局召开“守护消费”暨打击侵害消费者个人信息违法行为专项执法行动专题新闻发布会 (11. 18, 2019), https://www.samr.gov.cn/xw/xwfbt/201911/t20191118_308613.html.

⁵⁶ 2020 年 App 违法违规收集使用个人信息治理工作启动会在京召开，中国网信办？网信办？ (8. 25, 2020), http://www.cac.gov.cn/2020-07/25/c_1597240741055830.htm.

⁵⁷ 关于 35 款 App 存在个人信息收集使用问题的通告 2020, 11. 17, 2020, http://www.cac.gov.cn/2020-11/17/c_1607178245870454.htm.

⁵⁸ 个人信息保护法 38 条

在其他司法管辖区，双边协议或区域贸易协定一直是建立跨境数据流互操作机制的主要跨国合作形式。例如，欧盟推动双边合作，承认 15 个国家的个人数据保护水平已达到与欧盟相当的保护水平。⁵⁹美国利用亚太经合组织 (APEC) 跨境隐私规则 (CBPR)，这是其成员之间建立的自愿跨境隐私机制，以扩大自由数据流的范围。⁶⁰美国也在进一步扩大 CBPR：于 2022 年 4 月与加拿大、日本、韩国、菲律宾、新加坡和中国台北启动了全球 CBPR 论坛。中国虽然加入了 RCEP，并已申请加入 DEPA 和 CPTPP，然而这些区域贸易协定留下充足空间适用例外条款，为数据本地化留下了很大的漏洞。⁶¹因此，中国在进行贸易谈判和伙伴关系的同时，仍将对数据流保持相当大的控制。

为了进一步维护国家安全，中国开始将数据监管作为一种灵活的工具，以对抗美国主导的针对中国企业的“技术封锁”。2023 年 4 月，网信办以维护关键信息基础设施和国家安全的供应链安全为由，对美光科技启动了网络安全审查。⁶²这项调查被认为是对美国试图限制其获取尖端芯片的报复性举措。2017 年，美光指控中国国有存储器工厂制造商福建金华集成电路有限公司窃取其商业机密。⁶³这导致美国司法部起诉该公司和三名员工，随后特朗普政府实施制裁，切断了该公司与美国供应商的联系。⁶⁴某种程度上，中国严格的数据安全措施已演变成一种强有力的谈判工具，以制衡美国及其盟友对中国的技术限制。⁶⁵

三、数字主权实践的反思

各国数据监管模式体现出不同的数字主权叙事路径。虽然话语叙事不同，但都呈现出安全化的演进趋势。虽然数字主权不同于传统威斯特伐利亚式主权，但是各国实践仍然没有脱离传统主权观的束缚，采取内外有别、严格控制的规制思路，无法实现新的数字主权目标和意义。

（一）主权的安全化演绎

纵观美、欧、中三种数据监管实践，数据监管愈发朝向安全化监管模式发展。

⁵⁹ European Commission, Adequacy Decisions, https://commission.europa.eu/law/law-topic/data-protection/international-dimensions-data-protection/adequacy-decisions_en.

⁶⁰ Anupam Chander & Paul M. Schwartz, Privacy and/or Trade, 90 U. Chi. L. Rev. 49, 33–35 (2023).

⁶¹ Yik Chan Chin & Jingwu Zhao, Governing Cross-Border Data Flows: International Trade Agreements and Their Limits, 11(4) Laws 63, 9 (2022).

⁶² Chang Che & John Liu, China Strikes Back at Micron Technology Even as It Signals Openness, N.Y. Times (Apr. 4, 2023), <https://www.nytimes.com/2023/04/04/business/micron-china-investigation.html>.

⁶³ Lilian Zhang, Why China Launched a Cybersecurity Review into US Memory Chip Maker Micron Technology and What Could Happen Next, South China Morning Post (Apr. 2, 2023), <https://www.scmp.com/tech/tech-war/article/3215742/why-china-launched-cybersecurity-review-us-memory-chip-maker-micron-technology-and-what-could-happen>.

⁶⁴

US Department of Justice, PRC State-Owned Company, Taiwan Company, and Three Individuals Charged With Economic Espionage (Nov. 1, 2018), <https://www.justice.gov/opa/pr/prc-state-owned-company-taiwan-company-and-three-individuals-charged-economic-espionage>; David Lawder, U.S. Restricts Exports to Chinese Semiconductor Firm Fujian Jinhua, Reuters (Oct. 30, 2018), <https://www.reuters.com/article/us-usa-trade-china-semiconductors-idUSKCN1N328E>.

⁶⁵ China Starts “Surgical” Retaliation against Foreign Companies after US-led Tech Blockade, Fin. Times (Apr. 16, 2023), <https://www.ft.com/content/fc2038d2-3e25-4a3f-b8ca-0ceb5532a1f3?shareType=nongift>.

世界主要国家和地区对数字主权的认识主要基于不同层面的安全考量,促使全球形成割裂的数据流动秩序。主权国家陷入“二选一”的政策权衡困境:选择促进数据跨境流动或是加强数据本地化保护。这导致各国采取了差异化较大甚至完全相反的数据治理机制,引入不同模式的数据法律规则体系,或是采取保守的数据保护主义政策,或是采取激进的数据自由流动政策。然而,为防范国家安全风险,大国地缘政治竞争加剧,数据治理机制也普遍趋向安全化发展。数字社会国家安全问题重点体现在数字安全之上,其中尤以数据安全为重,这是数字技术的发展、国际环境变化、国家利益诉求等方面因素交织演变的结果。相较于其他领域的安全问题,数据安全更具复杂性和动态性。

首先,数据监管整体朝向安全模式演进。数据安全上,各国规制手段扩大,开启综合立法模式,通过多层次全方位的战略、法律、政策构造新的数据安全体系,包含了“安全——控制——利用”三个层次。⁶⁶ 欧盟、美国和中国在不同时间和阶段侧重点不同。从一开始欧盟重控制、美国重利用、中国重安全全面向侧重安全的模式演进。各国开始在内部统一数据市场,允许数据内部自由流动,对外出于防范外国政府获取数据的目的,严格把控数据走向,限制数字企业境外业务开展及伴随的数据跨境流动,构建数据“宽进严出”体系。特别是美欧为稳固自身实力,试图加大对本国数据范围和安全程度的掌控,将有竞争威胁的敌对国家排除在数据跨境流动的圈子之外。⁶⁷

其次,数据安全规制呈现出利益诉求多样化特征。总体而言,美、欧、中基于不同层面的安全欲求,数据安全规制模式也呈现出不同样态的安全理念和行动逻辑。虽然总体上各地数字安全规制的普遍呈现出外严内松趋势,但是从法律政策来看,其背后的规制逻辑和规制理念差异较大。欧盟规制理念以个人权利保护为中心,数据安全、网络安全等都围绕个人基本权利构建。监管和创新是欧盟数据政策的治理目标,然而二者之间存在很大的张力。监管过于严格可能增加市场主体的守法成本,抑制创新。过少的监管则可能放大市场主体的行为风险。在主权上则体现为经济主权和技术主权的紧张关系。欧盟规制模式试图弥合数字主权在不同层面的利益需求。美国规制理念表面以国家安全和利益为核心,实际是延续了以往的霸权叙事话语,是霸权收缩的泛安全化表征,是由扩张性霸权向收缩性霸权的战略性转向。中国规制理念则由传统国家安全观向新安全观转向,致力于构建安全与发展并行的总体安全观和数字安全规制思路。

最后,地缘政治竞争愈发激烈,国家安全规制渐趋泛化。国家安全具有多重面相,不仅是一种理念,也可以作为政策和工具。国家安全不再只是规制领域,而成为规制手段。各国数字规则在各个领域都开始转向安全模式。从数据、供应

⁶⁶ 洪延青.我国数据安全法的体系逻辑与实施优化[J].法学杂志,2023,44(02):38-53.

⁶⁷ 洪延青.数据跨境流动的规则碎片化及中国应对[J].行政法学研究,2022(04):61-72.

链、半导体、新能源...各个领域都在安全化。在数据监管领域，国际法上，安全例外原则的适用不断扩张。国内法上各国以数据安全为依据的限制和制裁也愈发增加。

（二）传统主权观的影响

主权是近代民族国家产生后形成的。自博丹提出近代意义的主权后，《威斯特伐利亚和约》确立了国家的主权原则。国家主权原则的核心是国家对领土内事物的决策权和控制权。因此，国家主权首先强调领土边界，国家的领土边界清晰确定，具有稳定性，主权范围由边界框定。地理界线决定了主权具有划分内外、甚至决定敌我的功能。其次，主权强调国家对内事物的控制权和对外部事物的自主权。国家拥有对领土和人民的绝对控制权。在领土主权观下，国家在领土范围拥有最高的权力，包括立法、司法和行政权。传统主权观主张国家主权的绝对性。

但这只是理论意义上的，实践中受到国际法、国际关系等多重因素影响。国家之间相互依存，需要国际合作。随着全球化的发展，国家之间的联系更加紧密，也使得主权受到更多限制。斯蒂芬·克拉斯纳将主权分为四类：国内主权、威斯特伐利亚主权、相互依赖的主权以及国际法制主权，并认为四个主权形态并非总是同时存在。数字社会网络空间的界限越发模糊，国家对相互依赖主权的依存度越来越高。然而，越来越多的国家承认的仍然是传统主权观的领土原则。《塔林网络战国际法手册》对网络空间中主权的界定表明网络主权是主权原则在网络空间的延续和运用。“主权意味着一国可以控制入境，且依据条约和习惯国际法在其领土上享有管辖的排他性权力和最高权威。”《网络行动国际法塔林手册 2.0 版》更是明确提出国家主权原则适用于网络空间。“一国在遵守其国际法义务的前提下，对其领土内的网络基础设施、人员和网络活动享有主权权威。”中国、俄罗斯等国家认可的仍然是传统主权观。2023 年世界互联网大会中国研究机构发表的《网络主权：理论与实践 4.0 版》中指明“网络主权是国家主权在网络空间的自然延伸。是一国基于国家主权对本国境内的网络设施、网络主体、网络行为及相关网络数据和信息等所享有的对内最高权和对外独立权。”

从数据监管实践来看，传统主权观仍然深刻影响着国家行动。在各国数据监管政策中都可以看到对数据流动的严格管控。即使是被视为数据自由主义代表的美国，其数据政策也开始逐渐限缩和保守。2023 年 10 月，美国贸易代表在 WTO 电子商务联合声明倡议会议上公开放弃数据跨境自由流动的立场要求。与此同时，2024 年美国出台第 14117 号行政令，首次限制数据向特定国家出境。这标志着美国数据治理政策的国家安全主义转向。

传统主权观引导之下，各国数据治理体系以国家安全为核心，重在风险控制，而忽视了在全球数字经济发展背景下国家之间、国家和企业之间相互关联，彼此共生的关系。国家安全正在成为一项规制手段愈发频繁地被应用到数据规制中，

特别是针对他国的大型数字企业。国家安全的武器化和泛化趋势严重损害了数据治理体系的稳定性和可预期性。传统主权观下，主权与安全一体两面。传统主权治理思路 and 理念将一国决策与他国、其他主体的决策割裂，但数据的流动性和跨国性、数据治理权力的分散性以及数据治理政策的外部性、关联性要求拓展传统主权理论，将重心转移到相互合作上，以此兼顾风险管控和经济发展。

（二）数字跨国公司成为关键节点

从各国数据监管实践可以看出，大型数字跨国公司成为国家主张数字主权的关键。由于数字领域主要以数字跨国公司主导，因此数字主权的主张通常发生在企业层面而不是政府之间。即使是政府主张的数字主权，同样更有可能针对外国数字企业。数字企业负责数字经济的各个要素，还负责提供一系列数字产品和服务。大部分跨国互联网治理主要由持有和管理我们数据的数字跨国公司负责。数字跨国公司越来越多地决定着互联网规则，以及在关键领域如何解释和应用这些规则。

大型数字跨国公司凭借其规模、技术和影响力，在地缘政治上愈发重要。数字空间的角力事实上形成了“国家-数字巨头-国家”的复杂三角关系。大型数字跨国公司成因此成为国家主张数字主权的重要切入点。但是这一三角关系引发了一个悖论：国家一方面试图加强对数字大型企业的监管，强化隐私保护和国家安全，遏制其规则制定权，另一方面国家又需要借助科技巨头的技术权力，强化对数据、算法等技术要素的管控，借助数字企业对内可以构建国家数字体系，对外可以进行国家数字竞争和安全博弈。随着国家地缘政治和意识形态竞争导致的国家安全监管加剧，各国对数字跨国公司规制呈现集中化趋势。没有大型数字跨国企业的国家和地区，比如欧盟，倾向于限制他国大型数字企业，以遏制他国潜在的数字技术威胁。美国、中国等拥有大型数字企业的国家，则开始围绕数字技术展开新一轮博弈。技术强国美国更是试图与本国大型数字企业深度合作，借助数字跨国公司的技术、经济实力建立全球霸权。

数字跨国公司由此在国家数字主权博弈过程中扮演不同的角色。数字跨国公司作为数字安全治理主体有其独立的利益主张和价值诉求，因而可能与各国政府、公民个人等其他治理主体之间存在利益冲突。与此同时，数字跨国公司也作为国家的规制对象、打击工具、和竞争战场，容易成为数字地缘政治斗争的工具和牺牲品。如何理解数字跨国公司在不同领域的角色以及其与母国和东道国之间的关系是数字主权研究和实践的难点和痛点。

四、多元化数字主权的理论探索和机制创新

传统威斯特伐利亚主权观无法适应数字社会的需要，数字主权应当探索更具包容性、关联性的面向。本文认为国家一方面可以发展合作式数字主权以指导数据治理实践，另一方面，可以创新数据治理机制以丰富数字主权理论内涵。

（一）发展合作式数字主权

从数据监管来看，数据具有流动特性，只有流动的数据才能具有价值。这决定了数字主权概念超越了传统主权“边界”“控制”的理念。数字主权需要考虑到数据跨境流动的客观需要，构建新的理论面向以指引数据治理实践。数字主权应当更加突出合作、包容、相互关联的理念。一方面，管理、控制数据的权力和能力已不再是国家独有，数字跨国企业事实上拥有更大的数据管控权，对数据跨境治理有着更重要的影响。另一方面，各国数据治理政策紧密相连，一国治理目标能否实现有赖于其他国家数据治理政策是否配合。也就是说，国家数据治理成效依赖于其他国家和企业。因此数据治理具有典型的关联性和外部性特征。国家的数据监管政策也应当将其他国家、企业的数据治理需求纳入考量。

数字主权也由此应当充分重视国家之间的相互影响，放弃内外区分的敌对思维，以及数据保护、数据流动二者对立的割裂思维。通过促进交流合作，风险管控，实现数据保护和数据流动的双重目标。这一合作式主权观在实践中也有体现。2024年9月联合国未来峰会通过的《全球数字契约》提出多层次推动全球数字合作，共同构建能够灵活应对数字时代挑战的治理体系。⁶⁸中国更是提出一系列数字合作倡议，联合有关国家发起《全球数据安全倡议》《全球人工智能治理倡议》《“一带一路”数字经济国际合作倡议》《携手构建网络空间命运共同体行动倡议》《金砖国家数字经济伙伴关系框架》《金砖国家制造业数字化转型合作倡议》。中国合作式主权观的根本依据是构建网络命运共同体。在第二届世界互联网大会期间，中国创造性地提出“构建网络空间命运共同体”。⁶⁹“中国愿同世界各国一道，把握信息革命历史机遇，培育创新发展新动能，开创数字合作新局面，打造网络安全新格局，构建网络空间命运共同体，携手创造人类更加美好的未来。”中国为数字时代国家主权概念的理论变革提供了新的方案，与数字空间的本质属性相契合，推动数字主权向更具包容合作性的方向重塑。

（二）创新数据治理机制

在理论创新之外，国家也可从实践出发推动数字主权提升现实适应性。数据需要跨境流动以释放其规模效应和范围效应，各国数据治理也应当更加关注于全球层面国际合作治理机制，以此引导助推数据跨境流动。在具体的数据监管合作机制构建上，OECD曾总结了11项合作方式，包括制度一致化、跨国监管伙伴关系、特定领域合作协议、国际组织领导的理监管合作、制度互认、跨国网路、软法之治等。⁷⁰各国主要的机制创新可以总结为监管合作、制度一致化、软法之

⁶⁸ 2024年9月联合国未来峰会《全球数字契约》。

⁶⁹ 习近平在第二届世界互联网大会开幕式上的讲话，2015.12.16。

https://www.gov.cn/xinwen/2015-12/16/content_5024712.htm

⁷⁰ OCED, International Regulatory Co-operation: Addressing Global Challenges, OECD Publishing, 2013, p.4.

治三个方面。⁷¹

首先,监管合作指的是国家以开放态度加强与其他国家和非国家行为体的广泛合作,特别是大型数字跨国公司。数字跨国公司愈益呈现出多层次、分布式的组织化运作方式,⁷²涉及“企业-国家-用户”三方主体间的复杂多边关系,整个数字生态嵌入不同层次的关系向度中。国家可以通过与数字跨国公司的监管合作,治理整个数字生态。欧美《安全港协议》是国家和企业“公-私”监管合作的代表。美国数字跨国公司可以选择接受欧盟独立结构的监管或者美国FTC的监督。《安全港协议》废止后,《隐私盾协议》仍然继承了这一“公-私”合作体系。

其次,制度一致化可以为各国数据治理合作提供便利。制度一致化是指各国数据治理制度的趋同,在实践中,欧盟数据规则产生的制度一致化效果是典型代表。一方面,欧盟数据流动的“充分性认定”原则,使加拿大、澳大利亚、日本、阿根廷等四十多个国家遵循了欧盟数据治理标准,实现了与欧盟制度的一致化。另一方面,欧盟数据规则的产生了布鲁塞尔效应,数字跨国公司对欧盟规则的遵守促进了欧盟规则的进一步传播。在全球数据治理领域,欧盟模式可以说是全球化最充分的一种治理范式。除了传统法律移植和欧盟模式自有的规范性力量,私人法律移植和布鲁塞尔效应也成为其全球化的重要基础。私人法律移植指的是通过私合同的法律移植。跨国公司在欧盟模式全球输出中发挥着独特作用,譬如通过自愿适用欧盟标准合同等法律工具,纳入公司全球运营,加速欧盟数据保护规则传播。布鲁塞尔效应指的是欧盟单方面监管全球市场的能力,也就是依靠市场欧盟数据保护规则得以传播到欧盟外的市场主体和监管机构。⁷³这根源于市场规模和跨国公司在全球范围内采用欧盟数据保护规则。

最后,国家可以通过理念倡议、国际标准等软性治理手段推动数据治理合作。通过与互联网工程任务组(IETF)、互联网名称与数字地址分配机构(ICANN)、联合国信息世界峰会(WSIS)等非国家行为体的合作,构建网络安全、信息安全、数据安全等领域的国际标准,直接或间接推动数据治理议题的解决。

结论

数字主权概念模糊,与国家安全概念相伴相生,在国际法中呈现不同的形式,既是一种概念、原则、也可以是一种规制工具、手段。数字主权兼具防御性与扩张性,与网络主权、技术主权、经济主权、数据主权等概念紧密相关。

从各国数据监管的动向来看数字主权,可以看到这一概念更为明显的变化和不同方向的走势。欧盟数据监管以个人数据权利为核心,旨在保障个人基本权利和维护统一数字市场,意图通过欧盟数字规则的全球化和标准化提升其数字主权

⁷¹ 贾开,赵静:“双目标”主权理论创新与数据跨境治理体系改革[J].学海,2024,(04):105-113+215.

⁷² Shoemaker, Pamela J., Timothy Vos. Gatekeeping Theory[J]. Routledge, 2009.

⁷³ Anu Bradford. The Brussels Effect: How the European Union Rules the World[M]. Oxford University Press, 2020.

能力，扩张数字主权版图。美国则是以数字霸权为核心，通过构建数字盟友圈打击不服从其强权的国家和企业，维持其既有的全球数据霸权地位，试图将霸权拓展至整个数字领域。中国数据监管以国家安全为核心，稳中求进，秉持总体国家安全观，强调数据安全秩序均衡有序，既重视自身安全，也重视共同安全。数据规制偏向保守，以维护国家经济和社会稳定。

总体而言，全球都在强化数据监管力度，以维护数字主权。主权与安全一体两面。在传统威斯特伐利亚主权观影响下，各国数据监管政策都在强调数据安全目标，忽视了数据只有在国家之间流动才能发挥其最大功效。而构建安全有效的数据流动机制一方面需要各国发展合作式数字主权观，加强各主体之间的多元合作，另一方面也需要创新数据治理机制，以进一步拓展数字主权内涵，从传统国家主权转向更具包容合作的新型数字主权。

网络空间与人工智能治理中国际法的作用：现状、原因与未来

于昊昆*

摘要：在国际法“数字化”转型背景下，国际法在网络空间与人工智能治理过程中面临着技术变革带来的挑战，同时也根据差异化的治理需求进行相应的法律回应。比较国际法在网络空间与人工智能治理过程中的作用，能够认识到两者之间存在的一定差异与共通之处。两者在现行国际法的适用性、国际法的规制模式特征上存在区别，同时也在治理偏好上存在相似的软法倾向。通过对上述三个方面进行不同程度的现状比较、原因分析，有利于为人工智能领域的国际法之未来发展提供启示和指引。网络技术与人工智能的技术特征差异引发了差异化的治理需求，进而影响了国际法的规则供给。国际法在未来应对人工智能领域技术变革过程中，应当朝着差异适配、平衡发展、硬法后进的路径发展。

关键词：网络空间、网络空间国际法、人工智能、人工智能治理、国际法

一、问题的提出：国际法治理的共同挑战

伴随着国际法的“数字化转型”，¹国际法在全球治理中面临着一系列挑战，网络技术与人工智能无疑是挑战最为明显的两个领域，在作为技术变革冲击国际法秩序的媒介方面具有较强的相似性。两者作为驱动当代科技进步的双重动力，已经深刻影响并重塑了全球经济、政治、安全和法律等多个领域。²随着互联网技术的发展与人工智能的广泛应用，二者在技术演进和社会结构变革中不仅改变了国际社会的格局，也带来了前所未有的复杂性问题。

在网络空间里，代码塑造了网络空间的架构，并构成了网络空间中的法律。³尽管网络空间在某种意义上被视作虚拟空间，但其对现实世界的影响却又是毋庸置疑的，这一点在国际法中也不例外。然而，网络空间的架构使得这些问题对国际法而言尤为特殊，削弱了国际法的有效性。首先，架构决定了网络活动跨越地域的特征，使网络活动的影响跨越多个管辖区。雅虎案等判决反映了网络空间中主权之间的竞争比以往任何一个领域都要普遍。⁴其次，架构增强了网络活动

* 于昊昆，男，河北秦皇岛人，武汉大学法学院、武汉大学国际法研究所博士研究生。联系方式：yuhakun1012@whu.edu.cn。

¹ 参见黄志雄，罗旷怡：《论国际法的“数字化转型”——兼评〈国际法的数字挑战〉白皮书》，载《武大国际法评论》2023年第3期。

² See Bhushan, Tripti, *Artificial Intelligence, Cyberspace and International Law*, 21 Indonesian Journal of International Law 281, (2024).

³ 参见劳伦斯·莱斯格：《代码 2.0:网络空间中的法律》，李旭等译，清华大学出版社 2018 年版。

⁴ 以“雅虎案”为例：雅虎作为一家美国公司在法国运营的过程中曾允许用户在网上拍卖纳粹纪念品，并触犯了法国禁止宣传纳粹的法律。然而，美国宪法第一修正案则要求保护相应的政治观点。法国、美国法律之间的差异导致了两国法院给出了截然不同的判决。其背后的本质是主权的竞争。See Superior Court of Paris, *LICRA and UEJF v. Yahoo!, Inc. and Yahoo France*, (20 November 2000)

的隐匿性，进而刺激了网络攻击的出现。爱沙尼亚网络攻击等事件构成了对国际法中使用武力与国家责任体系的巨大冲击。⁵最后，架构具有广泛性，物理空间中事物的“数字化”导致了网络活动的影响巨大。伊朗核设施攻击等事件恰恰印证了一次不起眼的网络行动将可能给国家行业乃至国家安全带来重大打击。⁶

就人工智能而言，其作为一项新兴技术同样降低了国际法的可规制性。特别是在致命性自主武器方面，一些武器已经进入了完全自主的状态，可以无需人类操控自动执行搜索和打击目标。⁷因此自主武器也被称为继火药和战争之后的第三次“战争革命”，并引发了伦理、法律等一系列问题。⁸国际法人道法需要对上述问题予以解决。在国际人权法领域，人工智能不仅引发对人权法中法律主体、人权法具体权利的质疑，同样也可能导致严重的侵犯人权行为。⁹此外，其还对国际法的居民制度、国际刑法、国际知识产权法、国际海洋法等其他具体领域带来了问题。¹⁰

毋庸置疑，网络技术与人工智能作为数字时代的产物，因其技术特征的不同，共同引发了对国际法的差异化治理需求，提出了国际法其他领域中不曾关注过的新问题。这些差异化治理需求需要法律予以回应。然而，在国际法对网络空间与人工智能治理进行回应时，回应虽有相似之处，但也有较大的差异。在国际法视阈下，网络空间国际法经过十余年的发展，已初步形成了一定的国际法律框架，¹¹尤其在国家主权¹²、网络犯罪¹³等领域，形成了若干行为规则和法律共识，虽然尚存在一些规则空白，但仍然为主权国家提供了行为模式上的指引。与之相比，人工智能国际法的进展则显得滞后且不尽完善。尽管联合国大会及一些区域性立法机构针对人工智能规制颁布了一些不具有法律拘束力的决议或效力范围有限的区域立法，¹⁴但全球范围内专门针对人工智能的国际法规范仍然寥寥无几。¹⁵

<http://www.lapres.net/yahen11.html>.

⁵ 有关爱沙尼亚网络攻击的实践涉嫌世界各地的黑客团体，关于俄罗斯支持此事的主张一直在流传，而俄罗斯则矢口否认，这体现了网络空间的归因困难问题。See *Cyber attacks against Estonia*, International Cyber Law: Interactive toolkit, (last visited 10 April 2025),

[https://cyberlaw.ccdcoe.org/wiki/Cyber_attacks_against_Estonia_\(2007\)](https://cyberlaw.ccdcoe.org/wiki/Cyber_attacks_against_Estonia_(2007)).

⁶ See *Stuxnet*, International Cyber Law: Interactive toolkit, (last visited 10 April 2025),

[https://cyberlaw.ccdcoe.org/wiki/Stuxnet_\(2010\)](https://cyberlaw.ccdcoe.org/wiki/Stuxnet_(2010)).

⁷ [美] 斯图尔特·罗素，彼得·诺维格：《人工智能：现代方法（第四版）》，张博雅等译，人民邮电出版社，第 838 页。

⁸ [美] 斯图尔特·罗素，彼得·诺维格：《人工智能：现代方法（第四版）》，张博雅等译，人民邮电出版社，第 836-839 页。

⁹ 参见赵骏，李婉贞：《人工智能对国际法的挑战及其应对》，载《浙江大学学报（哲学社会科学版）》2020 年第 2 期。

¹⁰ 参见赵骏，李婉贞：《人工智能对国际法的挑战及其应对》，载《浙江大学学报（哲学社会科学版）》2020 年第 2 期。

¹¹ 参见黄志雄：《“网络空间国际法治：中国立场、主张与对策”》，载《云南民族大学学报（哲学社会科学版）》2015 年第 4 期。

¹² UN, A/68/98, 24 June 2013, para 20; UN, A/70/174, 22 July 2015, para 27.

¹³ 在网络犯罪领域存在的国际法规则包括《布达佩斯公约》以及《联合国打击网络犯罪公约》，在一定程度上反映了主权国家打击网络犯罪的共同愿景。

¹⁴ UN, A/78/L.49, 21 March 2024; EU, Artificial Intelligence Act.

¹⁵ 罗旷怡译，黄志雄校：《国际法的数字挑战》白皮书，载《武大国际法评论》2023 年第 3 期，第 6 页。

这也或将引起人们的思考：差异与相似性的产生原因为何？国际法是如何在网络空间与人工智能领域中作用并应对技术变革的？考虑到网络空间国际法作为第二次世界大战后建立的以《联合国宪章》为核心的新的国际法体系¹⁶的一部分，其不仅反映了战后国际法发展的普遍趋势，¹⁷同时也展示了当代国际法如何在特定科技背景下塑造新的国际规则，网络空间国际法的法律发展历程或许能够为人工智能国际法的发展提供镜鉴。通过采取比较的研究方法，对这些问题进行观察，或有利于帮助我们更好的理解网络空间与人工智能领域中国际法的演进路径，并对人工智能国际法的未来发展提出展望。

有鉴于此，本文在第一部分明晰了网络空间与人工智能治理带来的共同挑战后，文章第二、三、四部分将分别从现状、原因与未来三个维度进行探讨。通过在第二部分比较“现行国际法的可适用性”“规制模式特征”以及“治理过程中的软法倾向”三方面现状特点，文章将在第三部分揭示国际法在网络空间与人工智能治理中作用差异与异同的原因，亦即“技术特征”反映“治理需求”进而影响“规则供给”的潜在脉络。最后，基于网络空间国际法的发展经验，本文将讨论上述现象可能对人工智能国际治理所带来的未来发展启示，推动国际法在应对技术变革中朝着差异适配、平衡发展、硬法后进的方向迈进。

二、国际法在网络空间与人工智能领域治理的现状

在治理现状方面，本文通过比较各自领域的现行法律、治理模式与造法偏好三个方面，认为网络空间与人工智能治理既有区别，又有联系。两者在现行法律与治理模式两方面存在较大差异，同时在造法偏好存在相同之处。申言之，现行法律上，网络空间治理更加依赖于现行国际法规则，人工智能国际治理更加依赖于创制新的规则。治理模式上，网络空间的国际治理往往倾向于对问题的事后规制，而人工智能治理则更加倾向于对问题的事前监管。网络空间与人工智能领域的治理差异致使了网络空间国际法领域已经形成了相对成体系化的国际法框架，而人工智能的国际治理仍然有限且分散。与此同时，在造法偏好上，网络空间与人工智能的治理都表现出了对软法的较大青睐。概言之，治理过程中的上述异同共同构成和塑造了网络空间与人工智能治理的现状。

（一）现行国际法的可适用性

在网络技术与人工智能对国际法提出差异化治理需求的背景下，首先遭受挑战的便是现行的国际法规则面对新技术的可适用性。在网络空间中，由于网络空间跨越地域的特点，含主权原则在内的现行国际法规则可适用性曾一度遭受争议。

¹⁶ 有关以《联合国宪章》为核心的国际法，参见黄进：《百年大变局下的国际法与国际法治》，载《交大法学》，2023年第1期，第10页。

¹⁷ 现代国际法的发展趋势包括国际法部门不断增多，区域造法活跃，国际法主体多重扩展等，参见古祖雪：《现代国际法的多样化、碎片化与有序化》，载《法学研究》，2007年第1期。

¹⁸伴随着后续网络空间国际规则讨论的逐渐展开，现行的国际法适用于网络空间这一共识才逐渐确立下来。¹⁹

在人工智能领域，亦有观点认为技术变革下国际法规则在人工智能领域的可适用性并未遭受质疑。²⁰《全球人工智能治理倡议》、《布莱切利宣言》、《关于发展包容、可持续的人工智能造福人类与地球的声明》等人工智能领域宣言都强调遵守国际法的重要性，也反映了国家在该领域遵守国际法的国际共识。据此，诸如《联合国宪章》、各类人权公约、武装冲突法等既存国际法规范为网络空间与人工智能技术的规制提供了一个很好的平台，并为解决技术挑战奠定了基础。²¹

然而，理论上的“可以适用”并不等同于实践中的“切实适用”。现行国际法的“可适用性”与其讨论其在应然层面能否适用，更应当讨论实然层面现行国际法框架、规则能否应对网络空间与人工智能所带来的挑战。在这一意义上，网络空间国际法与人工智能国际法实则仍存在较大的差异。

1.网络空间治理更加依赖于现有国际法框架

从实然层面讲，国际法在网络空间中的适用很大程度上遵循了应然层面的主张，是从现行国际法框架出发来解决问题的。美国作为这一观点的主要倡导者，其早在 2011 年《网络空间安全战略报告》中便指出：发展国家在网络空间中的行为规范既不需要重新发明新的习惯国际法，也不需要使现有的国际法过时。²²在和平与战争时期长期建立的国际规范同样适用于网络空间。²³面对网络技术的特点所带来的适用困境，则要通过规则进行解释来化解问题。²⁴上述观点在一定程度上反映了国际社会对现有国际规则适用于网络空间的实践情况，并在国际法框架与具体国际法规则上得到了落实。

网络空间治理的法律问题具有广泛性，涉及到管辖权重叠、网络干涉选举、网络战等多方面问题。因此，在国际法框架上，现行国际法框架中的主权原则、不干涉原则、禁止使用武力原则等国际法基本原则在很大程度上被网络空间国际法所承继。同时，针对网络技术在国家责任、海洋、外层空间、人权保护等领域产生的法律问题，网络空间国际法也侧重以既存的国际法框架为起点发展新的规则。上述特点很大程度上被反映在《塔林手册 2.0》中，手册在编写过程中便以现有国际法的体例为框架，在四个部分中分别讨论了一般国际法、特别领域的国

¹⁸ François Delerue, *Cyber Operations and International Law*, Oxford Press, 2010, p.1.

¹⁹ François Delerue, *Cyber Operations and International Law*, Oxford Press, 2010, p.1.

²⁰ Anne-Thida Norodom et al., *White Paper 16: Digital Challenges for International Law*, (2023), <https://www.ilaparis2023.org/wp-content/uploads/2022/08/Numerique-VHD-EN.pdf>.

²¹ Bhushan, Tripti, *Artificial Intelligence, Cyberspace and International Law*, 21 Indonesian Journal of International Law 281,286 (2024).

²² White House, *International Strategy for Cyberspace: Prosperity, Security, and Openness in a Networked World*, (2011), p.9,

https://obamawhitehouse.archives.gov/sites/default/files/rss_viewer/international_strategy_for_cyberspace.pdf.

²³ *Ibid.*

²⁴ *Ibid.*

际法、国际和平安全、武装冲突法在网络空间中的适用。²⁵

与此同时，网络空间的具体法律问题具有复杂性，需要基于现行国际法规则给予恰当的解释。以网络空间使用武力问题为例，其首先牵涉到何种网络攻击构成使用武力的判断。因此，在具体规则上，国际社会的讨论已然聚焦到“网络空间禁止使用武力的法律标准”之上，并形成了多种解释路径，²⁶包括目标说、装备说以及后果说等。²⁷又如在言论自由的国际人权保护中，为了更好的解决网络空间人权侵犯问题，人权事务委员会（Human Rights Committee）在《公民政治权利公约》的第 34 号公约评注中强调：有关表达自由的保护同样包括基于互联网的媒体，对网络言论的限制亦应当与公约第 19 条第 3 款第要求相兼容。²⁸这也反映了关于“个人在线下享有的诸多人权在线上也要予以保护”的国际解释立场。²⁹

概言之，现行国际法规则的可适用性在网络空间的实然治理中得到了较好的贯彻。在对新技术的治理过程中，网络空间国际法以现行国际法规则为基础，通过对现有规则的解释应对新的技术挑战。这一方法在最大限度上保障了国际法体系稳定性与可预测性，又能够通过最少的调整和修改应对技术挑战带来的新问题。在国际法自身规则更新迟滞，网络技术的挑战来临又如此迅速的背景下，采取此种法律适用路径不失为一种明智的选择。

2. 人工智能治理鲜有现行国际法适用的讨论

相比之下，尽管现行国际法对人工智能治理同样具有应然的拘束力，但是在实然层面，却鲜有牵涉现行国际法规则适用的讨论。以联合国主导的人工智能治理进程来看：在联合国大会通过的两份人工智能决议中，其中一份主要强调的是发展安全、可靠、可信赖人工智能与助力可持续发展，³⁰另一份则是强调了加强人工智能的能力建设国际合作。³¹两者均未将现行国际法在人工智能领域的具体适用标准问题作为讨论的重点。此外，联合国教科文组织出台的《人工智能伦理问题倡议书》则是以人工智能伦理为中心，未讨论现有国际法的适用问题。³²

即使是在理论与实践讨论较为健全的人工智能武器领域，在现行的国际人道法规则当然适用于人工智能武器的情况下，国际社会在此问题上也在相当程度上更加倾向于创制新的国际规则标准。自 2013 年《特定常武公约》缔约国会议就致命性自主武器展开首次讨论以来，人工智能武器便受到了相当高的关注。尽管在 2018 年的缔约国会议中，针对禁止致命自主武器的新规则制定因美国、俄罗

²⁵ See Michael N. Schmitt, *Tallin Manual 2.0 on the International Law applicable to Cyber Operations*, Cambridge University Press, 2017.

²⁶ 张华：《网络空间适用禁止使用武力原则的法律路径》，载《中国法学》2022 年第 2 期，第 284 页。

²⁷ François Delerue, *Cyber Operations and International Law*, Oxford Press, 2010, p.288-290.

²⁸ See Human Rights Committee, Draft general comment No. 34.

²⁹ Human Rights Council, The Promotion and Enjoyment of Human Rights on the Internet, para. 1.

³⁰ UN, GA Resolution 78/265, 21 Mar 2024.

³¹ UN, GA Resolution 78/311, 1 June 2024.

³² UNESCO, Recommendation on the Ethics of Artificial Intelligence, 23 November 2021.

斯、以色列、澳大利亚等国的反对而未能成功，³³但是针对致命性自主武器所确立的“有意义的人类控制则”却已经成为了各个国家的共识。³⁴

可以说，相较于依赖现有的国际法规则，人工智能领域的国际治理似乎更加倾向于创制新的国际法规则。这在一定程度上导致了在新的国际共识尚未达成时，人工智能领域几乎不存在具有拘束力的国际法。然而，人工智能领域并非是国际法真空，在人工智能领域加强现有国际法适用性的讨论也并非是不必要的。遗憾的是，至少到目前为止，上述工作并未在人工智能治理的过程中得到深入的讨论，人工智能的国际法治理仍待国际规则的进一步填补。

（二）国际法的规制模式特征

网络技术与人工智能治理在现状层面的第二个差异表现为国际法规制模式差异。从宏观层面检视，国际法对网络空间的规制内容更加侧重于“事后规制”，也更契合现行的国际法框架；而人工智能的规制则更加强调“事前预防”从而对现行国际法框架提出了更大的挑战。

1. 网络空间治理的事后规制模式

在网络空间中，当下国际法对于网络空间的治理主要采取了事后规制的模式。³⁵这种模式的特点在于：国际法的关注点往往集中于网络活动发生后，对其是否违反特定国际法规则进行合法性评价。一旦国际法规则遭到违反，国际社会将依据国家责任的相关规则，对不法网络活动进行归因，并要求相关国家承担相应的国际法律责任。在2022年的俄乌冲突前，恶意软件Whispergate曾在乌克兰展开了有组织性的破坏活动。³⁶事件发生后，乌克兰方面迅速对相关网络攻击进行了归因，将矛头指向俄罗斯。³⁷俄罗斯方面则对此问题加以否认，并试图规避其责任。³⁸

这一现象凸显了网络空间国际法在规则适用上的显著特征：其核心规则体系主要围绕网络活动发生后的合法性评价以及行为者的责任追究而构建。由于网络活动的隐匿性特征，行为的归因往往面临技术与政治层面的双重挑战。技术层面的挑战主要在于网络攻击的匿名性和跨国性，这使得针对网络行动准确识别行为

³³ 杨成铭，魏庆：《人工智能时代致命性自主武器的国际法规制》，载《政法论坛》2020年第4期，第138页。

³⁴ 冷新宇：《致命性自主武器系统议题下有意义的人类控制原则》，载《国际法学刊》2022年第2期，第3页。

³⁵ 此处所强调的“事后规制模式”主要涉及网络空间国际公法部分的行为讨论，并不涉及国际贸易、知识产权、国际私法等内容。

³⁶ Silas Cutler, *Whispers in the noise : A technical overview and the historic context of WhisperGate*, Stairwell Threat Research Blog (18 January 2022), https://stairwell.com/resources/whispers-in-the-noise/?utm_source=LinkedIn&utm_medium=post&utm_campaign=Threat+Research.

³⁷ Cynthia Brumfield, *Russia-linked cyberattacks on Ukraine: A timeline*, CSO United States (Jan.19, 2022), <https://www.csoonline.com/article/571865/a-timeline-of-russian-linked-cyberattacks-on-ukraine.html>.

³⁸ Ukraine says evidence points to Russia being behind cyber-attack, The Guardian (16 Jan. 2022), <https://www.theguardian.com/world/2022/jan/16/ukraine-says-evidence-points-to-russia-being-behind-cyber-attack>.

者变得极为困难。政治层面的挑战则体现在国际关系的复杂性上，国家往往擅长利用网络活动隐匿性的特点，否认或淡化自身在网络攻击中所从事的行为，并推卸原本应当由国家所承担的国际法律责任。外加黑客组织等非国家行为体在网络空间中的活动异常活跃，这一特点也进一步加剧了网络空间中国际责任归属问题的复杂性。可以说，恰恰是网络空间隐匿性的技术特征，激发了学界与国际社会在网络空间治理过程中针对责任分配问题的讨论。毕竟，若无法将国际不法行为归属于具体的国家，那么对国际规则违反的责任分配也就成了空谈。因此，在网络空间国际法中，有关网络行动发生后责任分配的国际规则建构讨论主张也就异常激烈。

从国际法的规则体系来看，这种事后规制模式与现行国际法的“初级规则”（Primary Rules）与“次级规则”（Secondary Rules）的结构特征高度契合。其中，初级规则主要规定国家在特定领域中的权利与义务，而次级规则则涉及规则未履行时的国家责任问题。³⁹在网络空间治理中，初级规则为国家在网络活动中的行为设定了规范，例如禁止未经授权的非法访问、数据窃取以及破坏性网络攻击等。而次级规则则为不法行为的归因与责任追究提供了法律依据，例如通过《联合国宪章》中的国家责任条款以及相关国际习惯法，明确网络攻击行为的法律责任。这种规则体系的有机嵌合，使得网络空间国际法在很大程度上依赖于现行国际法的框架，并强调对问题的事后规制。与此同时，这样一套规则适用模式也能使得网络空间相关的国际法规则更快的进行建构与适用。

2. 人工智能治理的事前预防模式⁴⁰

在人工智能（AI）治理领域，国际社会的法律关注点与网络空间治理模式存在显著差异。不同于网络空间治理中事后规制的主导模式，人工智能治理更加强调“事前预防”的理念。这种治理思路的核心在于通过法律框架的构建，预防人工智能技术可能带来的潜在风险，而非仅仅在问题发生后进行责任追究或损害补救。其既体现了国际社会对人工智能技术潜在风险的高度警觉，也反映了对技术发展与法律规范之间平衡的深刻思考。

以人工智能武器领域为例，正如联合国日内瓦办事处代理总干事迈克尔·穆勒所言：国际法在很多情况下只是对已经发生的暴行和所造成的痛苦做出反应，在对人工智能武器的规制方面，则要寻求采取“先发制人的行动”，以确保最终的生命决定权仍掌握在人类而非机器人手中。⁴¹在国际法对致命性自主武器的规制上，面对国际人道法中区分原则、比例原则、预先警告原则的潜在违反问题，国际专家组并未如网络空间的治理一样，强调国家违反国际人道法后承担国际责任

³⁹ ILC, Year book 1970/II, p.179.

⁴⁰ 考虑到人工智能国际治理过程中形成的硬法规则仍然十分有限，此处的国际法为广义，并将“国际软法”考虑在内。

⁴¹ 杨成铭，魏庆：《人工智能时代致命性自主武器的国际法规制》，载《政法论坛》2020年第4期，第138页。

的问题，而是在此之外亦从事前角度针对人工智能武器的特点创制新的规范，而是强调“有意义的人类控制”这一标准，从而在事前避免人工智能武器对重要国际法规则的违反。⁴²在此基础上，有意义的人类控制原则通过要求国家对武器进行全流程监管、并推广防扩散机制来实现其目的。⁴³

在人工智能治理的其他领域，国际组织也已采取了一系列重要措施应对人工智能技术可能带来的法律风险。联合国教科文组织（UNESCO）推出的《人工智能伦理问题建议书》（Recommendation on the Ethics of Artificial Intelligence）在多个政策领域提供了系统性伦理规范，旨在对人工智能的全生命周期进行全过程规制，从而预防因人工智能误用而引发的人权保护、隐私安全以及社会公平等方面的法律风险。⁴⁴与此同时，经济合作与发展组织（OECD）通过发布《人工智能原则》（OECD Principles on Artificial Intelligence），对人工智能技术的开发和应用提出了全面要求，强调了透明度和可解释性（Transparency and Explainability）、坚固安全可靠（Robustness, Security and Safety）等核心原则。⁴⁵这些国际组织的举措实质上体现了“预防为主”的治理理念，通过事前规制和系统性规范，力求在人工智能技术的快速发展中平衡创新与风险，确保其在法律框架内安全、可靠地服务于人类社会。

就两个领域的总体差异而言，上述人工智能相关国际法规则构建均以预防性措施为主，从而试图避免人工智能相关风险的产生。但这也使得现行国际法的规则不足以解决相关问题，需要针对人工智能技术的特点进行有针对性的构建。相比于基于现行国际法体系并侧重事后规制的网络空间治理模式，人工智能治理中采取的事前预防模式以及对新规则创制倾向，共同使得人工智能领域的国际规范在很大程度上没有依赖于对历史上国家间国际法共识的直接适用，而是选择了在过往未能涉及的国际法领域摸索新的规则。这直接导致了有关人工智能的国际法规则体系分散且不完善。

（三）国际规制的软法倾向

尽管如此，网络空间与人工智能治理在现状层面也有联系之处。其在治理上均表现出较为明显的软法偏好，进而使得发挥规制功能国际规范的法律性质主要属于国际软法而非硬法。就概念角度而言，国际软法往往被用来指代不具有法律拘束力却以规范方式表达并由国家和国际组织在国际关系中使用的文件。⁴⁶在对两者的治理中，恰恰是联合国决议、国家间倡议等不具有拘束力但却具有规范属性的文件发挥着重要作用。

⁴² 冷新宇：《致命性自主武器系统议题下有意义的人类控制原则》，载《国际法学年刊》2022年第2期。

⁴³ 冷新宇：《致命性自主武器系统议题下有意义的人类控制原则》，载《国际法学年刊》2022年第2期。

⁴⁴ UNESCO, Recommendation on the Ethics of Artificial Intelligence, 23 November 2021.

⁴⁵ OECD, Recommendation of the Council on Artificial Intelligence, 2024.

⁴⁶ Alan Boyle, *Soft Law in International Law-Making*, in Malcolm D. Evans, *International Law Fifth Edition*, Oxford Press, 2018, p.119.

在网络空间的治理中，国际软法直接推动了相关规则的发展和演进，促进了全球参与以及合作并影响了该领域的规则博弈。⁴⁷ 规制网络行动的实践也表明，与具有法律拘束力的规范相比，各国更倾向于讨论无拘束力的规范。⁴⁸ 前文所提及的 UNGGE、OEWG 等组织成果均以软法性文件加以规定，《塔林手册》等内容作为国际专家组报告虽然规则详细，评注全面，但是也同样属于不具有法律拘束力的国际软法。相比之下，除去网络犯罪领域的两项条约，⁴⁹ 俄罗斯等国采用条约等硬法方式进行造法的努力则在很大程度上受到了美国等欧洲国家的反对。⁵⁰ 其背后反映了网络空间规则博弈背景下路径选择的分歧，也部分揭示了国际硬法在国际共识难以达成背景下难以应对网络空间治理问题的窘况。

在人工智能领域，软法治理的偏好则要更加明显。截止到目前为止，除去欧盟《人工智能法案》⁵¹ 以及美英欧签署的《欧洲委员会人工智能、人权、民主和法治框架公约》⁵² 两项区域性立法及条约外，还尚不存在全球范围内的人工智能条约。与此同时，在国际社会对人工智能领域关注时间仍然较短的背景下，针对人工智能技术的国际习惯法也还难以形成，有关人工智能国际治理的硬法规制仍然处于“总体空白”状态。相比之下，前文所述及人工智能领域的众多里程碑式文件均为国际软法。人工智能领域的治理也更加倾向于通过人工智能安全峰会、人工智能首尔峰会、人工智能发展峰会、全球人工智能大会等一系列会议的形式凝聚国际共识，化解治理难题。总体而言，无论是网络空间治理还是人工智能治理，国际软法的作用较之其他传统国际法领域都大大加强了，并表现出来明显的软法偏好。

总言之，通过比较国际法在网络空间与人工智能领域的治理现状可以发现，在网络空间中，对现行国际法的依赖与对事后治理的青睐推动了网络空间国际规则体系的逐渐形成。然而，在人工智能治理中，对新规则创制的热情使其在某些程度上对现有国际法适用的讨论依然是不足的。同时，对人工智能风险事前预防的要求则进一步使得人工智能治理涉足到国际法此前未曾踏足的盲区和新领域。上述两方面因素共同导致了国际法在人工智能领域的治理也就显得分散和不成体系。此外，两者对软法的偏好使得不具有法律拘束力的国际规范在很大程度上主导了各自领域的治理，使得国际法对两者的治理在当前呈现出多方主体、多种方式共同介入的局面，而不再仅由传统的造法路径加以规制。

⁴⁷ 居梦：《论网络空间国际软法的重要性》，载《电子政务》2016年第8期，第35-39页。

⁴⁸ [法] 弗朗索瓦·德勒吕：《网络行动与国际法》，杨帆译，厦门大学出版社2024年版，第25页。

⁴⁹ 文中的两项条约指《布达佩斯公约》与《联合国打击网络犯罪公约》。

⁵⁰ [法] 弗朗索瓦·德勒吕：《网络行动与国际法》，杨帆译，厦门大学出版社2024年版，第27页。

⁵¹ Artificial Intelligence Act.

⁵² Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.

三、治理异同产生的原因

从原因角度讲,国际法在网络空间与人工智能领域的作用存在显著异同主要源于两者在技术特征上的结构性异同,这些技术特征上的因素提出了不同或相同的治理需求,并要求国际法给予相应的规则供给。“技术特征——治理需求——规则供给”三者锁链式的联动结构塑造了国际法在网络空间与人工智能领域的治理图景。

(一) 差异产生的原因

在网络空间与人工智能治理中,两者在现行法律与治理模式上的差异,由主要是因其技术特征的不同而产生的。在现行国际法的可适用问题上,网络空间的“领土性”与网络活动的“现实性”为现行国际法的适用提供了天然的基础。网络空间的物理层(硬件和网络基础设施)与社会层(参与网络行动的个人和团体等)通常位于各国的实际领土之内,⁵³这使得国际法的主权原则能够延伸至网络空间。主权国家的属地管辖权不仅体现在对境内互联网基础设施和网络行为主体的实际管理上,也体现在其不得允许本国行为主体损害他国利益的义务之中。尽管网络空间的逻辑层以代码的抽象形式存在,但网络活动的行为和后果最终都会对现实世界产生影响,这也使得现行国际法能够天然地适用于网络空间。因此,网络空间的“领土性”与网络活动后果的“现实性”共同构成了现行国际法得以适用的实践基础。

相比之下,人工智能技术虽然原则上应当受制于现行国际法,但由于其独特的技术特征,国际社会对其适用性的讨论相对较少。人工智能技术的“算法黑箱”与“自动决策”问题引发了国际社会对其可知性与可控性的深刻担忧,这在一定程度上分散了国际社会对现行国际法适用性的深入探讨。具体而言,“算法黑箱”这一概念作为一种隐喻,指代不为人知、无法打开且无法从外部观测的内部系统。⁵⁴在人工智能领域,“算法黑箱”凸显了技术对部分主体的不透明性,使得相关主体难以获得充分的了解或解释。⁵⁵这种不透明性可能导致人工智能技术从传统法律制度的监管框架中“逃逸”:一方面,国家可能无法及时发现算法决策中的潜在问题;另一方面,算法内部过程的不可知性也使得对其实施有效规制变得困难。

算法的自动决策进一步加剧了这一问题。在人工智能技术进行深度学习并输出决策的过程中,算法的运作可能使技术在无意间做出损害人类利益的行为。由于算法过程的高度复杂性,即使是算法的设计者也可能无法对其决策原因进行充分解释。⁵⁶可以说,算法黑箱与算法决策过程的复杂性共同加剧了人工智能技术

⁵³ 有关网络空间分层结构的标准来源于《塔林手册 2.0》。

⁵⁴ 谭九生,范晓韵:《算法“黑箱”的成因、风险及其治理》,载《湖南科技大学学报(社会科学版)》2020年第6期,第93页。

⁵⁵ 谭九生,范晓韵:《算法“黑箱”的成因、风险及其治理》,载《湖南科技大学学报(社会科学版)》2020年第6期,第93页。

⁵⁶ 参见王海燕:《算法可解释性的价值及其法治化路径》,载《重庆社会科学》2024年第1期,第126

的不可控性。即使现行国际法对人工智能技术的适用进行了详尽的规定，技术本身的不可知性和不可控性仍然限制了法律的适用效果。

因此，人工智能领域的国际法规则面临两个复杂问题：第一，人工智能的设计者有可能有意或无意的受到“算法黑箱”这一特点影响，使人工智能系统中存在不符合国际规范的潜在要求，这些有悖于国际规范的潜在要求难以被外界探查。第二，尽管设计者可能有意遵守现行国际法规范，技术本身却由于“自动决策”这一特性，未必能够完全遵循这些规范，进而造成对国际规范的违背。基于此，当前学术界和国际社会的讨论焦点与其放在解释不确定是否会被遵守的现行国际规范，更倾向解决其先决问题，即如何规制人工智能的开发和使用过程，以便使其更加透明、安全、可控，从而为其遵守现行国际法进行铺垫。

上述特征同样影响了两者在规制模式上的差异。在网络空间中，国家之间、国家内部网络的互联互通所带来巨大利益让各主权国家不得不在风险社会中承受遭受网络攻击的风险，国家无法切断本国网络之间的联系。因此国家无法用国际法改变网路空间的架构从而放弃巨大的开放利益要求对网络空间进行完全的事前监管。相比之下，恰恰是人工智能存在的算法黑箱与自动决策等问题，导致了人工智能较网络技术而言欠缺相应的可控性，需要将风险在最前环节进行阻却，亦即对人工智能本身进行事前预防。

时间因素进一步突显了网络空间与人工智能领域因技术特征不同引发的在国际法问题讨论上的差异。2023 年被视为人工智能技术发展“元年”，众多相关法律问题广泛涌现也是在此前后才逐渐进入国际社会的视野。因此，在短短两年（2023-2025 年）时间内，即使国际社会关注到了现行国际法规则在人工智能领域的适用与规制模式问题，也难以在短时间内形成充分的共识，更遑论着力解决人工智能领域存在的国际法问题。更何况，当前国际社会的讨论焦点更多集中在如何规范人工智能技术本身这一紧迫问题上，学界与实务界智力资源的集中，也就影响了对诸如现行国际法可适用性等其他问题的讨论。与之形成鲜明对比的是，网络空间的治理早在 20 世纪 90 年代便已开始，⁵⁷虽然相关国际法规则的确立主要集中在 2010 年之后，⁵⁸但网络空间的法律性质和治理模式的早期讨论早已展开。在彼时，网络技术的广泛应用早已使其深入人们的日常生活，这也为网络空间国际法的发展提供了丰富的实践基础和理论积累。因此，国际社会将有更多机会和更充裕的时间去准备和讨论国际法在网络空间中的适用问题。

此外，短时间内人工智能领域缺乏完善的发声渠道与协商机制也对差异的产生带来了一定影响。在网络空间治理过程中，联合国成立了国际安全背景下信息

页。

⁵⁷ Laura DeNardis et.al, *Researching Internet Governance: Methods, Frameworks, Futures*, MIT Press, 2020, p.60.

⁵⁸ 黄志雄：《网络空间国际法治：中国的立场、主张和对策》，载《云南民族大学学报（哲学社会科学版）》2015 年第 4 期，第 136 页。

和电信领域的发展的政府间专家组（UNGGE）和不限成员名额的工作组会议（OEWG），为国际社会就国际法的适用问题展开讨论提供了重要平台。尽管第五次 UNGGE 会议未能达成共识，但其失败的原因并非各国对国际法在网络空间中的适用产生根本性异议，而是对国际法如何具体适用存在分歧。从积极角度来看，这种分歧并非削弱了现行国际法适用的基础，反而体现了国际社会在这一问题上展开深入辩论的活力。⁵⁹如今，UNGGE 与 OECD 通过的文件已成为各国后续对话沟通的重要基础。在此基础上，各国纷纷通过发表立场性文件的形式，尝试阐述对现行国际法适用的理解，进一步推动了国际法在网络空间适用的讨论与问题解决。⁶⁰相比之下，人工智能技术在短时间内尚未形成这样一种较为完善的合作与协商机制。缺乏有效的国际对话平台和统一的规范性文件，直接导致了国际社会难以就国际法在人工智能领域的适用问题展开深入讨论并形成共识。这也解释了为何人工智能领域相关讨论较于网络空间而言较为滞后。

（二）联系产生的原因

就共性而言，网络空间治理与人工智能治理表现出的这种共同的软法偏好则与技术更迭速度快的共同特征紧密相关。是两者的共性技术特征推动了国际社会在治理中对软法的需求，并引发了国际法软法规则的供给。事实上，一般意义上的国际造法往往需要较高的造法成本并且仅具有较低的造法效率，⁶¹这使得快速发展的技术变革同缓慢的造法过程之间存在紧张冲突的复杂关系。就成本而言，各国在进行网络空间与人工智能治理过程中往往存在着巨大的立场分歧：以互联网自由为主的发展道路⁶²与侧重互联网安全为主的发展道路⁶³之间存在难以抉择的取舍关系。人工智能发展优先的政策主张⁶⁴与安全优先的政策主张⁶⁵亦存在难以调和之处。国际社会甚至一国内部就这些路径选择而言远没有达成共识。在分歧如此巨大的背景下，要求各国家共同协商并产生一份具有拘束力的文件是十分困难的。形成具有拘束力的法律文件不仅需要巨大的谈判成本，而且同样面临着徒劳无功的沉没成本。⁶⁶

⁵⁹ [法] 弗朗索瓦·德勒吕：《网络行动与国际法》，杨帆译，厦门大学出版社 2024 年版，第 19 页。

⁶⁰ 有关各个国家发布的国家立场文件，可参考网络法工具箱。Cyber Law Toolkit, (last visited 10 March 2025), https://cyberlaw.ccdcoe.org/wiki/List_of_articles#Common_and_national_positions.

⁶¹ 何志鹏：《逆全球化潮流与国际软法的趋势》，载《武汉大学学报（哲学社会科学版）》2017 年第 4 期，第 57 页。

⁶² Hillary Rodham Clinton, *Remarks on Internet Freedom*, U.S. Department of State (Jan. 21, 2010), <https://2009-2017.state.gov/secretary/20092013clinton/rm/2010/01/135519.htm>.

⁶³ 中华人民共和国国家互联网信息办公室，《国家网络空间安全战略》，载中国网信网 2016 年 12 月 27 日，https://www.cac.gov.cn/2016-12/27/c_1120195926.htm?from=timeline。

⁶⁴ The White House, *Removing Barriers to American Leadership in Artificial Intelligence* (January 23, 2025), <https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/>.

⁶⁵ *The Bletchley Declaration by Countries Attending the AI Safety Summit* (1-2 January 2023), <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>.

⁶⁶ 参见何志鹏：《逆全球化潮流与国际软法的趋势》，载《武汉大学学报（哲学社会科学版）》2017 年第 4 期，第 57 页。

就其效率而言，国际硬法，特别是条约的形成也往往需要较长的时间。以《联合国打击网络犯罪公约》为例，从 2019 年正式开启公约谈判议程，到 2024 年公约通过已达 5 年时间，未来公约何时能够满足生效条件仍有待时间检验。事实上，中国、俄罗斯等国主张在联合国框架下建立全球性的打击网络犯罪公约则可追溯到过更久。⁶⁷概言之，面对网络技术与人工智能技术的日新月异，国际硬法不仅需要极大的成本去订立新的国际条约，也有可能无法在有限的时间内形成规范及时应对，甚至可能无法达成共识形成新的国际硬法。

相比之下，国际软法虽然具有形式与效力均未法律化的固有局限，但是却能够在较短的时间内加以形成并发挥作用，具有较高的自由度和灵活性。2018 年举行的“世界经济论坛”便曾就快速变化的第四次工业革命提出了“敏捷治理”这一概念，要求形成一种更具有适应性、以人为本、包容且可持续的方法应对方式。⁶⁸国际法中的国际软法则恰恰在一定程度上满足了敏捷治理的需求，并得以适用于网络空间与人工智能技术的监管和规制之上。其通过宣誓人类对于网络空间与人工智能治理问题的一般价值观念和基本认识，不仅具有政治和道义的影响力，同时也将为后续的公约订立、习惯法形成以及法律原则确立提供有利条件。⁶⁹

与此同时，网络空间与人工智能领域治理过程中的多利益攸关方模式也对软法的广泛形成起到了推波助澜的作用。在由技术所主导的行业中，技术不仅掌握在国家手中，也掌握在企业手中。因此跨国公司等非国家行为体在技术行业的治理中可能比一些主权国家发挥了更大的作用。因而无论是在网络空间还是人工智能的治理过程中，都存在广泛的非国家行为体参与。例如 2005 年召开的信息社会世界峰会通过的《突尼斯议程》反映了网络空间治理进入多利益攸关方治理模式，并鼓励政府、私营部门、民间团体共同参与，群策群力。⁷⁰在此过程中，诸如 Microsoft 便曾发布一系列有关的国家行为规范，借以尝试降低网络空间中的冲突，⁷¹北约网络合作防御卓越中心特邀专家组编写的《塔林手册》、网络空间的国际法保护牛津进程系列⁷²也属于多利益攸关方共治的结果。在人工智能领域，由技术公司主导的自我约束承诺、⁷³由企业、学界、国际组织多边参与开展的人工智能安全峰会、人工智能行动峰会也体现了这一特征。然而，国际法的造法主

⁶⁷ 参见赵永琛：《〈联合国打击网络犯罪公约〉的制定问题研究》，载《武大国际法评论》2022 年第 5 期，第 1-3 页。

⁶⁸ See *Agile Governance Reimagining Policy-making in the Fourth Industrial Revolution*, World Economic Forum (April. 24, 2018), p.4, https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf.

⁶⁹ 万霞：《国际法中的“软法”现象探析》，载《外交学院学报》2005 年第 2 期，第 97-98 页。

⁷⁰ 蔡俊：《网络主权的法律问题研究》，武汉大学 2019 年博士学位论文，第 82 页。

⁷¹ Angela McKay et al, *International Cybersecurity Norms: Reducing Conflict in an Internet-Dependent World*, Microsoft (2015), <https://www.microsoft.com/en-us/download/details.aspx?id=45031>.

⁷² *The Oxford Process on International Law Protections in Cyberspace* (last visited on Feb. 28, 2025), <https://www.elac.ox.ac.uk/the-oxford-process/>.

⁷³ *Frontier AI Safety Commitments AI Seoul Summit 2024*, Gov.uk (7 February 2025), <https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024>.

体终究是国家，因此多利益攸关方治理模式所形成的成果在规范意义上无法构成国际硬法，但其作为国际软法仍然为领域的治理提供了有益补充，并使软法治理蔚然成风。

四、人工智能的国际法治理：未来展望

本文认为，通过对网络空间与人工智能领域的治理现状以及现状原因的梳理，或可以借此对未来人工智能的国际法治理进行合理的预测与展望。考虑到网络空间与人工智能治理在国际法数字化转型中技术背景方面相似性，网络空间国际法逐渐发展并解决问题的路径，或可为人工智能国际法的未来发展路径提供前瞻与镜鉴。本文认为，未来人工智能的国际治理或将围绕三个方面进行展开：在现行国际法适用问题上，伴随着讨论的深入，现行国际法在人工智能领域的适用或将达成更大的共识。然而，传统领域现行的统一国际法规则将进一步适应其技术特征，使国际法规则从“规则统一”到“差异适配”。在规制模式上，人工智能的国际治理也将从当前的“规制主导”，转向未来的“平衡发展”。在造法偏好上，人工智能领域国际法规则的产生方式将进一步摆脱传统的造法路径，由软法首先进行规制，并由硬法后续跟进，使国际造法从“软法先行”迈向“硬法后进”。

（一）从“规则统一”到“差异适配”

本文认为，就现行国际法的适用问题而言，伴随着时间的发展，人工智能治理领域的讨论也将如网络空间国际法一样，对现行国际法的适用展开更多的讨论。在解决新的治理问题同时，同样也应当强调现行国际法规则在人工智能领域的可适用性：例如探讨审慎义务(Due Diligence)可能在何种程度上要求国家采取措施预防或减轻危害、不干涉原则(Non-intervention)如何防止它过利用人工智能技术干涉选举进程等等。⁷⁴这种讨论的起点往往是从学术界先开始的。到目前为止，的确已有少量文章开始重视如何将环境法领域的预防原则（Precautionary Principle）适用于人工智能领域以预防未来的风险，⁷⁵此类讨论在未来将有可能进一步扩大。

然而，正如本文在第三部分所强调的那样，网络空间与人工智能领域的国际法治理具有明显的技术特征。在此过程中，网络空间与人工智能技术的技术特征为国家治理提出了差异化的治理需求，此种差异化的治理需求在某种程度上又影响了国际法规则的回应。面对上述现象，传统国际法中所存在的放之四海而皆准的“统一规则”在人工智能领域或将难以应对新的技术变革，国际法规则的差异适配或将成为趋势。以国际法的基石主权原则为例，现行的网络空间国际法虽然

⁷⁴ Talita de Souza Dias and Rashmin Sagoo, *AI Governance in the Age of Uncertainty: International Law as a Starting Point*, Just Security (Feb 23, 2025), <https://www.justsecurity.org/90903/ai-governance-in-the-age-of-uncertainty-international-law-as-a-starting-point/>.

⁷⁵ See Bryan H. Druzin et al., *Confronting Catastrophic Risk: The International Obligation to Regulate Artificial Intelligence*, 46 Michigan Journal of International Law 173,193-204 (2025).

承认了主权原则的可适用性，并依然对主权原则有着高度的依赖。然而在此前提下，主权原则又衍生出了对主权违反标准、跨境管辖、数据跨境等一系列新法律问题和标准讨论，从而在一定程度上突破了传统领土主权概念的领土边界，以差异适配网络空间的架构特征。在人工智能领域，现有的人工智能主权讨论则将其聚焦于国家对人工智能技术的独立发展控制之上，⁷⁶其更是在很大程度上同领土主权概念相偏离，以同人工智能治理的差异化治理需求相符合。可以说，国际法的“例外”将有可能越来越多。

在文章中，本文分别讨论了网络空间与人工智能领域国际法在现行国际法适用以及规制模式方面的特征与差异，其所反映的便是国际规则在两个不同领域的差异适配过程。网络空间对领土的依附以及归因难等特征分别决定了国际法在网络空间中的适用模式是通过对现行国际法的解释为基础，并以事后规制为主。与此同时，由于人工智能领域存在的算法黑箱与自动决策问题，使得国际法在人工智能领域适用的主要问题不在于具体规则为何，而在于规则是否能够被人工智能所遵守。因而在人工智能领域，国际法规则的供给则侧重于人工智能本身的透明度、可解释性以及安全风险等问题，同时强调事前预防的规制模式。尽管本文认为未来网络空间与人工智能的治理或将发生改变，从而使得网络空间治理出现更多的新规则创制，人工智能治理体现出更多的对现有国际规则的依附，但是上述新可能的存在也毫无疑问是以各自领域的特点进行差异适配的。

可以预见，对于一般国际法而言，伴随着规制领域的增加，国际法已经不再能够用简单的统一规则应对全部的国际社会议题。国际法应当察觉到人工智能领域差异化的治理需求，并对其进行差异化的规则供给。这一过程将进一步扩充现有的国际法规则，推动国际法规则的多样化，与此同时也会使国际法碎片化的进一步加。⁷⁷对于未来的人工智能全球治理而言，国际社会既应当强调国际法规则的差异适配，也应当继续应对国际法碎片化的挑战。

（二）从“规制主导”到“平衡发展”

第二，在规制模式方面，现行的人工智能治理机制呈现出以“事前预防”为主的风险治理。然而，在国际法进行差异适配应对技术问题的过程中，同样应当注重“技术规制”与“技术发展”之间的平衡。当下在对网络空间与人工智能治理的国际法规制过程中，无论是采取“事后规制”模式还是“事前预防”模式，都主要反映出了国际法对技术挑战与技术安全问题的隐忧。在人类对人工智能等新兴领域进行探索的艰难旅程中，如何避免国际法成为技术发展的泥泞和障碍，寻求规制与发展之间的平衡，却仍然有待进一步的探索和思考。2025年2月，美国副总统

⁷⁶ 参见赵骏，李婉贞：《人工智能对国际法的挑战及其应对》，载《浙江大学学报（哲学社会科学版）》2020年第2期。Rémi Sabonnadiere, *What Does AI Sovereignty Really Mean?* Artefact (Last Visited at: Mar. 2, 2025), <https://www.artefact.com/blog/what-does-ai-sovereignty-really-mean/>.

⁷⁷ 参见祖古雪：《现代国际法的多样化、碎片化与有序化》，载《法学研究》2007年第1期。

万斯在巴黎人工智能发展峰会上指出了对过度监管人工智能行业的担忧。⁷⁸中国也曾在《全球人工智能发展倡议》中指出了坚持“发展与安全并重”的原则，⁷⁹这些都凸显了国际社会针对技术规制与发展平衡的思索。

在文章中，本文强调了网络空间的事后规制模式，其本质上反映了国际法并未对国家进行网络活动的种类和方式进行限制，此举原因在一定程度上或可以归结于风险社会中网络空间的巨大利益，让人们得以对网络技术使用引发的特定风险给予包容。另一方面，人工智能技术采取的事前预防模式则侧重于对风险的规避。这或是由于人工智能技术特征使得人工智能技术的使用较网络技术使用或将引发更大且超出人类可接受范围风险，需要提前加以规制。然而也需承认，其在保障安全的同时确有可能影响并阻碍人工智能技术的未来发展。

因此，对于新兴技术而言，国际社会在对其进行规制的过程中或有必要首先依赖于科学方法，借由互联网、人工智能等相关学科专业人员的专业知识对技术的未来发展进行风险评估。在此基础上，国际法则可通过“法益权衡”等法学方法，在激发技术主主体活力的自由价值与进行合理有效规制的安全价值之间寻求平衡，对应当保护的主张赋予相应的等级和位序从而制定国际规则，并平衡规制与发展之间的关系。

（三）从“软法先行”到“硬法后进”

最后，考虑到网络空间与人工智能变化迅速的技术特征，在网络空间中国际规制对软法的偏好已经成为了这一领域的基本特点；⁸⁰在人工智能领域，亦有学者认为“超级软法”（supersoft law）将成为国际法规制人工智能领域的主要方式。⁸¹基于上文所述及的国际软法的相对优势，本文对上述观点总体持认可态度。面对飞速发展的数字技术变革，国际硬法自身存在着更新缓慢的局限性，在现阶段仍然有必要借由软法凝聚国家共识、并依其灵活性特点及时随着技术的指数化变革而加以动态调整应对。特别是在规则尚未成熟的人工智能领域，软法在未来长期时间内都可能占据主导作用。

然而，就人工智能国际治理的未来展望而言，应当认识到国际硬法存在的必要价值。适用于传统国际法各个领域的基本原则与重要规则，仍然有必要发挥其统摄功能。也应当有必要认识到，在国际层面，若要使国家遵守规则，能够为其提供正当性支持的，仍非国际法莫属。国际法作为国家间沟通的共同语言，为国家提供了更大程度的清晰度、可预测性与解决全球复杂问题的信心，国际法在全

⁷⁸ Aamer Madhani & Thomas Adamson, *JD Vance rails against 'excessive' AI regulation in a rebuke to Europe at the Paris AI summit*, American Press (12 February 2025), <https://apnews.com/article/paris-ai-summit-vance-1d7826affdcdb76c580c0558af8d68d2>.

⁷⁹ 《全球人工智能治理倡议》，载外交部网站 2023 年 10 月 20 日

https://www.fmprc.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml。

⁸⁰ 参见居梦：《网络空间国际软法的发展前景与方向》，载《情报杂志》2022 年第 7 期，第 26 页。

⁸¹ Thomas Burri, *International Law and Artificial Intelligence*, 60 German Yearbook of International Law 91, 105-107 (2017).

球治理中的核心作用仍不容忽视。⁸²从这层意义上讲，在人工智能的国际法治理中，软法同硬法之间将长期存在着一种微妙的互动关系。虽然当下人工智能领域对国际硬法的依赖仍然较小，但考虑到国际硬法相较于软法所具备的稳定性、可预测性与强拘束力等优势，国际社会在人工智能治理的重要议题和关键领域上也或将有国际硬法的需求，国际硬法“虽迟但到”。

因此，在硬法缺位的情况下，国家宣言、企业倡议等软法在未来将发挥规制的“排头兵”作用。待国际共识逐渐形成，国际硬法也会在更远的将来对软法欠缺拘束力的局限性加以克服，发挥其功能与效应，形成“软法先行，硬法后进”的治理图景。

五、总结

“法与时转则治，治与世宜则有功。”在国际法的“数字化”转型背景下，面对网络空间与人工智能技术所带来的技术挑战，国际法需要积极应对挑战并响应技术变革。本文通过考察网络空间与人工智能技术在三个方面的显著异同，认为网络架构的特征推动了网络空间将在较大程度上依赖于现有国际法，并基于现有国际法框架强调事后治理。相比之下，人工智能的技术特征则导致了对人工智能的规制将要从对人工智能自身出发，同时强调事前预防。但与此同时，两者技术发展迅速的特征则又影响了其在治理方式上对于国际软法的偏好。

从原因解释的角度讲，差异与共同之处存在的根源往往在于网络空间与人工智能技术的技术特征决定了其治理需求，国际法则需要对所提出的差异化治理需求予以因应。国际法回应技术变革的过程，亦是“法与时转”的过程。然而，在“法与时转”的过程中，在方法论层面，仍应当首先强调规则自身的“差异适配”，使规则供给同治理需求相结合。其次，在治理过程中，还应当诉诸价值权衡，风险评估等法学与科学方法，寻求规制与发展之间的平衡。最后，在造法路径上还应当做到“软法先行”，“硬法后进”，推动国际规则的合理发展。

⁸² Talita de Souza Dias & Rashmin Sagoo, *AI Governance in the Age of Uncertainty: International Law as a Starting Point*, Just Security (Jan. 2, 2024) <https://www.justsecurity.org/90903/ai-governance-in-the-age-of-uncertainty-international-law-as-a-starting-point/>.

美国网络空间立场文件中的塑造方法

——以主权的重塑和自卫权的强化为例

张予惟

摘要：当前，网络空间国际法正处于规范构建与规则博弈的关键时期。美国在这一过程中，善于根据自身利益灵活调整法律主张，及时适应国际法律环境的变化。以网络空间的主权与自卫权为例，美国关于主权的立场伴随技术进步、国际形势与各国反应而调整，从明确的规则逐步转向模糊的原则；而在自卫权问题上，其立场则逐步强化，并通过精细铺垫，使“前沿防御”（Defend Forward）战略在其语境环境下合法化。这种在主权和自卫权上的策略组合，不仅体现了美国网络空间法律政策的演进性与策略性，也揭示了其利用精妙的法律话语和软硬法结合方式，主动拓展国际法的解释空间和适用边界。对此，中国在深入分析美国经验与策略的同时，也需清晰揭示其背后的法理矛盾，积极探索应对之策，通过增强自身法律论证水平、深化国际合作机制建设，进一步提升在网络空间国际法领域的话语权和国际影响力。

关键词：国家立场声明 网络空间国际法 主权 自卫权

一、问题的提出

近年来，围绕网络空间国际规则的博弈日益激烈，各国纷纷发布立场文件，以期在规则制定中掌握话语权。然而，由于网络技术发展迅速，国家对网络空间治理的认知也在不断深化和演变，导致关于网络空间国际法治理的博弈周期变长。因此，对于任何国家而言，如何根据网络空间的技术发展趋势和国际法发展，及时调整和完善自身立场，以逐步形成更具优势且适应网络空间特征的治理策略，成为当前亟需解决的重要问题。

美国是在网络空间国际法治理中，发声最早且最积极的国家之一。早在 1999 年，美国国防部便发布了题为《信息行动中的国际法律问题评估》（*An Assessment of International Legal Issues in Information Operations*）的报告。这一文件明确提出了“国际法适用于网络空间”的基本前提，但同时敏锐地指出国际法律体系本质上具有被动回应的特点，即只有在特定事件发生并引起国际社会广泛关注之后，国际法才会作出相应的回应与调整。¹这一特性决定了各国若想在网络空间的规则

¹See U.S. Department of Defense, Office of General Counsel, *An Assessment of International Legal Issues in Information Operations* (2nd ed., 1999), in Michael N. Schmitt & Brian T. O'Donnell eds., *Computer Network Attack and International Law*, Newport: Naval War College, 2002, p. 463.

制定中占据主动地位，就必须密切关注国际法律环境的变化趋势，并在关键节点适时主动提出和推动符合自身利益的法律主张，抢占规则塑造的先机。换言之，国际法在网络空间问题上的解释和适用，必然随具体实践事件而逐步演变。

在这一背景下，美国如何积极适应国际法发展趋势，提出有利于自身的治理策略，值得深入探讨。本文以美国官方立场文件中关于主权和自卫权的论述为切入点，分析美国如何将网络空间的主权概念塑造成一项灵活的原则，同时不断强化自卫权的适用条件和范围，并最终铺垫出以“前沿防御”（Defend Forward）为核心的行动策略。

二、美国网络空间立场文件中对主权原则的重塑

1648年《威斯特伐利亚条约》确立了民族国家的领土主权观念。尊重领土主权逐渐发展至当代国际法的核心原则之一。根据1928年“帕尔马斯岛仲裁案”中的定义，“国家间关系中的主权意味着独立。独立指在该领土上排他性地行使国家职能的权利，排除任何其他国家的干预。”²这一表述奠定了现代国际法中主权概念的核心内涵，强调了国家对其领土的排他性管辖权以及领土内不受外部干涉的基本原则。

（一）传统的主权原则

一方面，主权原则作为国际法初级规则，当一国在未获得领土国同意的情况下采取行动时，前者即违反了对后者的尊重其主权的国际法义务，违反这一义务构成“国际不法行为”。国际法院的判决确定了侵犯一国主权造成的法律后果，并确定了其习惯国际法地位。主权国家享有对自身领土的排他性管辖权，未经一国同意，在其领土内实施行动的行为通常被视为侵犯主权。1927年“莲花号案”中，常设国际法院（PCIJ）明确指出，“国际法对国家施加的首要且最重要的限制是，除非存在与之相反的许可规则，否则国家不得以任何形式在他国领土上行使其权力。”³法院将这一原则视作保护主权国家领土完整和不可侵犯的强制性限制。在1986年的“尼加拉瓜案”中⁴，国际法院裁定，法院宣判美国向反政府武装提供援助不仅构成非法使用武力，也构成对尼加拉瓜领土主权的侵犯，从而违反了一般国际法原则和习惯国际法下所承担的义务。在该案中，法院明确区分了侵犯主权与非法使用武力，强调一项国家行为可能同时违反多个国际法原则。

另一方面，国际组织和国际条约在国家间主权平等原则的基础上，进一步衍生出一系列具体的法律义务。《联合国宪章》（以下简称《宪章》）第2条第1款中明确了联合国组织建立在各会员国主权平等的基础上。《宪章》还通过第2条

² *Island of Palmas Case (Netherlands v. United States of America)*, Award of 4 April 1928, II Reports of International Arbitral Awards (RIAA) 829, 838.

³ *Case of the S.S. "Lotus" (France v. Turkey)*, Judgment, 1927, P.C.I.J. Series A, No. 10, p. 18.

⁴ *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Merits, Judgment, ICJ Reports 1986, para. 190.

第4款和第7款对主权国家的固有领域（*domaine réservé*）提供了特别保护，明确禁止他国对一国的领土完整的侵犯和国内事务的干涉。此外，1970年《根据联合国宪章的国际法原则关于各国间友好关系与合作的宣言》进一步阐明了主权平等原则的权利与义务内涵。该宣言详细列举了主权国家所享有的核心权利（如独立自治权、排他性管辖权和不受外部干涉的权利）以及应履行的法律义务（如尊重其他国家的主权、不得对他国使用武力、通过和平手段解决争端的义务）。

由此可见，主权原则作为直接适用的国际法原则，具有习惯国际法性质。⁵与此同时，不干涉原则、禁止使用武力原则等核心规范均以主权平等原则为基础，并在其框架内得以发展和巩固。这些原则并非独立于主权原则而凭空产生，而是源于主权平等的法律逻辑和规范体系，共同构成了当代国际法的核心规则体系。

（二）美国承认网络空间适用主权

2012年9月，美国国务院法律顾问高洪柱发表了关于《网络空间的国际法》的演讲⁶（以下简称《高洪柱演讲》）。《高洪柱演讲》强调既有的国际法原则适用网络空间。《高洪柱演讲》在问题九中回答了关于网络空间的主权的问题，表示：在网络空间开展活动的国家必须考虑其他国家的主权，一国无论何时谋划在网络空间开展活动，都需要对其他国家的主权加以考虑。2013年6月，第三届联合国政府专家组（UN GGE）克服重重困难，达成了一份具有里程碑意义的报告。⁷该报告确认国际法，特别是《宪章》适用于信息通信技术的使用，并承认“国家主权及源自主权的国际规范和原则适用于国家在信息通信技术领域的活动。”2014年10月，美国向2014-2015年度的政府专家组提交了其官方立场文件⁸（以下简称《2014年美国官方文件》）。在这份文件中，美国肯定了前几届政府专家组的工作成果，尤其是对网络空间主权原则的确认表现出明确支持。文件进一步明确了主权与网络活动领土管辖的关联性，但强调主权的管辖不是无限制的，并以言论自由的案例进行了说明。

截至2014年，美国明确承认主权原则适用于网络空间，且并未提出关于主权在网络空间适用的特别法观点。在外空和海洋法领域，主权原则作为国际法初级规则仍然适用，属于一国的航天器和船只禁止他国管辖。但是各国通过条约、习惯国际法等，确定了各国不得通过主权要求、使用、占领或任何其他方式，将外层空间据为己有⁹、他国船只可以在一国领海无害通过等特殊法（*lex specialis*）。

⁵更详细的论述过程见 Schmitt, Michael N. and Vihul, Liis, Respect for Sovereignty in Cyberspace (November 3, 2017). Texas Law Review, Vol. 95, 2017, Available at SSRN: <https://ssrn.com/abstract=3180669>

⁶ Harold Hongju Koh, "International Law in Cyberspace", U.S. Department of State, 19 September 2012, <https://2009-2017.state.gov/s/l/releases/remarks/197924.htm> (visited 26 May 2024).

⁷ United Nations General Assembly, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, UN Doc A/RES/68/98 (2013).

⁸ CarrieLyn D. Guymon (ed.), Digest of United States Practice in International Law 2014, U.S. Department of State, 2015, <https://2009-2017.state.gov/documents/organization/244504.pdf>, visited 6 July 2024.

⁹ 《关于各国探索和利用包括月球和其他天体在内外层空间活动的原则条约》第2条（1967年1月27日开放签署，1967年10月10日生效）

UN GGE 2013 年的报告将主权和源自主权的规则进行了区分，这一声明不仅为后续的网络空间法律讨论奠定了基础，也为主权原则在网络空间中的适用提供了重要依据。根据 2013 年由北约赞助的《塔林手册》解读，网络空间中的国家主权“在遵守国际法义务的前提下，对其领土内的网络基础设施、人员和网络活动享有主权权威”。¹⁰《高洪柱演讲》的立场表明，美国承认无论是和平状态还是战时状态，主权原则对网络空间行为具有直接的规范作用，主权与网络活动领土管辖的关联性。互联网虽然具有虚拟性，但其运行依赖于物理基础设施，如服务器、数据中心和通信网络，这些设施归属各国的领土管辖范围。美国的立场承认，这些设施的管辖权应受到主权原则的保护，主权管辖的行使应符合《公民权利与政治权利国际公约》（ICCPR）等国际人权法的规定，以避免对基本权利的过度限制。由此来看，美国承认网络空间适用的主权规则，具有习惯国际法性质，是一项初级规则。

（三）美国对网络空间主权的重塑

2016 年美国国务院法律顾问布莱尔·伊根在其母校伯克利大学发表题为《国际法与网络空间的稳定》的演讲（以下简称《伊根演讲》）¹¹。在《伊根演讲》中，伊根紧接之前主权与网络活动领土管辖的思路，澄清美国尊重所有国家通过立法规范在其领土内的主权行为，违反这些法律的行为会违反一国的国内法，或者产生外交政策后果。但是他又首次提出，针对位于他国领土内的计算机或其他网络设备的远程网络行动本身不构成国际法的违反，尤其是没有造成影响或仅有微小的影响时，并以网络间谍为例。伊根认为，未经同意网络行动究竟何时侵犯了另一国的主权有待通过各国的实践和法律意见来解决。

2018 年 5 月，英国司法部长表示¹²，主权并不是一项独立的国际法规则。对于英国而言，只有达到强制干预或使用武力程度的网络行动才是国际不法行为。2020 年美国国防部总法律顾问保罗·内伊在美国网络司令部年度会议上发表的演讲（以下简称《内伊演讲》）中表达，国防部总法律顾问办公室的观点与英国政府在 2018 年的观点有相似之处。内伊提到“在分析拟议的网络行动时，国家主权原则是国防部法律专家要考虑的门槛性问题。”内伊表示国防部会长期考察主权对网络空间的影响，并最后表示“尽管似乎并不存在这样一项规则，即所有侵犯网络空间主权的行为必然涉及对国际法的违反。”

美国对网络空间的重塑分为两步。首先是弱化主权本身的约束力。伊根在领

¹⁰ Schmitt, Michael N., ed. *Tallinn manual on the international law applicable to cyber warfare*. Cambridge University Press, 2013.

¹¹ Egan, Brian J. "International Law and Stability in Cyberspace." Speech delivered at Berkeley Law School, November 10, 2016. <https://www.law.berkeley.edu/>, visited on 16 May 2024.

¹² Jeremy Wright QC MP, *Cyber and International Law in the 21st Century*, UK Foreign, Commonwealth & Development Office, 23 May 2018, <https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>, visited 16 May 2024.

土管辖的思路提出，诸如间谍活动这种未对物理设施造成影响，或仅造成微小影响时（在脚注），不构成国际法上的不法行为，习惯国际法上并未对此加以禁止。伊根选择了一个非常“狡猾”的案例，即各国之间的间谍行为。一是国内外学者¹³普遍认为间谍活动是习惯国家法上尊重主权的一个例外情况。习惯国际法由国家实践和法律确信（*opinio juris*）构成。一方面各国虽然都在隐秘地进行间谍活动，但另一方面各国纷纷通过国内立法，反间谍行为。所以看得见的国家实践是国内立法行为。而各国正是明知间谍行为的不合法性，才三缄其口。二是美国希望将自己在网络技术和供应链上的优势延伸至网络空间，为其主张网络间谍的合法化埋下伏笔。《内伊演讲》则更是在英国立场文件的基础上表明，网络空间主权不再作为国际法初级规则。

其次加强对源自主权规则的强化。伊根论述了不干涉原则对主权的影响：禁止各国采取强迫性行动来影响他国根据主权原则有权自由决定的事项，例如对于政治，经济，社会和文化制度的选择。他称不干涉原则为一条“相对狭窄的习惯国际法规则”，并以一国操纵另一国的总统选举作为举例。不干涉内政原则也是基于各国主权平等，每一国均有选择其政治、经济、社会及文化制度之不可移让，不受他国任何形式之干涉。¹⁴是各国公认的国际法基本原则之一，具有习惯国际法性质。在传统国际法语境中，不干涉原则主要由相对弱势国家提出，用以抵御强势国家的政治、经济或军事干预。通常而言，主张该原则的国家多处于国际权力体系的边缘，长期以来饱受以美国为代表的西方国家的非法干涉，因而希望借助不干涉原则作为抵御外部影响、维护国家主权的重要法理依据。¹⁵而在俄罗斯干预美国总统选举的背景下，《伊根演讲》却转而主张：一国操纵另一国选举的行为构成对不干涉原则的违反。这一立场巧妙地将“选举干预”纳入传统不干涉义务下的“保留领域”（*domaine réservé*），即仅由国家自主决定的核心事务范围。值得注意的是，伊根在演讲中明确指出：对网络主权的侵犯，并不必然构成国际不法行为；只有当行为干涉了一国的“保留领域”，或对关键物理设施造成实质性影响时，才可能违反国际法。通过上述论证，美国成功地将网络主权描绘为一种宽泛的指导性原则，而非具备明确法律约束力的规则。

违反基于此的特定规则，对物理设施造成影响或某种程度上干涉一国内政才造成一国不法行为。这无异提升了网络空间违法的门槛，造成了一国网络空间中的灰色地带。同时，网络空间不仅包括包括物理基础设施的管辖权，还包括互联

¹³Schmitt, Michael N. and Vihul, Liis, *Respect for Sovereignty in Cyberspace* (November 3, 2017). Texas Law Review, Vol. 95, 2017, Available at SSRN: <https://ssrn.com/abstract=3180669>; Russell Buchan, *When more is less: The US Department of Defense's statement on Cyberspace*, EJIL:Talk!, <https://www.ejiltalk.org/when-more-is-less-the-department-of-defenses-statement-on-cyberspace/>

¹⁴*Declaration on Principles of International Law concerning Friendly Relations and Co-operation among States in accordance with the Charter of the United Nations*, UN General Assembly Resolution 2625 (XXV), 24 October 1970.

¹⁵See Arnulf Becker Lorca, *Mestizo International Law: A Global Intellectual History* (2014), p.1842-1933, 341-352.

网的管辖。降低非物理损害的不干涉的门槛，使其容纳更多行为“软性损害”，做为利用国际法反击的武器。

三、美国网络空间立场文件中对自卫权的强化

（一）模糊“使用武力”和“武力攻击”之间的区别

2012 年《高洪柱演讲》在问题四中，承认一国可以通过行使《宪章》赋予的自卫权来应对计算机网络攻击。但在随后谈论到网络空间国际法面临的挑战和不确定性时，高洪柱重申了美国“长期以来的立场”：“固有的自卫权可能适用于任何非法使用武力的行为。我们认为，使用致命武力的行为无需达到某个门槛，就可以构成可进行武力应对的‘武力攻击’。”高洪柱承认部分国家将“武力攻击”作为“使用武力”的一个子集，严重程度更高“武力攻击”才可触发自卫权。高洪柱认为，关于二者的区分，“只不过是把老问题适用到最新的技术发展上了。”在网络空间模糊“使用武力”（use of force）和“武力攻击”（armed attack）之间的界限，不仅可能引发已有争议的再次出现，还可能随着国际法在网络空间中的适用而带来新的问题。

首先，模糊二者的界限可能导致网络空间自卫权的过度扩张。《宪章》中第 2 条第 4 款规定，各国不得使用武力。1996 年国际法院在关于“威胁使用或使用核武器的合法性”的咨询意见中指出：第 2 条第 4 款的规定并未对武器的类型作出限制，而是适用于任何形式的武力使用。《高洪柱演讲》中认为“如果网络攻击造成的物理后果相当于投掷炸弹或发射导弹造成的物理损害，该网络攻击同样应该被视为使用武力。”网络空间的网络攻击可以造成和武力攻击同样的物理损失和人员伤亡，明显属于《宪章》中“使用武力”适用范围内。

然而，并非所有“使用武力”都构成可引发自卫权的“武力攻击”。根据《宪章》第 51 条，只有当被禁止的“使用武力”行为上升为“武力攻击”时，受害国才可采取单独或集体自卫措施。在“尼加拉瓜诉美国军事行动案”中，国际法院的判决认定向叛乱分子提供武器、后勤或其他形式支持的援助，不构成武力攻击，但此类援助可能被视为对他国的武力威胁或使用，亦或构成对他国内政或外交事务的干涉¹⁶。国际法院的判决提出根据“规模 and 效果”区分“最为严重的武力使用”（the most grave forms of the use of force）与“较轻微的武力使用”，前者才构成“武力攻击”，从而可能引发自卫权。这一原则在后来的“石油平台案”中得到了进一步重申。国际法院指出，美国为证明其攻击伊朗平台的单独自卫行为在法律上是正当的，必须提供证据证明伊朗对其实施了攻击，并且这些攻击的性质达到《宪章》第 51 条中“武力攻击”的标准。目前已有 24 份网络空间立场文件明确指出，应区分构成“武力攻击”和“使用武力”的网络行动¹⁷。美国在其 2014 和 2021 年的立场

¹⁶Ibid.3, para. 191, 195.

¹⁷这 24 份立场文件分别是非洲联盟（2024）、欧盟（2024）、奥地利（2024）、古巴（2024）、丹麦

文件中提出，应围绕事件的背景、实施行为的行为者、目标及其位置、行为者的意图等“逐案考虑”某一事件是否构成网络空间的“武力攻击”或“使用武力”。然而，这两份文件均未明确区分“武力攻击”与“使用武力”之间的界线。尽管网络空间的攻击相较于传统动能攻击缺少直观的物理表现和量化标准，但缺乏对“武力攻击”与“使用武力”之间程度差异的认识，增加了国际法理解和适用上的不确定性，在所谓自卫权的框架下可能不合理地扩大武装攻击的适用范围，进而使其成为合法化攻击的借口，带来法律与实践上的潜在风险。

其次，由于现有评判标准的不完善及主权作为规则在网络空间中的缺位，低烈度网络攻击可能因此被合法化。《宪章》中未对“使用武力”或“武力攻击”进行定义，导致这些概念在网络空间中的适用变得更加复杂。在现有的法律确信中，26 份国家网络空间立场文件¹⁸普遍认为“使用武力”的认定应遵循国际法院在尼加拉瓜诉美国军事行动案中提出的“规模和效果论”，或认为其应达到与传统的动能“使用武力”相当的程度，即造成实质性物理损害或人员伤亡。然而区别于传统的物理空间，网络空间最大的特点在于其虚拟性，网络空间由数据、协议等构成，关键基础设施等实体映射仅占其中一部分。荷兰、丹麦等国家已开始考虑网络行动对经济活动的影响，但《宪章》的起草者明确拒绝了将禁止武力的范围扩大到超出军事或武装武力的建议。¹⁹这些立场文件中的观点反映出一个普遍被忽视的重要问题：当前对网络空间武力使用的讨论，仍然未能充分考虑网络空间的虚拟特性、其由数据组成的本质以及这一特质对国际法适用的深远影响。

最后，在“规模和效果论”的框架下，某些低烈度的网络行动，如仅对数据或虚拟“防火墙”造成破坏，若未能达到传统动能军事攻击的明确界限，便很难被认定为“使用武力”。例如，如果一国替换或删除另一国军用数据库的部分数据，且暂未产生实际后果，根据美国提出的主权不作为规则，此类行为很难被视为国际不法行为。正因如此，低烈度网络行动在国际法适用过程中，往往落入了和平与武装冲突之间的“灰色地带”²⁰。这一“灰色地带”的存在，使得低烈度网络攻击在缺乏明确判定标准的情况下，可能被视为合法化，从而导致网络空间中的国际法律规则滞后于实际情况，进一步加剧了网络行动的法律复杂性。

(2023)、爱尔兰(2023)、波兰(2022)、巴西(2021)、德国(2021)、意大利(2021)、日本(2021)、爱沙尼亚(2021)、瑞士(2021)、新加坡(2021)、英国(2021)、挪威(2021)、芬兰(2020)、澳大利亚(2020)、伊朗(2020)、以色列(2020)、新西兰(2020)、荷兰(2019)、法国(2019)、英国(2018)
18这 23 份立场文件分别是非洲联盟(2024)、欧盟(2024)、奥地利(2024)、丹麦(2023)、爱尔兰(2023)、波兰(2022)、瑞典(2022)、加拿大(2022)、美国(2021)、巴西(2021)、德国(2021)、意大利(2021)、爱沙尼亚(2021)、英国(2021)、挪威(2021)、罗马尼亚(2021)、美国(2020)、芬兰(2020)、捷克(2020)、澳大利亚(2020)、新西兰(2020)、以色列(2020)、荷兰(2019)、法国(2019)、英国(2018)、美国(2012)

19Documents of the United Nations Conference on International Organization, San Francisco, 1945, Vol. VI, Commission I (General Provisions), United Nations Information Organizations, 1945, p. 334.

20Gary Corn, *Cyber National Security: Navigating Gray Zone Challenges In and Through Cyberspace*, in *Complex Battlespaces: The Law of Armed Conflict and the Dynamics of Modern Warfare*, Oxford: Oxford University Press, 2019, p. 45-67.

（二）对非国家行为体的自卫行为

虽然国际公法的核心宗旨在于协调国家间的关系，赋予自卫权的《宪章》的约束对象也是国家。但 9.11 恐怖袭击后，美国积极主张将自卫权的行使范围扩大至非国家行为体，这一立场在网络空间的应用也同样得到了体现。美国在 2014 年的立场文件中，提出了“不愿意不能够”理论，指出“如果领土所属国不愿意或不能够阻止或防止在网络空间或通过网络空间发动的实际或即将发生的武装攻击，则该国可在未经同意的情况下采取行动。”²⁰2016 年的《伊根演讲》中，美国进一步阐述了为了打击伊斯兰国的指挥和控制系统，必须“中断其转移资金的能力、削弱其施暴和控制民众的能力，以及切断其招募人员的渠道。”在对非国家行为体发起的网络攻击中行使自卫权，特别是以传统动能方式进行反击，可能会带来新的问题。

首先，“不能够不愿意”理论降低了网络空间国际法中国家责任归因的门槛。当一国遭受非国家行为体的“武力攻击”时，除非能将此攻击归因于一个具体的国家，否则受害国无法在一般国际法的框架下行使自卫权²¹。在《国家对国际不法行为的责任条款》（Responsibility of States for Internationally Wrongful Acts，以下简称 ARSIWA）第二章第 4-11 条列举了可将不法行为归因于国家的情况。然而事实上，由于网络空间活动具有高度的隐蔽性和私密性，即使受害国能够通过网络溯源追踪到攻击源，但往往难以找到符合 ARSIWA 第 4 条至第 11 条要求的法律证据，将非国家行为体的攻击行为归因于具体国家。²²“不能够不愿意”理论最有力的法理依据来自“审慎原则”。国际法院在“科孚海峡案”中表明，每个国家都“有义务不允许其领土在知情的情况下被用于侵犯其他国家权利的行为”。审慎义务在网络空间中的应用也得到了广泛关注。2013 年和 2015 年的 UNGGE 报告，均提到了“负责任国家行为”，2015 年 UNGGE 的报告指出：各国应鼓励负责任地报道信通技术的脆弱性，采取合理步骤确保供应链的完整性，并防止恶意的信通技术工具、技术或有害隐藏功能扩散。但 UNGGE 并未将其作为强制性义务。但是，美国在其 2020 年的立场文件中提出，并不认为存在支持审慎已经构成国际法上的一般性义务的国家实践和法律确信。这与其提出的“不能够不愿意”理论造成了法律上的冲突。

其次，在未经一国同意对非国家行为体进行自卫，针对其管辖的非国家行为体行使自卫权，可能构成对国家主权的侵犯。在相关国家技术手段有限的“不能够”，或者拒绝旅行自身义务的“不愿意”的情况下受害国单方面实施域外武力干预，实质上是以自卫权之名行域外执法的“长臂管辖”之实。以传统动能回应网络

²¹参见 [德] W. G. 魏智通主编：《国际法》，吴越、毛晓飞译，法律出版社 2012 年版，第 605 页。

²²张华：《论网络空间自卫权的行使对象问题》，载《法学论坛》2021 年第 1 期，页 72-80

攻击，造成人员伤亡、财产损失、社会扰乱，可能会超过自卫权行使必须遵守的必要性和比例原则。而即便采用网络自卫手段，依靠过境国网络关键基础设施发动网络攻击的情况下，自卫权的行使极有可能“殃及无辜”，即影响与攻击无直接关系的第三国非国家行为体和网络，进一步侵犯其主权。

四、美国“前沿防御”的提出

2018 年，美国网络司令部提出了一种新的自卫策略——“前沿防御”(Defend Forward)。这一策略的核心目标是从源头上打击或阻止恶意网络行动，包括那些低于武装冲突门槛的网络行动。²³根据美国国会设立的网络空间索拉里安委员会 (Cyberspace Solarium Commission) 的报告：“前沿防御”符合国际法，是一种“防御性策略”，但在战术和操作层面上包含“进攻性成分”，这一策略要求“为了实现网络空间中的防御性战略目标，必须将力量和能力进行前置部署，无论是在地理上还是在虚拟空间中”。²⁴美国国家安全局局长保罗·中曾根 (Paul Nakasone) 解释道，美国的目标是“通过不断在敌方发动网络攻击的地方与其对抗，尤其是在低于武装冲突的层级上，来实现并保持在网络空间中的主动权。”

“前沿防御”的提出并非偶然，在之前的官方立场，已经多次铺垫使其合法化的法理依据。“前沿防御”不仅仅局限于被动自卫，而是由一系列多层次的行动组成，既包括主动获取对方信息、监测敌方活动等信息战手段，也涵盖了在必要时对敌对行动进行主动打击的军事行动。这一激进战略的背后体现了主权原则的运用与自卫权的强化铺垫的交织。首先，“前沿防御”战略涉及在对方攻击之前，进入其他国家或非国家行为体的网络获取“可以使经济实力的关键要素更具弹性和可防御性的见解和威胁信息”。²⁵这种网络空间行动，尤其是在未达到“使用武力”门槛的情况下，涉及在其他国家的网络中展开活动，有学者评论其行为超过了传统间谍活动的范畴。在美国之前的铺垫中，针对他国的此类行为若没有达到干涉他国内政的程度，与间谍活动类似，未必被视为违反国际法。然而，这一战略在实践中的核心问题在于，它虽未达到‘使用武力’的门槛，但已对他国网络空间实施渗透和干预，实质上削弱了国家主权原则的约束力。一旦此类行动的合法性被普遍接受，国际法关于网络主权的界限将变得更加模糊，甚至为未来更具侵略性的网络行动提供了法理依据。

其次，根据美国白宫官方文件，“前沿防御”的目标是“从源头破坏或阻止恶意网络活动，包括低于武装冲突级别的活动。”²⁶旨在建立削弱对手能力的能力的

²³U.S. Department of Defense, *Summary: Department of Defense Cyber Strategy 2018*, at 1 (2018) (emphasis added), p. 17.

²⁴Cyberspace Solarium Commission, *Report*, 33 (2020).

²⁵U.S. Cyber Command, *Achieve and Maintain Cyberspace Superiority: Command Vision for US Cyber Command* (2025).

²⁶Ellen Nakashima, "White House Authorizes 'Offensive Cyber Operations' to Deter Foreign Adversaries, *The Washington Post*, September 20, 2018.

网络行动是否会被视为使用武力。这一战略的关键问题在于，旨在削弱对手能力的网络行动是否会被视为“使用武力”。这一行动可能对敌方网络或其他相关基础设施造成影响，达到“使用武力”的门槛。美国通过模糊“使用武力”与“武装冲突”的界限，以自卫为名，对国家和非国家行为体实施攻击，突破了国际法对武力使用的限制，实质上削弱了《联合国宪章》确立的国际安全架构。传统上，自卫权的行使必须遵循“必要性”和“相称性”原则，并且通常要求先遭遇“武力攻击”或者美国自身立场文件中的对迫在眉睫的“武力攻击”进行自卫（immediate self-defense）。然而，“前沿防御”并非在敌方攻击发生后进行的被动防御，而是在威胁尚未明确发生之前采取主动行动。这种策略使得“前沿防御”更接近于预防性自卫（Anticipatory Self-Defense），甚至可能被视为先发制人的攻击（preventive attack），而非严格意义上的“自卫”。尽管美国政府为“前沿防御”的提出铺垫了一套法律解释，使其行动在某种程度上“自洽”，但其在国际法框架下合法性仍存在疑问。

五、美国立场文件的策略构建特点

（一）对习惯国际法的熟练运用

美国试图通过工具化解释与规则塑造双重路径主导网络空间习惯国际法的演进。一方面，对习惯国际法的解释具有鲜明的工具性特征，关注应用国际法规则的目的和意图，以适应国家利益和现实需要。新技术的出现，尤其是可应用于军事领域的技术，通常会引发关于现有法律框架是否足以涵盖这些技术使用，或它们是否完全超出现有规则范围的讨论。然而，为了避免法律上的“真空地带”和可预测性，大多数情况下，各国倾向于至少在概念上将现有框架应用于新技术。高洪柱在自己的论文中将这一过程可以被视为“翻译性操作”——即通过审视现有法律的“精神”，将其适应于“当今情境”。在关于侵犯一国主权的干涉总统选举和间谍行为，美国积极运用习惯国际法对情形进行化解。直指俄罗斯黑客干预美国大选为违反国际法的干涉原则，而却对自身网络情报行动，如“棱镜计划”，背后的对关键基础设施的渗透是否构成对主权构成侵犯进行开脱。凸显美国对习惯国际法解释的立场依赖性。

另一方面，运用网络空间的特点，积极促成新习惯国际法的形成。网络行动的特点是隐蔽性和难归因性。《内伊演讲》对习惯国际法未普遍禁止在另一国境内进行未经同意的网络行动的一个原因解释为“许多国家在公开场合对于大量已知的外国网络侵入事件保持沉默。”网络空间具有较强的隐蔽性，各国开展的秘密网络行动很难被其他国家察觉。秘密行动通常缺乏透明度，难以形成国家实践的公开记录，因此无法为习惯国际法的形成提供充分的证据支持。这种隐秘行为使国家在公开领域中的法律立场无法清晰呈现，进而影响国际法律话语的形成

和发展。美国却未承认这种沉默可能源于技术能力不对称（如发展中国家难以溯源攻击），而非对规则的默许。若套用“渔业案”逻辑（沉默即同意），可能加剧“技术霸权”对习惯法解释权的垄断。

（二）软法和硬法的综合

美国在网络空间治理中采取“硬法”约束与“软法”并存的策略，以保持其在国际规则塑造中的战略灵活性。根据《国际法院规约》第三十八条确立的国际法渊源体系，传统意义上的“硬法”规则一般指：条约、一般法律原则和习惯国际法。近年来，各国的国家实践日益倾向于通过不具法律约束力的政治性文件（如宣言、决议和行动计划）发布规范性声明，并在实践中赋予这些规范一定的遵守预期。这类规范性文件被统称为“软法”（soft law）。²⁷相比于“硬法”，达成普遍共识的“规范”或“共识”更具灵活性。然而，两者之间存在一个根本区别，即违反具有法律约束力的“硬法”会引发国际法律责任，而“软法”则不会。

在“硬法”缺位时，“软法”可以适时补充空白，良好过度。继 UNGGE 在 2015 年的报告中详细列举了 11 项“自愿、非约束性的负责任国家行为规范”后。《伊根演讲》解释美国确保网络空间的稳定中的三大支柱中就包括“自愿性、不具约束力的和平时期网络空间负责任国家行为规范达成的国际共识。”各国达成共识的“负责任国家行为规范”可以使各国在对于战时法、网络空间人道主义法在争论不休的前提下，以更低的政治和法律成本维护网络空间的和平。

“软法”可能逐渐发展成习惯国际法，进而长出牙齿。美国推动软法的方式体现了一种“先共识、后约束”的路径，即先通过软法规则塑造各国的行为模式，并在条件成熟时将其转化为更具约束力的法律框架。通过主导“自愿性规范”的议题设定，塑造全球共识，并逐步使这些规范在国家实践和法律确信的积累下演变为习惯国际法。例如说“负责任国家行为规范”中由西方国家主导的关于人权的表达自由，会导致发展中国家对忽视了对网络暴力、犯罪信息审查的需求。除了“负责任国家行为规范”，《伊根演讲》中还突出推广 4 条规范，其中一条包括关于窃取商业信息的规范呼吁。随后网络经济间谍的表述，出现于《七国集团伊势志摩领导人宣言》、《美印网络关系框架》等文件。美国还试图推动将“各国不应通过网络手段窃取知识产权、商业机密和其他敏感商业信息用于获取商业利益”确立

²⁷Andrew T. Guzman & Timothy Meyer, *Soft Law*, in *Economic Analysis of International Law* 123-154 (Edward Elgar Publishing 2016).

为一项新的负责任国家行为规范。硬法”约束与“软法”并存的策略，不仅规避了硬性约束，还通过主导规则演变路径，确保美国在未来国际法形成过程中占据主导地位。

五、结语

美国网络空间立场文件在间隔长时间中呈现出演进性和稳定性。美国关于主权的立场，随技术、自身利益和各国的反应进行调整，而关于自卫权的立场一步步得到强化。最后主权和自卫权的交织，铺垫了在美国立场文本中“合法”的“前沿防御”。

对中国而言，美国的行动提供了重要的启示。中国需要持明确的反对态度，指出其法理上的矛盾。同时，中国在维护国家网络安全和推进国际法律地位的过程中，可以学习借鉴美国的思路，采取更为积极和灵活的策略，通过增强与其他国家的合作、积极参与网络空间国际法的制定和改革，来保护自身利益并在国际法中争取更大的话语权。

未来，网络空间的国际法治环境将继续发展和变化，这要求各国不断调整和优化自己的战略以适应这些变化。通过本文的分析，希望能为中国及其他国家在这一过程中提供有益的参考和启示，共同推动构建一个更加公正和有效的国际法网络治理体系。

全球人工智能治理下的发展鸿沟与平权进路

候柔倩*

摘要：人工智能的演进于多个维度重塑了主权概念的内涵，引发数字主权范畴内广泛的竞争与合作。美国凭借其技术优势施行霸权主义策略，欧盟借助严格的监管架构践行规范性治理，中国着力统筹发展与安全的双重目标。在全球人工智能治理的框架下，数据资源的高度集中与垄断、发展机遇的非均衡分配、数字话语权的激烈角逐等数字不平等现象频发。为推进全球人工智能治理向普惠共享的平权模式转型，在权利平等维度，需搭建平衡数据控制权与发展权的制度架构，建立算法公平性的司法审查机制，探索全球算力治理的契约模式；在机会平等维度，应推进数字基础设施的共商共建共享，加大技术转移与援助力度，构建技术风险共担的市场准入机制；在规则平等维度，需通过多边协商推进数字规则制定，重构数字价值链的收益分配规则体系，优化数字主权冲突的法治解决路径。

关键词：人工智能 平权治理 数字鸿沟 普惠共享

一、问题的提出

人工智能作为战略性技术范式，凭借其跨国界、跨文化、跨领域的技术扩散特征，正成为重塑全球治理格局的核心动力。其发展逻辑与人类命运共同体理念存在深层耦合性，一方面，人工智能的技术外溢效应突破传统地缘政治边界，客观上要求构建新型国际合作框架；另一方面，人类命运共同体理念为人工智能治理体系构建提供了价值公约数。作为数字经济时代的核心生产要素，人工智能通过与 5G、区块链、物联网等新一代信息技术的深度融合，催生出多样化的应用场景，重构了社会生产生活方式。人工智能的成果共享，理应使不同发展阶段的国家或地区都能从人工智能的发展中受益，更重要的是，要使其尽可能地惠及每一个人，成为全人类的共同财富。¹基于此，将人工智能治理纳入全球治理体系改革议程已刻不容缓。2023 年 10 月，联合国秘书长宣布成立联合国高级别人工智能咨询机构，并发布《以人为本的人工智能治理》报告，提出建立国际科学小组、启动政策对话、创建标准交易所等七项核心建议。其中，人工智能治理政策对话和国际人工智能科学小组被纳入《全球数字契约》，以规制人工智能技术的发展成果惠及全人类。2023 年 11 月，首届全球人工智能安全峰会召开，包括中国在内的 28 国及欧盟签署《布莱切利宣言》，达成构建开放、公正、有效的人工智能

* 作者简介：候柔倩，女，中南大学法学院博士研究生，主要从事国际法学、数字法学研究。

¹ 参见潘万历：《共同担当人工智能全球治理责任》，载《经济日报》2025 年 3 月 14 日，第 12 版。

治理共识。²2024 年联大通过《抓住安全、可靠和可信赖人工智能系统带来的机遇，促进可持续发展》和《加强人工智能能力建设的国际合作》两份重要决议，强调弥合人工智能鸿沟、推动安全可靠的人工智能系统治理，提升发展中国家能力建设，为实现联合国 2030 年可持续发展议程提供方向指引。2025 年 2 月，人工智能行动峰会在法国巴黎闭幕，包括法国、中国、印度、欧盟在内的多个国家和国际组织共同签署了《关于发展包容、可持续的人工智能造福人类与地球的声明》，强调促进人工智能的可及性以减少数字鸿沟，确保人工智能开放、包容、透明、合乎道德、安全、可靠且值得信赖。³

既有研究探讨全球人工智能治理，多集中于以下三个方面：一是，对全球人工智能治理的理念与模式研究，如张吉豫提出应建设以法治为核心的赋能型人工智能治理机制以及法治统领下的各项具体机制；⁴如张凌寒认为人工智能的法律治理正处在规则形成期，需引入适应性治理理念统筹安全发展。⁵二是，对中美欧人工智能治理政策的研究，如丁晓东指出美国的人工智能立法强调市场主导与企业自我规制，而欧盟则急于发挥布鲁塞尔效应，对人工智能进行统一立法与风险规制；⁶如赵申洪阐明在治理倾向上，欧盟主要关注个人数据权，美国则担心过度监管扼杀创新，中国模式在很大程度上以国家为中心。⁷三是，对全球人工智能治理算力、算法和数据的单线研究，如袁曾认为人工智能治理的立论基础应修正先前以“算法解释”为核心的立法进路，构建以“可控制”为核心的治理标准和体系；⁸如程乐强调在迈向通用人工智能时代进程中，亟待建立多维度动态风险识别和评估体系，完善数据及算法安全规范。⁹目前学界对上述问题的研究，鲜少注目于国际层面国家之间逐步扩大的发展鸿沟和不平等现象。鉴于当前全球人工智能领域的战略博弈已步入关键转折期，本文将主要关注大国间持续升级的技术竞争和主权博弈，以及技术垄断使发展中国家难以公平地参与其中，进一步扩大数字鸿沟，以致对全球创新生态系统的重塑产生恶劣影响。为解决这一问题，亟需推动兼顾发展权与共享权的平权进路构建。

² 倪浩、纪双城：《首届全球 AI 安全峰会聚焦风险管控》，载《环球时报》2023 年 11 月 2 日，第 11 期。

³ 王卫：《巴黎行动峰会聚焦人工智能全球治理问题》，载《法治日报》2025 年 2 月 17 日，第 5 版。

⁴ 参见张吉豫：《赋能型人工智能治理的理念确立与机制构建》，载《中国法学》2024 年第 5 期，第 61-81 页。

⁵ 参见张凌寒：《人工智能法律治理的路径拓展》，载《中国社会科学》2025 年第 1 期，第 91-110 页。

⁶ 参见丁晓东：《全球比较下的我国人工智能立法》，载《比较法研究》2024 年第 4 期，第 51-66 页。

⁷ 赵申洪：《全球人工智能治理的困境与出路》，载《现代国际关系》2024 年第 4 期，第 116-137 页。

⁸ 参见袁曾：《算法应当被解释吗？——人工智能“可控制”的治理向度》，载《法学论坛》2025 年第 1 期，第 130-142 页。

⁹ 参见程乐：《“数字人本主义”视域下的通用人工智能规制鉴衡》，载《政法论丛》2024 年第 3 期，第 3-20 页。

二、全球人工智能治理下数字主权的理论阐释

（一）人工智能时代对主权内涵的影响

1.对传统国家主权空间的冲击

人工智能技术通过数据要素的全球流动与算法模型的跨境部署，对主权的地理边界构成实质性解构。传统主权理论以领土、领海、领空为物理载体，而数字空间的虚拟化特征使得主权管辖范围呈现“去领土化”倾向。数字空间的无界性突破了传统地理边界，数据要素的瞬时跨境流动与分布式存储架构削弱了基于物理空间的排他性治理效能，这种技术特性与主权管辖的时空错配，重构了人工智能时代的主权界线。例如，欧盟《通用数据保护条例》（General Data Protection Regulation，以下简称 GDPR）通过“效果原则”将其数据管辖权延伸至境外数据处理器，即使服务器位于欧盟境外，只要涉及欧盟公民数据即需遵守 GDPR 所设定的标准，这本质上构成对传统属地管辖权的突破。《中华人民共和国网络安全法》则通过数据本地化要求，将数据存储与处理设施纳入主权管辖范畴，形成“数据领土化”实践。此类立法实践表明，主权空间已从物理疆域扩展至数据流动的虚拟路径与存储节点，形成主权空间的双轨制。换言之，物理空间遵循领土原则，数字空间则以数据主权为核心构建管辖权。数字主权实质是国家主权的一种延伸，即将国家主权延伸至传统疆域之外的数字空间，对数字空间的活动及其结果行使权力。¹⁰然而，主权空间的扩展亦引发国际法适用的冲突。例如，美国《澄清境外数据合法使用法案》（Clarifying Lawful Overseas Use of Data Act，以下简称 CLOUD 法案）主张对本国企业控制的境外数据拥有直接调取权，与欧盟 GDPR 第 3 条对于管辖范围的规定形成直接对立，导致数据主权空间产生重叠冲突。究其根源在于，传统国际法框架基于威斯特伐利亚体系设计，难以适应数字空间的多层次治理需求。联合国专家组报告虽表明国家主权适用于网络活动，但始终未明确网络主权的具体边界，导致主权空间在虚拟场域中的适用性存疑。人工智能技术的介入进一步加剧了这一困境。深度学习模型的训练依赖跨境数据流动，而训练完成的算法模型本身可能蕴含他国数据特征，演变为“算法主权重叠”，以致主权管辖权的归属更加模糊。

2.对传统国家主权运作的革新

人工智能技术的深度嵌入促使国家主权运作机制发生结构性革新，其核心在于算法治理对传统科层制权力运行范式的解构与重构。传统主权运作以层级化决策程序、专业化行政分工及书面化规则体系为特征，其效能受限于信息处理速度与决策者理性边界。而人工智能技术通过数据实时采集、算法即时运算与决策自动化执行的三重机制，实现了治理效能的范式转换。从法理维度审视，这一转变体现为治理进程的技术化转向。申言之，主权行使者将部分决策权力让渡于算法

¹⁰ 黄其松：《数字时代的国家理论》，载《中国社会科学》2022 年第 10 期，第 60-77 页。

系统，从而构建起人机协同治理的新型权力架构。例如，欧盟《人工智能法案》通过风险分级制度，将高风险人工智能系统纳入强制性合格评定程序，其创新性在于构建了“技术标准前置审查”机制，使主权规制从传统的行为后果监管转向技术设计阶段的过程控制。此类立法实践表明，算法正在成为主权运作的新兴载体，传统以法律文本为中心的治理模式逐步转向法律和技术的混合规制体系。技术化治理的持续推进，亦对权力合法性构成深层次的结构性挑战。一方面，算法决策显著提升了治理效能与精准度，技术赋能使得主权权威从物理空间的强制威慑，转向数据空间的预防性干预；另一方面，主权运作的技术化推动公民对主权权威的认同从程序正当性转向技术可信性，而算法的“黑箱”特性难免引起公众的信任危机。这种矛盾状况揭示了技术治理的双重属性，它既可能成为强化国家治理能力的有力工具，亦可能对权力合法性根基造成侵蚀。在技术深度重构权力架构的时代语境下，法律体系亟需不断优化，旨在实现治理效能与法治价值间的动态均衡及有机整合。

3.对传统国家主权利能的变迁

人工智能技术的演进正在引发国家主权利能的范式重构，这种变迁不仅体现为权力作用领域的扩展，更涉及权力性质与合法性基础的深刻转型。传统国际法框架下，国家主权利能以《联合国宪章》第2条确立的领土完整、政治独立与不干涉内政原则为边界，其权能行使依赖于物理空间内资源与行为的直接控制。而在数字主权的理论语境中，主权利能向数据规制、技术标准与算力资源三个维度进行结构性延伸。一是，数据控制权的法律构造重塑了主权利能的作用维度。基于数据跨境流动的不可逆性，主权国家通过立法技术将数据资源纳入主权规制范畴，形成数字主权的新型权能形态。当前，数字主权的权能实现已形成双重路径，一方面通过数据本地化强化物理控制，另一方面借助长臂管辖扩展域外影响力。二是，技术标准制定权成为主权竞争的核心场域，其本质是国际法未完全规制的规则空白地带的权力博弈。例如，在5G通信标准领域，知名专利数据公司IPlytics所发布的《5G专利竞赛的领跑者》报告指出华为公司在5G专利排名中遥遥领先，这一技术优势进而转化为国际标准组织中的规则制定话语权，使得技术标准成为“软性主权工具”。此类博弈揭示出主权利能的实现方式已从传统军事与经济硬实力，转向技术规则软实力的制度性竞争。三是，算力资源的战略性分配重构主权利能的物质基础，引发算力地缘政治的新型权力形态。美国2022年对华实施的A100、H100芯片出口禁令，将算力资源转化为技术遏制的战略武器，可视为主权利能在算力维度的新型扩张。作为反制措施，中国“东数西算”国家工程计划构建全国一体化算力网，确立算力设施的自主可控要求，逐步在物理层面对冲外部技术依赖。

（二）人工智能领域的数字主权竞争

1. 技术自主权竞争

技术自主权作为网络主权在技术治理领域的延伸，直接关系到国家在人工智能时代的战略安全与产业主导能力，其法律属性体现为国家对关键技术的排他性管辖与支配权。技术自主权包含对内治理效力与对外防御效力的双重属性：对内表现为国家可自主制定人工智能产业政策、算法监管框架及数据资源供给机制；对外则体现为抵御域外技术封锁、维护跨境技术合作的独立性。¹¹例如，美国通过《芯片与科学法案》限制高端算力芯片出口，本质上构成对技术自主权的战略性运用，该政策不仅使中国人工智能算力基础设施面临供应链断裂风险，更强化了美国在算力资源分配体系中的结构性权力。在实践层面，技术自主权竞争呈现多维度特征。其一，基础技术层竞争聚焦于算法框架与算力基础设施。美国凭借英伟达、AMD 等企业的 GPU 技术垄断，构建起全球人工智能算力供给的技术围墙，而中国通过昇腾系列芯片研发及国家算力枢纽建设，力图实现算力供应链的“去美化”替代。其二，应用技术层竞争表现为产业生态主导权的争夺。欧盟通过《欧洲芯片法案》投入 430 亿欧元重塑本土半导体产业链，旨在降低对美亚技术依赖，其政策逻辑凸显技术自主权与产业安全的深度绑定。究其根本，技术自主权的法律冲突是主权绝对性与技术全球化不可调和的产物，其国际协调亟需构建新型治理范式，可考虑通过技术标准的互操作性设计平衡主权诉求与全球技术协作需求。

2. 数据控制权争夺

数据控制权的法律建构始于欧盟 GDPR 所确立的“数据控制者”概念，其核心在于通过目的与方式决定权的实现来划定数据处理全周期的责任分配。该制度设计将数据控制者定位为数据处理活动的决策中枢，通过义务链设计确保其对数据流向的实质掌控。而美国《加州消费者隐私法案》（California Consumer Privacy Act，以下简称 CCPA）则采取“消费者赋权”路径，通过知情权、拒绝权与删除权等消极权利配置，形成市场调节型的数据控制权体系。这种差异折射出，欧盟强调通过预设规则塑造控制权结构，而美国更依赖事后救济与市场竞争机制。以及，欧盟通过 GDPR 确立充分性认定机制，实质将数据主权边界扩展至境外数据处理者，这种规范性权力的行使引发大国在数据治理领域的制度性对抗。美国 CLOUD 法案所确立的数据控制者管辖原则与欧盟数据存储地标准形成直接冲突，两大经济体在数据主权的主张上具有不可调和性。相比之下，发展中国家在数据控制权争夺中面临着双重困境。一方面，数据本地化政策成为维护数字主权的必要工具，印度《数字个人数据保护法》要求关键个人数据必须境内存储，但这种

¹¹ 赵精武：《人工智能治理中的技术自主权》，载《中国社会科学院大学学报》2025 年第 1 期，第 94-112 页。

防御性策略导致跨国企业运营成本显著增加，客观上形成数字贸易壁垒。另一方面，全球数据流动规则制定权被发达国家垄断，亚太经济合作组织（Asia-Pacific Economic Cooperation，以下简称 APEC）跨境隐私规则体系的认证标准与欧盟 GDPR 存在制度性错配，使发展中国家陷入“规则追随者”的被动地位。

3.规则制定权角力

人工智能国际规则制定权的本质可视为“制度性权力”与“技术性权力”的耦合作用。¹²欧盟通过“布鲁塞尔效应”构建规则主导权，其《人工智能法案》确立的风险分级监管模式将人工智能系统分为不可接受风险、高风险、有限风险和最小风险四个等级，针对不同风险匹配以差异化的规制措施并尤其对高风险人工智能系统应用提出了详细而明确的规制要求。¹³但这种制度优势正面临美国“技术-规则”深层捆绑策略的挑战。2023 年 1 月，美国国家标准与技术研究院公布《人工智能风险管理框架》，用于指导组织机构在开发和部署人工智能系统时降低安全风险，提高人工智能可信度，¹⁴迫使使用美国云服务的企业被动接受其治理规则，这种技术设施规则化路径正在消解欧盟的制度性权力。全球南方国家在人工智能国际规则制定中的系统性边缘化，深刻揭示了国际法体系在权力配置、议程设置与制度包容性方面的结构性缺陷。规则制定权的未来走向呈现出多元体系竞合的特征。中国通过“数字丝绸之路”项目构建的替代性规则网络已初见成效，在《金砖国家数字经济伙伴关系框架》下，中国推动的人工智能治理弹性原则强调发展权优先，允许成员国根据基础设施水平自主设定人工智能伦理标准。这种“差异化合规”模式正在挑战美欧推行的“一刀切”规则体系。规则制定权的终极争夺将取决于三大要素：技术标准的先发优势、制度创新的扩散能力以及数字联盟的构建效能，这要求各国在“技术民族主义”与“全球数字公域”之间寻求动态平衡。

（三）人工智能领域的数字主权合作

1.基础设施共建

人工智能基础设施的共建是数字主权合作的核心环节，关键在于通过技术资源与物理载体的共享，破解全球算力分配失衡的结构性矛盾。根据国际数据公司（IDC）发布的《全球半年度人工智能基础设施跟踪报告》可以发现，全球人工智能基础设施支出呈现显著的地域集中性。美国以 59%的份额占据绝对主导地位，中国以 20%份额紧随其后，而欧洲、中东和非洲地区总和不足 10%。¹⁵这种“算力寡头化”现象不仅加剧了技术应用的南北差距，更导致发展中国家在算法训

¹² 参见王卫华、杨俊：《人工智能的资本权力批判与全球经济正义的追问》，载《广西社会科学》2021 年第 10 期，第 88-94 页。

¹³ 陈少威等：《比较政策研究视野下全球人工智能治理模式的差异、共识与改革启示》，载《中国行政管理》2024 年第 12 期，第 15-24 页。

¹⁴ 徐超：《人工智能监管的三重维度》，载《中国社会科学报》2024 年 12 月 31 日，第 5 版。

¹⁵ Artificial Intelligence Infrastructure Spending to Surpass the \$200Bn USD Mark in the Next 5 years, According to IDC, <https://www.idc.com/getdoc.jsp?containerId=prUS52758624>, last visited on February.20.2025.

练、数据存储等关键环节受制于跨国企业的技术垄断。为弥合数字和智能鸿沟，特别是帮助全球南方在人工智能发展进程中平等受益，中国提出的《人工智能能力建设普惠计划》为基础设施共建提供了制度创新样本。该计划通过南北合作、南南合作和三方合作等形式，重点推进联合实验室建设、产供应链整合及绿色算力中心布局。¹⁶例如，2024 年中非合作论坛北京峰会提出“中非携手推进现代化十大伙伴行动”，至今中国企业帮助非洲国家新增和升级输变电线路 6.6 万公里、电力装机容量 1.2 亿千瓦、通信骨干网 15 万公里，网络服务覆盖近 7 亿用户终端；建设大型数据中心，提供云计算服务，助力非洲数字化转型和人工智能发展。¹⁷此类实践表明，基础设施共建需突破传统援助范式，转向“技术主权共享”与“本地化适配”并重的路径。

2. 治理框架协同

人工智能治理框架的协同本质上是主权让渡与利益再分配的动态博弈过程。当前人工智能治理的国际法体系整体呈现出“硬法缺位、软法先行”的特征。联合国层面尚未形成具有约束力的公约，但《欧盟人工智能法案》、《美国人工智能倡议》及中国《生成式人工智能服务管理暂行办法》已构成三大区域性监管范式。这种“制度区块化”现象导致跨国企业面临合规成本叠加，典型如爱尔兰数据保护委员会曾要求 Meta 推迟使用 Facebook 和 Instagram 用户数据训练大型语言模型的计划，导致 Meta 暂停在欧洲推出人工智能助手。治理协同的突破口在于提炼不同制度的价值公约数。一是，人权保障方面需达成最低共识。联合国高级别人工智能咨询机构提出的“以人为本”原则，要求将《公民权利和政治权利国际公约》第 17 条¹⁸和第 19 条¹⁹嵌入人工智能全生命周期监管。二是，风险分类分级管理成为方法论基础。尽管欧盟采用“四阶分类法”（禁止/高风险/有限/最小风险），而中国侧重“服务分类+场景适配”，但双方均承认医疗诊断、刑事司法等领域的算法需实施特殊准入限制。三是，透明度义务呈现趋同态势。美国国家标准与技术研究院的人工智能风险管理框架与中国的算法备案制度，均要求披露训练数据来源及决策逻辑。这些共性为人工智能领域全球首个具有法律约束力的国际框架公约《人工智能与人权、民主和法治框架公约》的起草提供了法理支点。

¹⁶ 参见《人工智能能力建设普惠计划》，载外交部官网，https://www.fmprc.gov.cn/wjb_673085/zjzg_673183/xws_674681/xgxw_674683/202412/t20241218_11496414.shtml，2025 年 2 月 20 日最后访问。

¹⁷ 参见王珩：《中非人工智能合作：助力非洲可持续发展的新引擎》，载中国日报网，<https://world.chinadaily.com.cn/a/202503/07/WS67ca9955a310510f19eca509.html>，2025 年 2 月 20 日最后访问。

¹⁸ 《公民权利和政治权利国际公约》第 17 条：“一、任何人的私生活、家庭、住宅或通信不得加以任意或非法干涉，他的荣誉和名誉不得加以非法攻击。二、人人有权享受法律保护，以免受这种干涉或攻击。”

¹⁹ 《公民权利和政治权利国际公约》第 19 条：“一、人人有权持有主张，不受干涉。二、人人有自由发表意见的权利；此项权利包括寻求、接受和传递各种消息和思想的自由，而不论国界，也不论口头的、书写的、印刷的、采取艺术形式的、或通过他所选择的任何其他媒介。三、本条第二款所规定的权利的行使带有特殊的义务和责任，因此得受某些限制，但这些限制只应由法律规定并为下列条件所必需：（甲）尊重他人的权利或名誉；（乙）保障国家安全或公共秩序，或公共卫生或道德。”

3.南北技术转移

南北技术转移的实质是知识权力结构的重构，其核心矛盾体现为技术私有化与发展权保障的法益冲突。根据斯坦福大学与 Quid 合作编写的 2024 AI Index Report 可得，2013-2023 年期间全球人工智能初创公司主要集中于美国、中国和英国。²⁰这种创新垄断使全球南方国家陷入“数据原料输出—技术成品输入”的依附性循环。国际货币基金组织制定的“人工智能准备指数”表明，2023 年发达国家这一指数为 0.68，而新兴市场国家和低收入国家分别为 0.46 和 0.32，全球智能鸿沟或导致出现少部分国家收割技术红利，发展中国家沦为原始数据供应地的极端分化。²¹中国主导的《人工智能能力建设普惠计划》为破解该困局提供了制度创新，通过技术转移和能力建设的双轮驱动。其一，创新人力资源培育模式。例如，由肯尼亚马查科斯大学、天津城市职业学院和华为公司三方共建的肯尼亚鲁班工坊，是非洲大陆唯一提供人工智能、云计算和网络安全等数字技能培训的鲁班工坊，设有“智慧城市实践区”“数据中心实践区”“云网融合学习区”和“空中课堂授课区”。²²其二，构建南南技术转移网络。例如，中国—金砖国家人工智能发展与合作中心已于 2025 年 1 月正式启用，用以搭建金砖国家间的人工智能项目合作与市场需求对接平台，促进产业界与学术界的精准对接，推动技术转移与产业合作，助力企业出海。此类实践表明，有效技术转移需突破传统知识产权买断模式，转向知识共享和收益共享的共生范式。

三、全球人工智能治理下大国之间的主权博弈

（一）美国：推行霸权主义的数字主权

1.施行“美国优先”政策难掩垄断野心

在国际格局深度调整、美国国内矛盾凸显的背景下，近两届美国政府均以“美国优先”为核心理念参与全球治理，通过在经济领域重塑经贸架构以维护自身经济利益、在传统安全领域调整军事部署与联盟体系巩固地缘优势、在非传统安全领域依国内政治需求灵活转变立场，其种种举措虽形式有别，但本质皆为维护美国利益，将全球治理塑造为服务于大国竞争的工具。²³基于此，不难得出美国的人工智能战略本质属于其“数字霸权主义”在国际法灰色地带的扩张。2019 年《美国人工智能倡议》虽以促进创新为名，但诸如“Maintaining American Leadership in Artificial Intelligence”等条文均彰显出美国旨在维护其人工智能技术优势地位的核心意图，并随后在行政条令和立法层面进行诸多尝试。根据《2019 财年国防授

²⁰ See Artificial Intelligence Index Report 2024, <https://www.quid.com/2024-stanford-ai-index-report>, last visited on February.20.2025.

²¹ 参见寰宇平：《让人工智能成为全球发展的普惠引擎》，载《人民日报》2025 年 3 月 21 日，第 3 版。

²² 严瑜：《人工智能为中非合作打开新空间》，载《人民日报海外版》2025 年 3 月 20 日，第 6 版。

²³ 参见俞凤：《美国近两届政府参与全球治理的理念与实践》，载《当代美国评论》2023 年第 2 期，第 41-59 页。

权法》要求，美国正式组建国家人工智能安全委员会，随后《2021 财年国防授权法》要求美国制定并实施“国家人工智能计划”，以解决美国人工智能发展面临的一系列问题。2020 年美国出台《生成人工智能网络安全法案》，旨在确保美国制定并实施保护美国价值观和领导力的人工智能战略。2024 年拜登政府通过出台《关于推进美国在人工智能领域的领导地位、利用人工智能实现国家安全目标，以及促进人工智能的安全、安保和可信性的备忘录》等 5 个文件，强调通过创新、负责任的技术开发和治理确保美国在人工智能领域的全球领先地位，保障国家安全和经济竞争力，注重人工智能在国家安全中的应用以及技术的安全性和伦理性，利用其应对全球挑战。2025 年特朗普上台后，成立“星际之门”项目计划投资 5000 亿美元建设人工智能基础设施以推动“美国制造”的人工智能发展，撤销拜登政府相关行政令减少对企业约束，任命人工智能顾问助力决策，利用“能源紧急状态”加速人工智能相关电力建设，这些举措旨在增强美国人工智能领域竞争力、争夺全球领导地位，但也带来环境、能源等方面的潜在风险。²⁴于人工智能领域，美国不仅牢牢把控关键技术、占据行业制高点，且在可预见的短期内，维持着他国难以撼动的垄断地位。美国在人工智能方面的垄断权力，由技术垄断、标准垄断以及理念垄断协同构成，此已成为其谋求全球霸权的战略性工具。人工智能领域的垄断，为美国霸权提供了坚实的技术支撑，并与其他手段相互交织，进而强化了美国霸权的效能。特别是美国将技术垄断与金融垄断、知识垄断、平台垄断等多元手段综合运用，这些因素以复杂且多样的方式相互作用、层层叠加。凭借这一体系，美国得以在全球范围传播西方意识形态与文化价值观，试图对人们的思想认知以及全球资源施加控制，进而按照自身意愿形塑国家及国际制度。

2. 构筑“小院高墙”策略拉大发展鸿沟

美国在人工智能领域推行“小院高墙”策略，其核心逻辑在于甄别对国家安全具有关键意义、自身具备显著优势且中国相对匮乏的人工智能技术，将其圈定为“小院”，并构建严密的保护机制即“高墙”，对“小院”之外的技术则保留一定程度的对华交流空间。从经济维度审视，受“美国学派”经济思想回潮影响，美国为维持在人工智能领域的生产率优势，宁愿牺牲部分绝对经济效益，采取精准的“脱钩断链”政策以抑制中国相关技术的发展，同时加大本国在该领域的科研投入，推动创新体制改革。²⁵在政治层面，拜登政府的干涉主义在人工智能领域体现为构建诸如“民主科技联盟”等排他性多边机制，意图孤立并围堵中国，迫使盟友配合其遏制中国人工智能技术发展的战略。在决策逻辑上，美国延续了对苏、对日“二元并重”的科技竞争战略，在人工智能的高政治领域对中国实施打压遏制，同

²⁴ 《2024-2025 年美国 AI 发展战略风云：从政策布局到激进变革》，

https://www.szaicx.com/page00181?article_id=16339，2025 年 2 月 22 日最后访问。

²⁵ 参见黄日涵、高恩泽：《“小院高墙”：拜登政府的科技竞争战略》，载《外交评论（外交学院学报）》2022 年第 2 期，第 133-154 页。

时考量到与中国经济关系破裂可能对自身产业发展造成的损害，故而采取在关键领域精准脱钩、降低经济依赖等变革性举措。²⁶然而，这一战略存在多重局限性。首先，“小院”边界界定模糊，易引发寒蝉效应，对全球人工智能领域的投资、合作及贸易产生抑制作用；其次，美国肆意的限制超出了确保自身安全的必要范围，反而为中国人工智能公司在大多数国家占据人工智能应用主导地位创造机会；²⁷再次，美国官僚体系对政策落实存在阻碍，政治多元化格局也对该战略形成制衡；从次，美国盟友对参与这一战略持谨慎态度，“民主科技联盟”内部并非坚如磐石；最后，切断技术网络连接将产生“霸权侵蚀效应”，严重威胁美国在全球人工智能技术网络中的主导地位。实践反复证明，美国的“小院高墙”策略不仅违背全球科技协同发展的客观规律，未能达成预期政策目标，反而在一定程度上削弱了美国自身的科技竞争力与产业吸引力，加剧了全球贸易体系的碎片化与不稳定性。

3. 布局数字技术联盟体系强化主导优势

美国通过构建多极技术联盟体系强化其在人工智能领域的主导优势，本质上是以结构性技术权力重构为核心的全球霸权战略，其法律框架与技术治理模式呈现出体系化、分层化与排他性特征。美国通过技术标准垄断、供应链重组与数据治理规则输出，形成对人工智能全产业链的闭环控制。首先，在技术基础设施层面，2024年9月美国国务院发布《全球数字化时代关于海底电缆安全与韧性的联合声明》，声称当前海底电缆领域存在针对国家安全和公众利益的“风险”，需要联合管控，企图将中国企业排挤出全球海缆供应链，进而抢占国际公域规则制定权和话语权。其次，在技术供应链领域，《芯片与科学法案》与“友岸外包”战略相结合，依托“芯片四方联盟”（Chip4）形成半导体产业纵向垄断，试图对内重塑美国民主体制、强化芯片供应链弹性，对外重振同盟体系、遏制中国半导体技术发展，意欲借重塑“民主”认同之名，实现维持技术“霸权”之实。²⁸再次，在技术生态系统构建中，美国发布的《互联网未来宣言》及《外国合作伙伴标准指南》以民主价值观为名，实则重申其在互联网治理问题上的政策，即以意识形态画线，煽动分裂和对抗，破坏国际规则，并试图将自己的标准强加于人，形成对国际技术标准制定权的实质性劫持。²⁹最后，在数字治理规则层面，美国通过《欧美跨大西洋数据隐私框架》与《美日数字贸易协定》中关于算法保密、数据本地化限制等条款，构成对人工智能研发要素流动的单向控制。根据《联合国宪章》第55条，成员国负有促进“较高之生活程度，全民就业，及经济与社会进展”的义务，

²⁶ 参见易继明、陈孟麟：《“开门迎客”：应对美国“小院高墙”战略》，载《科学学研究》2025年第2期，第246-253页。

²⁷ Mark Kennedy, *America's AI Strategy: Playing Defense While China Plays to Win*, <https://www.wilsoncenter.org/article/americas-ai-strategy-playing-defense-while-china-plays-win>, last visited on February.22.2025.

²⁸ 韩召颖、刘锦：《拜登政府组建“芯片四方联盟”的逻辑悖论》，载《世界经济与政治论坛》2023年第3期，第70-91页。

²⁹ 参见裴轶：《美式“双标”践踏数字人权》，载《光明日报》2022年8月21日，第8版。

但美国所主导的技术联盟，导致发展中国家在人工智能的供应链中被边缘化，加剧全球技术鸿沟。2023年5月，美国发布《国家人工智能研发战略规划》虽在新增的战略9中提及“为人工智能研究的国际合作建立有原则和可协调的方法”，但其从不同维度加快推进的“民主科技联盟”却将中国等新兴经济体排除在人工智能标准制定之外，形成事实上的“技术治理隔离区”。这种法律实践的双重性，折射出美国在人工智能治理中“规则制定者”与“规则破坏者”的角色悖论：当其需要维持技术垄断时，便诉诸长臂管辖扩大国内法域外效力，而当多边机制妨碍其利益时，则通过“小多边”机制试图掌控人工智能领域的国际话语权。申言之，美国通过制度性权力垄断，将全球人工智能治理体系扭曲为维护其技术霸权的“私人俱乐部”。

（二）欧盟：践行严格监管的数字主权

1. 通过立法确认高标准监管模式

欧盟在人工智能法律监管范畴率先布局，相继颁布一系列兼具战略意义与监管属性的文件。2018年，欧盟委员会发布的《欧洲人工智能战略》，堪称欧盟人工智能发展的纲领性文献。嗣后，欧盟陆续推出《人工智能协调计划》与《人工智能白皮书》，搭建起人工智能监管的基本架构；发布《可信赖人工智能伦理准则》《人工智能、机器人及相关技术的伦理问题框架》，聚焦人工智能伦理议题，将可信赖人工智能的三大特性归纳为合法性、伦理合规性与稳健性。2021年欧盟发布的《人工智能法》作为全球首部综合性人工智能监管框架，系统性重构了技术开发与应用的合法性边界。一是，《人工智能法》采用横向立法模式，几乎囊括在欧盟市场投放或使用的所有人工智能系统，对人工智能系统的定义进行了清晰明确与细化，确保规制对象的广泛全面性。³⁰同时，法案基于风险分类原则，将人工智能系统划分为禁止、高风险、有限风险和最低风险四类，并针对不同风险等级制定相应强度的监管措施，其核心逻辑在于将《欧盟基本权利宪章》确立的人类尊严、隐私权、非歧视等价值嵌入技术生命周期。对于存在不可接受风险的人工智能系统，法案明确禁止“社会信用评分系统”“实时生物识别监控”等八类应用，并建立动态清单更新机制，确保监管与技术进步同步；针对高风险人工智能系统，则设定详尽的合规要求，包括构建风险管理体系、保障数据质量、满足透明度与可解释性要求、接受人工监督等，从多个维度实施严格监管。二是，在高风险人工智能系统的范围界定上，涵盖了作为产品或产品安全组件且需第三方合格评估的系统，以及附件三所提及的涉及生物识别、关键基础设施、教育、就业等多个重要领域的系统。针对通用人工智能模型新增了重点监管制度，当满足特定条件时，提供者需履行一系列义务，包括通知、评估、报告等，以此防范系

³⁰ 参见王天凡：《人工智能监管的路径选择——欧盟《人工智能法》的范式、争议及影响》，载《欧洲研究》2024年第3期，第1-30页。

统性风险。三是，通过引入监管沙盒机制，为人工智能系统提供商搭建了一个在监管之下进行开发、训练、验证和测试的受控框架。这一机制既有力支持了创新，推动人工智能生态系统的良性发展，又通过明确规则确保在实验过程中对健康、安全和基本权利的潜在风险得到充分缓解，实现了创新与监管的有机平衡，体现出高标准监管对新兴技术发展的规范引导作用。为确保执法威慑力，其对超大型平台设定了严苛的义务，在欧盟市场运营务须满足多方面合规要求，否则将面临巨额处罚。同时，虽然对中小企业设置了优惠和豁免措施，但仍然强调其需遵守相关义务，凸显出欧盟的监管模式整体呈现高标准态势。这种立法实践不仅重塑了欧洲人工智能产业生态，更确立了“技术服从于价值”的治理哲学，为全球数字人权保护提供了制度蓝本。

2.“布鲁塞尔效应”的全球投射

“布鲁塞尔效应”作为欧盟在新兴领域构建监管主导权并实现监管外溢的核心话语与规则塑造机制，充分彰显出欧盟凭借单边监管举措影响全球市场的能力。回溯已呈现“布鲁塞尔效应”的新兴领域，监管主体的市场规模或市场权力、监管能力、严格规则偏好、非弹性监管对象偏好以及采用不可切割标准，共同构筑了“布鲁塞尔效应”得以达成的核心要素与内在逻辑，为深入理解欧盟在新兴领域的监管影响力提供了关键视角。³¹首先，欧盟拥有庞大的消费市场，这一市场优势使其在全球经济格局中占据关键地位。《人工智能法》第2条明确规定：“在欧盟内投放人工智能系统或投入使用人工智能系统或投放通用人工智能模型的提供商，无论这些提供商是在欧盟内还是第三国设立或位于欧盟内或第三国，均需遵守本法规”，这种长臂管辖机制迫使跨国企业因欧盟市场准入压力，主动采纳其风险分类框架，形成事实性国际标准趋同。其次，在长期发展过程中，欧盟不断累积了丰富且成熟的监管经验，具备将市场力量有效转化为实际监管影响力的能力。在竞争法等领域，欧盟多次对谷歌等大型企业的垄断及不正当竞争行为采取严厉措施，开出巨额罚单，充分彰显强大的监管执行力。在人工智能监管方面，欧盟通过设立专门的监管机构，如欧洲人工智能委员会等，确保人工智能技术监管的有效性和法规适用的一致性，进一步强化了其监管能力。最后，欧盟在标准设定上秉持严格要求的理念，在人工智能领域制定了当前全球最为严格的监管规则。《人工智能法》涵盖的监管规则全面覆盖了人工智能系统的开发、使用以及数据治理等多个关键环节，对企业提出了较高的合规门槛。虽然这种严格的规则和标准在一定程度上增加了企业的合规成本，但也变相激励企业提升自身技术水平和管理能力，以适应欧盟的监管要求，对全球人工智能行业标准的整体提升起到了积极推动作用。这种规则输出策略使欧盟在算力与数据资源处于劣势的情况

³¹ 参见苏可桢、沈伟：《欧盟人工智能治理方案会产生“布鲁塞尔效应”吗？——基于欧盟《人工智能法》的分析》，载《德国研究》2024年第2期，第66-88页。

下，仍能通过制度性权力重塑全球人工智能治理秩序。

3.单一数字市场的技术主权突围

尽管欧盟在数字经济发展的初始阶段处于相对劣势地位，但这并不表明其缺乏战略影响力。欧盟采取的策略是将自身的市场优势转化为监管权力，以此参与全球数字生态系统的塑造，展现出其在数字领域的积极作为与战略意图。2021年3月，欧盟委员会正式发布《2030 数字罗盘：欧盟数字十年之路》计划，为欧盟到2030年实现数字主权的数字化转型愿景指出方向，旨在构筑一个以人为本、可持续发展的数字社会，增强欧洲的数字竞争力，摆脱对美国和中国的依赖。该计划将通过构建监管壁垒、产业政策和基础设施三位一体的技术主权来贯彻实行。其一，欧盟在传统数字经济领域的竞争力不足，使其转而通过监管规则输出实现“软实力”扩张。例如，欧盟通过《人工智能法》建立全球首个全面的人工智能风险分级监管框架，并要求所有进入欧盟市场的人工智能系统满足透明度、可解释性和人类监督等伦理标准。这种“预防性监管”模式不仅抬高了技术准入门槛，更通过将伦理规范嵌入技术标准，试图重塑全球人工智能治理范式。其二，面对半导体、量子计算等关键技术的供应链脆弱性，欧盟采取“选择性产业政策”突破技术依赖。在半导体领域，2023年7月欧洲议会通过了包含总额高达430亿欧元配套资金的《欧洲芯片法案》，目标是在2030年前，欧盟在全球半导体制造市场的份额从10%提升至至少20%，并大幅提升当地的芯片制造工艺，建立欧盟的半导体供应链。在量子技术领域，欧盟通过“量子旗舰计划”构建完整创新链条。该计划重点支持量子计算、量子通信、量子模拟和量子传感四大领域，2000-2023年期间所有量子技术领域的专利授权中，欧盟合计占量子技术授权专利总数的44%，位居榜首。³²其三，欧盟意识到数字主权的物质基础在于对数据存储、传输和处理链路的控制。Gaia-X云计划作为欧盟数字基础设施建设领域的重点项目，致力于构建一个以欧洲数据和人工智能驱动的生态系统，降低对国外技术的依赖程度。但该计划在实施过程中发生了导向转变，从最初以提升技术能力为主，转变为以制定标准与规则为主。在缺乏坚实技术实力支撑的情况下，欧盟单纯依靠监管手段，在数字技术相关议题的决策过程中难以掌握主导权。³³其四，在应对中美技术竞争时，欧盟采取“选择性耦合”策略。一方面通过“全球门户”全球基建计划与日本、印度等国家建立半导体供应链联盟，另一方面在“一带一路”框架下与中国开展数字基础设施合作。当前，欧盟的数字主权博弈呈现“规制现实主义”特征。在技术代差难以弥合的领域，通过监管权力重塑竞争规则；在关键瓶颈环节，以产业政策实现技术突破；在基础设施层面，构建排他性数据控制网络。这种三位一体战略虽在短期内提升了欧盟的规则话语权，但其内在张力也不容忽

³² 参见王海芸：《全球量子技术与产业发展态势》，载《光明日报》2024年9月26日，第14版。

³³ 参见宫云牧：《数字时代主权概念的回归与欧盟数字治理》，载《欧洲研究》2022年第3期，第18-48页。

视，严格的监管难免抑制技术创新，而巨额产业补贴面临财政可持续性考验。

（三）中国：统筹发展和安全的数字主权

1. 人工智能治理体系的法治化构建

中国人工智能治理体系的法治化构建以统筹安全与发展为战略导向，通过立法先行、标准支撑、技术赋能、伦理嵌入四位一体的制度设计，形成覆盖网络、数据、算法、算力、应用五大领域的全链条治理框架。在网络领域，以《中华人民共和国网络安全法》（以下简称《网络安全法》）为核心的法律体系确立了关键信息基础设施保护制度，结合《网络关键设备安全通用要求》（GB 40050-2021）等技术标准，构建了从设备安全到内容治理的立体化防线。例如，根据《网络安全法》第 37 条关键数据跨境传输需通过国家安全评估，而《推进互联网协议第六版（IPv6）规模部署行动计划》则从技术底层强化了网络自主可控能力。《中国 IPv6 发展状况白皮书（2024）》显示，截至 2024 年 5 月底，我国 IPv6 活跃用户数达 7.94 亿，移动网络 IPv6 流量占比达 64.56%，为人工智能应用的网络基础提供了安全保障。³⁴在数据领域，《中华人民共和国数据安全法》（以下简称《数据安全法》）与《中华人民共和国个人信息保护法》形成双轮驱动，前者通过数据分类分级制度严控核心数据出境，后者确立知情同意为核心的个人信息处理规则，而地方立法如《深圳经济特区数据条例》率先探索数据要素权属界定，允许数据加工使用权依法交易。在算法领域，《互联网信息服务算法推荐管理规定》首创算法备案与透明度披露制度；《生成式人工智能服务管理暂行办法》进一步要求具有舆论属性或者社会动员能力的算法推荐服务提供者应履行相应的备案手续，并后续相继在北京、上海等地试点监管沙盒的运作模式，允许企业在可控环境测试大模型伦理风险。在算力领域，《数字中国建设整体布局规划》提出算力基础设施“东数西算”工程，配套《新时期促进集成电路产业和软件产业高质量发展的若干政策》等政策文件，2023 年全国算力总规模超 230EFLOPS，其中智能算力规模达到 70EFLOPS，新增算力基础设施中智能算力占比过半，成为算力增长的新引擎。³⁵在应用领域，《移动互联网应用程序信息服务管理规定》第 11 条、第 12 条明确应用程序提供者开展应用程序数据处理活动，应当遵循合法、正当、必要和诚信原则，具有明确、合理的目的并公开处理规则，进而有效遏制人工智能系统内的数据过度采集。³⁶

2. 数据安全与跨境流动的平衡治理

在人工智能技术快速迭代与全球化竞争的背景下，中国针对算法训练数据跨境流动的特殊需求，构建起分类监管和场景豁免的动态治理体系。这一体系以技

³⁴ 参见《我国 IPv6 活跃用户数达 7.94 亿》，载《光明日报》2024 年 07 月 11 日，第 8 版。

³⁵ 参见《国家信息化发展报告（2023 年）》，第 12 页。

³⁶ 参见《移动互联网应用程序信息服务管理规定》第 11 条、第 12 条。

术风险分级为基础,以场景化风险评估为工具,突破了传统数据主权治理中安全与发展二元对立的困境,为人工智能产业的跨境协同创新提供了制度保障。一是,建立人工智能数据分类标准,划定算法安全的技术红线。中国基于《数据安全法》《生成式人工智能服务管理暂行办法》,构建起覆盖国家安全数据、公共数据、企业数据及个人数据的分类分级保护体系。针对不同类别的人工智能训练数据,制定差异化跨境流动规则,国家安全数据严格禁止跨境传输、核心公共数据出境实施严格限制措施、企业数据跨境流动采取审慎许可机制、强化个人敏感数据跨境合规管理,实现风险的分层管控。³⁷数据风险分级分类标准将技术特性纳入监管框架,使数据主权保护深度嵌入算法开发生命周期。以智能驾驶领域为例,对拟传输至境外的高精度地图数据应依法履行数据出境安全评估和对外提供审批程序。³⁸二是,构建全产业链条的跨境数据监管,以促进跨境协同创新。人工智能的数据处理过程涉及众多参与主体,明晰各主体的治理义务至关重要。跨境数据安全义务应全面覆盖产业链各阶段的数据处理主体,以防范训练数据来源方的权益受损。网络数据爬虫企业、合成数据企业以及数据中介机构等数据提供方,均应履行基础层面的跨境数据安全风险,确保数据来源合法、数据处理合规。在应用中,跨境数据安全规制遵循科技企业、行业组织与监管部门的多元主体协同机制。通过多主体协商,制定生成内容行为规范、细分应用场景的数据跨境指引、行业自愿性共识标准以及用户使用行为规范等,防止核心数据及重要数据被非法处理,依靠规则的有效实施引导行业有序发展。

3.凝聚包容可持续发展的国际共识

中国一贯以积极主动之姿,深度参与人工智能全球治理进程,并发挥着关键的引领性作用,为国际社会贡献出独具中国智慧的治理理念。首先,中国的治理理念以发展负责任的人工智能为核心,构建了全球领先的系统性伦理框架。2019年发布的《新一代人工智能治理原则——发展负责任的人工智能》确立了八项基本原则,形成技术伦理、社会价值和治理机制的三维治理范式。2023年10月,中国政府发布《全球人工智能治理倡议》,强调发展与安全并重的系统思维,倡导以人为本、智能向善的普遍共识,弘扬平等互利、尊重人类权益的价值理念,呼吁努力弥合“智能鸿沟”,确保“智能红利”惠及各国。³⁹《全球人工智能治理倡议》将先前的人工智能治理框架升级为国际治理方案,提出11项行动纲领,着重强调了数据主权与技术民主化的平衡机制。其次,针对全球数字鸿沟的结构性

³⁷ 参见张亮、陈希聪:《生成式人工智能背景下的跨境数据安全规制——基于 DeepSeek、ChatGPT 等主流 AI 的思考》,载《湖北大学学报(哲学社会科学版)》2025 年第 2 期,第 120-128 页。

³⁸ 关于印发《上海市智能网联汽车高精度地图管理试点规定》的通知(沪规划资源规〔2022〕9 号)第十七条(数据存储传输):存储高精度地图相关数据的服务器应当设在中华人民共和国境内。存储和传输高精度地图数据,应当使用相关技术保证数据安全。对拟传输至境外的高精度地图数据应依法履行数据出境安全评估和对外提供审批程序。

³⁹ 参见和音:《引领全球人工智能治理的中国强音》,载《人民日报》2023 年 10 月 28 日,第 3 版。

矛盾，中国开创了技术转移、制度赋能和人力资本培育的多维度弥合方案。例如，中国—金砖国家人工智能发展与合作中心，已形成建立技术专利池、构建标准互认机制、推动应用场景共享和培育治理人才的立体化合作模式，有效破解了传统技术援助的单向性缺陷。2024年12月，中国和赞比亚在纽约联合国总部正式建立人工智能能力建设国际合作之友小组，以期之友小组为加强人工智能能力建设合作、完善人工智能治理、弥合数字鸿沟注入新动力。⁴⁰最后，中国始终倡导多元的主体协同治理模式，在2023年7月联合国安理会就人工智能与安全问题举行的会议上，中国代表提出伦理先行、安全可控、公平普惠、开放包容和和平利用人工智能治理五项原则，释明中国始终以高度负责的态度参与人工智能全球合作与治理，携手各国共享人工智能技术成果。⁴¹这一立场在《人工智能全球治理上海宣言》中形成制度性突破，宣言支持联合国发挥主渠道作用，欢迎加强南北合作和南南合作，提升发展中国家的代表性和发言权；鼓励国际组织、企业、研究机构、社会组织和个人等多元主体积极发挥与自身角色相匹配的作用。⁴²

四、大国之间的数字主权博弈加剧数字不平等

（一）数据资源垄断与技术依赖性强化

1. 数据要素垄断构筑行业壁垒

在数字经济与人工智能技术深度融合的背景下，数据要素已成为驱动产业创新的核心资源，数据垄断作为大国博弈的核心工具之一，通过重构国际法秩序中的权力关系，形成系统性治理困境和行业准入壁垒。数据垄断的本质是科技巨头与国家权力的战略合谋。谷歌作为大型科技公司占据全球近90%的通用搜索市场份额，它在搜索市场的统治地位堪称其垄断的证据，⁴³美国借助CLOUD法案将企业数据控制权升格为域外数字治理权，实质上构成以私营部门为载体的数据主权实施策略，进而导致《联合国宪章》确立的领土主权原则被算法架构解构。当前，数据垄断行为呈现出多种复杂形态。一是数据壁垒的构建，人工智能的基础模型训练高度依赖大规模数据集，大型数字平台凭借庞大的用户基数与卓越的数据处理能力，在数据获取环节占据先机，而初创企业若试图收集特定数据，则通常面临较高的成本门槛；二是关键数据的拒绝开放问题凸显，掌握“关键数据”的运营商实施数据封锁策略，限制其他经营者获取数据，构成排他性滥用的典型表现；三是剥削性滥用现象广泛存在，包括但不限于数据收集过程中的不合规操

⁴⁰ 参见和音：《让世界共享人工智能发展红利》，载《人民日报》2024年12月5日，第2版。

⁴¹ 参见《中国代表：人工智能问题凸显构建人类命运共同体的重要性、必要性、紧迫性》，载《光明日报》2023年7月20日，第12版。

⁴² 参见贾平凡：《中国为全球人工智能治理贡献智慧》，载《人民日报海外版》2024年7月13日，第6版。

⁴³ 《美国一地区法院裁定谷歌非法垄断搜索市场》，载新华网，<https://www.xinhuanet.com/fortune/20240806/91578c8e0c56437fa9e75b691f198b48/c.html>，2025年2月28日最后访问。

作、开发商过度回收利用数据导致用户隐私泄露，以及不同相邻市场间的数据交叉滥用等。⁴⁴从国际竞争维度审视，数据要素垄断借由三重机制构筑起行业壁垒。其一，资源控制型壁垒。美国凭借其在信息技术领域的先发优势，依托 CLOUD 法案确立“数据控制者标准”，只要数据控制者为美国企业，即便数据存储于境外服务器，美国也将管辖权延伸至此，进而构建起全球数据单向流向美国的体系。这使得其他国家企业在数据获取与利用方面受到极大限制，阻挡了外部力量对美国所控数据资源的染指。其二，技术标准型壁垒。欧盟通过颁布《数字市场法案》，设立禁止不合理的数据使用限制、实现数据互通和可移植性、提供实时数据访问权限等规定。⁴⁵然而，该规则在制定过程中，与欧洲本土企业长期积累的技术能力深度耦合。非欧盟企业因技术基础、研发路径等方面与欧洲企业存在显著差异，这一规则客观形成针对非欧盟企业的合规阻碍，限制其在欧洲数字市场的拓展。其三，政策工具型壁垒。中国基于数据安全与国家主权考量，通过《数据安全法》《网络安全法》建立数据本地化存储制度，强化对跨境数据流动的监管。这一举措明确了中国在数据领域的主权边界，使得境外企业在中国开展业务时，必须严格遵循中国的数据管理政策，形成区域性的数据主权边界。这三大机制并非孤立运行，而是相互交织、协同作用，致使人工智能产业的全球竞争不再仅仅局限于技术、人才与资本的较量，而是深度演变为数据控制权的制度博弈。

2. 算法依赖压缩市场竞争空间

算法作为人工智能技术的核心驱动力，凭借其技术属性与资源依赖特性，正在全球市场中形成结构性垄断力量，并系统性压缩市场竞争空间。从形成机制看，算法依赖的垄断性源于三重技术经济逻辑，数据和算力资源的高壁垒性、算法迭代的正反馈效应以及技术标准的排他性控制。首先，深度学习等复杂算法的研发高度依赖海量数据与超级算力，据 Lambda，训练 GPT-3 模型的算力成本超过 460 万美元，而中美两国占据全球人工智能领域融资的主要份额，反观绝大部分发展中国家在数据基础设施与技术人才储备上存在显著的发展鸿沟。这种资源密集型特征天然形成市场进入壁垒，使头部企业构建数据采集、算法优化、用户锁定的闭环优势，例如 Meta 通过跨平台数据聚合算法持续强化用户画像精准度，其动态定价系统可实时调整广告竞价策略，形成对中小企业流量获取能力的压制。其次，算法的自我强化机制加剧垄断固化。机器学习模型通过用户行为数据的持续输入优化决策精度，形成数据越多、算法越优、市场份额越大的正向循环，根据亚马逊发布的技术白皮书，升级后的 A9 算法通过引入更细致的用户行为建模，提升了对买家兴趣动态变化的实时响应能力，使得平台累计复购率提升了 27%。最后，技术标准的主导权争夺进一步巩固垄断地位，欧盟《数字市场法案》第 3

⁴⁴ 参见张严：《生成式人工智能垄断风险问题探析》，载《科学决策》2024 年第 9 期，第 203-213 页。

⁴⁵ 参见杨瑜娴：《数据交易平台治理：欧盟制度经验与借鉴》，载《世界社会科学》2024 年第 4 期，第 65-79 页。

条第 1 款将提供核心平台服务的大型企业定义为“守门人”，并在第 3 条第 2 款明确认定标准包括过去三年市值至少达到 750 亿欧元，在欧盟年营业额超过 75 亿欧元，拥有至少 4500 万月活跃用户以及超过 1 万名年度活跃用户，至少在三个欧盟成员国提供核心平台服务等。⁴⁶这类企业通过预装算法系统、限制 API 互操作性等手段来排斥竞争者，以致形成事实上的技术标准垄断。技术垄断的深化直接引发市场竞争空间的系统性压缩，集中表现为准入壁垒刚性化、竞争行为隐蔽化与福利分配失衡化三大效应。其一，市场准入壁垒因算法依赖而呈现刚性特征。如要求服务提供者建立数据标注规则与安全评估机制等常规合规成本，对中小企业不免构成沉重负担，而头部企业凭借规模效应摊薄边际成本。其二，算法黑箱与自主决策能力使垄断行为更具隐蔽性。传统反垄断法在规制算法协同行为时面临取证难题，例如动态定价算法可能通过机器学习自发达成价格同盟，但因缺乏书面协议而难以纳入禁止垄断协议的规制范围。其三，消费者福利与市场竞争活力遭遇双重侵蚀。算法歧视通过用户画像实施差别定价，发展中国家消费者因数字素养不足，更易成为算法操纵的对象，这种结构性失衡使全球数字鸿沟扩大至技术垄断与市场失能的恶性循环。

3. 算力资源配置的结构性失衡

全球算力资源配置的结构性失衡本质上是国际数字权力分配失衡的具象化表现，这种失衡既源于技术垄断与制度壁垒的叠加效应，也折射出传统国际法框架在应对新兴技术治理时的范式局限。这种失衡不仅体现为技术基础设施的物理性差距，更通过数据主权让渡、算法权力异化、平台支配地位固化等机制，将技术差异转化为系统性权利剥夺，构成数字时代新型不平等的法理内核。根据中国信息通信研究院发布的《综合算力评价研究报告（2024 年）》⁴⁷数据显示，截至 2023 年底，全球算力总规模达到 910EFLOPS(FP32)，同比增长 40%，智算规模同比增长 136%。⁴⁸在全球算力供给中，美国、中国、欧洲、日本在全球算力规模中的份额分别为 41%、31%、15%和 4%，美国算力规模快速增长，并拉开与中国的差距，算力规模全球占比超过中国 10 个百分点。⁴⁹这种“算力寡头化”格局直接导致技术能力与数字权利的双重分化。在物理层面，发达国家通过《芯片和科学法案》等立法手段构建技术壁垒，限制高端 GPU 出口；在价值层面，如亚马逊 AWS、微软 Azure 等跨国云服务提供商凭借全球领先的公有云市场份额，迫使发展中国家形成算力依附关系，这种技术支配实质上构成了对《世界人权宣言》第 27 条“分享科技进步利益权”的实质性剥夺。⁴⁹数字不平等的本质在于，算力资源配置失衡重构了传统社会契约中的权利义务关系。当前，算力垄断已突破技术能力差异的范

⁴⁶ Digital Markets Act, CHAPTER II GATEKEEPERS, Article 3 Designation of gatekeepers.

⁴⁷ 中国信息通信研究院：《综合算力评价研究报告（2024 年）》，第 1 页。

⁴⁸ 中国信息通信研究院：《先进计算暨算力发展指数蓝皮书（2024 年）》，第 10 页。

⁴⁹ 参见《世界人权宣言》第二十七条第一款：“人人有权自由参加社会的文化生活，享受艺术，并分享科学进步及其产生的福利。”

畴，演变为新型权力与权利关系的制度性构建。一是，在数据权利层面，算力垄断者通过平台规则将用户数据转化为私有生产资料。数据生产者与数据所有者的分离必然导致数字权利异化，用户虽生成数据却丧失所有权，平台通过算法黑箱实现价值攫取。其二，在算法权力层面，算力优势赋予平台“第二立法权”，平台通过智能合约、推荐算法等工具重构社会规则，形成“算法利维坦”。其三，在发展权层面，算力鸿沟固化“中心-边缘”国际秩序。据 CNBC 报道，随着人工智能竞争的加剧，科技巨头计划 2025 年在人工智能技术上投资高达 3200 亿美元。⁵⁰ 全球提升算力技术的资金多流向高收入国家，致使基础设施与技能发展出现显著差距，发展中国家及其本土初创企业在竞争中处于明显劣势。其背后所蕴含着更深层的制度性危机在于，现行国际法体系难以应对算力垄断引发的权利重构，而这种制度缺位无疑纵容了“算力霸权”的扩张。与此同时，算力资源配置失衡正在重塑社会正义的认知基础。当平台通过算力优势将数据控制权转化为数字资本时，传统法律确立的机会平等原则被异化为算力准入平等，而忽视初始资源禀赋差异的平等恰恰构成实质不公。

（二）发展机会不均持续扩大数字鸿沟

1. 数字基础设施的全球分布不均

国际法规范于数字基础设施分配领域呈现出制度性失灵，其根源在于技术资本主义的兴起与扩张。在全球化进程中，技术资本主义凭借其强大的经济与技术优势，对作为传统国际法基石之一的主权平等原则进行了深刻解构。步入数字时代，技术巨头借助跨境数据流动、算法主导等方式，在全球范围内构建起复杂且具有重大影响力的数字基础设施网络，致使部分国家在数字领域的主权权力行使遭受冲击，国家间在数字基础设施分配中的传统平等地位被打破，由此造成国际法规范在该领域难以充分发挥其应有的调节与规范功能。《国际电信联盟组织法》第 1 条第 1 款与《联合国宪章》第 55 条的规定共同指向“国际社会的协同发展”，然而在技术标准私有化浪潮下，这些规定已异化为权力合法化工具。发达国家借以《专利合作条约》极大地提升了数字基础设施核心技术专利的集中率，使得《与贸易有关的知识产权协议》第 66 条第 2 款⁵¹的技术转让承诺沦为道德倡议。据澳大利亚战略政策研究所数据可得，Google、Meta、Microsoft 和 Amazon 等美国超大规模企业在海底电缆领域的主导地位重塑了该行业，并将关键基础设施置于风险之中。⁵² 海底光缆作为数字时代的基础设施命脉，地缘政治因素使其情况进

⁵⁰ Samantha Subin, *Tech megacaps plan to spend more than \$300 billion in 2025 as AI race intensifies*, By CNBC, <https://www.cnbc.com/2025/02/08/tech-megacaps-to-spend-more-than-300-billion-in-2025-to-win-in-ai.html?msockid=0f463874530c69680c8e2a14523c68c8>, last visited on February.28.2025.

⁵¹ 《与贸易有关的知识产权协议》第 66 条第 2 款：“发达国家成员应鼓励其领土内的企业和组织，促进和鼓励向最不发达国家成员转让技术，以使这些成员创立一个良好和可行的技术基础。”

⁵² See Laura Dobberstein, *Hyperscalers are carving up the ocean floor into private internet highways*, By The Register, https://www.theregister.com/2024/09/25/aspi_hyperscaler_cables/, last visited on February.28.2025.

一步复杂化，政府通常会资助其他国家的电缆，以此在该地区获得影响力。当前，大国几近对互联网核心基础设施和全球数据通信控制权形成绝对控制优势，这也造成对发展中国家数据主权的实质性剥夺。尤其是，美国依据 CLOUD 法案获取全球其他国家数据中心信息的司法实践，使得这一问题面临更大的数据主权争议。诚然，制度重构必须直面国际法在技术权力分配中的价值错位，将基础设施从“贸易商品”重新定性为“人类共同继承财产”。

2. 接入能力分化与使用效能落差

在人工智能技术加速迭代的背景下，国际社会中的接入能力分化与使用效能落差已成为全球数字鸿沟的症结之一。这种分化从技术基础、技术素养与制度支持三重维度展开。其一，技术基础的断层表现为硬件与算力分配的集中化，这一现象促使发展中国家被迫陷入技术依赖困境。国际货币基金组织构建了“人工智能准备指数”，旨在从量化角度对各国引入人工智能技术的准备程度进行评估。数据显示，截至 2024 年 4 月，发达国家的人工智能准备指数为 0.68，新兴国家和低收入国家的该指数分别为 0.46 和 0.32。⁵³通过对这些数据的比较分析可知，新兴国家与低收入国家在人工智能引入准备方面，与发达国家存在显著差距，其准备水平相对滞后。其二，技术素养的极化源于人力资本培育的制度性缺陷，《经济、社会及文化权利国际公约》第 13 条第 1 款⁵⁴关于“教育赋能社会参与”的承诺因人工智能能力框架的推广不足而功能虚置，导致发展中国家在规则制定与技术应用层面双重边缘化。其三，制度支持的断裂表现为创新生态系统的结构性排斥，发展中国家人工智能初创企业存活周期短，面临着跨境专利诉讼成本高、数据主权频频受制于境外云服务、融资难度大等系统性障碍，形成规则、资本和技术的复合性壁垒。同时，使用效能的落差则进一步将技术能力差异转化为实质性的权利剥夺与经济安全风险。在经济领域，高收入国家与低收入和中等收入国家之间的不平等随着创新步伐的加快而扩大，因为前者可以最大程度地受益，因为其技术领先和从事研发的劳动力比例更高。⁵⁵在社会领域，到 2033 年，人工智能的市值预计将达到 4.8 万亿美元，成为数字化转型的重要力量。然而，获得人工智能基础设施和专业知识的机会仍然集中在少数经济体。只有 100 家公司，主要在美国和中国，占全球企业研发支出的 40%。苹果、英伟达和微软等领先的科技巨头的市值都在 3 万亿美元左右，与整个非洲大陆的国内生产总值相当。国家和企业层面的市场主导地位可能会扩大技术鸿沟，使许多发展中国家面临错失其优势的

⁵³ AI Preparedness Index, By International Monetary Fund,

https://www.imf.org/external/datamapper/AI_PI@AIIPI/EME/LIC/ADVEC, last visited on March.2.2025.

⁵⁴ 《经济、社会及文化权利国际公约》第 13 条第 1 款：“本盟约缔约国确认人人有受教育之权。缔约国公认教育应谋人格及人格尊严意识之充分发展，增强对人权与基本自由之尊重。缔约国又公认教育应使人人均能参加自由社会积极贡献，应促进各民族间及各种族、人种或宗教团体间之了解、容恕及友好关系，并应推进联合国维持和平之工作。”

⁵⁵ Clovis Freire, *Is this time different? Impact of AI in output, employment and inequality across low, middle and high-income countries*, Structural Change and Economic Dynamics, Vol.73, 2025, pp.136-157.

风险。⁵⁶在安全领域，发展中国家需要确保他们开发的人工智能系统能够抵御各种安全威胁，包括确保网络安全免受潜在的黑客攻击、数据纵和其他形式的网络威胁。此外，发展中国家应该警惕使用人工智能进行军事攻击以及传播与当地价值观和文化不一致的意识形态。⁵⁷当前，人工智能领域面临着风险全球化却承担本地化的治理悖论。

3.治理能力缺失和创新生态失调

在全球人工智能治理框架下，数智发展鸿沟的持续扩大已演变为系统性权利剥夺与国际法治危机。首先，跨境数据流动监管失效加剧治理能力缺失，欧盟GDPR通过充分性认定机制构建排他性数据圈，迫使发展中国家在数据主权让渡与技术孤立间被动抉择，这种以隐私保护为名搭建起的“数字贸易壁垒”实质违反联合国《发展权利宣言》第8条的发展机会均等原则，促发技术霸权的制度性延伸。其次，算法歧视的司法救济陷入规范真空，算法决策的“黑箱”特性引发系统性歧视风险，但现有国际人权法框架缺乏针对性规制工具。意大利户户送有限责任公司算法歧视案中，博洛尼亚法院做出的判决被认为是世纪性判决，法院首次认定算法在其决策中并不是中立的，而是有意识的、具有辨别能力的，这种算法歧视实然构成对《欧洲人权公约》第14条与《欧洲基本权利宪章》第21条反歧视权的背离。⁵⁸当前，多数发展中国家尚未建立算法透明度强制披露制度，司法审查面临技术鸿沟与证据偏在的双重困境。最后，创新主体间协作失衡、创新资源分配不均和创新环境区域差异显著等问题突出，导致人工智能创新生态的结构性失衡，进一步制约了发展中国家的应用与拓展能力。以及，人工智能领域呈现出“赢者通吃”的竞争格局，这一特性使得反垄断法在推动创新与保障公平竞争的目标之间面临严峻挑战。头部企业依托先入市场所积累的优势及雄厚资源，深度介入专利标准化进程，主导行业标准的制定，构建起具有高度排他性的技术壁垒。在国际层面，发达国家通过《数字经济伙伴关系协定》强化数字知识产权保护，而发展中国家则主张建立人工智能技术共享池。这一分歧加剧了国际规则的碎片化态势，进一步割裂全球创新生态，致使创新资源难以实现自由流通与协同发展。

（三）规则话语权争夺与数字殖民主义

1.数字规则制定的排他性建构

在人工智能快速迭代的时代背景下，国际人工智能治理呈现出“规范性分层”的典型特征。依据《维也纳条约法公约》第31条第1款的善意解释原则，经济

⁵⁶ AI's \$4.8 trillion future: UN Trade and Development alerts on divides, urges action, By UNCTAD, <https://unctad.org/press-material/ais-48-trillion-future-un-trade-and-development-alerts-divides-urges-action>, last visited on April.3.2025.

⁵⁷ See TUHU NUGRAHA, *The Global Landscape of AI Security: A Guide for Policymakers in Developing Countries*, By Modern Diplomacy, <https://modern diplomacy.eu/2024/05/24/the-global-landscape-of-ai-security-a-guide-for-policymakers-in-developing-countries/>, last visited on March.3.2025.

⁵⁸ 参见罗智敏：《算法歧视的司法审查——意大利户户送有限责任公司算法歧视案评析》，载《交大法学》2021年第2期，第183-195页。

合作与发展组织发布的《人工智能原则》与联合国教科文组织发布的《人工智能伦理建议书》虽宣称普适性，但其效力生成机制存在落地难的问题。一是，议程设置的制度性排斥。欧盟《人工智能法案》第 43 条所确立的合格评定机构制度，将标准认证权收归于欧盟认可实体。发展中国家技术机构因未获相互承认协议资格，被迫接受单边技术验证结果。二是，数据主权的规则侵蚀。美国国家标准与技术研究院主导的《人工智能风险管理框架》将数据可追溯性定义为跨境传输的前提，实则通过 CLOUD 法案第 2713 条赋予执法机构域外数据调取权。当《人工智能风险管理框架》的技术合规要求与 CLOUD 法案的域外管辖权形成规制套嵌，将迫使发展中国家让渡数据主权以换取技术准入资格，构成对数据本地化自主权的侵蚀性解释。以美国和欧盟为典型代表，二者在数字规则塑造过程中，均呈现出显著的自我利益导向性。美国凭借其在数字技术领域的先发优势，先后颁布《为人工智能的未来做好准备》《人工智能应用监管指南》等法案，大力倡导企业自我监管模式，旨在构建契合自身利益的宽松治理架构，巩固其在全球人工智能领域的优势地位。欧盟则通过发布《欧洲人工智能战略》《可信人工智能伦理指南》以及《人工智能法案》等举措，着重突出政府的事前监管职能，打造全球人工智能技术监管的国际标准范式。欧美借助国际组织、区域机构以及双边协定等软法手段，将自身价值观与利益诉求深度嵌入数字规则体系。⁵⁹发展中国家在国际规则制定中的参与度较低，缺乏有效的发声渠道，难以充分表达合理的发展诉求，导致在人工智能领域的发展进程中，难以获得公平合理的规则环境支持。

2. 数字价值链的收益剥夺结构

数字技术驱动的全球化重构了传统价值链的收益分配逻辑，形成了以知识产权垄断、算法技术壁垒和数据主权不对称为核心的结构性剥夺体系。这一体系通过国际法规则的技术性包装，将技术优势固化为制度性霸权，导致发展中国家在全球数字价值链中面临系统性权利侵蚀。当下，全球数字价值链的收益分配呈现“中心—边缘”极化特征，其本质是发达国家借助知识产权国际规则与数字技术标准，对发展中国家实施制度性的利益剥夺。国际货币基金组织的国际收支统计数据 displays，大多数发展中国家向第三方支付的费用和特许权使用费远高于他们收到的，技术转让已被视为发展中国家获得发达经济体所拥有的急需但受法律保护的知识资产的主要战略。⁶⁰这种剥夺性结构的法律根源在于现行国际经贸规则对“数字公域”的私有化默许。例如，《与贸易有关的知识产权协定》第 39 条将未披露的信息纳入保护范畴，使得深度学习模型架构等算法核心技术成为企业排他性

⁵⁹ 参见蒋万胜、刘玲霞：《国际数字主权领域内的多元竞合与应对》，载《中国特色社会主义研究》2023 年第 5 期，第 64-73 页。

⁶⁰ See Roya Ghafele, Benjamin Gibert, *Promoting Intellectual Property Monetization In Developing Countries: A Review of Issues and Strategies to Support Knowledge-driven Growth*, By The World Bank, <https://documents1.worldbank.org/curated/en/191191468341930761/pdf/WPS6143.pdf>, last visited on March.3.2025.

资产。跨国企业通过专利常青化策略，延长技术垄断周期，迫使发展中国家陷入技术依附、支付许可费、创新能力萎缩的恶性循环。更值得警惕的是，世界贸易组织争端解决机制至今未将“算法歧视性定价”“数据跨境收益分配”等新型剥夺行为纳入管辖范围，导致发展中国家缺乏有效的国际法救济途径。同时，数据生产要素的“非互惠性”根植于全球数据治理规则的“单向合规”要求。发展中国家缺乏将原始数据转化为智能产品的技术能力，只能以低廉价格出口原始数据，再以百倍成本进口算法模型，形成数据与智能的不对等互换。这种“数字圈地运动”的法律实质，是对数据生产要素的产权制度进行不对称重构。现行规则未建立数据价值贡献度计量标准与收益分配机制，导致数据主权平等在数字空间被实质性架空。

3.长臂管辖与数字新殖民主义

数字时代的全球治理正面临前所未有的法律张力，长臂管辖的扩张性实践、数字殖民主义的系统性渗透与数据主权的渐进消解，共同构成了新型霸权秩序的法律工具箱。当下，长臂管辖已从美国民事诉讼中的“最低限度联系”原则异化为数字霸权扩张的核心武器。其法理悖论在于，一方面，美国通过 CLOUD 法案第 2713 条将数据控制者的“物理存在”标准替换为“数据关联性”标准，使得任何使用美元结算或美国服务器存储的数据均可触发管辖权；另一方面，欧盟 GDPR 第 3 条确立的“目标指向性”标准，通过充分性认定机制将第三国数据保护水平强制纳入欧盟法律框架。这种“双重长臂化”实践实质是将国内法作为“数字治外法权”投射工具，以“数据自由流动”之名行“规则单向输出”之实。当美国依据《芯片与科学法案》对中国芯片研发实施技术封锁，或欧盟法院在 Schrems III 案中单方面裁定第三国数据保护法“充分性缺失”时，本质上是以“合规审查”为名实施制度性胁迫，迫使发展中国家让渡数据治理权以换取技术准入资格。这种法律殖民主义的运作逻辑，恰是数字殖民主义在规范层面的具象化表达，即通过规则输出构建非对称依赖关系，使边缘国家沦为数据价值链的“数字附庸”。例如，在全球云服务市场，美国企业凭借先发优势与技术积累占据主导地位，借助长臂管辖，不受数据实际地理位置限制，最大限度获取存储于云服务器上的各类数据，严重侵犯他国数据主权，进一步推动数字殖民主义在全球的扩张。诚然，数据主权的消解并非技术中立的自然产物，而是数字殖民主义与长臂管辖协同作用下的制度性后果。数据主权兼具着“硬主权”与“软主权”的复合性特征，但世界贸易组织框架下的《服务贸易总协定》仍将数据归类为普通服务贸易标的，忽视其战略资源本质，这种制度性滞后为长臂管辖提供了规范缝隙。

五、迈向普惠共享的全球人工智能平权治理

（一）权利平等：数字发展权的规范化展开

1.平衡数据控制权与数据发展权

国际法义务的重构作为平衡权利的关键，规范创新的终极目标在于实现数据治理从“绝对主权”向“协同发展”的范式转换。在全球人工智能治理框架下，平衡数据控制权与数据发展权的核心路径在于构建法律赋权、技术赋能、利益再分配、治理协同四位一体的制度体系。首先，法律赋权需以国际法重构为起点，将数据发展权纳入《世界人权宣言》修订议程，明确其作为集体人权与个体权利的双重属性。在集体层面，依托《全球数字契约》建立数字基建援助机制，将全球数字服务贸易收入的定向比例投入发展中国家算力基建；在个体层面，强制科技平台向中小企业开放核心数据接口，以去中心化自治组织架构实现数据控制权向生产者的回归。其次，技术赋能需聚焦分布式架构创新。区块链与零知识证明技术的结合可在确保数据主权前提下实现验证信息跨境流动，联邦学习框架则通过“数据可用不可见”机制化解控制权垄断困局，而量子密钥分发网络可帮助发展中国家构建自主可控的数据安全体系。再次，利益再分配机制需突破传统经济模式。基于“共同但有区别的责任”原则，对全球数据收益领先的企业征收累进式数字服务税，并将税收注入全球数据发展基金，定向支持算力基建、人工智能人才培养与安全技术研发；同时建立“数据贡献度计量系统”，根据质量、数量与使用频次量化分配联邦学习框架下的模型收益。最后，治理协同需完善多边机制设计。设立联合国人工智能治理委员会整合碎片化机制，下设技术标准局、争议解决局与发展援助局等核心部门，推动“人类数据命运共同体”理念落地，通过《数字南南合作协定》允许发展中国家以数据抵偿技术引进费用，同时将算法透明度审查与数据垄断行为纳入国际刑事法院所规制的数字犯罪范畴。这一系统性路径的本质是通过法律划定数据主权底线、技术构建价值共享通道、经济矫正结构性失衡、多边重塑治理伦理，最终实现控制权与发展权的平衡治理，使人工智能成为普惠共享的文明力量而非加剧分裂的霸权工具。

2. 构建算法公平性司法审查机制

算法公平性的司法审查本质是程序正义原则在自动化决策场景中的适应性重构，其理论内核在于突破传统正当程序的形式化局限。在实体层面，可扩展适用《国际法院规约》第38条第1款“一般法律原则为文明各国所承认者”的规定，将算法公平性纳入强行法范畴。美国最高法院 *Carpenter v. United States* 案将合理隐私期待标准延伸至数据聚合领域，而欧洲人权法院在 *Bărbulescu v. Romania* 案则通过比例原则划定了算法监控的合法性边界，二者共同奠定了算法告知义务的实质化基准。德国联邦宪法法院“人口普查案”提出公民有关于个人信息的“信息自决权”，又经法国行政法院“公立医院评分案”的判决扩展解释，已衍生出“算法异议权”这一新型程序性权利，即当自动化决策影响基本权利时，相对人有权要求采用可解释人工智能模型重新评估。2021年联合国教科文组织正式推出《人工智能伦理问题建议书》，其中的“算法透明度三级标准”为此提供了国际软法框

架。基础透明度要求算法目的与风险评估的强制披露,可援引《在环境问题上获得信息、公众参与决策和诉诸法律的公约》中环境知情权的法理逻辑;过程透明度需解释关键参数选择与权重设置,可借鉴世界贸易组织争端解决的“合理性说明”义务;结果透明度则对应差异化影响评估报告的定期公开,其审查频率可参照《消除一切形式种族歧视的国际公约》第5条的定期审查机制。为使审查标准客观化,可进一步构建“四阶审查框架”。即,目标正当性须符合“民主社会必需”标准,手段必要性需验证是否存在歧视性更小的替代方案,程序可控性要求人类监督的有效介入,结果可逆性则保障错误决策的及时纠正。这一框架突破了传统比例原则的技术适用瓶颈,使算法审查从技术黑箱转向权利影响评估,最终在比较宪法学层面实现正当程序原则的数字化转型。

3. 重构全球算力治理的契约模式

约翰·罗尔斯的“正义即公平”理论所构建的社会契约论框架,为解决算力治理的全球性困境提供了深层的哲学支撑与制度重构方向。罗尔斯通过“原初状态”与“无知之幕”的假设,要求参与者在剥离既有优势地位的前提下,通过理性协商确立资源分配的正义原则。⁶¹这一理论框架可具象化为三项核心原则,差异原则允许算力禀赋差异存在但需补偿最弱势群体,开放参与机制突破国家中心主义以实现多元主体协同,动态调整条款通过技术演进系数实现制度弹性,这些原则为全球算力治理契约模式的构建奠定了三重理论基石。首先,差异原则的补偿性正义要求转化为分级配额分配机制。可依据国家发展阶段设定差异化义务,由发达国家承担“算力转移义务”,每年将不低于一定比例的算力资源向发展中国家进行帮扶;由新兴经济体履行“算力效能提升义务”,通过技术优化降低算力消耗;由最不发达国家则享有“普惠算力使用权”,优先接入区域算力调度网络。这一设计既承认算力禀赋的客观差异,又通过转移支付与技术赋能实现实质平等。其次,开放参与机制的包容性治理需嵌入技术共享契约框架。可通过强制许可条款允许发展中国家在公共紧急状态下使用专利算力技术,支付合理许可费以平衡知识产权保护与公共福祉;设立国际算力开源社区,要求跨国企业贡献可公开的算力算法代码形成公共技术池,如欧盟《数字市场法》强制科技巨头开放算力接口的模式;同时设立专项资金用于发展中国家算力人才培养与设备采购,解决技术转移机制失效问题。最后,动态调整条款的适应性治理需依托履约保障制度。可建立全球算力登记簿强制披露算力设施位置、能耗及服务对象,结合区块链存证技术固化履约证据,解决现有市场规范碎片化问题;在国际法院下设算力争端仲裁庭,采用技术中立原则裁决跨境争议;对违约国家实施算力出口限制,同时设立算力契约援助基金,提升发展中国家监管能力。此类机制既继承了通过技术转移承诺

⁶¹ 参见高景柱、王培培:《援助义务抑或全球分配正义?——评罗尔斯与世界主义者之争》,载《江苏行政学院学报》2021年第6期,第92-99页。

实现履约的成功经验，又通过动态调整避免主权让渡争议。通过差异补偿、开放共享和动态调整的三维制度嵌套，推进算力治理从零和博弈走向普惠共享的全球公域治理。

（二）机会平等：数字共享权的体系化构建

1. 推行数字基础设施的共商共建共享

数字基础设施的全球均衡布局是实现人工智能普惠发展的物质基础。当前，全球数字基础设施呈现“中心-边缘”分布格局，发展中国家在算法训练、数据存储等核心环节受制于人，形成技术依附性发展路径。为破解这一困局，法律路径创新可从三方面进行突破。第一，构建多边协商机制。借鉴欧盟立法机构正式通过的首部网络安全法《网络与信息系统安全指令》的协同治理模式，在联合国框架下建立数字基础设施规划委员会，通过《全球数字契约》明确各国建设义务。如大多数新兴经济体的政府都将私营部门作为基础设施发展的融资手段，该模式可升级为跨国政府与社会资本间的合作框架，即以恰当的方式安排私人参与人工智能基础设施建设并分担风险，是弥合基础设施差距的可行路径。⁶²第二，完善融资法律工具。可借鉴全球人工智能基础设施投资伙伴关系（GAIIIP）的融资机制，精准定位资金流向，将资金有针对性地投入到人工智能基础设施的新建与扩建项目中。同时，为助力发展中国家缩小与发达国家在人工智能领域的差距，可创新性地设立人工智能发展专项基金，允许发展中国家将自身丰富的数据资产进行专业评估与质押，以此获取低息融资贷款，用以建设和完善本国的人工智能基础设施。第三，强化技术标准互认。例如，2022年达成的《金砖国家数字经济伙伴关系框架》，明确鼓励成员国积极探索在数字基础设施领域开展联合行动。其中，特别强调推动量子技术在各国间的互认进程，并致力于相关标准的设定工作。通过系统性协作，可推动在世界贸易组织的框架体系下，逐步形成具有广泛影响力的全球数字技术互认协定，并建立“等效性审查+风险分级”的双轨认证机制，全方位为数字技术的跨境流通与合作筑牢坚实的制度保障。此路径需突破传统国际法中的技术中立原则，推进达成国际间的数字共享权公约，确立基础设施的“人类共同关切”法律地位，使共商共建共享从道德倡导升格为法律义务。

2. 加大技术转移力度与提供技术援助

为缩小全球智能鸿沟，国际社会应通过构建常态化协作机制，进一步加大对发展中国家的技术援助和技术转移力度。第一，建立强制性技术池。在《人工智能框架公约》中创设必要专利强制许可制度。诸如，医疗影像识别、农业智能监测等关乎民生与基础产业的人工智能技术，可通过认定列为“基础性公共产品”，权利人需以较低的许可费向发展中国家开放。同时，建立国际人工智能技术共享

⁶² See Avtar Singh Jadoun, *Public-Private Partnership in Indian Infrastructure Development*, IJRAR, Vol.9, 2022(4), pp.574-582.

平台，推动负责任技术流动，避免数据壁垒和技术垄断限制发展中国家的技术升级，确保发展中国家能公平利用数据资源。第二，创新技术援助模式。突破传统政府开发援助的捐赠模式，建立技术债务置换机制。申言之，允许发展中国家以碳排放权配额抵扣技术许可费；创设“技术特别提款权”，根据各国数字鸿沟指数分配技术获取额度。此外，全球人工智能合作也需加强国内外企业和学术机构的协作，推动跨国企业共建人工智能产业生态，促进全球人工智能人才交流，避免人才保护主义阻碍技术进步。第三，构建争议解决机制。在世界贸易组织争端解决机制下设立技术转移专门法庭，采用举证责任倒置规则。当发展中国家主张技术封锁时，由技术优势方证明其限制措施的合法性。此路径需修订完善联合国可持续发展技术促进机制的内容，将人工智能技术转移纳入履行国际义务范畴，并对拒不履约国家实施算力资源配额削减等新型制裁措施。在实际推进过程中，国际社会各成员应以开放、公平、合作的姿态，积极参与规则的制定与执行监督。发达国家应摒弃零和博弈的陈旧思维，深刻认识到技术传播能够带动全球经济协同共进；发展中国家则需着力提升技术吸纳与转化能力，更高效地承接技术转移并转化为自身发展的强劲动力；国际组织应充分发挥协调引导职能，定期对制度效果进行评估，根据评估结果动态优化调整，逐步缩小南北技术差距。

3.构建技术风险共担的市场准入机制

人工智能技术的广泛应用带来了伦理风险、算法歧视、数据安全隐患和军事化风险，国际社会需构建一个涵盖技术研发投入、收益分配以及风险承担等多维度的技术风险共担市场准入机制，确保人工智能发展符合全球社会责任，有效缩小不同国家和地区在数字技术获取与应用能力上的差距。第一，构建动态风险评估体系。充分汲取欧盟在人工智能风险监管领域的“四阶风险分类法”，结合全球人工智能发展的现状与趋势，制定普适性强的全球人工智能风险评级标准。在构建该标准时，创新引入“技术脆弱性指数”，从人工智能算法稳定性、数据安全性、技术可替代性等多个维度展开评估，量化技术层面的潜在风险；同步纳入“社会韧性系数”，全面考量不同国家和地区的社会结构、文化传统、政策法规等要素，评估社会对人工智能技术冲击的承受与调适能力，将其作为全面评估的关键参数，保障风险评估兼具科学性与精准性。第二，创新保险法律制度。在《服务贸易总协定》框架下，各国应协同合作，积极创设全球人工智能责任保险池。技术输出方作为关键参与者，需严格按照技术交易额的一定比例缴纳保费。这一比例可依据技术的复杂程度、潜在风险等级等因素科学确定，确保缴纳金额合理且具有风险共担意义。所筹集的保费资金将专款专用，专门用于赔偿发展中国家在技术引进与应用过程中可能遭遇的各类衍生损失，如因技术适配问题导致的额外成本投入、技术安全漏洞引发的数据泄露损失等，切实保障发展中国家在国际技术转移中的权益。第三，完善技术补救制度。建立“算法修正优先权”机制，当人工智能

系统在发展中国家引发系统性风险时，输入国可向技术提供方发出紧急通知，要求其在 72 小时的严格时限内，提交详尽且切实可行的补救方案。该方案需全面涵盖风险评估、问题根源剖析、修正措施步骤以及预期效果等关键内容。若技术提供方未能在规定时间内提交，输入国则有权按照既定程序，暂停该人工智能系统的运行，以避免风险进一步扩大，切实保障本国相关领域的稳定运行与安全。

（三）规则平等：数字命运共同体的法律建构

1. 推动数字规则制定的多边协商机制

数字规则制定的多边协商机制作为数字命运共同体法律建构的核心环节，其规范基础根植于《联合国宪章》确立的主权平等原则及国际法体系中的程序正义要求。当前全球数字规则制定权的结构性失衡已引发严重的治理赤字，发展中国家在规则制定中的失语直接导致现行规则体系与全球利益格局脱节。例如，欧盟《人工智能法案》中“高风险系统”分类标准的设计逻辑完全基于发达国家产业需求，却未纳入发展中国家在农业人工智能系统、教育人工智能系统等民生领域的技术适配性考量。为有效矫正这种权力失衡的局面，构建一个包容性的多边协商机制已成为国际社会亟待解决的重要课题。2024 年出台的《全球数字契约》第 38 条强调，“迫切需要在各级层面加强数据治理合作，确保所有国家能够有效、公平且有意义地参与其中，并积极与相关利益攸关方进行充分协商，以此充分释放数字和新兴技术的全部潜力”。在具体的制度设计层面，可参考世界贸易组织共识决策程序中的双否决权机制，数字规则提案需同时获得发展中国家与发达国家的多数支持方能通过。这一设计旨在平衡不同发展水平国家在规则制定中的权力，避免规则被少数发达国家主导，确保发展中国家的合理诉求能够在规则制定过程中得到充分表达和尊重。从法律建构的具体实现路径来看，依托联合国框架下的制度创新是切实可行的方案。例如，可考虑设立全球数字规则制定委员会作为核心决策机构。该委员会应由来自不同区域、不同发展水平国家的代表组成，通过民主的决策程序，共同商讨制定全球数字规则。同时，为保障委员会决策的有效执行，还需配套建立一套垂直贯通的国内对口机构体系。这些国内对口机构负责将全球数字规则转化为国内政策和法规，并在国内层面推动规则的实施与监督，确保各国在数字治理领域能够与国际规则有效接轨，形成上下联动、协同一致的全球数字治理格局。此外，协商成果的法律转化也需遵循一套科学合理的路径。首先，通过多边协商达成软法共识，这种共识虽然不具有强制执行力，但能够为各国在数字领域的合作提供基本的行为准则和价值导向，促进各国在理念和行动上的初步协调。在此基础上，鼓励区域内国家通过签订区域条约的方式，将软法共识进一步细化和强化，使其在区域范围内具有更强的约束力。随着时间的推移和实践的积累，逐步转化为习惯国际法，从而在全球范围内形成具有普遍强制力的数字治理规则。而针对协商过程中可能出现的僵局，可引入《国际法院规

约》第 38 条“公允及善良原则”设立数字规则调解庭，综合考虑各方利益诉求、数字技术发展趋势以及国际社会的整体利益，提出公正合理的调解方案，推动协商进程的继续进行，避免因僵局导致数字规则制定陷入停滞，保障全球数字治理的持续推进与完善。

2. 重构数字价值链收益分配规则体系

数字价值链收益分配规则体系的重构是破解全球数字鸿沟、实现普惠共享的关键一环，其法理基础根植于《经济、社会及文化权利国际公约》第 15 条第 1 款确立的“享受科学进步及应用之惠”与《世界人权宣言》第 17 条财产权条款的平衡性解释。这一体系的法律建构需突破“数据殖民主义”的桎梏，以贡献、风险和收益三位一体的衡平原则重塑规则。2024 年 7 月联合国贸易与发展会议发布的《2024 年数字经济报告》指出，发展中国家在全球数字行业蓬勃发展的当下，首当其冲地承受了数字化的环境成本，但却没有走在从全球数字经济中获益的最前沿。⁶³申言之，数据所有权与控制权的法律真空导致数字寡头垄断生产价值，却将不利影响转嫁给发展中国家来承担。基于此，理应将数据生产者的源生权利纳入知识产权体系，明确数据采集、清洗、标注等全流程贡献的法律价值。具体规则建构的路径，可围绕四个方面展开。其一，数据贡献度的法律量化。联合国 2025 年国民经济核算体系（2025 SNA）将数据明确纳入资产范畴，再次强调数据绝对的财产价值性，故而可考虑建立区块链智能合约驱动的分布式确权系统，对每一环节的数据贡献值加以明确。其二，收益分配的比例法定化。设定平台企业、数据生产者、主权国家之间的法定分配比例，并借助国际货币基金组织等推动制度的全球化实施落地。其三，税收权属的重构。鉴于传统国际税收规则通常低估发展中国家价值创造的制度缺陷，需将用户参与、数据积累等地域性特殊优势纳入应税利润分配因子。其四，争端解决的司法化。在世界贸易组织争端解决机制框架下增设“数字收益专家组”，适用数字问题争端解决特别程序规则。该程序可突破传统贸易救济的局限，允许对数据滥用、算法歧视等新型侵权行为实施“收益剥夺令”。以及，制度创新的边界需警惕反向制度歧视风险，唯有将法律规则的刚性与数字技术的弹性深度融合，方能在数据文明时代实现人人共享数字红利的平权愿景。

3. 优化数字主权冲突的法治解决路径

数字主权冲突的法治化解决，核心在于调和传统主权原则与数据全球化之间的法理矛盾。法律构建的法理基础应回溯至《联合国宪章》第 2 条所确立的主权平等原则，并吸纳《塔林手册 2.0》中的“实质性影响”管辖理论。制度创新可遵循，其一，实现预防性法治工具的系统化。建立数字影响评估机制，要求重大数字项目提交主权风险报告。例如，对于涉及跨国数据存储与处理的大型云计算项

⁶³ Rebeca Grynszpan, *Launch of the Digital Economy Report 2024*, By UN Trade and Development (UNCTAD), <https://unctad.org/osgstatement/launch-digital-economy-report-2024>, last visited on March.3.2025.

目，项目发起方需详细说明项目实施过程中的数据流向、存储地点及安全保障举措，全面评估对目标国数字主权可能造成的风险。同时，要求企业证明其技术部署不会损害目标国数字主权的完整性，可借助独立第三方审计机构对企业的技术架构、数据管理流程等进行审查，确保企业的技术应用符合目标国主权要求。其二，构建立体化的争端解决机制，形成司法、仲裁、调解三级体系。在国际法院框架内设立数字分庭，依据《国际法院规约》第 26 条受理主权争议案件。该分庭可配备数字法领域的专业法官，运用国际公法及相关数字规则对案件进行公正裁决。在仲裁方面，常设仲裁法院数字仲裁庭适用相关规则，借助区块链存证技术实现跨境证据的即时核验，解决跨境数字争端中证据收集困难、认证复杂的问题，提高仲裁效率与公正性。在调解层面，可建立“技术中立调解员制度”，选拔具备数字技术背景且中立公正的专业人员担任调解员，在充分理解数字技术原理及争议焦点的基础上，推动争议双方达成和解。其三，推动技术赋能的法律化。推广“监管科技”在主权冲突解决中的应用，利用人工智能、大数据分析等技术手段，对数字主权冲突中的海量数据进行快速分析，精准识别冲突根源、评估风险等级，为法律决策提供科学依据，提升主权冲突解决的智能化与精准化水平。整个优化路径以“数字主权边际效益”理论为支撑，致力于形成主权平等原则与数字技术特征深度融合的治理方案，既保障国家数据主权安全，又推动建立多边共治、利益平衡的全球数字秩序，为发展中国家抵御数字殖民、参与规则重构提供制度性解决方案。

六、结语

在全球人工智能技术飞速发展的大背景下，国家间围绕数字主权的博弈和治理权力的再分配，已成为重塑国际秩序的关键议题。联合国贸易和发展组织发布的《2025 年技术与创新报告》指出，2033 年人工智能将发展成为价值 4.8 万亿美元的全球市场，规模相当于德国当前的经济总量，但人工智能技术的发展正在高度集中于少数企业和经济体，全球范围内有 118 个国家几乎完全被排除在人工智能治理讨论之外，其中大多数来自全球南方。⁶⁴当前，全球技术红利的分配呈现出明显的“中心-边缘”结构差异。这种技术快速迭代与经济利益分配不均的状况，直接引发了大国间以数字主权为核心的规则竞争。美国凭借《芯片与科学法案》筑起技术壁垒，欧盟依靠《人工智能法案》强化区域监管主导权，中国则借助“东数西算”工程推进数据主权实践。在此情形下，尽管国际社会已达成《人工智能伦理建议书》等原则性共识，但治理框架的碎片化却加剧了“阵营化治理”的趋势。发展中国家在算力资源、算法话语权以及规则制定中的结构性缺失，使得数字不平等从技术鸿沟演变为系统性权利失衡。人工智能成果的广泛共享，需要

⁶⁴ 参见《联合国报告：人工智能将发展为 4.8 万亿美元市场》，载联合国新闻，<https://news.un.org/zh/story/2025/04/1137251>，2025 年 4 月 3 日最后访问。

突破地缘政治障碍，构建包容性的技术治理框架。作为全球性公共产品，人工智能的发展路径取决于各国能否在技术创新与伦理规制之间实现动态平衡。人类命运共同体理念所倡导的共商、共建、共享原则，为建立普适性国际合作机制提供了理论依据。通过提升技术转化效率和促进跨领域协同创新，充分发挥技术对社会发展的积极推动作用，最终形成兼顾效率与公平的全球治理模式，实现人工智能红利在全人类范围内的共享，确保人工智能始终朝着有利于人类文明进步的方向前行。

2025年中国国际法学会年会

丛林规则还是协同治理？AI 国际治理法律理念的构建

罗早西¹

摘要：在人工智能重塑全球治理格局的进程中，AI 治理面临着丛林规则与协同治理两种法律理念的分野。丛林规则下，技术霸权（如美国《云法案》长臂管辖）和数据殖民（发展中国家数据资源被掠夺）引发法理困境，对“共享”“共建”“共治”原则造成破坏，形成冲突。协同治理则秉持人类命运共同体理念，通过共享发展权（技术普惠）、共建规则（多边参与）、共治风险（责任分担）的三维重构，试图解决全球 AI 治理的公平性与可持续性问题。本文从法理批判入手，分析丛林规则的价值悖反，阐释协同治理的法理内核，并结合中国实践提出治理路径，主张以“共同体法理”取代“霸权法理”，构建全球治理框架，最终迈向共享共建共治的法律文明，为技术时代的全球治理提供中国智慧。

关键词：AI 国际治理；协同治理；共享共建共治

一、问题提出：AI治理的法律理念分野

在人工智能技术重塑全球治理格局的进程中，AI 治理面临着丛林规则下技术垄断、数据殖民等现实冲突的法理困境与人类命运共同体理念下共享发展权、共建规则、共治风险的协同治理路径分野，二者的价值张力凸显了技术时代全球法律治理体系转型的必要性与紧迫性。

（一）现实冲突的法理投射

1. 丛林规则

（1）技术霸权

美国《云法案》以长臂管辖之姿，悍然突破传统属地原则，于世界范围内率先确立网络空间数据管辖领域的“数据控制者”原则。²此行径严重冲击了传统国际法秩序，赋予美国对全球数据进行管辖的“长臂”权力，使得其能够凭借技术优势与国内法，肆意获取并控制他国企业与个人的数据，极大地损害了他国主权与数据安全。

这种技术霸权的法理投射体现了以下几个方面：

其一，管辖权的扩张。传统的属地原则在网络空间面临挑战，《云法案》的长臂管辖使得美国可以对全球范围内的数据进行管辖，突破了传统的地域限

¹ 华东政法大学国际法学院 24 级博士生罗早西

² 参见郭烁：《云存储的数据主权维护——以...断法案规制“长臂管辖”为例》，载《中国法律评论》2022 年第 6 期（总第 48 期）P72-85。

制。这意味着美国可以凭借其技术优势和法律手段，对其他国家的企业和个人数据进行获取和控制。

其二，权力的不平衡。技术霸权导致了权力的不平衡，拥有先进技术的国家可以通过法律手段获取更多的数据资源，从而在经济、政治和军事等方面获得更大的优势。这种权力的不平衡可能会引发国际间的紧张关系和冲突。

其三，隐私和安全的威胁。霸权可能会对个人的隐私和数据安全造成威胁。当一个国家可以随意获取其他国家的个人数据时，个人的隐私和数据安全就难以得到保障。此外，技术霸权还可能导致数据的滥用和泄露，给个人和企业带来巨大的损失。

（2）数据殖民

全球大量的训练数据源自发展中国家，然而这些数据的贡献国却未能获得相应收益分配。这一现象本质上是发达国家利用自身技术与经济优势，对发展中国家数据资源的掠夺，凸显了数据资源分配的极端不公平，同时也反映出国际数据治理规则的严重缺失。

这种现象在法理层面也有着深刻的投射：

其一，资源的不平等分配。数据被视为一种重要的资源，全球大量训练数据源自发展中国家，但发展中国家却没有获得相应的收益分配。这体现了资源的不平等分配，发达国家凭借其技术和经济优势，获取了大量的数据资源，而发展中国家却无法从中受益。

其二，知识产权的问题。数据殖民也涉及到知识产权的问题。发展中国家提供了大量的训练数据，但这些数据的知识产权却往往被发达国家的企业所掌握。这使得发展中国家在人工智能领域的发展受到了限制，无法充分利用自己的数据资源进行创新和发展。

其三，法律和监管的缺失。数据殖民现象也反映了法律和监管的缺失。目前，国际上对于数据的归属和收益分配还没有明确的法律规定，这使得发展中国家在面对数据殖民时缺乏有效的法律手段进行保护。

2.协同治理

秉持人类命运共同体理念，协同治理涵盖共享发展权、共建规则、共治风险三个维度。

（1）共享发展权--技术普惠

在人类命运共同体理念下，共享发展权意味着技术普惠，让不同国家和地区的人们都能享受到技术进步带来的好处。随着全球化的深入发展，技术的传播和应用已经超越了国界，但在现实中，由于经济、社会和政治等多种因素的影响，技术的分配并不均衡。

一方面，发达国家在技术研发和创新方面具有明显的优势，掌握着先进的技术和知识产权。这些国家往往通过技术垄断和知识产权保护来维护自己的利益，

限制了技术在全球范围内的传播和应用。另一方面，发展中国家由于经济实力和技术的限制，往往难以获得先进的技术和资源，导致技术差距不断扩大。

为了实现技术普惠，需要加强国际合作和交流，打破技术垄断和知识产权保护的壁垒。各国应该共同努力，推动技术的创新和应用，让技术更好地服务于人类的发展。同时，国际社会也应该加强对发展中国家的技术援助和支持，帮助发展中国家提高技术水平和创新能力，实现共同发展。

比如，在信息技术领域，互联网的普及和发展为全球信息交流和共享提供了便利。然而，在一些发展中国家，由于基础设施薄弱和技术水平低下，仍然有很多人无法享受到互联网带来的好处。为了解决这个问题，国际社会可以通过提供技术援助和资金支持，帮助发展中国家加强信息基础设施建设，提高互联网普及率，让更多的人能够享受到信息技术带来的便利。

同时，在教育、医疗、环保等领域，也可以通过技术普惠来实现共享发展权。例如，通过远程教育和在线医疗等技术手段，可以让发展中国家的人们享受到优质的教育和医疗资源；通过环保技术的推广和应用，可以实现全球环境的可持续发展。

（2）共建规则--多边参与

在人类命运共同体理念下，共建规则意味着多边参与，各国共同制定和遵守国际规则，以维护国际秩序和促进共同发展。随着全球化的深入发展，国际规则的重要性日益凸显。国际规则不仅是各国行为的准则，也是维护国际秩序和促进共同发展的重要保障。

然而，在现实中，国际规则的制定和执行往往受到少数国家的主导和控制，缺乏多边参与和公平性。一些发达国家往往利用自己的经济和政治优势，制定有利于自己的国际规则，对发展中国家进行限制和打压。这种不公平的国际规则体系不仅不利于发展中国家的的发展，也不利于全球的稳定和繁荣。

为了实现共建规则，需要加强多边合作和协商，让更多的国家参与到国际规则的制定和执行中来。各国应该在平等、互利、共赢的基础上，共同制定和遵守国际规则，以维护国际秩序和促进共同发展。同时，国际社会也应该加强对国际规则的监督和执行，确保国际规则的公平性和有效性。

（3）共治风险--责任分担

在人类命运共同体理念下，共治风险意味着责任分担，各国共同应对全球性风险和挑战，以维护人类的共同利益。随着全球化的深入发展，全球性风险和挑战日益增多，如气候变化、恐怖主义、传染病等。这些风险和挑战不仅影响着各国的安全和发展，也威胁着人类的共同利益。

为了应对全球性风险和挑战，需要加强国际合作和协调，各国共同承担责任。各国应该在平等、互利、共赢的基础上，共同制定和实施应对风险和挑战的策略和措施。同时，国际社会也应该加强对全球性风险和挑战的监测和预警，提高应

对风险和挑战的能力和水平。

比如，在气候变化领域，各国应该共同努力，减少温室气体排放，应对气候变化挑战。发达国家应该承担更多的责任，向发展中国家提供资金和技术支持，帮助发展中国家提高应对气候变化的能力。同时，发展中国家也应该积极采取措施，减少温室气体排放，加强环境保护。

此外，在传染病防控、恐怖主义打击等领域，也需要加强国际合作和协调，各国共同承担责任。例如，在传染病防控领域，各国应该加强信息共享和技术合作，共同研发和生产疫苗和药物，提高传染病防控的能力和水平。在恐怖主义打击领域，各国应该加强情报交流和执法合作，共同打击恐怖主义活动，维护国际安全和稳定。

二、法理批判：丛林规则与协同治理的价值冲突

（一）对“共享”原则的破坏——算力垄断和数据剥削

1. 算力垄断

算力作为现代科技发展的重要支撑，在人工智能等领域发挥着关键作用。算力的垄断会使得少数国家或地区在科技发展中占据了巨大优势。一方面，这限制了其他国家和地区在人工智能等前沿领域的发展。缺乏足够的算力支持，研究机构和企业难以进行大规模的模型训练和创新实验，从而阻碍了科技进步的步伐。例如，在人工智能的研究中，大规模的模型训练需要强大的算力支持。如果没有足够的算力，研究人员就无法进行高效的训练，从而影响到模型的性能和创新。另一方面，算力垄断也可能导致技术壁垒的形成。拥有强大算力的国家或地区可以通过技术封锁等手段，限制其他国家和地区获取先进的技术和知识，进一步加剧了科技发展的不平衡。

此外，算力垄断还可能引发一系列的经济和社会问题。在经济方面，算力资源的不均衡分布可能导致市场竞争的不公平。拥有强大算力的企业可以在市场竞争中占据优势地位，从而挤压其他企业的生存空间。在社会方面，算力垄断可能加剧数字鸿沟的问题。那些缺乏算力资源的地区和人群，将难以享受到科技发展带来的便利和机遇，进一步拉大与发达地区和人群的差距。

2. 数据剥削

数据在当今数字经济时代具有巨大的价值。然而，Meta 非洲用户数据被用于训练商业模型但无利益回馈³的现象，严重破坏了“共享”原则。

数据剥削首先侵犯了用户的权益。用户在使用数字平台的过程中，产生了大量的数据，这些数据包含了用户的个人信息、行为习惯等重要内容。平台在

³ 参见 APOLLO JNR:《元与非洲：真实的爱情故事还是垄断的尝试？》，载于网站 <https://hackernoon.com/lang/zh/%E5%85%83%E5%92%8C%E9%9D%9E%E6%B4%B2%E4%B8%80%E4%B8%AA%E7%9C%9F%E6%AD%A3%E7%9A%84%E7%88%B1%E6%83%85%E6%95%85%E4%BA%8B%E6%88%96%E5%9E%84%E6%96%AD%E7%9A%84%E5%B0%9D%E8%AF%95>

未经用户同意或未给予合理补偿的情况下，将用户数据用于商业目的，这是对用户隐私权和财产权的侵犯。例如，用户的浏览记录、购买行为等数据被用于精准广告投放，为平台带来了巨大的商业利益，但用户却没有得到相应的回报。

其次，数据剥削也破坏了公平竞争的市场环境。拥有大量用户数据的企业可以通过数据分析和挖掘，获得更多的市场信息和竞争优势，从而在市场竞争中占据主导地位。这使得那些缺乏数据资源的企业难以与之竞争，限制了市场的创新和发展。例如，在电商领域，拥有大量用户数据的平台可以通过分析用户的购买行为，为用户提供个性化的推荐服务，从而提高用户的购买转化率。而那些缺乏数据资源的平台则难以提供这样的服务，在市场竞争中处于劣势。

最后，数据剥削还可能对社会的发展产生负面影响。数据作为一种重要的资源，应该为社会的共同利益服务。然而，数据剥削使得数据资源集中在少数企业手中，这些企业往往只关注自身的商业利益，而忽视了数据对社会发展的潜在价值。例如，在医疗领域，大量的医疗数据可以用于疾病的研究和治疗，但如果这些数据被少数企业垄断，就难以实现数据的共享和利用，从而影响到医疗事业的发展。

（二）对“共建”原则的消解——规则制定权垄断

2024 年 3 月，欧盟正式通过《人工智能法案》，这是全球首部全面监管 AI 的法规；而欧洲企业观察站（CEO）的研究表明，大型科技公司在决定如何通过协调标准实施 AI 法案的关键规则方面占据主导地位。⁴

一方面，这种情况可能导致标准制定的不平衡。欧美企业在标准制定中占据主导地位，可能会使标准更倾向于满足他们的利益和技术需求，而忽视其他地区和企业的需求。例如，在某些特定的技术应用场景中，欧美企业可能更注重满足其本土市场的需求，而对于其他国家和地区的特殊需求考虑不足。这可能会限制其他地区企业的发展，降低全球市场的竞争活力。

另一方面，规则制定权的垄断也可能阻碍技术创新的多样性。当少数企业主导标准制定时，可能会形成技术路径的依赖，限制其他创新思路的发展。新的技术和创新往往来自不同的地区和企业，如果标准制定过于集中，可能会扼杀那些具有潜力的创新想法。例如，在人工智能领域，不同国家和地区的企业可能会有独特的技术创新方向，但由于标准制定权的垄断，这些创新可能无法得到充分的体现和推广。

此外，规则制定权垄断还可能引发贸易摩擦和技术壁垒。当标准主要由欧美企业主导时，其他国家的企业可能需要花费更多的成本来适应这些标准，或

⁴ 参见《Setting the rules of their own game: how Big Tech is shaping AI standards》，载于 <https://www.corporateeurope.org/en/2025/01/setting-rules-their-own-game-how-big-tech-shaping-ai-standards>

者在进入国际市场时面临更高的门槛。这不仅会影响全球贸易的公平性，也会阻碍技术的全球传播和应用。

（三）对“共治”原则的挑战

1. 责任规避

ChatGPT 跨境侵权事件⁵中，开发者、平台、用户三方责任界限模糊，难以追溯。这暴露出当前法律体系在应对 AI 跨境侵权时的滞后性与局限性，使得侵权行为难以得到有效规制，严重挑战了 AI 治理中的共治原则。

（1）开发者责任难以确定

ChatGPT 的开发者在跨境侵权事件中，其责任的界定存在一定难度。一方面，开发者可能会声称其只是提供了一种技术工具，对于用户如何使用该工具所产生的侵权行为不应承担直接责任。例如，在一些情况下，用户可能会利用 ChatGPT 生成的内容侵犯他人的版权，但开发者可能会认为他们无法控制用户的具体行为。另一方面，由于人工智能技术的复杂性，确定开发者在设计和开发过程中是否存在潜在的缺陷或故意引导侵权行为也并非易事。如果开发者在算法设计中没有充分考虑到可能的侵权风险，或者没有采取足够的措施来防止侵权行为的发生，那么他们是否应该承担相应的责任呢？目前，对于这些问题尚未有明确的法律规定和司法实践标准。

（2）平台责任难以明确

平台在 ChatGPT 跨境侵权中也面临着责任难以明确的问题。平台通常会声称自己只是提供了一个交流和信息传播的场所，对于用户上传和生成的内容不承担直接责任。然而，平台在一定程度上具有监管和管理的义务。例如，平台应该采取合理的措施来防止侵权内容的传播，及时处理用户的举报和投诉，并对违规用户进行处罚。但是，在跨境环境下，不同国家和地区的法律规定和监管要求存在差异，平台很难在全球范围内统一履行这些义务。此外，平台还可能面临技术上的挑战，如如何有效地识别和过滤侵权内容，以及如何应对大量的用户生成内容。

（3）用户责任难以追溯

在 ChatGPT 跨境侵权中，用户的责任也难以追溯。用户可能会利用 ChatGPT 生成的内容进行侵权行为，例如抄袭他人的作品、侵犯他人的商标权等。然而，由于互联网的匿名性和跨国性，很难确定具体的侵权用户身份。即使能够确定用户身份，不同国家和地区的法律对于用户侵权行为的认定和处罚也存在差异。此外，用户可能会声称自己并不知道使用 ChatGPT 生成的内容会

⁵ 参见《ChatGPT 侵权风波：德国音乐组织起诉，AI 版权争议再起！》，德国音乐版权管理组织 GEMA 于 2024 年 12 月向慕尼黑法院提起诉讼，指控 OpenAI 未经许可复制德国作者的歌词用于训练模型，并要求明确欧洲 AI 提供者支付版权许可的义务。此案可能成为欧洲首例由集体管理组织起诉 AI 公司的案件，判决结果将影响合理使用规则在 AI 领域的适用性。载于 https://www.sohu.com/a/833434161_122066679。

构成侵权，或者认为自己只是在合理使用范围内使用了该内容。这些因素都增加了用户责任追溯的难度。

（四）协同治理的法理内核：共享、共建、共治的三维重构

共享、共建、共治的三维重构，本质是法律对社会治理规律的制度回应。在国家治理现代化进程中，需以权利保障为逻辑起点，以程序正义为实现路径，以责任法治为根本保障，构建“主体平等、规则透明、资源共享、责任共担”的协同治理法律体系。这不仅是应对复杂治理难题的现实选择，更是将中国特色社会主义制度优势转化为治理效能的法理升华，为人类社会治理贡献中国智慧与制度方案。

四、结论：迈向共享共建共治的法律文明

（一）范式突破

以“共同体法理”全面取代“霸权法理”，在技术时代确立全球契约原则。摒弃单边主义与霸权行径，倡导各国基于平等、合作、共赢构建全球 AI 治理新秩序，重塑技术时代的国际法理基础。

1.“共同体法理”的内涵与意义

“共同体法理”强调各国在平等、合作、共赢的基础上共同构建全球治理秩序。在技术时代，尤其是人工智能迅速发展的当下，这一法理具有重要意义。随着人工智能的不断进步，其对全球的影响日益加深，人工智能帝国时代，其背后的压迫系统可能会加深和加剧科技行业的非人性和有害特征。因此，需要一种新的法理来规范人工智能的发展，确保其为人类带来福祉而非灾难。

“共同体法理”摒弃单边主义与霸权行径，这与当前国际形势密切相关。在全球化的背景下，各国之间的联系日益紧密，任何国家都无法独善其身。单边主义和霸权行径不仅会破坏国际秩序，也会阻碍全球的发展。中国近年来在多边外交场合积极倡导的“以国际法为基础的国际秩序”，与某些西方国家或地区所宣扬的“基于规则的国际秩序”形成鲜明对比。“基于规则的国际秩序观”存在合法性与正当性赤字，同时在实践中有被滥用的法律风险。而“以国际法为基础的国际秩序观”则能够彰显国际法与国际秩序之间的亲缘关系，发扬国际法在国际关系中的规制优势，亦有助于确保国际秩序的多边性、统一性和民主性。

2.全球契约原则在技术时代的重要性

全球契约原则在技术时代确立具有重要意义。随着技术的飞速发展，尤其是人工智能等新兴技术的崛起，全球面临着新的挑战 and 机遇。全球契约原则可以为各国在技术领域的合作提供一个共同的框架和准则。

例如，在人工智能治理方面，近期人工智能的发展以及其使用的增加，使得加强全球人工智能治理成为新的重点。然而，目前国际上的人工智能治理仍

存在不足，如治理赤字、现有倡议的不充分、治理格局中的差距以及在达成更合适机制方面的困难等。为了加强全球人工智能治理，可以采取两种途径：一是发展新的集中式国际人工智能机构；二是加强现有机构之间的协调和能力。作者认为，加强现有国际机构的弱“制度复合体”是更具政治合法性和可行性的前进道路。

3.构建全球 AI 治理新秩序的挑战与机遇

构建全球 AI 治理新秩序面临着诸多挑战。一方面，各国在技术发展水平、利益诉求等方面存在差异，这可能导致在治理规则的制定和实施过程中出现分歧。另一方面，人工智能技术的快速发展使得治理规则的制定难以跟上技术的步伐，存在一定的滞后性。

然而，构建全球 AI 治理新秩序也带来了机遇。首先，各国可以通过合作共同应对人工智能带来的挑战，实现共赢。其次，全球 AI 治理新秩序的建立可以促进技术的健康发展，为人类社会带来更多的福祉。

为了促进正义，我们不仅必须从根本上改变技术本身，而且必须从根本上改变我们对技术的看法，并从那些最有可能受到伤害的人的角度自下而上地发展它。这意味着在构建全球 AI 治理新秩序的过程中，需要充分考虑到不同国家和群体的利益，确保治理规则的公平性和有效性。

4.重塑技术时代国际法理基础的路径

为了重塑技术时代的国际法理基础，需要采取一系列措施。首先，各国应加强合作，共同推动国际法的发展和完善。在技术领域，各国可以通过签订国际条约、制定国际标准等方式，规范技术的发展和應用。其次，国际社会应加强对新兴技术的监管，确保其安全、可靠、可持续发展。数据跨境流动需要在法理层面反思，应将数据安全主权作为数据跨境流动的法理基础。数据安全主权秉持安全不可分割与人类命运共同体原则，既平等回应各国对数据安全的合理关切，又反对单边主义与霸权主义，可以成为国际数据跨境流动的重叠共识。

此外，国际社会还应加强对国际法的宣传和教育，提高各国对国际法的认识和遵守程度。只有各国共同努力，才能重塑技术时代的国际法理基础，实现全球的共享共建共治。

（二）制度抓手

1.共享：制定《人工智能技术转让协定》纳入 WTO 规则体系

在技术全球化的背景下，制定《人工智能技术转让协定》并将其纳入 WTO 规则体系具有重要的意义。一方面，这将促进人工智能技术在全球范围内的合理流动和共享，提高全球各国的技术水平和创新能力。另一方面，这也将有助于避免技术垄断和不公平竞争，促进全球贸易的公平、合理、可持续发展。

制定《人工智能技术转让协定》需要考虑到各国的不同利益和需求，确保协定的公平性和可操作性。协定可以包括技术转让的范围、条件、价格等方面的规定，同时还可以建立技术转让的监督和评估机制，确保技术转让的顺利进行。

2. 共建：推动《全球 AI 治理公约》缔约谈判

推动《全球 AI 治理公约》缔约谈判是构建全球人工智能治理体系的重要举措。《全球 AI 治理公约》将为全球各国在人工智能治理方面提供一个共同的框架和准则，促进各国在人工智能治理方面的合作和协调。

《全球 AI 治理公约》可以包括人工智能的发展和应用原则、安全和伦理标准、监管机制等方面的内容。公约的制定需要充分考虑到各国的不同文化、法律和社会背景，确保公约的普适性和可操作性。同时，公约的制定还需要充分听取各方面的意见和建议，确保公约的科学性和合理性。

3. 共治：建议设立世界人工智能组织实施算法审计

设立世界人工智能组织实施算法审计是加强全球人工智能治理的重要手段。该组织可以作为一个独立的国际组织，负责监督和评估全球各国在人工智能治理方面的情况，推动各国在人工智能治理方面的合作和协调。

算法审计将对人工智能算法的安全性、公正性、透明度等方面进行评估和监督，确保人工智能算法的发展和应用符合人类的利益和价值观。同时，算法审计还可以为各国在人工智能治理方面提供技术支持和咨询服务，促进全球人工智能治理水平的提高。

（三）文明意义

将儒家“天下大同”哲学融入 AI 治理，构建多元包容的法理基础

儒家“天下大同”思想强调人类社会的和谐、平等、互助和共同发展。在 AI 治理中，多元包容的法理基础将强调不同文化、法律和社会背景的国家 and 地区之间的平等对话和合作，促进全球人工智能治理体系的公平、合理、可持续发展。同时，多元包容的法理基础还将强调人工智能技术的发展和应该符合人类的共同利益和价值观，避免技术的滥用和负面影响。

总之，迈向共享共建共治的法律文明是人类社会在技术时代的必然选择。通过范式突破、制度抓手和文明意义的构建，我们可以为全球人工智能治理体系的建设提供重要的理论和实践支持，推动人类社会的和谐、稳定、可持续发展。

“规模与效果”路径在网络空间国际法适用中的扩张及方法论反思——基于国家立场声明的考察

方镇邦*

摘要：传统国际法适用于网络空间时，违法性门槛的界定存在不确定性。《塔林手册》将原本作为“使用武力”与“武力攻击”区分要素的“规模与效果”引入网络空间，成为了确定二者门槛的共同路径。该路径受多国青睐，在国家立场声明中被广泛援引，并在推动不干涉原则与主权原则适用于网络空间的进程中呈现出多样化形态。“规模与效果”路径可视为一种基于后果的国际法类比适用的方式，虽具备一定形式权威，能促进国际司法适用与国家初步共识的达成，但也面临逻辑层面说服力不足的内在缺陷，可能导致网络空间国际法适用的不对等、规则可预期性的削弱及其价值宣示功能的减损。我国需要警惕该路径在各个层面的扩张，分情况对待，并积极主张制定新国际条约作为补充。

关键词：规模与效果 禁止使用武力原则 不干涉原则 主权原则 类比推理

网络新兴技术的发展与跨国网络行动的频发给国际法规制带来了挑战。虽然传统国际法能够适用于网络空间在国际上已达成共识，¹但该如何适用仍面临较大不确定性，这也成为主权博弈与法律辩论的焦点。越来越多的国家和地区以立场文件的形式表达了对国际法适用于网络空间的看法。²这些文件不仅反映一国的政策主张，还可能构成网络空间国际法适用的国家实践与法律确信，³影响网络空间习惯国际法的形成与发展。在国家立场声明中，三项国际法原则尤其受到关注，分别是禁止使用武力原则、不干涉原则与主权原则。在上述原则违法性门槛的判定上，聚焦于网络行动后果的“规模与效果”（scale and effect）路径有着强大生命力，展现出了明显的演进与扩张趋势，映射出传统国际法在向网络空间延伸过程中所经历的重要重变革。

* 方镇邦，武汉大学国际法研究所博士研究生。

¹ 2013年6月，联合国信息安全政府专家组(GGE)达成了一份具有历史性意义的文件，确定了国际法在网络空间的适用，这一共识也被后来的开放式工作组(OEWG)所继承。

² 截至2025年3月28日，有35个国家和地区发表了立场文件，其中包括发布共同立场文件的非盟（含55个成员国）与欧盟，需注意的是欧盟既有共同立场，其成员国也有各自立场。另外，为了表述上的方便，本文将国家联盟发布的共同立场文件也统称为国家立场声明文件。

³ 参见黄志雄、罗旷怡：《各自为“辩”：网络空间新近国家立场声明的总体考察与中国因应》，载《云南社会科学》2023年第6期，第98页；See also Ori Pomson, *Methodology of identifying customary international law applicable to cyber activities*, 36 Leiden Journal of International Law 1023 (2023).

目前,国内外对于“规模与效果”路径的研究主要集中在禁止使用武力原则适用的层面,⁴尚未对其演进与扩张做出系统性考察,因而也难以窥得此路径背后的逻辑张力及其对网络空间国际法发展的深远影响。对此,本文将以网络空间国家立场声明为主要分析对象,描绘其“规模与效果”路径在网络空间的发展轨迹并对其进行规范性反思,以期在网络空间的国际法治建设提供参考,并为我国在国际网络空间治理中的立场表达与规则塑造提供理论支撑。

一、“规模与效果”路径的起源:从“尼加拉瓜案”谈起

在探究“规模与效果”路径在网络空间的适用之前,有必要回溯“规模与效果”路径的传统国际法根源,厘清其在国际法中的原初形态。从词源上来看,“规模与效果”最早出现在国际法院 1986 年“尼加拉瓜诉美国军事行动案”(以下简称“尼加拉瓜案”)的判决之中。法院在该案中对于许多国际法上的基本原则,如禁止使用武力原则、不干涉原则、主权原则等均作出了开创性的解释。后续实践在尼加拉瓜案的基础上有了一定发展,但需注意的是,该中关于“规模与效果”的判决本身就招致了众多质疑。

(一) 作为区分“使用武力”与“武力攻击”的要素被提出

“规模与效果”直接源自国际法院在“尼加拉瓜案”中对禁止使用武力原则的阐释,最初是作为区分“使用武力”与“武力攻击”的关键因素。在该案中,尼加拉瓜指控美国政府指使军人和拉丁美洲国家民众在尼加拉瓜港口布雷、破坏尼加拉瓜的石油设施和海军基地、侵犯尼加拉瓜的领空主权等军事和准军事行动,认为这些行为违反了禁止使用武力原则。而美国则主张尼加拉瓜对周边国家实施了跨境军事活动等行为,构成武力攻击,美国有权行使集体自卫权。本案争端双方以及国际法院均认可习惯国际法中的禁止使用武力原则与《联合国宪章》中相关条文在内容上的一致性。国际法院在审理过程中强调,有必要区分最严重形式的使用武力(构成“武力攻击”从而触发自卫权的行为)与其他更轻程度形式的使用武力。⁵最终,国际法院判决尼加拉瓜的跨境行动不构成使用武力,并在判决书中做出了著名论述:“在习惯国际法中,禁止武力攻击可能适用于一国对另一国派遣武装部队的情形,如果这种行动,因其‘规模与效果’,可以被定性为武力攻击,而不仅仅是由寻常武力所实施的边境事件。”⁶

由此可见,^{国际}法院在该案中明确认为,“规模与效果”是区分寻常武力所实施

⁴ 如黄志雄:《国际法视角下的“网络战”及中国的对策——以诉诸武力权为中心》,载《现代法学》2015 年第 5 期;张华:《网络空间适用禁止使用武力原则的法律路径》,载《中国法学》2022 年第 2 期;Marco Roscini, *Cyber Operations as Use of Force*, in *Research Handbook on International Law and Cyberspace*, Nicholas Tsagourias & Russell Buchan eds., Edward Elgar Publishing Limited, 2021, p. 300 等。

⁵ *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Merits, Judgment, I. C. J Reports 1986, para. 191.

⁶ *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Merits, Judgment, I. C. J Reports 1986, para. 195.

的边境事件与武力攻击的因素。也因此，“规模与效果”被认为是区分一般情况的使用武力（*use of force*）与武力攻击（*armed attack*）门槛的要素。⁷但遗憾的是，国际法院在该案中并没有对“规模与效果”的判断标准做出进一步的阐释。

（二）后续实践对“尼加拉瓜案”的继承与批判

“尼加拉瓜案”确立了“使用武力”与“武力攻击”相区分的法律性质，但后续国际司法实践在采纳这一区分时，并未统一强调“规模与效果”。如在 2003 年的“伊朗诉美国石油平台案”⁸（以下简称“石油平台案”）、2005 年“刚果诉乌干达军事活动案”⁹、2005 年的“厄立特里亚和埃塞俄比亚索偿案”¹⁰等案件中，国际法院与特设索偿委员会均认可了国际法院在“尼加拉瓜案”中关于“使用武力”与“武力攻击”的界分的观点，认为达到“武力攻击”的所需严重程度比“使用武力”更高，但并未明确援引“规模与效果”作为判断标准。¹¹在“石油平台案”中，国际法院甚至在裁决中表示，不排除给一艘军舰布雷的行为足以（构成“武力攻击”从而）触发自卫权的行使。¹²

同时，“规模与效果”在提出之初便引发了诸多法理与实践层面的质疑。一方面，“规模与效果”的国际法基础受到质疑。质疑声认为，尽管《联合国宪章》中“使用武力”（*armed attack*）与“武力攻击”（*use of force*）确属不同表述，但以此来判定自卫权的行使需达到较高严重程度“武力”过于表面。¹³从文义解释来看，“武力攻击”也可以是“使用武力”，措辞上的区别并非是立法者有意为之。¹⁴同时，国际法院在依据《联合国宪章》阐明禁止使用武力原则的习惯国际法时，未充分参考国家实践，“使用武力”与“武力攻击”之间的真实区分可能并不如国际法院所描述那般。¹⁵一国攻击的严重程度或许影响防卫性武力的适当范围（即比例原则的行使），但与自卫权行使的先决条件无关。¹⁶美国提出了相似质疑，其认为国际

⁷ See Marco Roscini, *Cyber Operations and the Use of Force in International Law*, Oxford University Press, 2014, p. 72.

⁸ Oil Platforms (Islamic Republic of Iran v. United States of America), Merits, Judgment, I. C. J Reports 2003, para. 72.

⁹ Armed Activities on the Territory of the Congo (Democratic Republic of Congo v. Uganda), Merits, Judgment, I. C. J Reports 2005, para.146.

¹⁰ Eritrea–Ethiopia, Partial Award, Jus Ad Bellum Ethiopia’s Claims 1–88, Eritrea–Ethiopia Claims Commission, 19 December 2005, para. 12 (emphasis added).

¹¹ 在“刚果诉乌干达案”中，国际法院避开对争议行为是否构成武力攻击作出判断，并认为即使争议行为构成武力攻击，乌干达也没有足够证据证明行为可以归因于刚果；在“石油平台案”中，国际法院认定不够成武力攻击的原因是美国无法证明伊朗有特定的攻击意图；而在“厄立特里亚和埃塞俄比亚索偿案”，特设委员会用了“量级”一词来区分“使用武力”与“武力攻击”。

¹² Oil Platforms (Islamic Republic of Iran v. United States of America), Merits, Judgment, I. C. J Reports 2003, para. 72.

¹³ John Hargrove, *The Nicaragua Judgment and the Future of the Law of Force and Self Defense*, 81 American Journal of International Law 135, 139 (1987).

¹⁴ See Thomas Eaton, *Self-Defense to Cyber Force: Combatting the Notion of 'Scale and Effect'*, 36 American University International Law Review 697, 720-722 (2020).

¹⁵ See Thomas Eaton, *Self-Defense to Cyber Force: Combatting the Notion of 'Scale and Effect'*, 36 American University International Law Review 697, 755-756 (2020).

¹⁶ See William H. Taft IV, *Self Defense and the Oil Platforms Decision*, 29 Yale Journal of International Law 295, 300 (2004); Värk, René, *The Legal Framework of the Use of Armed Force Revisited*, 15 Baltic Security & Defence Review 56, 69 (2013).

法院在“尼加拉瓜案”中以“规模与效果”为借口，人为地制造了二者之间的界限，不合理地提高了“武力攻击”的门槛。¹⁷

另一方面，也有观点质疑“规模与效果”的清晰度与实用性。有学者批判该判决并没有清楚阐述需要何种程度的“规模与效果”才能达到武力攻击的要求，以及在一项行动处于“使用武力”与“武力攻击”之间时，目标国可以做出何种回应。¹⁸此外，还有观点认为，“规模与效果”本身以及“武力攻击”与边境事件之间的界分过于模糊，缺乏实际可操作性。¹⁹

总而言之，作为区分“使用武力”与“武力攻击”门槛的要素，“规模与效果”的提出体现了国际法院对于自卫权行使的谨慎态度，但无论从后续的国际司法实践还是学说观点来看，均对其持有一定的怀疑态度，既有的争议也随之延续至网络空间中继续被讨论。

二、“规模与效果”路径在网络空间适用的主要场景：禁止使用武力原则

上个世纪 90 年代末，在科索沃战争等重大军事冲突中，网络手段的运用使得国际法，特别是禁止使用武力原则的适用开始受到关注。21 世纪后，爱沙尼亚等国受到的重大独立网络攻击进一步使得网络空间的武力适用得以激烈讨论。“规模与效果”在传统国际法中的适用场景为禁止使用武力原则，在网络空间同样如此。不同之处在于“规模与效果”的地位与作用在网络空间有着鲜明的变化。

（一）“规模与效果”作为构成“使用武力”与“武力攻击”判断路径的引进

1999 年，美国国防部发布了《信息行动中的国际法问题评估》，其中提到：“虽然尚不清楚国际社会是否会将计算机网络的攻击视为‘使用武力’或‘武力攻击’，以及反措施于自卫权能否适用。但是评估结果可能更取决于攻击的结果（consequences）而非手段（mechanisms）。²⁰这一报告对网络行为后果的关注，可以视为“规模与效果”路径适用主张的早期萌芽。

美国国防部的观点得到了《塔林手册》²¹的主编迈克尔·施密特（Michael Schmitt）等学者的支持，²²成为在禁止使用武力原则违反性门槛判定的主流观点。

¹⁷ Chatham House, *The Chatham House Principles of International Law on the Use of Force in Self-defence*, 55 International and Comparative Law Quarterly 963, 966 (2006).

¹⁸ Judge Abdulqawi A. Yusuf, *The notion of ‘Armed Attack’ in the Nicaragua judgment and its influence on subsequent case law*, 25 Leiden Journal of International Law 461, 465 (2012). 在“石油平台案”中，Simma 法官也在单独意见中批评了“尼加拉瓜案”中的相关论述，见 *Oil Platforms (Iran v. USA)*, Merits, Judgment, Separate Opinion of Judge Simma, I. C. J Reports 2003, para. 12.

¹⁹ Judge Abdulqawi A. Yusuf, *The notion of ‘Armed Attack’ in the Nicaragua judgment and its influence on subsequent case law*, 25 Leiden Journal of International Law 461, 465 (2012).

²⁰ See Department of Defense, *An Assessment of International Legal Issues in Information Operations*, p. 18, 1999, p. 21, 1999, <https://apps.dtic.mil/sti/pdfs/ADB257057.pdf>.

²¹ 至今，《塔林手册》存在两版，第一个版本为 2013 年出版的《网络战争国际法塔林手册》，此版本也被称为 1.0 版本；第二个版本为 2017 年出版的《网络行动国际法塔林手册 2.0》，进一步处理了国际法对平时时期网络行动的适用问题，此版本也被称为 2.0 版本。本文中出现的《塔林手册》，在无特殊说明的情况下指代后者。

²² See Michael N. Schmitt, *Bellum Americanum: The US View of Twenty-First-Century War and Its Possible Implications for the Law of Armed Conflict*, 71 International Law Studies 390, 401(1998). 有学者虽然不同意施

早期，国际上还存在“工具说”与“目标说”两种相竞争的观点。其中，“工具说”关注行为的方式，将“武力”与武器的使用相联系，侧重于比较所使用手段与传统武器的物理特性；²³“目标说”则聚焦于行为目标，认为只要网络行动所针对的目标是国家关键基础设施，不论其后果如何，均构成对禁止使用武力原则的违反。²⁴施密特是最早支持“后果说”的学者，其认为“在某些非动能性攻击（如生化攻击与放射性攻击）可以纳入使用武力的范畴之中的情况下，将造成与动能性攻击类似后果的网络攻击排除在禁止使用武力原则的范围之内是不合理的。”²⁵在衡量网络行动的后果上，其开始提出了六大标准²⁶，后来进一步修订为七要素²⁷。

后来，施密特的观点融入了《塔林手册》中，并在采用“尼加拉瓜案”判决术语的基础上，进一步演变为“规模与效果”说，成为了判断“使用武力”与“武力攻击”的共同路径。《塔林手册》规则 67 写道：“如果网络行动的规模与效果相当于使用武力的非网络行动，则构成使用武力”。²⁸《塔林手册》规则 71 认为：“网络行动是否构成武力攻击取决于其规模与效果。”²⁹在此基础上，《塔林手册》进一步提出了衡量网络行动规模与效果的八大因素：严重性、迅即性、直接性、侵入性、效果的可衡量性、军事性、国家介入程度与推定合法性。³⁰其中，严重性是最重要的因素，也八个因素中唯一可以单独构成使用武力的因素，³¹从前述因素也可以看出，“规模与效果”说侧重于后果，尽管也有学者坚持主张，只有大规模的网络攻击才构成武力攻击。³²《塔林手册》产生了广泛影响，“规模与效果”路径也随之走入了更广阔的国家立场声明之中，许多国家直接或间接援引《塔林手册》中的观点。

密特提出的衡量网络行动后果的几点要素，但仍支持“后果说”，See Stephenie G. Handler, *New Cyber Face of Battle: Developing a Legal Approach to Accommodate Emerging Trends in Warfare*, 48 *Stanford Journal of International Law* 209 (2012).

²³ See Marco Roscini, *Cyber Operations and the Use of Force in International Law*, Oxford University Press, 2014, p. 46-47.

²⁴ See Reese Nguyen, *Navigating Jus Ad Bellum in the Age of Cyber Warfare*, 101 *California Law Review* 1079, 1119 (2013).

²⁵ Michael N. Schmitt, *Cyber Operations and the Jus Ad Bellum Revisited*, 56 *Villanova Law Review* 569, 573 (2011).

²⁶ See Michael N. Schmitt, *Computer Network Attack and The Use of Force in International Law: Thoughts on a Normative Framework*, in Michael N. Schmitt eds., *Essays on Law and War at the Fault Lines*, TMC Asser Press, 2012, p.26-27.

²⁷ Michael N. Schmitt, *Cyber Operations and the Jus Ad Bellum Revisited*, 56 *Villanova Law Review* 569, 576-577 (2011).

²⁸ [美]迈克尔·施密特：《网络行动国际法塔林手册 2.0》，黄志雄等译，社会科学文献出版社 2017 年版，第 335 页。

²⁹ [美]迈克尔·施密特：《网络行动国际法塔林手册 2.0》，黄志雄等译，社会科学文献出版社 2017 年版，第 342 页。

³⁰ [美]迈克尔·施密特：《网络行动国际法塔林手册 2.0》，黄志雄等译，社会科学文献出版社 2017 年版，第 338-340 页。

³¹ Michael N. Schmitt & Anusha S. Pakkam, *Cyberspace and the Jus ad Bellum: The State of Play*, 103 *International Law Studies* 194, 205 (2024).

³² Marco Roscini, *World Wide Warfare – Jus ad bellum and the Use of Cyber Force*, 14 *Max Planck Yearbook of United Nations Law* 85, 130 (2010).

（二）“规模与效果”路径在国家立场声明中的援引与发展

“规模与效果”路径在国家立场声明中得以广泛援引。据统计，在目前已经发布立场声明文件的 34 个国家和地区中，有超过 20 个明确承认其可成为判断构成“使用武力”、“武力攻击”或二者的共同路径。³³ 尽管中国、俄罗斯、巴基斯坦等少数国家未对“规模与效果”路径的可适用性进行表态，但尚未有国家公开反对“规模与效果”路径在判断禁止使用武力原则门槛时的适用。国家关于“规模与效果”路径的具体主张又可以分为以下三类。

其一，物理损害的“规模与效果”构成“使用武力”与“武力攻击”已基本形成共识。物理损害包括人员伤亡、财产毁损等，如 2010 年的“震网”（Stuxnet）事件就造成了伊朗纳坦兹核设施内部的数以千计的离心机受损，2024 年阿塞拜疆的飞机遭受网络干扰而导致大量人员伤亡。在立场声明中，几乎所有援引“规模与效果”路径的国家均认可关于达到一定程度物理损害能够构成“使用武力”与“武力攻击”。同时，除美国明确反对外，许多国家依然坚持网络行动构成“使用武力”与“武力攻击”有程度上的区别，³⁴ 尽管这些国家中大部分未能对“使用武力”与“武力攻击”之间的差别进行更具体地分析。

其二，非物理损害的“规模与效果”构成“使用武力”与“武力攻击”的主张日盛。网络行动很多时候所造成的后果是非物理性的。如 2007 年对爱沙尼亚的分布式拒绝服务（DDoS）攻击未造成物理损害，但导致政府网站关闭和通讯中断，引发了巨大经济损失。国家立场声明中非物理损害的主要有三个方面：其一是功能性损害，如爱尔兰表示导致关键基础设施功能严重受损的网络行为将构成“使用武力”，³⁵ 意大利表示造成基本服务的中断能够构成“使用武力”；³⁶ 其二是国家职能（state function）³⁷ 的损害，如挪威³⁸、新西兰³⁹ 等国持有此观点；其三是经济影

³³ 另外，在一些发布了多个版本立场声明的国家中，我们可以看到国家对于“规模与效果”路径立场的转变。如美国在 2012 年的布莱恩演讲中表示，构成使用武力需要综合考虑目标与位置、目的与效果等相关问题，但是在 2020 年及以后的立场声明中转向了“规模与效果”。参见：[https://cyberlaw.ccdcoe.org/wiki/Use_of_force#United_States_\(2020\)](https://cyberlaw.ccdcoe.org/wiki/Use_of_force#United_States_(2020))。

³⁴ 如荷兰、非盟、意大利等国都在立场声明文件中作出了明确表示。

³⁵ Irish Department of Foreign Affairs, *Position Paper on the Application of International Law in Cyberspace*, July 2023, <https://www.dfa.ie/media/dfa/ourrolepolicies/internationallaw/Ireland---National-Position-Paper.pdf>

³⁶ Italian Ministry for Foreign Affairs and International Cooperation, *Italian position paper on 'International law and cyberspace'*, November 2021, https://www.esteri.it/mae/resource/doc/2021/11/italian_position_paper_on_international_law_and_cyberspace.pdf

³⁷ 各国的立场声明文件中并没有对何为“国家职能”作出一般性解释，但需要与后文对主权原则探讨中的“固有政府职能”相区分，至少从文义来看，后者的范围应当比前者的范围更小。国家豁免理论可以为“国家职能”的一般理解提供帮助。国家豁免理论将国家职能区分为统治职能（*de jure imperii*）与管理职能（*de jure gestionis*），前者是仅能由国家行使的职能，后者虽然也由国家行使、但国家在其中所发挥的功能类似于私主体。相关探讨可参见 Frédéric Mégret, *Are There “Inherently Sovereign Functions” in International Law?*, 115 *American Journal of International Law* 452 (2021).

³⁸ The United Nations, *Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States*, UNODA, A/76/136, August 2021, p. 69-70, <https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf>

³⁹ Department of the Prime Minister and Cabinet of New Zealand, *The Application of International Law to State Activity in Cyberspace*, December 2020, <https://www.dpmc.govt.nz/sites/default/files/2020-12/The%20Application%20of%20International%20Law%20to%20State%20Activity%20in%20Cyberspace.pdf>

响，荷兰与挪威表示严重的经济影响也可以构成“使用武力”。⁴⁰法国甚至明确表示严重的经济影响可以构成“武力攻击”。⁴¹

其三，累积性低烈度网络行动的“规模与效果”构成“武力攻击”的主张渐起。前两种类型仍聚焦于某单一行为本身的规模与效果是否达到了相关门槛。现实中很多网络行动烈度较低，虽未达到“武力攻击”的量级，但也给国家造成了一定干扰与损害。对此，新加坡、法国、奥地利三国主张，对于一系列单独不构成“武力攻击”的行为，可以综合考虑其规模与效果，若达到足够的严重程度，也可以构成“武力攻击”。新加坡进一步阐述了此种攻击构成“武力攻击”需要满足的条件，即要由“同一行为体发起、或由不同的攻击者协同发起”。⁴²

（三）对“规模与效果”路径在网络空间禁止使用武力原则适用中引入的反思

总之，与传统国际法相比，“规模与效果”在网络空间禁止使用武力原则中的地位与作用得到了显著强化，成为了确定国际法违反性门槛的主要判断路径，即在判断某一行为是否违反国际法时，将“规模与效果”作为唯一或最关键的要件，而忽视其他要件的重要性。简言之，“规模与效果”在进入网络空间后实现了从“区分要素”到“判断路径”的转变。这种转变有其合理性，但是也可能会造成新的问题。

1.“规模与效果”路径引入的合理性与不足

不可否认，“规模与效果”路径的引入存在一定的合理性，支持者的理由可以归纳为两点。第一，传统国际法的要件需要改造以增强其在网络空间的适用性。以《联合国宪章》为代表的传统国际法中的禁止使用武力原则采用工具主义路径，⁴³关注使用武力的方式，限于传统的武装力量（armed force），网络行动被排除在“武力”的范围之内，因而需要改变思路，转向效果路径。第二，这种改造是国际社会所默许的，因为国际社会共同关切行为后果而非方式。具体而言，禁止使用

⁴⁰ Government of the Kingdom of the Netherlands, *Appendix: International law in cyberspace*, 26 September 2019, <https://www.government.nl/binaries/government/documenten/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace/international-law-in-the-cyberdomain-netherlands.pdf>; The United Nations, *Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States*, UNODA, A/76/136, August 2021, p. 66, <https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf>

⁴¹ Government of France, *International Law Applied to Operations in Cyberspace*, November 2021, <https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf>.

⁴² The United Nations, *Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States*, UNODA, A/76/136, August 2021, p. 84, <https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf>.

⁴³ 主张传统国际法对禁止适用武力原则采取工具主义的论据主要有以下三点：一是《联合国宪章》的准备工作，在1945年的旧金山会议中，巴西代表团曾提出将经济制裁也纳入武力的范畴之中，但是这一方案最终没有通过；二是《联合国宪章》第41条中将“经济关系、铁路、海运、航空、邮、电、无线电、及其他交通工具、之局部或全部停止，以及外交关系之断绝”作为可能（may）排除在武力范围之内的方式；第三点是在《国际法原则宣言》起草过程中，有国家提出要将“武力”的范围扩展至“所有形式的压力，包括那些具有政治或经济性质的压力，都会威胁到任何国家的领土完整或政治独立”，这一观点也未被采纳。

武力的国际法旨在避免武力对国家社会所共同关注的价值（如生存、尊严等核心价值造）成损害，而国际社会并不直接关心使用武力的方式。⁴⁴施密特进一步认为，《联合国宪章》之所以采取工具主义而非后果主义的原因在于后果衡量的复杂性，而直接对方式进行规制是一种捷径。⁴⁵

然而，仔细分析后可以发现其中的缺漏。一方面，支持引入者曲解了传统国际法的构造。因为时代的原因，《联合国宪章》在制定时对禁止使用武力原则确实采取了工具主义路径，但其针对的是“武力”的种类，而非对某一行为构成“使用武力”或“武力攻击”的整体判定。传统的国际法实践在判断一项行为是否构成“使用武力”或“武力攻击”时，也并非仅依据行为的方式，还需要综合考虑行为的目的、目标、性质、效果等。⁴⁶另一方面，引入理由扁平化了国际社会的向善愿景。尽管不法行为的后果是国际社会所共同关注的重要问题，但这并不意味着后果是最为重要乃至唯一的问题。再者，如果将第二点理由的逻辑推而广之，是否意味着所有国际法原则与规则违反性门槛的判断都可以采用“规模与效果”路径？最后，施密特与《塔林手册》提出的对于网络行动“规模与效果”判断的八点要素实施效果可能也难以为其引入提供更合理的支撑。有学者对这些要素逐一分析后认为它们同样无法清楚地判断一项行为是否构成使用武力，其容易被解释者所操控，从不同的地缘政治目标出发，甚至可能得出正反迥异的结论。⁴⁷

2. “规模与效果”路径的引入的突出风险：网络空间自卫权的扩张

“规模与效果”路径影响着网络空间“武力攻击”门槛判定，也因此影响着自卫权行使的门槛的确定。自卫权可能在以下三方面进行扩张：

一是“武力攻击”与“使用武力”共同路径的适用可能导致二者门槛混同，进而导致自卫权行使门槛的降低。在传统国际法中，“使用武力”与“武力攻击”有着严格区分，这一区分也得到绝大多数国家的支持。但在网络空间中，二者的区分被弱化，只有部分国家会明确表明二者之间存在门槛区别。同时，对“规模与效果”路径讨论的语境也存在区别，许多国家如法国、瑞士、芬兰、英国等仅在“武力攻击”语境下对该路径进行讨论，彰显了在面对网络攻击时优先判断是否构成“武力攻击”进而行使自卫权的意图。出于保护本国利益之考虑，一国更可能将本只能达到“使用武力”的行为拔高定性为“武力攻击”。

二是“规模与效果”路径为“累积性自卫”的证成提供了理论上的便利，导致自

⁴⁴ See Andrew C. Foltz, *Stuxnet, Schmitt Analysis, and the Cyber "use-of-force" Debate*, 67 Joint Force Quarterly 40, 41-43 (2012); Michael N. Schmitt, *Computer Network Attack and The Use of Force in International Law: Thoughts on a Normative Framework*, in Michael N. Schmitt eds., *Essays on Law and War at the Fault Lines*, TMC Asser Press, 2012, p.23.

⁴⁵ See Michael N. Schmitt, *Computer Network Attack and The Use of Force in International Law: Thoughts on a Normative Framework*, in Michael N. Schmitt eds., *Essays on Law and War at the Fault Lines*, TMC Asser Press, 2012, p.23

⁴⁶ 参见黄志雄：《国际法视角下的“网络战”及中国的对策——以诉诸武力权为中心》，载《现代法学》2015年第5期，第。

⁴⁷ Reese Nguyen, *Navigating Jus Ad Bellum in the Age of Cyber Warfare*, 101 California Law Review 1079, 1123-1124 (2013).

卫权行使门槛的进一步降低。“累积性自卫”是指的由累积性“武力攻击”引发的自卫。⁴⁸传统国际法对于系列行动能否构成“武力攻击”持有克制态度，国际法院在“石油平台案”以及“刚果诉乌干达案”等案件中虽未否认低烈度行动累积构成“武力攻击”之可能，但也均未做出肯定判决。⁴⁹“累积性”的判断过于模糊，一系列本身不构成“武力攻击”的行为该如何累积、其后果该如何叠加才能达到“武力攻击”的门槛要求，难以提出统一的、量化的标准，也更容易面临政治操控。⁵⁰

三是“规模与效果”路径将给与网络相关的经济胁迫（下称“网络经济胁迫”）构成“武力攻击”造势，可能导致自卫权行使范围的扩张。在传统国际法中，经济胁迫被排除在“武力”的范围之内。但在网络空间的讨论中，不仅有国家主张造成物理损害与功能性损害的行动可以构成“武力攻击”，甚至还有部分主张带来严重经济影响与纯粹经济影响的网络行动也可以。网络经济胁迫造成的纯粹经济影响可能并不亚与通常意义上的网络攻击，如俄乌冲突中美国等对俄罗斯实施的断开金融电讯协会（SWIFT）国际结算系统的措施，因其对一国经济的巨大破坏而被成为“金融核弹”。⁵¹同时，网络经济胁迫对一国经济造成的巨大损害甚至也可能通过影响生活资料的供给而进一步对人的生命造成威胁。另外，值得一提的是，《塔林手册》在这一点上态度也有着鲜明的转变，在 1.0 时期，明确表示网络行动包含或者类似于经济或政治强迫的绝不构成使用武力，⁵²但在 2.0 中相关的表述被删除。

三、“规模与效果”路径适用的扩张：不干涉原则与主权原则

除战争背景下的网络攻击外，平时（peacetime）网络行动也引起了许多国际法上的争议。2014 年索尼影业被黑事件、2016 年的网络干涉美国大选事件等引发了关于网络空间不干涉原则与主权原则适用的激烈讨论。但是二项原则适用上的不确定性丝毫不亚于禁止使用武力原则，“规模与效果”路径作为一种解决方案被引入。“规模与效果”路径在不干涉原则与主权原则中的形态与地位有了进一步的变化。（“规模与效果”路径在国家立场文件中的具体援引情况见本章后的表 1）

（一）“规模与效果”路径在网络空间不干涉原则适用中的初步扩张

干涉原则被认为是国际法上最模糊与最有争议的领域之一，⁵³其适用本身就

⁴⁸ 参见张华：《网络空间适用自卫权的法律不确定性与中国立场表达——基于新近各国立场文件的思考》，载《云南社会科学》2021 年第 6 期，第 82 页。

⁴⁹ Oil Platforms (Iran v. USA), Merits, Judgment, I. C. J Reports 2003, para. 64; Armed Activities on the Territory of the Congo (Democratic Republic of Congo v. Uganda), Merits, Judgment, I. C. J Reports 2005, para.146.

⁵⁰ 参见张磊：《网络空间累积性自卫的理论解构与因应思考》，载《国际法研究》2024 年第 2 期，第 80-95 页。

⁵¹ Meredith Deliso, What Is SWIFT and Why It's Being Called the 'Nuclear' Option for Russian Sanctions, ABC News, 27 February 2022, <https://abcnews.go.com/International/swift-called-nuclear-option-russian-sanctions/story?id=83131068>.

⁵² Michael N Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Cambridge University Press, 2013, p.46.

⁵³ 陈一峰：《论当代国际法上的不干涉原则》，北京大学出版社 2013 年版，第 2 页。

存在很大的复杂性。根据“尼加拉瓜案”判决，不干涉原则的适用需要同时满足两个条件：一是所涉及事项须落入一国“保留领域”（*domain reserve*）的范围；二是行为要构成“强迫”（*coercion*）。其中，“保留领域”界定了该原则的横向适用范围，而“强迫”则决定了其纵向适用门槛。

1. “强迫”要件作为“规模与效果”路径扩张的媒介

国际法院在“尼加拉瓜案”中虽然提出了“强迫”要件，但未对其进一步分析，也给各种解释观点留下了空间。早期国际上对于“强迫”要件的标准较高，以国际法委员会前委员伍德（*Wood*）为代表的观点认为，“强迫”是旨在使行为国实施特定行为（包括作为或不作为）的压力，并且需要达到使目标国不能合理反抗（意志屈从）的程度。⁵⁴

在将不干涉原则适用于网络空间时候，学界对“强迫”要件的解释逐渐呈现多样性，“规模与效果”路径为其中一种重要观点。综合看来，对于“强迫”解释的主张可能包含主客观两方面。主观方面包括行为国的意图与目标国的意识，客观因素包括行为国对目标国施加压力的行为、行为所产生的后果。主客观要件的比重是“强迫”的争议焦点之一，⁵⁵而以行为后果为核心的“规模与效果”路径因此得以在不干涉原则的适用中扩张。

在衡量“强迫”后果的问题上，麦克杜格尔（*McDougal*）和费利西亚诺（*Feliciano*）提出要考虑三个维度，包括“受影响价值的重要性和数量，价值受影响的程度，以及价值观受到如此影响的参与者数量”。⁵⁶施密特也以网络干涉选举为例，再次主张将“规模与效果”路径引入网络空间不干涉原则的适用。他表示，在衡量“规模与效果”时，可以考虑影响范围与严重程度，选举的性质和重要程度（如地方选举与全国选举的区分），以及信息的真实性等。⁵⁷

2. “规模与效果”路径在国家立场文件中的援引与类型化分歧

目前，德国、爱尔兰、捷克等在其国家立场文件中关于不干涉原则的主张部分适用“规模与效果”路径。德国是最早在国家立场文件中将“规模与效果”路径引入至网络空间不干涉原则适用中的国家。德国认为，网络行动的规模与效果与非网络情境中的强迫相当时，将构成被禁止的干涉。爱尔兰也表示，网络环境中的干涉必须具有足够的严重性，在规模和效果上与非网络情境中的强迫性行动相当。⁵⁸捷克的观点也与前两者相似。

关于行为达到何种程度的规模与效果才违背不干涉原则时，各国有着不同的

⁵⁴ See Maziar Jamnejad & Michael Wood, *The principle of non-intervention*, 22 *Leiden Journal of International Law* 345, 346–350 (2009).

⁵⁵ Richard C. Visek, *International Law and Cyberspace: Building Consensus*, 97 *Temple Law Review* 1, 12 (2024).

⁵⁶ Myres S. McDougal & Florentino P. Feliciano, *International Coercion and World Public Order: The General Principles of the Law of War*, 67 *Yale Law Journal* 771, 782 (1958).

⁵⁷ Michael N. Schmitt, *Foreign Cyber Interference in Elections*, 97 *International Law Studies* 739, 748–749 (2021).

⁵⁸ Irish Department of Foreign Affairs, *Position Paper on the Application of International Law in Cyberspace* 6 July 2023, <https://www.dfa.ie/media/dfa/ourrolepolicies/internationallaw/Ireland---National-Position-Paper.pdf>.

偏好，总体来看可以分为“选择自由”与“控制自由”两大标准。“选择自由”从词源上可以追溯到“尼加拉瓜案”，国际法院在该案的判决中表示，当一国以强迫方式干涉另一国自由选择之权利时，构成错误干涉。⁵⁹“选择自由”得到了多数国家的支持。如加拿大认为，最严重的强迫形式可能表现为威胁或使用武力，但也可能表现为旨在剥夺受影响国家的选择自由（freedom of choice）的网络活动。⁶⁰欧盟新发布的共同立场也持有类似的观点，认为强迫意味着迫使某一国非自愿地采取或放弃采取某一行为。⁶¹

“控制自由”标准可以追溯到《奥本海国际法》，该著作认为，构成干涉的干预必须是武力的（forcible）、专断（dictatorial）的，或者其他强迫性（coercive）的，并且实际上剥夺了被干预的国家对有关问题的控制。⁶²目前，支持“控制自由”的国家有新西兰、英国、澳大利亚等国。英国前司法部长苏拉·布雷弗曼（Suella Braverman）在其演讲中指出，若对他国事务的干涉是强行的、专断的或在其他方面是强制的，剥夺了一国对国家主权原则允许其自由决定事项的控制自由，则构成非法干涉。⁶³英国进一步明确，特定领域（如紧急医疗运输、医院系统等八个重点领域）若遭受控制性干涉，均构成“强迫”。澳大利亚则在“控制自由”的基础上进一步扩展，认为凡是有效剥夺或意图剥夺另一国控制、决定或管理固有主权性质事项的能力的手段，均可构成“强迫”。⁶⁴

总而言之，与禁止使用武力原则下较为具体的三种适用模式相比，不干涉原则适用中的“规模与效果”路径采取了更加抽象的判断标准。同时，在禁止使用武力原则中，“规模与效果”路径已居于主导地位，但在不干涉原则领域，该路径的适用仍处于初步扩张阶段，仅表现出萌芽趋势。

（二）“规模与效果”路径在网络空间主权原则适用中扩张的新形式⁶⁵

1. “规模与效果”路径扩张的前提：主权原则在网络空间适用的规则化

传统国际法的主流观点中，主权原则的规则属性并不突出。除去上文提及的

⁵⁹ “Intervention is wrongful when it uses methods of coercion in regard to such choices, which must remain free ones.” See *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Merits, Judgment, I. C. J Reports 1986, para. 205.

⁶⁰ Government of Canada, *International Law Applicable in Cyberspace*, April 2022, https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/cyberspace_law-cyberespace_droit.aspx?lang=eng.

⁶¹ 需要注意的是，欧盟共同立场与各成员国之间的立场是平行的关系，欧盟共同立文件中也明确表示不影响成员国目前和未来的立场。

⁶² See Lassa Francis Lawrence Oppenheim, *Oppenheim's International Law: Volume 1 Peace* (9th Edition), Edited By: Sir Robert Jennings QC, Arthur Watts KCMG QC, p. 432. 值得注意的是，《奥本海国际法》在第九版中的相关定义发生了较大的变化，此前的版本仅认为干涉不能是专横的。《奥本海国际法》在第九版中说明了修改的理由，即此前版本中的定义过于僵化（rigid）。

⁶³ See Suella Braverman, *2022 Speech at Chatham House*, 2022, <https://www.ukpol.co.uk/suella-braverman-2022-speech-at-chatham-house/>.

⁶⁴ Australian Government, *Australia's Position on How International Law Applies to State Conduct in Cyberspace*, October 2011, https://cdco.org/uploads/2018/10/Australia_submission-on-international-law-to-be-annexed-to-the-report-of-the-2021-Group-of-Governmental-Experts-on-Cyber.pdf.

⁶⁵ 之所以说“规模与效果”路径在主权原则中的扩张是一种新形式，是因为其在主权原则的扩张中可谓“舍其表而留其神”，尽管国家在其立场声明文件中没有明确提及“规模与效果”，但是其内核已经得到充分体现。

禁止使用武力原则与不干涉原则外，主权原则往往是以其另一衍生规则——领土主权规则的形式出现。如在 1947 年的“科孚海峡案”中，英国未经同意在阿尔巴尼亚的领海内进行排雷活动，国际法院判决前者侵犯了后者的主权，此处实际上指领土主权。⁶⁶因而也有学者认为，侵犯主权原则的提法不过是对侵犯国家领土完整权的一种有争议的简化。⁶⁷但网络空间超出了传统意义上领土的范畴，一般意义上主权原则的规则化面临着激烈的讨论。⁶⁸网络空间中主权原则的适用，总体存在着“间接适用”与“直接适用”两种观点。间接适用论认为主权是一种背景权利，不直接在网络空间产生行为义务，而要通过从其衍生规则实现，而直接适用论则认为主权原则可以作为一种规则直接适用。⁶⁹ 不仅是其被用以直接规制网络行动的基础，也是“规模与效果”路径得以扩张，并成为判断网络行为是否构成违反主权的门槛的前提。

2. “规模与效果”路径在国家立场文件中的援引

在主权原则规则化的前提下，国家关于其如何适用存在两种观点。其一是强调行为本身的“权力介入说”，⁷⁰该观点认为行为国对目标国任何未经同意的权力介入都构成对主权原则的违反。法国、中国为此种观点的代表。如法国认为，任何对于计算机系统的渗入都将构成对于主权原则的违反；⁷¹另一类是基于行为后果的“最低门槛说”（*de minimis threshold*），该观点认为，违背主权原则需要满足最低程度的后果。第二种观点较第一种观点来说更为主流，挪威、荷兰、德国等持有此观点，“规模与效果”路径正是在此观点下得以发展。

围绕着何种行为后果会达到违反主权原则的最低门槛，国家的立场观点基本围绕着《塔林手册》中的两条路径展开：一是对目标国领土完整造成的损害程度，二是对固有政府职能（*inherently governmental function*）的干扰或者篡夺，二者满足其一即可。对于前一路径，其又可进一步细分为三个层次：物理损害、网络基础设施功能的丧失以及侵犯领土完整但未造成功能丧失的程度。⁷²目前，各国立场基本认同物理损害能够造成对主权原则的违反。前一路径得到了几乎所有国家的采纳。其中，对于功能丧失，加拿大、哥斯达黎加、丹麦等十多个国家与地区表示赞同，其中，哥斯达黎加进一步将功能损害的分为需要部件替换的功能损

⁶⁶ Corfu Channel Case (United Kingdom v Albania), Merits, Judgment, I. C. J. Reports 1949, p.36.

⁶⁷ Leonhard Kreuzer, *Sovereignty in Cyberspace—A Rule without Content?* In Antonio Segura Serrano eds., *Global Cybersecurity and International Law*, Routledge, 2024, p. 33.

⁶⁸ 领土主权（*territorial sovereignty*）与一般意义上的主权（*sovereignty*）的区分在国家立场声明中十分明显，采取不同观点的国家会在用词上保持较高的统一性。

⁶⁹ 参见王超：《主权原则在网络空间适用的理论冲突及应对》，载《法学》2021 年第 3 期，第 106-107 页。

⁷⁰ “权利介入说”与后面提及的“最低门槛说”在有的文献里会将其并入“纯粹主权说”（*pure sovereignty*）与“相对主权说”（*relative sovereignty*）。See Kevin Jon Heller, *In Defense of Pure Sovereignty in Cyberspace*, 97 *International Law Studies* 1432 (2021).

⁷¹ Government of France, *International Law Applied to Operations in Cyberspace*, November 2021, <https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf>.

⁷² [美]迈克尔·施密特：《网络行动国际法塔林手册 2.0》，黄志雄等译，社会科学文献出版社 2017 年版，第 65 页。

害与造成运营停止的损害；⁷³奥地利则进一步认为，导致功能的暂时性丧失也将违背主权原则。⁷⁴对于干扰或篡夺固有政府职能，奥地利、加拿大、瑞典等九个国家采纳了这一路径。在对固有政府职能进行定义时，各国会依据其本国所关切之事项进行列举，可能包括选举、医疗、税收、国防等。

总而言之，目前主权原则适用中的“规模与效果”路径具有复合性特征，融合了禁止使用武力原则与不干涉原则中相关路径的适用逻辑：前者聚焦于后果的严重性，与禁止使用武力原则中的路径相似；而后者强调对目标国行使独立国家权力能力的实际影响，与不干涉原则的运作路径雷同。⁷⁵

（三）对“规模与效果”路径网络空间国际法适用中扩张的反思

为了应对适用上的不确定性，发源于禁止使用武力原则的“规模与效果”向不干涉原则与主权原则的适用中进一步扩张。这种扩张是纵向的，因三项原则间具有纵向的联动关系。根据“尼加拉瓜案”，这三项原则效力有着重叠之处，使用武力在违反禁止使用武力原则的同时也将违反不干涉原则与主权原则。⁷⁶这一联动关系使得“规模与效果”路径的适用边界在不同原则之间形成了互为影响的局面。“规模与效果”路径已经是一个在禁止使用武力原则熟知的路径，“规模与效果”路径已被国际社会成员所熟知，使得其扩张具有一定的社会基础，但同时也会给这两项原则的适用带来难以一定程度的消极影响。

1. 加剧不干涉原则与主权原则之间界限的模糊性

不干涉原则与主权原则之间保护范围上的关系实则由“保留领域”与主权原则范围之间的关系确定。然而，这种关系在现有国际法实践中并未得到清晰界定。例如，在“尼加拉瓜案”中，“保留领域”被描述为“一国根据主权原则能够自由决断”的事项。⁷⁷可知，主权原则是“保留领域”确立依据，但依旧无法判断“保留领域”与主权原则在范围上的关系。《国际公法百科全书》没有给两者画上等号，而是谨慎地判断二者在某种意义上没有区别，其认为，一国主权由对内主权与对外主权构成，“保留领域”习惯性地用以指代对内主权的体现，但是侵犯对外主权，如领土完整，可能导致对对内主权的侵犯，从而构成对“保留领域”的干涉。⁷⁸《塔

⁷³ Ministry of Foreign Affairs of Costa Rica, Costa Rica's Position on the Application of International Law in Cyberspace, 2021, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_2021/Costa_Rica_-_Position_Paper_-_International_Law_in_Cyberspace.pdf.

⁷⁴ See Government of Austria, *Position Paper of the Republic of Austria: Cyber Activities and International Law*, April 2021, Position Paper of the Republic of Austria: Cyber Activities and International Law, [https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_2021/Austrian_Position_Paper_-_Cyber_Activities_and_International_Law_\(Final_23.04.2024\).pdf](https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_2021/Austrian_Position_Paper_-_Cyber_Activities_and_International_Law_(Final_23.04.2024).pdf).

⁷⁵ See Moynihan, Harriet, *The Application of International law to State Cyberattacks: Sovereignty and Non-Intervention*, Chatham House, 29 November 2019, p. 24, <https://www.chathamhouse.org/sites/default/files/publications/research/2019-11-29-Intl-Law-Cyberattacks.pdf>.

⁷⁶ Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), Merits, Judgment, I. C. J Reports 1986, paras. 205, 251.

⁷⁷ Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), Merits, Judgment, I. C. J Reports 1986, para. 205.

⁷⁸ Katja S Ziegler, *Domaine réservé*, Max Planck Encyclopedia of Public International Law, April 2013,

林手册》也存有模糊之处。其认为对固有政府职能（主权原则的定义路径之一）的干涉与篡夺与“保留领域”之间存在一定程度的重叠但又并非完全相同。⁷⁹

不干涉原则与主权原则之间保护范围之间若即若离的关系使得使二者的界限主要由“强迫”来划定，但网络空间“强迫”门槛的降低使得二者之间的界限变得更加模糊。如前所述，在传统国际法中，“强迫”的判断可能涉及主客观方面的多项要素，而网络空间“规模与效果”路径则注重行为后果的严重性。一方面，判断要素的减少使得二者之前的区分标准单一化；另一方面，不少国家主张降低“强迫”的标准，“选择自由”“控制自由”的提出使得“强迫”要求的后果严重性一步步降低。主权原则与不干涉原则之间的界限变得更加模糊。

2. 促使部分低烈度网络行动处于“规制真空”状态

如前所述，“规模与效果”路径在网络空间主权原则适用中的扩张是“最低门槛说”的体现，即只有造成一定程度的损害后果时，才能构成对主权原则的违背。在现实案件中，由于受害国主张权利时需要识别损害结果并证明因果关系，这就会导致一些低烈度的，后果不太明显的网络行动处于事实上的“规制真空”状态。

以网络间谍为例，据其作用机理，大体上可以分为入侵计算机网络、拦截计算机信息、利用网络公开信息三种方式。⁸⁰其中，入侵计算机网络在特定情况下可能造成显著损害，例如奥地利在立场声明中指出，若网络间谍行为导致信息技术系统必须完全关闭，并且在清理和恢复期间需要采取替代措施，则构成对主权的侵犯。⁸¹然而，在大多数情况下，网络间谍行为采取潜伏方式，对计算机系统本身并不造成物理或功能性破坏。此外，拦截计算机信息及搜集网络公开信息更是通常不会涉及对计算机系统的入侵，自然也不会对其造成可识别的损害。因此，网络间谍行为即便可能造成严重后果，其损害通常并非即时或直接的，而是通过泄密导致的预期性或间接性损害。在国际法实践中，受害国难以就这种损害主张权利，原因主要有两点。一是国际法实践中对预期或间接损害作为赔偿事由的适格性存在很大争议，在网络空间国家立场声明文件中支持预期与间接损害的国家也寥寥无几；⁸²其次，即使承认间接损害的法律地位，在实践中仍然面临极大的证明难度。受害国需要在高度复杂的网络环境中厘清损害的具体程度，并证明行

<https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e1398>.

⁷⁹ Michael N. Schmitt eds., *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, Cambridge University Press, 2017, p.24.

⁸⁰ 参见隆峰、谢宗晓：《网络空间间谍活动的特征、形式及应对》，载《中国信息安全》2021年第10期。

⁸¹ See Government of Austria, *Position Paper of the Republic of Austria: Cyber Activities and International Law*, April 2021, Position Paper of the Republic of Austria: Cyber Activities and International Law, [https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_2021/Austrian_Position_Paper_-_Cyber_Activities_and_International_Law_\(Final_23.04.2024\).pdf](https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_2021/Austrian_Position_Paper_-_Cyber_Activities_and_International_Law_(Final_23.04.2024).pdf).

⁸² 对于间接损害，在主权原则适用的场合，仅有德国明确在其立场声明文件中明确支持间接的物理影响可以构成对主权原则的违反；在禁止使用武力原则适用的场合，也仅有奥地利在其立场声明文件中赞同间接的物理损害与功能损害可以构成对禁止使用武力原则的违反；不干涉原则中强调间接“强迫”可以构成对不干涉原则的违反，但没有国家认为造成间接损害会导致对不干涉原则的违反。对于预期损害而言，仅在禁止使用武力原则适用的场合下有非盟、以色列、奥地利等少数几个国家与地区支持。

为与损害结果之间的因果关系，这在技术和法律层面均极具挑战性。因此，在“规模与效果”路径下，网络间谍及其他类似的低烈度网络活动往往难以被有效规制，使得这些行为游离于现行国际法框架之外，形成事实上的“规制真空”。

表 1：“规模与效果”路径在国家立场文件中的支持情况⁸³

国际法原则	表态国家数量	支持国家数量	相竞争的学说	路径的分化
禁止使用武力原则	26	21	无	物理损害、非物理损害、累积性低烈度损害
不干涉原则	28	5	目的说	损害自由控制、损害自由选择
主权原则	30	12	间接适用说、直接适用说中的权力介入说	损害领土完整、损害固有政府职能

四、“规模与效果”路径适用的整体反思：基于类比推理的展开

类比推理，与归纳推理与演绎推理一起并称为三大推理方式，是人类认识世界的重要渠道。作为法律适用方式之一，类比推理能够拓展法律的适用范围。在国际法领域，这种推理方式屡见不鲜。例如，战争法中的空战规则便是在对陆战规则进行类推适用的基础上发展出的新规则，国际组织责任的许多规则也源于对国家责任规则的类推适用。“规模与效果”路径在网络空间的适用过程中，显著依赖类比推理。主权国家通过类比网络行动与非网络行动的规模与效果，认为若二者相当，则该网络行动可能违反国际法。在禁止使用武力原则与不干涉原则的适用中，“规模与效果”路径的适用采用的是显性的类比推理的方式，可以轻易察觉其存在，而在主权原则的适用中则是更为隐性的方式。

（一）“规模与效果”路径涉及的类比推理过程

“规模与效果”路径的本质是一种基于后果的类比推理适用传统国际法的方式。类比推理的基本推导过程是“根据两个或两类对象某些属性相同或相似,从而推知它们的另一个属性也相同或相似的”。⁸⁴从类比推理的逻辑结构来观察，非网络空间中的国际规则适用是源域（或称为“类比物”），网络空间中国际规则的适用是目标域（或称为“认识对象”）。源域与目标域之间相似性比较的过程也被称作“映射”。行为后果是相似性比较的（唯一或最重要的）连接点，而“尼加拉瓜案”中的“规模与效果”这一术语则为这该连接点披上了一层“合法性”外衣。同时，源域与目标域均为完整的法律规范，因果关系在其中成为沟通行为模式与法律后果之间的重要桥梁，这一类比推理模型属于因果关系模型。网络空间“规模与效果”路径涉及的基本推导过程如下图：

⁸³ 图表系作者根据网络法工具箱（cyber law tooltik）网站上收集的国家立场声明文件整理。需要注意的是，表中所指的“规模与效果”路径与文章中的含义保持一致，而非单指“规模与效果”一词出现的场合。参见：https://cyberlaw.ccdcoe.org/wiki/Main_Page.

⁸⁴ 雍琦：《法律逻辑学》，法律出版社 2016 年版，第 286 页。

图 1：“规模与效果”路径的类比推理过程⁸⁵

(源域) 非网络空间中, 行动A具有a性质、b目的、c方式、d后果, 所以行为A违反了国际法规则X,
(目标域) 网络空间中, 行动B具有a'性质、b'目的、c'方式、d后果,
(结论) 所以行动B也违反了国际法规则X。

(二) “规模与效果”路径扩张的动因分析

与归纳推理与演绎推理等自然推理不同, 类比推理是一种目的导向型的推理, 其在国际法的发展中的适用同样如此。源域的选取、连接点的选择都是一个有意识干预的过程, 受到法律、政治等多方面因素的影响。⁸⁶

在法律方面, “规模与效果”路径的法律优势是其得以扩张的最直接的原因, 体现在两方面。其一是该路径具有形式上的权威性。“规模与效果”路径有着国际法院判决的背书。国际法院作为解释国际法的重要平台, 其所作判决对于国际法的适用有重大影响力, 具有一定形式上的权威性。国家在立场声明中关于禁止使用武力原则的部分援引“规模与效果”路径已经形成了路径依赖。施密特对此认为, 尽管有不同的观点认为有必要在区分“使用武力”与“武力攻击”的问题上适用不同的标准, 但国家仍然趋于采取共同路径是一有趣现象。⁸⁷而这种权威性又进一步向着不干涉原则与主权原则溢出。其二是该路径在一定程度上加强了传统国际法在网络空间的司法适用性。“规模与效果”路径以后果为判断基准, 可视为一种简化的客观主义路径, 减少证明要素, 降低证明难度。在“石油平台案”中, 因为美国无法证明伊朗的攻击具有特别针对 (specifically aim) 美国船只的意图, 国际法院否认了伊朗的行为构成武力攻击。网络空间的虚拟性, 技术的复杂性, 以及代理人的加入等将会使得行为方的意图更难以证明。尽管目前各国乐于披露与谴责敌对网络行动, 但尚未有网络行动在国际司法机构中被正式起诉。⁸⁸“规模与效果”路径的简化逻辑有助于推动未来可能的司法实践。

在政治方面, “规模与效果”路径贴合了国际社会意图尽快将网络空间纳入国际法规制之下的政治诉求。网络空间地缘政治中存在的“玻璃房困境”⁸⁹揭示了网络空间的“攻防脆弱性对称”, 促使着传统意义上的强国与弱国对网络空间的均抱有强烈的规制意向。而建立在类比推理基础之上的“规模与效果”路径的在填补规

⁸⁵ 本图系作者根据类比推理的流程自制, 其中, b'c'd'表示目标域所涉及的要素与源域中的bcd不同。

⁸⁶ 参见蔡从燕:《类比推理与国际法发展的逻辑》, 法律出版社2012年版, 第84-86页。

⁸⁷ See Michael N. Schmitt and Anusha S. Pakkam, *Cyberspace and the Jus ad Bellum: The State of Play*, 103 *International Law Studies* 194, 213 (2024).

⁸⁸ 目前, 仅有媒体报道, 国际刑事法院检察官正对乌克兰民用基础设施的网络攻击搜集证据, 准备起诉。See Anthony Deutsch et al., *Exclusive: ICC Probes Cyberattacks in Ukraine as Possible War Crimes*, Reuters, 14 June 2024, <https://www.reuters.com/world/europe/icc-probes-cyberattacks-ukraine-possible-war-crimes-sources-2024-06-14/#:~:text=A%20team%20at%20the%20Human,be%20charged%20as%20war%20crimes>.

⁸⁹ 网络空间“玻璃房困境”指国家在网络空间仿佛居住在“玻璃房”里, 每个国家都手握能够破坏他国“玻璃房”的“石头”, 但同时也担心自己的“玻璃房”被他国“石头”砸碎。“玻璃房困境”说明因为互联网的相互依存程度高, 即使是传统意义上的强国也担心自己受到网络攻击。See Kubo Mačák, *On the shelf, but close at hand: The contribution of non-state initiatives to international cyber law*, 113 *ASIL Unbound* 81, 82-83 (2019).

制空缺的作用时有其优势之所在。其一,“规模与效果”路径建立在源域中的既有规则之上,能够在短期内促进国家就某一问题达成初步一致,进而将网络空间从“数字荒野”尽快变得“有法可依”。如在将禁止使用武力原则类比推理适用于网络空间时,鉴于传统国际法中的相似要求,国家之间很快地对物理攻击或动能攻击构成使用武力与武力攻击的可能性达成了共识。其二,“规模与效果”路径能够在促进对某一领域进行治理的同时,更大程度地避免了直接对其中的争议问题进行界定时所面临的复杂性,提升了治理效率。如在不干涉原则的适用中,对“强迫”直接进行定义是各国公认的难题与争议的焦点,而此时诉诸于类比推理能够绕过这一难题,在保持较高灵活性与后续解释的空间的基础上推进规则的适用。这种避免争议焦点的路径选择,正是各国在“玻璃房困境”下的一种现实主义策略,即在缺乏共识的领域采取务实性规制路径,以便在确保一定法律约束力的同时,不损害自身未来的政策调整空间。

(三)“规模与效果”路径的内在局限

类比推理是一种或然性推理,其结论的真假具有不确定性,因此在法律适用中的应用需受到限制。在国际法的类推适用中,其适用需至少满足三个条件:“规制真空”的存在(决定前提),相类比的领域之间具备较高程度相似性(决定充分性与说服力),不与强行法产生冲突(决定底线)。⁹⁰其中,在相似性判断上,类比物与认识对象之间的可用来比较的相似点越多,相似点越重要,则类比推理的充分性与说服力就更高。从反面来说,也要关注二者之间的相异之处,并证明相异点无足轻重。⁹¹而源域规则的目的与所保护的价值则是判断相同点(相异点)相关性的重要指标。⁹²“规模与效果”路径的内在局限体现在以下两个方面。

首先,“规模与效果”路径在逻辑层面的说服力存在不足。“规模与效果”路径关乎则类比推理相似性比较这一重要环节,但其合理性受到较大挑战。传统国际法的适用会基于多方面的因素考量,如目的、目标、性质、方式、后果等,这些要素共同构成了国际法违反性门槛的判断标准,除后果外的其他要素都并非无足轻重。禁止使用武力原则中,行为的性质与目的是相关的,其是用以评价行为是构成侵略性行为还是自助(自卫)行为的重要方面。不干涉原则中,行为的性质与目的同样是相关的,传统国际法中不干涉原则的宗旨是保护一国对于其在制度选择上的自主性⁹³、不受他国干涉,这就要求行为国不能抱有使目标国意志屈从、并从中获益的想法。

而“规模与效果”路径却过分侧重于后果的严重性,将其作为最核心甚至唯一

⁹⁰ 刘碧琦:《论国际法类推适用的有效性——以〈塔林手册〉为视角》,载《社会科学家》2021年第5期,第130-131页。

⁹¹ See An Hertogen, *The Persuasiveness of Domestic Law Analogies in International Law*, 29 *European Journal of International Law* 1127, 1143 (2018).

⁹² See Joseph Raz, *The Authority of Law: Essays on Law and Morality*, Oxford University Press, 2009, p.203.

⁹³ 参见陈一峰:《论当代国际法上的不干涉原则》,北京大学出版社2013年版,第3页。

的判断依据，这大幅削弱了其推理的说服力。此外，从后果的严重性角度来看，网络行为所造成的后果与传统动能性行为仅在物理损害方面存在较高程度的相似性，因而国家在这一点上达成的共识较为广泛。但对于非物理损害，包括计算机系统功能损害、经济损害，以及更抽象层面的对固有政府职能的损害、剥夺选择自由、控制自由等，网络行为难以与传统动能性行为进行直接类比。因此，目前部分国家在立场文件中提出的关于“规模与效果”路径适用的主张，往往超出了类比推理应有的限度。

其次，“规模与效果”路径在填补规制空缺方面的作用具有时效性限制。首先，源域的有限共识难以应对目标域的实践演变。例如，尽管国家对动能性攻击违反禁止使用武力原则已达成一定共识，但随着网络非物理损害种类的不断增多，原有共识已显得难以适用。其次，源域规则中本身存在的争议在目标域中可能被放大。如前所述，关于“使用武力”与“武力攻击”门槛区分的争议在传统国际法领域仍未达成共识，而在网络空间，这一问题变得更加复杂，使得“规模与效果”路径的适用更加不确定。总而言之，过度依赖类比推理可能导致法律的发展受制于既有框架，缺乏创新动力。⁹⁴

（四）“规模与效果”路径对网络国际法发展的整体影响

网络空间国际法，作为国际法系统中的正在兴起的一个分支，其发展应当服务于国际法整体体系的稳定与协调。“规模与效果”路径以后果作为类比推理唯一或最重要的连接点，显著地改变国际法传统地适用方式，给网络空间国际法自身的发展带来深远影响。

首先，从本体的视角，“规模与效果”路径侧重于事后判断，有损网络空间国际法的可预期性。⁹⁵可预期性是法律的重要价值，对于国际法来说也同样如此。⁹⁶效果路径以后果为导向，事先难以预估行为所产生的后果，因此也难以为行为者提供合法性指引，在网络行动中更是如此。部分网络行为的后果难以在短时间内显现，损害难以量化，如网络间谍行为，往往不会立即造成可见的物理或经济损害，但可能在长期内削弱国家安全或技术竞争力。一些网络行为的后果则难以定性，如造成剥夺一国“选择自由”的后果。还有，即使是在一些较为明显网络攻击中，事先判断行为是否具有严重的损害后果有时也并非易事，其受到目标对象个性特征的影响严重。如一网络攻击者入侵了某金融机构的交易系统，初衷仅是窃取数据。然而，由于现代金融体系高度自动化，交易系统之间又存在复杂的联动机制，这一攻击意外地触发了一系列程序化交易的连锁反应，导致市场在极短

⁹⁴ Mariarosaria Taddeo, *On the Risks of Relying on Analogies to Understand Cyber Conflicts*, 26 *Minds and Machines* 317, 321 (2016).

⁹⁵ 参见张华：《网络空间适用禁止使用武力原则的法律路径》，载《中国法学》2022年第2期，第290页。

⁹⁶ Jutta Brunnée & Stephen J. Toope, *International law and the practice of legality: stability and change*, 49 *Victoria University Wellington Law Review* 429 (2018).

时间内出现剧烈波动，即“闪崩”现象。在这种情况下，该攻击的严重后果并非攻击者初衷，而是由目标系统自身的脆弱性所放大，进一步加剧了事前可预期性的缺失。

其次，从执行的视角，“规模与效果”路径可能导致国际法适用的不一致，使得网络能力不同的国家在国际法执行上出现不对等现象。各国的网络化程度与网络防御能力是决定某项网络行动后果的重要因素，但不同国家在这两方面存在显著差异。通常而言，一方面，网络化程度较高的国家更容易受到网络行动的影响。有研究表明，爱尔兰是欧洲网络化程度最高的国家之一，其经济和社会运行对数字基础设施的依赖程度远超许多其他欧洲国家。因此，相同的网络攻击行为可能在爱尔兰产生严重的经济和社会后果，⁹⁷而在网络化程度较低的国家，其影响则可能微乎其微。另一方面，网络防御能力较强的国家受到的网络攻击影响较小。例如，对美国或以色列这样拥有强大网络防御能力的国家发动网络攻击，可能被成功拦截或迅速修复，从而不会造成严重后果。而同样的攻击若发生在网络防御能力较弱的国家，可能导致关键基础设施瘫痪，影响其国家安全。上述情况将导致同样的网络行为在不同的国家会造成不同的后果，从而致该网络行为在一些国家会被认为是违背国际法，在另一些国家将认为不是。这种适用上的不一致在传统国际法中较为罕见，例如，对一个人口密集的国家发射导弹与对一个人烟稀少的国家发射导弹都将违背国际法。国际法适用不对等不仅有可能加剧国际法适用的不确定性，还可能对阻碍国际法公平价值的实现，尽管当前国际法学界对“公平”是否构成国际法的核心价值仍存在争议，但可以预见，“公平”应当成为继“和平与安全”等基本价值之后的、国际法价值变革的重要方向。⁹⁸

再次，从功能的视角，如果任“规模与效果”路径扩张，可能会造成国家在违法性边缘“试探”的道德风险，将会有损国际法的价值宣示功能。⁹⁹如前所述，施密特将“规模与效果”路径引入网络空间的重要理由为——国际社会并不直接关心违法行为的手段，而只关心其所造成的严重后果。这一理由在现实主义视角下具有一定的吸引力，在这种推断下，“规模与效果”路径甚至可以扩张至任何国际法中的新兴领域。同时，也如前所述，“规模与效果”路径可以带来一定程度上的司法适用之便利，这同样为其扩张带来驱动力。但是，国际法的功能，即使被认为不如国内法一般丰富，也不仅仅在于寻求一个在可能发生争议时的确定结果，大沼保昭（Onuma Yasuaki）认为，国际社会中的国际法还应当具有价值宣示功能，即反映国际社会对于一个良好国际秩序的共同愿望；罗莎琳·希金斯（Rosalyn

⁹⁷ Joshua Davis, *Hackers Take Down the Most Wired Country in Europe*, *Wired*, 21 August 2007, <https://www.wired.com/2007/08/ff-estonia/>.

⁹⁸ 参见江河：《论国际法的公平价值及其实现进路：从和平到正义》，载《政法论丛》2023年第1期，第80页。

⁹⁹ Onuma Yasuaki, *International law in and with international politics: The functions of international law in international society*, 14 *European Journal of International Law* 105 (2003).

Higgin) 也认为, 国际法的功能更在于提供一个能保护国际社会共同价值的运行系统。¹⁰⁰ 也正如此, 传统国际法中要判定行为的性质, 如是正义战争还是非正义战争; 传统国际法要将行为的目的纳入违反性门槛的考察之中, 如对于强迫的传统理解等。而“规模与效果”路径聚焦行为的后果, 可能会纵容行为国采取渐进式策略, 徐徐试探其行为的后果, 从而对国际秩序造成缓慢而持续的干扰与侵蚀, 这样并不利于国际秩序的安全与稳定。

最后, “规模与效果”路径还会导致网络空间国际法与传统国际法适用上的无端割裂, 使得国际法体系碎片化程度的加剧。国际法碎片化是指国际法呈现出的各分支领域之间的冲突与矛盾的状态。¹⁰¹ 国际法多样化发展背景下, 碎片化现象虽难以避免但并不值得倡导, 不同种类与层级的规范只有在协调有序才能提高适用效率, 为良好国际秩序提供保障。¹⁰² 在传统国际法中, 国家责任法未有将行为后果作为国际不法行为的必要要件一般要求, 是否需要后果要件取决于各初级规则的具体要求。¹⁰³ 以主权原则为例, 传统国际法中并未规定侵犯主权原则应当承担特定后果。例如, 一国军队或飞机未经许可进入另一国领土或领空, 无论是否产生物理损害或影响, 通常都被认定为侵犯该国的主权。在网络空间中强调后果要件则显得缺乏依据, 尤其是考虑到网络数据、系统都有其现实附着之载体。又如, 在传统国际法中, 如前所述, 经济强迫的方式被明确排除在“使用武力”的范畴之外。“规模与效果”路径在网络空间的引入将会冲击传统国际法对于经济强迫与使用武力之间的关系, 不仅网络经济强迫在该路径下有构成“武力”之可能, 也没有理由否认达到一定后果一般经济胁迫可以构成“武力”。

五、我国对于“规模与效果”路径的立场建议

“规模与效果”路径的扩张不仅是一国政治选择, 也具有极强的国际法意义。我国应当更充分地参与以国家立场声明为载体的法律辩论, 更积极参与网络空间国际法的塑造进程。虽然我国已于 2021 年就网络空间国际法规则的适用发表了声明, 但在国际法内容的广度与深度上相对有限, 与部分发达国家及发展中国家 (如哥斯达黎加、捷克、非盟等) 相比都有所不足。目前, 我国还未对“规模与效果”路径的适用表达立场, 应秉持“网络空间命运共同体”理念, 坚持主权、安全、合作、秩序等核心要义, 站在我国利益与全球利益的交汇点, 对“规模与效果”路径的适用进行理性回应。

具体而言, 对于“规模与效果”路径在不同层面的扩张, 我国应采取差异化应对措施。禁止适用武力原则是应当最为重点关注的领域。如前所述, “规模与效

¹⁰⁰ Rosalyn Higgins, *Problems and Process: International Law and How We Use It*, Clarendon Press, 1995, p.1.

¹⁰¹ 莫世健:《国际法碎片化和国际法体系的效力》, 载《法学评论》2015 年第 4 期, 第 118 页。

¹⁰² 参见古祖雪:《现代国际法的多样化、碎片化与有序化》, 载《法学研究》2007 年第 1 期, 第 142 页。

¹⁰³ ILC, *Commentary to the Articles on State Responsibility (2001)*, 2 Yearbook of the International Law Commission (Part II) 31, 36, commentary to Article 2, para 9.

果”路径容易导致网络空间自卫权的扩张与滥用，不利于网络空间的和平的安全。有学者曾认为我国应当避免直接引入“规模与效果”路径，并主张应在具体情境下进行个案分析。¹⁰⁴该观点实际上仍主张对“规模与效果”路径持一种偏向于观望的态度，这种观点在早期是一种务实的合理选择，但在“规模与效果”路径在禁止使用武力原则的适用已成为国际社会的绝对的主流观点，甚至有可能发展成习惯国际法的情况下，继续“无视”该标准可能会导致我国错失塑造国际规则的良机。对此，我国可以考虑直接引入该路径，但对其予以新的阐释，一方面，我国可以调整“规模与效果”的解释框架来弱化其影响。例如，可借鉴非盟的做法，将“规模”与“效果”拆解为两个独立要件，分别加以阐释，¹⁰⁵以增加类比推理的连接点、削弱行为后果在法律适用中的比重，从而增强该路径的逻辑合理性与可预期性等。另一方面，我国应当对“规模与效果”路径的适用场景予以限制，主张只有在造成物理损害或人员伤亡的情况下，才构成对禁止使用武力原则的违反，从而缩小其适用范围，也尽可能地避免了由此而导致的自卫权的扩张。

至于不干涉与主权原则，这两项原则肩负了对禁止使用武力原则涵盖不到的网络攻击的规制的使命。如何实现二者规制功能上的差异化与避免不当规制空白的出现应是对二者新阐释的发力点。如前所述，一方面，“规模与效果”路径的使用容易模糊二者适用的界限，导致不干涉原则的功能的淡化。另一方面，也容易导致网络间谍等低烈度网络行动出现规制空白。而我国长期经常遭受来自五眼联盟等国家的网络间谍行为的干扰，如，2022 年西北工业大学遭美国安全局（NSA）侵入，2023 年中国多个政府机构通过微软 Outlook 邮箱系统被美国情报机构“后门”攻击等。而美国却以不存在相关领域的习惯国际法为由，对其网络间谍行为予以合理化。¹⁰⁶对此，我国应对“规模与效果”路径在此两个层面的适用持否定态度。在具体表达上，我国应当突出不干涉原则中目的要件的重要性，在同时主张主权原则不应当有最低限度后果的门槛。

最后，我国还应坚持“条约订立”与“传统国际法适用”相结合的“两条腿走路”的策略。推动新的条约制定，以弥补现行国际法对网络空间规制的不足。有观点认为，制定新条约并不能解决所有问题，且面临快速过时的风险，¹⁰⁷但这种观

¹⁰⁴ 参见张华：《网络空间适用禁止使用武力原则的法律路径》，载《中国法学》2022 第 2 期，第 300 页。

¹⁰⁵ 非盟认为，“规模”要素需要考察攻击的持续时间、受攻击目标的性质与位置、使用的武器类型等要素，而“效果”要素则衡量攻击造成的损害程度。See African Union Peace and Security Council, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace*, 29 January 2024, [https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_\(2021\)/CAP_Communique_FULL_0e34eb5799.pdf](https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_(2021)/CAP_Communique_FULL_0e34eb5799.pdf).

¹⁰⁶ 参见美国国防部总法律顾问保罗·内伊在 2020 年的演讲。Paul Ney, DOD General Counsel Remarks at U.S. Cyber Command Legal Conference, <https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/>

¹⁰⁷ Henrietta M. af R., *Online Analogies: The Legal Uncertainties of Cyberspace: A Study on Cyber Operations and the Jus ad Bellum*, 2023, <https://www.diva-portal.org/smash/get/diva2:1804992/FULLTEXT01.pdf>.

点无法成为拒绝订立新条约的理由。首先，国际法在适用过程中可以通过演进解释等条约解释方法，使其与时俱进。其次，国际法的发展本就是一个螺旋式渐进的过程，新条约能够对解决新问题提供更具确定性的法律依据。《联合国打击网络犯罪公约》的通过已为网络空间国际法的发展提供了良好的示范，国际社会可以围绕当前共同关注的议题，如供应链保护、虚假信息治理、网络基础设施安全等展开谈判，寻求共识。相较于从一般角度对网络空间的国际法适用进行整体规制，就特定问题进行规制能够避免更大程度国际法适用上的割裂。但如何设计专门性条约，以及如何促成国家间的新共识，仍需要进一步观察与研究。

六、结语

为推动传统国际法在网络空间的适用，回应有关违法性门槛界定的难题，“规模与效果”路径应运而生，并在不断拓展的实践中呈现出多样化的形态。“规模与效果”路径的兴起，是网络空间规范发展过程中的一个特色现象，亦是传统国际法走向新领域时所面临的深层变革之一。“规模与效果”路径在一定程度上弥合了传统规则与新兴情境之间的空白，回应了实践层面的现实需求，但其对所适用的具体原则、对网络空间国际法体系乃至对国际法整体结构可能带来的影响不容忽视。类比推理作为其逻辑基础，在为规则扩展提供工具的同时，也因其内在的或然性与目的导向性，使“规模与效果”路径在逻辑上面临一定的说服力缺陷，并易被操控以服务特定政治目标。在愈发激烈的网络空间规则博弈中，中国不应被动接受既有路径，而应主动参与塑造规则，通过适当区分、合理限缩与新规则倡议，推动构建更加稳健、公正、具备前瞻性的国际法体系。这不仅是对当前网络安全挑战的回应，也是对国际法未来走向的积极参与。

Reflections on the Expansion of the "Scale and Effect" Approach in the Application of International Law in Cyberspace

——An Observation on State Position Statements

Fang Zhenbang

The application of traditional international law to cyberspace faces persistent uncertainty regarding the threshold for determining the illegality of state conduct. The Tallinn Manual introduces the "scale and effects" criterion—originally developed to distinguish between the "use of force" and an "armed attack"—into the cyber context, positioning it as a common framework for assessing thresholds in cyberspace. This approach has gained broad support among states and is frequently cited in national position statements. It also manifests in diverse forms during the expansion of principles such as non-intervention and sovereignty into the cyber domain. The "scale and effects" approach represents a consequence-based mode of analogical application of international law. While it possesses a certain degree of formal authority and facilitates both judicial application and the emergence of preliminary state consensus, it suffers from inherent logical shortcomings. Specifically, it may contribute to asymmetrical application of international law in cyberspace, undermine the predictability of legal rules, and diminish the normative and declaratory functions of international law. China should remain cautious of the expansive use of this approach across various contexts, adopt a differentiated and context-sensitive stance, and actively advocate for the development of new international treaties as a complementary framework.

Keywords: Scale and Effect; Prohibition of the Use of Force; Non-intervention; Sovereignty; Analogy

人工智能生成儿童性虐待材料的国际法治理困局与规则重构

高杨洋

摘要：本文聚焦人工智能生成儿童性虐待材料的国际法治理难题，揭示其通过“篡改型”与“合成型”技术对儿童权利的双重侵害——前者依托真实儿童生物特征实施数字剥削，后者通过算法虚构场景破坏儿童群体尊严。当前国际法治理面临三重困局：国际人权法因“生物身份”保护框架局限形成虚拟侵权规范断裂；主权管辖与平台技术权力的冲突导致跨境治理失效；“技术中立”原则下创新自由与儿童权利保护的价值排序陷入伦理悖论。破解路径需重构法律规范、平衡技术权力、重塑伦理准则，通过确立儿童数字身份法益、构建技术连接点管辖体系、实施儿童权利优先的技术伦理审查，推动全球治理从碎片化探索转向体系化规则重构，为数字时代儿童权利保护提供国际法解决方案。

关键词：生成式人工智能 儿童性虐待材料 国际法治理 技术伦理

一、引言

技术是人类社会发展的助推剂，但同时也会成为用来扩大剥削的工具。¹这在生成式人工智能的勃兴中表现为，其既为创意产业注入革新动力，亦沦为儿童性剥削的数字化推手。欧洲刑警组织“坎伯兰行动”(Operation Cumberland)揭示，人工智能生成的儿童性虐待内容已占自制儿童性虐待材料的显著比例。²儿童网络性剥削材料制作者可以从已有的材料中生成新的在线儿童性剥削材料，制作更多被剥削和虐待儿童的图像。³

国际社会已经意识到并开始重视人工智能技术对儿童带来的性剥削问题，国际人权法面临监管困境，一些国家开始从立法入刑、颁布政策、技术检测、平台管理等角度探索监管和预防措施。⁴然而这些举措大多仍处在探索阶段，从传统儿童色情制品治理到人工智能生成儿童性虐待材料，全球治理范式转型迫在眉睫。

¹ Hosting Journalist Ed. Team, Ransomware, DDoS Attacks and Child Abuse Among Key Cybercrime Threats, Says Europol, HOSTING JOURNALIST(April 12,2024), <https://hostingjournalist.com/cybersecurity/ransomware-ddos-attacks-and-child-abuseamong-key-cybercrime-threats-says-europol/>.

² 欧洲刑警组织(Europol), “25 arrested in global hit against AI-generated child sexual abuse material”, <https://www.europol.europa.eu/media-press/newsroom/news/25-arrested-in-global-hit-against-ai-generated-child-sexual-abuse-material>.

³ Christian Berg, The Main Challenge Is Victim Identification, NETCLEAN (May 14, 2019), <https://www.netclean.com/2019/05/14/christian-berg-the-main-challenge-is-victimidentification/>.

⁴ Brian Dolhansky et al., “The Deepfake Detection Challenge (DFDC) Preview Dataset,” ArXiv:1910.08854 [Cs], October 23, 2019, <http://arxiv.org/abs/1910.08854>; Vasileia Karasavva and Aalia Noorbhai, “The Real Threat of Deepfake Pornography: A Review of Canadian Policy,” Cyberpsychology, Behavior and Social Networking 24, no. 3 (n.d.): 23–209, <https://doi.org/10.1089/cyber.2020.0272>; Mika Westerlund, “The Emergence of Deepfake Technology: A Review,” Technology Innovation Management Review 9, no. 11 (January 1, 2019): 39–52, <https://doi.org/10.22215/timreview/1282>.

二、人工智能生成儿童性虐待材料的技术谱系与危害本质

人工智能生成儿童性虐待材料（AI-Generated Child Sexual Abuse Material, AI CSAM）是指借助生成式人工智能技术，包括深度伪造、文本到图像、视频模型、3D 建模等，以儿童为主体或模拟儿童特征创作的性虐待内容。⁵这类材料既包括对真实儿童形象的篡改、合成，也涵盖完全虚构的虚拟性虐待场景，其核心特征是通过技术手段对儿童人格尊严、身体自主权及社会形象造成数字化侵害。

（一）从“特征篡改”到“无中生有”的技术演进路径

根据技术实现路径，AI CSAM 的技术生成可分为两大核心分支：一类是篡改型生成技术，以真实儿童的生物特征为基础，如面部图像、身体轮廓、语音数据等，通过深度伪造算法进行扭曲、嫁接或合成，制造虚构的性虐待场景，例如利用 FaceSwap 等模型将成人面部“幼龄化”后嵌入色情画面，或通过 3D 姿态估计技术篡改真实儿童的肢体动作以模拟虐待姿态，此类技术对真实儿童构成二次剥削，即使未公开身份也会侵害其身体完整性与社会形象；⁶另一类是合成型生成技术，完全依赖算法生成虚拟儿童形象及性虐待场景，无需真实个体数据，典型方式包括通过 Stable Diffusion 等文本到图像模型按指令生成高逼真图像，或利用 3D 建模与对抗生成网络（GAN）构建虚拟儿童身体数据并渲染虐待场景。

（二）“真实关联型”与“虚拟建构型”的二元危害范式

根据是否存在物理世界的权利侵害载体，AI CSAM 同样可分为两类：真实关联型 CSAM 以真实儿童的生物特征为基础，通过换脸、身体合成、语音克隆等技术制造虚假性虐待内容，其法律本质是对具体受害者的直接侵害，技术手段加剧了危害的隐蔽性，受害者可能在不知情的情况下成为数字性暴力主角，且图像的无限传播导致伤害持续扩大；⁷虚拟建构型 CSAM 完全由算法生成虚拟儿童形象及场景，不涉及真实个体数据，包括特征组合型（如生成“8 岁亚洲女孩”虐待场景）与纯虚构型（如标准化虚拟儿童身体执行性动作），其核心法律挑战在于突破“无受害者即无犯罪”的传统构成要件，尽管不存在物理受害者，但其传播与消费会强化儿童性剥削合理化的文化想象，构成对儿童群体整体尊严的象征性侵害。

两类 AI CSAM 的本质分野在于权利侵害是否存在物理锚点。前者依赖真实儿童的生物特征，挑战生物特征与权利主体相对应的法律逻辑；后者通过算法建构虚拟受害者，颠覆“受害者可识别性”的证据规则。但二者共享技术滥用的核心危害，即通过视觉化、场景化的性虐待内容强化对儿童的物化、剥削与暴力想象，

⁵ See Internet Watch Foundation, "AI-Generated Child Sexual Abuse Material Surges", July 23, 2024, <https://www.iwf.org.uk>.

⁶ Citron, Danielle Keats, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security", California Law Review, Vol. 107, No. 6, 2019, pp. 1763-1820.

⁷ 参见刘艳红：《生成式人工智能安全监管的法治逻辑》，《中国法学》2024 年第 6 期，第 5-23 页。

侵蚀社会对儿童权利保护的价值共识。“我们保护全球联盟”（WeProtect Global Alliance）2023 年报告显示，2019-2023 年间全球 AI CSAM 报告量激增 87%，且受害者年龄呈现低龄化趋势，⁸凸显技术迭代对传统治理框架的降维打击。

三、国际法治理困局的三重解构

对于 AI CSAM 问题，当前全球治理体系正面临三重范式断裂。其一，国际法秩序仍锚定物理空间的真实儿童、直接伤害的二元框架，而 AI 犯罪已进化为虚拟生成、分布式传播的数字形态；其二，国家主权在算法跨境流动中呈现碎片化，平台技术权力却借“技术中立”主张形成治理黑洞；其三，伦理共识在技术创新自由与数字儿童权利的价值排序中陷入结构性矛盾。

（一）制度性困局：虚拟侵权的规范断裂与保护客体错位

AI CSAM 对国际人权法的制度性冲击，本质是技术生成的“数字身份”与传统法律“生物身份”的保护体系冲突。作为儿童权利保护的基础性国际法律文件，《儿童权利公约》（以下简称《公约》）第 16 条构建的隐私权保护框架，其规范构造以物理空间的可识别性为基础，将“隐私”范畴局限于物理身份的人格要素，形成生物特征关联的保护要件。⁹这种规范构造的核心在于，将权利侵害与具体物理主体的可识别性绑定，唯有存在真实儿童作为权利载体时，法律救济机制方可启动。

然而，生成式 AI 技术通过两种路径突破了上述规范边界：对于合成型 CSAM 而言，施害者利用 StyleGAN 身份剥离算法、文本到图像模型等技术，完全虚构与真实儿童无生物特征关联的虚拟形象并实施性虐待场景生成。此类行为因缺乏物理载体，导致《公约》第 16 条的生物特征关联要件无法适用，形成无实体受害者的权利真空；对篡改型 CSAM 而言，通过深度伪造技术对真实儿童图像进行面部年轻化改造、身体部位合成等处理，虽以真实儿童生物特征为基础，但侵害后果已从物理空间的人格要素延伸至数字空间的身份重构。由于《公约》未将“数字身份”纳入保护范畴，致使法律规制仅能针对图像传播行为，而无法追溯技术生成环节的核心侵权行为。

各国司法实践的殊化发展进一步凸显了国际规范的滞后性。在北京互联网法院 2020 年“AI 陪伴案”¹⁰中，法院虽承认未经授权生成自然人虚拟形象构成人格权侵害，但裁判逻辑仍依赖姓名、肖像等具体人格要素的可识别性，对不依附于真实主体的纯数字身份侵权问题未形成有效规制。美国联邦法院在“史蒂文·

⁸ WeProtect Global Alliance: Global Threat Assessment 2023 Data, <https://www.weprotect.org/>.

⁹ 《儿童权利公约》第 16 条，联合国大会第 44/25 号决议，1989 年 11 月 20 日通过。

¹⁰ “AI 陪伴”案：未经同意创造自然人的 AI 形象构成对人格权的侵害——何某诉某科技公司网络侵权责任纠纷案，北京互联网法院（2020）京 0491 民初 9526 号。

安德雷格案”¹¹中则展现了技术时代的分层治理思路。通过“斯坦利案”¹²确立的“家庭隐私神圣”原则与“自由言论联盟案”¹³的司法先例，明确区分“真实儿童”与“虚拟生成内容”的法律属性——即基于思想自由与家庭隐私的宪法价值，私人持有虚拟儿童淫秽内容受《第一修正案》保护，而生产传播涉及对儿童权益的潜在危害，可依据《联邦儿童淫秽物品法》1466A 条款¹⁴追责，无需证明涉及真实儿童。

这种由技术匿名性与规范滞后性交织形成的“合成豁免”困境，集中暴露了国际法在虚拟身份侵权保护中的结构性缺陷。《公约》未能构建“技术生成型侵权”的规制范式，对 AI 技术引发的“无实体侵害”特性缺乏应对方案，同时其机械的解释方法导致“隐私”范畴未能随技术进步扩展至数字空间，形成保护客体的历史性错位。

（二）结构性困局：主权管辖边界与平台权力的重构挑战

技术滥用的全球化特质凸显了国际协同治理机制的制度性失灵，传统主权国家基于属地原则构建的管辖权体系，在算法跨境流动与数据云存储模式下遭遇系统性挑战。自 20 世纪末互联网全球化浪潮以来，数据主权与平台权力的冲突即已显现。¹⁵如今，主权国家与跨国平台间的法律博弈在 AI CSAM 治理中进一步升级为主权让渡与技术权力扩张的深层矛盾。

一方面，犯罪行为突破物理空间限制，通过数字技术将侵害范围扩展至全球。例如，英国 2024 年尼尔森（Hugh Nelson）案中，犯罪嫌疑人利用 3D 角色生成器将真实儿童照片转化为性虐待图像，并通过互联网论坛向法国、意大利、美国等国家的客户出售，其犯罪链条涉及的受害儿童更是分布多国。¹⁶此类案件暴露出传统属地管辖原则在数字空间的适用困境，即便犯罪者已在行为实施地被定罪，受害者的跨国分布与证据链的碎片化仍导致司法救济难以实现。

另一方面，平台企业的技术架构与商业逻辑形成“治理黑洞”，主权国家难以穿透算法黑箱实施有效监管。2024 年 4 月欧盟根据《数字服务法案》(Digital Services Act)对 Meta 公司¹⁷是否可能在与未成年人保护相关的领域违反法律启动

¹¹ United States v. Anderegg, No. 3:24-cr-00064 (W.D. Wis. 2025).

¹² Stanley v. Georgia, 396 U.S. 557, 1969

¹³ Ashcroft v. Free Speech Coalition, 535 U.S. 234, 2002

¹⁴ 18 U.S.C. § 1466A.

¹⁵ 例如 2007 年比利时与雅虎公司的跨境数据纠纷案即为典型例证。该案中，比利时司法机关因一起网络诈骗案向美国雅虎公司发出数据调取令，要求其提供涉案邮箱用户信息，但雅虎以美国《电子通信隐私法》(ECPA)禁止未经司法部命令披露用户数据为由拒绝配合。最终比利时最高法院通过扩张解释属地管辖原则，认定雅虎在比服务构成司法管辖权基础。案件详情参见法治周末报道《雅虎，交出你的数据》，2019 年 2 月 13 日，https://www.legaldaily.com.cn/index/content/2019-02/13/content_7806226.htm（最后访问日期：2025 年 4 月 15 日）。

¹⁶ 案件详情参见英国《卫报》2024 年 10 月 28 日报道，<https://www.theguardian.com/uk-news/2024/oct/28/hugh-nelson-ai-child-sex-abuse-images-sentenced>，以及英国天空新闻 2024 年 10 月 28 日报道，<https://news.sky.com/story/paedophile-hugh-nelson-who-made-ai-child-abuse-images-from-real-pictures-sent-to-him-jailed-for-18-years-in-deeply-horrifying-landmark-case-13220848>。

¹⁷ 知名社交平台 Facebook 和 Instagram 的母公司。

正式调查，但 Meta 未公开核心参数，导致监管难以深入。¹⁸在后续的监管中，欧盟又于 2024 年 10 月再次要求 YouTube、Snapchat、TikTok 提交包括与未成年人保护相关的算法推荐参数及系统性风险评估。¹⁹尽管一再强化监管，但对于核心算法机制，平台仍掌握最终解释权。

平台企业的技术治理权能优势，已然引发国际法框架下的合法性危机。根据《儿童权利公约》第 19 条，缔约国负有保护儿童免受“一切形式的身心摧残”的义务，但平台企业基于商业利益的内容审核标准往往与国家义务相冲突。例如，为规避法律风险，平台可能过度删除敏感内容，而忽略儿童权益保护的复杂性。三星德国 2025 年数据泄露事件则表明，过度依赖平台自主治理也会导致数据安全风险激增。该事件中，黑客通过攻击第三方服务商 Spectos 获取了 27 万份三星德国客户数据，包括姓名、地址、支付详情等敏感信息，暴露了平台在数据安全管理体系中的漏洞。²⁰种种事例表明，数字时代的全球治理秩序正面临主权国家与跨国平台的二元权力结构的重构挑战。

（三）伦理性困局：技术创新自由与儿童权利保护的价值冲突

数字时代 AI CSAM 治理的伦理困局，本质在于技术工具理性与人类价值理性在儿童权利保护场域的深度碰撞。生成式 AI 技术对技术中立伦理边界的突破，使国际社会在技术创新自由与儿童权益保护的价值排序中陷入系统性分歧，暴露出技术伦理责任与法律规制的深层断裂。

跨国互联网企业往往以“技术中立”为名构建责任避风港，拒绝为算法推荐导致的 CSAM 传播担责。然而技术架构的文化偏见始终潜藏于所谓中立原则之上，国际社会对 AI 技术的工具性与价值嵌入性的责任归属一直存在争议。国际刑警组织 2025 年数据显示，Meta 公司的 AI 审核系统因训练数据中欧美儿童面部特征占比超过 80%，导致对亚非拉裔儿童图像的漏检率高达 34%，而发展中国家儿童在 AI 生成 CSAM 中的受害占比已超过 60%。²¹这种以“技术中立”包装的系统性偏见，实质是技术设计者文化认知局限的产物，却成为平台规避地域治理责任的借口，导致全球治理陷入“技术殖民”的伦理困境。联合国儿童基金会（UNICEF）2024 年报告指出，72% 的 AI CSAM 传播与商业平台的中立技术架构相关，呼吁

¹⁸ 由于算法黑箱的缘故，尽管 Meta 公司声称已经花费多年开发了 50 多种工具和政策，旨在保护年轻人的安全上网体验，然而欧盟监管机构“不认为 Meta 已经做了足够的工作来遵守 DSA 义务——以减轻其平台 Facebook 和 Instagram 对欧洲年轻人身心健康的负面影响”。参考 EU investigates Meta over Facebook and Instagram child safety, 16 May 2024 <https://www.aljazeera.com/news/2024/5/16/eu-investigates-meta-over-facebook-and-instagram-child-safety>

¹⁹ Shaping Europe's digital future, Commission sends requests for information to YouTube, Snapchat, and TikTok on recommender systems under the Digital Services Act, 02 October 2024, <https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-youtube-snapchat-and-tiktok-recommender-systems-under-digital>

²⁰ 三星德国数据泄露事件详情参见 ITBear 科技资讯 2025 年 4 月 9 日报道，<http://m.itbear.com.cn/html/2025-04/780529.html>

²¹ Interpol, Global Report on AI-Generated CSAM, 2025, p. 18.

建立技术应用的儿童权利影响评估机制。²²

技术检测与规避的博弈揭示了更深层的价值冲突，而 AI 技术的双重属性则进一步激化了这种伦理困境。生成式 AI 技术的快速迭代²³已突破传统检测算法的识别阈值，互联网监督基金会（IWF）对暗网传播的 3512 张 AI 儿童性虐待图像进行分析，发现其中 90% 的图像都相当逼真，肉眼难以分辨出与真实性虐待图像的区别。²⁴传统哈希值比对系统在对抗生成网络（GAN）制造的“隐形图像”面前效能大幅衰减。²⁵美国国家标准技术研究院（NIST）2025 年测评显示，主流 AI 检测工具对深度伪造儿童性虐待内容的误报率仍达 37%，漏检率为 12.4%。²⁶这种技术对抗本质上是算法透明度与隐私权的价值冲突，平台以保护用户隐私和商业秘密为由拒绝共享检测数据，而执法机构因缺乏技术细节难以精准打击犯罪，形成检测技术越先进，规避手段越隐蔽的治理死循环。

AI 技术“犯罪工具”与“监管工具”的双重属性还催生了“以恶制恶”的伦理争议。在犯罪分子利用 Stable Diffusion 等 AI 工具生成高度逼真的虚拟 CSAM 的同时，执法机构与公益组织也正探索技术反制路径。美国联邦调查局通过创建 AI 生成的虚构儿童色情网“Playpen”，获取下载者 IP 地址及交易信息，最终关闭该网站并逮捕 350 余名犯罪者；²⁷荷兰“Terre des Hommes”组织开发虚拟女孩“甜心”（Sweetie），通过 AI 对话技术诱捕聊天室中的潜在罪犯，协助国际刑警组织打击数千起网络儿童性剥削案件。²⁸这种以技术对抗技术的治理范式虽取得阶段性成效，却陷入制造技术恶果与利用技术手段的伦理困境，当监管者采用与犯罪者同源的 AI 技术实施诱捕时，数字正义的边界究竟何在？技术工具的中立性又在何种程度上被目的理性消解？再次凸显了数字时代工具理性与价值理性难以调和的根本矛盾。

四、国际法治理困局的破解路径

（一）制度性困局破局：从生物特征保护到数字身份治理

破解 AI CSAM 引发的制度性困局，需突破《儿童权利公约》以物理空间为中心的规范局限，构建生物身份+数字身份的双重保护体系，实现从实体侵害规制到技术风险预防的治理范式转型。

1、确认儿童数字身份法益地位

²² UNICEF, AI and Child Rights: A Global Assessment, 2024, p. 45.

²³ 如 Stable Diffusion v3.5 的多模态生成能力。

²⁴ 参见互联网监督基金会（Internet Watch Foundation, IWF）, 2024 Update: Understanding the Rapid Evolution of AI-Generated Child Abuse Imagery, p.7, <https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/>.

²⁵ Europol, "Technical Analysis of AI-Generated CSAM in 2025", EUROPOL Tech White Paper, 2025.

²⁶ NIST, "AI-Generated Child Sexual Exploitation Material Detection Benchmark", March 2025.

²⁷ U.S. Department of Justice, "Florida Man Sentenced to Prison for Engaging in a Child Exploitation Enterprise", May 1, 2017, <https://www.justice.gov>.

²⁸ Leslie Katz, "Meet 'Sweetie,' a virtual girl created to target child predators", CNET, <https://www.cnet.com>.

国际社会应通过《儿童权利公约》的动态解释与专项立法，确立数字身份的法益地位。依据《维也纳条约法公约》第 31 条“善意解释”原则，对《公约》第 16 条隐私权条款作目的性扩张，将通过技术手段生成的儿童虚拟形象纳入人格权保护体系，明确即使未依附于真实儿童的合成内容，若其形象可被特定化识别或具有儿童特征，即构成对儿童尊严权的侵害。

同时可通过制定专门议定书的方式，建构三类核心规则：其一，突破“真实儿童”要件限制，凡呈现儿童样态的性虐待内容均构成对“儿童群体尊严”的抽象侵害；其二，增设“数字身份完整权”，禁止未经同意修改儿童生物特征；其三，要求平台实施“生成式 AI 内容全生命周期审查”，包括训练数据清洗、哈希值比对与实时年龄分级。此类规范更新可弥合《公约》文本的静态性与技术发展的动态性之间的张力。

2、区分技术侵权二元逻辑

针对合成型与篡改型 CSAM 的不同侵权机理，应建立差异化的法律规制框架。对于合成型 CSAM，其核心危害在于通过算法虚构儿童形象实施象征性侵害，破坏社会对儿童尊严的伦理共识。对此，可参照美国《联邦儿童淫秽物品法》第 1466A 条对虚拟儿童形象的界定，²⁹在国际法中创设“拟制保护客体”制度，将具有儿童外观特征的虚拟形象纳入保护范围，无论是否存在真实儿童作为物理载体，只要生成内容具有性虐待性质，即构成违法。这种“形式认定标准”的建立，有效解决了“无实体受害者”的追责困境。对于篡改型 CSAM，其侵权本质是对真实儿童生物特征的数字化滥用，需强化对数字身份滥用的规制。可参考中国《民法典》第 1019 条关于禁止利用信息技术手段伪造侵害他人肖像权的规定，³⁰在公约中确立“技术处理行为违法性”认定原则，明确对真实儿童图像的深度伪造处理，即使未造成实际传播后果，技术生成行为本身即构成对儿童数字身份权的侵害，需承担民事乃至刑事责任。

3、建立全链条责任体系

国际法需突破单一环节规制的局限，构建覆盖技术开发、内容生成、跨境治理的立体化责任体系。在技术开发端，应落实 UNESCO《人工智能伦理问题建议书》提出的“责任链原则”，³¹要求生成式 AI 模型开发者履行儿童权利影响评估义务，对可能被用于 CSAM 生成的包括面部合成算法、身体生成模型在内的关键技术模块进行伦理审查，并向联合国儿童权利委员会提交风险评估报告，未通过审查的技术不得投入使用。这一机制旨在从源头阻断技术滥用的可能性。在内容生成端，需确立“无受害者侵害”的法律认定标准，借鉴欧盟《数字服务法》的预防性删除义务，规定任何利用 AI 生成儿童性虐待内容的行为，无论是否涉及

²⁹ 18 U.S.C. § 1466A (a)(1)-(2)

³⁰ 《中华人民共和国民法典》第 1019 条，2021 年施行。

³¹ UNESCO, Recommendation on the Ethics of Artificial Intelligence, 2021, Art. 5(c).

真实儿童，均构成对《公约》第 19 条“儿童免受身心摧残”义务的违反，缔约国应将此类行为纳入刑事犯罪范畴。在跨境治理端，针对 AI 合成型 CSAM 的跨国传播特性，可引入“技术来源地管辖”原则，赋予技术服务器所在国刑事追诉权，即使侵害后果未及于本国境内，仍可依据国际公约启动司法程序，避免因“无实体受害者”导致的管辖真空。

（二）结构性困局破局：从主权管辖冲突到平台责任平衡

数字时代的主权管辖困境与平台权力扩张，本质上是物理空间治理规则与数字技术跨境流动性的内在冲突。这一结构性困局的根源在于传统主权理论难以适应技术权力的全球化扩张，以及跨国平台通过算法架构形成的私权力对公权力的消解。基于既有实践与理论探索，解决路径需以重构主权与平台权力的互动关系为核心，通过技术赋权与制度创新实现治理效能的提升。

1、确立技术连接点管辖体系

传统威斯特伐利亚主权模式在数字空间遭遇属地管辖失效的困境，单纯依赖领土主权难以应对数据跨境流动的复杂性。网络主权的核心在于对代码层和基础架构的控制，而非简单的领土延伸。³²针对 AI CSAM 的跨国传播特性，国际社会应超越物理存在的单一管辖标准，建立以“技术连接点”为核心的多元管辖权规则。这一体系的核心在于将服务提供地、数据处理地、危害结果地等技术相关要素纳入管辖依据，弥补传统属地原则的适用盲区。在 AI CSAM 治理中，对于通过互联网向多国用户传播合成型 CSAM 的行为，用户所在地或服务接受地国家可依据“效果原则”³³主张管辖权；对于利用云服务器进行 AI 模型训练或内容生成的行为，数据处理地国家可基于技术实施地启动调查程序，避免因犯罪行为的地域化导致的管辖真空。同时逐步建立跨境数据托管与争议解决机制，借鉴国际海底管理局的中立托管模式，由全球数据治理委员会对涉及多国利益的数据实施技术中立分析，避免单一国家因数据主权争议陷入管辖僵局，防止跨国平台利用技术垄断规避法律责任。这种将技术要素与主权理论相结合的治理框架，为破解数字空间的管辖权冲突提供了法理支撑，使主权的刚性约束与技术的柔性流动形成治理合力。

2、升级平台风险控制义务

跨国平台的技术优势与商业逻辑不应成为儿童权益保护的治理黑洞，国际法需明确其作为数字空间管理者的积极义务，构建与技术能力相匹配的责任体系。首先，应确立平台的算法透明度义务，要求其向监管机构披露与儿童性虐待材料传播相关的核心技术参数，尤其是针对未成年人的内容推荐机制，防止平台以商

³² 沈伟伟，《网络主权视角下的数据治理比较研究》，载《法学杂志》2024 年第 4 期。

³³ 效果原则是美国法院在 1948 年的“美国铝公司”案中提出的，指当公司在外国行为对国内产生“效果”时，就对其行使管辖权。

业秘密为由规避监管。其次，建立内容审查的国际法标准，要求平台对 AI 生成内容实施分级分类管理：对于合成型 CSAM，需整合国际刑警组织的哈希值数据库进行实时比对过滤；对于篡改型 CSAM，需运用生物特征识别技术阻断对真实儿童图像的非法处理。前述英国尼尔森案中，犯罪者通过公开论坛销售 AI 生成图像的行为，暴露出平台在用户异常行为监测上的缺失，国际法应据此要求平台建立自动化审查机制，对涉及儿童图像的高频生成、传播行为触发人工复核程序。通过将技术能力转化为法律义务，可有效遏制平台因商业利益驱动导致的治理不作为。

3、构建全球协同治理机制

应对数据主权与平台权力的深层矛盾，需突破国家中心主义的单一治理模式，构建主权国家、跨国平台、技术共同体协同参与的治理架构。在主权协作层面，各国应通过缔结专门应对 AI 侵害的儿童权利保护国际公约，统一跨国取证程序、责任认定标准及司法协助机制，解决传统司法协助中证据链碎片化问题。例如，确立主要损害发生地国家优先管辖原则，同时要求缔约国建立 72 小时紧急数据调取通道，提升跨境执法效率。在技术共治层面，由联合国儿童基金会牵头成立一个“全球儿童数字保护联盟”，推动制定 AI 生成内容的技术标准，包括虚拟儿童形象的年龄识别算法、合成内容的数字水印技术等，为各国监管提供技术支撑。单个国家通过技术手段追踪虚拟儿童色情网站的实践，可转化为国际层面的技术共享机制，促进执法机构与科技企业的合作。此外，需建立平台责任的分层评估体系，区分包括云服务商在内的基础设施平台与内容服务平台的义务边界：前者需履行跨境数据协助义务，后者需承担内容审核与用户管理的直接责任，通过合规认证与市场准入的挂钩机制，形成全球性的责任约束网络。

（三）伦理性困局破局：从技术中立悖论到儿童权利优先

对于 AI CSAM 的伦理性困局破局之道，需要通过责任体系重构、检测网络协同与反制边界划定，将儿童权利保护内化为技术创新的核心约束条件，促使工具理性与价值理性形成治理合力。这一过程不仅需要技术层面的突破，更依赖全球治理从利益博弈向价值共识的根本性转变，最终构建既包容技术创新活力又坚守人类伦理底线的数字治理新范式。

1、重构技术伦理的责任范式

突破“技术中立”的伦理遮蔽与责任模糊地带，需构建以“儿童权利优先”为核心的技术伦理责任范式，将儿童权益保护内化为技术设计的元规则。国际社会应通过立法确立技术开发者与平台运营者对 AI CSAM 的“过错推定责任”，改变技术工具化的责任剥离逻辑，要求企业在算法开发阶段嵌入“儿童特征识别阻断模块”，并建立训练数据的文化多样性校准机制，强制规定发展中国家儿童面部特征在训练数据中的占比不低于地域人口比例，从数据采集与模型训练源头消除

系统性偏见。针对跨国互联网公司等平台审核系统的地域失衡问题，可推动建立跨区域儿童特征数据共享机制，确保全球不同种族儿童形象在技术架构中的公平覆盖。实施“儿童权利影响评估”刚性标准，要求所有生成式 AI 产品上线前通过独立第三方机构的全生命周期伦理审查，将合规性与企业信用评级、跨境数据流动权限挂钩，形成从技术设计、数据训练到应用准入的事前主动规避机制。此外，还可以借鉴医疗设备召回制度，对存在 CSAM 生成漏洞的 AI 模型实施强制修复与下架，通过伦理漏洞奖励计划活动鼓励社会力量参与技术伦理监督，推动技术主体责任从事后被动担责转向事前价值嵌入，实现技术工具理性与儿童权利保护的深度耦合。

2、重塑技术工具的价值赋能逻辑

化解技术工具理性与价值理性的冲突，需以“儿童最大利益不可逆”原则重塑技术应用的价值赋能逻辑，推动 AI 从“功能实现”向“价值共融”转型。在技术研发阶段嵌入儿童权利保护的刚性约束，建立类似绿色电力证书机制的伦理算力认证体系，通过计算资源分配优先支持符合儿童保护标准的技术开发，同时培育数字监护人生态，由各国政府主导购买公益组织的反制技术服务，形成技术向善的正向激励循环。针对以技术对抗技术的伦理争议，建立覆盖技术应用全流程的“目的限定”治理准则，事前由联合国儿童权利委员会等独立第三方机构审查 AI 诱捕技术方案，明确禁止生成真实儿童形象、诱导犯罪行为或过度采集隐私，如规定虚拟形象与真实儿童特征重合度 $\leq 15\%$ ；事后通过未成年人代表参与的监督机制定期评估技术效果，确保反制手段的伤害性控制在最小范围。另外需构建生成式 AI 分级监管体系，对高风险人物形象生成技术实施严格管控，强制植入“儿童形象自动阻断功能”并缴纳专项风险保证金；中低风险技术实行备案制并公开基础算法逻辑；低风险技术采取宽松监管，在创新自由与风险防控间建立弹性平衡，最终通过目的正当性限定、程序正义审查、效果动态评估的三重机制，将技术创新自由严格限定在儿童权利保护的伦理边界内，使 AI 技术既成为打击犯罪的利器，更成为守护儿童尊严的数字盾牌。

五、结语

人工智能生成儿童性虐待材料的治理危机，本质是技术革命对既有法律秩序的系统性挑战。当生成式 AI 突破物理世界的侵害边界，国际社会必须超越传统法律框架的路径依赖，在技术创新与权利保护的张力中重构治理范式，从生物特征保护转向数字身份治理，从主权管辖冲突走向技术权力平衡，从技术中立博弈升维至儿童权利优先的价值共识。

这一进程既需要《儿童权利公约》等国际法律文件的动态解释与规则更新，更依赖技术共同体、主权国家与跨国平台的协同共治，通过建立全球统一的技术标准、弹性的责任分配机制与包容的伦理审查框架，将儿童权利保护内化为技术

开发与应用的底层逻辑。唯有如此，方能在数字文明浪潮中筑牢儿童权利的防护堤，实现技术工具理性与人类价值理性的真正和解。未来研究可进一步聚焦具体治理工具的实效评估，为国际法规则的落地提供更精细的实践指引。

2025年中国国际法学会年会

《非盟共同立场文件》：“全球南方”参与网络空间国际规则博弈的新范式

林嘉悦*

摘要：发布国家立场文件已成为澄清并发展网络空间国际规则的一大态势，“全球南方”在此方面表现稍显沉默，而非盟延续了“抱团”集体发声的思路，制定并发布了凝聚 55 个“全球南方”共识的《非盟共同立场文件》。在制定方式上，非盟在能力建设与完善草案阶段均坚定践行“对外寻求援助、对内协调立场”的制定思路，从而促使资源整合最大化，提升组织分工运作效率并稳步提升了文件的法律专业性。在实质内容方面，《非盟共同立场文件》呈现出强烈的主权意识，对于主权原则、不干涉原则与禁止使用武力原则的相关问题的澄清提供了来自“全球南方”视角的阐述。作为区域集体立场的制度化表达，《非盟共同立场文件》不仅成功凝聚了非洲内部的共识，还为网络空间国际规则的发展做出了来自“全球南方”的积极贡献，是值得其他“全球南方”学习选取发声策略与酝酿发声内容的新范式。

关键词：网络空间 国际规则 全球南方 非洲联盟 立场文件

随着网络空间战略地位不断上升，网络空间的秩序构建与规则博弈在国际议程中的位置加速前移，网络空间建章立制的进程已然开启。作为陆海空天之后的第五疆域，网络空间为“全球南方”参与国际法制定与适用提供了全新的机遇，“全球南方”有机会从“局外人”转变为真正的国际法制定主体，将自身诉求转化为国际法话语，完成国际法从“破旧”到“创新”的转变，为网络空间国际法治化进程助力。非洲联盟和平与安全理事会于 2024 年 1 月发布了《关于在网络空间适用国际法的非洲共同立场》（Common African Position on the Application of International Law to the Use of ICTs submitted by African Union，简称《非盟共同立场文件》）。¹此文件凝聚了非洲联盟 55 个发展中国家的共识，既打破了非洲发展中国家在网络空间国际法事务中的沉默状态，也创造了“全球南方”参与国际法规则博弈的新范式。因此，本文以《非盟共同立场文件》为核心研究对象，通过对该文件发布背景、制定过程与核心内容的剖析，探寻“共同立场”这一范式为何应该且能够成为“全球南方”参与网络空间国际规则博弈的新范式，从而为“全球南方”切实从“规则接收者”向“规则制定者”转变提供启示。

* 作者简介：林嘉悦，女，武汉大学国际法研究所博士研究生，主要从事网络空间国际法研究。

¹ Peace and Security Council of Africa Union, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace*, PSC/PR/COMM.1196 (2024).

一、《非盟共同立场文件》的发布背景

（一）发布立场文件成为重要博弈方式，而“全球南方”表现不佳

网络空间国际规则博弈进入深水区，通过高管演讲或书面声明的方式发布立场文件已经成为网络空间国际规则塑造的一大途径。正如联合国信息安全政府专家组在其 2021 年报告所强调的，“各国在联合国就国际法的具体规则和原则如何适用于各国使用信通技术的问题继续进行集体讨论和交换意见，对于加深共同理解、避免误解、提高可预测性和稳定性而言是不可或缺的。各国之间的区域和双边意见交流，可以为此类讨论提供参考和支持。”²而网络空间新近国家立场声明作为各国之间的法律论辩，不仅仅是对现有国际法在网络空间适用的解释实践，同时也可能成为塑造新的习惯国际法规则的证据，对网络空间国际法的澄清与发展有着深远影响。³

然而，南北方国家在发布立场文件方面的表现差距较大。从数量来看，截至非盟发布《非盟共同立场文件》之时，共有 31 个国家和地区通过提交立场文件、高官演讲等形式阐述其对于国际法适用于网络空间的观点，而其中发布单方立场文件的“全球南方”仅有 10 个。⁴从目前各国立场文件法律意见的质量来看，可将世界各国大致分为三个梯队：处于领先地位的“北方国家”属于“第一梯队”，这些国家以高级官员讲话和提交材料的形式发表了立场声明，其中包括国际法现有各种原则和规则的适用，篇幅相对较长，法律推理也比较详细；⁵少数正在奋起直追的“全球南方”属于“第二梯队”，这些国家主要以书面形式表达本国立场，但立场较为原则，法律论证一般较为薄弱和简短；其他大多数根本无法有效参与和影响网络空间国际规则制定的“全球南方”则属于“第三梯队”。从整体角度观察，“全球南方”基本处于第二梯队和第三梯队。

（二）通过“共同文件”集体发声是非盟的惯常实践

受到非洲本土哲学思想乌班图（Ubuntu）的影响，非洲国家在国际问题上常常倾向“抱团”。⁶而通过外交谈判和建立共识来制定、推广和捍卫非洲立场，已成

² UNGA, Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, A/76/135, 14 July 2021, para. 72.

³ 黄志雄、罗旷怡：《各自为“辩”：网络空间新近国家立场声明的总体考察与中国因应》，《云南社会科学》2023 年第 6 期，第 81 页。

⁴ 在 2024 年 1 月之前，发布国家立场文件的“全球南方”有伊朗（2020 年）、中国（2021 年）、巴西（2021 年）、哈萨克斯坦（2021 年）、肯尼亚（2021 年）、俄罗斯（2021 年）、罗马尼亚（2021 年）、哥斯达黎加（2023 年）和巴基斯坦（2023 年）。

⁵ See U.N.G.A., Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States Submitted by Participating Governmental Experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security Established Pursuant to General Assembly Resolution 73/266, A/76/136, 13 July 2021.

⁶ 周鑫宇：《南非乌班图思想与新兴大国土政治思想崛起》，载《现代国际关系》2018 年第 2 期，第 58 页。

为成员国与非盟关系中的一重要实践。⁷“共同文件”是一种在非盟团结、统一和共识的核心原则与共同价值观影响下，并受到泛非主义和非洲复兴理想推动的集体发声方式。非盟制定了不少非洲共同政策立场，在全球舞台上产生了一定影响。⁸采用“共同发声”这一方式参与网络空间国际制定博弈，符合非盟的一贯实践，也契合非盟《2063 年议程》表达的“一个一体化、繁荣与和平的非洲，由其公民驱动，在国际舞台上代表一股充满活力的力量”的愿景。在此背景下，2022 年 11 月，非盟和平与安全理事会通过了一份里程碑意义的决议，“强调亟需就国际法在网络空间的适用达成非洲共同立场，以及非洲需要积极参与阐明这方面国际法的进程”。⁹由此开启了《非盟共同立场文件》的制定。

二、《非盟共同立场文件》的制定过程及其特点

（一）《非盟共同立场文件》的制定过程

与以往偏重政策表达性质的“共同文件”不同的是，《非盟共同立场文件》是一份偏重法律话语的意见表达。因此，非盟在意识到起草并发布《共同文件》的必要性与紧迫性的同时，也充分认识到自身能力的不足，采取了“对外寻求援助、对内协调立场”的制定思路。根据特别报告员的阐述，《非盟共同立场文件》的制定大致分为两个阶段，即能力建设培训与文件实际起草。

第一阶段侧重于能力建设。在能力方面的限制是非洲国家阐述国际法适用于网络空间的立场观点所面临的主要障碍。因此，起草工作第一阶段的重点是为非洲外交官、专家和政府律师组织能力建设项目。该计划由加拿大全球事务部（Global Affairs Canada）资助并与非洲联盟共同组织，分为三部分于 2023 年 3 月至 7 月举行，培训对象包括各成员国驻非洲联盟总部的代表、各成员国外交部和其他政府机构的专家以及非洲联盟委员会的官员，还有出席联合国大会第一委员会和第六委员会会议的非洲国家代表。¹⁰考虑到如果在制定非洲共同立场的过程中过度依赖外部合作伙伴提供资源，非盟成员国可能对《非盟共同立场文件》缺乏自主权。因此，为了确保参加培训的非洲学员也能听到非洲的声音，特别报

⁷ Bankole Adeoye, Common African Positions on Global Issues, ISS African, December 2020, p. 3.

⁸ 例如非盟 2014 年通过的《关于 2015 年后发展议程的非洲共同立场（The Common African Position on the Post-2015 Development Agenda）》、2016 年通过的《关于联合国大会世界毒品问题特别会议的非洲共同立场（The Common African Position on UN General Assembly Special Session (UNGASS) on the World Drug Problem）》、2017 年《关于安全、有序和正常移民的全球契约的非洲共同立场 The Common African Position on the Global Compact for Safe, Orderly and Regular Migration）》（均在联合国框架内产生了较大影响）。

⁹ Peace and Security Council of the African Union, Communiqué of 1120th meeting, held on 9 November 2022, on the Inaugural Engagement between the Peace and Security Council and the AU Commission on International Law, PSC/PR/COMM.1120.1 (9 November 2022), para. 4.

¹⁰ Mohamed Helal, The Common African Position on the Application of International Law in Cyberspace: Reflections on a Collaborative Lawmaking Process, EJIL: Talk! (5 February 2024), <https://www.ejiltalk.org/the-common-african-position-on-the-application-of-international-law-in-cyberspace-reflections-on-a-collaborative-lawmaking-process/> (Last Visited on 20 March 2025).

告员还邀请了非洲学者或非裔学者¹¹对培训人员所主张的立场进行批判性评论、提问和回应。而在此阶段，非盟也在总结以往“共同立场文件”经验教训的基础上¹²，做好了对内协调的相关事宜。一方面，非盟在系列决议中明确了《非盟共同立场文件》的主导机构为非盟核心机构，以及文件的起草流程与分工问题。具体而言，和平与安全理事会在决议中反复强调《立场文件》是由和平与安全理事会与非盟国际法委员会（AU Commission on International Law）共同主导，邀请特别报告员穆罕默德·赫拉尔（Mohamed Helal）主笔起草，特别报告员向非盟国际法委员会提交文件，而非盟国际法委员会通过草案后才将文件内容转交和平与安全理事会进行审议，最后由后者决定是否颁布以及如何颁布的问题。另一方面，在起草《非盟共同立场文件》的过程中，主导机构向各成员国发放了背景说明与问卷，鼓励各成员国对问卷中的问题进行回应，这一方式不仅可以提升确定《非盟共同立场文件》内容与框架的效率，同时也使得各成员国能够在问卷引导下，进一步回应对网络空间国际规则相关细节，降低参与难度，提升各成员国的参与积极度。

第二阶段是《非盟共同立场文件》的实际完善。特别报告员于2023年5月向非盟国际法委员提交了《非盟共同立场文件》初版草案。在此基础上，和平与安全理事会决定共同从对内对外两个方面推进《非盟共同立场文件》的起草与完善。在对内立场的协调方面，和平与安全理事会决定建立专家工作组，并授权该工作组对《非盟共同立场文件》草案进行审查；¹³与此同时，和平与安全理事会鼓励充分利用非洲区域内的相关资源，包括非洲警察合作组织（AFRIPOL）和非洲情报和安全局委员会（CISSA）¹⁴。而在对外寻求援助方面，非盟不仅寻求了德国和英国政府的财政援助，以支持专家工作组举行会议审查《非盟共同立场文件》草案，还寻求了美洲国家、红十字国际委员会等国际组织以及不同国家地区的专家学者的法律援助，以增强《非盟共同立场文件》的法律性与说服力。

（二）《非盟共同立场文件》的制定特点

在资源整合方面，《非盟共同立场文件》的制定展现出了广泛且多元的特性。为确保《非盟共同立场文件》能够具备充分的可行性与前瞻性，非盟充分利用各

¹¹ 包括瑞士日内瓦大学的马卡内·姆本格（Makane Mbengue）教授、荷兰莱顿大学的马马杜·赫比埃（Mamadou Hébié）副教授、南非约翰内斯堡大学的玛莎·布拉德利（Martha Bradley）教授和国际法院的提拉迪法官（Judge Tladi）。

¹² 由此前“共同立场文件”的实践可见，若对非洲共同立场牵头国家和倡导者的认可度可能不足，将削弱指导小组或技术委员会中牵头国家的积极性；此外，区域委员会和政府间机构参与有限，使得制定非洲共同立场的过程并不完整。See Bankole Adeoye, *Common African Positions on Global Issues*, ISS African, December 2020, pp. 7-9.

¹³ Peace and Security Council of the African Union, Communiqué of its 1171st meeting held on 24 August 2023, on “Updated Briefing on the Development of the Common African Position on Cyber-Security in Africa.”, PSC/PR/COMM.1171 (24 August 2023), para. 5.

¹⁴ 鉴于间谍活动的合法性是关于在网络空间适用国际法的辩论中富有争议的问题，非洲情报和安全局委员会的专家也全程参与了这一进程。

类资源。这其中不仅积极吸纳外国政府的援助，这些援助以资金注入、技术支持以及专业人才交流等多种形式呈现，为《非盟共同立场文件》的制定提供了坚实的物质与智力基础；同时，还大力动员全球非洲人参与其中，借助其在世界掌握的先进技术以及自身具备的专业人才优势，从更广泛的视角为文件制定贡献力量。

从组织运作角度而言，《非盟共同立场文件》的整个制定过程由非盟和平与安全理事会和非盟国际法委员会主导。非盟各机构依据自身专长与职能，分工明确且细致入微。在制定进程中，充分运用非洲联盟机制，通过完善的沟通、协调渠道，保障各机构间紧密协作、高效运转。这种协同模式的核心在于确保所有成员国在《非盟共同立场文件》的制定过程中，既能实现彼此间的互相包容，充分尊重不同国家的文化、政治与经济差异，又能切实维护各国在文件制定环节中的自主权，进而逐步凝聚起强烈的集体感，使得《非盟共同立场文件》能够真正代表非盟各成员国的利益与诉求。

在法律专业性保障上，《非盟共同立场文件》的制定极为重视国际法的深度介入。一方面，积极开展国际法培训活动，全面提升谈判代表与文件起草者的国际法专业素养，使其在工作中能够准确把握国际法律规范与准则；另一方面，邀请国际法领域的资深专家深度参与其中。这些专家凭借丰富的经验与深厚的专业知识，在促进谈判进程顺利推进的同时，对文件中的法律条款进行严格把关与精心雕琢，显著提升了文件在法律表达层面的准确性、严谨性与专业性，为文件在国际舞台上的权威性与有效性奠定了坚实的法律基础。

三、《非盟共同立场文件》的核心观点

《非盟共同立场文件》由序言、结论和十个章节组成，其中包括九个关于国际法权利与义务的主题章节和一个关于能力建设的章节。本文选取了主权原则、不干涉原则、禁止使用武力原则三个核心议题，对《非盟共同立场文件》的相关观点进行分析。选取这些议题进行分析，一方面是因为这些规则作为国际法的一般原则或至关重要的规则，在规范网络空间方面发挥着不可替代的作用；另一方面，作为仍然存在诸多适用争议的议题，《非盟共同立场文件》在这些议题上提出了一些值得深入分析的主张。

（一）主权原则

《非盟共同立场文件》肯定了主权是国家的属性，并将领土主权视为国际法的首要规则，并支持网络主权“纯粹”的路径。¹⁵根据这一观点，只要未经目标国同意或缺乏其他合法理由，对位于其境内的计算机系统进行渗透，就构成对国家主权的侵犯。这一观点与国际判例一致，¹⁶也与法国、瑞士、巴西、中国等越来

¹⁵ See Kevin Jon Heller, In Defense of Pure Sovereignty in Cyberspace, 97 International Law Studies 1444-1449 (2021).

¹⁶ Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), Merits, Judgment. I.C.J. Reports 1986, para. 251.

越多国家的立场，以及美洲国家组织的立场相符。关于主权侵犯的标准，《非盟共同立场文件》明确拒绝“最低门槛”（*de minimis*）标准。这一标准来源于《塔林手册 2.0 版》，¹⁷主要是指领土主权原则仅禁止在他国网络空间产生重大有害影响的网络行动。当前不少国家认同“最低门槛”标准的部分原因在于不希望自身开展低烈度网络行动被认定为侵犯了他国主权，通过提高网络空间的主权被侵犯的认定门槛，从而在法律方面为可能其为非法的低烈度网络行动拓展边界。然而，非盟在《非盟共同立场文件》中不仅明确反对了这一做法，更进一步的是，《非盟共同立场文件》强调基于领土主权，各国对位于其境内的网络空间组成部分拥有专属管辖权，包括立法、裁决和执行权；因此，任何在外国领土上的执法行动，无论是否造成物理或虚拟损害，构成对该国主权的侵犯。

值得注意的是，《非盟共同立场文件》诠释主权的理论基础是领土主权原则，而对于领土主权的过分强调通常会被视为是一种否认主权规则属性的观点。如以色列司法部副部长罗伊·舍恩多夫（Roy Schöndorf）在其 2020 年演讲中指出：“我们需要区分主权和领土主权，前者通常作为一个一般性概念，意味着独立，后者则是一项国际法规则。国家有时会指出需要保护其主权，泛指其政治意愿和自主权，而不一定是指一项法律规则。”¹⁸此外，美国也倾向于通过强调“领土主权”、淡化“主权”的方式来否定网络主权的规则性质。从 2016 年布莱恩·依根（Brian Egan）的公开演讲，¹⁹到 2017 年珍妮弗·奥康纳（Jennifer M. O'Connor）转呈的备忘录，²⁰再到 2018 年美国参联会公布的战略，²¹都隐喻地展现了美国对于主权规则性质与直接适用的否定态度。然而，必须指出的是，《非盟共同立场文件》对领土主权原则的运用则与美国和以色列的意图完全不同。《非盟共同立场文件》对领土主权的强调，一方面来自于其对主权原则的理解与认同是源于对土地上权力之间的对抗与争夺，各国的主权是一种“排他控制权”²²，这在领土上更为直接易见，而网络空间的虚拟性使得网络主权是一项需要结合想象力来界定的法律存在，通过将领土主权延伸适用至网络空间更符合信息通信技术发展较为滞缓的非洲国家对网络主权的理解，也更符合其当前对于网络主权的构想；另一方面，《非盟共同立场文件》通过强调领土主权，意在将保护网络主权提升至保护领土主权

¹⁷ 《塔林手册 2.0 版》国际专家组的大多数专家认为，要构成对领土主权原则的违反，网络行动至少必须干扰属于外国管辖的计算机网络和系统的功能。参见

¹⁸ Roy Schöndorf, *Israel's Perspective on Key Legal and Practical Issues Concerning the Application of International Law to Cyber Operations*, 97 *International Law Studies* 402 (2021).

¹⁹ See Brian J. Egan, *International Law and Stability in Cyberspace*, 35 *Berkeley Journal of International Law* 169, 174 (2016).

²⁰ See Sean Watts, Theodore Richard, *Baseline Territorial Sovereignty & Cyberspace*, 22 *Lewis & Clark Law Review* 771, 827-829 (2018).

²¹ See US Joint Publication 3-12, *Cyberspace Operations* (8 June 2018), pp. 13-14, <https://info.publicintelligence.net/JCS-CyberspaceOperations.pdf> (Last Visited on 20 March 2024).

²² See Peace and Security Council of Africa Union, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace*, PSC/PR/COMM.1196 (2024), para. 13.

一样的高度，²³强调网络主权的不可侵犯性，这不仅符合非盟主权观点的一贯逻辑，同时还是对网络主权的更全面保护。

（二）不干涉原则

在不干涉章节，《非盟共同立场文件》首先指出，禁止干涉他国内部与对外事务是一般国际法原则，同时也是适用于国家间关系的规则。参照国际法院在尼加拉瓜案中的定义，《非盟共同立场文件》解释称，不干涉原则禁止在国家的“保留领域”进行“胁迫”（coercion）。而什么构成“胁迫”是一个存在争议的问题，也是当前各国所争论的焦点之一。

一些国家将其狭义定义为禁止通过网络行动迫使各国采取或不采取国际法允许其自由决定的行动。例如，荷兰认为“胁迫意味着迫使一个国家开展它本来不会自愿采取的行动（无论是作为还是不作为）”²⁴。按照这种理解，不干涉原则保护的是国家的“选择自由”。而英国采取了更为激进的立场，即认为胁迫不仅包括剥夺选择，还意味着“剥夺一国对其依据国家主权原则有权自由决定的事务的控制权”。早在 2018 年杰里米·赖特的演讲中，就忽视了对“胁迫”行为烈度的考察，单纯聚焦对目标国“选择自由”的影响；²⁵而后 2022 年苏拉·布雷弗曼在此基础上进一步扩大解释，从“选择自由”扩大为“控制自由”。²⁶英国试图以这种方式（重新）定义胁迫，是因为其希望扩大不干涉原则的范围，确保该原则能够涵盖更广泛的恶意网络行动，以此弥补其拒绝将领土主权原则视为国际法规则的做法。

《非盟共同立场文件》将“胁迫”定义为“旨在限制外国意志的政策”，并强调不一定非要达到完全剥夺一国选择自由或迫使该国非自愿行事或不作为的程度，同时指出，即使胁迫威胁未得逞，根据国际法也是非法的。值得注意的是，《非盟共同立场文件》这一观点本质上也扩大了“胁迫”的范围，并降低了违反不干涉原则的门槛。因此，《非盟共同立场文件》的这一观点更接近那些主张将胁迫定义为侵害他国“控制自由”的国家的观点，这一定义比“选择自由”更宽泛。《非盟共同立场文件》对于不干涉原则的此种解释本质上是与其对主权原则与不干涉原则在网络空间的严格适用态度密切相关的，正如路易斯·亨金所说，“所有国家都坚持

²³ Russell Buchan & Nicholas Tsagourias, The African Union's Statement on the Application of International Law to Cyberspace: An Assessment of the Principles of Territorial Sovereignty, Non-Intervention, and Non-Use of Force, EJIL Talk!, <https://www.ejiltalk.org/the-african-unions-statement-on-the-application-of-international-law-to-cyberspace-an-assessment-of-the-principles-of-territorial-sovereignty-non-intervention-and-non-use-of-force/>.

²⁴ U.N.G.A., Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States Submitted by Participating Governmental Experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security Established Pursuant to General Assembly Resolution 73/266, A/76/136, 13 July 2021, p. 57.

²⁵ 在该演讲中，杰里米·赖特称：“这一原则（不干涉原则）的目的是确保所有国家在作为国家主权核心的政府事务方面（如选择自己的政治、社会、经济和文化制度的自由）不受外部强迫性干涉。”See Attorney General's Office, The Rt Hon Sir Jeremy Wright KC MP, *Cyber and International Law in the 21st Century*.

²⁶ 苏拉·布雷弗曼在其演讲中指出：“如果对另一国事务的干涉是强行的、专断的或在其他方面是强迫的，剥夺了一国对国家主权原则允许其自由决定事项的控制自由，那么这种干涉就是非法的。”See Attorney General's Office, The Rt Hon Suella Braverman KC MP, *International Law in Future Frontiers*.

不干涉的规范，但它其实是弱国和小国所偏好的规范，该规范有助于表达它们的平等地位，并保障它们的独立和自主。”《非盟共同立场文件》希望通过扩大不干涉原则在网络空间的适用范围，从而达到该原则“保护弱者、限制强者”的目的。

（三）禁止使用武力原则

禁止使用武力或以武力相威胁是一项“强行法规则，也是一般国际法的基本和核心规则”，只存在两个例外，即针对武力攻击进行自卫，以及获得联合国安理会根据《联合国宪章》第七章授权的情况。关于这一原则，当前学界与实务界的讨论聚焦于两个问题：一是什么样的网络行动被认为是国际法下被禁止的“武力”；二是什么样的网络行动构成能够触发自卫权的“武力攻击”。

关于“武力”的认定，《非盟共同立场文件》认同“规模与效果”标准，即“当网络行动的规模与效果与被禁止的传统暴力行为相当时，网络行动将属于禁止使用武力的范畴。特别是，根据其规模与效果，若某一网络行动预计会造成与被禁止的使用武力行为相当的实际损害、人员伤亡，那么该网络行动就构成使用武力。”²⁷而“规模与效果”标准适用的后置争议问题是未造成人员伤亡与物体损毁而造成诸如目标功能丧失后果的网络行动是否也属于“武力”。对此，巴西、丹麦等国持谨慎态度，而美国、以色列则表示明确反对，而哥斯达黎加、新西兰、挪威等国则认为“武力”的概念必须与时俱进，以满足国际社会不断变化的需求，并且即使网络行动不会造成实际损害，也可能产生严重的负面影响，故而应将“武力”的范围拓展至致使关键计算机网络和系统瘫痪的网络行动。《非盟共同立场文件》认为，“破坏、损害或永久性瘫痪一国内关键基础设施或民用物体的网络行动，根据国际法可能被视为构成使用武力”，但值得注意的是，《非盟共同立场文件》虽然支持了哥斯达黎加、新西兰与挪威等国的观点，但将相关标准提高至“永久性瘫痪”，这点与其他国家稍有不同。

关于“武力攻击”的认定，《非盟共同立场文件》采用了传统观点，即只有在发生武装攻击时才能行使自卫权，并且武装攻击是严重的使用武力行为。²⁸针对迫在眉睫的威胁进行自卫这一有争议的概念，《非盟共同立场文件》倾向于限制性解释，即只有“在发生武力攻击时”才允许自卫。²⁹有学者将《非盟共同立场文件》这一观点视为一种“保守”的倒退，³⁰但针对“迫在眉睫”的威胁进行预先性自

²⁷ Peace and Security Council of Africa Union, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace*, PSC/PR/COMM.1196 (2024), para. 39.

²⁸ Peace and Security Council of Africa Union, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace*, PSC/PR/COMM.1196 (2024), para. 41.

²⁹ Peace and Security Council of Africa Union, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace*, PSC/PR/COMM.1196 (2024), para. 42.

³⁰ See Russell Buchan & Nicholas Tsagourias, *The African Union's Statement on the Application of International Law to Cyberspace: An Assessment of the Principles of Territorial Sovereignty, Non-Intervention, and Non-Use of Force*, EJIL Talk!, <https://www.ejiltalk.org/the-african-unions-statement-on-the-application-of-international-law->

卫的观点具有鲜明的西方大国“进攻”色彩。如前所述，在制定《非盟共同立场文件》的过程中，各国外交法律代表接受了相关国际法培训，这一观点的转变应被视为非盟各成员国在了解对国际法实践发展并充分审视自身利益诉求下，所作出的政策选择。对预先性自卫的消极态度，本质上是充分考虑非洲国家缺乏对迫在眉睫威胁的感知与反应的技术能力，更是考虑到非洲各国更有可能成为预先性自卫的对象而非主体的现实。

四、《非盟共同立场文件》的意义与启示

（一）《非盟共同立场文件》的意义

作为区域集体立场的制度化表达，《非盟共同立场文件》堪称国际法表达的“南方叙事”典范，这标志着非洲作为“全球南方”的重要力量，其法律观念的觉醒以及表达水平的显著提升。长期以来，在网络空间国际规则的塑造中，“全球南方”往往处于边缘地位，而《非盟共同立场文件》开创了“全球南方”参与网络空间国际规则塑造的新形式。这种模式充分整合了“全球南方”的数量优势，借助集体的力量发声，打破了以往由少数发达国家主导规则制定的局面，增强了“全球南方”在全球网络空间治理规则制定过程中的影响力。

《非盟共同立场文件》成功凝聚了非洲内部的共识，在深入剖析自身需求与政策偏好的基础上，运用国际法话语体系，为网络空间国际规则的发展提供带有自身特色的规则。《非盟共同立场文件》体现了一种强烈的主权意识，具体而言：在主权方面，《共同文件》坚持主权的规则属性，将对网络基础设施的保护提升到与领土保护一样的高度；在不干涉方面，《共同文件》强调对“强迫”要件的宽泛解释，以尽可能降低不干涉原则的适用门槛，以保护各国主权；在禁止使用武力方面，《共同文件》特别强调了迫在眉睫的威胁并不构成对自卫权的触发，只有出现武力冲突时，自卫权才有了适用基础。

对于网络空间国际法的发展而言，《非盟共同立场文件》也为当前确定和阐明国际法在网络空间适用的进程做出了积极贡献。一方面，《非盟共同立场文件》中对各个原则的解读和应用，为其他国家和地区在探讨网络空间行为规范时提供了重要参考，拓宽了相关讨论的维度。另一方面，《非盟共同立场文件》也向世界展示了非洲国家在网络空间治理中的积极态度和建设性意见，吸引了更多国家关注非洲在网络空间的发展需求和利益诉求。这有利于促进不同国家和地区之间的交流与合作，共同探索更合理、更有效的网络空间国际规则，进而推动网络空间国际治理体系朝着更加公平、公正、包容的方向发展。

（二）对其他“全球南方”的启示

一是在全球治理规则的塑造过程中，“全球南方”不应保持沉默或被动接受，

to-cyberspace-an-assessment-of-the-principles-of-territorial-sovereignty-non-intervention-and-non-use-of-force/.

积极发声是至关重要且十分必要的。《非盟共同立场文件》一经发布，即受到整个国际社会的关注，不少国家在联合国开放式工作组的讨论中直接对《非盟共同立场文件》的制定与发布给予了肯定，同时《非盟共同立场文件》中呈现的诸多国际法观点已经成为学者与政策制定者所讨论与参考的对象。可见，只有主动表达自身的利益诉求、观点和立场，才有可能改变过去在国际规则制定中被边缘化的局面，在全球网络空间治理等领域获得与自身实力和地位相匹配的话语权。“全球南方”应认识到，自身在国际事务中有独特的视角和利益需求，积极参与国际规则的讨论和制定，是维护自身发展权益、推动全球公平正义的重要途径。

二是《非盟共同立场文件》的制定为其他“全球南方”如何充分运用能力建设达成目的提供了良好范例。首先，能力建设需要平衡外部援助与自主发展，对外寻求援助是提升能力的一种方式，但并不意味着要丧失自主性。在《非盟共同立场文件》的制定过程中，即便接受了西方的资金、技术等援助，但非盟仍需采取多种措施，确保自身利益在政策立场中得到充分体现。要对外部援助进行审慎评估和有效管理，明确援助的目标和方向，使其服务于本国的发展战略，避免因过度依赖外部援助而被他人左右决策。其次，在能力建设过程中仍坚持文化自信与创新，不能一味追随并全盘吸收西方的话语体系和行为方式。“全球南方”拥有丰富多样的本土文化，应深入挖掘本土文化中的智慧和价值，将其融入到国际事务的参与方式和规则制定中。从本土文化中汲取养分，不仅能够为提出创新性的观点和解决方案提供灵感，同时也能够促进“全球南方”对自身利益考量的进一步思考。最后，能力建设最需要重视的是人才培养，人才是实现国家发展和提升国际竞争力的关键。³¹“全球南方”必须加强自身的人才建设，培养一批熟悉国际规则、具备专业知识和创新能力的人才队伍。只有拥有了足够的专业人才，才能在国际事务中更好地表达本国立场、参与规则制定，摆脱在国际合作和决策过程中对外部人才的过度依赖。

三是运用法律表达技巧，借助国际平台传播，从而提升发声的影响力。在参与国际事务和制定相关文件时，要充分利用法律表达技巧。通过严谨的法律语言、合理的逻辑结构以及准确的概念界定，提升文件的法律专业性和可读性。一份法律性强且通俗易懂的文件，不仅能够更准确地传达自身的立场和主张，还能增加文件的说服力和影响力。在撰写关于网络空间治理的文件时，运用精确的法律术语和规范的法律条文表述，使文件在国际法律框架下具有更强的权威性和可信度。选择在国际论坛等重要平台发布文件，能够吸引国际社会的广泛关注。在这些平台上发布精心制定的文件，可以让更多的人了解“全球南方”国家的立场和诉求，引发国际社会的讨论和思考，进而推动相关议题的发展和国际规则的完善。积极

³¹ 参见黄志雄：《网络空间规则博弈中的“软实力”——近年来国内外网络空间国际法研究综述》，载《人大法律评论》2017年第3期，第261页。

与其他国家和国际组织进行交流与合作，争取在国际论坛上获得更多的发言机会和支持，也有助于提升文件的影响力和“全球南方”在国际事务中的地位。

2025年中国国际法学会年会

跨境远程访问之法律规制与中国因应

徐铭泽*

摘要：跨境远程访问是指通过外部网络连接从外部位置访问内部资源，且远程访问双方位于不同国家境内，其性质应当被认定为“服务”。国际社会 11 当前针对跨境远程访问的规制路径主要可分为“出口管制模式”以及“合规声明模式”，二者分别以出口管制和企业合规的方式实现其规制功能，并分别侧重于规制服务的出口与进口。针对跨境远程访问的不当规制不仅违反了 WTO 协定，同时构成了对于“长臂管辖”的滥用。此外，不当规制跨境远程访问还限制了中国人工智能产业发展，影响了中国科技企业的竞争力，并阻碍了中国产品海外市场的拓展。中国应当积极实施补贴政策保障关键领域的供应链安全，综合权衡提起 WTO 违法之诉的可能性，合理运用国内法机制以反制不当措施。

关键词：远程访问 GATS 出口管制 合规 算力。

伴随着信息技术的高速发展，远程访问逐渐应用于现代社会的各个领域，并在云服务以及工业物联网等典型应用场景中发挥了关键作用，极大便利了数据、算力以及云服务等数字产品的跨境流动。尤其是在人工智能、互联网等产业发展的过程中，跨境远程访问的重要性不断提升，不仅促进了技术的研发与应用，亦推动了产业协作模式与全球化布局的深刻变革。与跨境远程访问在技术及应用方面的重大突破相对立，在法律层面，各国虽然已在近期不断推进有关立法进程，但针对跨境远程访问的法律规制整体而言仍处于起步阶段，其具体的规制路径、可能的违法性、中国视角下的危害性以及应对措施等问题，尚需进一步研究。

一、跨境远程访问的概念辨析

（一）跨境远程访问的权威释义

近年来，为弥补法律漏洞并合理引导科技发展，各国针对远程访问的立法工作不断加强，并在各类官方文件中就远程访问进行了界定阐释。不仅如此，在各类权威技术标准以及大型公司的内部管理文件中，也相继出现了众多有关远程访问的内容，为确定远程访问乃至跨境远程访问的概念内涵提供了重要参考。

在美国众议员 Mike Lawler 于 2024 年 4 月 29 日提出的《远程访问安全法案》第 2 条中，针对远程访问进行了明确定义，即是指“外国人通过网络连接，包括互联网或云计算服务，从物项实际所在位置以外的地点访问受美国管辖的物

* 作者简介：徐铭泽，北京大学法学院博士研究生，主要研究方向为国际私法、国际经济法。通讯地址：北京市海淀区颐和园路 5 号。邮政编码：100871。邮箱：xumingze30@126.com

项”以及“商务部长颁布的条例中所规定的任何其他获取方式”¹。虽然该法案目前尚未通过，但作为针对远程访问行为的专门性立法，该法案已于 2024 年 5 月 16 日在美国众议院外交事务委员会（Committee on Foreign Affairs）通过，并于 2024 年 9 月 9 日在美国众议院“规则搁置”（Suspension of rules）程序下通过了口头表决，可见其在一定程度上体现了美国国会对于远程访问的官方释义。

在 2016 年 7 月美国国家标准与技术研究院（NIST）出版的《企业远程办公、远程访问和自带设备安全指南》中，以不同的表述重新定义了远程访问，并可将其大致总结为，“通过非组织控制的网络（如公共互联网），从外部位置连接到组织内部网络资源的能力”。作为美国联邦政府的下属机构，美国国家标准与技术研究院隶属于美国商务部，并在美国技术创新和国家安全中发挥着重要作用，其发布的标准与指南具有较高的权威性，在政府层面和私营领域得到了广泛采纳。

除了美国国会以及联邦政府的官方定义外，在国际标准化组织（ISO）和国际电工委员会（IEC）于 2022 年联合发布的《信息安全、网络安全和隐私保护——信息安全控制》（ISO/IEC 27002:2022）这一标准中，针对远程访问同样作出了规定，并将其解释为，“通过外部网络（如互联网）从非组织物理场所访问组织信息系统的过程”。国际标准化组织和国际电工委员会均为国际社会公认的权威标准机构，其标准被全球范围内众多政府、企业和国际组织所认可与接受，具有较高的行业影响力。

综合以上权威释义，其表述虽略有不同，但关于远程访问的核心概念和本质却是一致的，即通过外部网络连接从外部位置访问内部资源。更进一步讲，由于互联网突破了地理界限的阻隔并具有全球性的特征，当远程访问的双方位于不同国家境内时，该远程访问便具有了涉外性特征，属于跨境远程访问的范畴。

（二）跨境远程访问的应用场景

事实上，虽然跨境远程访问的应用愈发普遍，但法律层面的争议主要集中于以下两种应用场景，即远程获取境外先进技术和远程运维境外设备。

第一，云服务领域通过远程访问获取境外先进技术。由于美国在近年来不断升级其出口管制措施，限制向中国出口高端人工智能芯片，加之美国在高端芯片和先进算力领域的全球领先地位。众多被限制的中国实体转而通过亚马逊等美国公司提供的云服务，以获取其无法通过其他方式获得的先进美国芯片和人工智能能力。事实上，由于只有商品、软件或技术的出口或转让受到监管，通过云提供访问此类芯片或先进人工智能模型并不违反美国法规。²当前，美国英伟达等公

¹ 美国《2018 年出口管制改革法》第 1742 条第（10）部分规定，“除非另有规定，‘部长’一词是指商务部长。”作为对于《2018 年出口管制改革法》的修正，《远程访问安全法案》中所规定的“部长”一词在没有另行规定的情况下亦应当指“商务部长”。

² See Eduardo Baptista, Fanny Potkin & Karen Freifeld, *Exclusive: Chinese entities turn to Amazon cloud and its rivals to access high-end US chips, AI*, Reuters, <https://www.reuters.com/technology/chinese-entities-turn-amazon-cloud-its-rivals-access-high-end-us-chips-ai-2024-08-23/>, last visited 2025-2-6.

司在全球先进算力领域占据着的垄断地位，通过云服务等远程访问方式，能够在不违反美国出口管制规定的情况下，为有关中国实体获取先进算力等美国人工智能能力提供替代方案。

第二，工业物联网领域通过远程访问运维境外设备。随着企业的跨境合作和海外业务的拓展，企业设备逐渐分布至全球各地。传统运维方式面临着出差成本高、工作效率低、服务不及时等问题，通过远程访问进行运维极大克服了上述问题，近年来需求快速增长，并逐渐成为企业运维境外设备的重要方式。包括物通博联、³贝锐向日葵⁴等企业皆已推出其设备远程运维系统。2023年3月，国机集团经纬智能纺织机械有限公司利用其自主研发的远程运维系统，成功完成了马来西亚项目2台JWF1580超长智能型细纱机的升级维护。相比于传统的设备运维方式，此种远程运维具有显著优势，不仅提高了设备的运行和生产效率、缩短了设备故障的排除时间、降低了设备的停机损失，还避免了因地域、跨国境而导致的人力、物力以及时间成本的浪费，有效降低了设备的运维成本。⁵在远程访问技术的助力下，为企业“走出去”和全球化发展提供了高效的远程运维解决方案。

（二）跨境远程访问的定性分析

1. WTO 规则下跨境远程访问的性质界定

WTO 针对“货物”与“服务”两个领域的贸易规则作出了不同规定，并分别规定于《关税与贸易总协定》（GATT）以及《服务贸易总协定》（GATS）中。“远程访问”应当被归于“服务”并适用服务贸易的相关规则。

第一，在 GATS 第 1 条 3.(b) 部分，对“服务”这一概念作出了规定，即“包括任何部门的任何服务，但在行使政府职权时提供的服务除外”。此处的“任何服务”当然也包括提供远程访问的服务，更进一步来讲，通过远程访问这一技术手段进行的人工智能模型训练、设备远程运维等也皆属于“任何服务”的范畴。

第二，根据 GATS 第 1 条 2.(a) 规定，“服务贸易”包括“自一成员领土向任何其他成员领土提供服务”，即跨境交付，远程访问即属于此种服务贸易模式。一方面，在远程访问时物项和访问者的实际位置分处于不同国家的情形更是十分常见，符合“跨境”的要求。另一方面，访问者可以通过远程访问接受例如人工智能模型训练、设备远程运维等技术服务，符合“交付”的要求。

2. 学术界观点中跨境远程访问的性质界定

³ 参见厦门物通博联：《设备远程维护运维系统》，中国工控网，<https://www.gongkong.com/article/201612/72206.html>，2025年2月6日最新访问。

⁴ 参见向日葵：《如何运维管理上千门店的智能设备？泡泡玛特这样选择远控方案》，贝锐向日葵，<https://sunlogin.oray.com/news/36661.html>，2025年2月6日最新访问。

⁵ 参见中国恒天：《经纬智能研发的远程运维技术首次在海外项目中成功运用》，中国机械工业集团有限公司，https://www.sinomach.com.cn/xwzx/zgsdt/2023_qydt/202303/t20230329_414518.html?utm_source=chatgpt.com，2025年2月6日最新访问。

针对“货物”与“服务”之间的区别，学术界进行了深入研究，相关研究成果同样能够为判断“远程访问”的性质提供帮助。根据学术界的主流观点，远程访问应当属于“服务”。

有关“货物”与“服务”的一种常见区分方式在于，“货物”是有形的（tangible），而“服务”是无形的（intangible）。⁶若采取此种判断方式，由于“远程访问”具有“通过网络连接”这一特征，结合网络连接本身的特点，能够通过网络连接进行传输的仅能为“无形的”物项，因此，应当将“远程访问”归入“服务”的范畴。

Peter Hill 认为，从特征上来讲，“货物”是独立于其所有者而存在的实体，同时“货物”也具有可转让性，能够从一个经济单位转移到另一个经济单位。而“服务”则不是一个独立于生产者或消费者的实体，正因如此，“服务”本身亦无法像“货物”一样被储存或是转让。⁷Sam Fleuter 认为，结合《韦氏词典》（Merriam-Webster Dictionary）中有关“货物”与“服务”的定义，虽然在一定程度上同 GATS 一样存在循环论证的问题，但同时也说明了“货物”与“服务”二者之间在时间要素上的区别。即“货物”是先制造出来随后出售的，而服务则只存在于“由服务者提供”的期间。⁸

聚焦于“远程访问”而言，由于其本身并不能够脱离访问者和被访问者而独立存在，相较于是一种实体，“远程访问”更加偏向于是一种行为。这种“远程访问”既无法被储存，也无法在不同主体之间进行转让。不仅如此，“远程访问”在生产 and 交付的时间上也具有同一性，仅存在于访问者与被访问者“远程访问”期间。综上所述，“远程访问”更加符合学术界对于“服务”的定义。

二、跨境远程访问的规制路径

根据当前国际社会已有的立法实践，各国对于跨境远程访问的规制主要存在两种路径，即“出口管制模式”和“合规声明模式”，两种路径模式的规则基础和侧重点各不相同，分别适用于跨境远程访问的不同应用场景。

（一）出口管制模式

出口管制模式是指，通过将远程访问纳入本国的出口管制法律体系，并由此限制以跨境远程访问的方式获取本国特定物项。由于此种规制模式以本国的出口管制法为基础，并以出口管制的方式实现其规制功能，故而被称之为“出口管制模式”。

出口管制模式通常在“远程获取境外先进技术”这一跨境远程访问的应用领域发挥作用，侧重于规制本国算力等先进人工智能能力对外出口。避免在高端芯

⁶ 参见罗云开：《货物抑或服务：数字产品属性探究》，《社会科学研究》2023年第6期。

⁷ See Peter Hill, “Tangibles, intangibles and services: a new taxonomy for the classification of output”, *The Canadian journal of economic*, Vol.32, No.2, 1999, pp. 437-443.

⁸ See Sam Fleuter, “The Role of Digital Products under the WTO: A New Framework for GATT and GATS Classification”, *Chicago Journal of International Law*, Vol.17, No.1, 2016, p.163.

片等特定物项被限制出口的情况下，有关外国实体通过远程访问的方式直接使用或获取特定物项。其立法目的与出口管制的立法目的基本相同，是对于本国出口管制措施的进一步增强，以此服务于本国在政治、经济、军事和外交等领域的政策需要。

美国近期不断推进的《远程访问安全法案》是出口管制模式的典型代表。作为对现行《2018 年出口管制改革法案》的修订，《远程访问安全法案》授权美国商务部对相应物项的“远程访问”进行出口管制，并将“远程访问”与出口、再出口、境内转让并列使其成为一种全新的出口管制类型。该法案一方面授予了美国有关部门在出口管制方面以更大的权力，另一方面也在一定程度上对远程访问进行了规制，当用于获取特定物项时，远程访问也将受到相应限制。

（二）合规声明模式

合规声明模式是指，通过对特定行业提出合规要求，并在合规要求中加入限制跨境远程访问的规定，由此实现对远程访问的规制。由于此种规制模式以合规制度为基础，其规制功能的实现亦有赖于企业的自身合规，因此可被成为“合规声明模式”。

合规声明模式通常在“远程运维境外设备”这一跨境远程访问的应用领域发挥作用，侧重于规制外国远程操控国内设备，在一定程度上亦阻碍了设备远程运维服务的进口。避免关系到国计民生的重要设备，被外国实体以跨境远程访问的方式加以操纵和控制。其立法目的主要是为了维护本国在电力能源等关键领域的安全性和独立性，避免因使用外国设备而受到干扰，降低此类关键领域中系统的运行风险与威胁，进而服务于本国的国家安全战略。

立陶宛议会于 2024 年 11 月 12 日通过的《电力法》修正案便采取了合规声明模式。该修正案新增了第 733 条，即电力设备控制系统的安全性。在该条第 1 款中，针对装机容量超过 100 kW 的发电及储能设备的信息管理系统和生产控制系统，提出了安全要求，防止其被立陶宛国会认定的威胁国家安全的主体远程操控功率参数或启停设备。该条第 2 款则要求电网运营商制定相应的实施细则及合规声明格式，并在《电网使用规则》中明确技术规范。根据该条第 3 款的规定，设备所有者或其代表需向电网运营商提交合规声明，确认其满足该条第 1 款的安全要求。若检查发现不达标，运营商应通知国家能源监管委员会及用户，而国家能源监管委员会则可据此暂停相关许可证的效力。除此之外，若已发放许可证的设备被发现不符合安全要求，根据该条第 4 款的规定，国家能源监管委员会有权责令整改，并可要求运营商提供技术评估。由此可见，立陶宛的《电力法》修正案在合规的安全性要求中加入了限制远程访问的内容，并要求有关主体制定相应的配套制度，综合运用提交合规声明、合规检查、责令整改以及技术评估等方式，多措并举，共同助力于安全性要求的落实。这同样在一定程度上限制了对于本国

特定设备的远程访问，实现了对于远程访问的规制。

三、跨境远程访问不当规制的违法性分析

针对跨境远程访问的不当规制不仅阻碍了国际间的正常经贸往来，严重危害人类科技的发展进步，还有可能引发法律层面的质疑和挑战，存在相应的违法性。无论是吸收了出口管制模式的美国《远程访问安全法案》抑或是采纳了合规声明模式的立陶宛《电力法》修正案，二者均违反了 WTO 协定。除此之外，上述法案对于跨境远程访问的不当规制还构成了对“长臂管辖”的滥用，缺乏相应的管辖权基础。

（一）不当规制跨境远程访问违反WTO协定

综观 WTO 中 GATS 的规则条款以及 DSB 作出裁定的有关案例，不当规制跨境远程访问涉嫌违反 GATS 的序言、GATS 第 2 条的最惠国待遇、GATS 第 6 条的国内法规以及 GATS 第 16 条的市场准入。除此之外，不当规制跨境远程访问亦不符合 GATS 第 14 条之二的例外条款。

1. 违反 GATS 最惠国待遇原则

美国《远程访问安全法案》以及立陶宛《电力法》修正案均仅限制向中国等少数国家的部分企业提供远程访问服务，这一做法违反 GATS 第 2 条的最惠国待遇义务。一方面，上述法案限制他国远程访问服务的措施应受 GATS 规制，另一方面，上述法案违反了其给予“不低于其给予任何其他国家相同的服务和服务提供者的待遇”的义务。⁹

第一，限制他国远程访问措施应受到 GATS 规制。首先，限制“远程访问”属于影响服务贸易的措施。美国在《远程访问安全法案》中意图限制中国等 WTO 成员国以远程访问的方式使用美国的特定物项，如通过远程访问使用美国的算力等。由于“算力”本身可被归入“服务”的范畴，美国法案针对“远程访问”加以限制本质上是对“算力服务”的出口管制，在一定程度上将对国际间的服务贸易产生影响。立陶宛《电力法》修正案则直接限制了中国企业通过远程访问的方式远程运维其境外电力设备，而设备远程运维同样属于服务的一种，立陶宛的这一法案影响了中国企业对外服务贸易的正常开展。其次，最受惠国待遇原则规制的“任何措施”包括服务的进口措施和出口措施。WTO 服务贸易规则 GATS 中的最惠国待遇原则不同于货物贸易规则 GATT 中的最惠国待遇原则，二者在相关法律条文中的措辞具有明显差异。具体而言，根据 GATT 第 1 条的规定，“与进出口有关的所有规则和手续”都属于最惠国待遇涵盖的范围，而在 GATS 第 2 条中，最惠国待遇的范围则为“本协议所涵盖的任何措施”，并没有对措施的种类进行具体限定。因此，在没有明确排除的情况下，此处的“任何措施”应当同时包括服务的进

⁹ See Appellate Body Report, Canada – Autos, paras. 170-171. See also Panel Report, EU Energy Package, paras. 7.227 and 7.404.

口措施以及出口措施，美国《远程访问安全法案》中限制算力等服务出口的措施同样受最惠国待遇原则的规制。

第二，限制他国远程访问的措施违反了应给予“不低于其给予任何其他国家相同的和服务和服务提供者的待遇”的义务。GATS 第 2 条第 1 款的规定表明，GATS 最惠国待遇要求其将给予第三国待遇不低于给予任何其他成员方。在“欧共体香蕉案”（EC-Banana）中，专家组指出“不低于待遇”的解释包括法律上和事实上的不歧视。而确定“不低于待遇”的标准是“竞争条件相同标准”，即不提供较差的竞争条件。¹⁰而在“阿根廷金融服务案”（Argentina-Financial Service）中，上诉机构进一步指出，一项措施若改变了竞争条件，对任何其他成员的服务或服务提供者造成了不利影响，本身就足以构成对最惠国待遇违反的条件，而无需考虑措施背后的监管目标或关切。¹¹

美国《远程访问安全法案》和立陶宛《电力法》修正案既构成法律上的歧视，也构成事实上的歧视。就法律上的歧视而言，根据《远程访问安全法案》，被管制的对象可以参考美国《出口管理条例》第 744 条的实体清单。因此，美国只对来自中国、俄罗斯、伊朗、朝鲜等少数被列入实体清单的国家的的服务和服务提供商实施出口管制，为其创造了不利于竞争的条件，显然构成法律上的歧视。而立陶宛《电力法》修正案同样仅限制被立陶宛国会认定的威胁国家安全的部分主体，在法律上构成对少数国家的歧视。就事实上的歧视而言，虽然美国《远程访问安全法案》以及立陶宛《电力法》修正案并未明确指明其针对的对象，但是该法案通过的背景、立法者的表态以及新闻媒体的报道分析已充分说明，该法案在实践中必然会对中国相关实体造成实质上的歧视。例如，提出《远程访问安全法案》的美国众议员 Mike Lawler 表示，该法案将弥补美国出口管制法中被中国利用的漏洞，避免中国公司利用远程访问的方式使用美国的先进芯片并将其用于人工智能的开发和中国军事力量现代化的实现，不仅如此，该众议员还将中国称作是美国的“头号地缘政治敌人”。¹²而在立陶宛《电力法》修正案通过后，包括欧洲太阳能制造委员会（European Solar Manufacturing Council）官网在内的几乎所有新闻报导皆指出，该法案是对中国公司远程访问立陶宛太阳能、风能和储能设备的禁止。¹³上述法案皆体现了针对中国等少数国家在事实上的歧视，

¹⁰ See Appellate Body Report, EC - Bananas III, paras. 231-234.

¹¹ See Appellate Body Report, Argentina Financial Services, para. 6.109.

¹² 提出该项《远程访问安全法案》的美国众议员 Mike Lawler 在个人网站上阐述了相应的立法理由。See Mike Lawler, *Representatives Mike Lawler, Jeff Jackson, Rich McCormick, and Jasmine Crockett introduce Legislation to Prevent CCP from Acquiring American Tech*, Congressman Mike Lawler, <https://lawler.house.gov/news/documentsingle.aspx?DocumentID=1648>, last visited 2025-2-6.

¹³ See Žygimantas Vaičiūnas, *Lithuanian Parliament bans remote access of companies from China to Lithuanian solar, wind and storage devices*, The European Solar Manufacturing Council, <https://esmc.solar/lithuanian-parliament-bans-remote-access-of-companies-from-china-to-lithuanian-solar-wind-and-storage-devices/>, last visited 2025-2-6; See Energynomics, *Lithuanian Parliament bans remote access of companies from China to solar, wind and storage devices*, Energynomics, <https://www.energynomics.ro/en/lithuanian-parliament-bans-remote-access-of-companies-from-china-to-solar-wind-and-storage-devices/>, last visited 2025-2-6; See Patrick Jowett, *Lithuania bans remote Chinese access to solar, wind, storage devices*, pv magazine, <https://www.pv->

综上所述，对于被上述法案限制的部分企业而言，在获取先进算力和高性能芯片等美国垄断资源方面，以及远程运维境外设备的过程中，并没有授予这些被限制企业与其他大多数未被限制企业一样的自由度。尤其是针对《远程访问安全法案》而言，被限制的人工智能企业作为其自身客户的服务提供者，法案中所采取的相应措施即使是关于美国本国服务的出口管制措施，由于美国在先进算力等领域的垄断地位以及算力对人工智能发展的关键作用，在缺乏算力支持的情况下，也将影响这些作为服务提供者的人工智能企业正常提供服务及其服务贸易的开展。因此，对于跨境远程访问的规制属于 GATS 第 2 条最惠国待遇条款的规制范围，而不当规制跨境远程访问则违反了最惠国待遇原则。

2. 违反 GATS 国内规制条款

GATS 第 6 条为成员在服务贸易领域的国内规制设定了程序和实体的双重要求，体现 GATS 在成员主权和自由贸易之间的平衡。针对远程访问的不当规制违反了 GATS 第 6 条第 1 款的程序义务、GATS 第 6 条第 4 款和第 5 款的实体义务。

在程序上，GATS 第 6 条第 1 款对 WTO 成员的实施国内法规提出了“以合理、客观、公正的方式予以实施”的要求。根据“美国赌博案（US-Gambling）”及“欧共体能源案（EC-Energy Package）”，“合理、客观、公正”仅限于成员对国内法规的实施方式，而不涉及国内法规的实质内容。¹⁴具体而言，“合理”指的是符合普遍接受的理性标准和健全判断，不能无视明显的事实和标准做法。¹⁵“客观”要求排除不理性、情绪化或个人偏见。“公正”则强调不应给予特定当事方或经营者群体特殊考虑，应考虑对其他经营者产生的影响，对成员作出不平等待遇应给予正当理由。¹⁶

以美国《远程访问安全法案》为例，该法案虽然尚未通过，但是从美国在出口管制方面迄今的许可实践，尤其是美国对中国半导体出口管制的许可实践上，中国实体在美国出口管制实践中难以获得合理、客观、公正的待遇。尤其是在美国出口管制实体清单中，受控中国实体数量远远超过了列表中其他大多数正常国家，具有较强的中国针对性。而美国以国家安全为由，将如华为公司等大量的中国民用企业，甚至包括北京航空航天大学、北京邮电大学等在内的部分中国科研院校列入实体清单，实施出口管制措施，无视其并不威胁国家安全的明显事实，显然难言合理、客观。而对于立陶宛《电力法》修正案而言，该法案针对中国能源设备的制造商施加远程访问限制，并阻碍其正常运维境外设备。立陶宛以维护国家安全的名义限制光伏、风电等与国家安全关联性较低的新能源企业，且仅仅

magazine.com/2024/11/18/lithuania-bans-remote-chinese-access-to-solar-wind-storage-devices/, last visited 2025-2-6.

¹⁴ See Panel Report on United States – Gambling, para. 6.432

¹⁵ See Panel Report on Dominican Republic – Import and Sale of Cigarettes, para. 7.385

¹⁶ See Argentina- Hides and Leather.

针对中国等少数国家的企业，该项法案的实施并不具有合理性、客观性和公正性。

在实体上，虽然 GATS 第 6 条第 4 款是一个授权谈判条款，但第 6 条第 5 款已要求成员在已作出承诺的范围内，遵守第 6 条第 4 款的要求。第 6 条第 4 款则规定成员国设定的资格要求和程序、技术标准和许可要求的措施“不致构成不必要的服务贸易壁垒”。国内规制的纪律应旨在确保国内规制：“基于客观和透明的标准，诸如提供服务的资格和能力；除为保证服务质量所必需以外，不应成为负担；如是许可程序，则其本身不应成为提供服务的限制。”由此可见，GATS 第 6 条既要求国内规制的“必要性”，又对其标准、目的等要素进行了限制。

美国《远程访问安全法案》以及立陶宛《电力法》修正案虽以维护国际安全为目的，具有形式上合理的政策目标，但是其并没有明确决定受控对象的机制和标准，也没有提供受控对象申诉的程序和方法，有违客观透明的要求。其次，美国 and 立陶宛对于“远程访问”限制也并非基于通过远程访问所提供的服务能力本身，而是基于服务消费者或服务提供者的国籍，由此可见其并非是为了保证服务的质量。

综上所述，作为显著影响服务贸易的国内立法措施以及对于跨境远程访问的不当限制，美国《远程访问安全法案》和立陶宛《电力法》修正案违反了 GATS 第 6 条的规定。

3. 违反 GATS 市场准入条款

GATS 第 16 条规定了服务贸易市场准入问题。“美国赌博案”等多个案件指出，违反第 16 条要求：“作出限制的国家已在减让表中作出承诺；有证据显示该限制构成第 16 条第 2 款的六项措施之一。”¹⁷据此，美国《远程访问安全法案》违反 GATS 第 16 条下的义务。

第一，美国 and 立陶宛对计算机服务作出完全市场准入承诺。根据“美国赌博案”，服务贸易减让表中的“None”应理解为该成员国作出完全市场准入的承诺，即该国不得对该类型服务设置任何市场准入条件。¹⁸在美国 and 立陶宛的服务贸易减让表中，计算机服务等服务类型的市场准入一栏中均记为“None”，¹⁹因此美国 and 立陶宛在该类型不能作出限制市场准入的承诺。

其二，美国《远程访问安全法案》和立陶宛《电力法》修正案中的限制远程访问措施违反其完全市场准入的承诺。根据 GATS 第 16 条第 1 款，若成员已允许外国服务提供者在其本国或地区设立商业存在并提供服务，且其具体承诺减让表中未作出任何限制，该成员不得任意限制此种服务的出口，否则可能违反其市

¹⁷ See Appellate Body Report, US Gambling, para. 143. 8 Panel Reports, China Publications and Audiovisual Products, para. 7.1354; China - Electronic Payment Services, paras. 7.511-7.512, EU - Energy Package, para. 7.233 and Argentina - Financial Services, paras. 7.391-7.392.

¹⁸ See Panel Report, US Gambling, paras. 6.269-6.279.

¹⁹ See U.S. Schedule of Commitments under the General Agreement on Trade in Services (Investigation No. 332-354), p.37.

场准入的承诺。而限制“远程反问”的措施构成 GATS 第 16 条第 2 款第 2 项，即“以数量配额或经济需求测试要求的形式限制服务交易或资产总值”。例如，在《远程访问安全法案》第 2 条(4)(A)部分，对《2018 年出口管制改革法案》1754 条(a)部分进行了修订。该条款在修订后，商务部长不仅可以禁止未经授权远程访问受管制物项，同时，商务部长还可以酌情要求对受控物品的远程访问颁发许可证或其他授权书，构成对该项服务的“配额”甚至“零配额”，违反其完全市场准入的承诺。而对于立陶宛《电力法》修正案而言，该法案通过限制远程访问的方式变相禁止了中国企业向立陶宛提供设备远程运维服务，同样违反了 GATS 第 16 条第 2 款第 2 项的要求，一定程度上构成了对该项服务的“配额”甚至“零配额”。

综上，美国《远程访问安全法案》和立陶宛《电力法》修正案在其减让表未有列明的情况下，对其他 WTO 成员的服务和服务提供者施加了限制服务交易的措施，违背了美国和立陶宛所作出的服务贸易市场准入承诺，违反了 GATS 第 16 条的规定。

4. 不符合 GATS 安全例外条款

GATS 第 14 条之二是 WTO 服务贸易领域的安全例外条款，规定了 WTO 成员以维护国家安全为由免除服务贸易协定中所规定义务的特殊制度安排。然而，对于美国《远程访问安全法案》和立陶宛《电力法》修正案而言，相关情况并不符合 GATS 第 14 条之二所规定的安全例外的适用范围，不能因此免除美国在 WTO 服务贸易规则项下包括最惠国待遇、国内法规以及市场准入等方面的相应义务。

美国《远程访问安全法案》和立陶宛《电力法》修正案不符合 GATS 第 14 条之二安全例外条款的理由主要可以从三个方面予以阐明：

第一是文本分析层面。根据 GATS 第 14 条之二的规定，美国《远程访问安全法案》中的管制对象为广大中国民用企业和科研院所的远程访问行为，该行为既不与为军事机关提供给养的服务有关，又不与裂变和聚变物质或衍生此类物质的物质有关，且不属于美国为履行在《联合国宪章》项下维护国际和平与安全的义务而采取的行动，中美两国之间亦非处于战时或国际关系中的其他紧急情况。因此，《远程访问安全法案》不符合 GATS 第 14 条之二所规定的任何一种安全例外情形。而立陶宛《电力法》修正案则是将禁止远程访问的对象限定为中国的光伏、风电以及储能设备，此类新能源产业与国家安全亦不存在直接联系，同样不符合 GATS 第 14 条之二所规定的任何一种安全例外情形。

第二是历史分析层面。安全例外条款不仅被规定于 WTO 服务贸易协定中，在货物贸易协定乃至与贸易有关的知识产权协定中也皆有所体现。其中，WTO 货物贸易协定 GATT 脱胎于国际贸易组织（ITO）宪章，而安全例外条款则被规定于 GATT 第 21 条，该条款实际上来源于 ITO 筹备过程中美国的提案。根据谈

判的历史，该提案一经提出便引起了其他谈判方的担忧，认为安全例外如被滥用并服务于经济目的将成为多边贸易体系的巨大漏洞，而美国代表团的回应则认为安全例外应当关注真正的安全利益并尽可能地限制例外，并将真正的安全利益区别于为保护产业而采取的行动。美国虽然以国家安全的名义不断推进《远程访问安全法案》的立法进程，但根据提出该法案的众议员 Rich McCormick 表示，随着全球范围内人工智能技术的迅速发展及其在诸多行业进步过程中至关重要的作用，提出该法案的目的是为了确保美国在人工智能发展中的领导地位。²⁰由此可见，《远程访问安全法案》的提出显然是为了保护美国产业而非维护美国真正的安全利益，不符合美国自身在安全例外条款提出时的历史提案和相应表述。此外，虽然立陶宛《电力法》修正案并没有直接表明产业保护的意图，但根据 2024 年 5 月 27 日通过的欧盟《净零工业法案》并着力降低净零领域对其他经济体的依赖程度，²¹立陶宛《电力法》修正案确实一定程度上助推了上述立法目标的实现。因此，立陶宛《电力法》修正案同样存在产业保护之嫌，是否为了维护国家真正的安全利益尚有疑问。

第三是案例分析层面。2019 年 4 月 5 日，WTO 散发了乌克兰诉俄罗斯运输限制措施案（DS512）的专家组报告，该报告也是 WTO 争端解决机制就《关税与贸易总协定》中第二十一条安全例外条款的适用问题作出的首个裁决。本案专家组报告的裁定，根据国际公法的习惯解释规则，“基本安全利益”一词应由专家小组而不是各成员来解释。由于俄乌之间的紧张关系，专家组认为其构成了“国际关系中的紧急情况”。²²

除此之外，在 WTO 沙特阿拉伯知识产权措施案（DS567）的专家组报告中同样表示，“国际关系中是否存在战争或其他紧急情况”等国家安全例外情况应当由专家组“客观地”决定，而非由各个国家完全自行判断。更重要的是。专家组报告认为，一名 WTO 成员断绝与另一名 WTO 成员的“所有外交和经济关系”属于“国际关系中存在紧急状态的最终国家表达”。²³

在 WTO 美国钢铝产品案（DS544），专家组认为，“国际关系中的紧急情况”系指具有一定严重性的局势，以及在对国际关系的进行产生影响方面具有关键性

²⁰ 提出该项《远程访问安全法案》的美国众议员 Mike Lawler 在个人网站上阐述了相应的立法理由。See Mike Lawler, *Representatives Mike Lawler, Jeff Jackson, Rich McCormick, and Jasmine Crockett introduce Legislation to Prevent CCP from Acquiring American Tech*, Congressman Mike Lawler, <https://lawler.house.gov/news/documentsingle.aspx?DocumentID=1648>, last visited 2025-2-6.

²¹ 参见生态环境部环境与经济政策研究中心：《独家解读！欧盟〈净零工业法案〉对我国带来多大影响？如何应对？》，石家庄市生态环境局，<https://sthjj.sjz.gov.cn/cms/jsp/site001/article.jsp?fchannelidentity=8afaa16142796386014279efe11b0937&articleId=8afaa1618fe7d997019096c089220ca1&a1b2dd=7xaac>，2025 年 2 月 6 日最新访问。

²² See Panel Report on Russia- Measures Concerning Traffic in Transit, para 7.130.

²³ See Panel Report, Saudi Arabia — Measures Concerning the Protection of Intellectual Property Rights, para 7.245.

或严重性的国际紧张局势，其影响程度至少应与“战争”相当。虽然国际社会对产能过剩问题感到关切，但因其对国际关系的影响并不具有足够的严重性，从而不足以构成“紧急情况”。²⁴

在 WTO 美国产地来源标记案 (DS597)，专家组认为，虽然美国和其他成员高度关切中国香港的人权状况，但该状况并未升级到构成国际关系中的紧急情况的程度。本案情况与此前与安全例外条款相关的案件中的情形（例如战争等）不具有相似性，因此不足以构成“紧急情况”。

由此可见，是否存在国际关系中的紧急状况应当由专家组“客观地”决定，并由各个国家完全自行判断。中国与美国或立陶宛之间不处于“战时或国际关系中其他紧急情况期间”，双方之间并没有断绝“所有外交和经济关系”，亦不存在与“战争”相当的国际紧张局势，并不属于 WTO 专家组所认定的国际关系中的紧急状况。

综上，上述法案对于跨境远程访问的限制并不属于 WTO 安全例外条款的适用范围，并不能因此正当化其限制“远程访问”的举措或免除相应的协定义务。

（二）不当规制跨境远程访问滥用“长臂管辖”

美国《远程访问安全法案》和立陶宛《电力法》修正案意图限制发生或部分发在本国境外的“远程访问”行为，缺乏相应的立法管辖权基础，属于滥用“长臂管辖”的贸易霸凌主义行为。

具体以美国《远程访问安全法案》为例，该法案的可能适用场景是：新加坡公司将 100 台配备英伟达 H100 芯片的服务器装入位于新加坡的数据中心，但服务器实际是为中国公司处理 AI 算法；美国商务部以新加坡公司向中国公司提供“远程访问”服务为由对其进行处罚，要求停止。

根据美国《对外关系法第四次重述》的规定，管辖权基础共包括属地管辖、效果原则、积极属人管辖、消极属人管辖、保护性管辖以及普遍管辖与六种，上述适用场景并不符合其中的任何一种。

第一，远程访问全程皆位于美国境外，不符合属地管辖。远程访问者和远程访问服务提供者分别位于中国和新加坡，英伟达芯片以及开展远程访问的服务器位于新加坡，甚至是远程访问行为本身也是发生在中国和新加坡之间，各方面都未与美国领土产生任何联系。第二，远程访问行为帮助了中国公司处理其 AI 算法，影响主要集中于中国市场，并未对美国国内市场产生直接、重大与可预见的效果，不符合效果原则。第三，作为一家新加坡公司，被美国法案约束的远程访问服务提供者并非美国人，不符合积极属人管辖。第四，远程访问发生于中国公司和新加坡公司之间，并未对美国企业产生直接损害，不符合消极属人管辖。第

²⁴ See Panel Report, United States — Certain Measures on Steel and Aluminium Products, para 7.139.

五，中国公司为开展合法业务与新加坡公司进行远程访问，并未犯有侵害美国国家及其公民重大利益的罪行，不符合保护性管辖。第六，中国和新加坡公司之间进行的远程访问是为了处理 AI 算法，本质上促进了科技的进步与发展，远程访问本身及其目的皆从未出现普遍地危害国际和平与安全以及全人类共同利益的特定国际犯罪行为，不符合普遍管辖。

四、跨境远程访问不当规制的中国危害与影响

针对跨境远程访问的不当规制将对中国产生一系列严重危害与不利影响。一方面，美国《远程访问安全法案》限制了中国人工智能企业远程获取境外先进技术，限制了中国人工智能产业的发展并影响了中国科技企业的竞争力。另一方面，立陶宛《电力法》修正案限制了中国能源设备制造企业远程运维境外设备，阻碍了中国产品海外市场的拓展。

（一）限制中国人工智能产业发展

鉴于算力在人工智能产业发展中的重要推动作用，美国《远程访问安全法案》的提出标志着美国欲加大对中国高性能芯片领域的出口管制力度，并意图对中国企业通过远程访问方式使用美国先进算力进行封锁，以此限制中国的人工智能等高科技产业。

一方面，算力对于人工智能产业的发展至关重要，同时也是人工智能发展的技术保障和动力引擎。²⁵人工智能技术之所以在近年来实现突破，主要是由三大要素合力促成的，即丰富的数据资源、深度学习算法以及充足的计算力支持。其中，充足的算力得益于摩尔定律的不断演进发展，高性能芯片大幅降低了深度学习算法所需的计算时间和成本。²⁶

另一方面，美国在算力领域不断对中国施加出口管制措施，并以此遏制中国人工智能产业的发展。在美国《远程访问安全法案》通过前，美国《2018 年出口管制改革法案》已经就高性能芯片对中国的出口、再出口和境内转让进行了限制，但中国企业仍可以通过远程访问的方式获取美国的先进算力。²⁷美国《远程访问安全法案》将远程访问列入出口管制类型，进一步封堵了中国人工智能产业的算力获取渠道，²⁸阻碍中国人工智能产业的发展。

²⁵ 参见吴洁：《算力为何如此重要》，中国共产党新闻网，<http://theory.people.com.cn/n1/2023/0903/c40531-40069537.html>，2025 年 2 月 6 日最新访问。

²⁶ 参见尹首一、郭珩、魏少军：《人工智能芯片发展的现状及趋势》，《科技导报》2018 年第 17 期，第 45-51 页。

²⁷ 根据英国路透社 2024 年 8 月 23 日报道，中国公司长期通过云服务获取美国先进算力，并指出美国四月有一法案试图填补这一漏洞。See Eduardo Baptista, Fanny Potkin & Karen Freifeld, *Exclusive: Chinese entities turn to Amazon cloud and its rivals to access high-end US chips, AI*, Reuters, <https://www.reuters.com/technology/chinese-entities-turn-amazon-cloud-its-rivals-access-high-end-us-chips-ai-2024-08-23/>, last visited 2025-2-6.

²⁸ 提出该项《远程访问安全法案》的美国众议员 Mike Lawler 在个人网站上阐述了相应的立法理由。See Mike Lawler, *Representatives Mike Lawler, Jeff Jackson, Rich McCormick, and Jasmine Crockett introduce Legislation to Prevent CCP from Acquiring American Tech*, Congressman Mike Lawler,

（二）影响中国科技企业竞争力

鉴于中国对美国算力的高度依赖性，美国《远程访问安全法案》限制中国企业获取美国算力，将使中国企业在全球竞争中处于不利地位。近年来，美国英伟达公司主要主导了我国人工智能算力芯片的市场格局，并占据了 80% 以上的市场份额，且国产芯片在性能、生态等方面仍难以替代美国芯片，²⁹限制中国企业获取美国算力将使中国企业面临严重的算力短缺。对于人工智能企业而言，算力短缺将会削弱中国人工智能企业的产品竞争力，对人工智能算法的复杂度与性能提升、模型的训练速度以及应用部署的规模等产生不利影响。对于其他企业而言，算力获取不足也将阻碍其竞争力的提升。以消费互联网领域为例，算力对于电商平台的智能推荐、物流调度等应用发挥着支撑作用，并极大提升了用户体验和运营效率。³⁰在失去了充足算力的加持后，中国企业在其他领域的产品竞争力和影响力都将受到一定冲击。

（三）阻碍中国产品拓展海外市场

相比于设备的传统运维方式，远程运维具有成本低、效率高、服务及时等方面的显著优势，对于企业“走出去”以及产品海外市场的拓展发挥着积极的推动作用。立陶宛《电力法》修正案不仅限制了远程访问，同时影响了中国企业通过远程访问的方式运维位于立陶宛的境外设备，极大阻碍了中国相关能源设备拓展其海外市场。一方面，对于中国的能源设备制造企业来说，立陶宛《电力法》修正案对于远程访问的限制，使得中国企业仅能继续采取传统的方式进行运维，被迫承担因跨越地域与国境而造成的大量人力、物力与时间损失，大幅提高了中国产品在拓展海外市场时的运维成本。另一方面，对于能源设备的使用者来说，立陶宛《电力法》修正案对于中国产品进行了针对性限制，使得中国产品无法同其他国家产品一样进行更加高效的远程运维，进而造成中国设备的运行和生产效率降低、设备故障排除时间延长以及设备的停机损失增加等消极后果，严重削弱了中国能源设备相比于他国设备的产品竞争力。

五、跨境远程访问不当规制的中国因应

不当规制跨境远程访问不仅违法，同时也对中国产生了一定程度的消极影响，有必要实施更加积极的措施维护中国的国家利益。一方面，由于美国《远程访问安全法案》不当限制跨境远程访问所可能导致的先进算力获取不足，中国应当大力发展本国的芯片产业，并着力提高国产高端芯片与先进算力的供应能力，逐步

<https://lawler.house.gov/news/documentsingle.aspx?DocumentID=1648>, last visited 2025-2-6.

²⁹ 参见《应对算力短缺，国产芯片如何变挑战为机遇？》，人民网，<https://baijiahao.baidu.com/s?id=1775253810890108342&wfr=spider&for=pc>，2025 年 2 月 6 日最新访问。

³⁰ 参见米加宁、李大宇、董昌其：《算力驱动的新质生产力：本质特征、基础逻辑与国家治理现代化》，《公共管理学报》2024 年第 2 期，第 3 页。

降低人工智能算力领域对美国的依赖程度，并在法律层面积极应对美国的不当限制措施。另一方面，针对立陶宛《电力法》修正案不当规制跨境远程访问而对中国企业产生的消极影响，中国同样应当从国内法和国际法两种角度予以同时发力，通过法律手段维护己方权益。

（一）积极实施补贴政策，保障供应链安全

针对美国《远程访问安全法案》不当限制跨境远程访问，并以此阻碍中国使用美国的先进算力，中国应当实施积极的补贴政策大力促进本国芯片产业发展，逐步减少对于美国算力的依赖，保障本国的算力供应链安全。

当前，加大芯片领域的补贴力度已逐渐成为当前世界各主要经济体的共识。以美国《2022 年芯片与科学法》（CHIPS and Science Act of 2022）为例，美国政府将投入 527 亿美元的资金补贴芯片行业发展、为半导体行业投资提供 25% 的税收减免额度并授权多部门追加超过 2000 亿美元的科技研发资金。³¹

对于中国来说，应当正确认识芯片产业补贴政策存在的利端和弊端。一方面，补贴为芯片产业的创新发展发挥了重要的激励作用，有利于推动本国芯片产业的竞争力提升与技术突破。另一方面，在芯片产业补贴政策的制定和实施过程中，可能面临违反 WTO《补贴与反补贴措施协定》，触发 WTO 的反补贴机制，影响补贴政策的实效。现有 WTO《补贴与反补贴措施协定》将补贴被分为了禁止性补贴、可诉补贴与不可诉补贴三类。随着不可诉补贴条款的失效，除禁止性补贴外，其他补贴措施如果具有专向性都可能构成可诉补贴。针对芯片领域的补贴政策有可能违反产业专向性的要求，并承担被有关国家诉诸 DSB、征收反补贴税等法律风险。³²

（二）综合权衡，考虑提起WTO违法之诉的可行性

通过上文的分析，美国《远程访问安全法案》和立陶宛《电力法》修正案对于远程访问的不当规制违反了 WTO 服务贸易协定 GATS 中的最惠国待遇原则、市场准入规则、国内规制规则，且不符合 WTO 的安全例外条款。由于美国和立陶宛未能履行 WTO 协定，根据 GATS 第 23 条的规定，中国可以援用 WTO《关于争端解决规则与程序的谅解》（DSU）中第 3 条 8 款，向美国和立陶宛提起“违法之诉”。

针对中国来说，就美国的违法行为提起 WTO 之诉是一把利弊兼存的“双刃剑”。一方面，提起 WTO 违法之诉将有力维护中国在 WTO 协定中所享有的合法权益，提高美国和立陶宛实施违法措施的成本。另一方面，在诉讼的过程中，中

³¹ 参见史九领、洪永淼、刘颖：《美国〈2022 年芯片与科学法案〉对我国相关产业的影响与对策》，《中国科学院院刊》2024 年第 2 期，第 381 页。

³² 参见《论招商引资模式的改进——从反补贴的角度分析》，国际经济法网，<https://ielaw.uibe.edu.cn/fxlw/gjjjfl/gjtzf/11855.htm>，2025 年 2 月 6 日最新访问。

国也将面临裁决无法生效、不利于中国出口管制政策的未来实施等两个方面的风险。

从利端角度来看，中国提起 WTO 之诉是阻止美国和立陶宛在远程访问领域违法措施的有力手段，提高了美国和立陶宛在远程访问领域违法实施出口管制的成本。在提起 WTO 之诉并获得 DSB 的胜诉裁决后，中国可以依据胜诉裁决要求美国和立陶宛在合理期限内改正其违法做法。若美国和立陶宛无法在合理期限内改正其违法做法，中国也可以与美国和立陶宛进行谈判并要求得到相应补偿。在美国和立陶宛拒不执行裁决或双方之间没有就补偿问题达成协议的情况下，中国还可以请求 DSB 授权其对美国和立陶宛进行报复或交叉报复，中止对美国和立陶宛履行依照 WTO 协议所应承担的减让或其他义务。上述措施将大幅提升美国《远程访问安全法案》与立陶宛《电力法》修正案的实施成本，并迫使美国和立陶宛重新考虑并改变这一违法措施。

从弊端角度来看，中国提起 WTO 之诉，该裁决不仅有极大可能无法生效，并可能会对中国出口管制政策的未来实施造成不利影响。第一，由于 WTO 上诉机构停止运作，中国提起 WTO 之诉的裁决极有可能无法生效。2019 年 12 月，由于美国的阻挠，WTO 上诉机构因为法官人数不足而停止运作，至今尚未恢复。因此，在 DSB 中，虽然作为“一审”的专家组程序目前仍可以正常运转并作出裁决。但若 WTO 之诉中的其中一方提出上诉，由于上诉机构已经停止运作、上诉程序无法完成审理，该专家组裁决同样无法生效。³³截至 2023 年 2 月，WTO 一共有 29 起案件处于上诉待审的状态，而其中由美国提起的上诉待审案件就有 11 起，占据了极高的比例。³⁴由此可见，在 WTO 上诉机构停止运作的背景下，中国即使针对美国和立陶宛提起了 WTO 之诉并获得了专家组的胜诉裁决，美国和立陶宛也极有可能就此提出上诉并使该裁决无法生效。第二，即使中国获得了 DSB 的胜诉裁决，该裁决未来同样有可能阻挠我国限制跨境远程访问的政策实施。DSB 若裁定美国和立陶宛针对特定国家的远程访问限制违反了最惠国待遇、市场准入等 WTO 协定规则，这意味着“不得限制向特定国家提供远程访问”得到了 WTO 的官方承认。当前，中国的科技发展水平不断提升，中国在未来亦有可能产生限制外国远程访问并维护我国科技领先地位的政策需求。除此之外，科技的发展使得各国对于国家安全认识的不断变化，未来中国同样有可能在特定领域产生限制他国远程访问我国境内设备的政策需求。本案中的胜诉裁决将在一定程度上对于上述政策的实施产生阻碍。

³³ 参见都毫：《上诉机构停摆后的 WTO 争端解决》，《南大法学》2021 年第 1 期，第 101 页。

³⁴ 参见《WTO 上诉机构停摆近 4 年，全球贸易“裁判”还能归位吗？》，中国贸易救济信息网，<http://cacs.mofcom.gov.cn/article/flfwpt/jydy/cgal/202309/177731.html>，2025 年 2 月 6 日最新访问。

（三）合理运用国内法机制，反制不当措施

针对美国《远程访问安全法案》和立陶宛《电力法》修正案的不当措施，中国可以通过国内法机制予以坚决反制，积极维护中国国家 and 企业的合法权益。

当前，中国已经形成了以《阻断外国法律与措施不当域外适用办法》以及《反外国制裁法》等为代表的反制裁法律体系，在阻断外国法律与措施不当域外适用对中国的影响、阻断外国歧视性限制措施在中国的效力等方面发挥了重要作用。就美国《远程访问安全法案》以及立陶宛《电力法》修正法针对中国实施的一系列具有歧视性的远程访问限制措施，中国应当通过《阻断外国法律与措施不当域外适用办法》《反外国制裁法》等阻断法规范予以积极反制。具体于《阻断外国法律与措施不当域外适用办法》而言，根据该办法第 7 条而言，可以由国务院商务主管部门发布不得承认、不得执行、不得遵守美国《远程访问安全法案》以及立陶宛《电力法》修正案的禁令，还可以依据第 9 条的规定向法院起诉，要求遵守该美国 and 立陶宛法案的当事人赔偿中国实体遭受的损失。具体就《反外国制裁法》而言，根据该法第 4 条的规定，中国国务院可以将直接或者间接参与制定、决定、实施美国《远程访问安全法案》以及立陶宛《电力法》修正案的个人、组织列入反制清单，甚至可以将反制扩展到该法第 5 条所规定的范围，并实施该法第 6 条所规定的一系列必要反制措施。不仅如此，针对执行或者协助执行美国《远程访问安全法案》以及立陶宛《电力法》修正案并侵害中国公民、组织合法权益的组织和个人，根据《反外国制裁法》第 12 条的规定，中国的公民、组织也可以依法向人民法院提起诉讼，要求其停止侵害、赔偿损失。

针对美国《远程访问安全法案》和立陶宛《电力法》修正案适用中国阻断法将会带来利弊两个方面的不同影响。一方面，适用阻断法可以对美国 and 立陶宛的违法措施进行积极反制，并使中国实体的损失得到赔偿。另一方面，适用阻断法也有可能导致义务冲突及合规困境。

从利端来看，通过适用中国阻断法，美国 and 立陶宛被列入反制清单的个人和组织将被实施一系列反制措施，中国实体遭受的损失也可以通过国内诉讼的途径得到赔偿。此举不仅有效降低了该违法措施对我国的危害和影响，弥补中国实体由于远程访问被限制以及算力断供而导致的损失，在一定程度上也可以增加美国与立陶宛实施违法措施的成本，迫使美国 and 立陶宛对相关法案和措施予以重新考虑。

从弊端来看，适用阻断法可能导致当事人面临中国阻断法和美国《远程访问安全法案》、立陶宛《电力法》修正案之间的制度对立和义务冲突。若美国《远程访问安全法案》以及立陶宛《电力法》修正案被列入《阻断外国法律与措施不当域外适用办法》第 7 条的禁令范围，对于有关当事人来说，将仅能在外国法案和中国阻断法之间选择其一而遵守之，并陷入“矛盾”的选择困境，当事人的违法

风险也将显著升高。这种义务冲突对于国际贸易的发展将产生显著不利的影响，影响企业跨国业务的开展和正常的生产经营。尤其是对于美国《远程访问安全法案》而言，鉴于美国在关键技术领域的领先地位以及中美之间在综合国力方面的差距，对于外国当事人来说，面对中国阻断法和美国法案之间的冲突，当事人极有可能屈从于美国的压力而选择遵守美国法案，甚至选择减少在中国开展相应业务以避免阻断法的处罚，在一定程度上使中国在国际社会上更加孤立。

综上所述，针对美国和立陶宛不当规制跨境远程访问，中国可以通过产业补贴政策、WTO 违法之诉以及国内法机制予以应对，积极利用各项法律手段有效维护中国的合法权益。与此同时，也要考虑上述措施所可能带来的不利后果，在综合平衡、充分考虑的基础上谨慎作出决策。

WTO 规则下人工智能产品的归类

——挑战与应对

张鹏文¹

摘要：人工智能与国际贸易关系密切，在促进贸易发展的同时也给现行贸易规则带来诸多挑战。应用层的人工智能产品有两类：一类产品在物理载体中嵌入人工智能，另一类产品则是没有物理载体的纯人工智能服务。两类人工智能产品都是数智时代的重要贸易对象，但在 WTO 规则下却难以对其进行归类。归类困难表现在两个方面：一方面，第一类人工智能产品的货物和服务属性难以明确。此类产品具有货服聚合、产品新兴、自主学习等特性，这些特性使“基于特征的方法”、“技术中立的方法”和“HS、CPC 编码的方法”都难以判断其货服属性。另一方面，作为纯人工智能服务的第二类产品 and 可能被定性为服务的第一类产品，都面临确定具体服务部门和服务提供模式的难题。人工智能产品可能集成多种服务具备不同功能，难以明确其服务部门的归属。与载体打包流通的方式及其潜在的法律人格使两类人工智能产品难以适应 GATS 现有的四种服务提供模式。解决人工智能产品归类难题存在过渡性方式和根本性变革两种路径。过度路径主要是对现有规则的修补完善，可以借鉴《美国统一商法典》为货服归类提供更加清晰的指引，引入模式 5 明确人工智能产品的服务提供方式。根本性的变革则需要打破 WTO 规则下“货服二分”的基本设定，在初阶的尝试中可以借鉴 FTA 中的数字产品模式，一定程度上回避归类难题，在非歧视待遇上适用相同规则。而进阶模式则需要最大程度的实现货物贸易和服务贸易规则的趋同，建立一套统一的规则，从根本上解决归类的难题。

关键词：人工智能产品 货服二分 货物贸易 服务贸易

一、问题的提出

人工智能对贸易有着广泛而深刻的影响，人工智能可用于降低与贸易物流、供应链管理和法规遵从相关的贸易成本、甚至改变服务贸易的模式。²但同时人工智能也对现有的贸易规则，尤其是 WTO 规则提出了诸多挑战。WTO 框架下贸易的归类变得困难就是人工智能对 WTO 规则带来的典型挑战。

WTO 按照“货物——服务”二分的逻辑对贸易进行规制，不同的贸易类型适用不同规则。货物贸易适用《关税与贸易总协定》（General Agreement on Tariffs

¹ 张鹏文，海南大学法学院博士研究生，联系方式：2637701094@qq.com，18774002232

² WTO, Trading with intelligence How AI shapes and is shaped by international trade

and Trade, GATT), 服务贸易适用《服务贸易总协定》(General Agreement on Trade in Services, GATS)。对于货物, WTO 成员方通过关税减让承诺表对不同种类的货物削减相应的关税幅度。而对于服务贸易, 成员方按照服务提供的模式、服务的类别做出承诺, 成员方的义务与其做出的承诺密切相关。由 WTO 规制贸易的逻辑可以知晓, 对一项产品的贸易进行归类至关重要。这种归类可以分为两个层面: 首先, 要确定该贸易是货物贸易还是服务贸易; 其次, 若该贸易为服务贸易, 则要确定它所属的服务部门和服务提供的方式。在此将上述两个层面的问题均称之为贸易的归类。

随着人工智能等数字技术的发展, 货物和服务的边界变得日益模糊, 数字产品的归类变得困难。学界较早注意到了数字产品的归类问题, 并对数字产品的货服属性提出了相应的主张。有观点认为数字产品是货物³, 有观点认为数字产品是服务⁴, 还有观点认为数字产品应该是一种新的类别⁵。国内学者则从我国的现实利益出发, 指出在谈判中我国应该坚持将数字产品作为服务的立场。⁶但上述研究主要聚焦广义的数字产品归类, 没有针对人工智能这一特殊的数字技术引发的贸易归类问题进行研究。人工智能相较于其他数字技术, 具有特殊性, 这种特殊性引发的归类难题与普遍意义上的数字产品归类问题, 存在差异。因此, 需要对此进行研究。

二、人工智能使货物和服务难以区分

(一) 人工智能产品对“基于特征分类”的挑战

“基于特征的方法”是区分货物和服务的重要方式。WTO 建立时, 人们通常认为服务相较于货物有四个特征: 无形性、异质性(定制的、非标准化生产的)、生产和消费同时进行、不可存储。⁷在 WTO 的争端解决实践中, 专家组和上诉机构也曾使用这些特征区分货物和服务。但这些特征在区分货物和服务时, 并无次序上的先后。

1. 人工智能产品同时具有货物和服务的部分特征

在美国——加拿大软木案(DS257)中, 上诉机构指出有形性是货物的一个参考特征, 但不是唯一特征。⁸这意味着需要综合考虑四项标准来区分货物和服

³ FLEUTER S. The role of digital products under the WTO: a new framework for GATT and GATS classification[J]. Chicago journal of international law, 2016(17): 153-177.

⁴ WTO. Fifth dedicated discussion on electronic commerce under the auspices of the general council summary by the secretariat of the issues raised[R]. 2003.

⁵ HOJNIK J. Technology neutral EU law: digital goods within the traditional goods/services distinction[J]. International journal of law and information technology, 2017(25): 63-84.

⁶ 陈儒丹: 《WTO 框架下数字产品在线跨境交易的法律性质》, 载《法学》2008 年第 7 期; 刘佳奥: 《WTO 规则下数字产品的法律属性研究》, 载《国际法学刊》2022 年第 1 期。

⁷ Glenn Parry et al., Goods, Products and Services, in MAIRI MACINTYRE ET AL. (EDS.) SERVICE DESIGN AND DELIVERY 19, 19-23 (Springer, 2011).

⁸ Report of Appellate Body, United States – final countervailing duty determination with respect to certain softwood lumber from Canada, WT/DS257/AB/RW, 5 December 2005.

务。当与部分特征相符，但与另一部分部分不符时，则难以判断。在过去，通常而言一项产品或者满足全部特征，或者全部不满足；但人工智能的运用，使得一项产品仅满足服务部分特征的情形越来越多。以可以个性化定制的智能家具为例，其满足异质性的特征，却无形性不相符合。此时，无法轻易地将其定性为货物或服务。

2.同一特征下人工智能产品的不同部分或符合或不符

除此之外，就某一特征而言，嵌入人工智能的产品可能在某些部分符合，在某些部分又不符合。人工智能产品的核心特征在于其物理载体与无形功能的深度融合。传统货物以有形物理形态为存在基础（如机械设备），服务则以无形交互为核心（如咨询服务）。而人工智能产品往往同时具备硬件实体与软件算法的双重属性，导致其法律归类陷入两难。以自动驾驶系统为例：其硬件部分（传感器、芯片）属于典型货物，受 GATT 规则约束；而软件系统（算法、数据模型）则通过云端更新持续优化功能，具有服务的即时性与无形性。特斯拉的“全自动驾驶（FSD）”套餐即为此类混合产品的代表——消费者一次性购买车辆（货物），但需定期订阅软件服务以激活功能。此类产品在贸易中既涉及硬件关税，又涉及服务市场准入承诺，现行规则难以协调。

（二）技术中立的方法失灵或失效

技术中立在国际贸易法中尚未形成统一的定义，但从不同学者对技术中立的理解中可以发现：对相同的交易不因其技术手段的不同而适用不同的贸易规则是技术中立的应有之义。⁹技术中立在贸易归类问题上最通常的应用是将在线服务归类为其他同类服务所属的传统服务部门。虽然在 WTO 争端解决的实践中还未出现利用技术中立区分货物和服务的案例，但其内涵和在 WTO 规则中的地位都表明其可以用来区分货物和服务。技术中立在 GATT 和 TRIPS 中都具有事实上的原则地位，在 GATS 中虽有争议，但“美国诉中国出版物及视听制品案”中上诉机构适用演进解释间接承认技术中立原则的法理。¹⁰事实上，美国曾提出将数字产品定性为货物更符合技术中性的原则。¹¹技术中立要求相同的交易不因技术手段而适用不同的规则，这意味着同类产品不能因技术手段不同而分别适用 GATT 和 GATS。这在理论上可以解决数字产品的分类问题。然而在现实中，涉及人工智能的产品会因为难以判别同类产品，和自身自主学习的特性等原因，导致技术中立难以解决其货服归类问题。

1.难以判别同类产品

人工智能产品的同类产品难以判别有两方面的原因：第一，人工智能是一项

⁹ 参见范笑迎：《WTO 的数字化：技术中立原则的制度证成与展开》，载《国际法研究》2024 年第 4 期。

¹⁰ 参见范笑迎：《WTO 的数字化：技术中立原则的制度证成与展开》，载《国际法研究》2024 年第 4 期。

¹¹ 参见刘佳奥：《WTO 规则下数字产品的法律属性研究》，载《国际法研究》2022 年第 1 期。

新兴的数字技术，其既可以应用于传统的货物和服务之中，也可以用来创造出全新的产品。这些新产品是 WTO 在谈判时不存在的，这就导致很难找到其同类产品。第二，人工智能产品具有集成性。其本身是不同服务和物理载体的集成。从整体来看，这是全新的产品，没有与之完全对应的同类产品。从部分来看，不同的部分对应的同类产品在货服属性上也不尽相同。生成式 AI（如 ChatGPT）提供的内容生成服务无传统对应物，既非实体书籍（货物），亦非传统咨询服务（服务）。技术中立原则依赖的“相似性逻辑”在此失效。

2. 自主学习引发的挑战

人工智能具有自主学习能力，可以自我更新迭代，可以通过大数据的训练在自主学习中获得更多的功能。因此，从时间的动态维度来看，不同时期的人工智能产品其同类产品也会有所不同。这进一步加大了利用技术中立进行货服归类的挑战。

（三）人工智能产品的特征放大 HS 编码和 CPC 编码分类的局限

一种确定货物和服务界限的方法可能是参考分类清单。GATT 和 GATS 分别都采用数字分类系统，产品被分配不同的代码，然后列在成员方的清单中。¹²GATT 通常使用世界海关组织制定的《商品名称及编码协调制度》（HS）来对货物进行分类。而 GATS 则鼓励成员方采用乌拉圭回合谈判期间使用的《服务部门分类清单》（W/120），该清单基于更全面的联合国《产品总分类》（CPC）制定。一旦产品符合某个代码，根据相关分类法分配的代码，它就被归类为货物或服务。

1. 新兴的人工智能产品放大 HS 和 CPC 的过时缺陷

虽然这种方法对许多传统产品可能没有问题，但其更新速度远落后于技术迭代，导致人工智能产品陷入分类盲区，可能会引发争议。HS 每 5 年修订一次，难以匹配技术演进速度。新兴的人工智能产品可能在 HS 和 CPC 编码制定时尚未出现，例如自动驾驶系统的软件更新、神经网络模型的跨境传输等新兴交易形式就会没有准确的对应编码。人工智能产品的新兴特征放大了两类编码动态更新慢的缺陷，可能会引发以下两类归类问题：第一，某一人工智能产品可能在 HS 或 CPC 中同时拥有两个或两个以上的合适编码，以 CPC 为例，CPC 将服务分为 12 大类，但人工智能相关服务可能分散于多个类别，但同一个类别往往又难以完整覆盖一项具体的服务：计算机服务（CPC 84）：涵盖软件开发，但不包括 AI 训练服务，专业科技服务（CPC 85）包括数据分析，但未明确算法优化是否属于此类。更极端的情况是，人工智能产品可能既存在 HS 编码，又存在 CPC 编码，这使其在一种分类系统中被归类为货物，而在另一种系统中被归类为服务。出现这一情形除了编码未动态更新外，人工智能技术和服务被嵌入在物理载体中

¹² Smith, Fiona, & Woods, Lorna, (2005). distinction without difference: exploring the boundary between goods and services in the world trade organization and the european Union, 12(1)Columbia Journal of European Law, 1-52(2005).

的这一特性也是原因，这一特性放大了 HS 编码物理形态导向的缺陷。

2. 人工智能技术服务嵌入物理载体的特征放大 HS 编码物理形态导向的缺陷

HS 编码是物理形态导向的，其以物理属性（如材质、用途）为分类依据，无法涵盖无形组件。例如，HS 8471.30 条目仅覆盖“数据处理设备”，未区分是否具备 AI 功能。人工智能的技术和服务通常会嵌入到物理载体中，这一特性与 HS 编码的物理形态导向相结合，会导致任何有形的人工智能产品都有或者可能有 HS 编码。同时被嵌入的人工智能技术本身可能会提供服务，有相应的 CPC 编码。

人工智能产品自身的特性和 HS、CPC 编码的局限会引发实践中的归类混乱。云计算服务通常被归类为 CPC 84（计算机服务），但 AI 训练所需的算力租赁是否属于同一类别存在争议。3D 打印设计文件若通过电子传输，可能被视为服务（CPC 84）；若存储于 U 盘出口，则被视为货物（HS 8523.49）。综上，无法通过 HS 编码或者 CPC 来确定人工智能产品究竟属于货物还是服务。

三、人工智能使服务的所属部门和提供方式难以确定

（一）所属部门难以确定

1. 人工智能集成性与服务部门排他性的冲突

很多数字服务越来越多地将电信服务与其他服务捆绑在一起。人工智能更加剧了这一点。这些数字服务的功能具有多样化，通过各种形式的服务来提供全面的数字平台。¹³以人工智能律师为例，除了法律服务，它还可以同时被视为互联网电信服务(代码 842)、在线内容服务(代码 843)、数据传输服务(代码 8415)，甚至数据库服务(代码 844)。同时 GATS 的分类具有排他性，换言之，任何特定的服务应该都只属于一个类别。¹⁴这表明集成了多种服务的人工智能产品，即使被归类为服务，也仍然难以判断其具体属于哪一服务部门。各成员方对不同的服务部门承诺的开放水平存在差异。为了保护本国产业，成员方可能会倾向于将人工智能产品归类为开放水平低或未作出开放承诺的服务部门，而想要拓展市场的企业则倾向于将同样的服务归类为开放水平高的服务部门。这无疑会引发贸易争端。

2. 人工智能的安全风险使其难以被归类为“其他服务”

由于 W / 1 2 0 的大部分服务部门都包括了“其他服务”，因此，W / 1 2 0 未明确规定的新服务可以归入“其他服务”中。这些“其他服务”可以在不进行重大结构更改的情况下容纳现有归类系统未充分涵盖的“新服务”。¹⁵有许多成员方在承诺中对“其他服务”进行了开放的承诺，如果将人工智能产品归类为其他服务，就意味着要对其予以开放。但人工智能产品区别于其他服务的一个特点就在

¹³ 参见谭观福：《国际贸易法视域下数字贸易的归类》载《中国社会科学院研究生院学报》，2021 年第 5 期。

¹⁴ 参见戴艺晗：《国际贸易法视域下的人工智能规制》，载《上海财经大学学报》，2023 年第 2 期。

¹⁵ 参见谭观福：《国际贸易法视域下数字贸易的归类》载《中国社会科学院研究生院学报》，2021 年第 5 期。

于其存在较高的安全风险。基于这种担心，很多缔约方不愿意将其归类为“其他服务”。事实上，非洲集团就指出，成员在乌拉圭回合谈判中作出的承诺不能自动适用于3D打印、机器人技术、无人机交付或人工智能等新技术。成员在谈判承诺表时尚不存在的新服务不能事后适用，这些服务超出了W/120的涵盖范围。按照正面清单的逻辑，GATS不会自动涵盖新型数字服务，新义务的纳入需要成员进行积极协商并更新承诺。

（二）服务提供方式难以确定

1. 人工智能的潜在法律人格会模糊不同提供方式的边界

与其他数字技术不同，人工智能存在被授予法律人格的可能。欧洲议会早在2017年2月16日通过的一项决议中提出，在某些条件下，人工智能机器人有朝一日可能会被法律赋予“电子个人”的地位。¹⁶人工智能是否被授予法律人格（如欧盟讨论的机器人“电子人格”），将会影响对其服务提供方式的确定。

以法律咨询服务为例，有些国家在GATS的承诺表中会要求提供法律服务的是自然人或法人。日本在其承诺表中要求模式1和模式2下提供法律咨询服务的是在相关司法管辖区具有合格律师资格的“自然人”，在模式3下则要求是自然人或法人（专业的法律公司）。¹⁷如果是人工智能律师¹⁸提供法律咨询服务，其是否具有法律人格，具有何种法律人格，都会影响其服务的提供模式的辨别以及能否以该种模式进入日本市场。若规定人工智能律师不具有法律人格，则其无法以模式1、模式2和模式3的方式提供法律咨询服务；若认定其有法律人格，类似于ROSS的人工智能系统可以作为法人或法律规定的新的一类人获得更正式的承认，以承担法律权利和责任。但这种法律人格与自然人不具有同等地位，也无法以模式1和模式2提供服务。¹⁹

2. 服务嵌入物理载体超出了现有的服务提供模式

人工智能技术服务与产品通常聚合为一体而打包流通，²⁰当将其视为服务时，它的服务嵌入在了物理载体中并随着物理载体一起到消费者手中。这种服务的提供与GATS规定的四种服务提供模式都不能完全吻合。模式一为跨境提供，在模式一下服务提供者和消费者均不发生地理位置的移动，仅仅服务发生位移。人工智能服务与货物的打包流通的商业模式下，消费者不发生位移，服务发生了位移，

¹⁶ Resolution of 16 Feb. 2017 with Recommendations to the Commission on Civil Law Rules on Robotics, Eur. Parl. Doc. (2015/2103(INL)) ¶ 59(f) (2017) [hereinafter Civil Law Rules on Robotics].

¹⁷ 参见 Japan—Schedule of Specific Commitments, WTO Doc. GATS/S/C/46 (Apr. 15, 1994), <https://www.mofa.go.jp/policy/economy/wto/submit0304-s.pdf>.

¹⁸ 世界上第一个人工智能律师ROSS建立在IBM认知计算机沃森的基础上，被设计用来阅读和理解法律语言，在被问及问题时推测假设、研究，然后连同参考文献和引文回答问题，罗斯还能从经验中学习，其速度和知识储备会随着用户与其互动逐渐增长。<https://www.rossintelligence.com/what-is-ai>.

¹⁹ Han-Wei Liu and Ching-Fu Lin, Artificial Intelligence and Global Trade Governance: A Pluralist Agenda, 61 Harvard International Law Journal, 408-450(2020).

²⁰ 沈伟、赵尔雅：《数字经济和数字博弈双重背景下人工智能的国际法治理》，载《武大国际法评论》，2023年第4期。

但服务的提供者是否发生位移则难以辨别。当服务被嵌入到物理载体中时，该产品为消费者提供服务，但产品本身能否构成服务提供者存在极大争议。即使产品可以构成服务提供者，它也发生了位移，与模式一不符。模式二和模式四分别是服务提供者和消费者发生了位移。人工智能流通的过程来看，消费者没有移动，显然不符合模式四。至于模式二，人工智能产品之所以难以被认定为以模式二的方式提供服务，是因为人工智能的法律人格争议。模式三为商业存在，人工智能产品从一国运输至另一国，并没有在消费者所在地建立商业存在，因此也与模式三不符。

四、应对挑战的路径

（一）过渡性的路径——修补现有规则

1. 以判断混合交易性质的方法对人工智能嵌入产品定性

人工智能技术、服务嵌入到有形的物理载体中使两者构成了一个混合的整体。对这个整体的贸易实质上是一种混合性的交易。在 WTO 体系之外的其他法律中，也存在对混合性交易的定性问题。《美国统一商法典》（UCC）可以作为有益的参考。UCC 第 2 条的标题是“销售”仅适用于“货物”交易。在起草时，货物与服务是截然相反的。也就是说，商品是“可移动的”、有形的东西，而服务则不是。更具体地说，由于《统一商法典》第 2 条的适用范围仅限于货物，这就产生了一个问题，即同时包含货物销售和服务的合同是否应受《统一商法典》管辖。大多数法院只有在整个交易的“主要目的”是销售货物的情况下才适用第 2 条。在实践中，在处理具有双重性质的合同时，人们采用的是“全有或全无”的思路。²¹一旦法院确定合同的主要目的是货物，则该合同的性质为销售合同。在 WTO 法下对人工智能嵌入产品时也可以借鉴这一思路，确定该产品的主要目的是什么，确定主要目的后，该产品的货服属性在一定程度上可以得到明确。除 UCC 之外，OECD 也出台过相应的文件帮助判断兼具货服属性的产品的性质。²²

2. 创造模式 5 作为人工智能产品的服务提供方式

有学者提出建立 GATS 下的模式 5，适用于嵌入商品制造过程的服务。²³模式 5 这一术语仍处于起步阶段，其界限和范围并不精确。到目前为止，这一术语在经济文献中被用于探讨制造业出口中服务投入的重要性，在欧盟和其他地方均有相关研究。尽管这些研究具有一定的探索性，但这一探索能够制定更适合全球价值链中货物和服务互动方式的贸易规则。从现有的讨论中可以发现，模式 5 意在描述货物和服务捆绑在一起进行贸易流通的情形，这一情形也被称为“装在

²¹ Shin-yi Peng, A New Trade Regime for the Servitization of Manufacturing: Rethinking the Goods-Services Dichotomy, 54 Journal of World Trade, 699-726(2020).

²² OECD Trade Policy Paper, Interactions Between Goods and Services in International Trade: Implications for Customs Valuation and Rules of Origin.

²³ 参见戴艺晗：《国际贸易法视域下的人工智能规制》，载《上海财经大学学报》，2023 年第 2 期。

盒子里交易的服务”。这无疑与人工智能嵌入产品的交易情形相契合，人工智能技术和服务嵌入在物理载体中进行交易，也可以被视作“装在盒子里的交易”。因此，将人工智能产品定性为服务时，引入模式 5 能够在一定程度上明确其提供方式。但目前的模式是一个处在探索中的术语，在合理性和具体操作层面仍存在诸多不确定性。

首先，模式 5 与 GATT 和 GATS 之间的关系存疑。模式 5 描述是服务嵌入货物中交易的情形，旨在解决难以区分货物和服务的问题，将其作为 GATS 下的一种服务提供方式是否合理。曾有学者指出“模式 5”这一术语虽借用自 GATS 的四种服务提供模式，但并不一定意味着要将模式 5 作为 GATS 下的一种新的服务提供模式。如不将模式 5 作为 GATS 下的一种提供方式，那它与货物贸易和服务贸易的关系如何？是否要将其作为独立于货物和服务之外的第三类，仍然有待讨论。

其次，模式 5 的范围也需要界定。模式 5 是仅要求货物和服务在形式上打包交易，还是要求服务在最终产品中占有相应的份额。如若将最终产品定性为服务，则应该要求服务在最终产品中的价值份额超过物理组件的价值份额。

（二）根本性改革——打破货服二分

1. 初级模式——FTA 中的数字产品模式

自由贸易协定（Free Trade Agreement FTA）已经开始了打破货服二分的尝试。《全面与进步跨太平洋伙伴关系协定》（CPTPP）和《美墨加协定》（USMCA）分别在电子商务章节和数字贸易章节中引入了“数字产品”的概念，并规定了数字产品的非歧视待遇。两项协定都在脚注中明确表明数字产品的定义不反应缔约方对通过电子传输的数字产品贸易应被归入服务贸易或货物贸易的观点。²⁴采用数字产品的概念回避了缔约方间对电子传输的数字产品贸易归类为货物贸易或服务贸易的争论，在两个协定的电子商务/数字贸易章节中，与货物和服务待遇差异有关的困难得到了解决。在这一模式下，只要符合数字产品的定义，都适用相同的非歧视待遇。

但值得注意的是，CPTPP 的定义将数字产品限定为无形性的、可通过电子传输的产品。这一定义似乎难以包含嵌入到物理载体上的人工智能服务。因此，在 CPTPP 框架下，人工智能嵌入的产品可能无法直接被认定为数字产品，可能仍然需要在货物贸易和跨境服务贸易章节进行规制，此时仍然需要对其进行“货物——服务”的归类。因此，在 WTO 的改革和电子商务的谈判中，所选取“数字产品”、“数字贸易”或其他概念应该要包含人工智能等技术和服务嵌入物理载体的情形。

此外，在贸易规则体系中，除了非歧视待遇外，还有诸多其他规则。CPTPP 和 USMCA 等 FTA 仅仅实现了在非歧视待遇方面同类的数字产品享有相同的待

²⁴ 参见 CPTPP 第 14.1 条脚注 3

遇。至于同类数字产品在其他方面能否适用相同的规则,可能仍然取决于产品的货服属性。因此,彻底打破货物二分,需要实现货物贸易和服务贸易规则的趋同。

2.进阶模式——服务贸易和货物贸易规则的趋同

在 WTO“货物——服务”二分的架构下,对产品的贸易类型进行归类的意义在于明确其适用的规则。调整货物贸易的措施适用 GATT,调整和影响服务贸易的措施适用 GATS。随着人工智能等技术的发展,货物和服务高度融合的产品越来越多,单纯的货物产品和服务产品越来越少。与之对应,调整和规制这种混合性质产品的规则和措施也会越来越多,单纯针对或影响货物、服务的措施会相对变少。在这样的变化之下,越来越多的措施既涉及 GATT,又涉及 GATS。这无疑会导致规则和义务的累加。在这种情形下,将货物贸易和服务贸易规则趋同变得合理而必要。

参考文献

- [1] 戴艺晗. 国际贸易法视域下的人工智能规制——以 WTO 规则为视角 [J]. 上海财经大学学报, 2023, 25 (02): 122-136+152.
- [2] 李墨丝,应玲蓉,徐美娜. DEPA 模式数字经济新议题及启示 [J]. 国际经济合作, 2023, (01): 27-36+93.
- [3] 任宏达. 数字贸易国际规则的多元发展与中国元素——以中国申请加入《数字经济伙伴关系协定》为视角 [J]. 中国发展观察, 2021, (24): 46-48+52.
- [4] 沈伟,赵尔雅. 数字经济背景下的人工智能国际法规制 [J]. 上海财经大学学报, 2022, 24 (05): 123-137.
- [5] 沈伟,赵尔雅. 数字经济和数字博弈双重背景下人工智能的国际法治理 [J]. 武大国际法评论, 2023, 7 (04): 81-105.
- [6] 唐林垚. 公共治理视域下自动化应用的法律规制 [J]. 交大法学, 2022, (02): 17-27.
- [8] 张正怡. 论数字经济协定的造法“再平衡”走向及中国回应 [J]. 法商研究, 2022, 39 (06): 58-71.
- [9] 周念利,廖宁,黄宁. “贸易面向型”数字技术规则的发展演进研究——基于“美式模板”及“新式模板”对比的视角 [J]. 亚太经济, 2024, (02): 28-39.
- [10] Aik Hoe Lim, Trade Rules for Industry 4.0: Why the Technical Barriers to Trade Agreement Matters Even More, in Shin-Yi Peng, Ching-Fu Lin, and Thomas Streinz (eds), Artificial Intelligence and International Economic Law: Disruption, Regulation, and Reconfiguration. London: Cambridge University Press, 2021.
- [11] Anupam Chander. Artificial Intelligence and Trade, in Mira Burri(eds), Big Data and Global Trade Law. Cambridge: Cambridge University Press, 2021.

[12] Bryan Mercurio , Ronald Yu, Convergence, Complexity and Uncertainty Artificial Intelligence and Intellectual Property Protection in Shin-Yi Peng, Ching-Fu Lin, and Thomas Streinz (eds), Artificial Intelligence and International Economic Law: Disruption, Regulation, and Reconfiguration. London: Cambridge University Press, 2021.

[13] Gregory Shaffer , Trade Law in a Data-Driven Economy The Need for Modesty and Resilience in Shin-Yi Peng, Ching-Fu Lin, and Thomas Streinz (eds), Artificial Intelligence and International Economic Law: Disruption, Regulation, and Reconfiguration. London: Cambridge University Press, 2021.

[14] Han-Wei Liu , Ching-Fu , Lin Artificial Intelligence and Global Trade Governance: A Pluralist Agenda", Harvard International Law Journal, January 2020.

[15] Jaemin Lee(eds), Artificial Intelligence and International Law, Singapore: Springer Nature Press.,2022.

[16] KAMINSKI M E, IRION K, YAKOVLEVA S. Privacy Peg, Trade Hole: Why We (Still) Shouldn't Put Data Privacy in Trade Law [J] . University of Chicago Law Review, Online, 2023, 3 (27)

Paul Armstrong. From Law Office to Administrative Proceedings. The Judges' Journal, 59(1), 2020.

[17] S Frankel, "It's Raining Carrots: The Trajectory of Increased Intellectual Property Protection," in G Ghidini et al.(eds)

[18] S Yakovleva and K Irion,"Pitching trade against privacy: Reconciling EU governance of personal data flows with external trade" (2020) 10 International Data Privacy Law 201.

[19] Shin-Yi Peng, Ching-Fu Lin, and Thomas Streinz (eds), Artificial Intelligence and International Economic Law: Disruption, Regulation, and Reconfiguration. London: Cambridge University Press, 2021.

[20] Stephen M. Walt. Big Tech Won't Remake the Global Order, November 8, 2021.

智能汽车：规制措施适用国际商品贸易规则亦或服务贸易规则？

张素敏¹

摘要：人工智能技术的飞速发展正在深刻影响全球经济，其中，智能汽车作为人工智能与汽车制造业深度融合的典型代表，正在全球范围内加速商业化落地。然而，在国际贸易法律框架下，各国规制智能汽车的措施在 WTO 法框架下应当受到 GATT 还是 GATS 约束？选择不同的约束协定，将直接影响包括关税、知识产权保护、责任划分及争端解决机制，以及确定分类后的其他规则适用这与智能汽车的在国际贸易法下“商品”或“服务”的分类有关。从“物理特征——贸易经济——法律规则”三个层面入手，即考量产品的通常含义和特征，全球智能汽车技术发展状况和国际贸易活动，GATT 和 GATS 的目的宗旨以及成员方国内措施的法律内容性质，以智能汽车的类型化为基础针对这个问题做出相对明确的回答。

关键词：智能汽车；自动驾驶汽车；数字产品；国际贸易规则；人工智能

一、问题的提出

人工智能技术正加速变革全球经济，其中，智能汽车的快速发展成为最受关注的技术突破之一。全球各国纷纷推动自动驾驶技术的研发与商业化，但在国际贸易领域，智能汽车的法律定位尚未形成统一共识。传统汽车作为商品进行跨境交易，但当汽车集成了高级自动驾驶技术，并且涉及云端数据支持和软件服务时，是否仍应仅被视为商品，还是应归入服务？抑或两者皆有？智能汽车的分类问题如何影响国际贸易实践？以及现行国际贸易规则如何应对智能汽车的特殊属性？对上述问题的回应不仅关系到国际贸易规则的适用，也直接影响着各国在市场准入、关税壁垒、数据合规等方面的政策制定。目前学界在智能汽车特别是自动驾驶汽车的研究中，已指出其兼具“商品性”和“服务性”的复合属性，而当车辆的自动化程度越高，其“服务”的属性就越显著。但目前并没有适用于兼具产品和服务混合属性的国际贸易规则，而且其通常支持商品和服务贸易之间存在界限从而便利国际贸易的规制。本文通过文献分析、案例研究与比较研究的方法，首先探讨智能汽车商品与服务交织的属性，其次分析国际货物贸易与服务贸易规则的适用困境，最后提出相应的国际法应对策略。这将在理论尚澄清关于智能汽车适用国际贸易规则的分类困境，也将解决实践中智能汽车在国际贸易中的规制困惑，为

¹ 张素敏（1998 年—），女，山东聊城人，对外经济贸易大学法学院 2023 级博士研究生，主要研究方向：国际法、涉外法治。

我国参与 WTO 改革和新一代国际贸易规则制定、掌握国际规则制定话语权建言献策。

二、智能汽车：商品与服务的交织

随着人工智能和传感技术的快速发展，智能汽车尤其是自动驾驶汽车正逐步走向商业化应用。从国际贸易角度看，自动驾驶汽车具有明显的双重属性，即既具备商品的有形属性，又蕴含着服务的无形特征。

（一）智能汽车的概念和分类

智能汽车（intelligent vehicles）是指通过搭载先进传感器等装置，运用人工智能等新兴技术，具有自动驾驶功能，逐步成为智能移动空间和应用终端的新一代汽车，又被称为智能网联汽车、自动驾驶汽车。²智能汽车最大的特征是自动化，国际汽车工程师协会（Society of Automotive Engineers，简称为 SAE）根据智能汽车的自动化程度并结合其他相关参考因素将智能汽车分为六个等级³（参见图 1），也被称为智能汽车的“SAE 等级分类”，其已成为当前国际社会对智能汽车分类的通用模式。

智能汽车 L0-L5 类别的自动化驾驶系统包含从无自动化到不需要人工干预的完全自动化，所谓的自动化是指驾驶任务的部分内容被有意地委托给技术系统，从而使驾驶员从驾驶任务中解脱出来，但是不包括那些在后台活跃、准备在驾驶员无法应对驾驶情况即在紧急或即将发生碰撞的情况下时介入的系统。⁴根据 SAE 规定的自动驾驶级别，L0 级智能汽车和传统的功能汽车并无差别，具备货物的特征。L1 级智能汽车配备高级驾驶辅助系统（advanced driver assistant systems，简称 ADAS）分担驾驶员的部分工作，比如自适应巡航（Adaptive Cruise Control, ACC）、车道保持功能（Lane Keep Assist, LKA）、紧急制动刹车（Automatic Emergency Braking, AEB）等功能。L2 级智能汽车自动驾驶系统开始接管纵向和横向的多个动态驾驶任务，驾驶员无需持续地进行物理操作，但其仍需注意周边环境。L3 和 L4 级智能汽车由自动驾驶系统接管控制和监控周边环境，但是 L3 级相对 L4 级智能汽车，自动驾驶系统在次要任务（与驾驶任务无关的任务，比如看手机、阅读报纸等）方面存在限制并且仍然须驾驶员提供动态驾驶任务接管（fallback performance）。L5 级智能汽车实现了完全的自动化，自动驾驶系统代替人类驾驶员而当时当地地控制汽车，也被称为无人驾驶汽车⁵。根据美国国家

² 《智能汽车创新发展战略》（发改产业〔2020〕202 号）

³ Arne Bartels, Ulrich EBerle & Andreas Knapp, System Classification and Glossary, Adaptive-ip, (6 February 2015) <https://www.adaptive-ip.eu/index.php/AdaptIVe-SP2-v12-DL-D2.1-System%20Classification-file=files-adaptive-content-downloads-Deliverables%20&%20papers-AdaptIVe-SP2-v12-DL-D2.1-System%20Classification.pdf>.

⁴ 同上

⁵ 李磊：《论中国自动驾驶汽车监管制度的建立》，载《北京理工大学学报（社会科学版）》2018 年第 2 期，第 126 页。

公路交通安全管理局（National Highway Traffic Safety Administration, 简称 NHTSA）的，全自动或自动驾驶汽车（fully automated or autonomous vehicles）是指车辆运行时无需驾驶员直接控制转向、加速和制动，并且涉及驾驶员在自动驾驶模式下无需不断监视道路的汽车。由此，负责监控周边环境的主体是驾驶员还是自动驾驶系统是 ADAS 向自动驾驶过渡的核心标准。只有无需驾驶员监控周边环境的智能汽车可被称为自动驾驶汽车，所以 L3-L5 又可被称为自动驾驶汽车。

自动 驾驶 分级 (Level)	名称	定义	驾驶操作	操作环 境和周 边监控	动态驾 驶任务 接管 DDT fallback	系统能 力（系 统作用 域）
0	无自动驾 驶	人类驾驶员全时完成动态驾驶任务的所有方面，即使在警告或干预系统的帮助下也是如此	人类驾驶员	人类驾 驶员	人 类 驾 驶 员	无
1	辅助驾驶	驾驶辅助系统利用驾驶环境信息，根据驾驶模式执行转向或加速/减速，并期望人类驾驶员完成动态驾驶任务的所有其余方面	人类驾驶员和自动驾驶系统			部分驾 驶模式
2	部分自动 驾驶	一个或多个驾驶辅助系统利用驾驶环境信息，在人类驾驶员完成动态驾驶任务的所有其余方面的预期情况下，针对特定驾驶模式执行转向和加速/减速操作	自动驾驶系统			
3	有条件自 动驾驶	自动驾驶系统针对特定驾驶模式执行动态驾驶任务的所有方面，期望人类驾驶员对干预请求做出适当反应		自 动 驾 驶 系 统		
4	高级别自 动驾驶	即使人类驾驶员没有对干预请求做出适当反应，自动驾驶系统也能在动态驾驶任务的所有方面发挥特定驾驶模式的性能			自 动 驾 驶 系 统	
5	完全自动 驾驶	自动驾驶系统在人类驾驶员可控制的所有道路和环境条件下，全时完成动态驾驶任务的所有方面				全 部 驾 驶 模 式

（图 1：SAE 等级分类）

（二）智能汽车的双重属性

智能汽车作为集硬件、软件和数据通信于一体的新型科技产品，其构成和功

能的复杂性决定了其在国际贸易和智能交通领域的重要性。

1.产品构成与物理特征

智能汽车在基础上由硬件构成的实体产品，具备传统汽车的所有基本物理属性。智能汽车的硬件部分主要包括车辆主体、传感器系统、计算平台、执行机构以及通信设备等。车辆主体是自动驾驶汽车的物理载体，其设计和制造直接影响整车的安全性和稳定性。传感器系统是自动驾驶汽车的“眼睛”和“耳朵”，包括摄像头、激光雷达（LiDAR）、毫米波雷达、超声波传感器和 GPS 定位系统等。这些传感器通过采集环境数据，实时感知周围障碍物、车道信息、交通标志、行人和其他交通参与者的动态变化，为自动驾驶系统提供基础数据支持，传感器数据的准确性和实时性直接决定了车辆的环境感知能力。计算平台是自动驾驶汽车的“大脑”，主要由中央处理单元（CPU）、图形处理单元（GPU）和专用的神经网络加速器构成。这个平台负责接收、处理来自各类传感器的数据，并通过深度学习和机器视觉算法进行图像识别、目标检测和行为预测，从而实现对环境的全面理解。硬件平台的高性能和低延迟对于实时决策至关重要。此外，执行机构包括电控系统、制动系统、转向系统等，它们将计算平台发出的决策转化为车辆的实际行动，确保汽车在各种复杂路况下的平稳行驶。最后，通信设备则确保车内系统以及车与车、车与基础设施之间的信息互联互通，为车辆的协同驾驶和智能交通系统奠定基础。

2.数据和软件系统与服务特征

智能汽车在物理实体基础上依靠数据通信和决策系统是来实现智能驾驶。在此过程中，软件系统是链接智能汽车实体设备与数据通信和决策的关捩，是将实体设备收集的真实世界的信息转换为机器可理解的信息从而启动车辆的关键。

软件系统是自动驾驶汽车实现智能化的关键。它主要包括操作系统、传感器数据融合、路径规划、决策控制和系统安全等模块。操作系统作为车辆中所有软件模块的基础平台，负责资源分配和任务调度，为其他应用程序提供支持。传感器数据融合技术则将来自不同传感器的数据进行整合和校准，消除数据冗余和误差，提高环境感知的准确性和鲁棒性。路径规划模块是自动驾驶系统的重要组成部分，其核心任务是根据实时交通状况和道路条件制定最佳行驶路线。路径规划涉及到地图匹配、行驶轨迹预测和避障算法，要求系统能快速反应，并在紧急情况下做出最优决策。决策控制模块则在路径规划的基础上，结合车辆当前状态和周围环境情况，生成具体的驾驶指令，如加速、减速、转向等，确保车辆安全高效地行驶。此外，系统安全模块涵盖了对软件系统的防护措施，包括防止黑客攻击、数据篡改和系统故障等。自动驾驶汽车的核心竞争力在于内嵌的智能驾驶软件系统。该软件系统涵盖自动驾驶算法、深度学习模型、实时数据处理、路径规划、决策控制等功能，能够不断通过无线网络进行远程升级和优化。这部分无形

资产决定了车辆不仅仅是一件一次性销售的商品，而是一种持续性服务。

数据通信和决策系统在软件系统的支持下得以驱动智能汽车的运动。车辆内部通过车载网络（如 CAN 总线）实现各个传感器、计算平台和执行机构之间的高速数据传输；同时，车联网通信系统（V2X）支持车与内部乘客、车与车、车与基础设施等之间的通信，也为实现协同驾驶和智能交通管理提供了技术支撑。而支撑车辆行动的决策系统则依赖于大量数据的实时处理，通过深度学习和强化学习算法对车辆行驶环境进行解析，预测其他交通参与者的行为，并制定最佳驾驶策略。决策系统的高效运作离不开云平台 and 大数据技术的支持，通过不断更新的训练模型，车辆可以在复杂的道路环境中自适应调整，提高行驶安全性和效率。因此，消费者购买自动驾驶汽车后，往往还需要订阅软件更新、远程监控和数据支持等服务，这些服务特性使得自动驾驶汽车在交易中也具备服务属性。正因为如此，在国际贸易规则适用上，自动驾驶汽车面临着商品贸易与服务贸易双重监管的困境，如何在关税、技术标准、责任认定等方面找到平衡成为关键问题。

三、国际贸易规则：智能汽车适用的困境

智能汽车这一新兴技术产品不仅改变了传统汽车产业的生产和使用方式，也给国际贸易规则和法律监管带来了新的挑战，WTO 法监管框架下和区域贸易协定下尚无直接适用具有混合特性的智能汽车的贸易规则，同时针对划分为服务类别的全自动驾驶汽车的国际服务贸易规则规制上存在疏漏。

（一）智能汽车纳入WTO贸易框架的意义

智能汽车的双重属性在国际贸易领域引发了诸多争议。传统的贸易规则如《1994 年关税与贸易总协定》（General Agreement on Tariffs and Trade, 简称 GATT）和《服务贸易总协定》（General Agreement on Trade in Services, 简称 GATS）是规制有形商品和无形服务的两架支柱，但智能汽车同时具备两者的特征，这导致适用约束各国监管措施的国际贸易规则的不确定性。

智能汽车双重属性使其被归于货物或服务的不同而影响成员方的义务，相对比 GATT 相比 GATS 的自由化程度更高，成员方相对承担更严格的义务。⁶若汽车企业以整车销售为主，则主要采用传统的货物贸易模式，按照货物销售适用关税征收和产品认证的流程和规则，并适用 GATT 中贸易待遇制度，比如最惠国待遇、国民待遇、特殊差别待遇，平等地适用于所有国际货物贸易以保证自由市场准入。同时，各成员方采取的规制智能汽车的措施同时受到《原产地规则协定》（Agreement on Rules of Origin）《技术性贸易壁垒协定》（Agreement on Technical Barriers to Trade, 简称 TBT 协议）《反倾销协定》（Anti-dumping Agreement）《补

⁶ Fiona Smith and Lorna Woods, *A Distinction without a Difference: Exploring the Boundary Between Goods and Services in the World Trade Organization and the European Union*, 24 Yearbook of European Law 463, 465(2005).

贴和反补贴措施协定》(Agreement on Subsidies and Countervailing Measures, 简称 SCM 协定)等多边货物贸易协定的约束。

但越来越多的汽车企业开始采用“车+服务”打包销售的方式,即消费者购买车辆后,还可获得持续的软件更新服务、数据流量支持和远程监控服务。此种交易模式下,部分收入来自软件服务订阅或后续增值服务,因此,此种模式的智能汽车国际贸易适用 GATS 服务贸易规则,成员方在 GATS 下承担的义务同时却决于其在承诺表中具体承诺的受该协定约束的服务,同时涉及跨境数据流动⁷、服务质量监管等问题。这种双重模式不仅使交易结构更加复杂,也在国际贸易协定中产生争议,给分类和适用规则带来挑战。同时,监管难题还体现在智能汽车持续更新的技术上,自动驾驶软件的不断迭代和数据服务的实时性,使得车辆的使用寿命内不断享受服务更新,监管机构需要不断修订相关政策,保持对新兴技术的适应能力。

(二) WTO框架下国际贸易规则适用困境

WTO 是最直接的、公开的和透明的规制全球贸易的法律框架,其中, GATT 和 GATS 是 WTO 协定中两个实体性核心法律文件。GATT 位于 WTO 框架附件 1A 多边货物贸易协定下,其规制的是货物贸易;GATS 规制的贸易类型为服务贸易,并且 WTO 判例法表明两者之间存在界限,⁸但是尚不明朗的是两者之间的区分标准。在此基础上,部分具有混合属性特征的智能汽车又如何在 WTO 框架下规制。

WTO 相关案件的专家组和上诉报告中表明两者之间是存在的界限的,但是两者的区分是基于产品本身的固有特性还是规则的运作环境尚不明确。在加拿大——汽车产业措施案(Canada - Certain Measures Affecting the Automotive Industry)中,专家组的分析直接侧重于成员方违反规则的类型,而上诉机构则主张首先应当评估对受该措施影响的贸易类型,然后再分析其违反的实体性规则类型。“墨西哥——影响电信服务的措施案”(Mexico—Measures Affecting Telecommunication Services)中,专家组讨论墨西哥的电信竞争规则和实践是否违反了其在 GATS 下的承诺,即分析涉案的电信服务是否是跨境提供?本案专家组直接分析涉及的电信服务是否属于 GATS 下四种模式的服务提供之一。在美国——赌博和博彩服务案(United States—Gambling and Betting Service)中,专家组主张采用产品的分类表来确定其构成货物或服务贸易, GATT 采用商品名称及编码的协调制度(Harmonized Commodity Description and Coding System, 一般称为 Harmonized System, 简称 HS)对国际贸易商品归类。GATS 采用服务部门分类清单(Service

⁷ 徐子森:《智能网联汽车数据处理的法律规制:现实、挑战及进路》,载《兰州大学学报(社会科学版)》2022 年第 2 期,第 103 页。

⁸ Shin-yi Peng, A New Trade Regime for the Servitization of Manufacturing: Rethinking the Goods-Services Dichotomy, 54 Journal of World Trade 699, 2020 p.710-711.

Sectoral Classification, 又称 W/120), 其囊括了 GATS 所管辖的所有服务。⁹一旦产品被分类为相关代码就会根据代码所在的位置而被归为货物或服务, 这种分类方法是根据产品的固有特征来划分的, 具有一定的主观性和固定性。

智能汽车作为新兴科技产品, 其属性跨越了商品与服务的界限而具有两者的混合特征, 其兼具商品的有形特征和服务的无形特征, 这使得其在现行国际贸易体系下的分类变得复杂, 从而使得各成员方针对智能汽车的国内措施在现行国际贸易规则下难以得到精准规制。GATT 适用于“监管货物贸易的措施”¹⁰, GATS 的适用范围是“影响服务贸易的措施”¹¹。GATT 对“货物”以及 GATS 对“服务”并无直接明确的定义。¹²但是对 GATS 而言, 其根据服务贸易的提供形式, 将“服务贸易”分为跨境提供、境外消费、商业存在和自然人存在¹³的四种提供形式。其强调贸易的交易方式而非其固有特征, 从而“集中交易的经济要素或法律形式来界定定义的范围”¹⁴。同时, 服务是指“任何部门的任何服务”, 但将“行使政府职权时提供的服务”和“空运服务及与空运服务的提供直接相关的服务”排除在外。如果采用现代分类代码制度, 车辆受协调制度 (HS) 的规制而被划分到货物贸易, 除此, 代码中没有为自动驾驶车辆设置特殊分类, 其将与传统车辆一样被识为货物而受到 GATT 的规制, 这显然是不合理的, 因此对当前如何规制具备货物和服务混合属性的自动驾驶汽车的规则适用是不明朗的。

四、国际法路径：综合因素考量下的分类规制

(一) GATT和GATS的适用顺序

成员方的国内措施可能同时受到 GATT 和 GATS 的约束, 因此无论是针对产品的固有特征还是规则的适用环境, 成员方规制智能汽车的国内措施可能受到服务贸易规则与货物贸易规则的“双重规制”。虽然 WTO 并没有两者适用关系的直接解释文件, 但是根据相关规则解释、上诉机构在相关案件案件中的裁定意见和学者观点可知, 二者关系并没非黑即白的而是具有同时存在的可能性和合理性。

对 GATT 和 GATS 的适用顺序可以根据规则文本的解释入手, 对国际规则的解释要遵守《维也纳条约法公约》(Vienna Convention on the Law of Treaties, 简称 VCLT) 有关国际规则解释的要求, 以及“规则起草者的基本利益和确保规则

⁹ 汪震、刘萍：《世贸组织服务部门分类问题》，载《国际经济合作》2013 年第 8 期，第 13 页。

¹⁰ “measures which regulate trade in goods only”。

¹¹ GATS Article I.1: “This Agreement applies to measures by Members affecting trade in services.”

¹² 陈安主编：《国际经济法学新论》（第 5 版），高等教育出版社 2020 年，第 295 页。

¹³ GATS 第 1.2 条：就本协定而言，服务贸易被定义为提供服务：（a）自一成员领土向任何其他成员领土提供服务；（b）在一成员领土内向任何其他成员的服务消费者提供服务；（c）一成员的服务提供者通过在任何其他成员领土内的商业存在提供服务；（d）一成员的服务提供者通过在任何其他成员领土内的自然人存在提供服务。

¹⁴ Fiona Smith and Lorna Woods, *A Distinction without a Difference: Exploring the Boundary Between Goods and Services in the World Trade Organization and the European Union*, 24 Yearbook of European Law 463, 479(2005).

本身持续合法性需要”¹⁵。GATT 和 GATS 作为 WTO 规则体系的组成部分，都属于 VCLT 下的多边国际公约，受到 VCLT 的规则约束。从文义解释上看，GATT 适用于规制货物贸易的国家措施，而 GATS 则适用于规制服务贸易的国家措施，由于规则制定之处，货物贸易和服务贸易具有较为明显的界限，两者的选择适用比较容易判断。当融合货物和服务特征的新型产品的出现，在规则文本中也没有排斥两者合作适用的可能。同时国际规则的解释还要遵循“有效原则”，即“与其无效，不如使之有效”，该原则旨在约束规则的不合法地扩大解释并不排斥对同一规制措施的 GATT 和 GATS 的同时适用可能性。在“阿根廷——金融服务案”（Argentina-Financial Services）中，上诉机构的报告中指出，GATS 不仅涵盖直接“管理”或“监管”服务贸易的措施，也涵盖任何其他类型的“影响”服务贸易的措施，即使是为商品贸易涉及的措施，此类措施包括涉及与特定货物有关的服务或与特定货物一起提供的服务的措施。¹⁶同时，美国——大型民用飞机案（United States – Measures Affecting Trade in Large Civil Aircraft, DS535- 原件；上诉裁定 WT/DS316）上诉机构指出“与商品相反，服务的典型特征包括其非物质的、看不见的、不可储存的和短暂的性质；服务通常是同时生产和消费的，而商品不是，¹⁷……可能很难将商品和服务分开”。所以，GATT 和 GATS 在事实上存在着适用范围的交叉，成员方的规制智能汽车的国内措施可能恰好落在两者的适用范围交叉领域内。¹⁸

（二）智能汽车的类型化与国际规则适用

WTO 规则框架下成员方规制智能汽车的国内措施应该是适用 GATT 还是 GATS？下文主张从“物理特征——贸易经济——法律规则”三个层面入手，即考量产品的通常含义和特征，全球智能汽车技术发展状况和国际贸易活动，GATT 和 GATS 的目的宗旨以及成员方国内措施的法律内容性质，以智能汽车的类型化为基础针对这个问题做出相对明确的回答。

根据《布莱克法律英语词典》的定义，“货物”（goods），是指“除货币以外的有形或可移动的个人财产，特别是贸易物品或商品物品（articles of trade or items of merchandise）”并且还是“有价值的”；“服务”（service）是指“以人力、技能或建议等形式存在的无形商品”。也就是说，对“货物”来说，其本身有形的和可移动

¹⁵ Fiona Smith and Lorna Woods, *A Distinction without a Difference: Exploring the Boundary Between Goods and Services in the World Trade Organization and the European Union*, 24 Yearbook of European Law 463, 463(2005).

¹⁶ Appellate Body Report, Argentina-Financial Services, para 7.100, 54.

¹⁷ Appellate Body Report, United States-Measures Affecting Trade in Large Civil Aircraft, WT/DS353/AB/R, adopted 23 March 2012 (Large Civil Aircraft-2nd Complaint), para. 1295. In U.S.-Large Civil Aircraft, point out that “as opposed to goods, typical features of services include their immaterial, invisible, intangible, non-storable, and transitory nature; services are usually produced and consumed simultaneously, while goods are not.” The Appellate Body further stressed that “it may be difficult to separate goods from services, for instance where services are an input or processing step in the production of goods.”

¹⁸ （1）刘子平：《论 GATT 与 GATS 项下义务的累加性》，载《华东政法大学学报》2015 年第 1 期，第 35-36 页。

的,无论是传统的功能汽车还是具备自动驾驶功能的智能汽车在物理上都是有形的和可移动的。而“服务”是无形的,功能汽车不存在无形的物理属性,但是对需要高级自动驾驶功能的智能汽车,其智能驾驶服务可能是无形的。

由于智能汽车的发展依赖人工智能、雷达、激光雷达等高科技软件和硬件支撑,所以智能汽车的国际贸易格局相比较功能汽车的国际贸易格局更依赖各成员方的技术发展。全球智能汽车的发展格局上看,中美是具有竞争力的两级,中方的智能汽车车企有小米、华为、比亚迪等,美方的智能汽车车企以特斯拉(Tesla)、字母控股(Alphabet)等公司为代表,相比较而下,欧盟地区以及韩国、日本、印度、澳大利亚等成员方主要依赖进口。智能汽车这种具有明显不均衡性质的国际贸易格局在一定程度上会影响智能汽车的国内规制措施,成员方利用国内规制措施加强智能汽车发展竞赛抑或为促进国内智能汽车企业的发展树立规则壁垒,从而影响国际贸易的自由化,同时影响WTO下对国内措施的约束规则发展。智能汽车的不对称的国际贸易格局最终要以GATT和GATS实现贸易自由化的宗旨和目标为依归。存在些许差别的是,GATS项下的义务相对GATT来讲不具有普遍适用性,其需要结合各个成员国的承诺,以通过服务贸易自由化的手段促进所有贸易方的经济增长和发展中国家的发展,并最终实现贸易自由化。

同时,在实践中,由于智能汽车的发展尚未真正的达到L4-L5级别的程度,实现真正的无人驾驶还有一段距离,当前成员方已出台了相关的国内措施以促进和规范其发展,NHTSA在规则实践中重点关注智能驾驶汽车的安全和技术标准,将自动驾驶系统纳入机动车安全标准框架。中国工业和信息化部、公安部、交通运输部印发了《智能网联汽车道路测试与示范应用管理规范(试行)》(工信部联通装(2021)97号),该文件规定智能网联汽车道路测试和部署的条件,强调技术标准和安全要求,反映了商品贸易的监管模式。国际组织,比如联合国欧洲经济委员会(United Nations Economic Commission for Europe,简称UNECE)也在制定自动驾驶车辆的法规,重点关注车辆的技术标准,其属于商品监管框架。¹⁹总之,在实践规则中,成员方和国际组织规制智能汽车的规则措施偏向于商品监管框架。

本文探讨了智能汽车在国际商品贸易与服务贸易规则适用上的困境,并分析了各国法律与现实实践,提出了国际法应对策略。未来,需要进一步研究技术演进对国际贸易规则的影响,推动全球贸易法律体系适应科技发展的新需求。

¹⁹ UNECE is driving progress on Autonomous Vehicles, UNECE(2014), <https://unece.org/transport/road-transport/automated-driving-0>.

Intelligent Vehicles: international rules on trade in goods or services for regulatory measures?

ZHANG Sumin

Abstract: The rapid development of artificial intelligence (AI) technology is profoundly affecting the global economy, in which intelligent vehicles, as a typical representative of the deep integration of AI and automobile manufacturing, are accelerating commercialization worldwide. However, under the legal framework of international trade, should the measures of each country to regulate smart cars be bound by GATT or GATS under the framework of WTO law? The choice of a different binding agreement will have a direct impact on the application of other rules, including tariffs, intellectual property protection, the division of liability and dispute settlement mechanisms, as well as other rules after the classification is determined, which is related to the classification of intelligent vehicles as “goods” or “services” under international trade law. Starting from the three levels of “physical characteristics - trade economy - legal rules”, i.e., considering the usual meaning and characteristics of the product, the state of development of global smart car technology and international trade activities, GATT and GATS, the development of the smart car technology and international trade activities, the development of the smart car technology and international trade activities. In other words, taking into account the general meaning and characteristics of the product, the global development of smart car technology and international trade activities, the purpose of GATT and GATS, as well as the nature of the legal content of the domestic measures of the members, we can provide a relatively clear answer to this question based on the typology of intelligent vehicles.

Keywords: intelligent vehicles; self-driving cars; digital products; international trade rules; artificial intelligence.

人类命运共同体视角下的人工智能全球治理问题研究

陈苏¹

摘要：人工智能全球治理是国际社会通过政策协调、规则制定及多边合作，对人工智能技术的研发、应用及跨国影响进行系统性治理，国际社会目前已初步形成多元主体治理格局，且治理覆盖领域不断扩大。但仍存在治理机制共识缺失、治理标准不一致、治理经验滞后、治理监管机构不健全等多种问题。因此有必要建立有效的人工智能治理体系，要确保治理体系与科技发展同步。在此背景下中国提出人类命运共同体理念，通过构建多元的协同治理体系、完善的制度价值体系、透明的监管治理体系，破解全球治理的深层次困境，为人工智能全球治理注入新的活力。

关键词：人工智能 全球治理 现实困境 人类命运共同体

一、问题的提出

人工智能技术作为模拟人类智能行为的前沿技术体系，其核心目标在于开发具备复杂认知功能的智能系统。其快速发展驱动全球技术政治的深刻变革，深刻改变国家安全与国际和平的基本范式。²该技术主要包含两大方法论体系：基于符号逻辑的可解释性推理系统与基于数据分析的关联性学习系统，二者正呈现深度融合趋势。在技术实现层面，人工智能系统通过模拟人类感知、决策和行动机制，已形成涵盖自然语言处理、计算机视觉、智能机器人等多元应用场景的技术集群。当前人工智能技术已深度渗透农业现代化、智慧医疗、智能制造、社会治理等关键领域，引发全球范围内的技术革新浪潮。国内外围绕人工智能治理已取得阶段性成果，但在动态发展的技术生态中仍存显著理论缺口，尤其缺乏对人工智能全球治理中国角色及人类整体命运的关注。³现有治理框架呈现明显的地缘技术特征：西方国家依托技术先发优势，将价值伦理体系与算法霸权相绑定，试图构建具有排他性的技术治理范式。这种治理模式在机器伦理主体性界定、人机权责边界划分等核心议题上尚未形成突破性进展。值得注意的是，人工智能技术外溢已突破传统主权边界，其资源配置失衡、算法偏见扩散等负外部性效应正演变为全球性公共治理难题。在第四次工业革命纵深发展背景下，人工智能技术的指数级演进正重塑国际地缘格局与数字文明形态。技术双重效应带来的战略机遇

¹ 陈苏,女,24岁,大连海事大学,硕士研究生,研究方向为海商法,联系电话:13836087616,电子邮箱:cs13836087616@163.com.

² 参见鲁传颖:《人工智能:一项战略性技术的应用及治理》,载《人民论坛》2024年第1期.

³ 参见高奇琦:《全球善智与全球合智:人工智能全球治理的未来》,载《世界经济与政治》2019年第7期.

与系统性风险，迫切要求国际社会构建具有约束力的跨国治理框架，现有治理存在明显结构性缺陷。在此背景下，人类命运共同体理念为破解治理困境提供了创新思路，该理念既体现了中国作为负责任大国的技术伦理观，也为建立公平包容的全球人工智能治理秩序提供了实践路径。本文从人类命运共同体的出发视角，进一步研究全球人工智能治理的现状与困境，推动形成具有广泛共识的全球人工智能治理方案。

二、人工智能全球治理的内涵

（一）人工智能全球治理的概念

人工智能全球治理是国际社会通过政策协调、规则制定及多边合作，对人工智能技术的研发、应用及跨国影响实施系统性监管的治理框架。其核心目标在于平衡技术创新与伦理责任，通过建立道德规范、法律约束和技术标准，解决人工智能引发的安全风险、数据隐私及社会公平等全球性挑战，最终实现技术赋能与社会可持续发展的协同。

该治理体系包含三个维度：在主体层面，涵盖国家政府、国际组织、行业联盟及公私利益相关方的协同参与；在客体层面，既涉及算法透明度、数据安全等技术本体的治理，也需规范技术应用衍生的就业冲击、军事化风险等次生问题；在实施路径上，则依托伦理准则构建、关键技术突破（如可解释 AI）和跨境立法协同的三重机制。值得注意的是，其治理原则强调程序正义与多方协商，既包含国际条约等正式制度，也涉及行业共识等非正式规范，呈现出网络治理、环境治理等领域相似的跨域协同特征。

相较于传统治理模式，人工智能全球治理更具动态适应性，要求技术治理与地缘政治、经济格局深度耦合。当前治理实践已从单一的技术标准制定，逐步演变为涵盖数字主权博弈、全球公共产品供给的复合型治理架构，这反映出技术革命对国际关系范式的重构作用。⁴

（二）人工智能全球治理的发展现状

1、初步形成多元主体治理格局

自 2016 年起，全球范围内已有超过 40 个国家和地区将人工智能技术发展纳入国家战略规划。在治理实践层面，政府与市场协同、公共部门与私营机构协作等复合型治理机制，已然成为应对人工智能挑战的核心路径。当前国际社会正逐步构建起以主权国家为主导，国际组织、产业联盟、科研机构及科技企业等多元主体协同参与的立体化治理网络。

在国家层面，主权国家主导的硬性规制体系持续强化。以强制性法律规范为特征的国家中心治理模式，仍是当前全球人工智能治理的主导形态。主要经济体

⁴ 参见邓子纲：《人工智能的全球治理与中国的战略选择》，载《求索》2020 年第 3 期。

如中美等国，均通过立法建制、政策引导等手段强化对人工智能产业的规划与监管。该模式凭借行政权力的集中行使和强制性法律规范的实施优势，能够实现显著的治理效能。2023 年英国召开人工智能安全峰会，28 个参会国共同签署《布莱切利宣言》，初步确立人工智能安全治理的国际共识。

在国际组织层面，国际组织主导的伦理治理机制逐步完善。经济合作与发展组织、联合国教科文组织等机构在人工智能伦理准则制定方面发挥着关键作用。2019 年《经合组织人工智能原则》首次系统构建了以人为本的可信人工智能框架，2021 年联合国教科文组织发布的伦理建议书则确立了人工智能全生命周期应遵循的核心价值准则。至 2023 年，联合国安理会召开人工智能安全高级别会议，秘书长古特雷斯倡议设立专门国际机构强化全球治理。同年成立的联合国人工智能高级别咨询机构，标志着全球治理机制化建设取得实质性突破。

在企业层面，产业界主导的技术治理体系快速发展。科技企业、研究机构及非营利组织正通过技术标准制定、伦理框架建构等途径，逐步成为治理体系的重要参与方。⁵以亚马逊、谷歌为首的科技企业早在 2016 年即发起“人工智能合作伙伴计划”，迄今已形成涵盖 126 个成员的全球协作网络。2023 年世界经济论坛牵头组建的人工智能治理联盟，进一步拓展了行业协作维度。在技术标准制定方面，IEEE 等专业机构持续完善伦理标准体系，微软、谷歌等科技巨头联合中国计算机学会等组织，共同推进技术规范与伦理准则的国际化进程。政企间的知识共享机制，有助于构建更具实操性的技术标准体系与动态化监管框架。对企业而言，参与治理不仅关乎合规成本控制，更是获取市场准入资质、塑造行业话语权的战略选择。特别是在全球数字经济竞争格局下，头部企业通过主导技术标准制定，可实质性影响国际规则走向。

2、治理覆盖领域不断扩大

当前全球人工智能政策体系正经历结构性变革，呈现出从单一技术监管向全流程治理体系演进的显著特征。智能技术治理的覆盖维度已突破传统边界，延伸至医疗卫生、交通运输、金融科技及教育服务等关键民生领域，构建起涵盖算法研发、数据治理、场景应用及社会效应评估的全流程监管框架。以技术发展周期为观察维度，早期政策聚焦于技术安全防护与战略数据资产保护，而现阶段治理范畴已扩展至技术伦理审查、市场准入规制等领域。技术赋权效应重塑教育资源配置格局，智能教育平台的应用显著降低了优质教育资源的获取门槛。⁶政策工具链贯穿技术研发的数据合规审查到应用场景的市场行为规制，形成覆盖技术生命周期的闭环监管体系，通过外部性内部化机制设计，倒逼市场主体将治理成本

⁵ 参见张发林、杨佳伟：《统筹兼治或分而治之--全球治理的体系分析框架》，载《世界经济与政治》2021 年第 3 期。

⁶ 参见李益斌、李浩洋：《欧美人工智能监管规范比较研究》，载《当代世界与社会主义》2024 年第 5 期。

纳入经营决策。我国 2023 年颁布的《生成式人工智能服务管理暂行办法》通过设立研发备案、数据溯源等制度，促使企业将算法偏见矫正、数据隐私保护等社会成本转化为内部管理要素，覆盖技术研发、应用场景和数据管理等多个领域。该治理转型具有双重驱动逻辑：技术层面，算法黑箱、数据跨境流动等技术特性要求建立跨国协同治理机制；制度层面，各国正通过“规制竞争”探索最优治理路径。这种多维治理不仅需要平衡创新激励与风险防控，更需构建政府主导、企业担责、社会监督的协同共治体系，以实现技术红利最大化与社会风险可控化的动态平衡。

（三）人工智能全球治理的必要性

人工智能技术的全球化渗透正在重塑世界运行的基本逻辑。随着智能系统深度嵌入全球产业链，其影响已突破经济领域边界，向社会治理、国际关系等复杂领域持续扩散。这种技术扩散在释放生产力的同时，也带来了前所未有的治理挑战，构建有效的全球治理框架成为国际社会的紧迫课题。

当前人工智能引发的复合型风险正在多个维度显现：技术层面存在数据滥用与算法失控的隐患，经济领域面临自动化替代引发的就业结构震荡，军事安全方面遭遇自主武器系统对战争伦理的冲击。人工智能技术的快速发展已引发了社会各界的广泛担忧，而构建人工智能全球治理体系正是对这些担忧的有力回应。回顾世界历史，每一次世界大战或工业革命都迅速重塑了社会秩序与国际格局，而以人工智能为核心的第四次技术革命，其影响力更是前所未有的颠覆性。对这场变革的未知与恐惧，若不加以合理引导，可能会阻碍科技的进步。而有效的人工智能治理体系，正是缓解这种恐惧、控制其蔓延的最佳途径。正如核科学技术曾一度令人心生畏惧，但通过国际社会的共同努力，以核不扩散机制为核心的核技术治理体系得以建立，不仅避免了核技术引发的人类灾难，还在医疗等多个领域推动了生产力的飞跃。科技本身并无善恶，关键在于如何使用。我们既要积极探索科技的无限潜力，也要确保治理体系与科技发展同步，以降低潜在灾难的风险。因此，面对人工智能技术的快速发展，既要拥抱其带来的机遇，也要警惕其潜在的风险，通过构建全面、有效的人工智能全球治理体系，确保科技发展与人类社会的和谐共生。

三、人工智能全球治理的困境

（一）治理机制面临共识困境，排他治理突出

尽管各国已形成技术同步发展的原则性共识，但国际社会在人工智能全球治理领域面临深刻的共识建构危机，在治理路径选择、权责分配等实质层面仍存在显著的制度性共识困境。当前国际人工智能治理体系面临多种结构性矛盾。首先表现为制度性协同缺失困境。尽管联合国系统内设机构如教科文组织聚焦技术伦

理框架建构，世界知识产权组织专司算法专利确权，但跨机构协调机制的缺位导致治理呈现职能碎片化特征。这种机构间权责离散状态，使得全球层面难以形成统合技术标准、市场规制与伦理审查的协同治理架构。当前全球治理体系既缺乏具有约束力的多边协定，又缺少包容性对话平台。技术先发国家更倾向于将治理机制作为战略竞争工具，而非全球公共产品。这种“伪多边主义”实践不仅加剧数字鸿沟，更导致人工智能军事化应用、算法歧视等跨国风险处于监管真空。其次，技术先发国家正通过构建排他性规制同盟强化竞争优势，典型表现为美欧等技术经济体频繁运用出口管制清单、人才流动限制等非对称手段，试图将技术霸权转化为规则制定权。这些国家普遍采取双重标准策略——对内构建技术发展规范体系，对外推行排他性协商机制。2019 年经合组织发布的人工智能政策指南即为明证，其五项原则与美国“维持人工智能领导地位”的行政令高度耦合，刻意排除中国等新兴经济体参与。⁷这种规制政治化操作导致国际规则体系出现“制度性裂变”，催生出互斥的算法认证标准与数据跨境流动规则。这种制度性割裂引发的负外部效应呈现多维度扩散特征：技术层面，标准互认障碍加剧数据跨境传输的安全隐患；经济层面，规制性壁垒扭曲全球产业链布局，引发人工智能产品市场的非正当竞争；创新层面，技术民族主义政策严重制约知识共享与技术扩散效率。更值得警惕的是，这种“规则武器化”态势可能诱发治理体系的系统性失灵，使本应促进技术向善的治理框架异化为地缘竞争工具。⁸

（二）技术规范多元发展，缺乏统一的治理标准

人工智能全球治理的演进历程呈现出显著的范式变迁。这一变迁的核心特征体现在治理标准与技术规范的多元化发展态势上。这种多元格局的形成既为各国提供了因地制宜、灵活应对人工智能技术风险的制度弹性，又使得不同规则体系间的兼容性困境日益凸显，成为亟待解决的全球性难题。在全球治理的广阔舞台上，主权国家、区域性组织、跨国企业等多元主体通过各自差异化的路径积极参与到人工智能全球治理规则的建构中。这种多轨并进的治理模式，虽然在一定程度上促进了全球治理体系的丰富与多元，但也在客观上加剧了标准体系的碎片化，使得不同规则体系间的协调与统一变得愈发困难。从国际组织的实践来看，这种标准制定的多元化趋势尤为明显。国际电信联盟（ITU）作为联合国下属的专门机构，始终聚焦于通信基础设施的技术参数标准化工作，致力于推动全球通信技术的互联互通。而国际标准化组织（ISO）则更加注重构建质量管理体系，以提升全球产品和服务的质量水平。电气与电子工程师协会（IEEE）则另辟蹊径，侧重于制定伦理准则，为人工智能技术的行业应用提供道德指引。此外，欧盟的《人

⁷ 参见朱荣生、陈琪：《美国对华人工智能政策：权力博弈还是安全驱动》，载《和平与发展》2022 年第 6 期。

⁸ 参见叶淑兰、李孟婷：《全球人工智能治理：进展、困境与前景》，载《国际问题研究》2024 年第 5 期。

工智能法案》也独树一帜，强调决策逻辑的社会可理解性，旨在确保人工智能技术的决策过程符合人类的价值观和道德标准。这些国际组织在标准制定上的差异化实践，本质上反映了技术理性与价值理性之间的张力。技术理性追求的是技术的先进性和效率性，而价值理性则更加关注技术的社会影响和人类福祉。这种张力在人工智能全球治理的实践中表现得尤为突出，使得不同国家和组织在标准制定上产生了分歧和争议。

在治理实践层面，数据隐私规制成为了标准分歧的典型场域。不同国家和地区在数据隐私保护上采取了截然不同的策略。中国通过颁布《数据安全法》，构建了以国家安全为导向的垂直管理体系，强调数据的战略价值和国家安全的重要性。而美国则采取了一种以行业自律为主的分散治理模式，鼓励企业通过自我监管来保护用户隐私。欧盟则通过实施 GDPR（通用数据保护条例），确立了个人权利本位的严格监管范式，旨在保障个人数据的安全和隐私权益。这种规制差异不仅源于不同法律传统的分野，更深层次地折射出数字主权认知的差异。中国强调数据作为新型生产要素的战略价值，致力于推动数字经济的发展和创新。欧盟则更加侧重基本权利的保护，将个人隐私视为不可侵犯的基本人权。而美国则体现了科技资本主义的发展逻辑，将技术创新和市场竞争视为推动经济发展的关键动力。这种数字主权认知的差异，使得不同国家和地区在人工智能全球治理中采取了不同的策略和路径，进一步加剧了标准体系的碎片化和兼容性困境。⁹

（三）治理领域复杂，治理经验滞后

在网络空间内，技术的迭代速度远远超越了现行世界运行规则的更新步伐。正如有学者所言，当前全球化、信息化与网络化的迅猛进展正深刻重塑全球经济与社会结构，世界经济政治、社会形态及文化层面均经历着前所未有的全面转型。¹⁰技术的创新性与复杂性无疑为人工智能的全球治理提出了前所未有的挑战。人工智能技术的治理议题在国际社会尚处于伴随技术发展而不断扩散的过程中，众多新兴挑战与风险往往超出了先前的预估范畴，传统的治理结构在应对人工智能技术的开放性与不确定性时显得捉襟见肘。人工智能的发展展现出极强的开放特质，任何一名程序员或任意一家公司均可无门槛地投身于人工智能程序的开发与应用之中。¹¹人工智能治理几乎覆盖全部社会领域，全面覆盖的监管不仅会扼杀人工智能技术的内在发展动力，还会阻碍人工智能企业的技术深化与拓展，从而令人工智能技术丧失其核心意义。然而，从风险管控的视角出发，必要的监管至关重要。传统的治理结构通过局域性或阶段性地对特定领域实施统一监管的方式，

⁹ 参见：韩永辉、周港隽、徐翠芬：《人工智能全球治理的现状、困境和中国路径》，载《特区实践与理论》2024年第6期。

¹⁰ 参见卢静：《当前全球治理的制度困境及其改革》，载《外交评论(外交学院学报)》2014年第1期。

¹¹ 参见贾开、蒋余浩：《人工智能治理的三个基本问题：技术逻辑、风险挑战与公共政策选择》，载《行政管理》2017年第10期。

显然不适用于人工智能治理领域。如何在维护人工智能技术开放性的同时，实施有效的风险监管，这一全新治理领域向全球治理方式提出了严峻挑战。

鉴于人工智能技术的广泛适用性，其在众多领域均发挥着不可或缺的作用。例如，在金融领域，神经网络、专家系统、混合智能以及支持向量机等人工智能技术模型正被积极运用，以辅助金融工程师进行更为精确的分析与数据处理。这些技术手段显著提升了经济工作的效率与成果，但与此同时，也加剧了金融领域的风险与挑战。市场交易频率的提升意味着大量数据的采集与使用，而这些数据的采集过程在合法性与合理性方面存在诸多疑问。此外，一旦机器决策的金融过程中出现误差，可能导致整个金融系统的崩溃，带来无法估量的损失。类似的风险同样潜藏于国际贸易、知识产权、海关检验等多个领域。这意味着相关国际准则必须进行调整，以适应人工智能新技术的诞生。然而，如此大规模冲击国际准则的现象，在全球治理的历史经验中实属罕见。从国际社会的普遍实践来看，即便在单一领域中修正与调整国际准则已属不易，要对多个领域的国际准则进行修正与替换，需协调各方利益集团的矛盾与诉求，这无疑是一个庞大且艰难的过程，缺乏现成的经验可供借鉴。如何在治理进程中迅速且有效地应对人工智能技术引发的国际准则冲击，实现规范与技术的同步更新，是决定人工智能全球治理成效的关键所在。国际社会对于这一全新的治理领域治理经验几乎为零。例如，如何追溯不断流通的数据以确定其权属关系；如何在法律层面平衡算法在人工智能技术发展中的开放性与规范性；以及面对人工智能技术导致的失业问题，应如何实施治理？在人工智能时代到来之前，这些问题并不存在或并不突出；但随着人工智能的快速进步与广泛应用，它们的重要性日益凸显。

（四）人工智能治理监管机构不健全

人工智能技术在全球范围内的快速迭代与行业渗透，正面临着监管架构滞后于技术发展的困境。当前国际社会在监管主体构建层面呈现出“机构缺位”与“权责失焦”的双重矛盾。从组织架构维度观察，主要经济体普遍缺乏具有法定职权的专门监管机构，既有的标准制定体系与风险防控机制呈现显著的职能分散特征。美国在人工智能治理领域构建了多层次的制度框架，但其机构设置存在明显的功能碎片化问题。国家标准与技术研究院（NIST）主导制定的风险管理框架虽具技术权威性，却因缺乏强制执行力难以形成有效约束。其国家人工智能安全委员会（NSCAI）作为跨领域咨询机构，其政策建议与国防系统新设的人工智能安全中心存在职能重叠，这种多头管理格局导致战略资源分散与监管效能弱化。最新成立的人工智能安全研究所（AISI）虽被赋予风险评估职能，但其依托 NIST 的建制模式仍未突破技术标准机构与行政执法部门间的制度壁垒。¹² 欧盟《人工智能法案》的立法进程虽开创了全面立法规制的先例，但在监管主体建设方面仍显滞

¹² 参见王威：《〈布莱奇利宣言〉：人工智能国际合作监管的新起点》，载《服务外包》2023 年第 12 期。

后。法案虽原则性要求成员国设立监管机构，但除西班牙率先成立人工智能监管局（AESIA）外，多数成员国仍处于制度空白状态。¹³这种中央立法与地方执行之间的断层，使得欧盟层面难以形成统一的监管合力。中国虽较早成立国家人工智能标准委员会，但其职能定位仍停留在技术标准层面，尚未构建起涵盖研发应用全周期的监管体系。当前监管架构的深层矛盾体现在三个方面：其一，技术标准制定机构与行政执法部门存在制度性区隔，导致技术规范难以转化为具有约束力的监管措施；其二，跨部门协调机制缺失引发监管重叠与真空地带并存，典型表现为国防系统与民用领域各自为政的治理格局；其三，国际监管协作仍停留在原则性共识阶段，主要经济体在风险评估标准、算法审计规范等关键领域尚未建立实质性合作机制。这种制度性缺陷使得人工智能治理陷入“标准先行、执行滞后”的实践困境，加剧了技术风险防控的系统性挑战。

四、构建人类命运共同体：人工智能全球治理的中国方案

在全球化时代，人工智能产生的许多政策治理问题具有全球意义。对于全人类来说，集体共同行动解决全球人工智能治理所面临的现实问题极为重要。中国作为全球性大国，是人工智能领域的领跑者之一，中国所倡导的人类命运共同体不仅符合人类共同发展的内在逻辑，同时也为破解全球人工智能发展现实困局提供了可行路径。

（一）构建多元的协同治理体系

为防止人工智能治理领域寡头垄断地位的形成，需要充分发挥主权国家作用，推动多边主义治理。¹⁴全球人工智能的蓬勃发展触及众多利益相关方，这一现状内在地要求全球人工智能治理框架下的多方主体协同并进、共同参与。一方面，人工智能的发展根植于一个由多方合作构筑的、覆盖全球的协同治理网络。政府作为关键角色，需在私营与公共部门间铺设坚实的桥梁，确保合作与竞争并存的国家间格局得以稳健形成。私营部门则需主动出击，创造一个既能激励创新又能确保技术责任担当的良性环境，为新兴技术的构思、研发及应用铺设道路。同时，国际组织在凝聚国际共识、推进共同愿景及提升治理效能方面，扮演着不可或缺的桥梁与催化剂角色。另一方面，全球人工智能治理的共同体不仅限于发达国家的参与，其广度和深度更需发展中国家的积极融入。从全球视野审视，人工智能治理触及全人类的共同福祉，这要求我们务必关注技术的均衡发展，尤其是在发展中国家与发达国家之间构建更为平等的对话平台，通过缩减“数字鸿沟”与“能力差异”，确保人工智能技术能惠及全球每一个角落，推动其在全球范围内的有益应用。

¹³ 参见焦授松：《人工智能治理：欧美新合作还是新博弈》，《光明日报》，2023年12月22日。

¹⁴ 参见陈伟光、袁静：《人工智能全球治理：基于治理主体、结构和机制的分析》，载《国际观察》2018年第4期。

人类命运共同体的理念，其核心在于将全人类的命运紧密相连，将每一个“类存在”个体纳入这一宏大的共同体之中。¹⁵面对全球人工智能这一公共议题，我应凸显人类命运的休戚与共，不同主体间既要基于共同利益构建规范共识，也要深刻认识自身在全球集体角色与未来命运中的定位。在命运共同体的宏观框架下，全球人工智能治理应以开放与合作作为核心理念，充分激发多元利益主体的积极性，真正构建一个涵盖国家政府、私营部门、国际组织及公民社会等多方协同参与的治理体系。通过积极投身全球人工智能治理议程，联结并优化现有的治理平台与机构，逐步建立和完善更为宽广的国际合作网络。人类命运共同体理念致力于共同协商合作，中国参与和推动全球人工智能治理的重要内容就是全球多层次的协同合作。在国际实践中，中国秉持人类命运共同体理念，坚决抵制关键少数国家对人工智能领域的垄断和独霸，突出强调与所有国家的协同治理合作，特别是发展中国家的技术共享。2019年，中国所发布的《新一代人工智能治理原则》鼓励国际组织、政府部门、研究机构、教育机构、企业、社会组织和公众之间的协调和互动，促进人工智能的发展和治理；同时强调通过开展国际对话与合作，推动就国际人工智能治理框架、标准和规范达成广泛共识。这些原则将中国特色与全球命运相结合，向世界彰显了人工智能各利益主体“全球共治”的中国方案。

（二）构建完善的制度价值体系

人工智能治理本质上是技术文明与制度文明交融共生的系统工程，需要构建法律规范、伦理准则、社会价值三位一体的协同框架。中国作为数字文明新形态的探索者，应在制度创新层面发挥引领作用，通过构建动态响应的法律体系，将技术预见性融入立法进程。这要求立法机构建立与算法工程师、伦理学家、社会公众的常态化对话机制，运用数字孪生技术模拟法律实施效果，形成既能适应自动驾驶、生成式 AI 等技术迭代速度，又能坚守人权保障底线的弹性规制框架。在国际合作维度，可倡导建立“分层共识”协商机制：在基础层就数据主权、算法透明度等形成全球最低标准；在应用层允许各国根据文化传统制定特色规范，如东亚国家可侧重智能社会治理，非洲国家聚焦农业 AI 伦理。

技术治理的价值锚点应植根于文明传承与时代创新的辩证统一。“向上的兼容性”体现为将“天人合一”的东方智慧注入算法设计，在智能城市系统中保留市井烟火气；“时代的容涵性”要求构建数字时代的“技术礼治”，通过区块链存证实实现数据确权，使农民也能享有算法红利；“逻辑的展开性”意味着在司法人工智能部署中，既要提升办案效率，更要通过可解释性算法保留法官的自由裁量空间；“思想的开放性”则催生出“伦理沙盒”机制，允许宗教团体、原住民社群等多元主体参与智能医疗伦理审查，用文明多样性校准技术航向。人类命运共同体视角下

¹⁵ 参见郝立新、周康林：《构建人类命运共同体--全球治理的中国方案》，载《马克思主义与现实》2017年第6期。

人工智能全球治理价值引领的终极目标在于塑造“技术向善”的文明形态。这需要将中华民族民惟邦本的政治智慧转化为算法问责机制，例如要求自动驾驶系统在紧急避险决策中优先保护弱势道路使用者；把天下大同的理想追求具象化为全球AI治理基金，资助发展中国家建设本土化伦理审查体系。当智能医疗诊断系统融入中医“治未病”理念，当农业无人机遵循“万物并育”的生态逻辑，技术文明才能真正成为人类命运共同体的数字纽带，在硅基与碳基的共生演进中谱写新的文明篇章。¹⁶

（三）构建透明的监管治理体系

在全球化浪潮汹涌的今天，人工智能技术的迅猛发展既带来了前所未有的机遇，也伴随着一系列复杂且深远的挑战。如何在这一时代背景下，通过精准而有效的监管策略，最大限度地削弱技术革新可能引发的负面效应，成为了人工智能治理领域亟待解决的核心议题。面对监管与创新之间的微妙平衡，全球人工智能治理体系必须充分考量各国在“风险偏好”层面存在的显著差异，这不仅是对治理智慧的考验，更是对国际合作能力的深度挖掘。为了有效应对这一挑战，各国政府需展现出更高的政策灵活性与前瞻性，维持监管与创新之间的动态平衡，确保监管措施既不过度束缚创新的步伐，也不因监管缺失而放任潜在风险滋生，从而避免国家的监管错位成为阻碍人工智能健康、有序发展的绊脚石。同时，尽管各国均希望通过人工智能创新来塑造新的竞争优势，但孤立的、以国家利益为出发点的监管框架显然无法全面应对跨国界的技术挑战。因此，将国家层面的监管实践融入更为广阔的国际机制之中，成为了破解这一难题的关键路径。在此基础上，构建多边人工智能监管合作平台成为了迈向全球统一、透明、可解释的人工智能监管治理体系的必由之路。这一平台不仅能够促使各国在监管实践中扮演“合作监管者”的角色，共同制定和执行高标准、高效率的监管规则，还能确保各国在人工智能领域的“竞争参与者”身份得以彰显，通过公平竞争促进技术的持续进步与创新。这种双重角色的融合，不仅是对传统监管模式的超越，更是对全球治理体系的一次深刻革新。

人类命运共同体理念的提出，与人工智能等新兴技术所带来的社会变革紧密相连，它强调了全球合作与共同发展的重要性。¹⁷特别是中美等科技强国，在引领人工智能技术创新的同时，采取负责任的监管措施显得尤为重要。中国所倡导的人类命运共同体理念，正是基于对人类共同利益和未来命运的深刻洞察，为全球人工智能治理提供了具有前瞻性和指导性的思想资源。这一理念鼓励各国在推进监管细化时，以创新为导向，构建适应新科技革命和产业变革需求的监管体系，

¹⁶ 参见张东冬：《人类命运共同体理念下的全球人工智能治理：现实困局与中国方案》载《社会主义研究》2021年第6期。

¹⁷ 参见冷溶：《更加坚定推动构建人类命运共同体》，《人民日报》，2020年09月01日。

这不仅是对人类命运共同体理念的实践，也是对全球治理体系的积极贡献。习近平总书记多次强调多边主义在全球治理体系变革中的核心作用，这一思想同样适用于人工智能治理领域。各国应携手并进，将国际合作置于治理议程的核心位置，以人类发展和共同命运为着眼点，共同推动人工智能监管体系的构建与完善。中国在这一进程中扮演着积极而重要的角色。2017年7月发布的《下一代人工智能发展计划》，不仅为中国国内的人工智能监管治理提供了初步框架，也为全球人工智能治理体系的建立贡献了中国智慧。¹⁸该计划强调建立公开透明、双层监管结构的监管体系，实现对人工智能全生命周期的有效监管，同时，在国际层面，中国将积极推动国际社会的协同参与，加强共性问题的研究与国际标准的制定，为全球人工智能治理注入新的活力。

五、结语

人工智能技术的快速发展正在重塑人类文明的基本图景，从工业4.0时代的智能生产线到城市治理的算法中枢，从医疗诊断的深度学习系统到金融市场的量化交易模型，这场以数据为燃料、算法为引擎的技术革命已渗透至社会机体的各个领域。然而，在技术跃迁的耀眼光芒背后，暗藏着多重维度的系统性风险：算法黑箱正在侵蚀司法决策的透明度，深度伪造技术动摇着社会信任的根基，自动化浪潮下的结构性失业加剧了阶层分化，而数据殖民主义的新形态正在全球范围内制造着数字时代的“中心-边缘”格局。

当前全球治理体系在应对这些挑战时呈现出显著的制度性滞后。技术更新速度与治理响应效率之间的差距持续扩大，主权国家间的规范竞合导致治理规则碎片化，跨国科技巨头的算法权力与民族国家治理权威形成张力，使得人工智能治理陷入困境，既不能在技术萌芽期预见所有风险，又难以在技术成熟后实施有效规制。在此背景下，中国提出的治理范式展现出独特的文明智慧。以人类命运共同体理念为核心，在技术伦理层面构建起以人为本的价值锚点，在实践路径上倡导渐进包容的演化逻辑，在治理结构上主张多元共生的生态体系。通过制度性开放的新实践，将基础设施互联互通的“硬联通”转化为治理规则互鉴互融的“软联通”，致力于构建数字时代的“新多边主义”。这种新型治理范式既保持对技术主权的必要尊重，又通过共商共建共享机制培育全球公共产品，最终在数字文明共同体中实现技术红利的人类共享，为人工智能全球治理贡献中国力量。

¹⁸ 《国务院关于印发新一代人工智能发展规划的通知》，载于《中华人民共和国国务院公报》2017年第22期。

Application Of Principles of Responsibility to Space Cyberattacks

Constituting International Crimes

Manzoor Hassan, Kuan Yang*

ABSTRACT

Cyber-attacks on space systems can cause widespread chaos, humanitarian calamities, and severe harm to space infrastructure, human lives, and the environment on Earth and in space. Determining whether these attacks are internationally wrongful and whether they reach the threshold of international crimes is a complex challenge that requires a multidisciplinary approach. It needs to merge international space law, cyber law, public international law, and customary norms. The ambiguity in defining space cyber wrongs, the anonymity of attackers, and the overlap of multiple legal frameworks make this determination even more difficult. Current legal frameworks governing space activities, cyber operations, and state responsibility were not intended to handle the intricate nature and unique technical and legal complexities of space cyberattacks. This gap underscores the need to rethink and redefine international responsibility to address these evolving threats better. Therefore, this research explores internationally wrongful acts in the context of space cyberattacks. By analyzing recent instances of space cyberattacks, this research illustrates the practical challenges in attribution and applying legal principles. This research advocates for significant enhancements to the existing international responsibility regime by specifying obligations for preventing and combating cyber threats, including those from private actors. This includes defining what constitutes space cyber wrongdoing, clarifying legal obligations, strengthening the existing legal regime, establishing novel mechanisms for attributing space cyberattacks, and facilitating more precise accountability.

Keywords: Space Cyberattacks, International Crimes, International Responsibility, ARISWA, Space Law.

* Academy of Air and Space Policy and Law, School of Law, Beijing Institute of Technology, No.5 South Zhongguancun Avenue, Haidian District, Beijing, 100081, PR China.

1.Introduction

As humanity ventures further into the cosmos, ensuring the safety of space systems against cybercrimes is essential for the continued peaceful and beneficial use of outer space and human sustenance in space and on Earth. Rethinking international responsibility for these crimes is crucial to safeguard our shared future in space.

Nevertheless, the increasing dependence on cyberinfrastructure in space (hereinafter mentioned as space cyberinfrastructure), including space segments, link segments, and ground segments, is accompanied by the growing hazard of cybercrimes targeting space cyber infrastructures. These infrastructures, including satellite systems, space stations, and other interplanetary communication networks, act as the backbone for essential operations on Earth and space operations¹. These systems, providing smooth communication, navigation, data transmission, and other vital services, have become attractive targets for cybercriminals aiming to disrupt, sabotage, or misuse space infrastructures².

With the increasing dependence on these systems, so does the landscape of potential hazards, ranging from unauthorized entry into satellite systems to manipulation of data transmission via space-based communication systems. The range of potential cyber threats in outer space is wide-ranging and diverse. These potential threats put the integrity and functionality of space missions at risk, jeopardize national security and global stability, and, more importantly, threaten human lives in space and on Earth³. A recent example highlighting the severity of cyber threats to space infrastructure is evident in the Ukraine dispute, which may be regarded as the 1st space-cyber conflict, showcasing the potential and allure of attacking space infrastructure during disputes.

Following the commencement of the Russian invasion, Russia reportedly blocked GPS signals in Ukraine, impairing the Ukrainians' capacity to pinpoint location, navigate, or direct weapons to their targets⁴. Additionally, ViaSat, a provider of fast satellite internet services, experienced a service disruption affecting its Ka-Sat satellite, resulting in disruptions of internet services in 55 countries. It impacted various Ukrainian departments, including the armed forces, police, and intelligence service. The impact began in Ukraine and led to a widespread effect, disrupting 5,800 wind

¹ Vijay Varadharajan and Neeraj Suri, 'Security Challenges When Space Merges with Cyberspace', Space Policy, 67 (2024).

² Brianna Bace, Yasir Gökce and Unal Tatar, 'Law in Orbit: International Legal Perspectives on Cyberattacks Targeting Space Systems', Telecommunications Policy, 48/4 (2024).

³ Roy Peled and others, 'Evaluating the Security of Satellite Systems', 2023.

⁴ Theresa Hitchens, 'Local Russian GPS Jamming in Ukraine Hasn't Affected Us Support Ops so Far', Breaking Defense, 2022.

turbines in Germany and thousands of connections throughout Europe. However, details on the origin of this attack and attribution remain elusive⁵. While physical satellite destruction has not occurred in this conflict, the use of cyberattacks in contra-space warfare has been prominently demonstrated.

Dmitry Rogozin, the leader of the Russian space agency Roscosmos, announced that Russia would respond to any attack on its satellites as a justification for war⁶. This highlights the potential for a cyberattack on space infrastructure to cause a conventional warfare retort. In 2011, the United States published its “International Strategy for Cyberspace”, announcing that the US would treat any hostile action in cyberspace like any other national security risk⁷. Israel, for the first time, took similar action, targeting a building in Gaza in response to alleged cyberattacks by Hamas hackers against Israeli targets⁸, even though not against space assets but Israel's cyber infrastructures.

Additionally, groups reportedly associated with the private military company Wagner are said to have hacked a Russian telecommunications satellite serving the Federal Security Service (FSB) and Russian military units. According to Ukraine's State Security Service (SBU), Russia's GRU is trying to deploy custom malware against Starlink satellites to gather information on Ukrainian troop movements⁹. Given these potential severities of cyber threats to space infrastructures, it is crucial to assess whether such acts could be classified as internationally wrongful, either as crimes of aggression, genocide, war crimes, or crimes against humanity.

Satellites and other space-based assets have become integral to myriad functions, ranging from global communications and navigation to weather forecasting and national security. However, this increased reliance on space infrastructure has also introduced new vulnerabilities, particularly in cybersecurity. Cyberattacks in outer space are a growing concern, as they significantly threaten space assets' safety, security, and functionality. Unlike traditional cyberattacks confined to terrestrial targets, cyberattacks on or from space-based systems can have widespread implications. It can disrupt critical services, breach confidential information, and potentially lead to catastrophic failures of vital space infrastructure¹⁰. Their unique legal and technical challenges compound the complexity of space cyber threats. The existing international legal frameworks relevant to space and cyber operations were not designed to address

⁵ Thomas Rid, ‘Why You Haven’t Heard about the Secret Cyberwar in Ukraine’ (2022).

⁶ Reuters, ‘Russia Space Agency Head Says Satellite Hacking Would Justify War’, Reuters, 2022.

⁷ Executive Office of the President, ‘International Strategy for Cyberspace: Prosperity, Security, and Openness in a Networked World’ (2011).

⁸ ZDNET, ‘In a First, Israel Responds to Hamas Hackers with an Air Strike’, ZDNET.

⁹ “Significant cyber incidents since 2006,” Center for Strategic and International Studies (CSIS).

¹⁰ Logstail, ‘Cyber Security Challenges in Space’ (2024).

the intricate nature of cyberattacks in the space domain. This gap creates a critical need for rethinking and redefining the legal nature of space cyberattacks and the international responsibility and legal accountability for such attacks¹¹.

Understanding the legal nature of cyberattacks in outer space requires a multidisciplinary approach, encompassing international space law, cyber law, public international law, and international customary norms. The uncertainty in defining international space cyber wrongs, the anonymity of space cyber attackers, and the overlapping nature of various legal regimes further complicate the determination of wrongful acts and the attribution of responsibility¹². Attribution of cyberattacks in space is particularly complicated because of the technical complexities of tracing attacks back to their sources. The principles of state responsibility and the mechanisms for attributing cyber activities to state and private actors must be scrutinized and adapted to the context of outer space.

2. Legal nature of cyberattacks in outer space

The growing dependence on space systems for crucial functions like communications and surveillance amplifies the severity of cyberattacks¹³. If such attacks lead to substantial destabilization or damage, they might be considered acts of aggression, warranting legal scrutiny¹⁴. Furthermore, intentional cyberattacks causing widespread destruction, loss of life, or deprivation of essential services may escalate to the use of force, war crimes, or crimes against humanity, requiring a detailed assessment of the attack's intent, scale, and impact.

2.1. Space cyber use of force/ Armed attack

Despite the seriousness of these threats, no internationally recognized standards currently determine whether a cyberattack on space assets constitutes a use of force equivalent to a traditional armed attack, which could justify an armed response. The term “armed attack” lacks a globally recognized definition, especially concerning space cyberattacks¹⁵. Beyond the ambiguity of what qualifies as a space cyber-armed attack, questions arise regarding which provisions of current international legal frameworks apply to the conduct of war in the space cyber domain. Use of force or armed attack in

¹¹ Varadharajan and Suri, ‘Security Challenges When Space Merges with Cyberspace’, 101600.

¹² *ibid.*

¹³ Harold Hongju Koh, ‘International Law in Cyberspace’, Harvard International Law Journal Yale Law School, 54 (2012).

¹⁴ Michael N Schmitt, “The law of cyber warfare: Quo Vadis,” Stanford Law & Policy Review, vol. 25, pp. 269-300, 2014.

¹⁵ Andrew C. Foltz, “Stuxnet, Schmitt analysis, and the Cyber ‘Use-of-force’ debate”. Defense Technical Information Center, (2012).

cyberspace can be best explained by the then US legal advisor, Harold Koh, who stated, “Cyber activities that proximately result in death, injury, or significant destruction would likely be viewed as a use of force”¹⁶, which can include cyberattacks on space assets. However, the United States recognizes that non-kinetic cyberattacks can also qualify as an aspect of armed conflict in specific situations, giving rise to a proportional retaliation in response¹⁷.

The Charter of the United Nations (UN Charter) establishes a framework of collective security, prohibiting states' recourse to forces, with two exceptions of the UN Security Council authorized under Chapter VII of the UN Charter or in use of the right of self-defence¹⁸. The three interrelated thresholds are threat or use of force, armed attack, threat to peace, breach of peace, and act of aggression. However, the UN Charter does not enshrine these terms explicitly.

2.1.1. Constituents of armed attack

The definition of an armed attack remains a topic of debate, with the International Court of Justice (ICJ) considering only the gravest use of force to constitute an armed attack. The UN Security Council is responsible for identifying threats to peace, breaches of the peace, or acts of aggression¹⁹. Aggression is considered a severe and dangerous instance of unlawful use of force. Under the UN Charter, cyber operations are considered warfare only if they qualify the threshold provided under Article 2, paragraph 4.

Legal interpretation is crucial to determine when an activity, such as a cyber intrusion, becomes tantamount to an armed attack. Proposed frameworks, such as the Schmitt framework²⁰ and the Libicki framework²¹, can help determine when a computer intrusion constitutes an armed attack²². Furthermore, it is widely acknowledged through the principle of legal analogy that a cyberattack on a computer network resulting in physical damage to another state's assets would constitute an unlawful armed attack unless carried out in self-defense or under a UN-sanctioned operation. Cyberattacks conducted in self-defense under the UN Charter²³, or as part of

¹⁶ Catherine A. Theohary “Congressional research service in focus report, use of force in cyberspace” report no IF11995 Congressional Research Service, (2024).

¹⁷ *ibid.*

¹⁸ “Charter of the United Nations”, San Francisco, (1945).

¹⁹ *ibid.*

²⁰ Andrew C. Foltz, “Stuxnet, Schmitt analysis, and the Cyber ‘Use-of-force’ debate”. Defense Technical Information Center, (2012).

²¹ M. C. Libicki, “What is information warfare?” National Defense University Washington DC Institute for National Strategic Studies, (1995).

²² Andrew C. Foltz, “Stuxnet, Schmitt analysis, and the Cyber ‘Use-of-force’ debate”. Defense Technical Information Center, (2012).

²³ “Charter of the United Nations”, San Francisco, (1945).

a UN-sanctioned operation even in space are legal as long as the Law of Armed Conflict principles are followed²⁴. However, even when conducted in self-defense, a Cyber-attack by a state intending to cause physical damage to cyber infrastructures and works or installations holding hazardous forces, such as dams, dikes, and nuclear electrical generating stations, would be deemed as an unlawful armed attack under the 1977 Geneva Protocol I²⁵, including those conducted in a cyber domain or space cyber domain.

2.1.2. jus cogens Norms

The prohibition of the use of force represents fundamental norms of international law, recognized as both customary and jus cogens. The lack of a precise definition of space cyber force is challenging, requiring its interpretation to be carried out in good faith, in the ordinary meaning of the treaty's provisions, its object and purpose. The rise of space cyber-attacks as a novel risk to global peace marks a contextual evolution of the prohibition of the threat and the use of force but doesn't make this prohibition obsolete²⁶. The concept of force outlined in Article 2, paragraph 4, is flexible enough to include cyber-attacks²⁷.

2.2.Space cyber acts of war

Under Common Article 2 of the four Geneva Conventions of 1949, armed conflict can result from a formal declaration of war, a state's occupation, or any other armed conflict between states, even without a formal declaration of war²⁸. Armed conflict arises when one state employs force against another with sufficient scope, duration, and intensity that qualifies as an armed conflict²⁹. Such use of force is considered a state act that threatens another state's territorial integrity or political independence. Determination of a space cyber-attack as an armed attack requires legal analysis on a case-by-case basis³⁰. A state's space cyber intrusion may be considered an act of war if it threatens another state's territorial integrity or political independence³¹. Still, the United Nations Charter and international treaties struggle to adequately address space

²⁴ Michael N Schmitt, 'Computer Network Attack and the Use of Force in International Law: Thoughts on a Normative Framework', Columbia Journal of Transnational Law, vol 37 (1999), 885–938.

²⁵ Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (1977).

²⁶ Michael N Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations, 2nd edn (2017).

²⁷ "Charter of the United Nations", San Francisco, (1945).

²⁸ The Geneva Conventions of 12 August 1949 (1949).

²⁹ Andrew C. Foltz, "Stuxnet, Schmitt analysis, and the Cyber 'Use-of-force' debate", Joint Force Quarterly National Defense University, vol 67/4 (2012), 40–48.

³⁰ *ibid.*

³¹ Peter Pascucci, 'Distinction and Proportionality in Cyber War: Virtual Problems with a Real Solution', Minnesota Journal of International Law, 26/2 (2017), 419–60.

cyber-armed attacks and war. The Tallinn manual provides guidelines, but it also lacks clarity on proportionality and attribution. Experts advocate for a case-to-case approach to evaluating the threat and its level and extent³².

2.3.Space cybercrime as a crime of aggression

Aggression is the use of armed force by a State against another State's sovereignty, territorial integrity, or political independence or in any other manner inconsistent with the UN charter³³. Furthermore, Article 3 of “Definition of aggression” provides a list of acts of aggression, including

“(a) The invasion or attack by the armed forces of a State of the territory of another State, or any military occupation, however temporary, resulting from such invasion or attack, or any annexation by the use of force of the territory of another State or part thereof;

(b) Bombardment by the armed forces of a State against the territory of another State or the use of any weapons by a State against the territory of another State;

(d) An attack by the armed forces of a State on the land, sea, or air forces or marine and air fleets of another State;

(g) The sending by or on behalf of a State of armed bands, groups, irregulars, or mercenaries, which carry out acts of armed force against another State of such gravity as to amount to the acts listed above or its substantial involvement therein”³⁴.

The definition of aggression further highlights that these lists are not comprehensive or exhaustive. It allows the Security Council to identify other acts that qualify as aggression under the provisions of the Charter of the United Nations. Similarly, the listed armed attacks are also not exhaustive, and other acts may qualify as armed attacks³⁵.

2.4.Space cybercrimes against humanity

Crimes against humanity are mainly enshrined in the Rome Statute³⁶, which sets out the principles governing these crimes. Crimes against humanity involve attacks on civilian populations. Article 7(2)(a) of the Rome Statute enshrines that "an attack directed against any civilian population refers to a series of actions carried out against

³² ‘State Responsibility and Attribution of Cyber Intrusions After Tallinn 2.0, Texas Law Review’.

³³ Jordan J. Paust, ‘Substantive and Procedural Aspects of International Criminal Law: The Experience of International and National Courts.’, ed. by Olivia Swaak-Goldman, American Journal of International Law, 96/4 (2002), 1006–9.

³⁴ United Nations “Definition of Aggression” (1974).

³⁵ Tom. Ruys, “Armed Attack” and Article 51 of the UN Charter Evolutions in Customary Law and Practice’, Cambridge Studies in International and Comparative Law (2010), 616.

³⁶ ‘Rome Statute of the International Criminal Court’, U.N.T.S 2187 (1998).

civilians as part of a state or organizational policy”³⁷. For a cybercrime in outer space to constitute a crime against humanity, the attack must qualify the circumstantial elements of Article 7(1), which specifies the general characteristics necessary for crimes against humanity³⁸. Specifically, the cyberattack must be part of a “widespread or systematic attack directed against a civilian population”, and the attacker shall know the attack’s scale and impact.

For an attack to be considered "widespread" under Article 7, it must have a large-scale impact, typically determined by the count of victims or the geographical extent of the harm. In the context of outer space, cyberattacks targeting satellites or space infrastructure providing critical services—such as GPS, weather monitoring, or communication systems—can potentially disrupt millions of lives on Earth. The global reliance on space-based technologies means that an attack on a few critical satellites could result in widespread disruption, impacting entire populations, economies, and governments. For example, an attack that disables GPS satellites could disrupt air traffic, shipping, and emergency response services, leading to a cascade of accidents and failures affecting a large civilian population. The interconnected nature of space infrastructure means that cyberattacks could have a large-scale impact without directly targeting physical objects or causing physical harm, meeting the “widespread” criterion³⁹.

The Rome Statute further stipulates the attack to be systematic. It is part of a coordinated effort or follows a preconceived plan or policy, often perpetrated by a state or organized entity. Cyberattacks in outer space, particularly those targeting space infrastructure, almost always exhibit a systematic nature due to the technical and organizational coordination involved. Conducting a cyber operation on space-based assets requires sophisticated technology, expertise, and often state resources. It is rarely an individual's work but involves careful planning, preparation, and execution. These cyberattacks usually target specific infrastructure to cause widespread disruption, making them systematic⁴⁰.

In outer space, the civilian population on Earth remains at risk of attacks on space-based infrastructure, which is crucial for providing services like telecommunications, navigation, and weather forecasting. A cyberattack on space-based assets can be considered a crime against humanity even if it does not involve direct physical violence.

³⁷ *ibid.*

³⁸ *ibid.*

³⁹ Patrick Lin and others, *Outer Space Cyberattacks: Generating Novel Scenarios to Avoid Surprise*, (2024).

⁴⁰ R. Kesavamoorthy and P. Karthikeyan, ‘Cyberspace and Outer Space Security and Mitigation Strategies’, in *Cyber Space and Outer Space Security*, 1st edn (New York, 2024), 1–15.

For example, disrupting a communications satellite could lead to widespread outages affecting millions of civilians and cutting off their access to emergency services, healthcare, and essential communications ⁴¹.

Additionally, cyberattacks in space could result in psychological, economic, and societal disruptions that cause significant suffering for civilian populations. The broad definition of "attack" in the Rome Statute includes non-physical or non-violent actions targeting civilians that result in significant harm. An attack on satellites that disrupts financial transactions, water supply systems, or energy grids could lead to severe economic losses, societal chaos, and psychological trauma, potentially meeting the criteria for crimes against humanity.

2.5. Other international crimes in outer space

Certain specific acts enumerated in Article 7(1) of the Rome Statute could be carried out through cyber means in outer space. For example, cyber operations leading to loss of life or severe injury due to the disruption of critical services could be considered murder under Article 7(1)(a). Similarly, a cyberattack that disproportionately attacks a certain group based on nationality, ethnicity, or religion and deprives them of essential services could qualify as persecution under Article 7(1)(h). The residual provision of Article 7(1)(k) also covers "other inhumane acts," which can include significant economic or psychological suffering caused by cyberattacks⁴². The flexibility of the Rome Statute in defining crimes against humanity, particularly in its ability to encompass non-physical attacks, means that cybercrimes in outer space could fall within its scope. A cyberattack on space-based infrastructure that causes widespread harm, systematically targets civilians, and results in severe suffering or deprivation of rights could meet the legal requirements for prosecution as a crime against humanity.

There is a growing awareness that cyberattacks in space could be considered severe international crimes, including crimes against humanity, acts of aggression, war crimes, cyber terrorism, or even genocide. However, current international law does not address how these crimes apply in the unique realm of outer space. The absence of specific treaties or conventions targeting cyberattacks in outer space leaves space assets and critical infrastructure vulnerable to various forms of space cyber aggressions. While existing international legal frameworks such as the Rome Statute, cybercrime conventions, the UN Charter, and the Geneva Conventions may perhaps be practical to cybercrimes in outer space, they were not originally designed to address the novel

⁴¹ R. Kesavamoorthy and P. Karthikeyan, 'Cyberspace and Outer Space Security and Mitigation Strategies', 1–15.

⁴² "Rome Statute of the International Criminal Court", U.N.T.S 2187 (1998).

threats posed by cyber operations in this domain. As a result, the interpretation and application of these laws to space-based cyber activities rely heavily on analogies to terrestrial acts of war, aggression, and humanitarian violations.

3.Determination of the wrongfulness of cyberattacks in outer space and the international legal obligations of states for the prevention of space cyberattacks

When a state's acts do not align with its international obligations, it might be considered to have committed an internationally wrongful act, as per international law. This is regardless of the nature or origin of the act. A state breaches an international obligation when it does not adhere to what is required under that obligation⁴³. However, an act of a State is not considered a breach unless it is bound by that specific obligation when the act takes place. No specific international legislation currently addresses cyber risks in or to outer space, especially state-sponsored space cyber threats⁴⁴. However, various intersecting normative regimes are in effect, such as domestic law, International Space Law, International Cyber Law, public international law, and the law of state responsibility.

Space-related activities, including space exploration, space technology development, space applications, and technology results implementation, are governed by a set of specific international treaties and conventions⁴⁵. However, cyberattacks, primarily terrestrial, generally are not covered under this definition. Nevertheless, there are specific scenarios where cyberattacks intersect with space activities. For example, a state may use a satellite to jam or spoof other satellites as part of space activity, or space objects might be manipulated to achieve the attackers' goals, such as taking control of a satellite to disrupt or deorbit another satellite, integrating cyberattacks into space activities, either directly or indirectly.

3.1.Wrongfulness of space cybercrimes:

Determining whether cyber-attacks constitute the use of armed force or armed attack involves various approaches. The effect-based approach, as affirmed by ICJ in the *Nicaragua v. US* judgment, suggests that cyber activities resulting in effects comparable to kinetic attacks could be deemed as the use of armed force⁴⁶. The Tallinn

⁴³ United Nations, Responsibility of States for Internationally Wrongful Acts (2001).

⁴⁴ Thanapat Chatinakrob, 'Interplay of International Law and Cyberspace: State Sovereignty Violation, Extraterritorial Effects, and the Paradigm of Cyber Sovereignty', *Chinese Journal of International Law*, 23/1 (2024), 25–72.

⁴⁵ ST, United Nations Treaties and Principles on Outer Space, Related General Assembly Resolutions and Other Documents United Nations.

⁴⁶ "*Military and paramilitary activities in and against Nicaragua (Nicaragua v. United States of America)*", Case No. 70, International Court of Justice, Judgment of Jun. 27 (1986).

Manual's non-exhaustive list of factors—such as severity, immediacy, directness, invasiveness, and state involvement—provides a framework for determining whether a cyber attack's impact equates to a kinetic attack. However, there is no consensus on this issue, and the applicability of these standards to space cyberattacks remains complex⁴⁷.

Two key issues under public international law are prominent in space cyberattacks: whether cyberattacks qualify as threats to peace and whether states should bear due diligence obligations towards affected states. Article 2(4) of the UN Charter prohibits the threat or use of force, while Chapter VII addresses threats to peace and acts of aggression, which is relevant in this context⁴⁸. Applying these provisions to cyberattacks depends on whether such actions constitute a "threat or use of force" under Article 2(4)⁴⁹. This provision is a jus cogens norm, suggesting space cyberattacks to be strictly prohibited if considered a threat or use of force, and states responsible for such attacks bear international responsibility⁵⁰.

3.2. International legal obligations of states

The Outer Space Treaty (OST) is the foundation of space law and emphasizes state responsibility for all space activities, including those by private actors⁵¹. It mandates states to ensure adherence of their national space activities to international obligations by authorizing and continuously supervising it. However, the OST does not explicitly address cyberattacks, creating a regulatory gap that needs to be filled. The Liability Convention complements the OST and establishes liability for damage caused on Earth's surface or to aircraft and fault-based liability for other damage. Nevertheless, it does not explicitly cover cyberattacks, leading to ambiguity regarding non-physical damages like data breaches or service disruptions resulting from cyber incidents.

3.2.1. Obligations under international space law

International space law establishes a framework to govern activities in outer space, addressing issues including weaponization and cyber threats primarily through the Outer Space Treaty and other relevant international instruments. Article IV explicitly prohibits the placement of weapons of mass destruction in outer space and emphasizes that the moon and other celestial bodies should exclusively be utilized for peaceful

⁴⁷ M. N. Schmitt, "Tallinn manual 2.0 on the international law applicable to cyber operations", 2nd ed. Cambridge University Press, (2017).

⁴⁸ "Charter of the United Nations", San Francisco, (1945).

⁴⁹ 'Military and paramilitary activities in and against Nicaragua (Nicaragua v. United States of America)', Case No. 70, International Court of Justice, Judgment of Jun. 27, 1986.

⁵⁰ Du Li, 'Cyber-Attacks on Space Activities: Revisiting the Responsibility Regime of Article VI of the Outer Space Treaty', Space Policy, 63 (2023).

⁵¹ 'Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies' (1967).

purposes⁵². However, the interpretation of "peaceful purposes" varies, with some scholars and states advocating for a complete prohibition of military uses while others interpret it as permitting non-aggressive military uses⁵³. The Prevention of the Placement of Weapons in Space (PPWT), a proposed treaty, extends this discussion by aiming to ban space-based weapons, focusing on physical, kinetic, and nuclear weapons, but does not specifically address cyber weapons.

Article IX of the OST requires states to conduct space activities with "due regard" for the interests of other states and to avoid harmful interference. This provision while traditionally focused on physical impact, could be interpreted to encompass cyberattacks that disrupt space systems, humans and the environment in space and on Earth. The inclusion of cyber harms aligns with the broader interpretation of harmful interference, as highlighted in the radio regulations, which define harmful interference as any action that endangers radio navigation or safety services or significantly degrades, obstructs or repeatedly interrupts radio communication services⁵⁴. While Article IV does not explicitly address harmful radio interference due to cyberattacks, such interference undermines the peaceful use of outer space. Similarly, physical interference also jeopardizes the peaceful execution of space activities.

Article IX imposes preventive obligations on states, requiring them to undertake international consultations if an activity by themselves or their nationals might interfere with the activities of other states before the actual interference or harm occurs⁵⁵. This consultative mechanism serves as a preemptive measure, enabling states that might be affected to be aware of potential harm and take relative measures, including those resulting from cyberattacks. By requiring these, article IX established the principle of mutual regard and cooperation in space activities. In sum, articles IV and IX of the OST outline the international obligations of states to ensure peaceful uses of outer space⁵⁶.

3.2.2. Obligation under international cyber law

International cyber law, including the United Nations Convention against

⁵² United Nations General Assembly, Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies (1967).

⁵³ Bin Cheng, 'Properly Speaking, Only Celestial Bodies Have Been Reserved for Use Exclusively for Peaceful (Non-Military) Purposes, but Not Outer Void Space', International Law Studies Series U.S. Naval War College, 75 (2000), 81–118, Stephan Hobe, 'The Meaning of Peaceful Purposes in Article IV of the Outer Space Treaty', Annals of Air and Space Law (2015), 9–24.

⁵⁴ 'International Telecommunication Union Radio Regulations' (2022), 'Constitution of the International Telecommunication Union', in International Organization and Integration Annotated Basic Documents and Descriptive Directory of International Organization and Arrangements (1982), 1–2.

⁵⁵ Stephan Hobe, Kai-Uwe Schrogl and Bernhard Schmidt-Tedd, The Cologne Commentary on Space Law, 2009th edn (2009).

⁵⁶ Schmitt, 'Computer Network Attack and the Use of Force in International Law: Thoughts on a Normative Framework', 885–938.

Cybercrime and the Budapest Convention, primarily addresses terrestrial cybercrimes; its principles implicitly extend to the emerging domain of space cybercrimes⁵⁷. These conventions mandate states to combat cybercrimes, an obligation that extends to the emerging challenges posed by space cybercrimes. By committing to align with the purpose and the principle of the United Nations charter, states are required to address vulnerabilities created by cybercrimes. Although international cyber law does not define space cybercrimes as constituting international crimes, use of force, and armed attacks, states are mandated to adopt and implement appropriate measures and legislations to define and criminalize cybercrimes, implicitly including those targeting space cyberinfrastructure. The United Nations Convention against Cybercrime obliges states to respect human rights and protect individuals and entities affected by cybercrimes. Further, there is an obligation to deny haven to cyber criminals and prosecute crimes irrespective of origin. These principles extend to cyber offenses targeting space systems constituting international crimes.

3.2.3. Due diligence obligation for prevention of space cybercrimes

Due diligence obligations are significant aspects of addressing space cyberattacks within the public international legal framework. States are obliged to ensure that their territory is not utilized to harm other states, a principle affirmed in the *Island of Palmas*⁵⁸ and *Corfu Channel* cases⁵⁹. While this principle has been established in various domains, its application to cyberspace, including space cyberattacks, remains under discussion. The UN, through various Groups of Governmental Experts (GGE), has addressed the increasing impact of transboundary cyber activities, reflecting a rising acknowledgment of the importance and necessity of cooperative measures to mitigate cyber threats effectively^{60,61}.

Due diligence obligations extend not only to states from which cyberattacks originate but also to those through which the attacks transit or whose infrastructure is involved. While the final report does not explicitly differentiate the obligations of these states, it implies that both types of states should take appropriate measures⁶². Given the

⁵⁷ United Nations Convention against Cybercrime (2024) [accessed 6 December 2024]; Convention on Cybercrime (2001).

⁵⁸ REPORTS OF INTERNATIONAL ARBITRAL AWARDS RECUEIL DES SENTENCES ARBITRALES *Island of Palmas Case* (Netherlands, USA), 1928.

⁵⁹ “*Corfu channel case, Case concerning violations of sovereignty and international maritime law* (United Kingdom vs. Albania),” International Court of Justice, I.C.J. Reports (1949).

⁶⁰ Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security.

⁶¹ Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, December (2015).

⁶² Group of governmental experts on developments in the field of information and telecommunications in the context of international security”, United Nations, A/68/98, (2013).

technological disparities among spacefaring nations, evaluating the proportionality and effectiveness of responses must consider the specific capabilities and constraints of the states involved. Constructive knowledge of ongoing cyberattacks is crucial for a state's due diligence obligations. International jurisprudence and documents, including Chapter IV of the ARSIWA, emphasize that a state's knowledge of the circumstances of a wrongful act is essential for Responsibility⁶³.

The 2013 report stated, "States must not use proxies to commit internationally wrongful acts and must ensure non-state actors do not use their territories for unlawful Information and communication technologies (ICT) activities"⁶⁴. The 2015 report emphasized that states should prevent non-state actors from using their territory to commit such acts⁶⁵. This perspective is embedded in Norm 13(c) of the final GGE report, which states that states shall not knowingly permit their territories to be utilized for the commission of international wrongful acts involving ICTs. The norm outlines the threshold for triggering state due diligence obligations, including constructive knowledge of a cyber-attack and its occurrence on a state's territory. It suggests that states should take all appropriate steps to detect, investigate, and address such acts while acknowledging that complete monitoring is not expected. States should act proportionately and effectively and seek assistance if needed⁶⁶.

To hold a state accountable for its due diligence obligations in relation to space cyberattacks, several conditions must be met: the space object involved must fall under the state's jurisdiction, the object must be utilized to execute the cyberattack or to transmit orders for it, and the state must be aware that the attack is ongoing. Once the state is aware, it should take action to prevent the attack if possible; if prevention is not feasible, it should inform relevant states, particularly the target state⁶⁷.

3.2.4. Obligation to prevent international wrongful acts

The obligation of states under international law to prevent and address internationally wrongful acts, including those in outer space, is clear. International space law, public international law, and customary international law all impose duties

⁶³ International Law Commission, "Responsibility of states for internationally wrongful acts," U.N. Doc. A/56/10, (2001).

⁶⁴ "Report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security," United Nations, A/RES/70/237, (2015).

⁶⁵ "Group of governmental experts on developments in the field of information and telecommunications in the context of international security", United Nations, A/68/98 (2013).

⁶⁶ Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, Note by the Secretary General, (July 2021); Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, (adopted by December 2018).

⁶⁷ Li, 'Cyber-Attacks on Space Activities: Revisiting the Responsibility Regime of Article VI of the Outer Space Treaty'.

on states to uphold these obligations. However, the current legal frameworks do not explicitly address the wrongful nature of cyberattacks in outer space, leading to a reliance on analogies to terrestrial legal concepts. The existing treaties and conventions serve as a foundational framework for governing space activities and preventing harmful interference, but they do not specifically cover the complexities of cyber threats. This regulatory gap highlights the need for more precise legal provisions that address cyberattacks in space. While existing international laws offer some guidance, the unique nature of space cyberattacks necessitates explicit legal standards tailored to this domain.

4. Attribution of cyberattacks in outer space

A state is internationally responsible when it commits wrongful acts which violate its obligations under international law. The International Law Commission's (ILC) Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA) outline these principles. These principles establish that states bear international responsibility for wrongful acts, including those by state organs or agents⁶⁸. However, applying these rules to cyberattacks, especially in the context of space activities, is challenging due to the difficulties in tracing perpetrators and establishing direct links between attackers and states.

The OST extends international responsibility to states for all national activities conducted in outer space, including those conducted on the moon and other celestial bodies, irrespective of whether governmental or non-governmental entities carry out these activities⁶⁹. This responsibility ensures obedience to the Treaty. However, in public international law, states are typically not held responsible for the acts of their non-governmental entities unless they are acting in an official state capacity.

The rising involvement of private actors in space-related cyber activities potentially brings these actions under Article VI of the OST, extending state responsibility to these cyber activities⁷⁰. However, regarding space cyberattacks, Article VI's attribution standards are limited. The general rules of international law, as codified in ARSIWA, also apply but may conflict with the space law attribution regime,

⁶⁸ 'International Law Commission, "Responsibility of states for internationally wrongful acts, U.N. Doc. A/56/10, (2001).

⁶⁹ 'F. G. Von der Dunk, " The origins of authorization: Article VI of the outer space treaty and international space law," in *National Space Legislation in Europe*, Chapter 1, p. 3-28, Leiden, The Netherlands: Brill Nijhoff, (2011).

⁷⁰ Frans G Von Der Dunk, *The Origins of Authorisation: Article VI of the Outer Space Treaty and International Space Law and International Space Law*, (2011).

leading to differing outcomes. Article VI of the OST mandates that states bear international responsibility for national activities in outer space, even if non-governmental entities conduct these activities⁷¹. This requirement represents a departure from traditional international law approaches and directly holds states accountable for breaches of the OST, including those committed by non-governmental actors⁷².

4.1. Attribution challenges in national space activities

Challenges arise in defining a state's national space activities and clarifying terms such as "national activities" and "appropriate state." While governmental operations are attributable to the state, the involvement of private actors complicates attribution. According to the prevailing view in public international law, a state is responsible for non-governmental activities within its jurisdiction, including those by its nationals⁷³. However, this approach can lead to complications in cases of concurrent jurisdiction when multiple states have connections to a space activity.

Article VIII of the OST introduces quasi-territorial jurisdiction, adding complexity to determining state responsibility. Factors like dual nationality and multinational involvement make it difficult to identify the responsible state, particularly cyberattacks in outer space. Under space law, the state of registry has jurisdiction and control over space objects and personnel, as the OST specifies⁷⁴. This jurisdiction entails responsibility for the space object, but the Registration Convention of 1975 allows multiple states to claim registration of a space object⁷⁵. While the state of registry should be closely associated with the launch activity, this responsibility generally pertains to outer space operations and not those conducted on Earth.

4.1.1. Private space actors

Under Article VI of the OST, holding a launching state responsible for a satellite controlled by a private actor in a different state or after a transfer of ownership would be unreasonable. However, the Registration Convention allows for the transfer of jurisdiction and control to another state through agreements⁷⁶. Concerns about

⁷¹ United Nations General Assembly, "Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies" (Outer Space Treaty), Resolution 2222 (XXI), (1967).

⁷² Guoyu Wang, 'The Complex Neutrality of Commercial Space Actors in Armed Conflict', Humanitarian Law & Policy Blog, (2023).

⁷³ 'State Responsibility and Commercial Space Activities', in Oxford Research Encyclopedia of Planetary Science (2023).

⁷⁴ Guoyu Wang, 'The Complex Neutrality of Commercial Space Actors in Armed Conflict', Humanitarian Law & Policy Blog, (2023).

⁷⁵ Stefan A Kaiser and Llm McGill, When Cyber Activities Are Space Activities Definitions Are Key, (2020).

⁷⁶ *ibid*.

"registries of convenience" persist, as no standardized provisions identify the responsible state for space activities.

Privatizing and commercializing space activities create challenges in attributing responsibility, especially in multinational projects and corporations⁷⁷. Current interpretations of "national activities" and "appropriate state" do not fully consider the roles and ownership stakes of companies or the relationships between countries and their nationals. In the case of space cyberattacks, these complexities are magnified, potentially implicating multiple states without adequately resolving the issues of multinational involvement.

4.1.2. Attribution of private activities

Legal and technical challenges arise when attributing responsibility under the existing regime. Under public international law, a state bears direct responsibility for internationally wrongful acts attributable to it and may also hold indirect responsibility for conduct not committed by the state itself⁷⁸. ARSIWA codifies these rules, including attribution to state organs and agents and situations where acts of private actors may be attributed to it⁷⁹. However, these principles fail to address cyber threats due to difficulty in tracing perpetrators and proving attribution.

For severe cyberattacks, individual attackers might be held accountable as international criminals. Establishing a direct link between attackers and a state is challenging for other attacks. A state might be indirectly responsible under Article 16 of ARSIWA if it aids or assists another state in committing an internationally wrongful act⁸⁰. For instance, if one state hacks into another state's satellite control system and uses it to attack a third state's satellite, both involved states might be held responsible, with the assisting state potentially liable under ARSIWA's criteria for aiding or assisting wrongful acts⁸¹.

Applying Article 16 of ARSIWA to space cyberattacks is problematic due to ambiguous terms like "aid and assistance" and the complexities of attribution. The concept of "aid and assistance" is context-dependent, making it challenging to determine whether omissions qualify as such. Distinguishing between indirect

⁷⁷ 'D. Ireland-Piper, M. Fehlhaber, and A. Bonenfant, "State responsibility and commercial space activities," *Oxford Research Encyclopedia of Planetary Science*, (2023).

⁷⁸ Frans G Von Der Dunk, 'Scoping National Space Law: The True Meaning of "National Activities in Outer Space" of Article VI of the Outer Space Treaty', *International Institute of Space Law*, 62/3 (2019), 227–37.

⁷⁹ F. G. Von der Dunk, "The origins of authorization: Article VI of the outer space treaty and international space law," in *National Space Legislation in Europe*, Chapter 1, p. 3–28, Leiden, The Netherlands: Brill Nijhoff, (2011).

⁸⁰ International Law Commission, "Responsibility of states for internationally wrongful acts," U.N. Doc. A/56/10, (2001).

⁸¹ F. G. Von Der Dunk, "Scoping national space law: The true meaning of 'national activities in outer space' of article VI of the outer space treaty," *International Institute of Space Law*, vol. 62, no. 3, pp. 227–237, (2019).

responsibility for assistance and direct responsibility for failing to meet due diligence obligations remains difficult⁸².

Lanovoy's differentiation between omissions and due diligence obligations, such as the knowledge-based, character-based, and control-based distinctions, provide limited guidance for dealing with space cyberattacks. It is hard to prove a state's knowledge of manipulation during a cyber-attack and establish intent under the purpose-based criterion outlined in the Commentaries on ARSIWA. These complexities make the application of Article 16 to space cyberattacks particularly challenging⁸³.

When space activities and cyberattacks intersect, the *lex specialis* principle suggests that Article VI of the OST precedes ARSIWA's general rules. However, strict application of this principle can lead to questionable outcomes. Article VI of the OST does not mandate a direct connection of a state's instruction or control over a private entity, potentially holding a state responsible for cyberattacks beyond its control. Conversely, ARSIWA requires a higher threshold for complicity under Article 16, indicating a state cannot be set responsible for a cyber-attack if it does not have knowledge of the act⁸⁴.

Given these challenges, a more flexible, case-by-case approach to attribution rules for space cyberattacks may be necessary⁸⁵. Public international law and space law could supplement each other, especially in multinational contexts where Article 8 of ARSIWA might provide a solution. However, applying Article 8 presents difficulties, as case law, such as in Nicaragua, demonstrates varying standards for the extent of control essential to establish state responsibility.

In the *Tadic* case, the International Criminal Tribunal for the former Yugoslavia (ICTY) established an 'overall control' standard for state responsibility, diverging from the ICJ's approach⁸⁶. The ICTY ruled that for attribution to a state, it is sufficient that a group is under overall state control, including participation in planning and supervising military operations, without requiring specific orders for each action⁸⁷. This standard's consistency with general attribution rules is questionable, as underlying principles should be uniform across contexts, and varying standards must be justified by context

⁸² Li, 'Cyber-Attacks on Space Activities: Revisiting the Responsibility Regime of Article VI of the Outer Space Treaty'.

⁸³ Samuli Haataja, 'Complicity in Unlawful Offensive Cyber Operations Under International Law on State Responsibility', *Eccws*, 22/1 (2023), 232–38.

⁸⁴ Frans G Von Der Dunk, 'Scoping National Space Law: The True Meaning of "National Activities in Outer Space" of Article VI of the Outer Space Treaty', *International Institute of Space Law*, 62/3 (2019), 227–37.

⁸⁵ Guoyu Wang, 'The Complex Neutrality of Commercial Space Actors in Armed Conflict'.

⁸⁶ 'Prosecutor v. Duško Tadić Case', (1999).

⁸⁷ Luca Erhart and Maria Boutovitskai, 'Transforming Article VI of the Outer Space Treaty into an Effective Mechanism of Space Debris Mitigation'.

or purpose.

5.Regulatory gaps and challenges in state responsibility and attribution for cyberattacks in outer space

Currently, no specific international treaties or conventions directly address cybersecurity threats in outer space, especially those related to cyberattacks on space assets. No internationally recognized criteria exist for the determination of cyberattacks on space assets, whether it qualifies as a use of force equivalent to a traditional armed attack. This ambiguity complicates the legal classification of cyberattacks and their potential escalation to acts of aggression, war crimes, or crimes against humanity. While definitions from other international laws can provide some guidance, there are no specific definitions for cybercrimes in outer space. The current treaties do not clearly define or distinguish cyberattacks from other cyber incidents, making applying appropriate legal standards and responses difficult. Moreover, there is no comprehensive framework for attributing cyberattacks to states or private actors, complicating efforts to hold responsible parties accountable. Existing space treaties primarily focus on the physical elements of space activities and do not extend to cyber threats. Furthermore, there are no binding international standards for states to implement specific cybersecurity measures or protocols for space operations.

5.1.Ambiguity in defining Space cybercrimes

The ambiguity in defining wrongful acts in the context of space cyberattacks presents significant challenges. International legal obligations for preventing such attacks remain underdeveloped, and there are no clear guidelines for assessing and categorizing specific cyber operations as unlawful. Under international law, a state's conduct is internationally wrongful if it breaches an international obligation owed to another state or the international community. However, determining wrongful acts becomes complex when applied to cyberattacks in outer space. The absence of explicit international legislation addressing cyber risks in outer space highlights a regulatory gap.

The lack of clear definitions for space cyberattacks terms, such as data breaches or service disruptions, makes applying these treaties to today's advanced cyber threats difficult. For instance, the Liability Convention focuses on fault-based liability for physical damage but does not consider the consequences of non-physical damage caused by cyberattacks. This lack of specific definitions hinders the efficiency of the existing international legal instruments in addressing non-traditional forms of harm in

outer space. Non-physical damages like data breaches or disruptions of communication systems are increasingly common in cyber operations but are not adequately covered by current space treaties.

5.2.Unclear existing international legal framework

While the OST and other related treaties broadly cover space activities, they do not specifically tackle the unique challenges cyberattacks pose. Cybersecurity is only briefly mentioned in provisions such as Article VI of the OST, which assigns states responsibility for space activities conducted by private entities. However, the nature of cyber threats, which often involve non-physical damage like data breaches or system disruptions, is not explicitly addressed. This regulatory gap leaves the international legal framework inadequate to manage and prevent the rapidly evolving cyber threats that could significantly impact space operations. The lack of a specific legal mechanism for outer space cybersecurity creates a significant regulatory gap, particularly in terms of how states or private entities should respond to cyberattacks on space infrastructure.

5.3.Overlapping international legal frameworks

The principle of *lex specialis* suggests that space law should prevail over public international law in cases of overlap, leading to regulatory conflicts. Article VI of the OST holds states responsible for the activities of non-governmental entities, even when the state does not directly control or instruct them. In contrast, under ARSIWA, a state may not be held responsible for an attack unless there is evidence of its involvement or knowledge. This conflict between the broad responsibilities imposed by space law and the more specific attribution rules under public international law leads to uncertainty in applying responsibility and due diligence standards for cyberattacks affecting space operations. Furthermore, Article VI of the OST mandates state responsibility for all national space activities, including those carried out by private entities. However, it does not clarify how this responsibility extends to cyber activities or explain how states should handle breaches caused by cyber means. Terms like "national activities" and "activities of non-governmental entities" are not explicitly defined in the context of cyberattacks. This ambiguity complicates the attribution of cyber incidents involving space assets, especially when multiple jurisdictions or actors are involved.

5.4.Technical challenges in space cybercrimes attribution

The existing international legal framework, including the ARSIWA and the OST, faces significant difficulties in attributing cyberattacks due to the nature of cyber operations. Cyberattacks often use techniques like anonymization, encryption, and

proxy servers, which make it difficult to trace the true source of an attack. While ARSIWA provides a framework for state responsibility, it is based on traditional notions of state conduct. It does not specifically address the challenges posed by modern cyber activities in outer space. The anonymity of cyber-attacks presents a significant challenge in applying the principles of attribution outlined in ARSIWA. Specifically, the challenges lie in attributing the conduct of private entities to a state, which requires building the link that the entity was acting under the instructions of, or under the direction and control of, the state. Similarly, Article 16, which holds a state responsible for aiding or assisting another state in committing international wrongful acts, requires a clear connection between the assisting state's action and the wrongful act. The anonymous nature of the space cyber-attacks creates a significant legal gap in determining responsibility in these situations.

5.5.Multinational/ multi-actor space activities

OST and the Registration Convention mandate the registration of space objects by states, but the transfer of jurisdiction and control between multinational entities and different states adds ambiguity. Where if ownership or operational control of a satellite is transferred across borders, it is unclear which state bears responsibility for cybersecurity breaches that might occur during or after the transfer. Moreover, states can register space objects in various jurisdictions. Still, without a clear responsibility framework for cyberattacks, there is ambiguity about how responsibility should be attributed when multiple actors are involved. This lack of clarity becomes even more critical when considering the complexities of multinational ownership and the often-global nature of cyber operations. These corporations often operate across several jurisdictions, making it difficult for any one state to supervise their space-related activities effectively. This gap leaves states in a precarious position, where they are held responsible for cyber incidents involving multinational entities but without the tools or authority to effectively regulate those entities' activities.

5.6.Challenges in the attribution of space cybercrimes resulting in non-physical harm

The Liability Convention establishes fault-based liability for damage caused on Earth or to aircraft but does not cover non-physical damages resulting from cyberattacks. Modern cyber operations can result in significant non-physical harm, such as data breaches or the disruption of satellite services. Still, these consequences are not specifically addressed in the current legal regimes. This gap creates uncertainty in how to handle cases where cyberattacks disrupt space operations without causing physical

damage to space infrastructure, such as satellites or ground stations, leaving the legal regime ill-equipped to regulate cyber threats in space efficiently. Another significant challenge lies in proving complicity in cyberattacks. Under ARSIWA and the OST, proving complicity in cases where one state aids or assists another in conducting a cyberattack is a complex issue. Article 16 of ARSIWA allows for indirect responsibility if a state aids or assists another in conducting an internationally wrongful act. Still, the terms "aid and assistance" are vague in the terms of cyberattacks. For example, it is hard to assess whether providing software, technology, or infrastructure that is later used in a cyberattack constitutes "aid and assistance" under ARSIWA. Moreover, proving that a state had the knowledge or intent to assist in a cyberattack adds another layer of complexity to establishing responsibility under international law.

5.7.Ambiguities of the terms

Different thresholds, such as "overall control" established in the Tadic case versus the "effective control" from the ICJ's Nicaragua case, create inconsistencies in how responsibility for cyberattacks is applied. These different standards complicate attribution, especially in the context of cyberattacks, where the extent of control a state exercises over non-governmental entities or proxy actors can vary widely. In space activities, these inconsistencies make it harder to establish clear responsibility for cyberattacks, particularly when multiple actors or states are involved in complex space operations.

Article IX of the OST mandates states to avoid harmful interference with the activities of other states, but the treaty does not explicitly address cyber threats. The due diligence standard, which calls for states to prevent harm based on factors like foreseeability, importance of protected interests, and state capabilities, is flexible and can vary widely. Without clear criteria for assessing whether states have met their due diligence obligations, holding states accountable for failing to prevent cyberattacks becomes challenging. Moreover, determining the adequacy of a state's response to cyber threats, given its capabilities and knowledge, is equally difficult, leaving considerable room for inconsistent enforcement of preventive measures.

6.Improving international responsibility regime for attributing and preventing cyberattacks in outer space

While the existing legal frameworks were not initially intended to address cybercrimes in outer space, they provide a foundational base that can be expanded to address international responsibility for space-related cybercrimes. A significant issue is the lack of universally accepted definitions for key concepts such as "use of force," "aggression," and other serious international crimes within the context of cyber operations. It is difficult to accurately classify cyberattacks without clear definitions, which complicates determining whether such attacks amount to acts of aggression or other forms of international crime. However, this gap also presents an opportunity to enhance the existing legal framework. Definitions from other international laws can be adapted to the context of cyber operations in space, providing a means to establish clear criteria for assessing and addressing cyberattacks. By integrating these concepts into the space law framework, states can develop more precise and effective guidelines for responding to and managing cybersecurity threats in outer space.

6.1.Extending existing legal frameworks to address cyberattacks in outer space

As the cornerstone of space law, the OST creates state responsibility for all national space activities, including those by private actors. Although it does not explicitly address cyber threats, the principle of state responsibility can be extended to cover cyber operations. States can enhance supervision and implement cybersecurity measures to address vulnerabilities in space infrastructure, using the OST as a platform for ensuring accountability for cyberattacks.

The ARSIWA further provides a valuable framework for attributing responsibility for cyberattacks, which often lack clear attribution due to their decentralized nature. By applying ARSIWA's principles of direct and indirect responsibility to space operations, states can be held accountable for cyberattacks launched by non-state actors under their jurisdiction. This strengthens the international framework by ensuring states take responsibility for securing their space operations.

Similarly, the Registration Convention, which requires states to register space objects, can be adapted to include cybersecurity obligations. By integrating cybersecurity measures—such as encryption and secure communication protocols—into the registration process, states could ensure that space objects are better protected against cyberattacks. This would expand the responsibility of states from merely launching space objects to ensuring their ongoing cybersecurity, addressing the need

for a clearer framework for accountability in space.

Customary international law, including the prohibition of harmful interference, can also be extended to include cyber threats to space infrastructure. This would enable states to take legal action against cyber operations that disrupt or damage space assets, using existing legal norms to address cyber-attacks in space. The principle of *lex specialis* further emphasizes that space-specific regulations should take precedence over public international laws, necessitating the development of dedicated cyber regulations to address these emerging challenges. The due diligence standard, already present in international law, can be adapted to cybersecurity, requiring states to take reasonable measures to prevent cyberattacks originating from their territory. By integrating cybersecurity obligations into the due diligence framework, states would be obligated to monitor and prevent cyber threats, ensuring they uphold their international responsibilities.

These frameworks, when revised to the unique novel threats posed by cyber operations in space, offer a comprehensive approach to addressing the issue of international responsibility for space cybercrimes. However, more detailed guidance is needed to apply these principles effectively, especially as cyber threats evolve alongside space technologies. For this purpose, several measures need to be implemented to effectively address the remaining gaps in international responsibility for space cybercrimes. These measures should draw from foundational legal frameworks while adapting to the unique challenges posed by cybersecurity in outer space.

6.2. Developing new cybersecurity standards and mechanisms for space systems

Establishing comprehensive cybersecurity standards is essential for protecting space systems. Specific standards shall include mandatory encryption for satellite communications to prevent unauthorized access, secure software development practices to minimize vulnerabilities, real-time monitoring coupled with detection capabilities, prompt response protocols, and stringent authentication measures to ensure that only authorized entities can access critical systems. Furthermore, these standards should be incorporated into international agreements to establish a baseline for cybersecurity across all space infrastructures, thereby ensuring a consistent level of protection and mitigating the risk of breaches.

A specialized attribution mechanism with an international body dedicated to cyber-attack attribution should be established and equipped with advanced forensic technologies and satellite monitoring systems. This body would be responsible for

tracing cyber intrusions, using criteria such as attack origins, behavioral patterns, and digital footprints to identify and attribute cyberattacks to specific states or non-state actors. Such a mechanism would enhance the ability to hold responsible parties accountable and improve the overall security of space operations.

An addition of a protocol to the Registration Convention to include cybersecurity protocols can further bolster space asset protection. Mandating that space operators demonstrate compliance with basic cybersecurity requirements before registering space objects. This would involve implementing encryption for satellite communications, conducting regular cybersecurity audits, and performing vulnerability patching to address potential weaknesses. By integrating these cybersecurity obligations into the registration process, new space assets would be better protected against cyber threats from the outset.

Formalizing cybersecurity responsibilities in multinational space operations is crucial for ensuring collaborative security efforts. International agreements should define specific cybersecurity roles and responsibilities for each party involved in the operation of space assets. This would include securing respective segments of the network or infrastructure, establishing joint incident response teams, and conducting regular cybersecurity drills and assessments. Clear delineation of these responsibilities would help ensure the entire system's integrity and facilitate coordinated responses to cyber incidents.

States should be required to monitor and take preventive measures against cyber threats within their jurisdiction, involving developing early warning systems to detect potential cyberattacks, enforcing cybersecurity compliance for all space operators, and implementing robust national cybersecurity frameworks. States should be held legally accountable for failing to adhere to these standards, ensuring they actively contribute to preventing cyber threats that could impact space operations.

Regulating non-state actors is crucial to address the growing participation of private entities and individuals in space-related cyber activities. Extending international law to include provisions specifically targeting these actors is essential. States should ensure private space operators implement adequate cybersecurity measures, conduct background checks on key personnel, and cooperate in international cybercrime investigations. Moreover, states could be held responsible for negligence if they fail to control non-state actors using their territory to launch cyberattacks on space infrastructure.

Classifying space assets as critical infrastructure under international law would

provide enhanced protection. Designating certain space assets—such as satellites supporting global communications and defense—as critical infrastructure would mandate higher cybersecurity standards and impose stricter penalties for cyberattacks targeting these systems, ensuring that vital space resources are adequately safeguarded.

7. Conclusion

The growing dependence on space systems for global communication, navigation, defense, and scientific exploration underscores the need to address cyber threats to space systems, humans, and the environment in space and on Earth. As humans expand into outer space, addressing space cyber security and preventing space cybercrimes, as well as human safety and environmental sustainability, are not mere technical challenges but critical legal obligations.

Existing international legal frameworks, while foundational, were not designed to tackle the novel threats posed by cyber-attacks on space infrastructure, causing international crimes. To combat these threats and fill the legal lacunas, space actors can build a more comprehensive international legal regime by extending the principles from existing international legal frameworks such as the Outer Space Treaty, ARISWA, the registration convention and the customary international Norms. Developing clear space cyber terms, definitions, updated attribution mechanisms and formalizing the roles of space cyber security and the security of humans and the environment from space cybercrimes in international agreements are necessary steps to prevent and mitigate the threats posed by space cybercrimes. Moreover, regulation of non-state actors in space and international collaboration is crucial for a global effort to uphold safety, security and sustainability in space and on Earth.

将责任原则应用于构成国际犯罪的空间网络攻击

Manzoor Hassan, Kuan Yang

摘要:对空间系统的网络攻击可能引发广泛的混乱、人道主义灾难,并对空间基础设施、人类生命以及地球和空间环境造成严重损害。确定这些攻击是否构成国际不法行为以及是否达到国际罪行的门槛,是一项复杂的挑战,需要采用多学科方法。这需要融合国际空间法、网络法、国际公法以及习惯法规范。定义空间网络不法行为的模糊性、攻击者的匿名性以及多重法律框架的重叠,使得这一判定更加困难。当前规范空间活动、网络行动和国家责任的法律框架并未针对空间网络攻击的复杂性及其独特的技术和法律挑战而设计。这一空白强调了重新思考和重新定义国际责任的必要性,以更好地应对这些不断演变的威胁。因此,本研究探讨了空间网络攻击背景下的国际不法行为。通过分析近期空间网络攻击的实例,本研究阐明了归因和适用法律原则的实际挑战。本研究主张对现有国际责任制度进行重大改进,明确预防和打击网络威胁的义务,包括来自私人行为者的威胁。这包括定义何为空间网络不法行为、澄清法律义务、建立新的空间网络攻击归因机制,并促进更精确的问责。

关键词:空间网络攻击、国际犯罪、国际责任、《国家对国际不法行为的责任条款》(ARISWA)、空间法

论我国人工智能的风险规制路径：以负面清单制度为中心

徐世泽

摘要：数字时代，人工智能立法已经成为当前国际规则博弈的关键领域。欧盟在全球范围内率先通过欧盟《人工智能法》，创设“以风险为基础”的分级规制路径，区分了三种形式的人工智能系统，构成欧盟人工智能立法的底层逻辑。此种风险规制路径存在风险标准界分不明和僵化适用的困境。鉴于此，我国《人工智能法》应坚守促进型立法的立场，以负面清单制度建构自主的人工智能规制框架，而不应简单效仿欧盟。具体来说，国家人工智能主管机关对负面清单内的人工智能产品服务实施许可管理，对负面清单外的产品服务实施备案管理，以负面清单制度为中心全面赋能人工智能高质量发展。

关键词：人工智能法 负面清单 风险规制 促进型立法

一、问题的提出

人工智能作为第四次工业革命的核心技术已经得到社会的普遍关注。蛇年春节期间，中国人工智能企业深度求索（DeepSeek）发布的开源模型 DeepSeek-R1 备受全球瞩目，其低成本开源优势和出色的自然语言生成能力颠覆了国际社会对人工智能研发“高投入、长周期”的固有认知，是对全球人工智能发展格局的一次重塑。¹随着 DeepSeek 越来越多地被应用于公共决策和社会治理领域，如何确保其被正确使用，从而避免对社会造成系统性的风险，值得重视。事实上，囿于人工智能系统固有的算法黑箱、数据造假、数字鸿沟及难以问责的特征，²其可能被滥用于实施操纵剥削、数据监控和社会控制等行为，³亟需公法和私法的协同治理。鉴于人工智能附随的各种风险，国际组织和各国政府已经开始着重关注人工智能的风险规制路径。不论是 2024 年联合国大会一致通过《抓住安全、可靠和值得信赖的人工智能系统带来的机遇以促进可持续发展》决议，还是同年欧洲议会高票通过欧盟《人工智能法》，抑或是美国特朗普政府最新发布《消除美国在人工智能领域领导地位的障碍》行政命令，欧美国家和国际组织都在密集开展人工智能治理活动，旨在通过构建一套人工智能的规制框架来推动人工智能在法治的轨道上健康发展。我国一贯高度重视人工智能这一新兴领域的法律治理。早在 2016 年，全国人大常委会就通过《国民经济和社会发展第十三个五年规划纲要》，

¹ 参见吴琼：《中国用行动证明开放创新的浪潮无法阻挡》，载《法治日报》2025 年 2 月 10 日。

² Laurin B. Weissinger, *AI Complexity and Regulation*, The Oxford Handbook of AI Governance, 2022, pp.619-648.

³ 参见李训虎：《刑事司法人工智能的包容性规制》，载《中国社会科学》2021 年第 2 期。

要重点突破新兴领域人工智能技术。第十四届全国人大常委会通过 2024 年度立法工作计划,指出要制定网络治理和人工智能健康发展等方面的立法项目,视情安排审议。可以看到,作为未来人工智能领域的基础性立法,我国《人工智能法》正呼之欲出。⁴

在人工智能立法具体进程中,我国需要构建一套怎样的人工智能规制框架?这是摆在理论界和实务界面前亟待解决的问题。欧盟《人工智能法》中以“以风险为基础”的分级规制模式是否值得我国借鉴?我国如何创设自主的人工智能规制框架?本文提出以负面清单制度构建自主的我国人工智能的规制框架,提升政府在人工智能治理和数字治理中的规制能力,⁵旨在为我国数字经济注入“新质生产力”的活力,⁶并以中国智慧和方案积极参与全球人工智能治理。⁷

二、我国立法对人工智能规制的促进型立场

我国正处于人工智能立法的关键时期,虽尚未制定统一的《人工智能法》,但已形成了以部门规章和司法解释为主干,配合地方性法规的多层次人工智能规制体系,彰显了我国在人工智能规制领域秉持的促进型立场。

一方面,以部门规章和司法解释为基础的立法例为不同领域的人工智能系统提供了合法性基础。在部门规章层面,国家互联网信息办公室出台的《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》成为规范特定领域人工智能研发者和部署者合规的制度。其中《生成式人工智能服务管理暂行办法》是我国首个针对生成式人工智能产业的规范性政策,⁸该办法第三条明确规定,国家采取有效措施鼓励生成式人工智能创新发展,对生成式人工智能服务实行包容审慎和分类分级规制,充分表明了国家促进生成式人工智能发展的决心。⁹此外,2024 年修订的《互联网信息服务管理办法》虽未专门针对人工智能的具体条款,但其对互联网信息服务提供者的要求,如信息安全保障措施、信息内容合法性等,可以间接规制涉及人工智能的服务。在司法领域,最高人民法院系统梳理智慧法院建设的实践经验,总结我国人工智能司法应用的理论依据和应用场景,2022 年《关于规范和加强人工智能司法应用的意见》,明确人工智能司法应用应遵循安全合法、公平公正、辅助审判、透明可信和公序良俗等五项基本原则。值得关注的是,2024 年全国网络安全标准化技术委员会发布的《人工智能安全治理框架》1.0 版明确提出促进人

⁴ 参见张凌寒:《中国需要一部怎样的人工智能法?——中国人工智能立法的基本逻辑与制度架构》,载《法律科学(西北政法大学学报)》2024 年第 3 期。

⁵ 参见莫纪宏:《论数字权利的宪法保护》,载《华东政法大学学报》2023 年第 4 期。

⁶ 参见莫纪宏:《论新质生产力基础上的“新质民主”与“新质法治”的辩证统一》,载《政治与法律》2023 年第 6 期。

⁷ 参见邵长茂:《人工智能立法的基本思路》,载《数字法治》2023 年第 5 期。

⁸ 《生成式人工智能服务管理暂行办法正式施行》, http://www.news.cn/2023-08/16/c_1212256586.htm, 2025 年 2 月 15 日访问。

⁹ 参见吴大奎:《基于生成式人工智能服务风险的分级法律规制研究》,载《智慧法治》2023 年第 3 卷。

工智能创新发展为第一要务，构建各方共同参与、技管结合、分工协作的治理机制来推动其健康发展和规范应用。

另一方面，除了在国家层面通过一系列的人工智能规制立法，一些地方立法机关开始根据本行政辖区内人工智能产业实践的发展需要进行先行先试，为国家统一人工智能立法提供探索经验。2022 年，深圳市人大常委会通过《深圳经济特区人工智能产业促进条例》，上海市人大常委会通过《上海市促进人工智能产业发展条例》，都旨在促进地方人工智能产业高质量发展，推进人工智能对经济社会领域的深度融合应用，率先在全国范围内创设具有地方特色的人工智能治理方案。关于我国人工智能规制的规范性法律文件参见下表。

表：关于我国人工智能规制的规范性法律文件

规范性文件	发布主体	生效时间
《互联网信息服务管理办法（2024 修订）》	国务院	2024 年 12 月 6 日
《非经营性互联网信息服务备案管理办法（2024 修订）》	工业和信息化部	2024 年 1 月 18 日
《生成式人工智能服务管理暂行办法》	国家互联网信息办公室	2023 年 8 月 15 日
《互联网信息服务深度合成管理规定》	国家互联网信息办公室	2023 年 1 月 10 日
《互联网信息服务算法推荐管理规定》	国家互联网信息办公室	2022 年 3 月 1 日
《科技伦理审查办法（试行）》	科技部、教育部、工业和信息化部等部门	2023 年 12 月 1 日
《关于规范和加强人工智能司法应用的意见》	最高人民法院	2022 年 12 月 8 日
《深圳经济特区人工智能产业促进条例》	深圳市人大常委会	2022 年 11 月 1 日
《上海市促进人工智能产业发展条例》	上海市人大常委会	2022 年 10 月 1 日

由此可见，不论是中央层面的国家立法，还是地方层面的地方性法规，我国立法者采用了以问题为导向的立法模式，¹⁰在人工智能规制领域贯彻促进型的立场，积极统筹人工智能的安全性和效率性，按照鼓励创新的原则，对人工智能产业实行包容审慎规制。然而，由于我国尚无在基本法律层面确立人工智能的一体式规制方案，可能导致不同立法主体出台的法律规范性文件设立各异的人工智能规制模式，甚至在出现全新的人工智能系统时可能面临规制方案缺失的困境。基于此，我国可以学习欧盟《人工智能法》以风险为基础的分级分类规制思路，结

¹⁰ 参见董新义、梅贻哲：《“人工智能法总则”建构原则与理念——欧盟立法经验之镜鉴》，载《数字法治》2024 年第 2 期。

合我国国情，适时在未来人工智能立法中确立统一的人工智能规制模式，即以负面清单制度全面推动我国人工智能高质量发展。

三、域外经验：欧盟“以风险为基础”的人工智能分级规制路径

人工智能是一个快速发展的技术序列，如何通过统一立法对各式各样的人工智能系统进行规制，这是个法律技术性很强的世界共同难题。¹¹欧盟认识到“以风险为基础”的人工智能规制战略对于保证欧盟价值观实现的重要性，¹²以人工智能发生危害的可能性和危害的严重性原则上区分了三种形式的人工智能系统，其核心宗旨是确保价值维度的相称性和有效性，¹³在促进创新和就业的同时，促进对民主和法治、环境保护等基本权利的保障，使欧盟成为采用可信人工智能的领导者。¹⁴以风险为基础的人工智能治理构成欧盟人工智能立法的底层逻辑，¹⁵具体来说包括绝对禁止的具有不可接受风险的人工智能系统、高风险人工智能系统及其他人工智能系统，其中具有不可接受风险的人工智能系统在实践中被完全禁止，高风险用途的人工智能系统将受到执法部门的严格规制，而风险程度较低的其他人工智能系统将受到相对宽松的法律规制，参见下表。

表：欧盟“以风险为基础”的人工智能分级规制模式

风险程度	具体应用	规制措施
绝对禁止的具有不可接受风险的人工智能系统	严重威胁人的安全、健康和其他基本权利的人工智能系统； 利用潜意识操纵和欺骗用户； 利用个人因特定社会背景而具有的任何弱点的人工智能系统； 社会评分智能系统； 实时远程识别智能系统	禁止； 若违法，处以前一财年全球营业额最高6%的罚款
高风险用途的人工智能系统	关键基础设施； 教育和职业培训领域； 就业、工人管理和自由职业领域； 获得和享受基本的社会福利领域； 移民、庇护和边境管制领域； 刑事追诉领域；	上市前严格监控； 开展风险安全评估； 提供高质量数据； 处理行为可追溯； 提供关于系统和目的的所有必要信息；

¹¹ 参见蔡培如：《欧盟人工智能立法：社会评分被禁止了吗？》，载《华东政法大学学报》2024年第3期。
¹² White Paper on AI - A European approach to excellence and trust, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065>.
¹³ Nikolinakos, N. T., *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies*, 2023, pp.354-362.
¹⁴ Preface, paragraph 1 of Artificial Intelligence Act.
¹⁵ 参见陈吉栋：《公私交融的人工智能法》，载《东方法学》2024年第2期。

	司法审判领域； 人脸生物识别领域	上市后持续监督； 向用户提供充分且 清晰的信息； 严格执法和处罚
有限风险或最小 风险的其他人工 智能系统	允许自由使用人工智能的电子游戏 或垃圾邮件过滤器等应用	透明度义务 基本不做限制

（一）绝对禁止的具有不可接受风险的人工智能系统

欧盟《人工智能法》在第二章归纳了绝对禁止的人工智能实践所产生的不可接受的风险。根据该法第 5 条，欧盟禁止在以下四类情况下投放、提供或应用人工智能系统，因为这些人工智能系统将导致对个人权利和集体自由的严重侵犯，明显违背欧盟人权优先的基本立场。¹⁶第一，禁止利用潜意识操纵、欺骗和控制人工智能使用者。由于该类人工智能系统的目的在于损害个人或集体作出知情决定的能力，从而使一个人作出该人本来不会作出的决定，其方式可能会造成该人和集体权利的重大损害。第二，禁止在市场上投放和部署因个人年龄、残疾和特定社会背景不同而存在的任何脆弱性的人工智能系统，其目的或效果是实质性地扭曲该人的特定行为，并对当事人权利构成严重损害。第三，禁止利用人工智能系统根据个人或群体的社会行为或已知、推断或预测的行为特征来进行社会评分，同时此种待遇与最初收集数据的背景无关或此种待遇是与其社会行为严重不相称的。第四，禁止在市场上投放或使用应用生物特征识别技术的人工智能系统来从事以下特定行为，包括调查某人犯罪的可能性、创设人脸识别数据库和实时远程识别系统以及对人进行分类，除非在特定情况下对实现目标而言是绝对必要且相称的。例如，为了保证执法目的的实现，可以在以下情况下使用远程识别系统，包括有针对性地合法搜寻特定的绑架、贩卖人口、性剥削受害者或其他失踪人员；防止恐怖袭击威胁或防止其他自然人的生命安全可能遭到具体、重大和紧迫的威胁；以及出于公共安全目的对犯罪嫌疑人进行有效定位，以便对其进行刑事调查、起诉或执行刑事处罚。

（二）高风险人工智能系统

欧盟《人工智能法》在第三章创设了研发、部署和应用高风险人工智能系统的法律规则。该法第 6 条指出，只要人工智能系统满足以下两个条件便应该被视为高风险系统，包括人工智能系统被应用于特定产品的安全部件，或者人工智能系统本身就是附件一所列欧盟统一立法所涵盖的产品。此外，该法附件三所述的

¹⁶ Evgeni Aizenberg, Jeroen van den Hoven, *Designing for human rights in AI, Big Data & Society*, 2020.

人工智能系统也应被视为高风险系统,因为在以下八类情形中部署人工智能系统都有可能对公民的基本权利和自由造成高度的风险。具体而言,包括第一类关键基础设施领域,即人工智能系统作为组件被用于道路交通或供水、供气、供热、供管理以及其他关键公共基础设施。第二类是教育和职业培训领域,即当人工智能系统被应用于教育资源的录取和分配、评估一个教育机构的师资水平以及用于检测学生考试期间是否存在违规行为。第三类是就业和工作领域,特别是当人工智能系统被用于招聘或选拔雇员,以及有针对性的发布招聘广告和过滤求职申请。第四类是获得基本社会福利,即当人工智能用来评估自然人获得基本公共援助福利和服务包括医疗保健服务的资格。第五类是刑事追诉领域,即当人工智能用以评估自然人成为刑事案件中被害人的风险或评估自然人犯罪或再犯罪的可能性。第六类是移民、庇护和边境管制,即当人工智能用于评估当事人入境可能引发的安全风险、非正常移民风险或其他公共健康风险。第七类是司法领域,即当人工智能系统被应用于协助司法机关研究和解释事实和法律,并将法律适用于一系列具体事实。第八类是远程生物识别领域,例如根据个人敏感信息进行数据分析以确定特定的自然人。

值得注意,该法第6条第3款专门创设了高风险人工智能的例外制度,即人工智能系统在以下四类情况下不对当事人的安全、健康或基本权利造成重大损害风险,则不应被视为高风险系统。具体包括:人工智能系统旨在执行范围狭窄的程序性任务;人工智能系统旨在改进先前完成的人类活动的结果;人工智能系统旨在检测决策模式或偏离先前决策模式的情况;人工智能系统旨在执行与附件三所列用例有关的评估的准备任务。应当说,上述条款构成了高风险人工智能系统的分类识别与判断审查标准。¹⁷在此基础上,该法在本章第二节进一步细化了高风险人工智能研发者、部署者和使用者的合规义务,包括建立风险管理系统,进行数据管理,保证透明度、确保人类的监督以及保证稳健性和网络安全等义务。

(三) 其他人工智能系统

不符合高风险人工智能定义的人工智能系统通常视为有限风险或最小风险的人工智能系统。欧盟《人工智能法》虽未明确专门规定其他人工智能系统的研发、部署和使用规则,但关于人工智能系统的共同规则须得到普遍遵守。例如该法第50条规定,人工智能系统的提供者和部署者被课以透明度义务,即提供者应确保意图与自然人直接互动的人工智能系统的设计和开发方式应使有关自然人知道他们正在与一个人工智能系统互动,信息披露最迟应在首次互动或接触时以清晰可辨的方式提供给有关自然人,并符合无障碍要求。此外,该法鼓励有限或最低风险人工智能系统的提供者积极制定行为守则,在明知自愿的基础上主动

¹⁷ Girasa, R., Scalabrini, G., *Regulation of Innovative Technologies Blockchain, AI and Quantum Computing*, 2022, pp.73-77.

适用于高风险人工智能系统的特定强制性要求，并结合现有的技术方案和行业实践，根据本系统的预期目的和所涉及的风险进行合理调整。例如，积极适用欧盟《可信人工智能道德准则》、环境可持续性和人工智能系统包容性和多样化的要求以及在部署人工智能系统时让企业和民间社会组织、学术界、工会和消费者保护组织等利益攸关方参与。

（四）小结：欧盟人工智能风险规制模式之检视

作为全球首个对人工智能进行全流程规制的立法，欧盟《人工智能法》旨在通过创设以风险为基础的分级分类规制模式，在欧盟层面营造高标准值得信赖的、以人为本的人工智能生态环境，确保人工智能技术在开发和应用的全过程都遵循伦理规范，实现可信人工智能，¹⁸对国际人工智能治理产生深远影响。该法首先明确禁止绝对禁止的具有不可接受风险的人工智能系统，其次重点规制高风险人工智能系统，确立了高风险人工智能研发者、部署者和使用者应履行的一系列合规义务。该法强调需要在人工智能系统全生命周期内建立和维护一套高效的风险管理系统，保证系统的稳健性和安全性，避免出现重大系统运行故障问题。¹⁹最后，该法对具有有限风险或最低风险的其他人工智能系统作了原则性规定，为新技术和新产业保留充分的发展空间。²⁰值得注意的是，该法也专门确立了“通用人工智能模型”风险规制的法律路径，即通用人工智能模型提供者应积极履行模型安全评估和报告义务，包括根据体现最新技术的标准化协议进行大模型安全评估和及时编制大模型的技术文件并向欧盟人工智能办公室报告。

由此可见，随着国际战略博弈和人工智能规则霸权不断加剧，²¹欧盟率先通过人工智能法，其意图在于在这一新兴科技前沿领域引领国际规则制定，试图重现欧盟施行《通用数据保护条例》对全球规则治理产生的布鲁塞尔效应。²²然而，当前欧盟通过《人工智能法》仅仅是其在人工智能法律规制道路上迈出的第一步，²³后续的配套技术标准和成员国如何有效落实规制要求尚未正式启动。当前，各方对该法的质疑主要包括以下三个方面。

第一，欧盟《人工智能法》中“以风险为基础”的分级规制要求存在标准界分不科学与僵化适用的现实困境。欧盟议会在审议人工智能立法草案初期，主要关注作为产品构成的人工智能系统，而非作为独立系统的人工智能，如旨在辅助或替代人类决策的自动化决策智能系统。这些具有独立系统的人工智能与作为产品

¹⁸High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI, https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

¹⁹ Park, E., *The AI Bill of Rights: A Step in the Right Direction*, Orange County Lawyer Magazine, 2023, p.25.

²⁰ 参见高艳东：《数字法治中的立法消极与司法能动》，载《南京社会科学》2024年第2期。

²¹ 参见李猛：《人类命运共同体视角下人工智能风险全球治理的国际法规制路径探究》，载《宁夏社会科学》2024年第2期。

²² 参见俞胜杰：《通用数据保护条例第3条(地域范围)评注——以域外管辖为中心》，载《时代法学》2020年第2期。

²³ Neuwirth, R. J., *The EU AI Act - Regulating Subliminal AI Systems*, Routledge, 2023, pp.23-31.

的人工智能系统所产生的风险存在显著差异。如果说产品类人工智能所引致的风险主要涉及基于产品责任的私法风险，那么具备独立系统的人工智能则可能直接影响部署者和使用者的公法性权利。对于这些能够独立运作的人工智能系统在部署和应用于教育、医疗、就业和司法等领域时不仅需要符合形式上的合法性，还需符合实质上的合法性，即不得对个体的基本权利产生不利影响。其次，在重要性问题的评估上，该法专门区分了人工智能作为产品所引发的风险与人工智能系统直接引致的基本权利风险，并基于此确立了不同的评估方法。对于主要关涉基本权利的人工智能系统，除了远程生物识别系统之外，该法要求其部署者进行基于内部控制的符合性评估，并不要求独立第三方机构进行评估，换言之由该人工智能系统部署者进行自我评估。相反，对于产品类的人工智能系统在投放市场前需要接受独立第三方机构的安全评估并获得相应的通过证书。该法将这两类不同的风险一概纳入到统一立法中进行规制，凸显欧盟在人工智能立法领域的实用立场，因为事实上这两类风险在性质上存在显著差异。一般而言，产品类的人工智能风险可以基于市场秩序有无遭到损害以及损害到何种程度进行定量化的计算，而基本权利类的人工智能风险往往被认为是特定政治经济和意识形态下的价值衡量，其损害程度往往不能简化为具体的财产性损失，并且也难以整体计算。由此可见，欧盟《人工智能法》未针对产品类人工智能风险和基本权利型风险创设精细化的规制标准，不能充分反应人工智能技术的复杂性，这种基于风险的分级规制规则在实际操作中可能难以落实。在实践中随着算法、算力和基础大模型的不断突破，人工智能发展存在高度不确定性，技术的升级迭代会打破不同风险等级之间的界限，导致某项人工智能技术在同一场景中可能同时受到多重规制，体现的法律效果是“九龙之水”，导致该项风险分级规则“失效”。

第二，由于欧盟《人工智能法》在其立法过程中存在透明度不足的问题，欧盟法院难以有效结合立法文本进行目的解释，存在较大的法律续造空间。2024年3月欧洲议会快速通过由于经三方会议达成一致的人工智能法妥协草案，这种立法程序使得嗣后法律解释变得复杂，因为公众难以进行立法史溯源，难以基于历史解释进路了解立法者的立法意图，而这给作为司法解释者的欧盟法院留下了相对较大的解释空间，其司法判决在一定程度上可能偏离原初立法者的立法原意。以欧盟《人工智能法》规制的禁止性风险人工智能系统为例，不同的解释往往使得该类人工智能系统的风险分级发生重大变化，给其后续部署实施带来极大不确定性。在数字时代，来自抖音、淘宝和京东等大型互联网平台的各种个性化智能推荐系统可以有效捕捉用户的兴趣点，从而增强平台的用户粘性，这种商业模式可以促进产品要素流通和符合市场需求。如果对该法禁止性风险人工智能系统的相关条款进行严格解释，将这些智能推荐类人工智能系统评价为可能严重影响个人潜意识或损害特定弱势群体权利，那么许多正常的数字化商业模式将受到极大限制。由此可见，欧盟这种基于风险的规制要求还存在法律确定性不足的问题，

其中很多规定是否科学可靠有赖实践的进一步检验,也给司法机关创设了很大的弹性解释空间,给该法的落地实施带来较大不确定性。

第三,欧盟《人工智能法》的分级规制要求可能加剧中小微人工智能研发者的合规负担,给广大中小微企业和初创公司设置了较高的合规门槛,引发妨碍人工智能产业发展的担忧。以通用人工智能模型为例,该法要求所有通用人工智能系统的部署者都应履行向下游主体披露技术性文件并承担风险减轻的义务。然而,基于通用人工智能天然的多用途性和功能开放性,这种要求事实上是让通用人工智能模型提供者对其系统在不同应用场景中可能引致的法律风险进行事前预防。尤其是在当前很多通用人工智能大模型可以免费使用或开源下载的背景下,这种裹挟“预防性”特质的概括式要求直接加重了大模型提供者和部署者的负担,导致其不再免费开源或仅向付费用户开放。与此同时,该法要求向通用人工智能下游主体提供技术性文件的义务可能会对通用人工智能模型部署者的商业秘密造成威胁,导致其具有重要商业价值的算法和代码被外界获取。有不少学者批评认为,该法实施太过严苛,可能阻碍欧盟在全球经济地位提升的绊脚石。²⁴由此可见,欧盟目前所采取的规制措施呈现过度前瞻性,可能会加剧欧盟新兴人工智能产业的合规负担,进而对欧洲人工智能技术的创新研发造成严重阻碍。

四、构建以负面清单为核心的我国人工智能风险规制制度

(一) 欧盟人工智能治理规制框架对我国的启示

从域外经验来看,作为全球范围内首部人工智能综合性立法,欧盟《人工智能法》为确保对公民健康、安全以及《欧盟基本权利宪章》中规定的基本权利的高度保障,促进以人为本、值得信赖的人工智能应用,创设一套“以风险为基础”的人工智能规制框架。然而,基于不同人工智能技术算法原理的异质性特征,其运行底层逻辑千差万别,如何在不同应用场景中规制人工智能可能引致的风险,尚无普世性的一体规制框架,而智能场景化的类案分析。考虑到当前我国人工智能产业发展的现状,我国不应直接照搬欧盟的人工智能治理框架,而应在立足本土的基础上积极借鉴欧盟的先进立法思路,²⁵构建自主的人工智能规制框架。

首先,我国和欧盟在人工智能技术和产业基础方面存在较大差异。以人工智能初创企业数量为比较标准,虽然欧洲单个国家缺乏全球竞争力,但欧洲加强其数字单一市场,将成为国际人工智能竞争主要参与者。²⁶反观我国,由于人工智能技术一直是中美贸易战的战略焦点,美国长期打压我国科技企业,我国智能计

²⁴ 参见沃尔夫冈·多伊普勒:《欧盟人工智能法案的背景、主要内容与评价——兼论该法案对劳动法的影响》,载《环球法律评论》2024年第3期。

²⁵ 参见董新义、梅贻哲:《“人工智能法总则”建构原则与理念——欧盟立法经验之镜鉴》,载《数字法治》2024年第2期,第195-210页。

²⁶ 《欧洲在全球人工智能竞争中的地位及战略分析》, <http://lib.ia.ac.cn/news/newsdetail/1724>, 2025年2月15日访问。

算面临发展困境。国内智能计算生态薄弱，尚未形成一个有竞争力的人工智能技术体系，并且人工智能应用主要集中在互联网行业，未实现全行业的整体赋能。其次，基于风险的欧盟规制路径，其意图在于保障符合欧盟价值观的人工智能系统之应用，不符合我国人工智能规制的基本立场。欧盟将风险划分的义务更多地赋予人工智能研发者和部署者，将诱发相关主体的防御性合规倾向，即人工智能研发者和部署者为避免不可预见的法律责任，倾向于按尽可能高的等级来评判该系统的安全风险以确保合规冗余度。此类“风险上浮”机制虽在一定程度上发挥了风险预防的功能，但受制于风险防范标准的不确定性，人工智能企业可能面临矫枉过正的合规困境，进而形成过度抑制技术创新的制度效应。最后，以风险为基础的人工智能规制路径构成欧盟人工治理范式的底层逻辑，其宗旨在于促进以人为本、值得信赖的，同时确保对健康、安全和基本权利的高度保障，使公民免受欧盟内人工智能系统之害，其战略意图在于利用自身庞大市场规模优势，通过风险分级的规制模式率先实施“遏制型”战略参与人工智能国际治理的规则竞争，令其在全球人工智能规制方面获得先发优势。²⁷欧盟倾向于将中国视为长期竞争对手和安全风险对象，尽管在经贸领域与中国开展伙伴关系合作，但基于技术主权的立场通过设定法律规则保护人工智能技术产业链和价值链的安全，²⁸与我国坚持发展与安全并重的理念存在重大分歧。由于当前国际地缘政治冲突与地缘经贸关系相互叠加、不断蔓延，²⁹我国应在总体安全观基础上坚持中国特色国家安全道路³⁰，具体到人工智能治理领域，我国应统筹安全和发展，在安全价值的基础上促进人工智能，同时又以发展促安全。

由此可见，当前值得提倡的做法是我国合理参考欧盟《人工智能法》“以风险为基础”的分级规制制度，在立足我国国情的基础上，提炼既往我国对人工智能相关领域规制的治理经验，确立适合本土人工智能实践发展的规制框架，即通过构建科学高效的负面清单制度逐步推动人工智能高质量发展。

（二）我国人工智能立法中负面清单制度的效能与作用

在我国人工智能立法中构建负面清单制度，这是推动人工智能与数据要素成为先进生产力的重要内容，在实践中具有重大的制度效能与积极作用。

首先，相较于欧盟《人工智能法》所确立的以风险为基础的二元规制框架，负面清单制度对清单内的人工智能研发、提供和部署等活动设置较高的准入门槛和合规义务，这可以为人工智能产业提供更高的确定性，为其提供持续稳定的信赖预期，保证企业能够在清晰的规制框架内从事合规经营，减少人工智能技术应

²⁷ 参见施雯、缪其浩：《从两极到三强：欧盟人工智能的全球竞争战略分析》，载《中国科技论坛》2022年第6期。

²⁸ 参见宋黎磊、戴淑婷：《科技安全化与泛安全化：欧盟人工智能战略研究》，载《德国研究》2022年第4期。

²⁹ 参见刘敬东：《全球经贸关系演变中的国际法治危机及其应对》，载《中国法学》2023年第3期。

³⁰ 参见莫纪宏：《构建中国特色国家安全法治理论》，载《人民日报》2024年4月。

用分级分类标准时无法达成一直结论的情况。监管机构应在负面清单制度的基础上,建立定期评估机制确保清单内容的动态适应性,而在清单具体事项的制定上,应当尽可能凝聚各种利益相关者的最大共识,以动态均衡的视角回应人工智能实践中的迭代和更新。与此同时,此种制度设计可以与我国现行《促进和规范数据跨境流动规定》所构建的负面清单制度形成规范协同,在严守国家安全红线的基础上打造创新友好型监管范式,既保障了网络与信息法治体系的协同性,又通过压缩监管自由裁量空间为人工智能新技术创新保留了必要的制度弹性。

其次,负面清单制度有利于降低人工智能研发者和部署者的创新风险,化解他们在法律空白领域的风险性。人工智能领域近年来正在迎来一场由生成式人工智能大模型引领的爆发式创新,各种通用人工智能大模型如雨后春笋般地集体涌现,极大地推动了人工智能与实体经济的融合发展落实负面清单制度。通过在法治轨道上规范政府的行政审批行为,遏制自由裁量权等,有利于理清市场和政府的关系,让市场机制发挥主导作用,减少人工智能企业的创新风险。事实上,如果不借助负面清单的规制模式,这些人工智能企业所面对的往往是种类繁多、内容冗杂的各种行政审批许可等公法上的要求,无法让企业形成稳定的制度预期。与此同时,在人工智能领域引入负面清单制度亦是传统私法自治精神的重要体现。私法自治原则确认市场主体在法定范围内自主决定,而不受任何非法的干涉。负面清单模式符合“法不禁止即自由”的私法自治精神,为大量人工智能企业自主创新提供了广阔的实验空间,也彰显了我国一贯坚持的人工智能促进立法理念。

最后,作为一种快速适应技术迭代的动态治理范式,负面清单制度契合人工智能治理的底层逻辑,与我国一直推行的包容审慎规制具有理念上的一致性。传统的规制体系和规制治理手段,已不能适应人工智能产业的发展要求。负面清单制度旨在追求效率与安全的动态平衡,其彰显的是包容审慎规制理念,要求政府给予新技术和新业态必要的发展事件与试错空间,并根据公共风险的大小进行适度适时干预,以激励市场与政府探索创新,确保行政便宜性与最佳性有机统一。通过负面清单,划分已知风险与未知风险,构建事前预防和事中事后相结合的适应性治理机制,综合运用技术、技术标准、市场、法律和政策等多元治理工具,实现人工智能治理范式的革新,促进科技向上向善发展。

（三）我国人工智能立法中负面清单制度的规范建构

负面清单制度缘起于 2013 年中国（上海）自由贸易试验区对外商投资管理的创新模式,其是指通过采取负面清单的形式,让相关的市场主体在工商业活动中享有较大程度的自由权³¹,体现了由政府全面干预到市场意思自治的法治精神。作为一种市场准入机制,负面清单奉行“非禁即入”的市场主导理念,依法保障各类市场主体平等进入清单之外领域,对于进一步激发市场活力、遏制行政权扩张

³¹ 参见张淑芳:《负面清单管理模式的法治精神解读》,载《政治与法律》2014 年第 2 期。

以及促进营商环境法治化具有重要的现实意义。

虽然我国过去几年在人工智能算法、算力和数据方面取得了重要进展，但依然面临诸多发展困境。深圳市人工智能行业协会发布的《2023 人工智能发展白皮书》指出，我国人工智能行业起步比较晚，我国在人工智能基础算法创新、底座大模型能力和底座大模型训练算力方面发展滞后，并且在智能芯片和开源框架方面尚未形成繁荣的自主生态。我国的优势在于实体经济，制造业全球产业门类最齐全，体系最完整，特点是场景多、私有数据多。我国应加快推进《人工智能法》的出台，构建自主的人工智能治理体系。聚焦到人工智能规制领域，通过否定式列表的形式标明重大风险和高风险人工智能研发者和部署者进入的特定领域，对人工智能负面清单内的系统实施行政许可，其余的人工智能产品和服务则不再加以一一界定，将市场自由让渡给私法自治，确有需要的实施备案管理。例如在刑事司法领域，人工智能的研发和部署必须要有明确的禁区，并且要在人工智能辅助司法判案过程中保证这些人工智能系统具备充分的安全性、可靠性和可控性。³²我国实施人工智能负面清单制度的意义在于进一步优化资源的有效配置，为人工智能和实体经济的融合发展注入强劲动力，同时又避免过度超前创设新兴权利，为我国未来的人工智能应用划出底线、留出空间。

具体而言，首先我国可以设立国家人工智能办公室，将其作为人工智能治理的主管机关，将我国人工智能领域的规则制定、落地实施和监督管理等多项职能统筹协调为一体，防止出现欧盟人工智能监管领域交叉重叠可能导致的“九龙治水”难题。我国人工智能办公室可以根据人工智能在经济社会发展中的重要程度，以及一旦遭到攻击、篡改、破坏或者非法获取、非法利用，对国家安全、公共利益、社会稳定、环境保护，或者个人、组织合法权益、经济秩序造成的危害程度，牵头制定并定期更新人工智能系统负面清单。对于负面清单内的人工智能研发和部署活动创设较高的准入门槛和合规义务，对于清单之外的人工智能研发和部署活动仅需承担一般性的安全义务。正如中国社会科学院法学研究所权威发布的《人工智能示范法 2.0（专家建议稿）》第 25 条指出的，国家建立人工智能负面清单制度，对负面清单内的产品、服务实施许可管理，对负面清单外的产品、服务确有需要的实施备案管理。与此同时，为避免国家人工智能办公室行政权力的不断扩张，应当确立非必要不干预的行政谦抑性监管原则。既要保证人工智能主管机关设置监管程序法定化，明确人工智能执法检查的频次和执法标准，也要实现与不同人工智能风险相适配的监管措施，对清单内高风险行为进行许可准入的事前监管，对一般行为实施以备案为主的事后监管，对处于实验阶段的纯技术研发免于备案要求，并且不得以备案未完成阻碍正常的人工智能企业实践。

其次，可以在负面清单对高风险人工智能实施更严格监管的基础上，构建高

³² 参见黄京平：《刑事司法人工智能的负面清单》，载《探索与争鸣》2017 年第 10 期。

风险人工智能的特别规制体系。参考《关键基础设施安全保护条例》对关键基础设施运行者确立特殊的安全保护责任义务的做法，要求高风险人工智能研发者，尤其是通用人工智能基础模型研发者承担更高标准的“技术安全守门人”责任，并建立模型安全评估、漏洞修复及应急处置等综合协调机制。此外，可以根据主体从事活动的不同，将人工智能研发者、部署者和使用者的义务进行差异化配置。依照技术参与度进行划分，人工智能研发者主要负责技术伦理审查和安全合规义务，人工智能部署者承担产品全周期检测和风险即时报告义务，以及人工智能使用者的合规使用义务。

简言之，综合考量我国人工智能治理实践和人工智能发展需求，我国应提炼既往规制人工智能的治理经验，合理借鉴欧盟的规制框架，通过构建科学高效的负面清单制度来促进和规范我国人工智能发展，以构建我国在人工智能新兴领域自主的规制框架。

结语

当前，由人工智能引领的新一轮科技革命和产业变革方兴未艾。人工智能已经深入各国经济、社会发展的各个领域，并成为当前全球科技博弈和规则竞争的核心领域。人工智能立法是新时代加快新兴领域和重点领域立法的应有之义，对于统筹推进国内法治和涉外法治具有重要保障作用。³³我国人工智能法立法在创设自主的人工智能规制框架时既应与时俱进地反映时代潮流，合理参考、借鉴欧盟《人工智能法》的先进经验，也应明确回应本土人工智能实践的独特性。我国应继续坚守人工智能促进型立法的立场，通过整合技术标准、市场、法律和政策等多元治理工具，³⁴以负面清单制度构建自主的我国人工智能规制体系，在划定安全底线的基础上鼓励人工智能企业开拓创新，全面推动人工智能实现高质量发展。

（责任编辑：赵勇）

³³ 参见杨建军、张凌寒、周辉等：《人工智能法：必要性与可行性》，载《北京航空航天大学学报(社会科学版)》2024年第3期。

³⁴ 参见赵精武：《论人工智能法的多维规制体系》，载《法学论坛》2024年第3期。

人工智能对金融科技伦理的挑战及其国际法规制

公惟韬

摘要：在数字经济时代，金融科技伦理由于人工智能的赋能而在数据安全、算法透明和算力平等方面面临挑战。由于金融科技服务的域外影响以及发达国家和发展中国家之间相关技术的发展水平不同，仅靠国内法无法有效应对前述挑战，有必要在国际法中构建统一的治理体系。相较于国际硬法，国际软法在应对新兴技术带来的挑战时更有优势。然而，目前缺乏应对上述挑战的国际硬法和国际软法。对此，可参考金融科技较为发达的国家的现有法律实践，完善应对人工智能挑战的国际软法，并在此基础上形成“国际软法-国内软法-国内硬法-国际硬法”的良性循环，并通过国际组织、国内行业协会和政府机构多主体协作及区域国际合作，构建应对人工智能挑战的金融科技伦理国际治理体系。

关键词：人工智能 金融科技 科技伦理 国际法规制 良性循环

引言

数字经济时代，科技赋能金融行业，金融科技的概念应运而生。所谓金融科技指的就是将技术应用于金融领域，通过新兴前沿技术提供新软件、应用程序、流程或商业模式等金融服务。¹现阶段，人工智能已成为金融科技发展的新增长点。²人工智能由数据、算法和算力三大要素构成，是能让计算机系统具有和人一样的认知的智能技术。³通过引入人工智能，金融科技公司能够更高效地分析其所收集的数据并提供智能投资顾问等金融服务。

2022年3月，中共中央、国务院办公厅印发《关于加强科技伦理治理的意见》，要求将科技伦理贯穿于科学研究、技术开发等科技活动全过程，推动科技向善，确保科技活动风险可控，科技成果造福于民。金融科技同样需要遵循科技伦理，进而形成规范、调节金融科技活动中各利益相关方行动的金融科技伦理。此外，为确保人工智能的发展不会对人类造成负面影响，联合国教科文组织大会于2021年11月已通过《人工智能伦理问题建议书》（以下简称“《建议》”），对人工智能领域的科技伦理提出了数据安全、算法透明以及算力平等的要求。在人

¹ See Nimitz Jordan, “Shattering the Looking Glass: How a Section 101 Revision Could Save FinTech from Alice”, 30 *Federal Circuit Bar Journal* 55, 60 (2020).

² 参见刘斌、赵云德：《金融科技（人工智能与机器学习卷）》，机械工业出版社2019年版，第16页。

³ See Martz Jessica G., “Artificial Intelligence Is Here, Get Ready!”, 28 *Catholic University Journal of Law and Technology* 33, 33 (2019).

工智能赋能金融科技的背景下，人工智能在数据、算法、算力方面所需遵循的科技伦理要求也应被纳入金融科技伦理之中。⁴

人工智能赋能的金融科技作为一种新兴的服务业态，在不同国家、地区，无论是技术发展速度和规模，还是监管态度都存在差异。在部分技术较发达的国家被限制的行为，在其他技术相对落后的国家则可能由于对相关技术的不了解而被允许，导致法外之地的存在。因此，有必要从国际法层面协调监管。早在 1975 年 11 月，联合国大会就已通过《利用科学技术进展以促进和平并造福人类宣言》，要求所有国家采取措施，保障科学技术不被滥用、误用，以至危害国家安全和个人权利，而是应加强国际合作，提升各发展中国家的科学技术水平，这在人工智能赋能的金融科技领域也是一样的，但是，金融科技发达的国家为牟取垄断利益，缺乏制定全球统一的行业监管标准的动机，而趋利的金融科技企业又往往会利用人工智能技术的风险管控盲区来实现监管套利，以至于对人工智能赋能的金融科技的数据安全、算法透明和算力平等构成挑战。⁵目前，关于人工智能的国际法研究受俄乌战争影响，多从国际人道法角度出发，集中于对人工智能武器的实体法规制研究，此外也存在着将人工智能作用于国际商事仲裁等程序规范的研究，但在数字经济背景下，对人工智能之于金融科技领域影响的具体国际法规制的讨论相对不足。⁶金融科技公司为运行人工智能而收集的数据是否存在安全风险、基于所收集的数据提供的投资顾问服务的算法是否透明，以及在算力层面技术发达国家对后发国家的技术殖民等平等问题，都是国际法需要追问和解决的问题。

一、人工智能对金融科技伦理的影响

（一）金融科技的发展

金融科技的内涵反映的是科技行业的创新对金融行业发展的影响。最早的例子如印刷技术的普及使各国印刷出纸币，纸币因为更便于流通，而增强了金融行业的流动性，此后，1866 年，电报的出现以及第一条跨大西洋电缆的建成，加速了金融全球化进程，直至 20 世纪 70 年代，电报系统一直被用于银行之间的资金转账；20 世纪 80 年代开始，银行间结算系统开始利用大型计算机进行数据获取、记录和保存，电子贸易公司也为个人投资者提供电子交易系统。⁷目前，金融科技的主要业态包括移动支付、P2P（Peer-to-Peer Lending，个人对个人、伙伴对伙伴或点对点）网络贷款、投资管理等相关金融服务。⁸

⁴ 参见王晓青、许成安：《金融科技伦理的内涵、规制方法与研究前景》，载《江汉论坛》2021 年第 10 期。

⁵ 参见尹振涛、王美懿：《人工智能考验金融监管》，载《中国外汇》2023 年第 6 期。

⁶ 参见沈伟、赵尔雅：《数字经济背景下的人工智能国际法规制》，载《上海财经大学学报》2022 年第 5 期。

⁷ 参见宛俊：《图解金融科技：法律实务操作要点与难点》，法律出版社 2019 年版，第 9 页。

⁸ 参见宛俊：《图解金融科技：法律实务操作要点与难点》，法律出版社 2019 年版，第 7 页。

（二）人工智能赋能的金融科技

人工智能对金融行业的赋能从 2017 年初高盛宣布股票交易员将由最高峰时的 600 人裁减到 2 人,且大部分工作都将由人工智能完成中开始表露,同年 4 月,管理近 5 万亿美元资产的黑石集团宣布计划裁员 400 人,并将用人工智能取而代之。⁹自此,人工智能成为金融行业未来的重要战略发展方向。

就人工智能而言,从 1956 年达特茅斯会议至今,该技术经历了三次发展浪潮,1956 年至 1974 年的第一波浪潮以人工智能的数学模型存在缺陷、无法完成计算任务告终;其后至上世纪 90 年代,人工智能数学模型才出现包括多层神经网络等在内的重大进展,并在这第二波浪潮后,迎来持续至今的第三次发展浪潮,其间出现的深度学习和自然语言处理是人工智能得以在金融行业应用的关键算法。¹⁰深度学习算法的基础是第二波浪潮的多层神经网络技术——所谓多层神经网络技术,即根据人脑结构和功能特点设计出包含大量数据的“神经层”,包括在微信、微博、淘宝、百度等互联网平台上所产生的海量数据,以及每一笔交易在金融机构后台留下的数据都将被收集在这些“神经层”中,作为训练人工智能的数据。¹¹深度学习算法能够使人工智能学会由人制定的通用数据分类规则,并运用这些规则自行标注“神经层”中的数据,进而进行用户行为和金融市场的预测分析。自然语言处理算法建立在深度学习算法的基础上,能够实现人工智能与人之间有效的自然语言通信,并使人工智能有能力从海量、复杂的企业年报、行业研究报告等文字信息中提取出最重要的数据指标,进而帮助进行投资决策、识别市场风险。¹²在数据、算法之外,人工智能赋能的金融科技服务的提供也离不开算力。如果说数据和算法是运行人工智能的软件的话,那么算力就是运行人工智能的硬件。算力是应用程序所能实际获得的计算能力¹³,它关涉芯片、网络基站、存储介质等基础设施建设,比如更加先进的芯片将能够提供更高的算力,意味着人工智能将能更高效地对用户的要求适用数据和算法并做出回应,而如果要使用人工智能赋能的金融科技服务,则需要能够接入的网络基站、储存人工智能数据所必备的存储介质以及研发算法的必要人才,这些都是算力的内涵。¹⁴

⁹ 参见刘斌、赵云德:《金融科技(人工智能与机器学习卷)》,机械工业出版社 2019 年版,第 17 页。

¹⁰ 参见刘斌、赵云德:《金融科技(人工智能与机器学习卷)》,机械工业出版社 2019 年版,第 21-25 页。

¹¹ 参见刘斌、赵云德:《金融科技(人工智能与机器学习卷)》,机械工业出版社 2019 年版,第 13 页。

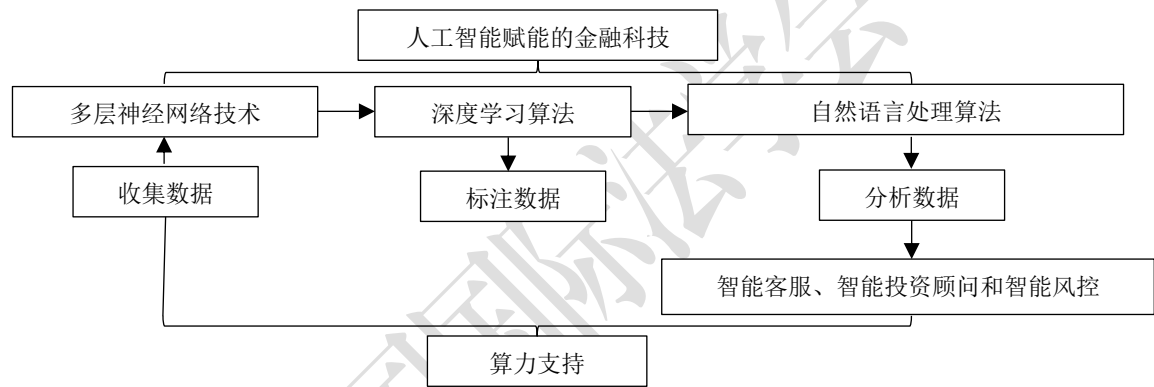
¹² See Grzegorz Mazurek, “Artificial Intelligence, Law, and Ethics”, 2023 *Krytyka Prawa* 11, 11-14 (2023).

¹³ 参见刘宇航、张菲:《计算概念谱系:算势、算力、算术、算法、算礼》,载《中国科学院院刊》2022 年第 10 期。

¹⁴ 参见陈寒冰:《数字经济时代算力网络建构的国际比较与镜鉴》,载《新疆社会科学》2021 年第 5 期。

在算力支持下，如今，智能客服、智能投资顾问和智能风控是人工智能赋能的金融科技的具体形式。¹⁵在前端，如工商银行 2016 年上线的“工小智”智能客服，就能帮助客户即时解决问题；在中端，Kensho 公司研发的 Warren 系统是智能投资顾问中的翘楚，用户只需在文本框中提问，Warren 系统便会搜索包括经济报告、货币政策变化、政治事件等在内的一切资讯以评估其对金融资产价格的影响；在末端，陆金所推出的 KYC（Know Your Customer，了解你的客户）系统和纳斯达克的 SMARTS 市场监察系统均能给出用户风险承受能力的评估，实时监控市场，提前发现市场操纵、内幕交易、异常定价等传统监管难以发现的异常。

16



（三）人工智能对金融科技伦理的挑战

金融科技伦理是各相关行业的参与主体和从业人员“能做”和“应做”的价值标准。¹⁷人工智能赋能的金融科技同样要遵循相同的伦理原则。然而，在数据、算力和算法层面，人工智能技术正在对前述安全、透明和平等的金融科技伦理形成挑战。

1. 数据层面：人工智能对金融科技伦理安全原则的挑战

数据是人工智能深度学习和自然语言算法的基础，但数据质量与数据存储是否合规却面临挑战。在数据质量方面，金融科技企业收集的数据包括第三方数据库中的数据，但这类第三方数据库的数据是否合规是存疑的。¹⁸此外，以 ChatGPT 为例，该类生成式人工智能的用户反馈强化学习模式，会将用户提供的所有反馈纳入数据库，成为人工智能作出决策的依据之一，但是这些用户反馈信息的真实、合法性未经核实，存在数据“投毒”风险。¹⁹人工智能的决策也会被投毒污染的数

¹⁵ 参见宛俊：《图解金融科技：法律实务操作要点与难点》，法律出版社 2019 年版，第 140-143 页。

¹⁶ 参见刘斌、赵云德：《金融科技（人工智能与机器学习卷）》，机械工业出版社 2019 年版，第 69, 139, 152, 155 页。

¹⁷ 参见何若水等：《金融科技伦理：特征、治理架构及实践》，载《科学学研究》2024 年第 7 期。

¹⁸ 参见罗世杰、贺国荣：《生成式 AI 阶段金融科技数据合规的法律化因应》，载《金融发展研究》2023 年第 11 期。

¹⁹ See Terrence J. Sejnowski, “Large Language Models and the Reverse Turing Test”, 35 *Neural Computation* 309,

据影响而偏离既定的训练目标和轨道，从而影响所提供的金融服务的质量与安全。

在数据存储方面，由于人工智能技术的特殊性，运用人工智能为客户提供金融服务的金融科技企业往往会掌握大量客户的个人隐私数据和公司的商业秘密数据。但是，有报道显示，在“暗网”上被泄露的个人信息中有 60% 以上来自金融行业，包括证券机构资金 50 万以上优质股民的个人信息页面截图及其商业银行银行卡、理财信息等数据；2021 年，WizCase 安全团队在扫描国际在线外汇交易平台服务器时也发现了严重的数据泄露情况。²⁰无论这些泄露是金融行业工作人员的主观行为，还是黑客入侵数据库所得，都对金融科技公司为运行人工智能而采集的数据的储存提出了更高的安全性要求。

2. 算法层面：人工智能对金融科技伦理透明原则的挑战

为提供智能投资顾问服务，人工智能需要对历史数据进行分析，才能预测未来可能出现的风险和机遇，并提供相应的避险策略和投资决策。但是，这一技术本身较为复杂，而且无论是避险策略，还是投资决策都属于公司的商业决策，具有一定的排他性，这导致客户更多时候只能获得金融科技给予的投资决策结果，而对于其过程是未知的，由此将带来算法黑箱风险，对金融科技伦理透明原则构成挑战。²¹简单地说，算法黑箱就是金融科技公司与客户之间的信息不对称，由于金融科技公司所掌握的信息、手段、条件对于其客户而言是未知的，这一未知状态就将用户置于与金融科技公司不平等的地位上。²²一些金融科技公司甚至会利用他们的信息优势，炮制新型 P2P 金融诈骗，即利用客户的不知情，打着通过人工智能帮助客户理财的幌子进行诈骗。目前，智能投资顾问技术并不成熟，对于相应的风险也尚未出台规范予以必要的预警和规制，而难以避免算法黑箱导致的人工智能决策不透明的风险。

3. 算力层面：人工智能对金融科技伦理平等原则的挑战

人工智能对金融科技伦理平等原则的挑战有三个方面：第一，根据世界经济论坛的报告，到 2028 年，23% 的工作岗位将会受到人工智能带来的颠覆性影响。²³这并不是说将有 23% 的工作岗位被人工智能取代，而是说这类岗位的工种将转化为如人工智能产品的研发、运行与维护的新型工种，故而对人才硬性技术素质提出更高的要求；此外，人工智能无法取代的人类工作岗位往往是那些强调沟通、逻辑与创造的岗位，这同样对人才的软性应用能力提出了更高的要求，但并非所有国家、地区都具备充足的教育资源来培养相应的人才。第二，人工智能运

309-342 (2023).

²⁰ 参见张超文等：《数十亿条个人信息明码标价“潜规则”盛行售卖泛滥成灾》，载《经济参考报》2021 年 4 月 19 日，第 A01 版。

²¹ 参见程雪军：《金融科技算法风险的体系化治理：欧美比较治理视角》，载《经济社会体制比较》2023 年第 6 期。

²² 参见程雪军：《金融科技算法黑箱风险的生成机制与多维治理》，载《金融论坛》2023 年第 8 期。

²³ See *The Future of Jobs Report*, World Economic Forum, 2023, p.6.

行时需要耗费大量的能源，比如训练一小时像 ChatGPT 这样的大型语言模型，就需要使用约 1300 兆瓦时（MWh）的电力，大约相当于 130 个美国家庭每年消耗的电量。²⁴因而并非所有国家、地区的金融科技公司都有必要的芯片、网络基站、存储介质等基础设施资源来确保包括金融科技服务在内的由人工智能赋能的服务可靠、安全。第三，技术发展水平不同的各国之间的技术鸿沟也会因为人工智能的应用而被扩大。所谓技术鸿沟是指工业化社会和发展中社会之间获取技术资源的差异，它关系到每个国家技术资源贫富之间的差距，意味着那些能够获取和不能获取技术来改善生活的人之间的差异。²⁵掌握数据、算法与算力优势的人工智能赋能的金融科技公司可以积聚更多资源，实现“赢者通吃”，甚至导致技术殖民主义，也就是这些公司所背靠的发达国家对发展中国家的掠夺，并通过游说、基础设施的投资以及向发展中国家输出硬件和软件等方式，实现对发展中国家的控制。²⁶前述三方面挑战的最终结果是全球社会不平等的加剧。

二、应对人工智能之于金融科技伦理挑战的国际法规制的必要性、可行性及其现状

人工智能在金融科技领域的应用中，可能会因数据滥用、算法不透明、算力资源配置不均衡，而对金融科技伦理构成挑战。部分监管者可能会认为已有的工具和政策足以监管科技领域的创新，但是金融科技，尤其是人工智能赋能的金融科技，具有其特殊性，它依赖自动化的人工智能算法，并会使用只有在数字时代才会存在的各类数据与算力资源，这导致原有的监管思路可能逐渐不再适用。²⁷

（一）应对人工智能之于金融科技伦理挑战的国际法规制的必要性

运用国际法对人工智能赋能的金融科技进行治理的必要性首先在于人工智能这一技术本身的特性。由于人工智能在金融领域的应用并不受国界所限，比如，对于金融科技后发的国家来说，一旦允许量化交易这一人工智能赋能的金融科技在证券领域运用，如未加以必要监管，则可能会扰乱证券市场的秩序。所谓量化交易，就是通过分析人的投资思想规则，并析出其中的变量以生成模型——一整套完整的、可以用历史数据加以分析验证的算法，并交付由算力支持的人工智能予以自动执行。²⁸由人工智能赋能的量化交易相较于普通的人工投资者来说，能够更快地结合历史数据做出决策，并下单或者撤单。对于已经长时间实践量化交易的金融科技较为发达的国家，一旦发现后发国家中在证券市场中引入量化交易，

²⁴ 参见万玉航：《AI“吃电”凶猛 ChatGPT 每天耗电已超 50 万千瓦时》，来源：https://tech.cnr.cn/gstj/20240312/t20240312_526624611.shtml，2024 年 4 月 6 日访问。

²⁵ See J. M. Spectar, “Bridging the Global Digital Divide: Frameworks for Access and the World Wireless Web”, 26 *North Carolina Journal of International Law* 57, 57-103 (2000).

²⁶ 参见王淑敏：《全球数字鸿沟弥合：国际法何去何从》，载《政法论丛》2021 年第 6 期。

²⁷ See Yesha Yadav, “Fintech and International Financial Regulation”, 53 *Vanderbilt Journal of Transnational Law* 1109, 1109-1146 (2020).

²⁸ 参见周方召等：《量化投资基金和股票市场稳定性》，载《金融经济研究》2023 年第 2 期。

其投资者便可能会进入这些后发国家，利用作为东道国的后发国家尚未完备的监管措施，牟取依据其母国法违法的利益。比如这些投资者可能会以低于市场卖价的价格挂出卖单，迫使卖家挂出更低的价格以寻求成交，而在逐渐出现更低的卖价后，这些投资者将会利用量化交易技术迅速撤单，转手开始做多，从而以更低的价格获取头寸，该行为在投资者母国很有可能已被当做市场操纵行为而获得了必要的事前监管，但东道国的监管措施并不完备才使得该行为得以可能。²⁹这种情况下，东道国就会面临来自他国投资者的域外操作的影响，哪怕相关的操作在投资者母国是违法的，东道国并不充分的国内监管机制就会为他国投资者提供攫取暴利的法外之地，而东道国又会受制于管辖规则而无法规制他国投资者。³⁰由此可见，人工智能在金融科技领域的应用不可避免地使得这一技术具有全球性³¹，因而不能忽视人工智能赋能的金融科技对国际规则的冲击，而需要在数据层面、算法层面和算力层面对其域外效应进行规制，在国际层面通过构建沟通、协调各国的全球治理体系进行指引，防止技术落后的国家沦为法外之地。

（二）应对人工智能之于金融科技伦理挑战的国际法规制的可行性

虽然尚未有直接规制人工智能赋能的金融科技的国际条约，但如前所述，联合国大会在上世纪就已通过《利用科学和技术进展以促进和平并造福人类宣言》，而且联合国教科文组织也已采纳《建议》，意味着人工智能及其对金融科技伦理的挑战是可以在国际法层面予以规范和调整的。此外，在区域国际实践中，欧盟以《通用数据保护条例》（General Data Protection Regulation，以下简称“GDPR”）和《金融工具市场指令 II》（Market in Financial Instruments Directive of European Union II，以下简称“MIFID II”）等为规则，建立了以人工智能为专题的立法及以数据为规制对象的战略政策。此外，《全面与进步跨太平洋伙伴关系协定》（CPTPP）中第 14 章专章规范的电子商务，以及《数字经济伙伴关系协定》（Digital Economy Partnership Agreement，DEPA）第 8.2 条所涉及的人工智能对数字经济之具体技术相关的国际贸易问题等均与规范人工智能赋能的金融科技相关。虽然相关的国际实践呈碎片化特征，但是这些实践本身足以证明在国际法层面规范人工智能赋能的金融科技的可行性。

（三）应对人工智能之于金融科技伦理挑战的国际法规制的现状及其不足

从上述可见，人工智能及其赋能的金融科技已经在国际法层面获得关注，但既有国际法规制尚存在如下三点不足。

²⁹ 参见冯永昌、张建民、刘天权：《国外量化投资经典案例介绍与法律分析》，载《清华金融评论》2016 年第 2 期。

³⁰ See Yesha Yadav, “Fintech and International Financial Regulation”, 53 *Vanderbilt Journal of Transnational Law* 1109, 1109-1146 (2020).

³¹ 参见沈伟、赵尔雅：《数字经济背景下的人工智能国际法规制》，载《上海财经大学学报》2022 年第 5 期。

1. 缺乏对人工智能之于金融科技伦理挑战的统一国际法规制

国际法体系在对人工智能之于金融科技伦理挑战规制上的滞后是由于国际法规则本身的局限性所致的。一方面，人工智能及其对金融科技伦理的挑战在国际法语境中系新的规制内容，但无论是国际条约还是国际习惯，他们的形成周期都较长，无法及时有效回应新出现的问题。³²此外，由于技术殖民主义等主观因素，导致国际法立法过程中的参与主体可能无法就人工智能及其对金融科技伦理的影响达成统一的国际硬法规则。技术发达的国家可能并不愿意与技术落后国合作，而是更希望通过对技术落后国进行控制与技术殖民以攫取垄断利益，而且各发达国家对人工智能及其在金融领域的应用的规制态度也不同，如欧盟国家的规制相对较严，而美国则力主较为宽松的技术流动，以充分发挥他们在相关领域的既有优势。³³因此，直接在国际层面上形成统一的国际硬法规则可能存在一定的难度。

另一方面，对于复杂、不确定、结果不可预见的新兴科技，如果只依靠固定、“刚性”的硬法，则可能难以跟上技术发展的节奏，而更为灵活的国际软法作为硬法的补充，反而更有利于完善整体的治理机制。³⁴但是目前相关的国际软法尚处于初步阶段，相关的国际软法不仅数量少，而且总体上并没有形成体系，也没有提出规制路线图。

2. 现有国际法体系中未明确人工智能赋能的金融科技产品是货物还是服务

既有国际法体系对人工智能赋能的金融科技产品的属性尚未予以明确，如CPTPP在数字产品的定义条款中就是通过注释直接回避了数字产品贸易归属于服务贸易还是货物贸易的争论，但这将导致世界贸易组织法律体系中仅适用于货物贸易的《关税及贸易总协定》《技术性贸易壁垒协议》和仅适用于服务贸易的《服务贸易总协定》等协定，在调整与人工智能赋能的金融科技产品相关的争端时面临困境。³⁵

3. 既有国际规则在人工智能对金融科技伦理构成挑战的数据、算法和算力层面存在规制路径不明的问题

就具体的国际条款而言，首先，在数据层面，既有国际规则缺乏关于数据“投毒”、存储等数据安全风险的规制。其次，缺乏透明度是算法层面面临的问题。对于算法的透明度，有两类监管方式：一是审计源代码的“白盒”方法，二是仅通

³² 参见沈伟、赵尔雅：《数字经济背景下的人工智能国际法规制》，载《上海财经大学学报》2022年第5期。

³³ See Peter Drahos, “Developing Countries and International Intellectual Property Standard-Setting”, 5 *Journal of World Intellectual Property* 765, 765-789 (2002).

³⁴ 参见朱明婷、徐崇利：《人工智能伦理的国际软法之治：现状、挑战与对策》，载《中国科学院院刊》2023年第7期。

³⁵ See Shin-yi Peng, Ching-Fu Lin & Thomas Streinz, *Artificial Intelligence and International Economic Law: Disruption, Regulation, and Reconfiguration*, Cambridge University Press, 2021, p.127.

过接口审计人工智能系统输入和输出而不审计源代码的“黑盒”方法，CPTPP、《美国-墨西哥-加拿大协定》等既有相关国际实践中并没有区分前述两种监管模式，而导致如果适用后者，也就是透明度更低的监管方式时，掌握源代码的金融科技公司一旦拒绝为其所提供的自动化应用决策予以准确解释，其所具有的信息与技术的双重优势地位就很可能为其带来监管套利的空间。³⁶最后，在算力层面，联合国贸易和发展会议在其 2021 年《数字经济报告》中指出，目前世界范围内的技术竞争，尤其是人工智能的竞争，可能会导致数字经济的碎片化，不平衡的人工智能发展将扩大技术鸿沟，因为利用人工智能在新兴市场快速部署系统和基础设施的能力目前都集中在发达国家的部分企业手中，而技术相对落后的国家很容易由于教育和科研水平的劣势陷入对发达国家的技术依赖，或者受发达国家所制而面临发展困顿，但是这些问题在技术殖民主义背景下的国际法体系中都被忽视了，导致全球社会不平等。

三、完善对人工智能赋能的金融科技的国际法规制

在国际法层面构建应对人工智能之于其所赋能的金融科技的挑战的国际法治体系，才能有效应对人工智能在金融领域应用时的域外效应、协调技术水平不同的国家之间的监管，并促进各国在相关技术领域内的共同发展。

（一）以国际软法-硬法协同治理为模式，构建应对人工智能之于其所赋能的金融科技的挑战的国际法治框架

从近现代国际法诞生以来，居于国际法法律渊源中心的一般都是条约、习惯等国际硬法，国际软法作为那些不能用强制力保证实施的法律规范，如宣言、建议等，系各方利益表达的协调机制和全球治理机制中国际硬法的补充。³⁷相较于国际硬法，国际软法具有治理功能上的优势，如前所述，国际软法更为灵活，在对人工智能等新兴技术的国际治理中可以填补国际硬法空白，此外，由于国际软法不需要像条约等国际硬法那样经历严格的立法程序转化为国内法，而能更快地引导相关领域的国内行为准则和专业标准的设立，因而更为高效。³⁸加之，由于各国发展水平不同，在无法形成统一的国际硬法规范的背景下，区域性国际软法还能因地制宜区分治理、分级监管，保障后发国家的技术发展。

³⁶ 参见程雪军：《金融科技平台算法黑箱的国际规制与中国因应》，载《国际经济法学刊》2023 年第 2 期。

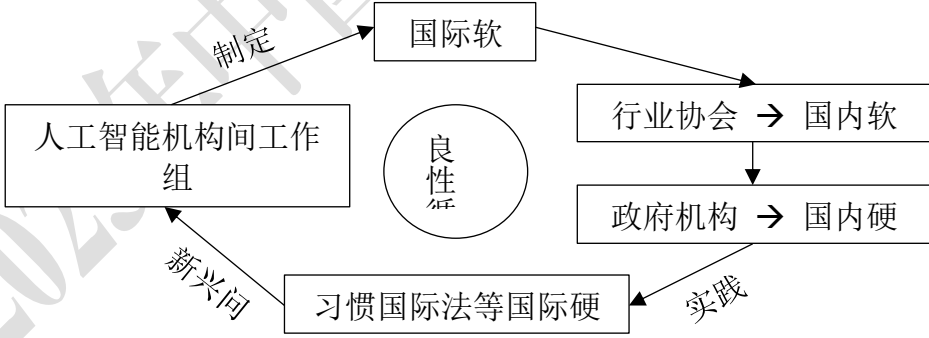
³⁷ 参见朱明婷、徐崇利：《人工智能伦理的国际软法之治：现状、挑战与对策》，载《中国科学院院刊》2023 年第 7 期。

³⁸ 参见朱明婷、徐崇利：《人工智能伦理的国际软法之治：现状、挑战与对策》，载《中国科学院院刊》2023 年第 7 期。

虽然国际软法作为国际硬法的协同治理手段具有上述优势，但对于国际软法的执行仍取决于各国的自愿，因而“国际软法-硬法协同治理模式”还需国际组织作为对接各国政府机构以外的国内行业协会的主体，在国际、国内双维度中协调配合。在人工智能领域，联合国经济社会理事会下已设有“人工智能机构间工作组”，对人工智能领域相关的国际法的制定与执行予以监督。在“国际软法-硬法协同治理模式”下，人工智能机构间工作组应首先整合相关的国际软法，以备在国内行业协会的帮助下将这些国际软法中的行为准则、技术标准转化为国内软法，因为行业协会作为更加专业的组织，能够对相关软法的制定进行把握，并对其执行进行全过程监督。³⁹同时，这些软法中的行为准则、技术标准最终仍需转化为法律而需立法机构的配合，而且各国内行业协会的设立也需要政府机构的帮扶政策来鼓励企业、非政府组织、科研机构等多元主体的共同参与，以期将国内软法最终落实为国内硬法，在此基础上，最终通过实践和各国政府的支持，使这些国际软法规则逐渐成为国际习惯或者国际法规范存在的证据，实现相关国际软法向国际硬法的转化。⁴⁰由此形成“国际软法-国内软法-国内硬法-国际硬法”的良性循环，完善应对人工智能对金融科技伦理挑战的国际法治理框架。

（二）在既有国际法体系中明确人工智能赋能的金融科技产品的属性为服务

在人工智能赋能的金融科技实践中，无论是在前端的智能客服如工商银行“工小智”，还是在中端的智能投资顾问如 Kensho 公司的 Warren 系统，抑或是在末端的智能风控如金所 KYC 系统和斯达的 SMARTS 系统，这些人工智能赋能的金融科技均具有明显的服务属性，而不是货物。因此，人工智能机构间工作组应与世界贸易组织等机构合作，在相关的国际法体系中明确人工智能赋能的金融科技产品的属性为服务。



（三）在数据、算法、算力方面完善国际法规制路径，应对人工智能对其赋能的金融科技伦理的挑战

前述以国际软法-硬法协同治理为模式，构建应对人工智能之于其所赋能的

³⁹ See John M. Conley, “A New Governance Approach to Regulating Human Genome Editing”, 22 North Carolina Journal of Law & Technology 107, 131-140 (2020).
⁴⁰ 参见安慧影、黄朝峰、李阳：《新兴技术伦理风险协同治理研究》，载《科技进步与对策》2024 年第 7 期。

金融科技的挑战的国际法治理框架的起点为国际软法，是因为软法在应对新兴技术挑战时具有更灵活、高效且更因地制宜的优势，因此，在数据、算法、算力方面完善国际法规制路径的起点也应为国际软法。具体路径则可参考在人工智能赋能的金融科技领域较为领先的美国、欧盟等地区的立法，在数据、算法、算力方面制定更为细化、全面的国际软法标准。⁴¹

1. 数据层面：甄别数据，应对人工智能对其赋能的金融科技的安全挑战

针对数据质量所面临的挑战，相关金融科技企业应当强化甄别数据的能力，并对数据分类进行细化，防止虚假、被投毒或可能侵犯隐私的数据被纳入人工智能的数据库。⁴²就用户数据储存所面临的挑战，应像 GDPR 中所规定的那样，落实保护数据安全的“知情-同意”原则。⁴³GDPR 中提出，当自动化决策产生的法律效力涉及数据主体或将其产生重大影响时，数据主体有权随时反对企业对其个人数据的使用，来保护自身的数据安全。为了运行人工智能，金融科技公司需要获取大量的用户信息来分析客户的偏好，而为了巨额利润，这些公司可能会滥用这些数据，甚至利用个人投资者和非专业金融服务用户的认知限制对其进行偏好操纵。⁴⁴对此，在 GDPR 之外，MIFID II 引入了大量实质性和规范性的披露要求，并规定金融科技公司应当向客户提供公平、清晰和不误导的信息的义务，为满足这些要求，金融科技公司被要求完成问卷调查，并采取其他步骤解释他们对数据的使用，以符合“知情-同意”原则。⁴⁵

在国际法体系中，为应对人工智能对金融科技伦理安全原则的挑战，在数据层面，可借鉴 GDPR 和 MIFID II 的规定，要求金融科技公司细化数据分类、提升所收集的数据质量与安全，并构建符合“知情-同意”原则的统一国际软法规则，确保被存储的数据不会被滥用、泄露而引发安全问题。

2. 算法层面：增加算法透明度，应对人工智能对其赋能金融科技的透明度挑战

人工智能所给出的决策结论往往无法用人们容易理解的术语解释，因为这些决策结论是由高度复杂的算法矩阵所生成的，在这个算法矩阵中包含数百、数千或数百万个连接节点，于其中，每个节点都包含大量的数据。对此，美国联邦政

⁴¹ 就世界各国的金融科技发展而言，美国一直都是全球金融科技的中心，其金融科技公司 Paypal 作为第三方支付平台的鼻祖，1998 年就在美国加州成立，不同于支付宝以手机号码进行用户识别，Paypal 主要依赖电子邮件进行用户间资金的转移，并对收款、提现收取一定的费用，而在欧盟，早在 2005 年，世界第一家 P2P 借款平台 Zopa 便已成立，提供个人消费贷款服务。参见宛俊：《图解金融科技：法律实务操作要点与难点》，法律出版社 2019 年版，第 23、27 页。

⁴² 参见马忠法、胡玲：《网络空间命运共同体构建中的金融安全国际法》，载《上海金融》2020 年第 10 期。

⁴³ See Ira S. Rubinstein, “Big Data: The End of Privacy or a New Beginning?” 3 *International Data Privacy Law* 74, 74-86 (2013).

⁴⁴ 参见埃米利奥·阿夫古利亚斯、亚历山大·谢莱塔基斯：《欧盟对数字金融价值链的监管》，载《国家安全比较研究》2023 年第 2 期。

⁴⁵ 参见[英]斯特凡·勒施：《监管科技：重塑金融安全》，林华等译，中信出版集团 2019 年版，第 226-242 页。

府为塑造透明、公平和可解释的人工智能算法，由其国防预先研究计划局通过“可解释人工智能”计划，要求在提升算法决策结果准确性的同时，兼顾算法决策的过程性，也就是决策的可解释性。⁴⁶在州层面，美国的华盛顿州全面禁止面部识别时可能存在的算法问题，而伊利诺伊州则出台《人工智能视频采访法》，对类似情况要求完善算法透明度而非全面禁止，为协调不同州之间对算法透明度监管的不同态度，美国众议院于 2022 年通过《美国数据隐私和保护法》，明确算法的定义并构建面向人工智能算法决策的法律框架，同时成立国家科学技术委员会人工智能特别委员会推进人工智能相关标准、指南与技术的研究与实施。⁴⁷

在国际法体系中，为提升人工智能决策的透明度，可将美国的法律实践考虑在内，由人工智能机构间工作组承担起和美国人工智能特别委员会一样的责任，明确算法的定义，构建面向人工智能算法透明度的统一国际软法标准，协调技术发展、监管态度不同的各国的法律规范。

3. 算力层面：通过区域性国际合作，应对人工智能对其赋能的金融科技的平等挑战

近年来，欧美国家的逆全球化思潮逐渐从思想和理论转变为一些西方发达国家的外交政策。⁴⁸部分发达国家为实现自身利益、技术霸权而重建小圈子，甚至为了实现自身在技术领域的利益与优势地位，滥用长臂管辖原则，或以意识形态、国家安全等为缘由，隔离、挤压新兴技术体和新兴技术国家，以至于在国际层面实现统一的国际合作面临困难。⁴⁹对此，只能以区域性的国际合作为手段替代统一国际合作，促成相关领域的全球治理的推进。如以“一带一路”为平台，与人工智能机构间工作组协作设立参与国的区域性人工智能技术评估机构，并结合参与国的非政府组织和相关金融科技企业，依据各国的技术水平，制定不同的技术标准和发展目标，并通过普惠教育培养人工智能方面的人才，促进国际科技合作。如 2023 年 10 月 19 日印发的《算力基础设施高质量发展行动计划》便支持我国企业“走出去”，以“一带一路”参与国家为重点，并以工业、交通、医疗等典型行业为主要场景，打造一批海外算力基础设施建设的优质项目，解决共建“一带一路”国家的实际需求，以期在“一带一路”平台的协调下，尽可能去除参与国内部受到的技术殖民主义的影响，弥合技术鸿沟，让发展中国家获得与发达国家相平等的技术发展机会。⁵⁰

⁴⁶ 参见肖红军、阳镇：《数字科技伦理监管：美国进展与中国借鉴》，载《财经问题研究》2023 年第 6 期。

⁴⁷ 参见肖红军、阳镇：《数字科技伦理监管：美国进展与中国借鉴》，载《财经问题研究》2023 年第 6 期。

⁴⁸ 参见李盛竹：《跨国公司国际竞争背景中的技术霸权现象——理论回顾与展望》，载《社会科学家》2011 年第 9 期。

⁴⁹ 参见[美]G.约翰·伊肯伯里：《一个民主的安全世界——自由国际主义与全球秩序危机》，陈拯译，上海人民出版社 2023 年版。

⁵⁰ 参见谢迪扬：《论“一带一路”倡议下数据安全共同体的规则建构》，载《国际经济法学刊》2023 年第 2 期。

结语

人工智能赋能金融科技能够在前端、中端和末端为行业带来革新的金融产品服务，但是该技术对金融科技的安全、透明和平等原则构成的挑战也不容忽视，而且由于技术发展水平不同的各国对相关技术的规制不同，而需在国际层面予以协调，防范域外效应及法外之地。在国际法层面，关于人工智能对其所赋能的金融科技，尚无统一规范，其定位也并不明确，就数据、算法、算力三方面所面临的挑战而言，更是缺乏具体的国际法规制，应由人工智能机构间工作组借鉴在相关领域实践较为先进的国内法，制定统一的国际软法，并协调相关的国际软法与硬法形成良性循环，协调各国及其国内行业协会、政府机构及国际统一实践与区域国际实践，构建针对人工智能对金融科技伦理挑战的国际治理框架，提升数据安全、算法透明度、均衡算力资源。于其中，为保障发展中国家在所述治理框架内拥有话语权，包括我国在内的发展中国家应同步完善各自相关的国内法，只有国内规则完善了，相关国家在国际上对于有关技术的国际规制路径的提议才具有信服力，进而才能通过区域合作应对技术鸿沟、化解发达国家的技术殖民主义挑战。

网络生物安全治理困境及破解路径

吴 惠

(上海海事大学 法学院, 上海 201306)

摘要:网络生物安全作为一个交叉的新型安全领域,面临着军事威胁、数据主权泄露、破坏公共卫生等风险。在治理上既存在理论困境也涉及实践和现实困境,其中法律规范不明确、评估标准缺失、治理主体碎片化、科技支撑不足等是网络生物安全治理中的巨大挑战。为此,必须立足于“人类命运共同体”理念,系统阐述网络生物安全的内涵,并提出完善网络生物安全治理战略规划、健全网络生物安全法律规范、完善网络生物安全风险评估和预防措施、加大科技投入力度等措施。

关键词:网络生物安全 网络安全 生物安全 网络生物命运共同体

一、提出问题

2020年2月14日,习近平总书记在中央全面深化改革委员会第十二次会议上指出“把生物安全纳入国家安全体系,系统规划国家生物安全风险防控和治理体系建设,全面提高国家生物安全治理能力”。¹2022年博鳌亚洲论坛年会开幕式上习近平总书记又首次提出“全球安全倡议”,坚持共同、综合、合作以及可持续的安全观,共同应对地区争端和恐怖主义、气候变化、网络安全、生物安全等全球性问题。²根据指示可知,国家高度重视生物安全和网络安全,并将其放在国家战略位置。随着人工智能技术的发展,生物安全在现代化、数字化、智能化的浪潮中涌现,使得网络科学技术与生物安全交叉融合为一个新领域——网络生物安全。当前,国家行为者、恶意非国家行为者和黑客可以通过各种手段利用数字世界和生物世界之间的互联,造成潜在的生物信息盗窃、传播不正确的信息或者致使各类活动的中断等危害。这种非法攻击行为对生物安全提出了严峻挑战,一旦这些数据受到非法入侵,将给全民健康、国家安全带来重大损失。网络生物安全是什么?网络生物安全会给国家带来怎样的风险?当前我国立法是否能够应对网络生物安全提出的挑战?这一系列问题均是文本探讨的范围。对此,文章从网络安全和生物安全治理出发,论证了当前网络生物安全的内涵并对当前网络生

¹ 人民网. 全面提高国家生物安全治理能力[EB/OL]. (2020-3-06) [2024-6-11].

<https://baijiahao.baidu.com/s?id=1660371548660398885&wfr=spider&for=pc>

² 央广网. 习近平提出全球安全倡议[EB/OL]. (2022-4-21) [2024-6-11].

<https://baijiahao.baidu.com/s?id=1730683164994473645&wfr=spider&for=pc>

物安全潜在的风险予以反思，进而阐述网络生物安全治理的制度困境、实践困境和现实困境，最后对网络生物安全治理制度予以重塑并从“理念设计—顶层设计—具体设计—关键设计”的破解路径试图回应上述挑战。

二、网络生物安全的基本内涵及风险

（一）网络生物安全的基本内涵

网络生物安全不同于生物安全和网络安全。生物安全通常包含生物安全（biosecurity）和生物安保(biosafety)，在广义上指一个国家有效应对生物威胁和相关因素的能力。³根据世界卫生组织的规定，生物安全是指在个人和机构层面执行一套实践和程序，以确保“防止病原体 and 毒素的丢失、盗窃、滥用、转移或故意释放”。⁴学界将其扩展为“通过预防自然发生或故意释放病原微生物带来的风险，以及阻止生物实验室产生的一系列危害动植物和人类安全等行为的战略和综合方法”。⁵网络安全作为非传统安全的下位概念指“网络系统的硬件、软件及系统中的数据受到保护，不因偶然的或恶意的原因而遭到破坏、更改、泄漏，系统连续可靠正常地运行，网络服务处于不中断状态，以及采取措施规避不法信息传播、防止域外的网络渗透及跨国网络犯罪、使网络信息传播处于可控范围之内，包括网络硬件、软件及携带数据的安全，也包括利用网络媒介传播信息的安全”。⁶一般来说，当生物信息被测量、监控或改变并转换为数字信息时，或者当数字信息被用来操纵生物系统，就会产生网络生物安全。⁷网络生物安全作为生命科学和信息技术融合的新型跨学科领域⁸，是从 Cyberbiosecurity 英文单词翻译而来，学术界对其并未形成一个明确的定义，且其概念引起大量学者的讨论，将不同的观点聚合到一个广泛且精确的定义中可以消除定义歧义。⁹我国国家计算机网络应急技术处理协调中心 2020 年发布的《中国互联网报告》中首次指出网络生物安全主要表现针对高价值的生物信息数据和敏感个人医疗健康数据主动或被动泄漏，到联网的生物基础设施和终端设备遭受网络攻击，网络空间安全与生物安全之间产生紧密关联互动，交叉融合形成的新的安全领域。最初 Cyberbiosecurity 这一概念在 2018 年被 Murch 在其文章中首次非正式提出，即“网络生物安全涉及对不必要的监视、入侵以及恶意和有害活动的理解、保护、调查和归因，这些

³ Zhou D, Song H, Wang J, et al. Biosafety and biosecurity[J]. Journal of biosafety and biosecurity,2019,1(1):15-18.

⁴ World Health Organization. Biorisk management: Laboratory biosecurity guidance[M]. World Health Organization,2006.

⁵ 阙天舒，商宏磊. 全球生物安全治理与中国的治理策略[J]. 社会主义研究，2022（02）：165-172.

⁶ 何阳，娄成武. 新时代中国网络安全再认识：基于网络空间属性视角[J]. 西南民族大学学报(人文社会科学版)，2021，42(08):118-125.

⁷ Mantle J L, Rammohan J, Romantseva E F, et al. Cyberbiosecurity for biopharmaceutical products[J]. Frontiers in Bioengineering and Biotechnology,2019,7:116.

⁸ Richardson L C, Connell N D, Lewis S M, et al. Cyberbiosecurity: a call for cooperation in a new threat landscape[J]. Frontiers in bioengineering and biotechnology, 2019, 7: 451363.

⁹ Adeoye S O, Batarseh F, Brown A M, et al. Understanding the Landscape of Cyberbiosecurity for Integrative Educational Programming[J]. Journal of the ASABE, 2023: 0.

活动可能发生在生命和医学科学的内部或界面上,涉及到安全、竞争力和复原力,他认为,网络生物安全处于生物安全、网络安全和网络物理安全的界面”。¹⁰与 Murch 不同, Lauren C. Richardson 等认为网络生物安全旨在解决生命科学和数字世界交界处对有价值的信息、程序和材料进行恶意破坏、滥用或利用的可能性或实际情况,¹¹并将网络生物安全扩展到生物安保与生物安全,认为网络生物安全是解决人工智能、生物安全和网络安全三个不同领域的交叉学科。¹²Peccoud 与 Lauren C. Richardson 一样在分析网络生物安全时分析了生物安全的两个领域 Biosecurity 与 Biosafety,但是他认为网络生物安全是在网络空间和生物学之间前沿地带出现的全新风险。¹³需要注意的是,Biosecurity 与 Biosafety 在中文描述时,广义上都被称为生物安全,但是 biosafety 强调的是防止非故意引起的生物技术及微生物危害,而 biosecurity 强调主动采取措施防止故意的,如窃取及滥用生物技术及微生物危险物质引起的生物危害。¹⁴J. Craig Reed 在此基础上进一步扩大了网络生物安全的内涵,他指出网络生物安全(Cyberbiosecurity)不同于网络生物安保(Cyberbiosafety),并认为网络生物安全作为一个新术语是用于描述与网络数据系统、实验室设备和设施安全及工程控制相关的网络漏洞,这些漏洞可能导致环境污染或对人类、动物和植物健康构成威胁。¹⁵我国学者对“网络生物安全”研究还处于初步阶段,对其作出的界定大多是引用 Murch 的描述,例如,边勇等指出“网络生物安全是保护生物安全、生物资源和生物基础设施系统不受恶意攻击、破坏、窃取、篡改等危害的一系列措施和方法,是生物安全和网络安全交叉融合的新兴安全领域”。¹⁶根据现有研究发现关于网络生物安全的概念并未达成共识。但是,有一点是明确的,即网络生物安全是应对以网络技术的形式对生物数据及设备等进行干扰、破坏、侵入、改变等一系列行为引发的风险(威胁)而保持安全的客观状态。为了更好地处理网络生物安全带来的威胁,必须明晰网络生物安全的内涵,以便尽快建立符合国情的网络生物安全治理机制。

(二) 网络生物安全的潜在风险

1. 网络生物安全面临的军事威胁

¹⁰ Murch R S, So W K, Buchholz W G, et al. Cyberbiosecurity: an emerging new discipline to help safeguard the bioeconomy[J]. Frontiers in bioengineering and biotechnology,2018:39.

¹¹ 同注⑥

¹² Richardson L C, Lewis S M, Burnette R N. Building capacity for cyberbiosecurity training[J]. Frontiers in Bioengineering and Biotechnology,2019,7:112.

¹³ Peccoud J, Gallegos J E, Murch R, et al. Cyberbiosecurity: from naive trust to risk awareness[J]. Trends in biotechnology,2018,36(1):4-7.

¹⁴ 王子灿. Biosafety 与 Biosecurity: 同一理论框架下的两个不同概念[J]. 武汉大学学报(哲学社会科学版), 2006, (02): 254-258.

¹⁵ Reed J C, Dunaway N. Cyberbiosecurity Implications for the Laboratory of the Future[J]. Frontiers in bioengineering and biotechnology,2019,7:182.

¹⁶ 边勇, 严晓, 黄正立. 网络生物安全治理: 问题缘起、现实困境与中国应对之策[J]. 中国国境卫生检疫杂志, 2023, 46 (04): 328-335.

尽管《生物武器公约》的签署消除了大部分生物武器在常规战争中的使用，但 2016 年，美国国家情报局 James Clapper 仍将基因编辑列为美国面临的与大规模杀伤性武器相关的最大威胁之一。¹⁷网络攻击是引起网络生物安全最重要的手段之一，我国在 2016 年制定的《国家网络安全战略》中明确了网络攻击会对整个经济社会造成重大破坏，使得能源、交通、通信、金融等基础设施瘫痪，危害国家经济安全。根据资料显示，美国自 2014 年开始就致力于探索网络生物安全研究，通过政府引导，公私合作等形式探索网络生物安全的覆盖范围、战略价值与风险隐患，美国不断扩大其“生物帝国”的版图并积极将其生物实验室部署扩展到其盟国，为发动网络生物“灰色地带行动”埋下了隐患。¹⁸生物信息价值高，汇集公民个人信息、临床数据、基因组学数据等庞大的信息，以人类基因数据为例，其中蕴含了人种、民族、人群基因、遗传疾病、基因武器等相关信息。¹⁹人工智能与生物学的互联使得数字化的生物信息成为网络攻击的新目标。以 DNA 数据为例，华盛顿大学的研究人员研究发现通过恶意软件可以轻而易举地入侵管控 DNA 计算机系统从而对 DNA 进行编码。²⁰此外，还可以通过计算机数据在生物学家毫不知情的情况下创造新的病毒或生化武器。早在 2010 年 J. Craig Venter 的团队就已经制造了历史上第一个不需要 DNA 物理样本直接利用计算机数据操作而成的细菌 phi X174 病毒 DNA 序列的字母串。²¹当前，支撑生物科学的数字基础设施都是人工智能主导的网络攻击，特别是对抗性数据的操纵，增加了将生物技术实验室和生物制造供应链武器化的潜力。²²这一行为相比传统生物战而言，无需实际进入某国的管辖范围，且无需物理接触生成病毒，即可实现对敌对国家或地区的侵略。

2. 生物信息泄露引发的数据主权风险

生物信息与人工智能、网络的整合正在成为地缘政治，可以将其定义为一个国家的数字主权，涉及国家安全和国际安全。据一项研究发现，超过三分之二的受访者认为相比传统生物技术行业，网络攻击和入侵是造成生物信息的最大威胁。²³2018 年，新加坡医疗系统内大约 150 万例患者的个人信息被窃取，遭遇了有史

¹⁷ Clapper J R. Statement for the Record Worldwide Threat Assessment of the US Intelligence Community Senate Armed Services Committee, 2016, 2 [EB/OL]. (2016-2-09) [2024-6-11] https://www.dni.gov/files/documents/SASC_Unclassified_2016_ATA_SFR_FINAL.pdf.

¹⁸ 穆琳. 网络生物安全威胁加剧警惕演化“灰色地带战争”[J]. 中国信息安全, 2022, (03): 89-91.

¹⁹ 穆琳, 李维杰, 陈海强, 周欣. 全球网络生物安全研究现状与启示[J]. 科技管理研究, 2021, 41(06): 28-32.

²⁰ Researchers are sounding the alarm on cyberbiosecurity [EB/OL]. (2018-2-09) [2023-11-30]. <https://www.c4isrnet.com/dod/2018/02/08/researchers-are-sounding-the-alarm-on-cyberbiosecurity/>.

²¹ Garrett L. Biology's brave new world: the promise and perils of the synbio revolution[J]. Foreign Aff., 2013, 92: 28.

²² Pauwels E. How to Protect Biotechnology and Biosecurity from Adversarial AI Attacks? A Global Governance Perspective[M]//Cyberbiosecurity. Springer, Cham, 2023: 173-184.

²³ Millett K, Dos Santos E, Millett P D. Cyber-biosecurity risk perceptions in the biotech sector[J]. Frontiers in bioengineering and biotechnology, 2019, 7: 136.

以来最强的信息泄露事件。²⁴根据 2020 年《中国互联网网络安全报告》统计, 2020 年我国国内基因数据通过网络出境高达 717 万余次, 涉及我国境内 2.4 万个 IP 地址, 覆盖 31 个省(自治区和直辖市), 其中美国 IP 接收我国生物基因数据的数量名列前茅。《2022 年勒索软件流行态势报告》显示数据泄漏和被勒索的医疗行业属于数据安全泄漏的典型行业。²⁵由于生物数据利用信息技术储存这一特征, 数据泄露为当前最大的挑战, 会对国家的数据主权造成严重威胁。这些生物数据在国家数据安全和网络军备方面具有重要战略地位。由于网络储存生物数据的属性, 使生物资源数据库极易遭到不法分子的破坏, 使国家的生物资源信息泄露, 对国家数据主权造成严重打击。

3. 网络攻击造成的公共卫生风险

网络生物安全还会给公众健康带来重大风险。一方面, 网络生物安全影响医疗领域, 对医疗和临床基础设施构成重大隐患。医疗机构设备, 被恶意行为人通过外部或内部的破坏, 使用网络操纵设备影响患者的健康。²⁶2019 年的内盖夫本古里安大学的研究人员设计了一种恶意攻击医院 CT 扫描中的癌症数据, 产生符合患者独特解剖学的假肺肿瘤, 欺骗医生及人工智能辅助设备从而导致误诊, 且误诊率超过 90%。²⁷此外, 哈佛大学的研究人员针对诊断皮肤癌症图像算法的对抗性攻击的一项研究表明此类攻击只需要修改原始图片中的几个像素即可损坏诊断结果。²⁸另一方面, 网络生物安全也会对公共卫生和生物制药制造工作造成严重影响。例如, 2017 年默克由于受到恶意软件的攻击, 使得生产研究工作瘫痪两周并影响 HPV 疫苗的生产。²⁹2021 年攻击者利用网络设备操纵佛罗里达州和加利福尼亚州水处理设备的化学水平, 对公众健康造成了严重的威胁。³⁰简言之, 当公共卫生设施受到非针对性的网络攻击时, 可能会导致基本药物或疫苗生产工作关闭或者延缓生产的进程。长久以来会导致疫苗或其他关键医疗对策的短缺, 导致疾病无法及时治疗、发病率和死亡率增加, 此类危害如同上述被网络攻击而造成误诊, 从而导致治理无效或构成谋杀。

²⁴ 新加坡医疗信息被盗后加强网络安保[EB/OL]. (2018-11-01) [2024-3-21].
https://www.xinhuanet.com/world/2018-11/01/c_1123649750.htm.

²⁵ 《2022 勒索软件流行态势报告》出炉, 教育和制造业或成勒索受害重灾区[EB/OL]. (2023-02-01) [2023-11-30]. <https://360.net/about/news/article63dc784d58f02a002a0d76e0>.

²⁶ Cyber can now create biowarfare effects, without a bioweapon[EB/OL]. (2022-02-23) [2023-11-30].
<https://breakingdefense.com/2022/02/cyber-can-now-create-biowarfare-effects-without-a-bioweapon/>.

²⁷ Mirsky Y, Mahler T, Shelef I, et al. {CT-GAN}: Malicious tampering of 3d medical imagery using deep learning[C]//28th USENIX Security Symposium (USENIX Security 19). 2019:461-478.

²⁸ Finlayson S G, Bowers J D, Ito J, et al. Adversarial attacks on medical machine learning[J]. Science, 2019, 3639(6433):1287-1289.

²⁹ 生物医药面对网络外患, 何以安也? 网络生物安全事件频[EB/OL]. (2023-06-16) [2023-11-30].
https://mp.weixin.qq.com/s?__biz=MzU3MzgyMTEyOQ==&mid=2247524961&idx=1&sn=1e131779e8c744b8acaa45b71d3a0353&chksm=f3d39fbc7ca4e72d133c0930d7027919d387eb30acf709ee0b56445d817f9aafac5e85a5434&scene=27.

³⁰ U.S. Water Supply System Being Targeted By Cybercriminals [EB/OL]. (2021-7-25) [2023-11-30].
<https://www.forbes.com/sites/jimmagill/2021/07/25/us-water-supply-system-being-targeted-by-cybercriminals/amp/>.

三、网络生物安全治理面临的困境

近年来,网络生物安全问题突出性愈发显著,对国际社会构成了严重的威胁,加剧了国际社会对全球网络生物安全治理的迫切需求。尽管我国已经出台了《中华人民共和国生物安全法》(以下简称“生物安全法”)《中华人民共和国网络安全法》(以下简称“网络安全法”)《中华人民共和国数据安全法》(以下简称“数据安全法”)等一系列法律,但网络生物安全治理依然面临一系列复杂的挑战。

(一) 网络生物安全治理的制度困境

1. 缺乏网络生物安全针对性的法律规定

网络生物安全的治理从不规范治理到有序发展的规制,离不开有效的治理机制。《中国互联网发展报告》(2023)显示我国的网络法治建设逐渐完善、数据安全保护体系也更趋完备,网络空间国际合作也在积极构建已经取得显著成效。目前为止,我国尚未在立法层面明确表述“网络生物安全”这一新术语,但是从过往的立法中可以看到涉及网络生物安全治理的相关条款。例如,2021年施行的《生物安全法》明确了对生物技术、病原微生物实验室、生物资源等做出了具体的规定,并赋予了任何单位和个人不得危害生物安全的义务。《生物安全法》的立法精神标志着我国生物安全法治迈出了里程碑的一步。但需要注意的是,《生物安全法》主要是针对传统生物安全犯罪手段予以规制的法律,在该法中并未涉及运用人工智能技术、互联网等技术侵害生物安全的具体规定,需要进一步从《网络安全法》的范围内找寻网络生物安全治理的法律依据。根据《网络安全法》第五条规定“国家采取措施,监测、防御……保护关键信息基础设施免受攻击、侵入、干扰和破坏……”;第二十六条、第二十七条规定个人和组织不得实施危害网络安全的义务,即“任何个人和组织不得从事非法侵入他人网络……,不得提供……窃取网络数据等危害网络安全活动的程序……”。对于违反《网络安全法》的规定,造成损害的应当承担民事责任,构成犯罪的,依法追究刑事责任。此外,《数据安全法》规定了关于数据安全保护义务,落实数据活动中的组织、个人的主体责任。在涉及网络生物安全治理时,往往会导致《网络安全法》《数据安全法》与《生物安全法》的竞合。当前,网络生物安全的立法仍存在空白,作为一个新型的交叉融合领域,仅仅依靠某一部法律难以充分解决网络生物安全。³¹如果借助上文所述以《生物安全法》为主,与《数据安全法》和《网络安全法》协同治理网络生物安全,必定会在保护法益上呈现碎片化,故为了应对网络生物安全犯罪,我国亟须在原有的法律基础上,弥补这一法律空白。

2. 缺乏明确的网络生物安全威胁和风险评估标准

³¹ 边勇,严晓,黄正立. 网络生物安全治理:问题缘起、现实困境与中国应对之策[J]. 中国国境卫生检疫杂志, 2023, 46 (04): 328-335.

管理任何行为造成的安全风险，都必须最大程度地了解现有的风险和威胁，并且在特定的框架内制定评估标准，科学地评估该风险的等级，以便预测未来可以多大程度降低风险的发生概率从而提出预防措施。当前网络安全评估和生物安全评估已经作出了相关规定。例如，《网络安全法》规定国家实施网络安全等级保护制度，采取数据分类保证网络免受干扰和破坏，防止网络数据泄漏或被窃取，并建立了网络安全监测预警和信息通报制度。《生物安全法》规定需要加强生物安全风险防控和治理，建立生物安全风险监测预警制度，开展生物安全风险调查评估等。但是网络生物安全风险评估仍处于初级阶段，并且人们认识到借用网络安全和信息安全等措施来规制是远远不够的。³²虽然网络生物安全与早期的互联网安全有相似之处，但由于网络生物安全不仅仅涉及网络犯罪的常见形式，还蕴含了生物犯罪等高科技的特征。在前沿性网络生物安全研究中发现“绘制网络生物安全的拓扑才刚刚开始……”，在某些情况下，特定的基础设施调查工具是有用的，但人们认识到全面界定网络生物风险，是相当具有挑战的。³³在面对网络生物安全评估时，我国缺乏具体的评估标准和应急措施，尚未形成针对性地对内保护、对外防御的系列保护举措。

（二）网络生物安全治理的实践困境

1. 治理主体的碎片化

治理主体的碎片化大大削弱了网络生物安全的协同治理效力。《生物安全法》规定由国务院卫生健康、农业农村、科学技术、外交等主管部门和有关军事机关负责国家生物安全相关工作，并由国务院部门组织建立相关领域、行业的生物安全技术咨询专家委员会提供咨询、评估、论证等技术支持。《网络安全法》规定网络安全治理由国家网信部门负责网络安全和监督工作，同时国务院电信主管部门、公安机关和其他机关以及网络运营者、网络提供服务者等均应采取措施保护网络运营安全。《数据安全法》又规定由中央国家安全领导机构负责国家数据安全工作，统筹国家数据安全建立国家数据安全协调机制，如果数据安全涉及网络安全时，由国家网信部门负责网络数据安全和相关监督工作。网络生物安全涉及了生物安全、数据安全、网络安全等各个方面，在交叉重叠领域现行法律并未明确负责网络生物安全的部门。由于欠缺明确且针对性的法律规定，互为关联的横向治理结构的缺失，导致各个主体难以参与到网络生物安全的监管。倘若按照网络生物安全的范围、性质、特征等与现有法律的密切程度分割而治，必然导致治理主体在对网络生物安全监管常态化下存在个人见解、主体意见不一、治理理念和治理措施的冲突，使网络生物安全治理权力分散在各个主体之间。

³² Bajema N E, DiEuliis D, Lutes C, et al. The digitization of biology: Understanding the new risks and implications for governance[J]. Emergence and Convergence Research Paper, 2018 (3).

³³ Richardson L C, Lewis S M, Burnette R N. Building capacity for cyberbiosecurity training[J]. Frontiers in Bioengineering and Biotechnology, 2019, 7: 112.

2.信息数据的碎片化

信息数据的碎片化阻碍了数据共享和部门合作,降低了网络生物安全治理效力。由于治理主体的碎片化,必然使各个部门之间因利益分割而存在“信息孤岛”“数字鸿沟”,从而严重影响国家网络生物安全治理的工作效率。例如,监管生物安全的部门在对生物安全治理的过程中,往往会根据自身的业务进行系统的优化和建设,只注重与生物安全相关的信息采集与保存。在制定生物安全评估标准时,仅按照生物信息的特征进行评估,而忽略了与网络监管部门的互联互通,不同的标准致使信息割据,严重影响网络生物安全治理的灵活性和整体性。

(三) 网络生物安全治理的现实困境

1.犯罪手段增加了打击难度

由于网络技术的加入,打击网络生物安全犯罪的难度大幅度提升。传统的生物安全具有鲜明的科技性、显著的传递性和高度的风险性。³⁴互联网犯罪的特点就是无国界性或跨国界性。网络生物安全犯罪和传统的生物安全犯罪在犯罪手段上呈现新的变化,其犯罪范围辐射更广、犯罪手段更加高科技和隐蔽。³⁵犯罪分子使用虚拟私人网络等技术掩盖自己的真实身份和位置,可以不受地理限制从世界各地发起攻击,使得监管者追踪和定位罪犯变得异常困难。网络生物安全犯罪不同于网络安全犯罪和生物安全犯罪,其采用以网络犯罪的方式达到危害生物安全的犯罪目的,由于犯罪手段的不同,加剧了对网络生物安全犯罪的打击难度。

2.科技支撑不足增加了风险预防难度

当前,网络生物安全的研究仍处于初步阶段,对于网络生物安全造成的风险认识有限。³⁶科技支撑不足是导致网络生物数据遭受攻击的源头。根据调查发现除了学术界、卫生机构等内部威胁以及合法的商业交易导致的数据泄露外,网络计算机系统漏洞是导致受保护的数据被非法攻击的主要原因。³⁷在对病原体基因数据库的网络生物安全研究中发现,目前还未开展针对关于基因数据库安全漏洞的系统研究。³⁸医疗机构信息数据泄露的状况中,其中四分之一的数据都是共享未加密从而导致未经授权的人随意访问、泄露患者记录和个人信息。³⁹早期有学者对转基因生物信息的安全进行研究,通过加密协议等方式降低信息泄露的危险。⁴⁰当前,我国的生物数据采集、分析和储存过度依赖国外的信息技术,由于技术短板的问题,使我国公民生物数据安全防御措施不完善。

³⁴ 焦艳鹏. 总体国家安全观下的生物安全刑法治理[J]. 人民论坛·学术前沿, 2020, (20): 36-45.

³⁵ 江溯. 打击网络犯罪的国际法新机制[J]. 法学, 2022, (11): 45-59.

³⁶ Shankar D D, Azhakath A S, Khalil N, et al. Data Mining for Cyber Biosecurity Risk Management—a comprehensive review[J]. Computers & Security, 2023: 103627.

³⁷ Lynch D J. Biotechnology: The US–China dispute over genetic data[J]. Financial Times, 2017, 30.

³⁸ Vinatzer B A, Heath L S, Almohri H M J, et al. Cyberbiosecurity challenges of pathogen genome databases[J]. Frontiers in bioengineering and biotechnology, 2019, 7:106.

³⁹ Bai G, Jiang J X, Flasher R. Hospital risk of data breaches[J]. JAMA internal medicine, 2017, 177(6):878-880.

⁴⁰ Mueller S, Jafari F, Roth D. A covert authentication and security solution for GMOs[J]. BMC bioinformatics, 2016, 17(1):1-8.

四、网络生物安全治理的具体破解路径

网络生物安全关乎国家安全与人类健康,对我国建立网络强国,全面提升国家生物安全治理能力具有重要意义,我国需直面网络生物安全治理的不足。本文试图从以下几个方面提出中国的破解办法。

(一) 从顶层设计完善网络生物安全治理战略规划

习近平总书记强调“要加强重点领域立法,及时反映党和国家事业发展要求,人民群众关切期待,对涉及全面深化改革、推动经济发展、完善社会治理、保障人民生活、维护国家安全的法律抓紧制订、及时修改”。⁴¹要加强网络生物安全治理,需从顶层设计的角度出发,以全局观的谋略,系统地制定网络安全治理机制,应以习近平新时代中国特色社会主义思想为先导,以《生物安全法》为基础,以“健康中国”为关键,以“国家安全”为核心,携手共建网络生物安全新战略。国家安全战略是筹划和指导国家安全全局的方略。⁴²只有全面提升国家治理体系和治理能力,将网络生物安全放在国家安全战略体系的重要组成部分,才能更好地立足国家实践。当前我国在生物安全治理方面缺乏发展战略规划,网络生物安全更无系统的指导文件。借助域外经验,2018年英国发布《英国生物安全战略》,将识别、预防、发现和有效应对作为生物安全的“四大支柱”。2022年澳大利亚根据2018年发布的《国家生物安全声明》发布了首个《国家生物安全战略》,提出共享生物安全文化、培养高技能劳动力、提高生物安全体系适应能力、加大合作等方面,详细阐述了澳大利亚的生物安全体系和面临的生物安全风险。美国也在2022年出台了《美国国家生物安全防御战略和实施计划》对生物安全的重要性作出了论述。根据现文件来看,中国目前尚未出台类似的生物安全战略规划。在此基础之上,应借鉴现有国际经验,确保国内生物安全治理与国际生物安全治理一致性。为更好地把握网络生物安全治理,应编纂符合中国国情的《中国国家生物安全战略》,将网络生物安全纳入生物安全战略规划中,明确网络生物安全的特殊地位,确保网络生物安全与生物安全治理能力的一致性。

(二) 从整体性视角具体设计网络生物安全治理机制

1.厘清网络生物安全的内涵

在基本法或一般法制定之后,通过出台相关的实施法规或规章来细化规定,推进法律的实施,是我国法治领域的惯常做法。⁴³中国应充分认识网络生物安全的概念和理论框架,科学定义网络生物安全的内涵。国家已颁布针对生物安全治理的法律,即《生物安全法》,作为基础性、全局性的框架法,该法并未直接涉

⁴¹ 许安标.习近平法治思想是新时代科学立法的根本遵循[EB/OL].(2022-10-12)[2024-6-11].
http://www.npc.gov.cn/npc/c2/c30834/202210/t20221012_319553.html.

⁴² 王林.我国新时代的国家安全战略:理论品格、实践展开和体系建构[J].河南警察学院学报,2024,33(01):16-29.

⁴³ 刘长秋.论《生物安全法》中的伦理审查制度及其完善[J].科学学研究,2023,41(09):1563-1569.

及网络生物安全的具体条款。规制具有多重属性的网络生物安全，是否须制定专门的《网络生物安全法》？这一疑问应明确网络生物安全的范畴及其秩序价值。根据前述网络生物安全内涵剖析，并结合《中国互联网报告》中关于网络生物安全的论述可以看出，构成网络生物安全的理论范畴至少包含网络生物安全客体、犯罪的表现形式、危害后果及防控措施等内容。第一，网络生物安全客体，包括生物信息安全和生物资源安全。前者涉及数据的存储、传输、访问控制、加密和完整性保护等；后者要求确保生物材料的安全存储、追踪、标识和访问控制。第二，犯罪的表现形式，网络只是作为一种犯罪手段存在于生物安全犯罪。第三，相关行为体（国家、公权力机关）能够实施相应举措，确保安全与利益免受外部威胁，保证网络生物安全不受外部入侵。根据网络生物安全的理论可知，网络生物安全本质上是为了保护生物安全，网络仅仅只是一种犯罪手段。

文章在前述分析的基础上提出两种方案：其一，在现有制度的基础上，细化现有《生物安全法》的规定，将网络生物安全作为生物安全的特殊类型做专章予以规定，包括网络生物安全的概念、风险评估、风险等级划分和预防应急措施等。其二，制定专门的《网络生物安全法》。笔者认为，根据网络生物安全的理论，应在《生物安全法》现有的框架和归责范围内，利用现有规则的模糊性、宽泛性来完善网络生物安全的治理，将网络生物安全看作生物安全的特殊类型予以防控。具体而言，需要在《生物安全法》第二条“生物安全”的基础上，进一步明确网络生物安全的内涵。主要分为两种形式：一，在“生物安全”原有范畴的基础上扩大其内涵；二，在第二条增加之一，专门对“网络生物安全”的内涵予以界定。笔者将其定义为“国家有效防范和应对生物数据、生物基础设施、生物安保、生物安全等不受人工智能、网络等技术非法攻击、破坏、窃取、篡改等一系列的综合措施及方法。”

2.完善网络生物安全风险评估和预防措施

生物风险管理（bio-risk management，BRM）是一种识别、消除或减轻生物安全和生物安保风险的管理系统方法。⁴⁴习近平总书记指出“必须把防风险摆在突出位置”。⁴⁵网络生物安全涉及重大国家利益、国民健康和国家安全，其牵涉范围的广泛性和不确定性会带来不可估量的严重后果，所以“防患于未然”的防控措施成为最佳的解决办法。保护生物数据免受网络攻击的政策是有限的，例如，《美国联邦法典》中关于选择病原体和毒素（SELECT AGENTS AND TOXINS）部分规定了生物信息安全、网络安全指南，并制定了防止实验室、设备和数据被

⁴⁴ Bowolaksono A, Lestari F, Satyawardhani S A, et al. Analysis of bio-risk management system implementation in indonesian higher education laboratory[J]. International Journal of Environmental Research and Public Health, 2021, 18(10): 5076.

⁴⁵ 中国共产党新闻网.下好先手棋 打好主动仗——习近平总书记关于防范化解重大风险重要论述综述[EB/OL]. (2021-4-15) [2024-6-11] <http://cpc.people.com.cn/n1/2021/0415/c64387-32078296.html>.

非法访问的安全评估。⁴⁶虽然国家在网络安全、生物安全等领域已经尽可能地保护生物数据安全，但是这些措施还远远不够，我们应该考虑如何将措施更好地应用于网络生物安全。针对新兴的网络生物安全问题，在一定程度上可以借鉴传统的生物安全防控措施。ISO31000《风险管理指南》将生物安全防控划分为风险评价、风险识别、风险分析三个环节。在网络生物安全防控时，借鉴生物安全防控环节，遵守风险识别、风险评价、风险分级和风险预防等流程。⁴⁷这些措施不仅可以帮助鉴别潜在的生物威胁，还有助于预防和减轻可能的危害。在考虑到生物、网络和数据安全风险时，必须明确：（1）生物和网络安全之间的合作；（2）端到端的风险评估；（3）针对特定数据的风险和脆弱性评估。⁴⁸

3.协调整合部门数据共享与合作

网络生物安全治理目标的实现需要强化各监管部门之间的互联互通、数据共享与合作。即使我国在制度层面建立针对网络生物安全的具体规定，但由于网络生物安全与风险管理、网络安全、生物安全、生物技术、伦理学等具有密切联系，这种交叉学科的问题需要各个领域的专业部门共同应对。在应对网络生物安全治理时，可以通过设立一个专门的统筹部门负责国家网络生物安全工作的决策和议事，建立国家网络生物安全工作协调机制、数据库和治理平台。在统筹部门的协调下，实现网络安全治理、数据安全治理和生物安全治理等各个权力部门之间的信息数据共享，打破“信息孤岛”和“数据鸿沟”。明确各部门之间的监管范围和职责，合理分配各个监管主体的管治权，构建整体性、系统性的网络生物安全监管网络，提高网络生物安全监管力度。此外，还有学者提出建立公共和私营部门自愿网络生物安全标准，政府和私营企业合作在较短的时间内制定网络生物安全评估标准、激励措施和处罚措施，共同防范网络生物威胁。⁴⁹

（三）提高网络生物安全治理能力

1.加大科技投入力度

技术创新是完善网络生物安全治理的核心驱动力之一。正如习近平总书记所言“我国科技界要坚定创新自信，坚定敢为天下先的志向，在独创独有上下功夫，勇于挑战最前沿的科学问题，提出更多原创理论，作出更多原创发现，力争在重要科技领域实现跨越发展，跟上甚至引领世界科技发展新方向，掌握新一轮全球科技竞争的战略主动”。⁵⁰与欧美国家相比，我国在网络生物安全方面存在认识

⁴⁶ 42 CFR Part 73[EB/OL]. (2024-6-23) [2024-7-2] <https://www.ecfr.gov/current/title-42/part-73>

⁴⁷ 刘光宇，付宏，李辉. 情报视角下的国家生物安全风险防控研究[J]. 情报杂志，2021，40（07）：94-100.

⁴⁸ Berger K M, Schneck P A. National and transnational security implications of asymmetric access to and use of biological data[J]. Frontiers in bioengineering and biotechnology,2019,7:21.

⁴⁹ George A M. The national security implications of cyberbiosecurity[J]. Frontiers in bioengineering and biotechnology, 2019, 7: 51.

⁵⁰ 中华人民共和国国家互联网信息办公室.中国要强必须有强大科技！19 句话读懂习近平科技强国布局[EB/OL]. (2018-1-09) [2024-6-11]https://www.cac.gov.cn/2018-01/09/c_1122230938.htm.

不到位、技术落后等缺点。美国对网络生物安全相关领域积极筹谋，使得中国在网络生物安全领域的核心设备和技术存在被卡脖子的状况，⁵¹科技发展和立法差距令我国在网络生物安全治理中面临极大的被动性。为保证网络生物安全科学治理，必须加大对科技创新机构的投资力度。此外，还须重视对综合型人才的培养，为保证科学、高效的网络生物安全治理，培养的人才需要契合网络生物安全犯罪的特征属性，既要是生物学的专业人员，也要具备互联网、计算机等高新技术，并辅助法学基础知识等综合型人才。科研人员提供专业意见，公权力机关需要评估这些专业意见从而以法律的形式予以规定，使专业意见得到确认。

2.增强网络生物安全威胁意识

国家的网络生物安全离不开公众的参与，仅仅依靠国家通过强制力的立法手段保护网络生物安全是远远不够的。应当增强企业、医疗机构、科学研究院、高新技术研发人员、学者、公众等主体方面的宣传工作，提高公民的网络生物安全风险防范意识，呼吁有关单位开展网络生物安全教育系列讲座，帮助公众认知网络生物安全的重要性。

（四）以人类命运共同体理念为指导共创国际合作平台

习近平总书记提出全球安全倡议，强调人类是不可分割的安全共同体，强调当今世界是一荣俱荣、一损俱损的命运共同体。⁵²人类命运共同体理念是对人类作为生命体终极关怀的一种最高价值诉求，是对人的生命、生存方式和意义的最高承认，关系到人们为何活着、怎样活着的共同求生之道、处世之道。⁵³在进行国际合作时，国家利益无冲突是合作的基础。大部分国家在网络生物安全领域将国家安全置于首要位置，国家为了确保自身生物安全利益不受损害的前提下，限制或者谨慎地参与国家合作。网络生物安全造成的危害不仅仅只在一国的范围传播，而是关系到整个人类的健康安全，从而发展为地区性、全球性的问题。例如，新冠感染在全球的肆虐就是一个很好的生物安全例子。当前，国际上已经形成了生物安全全球治理的总体结构。首先，国际条约提供了法律基础。《生物武器公约》是一个有效地阻止生物制剂作为武器的国际条约，但是仅仅依靠这一制度减少因生物安全引起的灾难性威胁是远远不够的。其次，国际组织作为积极力量推动生物安全的治理。在全球生物安全对话两次会议中提到了新出现的生物风险，第一次会议中提到与技术进步相关的故意和意外风险；第二次会议中提到在建立全球生物安全规范时，应重点关注新出现的生物风险，包括与技术进步相关的风

⁵¹ 王小理. 网络生物安全：大国博弈的另类疆域[EB/OL]. (2019-4-24) [2023-11-30]. https://www.cas.cn/zjs/201904/t20190425_4689832.shtml.

⁵² 中华人民共和国退役军人事务部. 携手推动“三大全球倡议”落地走实 引领人类发展迈向光明未来[EB/OL]. (2023-9-12) [2024-6-11]. http://www.mva.gov.cn/sy/zt/zt1/xxgcdsjsdjs/qwj/202309/t20230914_88627.html.

⁵³ 徐冬根. 论人类命运共同体理念的国际法寓意及贡献[J]. 广东社会科学, 2023, (05): 5-16+286.

险。⁵⁴当前，无论是以中国为中心的发展中国家，还是以欧美为主的发达国家，都在面临着网络生物安全带来的威胁。生物技术已经在全球范围内蔓延，人工智能也以势不可挡的形式快速发展，新一轮交叉领域的威胁正在向全人类涌来。这种威胁具备跨国界性，任何国家都不可能独善其身，单靠一个国家无法解决国际社会普遍关心的问题，各国应寻求在网络生物安全治理中国家利益的最大公约数，为实现人类福祉最大程度的加强国际合作，提高全球网络生物安全治理能力。我国应以习近平总书记提出的“人类命运共同体”为核心，发挥“人类命运共同体”先进理念的作用，贡献中国方案和中国智慧，为实现人类在网络生物安全领域的共同利益，构建“网络生物安全命运共同体”。

Cyberbiosecurity Governance Dilemmas and Paths to Cracking Them

(Shanghai Maritime University, Shang Hai 201306, China)

Abstract: Cyberbiosecurity, as a cross-cutting new type of security field, is characterized by risks such as military threats, data sovereignty leakage, and damage to public health. There are both theoretical and practical dilemmas in governance, among which unclear legal norms, lack of assessment standards, fragmentation of governance subjects, and insufficient scientific and technological support are the great challenges facing governance. In this regard, it is necessary to base on the concept of "community of human destiny", systematically elaborate the connotation of Cyberbiosafety, and put forward measures to improve the strategic planning of Cyberbiosafety governance, improve legal norms of Cyberbiosafety, improve Cyberbiosafety risk assessment and preventive measures, and increase scientific and technological inputs, and so on.

Keywords: Cyberbiosecurity; Cybersecurity; Biosecurity; Cyberbiodestiny Community

⁵⁴ 阙天舒，商宏磊. 全球生物安全治理与中国的治理策略[J]. 社会主义研究，2022（02）：165-172.

网络空间的武力使用与国际人道法中区分原则适用研究

陈熙杨

摘要：在网络技术高速发展的当下，网络攻击展现出与传统攻击截然不同的特性，但其引发的后果严重程度可与传统武力攻击相当，对国际和平与安全造成了极为严峻的威胁。“震网”病毒攻击伊朗核设施、乌克兰关键基础设施遭网络袭击等事件，便是典型例证。在现有国际法体系下，对于网络攻击能否被认定为《联合国宪章》框架内的“使用武力”，尚未达成统一认识。网络攻击的虚拟性、隐匿性，使得传统“武力”的界定标准难以直接套用。在网络空间中，军民两用目标的界限模糊不清，大量兼具民用和军事功能的网络设施，使得国际人道法中的区分原则在网络攻击场景中难以发挥实效。为应对这些问题，国际社会应制定专门规范网络攻击的国际人道法议定书。各国也需强化情报共享与国际协作机制，在技术交流、信息通报等方面加强合作，共同识别、应对网络攻击威胁。网络空间的法治化建设需要各国秉持合作精神，积极参与国际规则的制定与实施，综合运用法律、技术与合作手段，共同推动国际法在网络空间的适应性变革。

关键词：网络攻击 国际人道法 区分原则 武力使用 自卫权

一、引言

近年来，网络空间的法治化进程持续深化，但网络攻击的破坏性及其对国际法的挑战日益凸显。从历史事件来看，网络攻击的破坏性已不容小觑，其实际后果可与传统军事行动相提并论。2010年“震网”（Stuxnet）蠕虫病毒攻击伊朗核设施，导致大量离心机瘫痪并延缓其核计划进程；2015年以来，乌克兰能源、交通等关键基础设施多次遭受网络攻击，严重干扰社会正常运转并引发人道主义风险。当前，全球性网络攻击事件频发，新技术新应用不断涌现，国际地缘冲突延伸至网络空间，重大网络攻击事件层出不穷。¹据德国“Statista”网站统计，2023年勒索软件成为全球检测到的最频繁网络攻击类型，约占所有检测到网络攻击的70%。²这种全球性网络威胁凸显了国际合作的紧迫性。此类事件表明，网络攻击不仅会对国家关键基础设施造成严重破坏，还可能引发人道主义危机。

¹ 参见 桂畅旒,刘星.2023年国际网络空间形势回顾及发展动向[J].中国信息安全,2023,(12):19.

² 勒索软件有可能严重破坏业务运营并造成声誉和财务损失，因此仍然是全球最持久的网络威胁之一。这种类型的恶意软件会加密包含敏感信息的文件，要求支付赎金作为回报。数据来源参考：Statista. (2023). Global cyberattacks by type 2023. Statista. Retrieved January 21, 2025, from <https://www.statista.com/statistics/1382266/cyber-attacks-worldwide-by-type/>

同时，网络攻击的隐匿性和复杂性加剧了法律适用的难度。攻击者常借助匿名手段发动攻击，溯源和责任认定极为困难。其跨国性特征也使得国际社会必须共同制定规则应对挑战。网络空间的发展背景和网络攻击的特性，使得武力法和国际人道法在网络空间中的适用成为亟待解决的问题。本文采用案例分析法和文献研究法，通过剖析“震网”病毒攻击伊朗核设施、乌克兰关键基础设施网络攻击等典型案例，揭示网络攻击行为的复杂性、隐匿性和跨国性特征。结合国际法相关文献，梳理和借鉴已有研究成果，主要围绕网络空间武力使用情境下，国际人道法中“区分原则”的适用展开。旨在通过分析网络攻击的独特属性，探讨其是否构成《联合国宪章》框架下的“使用武力”，并进一步论证国际人道法区分原则在网络空间的适用困境与解决路径，以期国际法在网络空间的适应性改革提供理论支持。

二、网络攻击的武力属性界定与自卫权行使的法律框架

（一）传统武力与网络攻击的差异性分析

《联合国宪章》第2条第4款明确规定禁止国家在国际关系上使用武力，但并未对“武力”一词的内涵做出具体界定。基于《联合国宪章》的全文表述及立法背景等因素，多数学者认为《宪章》第2条第4款中的“武力”应当被限定为武装力量。

³所以，传统的“使用武力”概念通常是指通过物理手段实施的武装力量行为。

“尼加拉瓜案”中，相关军事行动造成了尼加拉瓜大量的人员伤亡和基础设施毁坏，这一严重后果成为判定该行为构成“使用武力”的关键因素。国际法院在该案判决里进一步确立了“武力”的判例标准，即行为需具有“规模与后果的严重性”，具体表现为直接造成人员伤亡或基础设施损毁等物理性破坏。⁴这一标准在传统武装冲突中具有明确的可操作性，例如军事打击、空袭、地面冲突等行为因其可见的破坏后果，能够被现有国际法框架清晰界定为“使用武力”。

然而，网络攻击的独特性对传统“严重性”标准提出了挑战。首先，网络攻击的虚拟性使其破坏不依赖物理手段，而是通过数据篡改或服务瘫痪实现。在2015年乌克兰电网攻击事件中，黑客通过恶意软件切断电力供应，虽未造成人员的直接伤亡，却导致数十万人陷入生存危机，其后果的严重性符合传统标准，但手段的虚拟性突破了原本《宪章》对“武力”的物理性预设。其次，网络攻击的隐匿性与跨国性导致责任归因困难。攻击者可通过僵尸网络或加密技术掩盖身份，使溯

³ 参见 黄瑶.论禁止使用武力原则——《联合国宪章》第2条第4项法理分析[M].北京：北京大学出版社,2003 年.

⁴ 尼加拉瓜诉美国军事与准军事活动案。尼加拉瓜于 1984 年 4 月 9 日向国际法院提出请求，指控美国在其境内对尼加拉瓜采取的军事和准军事行动，请求国际法院判定美国的行动构成非法使用武力和以武力相威胁干涉尼加拉瓜内政和侵犯尼加拉瓜主权的行为。国际法院最后在 1986 年 6 月 27 日，对本案作出判决。资料来源：<https://www.icj-cij.org/case/70/judgments>，访问时间：2025 年 3 月 9 日

源与归责陷入僵局，而传统武力因物理行动的可见性，如军队调动等，更易追踪。最后，网络攻击的复杂性表现为后果的延迟性与扩散性。例如，“震网”病毒对伊朗核设施的破坏效应在数月后才完全显现，与传统武力“即时可见”的严重性特征形成鲜明对比。

综上所述，传统“武力”使用方式更为直接，敌方难以忽视；而网络攻击具有虚拟性、隐匿性、跨国性和复杂性等特点。攻击者可以利用各种技术手段，如 IP 地址伪造、代理服务器、僵尸网络等进行攻击，隐藏自己的真实身份和位置，使得追踪溯源工作面临巨大挑战。攻击者身份追踪困难，攻击方式和后果难以预测，防御难度大。即使通过技术手段初步锁定了攻击来源，也很难在法律上明确将攻击责任归因于某个特定的主体，尤其是在涉及国家行为时。⁵二者在作用机制、后果评估与归责逻辑上的差异，凸显了现有国际法框架在网络空间的适用困境。

（二）网络攻击构成“使用武力”的法律标准

依据《联合国宪章》规定以及国际法院判例来看，判断网络攻击是否构成“使用武力”，不仅取决于攻击性质，还需考量其严重性及后果。若网络攻击造成的后果与传统武力破坏相当，如攻击国家关键基础设施，导致人员伤亡、大规模物资损失或严重威胁国家安全，则可能被视为“使用武力”。

国际法院在“威胁或使用核武器的合法性”1996 年的咨询意见中表示，对于是否构成武装冲突，无论使用何种武器，而是关注使用武力的后果。⁶这一司法观点也表明判断某一行为是否构成《联合国宪章》第 2 条第 4 款禁止的“使用武力”，关键在于考察其造成的破坏性后果是否与传统武力相当，而非拘泥于攻击手段的具体形式。正如核武器与传统武器的区别不影响其武力属性判定，网络攻击的虚拟性特征同样不应成为排除其“使用武力”性质的理由。⁷

2010 年美国实施的“震网”病毒攻击伊朗核设施，是一次高度针对性的网络战。此事件促使国际社会深入探讨网络攻击是否应被视为“武力使用”。⁸美国这一网络攻击行动，虽不符合传统“武力”标准直接造成人员伤亡，但其通过瘫痪伊朗核设施，实质削弱了伊朗的战略能力，并间接威胁地区安全。若根据“尼加拉瓜案”的“规模与后果严重性”标准，此类攻击可能被认定为“使用武力”。俄罗斯在 2015 年和 2016 年对乌克兰电力网发动的网络攻击，致使大范围电力中断，严重干扰了乌克兰民众的正常生活和国家经济运行。从后果看，尽管没有人员伤亡，但其破坏程度、目标的关键性以及乌克兰国家安全的影

⁵ 参见 左亦鲁.国家安全视域下的网络安全——从攻守平衡的角度切入[J].华东政法大学学报,2018,21(01):152.

⁶ International Court of Justice. Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, July 8, 1996, para 39.资料来源: <https://www.icj-cij.org/case/95>, 访问时间: 2025 年 3 月 25 日

⁷ 参见 Dan-Iulian Voitasac, Applying International Humanitarian Law to Cyber-Attacks, 22 LEX ET SCIENTIA INT'L J. 124 (2015).128.

⁸ 参见 何志鹏,王韵涵.网络攻击困境与应对的国际法审思——以“武力”为视角[J].哈尔滨工业大学学报(社会科学版),2023,25(05):50.

击具有类似武力攻击的特征。⁹网络攻击在性质上与传统“武力”不同，但其潜在的破坏性和影响，在严重情况下可被视为“武力使用”。

结合俄罗斯对乌克兰的网络攻击和美国对伊朗的网络战案例可知，尽管网络攻击形式与传统“武力”有所不同，随着网络攻击破坏力和影响范围扩大，当其后果对国家安全、社会秩序等造成重大影响时，有可能被认定为“使用武力”，网络攻击也可能符合《联合国宪章》对“武力”的定义。

从法理层面深入分析，《联合国宪章》禁止使用武力旨在维护国际和平与安全秩序。网络攻击若严重破坏这一秩序，如导致国家关键基础设施瘫痪，引发社会动荡，进而影响国际间正常的政治、经济交往，从维护国际秩序的本质目的出发，也应将其纳入“使用武力”范畴。网络攻击造成的社会秩序混乱与传统武力攻击具有等效性，两者都对国际社会的稳定发展构成威胁。

（三）网络空间自卫权行使的条件与限制

明确网络空间中“武力”的概念以及网络攻击是否构成“使用武力”，是探讨网络空间自卫权的重要前提。若网络攻击产生与传统武力相当的后果，尤其是对社会秩序、国家经济或安全造成大规模破坏，则可被视为“使用武力”。¹⁰在此情况下，受攻击国的自卫权如何适用成为关键问题。现有的国际法框架中，对自卫权适用的具体规定主要体现在《联合国宪章》第 51 条，自卫权的行使需满足即时性、必要性和比例性等条件，以确保其合法性与正当性。

1. 自卫权的法律依据

自卫权是一国使用武力反抗非法攻击而保护自己的权利。在国际法的框架下，自卫权的行使并非毫无限制，而是需要严格满足一系列特定的法律条件，这些条件共同构建起了自卫权合法行使的基石。

自卫权其核心依据来源于《联合国宪章》第 51 条，该条款明确规定：“联合国任何会员国受武力攻击时，在安理会采取必要办法以维持国际和平与安全以前，本宪章不得认为禁止行使单独或集体自卫之自然权利。”这一规定确认了国家在遭受“武力攻击”时享有固有的自卫权，但同时也对其行使设定了严格的条件：第一，自卫权的触发必须以“武力攻击已发生”为前提，传统解释认为，自卫权仅适用于实际发生的武力攻击，而非潜在的威胁；第二，国家在行使自卫权后，必须立即向安理会报告，且不得妨碍安理会采取必要措施；第三，自卫行动必须符合“相称性与必要性”原则，即自卫措施必须与攻击的规模和性质相匹配，且仅限于解除威胁的必要范围内。

2. 网络自卫权的合法性与限制

在网络攻击日益频繁且形式多样的背景下，网络空间作为国家主权在数字化

⁹ 参见 王韵涵.网络攻击中国际法适用问题研究[D].吉林大学,2024:21.

¹⁰ 参见 王岩.适用于网络空间行动的国际法[J].武大国际法评论,2019,3(06):145.

时代的关键延伸领域，与国家领土安全、军事安全同等重要，同样需要得到充分保护。网络攻击是否构成“武力攻击”并触发国家的自卫权，已成为国际法领域的重要议题。2007 年俄罗斯对爱沙尼亚关键基础设施的网络攻击，促使北约成立网络防御卓越中心，该中心编纂、修订、更新并出版了《塔林手册》与《塔林手册 2.0》。根据《塔林手册》及其 2.0 版本的规定，网络攻击若规模庞大且造成严重人员伤亡或财产损失，即可被视为“武力攻击”。¹¹

由此相关规定可得，网络攻击构成“武力攻击”并达到使用自卫权标准需满足以下条件，首先，攻击须达到造成重大人员伤亡或财产损失的程度；其次，攻击须能归因于国家行为体或受国家支持的非国家行为体；最后，自卫行动须具备即时性、必要性和比例性，以确保其合法性。

当网络攻击来自其他国家且符合上述“武力使用”标准时，网络攻击不再是简单侵扰，而是对主权和国家安全的实际威胁，受攻击国有权且具备合法基础行使自卫权。

基于上述理论，若国家在网络空间中遭遇破坏，且该破坏程度与传统领域战争中的武力攻击相当，应当援引国家自卫权以恢复和平。然而，在实际运用过程中，这一做法却面临着众多存在争议的复杂情形。

在国际法框架体系内，国家行使自卫反击权需具备合法性，为确保网络自卫权的合理行使，避免其被滥用，网络自卫权的行使必须遵循一系列限制条件，其中比例性原则尤为重要。在比例性原则要求下，网络自卫行为的强度和规模应与遭遇的网络攻击的严重程度精确匹配。比如，在遭受小规模网络数据窃取攻击时，受攻击国不应采取全面摧毁对方网络基础设施的极端反击方式，而应精准定位攻击者，采取针对性的措施进行反击，以最小的代价恢复自身安全状态，同时避免过度反应和无差别攻击，防止引发更广泛的网络冲突。

《联合国宪章》第 51 条规定，自卫行动必须及时向联合国安理会报告。然而，在网络自卫场景中，这一程序要求的适用性引发了广泛讨论。网络攻击的匿名性和瞬时性使得确定攻击者身份及攻击范围极为困难，受攻击国可能因忙于应对攻击而难以在第一时间履行报告义务。¹²而如果没有立即向安理会报告，是否会导致国家的自卫措施无效。这个问题曾经在“尼加拉瓜案”中提出，国际法院没有予以定论。施韦贝尔法官在其不同意见中认为，自卫权是一项固有权利，而向安理会报告是一个程序性条件，没有报告并不必然剥夺一个国家自卫的实质权利。¹³没有报告可能是不利于国家主张自卫权的因素之一，但不行为本身不应该损害自卫措施的合法性。尽管如此，《联合国宪章》规定的报告程序对维护国际法的

¹¹ 参见 袁泉,刘静.网络自卫权行使的理论依据与建构[J].江西社会科学,2021,41(06):212.

¹² 参见 陈颀.网络安全、网络战争与国际法——从《塔林手册》切入[J].政治与法律,2014,(07):155.

¹³ Schwebel.Dissenting Opinion Of Judge Schwebel[R].International Court of Justice,1986, para.222.

公正性与透明度至关重要，也不应因网络攻击的特殊性而被削弱。¹⁴

为平衡网络攻击的即时性与报告义务的履行，建议建立快速网络报告机制。例如，可通过联合国网络安全机构协调信息共享，或利用国际组织的技术手段协助受攻击国迅速确定攻击源并通报安理会。此类机制不仅有助于保障国际法在网络空间的有效实施，还能增强各国在网络自卫行动中的互信与合作。

三、国际人道法“区分原则”在网络攻击中的适用

（一）区分原则的核心要求

国际人道法是一系列旨在限制武装冲突所造成之人道后果的规则。¹⁵尤其是对战争中民众和战斗人员的保护的法律框架。国际人道法强调区分原则，要求交战各方在军事行动中明确区分军事目标和民用目标，这一原则的核心是保护无辜平民，防止其在冲突中受到不必要伤害。¹⁶《日内瓦公约》及其附加议定书规定，交战方有责任确保军事行动遵循该原则。

根据《日内瓦公约第一附加议定书》第 48 条的规定，区分原则要求冲突各方明确区分民用物体与军事目标，并且只可以攻击那些被认为是军事目标的物体。具体而言，第 52 条进一步规定，军事目标必须符合以下标准：“军事目标指那些通过其性质、位置、用途或使用对军事行动有实际贡献的目标”。这一规定不仅强调了目标的“性质”，还强调了目标的“用途”及“位置”，即目标在特定情境下的作用是否与军事目标的目的相符。

区分原则的核心目的是保护平民和民用设施免遭战争的无差别攻击。这一原则深刻影响了国际人道法对战时武力使用的规范性界定，特别是在针对网络攻击这一新型威胁时，传统的区分原则是否能适应现代战争的需求，成为了当前讨论的关键问题。

（二）区分原则在网络空间中的适用与挑战

《塔林手册》承认国际人道法在网络行动中的适用，区分原则同样适用于武装冲突中的网络攻击。¹⁷网络攻击作为一种非传统的武力使用方式，其性质和执行方式与传统的物理攻击大相径庭，给区分原则的实施带来了前所未有的挑战。

1. 军民两用基础设施的模糊性

在网络攻击背景下，应用区分原则面临诸多困难。在数字时代，网络攻击形式多样，攻击目标往往不直观，许多基础设施如电力和通讯系统兼具军民两用性，

¹⁴ 参见 余民才.国际法上自卫权实施机制[M].北京：中国人民大学出版社,2014.

¹⁵ 尼尔斯·梅尔策.国际人道法导论[M].日内瓦：国际红十字委员会,2016.资料来源：<https://www.icrc.org/zh/publication/4231-international-humanitarian-law-comprehensive-introduction>.访问时间：2025 年 3 月 12 日.

¹⁶ 马光.论国际法上网络安全的定义和相关国际规则的制定[J].中国政法大学学报,2019,(03):66.

¹⁷ 参见 张天舒.从《塔林手册》看网络战争对国际法的挑战[J].西安政治学院学报,2014,27(01):90.

攻击者在决策时难以明确区分军事与民用目标。¹⁸例如，针对电力供应网络的攻击可能短时间内达到削弱敌方军事能力的目的，但同时也可能导致大范围的民用损失，影响医院、学校和家庭的正常运转。¹⁹因此，需重新审视国际人道法中关于平民保护的原则，确保其在网络空间有效实施。

网络攻击对平民保护义务构成多方面挑战。因其常针对军事目标（如敌方指挥中心），但现代社会基础设施高度互联，致使攻击极易波及民用设施。以2017年乌克兰遭受的“NotPetya”勒索病毒攻击为例，该攻击借助恶意软件迅速传播，本可能针对军事或政府系统，却蔓延至民用领域，大量医院信息系统瘫痪，医疗设备无法正常运行，严重阻碍医疗服务供给，危及患者生命安全，²⁰这显然违背国际人道法对区分平民保护的基本要求。

2. 网络攻击后果的不可预测性影响

网络“武器”的不可控性也是对区分原则的重大挑战之一。网络攻击具有很高的匿名性和不可预测性，尤其是网络“武器”的扩散及其效果常常难以在攻击发生前完全掌控。

“WannaCry”病毒是由美国国家安全局（NSA）开发的网络武器，在2017年由于技术漏洞泄露，迅速蔓延至全球超过150个国家，并影响了包括非冲突国家在内的众多机构和组织。WannaCry病毒不仅针对特定目标，还波及到许多无辜的民用设施和机构，如英国国民健康服务系统（National Health System，简称NHS）在此期间遭遇到了有史以来最大的网络攻击。²¹这一病毒的传播方式和后果展示了网络攻击可能带来的“无差别攻击”效应，直接违反了区分原则。尤其是，当网络武器无法预见地扩散至非冲突方时，难以确定其是否仍能符合国际人道法的区分要求。

3. 网络攻击的隐匿性和复杂性

网络攻击常由匿名者发起，攻击者借助虚拟专用网络（VPN）、加密技术等手段隐匿身份，利用网络环境的复杂性实施隐蔽攻击。尽管当前存在流量分析、威胁情报共享等追踪技术，但攻击者频繁变换攻击路径、使用加密通信，极大增加了跨国追踪与问责的难度。²²而且，跨国追踪面临不同国家法律管辖权不统一、数据跨境获取限制等法律障碍，导致难以有效追究攻击者责任。由于攻击的匿名性，交战各方在实施网络攻击时难以预测后果，使得事前评估损害的难度剧增，遵循国际人道法中平民保护义务的要求变得更为复杂。

¹⁸ 焦园博.论军民融合趋势下网络武装冲突适用区分原则[J].国际法研究,2023,(04):60.

¹⁹ 克努特·德尔曼,洛朗·吉塞勒,蒂尔曼·罗登霍伊塞尔,等.国际人道法对网络战的适用性及其适用[J].国际法研究,2019,(04):5.

²⁰ Kristen E. Eichensehr, Ukraine, Cyberattacks, and the Lessons for International Law, 116 AJIL UNBOUND 145 (2022).

²¹ 唐岚.从 WannaCry 事件看网络空间国际规则的困境及思考[J].云南民族大学学报(哲学社会科学版),2019,36(06):145.

²² 参见 蔡高强,焦园博.“网络武装冲突”的国际法阐释[J].湘潭大学学报(哲学社会科学版),2017,41(03):20.

4. 国际法律框架的不完善

现有国际法在网络空间中的适用性不足是一个显著问题。国际人道法提供了区分原则，但网络空间中的网络冲突与实际攻击空间具有明显区别，能否适用以及如何适用国际人道法，以避免网络空间冲突中的不必要的平民利益损害成为不得不思考的问题。²³各国对网络技术和攻击方式的认知差异，使得在法律框架下处理跨国网络攻击时面临重大困难。网络攻击的迅速演变和技术更新速度超出法律制度的适应能力，进一步加剧了这些挑战。

当前的国际人道法在网络空间的适用存在法律空白，缺乏专门针对网络攻击中区分原则的具体规定。现有的法律条款在解释和适用上存在模糊性，对于“军事目标”和“民用目标”在网络空间的界定不够清晰。不同国家对这些概念的理解和实践标准存在差异，导致在处理网络攻击事件时，各国难以依据统一的标准判断攻击行为是否符合区分原则，难以有效保障区分原则的实施。这一法律框架的不完善，使得网络空间的平民和民用设施在面对网络攻击时缺乏足够的法律保护。

综上所述，网络空间的复杂性、基础设施的双重功能以及国际法律框架的不足，共同构成了网络空间中平民和非战斗人员保护的重重障碍。这进一步凸显了国际社会重新审视网络攻击与武装冲突关系、完善国际法适用方法以应对网络安全威胁的紧迫性。

四、应对网络空间区分原则适用困境的建议

在网络技术迅猛发展的当下，网络攻击的威胁与日俱增，在网络空间中的适用区分原则遭遇了前所未有的挑战。为有效应对这些挑战，维护国际网络秩序，保障平民和民用设施的安全，需从法律规范、技术手段、国家间合作以及国际组织作用等多个维度协同发力。

（一）完善网络空间领域国际法律规范

现有的国际人道法体系在传统战争场景下为区分原则的实施提供了指导框架，但在网络战争中，其局限性日益凸显。网络攻击的特殊性，如攻击手段的多样性、目标的虚拟性和攻击后果的不确定性，使得现有法律框架难以有效应对。因此，制定一部适用于网络空间的国际人道法议定书迫在眉睫。

该议定书应针对网络空间特性，制定明确的操作指南，指导冲突各方遵循区分原则。这些指南应当涵盖军事目标与民用目标的区分方法，以及网络攻击军事必要性和比例性的评估标准。同时，国际社会需要构建更明确的法律框架，界定和应对网络攻击中的“武力使用”问题，明确网络攻击“使用武力”的具体标准，在坚持《联合国宪章》禁止使用武力的基本宗旨和精神的前提下，根据网络攻击的特殊性质，专门构建适用于国际网络攻击行为的判定规范体系。²⁴此标准可以参

²³ 鞠徽.论国际人道法在网络攻击中的适用问题[J].中国公共安全(学术版),2018,(03):79.

²⁴ 刘仑.网络攻击构成“使用武力”的判定——对《联合国宪章》第2条第4款的阐释[J].社会科学战

考《塔林手册》中的“损害结果说”，即以攻击的规模和后果作为判断依据。通过明确的规范，有助于各国更精准地理解和适用区分原则，最大程度减少网络攻击中对民用设施的无差别伤害。这一标准的制定应旨在维护国际法的核心原则，引导各国在网络空间尊重主权原则，维护自卫权，并就国家安全事务做出明智的决定。²⁵通过明确的规范，有助于各国更精准地理解和适用区分原则，最大程度减少网络攻击中对民用设施的无差别伤害。

此外，议定书可以设立“关键民用基础设施清单”，明确列举医疗系统、能源系统、金融系统、水利系统等在战争期间对平民生存和社会稳定具有不可替代作用的民用基础设施。在网络攻击场景下，这些关键民用基础设施应受到特别保护，避免被错误地作为军事目标遭受攻击。制定该清单有助于国际社会在冲突期间达成共识，明确优先保护对象，增强对平民和民用设施的保护力度。

（二）加强技术手段支持

网络空间的独特性决定了技术手段在确保区分原则有效实施过程中具有关键作用。当今世界，科学技术的迅速发展提供了多种前沿技术，如人工智能、大数据和区块链等，为网络攻击的目标识别和防护提供了有力工具。

首先，利用实时数据流分析和网络流量监控技术，能够及时捕捉攻击目标的性质和目的信息。这些技术可以帮助决策者迅速区分军事目标和民用目标，确保网络攻击行动符合区分原则。借助大数据分析和人工智能技术，对网络流量中的特征数据进行深度挖掘和分析，从而准确判断目标的军事属性，避免误击民用设施。

其次，人工智能技术可通过对网络行为的分析，识别异常模式，辅助确定网络攻击目标是否符合“军事目标”标准。人工智能不仅能提高目标判断的速度，还能通过对大量数据的学习和分析，更精准地判断目标是否具有军民两用性，以及攻击可能对平民造成的影响。例如，通过模拟攻击后果，人工智能可以帮助决策者评估攻击的潜在影响，从而避免对民用设施的误伤。

最后，加密技术和身份验证机制在保障网络空间安全方面至关重要。这些技术有助于实现网络攻击的准确溯源，降低因网络攻击身份模糊和责任归属不清带来的问题，进而提高法律责任追究的效率。大力发展和广泛推广先进的网络防护技术，尤其是人工智能、实时监控和数据分析技术，能够为区分原则的遵循提供强有力的技术保障。

（三）提升国家间合作与协调

网络空间的无边界性使得网络安全问题和网络战争往往跨越多个国家的管

线,2016,(12):274.

²⁵ 参见 Catherine Lotrionte, *State Sovereignty and Self-Defense in Cyberspace: A Normative Framework for Balancing Legal Rights*, 26 EMORY INT'L L. REV. 825 (2012):832.

辖范围，因此，国家间的合作与协调对于维护网络空间秩序至关重要。

首先，各国应强化网络安全情报的共享机制，特别是在跨国网络攻击事件发生时，及时进行信息沟通和协作。通过共享情报，各国能够更快速地识别攻击源，准确判断攻击意图，降低误击无辜目标的风险。例如，美国现已将其联盟战略从物理空间平移到网络空间，美国与盟国实现了包括流量收集、设备获取、使用分析、密码分析以及人工智能、秘密行动、反间谍、地理空间情报、执法和交通安全等在内的广泛情报合作，建立起覆盖全球的侦察监视网络，增强网络空间集体安全的态势。在网络联盟内部定期举行网络联合演习和人才培养，磨炼各自网军力量，同时利用盟友的技术和资源提升自己的网络能力。²⁶这种多国协作充分体现了情报共享和联合响应机制的重要性。

其次，各国应积极在联合国等国际平台上推动网络空间国际规则和标准的制定工作，确保各国在网络攻击场景下能够遵循统一的法律框架和规范。这不仅能够增强网络攻击相关法律法规的约束力，还能为各国在面临网络战时提供一致的行动准则。

五、结论

随着网络技术的快速发展，网络攻击的形式日益多样化和复杂化，网络空间中的法律适用问题亟需国际社会的共同关注与协同应对。

网络空间的“区分原则”在适用中面临诸多挑战，尤其是军民两用基础设施的模糊性、攻击后果的不可预测性以及国际法律框架的不完善，使得区分原则的有效实施变得尤为困难。

未来，国际社会应持续关注网络技术的发展动态，深入研究新型网络攻击对区分原则适用的影响，进一步完善网络空间国际人道法的理论与实践。通过制定专门的国际人道法议定书、利用前沿技术辅助目标识别与攻击溯源、强化国家间合作与情报共享，国际社会可以更好地应对网络攻击，维护国际网络秩序与人类和平安全。

网络空间的法治化进程需要全球共同努力。国际社会应携手推动网络空间的规则制定与实施，构建一个和平、安全、有序的网络空间，为全球网络安全奠定坚实基础。

参考文献

1. 蔡高强,焦园博.“网络武装冲突”的国际法阐释[J].湘潭大学学报(哲学社会科学版),2017,41(03):16-23.
2. 陈颀.网络安全、网络战争与国际法——从《塔林手册》切入[J].政治与法律,2014,(07):147-160.

²⁶ 王桂芳,王南.美国网络联盟发展及其对我国的影响[J].中国信息安全,2019,(09):40.

3. 桂畅旒,刘星.2023 年国际网络空间形势回顾及发展动向[J].中国信息安全,2023,(12):19-23.
4. 何志鹏,王韵涵.网络攻击困境与应对的国际法审思——以“武力”为视角[J].哈尔滨工业大学学报(社会科学版),2023,25(05):49-56.
5. 黄瑶.论禁止使用武力原则——《联合国宪章》第 2 条第 4 项法理分析[M].北京:北京大学出版社,2003 年.
6. 焦园博.论军民融合趋势下网络武装冲突适用区分原则[J].国际法研究,2023,(04):57-70.
7. 鞠徽.论国际人道法在网络攻击中的适用问题[J].中国公共安全(学术版),2018,(03):79-82.
8. 克努特·德尔曼,洛朗·吉塞勒,蒂尔曼·罗登霍伊塞尔,等.国际人道法对网络战的可适用性及其适用[J].国际法研究,2019,(04):3-17.
9. 刘仑.网络攻击构成“使用武力”的判定——对《联合国宪章》第 2 条第 4 款的阐释[J].社会科学战线,2016,(12):271-274.
10. 马光.论国际法上网络安全的定义和相关国际规则的制定[J].中国政法大学学报,2019,(03):64-78+207.
11. 尼尔斯·梅尔策.国际人道法导论[M].日内瓦:国际红十字委员会,2016.
12. 唐岚.从 WannaCry 事件看网络空间国际规则的困境及思考[J].云南民族大学学报(哲学社会科学版),2019,36(06):145-156.
13. 王桂芳,王南.美国网络联盟发展及其对我国的影响[J].中国信息安全,2019,(09):39-41.
14. 王岩.适用于网络空间行动的国际法[J].武大国际法评论,2019,3(06):139-157.
15. 王韵涵.网络攻击中国际法适用问题研究[D].吉林大学,2024.
16. 余民才.国际法上自卫权实施机制[M].北京:中国人民大学出版社,2014.
17. 袁泉,刘静.网络自卫权行使的理论依据与建构[J].江西社会科学,2021,41(06):210-218.
18. 张天舒.从《塔林手册》看网络战争对国际法的挑战[J].西安政治学院学报,2014,27(01):87-92.
19. 左亦鲁.国家安全视域下的网络安全——从攻守平衡的角度切入[J].华东政法大学学报,2018,21(01):148-157.
20. Catherine Lotrionte, State Sovereignty and Self-Defense in Cyberspace: A Normative Framework for Balancing Legal Rights, 26 EMORY INT'L L. REV. 825 (2012).
21. Dan-Iulian Voitassec, *Applying International Humanitarian Law to Cyber-*

Attacks, 22 LEX ET SCIENTIA INT'L J. 124 (2015).

22. Kristen E. Eichensehr, *Ukraine, Cyberattacks, and the Lessons for International Law*, 116 AJIL UNBOUND 145 (2022).

2025年中国国际法学会年会

国际安全视域下人工智能武器的人道困境与因应

郝思源¹

摘要：人工智能技术随着人类在科学领域的不断探索而飞速发展，并逐渐在军事领域得到广泛应用，人工智能武器应运而生。在武装冲突中，人工智能武器可以在人类参与程度较低的情况下自主完成目标选择并实施打击，在未来甚至可以发展到完全不需要人类参与而作出关键决策的程度。这种区别于传统武器的自主性特点给国际人道法带来了一系列的挑战。首先，人工智能武器的法律地位模糊，高度自主性的人工智能武器属于战斗员抑或新武器并无定论；其次由于人类控制因素在关键决策中的缺位以及现有的技术限制，人工智能武器在使用中往往难以应对突发情况，难以评估攻击目标与强度，容易对区分原则与比例原则造成冲击；此外，人工智能武器的自主性和致命性特点还造成了国际不法行为责任中难以归因于国家，个人刑事责任中主观要件难以认定等归责困境。为了应对人工智能武器给国际人道法带来的挑战，弥补既有归责体系的空白，形式上应以增列CCW附加议定书和制定专门国际条约的方式填补现有法律规范的空白，进而建立健全的人工智能武器国际人道法规制体系；在内容上，要确立“有意义的人类控制”原则，禁止完全性自主武器的发展，确保人类控制参与人工智能武器的关键决策环节；在审查机制上，制定统一的武器法律审查标准、确定审查范围，推进建立强制性国内法律审查机制；在归责路径上，通过拓宽责任链、降低心理要件标准等方式，构建多元化的归责体系。以此促进人工智能武器的良性发展，避免产生更严重的人道主义危机。

关键词：人工智能武器；国际法规制；国际人道法；国际安全；安全治理

引言

随着世界经济水平不断提高，各国越来越重视科学技术的发展，人工智能技术尤为引人注目，逐渐从众多新兴技术中脱颖而出。近年来，人工智能技术逐渐被应用到各个领域，包括医疗、教育、金融、交通等方面，展示了其巨大的潜力和广泛的应用前景。然而，随着人工智能技术的迅猛发展，许多新的挑战也相继而来，尤其是在人工智能武器方面，问题显得尤为突出。人工智能武器，特别是致命性自主武器系统（lethal autonomous weapon systems, LAWS），已经引起了国际社会的广泛关注，并成为多个国家间进程的讨论主题。在人工智能已然成为全球范围内“新风口”的背景下，各国纷纷将融合多学科发展的人工智能技术研究作

¹ 作者简介：郝思源，西南政法大学国际法学院硕士研究生，19823839850，993370865@qq.com。

为国家重点项目，以期在这一领域占据一定的话语权，这种军备竞赛也给国际安全环境带来了潜在威胁。在现有的国际人道法视域下，以人工智能武器为探讨重点，深入分析人工智能武器在使用过程中可能出现的问题，研究其对当下国际人道法带来的各种挑战，并提出合理的规制建议，提出应对策略，这不仅为人工智能武器的研发与使用提供规范建议，还丰富了国际人道法研究的理论基础，有利于推进关于人工智能武器治理的国际安全合作。

人工智能的重要特点在于其机器学习能力，即通过算法基于经验进行自我改进，这赋予了人工智能武器显著的自主性和智能性。这些根本属性使得人工智能武器在实际应用中具有高度的复杂性和不可预测性，进而对现有的国际人道法基本原则以及归责机制带来一系列挑战。现有的国际法框架下，如何确定人工智能武器的使用责任、如何确保其遵守国际人道法的基本原则，以及如何在战争中合理使用这些武器，都成为亟待解决的重要问题。

一、人工智能武器概述

随着各国科学技术的不断发展，人工智能技术（Artificial Intelligence, AI）逐渐被应用到各个领域，人工智能武器化的问题也在世界范围内受到广泛关注。然而，对于人工智能武器的定义等方面国际社会仍未能达成共识，这给人工智能武器的发展和规制带来了阻碍。

（一）人工智能武器的定义

人工智能武器应如何定义，目前世界各国和相关国际组织还不能给出统一的结论，各国专家学者也未能给出统一的意见。若单从字面上分析，人工智能武器可以包含所有搭载了人工智能技术的武器或者武器系统。然而针对人工智能的研究缺乏统一的标准和路径，且由于“人工智能效应”的存在，人工智能技术的标准也在不断变化，或者说是在不断提高，因此也就无法准确定义什么是人工智能技术，²这就导致无法通过这种方式给出一个普遍的定义。从人工智能目前的武器化程度分析，人工智能武器基本由识别系统、执行系统、深度学习系统、知识库和数据库系统等辅助系统结合而成，其工作运行是以提前植入系统的算法程序为基础。依靠这些辅助系统和算法程序，人工智能武器得以准确识别和锁定目标，并实施打击行为。

早于《特定常规武器公约》(CCW)框架下对人工智能武器这一议题的讨论，³美国国防部于2012年11月起就颁布了3000.09号指令，作为世界上第一个在

² Santoro, Erik, & Benoît Monin. *The AI Effect: People rate distinctively human attributes as more essential to being human after learning about artificial intelligence advances*, 107 *Journal of Experimental Social Psychology* 104464 (2023).

³ 在《特定常规武器公约》(CCW)框架下，首次对人工智能武器进行探讨是在2014年。第一次非正式专家会议于2014年5月13日至16日在日内瓦召开，由法国大使让-于格·西蒙-米歇尔(Jean-Hugues Simon-Michel)担任主席。这次会议标志着CCW缔约国首次集中讨论人工智能武器(Autonomous Weapon Systems, AWS)的相关问题。会议为期四天，全球各国学界、政界以及军方相关人士参加了讨论，围绕技术的各个方面、伦

国内法意义上专门规制自主武器系统的法律文件，该文件规制了有关自主武器系统的研制、部署、使用等问题，在该文件中美国国防部给出了自主武器系统的定义：“一旦激活，就可以自主选择目标并与之交战，而无需人类操作员干预的武器系统”。⁴根据美国空军上校约翰·博伊德（John Boyd）于二十世纪 60 年代提出的“OODA 循环”（OODA Loop）理论，一次完整的攻击流程包括观察（Observe）、评估（Orient）、决策（Decide）、行动（Act）四个步骤，循环往复便构成了回路，人工智能的出现使这个循环速度大大加快，在此基础上，依据人在该循环中的参与程度，换言之武器的自主程度，该指令又进一步将自主武器系统分为三种类型：一是“人在回路中”（human in the loop），即在军事行动中，武器系统依靠人的指令选择和打击目标；二是“人在回路上”（human on the loop），这种武器系统在人类的监督下独立选择和打击目标；三是“人在回路外”（human out of the loop），这种武器系统可以在完全没有人类参与的情况下做出军事行动，随着人类科技水平的不断提高，人工智能武器正不断朝着这种方向发展，接近该类型的自主武器系统相继出现，如韩国部署在韩朝边境非军事区的“SGR-A1”型哨兵机器人可以探测投降动作并决定是否进行攻击，⁵IAI（以色列）开发的 Counter IED 和 Mine Suite（CIMS）能够自动执行从检测爆炸装置到将其破坏的整个过程。⁶

不难发现，人工智能武器区别于传统武器的本质属性是“自主性”，国际上大部分涉及人工智能武器的场合也都在使用“自主武器”（Autonomous Weapon, AW）或者“自主武器系统”（Autonomous Weapon Systems, AWS）的概念。目前，人工智能武器仍处于快速发展的阶段，而在事物刚开始发展的阶段就给出一个明确的定义很难实现，也非明智之举，因此通过关注其本质特征进而界定人工智能武器的范围才应当是一种被广泛采取的定义方法。红十字国际委员会（the International Committee of the Red Cross, ICRC）在 2015 年起草的新的《国际人道法和当前武装冲突中的挑战》研究报告中指出，“国际上对于自主武器系统并无一致同意的概念，但各种被提出的定义的共同见解是认为它是一种能够独立选择和攻击目标的武器系统”；同时，该报告认为“自主武器系统”是一个总括术语（umbrella term），包括任何一种在“关键功能”上具备自主性的武器系统。⁷可以说这是一个相当宽泛的定义，但是对于新兴且在不断发展的人工智能武器来说，不论是已经出现的还是在未来可能出现的在“关键功能”中使用了人工智能技术的武器或武器系统都可以被“自主武器系统”的外延包含。本文也将采用这种定义

理与社会问题、国际人道主义法及其他领域问题进行了广泛交流，并将研讨成果形成报告递交联合国。

⁴ Department of Defense Directive 3000.09, Autonomy in Weapon Systems, version II, effective: 25 January 2023, p.13.

⁵ 兰顺正：《观察 | 无人机首次自主杀人，机器杀人的“潘多拉魔盒”打开？》，澎湃新闻网，2021 年 6 月 2 日，https://www.thepaper.cn/newsDetail_forward_12931689。

⁶ ELTA Revealed SAR-based Counter-IED/Mine Suite, Techtime, August 11, 2021, <https://techtime.news/2021/08/11/iai-9/>.

⁷ ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, 32IC /15 /11, October, 2015, p.44.

方式进行进一步的讨论。

（二）人工智能武器的特征

由于具有人脑的部分功能，人工智能武器在选择打击目标时会像人类一样进行思考并作出判断，具有指挥高效化，打击精确化，操作自动化，行为智能化等特点。⁸其中，“自主性”和“智能性”最能体现人工智能武器与传统武器的根本区别。

1. 自主性

“自主性”（autonomy）是人工智能武器的本质特征之一，当前人工智能武器的研发也基本上都聚焦于自主性功能，对于人工智能武器自主性的高度重视也是其在许多场合被称为“自主武器系统”的原因。为了将其与传统武器区别开来，进而确定相关的国际人道法义务，以便在严重的损害结果发生后更好地进行归责，对人工智能武器的“自主性”特征进行清晰的认识和把握是十分必要的。

自主即武器系统在探测、识别、做出决策、打击目标的过程中拥有的自由，自主性程度越大意味着武器系统在决策时更自由，包括选择更多的任务、接受更少的监督等。目前，国内外无人系统自主性评估方法主要包括描述法、坐标轴法、查表法、公式法、图形法等，⁹但无论是哪种等级量化方式，在实际操作中每种因素应如何选择、量化和加权仍存在一定的不确定性，存在相当大的困难。而在国际人道法视域下要区分传统的自动化武器与自主武器，根本应在于定性而非定量。自动化与自主性并没有绝对的界限，自主性源于自动化，但两者又存在本质区别。武器的自主性主要取决于其在战场中的表现，由于自动化武器高度依赖基于规则的逻辑算法，因此其运行过程和打击结果是完全可预测的，而基于学习算法的自主武器系统则会基于战场各种因素的变化会做出不同的判断和决策。

同时，国际社会普遍认为对于人工智能武器“自主性”的理解应围绕“关键自主功能”展开。正如 ICRC 在 2018 年的一份报告中指出，任何在关键功能（critical functions）上具有自主性的武器都应被视作自主武器系统，而这些关键功能包括了本应有人类控制的选择和攻击目标的关键功能。¹⁰

2. 智能性

如果说“自主性”是人工智能武器在功能上的本质特点，那么“智能性”就是其在技术上的最大特征。自主武器系统做出决策的各个环节均离不开人工智能技术的支持。由于人工智能的本质特征是“智能”，那么高度运用了人工智能技术的人工智能武器也可以被认为具有智能性的技术特征。随着人工智能的不断进步，人工智能武器已经逐渐突破了基于固定算法从而做出特定反应的限制，而逐步转向

⁸ 侯嘉斌，李军：《人工智能武器：法律风险与规制路径》，《中国信息安全》2019 年第 12 期，第 90 页。

⁹ 王菡，郭虎生，柏航等：《基于人机协同的无人系统自主性评估方法》，载《无人系统技术》2023 年第 6 期，第 1-11 页。

¹⁰ Neil Davison, *A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law*, UNODA Occasional Papers No.30, November 2017, p5-18.

通过经验学习逐步优化自主性能，进而依据不同的战场情况做出不同的反应。

此外，还有一些学者提出了其他如“致命性”等作为人工智能武器的重要特征，但本文认为，“自主性”和“智能性”两个特征已经能很好地将已经出现甚至未来可能会出现的自主武器或自主武器系统纳入人工智能武器的范围内，已经是两种根本特征，若再将其他特征纳入根本属性的范围内，可能会导致当前大部分武器系统被排除在自主武器系统的范围之外，由此可能会引发各种自主武器的滥用问题，给人工智能武器的规制带来困难。

（三）人工智能武器的发展趋势

人工智能武器具有的优势及其先进程度都是传统武器无法比拟的，就目前形势来看，人工智能武器在借助人工智能技术的迅猛发展而不断地更新升级，其智能化和自主化程度也随之增强。然而由于国防军事领域本身具有的特殊性，各国并不会公开本国对人工智能武器的最新发展研究现状，但是从世界范围内各国已经投入使用的人工智能武器和经费投入来看，尽管人工智能技术不可预测且存在道德风险，各国都不约而同地将人工智能武器的研究与发展作为提高本国军事实力的主要方向。

1. 人工智能武器的发展现状

现如今，人工智能武器已经不再停留在宏观战略与作战概念的层面，而是逐步被应用在实战场景。与以往大多数新兴军事技术不同，人工智能技术并非只被应用于某一特定作战空间或领域，而是可以被应用在各个领域中并提升武器装备性能。一些学者将人工智能武器分为两类，一是“弱人工智能武器”，即仍需人类进行部分操作，自主化程度较低的一类人工智能武器；二是“强人工智能武器”，即已经达到甚至超过人类智慧水平，可以完全脱离人类控制的武器系统。目前来看，已经出现的大部分人工智能武器都属于第一种，强人工智能武器的出现依然较为遥远。

德国早在二十世纪四十年代前后就研发出遥控扫雷车并投入战场使用，是目前已知的最早研发人工智能武器并投入战场使用的国家。就世界范围内来看，地面无人车辆、战斗机器人、无人机、无人潜水器、无人船等武器已经列装部队并应用于战场。俄罗斯研发的“天王星”系列和“平台-M”“阿尔戈”等型号的作战平台，可执行战场巡逻、侦察、排雷、近距离火力支援等任务。¹¹在2015年12月，俄军机器人部队支援叙利亚政府军成功攻占极端组织“伊斯兰国”控制的拉塔基亚754.5高地，这也是全球范围内成建制机器人投入实战的首个战例。¹²中国也在2017年利用人工智能成功实现119架小型无人机飞行试验，名为“海翼1000”的

¹¹ 黄国志：《讲武谈兵 | 另辟蹊径：俄军打造“一枝独秀”无人战车家族》，澎湃新闻网，2017年9月12日，https://www.thepaper.cn/newsDetail_forward_1791113。

¹² 李驰江：《人工智能在军事领域的应用及全球治理》，载《人民论坛·学术前沿》2021年第10期，第24页。

水下无人滑翔机也已投入使用。以色列也研发出一款名为“哈比”（Harby）的巡飞弹，它能够探测、瞄准和攻击敌人的雷达设施而无需人类监督。同时，世界各地不断升级的冲突既是人工智能战争的催化剂，也是试验场，这也进一步体现了人工智能武器的不受监管程度。在俄乌冲突早期，俄军主要动用“海雕”-10、“海雕”-30、“前哨”-R、“猎户座”等无人机及天王星扫雷机器人等人工智能武器执行侦察监视、实施打击、火力校射、反炮兵/车队伴随侦察等任务。在前期，乌军主要使用“旗手”TB-2 执行对俄军燃油车、地面输油装置、弹药补给车等后勤保障节点的打击任务。随着冲突持续，大量自杀式无人机也被投入战场使用。俄军使用的“扎拉”自杀式无人机（KUB-BLA）融合了人工智能视觉识别技术，增强了无人机的实时杀伤力和自主性。乌克兰也在试图将使用人工智能驱动自杀式无人机“神风”引入战场。¹³

2. 人工智能武器的发展前景

正因为在实际应用中表现出了传统武器难以比拟的巨大优势，人工智能武器早已成为各国争相研究发展的对象。俄罗斯总统普京直言：“人工智能技术不仅仅是俄罗斯的未来，也是全世界的未来。谁能成为人工智能领域的领导者，谁就将主宰世界”。¹⁴同时俄军方已经出台一系列发展规划来支持人工智能技术的发展。目前美国军方有 800 多个活跃的人工智能相关项目，仅在 2024 年预算中就要求为人工智能提供价值 18 亿美元的资金。五角大楼计划在 2025 年前投入 10 亿美元用于“复制计划”（Replicator Initiative），该计划旨在开发大量无人机用于寻找潜在威胁。¹⁵美国国防部也已经表明，计划到 2035 年初步建成智能化作战体系，到 2050 年实现作战平台、信息系统、指挥控制的全面智能化，借此实现对手的“技术代差”。英国和韩国也都相继表明将在人工智能武器领域投入大量资金，以推动武器装备研发和应用。我国国务院也在《新一代人工智能发展规划》中提到尽快将人工智能技术应用于我国军事领域中，加强人工智能领域的军民融合，进一步推动国防军事实力的发展。¹⁶

由此可见，凭借科学技术飞速发展的时代背景，世界各国都在尽可能早地开始了对人工智能技术的研究并应用于军事实践，并将人工智能武器的研究与发展作为本国军事发展计划的核心要点，以求本国在该领域占据优势地位。由此可见，人工智能武器具有十分广阔的发展前景和巨大的发展潜力。

（四）人工智能武器的法律地位

要分析人工智能武器给国际人道法带来了何种挑战，以及如何应对这些挑战，

¹³ 兰顺正，邵丰顺：《俄乌冲突两周年|引发军事变革的武器和技术①：AI 狂飙》，载澎湃新闻网 2024 年 2 月 27 日，https://www.thepaper.cn/newsDetail_forward_26470186。

¹⁴ *Whoever Leads in AI Will Rule the World: Putin to Russian Children on Knowledge Day*, RT World News, September 1, 2017, <https://www.rt.com/news/401731-ai-rule-world-putin/>

¹⁵ Nick Robins-Early, *AI's 'Oppenheimer moment': autonomous weapons enter the battlefield*, The Guardian, July 14, 2024, <https://www.theguardian.com/technology/article/2024/jul/14/ais-oppenheimer-moment-autonomous-weapons-enter-the-battlefield>

¹⁶ 中共中央国务院：《国务院关于印发新一代人工智能发展规划的通知》，中国政府网，https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

就要分析人工智能武器在国际人道法中的法律地位。然而，目前各方就人工智能武器在国际人道法中处于何种法律地位并未达成共识，主要存在两种观点：第一种观点认为人工智能武器属于“武器”，或者可以进一步归类为“非常规武器”；第二种观点认为可以将人工智能武器归类为“战斗员”，但两种观点都具有一定的缺陷，不利于对人工智能武器进行合理的规制。

首先，人工智能武器被归类为“武器”或“非常规武器”有失偏颇。根据 ICRC 的规定，“武器”是指“具有旨在伤害、损害、摧毁或使人致残、使物体失去效用后果的所有军械、弹药、物质、工具、机械装置或设备”，¹⁷然而，人工智能武器由传感器组成的识别系统、算法程序构建的控制系统和决策系统以及搭载的各种武器共同构成，其中由算法程序构成的各种软件系统显然难以归为传统意义上的武器。“非常规武器”具有大规模杀伤性，并且人类对其使用结果几乎无法进行预测和控制，常见的非常规武器有生化武器、环境武器等。非常规武器的使用通常会导致不必要的严重伤害和痛苦，甚至可能持续对环境造成严重而广泛的损害。¹⁸而人工智能武器与非常规武器不同，是根据算法和经验学习有针对性地自主选择打击目标，而非“不分皂白的”攻击（undistinguished attack），因此也不能被归类为“非常规武器”。

其次，人工智能武器的自主性和智能性看起来可以使其被归类为具有主观能动性的战斗员，然而这种观点也并不合适。依据国际人道法等法律规范，战斗员享有各种如依法发动攻击的权利和被俘后享有的战俘法律地位等权利和待遇。而目前人工智能武器的攻击能力虽然看起来具有一定的自主性，但都是由人类提前设定并由算法程序决定的，也并非同战斗员一样的完全自主。同时，若人工智能武器被“俘”，其面临的不会是像人类战斗员一样的战俘待遇，只能是被摧毁或拆解，即便想要给予人工智能武器战俘地位也难以真正实现。再进一步讲，不同于人类战斗员可以承担违反国际人道法的个人刑事责任，若人工智能武器“实施”了严重违反国际人道法的行为，也无法向其追责，何况向一段算法代码追责也毫无意义。如果随着科学技术的发展，在未来的某一天人类可以研制出智慧水平达到甚至超过人类的人工智能武器，且开发出某种意义上的“算法监狱”，到那时再讨论人工智能武器的战斗员法律地位也为时不晚。

综上所述，就目前看来，人工智能武器在国际人道法中的法律地位其实并不明晰，给国际人道法带来了一系列挑战。因此，结合人工智能武器具有自主性等特点，本文建议将人工智能武器界定为新型武器系统，这样即避免了在传统武器框架下对其规制产生的问题，符合 ICRC 将其置于“新的作战手段和方法”中讨论

¹⁷ 凯瑟琳·拉万德：《新武器、作战手段和方法法律审查指南》，《红十字国际委员会资料汇编——国际人道法》2015 年，第 487-516 页。

¹⁸ Coupland, Robin M. *The effect of weapons: defining superfluous injury and unnecessary suffering*, *Medicine and Global Survival* 3, 1996.

的做法，¹⁹又便于对其进行分析并提出应对措施。同时，为了更有针对性地分析人工智能武器对国际人道法提出了何种挑战，本文将讨论的人工智能武器种类限定在具有较高自主性、人类参与程度较低的自主武器系统以及具有致命性的自主武器系统的范围内，从而避免产生歧义和疑惑。

二、人工智能武器带来的国际人道法挑战

国际人道法（International Humanitarian Law, IHL）起源于 20 世纪 50～70 年代，是规制战争如何合法进行的法律规则，其宗旨一是为了保护战争的受难者，二是为了限制作战手段和作战方法。²⁰国际人道法在发展过程中形成了日内瓦法体系和海牙法体系。现行的日内瓦法体系主要包括 1949 年制定的“日内瓦四公约”和 1977 年制定的补充和重申上述四公约的两个附加议定书。海牙法体系是以 1907 年的《海牙公约》为代表和开端的关于规范作战手段和方法的条约和习惯。在国际人道法不断完善的过程中，为了广泛传播国际人道法的精髓，以及在战争或武装冲突中准确、迅速地落实国际人道法，一些基本原则逐渐被确立。1979 年，ICRC 专家小组将有关国际人道法的基本原则浓缩在七个要点上。1996 年，国际法院在审查有关核武器的合法性问题时为了将问题说清楚，也列出了一些“构成人道法制度的核心原则”，如“区分原则”“避免不必要痛苦原则”和“比例原则”等，²¹目前这些国际人道法基本原则仍具有极其重要的地位。而在人道法刚出现时的战争形态是传统的机械化战争，而随着人工智能技术逐渐被广泛应用于军事领域，并促进着战争样态从传统的机械化战争转变为高科技战争，这也对这些国际人道法基本原则在实践中如何正确适用造成了冲击和挑战。

（一）人工智能武器对区分原则的冲击

区分原则（principle of distinction）是一项有关敌对行为的国际人道法基本准则。区分原则是指在战争或武装冲突中，交战方应当对军事目标与民用目标，战斗员与非战斗员，武装部队与平民进行区分，并明确只有军事目标和战斗员才是合法的攻击对象。²²作为国际人道法的一个基本原则，区分原则首次在 1864 年订立的《圣·彼得堡宣言》中被提出，该宣言认为在发生战争或武装冲突时，各国作战手段的目标只能是削弱敌方的军事力量。²³随着国际人道法的发展，1977 年《日内瓦公约第一附加议定书》（以下简称《第一附加议定书》）详细规定了区分原则以及对平民和民用物体的保护。²⁴

¹⁹ 红十字国际委员会，红十字会与红新月会国际联合会：《人道的力量——第 32 届红十字与红新月国际大会报告：国际人道法及其在当代武装冲突中面临的挑战》，2015 年 12 月 8-10 日，第 37-56 页。

²⁰ Crawford, Emily, and Alison Pert, *International humanitarian law*, Cambridge University Press, 2024.

²¹ 曾令良主编：《国际公法学》（第二版），高等教育出版社 2018 年第 2 版。

²² Kasher, Asa. *The principle of distinction*. *Journal of Military Ethics* 6.2, 2007, p152-167.

²³ Crawford, Emily. *The enduring legacy of the St Petersburg Declaration: Distinction, military necessity, and the prohibition of causing unnecessary suffering and superfluous injury in IHL*. *Journal of the History of International Law/Revue d'histoire du droit international* 20, no. 4, 2019, p544-566.

²⁴ 《日内瓦公约第一附加议定书》第 48 条明确指出了区分原则的定义，第 51 条和第 52 条分别规定了对

国际人道法规定区分原则的目的就在于促使参战各方区分战斗员与非战斗员、军事目标与民用目标，从而保障平民以及民用目标的安全。²⁵作为新形态战争的重要组成，人工智能武器自然也应受到区分原则的规制。若无法遵守区分原则，人工智能武器就会成为违反国际人道法的非法武器。然而，即便在相对传统的机械战争中，战场环境也是瞬息万变的，受到各种因素的影响，就连人类战斗员也很难精准区别战斗员与平民，也可能无法准确识别军事目标与民用目标。这对目前技术并不完全成熟的人工智能武器来说也绝非易事，因此在其被投入战场使用后可能会出现较为严重的区分困境。

1. 对战斗员与非战斗员的区分困境

《第一附加议定书》规定，除《日内瓦第三公约》第三十三条规定所包括的医务人员和随军牧师外，冲突一方的武装部队人员是战斗员。²⁶对于平民的定义，依据《日内瓦第三公约》第四条和《第一附加议定书》第五十条分析，则采用了排除式定义，即除了直接参加敌对行动的战斗员以外的其他人。因此，平民也应属于非战斗员的范畴，直接参加敌对行动时是平民不受保护的唯一一种例外情况，但当平民一旦放弃直接参加敌对行动就应重新受到国际人道法的保护。²⁷这就导致在实践操作时，战斗员与非战斗员、平民的区分仍然存在困难，因此将二者准确地区别开来对人工智能武器来说更是不小的挑战。尽管各国都在大力开展人工智能武器的研究与发展，但是军事自主系统判断错误，甚至导致意外交火的事故仍屡见不鲜，而造成这种错误的原因主要为目标识别错误。例如，在 1988 年 7 月，搭载了以电脑化的指挥决策与武器管制系统为核心、具有决策辅助功能的宙斯盾系统的美国海军文森斯（Vincennes）号巡洋舰误将一架伊朗 655 号民用客机识别为 F-14 雄猫式战斗机并对其进行攻击，导致机上 290 名乘客和机组人员全部遇难。²⁸

首先，平民身份难以通过外观直接判断。虽然随着技术发展，传感器已经具备了在没有额外输入信息的情况下准确识别物体的能力，并在战场上被广泛应用，人工智能武器也依靠各类传感器进行识别和锁定目标。²⁹然而，要准确地识别目标是战斗员还是平民，仅依靠对外貌、衣着等客观因素识别是远远不够的，还需加入经验判断，而这对于识别系统来说十分困难。况且当前战争形式愈加复杂，

平民和民用物体的保护。

²⁵ Mueller, Matthew T. *The drone question: Legality, ethics, and the need to recognize transnational armed conflict*. 2017, https://commons.lib.jmu.edu/cgi/viewcontent.cgi?params=/context/honors201019/article/1286/&path_info=muellmt.pdf

²⁶ 《日内瓦公约第一附加议定书》第 43 条第二款。

²⁷ 《日内瓦公约第一附加议定书》第 51 条第三款。

²⁸ 韩雨，葛悦涛，赵珈玉等：《军事自主系统误判事故回顾与分析》，载《飞航导弹》2021 年第 3 期，第 25 页。

²⁹ Alan L. Schuller, *At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law*, Harvard National Security Journal 8, no. 2, 2017, p379-425.

各种战术的使用也使战斗员和平民前所未有地混杂在一起,这无疑使身份区分的难度激增。而平民随时可能会参加敌对行动的情况进一步考验了人工智能武器在区分狭义的平民和直接参加敌对行动的平民方面的能力。

其次,人工智能武器很难准确识别“直接参加敌对行动”的行为。ICRC在2009年发布的《国际人道法中直接参加敌对行动定义的解释性指南》(以下简称《解释性指南》)中规定了构成“直接参加敌对行动”的三个要素,³⁰但要在错综复杂的战场环境中具体分析各个要素成立的情况,尤其是涉及因果关系的判断,复杂程度要远大于单纯的理论判断,因此实际判断难度对不具备人类特有的思维能力的人工智能武器来说更是可想而知。³¹例如,人工智能武器可能会瞄准并攻击正在使用暴力保卫自己免遭敌方士兵非法攻击的平民。更进一步,人工智能武器可能也无法准确判断目标的主观意图,被迫直接参加敌对行动的平民或未满合法征兵年龄的儿童也可能因自主武器系统错误的判断结果而遭到攻击。因此,在这种需要通过细微线索才能做出的判断与识别的情况下,由于仅根据设定的决策过程处理信息的底层逻辑并缺乏大量经验和直觉判断,人工智能武器很难做出正确的判断。

2. 对军事目标与民用目标的区分困境

区分军事目标要满足两个要件:一是“实际贡献”,二是“明确的军事利益”,目标只有同时符合两个要件才能被称作军事目标。³²一些目标如军火工厂等本质上即具有军事性质,还有一些目标则是因其位置和用途而具有军事性质,例如被用于军事目的的教堂和学校。³³民用物体亦使用排除法确定,即不是军事目标的物体均为民用物体,且在对目标性质产生怀疑时应一律推定为民用目标。³⁴诚然,以当前的人工智能发展水平,人工智能武器可以准确识别出一个普通的军事目标并进行打击。然而随着科学技术的发展,战争形式已然转变为高科技战争,许多“军民两用目标”如卫星通讯系统、交通基建、工业生产设备等也越来越多地出现在战争中,平民也越来越多地参与到这些系统的维护和操作。³⁵识别这些两用物体是否属于军事目标无疑给人工智能武器带来了更大的挑战。同时军事目标的伪装、军事目标和民用物体之间随时可能发生的转换也都给人工智能武器的识别功能带来了严峻的挑战,因为人类的评估和判断往往是正确区分这些目标的关键要

³⁰ “直接参加敌对行动”相关行为必须同时满足三个要求才能构成直接参加敌对行动:(1) 损害下限很可能是由该行为导致的;(2) 行为同预期损害之间存在直接因果关系;(3) 该行为同武装冲突各方所开展的敌对行动之间存在交战联系。且这三个要素应当被同时满足。参见尼尔斯·梅尔泽:《国际人道法中直接参加敌对行动定义的解释性指南》2009年,第44页。

³¹ Shea, Annemarie. *The Legal and Ethical Challenges Posed by Lethal Autonomous Weapons*. Trinity CL Rev. 24, 2021, p117.

³² 《日内瓦公约第一附加议定书》第52条。

³³ R Baxter, *Law and Responsibility in Warfare: The Vietnam Experience*, North Caroline: United States of America University of North Caroline Press, 1975, p 119.

³⁴ Nwotite, Anita. *The Principle of Distinction in the Light of Civilian Protection in International Armed Conflict*. AJLHR, 4, 2020, p78.

³⁵ E A Oji, *Responsibility for Crimes under International Law*, Lagos: Odade Publishers, 2013, p9.

素，并且需要在掌握真实且足够的情报信息后，才能做出准确的判断。在这种情况下，人工智能武器是否能做出符合区分原则的判断仍旧无法确定。

总之，要真正做到符合区分原则，只有先进的识别、打击能力是远远不够的，更重要的是根据战场情况做出具体判断的能力。然而就目前来看，再先进的人工智能武器本质上都只是一段算法代码，以目前的技术依然难以达到区分原则对于经验推理和战场意识等因素的要求。也就是说，人类的情感和意识很难在机械上得到体现，就连人类在瞬息变幻的战场环境下都有可能做出错误的判断，要使人人工智能武器做出比人类更理性的判断更是十分困难。

（二）人工智能武器对比例原则的冲击

比例原则（principle of proportionality）是国际人道法中另一项有关敌对行为的基本原则。在将区分原则引入国际人道法的同时，《圣·彼得堡宣言》也彰显了比例原则的精神。比例原则指实施军事行为的一方应使军事行动所造成的附带伤害与预期的军事利益成比例，最大限度地减少对平民和民用物体造成的伤害，禁止过分的攻击和引起不必要的痛苦。³⁶可以看出，比例原则对在武装冲突中使用武力进行了限制，但它也承认在武装冲突中的平民伤亡是不可避免的，甚至有军事需要是合理的。³⁷包含“附带损害”和“军事利益”两个核心要素，然后通过对实际情况中两个要素的比较，作出打击行为的附带损害相对于预期达到军事利益是否“过分”的判断。值得注意的是，判断损害是否“过分”并没有一个统一的评判标准。如国际刑事法院（International Criminal Court, ICC）适用的是“理性指挥官标准”，³⁸ICRC 则认为应该同时满足“常识”和“善意”两个条件，然而这两个标准都具有很强的主观性。但是 ICC 和 ICRC 都认为只有具备丰富的作战经验和常识并详细了解现场作战情况的人类指挥官才能做出尽可能正确的选择。

比例原则要求人工智能通过预先设定的程序来感知和衡量众多相关要素和变量，进而判断附带损害是否“过分”，并在可能违反比例原则的情况下有能力取消攻击行动。³⁹从技术层面来说，这甚至比要求人工智能武器遵守区分原则的难度更大。因为“军事利益”本就是抽象概念，需要人类指挥官的政治意识等才能做出判断，如前所述各种判断“过分”的标准和要素都难以完全进行量化，⁴⁰况且战场情况本就瞬息万变，比例原则能在多大程度上被转化为算法程序在技术上仍有待考证。而即便这些要素能完全被量化，平民伤亡和预期军事利益等涉及道德伦

³⁶ 《日内瓦公约第一附加议定书》第 51 条。

³⁷ Fenrick, William J. *The Rule of Proportionality and Protocol in Conventional Warfare*. Mil. L. Rev. 98, 1982, p91.

³⁸ 《国际刑事法院罗马规约》第 8 条第 2 款第 b 项（iv）规定：“故意发动攻击，明知这种攻击将附带造成平民伤亡或破坏民用物体或致使自然环境遭受广泛、长期和严重的破坏，其程度与预期得到的具体、直接的整体军事利益相比显然是过分的。”

³⁹ 张卫华：《人工智能武器对国际人道法的新挑战》，载《政法论坛》2019 年第 4 期。第 144-155 页。

⁴⁰ Fenrick, William J. *Attacking the Enemy Civilian as a Punishable Offense*. Duke J. Comp. & Int'l L. 7, 1996, p545.

理价值的判断也绝不仅仅是只靠数字就能衡量的,仅当人类指挥官具备丰富的经验,并在常识的基础上考虑战场实际情况时,他们才能提出一个更符合比例原则的解决方案,即便如此,这种解决方案可能也不是完美的。⁴¹

因此,比例原则涉及定性问题以及大量的主观价值判断,同时要结合具体的战场情况。对于人类而言,这种需要兼顾战场环境客观因素和社会政治等主观因素来进行高度的自主性评估都是一项极具挑战性的任务,何况是高度依赖算法程序的人工智能武器。

(三) 人工智能武器对人道原则的冲击

人道原则起源于马尔顿条款,是马尔顿条款的重要内容之一。作为国际人道法的“灵魂”,马尔顿条款首次由俄罗斯国际法学家马尔顿(Fyodor Fyodorovich Martens)在1899年的海牙和平会议上提出,并最终被写入1899年《海牙第二公约》与1907年《海牙第四公约》的序言中。⁴²马尔顿条款将“人道原则”与“公众良知”结合在一起,其最主要的价值就是作为一项兜底条款来对抗法律法规的滞后性,使某些没有被国际公约或其他协定规范的领域仍然要受到国际人道法原则的规制,这反映了国际人道法对人道主义精神的崇尚和对道德底线的坚定维护。时至今日,马尔顿条款不但在多个国际人道法条约中被重申,而且在国际法院、纽伦堡国际军事法庭甚至一些国内法院多出的判决中被多次援引,对国际人道法的发展产生了深远的影响。作为马尔顿原则的核心要素之一,人道原则从法律层面防止了“法无禁止即自由”的假设,⁴³甚至已经发展成为新的国际法渊源。

由于科学技术的不断发展,各种新型武器也不断涌现,而人道原则作为所有武器以及作战方法的共同参照点,自然也将人工智能武器纳入了规制范围。随着致命性自主武器系统越来越多地出现,在人道主义的审视下,缺少人类控制和干预的致命性人工智能武器是否被允许决定人类生命存亡也需要审慎考虑。然而,生命权不得被任意剥夺,永远不能被暂停的国际人权法基本条款在任何时代都应当适用。⁴⁴尽管合法地在战场上伤害或终止敌方战士的生命是可以接受的,但决定战斗员是如何受到伤害或失去生命的过程却是极其关键的。⁴⁵因此,暂且不论人工智能武器是否能作出符合区分原则、比例原则的决策,即便某一天技术可以达到使其作出和人类同样的选择,人工智能武器也会由于没有人类的情感和意识

⁴¹ Trumbull IV, Charles P. *Re-Thinking the Principle of Proportionality Outside of Hot Battlefields*. Va. J. Int'l L. 55, 2014, p542.

⁴² 《陆战法规和惯例公约(海牙第四公约)》序言:在颁布更完整的战争法规之前,缔约各国认为有必要声明,凡属他们通过的规章中所没有包括的情况,居民和交战者仍应受国际法原则的保护和管辖,因为这些原则是来源于文明国家间制定的惯例、人道主义法规和公众良知。

⁴³ Michael Salter, *Reinterpreting Competing Interpretations of the Scope and Potential of the Martens Clause*, Journal of Conflict and Security Law, vol. 17, no.3, p.405.

⁴⁴ 联合国《公民权利和政治权利国际公约》第6条。

⁴⁵ Trumbull IV, Charles P. *Autonomous weapons: How existing law can regulate future weapons*. Emory Int'l L. Rev.34, 2019, p533.

而被质疑是否适合作出这样的决定，因为将人的因素从 OODA 回路中排除实际上是对人的生命权的忽视，这违背了人权法的核心精神，将会引发严重的伦理道德问题。

同时，使用具有高度自主性的人工智能武器还可能构成“非人道待遇”。《国际刑事法院罗马规约》（以下简称《罗马规约》）附件《犯罪要件》中规定了“不人道待遇”，即使人遭受身体或精神上的重大痛苦。⁴⁶在 Delalic 案中，前南斯拉夫国际刑事法庭（International Criminal Tribunal for the former Yugoslavia, ICTY）将“非人道待遇”更宽泛地解释为所有构成酷刑、故意对身体和精神造成重大伤害或痛苦，以及严重侮辱人的尊严的行为。⁴⁷在现代战争和地区冲突中，无人机攻击导致大规模死伤的案件层出不穷，而掌握人类生命的却是没有人类情感和道德判断的算法程序，这无疑是对人类尊严的侮辱和精神的摧残，显然是非人道的。因此人工智能武器取代人类而掌握生杀大权对人道主义立场提出了巨大的挑战，只有让人类作出有关人类生命权的决定才能真正实现人道原则的精神内涵。

三、人工智能武器引发的国际法归责问题

归责是军事行动和军事道德的关键。相较于人工智能武器本身对国际人道法原则的挑战，人们似乎更关心由人工智能武器引发的归责问题。无论从事的战争是正义的还是非正义的，国家或个人都要为在战争中实施的非正义行为负责。⁴⁸如前文所述，人工智能武器的应用造成了“归责空白”，但是问责仍是亟待解决的问题。同时，问责对于平民的保护、维护受害者获得赔偿的权利和防止战争爆发等都至关重要。⁴⁹

（一）国际不法行为责任及归责问题

国际不法行为责任指一个国家对其所做的国际不法行为应承担的法律责任。国际不法行为责任是由于国家未履行其国际义务而产生，其核心目的是纠正这些不法行为，最大限度地减少不法行为的一切后果，敦促各国遵守其应承担的国际法义务，以确保国际社会的正常秩序。⁵⁰

在联合国国际法委员会通过的《国家对国际不法行为的责任条款草案》（以下简称《国家责任条款草案》）中，对国际不法行为的各个构成要素进行了详尽的阐述。根据《国家责任条款草案》第二条的表述，国际不法行为责任的构成包括两个方面：可归于该国的行为和对该国国际义务的违背。与个人刑事责任不同，国际不法行为责任的构成不包括心理要素，两个要素均未要求故意或者过失的主

⁴⁶ 《犯罪要件》第 8 条第 2 款 a 项(ii)-2。

⁴⁷ ICTY. Prosecutor v. Zejnir Delalic. Case No.: IT-96-21-T, Judgement, 16 November 1998, para. 1329.

⁴⁸ Buchanan, Allen, and Robert O. Keohane. *Toward a drone accountability regime*. Ethics & International Affairs 29, no. 1, 2015, p15-37.

⁴⁹ Afonso Seixas-Nunes, *Autonomous Weapons Systems and the Procedural Accountability Gap*, 46 Brook. Journal of International Law, 12-1-2021, p.421.

⁵⁰ FA. Mann, *The Consequences of an International Wrong in International and National Law*, British Yearbook of International Law, 48, 1976-77, pp. 1-65.

观因素，因此，要判断国家是否做出了国际不法行为较为简单。根据《国家责任条款草案》的规定，只要该国被认定存在国际不法行为，就要通过限制主权、进行损害赔偿等方式承担国家责任，通过这种方式受害者也能获得一定的赔偿。

虽然理论上国际不法行为责任的划分非常明确，但人工智能武器的使用还是给归责带来了不小的挑战。人工智能的技术问题可能成为国家逃避承担责任的主要借口之一。由于人工智能技术的发展尚不成熟，始终处于不断研发进步的阶段，且由于缺少相应国际条约的规制，对技术的审查机制也不够完善，因此一些尚未成熟的技术可能会被过早地投入战场，人工智能武器的所属国和使用国完全有可能以技术障碍或技术失去控制为由推卸法律责任。同时，人工智能武器的生产和贸易也给国际不法行为责任的划分带来挑战。虽然《武器贸易条约》（The Arms Trade Treaty, ATT）规定了六种禁止和限制武器出口的情况，但是由于武器审查的标准不统一、进口国隐瞒进口武器的真实目的等因素的影响，一旦造成严重的损害结果，很难将损害结果认定为可归于出口国的行为，进而对出口国国际不法行为的归责造成阻碍。且人工智能技术与自主武器系统存在可分离性，如相关人工智能技术可独立于自主武器的硬件独立存在，而目前没有国际公约对人工智能技术的出口作出限制，这也给《武器贸易条约》的规制作用带来挑战，而一旦进口国使用进口的人工智能技术对本国武器进行升级后作出违反国际人道法的行为，出口国是否应当承担国际不法行为责任也难以确定。⁵¹另外，鉴于国际政治环境的高度敏感性和复杂性，各国通常会出于对自身国家形象和国家利益的考虑，对本国违反国际人道法的行为予以否认，甚至通过依靠其国内法的方式规避国际不法行为责任。⁵²

因此，虽然在当前国际法体系下，各国对因其使用人工智能武器而引发的损害结果承担责任已经成为国际社会的共识，但是囿于人工智能技术仍处于发展阶段和现行国际法的相应规制存在缺位，这种国家问责制在实际操作中仍然难以实现。

（二）个人刑事责任及归责问题

如前文所述，即便具有很高的自主性，但是最低程度需要人类干预的人工智能武器也需要人类激活并投入战场使用，这个过程离不开人类决策。同时人工智能武器又不具有战斗员的法律地位，当其造成严重损害结果而违反国际人道法时无法追究其刑事责任，追究其责任也并无意义，因此除了对国家的国际不法行为进行归责外，还涉及个人刑事责任的归责问题。

不同于国际不法行为责任的认定，除了客观要件外，主观要件是认定被告人是否构成战争罪的必要构成要件。除另有规定外，ICC 管辖在“明知”和“故意”两

⁵¹ 冷新宇：《论自主武器军控的国际法及政策要点》，载《国防科技》2021年第4期，第48页。

⁵² Y. Matsui, *The Transformation of the Law of State Responsibility*, Thesaurus Acroasium, 20, 1993, pp. 5-65.

种主观心态下实施了犯罪物质要件的犯罪。⁵³所谓“明知”是指行为人意识到某种结果的发生或某种情况的存在；“故意”则是要求行为人有意实施行为并有意造成某种结果，或者意识到随着事态发展会导致该结果的产生。⁵⁴因此，《罗马规约》对于行为人犯战争罪的必要心理要素门槛很高。同时 ICC 在客观要素上也作出了一定的要求，《罗马规约》第六条至第八条及附件《犯罪要件》中规定了灭绝种族罪、危害人类罪以及战争罪的客观构成应包括危害行为、危害结果、因果关系及时空要素。而人工智能武器的出现使主观要件的认定、因果关系链条的连续性和时空要素的关联性都受到了巨大的挑战。

1. 设计者、生产商和经销商的责任认定

设计者、生产商和经销商在个人刑事责任的认定中处于一种较为“尴尬”的地位。一方面他们有可能成为人工智能武器实际控制人转嫁责任的对象，因为对于人工智能武器的所有者和实际控制人来说，以产品存在技术缺陷和故障为由，将责任转嫁给设计者、生产商和经销商是十分容易的。但设计者在设计程序时可能仅仅负责整个武器系统程序的一小部分，他们可能根本不知道自己设计的程序要被用于人工智能武器的研发。且由于贸易全球化的不断深入推进，一种商品的生产可能涉及多个国家的不同生产商，每一个生产商可能仅负责生产一种武器的一部分零件。对于经销商来说，他们完全有可能不知道自己所销售的武器系统最终会由谁实际控制并使用，况且即便已经充分向购买者充分披露了其所销售武器系统的功能、使用范围和使用方法，购买者也可能既不会向经销商透露其真实使用目的，也不按照使用说明使用人工智能武器。而在这种情况下，要使设计者、生产商和经销商承担责任是极不合理的。

另一方面，设计者、生产商和经销商又确实可能会利用设计、制造以及销售人工智能武器的手段造成特定的损害结果，而根据现行国际刑法的规定又难以追究其个人刑事责任。由于《犯罪要件》中规定了战争罪的客观构成中包含了时空要素，即行为需发生在战场环境中。然而，设计、生产和销售的行为都几乎不可能直接发生在战场中，这就导致即使设计者在设计程序时故意输入错误的代码或加入了包含“算法歧视”的程序代码，使武器系统在激活后可以进行违反国际人道法规定的攻击并造成严重的损害结果，也无法以战争罪追究其刑事责任，这就在一定程度上引发了“问责空白”，但实际上，即便用最朴素的价值观判断，设计者也应对这种行为承担责任。⁵⁵同时在设计者和生产商已经知晓其正在参与研发人工智能武器并作出符合预期功能生产设计行为的情况下，人工智能武器的高度自主性和复杂的战场情况仍会给这种主观心态的认定带来相当大的挑战。由于人工

⁵³ 《国际刑事法院罗马规约》第 30 条。

⁵⁴ 李世光，刘大群，凌岩主编：《国际刑事法院罗马规约评释（上）》，北京大学出版社 2006 年版，第 208 页。

⁵⁵ Singer, Peter W. *Wired for war: The robotics revolution and 21st century conflict*. Language 2020, 2012, p4-16.

智能武器可以依靠积累经验而不断开发学习，加之战场环境又会不断变化，人工智能武器在真正运行后会造成怎样的潜在后果是较难确定的，在这种情况下，设计者和生产商的主观心态最多属于“知道或应当知道”，其对潜在的危害结果仅有放任发生的心态，此时是否应追究其个人刑事责任又会产生争议。

况且，追究设计者、生产商及经销商的责任在程序上也较难操作。人工智能武器从研发生产到正式投入战场使用的过程中必定会涉及大量人员，而追责每个相关人员的责任势必会导致责任分散，从而失去实际效果。正如一句英谚所说：“如果所有人都承担责任，那么就没有人承担责任”。并且如此大范围的追究责任还可能使司法成本迅速增加，司法效率显著降低，是否符合比例原则仍值得考量。

2. 指挥官责任认定

相较于设计者、生产商和经销商的责任认定，指挥官的责任认定相对容易，不需要那么长的逻辑链条。《罗马规约》第二十八条详细规定了指挥官应对其有效指挥和控制下的部队的行为而负责的情况：一是知道或应当知道部队正在或者即将实施犯罪行为；二是没有在其权力范围内为了制止犯罪的实施而采取必要合理的措施或报请主管当局审查或起诉。从条文中可以分析得出，对指挥官或上级人员的责任认定建立在下级人员仍为人类的基础上。若人工智能武器在关键决策功能上仍有赖于战斗员的操作，则此时不存在人类下属缺位的情况。但是随着人工智能武器不断突破现有技术，其高度的自主性正在使其逐渐脱离人类控制和干预，作为指挥官的地位在某种程度上就会逐渐趋于与操作员类似，换言之指挥官的下级人员将逐渐从人类转变为机器，造成人类下属的缺位，此时若由指挥官部署控制的人工智能武器作出违反国际人道法的行为，该种情况下便无法按照《罗马规约》第二十八条追究其指挥官责任，因为该条要求指挥官对其下属能够有效管辖和控制，而在人工智能高度自主的情形下，指挥官的有效控制难以满足。此时主观心态又再次成为认定其是否构成犯罪的重要因素，只有满足“明知”或“故意”的心理要件才能构成上述罪名，而由于战场环境瞬息万变以及人工智能武器依据学习经验算法作出决策的不可预测性，这可能导致指挥官难以预测人工智能武器会作出何种打击行为，这种情况下指挥官出于过失心态而导致的严重损害结果仍不足以被认定为构成上述罪名，“责任真空”便再次产生。

综上所述，人工智能武器的应用确实在归责问题上给国际人道法的实施带来了新的挑战。目前基本上所有重要的国际人道法条约在制定时都无法考虑到人工智能武器不同于以往任何武器的特点，由此产生可归责性问题迫切需要解决，但是由于不同种类人工智能武器的不断出现，给出一个统一的解决方法似乎难以实现，因此只能寄希望于国际社会能构建出多元立体的规制体系，从而更好地填补可归责性的间隙和空白。

四、人工智能武器的国际人道法规制路径

人工智能武器的出现加速了现代战争形态的改变,其具有的传统武器无法比拟的优势无疑会使其成为未来战争中各国争相发展研究的对象。但同时,人工智能武器在决定未来战争走向的同时也给当前国际人道法体系带来了严峻的挑战,人工智能技术的军事化应用也在迅速发展,因此在国际法体系内对这类高科技武器的研发和使用采取全面的手段进行管控刻不容缓。⁵⁶当前,对人工智能武器带来的国际人道法挑战的担忧更多集中在对人道原则的挑战上,人们对于机器是否应该掌握对人类生命的生杀大权争论不休,并且这种担忧随着完全自主的致命武器系统出现可能性的增加而不断加剧,归责问题也因此争议不断。所以,要应对上述挑战,就要规划出合理的规制路径,可以从制定国际公约、完善法律审查机制以及构建并完善多元的归责体系等角度出发,从而对人工智能武器的研发和使用进行有效的规范和限制。

(一) 制定专门针对人工智能武器的国际公约

1. 增列新的 CCW 附加议定书

一直以来,国际社会关于武器问题的讨论主要集中在 CCW 框架下,缔约国会议是进行讨论的最主要平台。⁵⁷CCW 中强调了两个国际人道法基本原则:一是禁止使用不区分目标的武器进行不分皂白的攻击;二是禁止在使用可能造成这种伤害和痛苦的武器的情况下,造成不符合比例原则的超出限度的伤害和不必要的痛苦。⁵⁸CCW 本身并未对武器的使用作出具体的限制性规定,而是通过增列议定书的方式,在议定书中规定了关于不可检测碎片的武器、地雷、陷阱或其他类似设备、燃烧武器、激光致盲武器等的禁止或限制,由此构成了一种极具开放性的限制武器使用的国际条约体系。

由于人工智能武器日益引起国际社会的广泛关注,CCW 从 2014 年起就已经陆续开展了 4 次有关致命性自主武器系统 (lethal autonomous weapon systems, LAWS) 的非正式专家会议,并在 2016 年设立了一个关于讨论 LAWS 在研发和使用中有关问题的不限成员名额政府专家组 (Group of Governmental Experts, GGE),并在 2017 年 11 月举行了第一次会议,计划每年举行一次。而像 GGE 会议或缔约国审查会议等都在提供一个可以讨论和谈判与 LAWS 的规制有关问题的平台方面发挥了重要作用。⁵⁹目前, GGE 会议已经将制定一个有关 LAWS 的

⁵⁶ Sehrawat, Vivek. *Autonomous weapon system: Law of armed conflict (LOAC) and other legal challenges*. Computer Law & Security Review 33, no. 1, 2017, p38-56.

⁵⁷ CCW/GGE Report of the 2018 Group of Governmental Experts on Lethal Autonomous Weapons Systems, .2/2018/3, 31 August 2018, [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/20092911F6495FA7C125830E003F9A5B/\\$file/2018_GGE+LAWS_Final+Report.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/20092911F6495FA7C125830E003F9A5B/$file/2018_GGE+LAWS_Final+Report.pdf), pp.5, 8.

⁵⁸ ICRC Advisory Service on International Humanitarian Law, 1980 Convention on Certain Conventional Weapons, Fact Sheet, International Committee of the Red Cross, 30 June 2018.

⁵⁹ Michael Möller, *Opening Remarks: Convention on Certain Conventional Weapons (CCW)*, Speech, Meeting of Experts on Lethal Autonomous Weapon Systems, 13 May 2014.

新议定书作为下一步讨论的重点。同时，ICRC 习惯国际人道法研究会已经将 CCW 的一系列议定书列为习惯国际法，因此所有来自 CCW 的规定对所有国家都具有约束力，尽管有些国家并不是 CCW 的缔约国，这也使新增列议定书得到广泛遵守的可能性大大增加，因此这一路径具有高度可行性。⁶⁰

2. 制定专门的国际条约

除了增列 CCW 附加议定书，制定专门的国际公约也是一种理想的选择。目前在世界范围内，还没有专门用于规范人工智能武器的国际条约，也没有依公约成立的政府间国际组织。由于人工智能武器具有不同于以往任何武器的特殊性，且人工智能技术仍处于发展上升阶段，制定专门的国际公约有助于对人工智能武器的发展作出指导，在各种标准和机制上制定统一的规范，同时对人工智能武器的研发和使用作出禁止或限制。在制定专门针对人工智能武器的国际公约时，应当将程序性和指导性原则包含进来，随着人工智能武器不断发展，可以通过增列附加议定书的形式，进而构建起专门规制人工智能武器的公约框架并不断丰富完善，使整个体系在人工智能武器的发展过程中不断完善并发挥作用，同时应当设立组织大会、秘书处、审查执行机构等保障公约的有效执行。这种方式相比于在 CCW 框架下增列议定书的方式更有利于国际社会对人工智能武器的规制，更具有针对性和有效性。若公约框架能够建立起来，就可以进一步凝聚各方共识，促进国家间技术交流与合作，达成对自主武器系统的共识，这无疑对保障自主武器系统的良性发展和防止滥用具有重要意义，同时在全球范围内更好地应对人工智能武器对国际人道法等各个方面带来的挑战。⁶¹

（二）确立“有意义的人类控制”原则

“有意义的人类控制”（Meaningful Human Control, MHC）原则意在强调任何攻击都应该在人类的控制下发生。尽管对于这一原则的准确定义仍有待明确，但该原则已经被国际社会普遍认为是应对人工智能武器挑战的潜在手段，当前国际社会已经提出的如“人机互动、武器法律审查”等手段其实都是由“有意义的人类控制”衍生出的具体措施。通过对国际社会对于“人类控制”的讨论可以看出，在定义和适用这一原则时需要考虑四个重要因素：国际人道法的原则和规则；LAWS 的不同种类；不同类型的 LAWS 所拥有的不同程度的自主性；LAWS 整个生命周期中的不同阶段。⁶²

确立这一原则其实是在另一层面上否定了具有完全自主性人工智能武器的发展，在实践中也能确保在关键决策上保留人类的控制，这不但有助于自主武器

⁶⁰ 徐能武，龙坤：《联合国 CCW 框架下致命性自主武器系统军控辩争的焦点与趋势》，载《国际安全研究》，2019 年第 5 期，第 108-132 页。

⁶¹ 王蕾，《历史比较视野下自主武器军控规范争论及其走向》，载《国际安全研究》2024 年第 5 期，第 109-134 页。

⁶² Yasin, Tarisa. *Effective Human Control over Lethal Autonomous Weapon Systems and Compliance with International Humanitarian Law*, Doctoral dissertation, Bond University, 2023, p208.

系统的良性发展，为其遵守国际人道法原则和规则提供保障，还能在一定程度上解决归责问题。

1. 禁止完全性自主武器的发展

以目前的人工智能技术发展水平和速度来看，在未来出现完全性自主武器的可能性是非常大的。如前文所述，在当前人工智能水平发展没有那么高的情况下，人工智能武器都已经对国际人道法基本原则提出了挑战，甚至其作为非生命体即拥有自主决定剥夺人类生命的行为也对“马尔顿条款”的核心价值造成了猛烈冲击，更不用说完全性自主武器出现后将会引发更严重的伦理问题。而一旦不加禁止，研究发展完全性自主武器就会成为各国军备竞争的重中之重，甚至可能会使人工智能失去控制，威胁人类的整体生存。虽然当前国际社会没有用于禁止完全性自主武器的国际公约，但可以依照《不扩散核武器条约》《禁止生物武器公约》《禁止化学武器公约》的形式对其作出禁止和限制性规定，这些公约限制的都是国际社会公认应避免使用的武器，并且具有严格的法律约束力和道德禁忌规范。

63

2. 关键性决策过程中保留人类控制

如前文所述，人工智能武器对国际人道法基本原则的挑战也来源于人工智能技术水平的限制，而人类的经验、道德与伦理判断可以对技术上的缺陷进行弥补。在涉及人工智能武器的关键性决策阶段中加入“有意义的人类控制”，便可以在最大程度上保证军事利益和人道主义的相对平衡，尽可能避免违反国际人道法。

关键性决策中的人类控制主要体现在“关键功能”和“关键环节”两方面。在“关键功能”上，必须保留人类对于人工智能武器核心功能的有效控制。在人工智能武器选择任务类型、攻击目标以及部署环境时，必须由人类决定最终行动是否实施，即使要发展人工智能武器的高度自主性，也至少要保证最低水平的有效人类控制。在“关键环节”上，应当在人工智能武器生命周期的各个阶段，以合理的形式对武器系统进行不同程度的控制，包括设计、研发、测试、部署、使用和使用后的各个阶段，以达到对其进行全面合理规制的目的。

（三）完善对人工智能武器的法律审查机制

“武器合法性审查”并非最新提出的专门针对新型高科技武器的机制。在《圣·彼得堡宣言》中就已经被首次提出，奠定了新武器法律审查机制的基础。《第一附加议定书》第三十六条表明了各缔约国应履行“法律审查”的义务，但并没有进行更为具体的解释，虽然在 2006 年 ICRC 通过了《新武器、作战手段和方法法律审查指南》，但该份指南仅仅是一份建议性文件，不具有法律约束力，因此法律审查的机制并未在国家层面很好得到实施和落实。然而在目前国际公约

⁶³ 李享，罗天宇：《人工智能军事应用及其国际法问题》，载《信息安全与通信保密》2021 年第 1 期，第 99 页。

由于密切牵动各国战略意义而难以签订的情况下,建立具有包含建议性指导原则的法律审查机制并逐渐要求各国推进落实可能是实现人工智能武器全球安全治理的关键一环。⁶⁴

1. 制定统一的审查标准

一套统一的标准对于法律审查机制来说至关重要。由于人工智能技术可以搭载其他种类的武器或与其他武器进行结合,那么首先不能违反现有的武器种类禁止性规定,如《禁止化学武器公约》《禁止生物武器公约》等,若搭载了这些被禁止的武器,那么该武器本身就是非法的。如果不涉及上述情况,审查标准应聚焦于人工智能武器系统是否遵守区分原则、比例原则等国际人道法基本原则,以及是否违反了具有兜底作用的“马尔顿条款”。

2. 确定审查的范围

对于武器本身来说,审查应涵盖该种武器的所有组成部分,从识别系统到算法编程,再到开发预计用途和效果。在时间范围上,对于人工智能武器的法律审查应该覆盖其全部生命周期,在最大程度上确保每一类人工智能武器在部署使用前都能得到全面的法律评估,从根源杜绝非法的人工智能武器出现。另外考虑到人工智能武器具有的学习能力,对其审查也应延续到被投入战场使用后,在其使用周期内也定期进行法律审查,确保算法没有自我开发出以违反国际人道法规定的方式做出决策的能力。

3. 建立强制性国内法律审查机制

早在 2013 年,海恩斯就呼吁“在机器人武器系统发展的所有方面实现完全透明”。⁶⁵但是由于军事领域的特殊性,建议性的武器审查机制很难得到各国的履行,但各国对于人工智能的不断探索又会成为人工智能武器违反国际人道法的巨大风险之一。因此,推进强制性国内法律审查机制的建立势在必行。从国内层面来看各国应当依照国际统一标准探索建立国内法律审查机制,美国国防部 3000.09 以及 5000.01 号指令均可供参考。⁶⁶同时,敦促各国在保障国家安全及商业秘密的情况下在合理范围内履行披露义务,共享相关经验以及最符合国际人道法原则和规定的方法,促进国际交流与合作,维护国际和平与稳定。在国际层面,可以通过多边条约确立强制性国内法律审查机制,若有必要应成立专门性的人工智能武器的审查监督机构,对怠于履行国际义务的国家进行制裁和追责。

(四) 构建多元的归责体系

人工智能武器的出现引发了“问责空白”,若不对现行的归责制度进行完善和改进,国际法的作用将会受到损害,随着技术发展,人道主义也会受到巨大的挑战。因此,应当构建出一个多元的责任体系,适应现代高科技战争的快速变化,以填补人工智能武器引发的责任空隙,维护国际人道法的权威和尊严。

1. 完善人工智能武器的国家责任体系

⁶⁴ 刘杨钺:《全球安全治理视域下的自主武器军备控制》,载《国际安全研究》2018 年第 2 期,第 49-71 页。

⁶⁵ Ohlin, Jens David, *The Combatant's Stance: Autonomous Weapons on the Battlefield*, International Law Studies, 2016.

⁶⁶ 孟誉双:《美国规制自主武器系统的法律政策及其启示》,载《战术导弹技术》2021 年第 5 期,第 43 页。

如前所述,现行的国家责任体系在理论上仍然可以对人工智能武器引发的归责问题适用,只是具体的困难体现在实践操作时难以落实。而由于追究个人责任变得困难,国家责任便可以更好地发挥作用,虽然国家可能不会直接承认其应承担国家责任,但国家由于具有财政支持能够及时进行适当的人道主义补偿,且一旦国家承认了其应承担的国家责任,受害方就很容易获得赔偿,这对受害方来说无疑具有积极的作用。况且国家责任的形式又不仅仅包括赔偿,还包括保证不再犯等多种责任形式,这表明国家责任可以适用于人工智能武器生命周期的各个阶段,从设计研发到使用后评估,而不仅仅局限在损害结果发生后,但是这也依赖于专门国际条约的制定和统一法律审查标准的出台。同时,虽然《武器贸易条约》已经对人工智能武器的国际贸易作出了规制,要求出口国在明知进口国将被用于实施违反国际人道法行为的情况下承担国家责任,但人工智能技术的国际贸易却并没有进行规制。由于人工智能技术可将非自主性武器升级为具有自主性的人工智能武器,也可能导致违反国际人道法的损害结果出现,因此明知人工智能技术将被用以实施非人道目的行为的出口国也应承担相应的国家责任。若在出口国不知道其出口的人工智能武器或技术将被用于违反国际人道法的目的,而使用国无力对其造成的严重损害结果承担责任时,出口国应当对使用国造成的损害结果承担连带责任,这样既可以保障受害方的利益,又可以敦促出口国在出口人工智能武器和技术时积极履行审查义务,确定进口国的使用用途。其次,在一些联合行动中,多国联盟中的一个国可能会多次部署并使用非法的人工智能武器,而其他盟国在意识到这种行为的重复和错误性质后仍放任该行为的发生,在这种情况下除了人工智能武器部署国应承担相应责任外,其他盟国可能也需要承担国家责任。另外,技术问题也不应成为国家承担责任的法定免责事由,这也有助于国家在开发、审查和使用人工智能武器时保持更审慎的态度。

2. 完善人工智能武器的个人刑事责任体系

从理论上来看,个人刑事责任体系相比国家责任体系来说面临着更大的挑战。首先在客观方面,由于现行体系下入罪需要满足的时空要素门槛较高,而人工智能武器提前编写置入的算法和可远程操作的功能都大大削弱了因果关系链条,导致在客观方面无法追究个人刑事责任。因此,应当针对人工智能武器的设计适当降低时空因素的认定,若人工智能武器的开发者明知算法有瑕疵、带有歧视性因素或违反国际人道法的原则和规范,仍在人工智能武器的运行程序中编写该算法,从而导致违反国际人道法的严重损害后果发生,在该种情况下开发者就应承担个人刑事责任。同样地,若人工智能武器生产者和经销商明知其所生产销售的武器具有算法瑕疵或将被以违反国际人道法的目的投入使用,生产者和经销商也应当承担相应的责任。

其次,在主观方面,针对人工智能武器有关的犯罪在主观心态的入罪标准认定上也应该适当降低。指挥官和操作员主观心态为明知或故意时应承担个人刑事

责任不必赘述。但当其主观心态若为无法或难以预见人工智能武器依据其学习经验作出超出必要限度打击行为的过失时，则会出现无人担责的困境。因此将过失心态作为构成该类责任的要件是有必要的，但是可以适当减轻责任，以便与故意心态的犯罪区别开来。

结论

目前，人工智能技术的飞速发展正引领着军事发生巨大变革，由于相较于传统武器的巨大优势和经济政治等影响因素，人工智能武器正越来越多地被投入战场使用。随之而来的就是人工智能武器对国际人道法基本原则的挑战，也带来了新的国际法归责困境，这都对现行的国际法体系造成了强烈的冲击，使国际人道法的目的和宗旨难以实现。因此，在现有基础上对人工智能武器进行进一步规制刻不容缓。这不仅关乎国际人道法的尊严与权威，更与国际安全合作的稳定与发展紧密相连。

要想真正实现国际人道法的宗旨和目的，就要禁止完全性致命自主武器系统的发展，并根据人工智能武器的自主程度、执行任务种类等因素的不同作出区别化规制。在规制路径的表现形式上，可以在现有的 CCW 框架下增列第六附加议定书，用于针对人工智能武器作出更细致的禁止和限制性规定。从长远利益来看，可以制定独立的国际公约，成立专门的国际组织，并在该公约框架下不断细化，通过增列附加议定书的方式不断完善程序和实质内容。在规制内容上，要确立“有意义的人类控制”原则，保障人类在人工智能武器的各个关键性决策和整个生命周期中都保留控制权。同时，要统一人工智能武器的法律审查标准，明确法律审查的范围，推进建立强制性国内法律审查机制，以确保人工智能武器可以被合法研发与使用。此外，还应构建多元的归责体系，通过完善开发者、生产商、经销商和指挥官的个人刑事责任认定，同时完善国家在人工智能武器国际贸易中的归责机制，强化国家责任在归责体系中的兜底作用，避免出现归责空白。

总之，在复杂的全球安全形势下，国际安全合作在应对人工智能武器挑战中至关重要，这是应对人工智能武器挑战、维护全球安全稳定的必然选择，各国应积极参与，共同构建安全的国际环境。

Abstract

Artificial intelligence technology has developed rapidly with the continuous exploration of human beings in the scientific field, and has gradually been widely used in the military field, and artificially intelligent weapons have emerged. In armed conflicts, AI weapons can choose targets and strike them independently with low human participation, and even develop to the point where they can make key decisions without human participation at all in the future. This independent feature, which is different from traditional weapons, has brought a series of challenges to international humanitarian law. First of all, the legal status of AI weapons is vague, and it is inconclusive whether highly autonomous AI weapons belong to combatants or new weapons; Secondly, due to the absence of human control factors in key decision-making and the existing technical limitations, it is often difficult for AI weapons to cope with emergencies, to evaluate the target and intensity of attacks, and to easily impact the principle of distinction and the principle of proportionality; In addition, the autonomy and lethality of AI weapons have also caused the imputation dilemma that it is difficult to attribute the responsibility of international wrongful acts to the state and the subjective elements of individual criminal responsibility are difficult to identify. In order to cope with the challenges brought by artificial intelligence weapons to international humanitarian law and fill the gaps in the existing liability system, we should formally fill the gaps in existing legal norms by adding additional protocols to CCW and formulating special international treaties, and then establish a sound international humanitarian law system for AI weapons. In content, it is necessary to establish the principle of "meaningful human control", prohibit the development of completely autonomous weapons, and ensure that human control participates in the key decision-making links of AI weapons; On the review mechanism, we should formulate unified standards and determine the scope of review of weapons laws and promote the establishment of a mandatory domestic legal review mechanism; On the path of imputation, a diversified imputation system is established by broadening the responsibility chain and lowering the standard of psychological elements. In order to promote the benign development of AI weapons and avoid a more serious humanitarian crisis.

Key words: Artificially Intelligent Weapon; Regulation of International Law; International Humanitarian Law; International Security; Security Governance

论《联合国打击网络犯罪公约》的刑事定罪

刘龙*

摘要：《联合国打击网络犯罪公约》是第一个联合国框架下专门打击网络犯罪的国际公约，它突破了立足于西方立场的《布达佩斯网络犯罪公约》体系，有着参与度广、强调国家主权和扩张网络化传统犯罪罪名等优势。作为其核心部分的刑事定罪条款，详细规定了网络犯罪的行为、法人责任、犯罪形态、追诉时效以及起诉、审判和制裁相关问题，体现出了明显的规范化、灵活性与尊重国家主权等特征；同时，也对强调个人自由的欧美立场作出妥协，对非网络依赖型犯罪罪名进行了限缩。对比公约，我国《刑法》涵盖了“非法访问”、“干扰数据”等网络依赖型犯罪（典型计算机犯罪）和“利用通信技术欺诈、造假”等网络使能型犯罪（网络化的传统犯罪），但在持有儿童性虐待或性剥削材料犯罪、洗钱罪的上游犯罪等方面存在缺憾，可以通过立法或司法解释来增设罪名、修改犯罪要件等做法来完善。此外，通过加强国际合作、借鉴优秀经验，积极完善我国刑法中的电子数据取证、跨境司法合作等内容，不但有助于公约在我国有效适用，亦能使我国刑法有效应对更加新型、复杂的网络犯罪。

关键词：《联合国打击网络犯罪公约》 《布达佩斯网络犯罪公约》 网络犯罪
刑事定罪

在当今信息和通信技术飞速发展的背景下，网络犯罪日益猖獗。从最初的简单黑客攻击，到如今的复杂跨国网络诈骗、网络赌博、数据泄露、网络恐怖主义等，网络犯罪的形式和手段不断翻新，对国家安全、社会稳定和个人权益构成了严重威胁。比如我国遭受的跨境电信网络诈骗呈现高发态势，涉案人员与涉案财产均数量巨大，严重威胁人民群众的财产安全与社会秩序¹；又如某些犯罪组织通过勒索软件或恶意软件攻击政府、企业服务器，以获取非法利益或瘫痪行政运行。²这类以网络为手段或针对计算机的犯罪不仅具有高度的隐蔽性，也具有天然的跨国性甚至全球化特征，单个国家难以独立进行有效防治。犯罪分子往往利用不同国家的法律差异和司法管辖权漏洞，逃避法律制裁。因此，加强国际合作、共同打击网络犯罪是各国应对这种新型全球化犯罪的必要措施，也已经成为国际

* 刘龙，南昌大学法学院，国际法学硕士研究生。

¹ 参见李学军、贺娇：《推定在电信网络诈骗及其关联犯罪案件证明中的适用与规制完善》，载《法学论坛》2024年第2期，第119-130页。

² 参见杨帆、林炳杰：《美国反制勒索软件新举措的合法性评析》，载《美国研究》2023年第4期（第37卷），第73-94+6-7页。

社会的广泛共识。

尽管国际社会很早就意识到网络犯罪的严重性,并制定了一些全球性或区域性的公约、协定,比如欧洲委员会 2001 年通过的《网络犯罪国际公约》(又称《布达佩斯网络犯罪公约》)³、阿拉伯国家联盟 2010 年通过的《打击信息技术犯罪公约》、独立国家联合体 2001 年制定的《独立国家联合体打击计算机信息领域犯罪合作协定》⁴,但这些公约或协定往往局限于特定区域或国家,缺乏可供所有国家参考的全球性统一标准。由此,在已经存在的各国法律体系、司法程序和执法能力差异的前提下⁵,全球性规范的缺位进一步阻碍了国家间通过合作打击网络犯罪。因此,制定一个全球性的打击网络犯罪规范成为迫切需求。以我国为代表的发展中国家跨出了关键的一步,倡议制定联合国层面的打击网络犯罪规范并使其得到落地实施,取得的阶段性成果表现为联合国网络犯罪问题特设委员会于 2024 年 8 月通过了《联合国打击网络犯罪公约》(United Nations Convention on Cybercrime)。

作为《联合国打击网络犯罪公约》核心之一的刑事定罪条款,涵盖了网络犯罪的多种形态,如数据盗窃、网络诈骗、网络恐怖主义等,还为缔约国提供了一定的刑事责任和追诉程序参考,以确保对网络犯罪的精准有力打击和促进各国合作。对《联合国打击网络犯罪公约》(以下简称《公约》)的刑事定罪条款进行研究,在理论上,有助于深入理解该公约的价值取向和网络犯罪定罪范围⁶;在实践上,有助于我国将来的公约衔接工作顺利开展,来履行作为缔约国的义务而对我国刑法进行解释或修订。同时,也有助于发现和完善公约中存在的不足之处,为未来的国际磋商和公约修订提供理论依据。

一、《公约》刑事定罪的制定与比较研究

《公约》刑事定罪条款的诞生历程充满了挑战和争议,体现了国际社会对网络犯罪问题的高度关注和合作治理需求,经过多年的谈判和协商,终于对联合国框架下的全球性打击网络犯罪公约条文达成初步一致。对比作为“打击网络犯罪里程碑”⁷的《布达佩斯网络犯罪公约》,它在普适性、前瞻性等方面具有优势,是以中俄为代表的发展中国家,在制定国际公约领域上发声、突破欧美立场的一次成功实践。

(一) 曲折的制定过程

³ 参见江溯:《打击网络犯罪的国际法新机制》,载《法学》2022 年第 11 期,第 45-59 页。

⁴ 参见王肃之:《打击网络犯罪国际规则的博弈与中国方案》,载《法学论坛》2021 年第 1 期,第 81-92 页。

⁵ 参见杨军:《论诉的利益之定位——基于大陆法系诉的利益与英美法系救济法的比较视角》,载《社会科学家》2014 年第 7 期,第 94-98 页。

⁶ 参见赵永琛:《〈联合国打击网络犯罪公约〉的制定问题研究》,载《武大国际法评论》2022 年第 5 期,第 1-20 页。

⁷ 转引自周振杰:《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》,载《中国刑事法杂志》2024 年第 5 期,第 105-120 页。

《公约》刑事定罪的制定过程是一个历经多年、多方参与磋商、充满争议和妥协的复杂历程。早在上世纪 90 年代国际社会便注意到了网络犯罪，期间通过各种国际文件与会议深化了对网络犯罪的合作治理，在 2024 年通过的联合国层面公约是对前述努力的重大巩固与发展。

1.初步探索与倡议

早在 20 世纪 90 年代初，联合国便开始探讨网络犯罪治理问题，并通过一系列决议和倡议推动国际社会对这一问题的关注和应对。如在 1990 年第八届联合国预防犯罪和罪犯待遇大会上，通过了关于计算机犯罪立法的决议，⁸为后续公约刑事定罪的制定奠定了基础；在 1994 年制定了《联合国关于预防和控制计算机相关犯罪手册》，指出此类犯罪的潜在范围之广，并强调了打击计算机犯罪方面国际合作的重要性。

2.综合性研究与政府间合作

在计算机与信息技术快速发展的背景下，国际社会为应对形式多变的网络犯罪，展开了对网络犯罪的综合研究与国际合作⁹。在 2000 年第十届联合国预防犯罪和罪犯待遇大会形成了《维也纳宣言》，呼吁各国制定行动政策，提高预防和打击高科技及计算机相关犯罪的能力；在 2005 年曼谷召开的会议上，有国家建议制定打击计算机犯罪的普遍公约，尽管未达成一致，但推动了针对网络犯罪的专门研究；在 2010 年形成《萨尔瓦多宣言》，建议成立政府间专家组，全面研究网络犯罪问题及其对策，次年第一次网络犯罪政府间专家组会议在奥地利维也纳召开，标志着联合国打击网络犯罪工作进入实质性和专门性阶段。

3.公约特设委员会的谈判与争议

为了回应各方全球性打击网络犯罪公约的需要，在 2019 年的联合国大会上通过了一项决议，决定建立公约特设委员会，针对为犯罪目的使用信息和通信技术的行为开展新的国际公约的起草工作，正式开始了《公约》刑事定罪的制定工作。然而最困难的部分在于谈判过程，特委会先后举行了多次谈判会议，包括明确公约框架和谈判路线图、充分听取各方意见和建议、着眼共识提出公约草案、聚焦争议问题达成一揽子妥协方案等阶段¹⁰。

各国在形成《公约》刑事定罪条款草案中的博弈十分激烈，在实体和程序上均存在较多争议，比如在作为本文探讨内容的实体法上，公约应关注的网络犯罪类型、是否涵盖主要依赖信息网络技术实施的犯罪、是否应该包含“故意”这一主观要素等；又如在程序法部分，相关强制措施的适用范围、犯罪案件程序性管辖

⁸ 参见刘守芬、房树新：《八国网络犯罪立法简析及对我国立法的启示》，载《法学杂志》2004 年第 5 期，第 17-19+22 页。

⁹ 参见张樊、王绪慧：《美国网络空间治理立法的历程与理念》，载《社会主义研究》2015 年第 3 期，第 146-153 页。

¹⁰ 参见裴炜、黄志雄、冯俊伟等：《〈联合国打击网络犯罪公约〉：背景、内容与展望》，载《数字法治》2024 年第 5 期，第 9-27 页。

权的连接点、数据存留措施是否设置合理怀疑条件、第三方主体的义务和责任范围以及数据监听等都有不同意见¹¹。

4.《公约》的最终通过

通过长久且艰难的磋商谈判，《公约》草案在各方的妥协、让步中，在2024年8月8日获得了联合国网络犯罪问题特设委员会一致投票通过。¹²在随后的2024年12月24日，联合国大会以协商一致同意的方式通过具有法律约束力的《联合国打击网络犯罪公约》。虽然公约获得了通过，但并未生效，公约将于2025年在越南首都河内举行的正式仪式上开放签署，随后向联合国会员国开放签署并持续到2026年12月31日，只有在第40个签署国批准后90天，公约才会正式生效。此外，要注意条约签署后的批准并非必要的，且是否批准由缔约方决定¹³，所以公约是否能切实在各国得到适用，取决于各国国内的积极衔接。

（二）与《布达佩斯网络犯罪公约》的比较

《布达佩斯网络犯罪公约》简称《网络犯罪公约》，是2001年由欧盟牵头制定，美国、加拿大、日本等国参与起草的国际公约，是世界上第一部专门针对网络犯罪的国际公约¹⁴。因其操作性强、能有效应对网络犯罪的突发与即时等优势¹⁵，被誉为“打击网络犯罪历史上的里程碑”。自2001年11月23日在布达佩斯开放供签署以来，该公约已成为对全球地区具有重要影响力的文件，截至目前，已有68个国家批准了该公约，另有两个国家（爱尔兰和南非）已签署但尚未批准，韩政府拟加入《布达佩斯网络犯罪公约》。

以欧美国家为代表的部分国家，强烈建议将其作为国际通行标准，欲使其成为真正的全球性公约而对非缔约国生效。但《布达佩斯网络犯罪公约》毕竟只是个代表欧美国家利益的区域性公约，相较于普适法更为狭窄¹⁶，其在管辖权、罪名设置、电子取证等问题上不能得到全部国家的支持，尤其是与中俄等发展中国家的一些立场存在矛盾¹⁷，所以制定一个在联合国框架下的打击网络犯罪全球性公约是十分有必要的。《公约》与《布达佩斯网络犯罪公约》在制定背景、定罪

¹¹ 比如，《公约》中很多调查取证措施的实施或者执行，就和网络服务提供者密切相关。数据快速存储、数据快速保全、数据调取、数据搜查扣押、数据实时收集、数据拦截等，都离不开网络服务提供者的参与和介入。

¹² 参见叶伟、姜博谦：《〈联合国打击网络犯罪公约〉开启网络空间国际规则新篇章》，载《中国信息安全》2024年第8期，第49-52页。

¹³ 参见【英】蒂莫西·希利尔著：《国际公法原理》（第二版），曲波译，中国人民大学出版社2006年版，第48页。

¹⁴ 参见Filippo S. International cooperation and protection of victims in cyberspace: welcoming Protocol II to the Budapest Convention on Cybercrime[J]. ERA Forum, 2022, 23(1): 101-108.

¹⁵ 参见皮勇：《论欧洲刑事法一体化背景下的德国网络犯罪立法》，载《中外法学》2011年第5期，第1038-1060页。

¹⁶ 参见【英】劳特派特修订：《奥本海国际法》上卷，王铁崖、陈体强译，北京商务印书馆1989年版，第34页。

¹⁷ 事实上，在打击网络犯罪领域，欧美国家没有意识到要去制定一个新公约，因为其早有《布达佩斯网络犯罪公约》。也正是因为有这一公约的存在，争议的焦点是“要不要制定新的全球化公约”，而非讨论具体内容。

范围与其他内容、国际影响力与参与度以及双方的不足之处与争议等方面均存在差异，两者各有优势与价值倾向，但凭借有效应对新型化网络犯罪的优势与处于联合国框架下的地位，公约具备巨大的潜能且应受到更多的关注。

《公约》与《布达佩斯网络犯罪公约》的比较主要体现在以下几个方面：

1.制定背景与主体

《公约》是在当今信息和通信技术迅猛发展且其他规范无法满足需求的背景下，为应对给各国经济、安全和社会秩序带来严重威胁的网络犯罪，尤其是在其复杂性不断提高的情况下制定的。是由中俄等金砖国家倡导并引领推动促成，是首个在联合国框架下打击网络犯罪的全球性公约。它广泛吸收了《联合国打击跨国组织犯罪公约》《联合国反腐败公约》《欧洲委员会网络犯罪公约》（即《布达佩斯网络犯罪公约》）等规范的先进经验，旨在加强国际合作以共同打击网络犯罪，并为此提供系统性的法律框架。而《布达佩斯网络犯罪公约》由欧洲委员会于 2001 年发布，主要应对当时网络犯罪的新挑战，是数字时代首个专门针对网络犯罪的区域性国际公约，它更侧重于解决欧美地区的网络犯罪问题。

2.罪名与其他内容

《公约》内容全面，涵盖了网络犯罪的刑事定罪、管辖权、程序措施和执法、国际合作、预防措施、技术援助和信息交流等多个方面。它首次将网络主权理念转化为具有约束力的法律规则，并明确规定各国在履行公约义务时应恪守主权平等、领土完整和不干涉内政原则¹⁸。而《布达佩斯网络犯罪公约》虽然也涉及网络犯罪的多个方面，其罪名规定大多局限在传统的计算机犯罪，如非法访问和拦截、滥用程序或数据伪造等，但以计算机作为工具的罪名仅有少数欺诈与儿童色情。此外，其存在无需他国同意便可进行跨境取证的条款，这种牺牲主权换取实效的做法也不能被普遍接受。¹⁹并且随着传统犯罪网络化的趋势，其部分条款已无法实现最初制定时的打击网络犯罪宗旨。

3.国际影响力与参与度

《公约》作为联合国主持制定的首个网络犯罪公约，该公约具有广泛的国际影响力和参与度²⁰。它的通过标志着联合国在网络和数字领域首次制定普遍适用的国际公约，为全球范围内的打击网络犯罪提供了重要的法律基础。《布达佩斯网络犯罪公约》虽然也在一定程度上推动了国际社会对网络犯罪的关注和打击，

¹⁸ 《布达佩斯网络犯罪公约》中规定了跨境数据提取制度，虽然服务器位于缔约国境内，只要其他缔约国有能力进行提取，那就可以不经过授权而获取，这属于严重侵犯一国的信息主权。这种跨境数据提取对于欧美国家十分有利，因为大多数实力强劲的互联网企业都属于欧美国家，且其分公司也是遍布全球。

¹⁹ 参见李本灿：《〈联合国打击网络犯罪公约〉核心议题述评》，载《法学论坛》2025 年第 1 期，第 92-103 页。

²⁰ 虽然具体参与国家的名单在公约正式签署和生效前尚未完全确定，但可以预见的是，由于公约是在联合国框架下制定的，因此其参与国家将包括联合国的广大会员国。值得注意的是，中国是公约谈判的首倡者和坚定支持者，在公约的制定过程中发挥了重要作用。此外，其他国家如俄罗斯、德国、美国等也积极参与了公约的谈判和制定工作。

但其影响力和参与度相对有限，主要局限于欧洲地区和部分其他国家。此外，公约在与其他国际规范的关系上也具有包容性，不禁止缔约国适用其他规范开展打击网络犯罪的合作，公约明确规定：如果缔约国就公约规定的事项缔结或其他协定或条约，有权适用此类协定或条约。²¹

4. 不足之处与争议

尽管《公约》在打击网络犯罪方面具有重要意义，但也存在一些不足之处和争议。例如，认为其定罪范围较窄，未能充分有效回应犯罪日趋网络化的现实需要，《公约》刑事定罪范围整体上采取了扩张和限定的折中，目的是为了争取“最大公约数”以实现公约各国合作的宗旨。一方面对一些各国均成熟的罪名认为无须纳入而排除在外，体现了“窄定罪”模式；一方面也突破了“存粹网络犯罪”以便更好应对网络犯罪现状，体现了“宽定罪”模式²²。同时，公约的人权保护和拒绝合作的条款过多等可能影响未来开展合作的成效²³。例如，部分西方人权组织和科技公司对该公约的一些条款表示担忧，认为其可能侵犯人权和隐私。而《布达佩斯网络犯罪公约》同样存在一些争议和不足之处。例如，其实体层面定罪范围过窄、程序层面跨境取证侵犯主权等问题在一定程度上限制了其打击网络犯罪的效果。同时，由于该公约由欧美国家主导，其全球性和广泛代表性也受到一定质疑。

二、《公约》中的刑事定罪部分

《公约》的“刑事定罪”部分（第 7-21 条）是公约核心内容之一，条文数量位居各章节第二。它全面规定了网络犯罪行为、法人责任、参与和未遂、追诉时效以及起诉、审判和制裁等相关条款，对确立网络犯罪治理的国际标准与促进合作机制具有重大意义。以下为“刑事定罪”的主要内容：

（一）网络犯罪行为

1. 网络依赖性犯罪和网络使能型犯罪

《公约》第二章共 15 个条文，其中第 7 条至第 17 条分别对 11 类网络犯罪的构成要件行为作出了具体描述。根据对网络的利用关系的不同，可以将公约规定的犯罪行为分为网络依赖型犯罪和网络使能型犯罪。网络依赖型犯罪指必须要以计算机为载体，离不开计算机网络；网络使能型犯罪指以计算机为工具，计算机网络只是一种手段、工具，其他非计算机网络手段亦可成立犯罪。学界亦称为“以计算机作为对象”和“以计算机作为工具”²⁴，前者实际上指针对计算机系统实

²¹ 参见商震、叶伟、姜博谦：《〈联合国打击网络犯罪公约〉若干重点问题评析》，载《中国应用法学》2025 年第 1 期，第 181-199 页。

²² 参见李本灿：《〈联合国打击网络犯罪公约〉核心议题述评》，载《法学论坛》2025 年第 1 期，第 92-103 页。

²³ 参见毛俊响：《国际人权话语权的生成路径、实质与中国的应对》，载《法商研究》2017 年第 1 期，第 153-163 页。

²⁴ 参见于志刚：《网络思维的演变与网络犯罪的制裁思路》，载《中外法学》2014 年第 4 期，第 1045-1058 页。

施的犯罪，典型的如破坏计算机系统罪；后者指“网络的工具化”或“传统犯罪的异化”，如电信诈骗。²⁵电信诈骗能够纳入公约，是中国联合其他国家持续努力的结果。中方代表团在谈判过程中多次强调，现今的电信诈骗存在传播迅速、证据易灭失、财产易转移等特性，危害远超传统的诈骗手段，最终引起委员会的重视并将广义电信诈骗纳入公约。²⁶

网络依赖型犯罪包括以“没有授权”为核心的非法访问（第7条）、非法拦截（第8条）；干扰电子数据（第9条），行为包括删除、更改电子数据；干扰信息和通信技术系统（第10条）；体现为“获取、制作、销售”等行为的滥用装置（第11条）五个条文。网络使能型犯罪包括与信息和技术系统有关的造假行为（第12条）、与信息和技术系统有关的盗窃或欺诈（第13条）、涉及网上儿童性虐待或儿童性剥削材料的犯罪（第14条）、为对儿童实施性犯罪而进行教唆或诱骗（第15条）、未经同意传播私密图像（第16条）以及对犯罪所得的洗钱行为（第17条），其中有条都是涉及儿童性犯罪，体现了西方国家对儿童性权利的重视及对此种法益侵害行为的严厉打击²⁷。缔约国有义务对这些被定义为刑事犯罪的活动进行预防、监测、侦查和起诉。

2. 依据犯罪内容进行分类

除了依据对计算机依赖程度进行分类，还可以依据犯罪行为的内容进行分类，也即传统的罪名分类。第一，数据、装置、信息和通信技术系统关联犯罪：涉及与信息和技术系统有关的造假行为、盗窃或欺诈等行为。这些规定直接保护的是计算机网络或数据，间接保护的是当事人的人身或财产权益；第二，传播、持有网络色情犯罪：包括生产、提供、发行或传送、取得及持有儿童的色情资料等犯罪行为。其中重点是保护儿童的性权利，也规制未经同意传播私密图片，这种私密图片依据其定义可以被我国刑法中的“淫秽物品”所涵盖²⁸；第三，网络洗钱犯罪：将为相关网络犯罪洗钱的行为定为刑事犯罪，并要求成员国提供政府、银行、金融、公司或企业的记录，以识别和追踪犯罪所得。关于洗钱罪，《公约》规定“各缔约国均应将本公约第7至第16条所确定的相关犯罪列为上游犯罪”，表明其他网络犯罪为洗钱罪之前置。

3. 未被纳入《公约》的罪名

在《公约》草案的谈判过程中，各国也提出了一些与计算机网络相结合的传统罪名，但与一些国家注重个人隐私、个人自由等的价值取向相悖，这些罪名虽

²⁵ 参见于志刚主编：《网络犯罪定性争议与学理分析》，吉林人民出版社2001年版，第112页。

²⁶ 参见商震、叶伟、姜博谦：《〈联合国打击网络犯罪公约〉若干重点问题评析》，载《中国应用法学》2025年第1期，第181-199页。

²⁷ 参见牛旭：《美国网络儿童色情物品持有罪的立法变革》，载《当代青年研究》2016年第3期，第116-121页。

²⁸ 参见蒋小燕：《淫秽物品的“淫秽性”之判断标准——以社会通念为基点》，载《法学评论》2011年第1期，第139-144页。

经讨论但最终未全面纳入。以欧美为首的西方国家一贯倾向于人权保护，乃至淡化国家主权，对公约中的某些刑事定罪和程序性条款抱有敌意，认为其存在侵犯隐私、言论自由的问题。以致美国呼吁未来的缔约国应当在其国内法中采取相关措施，确保公约的适用不会排斥人权保护，例如拒绝违背人权保护的司法协助。²⁹这些经讨论但未被纳入的罪名有公民个人权益保护相关罪名：如侵犯个人信息（数据）等，虽然这些行为普遍被各国刑法犯罪化，但因入罪标准差异较大，公约未能对此进行具有包容性和指引性的规定；又如网络恐怖主义、极端主义相关罪名：由于“恐怖主义、极端主义”的定义存在争议³⁰，将其列入也容易成为欧美国家指责或打击其他国家的借口，于是相关罪名被删除，但打击网络恐怖主义、网络极端主义代表整个国际社会的共同需求，在未来可能会被公约所涵盖。这也体现出公约是在各国的让步与妥协中形成的，既有网络化的传统犯罪列入，也对各国主张的罪名进行了合理限缩。

（二）其他规定

除了核心的网络犯罪行为，与之相配套的犯罪主体、共犯形式及程序性规定也被纳入《公约》，给各国设定公约义务和为适用公约来打击网络犯罪提供参考。第一，法人责任：公约第 18 条规定了法人责任，要求各国要制定关于法人实施网络犯罪的规则，其中法人的责任不局限于刑事责任，可以扩展到行政、民事责任，同时明确规定“法人责任不得影响实施此等犯罪的自然人的刑事责任”，确保了相关自然人无法借助法人担责以逃避制裁。值得注意的是，关于法人责任性质的立场采用《布达佩斯网络犯罪公约》的立场，即肯定了“法人可以犯罪”的观点³¹，这也是公约争取缔约国的一种努力。

第二，犯罪形态和共犯形式：首先，《公约》规定各国可以把网络犯罪的任何未遂行为乃至预备犯纳入惩罚范围，注意其用词是“可以”，表明这种公约义务可以由缔约国自由选择是否承担，并非强制性规定。其次，公约也规定各国应惩罚参与网络犯罪的“共犯、从犯或教唆犯”等任何身份犯罪人，注意这种共同犯罪的分类也是和我国有区别的³²，这种强制性规定表明公约对共犯的低忍耐力；第三，追诉时效：公约第 20 条对追诉时效进行了规定，要求各国考虑到犯罪的严重性，规定一个较长的追诉时效期，也要针对逃避司法处置的情形作特别规定（即确定更长的追诉时效期或作中止追诉时效）；第四，起诉、审判和制裁：公约第 21 条涉及起诉、审判程序和制裁措施。首先确定了罪刑相适应原则，确保实现“劝诫”和“震慑”效果，也规定对影响关键信息基础设施设定加重情节，规定假释、

²⁹ 参见吴沈括：《〈联合国打击网络犯罪公约〉的治理体系与中国因应》，载《中国法律评论》2025 年第 1 期，第 214-226 页。

³⁰ 参见何秉松、廖斌：《恐怖主义概念比较研究》，载《比较法研究》2003 年第 4 期，第 44-54 页。

³¹ 参见谢治东：《论单位(法人)刑事责任之本质——兼论我国单位犯罪立法模式之完善》，载《湖北社会科学》2011 年第 10 期，第 133-137 页。

³² 参见陈兴良：《共同犯罪论》，载《现代法学》2001 年第 3 期，第 48-57 页。

“保外候审裁决”制度，保护儿童犯罪人的权利。

毋庸置疑，《公约》的“刑事定罪”部分在提供明确的法律框架、较为全面的犯罪类型覆盖以及强调国际合作等方面具有显著优点。但也要承认，公约刑事定罪条款实用性问题、前瞻性不足以及证明难题³³（比如跨国网络犯罪的证据收集与整合）等缺点也需要得到正视。在实施公约的过程中，各国需要根据自身实际情况进行适当调整，达到扬长避短的效果。以下是公约的优缺点：

（一）优点

第一，明确了法律框架：《公约》涵盖了多种网络犯罪行为，包括数据、装置、信息和通信技术系统关联犯罪、网络儿童色情犯罪以及网络洗钱犯罪等，兼具以计算机“对象型犯罪”和“工具型犯罪”，初步明确了网络犯罪的定罪犯罪和标准，能够推动各国的司法实践朝着一体化行进，为国际社会打击网络犯罪提供了全面的法律框架；第二，具有预防性：公约的刑事定罪条款体现出了明显的预防性特征，有助于各国提前采取措施，防止网络犯罪的发生。公约规定的绝大多数是易于产生风险的行为犯，也存在明显的“预备行为实行化”和“帮助行为正犯化”³⁴的取向，比如公约规定“将故意为实施根据本公约确立的犯罪做准备的对行为确立为刑事犯罪”，即制裁预备犯。

第三，灵活性：《公约》在罪名设定与构成要件设置上保持了一定的灵活性，也对义务进行了选择性和强制性之分³⁵，很大程度上提高了包容性，以便各国根据自身情况进行调整和适用。比如在处理共犯的问题上，考虑到德日刑法不处罚所有的独立教唆行为³⁶，公约在确立教唆行为为犯罪的同时，规定了这种教唆必须具有可罚³⁷性这个条件，又如在規定对儿童实施性犯罪进行教唆和诱骗时，允许缔约国作“更进一步”的行动，也在具体罪名衔接上规定“根据本国法律”和在起诉、审判和制裁条款中规定罪刑相适应原则；第四，尊重国家主权：公约在起草之初就非以《布达佩斯网络犯罪公约》为直接标准，在第5条明确规定了“保护主权”，规定“概不赋予缔约国在他国境内行使管辖权和履行该另一国国内法律所规定的专属于该国主管机关的职能的任何权利”，履行“不干涉”义务³⁸来维护国际法基本原则，也是排除了司法上的“长臂管辖”³⁹。例如在具体的跨境提取数据的问题上，排除了欧美主张的“未经许可直接提取缔约国境内的计算机存储数据”，贯彻了保护一国的信息主权原则。

³³ 参见张迪：《网络犯罪数额证明难题及其应对》，载《现代法学》2024年第5期，第162-176页。

³⁴ 参见熊亚文：《从共犯论到法益论：帮助行为正犯化再审视》，载《法学》2021年第8期，第83-95页。

³⁵ 参见梁西主编：《国际法》（修订第二版），武汉大学出版社2003年版，第286-288页。

³⁶ 参见方加亮：《论德、日刑法中的“组织犯”》，载《中州学刊》2002年第2期，第188-190页。

³⁷ 参见周光权著：《刑法总论》（第四版），中国人民大学出版社2021年版，第363-368页。

³⁸ 参见【美】约翰·罗尔斯著：《万民法》，张晓辉等译，吉林人民出版社2001年版，第40页。

³⁹ 参见李庆明：《论美国域外管辖：概念、实践及中国因应》，载《国际法研究》2019年第3期，第3-23页。

（二）缺憾

第一，实用性不足：虽然《公约》处于联合国框架下，对比《布达佩斯网络犯罪公约》有更广的参与度，但在磋商过程中大范围地删除了更广泛的网络化传统犯罪罪名，导致公约在公民个人权益保护和应对网络恐怖主义、极端主义方面的规定缺乏足够的普适性，以致公约在实践中的实用性不大，比如我国在应对跨境电信诈骗时，依旧得适用传统的国内法域外适用手段⁴⁰，在调查取证、引渡等方面受限，也导致公约强调的加强国际合作宗旨无法实现；第二，前瞻性略显不足：随着网络技术的不断发展，新型网络犯罪形态与附随出现的犯罪形态层出不穷。公约在制定时连已经出现的网络化传统罪名都没能有效涵盖，更何况能充分规制例如“帮信犯”⁴¹等犯罪形态，因此在前瞻性方面存在不足。又如，公约未对区块链、人工智能等新技术及此类新技术支撑下的网络黑灰产业链催生的新型犯罪形态作出探索性规定。针对公约罪名较少的问题，即使公约已经通过但仍有补足的余地，公约谈判委员达成了一项共识：在公约通过两年后谈判拟定补充协议书草案，以酌情处理其他刑事定罪问题。但附加议定书的产生也十分严苛，除了启动议定书的谈判时间点限定在两年后，议定书的通过必须具备 60 个以上缔约国参会这个条件，这些约束都是公约谈判艰难的缩影⁴²；第三，证明难题：公约中某些罪名的证明标准较高，给实际执法和司法工作带来了一定的困难，这点主要是指作为定罪前提的术语条款，比如公约关于“订户信息”系统的定义，除了列举繁多的“邮政信息、地理地址、电话或其他接入号码”等，还包括“任何其他信息”的兜底内容，使得在收集和固定证据方面面临巨大挑战，这种涉及高度技术化的手段超出了传统刑事司法证明的能力范围。

三、《公约》刑事定罪的中国思考

《公约》的刑事定罪部分与我国《刑法》在打击网络犯罪方面既有相似之处，也存在一些差异，比如都涵盖了非法访问、干扰电子数据等破坏计算机犯罪，但公约规定的某些持有型犯罪如持有儿童色情材料却不被我国刑法所约束。这些对比引发了关于公约在中国适用的思考，亦启发了中国刑法的进一步完善。

（一）相通之处

1. 罪名设置

我国《刑法》能涵盖《公约》规定的大部分犯罪。比如我国刑法也规定了网络依赖型犯罪和网络使能型犯罪。公约中规定的如“非法访问”、“非法拦截”、“干

⁴⁰ 参见贾济东、胡扬：《论我国反洗钱法域外适用的困境与出路》，载《华中科技大学学报(社会科学版)》2021年第2期，第116-126页。

⁴¹ 参见李明鲁、皮勇：《网络空间犯罪中的公共秩序法益研究——从帮助信息网络犯罪活动罪的独立性切入》，载《海南大学学报(人文社会科学版)》2023年第5期，第132-144页。

⁴² 参见商震、叶伟、姜博谦：《〈联合国打击网络犯罪公约〉若干重点问题评析》，载《中国应用法学》2025年第1期，第181-199页。

扰电子数据”等网络依赖型犯罪，与我国刑法中的计算机犯罪⁴³相对应，如非法侵入计算机系统罪、破坏计算机信息系统罪。在具体的犯罪构成来看，刑法第 285 条规定的被侵入的计算机系统，分为涉及国家利益的例如国防建设的计算机系统和其他信息系统，但公约规定的非法访问、拦截等，未对对象进行区分，在量刑上也未作特殊规定。不过，公约允许缔约国将“违反安全措施”作为犯罪构成，那么显然就可以被“危害国家安全”所涵盖。对于干扰信息行为的描述亦存在差异，如公约中表述为“输入”电子数据，而刑法中表述为“增加”，公约使用了带价值判断的“劣化”，刑法则单纯表述为“修改”；公约中的部分网络使能型犯罪，如“与信息通信技术系统有关的造假行为”、“与信息通信技术系统有关的盗窃或欺诈”等，也在我国刑法中有相应的体现，而且是表现为第 287 条的提示性条款⁴⁴，表明对网络使能型犯罪的重视。关于计算机诈骗，有观点认为我国刑法应效仿德日设置计算机诈骗罪，因为对“机器能否被骗”存在争议，但《公约》的计算机欺诈是指以计算机作为工具、手段，而非以计算机作为对象而进行犯罪。⁴⁵

在有关性犯罪条款方面，《公约》第 14 条规定的与网上儿童性虐待或儿童性剥削材料有关行为包括制作、销售、展示和出版等，第 16 条规定的“未经同意传播私密图像罪”行为包括利用信息和通信技术系统销售、传播，这些散布儿童性犯罪材料的行为都能被我国《刑法》的“制作、贩卖、传播淫秽物品罪”所制裁。此外，公约关于私密图像的定义“包括照片或录像在内的任何方式对一个年满 18 岁的人进行的具有性性质的视觉记录，其中该人性器官暴露，或录制时该人正在进行私密的性活动”，显然也属于我国刑法中被表述为“具体描绘性行为或者宣扬色情”的淫秽物品。但在保护儿童方面，公约和刑法的侧重不同，公约是禁止“持有儿童性犯罪材料”⁴⁶来防止诱发针对儿童的性剥削，⁴⁷而刑法是通过加重处罚“向未成年人传播淫秽物品”来防止毒害儿童心灵，乃至在将来，我国刑法也可以将“传播儿童色情”作为传播淫秽物品犯罪的从重处罚情节。

如上所述，《公约》与我国《刑法》在这些犯罪的具体成立要件上存在差异。比如公约 规定的非法拦截、干扰电子数据等罪名都是行为犯，成立犯罪不要求有实害结果，也没有规定需要主观犯罪意图，虽然规定缔约国可要求这种犯罪“以获取电子数据或有其他不诚实意图或犯罪意图”，但这属于非强制性义务。而我国刑法规定的破坏计算机信息系统等犯罪，在犯罪成立要件上与公约的规定不同，

⁴³ 我国的网络犯罪，位于《刑法》妨碍社会管理秩序罪部分的扰乱公共秩序罪中，是以第 285 条至第 288 条为核心的网络犯罪罪名体系。

⁴⁴ 参见谭莹：《论网络共犯的结果归责——以〈刑法〉第 287 条之二为中心》，载《中国法律评论》2020 年第 2 期，第 115-129 页。

⁴⁵ 参见李本灿：《〈联合国打击网络犯罪公约〉核心议题述评》，载《法学论坛》2025 年第 1 期，第 92-103 页。

⁴⁶ 参见 Canning M, Keenan M, Breslin R. The Sexual Exploitation of Children in the Care of the Irish State—Professionals' Perspectives[J]. Child Abuse Review, 2024, 33(6): e70013-e70013.

⁴⁷ 参见赵合俊：《禁止儿童性剥削——国际法与国内法之比较研究》，载《妇女研究论丛》2013 年第 1 期，第 40-46 页。

大多要求“情节严重”、“造成计算机系统不能正常运行”和“后果严重”等条件⁴⁸。但公约也没有剥夺缔约国的保留权利，允许缔约国可以选择性地附加条件，比如公约明确了各缔约国可以规定在“造成了严重损害”的情况下才成立犯罪。因此，我国刑法的现有罪名设置是公约可以涵盖的，或是公约允许进一步细化的。

2.其他规定

在法人责任方面，《公约》要求的犯罪法人责任承担在我国《刑法》已有规定。我国刑法不仅在 30 条规定了单位犯罪，更在第 31 条规定了具体的责任承担方式（包括法人共同犯罪的责任承担）⁴⁹，即“双罚制”，确保对法人的制裁是“有效，适度，而且具有警戒性”。同时，较之公约，我国刑法所规定的法人应更为宽泛⁵⁰，不仅包含“公司”、“企业”等传统法人，更有“团体”这种较为模糊的表述。此外，针对公约规定的法人承担刑事责任之外的民事、行政责任，我国刑法也作了兜底，即第 31 条中的“本法分则和其他法律另有规定的，依照规定”；在犯罪形态与共犯形态方面，公约要求的制裁具有“可罚性”犯罪预备、未遂等犯罪停止形态，以及处罚共犯、从犯或教唆犯等参与人，我国刑法中的“犯罪的预备、未遂和中止”以及不能犯⁵¹、“共同犯罪”部分也都能满足。

在追诉时效方面，《公约》第 20 条规定各缔约国应酌情在国内法中规定一个较长的追诉时效期，同时，为了应对犯罪人逃避刑事司法处置，要确立更长的追诉时效或认定为中止追诉。我国《刑法》第 87 条至第 89 条规定了追诉时效，包含针对无期徒刑和死刑的 20 年长追诉期限，乃至针对“经过 20 年任有必要追诉”的情况可以请求最高检突破最长追诉期限，以及应对逃避司法可以不受追诉期限的限制，都可以衔接公约的规定；在犯罪构成论证方面，公约有“主观内容客观推定”的倾向，如公约第 17 条规定的洗钱罪中，对于“明知、故意或目的”的推定，可以“客观实际情况”为依据。针对这一点，我国也通过了司法解释进行涵盖，在对行为人是否“知道”犯罪获利进行判断时⁵²，以“犯罪所得的转移方式、人员关系、职业经历”等客观情况为根据。

（二）我国《刑法》未包含的部分

1.持有型犯罪与洗钱罪

存在一些《公约》中规定的犯罪行为在我国《刑法》中尚未明确覆盖的情况。以持有型犯罪为典型，公约中规定了持有儿童性虐待或儿童性剥削材料犯罪，要求各缔约国将通过通信技术“访问”、“获取”、“拥有”或“控制”儿童性虐待或性剥削材料的行为规定为犯罪，而我国刑法对单纯的持有此种材料并不规定为犯罪，

⁴⁸ 参见于志刚：《中国网络犯罪的代际演变、刑法样本与理论贡献》，载《法学论坛》2019 年第 2 期，第 5-15 页。

⁴⁹ 参见陈兴良著：《共同犯罪论》（第三版），中国人民大学出版社 2017 年版，第 295-296 页。

⁵⁰ 参见魏远文：《论单位犯罪的“单位”与单位人格否认》，载《北方法学》2019 年第 4 期，第 99-113 页。

⁵¹ 参见张明楷著：《刑法学》（第六版）上册，法律出版社 2021 年版，第 454-465 页。

⁵² 参见张明楷：《洗钱罪的争议问题》，载《政法论坛》[2024-12-30]，第 1-17 页

只有利用网络进行了传播、售卖等造成危害的行为才能以传播、贩卖淫秽物品罪进行定罪⁵³。对于持有、访问儿童性虐待、性剥削材料行为，有学者认为我国刑法无必要进行专门立法，因为对比持有枪支、假币等持有性犯罪，这种行为产生的对法益的危险较为缥缈虚无；即使是增设了相关条文，也应当做到降低定罪门槛，如可以设定作为定罪阻却的悔罪条款或数量要求。⁵⁴其他国家对于公约中的网络儿童色情犯罪也存在争议，比如日本主张儿童色情材料应该限定在记录真实人体的范围内，而将其国内发达的动漫等拟人创作材料排除在外；欧盟某些成员国认为儿童自行制作色情材料不就有可罚性，倾向于对此作出法律保留。⁵⁵又如洗钱罪，我国刑法也规定了成立洗钱罪的前置条件，即要以毒品犯罪、恐怖活动犯罪、贪污贿赂、破坏金融管理秩序等特定犯罪作为“上游犯罪”，但是公约要求至少把第7条至第16条所确立的罪名如非法访问、拦截等作为上游犯罪，而这些罪名不能触发我国刑法的洗钱罪。当然，虽然《公约》规定的这些上游犯罪所带来的洗钱行为无法被我国刑法所规定的洗钱罪所制裁，但可以从掩饰、隐瞒犯罪所得的角度评价，故此处所指的未涵盖范围是洗钱罪条文。

2. 无授权提供私密图像

对于一些《公约》规定的非强制性义务，我国《刑法》也未有涵盖。比如公约规定的“通过网络无授权传播隐私图像”行为，即第16条所规定的利用信息技术“出售、分销、传送、发布或以其他方式”提供私密图像且“未经图像中所描绘的人同意”，各缔约国可在有“造成伤害”的意图下对其进行制裁，我国在淫秽物品相关罪方面规定了以“传播牟利”为目的的传播、贩卖淫秽物品牟利罪，而无伤害目的⁵⁶。同时，“无授权散布私密图像”的行为也难以归为刑法第237条的“以其他方法强制猥亵他人或侮辱妇女”⁵⁷和第246条的“以其他方式公然侮辱他人”，且这两种犯罪需要有的“严重或恶劣情节”不同于伤害意图⁵⁸，故我国刑法对此没有覆盖。此外，关于此行为的“伤害”意图，公约和我国刑法也有差异，公约中的伤害应是间接的精神伤害，因为散布图像并不会对受害人的身体完整和功能健全造成直接伤害，但我国刑法中的伤害通常是指身体伤害，不包括被害人的精神伤害以及由精神损害造成的间接伤害。有学者认为这种“未经同意”传播私密图像的行为，侵犯了受害者的性隐私权。首先不属于强制侮辱罪，因为这种行为不属于侵犯性自主权或直接的性羞耻，也不包含强制因素；也无法评价为寻衅滋事和侮辱罪，

⁵³ 参见皮勇：《互联网上淫秽信息涉罪的几个问题》，载《法学评论》2002年第3期，第144-149页。

⁵⁴ 参见李本灿：《〈联合国打击网络犯罪公约〉核心议题述评》，载《法学论坛》2025年第1期，第92-103页。

⁵⁵ 参见吴沈括：《〈联合国打击网络犯罪公约〉的治理体系与中国因应》，载《中国法律评论》2025年第1期，第214-226页。

⁵⁶ 参见黎宏著：《刑法学各论》（第二版），法律出版社2016年版，第482-491页。

⁵⁷ 参见陈家林：《〈刑法修正案(九)〉修正后的强制猥亵、侮辱罪解析》，载《苏州大学学报(哲学社会科学版)》2016年第3期，第68-75页。

⁵⁸ 参见陈兴良主编：《罪名指南》（第二版）上册，中国人民大学出版社2008年版，第687-692页。

因为未必会侵害公共秩序和个人的名誉、人格尊严。因此，这种侵害性隐私权行为在确实可以识别个人身份的情况下，可以评价为侵犯公民个人信息罪。⁵⁹

3.其他差异

在一些技术细节和具体行为的描述方面，《公约》和《刑法》亦有不同。例如，公约将侧重无授权的“非法访问”规定为犯罪，而我国《刑法》中打击的是侧重获取或控制数据的“非法侵入”⁶⁰；又如上述，公约将“输入、传输、破坏、删除、劣化、更改或隐瞒电子数据”规定为“干扰信息和通信技术系统”的行为，而我国刑法中的破坏计算机系统罪仅就“删除、修改、增加、干扰”加以规定。又如上述公约中的“行为犯”，对于同一罪名，我国刑法规定必须要有相应的危害才能成立犯罪，而公约只需要有相应行为即可，且缔约国是否附加“犯罪意图”条件，是属于选择性的非强制义务。

（三）我国对《公约》刑事定罪的应对

1.调整现有部分

在《公约》生效后，我国相关部门应及时展开对公约生效和批准后的义务履行工作，学界与实务部门应对公约中的刑事定罪条款进行全面深入的研究，理解其预防性、灵活性、客观性、延展性与主权性特征⁶¹，并对比我国现行刑法的相关规定，明确覆盖和未覆盖的方面，确保我国《刑法》与公约的具体规定实现体系性衔接。对于公约中已经覆盖的犯罪行为，如非法访问、非法拦截，以及与信息和通信技术系统密切结合的造假行为、盗窃或欺诈行为等网络化的传统犯罪，我国应通过司法解释或立法来继续细化或调整相关法律规定，确保与公约相适应，此外，在通信技术快速发展的背景下，加强刑法的惩罚力度也助于有效打击新型网络犯罪。

2.增设未覆盖部分

对于那些目前我国《刑法》无法履行的义务，即未覆盖的部分，可以通过合理扩大我国刑法的罪名范围来实现，不仅能推动《公约》在我国的有效适用，亦能推动我国刑法罪名体系的更新。例如，针对公约中涉及持有儿童性虐待或儿童性剥削材料犯罪，我国可以考虑在刑法中增设相应的罪名或调整现有罪名的构成要件，以确保对这类行为的有效打击，有助于更好地保护儿童权利。司法解释明确对“隔空猥亵儿童”进行处罚，就是在保护儿童权利方面接轨国际的一次实践⁶²；又如洗钱罪，针对其上游犯罪列表，我国刑法可以立法或司法解释对其进行扩张，

⁵⁹ 参见李本灿：《〈联合国打击网络犯罪公约〉核心议题述评》，载《法学论坛》2025年第1期，第92-103页。

⁶⁰ 参见刘宪权：《数据犯罪刑法规制完善研究》，载《中国刑事法杂志》2022年第5期，第20-35页。

⁶¹ 参见周振杰：《〈联合国打击网络犯罪公约〉刑事定罪条款与中国刑法应对》，载《中国刑事法杂志》2024年第5期，第105-120页。

⁶² 参见朱光星：《网络隔空猥亵儿童的定罪研究——以保护儿童为分析视角》，载《中国政法大学学报》2022年第4期，第206-217页。

把公约要求的相关网络犯罪列入其中,也是在实践需要的情况下合理扩大刑法打击面。

3.合作与术语明确

同时,针对《公约》完善了“打击网络犯罪的国际合作”这个优势,我国应加强与其他国家在打击网络犯罪方面的交流与合作,借鉴他国的有利经验,正如早有呼声的刑事司法协助法成为现实⁶³,理论与实务界应一起完善国内法关于“电子数据跨境取证”⁶⁴、“网络犯罪管辖权”等内容,中国与各缔约国有信心与决心共同应对网络犯罪的挑战。此外,作为网络犯罪刑事定罪前提的术语内容,在我国《刑法》上没有很好地得到体现,但在公约中却占了较大的篇幅。例如关于“流量数据”的定义,不仅包括“由通信技术进行生成”或“作为通信链一部分”等特征,也对“起点”、“路径”和“持续时间”等显示内容进行了列举。有的学者认为,我国刑法关于某些网络犯罪术语的表述还停留在早期计算机犯罪时代,例如“计算机系统”“计算机信息系统”等术语的描述内容无法与公约相匹配。⁶⁵我国刑法可以借鉴其他国家的经验,在刑法或其他相关法律中对涉及网络犯罪的法律术语或证明规则进行明确⁶⁶。

四、结语

《联合国打击网络犯罪公约》是各国长期谈判和博弈的产物,它建立了联合国层面的网络犯罪和合法数据访问法律框架,是对代表西方立场的《布达佩斯网络犯罪公约》的突破,被称为迄今唯一的全球性网络犯罪“国际硬法”,它将于2025年经至少40个签署国批准后生效。届时,作为公约倡导者的我国应该起到引领作用,积极履行公约赋予的义务。公约通过离不开以中国为代表的广大发展中国家的共同努力,将《布达佩斯网络犯罪公约》视为黄金标准西方国家,一开始反对制定新的网络犯罪公约,后来转变为参与谈判并努力促成公约的通过,不仅反映了现今应对复杂网络犯罪的需求,也体现了广大发展中国家打击网络犯罪的决心。⁶⁷公约的刑事定罪部分给我国刑法带来许多启示,是否应当增设罪名或者作出相应的司法解释,比如肯定西方国家早已定为刑事犯罪的“持有儿童性剥削或性虐待材料”的可罚性,⁶⁸又或者扩张洗钱罪的上游犯罪,这都是衔接公约应关注的问题,也是我国《刑法》与时俱进必须要进行的考虑。在作为公约优势的国际

⁶³ 参见张志勋:《刑事普遍管辖权的发展趋势与我国的对策》,载《江西社会科学》2011年第9期,第172-176页。

⁶⁴ 参见潘美玉、高慧、戴奎著:《刑事证据审查手册》,法律出版社2021年版,第254-277页。

⁶⁵ 参见吴沈括:《〈联合国打击网络犯罪公约〉的治理体系与中国因应》,载《中国法律评论》2025年第1期,第214-226页。

⁶⁶ 参见皮勇:《〈网络犯罪公约〉中的证据调查制度与我国相关刑事程序法比较》,载《中国法学》2003年第4期,第146-161页。

⁶⁷ 参见商震、叶伟、姜博谦:《〈联合国打击网络犯罪公约〉若干重点问题评析》,载《中国应用法学》2025年第1期,第181-199页。

⁶⁸ 参见 Man Charged with Sexual Exploitation of Children and Distribution, Receipt, and Possession of Child Sexual Abuse Material[J].Department of Justice (DOJ) Documents / FIND,2023.

刑事司法合作方面，我国刑法应立足于本国、借鉴他国优秀经验，积极在电子证据获取、证明规则等方面进行更新，这是网络犯罪日益多元背景下的重要完善方向。

2025年中国国际法学会年会

全球卫生治理中人工智能算法监管的法律困境与突破

陈海玲*

(华南师范大学法学院 广东广州 510006)

摘要： 本文认为全球卫生治理中人工智能算法监管至关重要，但其面临法律框架不完善、监管主体不明确、责任界定模糊、数据隐私保护不足等困境。这些困境是由技术发展快、法律滞后、科技公司逐利以及国际合作不畅导致的。文章深入分析问题后针对性提出完善法律框架、明确监管主体、清晰界定责任、加强数据隐私保护和促进国际合作等突破对策，并强调中国应积极参与国际规则制，为人类卫生健康共同体的建设提供有力的法律支持。

关键词： 全球卫生治理；人工智能；算法监管；《国际卫生条例（2005）》修正案；人类卫生健康共同体

Abstract

This article assumes that regulation of the AI algorithms in global health governance has been the most important at this moment. But it faces a range of legal obstacles, including the incompleteness of law framework, the uncertainty of regulatory body, the ambiguity of the liability definition and the deficiency of data privacy protection. The problems emerged from fast technological advances, slow response of the laws, profit-seeking action of tech enterprises as well as insufficient international cooperation. Through in-depth analysis, the article proposes specific policy measures including improvement of law system, confirmation of regulatory responsibilities, clarification of legal responsibility, strengthening data privacy protection, and the promotion of international cooperation. It also reflected China's initiative in the formulation of rules for the world to provide firm legal backing for building a Human Health and Wellness Community.

Keywords: global health governance; Artificial intelligence; Algorithm supervision; Amendments to the international health regulations (2005); Human health community

*作者简介：陈海玲，女，华南师范大学法律硕士。

人工智能算法正在全球卫生中应用，为疾病的预防控制、医疗服务质量优化等方面发挥积极作用，但同时，算法的“黑箱”以及数据隐私风险等诸多问题不断爆发，对人工智能算法发挥积极作用、助力全球卫生发展带来严重的负面影响。如何在确保公共卫生安全的前提下有效监管人工智能算法，充分发挥算法优势，已经成为全球卫生法的热点问题，本文通过对当前监管困境的分析探索破解良策，以为促进全球卫生治理的智能化发展奠定坚实的法律基础。

一、全球卫生治理中人工智能算法监管的法律困境

（一）法律框架的不完善

人工智能算法监管在大数据背景下遇到的重大法律难题之一是缺乏相关的法律框架。从当前来看，无论是发达国家还是发展中国家，均尚未形成完整且切实的全球卫生治理中人工智能算法监管的法律框架，人工智能算法相关监管的法律法规均存在较大的差异性，这种差异性直接在跨国性的卫生治理实践中，产生监管真空与重复监管的并存。在面对人工智能算法在具体医疗实践过程中的各种挑战之时，有关部门很难做出有效的干预。现存的法律法规框架，均主要是对传统医疗事故、大数据保护等方面的法规治理，缺乏人工智能算法相关风险问题的具体与针对性的规定，实在难以顾及到目前实际的全球卫生治理需求。

从世界的角度看，经济合作与发展组织（OECD）颁布《人工智能原则》，虽然也在一定程度上对人工智能发展给予了原则性规范，突出强调人工智能发展应有的伦理道德要求、透明度和解释可及性等原则，但是这些原则主要均是建议性的、并非强制法律约束。在世界卫生治理中，面对现实的人工智能算法应用时，这类原则很难转化成为有效的监管对策。就卫生领域而言，不同国家在人工智能算法在处理医疗数据、疾病预测模型等方面在法律上的规定各有不同，比如在数据出境方面，一些国家对于医疗数据出境禁止得很严格，也有国家规定不够严格，这显然都会给跨地区的人工智能医疗协作带来很大阻力，也使得人工智能对这种全球统一的法律秩序显得格外稀缺。¹

以欧盟《通用数据保护条例》（GDPR）和美国《健康保险流通与责任法案》（HIPAA）为例，分别就各自关注的重点进行条款规定。前者关注数据主体权利保护，例如第15条规定强制执行详细的数据处理信息；后者主要聚焦于医疗保健行业有关的信息安全。但在全球卫生治理情境下，人工智能算法往往存在跨国界的数据流以及多主体的交互问题，这两大法典难以顾及不同国家/地区在算法治理方面的协调差异，监管空位与重叠不断显现。²

¹ 参见北京大学公共卫生学院：《〈国际卫生条例（2005）〉修订通过引起关注》，载中国法院网，2024年9月6日，<https://www.chinacourt.org/article/detail/2024/09/id/8102627.shtml>，最后访问日期：2025年2月24日。

² 参见魏健馨：《生物安全法》的宪法逻辑[J]，上海政法学院学报（法治论丛），2022，37(3):63-75。

2024年6月1日,《国际卫生条例(2005)》修正案经过通过,在全球公共卫生领域建立诸如“协调财政机制”和“实施和履约委员会”等新的协调制度,极大提升了全球卫生治理的协调程度,提高了资源配置效率。但在人工智能算法监管领域,有效且具有广泛约束力的成体系国际公约并未产生,在各国法律对待数据跨境流动、算法责任界定等方面存在较大差异的情况下,即使是最具约束力的欧盟GDPR和美国HIPAA在数据跨境流动、算法责任界定等关键问题上也分歧显著,在全球公共卫生治理领域存在明显难以合作点,这使得人工智能算法监管面临很大困难,也进一步凸显了统一的法律框架构建的重要性和迫切性。我国《生成式人工智能服务管理暂行办法》规定了生成式人工智能服务的训练数据处理、服务规范,其中明确使用合法来源数据和基础模型等规定,但在全球卫生治理的语境下与国际法律的衔接还需要进一步探讨,各类问题的纷起,也都体现了建立统一法律框架的必要性。

在生物安全领域,国际上已形成以《生物多样性公约》为代表的一系列国际公约,构建起相对完善的生物安全国际法体系,涵盖生物安全多个领域,并为成员国设定了相应义务。但在人工智能算法监管方面,国际层面缺少类似的具有广泛约束力和系统性的公约。当前各国法律差异较大,在跨国卫生治理中,人工智能算法监管面临诸多问题。例如欧盟的GDPR虽然对数据保护规定严格,但在算法监管的其他方面与美国等国家法律存在差异,导致在全球卫生治理中,人工智能算法监管难以形成统一标准,出现监管空白和重叠,难以有效应对人工智能算法带来的挑战。

欧盟《通用数据保护条例》(General Data Protection Regulation GDPR)明确规定数据主体享有一项权利,即有权从数据控制者处获得有关数据的控制者对个人数据进行处理有关详细的信息,这些信息包括处理的目的、个人数据的种类等内容;美国《健康保险流通与责任法案》(Health Insurance Portability and Accountability Act, HIPAA)要求医疗保健提供者应采用适当的行政、物理和技术保障措施保护电子的受保护的健康信息保密性、完整性和可用性,等等。但是当应对人工智能算法在应对全球卫生治理带来的独特风险时,这些法律都具有局限性。例如,在采用人工智能算法以建构疾病预测模型时,训练模型时的数据可能来自不同国家,并且涉及数据保护的标准不同。这将导致数据整合和使用面临着巨大挑战。根据GDPR,数据主体有权在任何时候撤回个人数据处理的同意。在跨国疾病预测的场景中,如果一个数据主体撤回同意,这将扰乱模型的整体训练数据的完整性,影响算法的精确度。同时,HIPAA主要是针对国内医疗保健行业的数据跨境处理,然而并没有相应的规定。在处理跨国企业间的合作中,数据如何流动并不十分清晰。因此我国可以借鉴《生物安全法》中系统性立法的原则,制定专门的人工智能算法监管法律,以此清晰划定从算法制备阶段、数据采

集和使用阶段以及算法应用的全过程规则。例如我国在立法中要求算法开发者在涉及全球卫生治理问题下，将所用数据的来源信息嵌入于算法当中，以达到数据跨境可追溯的目的，并且填平在国际或国内法上的空白。³

目前，我国尚未出台专门针对人工智能治理的综合性法律，不过已针对人工智能关键要素，即数据、算法以及算力，构建起了不同层级的法律法规体系。以算法监管来说，《生成式人工智能服务管理暂行办法》正式施行，明确国家对生成式人工智能服务采取包容审慎且分类分级的监管原则。办法要求提供者开展训练数据处理活动时要依法依规，采用合法来源的数据与基础模型，涉及知识产权的不得侵权，涉及个人信息需获同意或符合法定情形，同时提升数据质量。在服务规范上，提供者需依法担责、履行安全义务，这些规定规范了相关活动，为全球卫生治理中人工智能算法监管提供了有价值的法律实践经验。⁴

同时，人工智能应用到全球卫生治理中也会出现诸多法律挑战。例如在数据隐私保护问题上，欧洲地区 GDPR、美国 HIPAA 等，在实际应用时在管辖权方面或同意方面具有模糊地带⁵，《生成式人工智能服务管理暂行办法》第七条规定，生成式人工智能服务提供者开展预训练、优化训练等训练活动，应当使用具有合法来源的数据和基础模型；涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形。这为人工智能算法应用于全球卫生治理中的卫生数据处理，给出了规制性的内容，合理规范了算法处理卫生数据的范畴。⁶

此外，《欧盟人工智能法案》依据潜在风险程度，将人工智能系统划分成禁用级、高风险级、有限风险级、最低风险级四类。处理个人数据的系统，同时受该法案与《通用数据保护条例》规制，违规最高罚 3500 万欧元或年营业额 7%，研究它能分析对全球卫生治理算法监管的意义与挑战。⁷

从整体情况来看，欧盟的 GDPR 与美国的 HIPAA 在数据隐私及保护方面都制定了较为严格的规定。然而，当人工智能算法应用于全球卫生治理领域时，这些法规却在实际施行过程中遭遇了众多挑战，特别是在跨国数据流动以及不同主体数据权益的界定方面存在着不少模糊不清之处。此现象迫切需要进一步予以明确和完善。我国在人工智能制度建设进程中持续发力陆续颁布了《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等法律法规。《生成式人工智能服务管理暂行办法》明确指出国家秉持发展与安全并重、促进创新

³ 参见田晨，杨水华，杨克虎，等. 构建全球卫生政策与管理证据库，助力循证决策与卫生治理可持续发展[J]. 协和医学杂志，2024.

⁴ 参见许丽. 论通用人工智能治理中管制与反垄断的协同[J]. 上海政法学院学报（法治论丛），2024，39(1):117-132.

⁵ 参见曹建峰. 迈向可信 AI：ChatGPT 类生成式人工智能的治理挑战及应对[J]. 上海政法学院学报（法治论丛），2023，38(4):28-42.

⁶ 参见张薇薇. 美国的人工智能算法治理：理念、举措及影响[J]. 当代美国评论，2024，8(3):108-127+133-134.

⁷ 参见陈永森，陈博影. 人类卫生健康共同体释义及其构建的理由[J]. 广西大学学报（哲学社会科学版），2023，45(5):135-144.

与依法治理相结合的原则，对生成式人工智能服务实施包容审慎和分类分级监管。截至 2024 年，国家网信办已先后发布七批深度合成服务算法的备案信息。其中备案项目达 1919 项，涉及 1239 个备案主体。这表明我国在人工智能领域的监管工作正不断深入。但在全球卫生治理的大背景下，这些国内法规在跨国数据流动、国际协同监管等方面存在一定局限性，还需我们进一步探索国际合作与规则协调的有效路径。⁸

（二）监管主体的不明确

在全球卫生治理进程里，人工智能算法监管暴露出监管主体不明确的突出难题。全球卫生治理的范畴极为广泛，涵盖众多国家以及国际组织，而在人工智能算法监管环节，各监管主体间的职责划分与工作分工处于模糊状态，这无疑为监管重叠和监管漏洞的产生埋下隐患。

从国际组织的视角来考察，世界卫生组织(WHO)、国际电信联盟(ITU)以及国际标准化组织(ISO)均有所行动，但三个国际组织在人工智能与卫生的相关工作上各有千秋：ITU 积极推进信息通信技术应用至卫生领域的发展，但在人工智能算法的监管方面缺乏与 WHO 的协调合作的机制，导致在技术标准出台以及相关监管的执行方面存在脱节；ISO 出台相关标准用于规范人工智能用于健康。但是，在标准的全球推广以及执行方面，缺乏有力的监督，各个国家对于标准的采纳程度不一样。在实际工作中，三者相互交叉，缺乏明确的工作分工合作机制，监管起来难免会出现脱节的现象，导致监管主体职责的模糊。⁹

以远程医疗中人工智能算法应用情形来看，世界卫生组织（WHO）着重考量医疗服务的安全性及有效性，国际电信联盟（ITU）则更偏向聚焦通信技术标准。因二者缺少协调，致使远程医疗算法在技术标准制定和医疗服务监管执行上出现不一致状况。科技公司作为主要的人工智能算法开发者与使用者，在监管体系里地位重要。但当前其参与监管方式和应担责任不明，导致监管存在漏洞，难以形成统一有效的监管合力。

从各国政府视角看，不同国家的监管重点不同。发达国家多着眼于技术创新与产业发展，监管宽松；发展中国家由于受技术、资源限制，在监管过程中多是勉力支撑，难以在全球卫生治理的人工智能算法中构筑一个有效监管体系，况且算法本身的复杂专业导致了监管主体的技术能力及资源不足，使得监管也难以发挥作用。2023 年，WHO 颁发的《人工智能医疗健康监管报告》指出医疗领域人工智能的发展挑战，列出六大监管领域，同时提出了重视透明度、文本记录以及风险管理的重要性，但目前全球卫生治理中人工智能算法的监管仍存在大量法律

⁸ 参见张蕾. 网络化治理与东盟区域卫生合作的延展[J]. 东南亚研究, 2024(2):25-44+154.

⁹ 参见刘诚, 夏杰长. 线上市场、数字平台与资源配置效率：价格机制与数据机制的作用[J]. 中国工业经济, 2023(7):84-102.

空白、规则不协调的问题，难以有效面对新的挑战。¹⁰

（三）责任界定的模糊性

在全球卫生治理体系下，人工智能算法监管存在着一系列法律困境，其中责任界定模糊这一难题尤为突出。

在技术维度上，人工智能算法缺乏透明性的“黑箱”性质，导致人工智能算法所做出的判断难以对其作出回应和展开追责，一旦算法损害事件发生，就不可避免地进入责任不明的状态。实践中，在全球卫生治理中，人工智能算法在医疗领域被应用，导致损害人类健康权益，例如在医疗领域的诊断错误、药品致病等问题屡见不鲜，而由于责任不明，难以识别具体的行为主体担责。

例如 2023 年某外资药企事件，药企利用算法锁定药靶，科研机构根据结果设计临床实验，众多受试者发生严重不良反应。追根溯源发现算法对基因数据进行加工时对罕见的基因变异误判，导致错靶。药企、科研院所互相甩锅，这是在全球健康治理中的合作中，人工智能算法问责权责不清亟待解决的问题，需要有明确的权责划分标准。¹¹

2021—2023 年，IBA 对涉及人工智能医疗算法的纠纷案件开展专题研究后明确指出，40%的案件（大致因难以准确定位侵权责任主体），使得遭受侵害的患者在寻求救济的道路上步履维艰。一些患者苦等多年而仍未获得赔款与救济的公平正义，陷入了经济和健康上双重不的窘境。例如，在某跨国远程医疗项目中，一家欧洲科技公司的人工智能诊断算法，用于南美洲医疗机构远程进行皮肤病诊断。一位患者依据该人工智能给出的诊断结果开始治疗，然而患者的病情非但没有好转，反倒恶化。经过调查发现，作为人工智能诊断算法研发者的这家欧洲科技公司，以及相关医疗机构和数据供应企业在使用算法时都有可能出现地域差异，这就使得责任主体的界定变得复杂。由于各个主体无法明确各自应承担的责任，这就给受害者带来了极大麻烦，他们在追索权益时往往得不到预期的结果。

根据具有权威性的跨国法研究机构统计显示，2020—2023 年的全球人工智能医疗算法责任纠纷案件激增，且在大量的案件中，人工智能算法的责任认定困难的案件比率也逐渐增高。2023 年，全球在该类案件中的责任认定困难争议比率高达 75%，充分说明全球范围内因人工智能算法导致的难以进行责任划分的现象存在严重的不足，严重伤害了患者的权益，严重破坏了全球医疗的信任体系。

2024 年，某国著名医疗研究机构参与的一项国际医疗合作研发出了人工智能算法，用以辅助确诊罕见病。该算法融合了来自若干参与国的患者的基因信息、临床症状以及之前的诊断信息等海量数据。后将该算法用于患者实际诊疗，一位患者的诊断结果却遭到该算法误判，并被安排错误治疗，导致严重身体损伤。经

¹⁰ 参见唐要家，唐春晖. 基于风险的人工智能监管治理[J]. 社会科学辑刊，2022(1):114-124+209.

¹¹ 参见 车丕照. 中国的法律大国地位问题[J]. 上海政法学院学报（法治论丛），2023，38(4):1-12.

核实，该算法开发团队对于如何处理来自不同国家的不同数据格式以及各数据标准出现误差，并没有良好的解决方案，另外在模型训练中对于一些罕见病特性的学习也不够精准。至于责任追究的问题，各方当事人争议较大，人工智能算法开发团队认为是由于数据格式复杂、质量参差不齐致使算法出现误诊，数据提供方则认为是算法设计有误，参与诊疗项目的医院则表示依照算法的推荐进行诊疗，不清楚自身的责任所在。上述案例充分说明人工智能算法责任界定不明确的问题急需得到改善。¹²

根据最新国际医疗纠纷研究报告，在 2024 年上半年由人工智能医疗算法引发的责任纠纷案件数量相较于 2023 年同期增长了 20%。其中责任认定困难的案件占比高达 73%，。这些数据和案例有力表明责任界定模糊严重损害患者权益，并阻碍全球卫生治理中人工智能技术健康发展。这促使国际社会加快行动，建立明确责任界定机制，以此保障全球卫生治理公正有效。¹³

结合当前全球卫生治理的实际情况，算法在人工智能的应用中会越来越频繁，更多样化的算法应用案例不断出现，这就愈发急切了，即要通过努力打造责任界定机制的清晰性，各国必须通力合作，建立共识的责任判定规范，避免在产生问题以后各方互相推脱责任的情况出现，从而最终能够保障受害者对患者的权益。¹⁴实际案例中，由于算法可解释的决策难以追责，受害者的求偿困难，清晰的责任界定意义更得以显现。

（四）数据隐私保护的不足

在全球卫生治理体系下，人工智能算法的运行依赖于海量的数据作为支撑，而这些数据之中常常包含个人的隐私信息。数据隐私保护在全球卫生治理领域占据着重要地位，然而当下的法律规定以及技术手段在数据隐私保护层面存在着明显的不足。

再从业务流程上看，有采用 AI 算法的产品进行数据采集存在过度采集的数据。2024 年，全球数字隐私报告在对全球下载量领先的 5000 个人工智能健康应用程序进行分析时发现，超过 70% 的人工智能健康应用程序在用户同意数据采集后，存在过度采集数据的现象。例如，一款号称可以提供简单的健康建议的应用，要求采集用户的身高、体质量、血压等核心健康参数，同时还采集浏览历史信息、通讯录信息等健康建议功能无关数据，并存在用户隐私泄露的现象。与此同时，数据采集时，仅有 30% 的应用可以用消费者能理解的方式向用户说明数据用途和共享对象，导致用户在数据被采集过程毫无知觉，其个人数据有可能被滥

¹² 参见江时学. 论构建普遍安全的世界：必要性、路径及中国的贡献[J]. 亚太安全与海洋研究, 2024(6):1-19+133.

¹³ 参见许嘉驿, 陈卓, 江映珠, 等. 人工智能医学软件产品研发工艺及监管政策研究[J]. 中国医疗器械信息, 2022, 28(21):5-7+14.

¹⁴ 参见 张凌寒. 算法评估制度如何在平台问责中发挥作用[J]. 上海政法学院学报（法治论丛）, 2021, 36(3):45-57.

用。

存储环节,医疗相关机构以及科技企业采用的技术类数据存储设施存在数据安全漏洞,相关加密保护等技术防范措施不到位,增加了数据失窃风险。Verizon 公司 2023 年公布的《数据泄露调查报告》指出,35%的卫生相关数据泄露事件与存储系统存在安全漏洞有关。使用环节,数据在不同的使用者之间流转时没有相应的授权和监督机制。IAPP 的调查显示,全球约 60%的人工智能类健康管理应用和相关数据获取过程中,并没有就数据的具体使用和流入情况明确告知用户。一项针对 500 家机构在处理和使用医疗大数据时采用人工智能算法的情况的调查中,不到 40%的机构形成了完整的数据使用授权和监督机制,说明数据使用环节规范缺失的情况较普遍。¹⁵

种种迹象表明,全球卫生治理中数据隐私保护在各个环节都存在不足,形势极为严峻,亟待解决。

二、全球卫生治理中人工智能算法监管的困境分析

当前世界卫生事务治理面临着复杂局面,人工智能技术发展十分迅速,发展快于法律条文的制定和完善,导致了法律在面对新技术在人工智能下所带来的挑战时,会出现一定程度上的滞后。人工智能算法呈现出较强的复杂性和技术创新性,在一定程度上导致了现有的法律条文无法具体规定权利义务,进一步加大对监管的复杂难度,同时加上技术的更新变化快,技术方面的监督部门在技术层面无法实现紧跟趋势,使得其算法难以得到监管评估的合理规范。

科技企业全速创新和追求经济效益所决定的关注点高度倾向于算法技术性能和效率提升上,相对较少关注甚至在某种程度上忽视算法技术安全性、公正性。在利益引导下的创新很容易使算法产生偏见、歧视等一系列负面影响并损害公众健康新权利。同时,在人工智能算法市场竞争剧烈的驱动下,某些企业以谋求商业利益为由采取垄断数据、算法操控等市场不正当竞争行为会给全球卫生治理埋下潜在的风险。

从宏观角度来看,全球卫生治理是个由很多国家、国际组织共同参与的国际合作的过程。针对人工智能算法的治理,各国法律制度、监管模式显著不同,造成人工智能算法监管国际合作之路难以走远。数据跨境流动限制、监管标准不统一、追责困难成为全球合作的梗阻。世界卫生组织(WHO)在 2023 年发表《人工智能医疗健康监管报告》,提出卫生领域人工智能的监管方向有 6 大项,分别是:透明度和文本记录、风险管理和创新证据等。但各国因为人工智能发展状况、监管理念、法律制度的不同,在落实 WHO 监管框架方面的深度参差不齐,难以实现有效的协同监管。《国际卫生条例(2005)》修正案对于全球突发公共卫生事件

¹⁵ 参见张宇晖. 算法决策的结构化问责机制: 欧盟经验与中国借鉴[J]. 上海政法学院学报(法治论丛), 2022, 37(4):125-143.

的应对突破虽然有所创新，但是对于人工智能算法监管的合作治理却是留白，且缔约国之间应对突发大流行病的合作机制，同样无法照搬用于人工智能算法监管的国际合作。例如“大流行协定”的磋商，因为对新冠疫苗的分配不满，英国等国家对协定报以冷淡态度，这清晰地反映出了全球卫生治理相关事务上各国立场的差异，而在人工智能算法监管的国际合作过程中，这种差异可能加剧，合作之路艰险。

三、全球卫生治理中人工智能算法监管的突破对策

（一）完善法律框架

加强国际协调、构建人工智能算法监管的国际规则是重中之重。在当前全球卫生治理的大背景下，应当建立和完善人工智能算法监管的国际规则，即全球的通用统一的人工智能算法监管规则，对人工智能算法监管的目标、原则和具体路径进行规定。

因此需要鼓励各国积极开展国内立法，完善与人工智能算法监管相关的国内立法。人工智能算法监管相关领域的立法可以结合人工智能算法的性质和风险特征，对算法开发者、使用者和监管者的权利义务进行明确，强调和加强对算法的透明性、可解释性和公正性等相关要求，以筑牢算法监管的法治基础。例如，我国制定《生物安全法》，是统领包括数据安全在内的各个相关领域的生物安全领域的基本法律，它系统明确了生物安全的基本概念、基本原则和基本制度，我国人工智能算法监管国内立法可以结合此系统性立法思路，对算法监管进行立法。当前，我国人工智能领域虽然有相关立法，但相对零散，比如数据保护方面的规定，针对同一问题不同部门的规定侧重点各有不同，有些方面缺乏协调。美国立法主要考虑的是产业发展和创新，欧盟主要考虑的是人权与隐私保护，中国应该在结合本国国情和产业与社会发展需求的前提下，在立法中统筹好算法监管和算法发展，保障人工智能算法在促进国际卫生治理上的广泛运用，同时能够防控好其风险。此外，随着人工智能技术不断发展，现有法律法规可能出现不适应的情况，各国需及时修订完善，以有效应对新技术带来的挑战。

（二）明确监管主体与界定责任

在跨境卫生治理层面，人工智能算法监管同样需要多元化的主体协同治理。应该建立起监管的多方参与机制，对参与各方的责任和义务进行明确，加强信息共享和沟通协调，形成协作监管局面。相关政府机构应加强主导调控，制定相应的政策和法律制度，对于企业的监管指导加大力度，加强其研发监管制度提升算法监管的效率。行业协会进行行业自律，建立制度和标准，规范成员遵守法律法规，做好算法风险的管控。在算法对公众健康造成伤害的情况下，有必要建立责

任追究制度，并将责任主体定位准确依法问责，加大违法成本。算法需增强算法易解释性和透明度，代码中开发者给出解释，监管机构应加强对监管质量的审查，当然，引入第三方独立审查机构对风险进行评估与提出建议。

（三）加强数据隐私保护

立足于全球的角度来看，各国都应推进数据保护立法进程，细化数据全生命周期规范，以加强对数据主体权利的保障。互联网企业同时需要增加在技术上的研发力度，通过加密、访问控制等方式来保障数据安全，加强管理控制、及时排除数据泄露的风险，各国还要加强国际交流，共同应对挑战。

（四）促进国际合作

在全球卫生治理里人工智能算法监管的国际规则制定进程中，中国应当主动且积极地投身其中。凭借自身的思考与实践，提出贴合实际且具有建设性的主张和建议，为全球卫生治理迈向法治化贡献独特的中国智慧。参考《国际卫生条例（2005）》修正案通过时各国踊跃参与并发表意见的模式，中国可联合其他国家，在人工智能算法监管国际规则制定里共同发声。比如在数据跨境流动规则制定方面，结合我国数据保护与利用的实践经验，给出兼顾隐私保护与技术发展的方案。同时，积极推动设立类似《国际卫生条例（2005）》中“实施和履约委员会”的国际协调机构，以保障规则的有效施行。

在对外合作方面，中国可以更加积极参与国际合作，在人工智能算法方面的监管上要与他国进行沟通与交流，积极地分享他国在监管方面的经验和技巧，共同解决好全球卫生治理问题，努力参与到国际合作组织中去，促进全球卫生治理体系的改革和建设，反对强权主义，维护各国共同利益，为发展中国家提供更多援助，提高其参与能力，例如 2023 年世卫组织《人工智能医疗健康监管报告》。

16

四、结论

本文侧重探讨全球卫生治理下的人工智能算法监管问题，分析存在困境并提出对策。从法律规范角度，由于缺乏国际统一标准，国内法律不仅要构建完善，而且要与国际接轨；从监管主体角度看，监管主体责任不清晰，需要构建联动监管机制；主体责任不清，致使损害事件调节难度提高；从隐私数据角度来看，相关环节均存在缺陷。对于问题建议完善制度、厘清主体、定义责任、隐私保护和国际协作，有利于提升当下监管水平。

然而，研究较为片面，忽视了跨国司法协助的具体措施研究，在跨国纠纷中难以保证权益；缺乏算法解释性技术标准，监管缺乏精准指南；不重视发展中国

¹⁶ 参见张钰宁，林登辉，许恒. 生成式人工智能对金融监管框架的重构：技术、法律与市场的三维分析[J]. 金融市场研究，2024(6):42-54.

家差异化诉求，不利于全球协同治理。今后研究应注意结合案例经验研究动态监管工具，研究新兴技术的可能性，加强跨学科之间的交流与合作，建构全球治理框架，以应对挑战。

2025年中国国际法学会年会

联合国框架下的深度伪造犯罪治理与中国因应

杨翔宇¹

湖南师范大学法学院 长沙 410081

摘要：深度伪造技术的快速发展催生了具有跨国化、难识别性及侵害法益多样性特征的新型犯罪——利用深度伪造技术实施的犯罪。基于网络主权与国际礼让原则，当前国内法规制不足与国际治理碎片化的现状亟需联合国主导下的深度伪造犯罪合作治理框架形成。然深度伪造犯罪的国际法规制面临三重困境：传统刑事司法程序不适用深度伪造犯罪跨国化新发展，管辖权冲突多发且跨国电子取证受阻；深度伪造犯罪规制难以达成国际合作共识，各国法律制度与规制理念差异较大且缺乏有效的国际合作机制，《联合国打击网络犯罪公约》制定过程中大国博弈激烈。为此，应通过实体整合、程序完善以及全球共治强化的多层级进路，实现对跨国深度伪造犯罪的有效规制。

关键词：深度伪造；跨国犯罪治理；国际合作；《联合国打击网络犯罪公约》；数据跨境流动

人工智能在大数据时代的发展中实现了深度学习的突破，进一步催生了深度伪造技术的瀑布式发展。深度伪造技术以深度学习模型为基础，借助生成式对抗网络优化编码模型，实现从原始图像到目标图像的人脸信息高仿真处理。但科技的发展使深度伪造技术的适用变得低成本低门槛，深度伪造犯罪层出不穷，也逐渐演变出跨国化发展趋势。深度伪造欺诈攻击在 2023 年增加了 3000%，AI 换脸等合成身份欺诈行为在诸多跨国科技应用场景中频繁发生²。类似事件频繁发生于全球网络空间中，深度伪造犯罪已经成为全球犯罪治理共同面对的课题，亟需加强跨国深度伪造犯罪治理的国际合作，但目前国际层面暂无促进跨国深度伪造犯罪治理各国合作的有效框架。2024 年 12 月 24 日，联合国大会通过了谈判历时五年之久的《联合国打击网络犯罪公约》（以下简称《打击网络犯罪公约》）；加强对打击网络犯罪的国际合作。本文旨在证成联合国框架规制跨国深度伪造犯罪的基础上，揭示目前国际治理框架在应对跨国深度伪造犯罪时面临的困境，后结合《打击网络犯罪公约》提出在联合国框架下加强跨国深度伪造犯罪治理的中

1 作者简介：杨翔宇，女，汉族，江西瑞金，湖南师范大学 2024 级国际法学硕士研究生（硕士在读，研一），研究方向：网络空间法、国际法。E-mail: yxy1416415804@163.com。

2 《红色警戒！深度伪造欺诈蔓延全球，ADVANCE.AI 助力出海企业反欺诈新升级》，<https://blog.advance.ai/zh-cn/deepfake-tech-fraud>，最后访问日期 2025 年 4 月 10 日。

国因应。

一、深度伪造犯罪之基础概念厘清

（一）深度伪造犯罪概念

由于深度伪造犯罪是在网络空间中滋生的犯罪，首先需要对深度伪造犯罪的上位概念网络犯罪进行明确。国内外普遍认同对网络犯罪的类型划分，包括将信息网络作为对象的犯罪行为和利用信息网络作为作案工具的犯罪行为。³科学技术更新带动网络发展，丰富了网络犯罪的外延。于志刚⁴从犯罪行为方式演变的视角出发提出了网络犯罪还存在第三种基本类型，即网络作为犯罪空间；李晓明、李文吉⁵在重新界定网络犯罪概念时也明确“以互联网为工具或以互联网为空间”。在网络空间时代的新背景下，不少学者对网络犯罪的概念进行了重新明确。

深度伪造犯罪是由深度伪造技术衍生出的不法犯罪行为，也即行为人利用深度伪造技术实施的犯罪。我国刑法并未直接规定深度伪造犯罪，但行为人利用深度伪造技术实施犯罪的行为可以被纳入刑法分则对应罪名进行处罚。这决定了深度伪造犯罪并不是一个如刑法分则其他犯罪一样，能够从四要件进行明确定义的犯罪概念；而是笔者为研究便利，对利用深度伪造技术实施的犯罪所进行的简要概括提炼。深度伪造犯罪作为本文研究对象，仍应当对其概念及范围进行具体把握。在技术层面，深度伪造技术通常指利用人工智能生成式算法特别是深度学习技术，篡改或合成原始图像而生成虚假信息的过程。深度伪造技术作为计算机科学技术，是互联网时代人工智能发展的产物，依托于网络空间得以运行，同样具备网络虚拟性、隐匿性、传播范围广泛等特征。因此深度伪造犯罪是行为人违反刑法规定，利用深度伪造技术，以互联网为工具或将互联网作为实施空间的行为。

本文的研究对象是跨国深度伪造犯罪。跨国深度伪造犯罪涉及运用深度伪造技术，通过电子手段对网络系统、数据信息以及网络服务进行的犯罪活动，这些活动跨越了国家或边境。其主要形式包括：从境外对境内发起的犯罪（网外人内型）、从境内对境外发起的犯罪（网内人外型），以及涉及境外人员和资源的犯罪（人网皆外型）⁶。其跨国（境）的性质决定了这类犯罪活动将牵涉到多个国家和地区的法律体系。然而，不同国家对于同一犯罪行为的定罪标准和量刑规定可能存在显著差异，国际层面也尚未形成具备普遍共识的解决方案，导致在实际操作中加大了打击滥用深度伪造技术实施的犯罪行为的难度。

（二）深度伪造犯罪特征

1. 跨国化

3 刘宪权. 网络犯罪的刑法应对新理念[J].政治与法律, 2016(9):4.

4 于志刚. 网络思维的演变与网络犯罪的制裁思路[J].中外法学, 2014(4):1051.

5 李晓明、李文吉. 跨国网络犯罪刑事管辖权解析[J].苏州大学学报（哲学社会科学版）, 2018(1):91.

6 楼伯坤、李想. 跨国网络犯罪刑事管辖权：困境、成因与破解[J].法治研究, 2024(6):43.

网络在现实世界之外构建了模糊国界的数字虚拟空间。现实世界的传统刑事犯罪往往受到地域限制，而互联网的跨国性本质使得相伴而生的深度伪造犯罪等网络犯罪也呈现出跨国化发展趋势，消解了传统犯罪的地域边界。网络平台具有交互的信息传导模式，利用深度伪造技术生成的数据通过互联网用户的交互参与可以在短时间内实现世界范围内的传输⁷。同时，网络平台数据流动的匿名化加之深度伪造犯罪的跨国化趋势，使得追踪和定位跨国深度伪造犯罪源头的难度增大，给执法机构打击此类犯罪带来了巨大挑战。由于各国对深度伪造犯罪的法律规制不尽相同，在处理跨国深度伪造犯罪时往往存在法律适用的争议，常引发各国间的管辖权冲突。

2. 难识别性

深度伪造技术的核心是深度学习技术和生成式对抗网络模型，在生成式对抗网络的作用下，生成器与鉴别器运用的算法互相改进，实现原始图像与目标图像之间的不断学习、验证，从而生成具有高仿真度的深度伪造生成物⁸。且深度伪造识别技术的发展相较于深度伪造技术本身始终具有滞后性，目前未有网络平台探索研发出有效应对滥用深度伪造技术行为的检测手段。同时由于人工智能的运行机制具有复杂性及不透明性的特征，使得各国缺乏对于新兴技术的监管应对能力，在现有制度上难以实现有效监管⁹。如 Facebook 曾于 2019 年举办深度伪造检测挑战赛，参赛的 3.5 万种深度伪造技术检测算法最终达到的准确率最高也仅 65.18%¹⁰。由于深度伪造技术的高仿真性，生成的虚假信息难以被普通用户识别，也即跨国深度伪造犯罪行为难以被识别，从而增加了深度伪造技术侵害法益的可能性。如何有效识别和防范跨国深度伪造犯罪成为各国与国际社会亟需解决的问题。

3. 侵害法益多样性

深度伪造犯罪是利用深度伪造技术实施的犯罪，深度伪造技术仅仅是犯罪的工具；同时深度伪造犯罪常发生于多类型场景中，侵害的法益种类涵盖个人利益、社会秩序、国家主权与安全、国际稳定等。所以，我国刑事法律体系选择通过危害行为及侵害法益具体判断构成刑法分则的何种犯罪，强调对深度伪造犯罪行为的实质危害进行评估，而不是根据深度伪造技术本身。例如，当行为人利用网络传播深度伪造技术生成的虚假信息，诱使受害人做出本不会采取的行动，从而导致损失时，可能涉嫌构成诈骗罪。2024 年初，一家跨国设计工程公司的香港分公

7 曾磊、王争辉. 网络主权视域下网络犯罪国际合作治理的中国路径——以<联合国打击网络犯罪公约>的起草为契机[J].信息安全与通信保密,2023(2):143.

8 王禄生. 论“深度伪造”智能技术的一体化规制[J].东方法学,2019(6):59.

9 张永忠. 论人工智能透明度原则的法治化实现[J].政法论丛,2024(2):124.

10 《Deepfake 识别赛排名出炉！3.5 万种算法准确率最高 82.56%，Facebook 要在 CVPR 上“放大招”》，<https://baijiahao.baidu.com/s?id=1669371480309256536&wfr=spider&for=pc>，最后访问日期 2025 年 4 月 10 日。

司员工就遭遇了深度伪造技术的骗局，结果导致公司财务损失高达 1.95 亿港元¹¹。

二、在联合国框架下规制跨国深度伪造犯罪之证成

（一）理念基础

1. 网络主权原则

主权是国家在其管辖范围内享有的独立的自治自决权，国家主权原则强调国际社会交往中各国对彼此主权的尊重。网络空间时代主权的范围延伸到虚拟空间——网络，即网络主权¹²。中俄等国在 2011 年向联合国大会递交的《信息安全国际行为准则（草案）》（以下简称“《信息安全准则草案》”）主张应在网络空间治理合作中使用网络主权这一概念¹³。其中《信息安全准则草案》第 6 条也首次提出与互联网有关的公共政策决策权来自各个国家的主权¹⁴，网络主权的概念在《信息安全准则草案》中被视为是国家主权在网络空间语境下的延伸表达。北约主持编写的《可适用于网络行动的国际法的塔林手册 2.0 版》（以下简称《塔林手册 2.0》）体现了网络主权观念，在第一部分“国际网络空间法”第一章主权问题分析中就指出“国家主权原则适用于网络空间”¹⁵。这体现了中俄此前提出的网络主权概念已逐渐被其他区域性治理规则甚至国际规则接受，国际层面就互联网时代国家主权向网络空间的延伸发展形成了广泛共识。

一国坚守网络主权和尊重他国的网络主权，是在深度伪造犯罪等网络犯罪治理国家合作中应当坚持的根本原则。这有助于维护各国的网络空间安全，也是国际社会共同应对深度伪造犯罪调整的首要前提。考虑到网络主权是现实世界国家主权“嫁接”到虚拟网络空间的产物，网络空间的独特属性使得许多传统空间犯罪治理的规则出现了适用冲突¹⁶。例如随着深度技术的快速发展以及数据流动的范围扩大，网络空间的边界模糊使得犯罪呈现出跨国界化特征，国家对于发生在网络空间的深度伪造犯罪的控制力面临挑战。我国应当在网络主权原则的指导下，加强网络空间国际法律框架搭建的沟通交流，在此基础上解决司法管辖权冲突等问题，通过国际合作和对话解决争端。

2. 国际礼让原则

11 《悉尼歌剧院、北京鸟巢设计公司被骗：香港公司损失约 1.95 亿》，

<https://news.qq.com/rain/a/20240518A047KJ00>，最后访问日期 2025 年 4 月 10 日。

12 徐凤. 网络主权与数据主权的确立与维护[J].北京社会科学,2022(7):55.

13 曾磊、王争辉. 网络主权视域下网络犯罪国际合作治理的中国路径——以<联合国打击网络犯罪公约>的起草为契机[J].信息安全与通信保密,2023(2):141.

14 张喆：《中俄推行网络准则抗衡美国：网络政策是各国主权》，

<https://news.sohu.com/20110914/n319297564.shtml>，最后访问日期 2025 年 4 月 10 日。

15 TALLINN MANUAL ON THE INTERNATIONAL LAW APPLICABLE TO CYBER WARFARE.

“SECTION1 SOVEREIGNTY, JURISDICTION, AND CONTROL , Rule1 Sovereignty”,

<https://www.onlinelibrary.iihl.org/wp-content/uploads/2021/05/2017-Tallinn-Manual-2.0.pdf>，最后访问日期 2025 年 4 月 10 日。

16 古俊峰、施涵佳. 主权原则对网络域外执法的规制[J].中南民族大学学报（人文社会科学版）,2024(2):129.

国际礼让原则经常出现于国家间法律冲突解决的应用场景中，发端于优利克·胡伯的法律冲突三原则中对礼让概念的阐述。在当时的语境下，国际礼让原则主要指的是对外国法域外效力的承认取决于各国的主权考虑¹⁷。此后随着斯托雷三原则等学说的发展，逐渐演变成对外国主权的尊重，美国维生素 C 案的重审判决就是这种对主权尊重的鲜活展示。美国第二巡回上诉法院在重审时明确指出，法律意义上的国际礼让原则并非义务、礼貌或善意；相反，它是一国在充分考虑自身国际责任以及本国公民权利的基础上，对另一国立法、行政或司法行为所表现出的尊重¹⁸。

如今跨国深度伪造犯罪治理面临的司法管辖冲突激增，法律冲突其实是国家主权冲突的体现，面对冲突各国应当秉持国际礼让原则，展现出对他国网络主权的承认与尊重。深度伪造犯罪由于其跨国化特征，经常出现两个或多个国家的法院对同一案件享有管辖权的情形。此时法院基于国际礼让原则进行礼让分析，判断各国法律互相之间对争议法律关系是否存在真正的冲突，发挥该原则的协调作用，在尊重各国网络主权的基础上作出管辖权选择。但实际上法院在面临司法管辖权冲突判决时利用国际礼让原则进行判决的次数越来越少，在涉及己国与别国的主权利益时法院跟多的偏向于国内利益依照国内法进行裁判¹⁹。在全球化的大背景下，各国主张的主权利益存在差异，加之法律体系的多样性，国际礼让原则在深度伪造犯罪的国际治理中面临诸多挑战。例如，跨国深度伪造犯罪的治理与电子取证过程紧密相关，而数据跨境流动是这一过程的关键；然而，由于各国对网络主权边界的界定不尽相同，这在进行国际司法协助时，往往会对取证过程造成阻碍。尽管如此不可否认的是，国际礼让原则仍然是确保各国在犯罪治理方面顺利进行国际协商的基础。

（二）实践需求

1. 跨国深度伪造犯罪国内法规制不全面

我国目前对深度伪造犯罪的规制主要依据刑法分则的各项罪名展开，深度伪造犯罪根据行为人的不同行为常涉及诽谤罪、帮助信息网络犯罪活动罪、侵犯公民个人信息罪等刑法分则罪名。所涉罪名均以伪造生成信息的传播数量作为传播范围的考量标准，从而定罪。但在深度伪造技术生成式算法的自动推荐功能的作用下，深度伪造信息的传播范围出现了技术性扩大。算法推荐通过对用户的互联网行为踪迹进行追踪，运用数字算法模型计算出对应用户的个人深度特征等信息，并据此推算出用户可能产生兴趣偏好的内容²⁰。深度伪造技术以其高度仿真的模

17 李双元、欧福永.国际私法（第六版）[M].北京大学出版社,2022(6):46.

18 知产财经：《附判决书美国对华反垄断第一案：华北制药重审胜诉》，https://www.ipeconomy.cn/index/news/magazine_details/id/3501.html，最后访问日期 2025 年 4 月 10 日。

19 吴琦. 网络空间中的司法管辖权冲突与解决方案[J].西南政法大学学报,2021(1):59.

20 陈昌凤、石泽. 技术与价值的理性交往：人工智能时代信息传播——算法推荐中工具理性与价值理性的思考[J].新闻战线,2017(17):71.

拟生成能力能够生成难以识别或检测的虚假信息，又在算法推荐机制的加成下更容易能够达到入罪标准。但此时深度伪造生成物的传播对法益造成的侵害显然无法与相同传播次数下传统案件所造成的严重危害后果等同。且目前深度伪造犯罪出现跨国化发展趋势，国内法规制体系的现有规定并不能很好适配跨国深度伪造犯罪的规制。在我国国内法规制不尽如人意的情况下，可以将规制深度伪造犯罪的目光扩大到国际法领域，更要发挥联合国等国际组织促进各国交流合作的作用。通过国际条约等国际法统一治理框架的形成，推动各国在深度伪造犯罪立法、执法以及司法等方面的合作；同时，我国应积极参与国际社会层面的技术交流与数据共享，提高深度伪造生成物识别技术与应对能力，以此促进我国国内法治治理体系的不断优化，实现对跨国深度伪造犯罪更有效地打击。

2. 跨国深度伪造犯罪国际治理碎片化

深度伪造犯罪尚未形成国际统一治理框架，缺乏在全球范围内各国达成普遍共识的有效国际公约，相关合作仅在制定签署了区域性条约的特定范围国家之间展开；在《打击网络犯罪公约》生效前，联合国的主导协调作用也未得到发挥。深度伪造技术的非正当运用会导致隐私泄露等负面反馈，但在数据智能社会背景下，全球范围内治理人工智能伦理的国际硬法缺位、国际软法主体间合作不稳定，跨国深度伪造犯罪的国际治理呈现出碎片化的治理趋势²¹。各国在展开国际深度伪造犯罪治理合作时，首先考虑的对象是与所在地区的区域性国际组织或利益诉求相同、法律规制理念相近的国家。目前最有影响力的为 2001 年欧洲委员会于布达佩斯制定的《网络犯罪公约》（以下简称《布达佩斯公约》）；该公约的成员国主要来自欧洲委员会这一区域性政府间组织，同时美国、澳大利亚等国也加入了该组织，构成了英美等发达国家主导的区域性犯罪治理框架。这些规则仅在各自缔约国的有限范围内发生效用，难以覆盖全球范围内的所有国家，在跨国深度伪造犯罪涉及两个不处于同一条约体系内国家的情况下无法作出回应，导致在打击跨国深度伪造犯罪时经常产生司法管辖权冲突以及执法管辖权的行使冲突。从《布达佩斯公约》的制定国家与成员国可以看出该公约为了维护欧美等发达国家的利益，而缺乏对广大发展中国家的价值追求和利益主张的考量，在如今全球文化多元化的环境与国际合作的背景下显然有失公允²²。在这样的背景下，更需要国际社会共同努力构建一个联合国框架下的全球性的跨国跨国深度伪造犯罪治理体系。

三、跨国深度伪造犯罪的国际法规制困境

（一）传统刑事司法程序不适应犯罪跨国新发展

1. 跨国深度伪造犯罪管辖权冲突多发

21 朱明婷、徐崇利. 人工智能伦理的国际软法之治：现状、挑战与对策[J].政策与管理研究,2023(7):1038.

22 徐展鹏、丁丽柏. 网络犯罪治理国际合作：发展趋势、全球协作与中国方案[J].学术论坛,2024(2):126.

管辖权冲突是跨国深度伪造犯罪刑事管辖权面临的典型问题,深度伪造犯罪的跨国化发展对传统刑事司法管辖权产生了冲击。例如深度伪造犯罪的犯罪策划地、行为地以及危害结果发生地可能处于三个不同国家,这样多国对于同一刑事案件都有权主张管辖权,进而形成了管辖权冲突。同时,冲突的存在给了犯罪分子选择管辖权的可乘之机,趁虚而入钻法律漏洞,不利于我国惩治违法犯罪行为根本目的的实现。

我国传统刑事犯罪管辖权原则包括属地管辖原则、属人管辖原则、保护管辖原则与普遍性管辖原则。若行为人在犯罪事实过程中利用位于我国境内的网络服务器或计算机终端等设备实施了犯罪行为,亦或是深度伪造技术生成物造成的危害结果全部或部分发生在我国境内,则我国根据属地原则依法具有对跨国深度伪造犯罪行为的管辖权。但虚拟的网络空间是跨国深度伪造犯罪的滋生地,世界各国包括我国目前的属地管辖则主要聚焦于实在的传统物理空间,传统属地原则对网络空间的深度伪造犯罪是否能够直接适用并形成有效规制?同时跨国深度伪造犯罪的行为人与受害者之间被网络空间隔开,属人管辖原则的行使也受到局限。深度技术操作者作为藏身于网络空间背后的潜在“幕后黑手”,能够轻易在本国对他国及其公民实施严重危害他国公民权益、社会安定或国家主权的行为;同时各国针对深度伪造犯罪的法律规制差异较大,属人原则难以发挥其效用,受害国因行为人的行为未被其本国法定性为犯罪而面临管辖权的缺失。考虑到法律规制差异的原因,如在满足双方可罚原则的情况下保护管辖也面临适用难题;同时根据前文分析,由于暂且缺乏世界范围内的有效国际公约导致普遍管辖也难以发挥效用。

面对现有管辖权规定无法解决国际层面面临的各国管辖权冲突的问题,涌现了“实害联系原则”“网络地址来源国原则”“有限管辖原则”等许多跳脱出原有框架的管辖权原则的新理论。不可否认这些新理论为解决跨国深度伪造犯罪的管辖权冲突提供了新的思路。有限管辖原则强调在尊重各国主权的基础上限制管辖权的行使;实害联系原则强调要根据对现实生活的实际损害作出是否具有管辖权的判断;“网络地址来源国原则”则聚焦于网址这一稳定因素来确定管辖²³。但这些理论实质上仍然是传统管辖理论扩张化趋势的体现,无法彻底解决跨国管辖权冲突²⁴。

因此,为了有效应对跨国深度伪造犯罪带来的挑战,我国迫切需要对现行管辖权规定进行完善,国际社会尤其是联合国应发挥主导作用促进国际层面形成明确、统一且具有普遍适用性的管辖权规则。这些规则应当具备跨法域的兼容性,确保在全球范围内对这类犯罪行为进行有效监管和打击。

23 姜源丰. 网络犯罪刑事管辖权研究[D]. 中国人民公安大学硕士学位论文,2019:15-18.

24 李晓、李文吉. 跨国网络犯罪刑事管辖权解析[J]. 苏州大学学报(哲学社会科学版),2018(1):93.

2. 犯罪治理的国际数据流动机制缺失，电子取证受阻

跨国深度伪造犯罪多发生于网络空间，电子证据是主要证据类型。我国 2012 年将电子证据规定为法定证据类型之一，刑事诉讼定罪量刑过程中证据是关键，然而电子证据在如今跨国犯罪治理的环境中却面临着取证受阻的困难。电子证据属于数据，具有隐蔽性、易销毁性、分布广泛性等不同于传统实物证据的特征；证明跨国深度伪造犯罪的难点在于电子证据很容易被更改、操纵和删除²⁵。

跨国深度伪造犯罪治理过程中收集电子数据经常涉及数据跨境流动，数据跨境流动需要遵循一定的规则。但目前国际层面并未形成数据跨境流动的统一做法，各国关于数据跨国均形成了自己的一套模式，较为典型的有美国的“数据控制者模式”以及欧盟的“数据灵活出境模式”等。美国的数据跨国规制模式体现了其遵循的数据自由流动理念，在保障数据安全的前提下通过流动最大限度地释放数据价值，通过与其他国家间签订双边或多边协定以及构建数据流动框架来实现对数据跨国的规制。但美国的霸权主义与长臂管辖对其他主权国家的数据规制权利以及数据主权造成威胁。美国虽然在国际上呈现出数据自由流动奉行者的形象，其积极接受数据流入又严格限制数据流出、将数据控制存储于本国境内的行为在 2018 年《澄清域外合法使用数据法案》（Cloud Act）上得到完美体现；同时跨国隐私规则体系（CBPR）强制各加入国在数据跨国时对美国较低标准的数据保护水平予以认同，达到吸引数据流入美国本土的目的²⁶。而欧盟对于数据在成员国之间以及与非成员国之间的跨国流动进行了灵活的区分规制。欧盟利用成员国之间的紧密联系形成了内部高度自由开放的数据交流共享模式，电子数据在欧盟内部能够轻松实现自由流动²⁷；但是在我国等非成员国在跨国深度伪造犯罪治理过程中需要与欧盟展开司法协助、进行跨国电子数据取证时，则会因为欧盟规定的“充分性保护”认定标准等高要求而受到阻碍²⁸。也即在国际司法协助程序开展过程中，容易出现部分国家拒绝申请国对部分电子证据的调取请求。若我国作为调出国，则需要严格按照我国数据出境程序审核；若我国作为调入国，则需要满足他国关于数据保护的一系列要求。

此外，一些国家根据其国情形成了在主权与安全等的方面的特殊化需求，这些国家的国内法可能会基于对人权保障、数据安全、地缘政治以及经济贸易等因素的考量而限制部分数据的跨国调取与流动，对申请国提交的数据调取指令进行防御性抵制，从而造成了如今跨国深度伪造犯罪司法协助程序中取证流程的难推进现状²⁹。例如《布达佩斯公约》的解释性报告在对该公约第 18 条关于出示令内

25 Wiarti J. Legality of electronic evidence in cyber crime cases[J]. Ahmad Dahlan Indonesian Law Journal, 2023, 1(2): 11-19.

26 欧福永、范知智. 个人数据跨境流动规制中的“长臂管辖”及我国的对策[J]. 时代法学, 2023(1): 100-101.

27 元轶、纪钊洋. 大数据时代跨国电子取证的治理困境与应对[J]. 中国政法大学学报, 2024(4): 186.

28 刘金瑞. 迈向数据跨境流动的全球规制：基本关切与中国方案[J]. 行政法学研究, 2022(4): 75.

29 梁坤. 网络犯罪国际公约中的跨国调取数据条款[J]. 当代法学, 2024(2): 128.

容的解释中提到,出示令在某些情况下是涉及到人权保障数据被允许调取的前提³⁰。

在面对国际数据跨境流动机制缺失的挑战时,国际社会需要寻找新的合作模式和解决方案。联合国等国际组织应推动建立统一的数据跨境流动标准或在国际公约中规定统一的数据跨国调取条款,减少各国因数据跨境流动规范存在差异造成的电子取证困难,提高跨国深度伪造犯罪的打击成效。

(二) 跨国深度伪造犯罪规制国际合作共识难达成

1. 各国深度伪造犯罪规制法律制度与规制理念差异较大

美国与欧盟两个典型地区在规制路径上提供了不同答案,美国的深度伪造立法规制为结果导向,倾向于关注以受害者为主的后期阶段,而欧盟则倾向于关注深度伪造技术规制过程的早期阶段,以预防性的视角更多地进行技术、数据的保护³¹。美国对于深度伪造技术可能影响选举、威胁国家安全等方面高度关注,因此在相关立法中重点强调对这类行为的打击。加利福尼亚州针对深度伪造色情作品与政治选举制定了法律。2019年AB602法案与AB730法案先后得到加州政府的批准并被签署通过。两项法案从不同方面对加州深度伪造技术的运用进行了规制,AB602法案针对色情领域深度伪造视频,禁止未经当事人同意的色情活动的实施;而AB730针对政治领域的深度伪造视频,规定任何人不得出于损害候选人声誉、欺骗选民的意图散布有关候选人的深度伪造视频³²。而欧洲国家更侧重于从个人信息保护和虚假信息治理的角度来立法。基于技术本身在规制过程中的决定性作用,欧盟更加侧重于在保护技术发展的过程中实现对公民个人数据的保护。欧盟于2018年5月通过综合性法律《通用数据保护条例》(GDPR)并实施,将易成为深度伪造对象的公民的个人数据如面部信息作为一种数据纳入该条例进行保护。不少学者也认为除了立法规制以外,可以在技术运行方面通过不断地技术更新实现对深度伪造的规制。Penelope Thornton³³认为需要发挥互联网企业等平台在技术发展方面的优势;Karolina Mania³⁴提出2019年Facebook发布的“深度伪造检测挑战”以及禁止操纵人工智能的新政策等是可借鉴的技术进路。

我国刑事法律体系并未针对滥用深度伪造技术的行为治理专门立法,而是将其纳入现有刑法分则罪名中进行规制。各国基于不同理念的立法差异,导致在跨国合作中对某些滥用深度伪造技术行为的定性、量刑是存在分歧,增加了合作的

30 《关于〈网络犯罪公约〉的解释性报告》, <https://rm.coe.int/cets-185-and-expl-rep-chinese/1680a39f77>, 最后访问日期 2025 年 4 月 10 日。

31 Yinuo Geng, “Comparing” Deepfake” Regulatory Regimes in the United States, the European Union, and China”, *Geo. L. Tech. Rev*(2023):157-178.

32 《加州新法案: 未经当事人同意, 不得用 AI 换脸技术制作色情视频》, https://www.sohu.com/a/345766506_161795, 最后访问日期 2025 年 4 月 10 日。

33 Penelope Thornton, Aissatou Sylla, Patrick Fromlowitz, “The War against Deepfakes”, *Managing Intellectual Property*(2020):29-30.

34 Karolina Mania, “Legal protection of revenge and deepfake porn victims in the European Union: Findings from a comparative legal study”, *Trauma, Violence, & Abuse*(2024):1-9.

难度。

2. 缺少有效打击跨国深度伪造犯罪的国际合作机制

深度伪造犯罪所具有的跨国性以及网络空间的虚拟本质均揭示针对跨国深度伪造犯罪的治理工作亟需多个主权国家之间的协作³⁵。单一主权国家的独立行动不足以应对当前治理难题，唯有在国际社会中构建起切实可行的合作机制，方能达成对犯罪行为的有效打击。网络犯罪具有的复杂性、挑战性和高技术含量，彻底颠覆了个别国家真正解决和应对这些风险的治理能力，在国际层面形成有效的国家合作对于面对跨国网络犯罪带来的调整至关重要³⁶。

中国的统计数据表明，在调查的网络犯罪案件中，超过 90% 的诈骗、钓鱼、色情、赌博等违法网站和超过 70% 的僵尸网络控制端位于国外³⁷。目前我国在打击深度伪造犯罪的国际合作机制上存在欠缺，与其他国家在跨国取证、刑事司法协助等方面的合作频繁受阻，在联合国框架下缺乏一个统一的、权威性的国际合作平台。欧洲委员会主持制定的《布达佩斯公约》处处透露着欧美发达国家的利益需求底色，并非真正处于全球化犯罪打击的视角，缺乏对发展中国家网络犯罪规制价值取向的考虑。《布达佩斯公约》对全球其他区域化网络犯罪治理也产生了辐射影响，如非洲联盟《网络安全和个人数据保护公约》、阿拉伯国家联盟《打击信息技术犯罪公约》等均是以布达佩斯条约为蓝本，有些条款甚至直接移植³⁸。

此外，我国目前的深度伪造犯罪主要依据国内法律体系进行应对，但尚未制定出针对滥用深度伪造技术的全球性犯罪问题的国际合作治理策略。而我国与他国之间的跨国犯罪打击合作也主要是通过双边或多边的协商、签订合作协议等方式进行，如我国与俄罗斯于 2016 年签订了《关于在保障国际信息安全领域合作协定》³⁹。但这些方式缺乏系统性和稳定性，难以形成有效的合作机制。在面对复杂的跨国犯罪案件时，无法及时、高效地进行信息交流和协作办案；且对跨国深度伪造技术犯罪的打击范围仅限于参与制定条约的国家间，规制具有地域性、分散性，在国际法层面缺少统一性规范。

3. 《打击网络犯罪公约》制定协商下大国博弈激烈

其实联合国层面自 2017 年起便开始了构建国际层面网络犯罪治理框架的尝试。2024 年 8 月 9 日，联合国特设委员会经过投票一致通过了首个全球范围内的网络犯罪公约草案《打击网络犯罪公约》；并于同年 12 月 24 日在联合国大会经各国投票通过，联合国秘书长古特雷斯表示“这是 20 多年来通过谈判达成的第

35 李晓明. 跨国网络犯罪刑事管辖权解析[J]. 苏州大学学报(哲学社会科学版), 2018(1):96.

36 Chandra A, Devi S. The Need for International Co-Operation in Combating Cyber Crime[J].

37 李哲、朱晓琴. 《联合国打击网络犯罪公约》的中国立场与核心问题[J]. 北京师范大学学报(社会科学版), 2024(5):129.

38 徐展鹏、丁丽柏. 网络犯罪治理国际合作：发展趋势、全球协作与中国方案[J]. 学术论坛, 2024(2):126.

39 《中华人民共和国政府和俄罗斯联邦政府关于在保障国际信息安全领域合作协定》，
<http://treaty.mfa.gov.cn/tykfiles/20180718/1531877059893.pdf>，最后访问日期 2025 年 4 月 10 日。

一项国际刑事司法条约”⁴⁰。该公约最早可追溯至 2017 年由俄罗斯向联合国大会提交的《关于解决为犯罪目的使用信息和通信技术提案》⁴¹。

《打击网络犯罪公约》的主要推动国为中俄等国，在协商过程中欧美等西方发达国家强势地想让《布达佩斯公约》中体现其关切利益的相关内容融入草案，这也是中俄等国抵制《布达佩斯公约》的重要原因。当前各国构成了网络犯罪国际规制机制上的两大阵营，形成了以欧美等发达国家为主的霸权主义阵营与我国等广大发展中国家的主权主义阵营之间的博弈。这从最终通过的《打击网络犯罪公约》全称可见一斑，该公约全称为《联合国打击网络犯罪以及为打击使用信息通信技术系统实施的某些犯罪并共享严重犯罪电子证据而加强国际合作公约》⁴²；这是对两大阵营主张利益的艰难调和。

两大阵营的对峙焦点之一即为跨国调取电子数据相关问题。欧美等国认为应当赋予执法机关针对境内外互联网企业的调取权，以应对网络犯罪规制过程中调取电子数据的需求。该阵营借助《布达佩斯条约》第二附加协定试图在世界范围内推广执法机关与受调查国网络服务提供商直接合作的跨国取证模式，大肆开展长臂管辖弱化其他主权国家对境内数据的掌控，对他国数据主权形成威胁⁴³。但南非、伊朗等主权阵营国家认为，应当在跨国电子数据调取过程中首先考虑的应当是证据所在国的主权。我国也在联合国网络犯罪政府间专家组第五次会议中表示“各国跨国调取电子证据应尊重证据所在地的国家主权，不得采取侵入性和破坏性技术侦查手段跨国获取电子证据”⁴⁴。

但在《打击网络犯罪公约》的草拟过程中，即使主权主义阵营坚决反对欧美阵营在《布达佩斯公约》中采取的“数据控制者模式”，草案的最初合并谈判文件仍然与《布达佩斯公约》中的部分条款几乎完全相同⁴⁵。而通过《打击网络犯罪公约》最终提交至联合国大会表决的草案版本与《布达佩斯公约》的内容对比（见下表一），也不难发现草案规定的部分内容与《布达佩斯公约》一致，广大发展中国家在联合国治理网络犯罪框架搭建过程中提出的主张并未得到完全支持。

《打击网络犯罪公约》是在目前国际层面规制跨国深度伪造犯罪体系之上作出的重大改进，但仍然透露出西方发达国家的霸权主义政治色彩。在这一过程中，主

40 联合国官网：《联合国大会通过具有里程碑意义的〈打击网络犯罪公约〉》，<https://news.un.org/zh/story/2024/12/1134601>，最后访问日期 2025 年 4 月 10 日。

41 吴沈括：《〈联合国打击网络犯罪公约〉的规范涉及与制度影响》，<https://opinion.caixin.com/2024-10-08/102242866.html>，最后访问日期 2025 年 4 月 10 日。

42 联合国大会：《联合国打击网络犯罪以及为打击使用信息通信技术系统实施的某些犯罪并共享严重犯罪电子证据而加强国际合作公约》，第七十九届会议 A/RES/79/243，2024 年 12 月 31 日，<https://documents.un.org/doc/undoc/gen/n24/426/73/pdf/n2442673.pdf>，最后访问日期 2025 年 4 月 10 日。

43 方芳.坚持在联合国框架下制定电子证据国际标准——联合国毒品犯罪办公室第五届网络犯罪政府间专家组会议研究[J].信息安全与通信保密,2019(5):18.

44 联合国毒品和犯罪问题办公室：《联合国网络犯罪政府间专家组第五次会议书面评论意见》，<https://www.unodc.org/documents/Cybercrime/Chinese.pdf>，最后访问日期 2025 年 4 月 10 日。

45 李哲、朱晓琴.《联合国打击网络犯罪公约》的中国立场与核心问题[J].北京师范大学学报（社会科学版）,2024(5):131.

权阵营国家的合理关切往往被边缘化，导致公约在某些关键问题上未能充分反映全球多数国家的意愿和利益。例如，草案中关于数据控制者模式的坚持，实际上强化了数据控制者所在国的法律管辖权，而忽视了数据所在国的主权和法律权益。

	《布达佩斯公约》	《打击网络犯罪公约》
主持组织	欧洲委员会	联合国毒品和犯罪问题办公室
代表成员国	美国、英国	中国、俄罗斯
宗旨目的	开展在打击网络犯罪方面的合作	打击为犯罪目的使用信息和通信技术的行为
主要内容	第一章 术语的使用	第一章 总则
	第二章 国家一级应采取的措施：第一节 实质性刑法（网络犯罪行为）	第二章 网络犯罪的刑事定罪（非法访问、非法拦截、干扰电子数据、干扰信息和通信技术系统、滥用装置、与信息和通信技术系统有关的造假或盗窃欺诈行为、涉性（儿童）犯罪、洗钱行为）
	第二章 国家一级应采取的措施：第三节 管辖权	第三章 管辖权
	第二章 国家一级应采取的措施：第二节 程序法——（篇目2/4/5）所存储计算机数据的保存、搜查和扣押、收集	第四章 程序法（所存储计算机数据的保全、搜查和扣押、收集、拦截）
	第二章 国家一级应采取的措施：第二节 程序法——（篇目3）出示令	第四章 程序法（提交令）
	第三章 国际合作（引渡、国际司法互助）	第五章 国际合作（保护个人数据、引渡、人员移管与诉讼移交、跨境司法协助）
		第六章 预防措施
		第七章 技术援助和信息交流
	第四章 最后条款	第八章 实施机制

表一 《布达佩斯公约》与《打击网络犯罪公约》内容对比

四、联合国框架下跨国深度伪造犯罪规制的中国进路

（一）实体整合：公约入罪，构建跨国深度伪造犯罪规制体系

联合国大会文件 A/RES/79/243 即 2024 年 12 月 24 日决议中提到，决定打击利用信息和通信技术进行犯罪的全面国际公约特设委员会（以下简称“特设委员会”）定期举行会议、继续开展工作，以期谈判拟订一项《打击网络犯罪公约》补充议定书草案，特别用以酌情处理其他刑事犯罪问题⁴⁶；该决议为未来通过国际法直接规制深度伪造犯罪提供了可能性。打击网络犯罪公约》最终文本中，并未明确规定对深度伪造犯罪的直接规制。虽然根据上文分析，对深度伪造犯罪的规制可通过打击网络犯罪进行，但深度伪造犯罪自身也具有其特殊性。深度伪造犯罪作为一种新兴的犯罪形式，具有高度社会危害性与复杂性；利用深度伪造技术生成的高仿真虚假图像或音视频，常被用于实施诽谤、诈骗等犯罪或操控舆论环境、破坏社会安定或国家安全等目的。

在公约制定过程中，一些国家和国际组织提出了对深度伪造技术的关注。例

46 联合国大会：《联合国打击网络犯罪以及为打击使用信息通信技术系统实施的某些犯罪并共享严重犯罪电子证据而加强国际合作公约》，第七十九届会议 A/RES/79/243，2024 年 12 月 31 日，<https://documents.un.org/doc/undoc/gen/n24/426/73/pdf/n2442673.pdf>，最后访问日期 2025 年 4 月 10 日。

如，欧盟在于 2024 年 8 月生效的《人工智能法案》中将人工智能系统所涉风险划分为四个等级，其中可能对健康、安全或基本权利造成严重风险的 AI 应用被归为高风险，并提出了相应的规制措施⁴⁷。欧洲刑警组织创新实验室观察站的报告中详细概述了深度伪造犯罪的犯罪用途及潜在用途，并强调政策制定者需要制定政策适用不断变化的技术现实⁴⁸。此外，联合国毒品与犯罪问题办公室 2024 年的一份报告中提到，深度伪造技术可能被用于实施网络犯罪，亚洲犯罪集团已将新的基于服务的商业模式和技术（包括深度伪造）融入其运营中；指出目前涉及深度伪造的人工智能犯罪有所增加，2024 年上半年针对东南亚犯罪集团的在线平台深度伪造内容的提及增加了 600% 以上；并呼吁加强国际合作以应对这一挑战⁴⁹。

因此，在《打击网络犯罪公约》通过但未生效的背景下，公约补充议定书是联合国框架下国际社会将深度伪造犯罪纳入国际刑事法律规制框架的可取之策。

（二）程序完善：加强电子数据分级保护，促进数据跨境流动

《打击网络犯罪公约》第 27 条针对用户数据取证设置了提交令制度，与《布达佩斯公约》中出示令的规定相一致，要求各缔约国授权主管机关下令境内用户提交其拥有或控制的存储在网络介质中的指定电子数据，以及在境内提供网络服务的服务提供商提交其所拥有或控制的与此类服务有关的用户信息。但站在我国数据主权的保护立场下，相较于根据提交令或出示令的规定接受各国通过我国网络服务提供者调取电子证据，更偏向于通过国际私法协助和执法合作来优化跨国电子取证流程。数据主权是大数据时代在国家主权、网络主权等概念上的衍生概念。在坚持主权独立不容侵犯的根本理念基础上，数据主权包含国内数字空间独立管辖的硬数据主权与国际事务共同治理的软数据主权⁵⁰；因此，为解决当前跨国电子取证面临的难题，我们应当在联合国领导的国际体系和以国际法为根基的国际秩序框架内制定一套全球统一的跨国电子取证规范。结合我国的数据主权主张，还应当考虑跨国数据调取中体现的国家主权关切，对网络服务提供者所占有或掌握的数据，采用分级分类的数据出境制度，在一定程度上减少对特定类型数据的严格出境评估措施⁵¹。例如，对于一些对国家主权与数据主权、社会公共利益以及个人隐私影响相对较小的电子数据类型，在确保数据出境基本合规的前提下可以减少出境要求，从而提高数据跨境流动以及跨国电子取证的效率，促进跨

47 欧盟官网：“家-政策-人工智能-欧洲对人工智能的态度-《人工智能法案》”，<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>，最后访问日期 2025 年 4 月 10 日。

48 欧洲刑警组织：《面对现实？执法和深度伪造的挑战》，2024 年 3 月 13 日，<https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes>，最后访问日期 2025 年 4 月 10 日。

49 联合国毒品和犯罪问题办公室东南亚和太平洋区域办事处：《随着犯罪分子采用新技术，价值数十亿美元的网络欺诈行业在东南亚扩张》，2024 年 10 月 7 日，<https://www.unodc.org/roseap/en/2024/10/cyberfraud-industry-expands-southeast-asia/story.html>，最后访问日期 2025 年 4 月 10 日。

50 刘妍.数据主权的演进、挑战与层级治理路径[J].中国科技论坛,2023(6):144.

51 梁坤.网络犯罪国际公约中的跨国调取数据条款[J].当代法学,2024(2):131.

国深度伪造犯罪的合作治理顺利实现。

《打击网络犯罪公约》草案中第 25 条的规定体现了对电子数据进行分类保护的尝试,其中第 25 条将电子数据划分为流量数据、内容数据与用户信息三类。为了确保该公约中关于跨国数据调取的条款能够与我国国内数据出境规则进行顺利对接适用,我国应在由《中华人民共和国数据安全法》(以下简称《数据安全法》)为主的数据分级分类保护框架下,对跨国电子取证中可能涉及的网络服务提供者掌握的各类数据资源进行对应的分类整合。同时,在各地区、部门以及相关行业、领域的主管部门制定的重要数据具体目录中进行针对性完善,确保我国网络服务提供者在面对他国为跨国深度犯罪治理提出的电子证据调取需求时,能够实现对数据安全性与隐私的更好保障。

另外,《联合国网络犯罪问题政府间专家组第六次会议报告》(以下简称《第六次会议报告》)第 18 条还强调了会员国应当加强电子证据收集方面的合作,其中(d)项提到可通过公私伙伴关系与服务提供商接触,以建立执法、网络犯罪调查和证据收集方面的合作模式⁵²;虽然在《打击网络犯罪公约》最终提交大会表决的草案中却未明确公私合作关系构建的概念,仅在序言中出现了“各国应加强网络犯罪打击方面与相关国际和区域以及非政府组织、民间社会组织、学术机构和私营部门实体的支持和参与下开展合作”的概括性表述。若各国能在跨国取证的公私合作方面达成一致,认可执法管辖权的行使,有助于在国际合作的框架内推动数据价值的实现。公私伙伴关系的建立可以推广跨国犯罪治理的最佳做法,开发技术层面的解决方案,并提高企业和公众对预防网络犯罪的认识。缔约国为公、服务提供商为私形成的公私伙伴合作模式,是对国际社会合作治理犯罪时代背景下跨国电子取证命题的回应,能够在一定程度上促进国际层面的数据跨境流动,消解各国数据出境规则差异带来的流动阻滞。

(三) 全球共治强化: 联合国主导的深度伪造犯罪治理多边合作框架

1. 积极参与公约附加议定书谈判, 促进公约与我国国内法的对接

联合国在跨国深度伪造犯罪治理上提供了一个各国协商交流的平台,通过对相关事项的沟通得以在一定程度上消解各国法律制度、意识形态等方面存在的差异产生的管辖权冲突等法律冲突与执法障碍。联合国毒品和犯罪问题办公室执行主任表示,《打击网络犯罪公约》是在网络空间威胁迅速增长之际第一个联合国打击网络犯罪公约⁵³。《打击网络犯罪公约》在第二章第 12、13 条规定各国应将信息与通信技术系统有关的造假行为、盗窃或欺诈行为确立为刑事犯罪。跨国深度伪造犯罪可能在我国构成诈骗罪或盗窃罪等罪名,以上规定均是《打击网络

52 《2021 年 4 月 6 日至 8 日在维也纳举行的全面研究网络犯罪问题专家组会报告》, <https://documents.un.org/doc/undoc/gen/v21/025/94/pdf/v2102594.pdf>, 最后访问日期 2025 年 4 月 10 日。

53 联合国毒品和犯罪问题办公室:《联合国: 成员国敲定新的网络犯罪公约》, https://www.unodc.org/unodc/frontpage/2024/August/united-nations_-member-states-finalize-a-new-cybercrime-convention.html, 最后访问日期 2025 年 4 月 10 日。

犯罪公约》可能对跨国深度伪造犯罪形成有效规制的基础。同时联合国大会授权特设委员会进行会议，就公约的补充议定书草案进行谈判⁵⁴。我国应当在公约最终生效后，积极参与后续工作进行，促进我国利益主张在公约中的体现。

我国的法律法规体系中尚未直接将我国缔结、签署或批准加入的条约纳入其中，我国签署的国际条约在国内生效前需通过国内立法进行内容转化。因此我国在签署《打击网络犯罪公约》及该公约最终生效后应积极促进公约相关规定与我国国内法的衔接，最大限度地发挥公约在促进各国打击网络犯罪合作层面的效能。

在这一过程中，需要全面梳理我国刑事法律体系中对深度伪造犯罪规则的条款以及《数据安全法》《网络安全法》等规制数据流动与网络空间的法律法规；并挖掘现有规范与《打击网络犯罪公约》内容存在的差异或规制空白，结合我国主权主张对公约新增的条款进行可行性分析。例如我国数据出境制定体系目前由数据出境安全评估、数据出境标准合同以及个人信息保护认证三大路径构成，并无他国与我国网络服务提供者之间的数据提供规则。如果在《打击网络犯罪公约》生效后，我国需要将草案中规定的跨国数据调取条款即针对用户数据的提示令制度纳入我国法律体系中，就需要我国在统筹现有数据出境审查方案的基础上思考并通过立法修改为提示令制度创造一个合理使用空间。提示令制度涉及到国家主权与网络、数据主权以及公民隐私保护等多方面的平衡问题，我国需要在立法修改中充分考量如何界定提示令的适用范围，是仅限于与公约相同的用户信息范围还是扩充到所有涉及跨国深度伪造犯罪治理的“电子数据”？如果我国对跨国数据调取条款仅适用于用户信息表示认同，那么在《打击网络犯罪公约》的其他缔约国因犯罪治理需要向我国申请调取该类数据时，就无需再适用我国数据出境的安全审查制度⁵⁵。

2. 加强打击犯罪国际合作，在网络主权语境下解决管辖权冲突

跨国深度伪造犯罪治理国际合作的开展必须考虑各国的网络主权，只有在尊重各国主权与网络主权的基础上开展的合作才能实现真正的互利共赢。我国解决跨国深度伪造犯罪管辖权冲突方案的出发点就是在尊重他国主权的基础上加强国际合作与协调⁵⁶。刑事管辖权是国家主权的典型体现，因此脱离了网络主权语境去讨论如何解决跨国深度伪造犯罪的管辖权冲突无异于空中楼阁。

欧美国家极力向国际推崇的《布达佩斯公约》中针对管辖权冲突的解决也仅规定了在有利于犯罪追诉基础上的协商机制。纵观《打击网络犯罪公约》草案，其中第22条第5款概括地确立了管辖权冲突的磋商机制，规定若多个缔约国针

54 联合国毒品和犯罪办公室：快速链接-打击网络犯罪公约，<https://www.unodc.org/unodc/en/cybercrime/convention/home.html>，最后访问日期 2025 年 4 月 10 日。

55 梁坤.网络犯罪国际公约中的跨国调取数据条款[J].当代法学,2024(2):133.

56 徐展鹏、丁丽柏.网络犯罪治理国际合作：发展趋势、全球协作与中国方案[J].学术论坛,2024(2):128.

对同一行为展开管辖时应及时展开相互协商确定具体管辖权归属，与《布达佩斯公约》的规定类似。另虽然该公约通过第 22 条第 2 款(c)项明确的普遍管辖原则仅适用于跨国洗钱犯罪，但这意味着大部分国家已经意识到了应当跳出传统的属地、属人管辖规则采取普遍管辖对应跨国深度伪造犯罪治理的重要性。然而，普遍管辖原则若在跨国深度伪造规制上得以确立，国家在没有直接联系的情况下仅依据加入公约的刑事定罪条款对深度伪造犯罪行为进行管辖，容易引发国家间的主权争议；因此普遍管辖原则的适用范围和条件需要明确界定，明确普遍管辖原则在跨国深度伪造犯罪中的具体应用，确保在实现有效打击犯罪目标的同时又不会侵犯国家主权。

随着全球范围内深度伪造犯罪的日益猖獗以及深度伪造技术的持续发展进步，将国家主权理论融入到打击和惩治跨国深度伪造犯罪的国际合作机制构建中，已经成为世界各国以及国际社会治理跨国深度伪造犯罪命题的应有之义。我国在参与国际机制建立的讨论和协商过程中，应当积极主动地提出以网络主权为基础的管辖权冲突解决方案。我国还应该针对网络主权和刑事管辖权的主张，进行多层次、多形式的系统阐述和深入讨论；在国际层面推动开发与构建全球网络犯罪 Intelligence-Sharing 平台，促进执法机构、网络安全公司以及来自不同国家的相关利益方之间的交流，在识别和逮捕跨国深度伪造犯罪分子上加强合作和援助⁵⁷。

3. 完善区域性跨国深度伪造犯罪治理机制，构建网络空间命运共同体

在联合国主持制定的《打击网络犯罪公约》被联合国大会表决通过并生效之前，国际层面暂时缺少跨国深度伪造犯罪的全球性公约。在联合国框架内网络犯罪国际公约生效之前，考虑到区域治理目标相对于全球统一治理模式更容易实现，我国亟需完善现行的区域性网络犯罪治理国际合作机制，以确保网络空间的健康与有序发展；加强与周边国家的沟通与协调，共同制定区域性深度伪造网络犯罪治理的规则和标准，提升打击跨国深度伪造犯罪的效率。上海合作组织成员国于 2006 年签署通过《上海合作组织成员国元首关于国际信息安全的声明》，指出须加强双边、地区和国际层面的合作，加大各国保障信息安全的力度；后于 2020 年签署《上海合作组织成员国元首理事会关于保障国际信息安全领域合作的声明》，呼吁国际社会在信息领域紧密协作，共同构建网络空间命运共同体⁵⁸。在《第六次会议报告》以及《打击网络犯罪公约》草案中也均提到“联合国积极为需要援助的所有会员国特别是发展中国家在网络犯罪调查在电子证据处理等方面开展能力建设”。该措施能够有效消解跨国电子取证过程中，因部分国家国内的网络治理能力严重不足导致的网络主权、数据主权受损现象。这与我国对构建网络

57 AllahRakha N. Global perspectives on cybercrime legislation[J]. Journal of Infrastructure, Policy and Development, 2024, 8(10): 6007.

58 邓浩、李天毅：《上合组织信息安全合作：进展、挑战与未来路径》，https://www.ciis.org.cn/yjcg/sspl/202109/t20210924_8175.html，最后访问日期 2025 年 4 月 10 日。

空间命运共同体的理念相呼应，在习近平总书记提出的人类命运共同体理念的理論指导下，我国应当积极倡导和推动国际合作，以合作共赢为宗旨，努力构建一个共同商议、共同建设、共同治理的网络空间命运共同体。

在区域性深度伪造网络犯罪规则的制定过程中，我国可借鉴《塔林手册 2.0》的编写方式，积极倡导网络空间领域的专家群体发挥在国际网络规则制定中的主导作用⁵⁹。《塔林手册 2.0》在编写过程中专门组织了两次政府代表咨询会议，广泛听取各国代表对手册内容的意见与建议，手册初稿的撰写者包括国际专家组成员以及其他专业人士⁶⁰。通过这种方式可以确保区域性规则的制定更加贴近实际需求，同时兼顾国际社会的多元意见。在规则制定过程中，还应注重平衡不同国家的利益，尤其是发展中国家和发达国家在网络技术发展水平上的差异。

Governance of Deepfake Crimes under the UN Framework and China's Response

Yang Xiangyu

Hunan Normal University, Changsha, Hunan 410000

Abstract: The rapid advancement of deepfake technology has spawned a new form of criminal activity—deepfake-enabled crimes—which exhibit cross-border operations, high deceptive potential, and multifaceted violations of legal rights. Based on the principles of cyber sovereignty and international comity, the current inadequacy of domestic regulations and the fragmentation of international governance urgently call for the establishment of a cooperative governance framework for deepfake crime under the leadership of the United Nations. However, the international legal regulation of deepfake crime faces three major dilemmas: Traditional criminal justice procedures are ill-suited to the transnational nature of deepfake crime, leading to frequent jurisdictional conflicts and obstacles in cross-border electronic evidence collection; It is difficult to reach an international consensus on regulating deepfake crime due to significant differences in national legal systems and regulatory philosophies, as well as the lack of an effective international cooperation mechanism; The drafting process of the UN Convention on Cybercrime has been marked by intense power struggles among major nations. To address these challenges, a multi-tiered approach—incorporating substantive integration, procedural refinement, and strengthened global governance—

59 虞文梁：《论网络犯罪刑事管辖权的国际冲突与规制》，《环球法律评论》2022 年第 5 期，第 191 页。

60 《塔林手册 2.0 第一次有了中国专家的声音，九个问题让你认识它|专访黄志雄》，<https://www.shobserver.com/detail/45198.html>，最后访问日期 2025 年 3 月 24 日。

should be adopted to achieve effective regulation of transnational deepfake crime.

Keywords: Deepfake; Transnational Criminal Justice; International cooperation;
United Nations Convention against Cybercrime; Cross-border Data Flow

2025年中国国际法学年会

网络空间绝对主权论下的网络间谍行为的规制

高彬彬*

摘要：近二十年来，网络空间主权观念逐渐取代“无主权论”成为国际共识，但主权适用路径仍存在绝对主权与相对主权的理论分歧。通过解析主权原则的内涵、网络空间与现实空间的联系以及相对主权论的局限性，可以肯定绝对主权立场的正确性。同时，网络间谍活动作为典型的低烈度网络行为，直接侵犯他国网络主权，且因归责困难、各国立法差异大、国际反间谍规则的缺失面临着难以规制的障碍。为应对这一挑战，需在完善现有国家责任体系的基础上，逐步推动构建以联合国为主导的新型网络治理国际机构，并通过渐进式国际合作强化反间谍情报共享与协同执法。对中国而言，坚持网络主权原则是抵御网络间谍侵害的基石，需主动抵御网络间谍的侵害，同时坚持以联合国为主渠道推进网络治理，参与推进网络间谍的国际法规制。

关键词：网络空间 网络主权 绝对主权 网络间谍 中国

纵观历史脉络，网络空间的发展轨迹鲜明地划分为三个阶段：起始于“无拘无束”的自由发展阶段，继而转向“本土化监管”的国内规制时期，最终迈向了“跨国协同治理”的全球管理新阶段。主权原则适用于网络空间目前已成为国际社会公认的共识，但该如何适用于网络空间的话题一直饱受争论，不同立场对同一行为可能产生不同的结论。在当今的数字化时代，网络间谍也发展得愈加迅速与隐蔽，其是否会对网络空间主权造成威胁关系到如何采用国际法对网络间谍进行规制的问题。有鉴于此，本文拟从网络空间主权的性质及范围的界定出发，结合国内外相关法律规定与实践，对网络间谍行为的国际法规制作出探讨。

一、网络空间主权的性质及其争论

（一）无主权论的落败：网络空间主权的证成

20 世纪 60 年代末，互联网诞生并逐渐应用于社会生活，人们一度认为互联网将缔造一种摆脱束缚的自由形态。当时，网络空间无主权论的支持者主要是美国，秉持网络自由主义理念的学者们主张网络空间是一个没有国界、没有主权管辖的自由领域。网络的参与主体主要是非政府组织和个人，早期网络空间的各种公共问题都在无政府状态下得到解决，形成了一套互联网秩序。20 世

* 福州大学法学院国际法学硕士研究生。

纪 90 年代中后期，美国学者巴洛发表了《网络独立宣言》，从该宣言中可知，网络无主权论的主要观点有：一是政府权力在网络空间没有受到被统治者的同意，无权力来源的根据；二是网络空间是流动的，无界限的，不存在领域的概念；三是网络空间有自己的社会契约和治理模式¹。网络无主权论排斥政府的管理，淡化网络空间的主权性质，其本质是在维护企业、社团等非政府组织的利益。美国鼓吹网络自由主义，实际上也是在扩张自己的霸权主义，掠夺没有明确国家属性的空间资源，网络若无国家主权的规制，国与国之间的竞争就等同于网络能力的竞争，对网络能力较弱的国家来说，其国家网络、政治、文化安全岌岌可危。

自 20 世纪 90 年代以来，网络安全问题频繁出现，于互联网而言，最大的风险不在于政府的过度干预，而是政府的不作为²。不仅个人隐私、财产、安全会受到威胁，网络攻击、网络窃密、网络诈骗等网络犯罪行为也严重威胁着国家的政治、经济、军事等各个领域的安全。网络安全就是国家安全，站在国家安全的高度来看，互联网主权的理念，作为界定网络空间治理权限与国际秩序的关键原则，正日益获得全球范围内多数国家的普遍认同与支持，尤其在网络技术迅猛发展、新兴国家不断崛起的时代背景下，其存在正当性更是被广泛接受，网络无主权论逐渐失去话语权。

网络空间主权是传统主权的延伸与发展，从法理基础出发，主权是一个动态的概念，《威斯特伐利亚和约》中明确领土范围是以现实存在的、有形界限为依据的陆地上的边界。而领空、领海的界定受到多种因素的影响，从技术发展脉络来看，在 17 到 18 世纪，先后有国家以枪和炮火的射程为领海边界，如今，领空仍以航空器最大飞行高度为边界，地下以技术能达到的最大深度为边界³。技术进步扩展了国家的边界，延伸了主权的管辖范围，所以主权是动态的、开放的概念，同理，网络主权不是全新的主权，也是对威斯特伐利亚主权概念的发展。

从国际以及各国实践来看，网络空间中已客观存在了立法权、管辖权、防卫权等客观要素⁴。联合国信息安全政府专家组（GGE）在多个报告中均确认，国家主权和在主权基础上衍生的国际规范及原则适用于国家进行的信息通信技术活动，以及国家在其领土内对通信技术基础设施的管辖权。这表明主权原则在网络空间的适用性在国际层面得到了广泛认可。2021 年 2 月，联合国信息安

¹ John Perry Barlow, A Declaration of the Independence of Cyberspace. <https://www.eff.org/cyberspace-independence>, 2025-03-07.

² 王佳宜：《从分歧到共识：网络空间国际规则制定及中国因应》，《中国科技论坛》，2021 年第 8 期，第 168-177 页。

³ 刘文富：《网络政治——网络社会与国家治理》，商务图书馆，2022 年版，第 184 页。

⁴ 盛文楷、肖光荣：《网络空间主权的唯物史观逻辑——兼论网络空间主权的正当性》，《贵州省党校学报》，2023 年第 2 期。

全开放式工作组（OEWG）以协商一致的方式通过的最终报告确认主权适用于网络空间，就这一关键问题达成一致意见是一项重要成就，因为 OEWG 比之前规模较小的专家组不同，其向所有联合国成员国开放，并不限定参与国的范围，除了联合国成员国，OEWG 还允许私营企业、民间组织、技术社群、公民个人等多方参与讨论，这个协商一致的报告说明主权适用于网络空间已经达成了较为广泛的共识。一些区域国际文件，如《中非携手构建网络空间命运共同体倡议》，也明确提出了在尊重各国网络主权的前提下进行合作。我国也在国内法层面积极落实网络空间主权原则，《国家安全法》首次确立了网络空间主权原则，此后发布的《国家网络空间安全战略》明确提出“国家主权拓展延伸到网络空间”，并将网络空间主权作为国家主权的重要组成部分。此外，众多非官方行为体在网络空间规范发展进程中也明确了网络空间存在国家主权，例如，北约网络防御卓越中心组织国际专家编撰的《塔林手册 2.0 版》专章确立了网络主权原则，《塔林手册 2.0》虽不属于正式的法律文件，但其在网络空间国际法领域具有较高的学术和实践参考价值。

可见，网络空间主权理论不仅在法理层面具备充分的理论依据，更在国际实践中获得了普遍认同与有效实施，从法理和实践两方面深刻印证了国家主权在网络空间中适用的合理性和合法性。网络空间主权的发展表明其既是对传统国际法原则的继承与发展，也是应对网络空间特殊性而进行的必要调适。

（二）绝对主权论与相对主权论之争

主权如何适用于网络空间的问题本质上就是研究什么样的网络行为会侵犯国家主权，在此，各国持绝对主权与相对主权两种立场。可以肯定的是，某些网络行为在国际法上显然是非法的，如达到一定“规模 and 效果”即其造成的后果与物理世界中的武装攻击相似的网络行为以及一国通过网络行动干涉他国内政的行为，但对于某些渗透到另一国领土上的计算机系统，但不会造成任何物理损害且未上升到使用武力或禁止干涉的程度的行为，（以下简称低强度的网络行为）是否侵犯主权，存在不同的看法。低强度网络行为中的典型代表是网络间谍行为，其目的在于窃取目标国的机密信息，一般不会造成物理损害。

就以上争议，《塔林手册 2.0》在总结了专家组的意见后，对是否构成侵犯主权的网络行动设定了条件：（1）侵犯目标国领土完整的程度（2）是否存在干涉或篡夺固有政府职能的行为⁵。对于侵犯目标国领土完整的行为，可以分为三种程度：第一种是造成物理损坏的网络操作，例如恶意软件的入侵导致设备中的冷却元件发生故障，进而引发设备过热现象，最终造成关键组件的熔化与损毁，对于这种行为，绝对主权和相对主权论都认为这样的行为侵犯了主权；第二种是导

⁵ [美]迈克尔·施密特总主编，黄志雄等译：《网络行动国际法塔林手册 2.0 版》，社会科学文献出版社，2017 年版：规则 4，第 65 页。

致领土国的网络基础设施失去重要功能的网络行为，如 2012 年的沙蒙病毒，导致沙特阿美公司的数千个硬盘无法使用，绝对主权与相对主权论也都支持这样的行为侵犯了主权；第三种行为就是不会造成物理损害也不会导致网络基础设施失去重要功能的低强度网络行为，对于这种程度的行为，专家组未有共识。绝对主权主义立场的国家倾向于采取一种直接且明确的立场，即任何未经授权侵入位于另一国领土内计算机系统的网络行为，均被视为对目标国家主权的侵犯。相反，相对主权主义立场的国家则持反对意见，它们认为，单纯因网络行为侵入另一国的计算机系统，并不自动构成对该国主权的侵犯，在相对主权主义的视角下，网络行为必须至少对目标国造成实质性的伤害或损害，方能被视为国际不法行为。

对于干涉或篡夺固有政府职能的行为，比如非法的域外执法行为、篡夺或删除涉及提供社会服务、选举行为、征税或重要国防活动的数据的行为⁶，绝对主权与相对主权论都认为这种行为侵犯了国家主权。可以看出，绝对主权与相对主权主义只在低强度的网络行为是否侵犯主权存在歧义。

（三）捍卫网络空间的绝对主权

在荷花号案中，国际法院在判决中分析道：“国际法加之于国家的第一个、也是最重要的一个限制是：国家不得以任何形式在他国领土内行使权力，不存在相反的允许性规则。”⁷依据主权原则，各国在其领土上行使国家权力具有排他性，在没有习惯例外允许的情况下，仅仅渗透一国的领空、领海或陆地就违反了主权原则，不需要任何形式的物理效应。既然主权原则也同样适用于网络空间，那么未经他国允许，侵入到他国领域上的计算机系统也构成对主权的侵犯。此外，在国际法体系中，有一些规则明确允许国家在特定例外情况下合法进入另一国家的领土主权范围，无害通过权即为其中之一，它赋予国家船舶在遵守一定条件下穿越他国领海而不受阻碍的权利；《国际民用航空公约》亦明确规定，当一国民用飞机在另一国领土内遭遇紧急或危险情况时，民用飞机所属国有权且应提供必要的援助，但没有任何公约或条约明确或隐含地允许一国可在另一国从事任何网络行为。从物理层面看，任何网络行为都需依托在有形的基础设施上，比如计算机硬件、电缆等，这些都是在主权明确划分的自然空间实际存在的，网络活动的发起人或发起国虽然具有隐蔽性，但实施该行为的个人或实体始终是受一国或多国管辖的，而且该网络活动对某一国家所造成的影响是可确定的，切实发生在某国的领土范围之上。

相对主权主义不赞成低强度网络行为都侵犯了国家主权，但其又不能在侵犯主权的低强度网络行为与不侵犯主权的低强度网络行为中划出一条明确的红线。其次，在实践中，低强度网络行为通常进行情报、信息收集，而情报、信息收集

⁶ 黄志雄：《网络主权论——法律、政策与实践》，社会科学文献出版社，2017 年 7 月第一版，第 238 页。

⁷ See *S.S. Lotus (Fr. v. Turk.)*, 1927 P.C.I.J. (ser. A) No. 10 (Sept. 7).

与域外执法的界限是模糊的，情报部门与执法部门可能进行信息共享，因此无法确定收集到的情报是否只用于一个目的，而不造成对政府固有职能的篡夺。可见，相对主权主义的观点中存在着无法清晰的“灰色地带”。

从主权原则的内涵、网络空间与现实空间的联系、相对主权论的不足来看，绝对主权的立场是正确的，对计算机系统的任何渗透、入侵，即使是不造成任何损害的侵入，也侵犯了计算机系统所在国家的主权。

二、网络间谍侵犯国家主权之理论分析

传统间谍是指被间谍情报机构秘密派遣到对象国（地区）从事以窃密为主的各种非法谍报活动的特工人员，又指被对方间谍情报机构暗地招募而为其服务的本国公民。国际法上对间谍的定义主要是在武装冲突法中，其将间谍行为限制于军事领域，国际法著作中也不乏有关间谍的定义，如《奥本海国际法》将间谍界定为“一国为了秘密获取军事和政治情报而派往国外的秘密人员”⁸。传统间谍行为主要指收集政治、军事情报的行为，其行为的核心要素都是“秘密”、“伪装”、“窃取”。随着国家间经贸活动和经济往来的增加，经贸领域日益成为间谍活动的高发区。间谍逐渐发展为一国为秘密获取军事、政治、经济、科技等各种关乎国家安全的情报而被派往境外的秘密人员。

国际法目前还没有对网络间谍下一个明确的定义，根据《塔林手册》2.0 中的定义，网络间谍是指利用网络能力，以秘密或欺诈的方式收集或试图收集信息的行为。这包括但不限于利用网络能力进行监视、监控、采集或窃取通过电子传输或存储的通讯、数据或其他信息⁹。网络间谍主要活动在网络虚拟空间，利用网络技术和手段进行情报的窃取、刺探和传递，且不会对目标领土的计算机系统或硬件设施造成任何物理损害，也不会导致网络基础设施丧失重大功能，属于上述所提到的低强度网络行为。

2013 年 6 月 5 日，英国《卫报》首次披露了美国正在开展一个代号为“棱镜”的秘密项目，美国前中央情报局雇员爱德华·斯诺登向英国《卫报》和美国《华盛顿邮报》等媒体揭露了“棱镜”计划的真相，详细揭示了美国国家安全局（NSA）和联邦调查局（FBI）自 2007 年起启动的这项绝密电子监听计划，“棱镜计划”在曝光前竟秘密进行了 6 年之久¹⁰。2023 年，中国国家计算机病毒应急处理中心在会同 360 公司配合侦办西北工业大学遭受网络攻击案过程中，成功提取了名为“二次约会”的间谍软件多个样本，该软件为美国国家安全局开发的网络“间谍”武器，并在上千台遍布全球多国的网络设备中隐蔽运行¹¹。从目前的网络间谍事例

⁸ 詹宁斯等修订，王铁崖等译：《奥本海国际法》，中国大百科全书出版社，1998 年版。

⁹ [美]迈克尔·施密特总主编，黄志雄等译：《网络行动国际法塔林手册 2.0 版》，社会科学文献出版社，2017 年版：规则 32，第 193 页。

¹⁰ 中国日报网：揭秘：棱镜计划，https://www.chinadaily.com.cn/hqzx/2013-06/19/content_16637738.htm, 2025-03-27.

¹¹ 央视新闻：“二次约会”间谍软件分析报告：网络攻击西北工业大学，美国相关人员真实身份被锁定，

可以看出，网络间谍行为具有高度隐蔽性、持续时间长、效率高等特点。

网络间谍在国际法领域不仅缺乏明确的概念界定，在治理方面也存在局势不明的状况。美国鼓吹网络间谍的“两分法”，将网络间谍分为网络政治间谍和网络经济间谍，认为只有网络经济间谍侵犯国家主权¹²。一些专家组成员不赞同网络空间的绝对主权论，也是基于目前国际社会对网络政治间谍的容忍，如果主权不能禁止现实世界的间谍活动，那么间谍活动在网络空间也不能被主权所禁止。少数学者宣称，间谍行为在国际事务中如此常见，因此可以认为国家的实践承认间谍活动的合法性，甚至将间谍活动上升为一种国际习惯¹³。也有不少学者以“荷花号案”中确立的“法不禁止则允许”的原则，来证明网络间谍在国际社会的合法性¹⁴，毕竟目前没有任何一个国际规则明确禁止网络间谍行为，尤其是网络政治间谍行为。

从手段来分析，网络经济间谍与网络政治间谍使用的手段是一致的，均是以“秘密窃取”的方式，相比之下，政治、军事情报泄露的严重性与可能造成的损失并不比经济领域的情报泄露来得轻或少。虽然国际法没有明确禁止网络政治间谍，但未经同意、授权收集且泄露他国军队部署、政策实施细节等关乎国家发展安全的敏感信息，显然违反不干涉原则、领土主权原则等国际法原则¹⁵。可以看出，“二分法”只是美国为了维护其互联网技术、人工智能技术以及经济领域的超越地位，同时又不愿全面禁止窃取他国情报的行为的一种说辞罢了。

此外，与国际社会对网络间谍行为保持沉默、甚至默许的状态形成鲜明对比的是，许多国家都对网络间谍进行了谴责。最典型的例子是，欧盟-南方共同市场联合新闻声明（EU-Mercosur Joint Press Statement）中表明：“强烈反对(美国国家安全局)在我们国家进行的电信窃听和间谍活动”，理由是这种“不可接受的行为……侵犯我们的主权。”¹⁶这也恰巧说明将网络间谍行为视为一种国际习惯的观点是站不住脚的，国际习惯的形成不仅需要国家实践还需有法律确信，当前国际社会既未通过成文法将网络间谍行为合法化，也未能形成承认其合法性的普遍法律共识，因此间谍活动司空见惯的这一残酷事实并不能成为领土主权的习惯例外。

如前所述，在网络间谍的合法性问题上，绝对主权主义者认为所有网络间谍活动都是非法的，因为它必然涉及未经同意而侵入另一国领土上的计算机系统。相比之下，相对主权主义者将干扰或篡夺政府的固有职能的行为与网络间谍活动

<https://www.chinanews.com.cn/gn/2023/09-14/10077475.shtml>, 2025-03-27.

¹² 黄志雄：《论间谍活动的国际法规制——兼评 2014 年美国起诉中国军人事件》，《当代法学》，2015 年第 1 期，第 141-142 页。

¹³ Jeffrey H. Smith, Symposium: State Intelligence Gathering and International Law: Keynote Address, 28 Michigan Journal of International Law 543, 544 (2007).

¹⁴ Oğuz Kaan Pehlivan, *Confronting Cyberespionage Under International Law*, Routledge Focus, 2019, p.111.

¹⁵ 李彦：《场域视角下网络间谍的国际法规制》，《政治与法律》，2023 年第 10 期，第 162-176 页。

¹⁶ See Note verbale dated 22 July 2013 from the Permanent Mission of the Bolivarian Republic of Venezuela to the United Nations addressed to the Secretary-General, undocs.org A/67/946.

进行了区分,认为为执法目的而窃取数据的行为,比如窃取刑事证据,侵犯了主权,而以间谍活动为目的窃取同样的信息的行为却不侵犯主权,因为间谍活动不会干扰或篡夺目标国政府的固有职能。这是一种奇怪的观点,禁止一个国家的执法官网入侵位于外国的犯罪嫌疑人的电脑,但却允许一个国家的情报部门侵入外国的防御系统窃取诸如核武器密码等战略情报,这样的逻辑颇为悖谬。《塔林手册》2.0 中将网络间谍分为“远程网络间谍行为”以及“近距离网络间谍行为”,后者特指间谍人员身处目标国领土范围内,通过计算机网络实施的情报窃取活动。持相对主权主义立场的专家组成员主张,仅后者构成对国家主权的侵犯,但其实作出这种区分是无意义的,因为这两种行为所入侵的计算机系统均在目标国领土上,行为所达到的最终效果也是一致,如果“近距离网络间谍”行为侵犯了主权,那么“远程网络间谍行为”也一定侵犯了主权。在网络间谍问题上,网络空间绝对主权论具有充分正确的适用性。经过上述多角度的剖析与论证,可以清晰地得出结论:无论采取何种技术手段或行为方式,网络间谍行为本质上都是对国家主权的明确侵犯。

三、规制网络间谍行为之困境

(一) 网络间谍行为的归责问题

对网络间谍行为进行打击与规制的首要问题是明确该网络间谍行为的归属,网络间谍具有高度隐蔽性,其发起来源难以确定,即侵害主体无法直接被监测到,因此网络间谍可以不受惩罚地任意实施间谍技术去损害他国的主权,而无法援引国际法或国际规则加以规制或制约。网络间谍行为的溯源问题是首要的一大障碍,很多情况下,即使追查到了网络间谍的 IP 地址,也无法确定作案人。

此外,构成国际不法行为的主观要件是某一行为可归因于有关国家而被视为该国的行为,即某一行为对于有关国家具有“可归因性”,因此只有可归因于国家的网络间谍行为才构成国际不法行为,进而使其承担国家责任,然而,这一归责原则在实际应用中面临显著挑战,因为众多网络间谍组织实为非国家性质的民间实体,它们缺乏明确的国家归属与特性,导致在判定网络间谍行为的“可归因性”时存在极大的复杂性和不确定性。即便某些非国家行为体的网络间谍活动客观上有利于特定国家,也无法因此判定该非国家行为体处于该国家的控制之下,且从非国家行为体从事间谍行为的“初心”出发,它们通常不会承认或轻易暴露与特定国家实体的关联。而当网络间谍行为由民间组织或个人实施时,国际法的规制面临诸多挑战,因为传统上国际法主要聚焦于国家行为,而民间组织、个人的行为往往通过国内法或区域合作来规制。

(二) 全面的反间谍国际公约难以达成

条约作为两个或两个以上国家间正式缔结的书面协议,被广泛视为衡量国际法效力的黄金标准。与其他国际法渊源相较,条约的独特之处在于其以书面形式

明确界定了缔约国间的法律权利与义务，并基于各国的明确同意而具备约束力。鉴于条约在国际法体系中的这一核心地位，当前存在一种普遍期望，即在网络空间这一新兴且日益重要的全球领域中，也应通过签订广泛的双边或多边条约，最终达成广泛的国际共识来规范国家行为¹⁷。然而，如前所述，各国对间谍行为的态度莫衷一是，甚至在网络间谍领域出现合作阵营化、区域化，例如1948年“五眼联盟”（FVEY）的成立。该联盟由美国、英国、加拿大、澳大利亚及新西兰五个国家共同构建，旨在通过签署电子间谍网络协议，强化成员国之间的情报共享机制与联合行动能力，以更有效地拦截与分析敌对国家情报，随着国际政治经济格局的演变，“五眼联盟”的活动范围与影响力逐渐扩大，其行动亦被指向对中国、俄罗斯等特定国家进行情报压制与网络空间的攻击与对抗。此外，在网络领域，因为各国的技术能力不同、政府对互联网的管理力度不同、言论自由的范围边界也不同，很难促成各国基于互利的原则达成协议。事实上，各国其实是希望制定网络空间内各类活动的规范标准的，但是仍有众多国家认为，其他国家所倡导的规范往往与他们自身的利益存在冲突或对立，以我国为例，2015年《国家安全法》首次明确了“网络主权”的概念，并在国际舞台上积极推动相关共识的形成。然而，这一提议并未获得广泛支持，主要原因在于“网络主权”与美国所提倡的“多利益攸关方”互联网治理模式存在利益不一致之处。美国的“多利益攸关方”模式强调非国家行为体在网络治理中的主导作用，而中国的“网络主权”则强调国家在网络空间的主导权和主权。这种利益冲突使得通过国际条约的形式达成全球性的反间谍行为规则或标准面临巨大的局限性和挑战性。

（三）司法、执法层面的阻力

当网络间谍行为的实施者为无法归因于国家的非国家行为体时，受害国对他国民间组织、个人的追责往往面临着司法、执法层面的阻力。当前国际网络空间治理体系面临反间谍规则供给不足的结构性的困境。如前所述，在全球层面，尚无法形成统一的反网络间谍多边公约，导致各国通过单边立法争夺网络空间规则制定话语权。这种单边立法现象使得国际社会对网络间谍行为的规制呈现碎片化特征，并引发国内法域外适用的合法性争议。以美国的《经济间谍法》为例，其明确效力范围可适用于美国境外，由此开创了商业秘密刑法域外适用的立法先例，但该立法因有“长臂管辖”之嫌而饱受争议。

在跨国司法协作层面，网络间谍行为的规制面临三重障碍：其一，管辖区冲突。由于网络的跨国性特点，同一网络行为体可能因属人、属地等连接因素，同时受到多个国家法律的管辖，从而引起管辖权冲突。其二，跨境取证机制存在制度性缺陷。不同国家的法律制度和司法程序差异，导致跨境取证的效率低下，甚至可能被拒绝合作。其三，信息共享存在制度壁垒。信息共享机制的不完善使得

¹⁷ Rebecca Crootoof, INTERNATIONAL CYBERTORTS: EXPANDING STATE ACCOUNTABILITY IN CYBERSPACE, Cornell Law Review.565,(2018).

关键证据难以及时、准确地传递和共享¹⁸。关键证据的传递往往受制于数据本地化存储要求。这种制度性障碍导致国际社会在应对 APT29 (Cozy Bear) 等国家级网络间谍组织时, 难以形成有效司法、执法合力。

四、网络间谍行为规制之路径

(一) 国家责任原则的具体化适用

将国家责任原则适用于引发重大跨界损害的网络行动可以带来多重积极效应, 要求国家对其网络行动所导致的有害后果承担相应责任, 能够将这些潜在风险活动引入国家行动的成本考量, 从而构成一种有效的威慑机制, 遏制不负责任的网络行为; 同时, 这一原则的实施还极大地增强了因网络间谍行为而受损的国家获得充分赔偿的可能性。在网络空间中适用国家责任原则需明确两个问题: 一是非国家行为体行为的归因问题、二是造成损害的衡量标准。

对非国家行为体的国际不法行为追责的最高效路径就是将其转化为国家责任。根据《国家责任草案》的 8 条的规定: “如果该人或一群人的行为实际上是在该国的指示下或在该国的指示或控制下进行的, 则根据国际法应被视为国家的行为。”可以看出, 非国家行为体若处于国家的“有效控制”之下, 则该非国家行为体所做的网络间谍行为可以归责于国家。

另外, 《国际责任草案》中还规定国家可能因未采取必要措施防止或未阻止非国家主体的违法行为而承担国际责任, 即国家负有审慎义务。将国家审慎义务引入网络空间可以进一步扩大网络间谍行为中国家的责任范畴。《塔林手册 2.0》在第六条中明确指出了各国在网络空间中需要承担审慎义务, 即国家不得在知情的情况下允许位于其领土内的网络基础设施或者处于其政府控制下的网络基础设施对其他国家造成不利影响¹⁹。在一定程度上肯定了国家在网络空间中的积极作为义务——即需采取必要措施防止和阻止此类行为的发生, 还隐含了国家对于其管辖范围内网络活动的监督与管理责任。将审慎义务引入网络空间, 不仅能够加强国家对位于其领土管辖范围内的网络行为的监督与制约, 也为受害国寻求国际救济提供了法律依据。但是, 审慎义务的适用也必须谨慎, 以避免其滥用而导致类似反恐活动中“安全港理论”归因扩大化的不良后果。为此, 必须将审慎义务与归因标准明确区分开来, 这意味着当非国家行为体实施了网络间谍行为时, 应先审查其与国家之前是否存在“有效控制”关系, 若归因失败, 再审查国家是否尽到审慎义务, 若国家未尽到审慎义务, 只需对损害的扩大负责。而不宜以审慎义务为理由, 扩大国家责任的范围²⁰。

¹⁸ 楼伯坤, 李想:《跨境网络犯罪刑事管辖权: 困境、成因与破解》,《法治研究》, 2024 年第 6 期, 第 43-56 页。

¹⁹ [美]迈克尔·施密特总主编, 黄志雄等译:《网络行动国际法塔林手册 2.0 版》, 社会科学文献出版社, 2017 年版: 规则 6, 第 73 页。

²⁰ 马光, 杨琳:《非国家行为体网络行动的国际法责任认定研究》,《福建江夏学院学报》, 2021 年 12 月第 11 卷第 6 期, 第 44-45 页。

网络间谍行为所造成的损害关系到国家责任的承担，应根据实际情况来分析。网络政治间谍行为所造成的损失有时会十分抽象、难以直接计算其经济损失，这类损失包括但不限于干扰国家选举进程、阻碍军事行动的有效性等，为了更全面地评估其影响，可以将因对抗此类行为、恢复被阻碍的进程以及应对后续影响所支出的经济成本和因该政治、军事情报而获得的受益纳入考量范围，以间接反映其带来的综合经济损失。网络经济间谍活动经常导致企业、研究机构等的知识产权和商业机密被窃取，这些损失有时也难以直接量化，但可以通过评估被盗取信息的市场价值、研发成本以及潜在的经济收益来估算。

（二）逐步推动建立独立的国际机构

如前所述，网络间谍行为追查难、损害结果的衡量具有个案性，跨国司法、执法也面临着管辖权冲突等问题，在这种背景之下，构建一个拥有独立调查、裁决与规范建设能力的国际机构，或许能成为应对这些难题的一种较为理想且有效的解决方案。建立独立的网络活动国际机构并非全然新颖的构想。事实上，国际间已不乏成功典范，如国际原子能机构，便专注于核技术合法使用的监督与核查工作；联合国赔偿委员会、世界贸易组织等，均属于先例，这些机构展现了跨国界解决争端、维护秩序的能力。

首先，在网络空间中，精确识别网络行动的发起者构成了一项技术上的艰巨挑战，这往往要求具备高度专业化的计算机技术能力。从专业能力构建的角度来看，独立机构能够集中吸引并保留那些掌握必要专业知识与技能的各国专家，这些专家在不受单一国家利益或偏见影响的环境下工作，能够专注于技术细节与证据分析，从而确保调查过程的公正性与深度。其次，通过独立的机构对网络间谍行为所造成的损失进行衡量，能够保证衡量标准的一致性，确保网络空间国家问责制度发挥积极作用。再者，独立的网络行为管理国际机构可以推动网络空间国际规则的制度建设，由一个中立的机构对各个国家的利益进行调和，而且形成与网络间谍有关的国际习惯法是很困难的，这主要由于国家在网络空间中的行为往往具有高度的隐秘性，缺乏足够公开且一致的实例来证实国家以明确的方式行事，并视此种行为为法律所许可或强制要求。为了克服这一困境，并避免在漫长等待中寻求国际习惯法的形成，一个有效的解决途径是依赖独立的权威机构来制定规制网络间谍行为的成文法。同时，成文法的形成将促使各国阐明其对与网络间谍有关的法律义务的理解，这反过来也将促进学术与实践发展。进一步地，由统一的权威机构颁布网络行为准则，能够促使网络空间法规以更加连贯且灵活性强的方式不断演进和发展。

当然，建立独立的管理网络行为的国际机构仍面临着诸多挑战，推动国际机构的建立需各国政府、国际组织达成共识、形成联合国决议等，这一过程本身也伴随着前述所提及的复杂利益冲突，使得前期筹备工作尤为艰巨。因此，独立的

管理网络行为的国际机构的成立需渐进式推进，在短期内应先支持现有国际机构，加强联合国、国际电信联盟等国际组织在网络空间治理中的作用，支持其发挥领导力和协调作用。在中期阶段可推动建立“专项国际工作组”，聚焦网络间谍取证、归因和应急响应，逐步积累信任。同时，推动技术创新与资源共享，促进各国在网络安全技术研发、应用和推广方面的交流与合作。

（三）加强反间谍国际合作

2015 年，我国国家主席习近平在第二届世界互联网大会首次提出“构建网络空间命运共同体”理念。其中表明，国际网络空间治理，应该坚持多边参与，各国应该加强沟通交流，完善网络空间对话协商机制，研究制定全球互联网治理规则。在逐步推进建立独立的反间谍国际机构的过程中，网络间谍活动的追踪溯源工作以及司法、执法的运转迫切需要国际社会的合作与协调，在全球化的背景下，网络间谍行为往往超越单一国家的边界，涉及多个国家的关键基础设施和敏感数据。因此，加强国际合作不仅是应对网络间谍行为的有效途径，也是维护国际关系的必然选择，在网络空间命运共同体的视角下，国际合作不仅是技术层面的协作，更是价值观念和行为规范的共同塑造。为了实现这一目标，国际社会可以采取一系列措施：首先，通过建立区域性的多边协议和条约，至少在区域范围内达成一致的法律框架，为信息共享、联合行动和争端解决提供基础。其次，各国应致力于共享网络威胁情报，包括攻击模式、恶意软件特征等关键信息，以提高追踪溯源的效率和准确性。同时，国际合作还应包括联合研究与技术开发，以创新追踪溯源技术和方法。通过这些综合性措施，国际社会可以更有效地合作追踪网络间谍活动，共同维护国家主权与网络空间的安全与稳定。我国作为网络主权的坚定倡导者和积极实践者，不仅要提升自身的反间谍的技术能力，同时也积极推进各国反间谍合作，推动构建网络空间命运共同体。

五、结论

主权原则适用于网络空间具有法理和实践基础，更进一步，网络空间应坚持绝对主权原则，低强度网络行为的实施未经国家的允许，必然会侵犯国家主权。中国政府于 2022 年向联合国信息安全开放式工作组提交的《中方关于网络主权的立场》中明确主权平等原则也适用于网络空间，更要求各国在享有网络主权所衍生权利的同时，应遵守国际法一般原则和基本规则，切实履行国际法所规定的相关义务，包括不侵犯他国、不干涉他国内政、审慎预防义务和保障义务。虽然没有明确说明，但从这份文件的相关表述可以看出我国对网络间谍行为的态度：网络间谍行为属于未经许可入侵他国领土或管辖范围内的网络系统的行为，违反了主权原则，构成国际不法行为。网络间谍行为作为低强度网络行为的典型代表，从行为目的、结果、实施手段等各个方面分析，都对国家主权造成极大威胁。为更好地规制网络间谍行为，国际社会应积极合力完善网络空间治理中的国家责任

制度，同时逐步推进独立的网络治理国际机构的建立，以协调各国关于网络间谍行为的分歧，将反间谍行为相关立法提上日程。

2025年中国国际法学会年会

智能船舶保险法律问题研究

李依霓¹, 陈雅灵²

摘要：智能船舶的蓬勃发展为航运业带来了技术革新，同时也对传统的法律和保险制度提出了新的挑战。本文聚焦于智能船舶保险领域的三大法律困境：适航性认定、新型网络风险以及碰撞责任主体不明。首先，智能船舶的无人化和网络化特性使得传统适航性定义难以直接适用，需将适航性扩展至网络适航，并将远程操控人员和智能系统纳入适航性评估范围。其次，智能船舶面临的新型网络风险尚未被传统保险制度充分覆盖，需明确网络风险的定义、承保范围及举证责任。最后，智能船舶碰撞事故中责任主体多元化导致责任划分复杂化，需构建先行赔付与代位求偿相结合的保险模式，以保障受害人权益并明确各方责任。

关键词：智能船；船舶保险；网络风险；船舶适航；碰撞责任

智能船舶技术的迅猛发展对现行法律体系提出了更新和完善的迫切需求。从设计研发到实际运营，无人船舶的整个生命周期不仅投入巨大，还伴随着较高的风险。对于智能船产业来说，保险作为一种有效的风险分散机制，其重要性不言而喻。³智能船舶的机械设计、试验以及未来航行与运营面临的风险与传统的有人驾驶船舶显然在许多方面截然不同，因此就需要在法律层面，特别是保险相关法律方面进行约束，分散各方主体的风险，以期推动智能船舶的发展，但现有的船舶保险相关条款在制定时均未考虑到智能船舶的出现，因此亟需在法律层面解决智能船舶的海上保险相关法律问题。

一、智能船舶与保险发展现状

（一）智能船舶定义与发展现状

2018 年，国际海事组织海事安全委员会（MSC）同意成立一个海上自主水面船舶工作组（MASSWG），开展范围界定工作，旨在确定哪些涉及海上安全的现有国际文书需要修改，以及需要制定哪些新的法律文书。目前对于智能船舶，国际上仍然没有一个统一的定义，国际海事组织、中国船级社、英国劳氏船级社等国际组织对其仅仅作出一个初步的定义，概括来说：智能船舶是指运用人工智能、通信技术等技术手段，使船舶具有自我控制、自主航行、自我运营的智能化

¹ 李依霓，大连海事大学法学院硕士研究生，电子邮箱：Lynjenny@126.com。

² 陈雅灵，大连海事大学法学院硕士研究生，电子邮箱：alingisok@icloud.com。

³ 韩立新，姜亚迪：《无人船舶责任保险制度研究》，载《大连海事大学学报(社会科学版)》2021 年第 1 期，第 1 页。

船舶。⁴从已经取得试航成功的无人船舶来看，当前阶段的无人船舶具有以下特征：首先，现有无人船舶的自动化等级正在不断向更高等级迈进；其次，尽管现有无人船舶已经具备自主航行的能力，并且已有在完全脱离人为干预的状态下完成自主航行的先例，但试航中仍保持船员或岸基操作人员对无人船舶的可控性，并且注重岸基中心的远程监控，以平衡无人船舶自动驾驶技术的不成熟带来的安全风险。

（二）智能船舶保险发展现状

由于智能船舶的承保风险难以量化、责任划分复杂，加之相关司法实践和执法案例的缺失，保险公司在推出相关产品时持谨慎态度。一方面，立法和前期数据的不足使得风险评估困难；另一方面，智能船舶技术和运营模式的新颖性要求保险公司在承保前进行额外的第三方评估和尽职调查，这进一步提高了承保门槛和难度。此外，保险公司还需为智能船舶制定专门的保险条款。尽管如此，随着智能船舶技术的不断进步和产业的快速发展，保险行业必须积极应对新需求，适应技术变革。⁵

国际上已有部分保险公司开始为智能船舶提供保险服务。例如船东互保协会为智能船舶提供一切险。该保单还明确将因船舶或远程操控系统遭受网络攻击而引发的第三方损害纳入保障范围。⁶作为全球最大的船舶保险提供商，Gard 于 2018 年公开表示愿为智能船舶提供保险服务，包括船体、机械和损益保险。然而目前国内关于智能船舶保险的研究仍以理论探讨为主，尚未有保险协会或公司推出专门针对智能船舶的保险产品。《海商法》中的海上保险合同以及《中国人保船舶保险条款》也未能及时修订以涵盖智能船舶的相关内容。由于相关法律和保险制度的空白，智能船舶目前只能被纳入普通船舶的保险框架中进行处理。⁷值得注意的是，部分保险公司在船壳险保单中通过特别约定引入了“网络攻击除外条款”（CL380）。根据该条款，只有当网络攻击被认定为战争险中列明的具有政治目的的行为时，相关损失才属于战争险的承保范围。这种特别约定方式使得保险公司几乎完全排除了对网络安全风险的承保，即使是具有政治动机的网络攻击也不例外。⁸

二、智能船舶保险的法律困境

（一）智能船舶的适航性不明

⁴ 段立金，凤勇：《基于智能船舶下航海技术专业人才培养研究》，载《中国水运》2022 年第 9 期，第 32 页。

⁵ 吉鹏成，吴惟予：《无人船舶碰撞责任保险制度法律问题研究》，载《〈法学前沿〉集刊——新时代航运法治建设研究文集》2023 年第 4 卷，第 176 页。

⁶ 桑家军，严忠伟：《智能船舶对船舶保险条款的适用性研究》，载《中国海事》2024 年第 5 期，第 22 页。

⁷ 《颠覆船舶行业！全球首家无人船公司诞生》，载国际船舶网，https://www.eworldship.com/html/2018/Ship_Owner_0405/137884.html。

⁸ 刘晟昊：《远程控制无人船之船舶保险条款研究》，载《珠江水运》2023 年第 6 期，第 71-72 页。

智能船舶从设计、建造、操作、管理、供应、基础设施等各环节均与传统船舶存在差异,它的应用前景产生了一个非常基础的问题——船上船长和船员的作用,⁹没有了船长和船员,还增加了两个新的法律主体,即远程岸基操控人员和智能航行系统服务的提供者。此时就出现了一个问题,即“适航性”的传统定义是否适用于遥控和自主船舶。其次,“适航”的定义是否延伸到网络适航?换言之,如果船舶采用的网络保护机制被证明是不充分的,并因此发生了网络攻击,那么能否主张该智能船舶不适航?

船舶保险对船舶的适航性提出了明确要求。根据《海商法》第244条,若船舶在开航时因不适航而导致损失,保险人将不承担赔偿责任。自船舶航运业兴起以来,船长船员始终全面参与船舶的运营活动,其中,适当的配员被视为船舶适航性的核心要素,同时也是确保船舶得以在国内及国际水域合法运营的先决条件。对于受遥控的智能船舶,若岸基操作人员不被视为船员,则将无法满足适航性的要求;而自主化智能船舶的完全自主特性,更是与最低配员规定形成了直接的矛盾。此外,就新风险而言,根据劳埃德船级社的报告,网络赋能的无人船舶所面临的风险及其潜在危害,其性质与程度均可能经历显著的变化。¹⁰智能船舶拥有一个完整的网络保护机制至关重要,无论是受遥控船舶或是自主化船舶,其智能航行都依赖于网络的信息交互功能。如果没有一个完好的网络保护机制,网络数据面临的丢失、篡改或泄露风险,这些由于信息交互与共享所固有的“系统性”特征而加剧的风险,它们如同潜流般渗透于航行活动的每一个环节之中,¹¹没有安全的网络保护机制的智能船舶甚至无法保证正常航行,网络安全保护机制是否应作为智能船舶适航的必要条件。船舶不适航作为保险人的免责条款,一旦船舶未能符合适航性要求,保险人即有权依据被保险人违反船舶适航义务的理由,拒绝承担损失赔偿责任。

(二) 智能船舶面临的新型风险

智能船舶可以有效降低人为操作的事故率,减少船员受伤死亡以及海盗带来的风险。但并不意味着船舶自身所面临的风险就随之减少。¹²在智能船舶行驶过程中,岸上的控制人员需实时在线监控从船体搜集到的各类数据,这些数据涵盖了船舶的运行状态、位置信息、设备参数等关键内容。通过对这些数据的分析,控制人员能够全面掌握船舶的航行状况,并依据实际情况作出适时且恰当的反应。例如,当发现船舶出现异常振动或设备故障预警时,控制人员可

⁹ Van Hooydonk Eric, The Law of Unmanned Merchant Shipping-An Exploration, 20 Journal of International Maritime Law 408 (2014).

¹⁰ YEOMANS G. Autonomous vehicles - handing over control: opportunities and risks for insurance[EB/OL]. [2020-07-11]. [www.lloyds.com/news-and-risk-insight/risk-reports/library/technology/autonomous Vehicles](http://www.lloyds.com/news-and-risk-insight/risk-reports/library/technology/autonomous%20Vehicles).

¹¹ 韩立新,姜亚迪:《无人船舶责任保险制度研究》,载《大连海事大学学报(社会科学版)》2021年第1期,第3页。

¹² 蔡莉妍:《智能船舶法律规制的困境与突破》,载《华南理工大学学报(社会科学版)》2020年第6期,第5页。

以及时通知船上人员采取措施，或者远程指导进行故障排查与修复。正因如此，航行中应对网络攻击与数据窃取行为产生的风险至关重要。智能船舶高度依赖网络通信技术来实现岸上控制与船上设备的互联互通，这使得船舶成为网络攻击的潜在目标。黑客可能会试图入侵船舶的网络系统，篡改航行数据、控制船舶设备，甚至窃取敏感信息，这不仅会危及船舶的安全航行，还可能导致严重的经济损失和信息安全问题。而传统的承保风险已经相对固定，且关于网络风险的市场数据较少，加之对网络风险的责任不明，网络风险的纳入存在许多问题。

首先是对网络风险的定义不明确，其外延没有明确的界限，会影响承保范围的认定，网络风险涵盖了多种类型，但目前并没有一个统一、明确的定义来界定哪些行为属于网络风险范畴。这种模糊性会影响承保范围的认定，保险公司难以准确判断哪些风险可以纳入保险保障范围，哪些风险需要排除在外；其次是现有的实践案例较少，由于网络风险是近年来才逐渐凸显的问题，相关的保险理赔案例相对匮乏，保险公司缺乏足够的数据来评估网络风险的发生概率和损失程度，从而难以制定合理的保险费率，对于可保利益的界定也存在困难，如何确定船舶所有者在网络风险中的实际损失，以及如何衡量这些损失的价值，都需要进一步的研究和明确；再次是网络风险具有多样性和延续性，其具体形态难以通过法律进行规制且具体损失数额无法即时进行计算，网络攻击造成的损失可能会在一段时间后才逐渐显现出来，如数据泄露可能导致的商业机密泄露、客户信任度下降等间接损失，这些损失数额难以在事故发生时即时进行准确计算，可能会产生后续的不利后果；最后是举证责任的承担问题，部分网络攻击可能是无痕的，被保险人需要具有一定的技术知识和网络风险的甄别能力，这无疑增加了被保险人的举证难度以及索赔成本和时间，给被保险人带来不必要的经济损失。

（三）智能船舶的碰撞责任主体不明

智能船舶作为新兴事物，在其营运过程中不可避免的产生了需要法律进行调整的主体，如智能系统的生产商以及远程操作人员等。这些新兴主体需要承担何种程度的责任尚未在法律中明确。一旦智能船舶运营后发生碰撞事故，相关主体将陷入责任的推诿中，受害人难以获得赔偿。同时对于保险公司来说若先行赔付，此后想要实现其代位求偿权也比较困难，因此难以评估智能船投保的风险，可能提高保费或拒绝承保，不利于智能船的商业化运营。因此明确船舶承运人与新兴主体的责任划分对于智能船发展至关重要。

首先，在与船舶所有人责任划分问题上，最具争议的可能就是因开发智能系统而承担产品责任的系统制造商。在商业模式上来说系统制造商可以分为两类，一种是集船舶设计、开发、制造者及承运人于一体的新型航运科技公司，

另一种是单纯生产智能船舶或者提供智能系统系统的智能船舶生产商和系统制造商。前者本身是船舶生产者和所有者，由该船舶产生的责任自然由其承担。¹³后者与船舶所有人的责任划分比较复杂，这种情况下由于智能系统系统技术具有高度的专业性，船舶所有人无法最大限度的控制和掌握船舶在专业技术方面的具体情况与详细信息。而系统制造商仅仅参与智能船舶的运营中涉及的技术问题，可能无法实时监控船舶的整体运行状态，需要双方合作才能管理船舶。在此种背景下，智能船舶生产商和智能系统系统制造商依据《民法典》侵权责任编应当承担的产品责任和船舶所有人依据《海商法》应当承担的碰撞事故产生的侵权责任的产生发生竞合，如何确定责任划分需要法律进步明确。

其次，远程操控中心的出现必然带来新的主体——远程操控人员的出现。远程操控人员根据雇佣方式可以分为两类，第一类是由承运人雇佣的受雇型远程操控人员，根据转承责任理论其因过失产生的赔偿责任最终由承运人承担；第二类是由第三方公司雇佣的独立型远程操控人员，其责任划分与其工作模式有关。若是在船舶所有人监督控制下工作，独立型远程操控人员地位类似于劳动派遣中被派遣员工，因其过失产生的侵权责任最终由船舶所有人承担。随着智能船的普及，极有可能出现第三方专业公司单独负责远程操控业务，也就是说此时独立型远程操控人员工作是在第三方公司的监督与指挥下进行。此种情况下的船舶所有人似乎早已与传统意义不同，其角色可能更接近于“资产所有者”而非“运营管理者”。但先不论智能船所有人是否愿意将价值巨大的智能船完全交给第三方运营，目前智能船距离大规模投入使用尚有待一段时间，谈论第三方专业公司单独负责远程操控业务的商业模式能否出现并广泛使用可能为时尚早，因而本文不再过多讨论。

完善智能船舶保险制度的建议

（一）延伸船舶的适航定义

从“功能等同”的视角审视，在判定船舶的海上适航性时，很难将远程岸基操控人员和智能航行系统服务的提供者这些在船舶航行中扮演核心角色的个体排除在船舶适航要素之外。远程操作中心（ROC）的这些人员实质上承担了传统船长与船员在船上的职责。因此，合理配备远程操作中心及其操作人员应被视为等同于妥善配备船员的措施。更不必论高度依赖智能航行系统实现自主化航行的智能船舶，智能航行系统的完备与否对该自主化船舶的航行安全具有重要的意义。船东在委任这些人员或者装载系统时，需确认他们具备充分的能力以驾驭船舶航行。甚至，我们可以从一些判决所揭露的信息获得支持，为使船舶适应特定贸易而在船上采取的临时装备属于“船舶”的范畴，并有助于使船舶

¹³ 吉鹏成,吴惟予:《无人船舶碰撞责任保险制度法律问题研究》,载《法学前沿》集刊2023年第4卷:新时代航运法治建设研究文集,第177页。

适航。在谷物贸易中用于货物正常运输的垫子，即使在损失发生时没有固定在船上或没有使用，仍然是“被保险船舶的配件”。¹⁴以此类推，当然可以认为，对船舶运行至关重要的控制中心和航行系统是该船舶的一部分，以确定其适航性。因此，当船东在开航前不根据船舶的不同智能程度，以安排适格的远程岸基操控人员或者完备的航行系统，船舶开航时难以保障船舶安全航行时，保险公司即可主张该船舶不适航并拒绝承担保险责任。

在当代航运业中，对于依赖船上网络实现自主航行功能的智能船舶，处理网络安全的系统被视为安全航行的重要组成部分。因此，网络安全处理系统已被广泛视为确保智能船舶安全航行不可或缺的组成部分。鉴于此趋势，船舶的适航性定义应当与时俱进，扩展至涵盖其抵御网络风险的能力。传统上，船只只在航行中借助于地图、指南针、星盘和卡尺，但今天，没有雷达、陀螺仪指南针或自动雷达标绘辅助设备，任何船只都不会考虑航行；且这些设备的配备是满足国际公约要求、确保船舶适航性的必要条件。¹⁵同理，网络作为智能船舶航行必不可少的“设备”之一，当船东采取的网络保护机制不充分时，可以认为船东没有尽到适航义务，保险公司可以基于此拒绝赔偿保险金。

（二）扩大智能船舶保险的承保范围

海事网络风险兼具聚集性和广泛联结性的特征，这使得《保险法》规定的“法律上承认的利益”变得更为空泛，在财产保险中保险利益的判断标准应当进一步结合具体的网络风险加以明确。¹⁶虽然目前大多数海上保险并不承保网络风险，但基于智能船舶的特殊性，其所面临的网络风险可以被涵盖于海上风险中，因此对于智能船舶的承保范围可以进行适当的扩大，但应当注意的是，智能船舶的网络问题产生的原因复杂多样，如果是由于非恶意的行为，如系统升级、编程错误等非人为故意的情形，若根据前文所述岸基操控人员或中心可以代替船长船员行使权利履行义务，则在一定程度上可以将其归入传统海上保险中关于船长船员疏忽或过失的保险条款，但值得注意的是岸上人员并不能像在船人员那样直面海上风险与人身安全隐患，归入前述保险条款也可能不利于保险的公平；而如果风险的发生是由于恶意行为所引起，如黑客对智能船舶的网络系统进行攻击，则需要通过新型的网络风险对其进行定义或者对海盗行为进行扩大解释以适用。针对智能船舶所产生的新型风险，保险人可以运用一些保险技术来寻求对价平衡进而承保上述风险，如提高保费、提高免赔额、设立更为严格的除外责任等。¹⁷

（三）构建智能船舶碰撞责任保险制度

¹⁴ 参见 *Hogarth v. walker*[1900] 2 QB 283

¹⁵ D. Rhidian Thomas, *Modern Law of Marine Insurance Volume Five*, Routledge, 2023, p.27.

¹⁶ 方阁，初北平：《海事网络风险保险的法律治理研究》，载《江西社会科学》2020年第5期，第186页。

¹⁷ 苑宏宇：《海上保险法领域对价平衡问题研究》，大连海事大学2021年硕士学位论文，第39页。

明确碰撞事故中的承保模式,先由保险人承担赔偿责任,若存在第三方责任,则由保险人向第三方行使代位求偿权是较为合理的承保模式。海上碰撞事故责任认定由于专业化、技术化等特点本身就较为复杂,然而在智能化、网络化的背景下由于智能系统的介入、数据依赖性增强、电子证据鉴定难度大等问题,对于事故调查取证更加困难。同时由于出现新兴主体、责任主体多元化的特点,容易出现互相推诿责任的情况。例如前文所述的智能船舶因系统出现故障发生碰撞事故时,产品责任和碰撞责任相竞合的情况。智能网联汽车也出现类似的问题,部分地方立法给出了解决方案,例如《深圳经济特区智能网联汽车管理条例》第54条、《上海市浦东新区促进无驾驶人智能网联汽车创新应用规定》第29条、《武汉市智能网联汽车发展促进条例》第37条均规定,因智能网联汽车存在缺陷发生交通事故并造成损害的,车辆驾驶人或者所有人、管理人先行赔偿后,再向生产者、销售者请求赔偿。这一规定是为了更好保障受害者的及时取得赔偿的权利。但是海上碰撞事故所产生的赔偿责任通常远大于陆上交通事故,如果同样要求船舶所有人先对外承担系统故障产生的赔偿责任,对船舶所有人苛加的负担过重。¹⁸这一问题在保险先行赔付的情况下可以得到较好的解决,保险人的先行赔付可以弥补对方的损失,此后保险人向系统制造商追偿即可。因此如果需要指定一个主体首先对外部侵权关系中的受害人承担赔偿责任,来避免责任主体之间责任推诿,受害人举证难的困境,保险人作为各责任主体之间较为强势的一方,有良好的资金赔付能力,且有专门的部门进行追偿,以其对外承担赔偿责任较为合适。¹⁹这种承保模式也是英国的《汽车技术和航空法案》《自动与电动汽车法案》均使用的“单一承保模型”。²⁰智能船舶也可以采用此种方式,当智能船舶发生碰撞事故时,可以由保险公司对第三方受害者先行赔付,然后再根据产品责任等现行法律规定向系统制造商而非船舶所有人行使代位求偿权。

四、结语

数世纪以来,保险市场一直以其创新性和解决商业问题的能力而闻名。毫无疑问,保险市场在应对新兴技术方面同样具有得天独厚的优势。现有的保险法原理并未对为采用新技术的船舶提供保险构成任何障碍,然而值得注意的是,以往为传统船舶设计的标准保险条款已难以有效应对智能船舶所带来的新型问题和风险。因此,技术革新正驱动保险行业开发一种全新的保险产品,以适应船舶智能化领域的新发展。这一创新产品应全面考虑新技术应用所带来的

¹⁸ 韩立新,姜亚迪:《无人船舶责任保险制度研究》,载《大连海事大学学报(社会科学版)》2021年第1期,第3页。

¹⁹ 吉鹏成,吴惟予:《无人船舶碰撞责任保险制度法律问题研究》,载《法学前沿》集刊2023年第4卷:新时代航运法治建设研究文集,第177页。

²⁰ 袁曾:《无人驾驶汽车侵权责任的链式分配机制——以算法应用为切入点》,载《东方法学》2019年第5期,第38页。

各类新型法律问题，并将网络风险与海上风险紧密融合，作为产品的重要组成部分。尽管目前智能船舶尚未实现大规模商业化应用，但法律创新在此领域有望超前于应用创新，为智能船舶的未来发展奠定坚实基础。

2025年中国国际法学年会

论国家主权平等原则在网络空间的适用

林世钦¹

摘要：进入二十一世纪以来，随着互联网的显著进步以及计算机和其他网络设备的日益联通，越来越多的国家正在通过网络空间发展经济、政治实力。然而，某些国家利用网络手段在其他国家实施的恶意网络活动，可能对他国主权造成严重威胁。面对这种挑战，作为“国际共识”的国际法基本原则，能够最大限度统筹各个国家的立场，对网络主权威胁达成有效的解决方案。其中，作为国际法基础原则体系的核心，国家主权平等原则强调每个国家机会平等、地位平等，是建立网络空间国际法治的基础，也是国际治理的需要，还是国家参与国际治理的前提条件。正是因为主权平等原则特殊且重要的地位，才使得其在网络空间的适用成为建立网络空间法治以及完善国际治理的重要议题。

关键词：国家主权平等原则 网络空间 对内主权 对外主权 网络治理方式

一、国家主权平等原则的内容与特性

主权平等原则的概念最初被国际社会广泛采纳是在 1928 年的“帕尔马斯岛案”²，该案仲裁裁决中对主权做出了清晰定义：“国家之间的关系相互独立，一国有权对地球上的一部分行使排他性的国家职能。”随着第二次世界大战后第三世界涌起的独立浪潮，主权平等原则成为了各国加入国际治理的“入场券”，其内容和特性不仅鼓励弱小国家行使自身的发展权益，还抑制了霸权主义国家模糊主权干涉他国内政的尝试。可以说，主权平等原则是国际社会关系正式向现代化转型的“号角”。

（一）内容：

早在 1945 年，《联合国宪章》第一章第二条规定：“本组织系基于各会员国主权平等之原则。”这是主权平等原则的内容首次以法律文件明文出现。随后在 1965 年的《关于各国内政不容干涉及其独立与主权之保护宣言》明确指出国家主权平等，强调各国的内政不容干涉，独立与主权应受到保护，后续例如《美洲国家组织宪章》、《亚非会议最后公报》等国际法文件虽然也对主权平等原则进行了一定程度的细化和拓展，但总体上没有脱离《国际法原则宣言》设立的基本内容。

笔者认为，根据主权平等原则的内容，大致可以分为对内主权平等和对外主

¹ 福州大学研究生 林世钦 18050899061

² 王军敏.帕尔马斯岛案对国际法的影响[J].理论视野,2009,(09):50-51.

权平等两大原则。

1.对内主权平等

在讨论对内主权平等之前,需要先对“对内主权”做出合理的解释。一般认为,“对内主权”指国家在本国领土范围内拥有最高和最终的政治权威,具有绝对的统治力。在此基础上,笔者认为,对内主权平等主要包含国家的最高统治权与公民的基本人权之间的平等、国家管辖权与公民自由权的平等。

学界对于“对内主权”有众多解释,但对于“对内主权平等”的研究极少。究其原因,传统学说认为主权是相对于国际社会其他国家而言的,对于国内领域,主权至高无上,自然也不存在平等的说法。笔者认为,虽然主权的至高无上是为了证明其统治国家的合理合法,但对于国家公民而言,国家主权可能会侵害自身基本权利。因此,“对内主权平等”是指在涉及公民基本权利的范畴时,国家不能以主权至高无上一言以蔽之,需要切实考虑公民的基本人权,必要时可以牺牲部分国家权益。例如:当发生重大社会事件时,公民需要发表自身意见,国家不能以危害国家主权为由禁止公民行使言论自由的权利。

虽然“对内主权平等”在国际法范畴的适用情境较少,但是会间接影响一个国家在国际社会的平等独立地位,如果一个国家在国内以主权频繁侵害公民人权,将会透支自身的国际形象,遭到国际社会其他国家的孤立和排挤,因此对内主权平等不容忽视。

2.对外主权平等

对外主权指国家间相互尊重、彼此独立和平等相待,意味着本国具有保持独立平等,不受任何侵犯的权利,即其他国家和组织都必须尊重一国的政策和选择,不得随意干涉³。该内容可以从“独立”“平等”两大部分进行理解,作为国际社会的主权国家必须先获得独立的国际法律地位,不受他国在经济、政治方面之间或间接的控制,才能在国际治理、国际组织等方面与其他国家一样享受平等的权利。具体表现为:(1)平等的代表权,各国在国际组织中享有平等的代表权,可以平等地参与决策;(2)平等的发言权,各国在国际会议和论坛中享有平等的发言权,可以自由表达自己的立场和诉求;(3)平等的经济权利,各国在国际贸易、投资等领域享有平等的权利,例如最惠国待遇原则。

(二) 特征:

国际法的基本原则与国内法基本原则在具体案件的适用中有所不同,国内法的基本原则起到价值引领作用,基本原则的精神体现在法律规则中,在具体案件中一般不引用基本原则作为审判依据。但在国际社会,由于没有凌驾于国家至上的权利,因此国际法基本原则作为各国形成的共识,在解决具体国际争议中能够发挥重大作用。主权平等原则强调的主权是一个国家在国际的立国之本,缺失

³ 王丽莉. 国际法原则在网络空间国际治理中的适用[D]. 沈阳师范大学, 2023.

主权的国家就难以称作独立的现代国家，因此主权平等原则即使在国际法基本原则中也是核心的地位。此外，主权平等原则适用于所有主权国家，不仅能够为各国主张平等主权提供理论支持，还能够制止某些不正当限制国家主权的行为。

1.核心性

主权平等原则作为国际法体系的核心基石，贯穿于国际法的历史发展、规范构建和实践运作中。其核心地位不仅体现在国际法的基本原则体系中，更反映在对国际秩序的根本性塑造和对国家间关系的根本性规范。

首先，在法理方面，主权平等原则的逻辑起点具有正当性。主权平等原则赋予国家独立地位，确保国际法在分散权力结构中可操作。《维也纳条约法公约》第 34 条：条约非经第三国同意，不为其创设义务，体现主权国家意志优先性。主权平等赋予国家独立地位，国家自愿行为国际法效力背书。否定主权平等，国际法将成强权政治附庸，丧失合法性基础。

其次，在规范体系方面，主权平等原则具有核心表现。（1）主权平等原则是其他国际法原则的逻辑前提。例如，主权平等原则的逻辑延伸是不干涉内政原则，主权国家的独立性和自主性要求其他国家不得干涉其内政事务，包括政治、经济和文化等领域。（2）主权平等原则是国家基本权利的来源。国家的权力在国际法中得到确认和保障，而主权平等原则正是国际法的基础。主权平等原则作为国际秩序的基石，它确保国家之间的平等和稳定，国家的权力正是在这种秩序中得以行使和维护。（3）国际组织的权力来源于成员国主权的部分让渡，但其行动不得超越授权范围。例如，世界卫生组织（WHO）可制定全球卫生标准，但无权强制成员国执行；国际刑事法院（ICC）的管辖权需以国家同意为基础（《罗马规约》第 12 条）。

2.优先性

国际法基本原则相较于国际法规则：（1）在规范层级方面，原则对规则进行指导与约束。第一，国际原则为规则的解释提供价值导向，当规则存在模糊或冲突时，需依据原则进行目的解释。国际法院就曾在“核武器使用合法性咨询意见”（1996）中，以“人道主义原则”解释《联合国宪章》禁止使用武力条款，强调武力使用需符合必要性及相称性。第二，当规则缺位时，原则可直接作为裁判依据，这也是国际法原则与国内法原则最大的不同点。在“科孚海峡案”（1949）中，国际法院依据“人道主义原则”和“国家不得故意允许其领土用于损害他国权利的行为”，填补了当时海洋法规则的不明确性。第三，原则可防止规则被机械适用而导致不公正结果。尽管《联合国宪章》第 51 条赋予国家自卫权，但国际法院在“尼加拉瓜诉美国案”（1986）中强调，自卫行动必须符合“必要性”和“相称性”原则，避免滥用规则。（2）在实践应用方面，国际原则相较于国际规则的优先性具有诸多具体体现。第一，强行法原则（如禁止酷刑、禁止奴隶制）高于一切条约

和习惯规则，任何与之冲突的规则均无效。《维也纳条约法公约》第 53 条就规定：“若某国签订允许种族隔离的条约，该条款因违反强行法而自始无效。”第二，尽管“特别法优于一般法”是规则冲突的解决原则，但特别法不得违反基本原则。WTO 争端解决机制中，环保措施（如《濒危野生动植物种国际贸易公约》）的适用需符合“非歧视原则”等 WTO 基本原则。第三，国际法院、国际刑事法院等常将原则作为裁判的核心依据。前南斯拉夫国际刑事法庭（ICTY）在“塔迪奇案”（1995）中，依据“人道主义原则”将武装冲突中的性暴力定性为反人类罪，推动国际刑法规则的发展。

3. 普遍约束性

国际法中的主权平等原则具有普遍约束性，其效力覆盖所有主权国家，无论国家的发展水平、政治制度或国际地位如何。这种普遍性源于主权平等原则作为国际法体系的基石地位，其约束力通过法律渊源、国家实践、国际司法判例及强行法规范得以确立。

（1） 法律渊源方面，主权平等原则在国际法体系中具有普遍效力

《联合国宪章》第 2 条明确将“主权平等”列为联合国及其会员国必须遵守的基本原则。截至 2023 年，联合国会员国已达 193 个，几乎涵盖全球所有主权国家，使得主权平等原则成为最广泛接受的国际法规范。主权平等原则不仅是条约法的基础，更是被普遍认可的习惯国际法。即使非联合国会员国或未签署特定条约的国家，仍需遵守主权平等原则，因其已成为国际社会的共同法律确信和国家实践。国际法院在“北海大陆架案”（1969）中指出，习惯国际法对所有国家具有约束力，除非国家一贯明确反对。此外，主权平等原则中的核心内容（如禁止侵略、领土完整）属于强行法规范，具有最高法律效力，任何与之冲突的条约或行为均属无效。

（2） 国家实践方面，主权平等原则具有得到普遍遵守的现实基础

首先，国家在缔结条约、参与国际组织或制定外交政策时，均以主权独立为前提。世界贸易组织（WTO）的“一揽子承诺”机制要求成员国自主决定开放领域，体现对国家经济主权的尊重。其次，国际社会对武力改变领土现状的行为普遍谴责，并通过联合国机制实施制裁。2022 年俄罗斯对乌克兰的军事行动引发 141 国支持联合国大会 ES-11/1 号决议，要求俄撤军并尊重乌克兰主权。以上情况证明在当今国际治理的环境，主权平等原则已经是所有国家默认遵守的前提。

（3） 主权平等原则的普遍约束性也得到国际司法确认

国际法院多次通过判例重申主权平等原则的普遍约束力：科孚海峡案（1949）：确认国家对其领土的排他控制权，不得容忍他国侵犯。尼加拉瓜诉美国案（1986）：裁定美国支持反政府武装构成非法干涉，违反主权平等原则。查戈斯群岛咨询意见（2019）：要求英国尊重毛里求斯领土完整，结束殖民统治。

二、网络空间的概念与特性

（一）概念：

2009 年公布的《网络空间政策评估》中将网络空间定义为，包括“互联网、电信网络、计算机系统和嵌入式处理器、控制器组成的相互依赖的信息基础设施”。

北约网络合作防御卓越中心特邀国际专家组编写的《网络行动国际法塔林手册 2.0 版》，其给出的网络空间定义相对中立且简单，即“网络空间又译为赛博空间，由物理和非物理组件构成，利用计算机网络存储、修改和交换数据的环境”⁴。

综上，本文认为，网络空间是指包含网络服务器、各种终端设备（包括联网芯片）等网络硬件基础设施，以及基于这些物理硬件而产生的社会交往关系的总和。基于此定义，网络空间按照功能可以划分为物理、协议和内容三个层面。物理层主要指服务器、计算机、手机等基础物理设施，是整个网络空间结构中最基础且最重要的部分；协议层主要指各种网络传输协议和标准，物理硬件在协议的支持下得以联通，为网络空间的实时互联提供保障；内容层是人类在网络空间以“网络人格”生产创造的言论、图片、影像等具有网络传播性的内容成果。

（二）特性：

1.物理性

网络空间的构建需要大量的物理设备作为载体，如服务器、路由器、交换机、光纤等。这些设备分别承担算力、数据存储和数据交换等基础性功能，它们的运行状态直接影响网络空间的稳定性和可用性。

此外，网络空间虽然不存在现实世界中的高度、深度和长度，但它具有几何性质。电磁频谱作为网络空间的基础，具有可以测量的物理范围，可以用能量、波长或频率等术语来表述，这种几何度量性使得网络空间可以呈现各种不同形式。

不仅网络空间所需要的硬件设备和传输媒介具有物理性，这些硬件设备的分布也具有现实地理特征。网络元件在陆地、空中、海上或太空中的位置构成了网络空间的地理组件，这些位置关系影响着网络的传输效率和安全性。

2.无边性

在现实世界中，从疆域分界线到领海分界线，从降水分布线到地势等高线，再从地球经纬线到国际日期变更线，或天然或人为，都可以通过“划线”以明晰等级和边界。但在网络世界，这种情况发生了变化。

首先，网络空间打破了现实世界中的地域、国家、时间的疆界，使得人类活动在很大程度上突破了现实疆域的限制。在网络空间中，划分区域的标准是 IP 地址和与之对应的域名，而不是传统的地理边界。

其次，网络空间中的信息流动不受地理边界的限制，网络空间中的信息传播几乎是实时的，用户可以在瞬间将信息发送到全球各地，并可以在任何地方访问

⁴ 刘碧琦. 国际法在网络空间的适用问题研究[D]. 武汉大学, 2020.

全球的网络资源，为全球用户提供了平等的访问机会。

3. 虚拟性

网络空间的虚拟性是相较于现实生活中任何有形物而言的，虽然上文提到网络空间的基础设施、传播媒介和地理分布都具有物理性，但在普通民众眼中，硬件设施仅仅意味着数据载体和互联网终端入口，至于传输媒介和地理分布更是抽象的概念，与网络空间呈现出的丰富内容大相径庭。

归根结底，普通民众眼中的“网络空间”更聚焦于网络空间非现实性的一面，人们可以在网络空间中回溯过往的信息，可以突破地理限制进行实时沟通，甚至可以通过数字技术模拟一个全新的数字身份，在网络空间过上另一种生活。这些与传统现实相抽离的情形无不凸显网络空间的虚拟性特征。

（三）网络主权

随着网络空间对国际社会的发展与稳定产生日益重要的影响，互联网领域立法已将设立网络主权确立为核心原则，这不仅可以视为传统的主权平等原则在新兴网络领域的有效延伸，还能嗅见各主权国家在网络空间中寻觅重新洗牌国际地位的野心，具有极强的政治意义。因此，由于各国在网络空间意识形态方面存在不同的立场，实际上都制定了维护自身主权的法律政策。

1. “主权”的概念和内涵

近代国家主权体现了两个关键理念：一方面，主权是国家的固有属性。国家作为一个独立的政治实体，拥有管辖权和自卫权，能够在自己的领土内独立行使主权。另一方面，国家的边界具有稳定性。边界是国家主权的物理界限，主权的范围由边界所限定。国家边界不应因王朝战争或主权掌控者的个人原因而重新或随意划定。⁵

在网络空间里，互联网的互联互通特性对国家管控跨境行为的能力构成了挑战，进而深刻影响了国家的主权。针对网络空间对传统主权理论的突破，世界不同国家提出了不同的主张，有的国家力求彻底摆脱传统主权，实行互联网“无主权”形态，有的国家认为应该在网络空间划定“主权范围”以保护国家利益免受网络空间入侵带来的负面影响。不同的网络主权理念不仅代表各国在网络空间的对外政策倾向，还影响了网络空间技术发展占国家战略的权重，进而影响综合国力的发展方向。

2. 中国的“网络主权”理论

根据相关网络主权内涵的研究与我国网络空间治理基本原则，当今主流的中国“网络主权”理念定义为：一国以该国所管辖的信息通信技术系统为基础，以能够直连他国网络设备的该国网络设备端口集合为作用边界，拥有对构成该国网络

⁵ 熊光清，王瑞. 全球互联网治理中的网络主权：历史演进、国际争议和中国立场[J]. 学习与探索, 2024, (04): 24-33+176.

空间的平台、平台上的行为活动及产生的相应数据的司法与行政管辖权利，有权对该国境内的网络设施、网络主体、网络行为及相关网络数据和信息等制定法律法规。

在军事方面，允许一国拥有保护本国网络空间不受非法入侵和攻击的防御权利和军事能力。在网络空间国际关系中，一国可以平等参与网络空间国际治理，同他国共同制定网络规则而不受排斥。主权国家之间不干涉彼此网络发展道路、治理模式和公共政策，不进行网络入侵活动⁶。

3.其他国家的“网络主权”理论

美国以“互联网自由”为核心，主张开放网络空间，反对“网络主权”概念，认为其将导致互联网割裂与封闭。2011年，奥巴马政府发布《网络空间国际战略》，提出“基本自由、隐私、信息自由流动”三原则，批评他国防火墙等安全措施阻碍开放互通的网络生态。2023年，美国担任“自由在线联盟”主席国，聚焦抵制数字威权主义、规范人工智能应用及增强数字包容性，推动全球网络人权保护。其立场反映对技术霸权与意识形态输出的双重诉求。

欧盟为降低对美国数字技术依赖，近年力推“数据主权”“数字主权”及“技术主权”战略。2018年实施《通用数据保护条例》（GDPR），以严苛罚则强化个人数据跨境管控，要求所有与欧盟交易方遵守其规则，并重新定义数字主权——承认超国家机构可共享权力。2021年发布《2030年数字十年政策方案》，加速本土数字技术研发与网络安全建设，旨在摆脱美技术垄断，构建自主可控的数字生态。

三、主权平等原则在网络空间的适用困境

2007年“爱沙尼亚网络战争”、2010年“Stuxnet 震网病毒袭击”、2017年“NotPetya 网络袭击”、2018年“脸书泄密事件”，以及众多低烈度网络攻击行动的接连发生。因此，为了应对网络空间对于主权的威胁，自2010年代以来，网络空间已经步入国际法治阶段。然而，作为国际法基本原则的核心，主权平等原则中的对内主权和对外主权在网络空间的适用过程中均出现了诸多问题。

（一）网络空间管辖乏力

对内主权作为面向国内至高无上的统治权，对境内事物行使管辖权自然包含在对内主权的权力范围内。然而，上文提到网络空间的硬件设施、传输媒介和地理分布具有物理性，既然是物理性事物，就必然存在国家管辖的问题。此外，网络空间的无边性使得管辖不能只着眼于物理性事物，还要看到背后无形的数据，而数据并不受限于现实世界的地理疆域分布，这是关于网络空间管辖的最大争议点。

1.网络空间的物理性导致对内主权不能充分管辖

上文说到，网络空间的构建需要大量的物理设备作为载体，如服务器、路由

⁶ 孙梦寒. 中国“网络主权”理念国际认同的影响因素与提升路径研究[D]. 北京邮电大学, 2024.

器、交换机、光纤等。以服务器为例，全球共有 13 台 IPv4 根服务器，其中 1 台主根服务器位于美国，由美国互联网机构 Network Solutions 运作。其余 12 台均为辅根服务器，其中 9 台在美国，2 台在欧洲（位于英国和瑞典），1 台在亚洲（位于日本）。所有 IPv4 根服务器均由美国政府授权的互联网域名与号码分配机构 ICANN 统一管理，负责全球互联网域名 IPv4 根服务器、域名体系和 IP 地址等的管理。而美国政府又享有对 ICANN 根区的全面政策权威，这就使得任何对根区文件的修改，均需经获得美国商务部审核同意。因此，美国可以凭借对根服务器的绝对掌控使得其他主权国家的“对内网络管辖权”形同虚设。2003 年伊拉克战争中就出现伊拉克全境网络瘫痪；2005 年，美国终止对利比亚网络的解析服务，使得利比亚网络瘫痪，这无不展示了美国在网络空间里“单边主权”的力量。除了根服务器，在传输媒介方面，全球光缆业务几乎被西方发达国家所主导⁷。

这些案例都表明，网络空间的物理性决定了各国对网络空间的管辖权天生就是不平等的，管辖权不平等也深刻影响着主权不平等。任何一个主权国家想要对内行使管辖权要满足两个条件：其一是该管辖权高过其他任何国内权力；其二是其他国家不可进行干涉。在网络空间这两个条件同样可以适用，但现实情况往往不尽人意。首先，网络空间的管辖权与技术实力挂钩，外部黑客侵入国家数据的案例比比皆是，并不是所有国家都有能力及时修复网络漏洞并长期维护国内网络安全性。其次，网络技术落后的国家需要网络发达国家的帮助以搭建国内网络，这种帮助的代价就是根服务器、传输材料等关键硬件由他国掌握，国际关系风谲云诡，这无异于将国内网络空间的命脉交予他人之手，很难说拥有实质的网络空间管辖权。

2. 网络空间的无边界性导致对内主权的管辖边界模糊

从静态的角度来看，网络空间的区域划分标准是 IP 地址和与之对应的域名，对应的是上文“网络空间概念”的协议层面，这些协议标准本身就不带有主权色彩，使得国家想以管辖权“划分”网络空间使之与现实主权范围一一对应的想法缺乏支撑。

从动态的角度来看，网络空间中的信息可以实时传播到到全球各地，全球用户也都可以平等访问网络信息，对应的是上文“网络空间概念”的内容层面。网络空间的“内容”可以视为人类自我意识的数字化表达，既然是个人自我意识的体现，那么每个人不仅可以决定在网络空间创造的内容是什么，还能够决定内容发往何处。例如跨国聊天虽然是实时沟通，但是行为的发生地却同时包含了现实世界中主权范围不同的两个国家。如果国家尝试管辖网络空间中内容的生产和传播，运用何种理论来提供管辖的正当性是需要解决的首要问题。

3. 总结

⁷ 沈逸. 网络主权与全球网络空间治理[J]. 复旦国际关系评论, 2015, (02): 22-32.

网络空间的物理性和无边界性都是国家管辖面临的挑战。但可以明确的是，网络空间的管辖并不适合“一刀切”的解法。比如，美欧国家曾主张：主权国家有权对在境内产生实质性影响的网络活动行使属地管辖权。这种“唯后果论”当将可能导致管辖权的无限扩张，进而引发诸多问题⁸。结合多种管辖原则分别适用于复杂多样的网络空间情形，才是解决网络空间管辖模糊的有效做法。

（二）国家以维护网络空间主权平等的名义侵犯人权

当今时代，许多公民权利，例如言论表达自由、和平集会与结社自由等，常常借助网络得以实现。在网络空间中，虽然不存在对身体、健康等实体人身的直接威胁，但对人格尊严、隐私、言论自由等权利的侵犯行为却经常发生。本文讨论的主要是国家在网络空间行使主权侵犯本国国民人权的情形。

高洪柱在 2012 年的演讲中提到：“支撑互联网和网络活动的物理基础设施通常位于主权国家领土内，受领土国家的管辖。”然而，他进一步强调：“但领土国家行使的管辖权并非绝对无限的；它必须符合可适用的国际法，包括国际人权义务。”⁹。现实中，一些国家常常以反恐等借口，援引国家主权的概念，对公民进行过度审查。因此，各国在网络空间援引主权平等原则时应注意符合该国在国际人权法下所承担的可适用义务。

以著名的“棱镜门”事件为例，NSA 主要通过一个代号为 PRISM 的监视项目来实现对全球通讯的秘密监视。根据相关文件的泄露，NSA 通过 PRISM 项目获得了来自包括谷歌、脸书、微软、雅虎等公司的用户数据，并对其进行了大规模的监视和收集。NSA 通过 PRISM 项目获得的信息包括电子邮件、语音通话、视频、照片、文件等。更为重要的是，PRISM 项目的实施主要依靠美国《外国情报监视法》（FISA）及其修订版《外国情报监视法修正案》的授权，这也就意味着该窃听计划有着代表美国主权的美国法律作为背书，这也印证了当国家主权的无形巨手侵入网络空间时，普通网络用户的人权如同草芥。

综上，在国际人权保护的语境下，国家维护网络空间主权虽有必要，但不能背离初衷。主权虽是至高无上的统治权，但在社会契约中，国家、政府的所有公共行为都需要符合将它托举出来的人民群众的利益。诚然，人民群众在选举政府换来公共生活便利的同时需要牺牲自己的权力，但是这种牺牲应该以基本人权作为底线。国家本应通过确立网络主权排除任何对本国人民权益的侵害行为，但如果国家以“主权至上”的名义随意侵犯人权，这便是一个“屠龙者终成恶龙”的故事。

（三）主权平等原则适用于网络空间的可行性存在争议

网络空间侵害主权行为判定标准，就是要回答“国家主权是否能够适用于网

⁸ 孙尚鸿. 传统管辖规则在网络背景下所面临的冲击与挑战[J]. 法律科学(西北政法大学学报), 2008, (04): 160-168.

⁹ 布莱恩 J. 依根, 冯野, 应瑶慧. 国际法与网络空间的稳定[J]. 汕头大学学报(人文社会科学版), 2017, 33 (07): 132-138+5.

络空间”，对于这个问题，主流观点分为三个：

第一，网络空间无主权论。该理论主张国家在网络空间不享有主权，认为人权高于主权，应保障人们在网络空间的自由权利，一旦国家主权介入网络空间治理，就会限制人们在网络空间的自由。例如，2018年5月，英国司法部长杰里米·赖特在公开指出，主权不是一项具体规则，也无法从主权平等原则中“推断”出来。换句话说，除非构成其他国际不法行为，否则主权平等原则能够被违反，且不会产生国际法律责任¹⁰。

第二，网络空间绝对主权论。该观点主张网络主权应与现实国家主权别无二致，主张将国家主权照搬进网络空间。这种观点一般为网络技术落后的国家所主张，由于自身网络水平落后，本国网络空间难免受到其他强国直接或间接的控制，主张网络主权与现实主权对等将有利于其在网络领域获得独立自主地位，免受他国舆论和意识形态干涉。

第三，网络空间相对主权论。该观点认为：国家在网络空间中仍然享有主权，但这种主权是相对的，而非绝对的。国家的主权在网络空间中受到技术、国际法和全球化等因素的影响，呈现出多维性和动态性。

（四）主权平等原则难以融入国际网络治理方式

国际网络治理方式，主要涉及“各国在国际网络空间治理应该扮演什么角色”的问题。“各国”指的是拥有独立主权并得到国际社会广泛认可的国家，“治理”这一词是从决策者的角度出发如何管理并解决网络空间大大小小的问题。因此，该问题天然面临着“各国”的平等性与“治理”的决策者居高临下性之间冲突。如果只有少数国家参与决策，那么是否违背了国家之间的平等地位？如果所有主权国家一起参与决策，如何商讨出一个令所有人都满意的结果以及如何解决效率问题，都是亟待解决的问题。

当下主流的网络空间治理模式为欧美主推的“多利益攸关方治理模式”，该模式首次被记录在《突尼斯议程》中，主要内容是强调政府、私营部门和民间团体等各方共同发挥作用，共同制定互联网规则，通过多方参与、协商和合作来解决网络空间复杂问题。

落实到实践中，存在三种主流“多利益攸关方治理模式”：ICANN模式、联合国框架下的互联网治理论坛（IGF）以及信息社会世界峰会（WSIS）。其中，ICANN是由美国提出并主导的非营利性公益机构，成立于1998年，参与者包括全球范围的国家政府、非政府组织、民间团体甚至个人。该组织曾通过与美国商务部签订协议，获得了根区服务器的管理权。IGF（Internet Governance Forum）不是决策机构，而是一个开放的讨论平台，参与者同样来自社会各界，每年举行一次年度会议，会议期间会讨论各种互联网治理议题。WSIS（World Summit on the

¹⁰ See Jeremy Wright, Cyber and International Law in the 21st Century, Gov.UK., 23 May, 2018, <https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>.

Information Society)是一个由联合国发起并主办的多阶段峰会,第一阶段于2003年12月在日内瓦举行,第二阶段于2005年11月在突尼斯举行,旨在国家、区域和国际层面高效商讨和解决由信息通信技术(ICT)引发的各类问题。

诚然,“多利益攸关方治理模式”的许多出发点都是好的。比如该治理模式不排除国家政府以外的实体参与网络空间治理,此低门槛要求允许更多社会实体为网络空间治理贡献多方力量。此外,该治理模式做出的决策都是面向世界的,透明的决策直面各界监督以保证决策公正性。再比如,该治理模式的决策方式是参与主体达成共识,并不形成强制性决策。

虽然现行的网络治理方式勉强在各主权国与国际决策之间找到了一个平衡点,但仍然难以说它就是完美的解决方式。反之,“多利益攸关方治理模式”更像是一把“双刃剑”,在多年的实践中已经褪去光泽,反而像背离初衷的方向脱缰而去。

四主权平等原则在网络空间适用的解决方案

(一) 多种管辖原则综合运用

面对网络空间物理性和无边界性造成的管辖权困境,上文我们已经批判了“唯后果论”将会造成管辖权泛化的结果,进而会造成各国管辖权冲突更加激化。基于此,既然各国行使管辖权的目的是稳定国内网络空间秩序,那么采取保守的管辖权措施可以在专注于管辖本国网络空间的同时避免与他国管辖权产生冲突。

在讨论保守的管辖权之前,我们应该明确:管辖权可以进一步分为属地管辖权、属人管辖权、保护性管辖权以及普遍管辖权¹¹。其中,属地管辖权规定一个国家有权对在其领土范围内的一切人、物和事件行使管辖权。基于这个定义,保守的管辖权策略中“专注于本国网络空间”的主张与属地管辖原则不谋而合。讨论属地管辖权的适用,需要分成两个层面来看。

第一层面,对于构成网络空间的硬件设备(例如计算机、路由器、光纤),应以硬件所在地来判断是否对其具备管辖权,传统主权理论认为:国家对其领土范围内的物品拥有管控和治理的权力。网络硬件设备作为整个网络空间少有的有形体部分,根据硬件所在地决定管辖权归属,不仅能够赋予国家对内进行有效管辖的物质基础条件,还能够在法理上阻断网络强国通过硬件实施网络霸权的设想。此外,对于在本国境内从事网络活动的人员,国家同样可以主张属地管辖权,在一国境内从事网络活动的人员与其他从事传统活动的人员在本质上并无不同,只是活动内容和方式有所区别。他们同样需要遵守所在国的法律法规,接受所在国的管辖。

第二层面,国家对在境内产生或完成的网络活动享有管辖权。《塔林手册2.0版》规定了国家可以对境内网络设施、从事网络活动的人员,以及在境内产生或

¹¹ 杨秀琴:《税收相关法律》,中华工商联合出版社2001年版,第122页

完成，或者在境内造成实质影响的网络活动行使管辖权¹²。首先，因为判断一项网络活动是否产生或完成往往缺乏明确的标准，存在较大的模糊性，并且网络活动的结果是发散的、不确定的，在网络活动结果所在地行使管辖权需要耗费大量人力物力，还会受到结果所在国的阻碍，并且在法理上占不到优势。这便是属地管辖权选择以网络活动的发生地为准的理由。此外，不能完全排除网络活动的结果发生地，倘若确定结果发生地在本国境内，不论基于属地原则还是最密切联系原则，都说明该活动对本国产生的影响最大，自然可以行使属地管辖权。

总而言之，保守的管辖权策略选择以属地管辖原则为主，这可以让国家专注于境内网络空间，将有限精力最大限度用在维护国内网络秩序稳定，同时还要注意排除侵权结果发生地在境外的情形。当然，属地管辖权的行使还应以网络活动与相应国家存在真实且充分的联系为前提。

（二）守住保护人权原则的底线

国家强权与公民人权的斗争作为亘古不变的议题，今天在网络空间领域也同样上演着。之所以要划定国家在网络空间适用主权平等原则的同时保护人权的底线，是因为历史证明有太多普通人的基本人权在国家“利维坦”贯彻网络空间主权时被视作草芥。大众熟知的“棱镜门事件”(PRISM)，便是美国国家安全局(NSA)自2007年小布什时期起开始实施的一项绝密电子监听计划，该计划允许NSA和FBI直接进入包括微软、谷歌、苹果、雅虎等在内的9家国际网络巨头的服务器收集用户的电子邮件、聊天记录、视频、照片等数据。单就普通的一天NSA就收集到“444743份来自雅虎的、105068份来自Hotmail的、82857来自Facebook的、33697来自Gmail的，还有22881来自不能确定的提供商处的”用户通讯录¹³。

由于讨论的是网络空间范畴的人权保护，笔者将从网络空间最重要的三种人权进行讨论，分别是言论自由权、隐私权和被遗忘权。

1. 言论自由权

言论自由权是公民的基本权利之一，旨在保障公民能够自由地表达自己的思想、观点和意见，而不受政府或其他权力的不当限制。在形式上，言论自由权包括但不限于发表言论、集会、游行、示威、出版、艺术创作等形式。在传播媒介方面，言论自由权可以通过各种媒介行使，包括新闻、书籍、广播、电视等，本文提及的网络空间亦是言论自由权重要的行使媒介之一。

国家法范畴中对言论自由权同样重视。联合国《公民权利和政治权利国际公约》第三编第19条规定：“人人有发表自由之权利；此种权利包括以语言、文字或出版物、艺术或自己选择之其他方式，不分国界，寻求、接受及传播各种消息

¹² 甘勇.《塔林手册2.0版》网络活动国际管辖权规则评析[J].武大国际法评论,2019,3(04):117-135.

¹³ 孙冰.中国确认棱镜项目对华窃密：微软谷歌都有配合.中国经济周刊,2014,06,03,<http://www.chinanews.com/mil/2014/06-03/6236003.shtml>

及思想之自由。”2012年7月，联合国人权理事会通过第20/8号决议，明确该文件中关于言论自由的规定同样适用于网络空间。2014年6月，联合国人权理事会通过第26/13号决议，呼吁各国保障公民在互联网上的言论自由、结社自由以及隐私权，并履行国家在国际人权保护方面的义务。

在网络空间，国家如果不想让主权成为公民行使言论自由的阻碍，最重要的是完善相关法律。其一是提升立法层级，将网络言论自由的法律保障提升到宪法层面，确保其权威性和重要性。同时，制定或修订相关法律法规，明确网络言论自由的定义、性质、限制和保障措施，为实践提供明确的法律依据。其二是确保法律的可预测性，保证法律清晰、可预测，公民可以通过阅读法条明白其中所表达的意思，避免因法律模糊而导致的不确定性。

2. 隐私权

隐私权是个人对其私人信息和生活拥有控制权的权利，包括个人信息、私人活动和私人空间等。网络隐私权是隐私权在网络环境下的延伸，主要包含个人信息的保护、个人行为的保密和个人数据的安全。

人格尊严人人平等，每个人应受到的人权保护力度也应该平等，普通人也应该享受等同于身居要职的官员的信息保护力度。更何况国家对人权的保护绝不止于消极不侵犯，更要积极去维护。这就要求国家在网络空间主张主权平等的过程中，不仅要做好不侵犯公民网络隐私的自觉，还要加强防护以避免公民隐私遭受第三方侵犯。

目前在隐私权保护方面，美国和欧盟建立了相对完善的个人数据隐私权保护法律体系。美国采用分散性立法模式，将相关内容分布在多个独立法规中。其立法对公权和私权领域的个人数据采取不同态度：对公权领域的数据保护主要依靠行业自律，而对私权领域的数据保护则更多体现国家强制力的干预。欧盟的立法则以保护自然人的网络基本权益和数据隐私权为基本宗旨，保障个人数据在欧盟成员国之间的自由流动，并强调保护个人数据的完整性¹⁴。

3. 被遗忘权

被遗忘权（Right to be Forgotten）是指个人有权要求搜索引擎或其他数据控制者从其搜索结果或数据库中删除特定的个人信息，使其不再被公众获取或传播。相较于上述两种权利，被遗忘权更强调公民主动主张权利。与之对应，国家如果接收到公民的被遗忘权主张，不应让国家主权成为被遗忘权实现的阻碍，还要积极推动其实现。

被遗忘权的法律保护在不同国家和地区存在差异。欧盟通过《通用数据保护条例》（GDPR）等立法明确了被遗忘权的法律地位。美国则在特定领域有相关立法和实践，如加州的“橡皮擦法案”，允许未成年人删除其在社交媒体上发布的内

¹⁴ 蔡俊. 网络主权的法律问题研究[D]. 武汉大学, 2019.

容。

在中国，虽然尚未有明确的被遗忘权法律规范，但个人信息保护的相关立法和司法实践为引入被遗忘权提供了可能性。随着大数据时代的到来，被遗忘权的构建成为国内外学者和法律界关注的热门议题。

4.总结

综上，在讨论主权平等原则在网络空间应该如何适用的同时，不要忘记保护人权原则也是国际法基本原则之一，通过国家自身的努力和国际合作、条约和监督机制，国际社会正在不断推动人权保护的实现与进步。

（三）明确主权平等原则在网络空间的应用

关于主权平等原则能否在网络空间适用，是本文的核心问题，倘若主权平等原则不能够在网络空间适用，那么前文所述均已失去实际价值。因此，笔者认为，主权平等的原则能够在网络空间得到应用应成为全球共识。

首先，无主权论在网络空间“拓荒期”大行其道，与美国的“西部拓荒”类似，在新事物伊始的无序阶段，早入局者需要该观点作为在“新兴领域蓝海”大肆获利的理论支撑。但是，与“西部拓荒期”结束后的美国类似，世界上并不存在绝对的自由，一旦新事物的参与人数逐渐饱和，就意味着权利矛盾将逐渐激化，早入局者或许可以守住既得利益，但是如果想凭借“无限自由”理论凭借早入局的优势地位侵占他人权益，将受到他人的联合抵制。正因为如此，建立合理的法律体系才是稳定社会秩序的关键所在，也是新兴事物成熟的象征。无主权论实际上是网络霸权国家试图凭借其在网络技术的优势地位继续从他国榨取利益的借口。从微观角度，无主权造成了网络弱国在与网络强国的谈判中毫无回旋的余地，没有主权这一基本支撑，国无国格，将会使得弱小国家的网络空间成为强国的“赛博殖民地”；从宏观角度这将会加剧网络世界主权不平等，造成网络主权比重与现实主权比重的割裂。

因此，唯有在网络空间也能适用主权平等原则才能保证国际社会的健康发展。然而，支持主权适用的观点中，绝对主权论太过激进，缺乏可行性。

绝对主权论虽然出于朴素的国家平等观点，但在实践过程中却存在许多矛盾之处。首先，强国帮助弱国完成网络搭建不可能是为了做慈善，国际关系的利益驱动性决定了强国一定想从弱国获取好处，即使不是想掌控弱国的网络意识形态，也可以利用弱国国内的空白网络市场榨取经济利益，因此想要以主权为理由主张网络独立不切实际。其次，网络空间的无边界性也决定了网络主权无法“绝对化”，除非是通过技术手段完全封锁，但这也失去了网络最珍贵的“互联互通”价值，于自身并无好处。网络弱国需要网络强国的软硬件扶持是现实所需，绝对主权将使得弱国得到网络建设援助的可能性进一步降低，实际上是网络弱国“自断后路”。

作为折衷的相对主权论先让更具备实操性。例如，对于网络主权中至关重要

的管辖权，该观点认为国家仍然可以基于属地原则对其境内的信息和通信技术（ICT）基础设施以及网络行为行使管辖权。同样的，上文关于管辖权的问题在这里也同样存在，因此这种观点中网络主权的“相对性”便体现在除了考虑自身可享有的管辖权之外，不能侵犯他国的管辖权，这与上文提到的管辖权“纯后果论”导致本国管辖权与他国管辖权相冲突的局面形成鲜明对比。该观点的优点在于看到了网络空间主权的动态性，能够在尊重主权的同时，既承认国家在网络空间中的主权地位，又适应了网络空间的全球性和无边界性特点，通过国际法和多边合作来解决网络空间中的管辖权冲突，为网络空间的治理提供了新的视角。

因此，笔者认为，相较于无主权论和绝对主权论，相对主权论作为主权平等原则在网络空间适用具有正当性的根据更为妥当。

（四）多方共治以排除非国家实体的主导

在“多利益攸关方治理模式”中，参与决策的门槛虽然降低，但也导致参与决策的队伍鱼龙混杂，许多个人或者组织人微言轻，提出的建议不能得到平等的对待，相反更多的提案目的并不是为了营造更好的国际网络空间，而是背后势力为了自身利益另有所图，这种无效提案大大降低了决策质量，有时还会适得其反。其次，虽然决策结果对外透明，但是难免受到政治色彩的“染色”。以 ICANN 为例，在其取得根服务器管理权之后，在伊拉克战争期间仍然出现了全境网络瘫痪的事件，这不免让人担心所谓的“公正决策”其实只是强国实现霸权意志的“白手套”。最后，由于决策达成的是共识而非强制性规定，这在本就没有上位权力存在的国际社会根本没有效力可言，如果决策没有执行的强制保障，那么也就失去了它所存在的意义。

事实上，随着新兴国家网络实力的崛起带动的国际网络治理格局的变化，这种微妙的“平衡点”已悄然发生偏移，这就需要找到一种新的同时满足主权国家参与感和治理决策有效性的治理模式。

多边治理模式存在多种方案，其中具有代表性的主要有三种，分别是针对 ICANN 治理模式的改进、改革和颠覆：

第一种方案主张对 ICANN 治理模式进行改进。该方案通过分离权限的方式削弱 ICANN 的职能，对 ICANN 和 Verisign 公司管辖的顶级根区服务器进行限制，并确保 ICANN 独立管辖服务器，同时保障发展中国家各相关方的参与权。

第二种方案对 ICANN 治理模式的改革则更为激进。该方案并不完全反对美国在网络空间的霸权地位，而是希望从中分得一杯羹，让主权国家在网络空间获得与其地位相匹配的管理权限¹⁵。

第三种方案主张由国际电信联盟接管网络空间治理。这一主张旨在打破网络发达国家一家独大的现状，重塑网络空间的国际秩序，因此遭到了美国的强烈反

¹⁵ 王丽莉. 国际法原则在网络空间国际治理中的适用[D]. 沈阳师范大学, 2023.

对。在各种理论思潮涌动的大背景下，中国创新提出了“多边参与、多方参与”的共治模式，该模式坚持主权国家主导，避免非国家行为体过度参与¹⁶。该模式有效弥合了西方国家所主张的治理模式所造成的理论与实践之间的割裂。2017年3月1日，我国在《网络空间国际合作战略》中指出：“网络空间应遵循共同治理原则。作为国际社会的平等成员，各国都有权通过国际治理机制和平台，平等参与网络空间国际秩序和国际规则的建设。”这表明我国公开支持网络空间治理应采用国家主导下的多边治理模式。这种模式是最能适应当前网络发展状况、真正实现平等治理且最为有效的治理模式。

结语

随着网络空间的不断发展与演变，国家主权平等原则在网络空间的适用面临着诸多挑战与困境。然而，这一原则作为国际法的核心基石，其在网络空间的合理适用对于维护国家主权、保障网络安全以及构建公正合理的网络空间国际秩序具有至关重要的意义。通过对主权平等原则在网络空间适用困境的深入剖析，以及对多种解决方案的探讨，能够为网络空间国际法治的推进提供有益的理论参考。

从网络空间管辖乏力到网络主权与人权的权衡，从网络空间侵犯主权利行为危害性判定标准到主权平等原则难以融入国际网络治理方式，这些困境反映出网络空间国际治理的复杂性与紧迫性。面对这些问题，多种管辖原则的综合运用、人权原则的坚守以及多方共治模式的探索，为解决网络空间中的主权平等原则适用问题提供了可能的路径。

在未来的网络空间国际治理中，各国应秉持开放合作的态度，通过多边协商与合作，共同构建符合时代发展需求的网络空间国际规则体系。同时，必须充分认识到人权在网络空间中的重要性，确保网络空间的治理既尊重国家主权平等，又保障公民的基本权利。此外，多方共治模式的推进应注重平衡国家与非国家行为体之间的关系，避免非国家实体在网络空间治理中的过度主导，从而实现网络空间的和平、稳定与可持续发展。

综上，国家主权平等原则在网络空间的适用是一个动态发展的过程，需要国际社会的共同努力与智慧。只有在尊重国家主权平等的基础上，充分发挥国际法的规范作用，才能在网络空间中实现真正的公平与正义，为全球网络空间的繁荣与稳定奠定坚实的基础。

¹⁶ 王向阳：《网络空间治理的国际规范研究》，载《情报杂志》2021年第7期，第80-85页。

ON THE APPLICATION OF THE PRINCIPLE OF SOVEREIGN EQUALITY IN CYBERSPACE

【ABSTRACT】

Since the beginning of the 21st century, with the remarkable progress of the Internet and the increasing connectivity of computers and other network equipment, more and more countries are developing their economic and political power through cyberspace. However, malicious cyber activities carried out by certain countries in other countries by means of cyber means may pose a serious threat to the sovereignty of other countries. In the face of such challenges, the basic principles of international law, as an "international consensus", can coordinate the positions of various countries to the greatest extent and reach an effective solution to the threat of cyber sovereignty. Among them, as the core of the system of basic principles of international law, the principle of sovereign equality of states emphasizes the equality of opportunity and status of all countries, which is the basis for the establishment of international rule of law in cyberspace, the need for international governance, and the prerequisite for state participation in international governance. It is precisely because of the special and important status of the principle of sovereign equality that its application in cyberspace has become an important topic for establishing the rule of law in cyberspace and improving international governance.

【KEYWORDS】 the principle of sovereign equality of states cyberspace
internal sovereignty external sovereignty internet governance models

神经数据保护的法理基础与规范构造

汪雨馨¹

摘要：在全球科技高速发展的背景下，神经技术作为新兴技术产物逐渐引入人们的视野，但随之而来的风险也越发凸显。神经技术对人脑中的神经数据进行提取、传输、解码与加工这一过程使神经数据从大脑完全隐蔽的储存空间进入到公共空间。神经数据若在这一处理过程中遭受到脑黑客袭击，或未进行匿名化处理的泄露等问题，会对脑隐私、神经自主权等带来风险。本文明析神经数据的属性与其保护的法理基础，综合了国内外学者有关神经数据保护的相关学术观点，并提出了全新规范模式以保护神经数据，旨在构建一个符合技术发展且灵活使用的规范框架。

关键词：神经技术 神经数据 神经权利 规范构造

一、引言

神经数据(neurodata)是指直接通过监测大脑活动所收集的数据集合，这些数据提供了有关大脑结构和功能的信息²，其是通过侵入性或非侵入性神经技术收集的，这些技术采用不同方法，对人类大脑进行记录和/或提取图像³。而“神经技术”指的是与大脑和整个神经系统相互作用的各种方法和工具，或者被动地——通过监测大脑活动——或者主动地——通过改变这种活动。这些技术包括神经成像技术、神经调节技术、脑机接口和神经干细胞⁴。在过去的二十年里，处理大量数据（大数据）的新可能性和人工智能工具的使用推动了神经技术的进步，这使得以前所未有的速度分析此类数据变得更加容易，并在一定程度上允许识别神经活动模式，并间接地“解读”思想⁵，而神经数据就具有读取人的思想并预测其意图的特性⁶。

随着脑机接口等神经技术的进步，神经数据的获取变得更加便捷，随之带来了隐私和安全等保护挑战，其中包括：人们的精神隐私与精神完整性在神经数据的收集、传输与处理过程中受到威胁；人的精神自主性伴随外界神经数据的写入而受到干扰；此外，在获取神经科学数据时，也存在平衡个人隐私保护与科

¹ 汪雨馨，中南大学法学院 2024 级国际法学硕士研究生，邮箱：2954229651@qq.com。

² D. Hallinan, P. Schütz, M. Friedewald, P. Hert, Neurodata and Neuroprivacy: Data Protection Outdated?, in Surveillance & Society, vol. 12, n.1, p. 55.

³ D. Hallinan, P. Schütz, M. Friedewald, P. Hert, Neurodata and Neuroprivacy, cit., P. 58

⁴ Cenolli I. Literature Review: Neurotechnology[J]. 2024.

⁵ Andorno, R. (2023). Neurotecnologías Y Derechos Humanos En América Latina Y El Caribe: Desafíos Y Propuestas De Política Pública. University of Zurich; UNESCO.

⁶ MacKellar, C. (2019). Cyborg mind: what brain-computer and mind-cyberspace interfaces mean for Cyberneuroethics. Oxford, New York: Berghahn Books, 49-57.

学研究需要的问题⁷。

面对神经数据的独特挑战，国际社会一些组织和国家率先采取了倡议措施或立法，以探索保护路径。2024 年联合国教科文组织拟定的《神经技术伦理建议书（初稿）》第 49 条指出，神经数据的收集、处理、修改和共享必须在征得自由和知情同意的情况下进行，在方式上必须遵守本建议书概述的伦理和人权原则。⁸国际公益组织神经权利基金会 2024 年发布的报告《保护大脑数据:评估消费者神经技术》中认为神经数据的保护随着神经技术的发展具有了迫切性。⁹2025 年 3 月 12 日，联合国人权理事会第 58 届会议收到一份报告，建议制定神经技术和数据处理的监管框架，以确保负责任使用并强化知情同意教育。¹⁰

除国际组织的倡议外，各国也在立法层面积极探索神经数据保护的具体规范。以智利为代表的国家选择通过另立规范规范神经数据的保护。其国民议会通过并由总统签署的宪法修正案，“神经权利”作为新的人权列入宪法，与该修宪行为相对应，国会目前正在考虑一项超越宪法修正案广泛原则宣言的法案，该“神经保护”法案规定，所有神经技术设备都要受到与医疗设备相同的监管，神经数据将被视为等同于人体器官，禁止购买或出售此类数据。¹¹与智利不同，以巴西为代表选择在原有的法律框架的基础上进行修改以达到规范神经数据的目的。2021 年开始，巴西国民议会开始分析修改 2018 年 8 月 14 日第 13709 号法(LGPD)的提案，目的是将神经数据概念化，并规范与保护相关的程序和措施。2022 年提交给众议院的第 522/2022 号法案案旨在规范对神经数据的使用和处理的保护。¹²科罗拉多州通过《隐私法》将神经数据纳入“敏感数据”范畴，采取与巴西类似的改造路径。¹³

由此可见，国际社会在神经数据保护的具体规范上的分歧愈发显著：智利以另立规范凸显其独特性，但其严格性可能限制神经技术的商业应用与创新；而巴西和科罗拉多州则通过改造现有框架追求灵活适用，但现有框架可能难以应对神经数据涉及的“思想隐私”等独特风险，保护力度或显不足。

与国外的现存规范相比，我国对于神经技术与神经数据的保护仍处在倡议阶

⁷ Cornejo Ya. (2024). Neurorights, Neurotechnologies and Personal Data: Review of the Challenges of Mental Autonomy. *Journal of Digital Technologies and Law*, 2 (eng) (3), 711-728.

⁸ 联合国教科文组织,《神经技术伦理问题建议书（初稿）》
https://unesdoc.unesco.org/ark:/48223/pf0000391444_chi, 最后访问日期: 2025 年 3 月 25 日。

⁹ Genser J, Damianos S, Yuste R. Safeguarding Brain Data: Assessing the Privacy Practices of Consumer Neurotechnology Companies[J]. 2024.

¹⁰ 联合国专家呼吁监管神经技术以保护隐私权,
<https://www.ohchr.org/en/press-releases/2025/03/un-expert-calls-regulation-neurotechnologies-protect-right-privacy>, 最后访问日期: 2025 年 3 月 25 日。

¹¹ E. Strickland and M. Gallucci, "First Win for the Neurorights Campaign: Chile plans to regulate all neurotech and ban the sale of brain data," in *IEEE Spectrum*, vol. 59, no. 1, pp. 26-58, January 2022.

¹² Rainey S and Dalese P. An alternative focus on data in the neurorights discussion – Lessons from Brazil [version 2; peer review: 1 approved, 3 approved with reservations]. *Open Res* 2024, 1:3

¹³ 科罗拉多州隐私法将神经数据定义为受保护的敏感数据, Cenolli I. Literature Review: *Neurotechnology*[J]. 2024, 最后访问日期: 2025 年 3 月 24 日。

段，且主要集中在脑机接口技术领域。近日，《可穿戴式脑机接口专用脑电采集分析系统技术要求及测试方法》T/CAS 976—2024 团体标准（以下简称“标准”）正式发布，该标准为中国首个可穿戴脑机接口产品标准，规范了“可穿戴式脑机接口专用脑电采集分析系统”的技术要求及测试方法，填补脑机接口领域标准空缺¹⁴。2024 年工信部等七部门联合印发《关于推动未来产业创新发展的实施意见》，脑机接口就作为“创新标志性产品”之一名列其中。国家科技伦理委员会下属的人工智能伦理分委员会也研制编写了《脑机接口研究伦理指引》，为脑机接口的研究提供了重要参考，此外，2023 年脑机接口产业联盟在《脑机接口伦理治理倡议书》（征求意见稿）中提出了脑机接口伦理治理的十条准则，从伦理层面规范了脑机接口的发展¹⁵。

这种分化不仅源于法律传统的差异，更反映了对神经数据风险认知的分歧，从而引发学界与实践界的激烈争论。究竟哪种路径能更好地平衡隐私保护与技术创新，仍是亟待解决的核心问题。

二、神经数据保护的应然性与技术背景争议

如前文所述，神经数据作为量化脑结构、活动和功能的信息，因其与心理状态的直接关联，被认为具有高度敏感性¹⁶¹⁷。面对这一新兴挑战，如何构建有效的法律和伦理保护路径，成为学术界和政策制定者的核心议题。本节首先围绕神经数据与心理隐私的理论争议与技术能力现状评估的不同观点，聚焦怀疑派、保守派和自由派三种立场，进行探讨与分析。

（一）神经数据解码对心理隐私产生威胁的分歧

神经数据保护的理论争议源于对神经数据对心理隐私是否产生威胁的不同理解。心理隐私通常指个体对其神经数据及衍生心理信息的控制权¹⁸¹⁹。然而，神经数据是否等同于心理信息并对心理隐私产生威胁，以及是否具有特殊保护的应然性，成为争议的起点。这一分歧直接决定了后续保护路径的理论基础。

1、怀疑派：神经数据解码对心理隐私无独特威胁。

怀疑派否认神经数据对心理隐私构成独特威胁，认为其无需特别保护。哲学上，Gilead 主张，主观心理状态因其不可访问性，无法被技术直接读取，神经数据仅反映第三人称层面的脑活动，而非心理内容²⁰。基于此，Ryberg 指出，人类

¹⁴ <https://www.ttbz.org.cn/StandardManage/Detail/131192>，最后访问日期：2025 年 4 月 7 日。

¹⁵ 崔中良, 朱晨恩. 当前我国脑机接口技术的伦理进展与治理应对[J/OL]. 中国医学伦理学, 1-12.

¹⁶ Ienca, M., et al. (2022). Towards a governance framework for brain data. *Neuroethics*, 15(2), 20.

¹⁷ López-Silva, P., et al. (2024). Neurotechnological applications and the protection of mental privacy: An assessment of risks. *Neuroethics*, 17(31).

¹⁸ Wajnerman Paz, A. (2022). Is your neural data part of your mind? Exploring the conceptual basis of mental privacy. *Minds and Machines*, 32, 395-415.

¹⁹ Lavazza, A., & Giorgi, R. (2023). Philosophical foundation of the right to mental integrity in the age of neurotechnologies. *Neuroethics*, 16(10).

²⁰ Gilead, A. (2015). Can brain imaging breach our mental privacy? *Review of Philosophy and Psychology*, 6(2), 275-291.

通过日常行为推断他人意图的“理论心智”并未引发伦理危机，技术性读心不过是类似过程的延伸，不应被特殊化²¹。实证上，Meynen 强调，当前神经技术的解码受限于实验条件，难以威胁真实隐私²²。因此，怀疑派认为，神经数据仅是个人数据的一种，现有隐私框架已足够²³。这一立场将神经数据置于传统隐私概念之下，质疑任何额外保护的必要性。

2、保守派：神经数据的保护融入既有隐私保护框架

与怀疑派不同，保守派承认神经技术对心理隐私的潜在风险，但主张通过现有框架的调整即可应对，而非另起炉灶。Hertz 提出，《欧洲人权公约》（ECHR）第 9 条的“思想自由权”可扩展解释为保护所有心理过程，包括神经数据²⁴。Bublitz 进一步论证，欧盟《通用数据保护条例》（GDPR）的“敏感数据”类别中如健康数据已涵盖神经数据，无需新权利²⁵。Susser 和 Cabrera 补充，隐私权的本质在于保护“人”而非特定技术，神经数据与基因或行为数据并无本质差异，均可受 ECHR 第 8 条“私人生活权”约束²⁶。保守派批评自由派的新权利提案为“权利膨胀”，可能削弱现有体系的效力²⁷²⁸。相较于怀疑派的彻底否定，保守派在承认风险的基础上，强调现有框架的适应性。

3、自由派：神经数据的特殊性需要新权利来保护

自由派则从神经数据的独特性出发，认为现有法律不足以应对其新风险，主张引入神经权利。Ienca 和 Andorno 提出，神经数据因直接关联心理状态、具备高时间分辨率和读写能力，超越传统数据范畴，需通过心理隐私权等新权利加以保护²⁹。López-Silva 等人补充，神经数据比基因数据更具行为预测性，且可实时操控心理状态，现有法律（如 GDPR）未预见此类威胁³⁰。Lavazza 和 Giorgi 从哲学视角强化这一观点，强调神经技术威胁到意向性、第一人称视角和道德自主性等传统不可及的心理特征，需心理完整性权等新机制³¹。自由派以智利宪法改革（2021）和巴西数据保护法修正案（2021）为实践依据，呼吁全球神经特异

²¹ Ryberg, J. (2017). Neuroscience, mind reading and mental privacy. *Res Publica*, 23(2), 197-211.

²² Meynen, G. (2019). Ethical issues to consider before introducing neurotechnological thought apprehension in psychiatry. *AJOB Neuroscience*, 10(1), 5-14.

²³ Bublitz, J. C. (2022). Novel neurorights: From nonsense to substance. *Neuroethics*, 15(1), 7.

²⁴ Hertz, N. (2023). Neurorights – Do we need new human rights? A reconsideration of the right to freedom of thought. *Neuroethics*, 16(1), 5.

²⁵ Bublitz, J. C. (2022). Novel neurorights: From nonsense to substance. *Neuroethics*, 15(1), 7.

²⁶ Susser, D., & Cabrera, L. Y. (2023). Brain data in context: Are new rights the way to mental and brain privacy? *AJOB Neuroscience*, 1-12.

²⁷ Hertz, N. (2023). Neurorights – Do we need new human rights? A reconsideration of the right to freedom of thought. *Neuroethics*, 16(1), 5.

²⁸ Ligthart, S., et al. (2023). Minding rights: Mapping ethical and legal foundations of ‘neurorights.’ *Cambridge Quarterly of Healthcare Ethics*, 1-21.

²⁹ Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, 13(1), 5.

³⁰ López-Silva, P., et al. (2024). Neurotechnological applications and the protection of mental privacy: An assessment of risks. *Neuroethics*, 17(31).

³¹ Lavazza, A., & Giorgi, R. (2023). Philosophical foundation of the right to mental integrity in the age of neurotechnologies. *Neuroethics*, 16(10).

性框架³²。与保守派的渐进策略相比，自由派突出神经数据的突破性挑战，要求法律体系的根本性变革。

怀疑派以心理状态的不可访问性为前提，否定神经数据的特殊性；保守派则在承认风险的基础上，依托既有框架的灵活性；自由派从数据的独特属性和潜在威胁出发，推导新权利的必要性。这一界定分歧不仅是概念之争，更奠定了技术评估和法律路径的逻辑基础。

（二）技术能力现状评估的立场争议与“读心”炒作

界定分歧进一步延伸至对神经技术“读心”能力的评估。学者们在技术现状、未来潜力及其实用性上的不同判断，直接影响保护路径的紧迫性和方向。是否存在神经技术“读心”炒作，成为连接理论与实践的关键争议。

1、怀疑派：神经技术读心的局限否定其威胁性

怀疑派认为，“读心”能力被严重夸大，无需特别应对。Gilbert 和 Russo 分析 569 篇学术文章，发现 91%提及读心可能性，但仅 46%认为已实现，且定义混乱³³。Brown 指出，强读心（精确解码语义和实时思想）因脑个体差异和解释差距而遥不可及，fMRI 等技术需受试者配合³⁴。Mecacci 和 Haselager 补充，实验生态效度低，无法推广至非自愿场景³⁵。怀疑派据此质疑自由派基于“读心”威胁的新权利需求，认为其建立在虚构前景上。这一立场将技术局限作为否定特殊保护的依据，与界定分歧中的“无特殊性”论点一脉相承。

2、保守派：谨慎评估下的渐进应对

保守派对技术能力持谨慎乐观态度，认为当前技术未达读心水平，但未来潜力值得关注。Hertz 承认 fMRI 和 BCI 可推断部分心理状态，但强调其依赖行为数据和上下文，尚不足以构成独特威胁³⁶。Lighthart 等人分析欧洲案例法，指出神经数据采集已有法律先例，现有框架可逐步适应³⁷。相较于怀疑派的彻底否定，保守派在技术评估上更具开放性，认为无需新权利，但需根据发展调整法律。这种立场与保守派在隐私界定中的“融入既有框架”策略逻辑一致，注重现实与未来的平衡。

3、自由派：未来潜力驱动的预防性保护

自由派聚焦技术潜力，强调其可能实现的全面读心威胁，主张预防性立法。

³² López-Silva, P., et al. (2024). Neurotechnological applications and the protection of mental privacy: An assessment of risks. *Neuroethics*, 17(31).

³³ Gilbert, F., & Russo, I. (2024). Mind-reading in AI and neurotechnology: Evaluating claims, hype, and ethical implications for neurorights. *AI and Ethics*, 4, 855-872.

³⁴ Brown, C. M. L. (2024). Neurorights, mental privacy, and mind reading. *Neuroethics*, 17(34).

³⁵ Mecacci, G., & Haselager, P. (2019). Identifying criteria for the evaluation of the implications of brain reading for mental privacy. *Science and Engineering Ethics*, 25(2), 443-461.

³⁶ Hertz, N. (2023). Neurorights – Do we need new human rights? A reconsideration of the right to freedom of thought. *Neuroethics*, 16(1), 5.

³⁷ Lighthart, S., et al. (2023). *Minding rights: Mapping ethical and legal foundations of ‘neurorights.’* Cambridge Quarterly of Healthcare Ethics, 1-21.

Ienca 预测, AI 与 BCI 融合将使实时心理状态解码成为现实³⁸。Yuste 和 Quadra-Salcedo 补充, 潜意识活动也可被解码, 威胁自主性³⁹。Gilbert 和 Russo 虽承认当前局限, 但认为 AI 和 BCI 的“炒作曲线”正上升, 可能很快超越预期⁴⁰。自由派据此提出新权利, 如智利法案禁止非自愿神经数据采集。与保守派的渐进态度相比, 自由派的技术评估更具未来主义色彩, 直接支持其界定中“特殊性需新权利”的结论。

怀疑派以技术局限否定炒作的现实性; 保守派承认炒作但认为可控; 自由派则将炒作视为前瞻性保护的依据。Gilbert 和 Russo 发现, 21% 文章质疑读心炒作, fMRI 炒作下降, AI/BCI 炒作上升⁴¹。这一分歧不仅反映技术判断的差异, 更将界定争议延伸至法律策略的实践层面。

三、神经数据保护具体规范的构造与伦理争议

(一) 具体保护规范的构造分歧

基于上述对于神经数据属性的界定和技术背景评估的分歧, 三种不同立场促使在神经数据保护规范的选择与构造上形成截然不同的观点。

1、怀疑派: 现有有关法律框架的保护范围能够涵盖神经数据, 无需进行额外的更改或立法。

怀疑派主张, 神经技术对于神经数据所带来的威胁程度被过分夸大了, 实际上现有的法律体系已经足够应对相关的风险。Bublitz 认为, GDPR 已经涵盖健康和生物识别数据, 而神经数据的特征与属性可以将其纳入到 GDPR 特殊数据的保护范围内, 因此神经数据无需特殊处理⁴²。Ryberg 则指出, 类似于数字画像一类的隐私侵害问题早已存在, 神经技术对于神经数据的产生的威胁归根到底还是隐私威胁, 其应用并未引入新问题⁴³, 所以无需单独干预。怀疑派强调, 立法应基于实际的证据与数据, 而非有关问题的假设, 这样才能避免不必要的法律资源浪费。这一规范构造方式的选择直接来源于他们对神经数据面临风险的问题的界定, 即他们认为神经数据没有特殊性, 以及在技术评估中认为并没有带来现实的威胁。其立场更倾向于维持现状, 不主张进行大规模的法律或政策变革。

2、保守派: 调整现有法律规范框架而不是重新设立新权利

保守派倾向于在承认潜在风险的前提下, 主张通过将现有的法律框架进行解释或改造来应对这些挑战。Hertz 提出建议, 他主张将欧洲人权公约 (ECHR) 中

³⁸ Ienca, M. (2021). On neurorights. *Frontiers in Human Neuroscience*, 15, 701258.

³⁹ Yuste, R., & Quadra-Salcedo, T. (2023). NeuroRights and new charts of digital rights: A dialogue beyond the limits of the law. *Indiana Journal of Global Legal Studies*, 30(1), 15-37.

⁴⁰ Gilbert, F., & Russo, I. (2024). Mind-reading in AI and neurotechnology: Evaluating claims, hype, and ethical implications for neurorights. *AI and Ethics*, 4, 855-872.

⁴¹ Gilbert, F., & Russo, I. (2024). Mind-reading in AI and neurotechnology: Evaluating claims, hype, and ethical implications for neurorights. *AI and Ethics*, 4, 855-872.

⁴² Bublitz, J. C. (2022). Novel neurorights: From nonsense to substance. *Neuroethics*, 15(1), 7.

⁴³ Ryberg, J. (2017). Neuroscience, mind reading and mental privacy. *Res Publica*, 23(2), 197-211.

的第9条“思想自由权”所适用的范围扩展至脑部空间及神经数据，并辅以软法来进一步解释规范的适用范围⁴⁴。另一方面，Ligthart 等人建议，在现有的 GDPR 框架中明确神经数据为“特殊类别数据”的地位⁴⁵，以明确适用 GDPR 来应对未来可能产生的风险。此外，保守派还对自由派提出的神经权利概念持批评态度，他们认为这种做法可能会导致法律体系的碎片化，并且在实际操作中会遇到难以克服的难题。⁴⁶⁴⁷。此种规范构造的观点既体现了他们在界定问题时所倡导的“融入现有法律框架”的理念，也反映了其对在技术评估中所持的“谨慎乐观”的态度。这种规范构造设想的核心在于寻求法律的渐进性与实用性的平衡，以确保其适应性和有效性。

3、自由派：通过新的法律来填补保护神经数据规范的空白

自由派坚持认为，规避神经数据可能承受的独特风险远远超出了现有法律的覆盖范围，因此需要制定新的权利来应对这些风险。Ienca 和 Andorno 两位学者认为，心理隐私权的概念应当提出，从而来保护个人免受未经同意的神经数据采集和使用行为的侵害⁴⁸。Lavazza 则主张应当确立心理完整权，以此来防范任何形式的非自愿脑干预行为⁴⁹。López-Silva 等人建议将神经数据单独视为一种特殊敏感的数据类型加以规定，从而加强对其的保护措施⁵⁰。自由派倡导以智利的神经保护法案为蓝本，呼吁在全球范围内建立一个全新的法律框架。这种规范构造的设想逻辑基础在于他们对神经数据“特殊性”的界定，以及在技术评估中对“未来威胁”的预见，旨在强调神经数据现在以及未来可能面临的风险需要进行预防性的变革来应对。

（二）具体保护规范构造下的伦理争议

从伦理的角度来看，上述的分歧被进一步深化，这涉及到隐私的本质问题、技术与社会风险以及新权利双重性的考量，从而更深入地揭示了神经数据保护规范构造过程中所存在的内部逻辑。

1、有关隐私本质的哲学分歧

在神经技术的快速发展的背景下，Lavazza 和 Giorgi 两位学者强调了神经技术对个体意向性及第一人称视角隐私存在潜在威胁，并呼吁建立新的权利以保护

⁴⁴ Hertz, N. (2023). Neurorights – Do we need new human rights? A reconsideration of the right to freedom of thought. *Neuroethics*, 16(1), 5.

⁴⁵ Ligthart, S., et al. (2023). Minding rights: Mapping ethical and legal foundations of ‘neurorights.’ *Cambridge Quarterly of Healthcare Ethics*, 1-21.

⁴⁶ Bublitz, J. C. (2022). Novel neurorights: From nonsense to substance. *Neuroethics*, 15(1), 7.

⁴⁷ Susser, D., & Cabrera, L. Y. (2023). Brain data in context: Are new rights the way to mental and brain privacy? *AJOB Neuroscience*, 1-12.

⁴⁸ Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, 13(1), 5.

⁴⁹ Lavazza, A. (2018). Freedom of thought and mental integrity: The moral requirements for any neural prosthesis. *Frontiers in Neuroscience*, 12, 82.

⁵⁰ López-Silva, P., et al. (2024). Neurotechnological applications and the protection of mental privacy: An assessment of risks. *Neuroethics*, 17(31).

个体的“心理领域”。他们认为，随着神经技术的进步，个体的心理状态和思维过程可能面临被外部解读的风险，这要求法律和社会提供新的保护机制⁵¹。Wajnerman Paz 提出质疑，如果神经数据构成了个体心理身份的一部分，则可能需要超越传统的身体完整性保护范畴。她主张，心理身份的保护可能需要一种全新的法律框架来应对⁵²。然而，Gilead 作为怀疑派，提出了反驳意见，他指出主观体验的不可知性使得神经技术对个人隐私的威胁被过分夸大，由于我们无法完全理解他人的主观体验，因此神经技术带来的隐私侵犯风险可能并不像某些人所担忧的那样严重⁵³。这一系列的分歧和争议，实际上承接了关于界定神经技术影响范围的争议，进而影响了法律路径的哲学基础，即如何在尊重个体隐私和促进技术发展之间找到平衡点。

2、有关技术与社会风险程度争议

López-Silva 等人提出五种神经技术风险包括但不限于歧视和强制自证等。他们主张需要新设立的权利来防范这些技术的滥用⁵⁴以达到保护神经数据的目的。与此同时，Brown 认为数字读心技术已经带来了与神经技术相似的风险，因此，对于神经技术的使用，也需要一个统一的责任框架来规范⁵⁵。Gilbert 和 Russo 则发出警告，他们认为过分炒作读心会带来社会风险可能会误导伦理讨论，并且可能会导致人们忽视一些现实上的问题⁵⁶。这场关于神经技术的争议，不仅连接了技术评估的领域，而且也凸显了在社会风险方面存在的分歧。

3、有关神经权利的双重效应

自由派倾向于将视神经权利视作保护神经数据的一种关键机制⁵⁷，但 Fins 通过对智利医疗改革的深入分析，提出了一个具有启发性的观点，指出改革措施可能在无意中对那些无法自主表达同意的患者，如慢性肌营养不良（CMD）患者，造成限制。⁵⁸这些患者由于疾病的影响，可能无法清晰地表达自己的意愿，从而在改革中处于不利地位。Fins 的这一观点揭示了神经权利在实际应用中可能产生的双刃剑效应，即在保护一部分患者的同时，也可能对另一部分患者造成限制。另一方面，保守派则强调，现行的伦理原则，尤其是知情同意原则，已经是一种

⁵¹ Lavazza, A., & Giorgi, R. (2023). Philosophical foundation of the right to mental integrity in the age of neurotechnologies. *Neuroethics*, 16(10).

⁵² Wajnerman Paz, A. (2022). Is your neural data part of your mind? Exploring the conceptual basis of mental privacy. *Minds and Machines*, 32, 395-415.

⁵³ Gilead, A. (2015). Can brain imaging breach our mental privacy? *Review of Philosophy and Psychology*, 6(2), 275-291.

⁵⁴ López-Silva, P., et al. (2024). Neurotechnological applications and the protection of mental privacy: An assessment of risks. *Neuroethics*, 17(31).

⁵⁵ Brown, C. M. L. (2024). Neurorights, mental privacy, and mind reading. *Neuroethics*, 17(34).

⁵⁶ Gilbert, F., & Russo, I. (2024). Mind-reading in AI and neurotechnology: Evaluating claims, hype, and ethical implications for neurorights. *AI and Ethics*, 4, 855-872.

⁵⁷ Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, 13(1), 5.

⁵⁸ Fins, J. J. (2022). The unintended consequences of Chile's neurorights constitutional reform: Moving beyond negative rights to capabilities. *Neuroethics*, 15(3), 26.

成熟的机制，能够有效地平衡个体利益与潜在风险之间的关系⁵⁹⁶⁰。保守派的这一立场强调了对传统伦理原则的尊重和维护。

这场关于神经权利的讨论，不仅仅局限于伦理原则的探讨，它还承接了法律策略的讨论。法律专家和政策制定者在制定相关法律和政策时，必须考虑到这些伦理原则如何在实际操作中得以体现，并且要预见不同路径选择可能带来的伦理后果。这不仅关系到法律的公正性，也关系到社会对弱势群体的关怀和保护。因此，这场讨论不仅揭示了在选择不同路径时可能面临的伦理后果，也反映了社会在进步中不断寻求平衡个体权利与社会责任的复杂过程。

四、神经数据的法律属性探析及保护的法理基础

神经数据作为一项具有广泛争议的保护对象，明确其法律属性是构建其保护规范的先决条件。在深入探讨神经数据的保护路径之前，必须首先对其法律属性进行细致的分析和界定。

（一）神经数据的法律属性探析

神经数据的法律属性是保护规范的基石，其界定不仅关乎其在法律体系中的定位，还直接影响适用规则的设计和保护路径的选择。通过对现有研究的系统梳理，神经数据呈现出生物性、敏感性和心理性的递进特征，既与传统个人信息保护框架存在交集，又因其与思想和人格的直接关联而突破常规界限。这种多重属性的层次性要求法律定位超越单一归类，以动态、综合的视角回应其复杂内涵。

首先，神经数据的生物性构成了其作为人体信息延伸的基础属性。这一特性源于其采集方式和技术基础，通常通过脑电图（EEG）、功能性磁共振成像（fMRI）或脑机接口（BCI）等手段直接记录大脑的生理活动，具有客观可测性。侵入式脑机接口利用微电极阵列或深部脑刺激技术，能够捕获高精度的神经信号，用于治疗帕金森病或癫痫等神经疾病⁶¹。这种数据的生物来源使其与基因数据或健康数据共享相似的生理根基，揭示了个体的身体状态和功能特征。其生物性为法律适用提供了初步锚点，使其在个人信息保护体系中具备可操作性。然而，若仅将其法律属性限定于生物性，可能忽视其超越生理层面的更深意涵，导致保护范围过于狭窄，无法应对技术进步带来的新挑战。

在此基础上，神经数据的敏感性进一步凸显了其超越传统个人数据的隐私风险，构成其法律属性的第二层次。敏感性源于其高时间分辨率和潜在的预测能力，使其不仅限于静态的生理记录，还能动态反映个体的行为倾向、情感状态甚至决策过程。闭环脑机接口通过实时解码神经信号，已实现高达 90 字符/分钟的打字

⁵⁹ Hertz, N. (2023). Neurorights – Do we need new human rights? A reconsideration of the right to freedom of thought. *Neuroethics*, 16(1), 5.

⁶⁰ Lighthart, S., et al. (2023). Minding rights: Mapping ethical and legal foundations of ‘neurorights.’ *Cambridge Quarterly of Healthcare Ethics*, 1-21.

⁶¹ Zhao, Z.-P., et al. (2023). Modulating Brain Activity with Invasive Brain-Computer Interface: A Narrative Review. *Brain Sciences*, 13:134.

速度，展现了其在控制外部设备时的惊人潜力⁶²。这种技术特性不仅提升了医疗和商业领域的应用价值，也显著增加了数据滥用的可能性。研究指出，神经数据市场在神经营销或消费者行为分析中的广泛应用，进一步放大了隐私威胁⁶³。与传统个人信息（如姓名、地址）相比，神经数据的敏感性在于其可能暴露个体不愿公开的内在状态，甚至在未经同意的情况下被用于推断其偏好或意图。现有法律虽涵盖健康数据，却未明确神经数据的特殊敏感性，导致保护力度不足，可能威胁个体隐私权乃至思想自由⁶⁴。敏感性使神经数据从一般的个人信息升格为高度敏感信息，促使法律体系应对其采集、存储和使用的全过程施加更严格的约束。更深一层，神经数据的心理性构成了其最独特且最具争议的法律特征，直接关联思想和人格的核心领域，成为其属性的第三层次。通过量化脑功能和神经活动，神经数据能够映射个体的心理状态，例如情感波动、认知偏好甚至潜意识倾向，这使其远远超越基因数据或行为数据的间接性⁶⁵。例如，钙成像技术记录的神经信号可用于分析小鼠的行为偏好，而类似技术在人类中的应用可能揭示更复杂的心理内容⁶⁶。这种心理性使其可能触及传统上被视为不可侵犯的心理疆界，威胁到个体对自身思想的控制权和隐私权。有学者提出，神经数据不仅涉及生理信息，还可能影响心理隐私、个人身份和自由意志，因而需要通过新的权利形式加以保护⁶⁷。然而，这种特性也带来了新的风险，尤其是在技术资源分布不均的背景下，发展中国家可能因缺乏监管能力而面临跨国公司对神经数据的掠夺性使用，加剧隐私剥削和社会不平等⁶⁸。心理性不仅强化了神经数据的敏感性，还为其保护引入了人格尊严和个体自主性的法理维度，挑战现有法律框架的适用边界。

上述属性的递进揭示了神经数据法律定位的复杂性与动态性。生物性为其嵌入个人信息保护体系提供了基础，使其与健康数据等传统类别相容；敏感性要求法律提升保护标准，以应对其超越常规个人信息的高风险特性；心理性则推动法律向思想和人格领域延伸，突破既有框架的覆盖范围。这种复合特征引发了界定上的理论争议。一部分观点认为，神经数据的生物性和敏感性足以使其纳入现有框架，如巴西 LGPD 或科罗拉多州《隐私法》的敏感数据范畴，通过调整即可应对⁶⁹。另一部分观点则强调其心理性，认为现有法律无法充分覆盖“脑黑客”或非

⁶² Sorrell, E., Rule, M. E., & O'Leary, T. (2020). Brain-Machine Interfaces: Closed-Loop Control in an Adaptive System. 未正式出版，提交于 2020 年 8 月。

⁶³ Stark, C. (2023). Is a Global Governance Framework Necessary for Neurotechnology? *Drexel Law Review*, 15(4).

⁶⁴ Moral, N. P. B., & Vidotti, S. A. B. G. (2024). Neurodados: do necessário enquadramento como dados pessoais sensíveis. VII Workshop de Informação, Dados e Tecnologia.

⁶⁵ Pierré, A., et al. (2024). A Perspective on Neuroscience Data Standardization with Neurodata Without Borders. *Journal of Neuroscience*, 44(38).

⁶⁶ Pierré, A., et al. (2024). A Perspective on Neuroscience Data Standardization with Neurodata Without Borders. *Journal of Neuroscience*, 44(38).

⁶⁷ Stark, C. (2023). Is a Global Governance Framework Necessary for Neurotechnology? *Drexel Law Review*, 15(4).

⁶⁸ Li, Y. (2023). Digital Development Rights in Developing Countries: Where the Governance Rules for Cross-Border Data Flows. *Journal of Human Rights*, 22(5).

⁶⁹ Moral, N. P. B., & Vidotti, S. A. B. G. (2024). Neurodados: do necessário enquadramento como dados pessoais

自愿读心等新兴威胁，需引入新的权利形式⁷⁰。还有观点质疑心理性威胁的现实性，认为当前技术（如 fMRI）仅能在特定条件下推断有限状态，尚不足以改变其法律属性⁷¹。综合分析表明，神经数据的法律属性并非单一维度，而是由生物性奠基、敏感性提升、心理性突破的多层次结构，法律定位需以动态视角回应其复合性，既利用现有体系的稳定性，又为新风险预留扩展空间。

（二）神经数据保护的法理基础

基于神经数据的法律属性，其保护的法理基础从隐私权出发，逐步融入人格权和自主性保障，形成由浅入深、相互支撑的理论体系。这一体系不仅回应属性的多重性，还为构建规范提供逻辑依据。

1.从个人信息到思想隐私的递进

隐私权是神经数据保护的首要法理根基，其保护范围经历了从传统个人信息到思想隐私的递进过程。传统隐私权旨在保护个人信息免受非法采集和使用，例如姓名、地址等基本信息，而随着技术发展，这一权利已扩展至健康数据和生物数据的保护⁷²，神经数据的生物性和敏感性使其自然适用这一框架⁷³。在数据标准化和共享的过程中，需遵循严格的隐私规范，确保个体对其数据的控制权，如在跨实验室协作中避免未经同意的披露⁷⁴。在医疗场景中，医院采集患者的 EEG 数据用于诊断，必须事先征得明确同意并限制数据用途，这与现有个人信息保护的原则一致。然而，神经数据的心理性突破了传统隐私的边界，其可能揭示思想内容或情感状态的特性，使其威胁到所谓的“心理隐私”⁷⁵。为此，一些实践将“心理身份”纳入保护范围，明确禁止非自愿采集，通过立法将其视为不可侵犯的领域⁷⁶。另有观点建议，法律应明确神经数据的敏感地位，强化知情同意和数据最小化原则，以防止其被滥用于推断个体意图或偏好⁷⁷。

2.从生物载体到心理核心的深化

人格权作为保护个体尊严和完整性的法理基础，在神经数据语境下从生物载体的保护深化至心理核心的保障。神经数据的生物性使其最初被视为身体信息的

sensíveis. VII Workshop de Informação, Dados e Tecnologia.

⁷⁰ Stark, C. (2023). Is a Global Governance Framework Necessary for Neurotechnology? *Drexel Law Review*, 15(4).

⁷¹ Rainey, S., & Dalese, P. (2024). An alternative focus on data in the neurorights discussion – Lessons from Brazil. *Open Res*, 1:3.

⁷² Moral, N. P. B., & Vidotti, S. A. B. G. (2024). Neurodados: do necessário enquadramento como dados pessoais sensíveis. VII Workshop de Informação, Dados e Tecnologia.

⁷³ Unger W. Stay Out of My Head: Neurodata, Privacy, and the First Amendment[J]. *Wash. & Lee L. Rev.*, 2023, 80: 1439.

⁷⁴ Reimer, M. L., et al. (2021). Core principles for the implementation of the neurodata without borders data standard. *Journal of Neuroscience Methods*, 348:108972.

⁷⁵ Stark, C. (2023). Is a Global Governance Framework Necessary for Neurotechnology? *Drexel Law Review*, 15(4).

⁷⁶ Strickland, E., & Gallucci, M. (2022). First Win for the Neurorights Campaign: Chile plans to regulate all neurotech and ban the sale of brain data. *IEEE Spectrum*, 59(1), 26-58.

⁷⁷ Moral, N. P. B., & Vidotti, S. A. B. G. (2024). Neurodados: do necessário enquadramento como dados pessoais sensíveis. VII Workshop de Informação, Dados e Tecnologia.

延伸，与基因数据或器官数据类似，属于人格权的物质基础。使用侵入式脑机接解码神经信号并通过电刺激调控脑活动用于治疗抑郁症或恢复触觉，这一过程直接涉及个体的生理完整性⁷⁸。然而，其心理性使其保护需求进一步深化，因为神经数据的“读写”能力可能影响人格特质甚至心理状态，威胁到个体的内在尊严⁷⁹。若某技术未经同意操控用户的情感反应，可能改变其自我认知或行为模式，这种干预超越了传统隐私侵害，直接触及人格的核心。有学者提出，“认知自由”应作为人格权的延伸，确保个体免受非自愿脑操控的影响，这一权利旨在维护心理领域的不可侵犯性⁸⁰。另有观点建议通过程序性规范，如强制性伦理审查或技术评估，来保障人格完整性，尤其是在涉及脑干预的场景中⁸¹。

法理基础从生物性深化至心理性要求法律赋予神经数据更高的保护等级。这种深化不仅回应了神经数据的复合属性，还为法律体系注入了伦理维度，确保保护不仅停留于形式，而是深入个体的人性本质。

3.从知情同意到技术责任的整合

自主性是神经数据保护的伦理与法理支柱，强调个体对其数据的控制权，并通过技术责任实现全面保障。这一基础从程序正义的知情同意扩展至实质正义的技术设计，形成一个整合性框架。神经数据的采集和应用高度依赖个体的参与和神经可塑性，例如闭环脑机接口的实时控制需用户自愿配合，其效果依赖于大脑的适应能力⁸²。国际伦理原则明确要求，神经数据的处理必须基于自由和知情同意，并符合人权标准，确保个体在数据采集前充分了解用途和风险⁸³。在数据标准化实践中，这一原则同样适用，如通过技术规范确保个体知晓数据的共享范围和潜在影响⁸⁴。然而，自主性的保障不仅限于程序性同意，因为技术复杂性和权力不对等可能削弱个体的实际控制力。如在发展中国家，由于技术鸿沟和监管不足，个体可能无法有效抵御跨国公司对其神经数据的掠夺性采集，这种情境下单纯的同意机制难以保障自主性⁸⁵。自主性保障整合了知情同意的程序正义与技术责任的实质正义，从被动的权利主张转向主动的风险预防。这种整合不仅回应了神经数据的敏感性和心理性，还为法律体系提供了从个体赋权到技术约束的双重

⁷⁸ Zhao, Z.-P., et al. (2023). Modulating Brain Activity with Invasive Brain-Computer Interface: A Narrative Review. *Brain Sciences*, 13:134.

⁷⁹ Stark, C. (2023). Is a Global Governance Framework Necessary for Neurotechnology? *Drexel Law Review*, 15(4).

⁸⁰ Stark, C. (2023). Is a Global Governance Framework Necessary for Neurotechnology? *Drexel Law Review*, 15(4).

⁸¹ Moral, N. P. B., & Vidotti, S. A. B. G. (2024). Neurodados: do necessário enquadramento como dados pessoais sensíveis. VII Workshop de Informação, Dados e Tecnologia.

⁸² Sorrell, E., Rule, M. E., & O'Leary, T. (2020). Brain-Machine Interfaces: Closed-Loop Control in an Adaptive System. 未正式出版，提交于 2020 年 8 月。

⁸³ UNESCO. (2024). 《神经技术伦理建议书（初稿）》。
https://unesdoc.unesco.org/ark:/48223/pf0000391444_chi.

⁸⁴ Pierré, A., et al. (2024). A Perspective on Neuroscience Data Standardization with Neurodata Without Borders. *Journal of Neuroscience*, 44(38).

⁸⁵ Li, Y. (2023). Digital Development Rights in Developing Countries: Where the Governance Rules for Cross-Border Data Flows. *Journal of Human Rights*, 22(5).

保障路径，确保保护措施既具可操作性，又能适应技术社会的复杂需求。

五、神经数据保护规范构造的展开

在上述背景下，神经数据保护的规范构造必须直面核心问题：究竟应另立规范、在现有规范框架的基础上改进，还是结合二者形成一个全新的保护路径？本文主张选择结合另立规范与现有法律框架改造相结合的混合路径，并认为其在当前技术与伦理情境下最具合理性与可行性，因此，本节从规范目标的确立、混合路径的选择依据、具体规范构造及实施挑战四个方面，系统阐述混合路径的设计逻辑与实践框架，旨在为神经数据保护提供一个全面而可操作的框架。

（一）规范目标的确立：平衡现实风险与未来威胁

神经数据保护规范的核心目标是在当前技术限制和未来潜力之间找到平衡，既要应对已知风险，也要为潜在威胁留出保护空间。由于对神经数据威胁程度和技术能力的评估存在分歧，规范需整合不同观点，构建一个立足现实、面向未来的保护体系。

规范的起点是识别和应对现实风险。尽管当前技术如 BCI 尚未实现精确的“强读心”，但能在特定条件下推断部分心理状态⁸⁶。这些技术在医疗和商业领域已有应用，就存在一定的数据滥用和隐私侵害的风险。因此，规范应首先确保个体隐私和自主性在现有技术水平下得到有效保护。

规范的长期导向是预防和预备未来威胁。尽管当前技术能力有限，但未来神经技术与 AI 的融合潜力巨大，其能够实时解码个体的无意识大脑信息，从而挑战着传统信息隐私的自主性。⁸⁷规范需具备前瞻性，为未来技术突破预留保护空间，避免风险显现时措手不及。

规范的目标需平衡现实风险与未来威胁。混合路径以现实风险为底线，确保规范的可操作性；以未来威胁为导向，赋予规范适应性和预防性。这种双重目标旨在构建一个既务实又具有弹性的保护框架，为后续规范构造奠定坚实基础。

（二）混合路径的选择依据：理论分歧的调和与实践需求

混合路径的选择基于理论争议的调和与现实实践的需要。怀疑派、保守派与自由派在神经数据性质、技术能力及法律策略上的分歧，揭示了单一路径的局限性。混合路径整合各方优势，既避免过度立法，又填补现有框架的空白，满足技术社会的实际需求。

怀疑派的审慎启示为混合路径提供了底线约束。怀疑派认为，神经技术未对神经数据与心理隐私构成独特威胁，现有法律已足够应对。其核心论点是，现有技术的实际能力限制了对主观心理状态的访问，应避免因技术炒作导致不必要的

⁸⁶ 周彦每.脑机接口技术嵌入的风险表达及法律规制逻辑[J].江汉论坛,2024,(08):125-129.

⁸⁷ 陈鲁夏.精神隐私可以成为新兴权利吗?[J].环球法律评论,2024,46(05):89-104.

权利膨胀或资源浪费。混合路径确保新规范以现实证据为依据，避免脱离技术现状的空想立法。

保守派的实用贡献为混合路径提供了实施基础。保守派主张通过调整现有法律框架即可应对神经数据的潜在风险。混合路径利用既有框架的稳定性，降低另立规范的制度成本，同时通过微调增强其针对性，避免保守派可能忽视的未来风险。

自由派的前瞻洞见为混合路径补充了未来视野。自由派突出神经数据的独特性，认为现有法律无法应对新兴威胁。他们主张引入神经权利，以智利和巴西的立法实践为例，呼吁全球性的新框架。混合路径承认技术进步可能突破现有法律的保护范围，通过另立规范填补空白，确保规范能够应对未来技术场景下的伦理与法律挑战。

实践需求的驱动将理论转化为行动指南。技术社会的现实需求进一步支持混合路径。神经数据应用已渗透多个领域，涉及多方利益主体。单一路径难以满足这些场景的多样化需求：另立规范可能因定义模糊而难以实施，单纯调整现有框架则可能滞后于技术进步。混合路径通过结合二者，既利用现有体系的执行力，又通过新规范适应技术社会的复杂性，确保保护措施的全面性与可行性。

（三）混合路径的具体构造：规范框架的设计

神经数据的独特性在于其与心理状态的直接关联，既涉及当前的现实风险，又潜伏着未来的技术威胁。单一的法律或伦理手段难以应对这种复杂性，因此混合路径通过整合现有法律的稳定性、另立规范的针对性以及伦理与社会机制的灵活性，构建一个多维度的保护框架。这一框架不仅回应技术现状的迫切需求，还为未来的不确定性预留调整空间，从而实现保护与创新的动态平衡。

1. 现有法律框架改造：以欧盟《通用数据保护条例》与《欧洲人权公约》为例

在现有法律框架改造方面，一个切实可行的起点是将神经数据明确纳入欧盟《通用数据保护条例》（GDPR）的“特殊类别数据”范畴，从而为其提供更高的保护标准，并通过强化现有机制快速应对当前的隐私侵害风险。GDPR 作为欧盟数据保护的核心法律，已将健康数据、生物识别数据等列为“特殊类别数据”，要求处理此类数据时必须满足更严格的条件⁸⁸，例如获得数据主体的明确同意、遵循数据最小化原则，并确保处理的合法性与透明性。若将神经数据纳入这一范畴，意味着其处理将受到类似的高标准约束。在医疗场景中，医院使用脑电图（EEG）采集患者神经数据以诊断癫痫，需事先征得患者明确同意，并确保数据仅用于诊断而非其他目的；在商业场景中，企业通过消费级脑机接口（BCI）设备（如智能头显）采集用户神经数据用于行为分析，则需向用户清晰披露数据用途、存储

⁸⁸ 欧洲议会. 欧盟一般数据保护条例 [M]. 任虎, 译. 上海: 华东理工大学出版社, 2018: 22-24.

期限及潜在风险，并在获得自愿授权后方可实施，否则将面临高额罚款。这种改造充分利用了 GDPR 的成熟性，通过明确神经数据的特殊地位，增强了个体对其数据处理的控制权。为进一步应对数据再利用和再识别风险，若企业将医疗数据转用于商业预测或通过匿名数据追溯个体身份，可强化现有机制来要求任何二次使用必须经过国家数据保护机构（如欧盟各国的 DPA）的伦理与法律评估，并强制实施高级匿名化标准。具体而言，可采用差分隐私技术等措施实施匿名化，虽然数据价值因匿名化有所减弱，但可最大程度防止数据被再次识别⁸⁹，这种技术性保障措施显著提升了数据处理的透明性与可控性。

此外，2024 年，Clearview AI 在未获得法律依据的情况下，处理了荷兰境内数据主体的个人数据，包括生物数据，违反了《通用数据保护条例》（GDPR）的多项条款，荷兰监管机构对 Clearview AI 处以 3050 万欧元的罚款⁹⁰。该案例显示，若将神经数据纳入这一框架，GDPR 在类似场景中的有适用的潜力，可沿用既有判例，进一步强化其保护效力，确保数据处理全程符合隐私保护要求。

此外，《欧洲人权公约》（ECHR）提供了一个更广泛的法律基础，《ECHR》第 8 条承认数据隐私。但数据隐私是否意味着并包括精神数据隐私的问题依然存在⁹¹，可通过演绎其条款为神经数据保护注入人权维度，进一步完善现有框架的覆盖范围。第 8 条关于“私人生活权”的条款作为补充，能够保护由神经数据衍生的隐私信息。通过脑活动数据推断出的情感状态、行为倾向或心理特征，属于个体私人生活的重要组成部分，若被政府或企业未经授权使用（如在招聘中进行神经筛选），则可依据第 8 条主张权利保护。与此同时，欧洲人权法院可以通过已有的判例使《欧洲人权公约》适应新的社会环境⁹²。根据欧洲人权法院，DNA

和指纹包含与已识别或可识别个人有关的独特信息，因此包含受保护的个人信息⁹³，神经数据同样具备生物性与个体独特性，且涉及心理状态的高度敏感性，因此可沿用此判例获得类似保障。这种双重保障机制充分利用了 ECHR 的成熟性与灵活性，能够有效应对当前的隐私侵害风险。

在具体实施中，这一改造路径还可通过细化 GDPR 和 ECHR 的现有程序增强其针对性。在 GDPR 框架下，可制定专门的“神经数据处理指南”，由欧洲数据保护委员会（EDPB）发布，明确神经数据采集的合法条件、使用限制及匿名化要求，确保各缔约国在执行时的标准一致。同时，在 ECHR 框架下，可通过案例

⁸⁹ 陈兵,周裕喜.数字经济下企业破产个人数据处置路径再探[J/OL].兰州学刊.

⁹⁰Dutch Supervisory Authority imposes a fine on Clearview because of illegal data collection for facial recognition, https://www.edpb.europa.eu/news/national-news/2024/dutch-supervisory-authority-imposes-fine-clearview-because-illegal-data_en, 最后访问日期: 2025 年 4 月 2 日。

⁹¹Silvia Inglese & Andrea Lavazza, What Should We Do With People Who Cannot or Do Not Want to Be Protected From Neurotechnological Threats? 15 FRONT HUM NEUROSCI. (2021).

⁹² Hallinan D, Akintoye S, Stahl B C, et al. Neuroexceptionalism: Framing an emergent debate[J].2021.

⁹³ ECtHR (GC) S. & Marper v. UK, appl.nos. 30,562/04, 30,566/04, § 72, 75, 84 (4 December 2008); ECtHR Aycaguer v. France, appl.no. 8806/12, §33 (22 June 2017).

法的逐步积累，明确神经数据相关的侵权判定标准。这种细化措施不仅增强了法律的适用性，还通过既有机构的协调作用降低了改造成本。EDPB 和欧洲人权法院的现有基础设施可直接用于神经数据保护，无需新建专门机构，从而提高了实施效率。

2.另立规范：缔结《国际神经数据保护与隐私公约》

在另立规范方面，可在联合国框架下缔结《国际神经数据保护与隐私公约》（以下简称《公约》）来填补现有规范的空白，为神经数据的保护提供普适性与前瞻性兼备的全球规范基础。该公约旨在为神经数据的采集、使用和跨境流动建立清晰的法律边界，涵盖定义与范围、权利与义务、数据规范、技术干预及监督执行五个主要方面。在定义与范围上，《公约》应明确“神经数据”为通过神经技术（如 BCI、EEG）采集的反映脑活动的信息，包括原始数据和衍生心理信息，适用范围涵盖缔约国境内及跨境流动的神经数据，涉及个人、政府和企业行为。在权利与义务上，公约规定个体对神经数据的隐私权和自主权，以及缔约国保护这些权利的义务，同时明确企业与科研机构在数据处理中的法律责任。在数据采集与使用规范上，公约要求任何神经数据活动须经个体明确同意，禁止非自愿读心或脑操控，并设定跨境数据流动的统一隐私标准。在技术干预与伦理约束上，公约针对脑干预设界限，区分治疗性和非治疗性应用，仅允许前者在严格监管下实施。在监督与执行上，公约设立国际监督机构，监测执行情况并协调争端解决。《公约》的基本条款包括目的与原则（如保护隐私与促进技术公平发展）、定义（如“神经数据”“心理完整性”）、适用范围（如缔约国及跨境行为）、缔约国义务（如制定国内法与国际合作）及个体权利（如控制数据与免受干预）。核心条款则聚焦知情同意要求、禁止非自愿读心与脑操控、数据不可交易性、心理完整性保护及跨境数据流动标准等。公约可借鉴智利神经保护立法，规定神经数据的“不可交易性”，禁止除科研外的商业买卖，防止技术发达地区通过数据交易剥削个体隐私，同时保护发展中国家免受跨国公司数据掠夺。

冲突解决机制是《公约》执行的关键，处理缔约国间或缔约国与个体/企业间的争端。公约应规定，首先通过缔约国间直接协商解决争端，协商不成可提交 NDPC 调解，由中立调解员在 6 个月内提出非约束性方案。若调解失败，争端可提交常设仲裁庭（PCA），由三名仲裁员（各方各选一名，第三名由 NDPC 指定）在 12 个月内作出约束性裁决。对涉及重大原则性问题（如公约解释）的争端，可诉至国际法院（ICJ）。为保障执行，NDPC 监督裁决落实，未履行的缔约国可被提交联合国安理会采取进一步措施。同时，建立“神经数据保护基金”，由缔约国捐款，用于支持争端解决及受害者赔偿，确保弱势群体的救济途径。

伦理与社会机制的辅助是法律框架的必要补充，其作用在于通过软性约束增强规范的灵活性与社会认同，确保保护路径不仅具有法律效力，还能契合技术社

会的多元价值需求。在伦理层面，一个核心措施是强化知情同意原则，使其成为神经数据保护的基石。在医疗和研究场景中，个体应被详细告知神经数据采集的目的（如疾病诊断）、范围（如采集哪些脑区数据）及潜在风险（如数据泄露导致的隐私侵害），确保同意的真实性和自愿性。在脑机接口临床试验中，研究者需向患者提供清晰的书面说明，并留出充分的决策时间，避免因信息不对称导致的被动同意。同时，可以确立技术责任原则，要求企业和开发者在设计神经技术时遵循“最小侵害”理念，如默认启用数据加密、限制数据采集范围至必要最小值，并通过技术手段（如区块链）确保数据的可追溯性与安全性。为进一步落实这一原则，可设立独立的伦理审查委员会，由跨学科专家组成，负责评估技术应用的道德影响。若某公司开发的神经营销设备可能通过潜意识刺激影响消费者决策，委员会可要求其修改设计或暂停部署，以保护个体自主性。在社会层面，规范的实施离不开公众的支持与参与。通过公开讨论会或公民论坛，可以吸纳公众意见，如讨论神经数据在教育（如脑增强学习）或就业（如神经筛选）中的合理使用边界。这种参与机制不仅提升规范的合法性与接受度，还能避免技术与伦理脱节。这种伦理与社会机制的辅助弥补了硬性法律规范在灵活性上的不足，确保保护路径在技术理性之外，充满人文关怀。通过与法律框架的协同作用，这一机制能够在技术社会的快速变化中，为神经数据保护提供持续的社会支持与伦理约束。

（四）混合路径的实施挑战与应对策略

混合路径通过整合优化法律框架并辅以伦理与社会机制，为神经数据保护提供了一个兼顾现实与未来的规范体系。然而，其实际实施面临多重挑战，包括技术不确定性、法律协调难度、合规成本以及社会接受度的考验。这些挑战源于神经技术的快速发展、法律体系的多样性及社会对新兴技术的复杂态度。为确保混合路径的有效落地，必须设计针对性的应对策略，通过动态调整、国际协作和公众参与，将理论设计转化为实践成果。

技术不确定性是混合路径实施的首要障碍。神经技术如脑机接口（BCI）与人工智能（AI）的融合发展迅速，从当前有限的心理状态推断到未来可能的潜意识解码，其技术边界难以预测。技术的发展导致神经数据的解码能力可能超越 GDPR 当前设计的预期⁹⁴，将神经数据纳入 GDPR 的“特殊类别数据”或 ECHR 的保护范围，可能因技术迭代而滞后。《国际神经数据保护与隐私公约》（以下简称《公约》）的干预界定也可能因技术进步而滞后。为应对这一挑战，可建立技术监测与动态调整机制，依托欧洲数据保护委员会（EDPB）或联合国教科文组织（UNESCO）的伦理框架，定期评估技术对隐私的影响。每两年发布技术风险报告，更新相关神经技术对神经数据产生风险的相关数据，并更新 GDPR 匿名化标

⁹⁴ Ienca M, Malgieri G. Mental data protection and the GDPR[J]. Journal of Law and the Biosciences, 2022, 9(1): lsac006.

准或《公约》条款。

法律协调难度是另一关键挑战。将神经数据纳入 GDPR 并扩展 ECHR 适用范围,需欧盟成员国统一执行标准,但各国数据保护实践的差异可能导致跨国企业面临多重要求。《公约》的全球执行则因各国法律传统与技术水平差异而复杂化,发达国家可能倾向宽松条款以促创新,发展中国家则强调隐私防护。为解决这一问题,可设立跨部门协调机制,在欧盟由 EDPB 牵头,制定“神经数据处理指南”,统一采集、匿名化和跨境传输标准。在国际层面,《公约》可依托联合国框架下的“神经数据保护委员会”(NDPC),通过多边协商协调立场,提供技术援助支持发展中国家合规能力,要求发达国家承担更多责任。这种机制从区域扩展至全球,降低冲突风险,确保实施的连贯性。

合规成本构成进一步挑战。将神经数据纳入 GDPR 并强化审查,要求企业作为生产要素的神经数据满足知情同意、加密及匿名化标准。中小型企业开发 BCI 设备时,需额外进行伦理评估与技术改造,可能削弱其竞争力,导致行业集中。GDPR 中数据最小化原则的表述较为笼统和模糊,并未在技术层面做出具体指导,⁹⁵再加之神经数据的多维度特性使得在收集数据时无法明确是否满足最小化原则的相关性、充分性与必要性要求⁹⁶。为应对这一挑战,可通过欧盟创新基金补贴企业合规成本,特别是中小型企业,确保市场公平。

社会接受度的考验是实施的关键外部障碍。在国外的一项调查中,参与的受访者对 BCI 技术有适度的道德接受度和使用意愿,但更多受访者明确反对使用,而非明确表示会使用⁹⁷。这表明公众对神经技术的期待与担忧并存,可能会误解 BCI 技术读取与分析神经数据的能力。企业则可能因商业利益反对高合规要求。这种接受度危机可能削弱规范执行力。为应对这一挑战,可通过科普教育澄清技术现状与风险。同时,引入利益相关方协商机制,在《公约》起草或 GDPR 修订时举行公民论坛,讨论神经数据使用边界,加入公平性保障以回应社会不平等担忧。此外,可设立申诉渠道,由 EDPB 或 NDPC 处理违规报告,增强公众信心。

六、结语

神经数据保护是技术时代无法回避的课题,其复杂性与紧迫性在国际社会的探索与争鸣中日益显现。从智利以另立“神经权利”规范捍卫思想隐私的坚定尝试,到巴西和科罗拉多州依托现有框架灵活调整的务实路径,再到联合国教科文组织等国际组织提出的伦理倡议,这些努力共同揭示了一个核心难题:神经数据的独特属性——既承载生物信息,又直指心理状态与思想疆界——使其兼具科学价值与社会风险,单一路径难以在隐私保障与技术创新之间找到平衡点。各国实践的

⁹⁵梅傲,苏建维.欧盟数据最小化原则研究[J].国别和区域研究,2024,(01):64-83+310.

⁹⁶戴龙.论数字贸易背景下的个人隐私权保护[J].当代法学,2020,34(01):148-160.

⁹⁷ Sattler S, Pietralla D. Public attitudes towards neurotechnology: Findings from two experiments concerning brain stimulation devices (BSDs) and brain-computer interfaces (BCIs)[J]. PLoS one, 2022, 17(11): e0275454.

局限性，叠加学术界在神经数据法律属性与保护法理上的激烈争论，以及技术迭代、跨国性与伦理风险的交织，凸显了这一议题的多维挑战，呼唤更具包容性与前瞻性的解决方案。

本文提出混合路径，将现有框架改造与另立规范相结合。这一路径源于对神经数据法律属性的系统剖析——从生物性、到敏感性再到心理性，确立了其既可嵌入个人信息体系又需扩展至思想领域的定位；同时依托隐私权、人格权和人的自主性保障的法理基础，构建多层次理论支撑。在实践层面，混合路径整合 GDPR 与 ECHR 的成熟机制，通过将神经数据纳入“特殊类别数据”并扩展“私人生活”的边界，快速应对现实中的隐私与数据滥用风险；缔结《国际神经数据保护与隐私公约》则填补现有框架对未来威胁的治理空白，确立数据不可交易性与心理完整性保护，确保全球普适性。这一方案不仅调和了怀疑派对现状的坚持、保守派的渐进调整与自由派的前瞻洞见，还回应了智利、巴西等实践中的现实矛盾，并为国际倡议缺乏执行力的短板提供了制度补充。通过法律监管、技术支持与国际协作，混合路径从理念走向现实，为神经数据保护提供了兼具深度与弹性的框架。

然而，这一路径只是探索的起点，而非终点。随着脑机接口与人工智能融合的加速突破，神经数据的应用边界可能进一步扩展，例如读取记忆、调控情绪甚至预测行为倾向，这些潜在场景将持续考验治理体系的适应性与前瞻性。此外，全球范围内的法律传统差异、技术水平落差以及跨国企业在数据处理中的合规策略博弈，要求更精细的协调机制以确保规范的统一性与执行力。学术界对神经数据法律地位的讨论——究竟是个人信息范畴的延伸抑或应确立为独立权利——也未达成共识，这一争论可能随着技术演进与社会实践的深化而获得新解。未来研究需在此基础上进一步深化，针对不同场景设计具体规则，优化技术标准以匹配神经数据的采集与解码能力，或探索更具操作性的国际协作模式。只有通过持续的动态调整，才能确保保护措施既不过度限制创新，又不因滞后而失守隐私底线。

神经数据保护不仅是法律与技术的交融，更是伦理与社会的深刻对话。混合路径通过多层次设计，为这一对话提供了一个兼具理论深度与实践弹性的框架。然而，面对技术的未知变量与全球化的复杂挑战，治理之路依然漫长而艰辛。在这个时代，守护思想隐私、释放技术潜力，既是法律的使命，也是人类的共同课题。唯有以开放的心态、协作的精神和不懈的努力，才能为神经数据的未来找到既安全又充满可能的答案，真正实现个体权利与社会进步的和谐共存。

总体国家安全观视角下我国人工智能治理的困境与对策

胡琳英¹

摘要：在总体国家安全观视域下，人工智能技术的爆发式发展既带来生产力革新，也对国家安全构成系统性风险。本文通过分析人工智能技术的三层风险，揭示数据投毒、算法黑箱、深度伪造等技术风险、在场景中的传导性及对意识形态的影响如何引发传统与非传统风险交织的复合风险。目前，我国人工智能法治面临三重困境，数据、算法规制不足、监管体系缺陷、国际规则制定话语权缺失。对此，提出坚持科技向善的治理宗旨、构建数据要素产权分置制度与算法可解释性框架、运用场景化治理与监管沙盒机制并通过技术标准输出与发展中国家联盟破解规则霸权。须以总体国家安全观为统领，在保障技术自主创新与维护国家安全之间形成动态平衡，为构建人工智能治理中国方案提供理论支撑。

关键词：总体国家安全观；人工智能；数据治理；算法治理；意识形态安全

引言

2025 年开年，由杭州深度求索基础技术有限公司自主研发的生成式人工智能大模型 deepseek，以媲美美国 chatgpt o1 模型的性能而远低于 o1 模型的训练成本以及完全开源性迅速在海内外引发轰动。deepseek 仅在一周时间内，用户增长近一亿。并且由于其高性价比、国产模型、开源的特性，各大企业、政府机关、各行业机构、甚至个人使用自身服务器进行本地化部署。deepseek 的出现极度降低了民众对于人工智能应用使用的门槛，使人工智能应用走向千家万户及各行各业。

在人工智能技术重塑社会结构、经济的同时，人工智能对于国家安全的影响同样不可小觑。人工智能现在仍存在数据垄断、算法歧视及幻觉等风险，尤其是目前人工智能在社会层面的推广，风险更处于叠加状态，监管迫在眉睫。然而，人工智能技术在各国之间进行竞争式发展，学者沈伟提出“数字博弈”与“数字霸权”的概念，指出人工智能影响各国军事、经济、政治等各方面，具备国家安全属性，成为国际政治权力的重要内容，又将引发规则制定话语权之争。基于此，本文将在总体国家安全观的视角下，分析人工智能特性及其可能导致的国家安全风险，梳理我国人工智能法治存在的问题及对策。

¹ 西南政法大学学生，19967200337

一、总体国家安全观与人工智能的内在逻辑

（一）总体国家安全观为人工智能法治提供理论指导

2014年4月15日，习近平总书记在中央国家安全委员会第一次会议上指出当前我国国家安全内涵、时空领域、内外因素比历史上任何时候都丰富、复杂、宽广，并首次要求“贯彻落实总体国家安全观”，中国共产党对国家安全治理的理论创新进入新阶段。²2015年7月1日，《国家安全法》正式通过，并于第二条中明确将国家安全定义为“国家政权、主权、统一和领土完整、人民福祉、经济社会可持续发展和国家其他重大利益相对处于没有危险和不受内外威胁的状态，以及保障持续安全状态的能力。”³

总体国家安全观核心理论由“一个总体”、“五对关系”“十个坚持”构成。坚持总体国家安全观要坚持“一个总体”，即大安全理念。这强调国家安全是一个相互关联的有机整体，涵盖政治、经济、军事、科技、网络、生态等20余个领域其核心要义在于打破传统安全与非传统安全的界限，强调安全问题的联动性与整体性。这种“大安全”理念要求将国家安全视为动态平衡的过程，既维护静态安全状态，更注重持续安全能力的建设。坚持总体安全观要坚持“五对关系”的辩证统一。统一外部安全与内部安全，既要抵御外部威胁如西方意识形态渗透，又要化解国内社会矛盾与风险叠加；统一国土安全与国民安全，在捍卫领土主权的同时，更强调“以人民安全为宗旨”，将科技为民纳入安全范畴；统一传统安全与非传统安全，在确保军事、政治等传统安全基础上，重点应对网络安全、人工智能风险等新型挑战；统一发展问题与安全问题，发展是安全的基础，安全是发展的条件，要求通过高质量发展提升抗风险能力；统一自身安全与共同安全，倡导“共同、综合、合作、可持续”的全球安全观，推动构建人类命运共同体。

人工智能风险具有复合性、传导性、扩散性，具有“牵一发而动全身”的影响，对国家安全具有强威胁，已无法在传统安全观下予以解决。譬如，数据安全已从单纯的技术可靠性问题演变为涉及数据主权、意识形态安全和国际竞争力的复合型安全议题；生成式人工智能引发的深度伪造、认知操控和军事智能化风险，形成了跨物理领域、信息领域和认知领域的新型安全威胁，这印证了传统安全治理范式的局限性。人工智能风险已涉及国家安全的多个领域，因此，从总体国家安全观的视角审视人工智能风险，不仅是应对技术变革的必然选择，更是维护国家主权、安全和发展利益的战略需要。这一理论体系突破传统安全与非传统安全的机械二分法，以系统思维把握技术风险传导规律，以动态平衡协调创新发展与安全保障，以内外统筹推进国内治理与国际规则协同，构建起具有中国特色的人工智能风险法治模式。

² 习近平：《坚持总体国家安全观 走中国特色国家安全道路》，载《人民日报》2014年4月16日，第1版。

³ 《中华人民共和国国家安全法》，中国法制出版社2015年版，第3页。

（二）人工智能技术特征对国家安全的深刻影响

人工智能是指以数据与算法为基础,模拟人脑思维,对数据进行分析,理解、推理、深入学习特定内容的类人智能,拥有一定的自主性。⁴人工智能依据海量的数据与算法深入理解人类自然语言,并基于海量网络信息及自主思考生成输出内容,其互动性、便利性使其得以与人类社会迅速深刻连结。也正因如此,其技术特性与国家安全体系形成了深深连结的映射关系。这种映射关系既非简单的线性对应,也非静态的机械关联,而是呈现出动态交织、层级传导和系统演化的特征。从数据要素到智能武器,从算法权力到基础设施,人工智能的每个技术要素都如同精密齿轮般嵌入国家安全机器,可能引发针对国家安全的传导效应。

数据是人工智能运行的基础,可能引发科技安全与经济安全风险。人工智能对于数据的严重依赖也导致了数据质量对输出结果的决定性影响,训练数据在合法性、安全性上的风险将直接传导至输出端。数据投毒是一种常见的数据威胁,在训练阶段,有毒数据样本被投入沦陷数据中,阻碍训练收敛、破坏模型泛化能力、消解模型可用性。⁵除此之外,影响国家安全的数据风险源自于其强大的数据抓取能力,它对隐私数据、涉密数据、政府内部数据的无意抓取以及对大量数据的分析,可能导致关键信息或信号泄露从而危及国家安全。当海量数据通过智能传感器、社交平台 and 物联网设备持续涌入云端,数据主权便超越威斯特伐利亚体系的传统主权概念,模糊了传统的领土界限。但数据的跨境流动如同双刃剑,既可能催生科技创新,也可能导致关键产业数据泄露。美国科技巨头凭借全球数据采集网络形成数据优势,使众多发展中国家面临经济命脉被算法模型解构的风险。数据要素分布的不对称,不仅威胁着国家的科技安全根基,更可能通过产业链数据垄断动摇经济主权的独立性。

算法是根据其内在的数理逻辑进行信息筛选与决策生成的机制⁶,模型的不可解释性、偏见与歧视问题构成核心风险,可能引发政治安全与社会安全风险。自主决策的算法存在“算法黑箱”问题,仅凭输入内容与输出结构难以解释其决策逻辑,缺乏透明度,难以追溯风险成因,还存在“幻觉”问题,导致技术可控性、可信度存疑。这可能使算法成为影响国家安全的“武器”。例如,美国认为 tiktok 算法有引导美国民众改变选举选择的问题。而算法偏见问题更具渗透性,它由训练数据不完全开始又以算法歧视结束,原因是某些弱势群众的数据被边缘化,继而被算法忽视。在谷歌图像识别训练数据库中,仅有百分之三的数据来自占世界人口多数的中国与印度,其他数据基本来源于欧美国家,而这些被遗忘的主体在权利保护上在无形中受到歧视,⁷从而产生一系列社会风险。

⁴ 约翰·马尔科夫. 人工智能简史 (M). 郭雪, 译. 杭州: 浙江人民出版社, 2017.

⁵ 姜毅, 杨勇, 印佳丽, 等: 《大语言模型安全与隐私风险综述》, 载《计算机研究与发展》1-41.

⁶ 佩德罗·多明戈斯. 终极算法: 机器学习和人工智能如何重塑世界 [M]. 黄芳萍, 译. 北京: 中信出版集团, 2022: 3.

⁷ 郭小平, 秦艺轩: 《解构智能传播的数据神话: 算法偏见的成因与风险治理路径》, 载《现代传播(中国传媒

深度伪造实际上是算法问题的进一步延伸，但由于其普遍性与问题的严重性，需加以密切的关注，其可能引发文化安全与意识形态安全风险。人工智能正潜移默化地渗透民众认知，演变为新型信息战载体，使之风险蔓延至认知领域。人工智能输出内容对人有引导作用，在全球第一例人工智能致死案中，14岁少年的母亲指控AI公司 character.ai 存在管理疏忽，导致青少年在其管理的聊天机器人的诱导下开枪自杀，延伸至国家安全领域，人工智能可能通过制作虚假文本、音频等深度伪造工具，或是在被有关国家故意为人工智能输入特定意识形态价值观，动摇我国主流价值观。⁸西方技术霸权国家通过影视算法推荐、社交机器人评论形成的文化渗透，对其他国家国家进行文化价值观的传导，这种隐性的意识形态植入比传统宣传更具破坏性。

人工智能技术在智能武器上的进化使其逐渐进入伦理与法理的无人区，引发军事安全与国防安全风险。由于致命性自主武器系统的实战部署，算法决定生死的战争伦理危机迫在眉睫。所谓“致命性自主武器系统”指的是一种基于AI技术且具有自主决策性能的“杀手机器人”武器攻击平台，它可以在不经过人类操作员干预的情况下，独立搜索、识别目标并自主性地做出攻击决策，其设计初衷是寄望于以智能机器人取代人类战士从事战场侦察、日常巡逻和执行危险性战斗任务。伴随着各国争相研发、部署和使用此类武器系统，未来战争的“去人类中心化”和“去人类控制化”趋势已然显现。在人工智能技术的驱动下，人类自身的生存从未像今天这样面临自我毁灭之风险，人类的创造物-自主武器系统-正在取代人本身而成为生杀予夺的决策中心，基于机器算法逻辑的自动搜索与自主攻击正在颠覆传统战争理念，未来机器人战争将会拯救更多的士兵生命还是会波及伤害更多的无辜平民，目前已成新军事革命和国际人道主义规范重塑的前瞻性话题。

二、我国人工智能法治发展的困境

2017年，我国国务院发布《新一代人工智能发展规划》，规定2020年初步建立部分领域的人工智能伦理规范与政策，2025年初步建立人工智能法律法规，2030年建成更完善的人工智能法律法规与伦理规范。截至目前，除了《发展规划》《治理原则》《伦理规范》《治理原则》等顶层设计外，已有《网络安全法》《个人信息保护法》《数据安全法》《科学技术进步法》从基础设施、数据要素等方面作为人工智能基础治理法。部门规章方面，2021年我国制定《算法推荐管理规定》，强化平台责任；2022年，我国针对深度合成问题发布《深度合成管理规定》；影响较大的主要是2023年发布的《生成式人工智能服务管理暂行办法》，其规定了数据来源合法性、内容标识等内容。

（一）作为基础的数据与算法规制不足

大学学报)》2019年41卷09期19-24页。

⁸ 中国信通院：《人工智能治理蓝皮书（2024年）》，2024年12月。

数据方面，主要体现在如下几点：其一，数据权益的性质目前尚无定论。虽然《民法典》127条规定：“法律对数据、网络虚拟财产保护有规定的，依照其规定。”但这并非将数据性质认定为财产权，数据基于难以物权公示及流转利用的社会效益等原因无法被划分为财产权。并且，数据性质相当复杂，有学者认为，比起单一的一类权利，更像是多项权利集合的“权利束”，无法通过传统的物权保护方式使之有序流通、开发、利用。这将影响数据流通与共享的效率。其二，数据与个人信息保护的冲突。高质量中文语料的不足是我人工智能发展的一大难题，但1月17日，中国互联网络信息中心发布的第55次《中国互联网络发展状况统计报告》显示，截至2024年12月，我国网民规模达11.08亿人，互联网普及率达78.6%。⁹可见，我国还有丰富的数据资源亟待开发，但这其中涉及大量个人信息，而目前我国个人信息保护的相关法律规定与数据收集要求相冲突。我国《个人信息保护法》第十三条规定，通常情况下处理个人信息应当取得个人同意。但在实践中较难实现，一方面，人工智能服务提供者与个人信息提供者之间的沟通渠道有限甚至没有沟通渠道，另一方面，人工智能训练需要的数据极大，则服务提供者需与大量个人信息提供者进行沟通，沟通成本极高，因此，经过个人同意的规则将使人工智能服务者无法得到充足的训练语料。同时，法律还规定了个人同意的撤回权，但此种撤回与人工智能利用集合数据推理分析的功能相悖，一旦数据进行训练库就难以抽离，否则将影响模型稳定性。¹⁰不可忽视的是，人工智能尤其是联网功能对数据的自动化抓取也与个人信息保护相矛盾。当然，我们也不能一味牺牲公民个人信息权益，但在发展的巨大需求下，需要理清法律中存在的矛盾点实现公民权益与发展创新的平衡点。其三，我国对跨境数据流动的管理较为严格，这在一定程度上限制了国际数据合作和技术交流。当前我国人工智能数据治理存在的问题，客观上制约了高质量数据的有效供给与流通，在数据不足或质量存疑的情况下，人工智能可能产生推理偏差与决策失误，而大量数据不仅给人工智能提供大量语料，也能更好的稀释数据投毒问题，增强人工智能鲁棒性。

算法层面，算法透明度以及算法问责研究的模糊性是重要的发展障碍。算法黑箱的存在已是不争的事实，故而为实现算法的透明度，“算法解释”的概念兴起，是指通过人类与机器的交互以及人类可理解的方式揭示算法决策逻辑、运行机制与结果输出的过程，核心目标是破解算法黑箱，这对维持人们对技术的信任以及促进社会稳定上有重要意义。欧盟在2018年出台的《通用数据保护条例》中规定数据主体获得解释的权利，美国在2023年发布的《人工智能风险管理框架》中也规定了算法可解释性的要求。但我国目前对于算法解释的规定刚性不足。实际上，算法解释的权利化面临难题，算法权利能否单独作为一项诉讼权利、如何

⁹ 《中国互联网络发展状况统计报告》，载中国互联网络信息中心2025年1月17日。

¹⁰ 杨清望,唐乾：《生成式人工智能与个人信息保护法律规范的冲突及其协调》，载《河南社会科学》2024年32卷12期81-93页。

判定可理解性程度、如何是为侵权、如何对权利受侵害提供救济以及如何对侵权人进行惩戒都是理清算法解释权的关键问题,否则,算法可解释权将沦为象征性权利。¹¹与之相关的还有算法问责问题。算法对决策具有决定性影响,但其固有的技术特性却导致传统责任体系失效:当自动驾驶系统因感知模型误判引发事故时,责任应归属于算法开发者、数据提供方还是车辆制造商?人工智能主体人格也是学者争论问题之一。算法的自主性与黑箱性加剧了因果关系判断难题,难以追溯解释,多责任主体也会逃避追责,平台企业常以技术中立为由规避审查义务,开发者则强调"算法自主性"推卸控制责任,使之难以定位具体责任环节。算法透明度与问责的治理问题相互影响并可能逐步深化,例如算法黑箱下,深度伪造内容影响社会安全,而问责的困难又放任这一现象出现,导致愈演愈烈。

(二) 人工智能风险治理不足

欧盟率先搭建人工智能监管制度,其中,人工智能应用将禁止运用于不可接受风险,而有条件地限制应用于高风险上,因此欧盟对于人工智能风险治理的突出特点在于风险分级分类管理。另外,欧盟还重视多元的监管主体,除立法、执行、行业协会的监管外,还注重鼓励多元主体的参与。而美国的监管主要以审慎监管为主,监管力度较为宽松,注重于人工智能的创新与发展,强调以评估潜在监管措施对人工智能的影响为决策依据,降低对人工智能发展的阻碍,保持其技术全球领先的地位。¹²

但这两种监管方式皆有弊端。欧盟的风险分级分类方式中关于级别与类型划分的标准具有一定的主观性,欧盟《人工智能法》对不可接受风险的禁止清单被学者指出存在"价值观驱动的政治判断"。并且,这种固定化的级别分类难以适应迭代的人工智能风险,动态适应性不足。根据欧盟确立的技术监管框架,系统性风险判定采用浮点运算次数(FLOPs)作为核心量化指标。当通用人工智能模型的训练计算量超过 10^{25} FLOPs阈值时,即可直接推定其具有系统性风险。但这一标准也可因为技术突破而迅速过时。虽然技术指标划分人工智能模型的人工智能监管具有标准化的特点,但其高昂的合规成本也将抑制中小企业的创新发展。与之相反,美国人工智能监管治理呈现出显著的"创新优先"导向与"软性约束"特征,其核心逻辑在于通过弱监管维持技术竞争优势,但这一方式面临治理效能不足的困境。联邦层面长期采用"原则性框架+行业自律"模式,拜登政府虽发布《安全、稳定、可信的人工智能行政令》要求头部企业备案安全测试结果,却因缺乏强制执行力度导致合规依赖企业自愿承诺。

我国初步形成了以风险分级分类为核心、包容审慎为导向的监管模式。基于

¹¹ 戴维,王锡梓:《算法透明机制的局限性及其克服》,载《华东政法大学学报》2025年28卷01期33-44页。

¹² 和军,杨慧:《ChatGPT类生成式人工智能监管的国际比较与借鉴》,载《湖南科技大学学报(社会科学版)》2023年26卷06期119-128页。

《生成式人工智能服务管理暂行办法》等规范性文件，监管部门通过构建风险评估、风险管理和风险沟通的三环节框架，实现了对算法推荐、深度合成等高风险场景的重点管控。这种治理路径借鉴了欧盟基于风险规制的经验，但又注入了本土化改良，如率先建立生成合成内容强制标识制度，将科技伦理审查前置到研发环节，体现出安全与发展并重的价值取向。

然而，现行监管体系面临诸多挑战。首先，监管架构呈现碎片化特征，缺乏统一权威的治理机构。当前人工智能监管权分散于网信办、工信部、科技部等多部门，导致监管标准衔接不畅，协同成本高企，不同部门规章存在规则重叠问题。其次，治理工具迭代滞后于技术迭代，监管部门仍倚重传统的事前审批与事后追责机制，对监管沙盒、伦理设计等新型治理工具应用不足，难以应对大模型自主进化带来的“黑天鹅”风险。最后，现有监管过度聚焦技术应用端风险，对基础模型层和算力基础设施层的治理关注不足。这些缺陷表明，我国人工智能风险治理亟需从分散式监管向体系化治理跃迁，通过立法整合监管资源、升级监管工具箱，实现发展与安全的动态平衡。

（三）数字博弈背景下我国人工智能规则制定话语权不足

人工智能作为新一轮科技革命的核心驱动力将重塑全球经济格局，在这场决定未来百年方向的规则博弈中，国际话语权的争夺已演变为没有硝烟的战场。中国虽在人工智能技术研发、产业应用等领域跻身世界前列，却在规则制定层面陷入“技术上的巨人、规则上的矮子”的困境，这也将使我们在数字时代的国际竞争中陷入被动局面。根据世界知识产权组织发布的报告，2014 年—2023 年，中国生成式人工智能专利申请量超 3.8 万件，居世界第一，是第二名美国的六倍。¹³然而，我国在规则制定领域却频频失语。内在原因看，我国国内治理框架存在碎片化未能形成统一的规范框架，技术标准体系也相对之后，因而难以形成具有国际影响力的治理范式。另一方面来源于美西方国家对于国际规则制定话语权的垄断。例如，七国集团（G7）联合声明提到，要以价值观划线来推动前沿科技的发展，推动实施人工智能治理，意图构建科技“小圈子”，阻止中国参与人工智能等技术标准的制定。¹⁴此类对我国规则制定参与的排除，使我国在关键议题上丧失议题设置能力。

而技术治理之争已演变为技术霸权之争，暗含本国价值观、利益的输出。欧盟《人工智能法案》将社会信用频分与远程生物识别应用列为不可接受风险，可看出对中国相关制度与社会治理实践的误读。¹⁵美国主导的《布莱切利宣言》签署国中，新兴经济体仅占席位三分之一，暴露了全球南方在规则制定中角色的边缘化。在数字鸿沟如此之大的背景下，发展中国家可能面临“数字殖民”。并且，

¹³ WIPO:Patent Landscape Report - Generative Artificial Intelligence (GenAI)

¹⁴ 中国日报网：《推进人工智能全球治理，推动 AI 技术造福全人类》2023 年 10 月 19 日。

¹⁵ 丁晓东：《人工智能风险的法律规制--以欧盟〈人工智能法〉为例》，《法律科学》2024 年第 5 期。

“规则武器化”的趋势也愈加明显，美国通过“云法案”构建数据长臂管辖，欧盟借《通用数据保护条例》树立数字贸易壁垒。这种隐形的规则霸权，比技术封锁更具杀伤力。技术优势若不能转化为规则制定权，终将在国际博弈中沦为附庸。我国面临的规则话语权困境，既是新兴大国突破制度围堵必须跨越的鸿沟，也是构建更加公正合理国际秩序的必经之痛。如何在技术突破与规则创新之间找到平衡点，不仅关乎我国在数字时代的国际地位，更决定着全球人工智能治理体系的未来走向。

三、总体国家安全观下人工智能治理路径构建

在总体国家安全观指导下，人工智能治理需要突破传统安全治理的碎片化思维，构建涵盖数据主权、算法权力、风险传导、国际规则的全方位治理体系。面对数据权益模糊、算法透明度缺失、风险治理滞后、国际话语权不足等核心困境，应当以“大安全”理念为统领，以人民安全为导向，统筹国家安全各领域、协调国内国际两个大局、平衡安全与发展双重目标，形成具有中国特色的治理范式。

（一）坚持“一个总体”安全体系，引领科技伦理先行

2024年11月，习近平主席在二十国集团领导人第十九次峰会上指出：“要引领数字化转型、数字经济和实体经济深度融合、新兴领域规则制定，加强人工智能国际治理和合作，确保人工智能向善、造福全人类。”

贯彻科技向善理念是破解人工智能国家安全风险的核心路径，系统整体规制人工智能技术，维护“大安全”体系。科技向善强调技术发展应以服务人类福祉、保障公共利益为根本导向，通过伦理规范、法律框架与技术治理的协同，引导人工智能向安全可控的方向演进。故而，需构建“伦理先行”的价值规范体系，将国家安全、社会公平、隐私保护等核心价值嵌入人工智能研发与应用全流程。2022年3月，中办、国办印发《关于加强科技伦理治理的意见》，提出了完善科技伦理审查规则流程，健全科技伦理（审查）委员会设立标准、登记制度，建立科技伦理审查结果专家复核机制等具体要求。通过设立伦理审查委员会，对涉及敏感领域的AI技术开展前置评估，防范技术滥用对国家安全造成的系统性威胁。同时，强化多方协同治理机制，推动政府、企业、科研机构及公众共同参与风险治理。政府可借鉴欧盟《人工智能法案》经验，明确禁止人工智能干预选举、制造虚假信息等危害国家安全的行为；企业则需完善内部合规体系，落实数据安全审计与算法透明度要求，如《生成式人工智能服务管理暂行办法》中强调的“安全可控、透明可信”原则。科技向善不仅是伦理要求，更是国家安全战略的重要组成部分，唯有通过价值引领与制度约束，方能实现人工智能发展与国家安全保障的平衡统一。

（二）统筹发展与安全问题，平衡数据算法创新与规制

数据治理的核心矛盾在于如何平衡数据要素的市场化配置与安全保护。首先需破解数据权属不明的根本性问题。数据作为新型生产要素，其法律属性的模糊性不仅导致交易主体权利义务关系不清，更阻碍了数据要素的市场化流通。《关于构建数据基础制度更好发挥数据要素作用的意见》提出的“数据产权结构性分置”思路，将数据占有权、收益权、处分权利用权作为基础性权利，将这几项基础权利进行结构化分离。¹⁶确保真实数据提供者获取收益，合成数据生成者拥有完整权益，模型开发者享有使用权，从而厘清数据权属关系。其次需建立差异化的个人信息处理机制。譬如通过信息分类与场景化风险评估实施分层同意制度：对于生物识别、医疗健康等敏感信息采用严格的事前明示同意，对匿名化处理后的行为轨迹等低风险数据允许概括同意，而对公共治理所需的城市交通流量等非识别性数据则可探索默示同意机制。这种动态平衡既能缓解语料短缺压力，又可避免《个人信息保护法》的刚性约束对技术创新的过度抑制。跨境数据流动管理则需改变“一刀切”式管控，在自贸试验区先行试点数据分级流动制度，对涉及国家安全的核心数据实施绝对禁止出境，对重要数据建立安全评估与跨境传输协议双轨机制，对一般数据探索负面清单管理模式，如此既可保障数据主权，又能为国际科研合作与商业应用开辟通道。

算法治理的深化需要构建透明、可追溯、可问责的技术监管体系。针对算法透明度不足的顽疾，应当确立分层次的算法披露义务：基础层算法需向行政主管部门备案核心参数与训练逻辑，应用层算法应向用户公示决策维度和影响权重，而涉及公共利益的医疗诊断、司法评估等高风险算法则需接受第三方技术审计。算法透明不仅要求技术代码的公开，更需实现社会价值的可解释性。可借鉴欧盟《人工智能法案》的经验，要求开发者在算法设计阶段嵌入伦理评估模块，通过可视化界面展示数据偏差检测结果与纠偏路径，使普通用户能够理解算法决策的因果关系链。在归责机制层面，学者王莹提出的“算法侵害责任框架”为破解技术黑箱难题提供了新思路：对具有明确因果关系的算法歧视或错误输出适用过错推定原则，由运营者自证技术合规性，遭受严重损害时拓展使用集合式责任要求各技术主体承担责任；对深度学习模型自主演化产生的非预期损害，引入信义与公共义务，要求算法设计、应用者对信息数据及防范风险外部化承担责任。¹⁷针对深度伪造等技术滥用问题，除强化《生成式人工智能服务管理暂行办法》中合成内容标识义务外，更需研发数字水印、区块链存证等技术手段，构建从数据采集到内容生成的全生命周期溯源体系。

（三）坚持安全动态平衡，构建场景治理与敏捷治理机制

动态平衡理论要求根据国家安全形势的变化，在维护政治安全根本地位的

¹⁶ 张新宝：《论作为新型财产权的数据财产权》载《中国社会科学》年 2023 第 4 期 144-163 页。

¹⁷ 王莹：《算法侵害责任框架刍议》载《中国法学》2022 年 03 期 165-184 页。

前提下，灵活调整各领域安全策略，实现静态安全状态与动态治理能力的有机统一，最终形成内外联动、层次分明、持续演进的国家安全格局。场景治理与敏捷治理便是动态平衡理论的表现。

应用场景是数字技术的末端应用环节。不同场景可能产生的风险不同。数据与技术在某些场景下敏感而在其他场景下则不然，譬如人脸识别技术主要运用于核查身份、身份背景研判、追捕追逃这三类场景，而这三类场景对技术运用程度、对个人权利的干预与社会的影响各不相同，[颜飞,刘文琦：《人脸识别技术在刑事司法中的场景化应用、风险及治理》，载《学术交流》2023 年 04 期 59-73 页.]可见不同的场景所需治理的侧重点不同。故而，场景治理强调以特定情境的特殊性为导向考察人工智能应用风险水平，相应制定与特定场景的公共治理需求与治理问题相适配的规制策略，通过“将即时精准的个性化信息融于场景的丰富体验之中”，实现了技术赋能与社会治理需求的动态适配。并且，场景治理的交互式运行特征决定了其能够深入用户生活情境，与用户建立连接和互动，产生新的治理红利，这种实时互动机制为防范人工智能技术滥用提供了动态监测的可能。

但人工智能技术的跨界融合特性使得单纯场景治理面临“新应用场景不断涌现”的挑战，这就需要敏捷治理机制提供动态调整能力。2022 年 3 月，中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》，对科技伦理治理进行了顶层设计，明确提出了敏捷治理的要求。敏捷治理是一种具有敏捷性、适应性、持续性、自组织性和包容性的治理过程。¹⁸强调构建事前预防与时候应对相结合的适应性治理机制。敏捷治理中，学习利用监管沙盒制度等弹性监管技术至关重要。监管沙盒可在风险可控的前提下建立真实、可控的试验环境，动态且常态地检验人工智能风险，在保障国家安全底线的同时为技术创新留出试错空间。例如将算法黑箱、数据泄露等常见风险纳入快速小版本迭代之中。这种动态治理能力对于维护国家安全至关重要，特别是传统风险评估手段失效时，敏捷治理的“红队测试”和“基准测试”等技术机制能够快速响应新型安全威胁。

场景治理与敏捷治理的协同效应尤为突出。当技术民族主义导致“算法军备竞赛”加剧时，我国通过场景化治理构建的“数字高边疆”，既防范了“网络攻击主体凭借算法优势进行入侵”，又通过监管沙盒培育自主技术生态。这种双重治理机制有效平衡了安全与发展诉求。在金融、医疗等关键领域，场景化治理划定安全边界，而敏捷治理则通过“技术监督技术”提升监管穿透力，既避免过度监管扼杀创新，又防止技术失控危及国家安全。

（四）统一自身安全与共同安全，构建国际规则制定话语权

¹⁸ 陈万球：《科技伦理敏捷治理的三重维度》，载《光明日报》2023 年 9 月 18 日 15 版。

获得国际话语权是扩大自身理念与发展道路国际影响力的过程，除此之外，我国人工智能发展坚持为全球南方发声，也为共同安全作出贡献。目前全球人工智能规则博弈呈现碎片化、软法主导、阵营分化的特征，为我国争夺国际规则治理话语权提供了空间。2023年10月18日，习近平主席在第三届“一带一路”国际合作高峰论坛开幕式主旨演讲中提出《全球人工智能治理倡议》，发出了引领全球人工智能治理的中国强音。倡议指出我国致力于构建一个平衡各方利益、共同应对人类调整的人工智能治理新模式，并反对“小圈子”，反对特定国家将治理机制工具化，并站在广大发展中国家的立场上。¹⁹我国应当以《人工智能治理倡议》为支点，在技术标准制定、治理规范创造上构建国际话语权。

值得注意的是，我国目前占据人工智能技术发展的优势，其中我国数字技术企业的作用不容忽视。深度求索公司的 deepseek 模型已说明我国在算法优化与算法效能领域形成了独特竞争力，而我国其他科技公司在技术上也呈现百花齐放的特点。而数字企业可将技术优势转化为规则优势，以谷歌、微软为代表的超级技术企业，凭借超高的研发投入，在人工智能规则形成初期就实现了技术标准即治理规则的渗透。在 CEN 和 CENELEC 联合成立的人工智能技术委员会（JTC21）中，科技公司和咨询公司的代表成员数量占已确定的 JTC21 成员的一半以上。因此，我国也应该贯彻多元共治的理念，鼓励我国科技企业参与规则制定，尤其是软法的先行制定，这种企业先行、国家背书的规则输出模式，有利于代表我国科技发展利益的规则的快速推广。

我国作为主权国家也应该积极参与国际规则的制定。这包括两种路径：一是直接参与人工智能国际规则制定平台。二是将国内规则转化为国际规则。平台机制是话语权的载体，积极参与具有权威及主导性、基础深厚、讨论议题及规则于己有利的平台机制。目前制定中国人工智能国际规则的全球性平台主要包括联合国教科文组织、经合组织等，我国应积极发挥作用。另一方面，我国应积极构建有自身特色的人工智能治理规则，如明确定位，在反思美西方国家技术霸权的情况下，倡导体现发展中国家利益的国际法，并强调我国以人为本、科技向善的理念，将自身理念上升为国际社会认可、接受的国际共识。

基于目前人工智能规则制定的小圈子属性，美西方国家在规则制定上的垄断，我国应积极构建全球南方话语联盟，建立基于共同发展诉求的规则共同体。全球南方对发达国家在人工智能技术上的依赖性进一步加剧了全球权力格局的不平等性，最后将面临真实的安全威胁。全球南方本身就更易受到算法偏见的影响，治理规则上的偏向更不利于各发展中国家维护自身利益。中国作为新兴技术大国，在重塑全球南方话语权格局中发挥着枢纽作用。一方面，通过

¹⁹ 鲁传颖：《全球人工智能治理倡议》中的中国智慧，载光明网 2024 年 1 月 23 日。

技术合作打破“数字殖民”依赖。中国以防止人工智能技术垄断与阻断供应链为核心诉求，推动“一带一路”框架下的技术标准共建，肯尼亚等国在与中国的合作中逐步摆脱对美欧标准的单向依附。另一方面，中国积极构建多边发声平台，巴西总统卢拉在 2024 年联合国大会上疾呼“将数字技术纳入普惠性发展议程”，尼日利亚科技领袖阿布拉主张“发达国家应以公平方式投资非洲数字经济生态”，这些诉求通过金砖机制、中非合作论坛等中国主导的多边渠道得以发出。

总而言之，掌握国际规则话语权能使我国在技术竞争上占据一定的优势地位，同时为我国与发展中国家抵御美西方技术治理霸权、意识形态及伦理输出，统一自身安全与共同安全具备重大意义。

四、结语

人工智能技术的迅猛发展正以前所未有的速度重塑全球格局，deepseek 模型的崛起即是我国人工智能技术发展的彰显，也催生了更多与国家安全密切相关的风险。本文基于总体国家安全观的系统思维，揭示了人工智能技术特性与国家安全的关联，剖析了我国在数据权属界定、算法透明度缺失、风险治理滞后、国际规则话语权不足等领域的深层困境，并提出了以总体国家安全观理论为基础的治理路径方案。。

面对人工智能技术带来的复合性安全挑战，我国需以“大安全”理念为统领，将科技向善的伦理准则嵌入技术研发全流程，通过数据权益分置改革破解要素流通瓶颈，以场景化治理与敏捷监管应对技术迭代风险，在筑牢国家安全屏障的同时释放创新活力。在国际规则博弈中，既要依托《全球人工智能治理倡议》凝聚发展中国家共识，推动技术标准共建与治理经验共享，也要鼓励科技企业参与全球技术生态构建，将技术优势转化为规则话语权优势。唯有坚持安全与发展的动态平衡、自主创新与国际协同的统筹推进，才能在全球数字博弈中占据战略主动。

展望未来，人工智能治理体系的完善不仅是技术命题，更是国家治理能力现代化的时代答卷。在总体国家安全观的指导下，中国有望以包容审慎的治理智慧、协同创新的制度设计、共商共建的全球视野，为破解人工智能治理的“阿喀琉斯之踵”提供中国方案，在守护国家核心利益的同时，推动构建公平普惠、安全可控的全球人工智能治理新秩序，为人类命运共同体建设注入科技向善的持久动能。

人工智能全球治理：机制构建、挑战与中国的创新路径

付译瑶¹

摘要：人工智能的全球治理在其迅猛发展背景下至关重要。本文深入剖析其现状、困境与中国路径。当前，人工智能全球治理处于秩序形成关键期，呈现国际规制性立法走强、多边合作架构渐成等特点，但也面临发展中国家参与度低、现有治理机制协调性不足、技术规则协调困难以及治理实施效果不佳等困境。中国参与人工智能全球治理，以构建包容共享的治理体系为愿景，秉持发展、安全、公平与合作并重的原则，通过制度、技术、能力与伦理协同推进的行动来支撑。在实践路径上，中国深化多边合作机制建设，聚焦技术安全与伦理标准制定，注重数据跨境流动规则建设并加快落实治理承诺。中国方案为全球治理提供重要借鉴，未来应强化国际协调，中国也需持续发挥桥梁作用，推动人工智能技术红利普惠共享，迈向智能时代“善治”未来。

关键词：人工智能；全球治理；中国路径；治理困境

1 引言

党的二十届三中全会审议通过的《中共中央关于进一步全面深化改革、推进中国式现代化的决定》高屋建瓴，强调要扎实落实“三大全球倡议”，深度参与并有力引领全球治理体系的变革与重塑，妥善化解外部风险挑战。其中，针对人工智能领域，明确提出要“建立人工智能安全监管制度”，鼓励并支持中国在人工智能国际交流合作中发挥积极作用。在当下全球化浪潮中，人工智能作为极具变革性的新兴技术，其全球治理的重要性愈发凸显。人工智能技术的发展日新月异，早已突破国界藩篱。通过科学合理的全球治理，能够为人工智能营造健康的创新环境，推动技术在全球有序传播，让世界各国共享技术进步带来的丰硕成果，为全球经济的持续增长和可持续发展提供坚实支撑。然而，伴随人工智能的飞速发展，诸多人工智能全球治理困境和挑战也逐渐显现。在国际治理对话平台上，发展中国家的声音常常微弱，难以充分表达自身诉求；规则制定沦为部分国家的竞争工具，呈现出被武器化的危险态势；各类治理机制各自为政，缺乏协同联动；治理措施在实际推行中大打折扣，与预期承诺相去甚远。这些难题如同一层层阻碍，严重制约着人工智能技术红利的充分释放，对构建全球“善治”的理想格局形成巨大挑战。

¹ 付译瑶，女，2001年7月6日出生，河南财经政法大学民商法学院，民商法学专业在读硕士研究生，1226728178@qq.com，17803825111，河南省郑州市金水区金水东路180号河南财经政法大学郑东校区，450046

在人工智能全球治理中，中国积极推动全球人工智能研究资源的交流互动、深度合作，大力倡导开放共享。2024年7月4日，上海世界人工智能大会发布《人工智能全球治理上海宣言》，从人工智能发展、人工智能安全、人工智能治理体系、人工智能社会、社会福祉五个方面，提出对人工智能全球治理的政策框架进行改进。如今国际科技竞争日趋白热化，技术规则壁垒不断加固，科技发展差距也在持续拉大。在这样的大环境下，中国关于人工智能全球治理的愿景构想与治理架构，犹如一盏明灯，为未来人工智能全球治理照亮了前行的道路，提供了极具价值的有益借鉴。因此，梳理人工智能全球治理的现实状况和面临的困境，融合中国在实践中提出的极具针对性与可行性的“中国方案”，不仅是推动人工智能技术真正造福全人类的关键之举，更是积极践行构建人类命运共同体的必然要求。

当前，人工智能全球治理成为学界焦点，多数研究聚焦于分析全球人工智能治理倡议，深入剖析倡议中的核心内涵，探寻其中所蕴含的中国智慧与中国方案。从国际关系层面来看，众多研究主张构建人工智能全球性共识，并针对信任困境探索解决办法；而在全球经济领域，研究人员依据人工智能特性来解释分析其对全球经济治理秩序的冲击机制，同时在全球治理的框架下探讨中国的应对实施路径。然而，当前现有的相关研究存在一定局限性：一方面，研究大多停留在宏观层面，侧重于论述人工智能全球治理的理论重要性与抽象内涵，对微观层面的关注较少；另一方面，所提出的实践路径多从国内视角出发，从全球角度给出的治理方案则相对匮乏。鉴于此，本文将宏观与微观视角进行有机结合，深入剖析人工智能全球治理所面临的新挑战，并从国际视角提出兼具合理性与可行性的中国路径，为人工智能全球治理贡献新思路。

2 人工智能全球治理的内涵与现状

2.1 人工智能全球治理的内涵

在科技浪潮的持续奔涌下，人工智能强势崛起已然成为引领人类大步迈入智能时代的关键驱动力。回首历次工业革命，每一项新兴技术都如同双刃剑：一方面，为人类社会的发展开辟崭新道路，推动其跃上新台阶；另一方面，也向社会经济治理发起全新挑战。如今，人工智能在带来划时代变革的同时，诸多社会问题与安全隐患也相伴而生。从隐私泄露风险到算法偏见导致的不公平，再到就业结构面临的巨大冲击，这些问题都迫切呼唤着人工智能全球治理体系的完善。所谓人工智能全球治理，就是国际社会借助一系列政策、规则、标准、机制以及合作形式，对人工智能技术在全球范围内的开发、部署与应用进行全方位协调、监督和管理。在这一进程中，各国携手应对人工智能引发的跨国界挑战，充分挖掘并发挥人工智能的优势，以此促进全球科技公平、安全且可持续地发展，推动社会不断进步，这已然成为人工智能全球治理的核心方向。

人工智能的发展浪潮席卷各个领域，并不限制于单一产业，这也使得人工智能全球治理暴露出诸多负面特征。首先，人工智能技术治理尚不完善。人工智能算法存在被恶意篡改或本身有缺陷的风险，这对大量个体信息、企业秘密信息甚至国家机密的安全构成严重威胁。在人工智能完整的生命周期内存在着诸多问题，如研发设计时容易出现模型缺陷、训练阶段可能产生数据的偏差以及部署运营后又面临适应性与更新维护问题等，而目前对于这些问题，尚未形成体系化、制度化且成熟的解决方案与治理模式。其次，人工智能的应用版图不断拓展，发展态势迅猛，其负面影响存在全球范围内快速蔓延的风险。在信息流通不畅、各方掌握信息不对等的情况下，人工智能产生的负面溢出效应和下游影响会跨越国界，波及全球。而且人工智能技术成果在全球传播和应用的速度极快，任何一个国家在人工智能领域取得的重大突破或出现的失误，都会如同在平静湖面投入巨石引发连锁反应，通过信息交流等渠道迅速扩散到其他国家，产生难以预估的影响。最后，人工智能全球治理存在显著差距。随着人工智能在全球的广泛应用，一系列伦理道德和社会公平问题也随之而来，这些问题已然突破了地域限制从而呈现出显著的全球化特征；算法偏见等负面因素逐渐凸显，可能左右着不同种族和性别在就业、教育等资源分配中的结果，致使不公平现象的产生。不仅如此，全球范围内的人工智能倡议与相关机构各自为政，缺乏统一的协调机制，正将世界分裂为多个相对独立的人工智能治理体系，阻碍全球协同发展。而在政策法规层面，不同国家之间的巨大差异也给跨国人工智能业务带来了重重阻碍。企业在开展跨国业务时，常常会陷入监管套利的漩涡或是面临合规困境。因此，确保人工智能技术在全球范围内遵循法律、合乎规范、契合伦理道德框架发展，已成为人工智能全面治理的必要前提。

在人工智能技术蓬勃发展的当下，与之相关的伦理与安全风险日益凸显。在此形势下，人工智能的全球治理体系逐渐形成，其以法律法规为核心，行业标准、倡议以及国际合作作为重要补充。国际上涉及人工智能伦理与数据保护等方面的法律文件、共识性建议书，已然成为推动人工智能全球治理不可或缺的工具。²

2023 年，全球主要经济体如美国和欧盟纷纷在人工智能监管法规制定方面发力。以美国为例，其着眼于人工智能在社会治理方面的影响，于 2023 年 1 月 26 日正式发布《美国 AI 风险管理框架 1.0》（AI RMF 1.0）。该框架的制定集合了政府、私营企业、科研机构及行业协会等多方力量，利用立法手段针对人工智能的重点领域进行规范。治理范畴覆盖了人工智能系统全生命周期可能产生的风险，不仅涉及对人类、组织造成的危害，还将生态系统纳入考量范围，同时高度关注 AI 的可信度，致力于保障其有效运行。在政策框架构建上，AI RMF 1.0 提

² 苗逢春.从“国际人工智能与教育会议”审视面向数字人文主义的人工智能与教育[J].现代教育技术,2022,32(02):5-23.

出了“三位一体”的创新模式，即风险评估、信息共享与坚持问责。具体而言，设立专门的风险评估机制，组织专业力量对 AI 系统风险展开全面评估并给出针对性建议，进而制定技术操作和管理措施；搭建信息共享平台，促进各方交流与合作以实现组织内外部风险与决策信息的流通；严格落实问责机制，清晰各参与方的职责，确保 AI 系统一旦出现风险能够追根溯源。

无独有偶，中国网络安全标准化技术委员会在 2024 年 9 月发布了《人工智能安全治理框架》。与美国不同的是，中国这一框架的制定广泛吸纳了政府部门、企业、科研院所、民间机构以及社会公众的意见，其治理内容既涵盖人工智能内生安全风险，也囊括应用安全风险。中国的治理框架在理念上更强调包容、协同发展等共享型理念：一是秉持包容审慎、确保安全的原则，在积极推动创新的同时，绝不放松对安全底线坚守；二是以风险为导向，灵活开展治理工作，时刻紧盯风险动态从而能够及时调整治理策略；三是强调技术与管理相结合，协同各方力量共同应对，通过综合运用技术与管理手段，明晰各参与方在人工智能发展中的责任。

2.2 人工智能全球治理正处于秩序形成关键期

随着科技呈爆发式发展，全球一体化进程稳步向前，人工智能全球治理领域正经历深刻变革，呈现四大鲜明特点：国际规制结构加快更迭、多元人工智能框架逐步构建、政策核心朝着多维度治理过渡、治理主体日益多元。可以说，重塑人工智能全球治理格局已步入关键时期。

2.2.1 国际规制性立法走强，扩张性政策缩量

近年来，人工智能领域的政策导向正发生着深刻转变，国际规制性立法持续强化，重点聚焦于保障技术安全与伦理规范，旨在为人工智能的发展筑牢底线；而扩张性政策逐渐式微。

早期各国的人工智能法案多以扩张性为主，目的是快速提升本国在人工智能技术领域的实力，在全球竞争中占据优势。但随着人工智能与社会各领域的融合不断深入，数据泄露、算法偏见、伦理困境等潜在危害逐渐暴露，政策制定者们开始深刻认识到规范人工智能发展刻不容缓。如今，国际规制性立法朝着保障技术安全与伦理的方向稳步推进，通过对人工智能的使用加以限制，为其发展划定清晰的规则与界限。2023 年，全球范围内通过了 28 项 AI 相关法案，其中 18 项为中等相关度的监管法规，这充分说明了各国已将 AI 风险管控视为重中之重，开始通过立法手段为人工智能的发展戴上“紧箍咒”。以欧盟为例，2023 年 12 月颁布的《人工智能法案》构建了基于风险的监管框架，依据风险等级对人工智能系统进行分类管理：严格监管高风险系统，从技术标准到应用程序，提出了多项合规要求；对于低风险系统则仅要求其履行基本的透明义务，同时明令禁止存在过高风险的 AI 系统。

由此可见在人工智能全球竞争白热化的今天，仅仅利用扩张性政策已难以维系竞争优势。严格的规制性立法，不仅能提升人工智能产业的门槛与标准，更有利于维护国家利益。目前，各国政府与国际组织在人工智能全球治理方面正逐渐达成共识，愈发注重对人工智能技术发展与应用的规范监管，致力于保障技术安全、可靠且符合道德伦理准则。

2.2.2 多边合作架构渐成，“分而治之”倾向加剧

在人工智能蓬勃发展的当下，构建多边合作框架已成为国际社会的重要议题。各国基于共同利益积极探索人工智能治理共识，多边合作框架正在稳步搭建。2023 年 11 月，在英国人工智能安全峰会上中美等 28 国共同签署的《布莱切利宣言》，标志着国际社会在人工智能治理方面初步形成共识。与此同时，联合国等国际组织也把人工智能全球治理纳入工作范畴，推动治理朝着合理有序方向发展。

然而在多边合作框架构建进程中，“分而治之”趋势则日益显著。由于各国各地区的国情与发展阶段不同，在治理需求和理念上也存在着显著差异。欧盟颁布了《通用数据保护条例》，以严格的法律约束规范企业数据使用，对违规企业实施高额处罚；而美国加利福尼亚州推行的《加州消费者隐私法案》更加侧重于行业自律，数据隐私的法律规定则相对宽松。

这种差异不仅体现在法规层面，更深植于各国的发展战略与利益诉求中。发达国家凭借技术优势，希望通过人工智能巩固全球科技与经济领先地位，在国际规则制定中占据主导。它们尤为重视知识产权保护与数据安全，力求在人工智能发展中维护自身核心利益。发展中国家则更关注人工智能技术的普及与应用，期望借此提升经济发展水平与社会治理能力，保障国家数据主权，因此在治理时强调技术的公平性与可及性。

国家利益与价值观念的分歧进一步加剧了人工智能治理模式的差异。美西方部分国家担忧中国在人工智能领域的快速崛起，会冲击其长期把持的技术优势与全球主导权，因而采取技术封锁等手段，限制本土企业提供先进的人工智能和半导体技术给中国。这种单边主义行径不仅破坏了国际合作氛围，还使得全球人工智能治理关键问题上的分歧日益扩大，严重阻碍了全球治理机制的顺利建立。

2.2.3 政策转向多维度治理模式，覆盖领域逐步拓宽

人工智能政策导向历经深刻变革，政策重心逐步从单一维度转向多维度治理，治理覆盖领域也随之不断拓宽。这一转变不仅反映了人工智能技术的快速发展，也体现了社会对其广泛应用所带来影响的全面考量。

早期，全球人工智能政策主要聚焦于技术安全与信息数据保护，防止关键战略资源的泄露；随着人工智能在医疗、交通、金融、教育等行业的广泛渗透，政策关注点逐渐扩展到技术研发、数据管理、应用推广及社会影响评估等全生命周

期环节。如今的人工智能已深度融入于人类生产生活的方方面面，从便捷的知识获取到复杂的市场监管，其发展的复杂性与全球性迫切需要多维度、全方位的治理体系以维护人类利益与社会稳定。

这种多维度治理的趋势不仅体现在政策关注点的扩展上，更体现在各国为将人工智能外部性影响内部化所采取的具体措施中。例如 2023 年 7 月中国发布的《生成式人工智能服务管理暂行办法》全面规范了技术研发、应用场景和数据管理等多个领域。这一举措促使企业调整业务模式与管理策略并增加合规成本，从生产源头控制人工智能风险，有效降低其可能产生的负面外部效应。

2.2.4 政企合作初显成效，多元主体共建治理新生态

人工智能飞速发展的时代浪潮下，一个全新的治理格局正悄然形成——政企合作治理模式初露锋芒，治理主体也日益迈向多元化的发展方向。

过去的人工智能治理多依赖政府或者国际组织单打独斗，如今政府与企业合作的深度融合，共同为人工智能治理添砖加瓦。随着这种合作模式的兴起，人工智能治理的舞台上不再是政府或国际组织的“独角戏”，企业手握前沿技术创新成果与丰富实践经验，政府则在政策制定与监管层面独具优势，二者强强联合、互联互通，为人工智能治理体系注入了前所未有的强大活力。谷歌、微软、Meta 等七大 AI 行业巨头在 2023 年 7 月同白宫达成自愿承诺，明确表示将在 AI 系统发布前严格开展全面的安全评估工作并积极分享风险信息。而这一标志性事件，正是多方协同参与人工智能治理趋势的有力印证。

人工智能技术以其高度的复杂性著称，作为研发主力的企业，对其技术原理、应用场景以及潜在风险较于政府有着更为深刻的理解。政企合作让政府得以洞悉技术底层逻辑，从而制定出更有针对性、切实可行的技术标准与监管措施。对企业来说，人工智能产业蕴藏着无可估量的经济潜力，积极投身治理既能助力规范发展，也有助于企业树立良好的品牌形象，抢占市场竞争的有利地位。

值得一提的是，人工智能产品与服务的应用已突破国界、走向全球；跨国企业业务遍布世界各个角落。面对如此庞大且复杂的全球化业务体系，单个国家在监管跨国人工智能业务时往往显得力不从心。要确保数据存储、处理等关键环节契合各国法律与安全标准，就迫切需要多国政府与企业打破地域和组织界限，开展深度合作、协同共治。在跨国企业业务蓬勃发展的驱动下，政企合作治理模式逐渐成型，治理主体多元化格局顺势而生，为人工智能的未来发展奠定了坚实的治理基础。

3 人工智能全球治理的挑战

3.1 全球治理参与代表性不足

在人工智能全球治理中，发展中国家参与度低，代表性严重缺失。当前人工智能领域的规则制定与战略研讨由少数发达国家主导，国际会议、标准设定组织

以及决策平台上，发展中国家声音微弱。³这是因为发展中国家在人工智能技术研发、基础设施建设和专业人才储备等方面基础薄弱，缺乏深度参与所需的技术话语权。并且，受经济实力制约，它们难以大规模投入资源开展研究和国际交流合作，在治理体系中的影响力不足。以 7 个主要的非联合国 AI 治理倡议为例，仅有 7 个发达国家参与了全部倡议，而 118 个主要为全球南方的国家完全未参与。美国等西方国家还限制人工智能技术向发展中国家扩散，遏制这些国家在高科技领域的发展，掌控技术发展方向和创新路径。发展中国家独特的需求与视角被忽视，会引发数据隐私、伦理道德标准方面的差异与冲突，导致 AI 发展红利分配不均，扩大科技鸿沟，阻碍国际政治经济格局向公平包容方向发展。

3.2 现有治理机制协调性不足

3.2.1 国际组织协调缓慢

国际组织在调和各国利益、促进协同治理方面进展迟缓，各组织间缺乏有效的沟通协调机制。在联合国系统内，诸多机构虽涉及 AI 治理，但职责划分不明确，没有具体机构能全面处理或主导 AI 治理问题。例如，联合国教科文组织侧重于从伦理和教育文化角度规范引导人工智能，却较少涉及技术标准制定和市场监管；世界知识产权组织主要关注人工智能相关的知识产权问题，无法处理伦理道德问题。

3.2.2 规则竞争化与武器化

各国将人工智能视为关键技术竞争领域，发达国家掌握主导权，倾向选择或创建符合自身利益的排他性全球或地区治理机制。⁴美国、欧盟等凭借技术优势，频繁出台措施限制 AI 技术及相关产品出口，颁布法案限制人工智能人才交流，在国际人工智能合作中设置障碍，企图主导全球 AI 治理规则制定。这种行为使得规则竞争大于规则对接合作，规则武器化趋势明显。其结果可能导致世界分裂为互不兼容的 AI 治理体系，技术标准差异化加大，数据传输安全风险增加，监管政策不同使企业转移业务，造成安全监管空白，阻碍技术交流和创新扩散，还会引发不正当市场竞争。

3.3 技术规则协调存在困难

3.3.1 标准规范差异大

全球多个国家、地区以及国际组织参与人工智能全球治理规则制定，由于各方对人工智能的认识、发展需求以及应用行业特点不同，导致治理标准和规范存在差异。国际电信联盟、国际标准化组织、国际电工委员会和电气与电子工程师协会等机构都在制定 AI 标准，但这些标准之间缺乏共同语言，许多关键概念如

³ 凌敏洋,邹举.人工智能全球治理挑战与中国应对之策[J].中国军转民,2024(24):47-48.

⁴ 阙天舒,张纪腾.后疫情时代下全球治理体系变革面临的挑战及中国选择——基于实验主义治理视角的分析[J].国际观察,2021(04):125-156.

公平性、安全性、透明度等没有统一定义。

3.3.2 缺乏协调整合机制

国际上缺乏权威、统一的协调机制来整合各国治理标准和技术规范，难以达成广泛有约束力的国际协议。不同国际组织的主要职责、价值观、治理需求存在差异，使得所制定的技术规范 and 行业标准具有局限性和偏向性。国际电信联盟侧重数据传输等技术规范问题，而电气与电子工程师协会的治理标准往往反映行业内技术领先企业和专家群体的利益和技术倾向。在缺乏强制执行力的国际组织决议情况下，人工智能全球治理标准无法有效整合。各国在数据隐私方面的法律框架差异显著，如中国颁布《中华人民共和国数据安全法》构建了相关法律框架，而美国的数据隐私规则相对碎片化，在联邦层面没有统一法律。技术规范和标准的多元化加剧了国家间的竞争和战略博弈，促使地缘政治格局围绕人工智能治理进一步分化重组，影响全球治理体系的和谐稳定与有效运行。

3.4 治理实施的效果不佳

3.4.1 治理原则难以落实

国际组织提出的治理原则往往停留在理论宣示阶段，缺乏具体的实施细则与强制执行机制，难以对各国及各类主体的行为形成有效约束。各国在人工智能治理上的步调不一致，导致治理实施效果与承诺存在差距。许多治理措施仍停留在自愿性遵守层面，如日本发布的《人工智能原则》多为倡导性建议，缺乏强制执行力，企业和机构在开发应用人工智能时可选择性遵守，使得治理效果受限。

3.4.2 实际应用问题频发

尽管有安全标准和测试要求，但在复杂的现实场景中人工智能系统仍可能出现故障。由于人工智能治理滞后，其在众多应用领域面临诸多伦理与法律争议，如面部识别技术应用涉及隐私权问题，导致国际商业机器公司停止提供用于大规模监控的技术，限制了该技术在安防等领域的推广应用，治理红利无法充分释放。人工智能涉及多领域知识，专业人才短缺，且当前的监管技术难以对其进行有效监控。治理红利释放不足使企业难以充分利用人工智能创新，导致人工智能行业发展在不同国家或地区失衡，甚至对国际安全秩序构成潜在威胁。

4 人工智能全球治理的中国逻辑与中国路径

4.1 中国参与人工智能全球治理的逻辑基础

4.1.1 愿景引领：构建包容共享的治理体系

中国积极倡导构建平等包容、共建共享的人工智能全球治理体系，这一愿景在《全球人工智能治理倡议》中得到充分体现。其核心目标是打破国家和地区间在人工智能发展上的阻碍，促进不同经济和科技水平国家的交流合作，让前沿技术研发成果与应用场景探索经验能够自由流通。通过这样的方式，实现人工智能

技术对全人类的普惠，推动全球人工智能治理体系的完善，提升技术的安全性、可靠性、可控性与公平性，为全球人工智能发展营造良好环境。

4.1.2 原则导向：发展、安全、公平与合作并重

（1）发展与安全并重原则：中国将人工智能视为推动国家发展的重要技术力量，在国家战略层面积极布局，强化顶层设计和统筹规划，加大对人工智能研发的投入，力求在国际科技竞争中赢得主导权。与此同时，高度重视人工智能带来的安全风险，从技术、应用和生态等多维度加强管控，确保人工智能系统的稳定可靠运行，防止技术被滥用，保障国家和人民的安全，引导人工智能朝着有利于人类文明进步的方向发展。

（2）包容合作原则：充分尊重各国在人工智能政策和实践上的差异，倡导摒弃技术保护主义和地缘政治竞争思维。秉持全人类共同价值理念，承认并尊重各国自主选择的制度模式和发展道路，鼓励各国以开放包容的心态参与全球人工智能治理。积极推动多利益攸关方共同参与，通过对话与合作凝聚共识，接纳多元发展路径与模式，促进全球人工智能治理领域的交流与协作。

（3）公平正义原则：着重关注发展中国家在人工智能全球治理中的权益，致力于缩小数字鸿沟，推动全球人工智能治理的民主化进程。增强发展中国家在治理中的代表性和发言权，确保各国在人工智能发展与治理方面享有平等的权利、机会和规则。通过开展国际合作与援助，帮助发展中国家提升人工智能技术水平和治理能力，促进全球人工智能产业发展红利的合理分配。

4.1.3 行动支撑：制度、技术、能力与伦理协同推进

（1）制度建设行动：积极参与国际人工智能治理规则的制定，在国际舞台上发挥建设性作用，推动建立科学合理的全球人工智能治理框架。同时，不断完善国内人工智能监管制度体系，加强对人工智能研发、应用等环节的规范和管理，为人工智能产业的健康发展提供坚实的制度保障。⁵

（2）技术创新行动：持续加大对人工智能核心技术的研发投入，鼓励科研机构和企业开展前沿技术研究，推动人工智能产业的蓬勃发展。通过技术创新，提升中国在人工智能领域的核心竞争力，为全球人工智能治理提供具有创新性的中国方案和智慧，助力全球人工智能技术的进步。

（3）能力建设行动：重视人工智能人才培养体系的完善，加强高校、科研机构与企业之间的合作，培养适应人工智能发展需求的专业人才。积极促进国际交流合作，通过举办培训、研讨会等活动，帮助发展中国家提升人工智能治理能力，提高全球人工智能治理的整体参与水平。

（4）伦理规范行动：推动建立基于人类命运共同体理念的人工智能伦理框

⁵ 赵磊,刘常瑜.中国参与全球人工智能治理的现实逻辑与实践路径[J].中国井冈山干部学院学报,2024,17(05):110-117.

架，明确人工智能发展应遵循的伦理原则和道德规范。积极分享中国在伦理规范制定与实践方面的经验，引导全球人工智能的可持续发展，确保人工智能技术符合人类价值观和国际规范，避免技术发展带来的伦理风险。

4.2 中国推动人工智能全球治理的实践路径

4.2.1 深化多边合作机制建设

（1）积极参与国际规则制定：中国深入参与国际多边合作框架下的人工智能议题探讨，在各类国际组织和多边论坛中发挥积极作用。积极倡导改革国际组织的成员参与机制，增加发展中国家在人工智能治理决策中的投票权和话语权，使治理规则能够更好地反映各国的利益和诉求。通过参与国际标准制定，将中国的技术和理念融入其中，提升中国在全球人工智能治理领域的影响力。

（2）促进国际交流与合作：定期举办人工智能全球治理主题论坛，邀请各国政府官员、企业界人士和专家学者共同参与，为各方提供交流观点和经验的平台。鼓励发展中国家代表分享自身在人工智能发展和治理中的独特视角和实践经验，促进不同国家之间的相互理解和学习。同时，加强与发达国家在人工智能安全、伦理等关键领域的对话与合作，学习借鉴其先进的治理经验和技術，推动全球人工智能治理的协同发展。

（3）完善国内治理体系：借助全球“AI 热”的契机，从多维度完善国内人工智能治理体系。加强制度监管，制定和完善相关法律法规和政策措施，规范人工智能产业的发展秩序。提升技术智治水平，利用先进的技术手段加强对人工智能的监测和管理。通过建立强大且适应性强的监管实践，增强全球对中国人工智能治理的信任。设立人工智能全球治理专项基金，支持发展中国家开展人工智能相关项目和研究，帮助其优化人工智能资源配置，完善国内治理体系。

4.2.2 聚焦技术安全与伦理标准制定

（1）构建技术伦理标准体系：以构建兼顾安全与创新的技术伦理标准体系为核心目标，广泛征求全球各界对人工智能安全规范的意见，确保标准的科学性和合理性。将中国传统哲学思想中的有益元素融入人工智能伦理标准，如“仁智合一”思想，为国际标准制定提供独特的中国视角和智慧。通过推动技术伦理标准体系的建设，引导人工智能技术在安全、可靠的轨道上发展。

（2）推动国际产业创新合作：积极引领国际合作平台的搭建与规则塑造，汇聚全球智慧和资源，推动人工智能产业的持续创新。鼓励企业加大在深度合成和大模型技术等前沿领域的研发投入，引导企业开展良性竞争，增强人工智能知识外溢效应。通过建立人工智能创新合作园或产业集聚区，促进技术交流与合作，形成技术优势，推动人工智能产业的创新发展，提升全球人工智能产业的整体竞争力。

（3）促进国家与全球治理统一：顺应全球技术与产业革命的发展趋势，推

动人工智能的国家治理与全球治理的有机统一。在强化自身人工智能技术发展优势的基础上，制定和完善扶持优势智能产业的政策工具箱，助力企业推进数字化转型与智能化改革。通过政策引导和支持，促进人工智能技术在国内的广泛应用和健康发展，同时积极参与全球人工智能治理，为全球人工智能发展贡献中国力量。

4.2.3 注重数据跨境流动规则建设

（1）构建多层次监管框架：为保障国家数据安全，中国构建多层次的数据跨境流动监管框架，平衡数据开放与安全的关系。通过与主要贸易伙伴和数据大国签订双边或多边数据跨境流动协议，积极参与全球数据治理规则的制定，推动建立公平合理的数据跨境流动机制。在国内，建立必要的数据安全审查制度，以分类分级、风险评估为核心手段，制定分层级的数据跨境流动分类管理规则，确保关键数据和核心技术的安全。

（2）加强数据安全基础设施建设：加大对国内数据安全基础设施建设的投入，建立数据备份与恢复中心，提高数据存储和处理的安全性和可靠性。通过构建灵活性与全面性结合的跨境数据流动统一架构，加强对数据跨境流动的监管和管理。确保在面临自然灾害、网络攻击等意外事件时，数据的完整性和可用性得到保障，维护国家数据安全平稳运行，为人工智能的发展提供安全可靠的数据环境。

4.2.4 加快落实人工智能治理承诺

（1）推动标准规范实施：积极推动各国政府和国际组织共同制定和实施人工智能治理标准和规范，减少企业在全球范围内开展人工智能业务的障碍和风险。通过制定具体的实施计划和时间表，确保人工智能治理承诺得到有效落实。加强对人工智能系统安全性和可靠性的监管，强化监管技术的研发和应用，提高监管效能，保障人工智能技术的安全应用。

（2）加强宣传教育与监督评估：支持各国政府和国际组织开展人工智能治理宣传和教育活动，提高公众对人工智能治理的认识和理解，增强公众的安全意识和风险防范意识。加强对企业和机构人工智能治理实施情况的监督和评估，确保其遵守相关标准和规范。通过建立健全监督评估机制，及时发现和解决人工智能治理中存在的问题，推动人工智能治理的有效实施。

（3）培养专业人才与促进国际合作：加快培养具备专业素养的人工智能治理人才，加强高校和职业教育机构在人工智能相关专业的建设，培养既懂技术又懂治理的复合型人才。推动人工智能监管技术的研发和应用，促进人工智能治理的国际合作和交流。通过分享经验、技术援助等方式，帮助发展中国家提升人工智能治理能力，共同应对人工智能带来的全球性挑战，推动人工智能治理的全球协同发展。

5 结 语

人工智能的迅猛发展深刻重塑了全球政治、经济与社会格局，其治理的复杂性与全球性已成为国际社会的共同挑战。本文系统梳理了人工智能全球治理的现状与困境，揭示了当前治理体系在代表性、协调性、技术规则统一性及实施效果等方面的深层矛盾。与此同时，中国作为新兴技术大国，以构建人类命运共同体为核心理念，提出了兼具包容性与创新性的治理方案，为全球治理提供了重要启示。

中国路径的独特价值在于其“发展、安全、公平与合作并重”的原则导向，以及制度、技术、能力与伦理协同推进的行动框架。通过深化多边合作机制、聚焦技术安全与伦理标准、构建数据跨境流动规则，中国不仅推动了国内治理体系的完善，更积极参与全球规则制定，助力发展中国家提升治理能力。上海宣言、《生成式人工智能服务管理暂行办法》等实践成果，彰显了中国在平衡技术创新与风险管控、推动全球协同治理中的建设性作用。

然而，人工智能全球治理仍面临技术规则碎片化、地缘竞争加剧、治理承诺落地难等挑战。未来需进一步强化国际协调机制，弥合发达国家与发展中国家的治理鸿沟，推动技术标准与伦理框架的全球统一。中国应继续发挥桥梁作用，倡导多利益攸关方合作，将“共商共建共享”理念融入治理实践，为全球人工智能的可持续发展注入更多确定性。唯有各国摒弃零和思维，在包容互鉴中凝聚共识，才能真正实现人工智能技术红利的普惠共享，共同迈向智能时代的“善治”未来。

不干涉原则在网络空间国际治理中的适用困境

李海旭¹

摘要：不干涉内政原则作为国际法基石，在网络空间的适用面临虚拟性、跨国性及隐蔽性等挑战。本文通过分析网络干涉行为的法理与现实困境，指出因“内政”边界模糊、胁迫性认定标准缺失及执行机制不足等问题，需通过国际合作与法律框架重构实现治理突破。法理层面，非国家行为体的干涉行为难以归因于国家加剧了责任追究的复杂性；网络空间的“内政”概念因全球性和高技术性特征而模糊，传统主权边界被解构；国际法对“胁迫性”网络干涉行为缺乏明确定义，导致认定标准主观化。现实层面，网络干涉行为的隐蔽性使得证据收集与证明困难，非物质性损害难以量化；执行机制存在执行短板，《塔林手册2.0》等软法约束力不足，区域合作机制亦受限于碎片化与执行力缺失。国际社会应当通过联合国框架凝聚共识，明确“胁迫性”标准，倡导审慎原则，加强跨国信息共享与追责机制，并推动技术赋能与国际法动态改革。

关键词：网络空间；网络干涉行为；不干涉原则

一、问题的提出

随着全球化进程的推进和网络技术的迅速发展，网络空间已经成为全球政治、经济和社会活动的核心场域，成为新的国际治理领域。数字时代为各国提供了巨大的发展机会的同时，也带来了新的挑战，网络空间的虚拟性、隐蔽性与互联互通性等特征，使得网络干涉行为不断滋生，干涉行为对国家主权的侵犯已经成为各国面临的严峻问题。这些网络干涉行为通常未达到“武力威胁”的标准，但仍有可能干涉他国内政，影响其政治、经济和文化等核心事务。这种干涉的隐蔽性特征使得其界定和治理变得复杂。因此，探讨不干涉原则在网络空间的适用具有重要的理论和现实意义。

不干涉原则，又称“不干涉内政原则”，是国际法中的基本原则之一，其核心内容是禁止一国干涉另一国的内政，在国家主权保护领域一直发挥着重要的威慑作用。随着网络空间的逐步“再主权化”，网络空间的主权原则已经得到了国际社会的广泛认同。网络空间的主权原则要求各国在网络空间内拥有自主控制的权利，这使得不干涉原则成为主权原则的延伸。因此，网络空间中的不干涉原则适用是对国家网络主权的合法保护，维护各国在网络空间的独立性和自主性，避免外部力量通过网络手段对内政的干扰。不干涉原则适用于网络干预行为是坚持国家主

¹ 作者简介：北京化工大学文法学院在读硕士研究生，北京市昌平区，邮政编码 102200。

权原则和维护国家利益的需要,是规制当前和未来网络空间干涉行为的必要选择,也是实现网络空间国际法治的必要环节。

尽管不干涉原则的适用已得到国际社会的普遍承认,但在网络空间的具体应用中仍然面临诸多困境。首先,各国关于不干涉原则保护范围的界定存在分歧,且对于如何界定网络行为是否构成“干涉”,特别是区分胁迫性与非胁迫性行为,仍然缺乏明确的法律标准。其次,网络空间的隐蔽性和跨国性使干涉行为的源头难以追溯,难以找到有效证据来证明,以至无法归因于某一国家。此外,国际社会没有统一执行机制也成为亟待解决的难题。

二、 不干涉原则在网络空间国际治理中适用的法理困境

国际社会最早在《联合国宪章》第2条规定了联合国组织与国家之间的不干涉原则,并在此后通过联大的决议将国家之间的不干涉原则在《1970年国际法原则宣言》中确认下来。《1970年国际法原则宣言》提出了各民族和每一个国家所具有的“不可移让之权利”,明确指出一切形式的干涉均属于违反国际法,具体规定了国家应承担的不干涉的义务,还对属于不干涉的行为进行了解释与界定。²

国际法学界的许多学者往往将国际法院在对“尼加拉瓜案”的判决中指出的两大基本要素作为认定非法干涉行为的标准³:一是行为应当具有胁迫性,二是行为针对一国根据国家主权原则而允许该国自由决定的事项,即“内政”。干涉行为的具体要素为四个,即由国家或受其控制的实体实施,针对他国内政,行为具有“胁迫性”,非法性无法被排除。

(一) 各国关于不干涉原则保护范围的界定分歧

(1) 非国家行为体的行为归因性

网络干涉行动作为一种低强度、隐蔽性强的干涉手段,其实施主体具有多样性,由于网络干涉行为通常采用隐蔽性技术,导致对该行为进行追责面临着巨大的技术挑战。传统国际法主要适用于国家之间的法律关系,国家被视为唯一的法律主体,拥有承担责任和义务的能力。然而,在网络空间中,诸多网络干涉行为由非国家行为体实施,这些非国家行为体虽然在网络空间中扮演重要角色,但它们不属于传统国际法中的主体,传统国际法上的主体只包括国家和国际组织。非国家行为体往往采取匿名手段在网络空间实施干涉行为,通过技术手段隐藏其真实身份和所在地,难以被受害国或国际社会追踪到其源头。即使能够追溯到某个具体地址或身份,仍难以确认该行为体是否与某个国家有直接的联系,或得到某个国家的支持。这些非国家行为体的活动通常不受传统国际法的直接约束,即使他们的行为对他国主权造成了影响,现有的国际法律规则也无法进行追责。

网络行为的隐蔽性和互联互通性特征,使得追踪网络行动的实施者几乎不可

² 参见邵沙平:《国际法(第四版)》,中国人民大学出版社,第67页。

³ See William Ossoff, Hacking the Domain Reserve: The Rule of Non-Intervention and Political Interference in Cyberspace, Harvard International Law Journal, Vol.62:1, p.304(2021).

能仅依赖传统的归因方法，传统的“有效控制”标准也在网络空间的背景下面临适用困难。此标准要求证明国家对非国家行为者发布明确指令，并对其进行有效控制，在网络干涉行为中，虽然有些国家提供了资金、资源或技术支持，但无法证明非国家实体的行为受国家的有效控制。国际社会对于如何界定和追责这些非国家行为体仍存在较大争议，这使得归责问题更加复杂。

（2）网络空间“内政”概念及范围界定不清

当数字技术将人类活动延伸至由数据和算法构成的网络空间时，网络行为的全球性、高技术性、易接入性以及治理理念的高度意识形态化，使得传统国际法中的“内政”概念被解构，不干涉原则的适用陷入规范困境。国家在网络空间的对内和对外事务很难界分，在去边界化的网络空间，一国的网络政策可能会产生波及全球的影响力，这也是美国对根域名的管控措施，曾一度引发全球关注乃至激烈的国际性根域名管辖权争夺的原因所在。⁴

网络空间的独特性，使得“内政”的范围变得模糊。在现实空间中，某些行为可能仅对国家本身产生影响，但在网络空间中，因其全球性和易接入性，这些行为可能扩展为跨国或甚至全球性的影响。因此，传统国际法上定义为“内政”的事项，在网络空间中可能会被认为涉及外部干预，导致“内政”的外部界限难以厘定。尽管各国普遍认可“内政”包括国家的内部与外部事务，但在网络空间的背景下，不同国家对于“内政”范围的界定仍存在一定差异，因此，在网络空间中，如何界定何种行为属于“内政”，尚无明确的标准。随着技术进步和国际政治的变动，“内政”外延呈现出动态变化的特征。原本被认为仅仅属于一国内政的事务，在网络环境下，其干预可能会波及到其他国家，甚至在全球范围内产生影响。此种背景下，如何界定网络空间中的“内政”边界，成为网络空间国际治理中的难题之一。

（二）网络干涉行为胁迫性认定标准模糊

首先，胁迫性与非胁迫性难以进行有效界分。国际法学界普遍认可胁迫性是界定非法干涉行为的重要指标，但国际社会对何为胁迫并未形成定论，“胁迫”一词并没有明确的定义，这导致胁迫性与非胁迫性网络干涉行为之间的区分标准模糊不清。尽管网络行为的严重性可以按程度划分，从“武力攻击”到单纯侵犯主权的行为，但对于是否构成“干涉”，关键在于是否存在“胁迫”行为。由于“胁迫”没有清晰界定，同样的网络行为在不同情境下可能产生不同后果，这使得胁迫性和非胁迫性网络行动的区分变得更加复杂。在通过远程制造并散布虚假的、具有诱导性或操纵性的信息，足以影响人民意愿和政府权威的形成与表达，进而干扰主权事务的条件下，网络空间干涉行动往往具有间接性和较低水平的“胁迫”性。这既是认定“胁迫”的困境所在，也是“干涉”在网络空间适用充满争议的根源。

⁴ 参见刘晗：《域名系统、网络主权与互联网治理历史反思及其当代启示》，载《中外法学》2016年2期，第518-535页。

其次，判断是否构成“胁迫”具有较高的主观性，网络行动的间接性和隐蔽性使得“胁迫”的认定变得更加模糊。在此条件下，国家在适用不干涉原则时将拥有更大的主观裁量权，从而加剧该原则在网络空间适用的任意性和不确定性。国家的“胁迫意图”是界定网络空间干涉行为的重要要素，这是为了将网络活动具有实际上的胁迫效果的情况同那些一国法律上意在胁迫的情况区分开来。在国际法中，判断是否构成干涉行为时，国家的“胁迫意图”至关重要。然而，确立国家是否有意图通过网络行为对他国施加胁迫，往往是一个复杂的问题。国际法院在“尼加拉瓜案件”中指出，即使没有明确证据表明一国有意推翻他国政府，只要其行为符合干涉的标准，仍然可能被认定为干涉行为。在网络空间中，国家可能通过间接支持或资助非国家行为体来实施干涉，但如果没有充分证据证明国家有明确的“胁迫意图”，就很难归责于该国。因此，“胁迫意图”的模糊性和归因困难，造成了实践中对网络空间干涉行为的界定和责任归属的障碍。

对于网络干涉行为的胁迫性认定，关键在于是否能够确认加害国的“胁迫意图”。然而，在网络空间中，许多国家的网络行为往往是通过间接手段、匿名的方式进行，难以直接证明是否存在胁迫他国的意图，国家实体可能通过资助或支持黑客组织进行网络攻击，但由于行为主体的隐蔽性，难以确定该国是否有胁迫他国改变其政治、经济或社会结构的主观意图，使得干涉行为的胁迫性认定变得更加复杂。

三、不干涉原则在网络空间国际治理中适用的现实困境

（一）证明困境

当前，国际法对网络干涉行为的证明标准不够明确，现有的国际法律框架对网络干涉行为的证明标准缺乏统一的规范，许多国家在如何证明网络干涉的“非法性”上仍存在较大的分歧，如何界定一国对另一国的网络行为构成干涉，并确定其胁迫性或胁迫性，尚未形成定论，国际法没有专门的法律规则来规制网络干涉行为。

由于网络空间的匿名性、虚拟性及其跨国特性，追踪网络干涉行为的具体证据变得异常困难。很多网络干涉行为，尤其是低烈度的干涉行为，如信息操控、数据窃取等，通常隐蔽性极强，且行为主体常常通过技术手段隐藏其身份。这使得受害国在追究责任时，缺乏直接的证据链，难以确认具体的行为人和行为动机，从而影响了国际法上责任的追究。

不干涉原则在网络空间中适用时的证明困境具体体现在以下两方面：第一，非物质性损害的证明困难。网络干涉行为的影响不仅仅局限于物理破坏，许多干涉行为还可能带来非物质性损害，如对政治进程、社会稳定或公众信任的影响。这类非物质性损害的证明难度更大，传统的证据标准难以覆盖这种类型的损害。第二，非国家行为体的网络行动难以找到有效证据而被归因于国家。尽管这些行

为体可能在某国政府的支持下实施网络干涉，但国际法对这类非国家行为体的责任归属尚未明确，即使证据表明这些行为对他国造成了干涉，也难以将其直接归因于某一国家或非国家行为体，从而造成了归责的法律障碍。非国家行为体没有直接执行国家指令，许多情况下国家可能间接支持、资助或容忍这些行为体的活动。虽然国家未明确指示黑客组织进行网络攻击，但却为其提供资金、技术支持或庇护，甚至在某些情况下，国家可能通过默许的方式，允许这些非国家行为体进行网络干涉，这种间接支持的情况使得归因性的证明变得更加复杂。

（二）执行机制缺失

国际法等软法无法对国家行为起到有效约束作用，尽管《塔林手册 2.0》等文件尝试将传统战争法适用于网络空间，但此类文件仅代表部分国家立场，不具备强制拘束力。各国对“不干涉”的解释仍基于自身利益，导致规则执行弹性过大。

此外，国际法院、联合国安理会等机构缺乏处理网络争端的技术能力和政治共识。安理会常任理事国的否决权使得针对大国的干涉行为难以追责，国际法院、联合国安理会等机构缺乏专业的网络技术团队与实时数据监控能力，难以对网络攻击进行快速溯源和损害评估。例如，在 2022 年俄乌冲突中，双方互相指控对方实施网络攻击，但安理会因技术证据不足及政治分歧未能通过有效决议。部分区域组织尝试构建网络治理机制，但成效有限：欧盟虽通过立法强化成员国间网络安全合作，但其域外执行依赖双边协议，难以约束非欧盟国家；东盟《网络安全合作框架》强调信息共享与能力建设，但缺乏统一制裁机制，成员国技术能力差异导致合作流于形式。

当前，网络空间的治理主要依赖各国在联合国等国际组织框架内的合作，然而，缺乏具有强制性和执行力的国际监督机制，使得不干涉原则的执行面临障碍。国际法对网络空间的规定大多是指导性文件，并未制定具体、细化的操作标准。

网络空间国际治理需要各国之间的合作与协调，但在现实中，国际合作往往受到多种因素的制约。不同国家在网络空间治理上的利益诉求和立场分歧难以弥合，导致在制定和执行国际规则方面进展缓慢。一些国家凭借其在信息技术领域的优势，试图在网络空间中建立霸权，通过网络攻击、网络间谍活动等手段干涉他国的内政和经济，严重破坏了网络空间的国际秩序。除此之外，缺乏有效的国际合作机制和执行机构，也使得即使达成了某些国际共识和规则，也难以得到有效执行。

四、 不干涉原则在网络空间国际治理中的适用体系构建

（一）促进各国在不干涉原则的保护范围达成共识

首先，应加强联合国框架内的国际合作。联合国作为国际合作的重要平台，应当发挥其在推动不干涉原则保护范围达成共识中的核心作用。通过在联合国信息安全政府专家组和开放式工作组等多边论坛中，集中讨论网络空间中的不干涉

原则，促使各国在法律框架和实践问题上达成一致。联合国的多边机制提供了一个包容性的对话平台，有助于解决各国在不干涉原则适用上的分歧，尤其是在低烈度网络行动和非国家行为体干涉等新型问题上达成共识。

其次，推动各国对非国家实体行为的归因性标准达成一致意见。目前在对于网络空间干涉行为归因的问题中，学界存在四种标准，其中，“有效控制”标准和“全面控制”标准是国际法中传统的归因标准，而“实际控制”标准和“转嫁责任”标准是国际法学者提出的新标准。从《塔林手册 2.0》的内容来看，现有国际法仍以传统标准为主，其对“有效控制”标准进行了适度的调整，保留了其核心内涵，即“执行和方向”“开始和中止”“内在部分”，通过识别关键点来判断网络非法干涉行为的归责主体，同时，以“全面控制”标准作为补充。本文认为各国应进行充分协商，参考传统归因标准，结合网络空间的特性，细化国家与非国家行为体之间关系的认定标准。

最后，促进各国对网络空间“内政”概念及范围达成一致，应当以法律明确性为基础、技术透明性为支撑、多边包容性为纽带，通过联合国框架凝聚核心规则，利用区域性试点积累经验，借助技术手段减少模糊性，最终构建“主权自主+全球共治”的弹性框架。短期内可通过软法引导实践，长期则应当推动强行法落地，并建立动态调整机制以应对技术迭代。

明确“胁迫性”标准并倡导审慎原则的适用

（1）明确“胁迫性”标准的定义

“胁迫性”是指一种通过威胁或施加压力，迫使他国改变其政策或决策的行为。在网络空间中，胁迫性网络干涉行为通常通过虚假信息传播、网络攻击或舆论操控等方式进行，往往未触及武力使用的标准，却对目标国的主权、政治稳定及社会秩序造成严重影响。因此，明确“胁迫性”的标准，对于判定某一网络行为是否属于非法干涉至关重要。

目前，关于“胁迫性”行为的界定存在一定的模糊性，尤其是在“胁迫性”与“非胁迫性”网络行动的划分上。国际法未对“胁迫”行为作出明确的定义，导致在实际操作中难以对某些网络行为进行准确的界定。为了解决这一问题，国际社会需要制定国际法规则以明确何种干涉行为构成“胁迫性”干涉。本文认为，可参考传统国际法中对于“胁迫”的定义，如是否有意图改变他国政策，是否通过施加压力或威胁来实现该目标。同时，还应根据网络空间的特殊性，考虑行为的方式、手段、规模和效果等因素，形成一个清晰的标准。

（2）推动“胁迫性”和“非胁迫性”网络行动的区分

在网络空间中，许多低烈度网络行动虽然未达到武力攻击的标准，但其影响力和干预效果仍然可能对目标国的政治、经济、社会稳定产生深远影响。因此，如何区分“胁迫性”与“非胁迫性”网络行动，也是明确“胁迫性”标准的一个重要问

题。为了准确划分两者，应当在联合国的框架下，组织各国就这一问题进行深入探讨，并根据各国的立场和实践形成共识，并将网络干涉行为的具体影响纳入考量，而不仅仅依赖于行为是否达到某个既定的技术标准。通过明确影响的规模 and 效果，帮助区分胁迫性和非胁迫性网络行动。

（3）倡导审慎原则的适用

《塔林手册 2.0》规则 5 明确了审慎原则的内容，“一国应采取审慎态度，不得允许其领土，或处于其政府控制下的领土或网络基础设施，被用于实施影响他国行使权利，和对他国产生严重不利后果的网络行动。”这标志着国际法学界开始普遍承认审慎原则在网络空间领域的适用。审慎原则一般包括两层义务：一般警戒义务，以及当国家通过警戒，知道或应该知道有害活动时，一般的警戒义务转变成的具体行动义务。⁵审慎原则要求各国在行使权利和义务时应谨慎、合理地采取措施，以免对他国的主权、政治独立和领土完整造成不必要的损害。在网络空间中，审慎原则尤为重要，因为许多网络干涉行为具有隐蔽性和间接性，往往难以立即显现其后果。在评估网络行动是否构成“胁迫”时，审慎原则应被广泛适用，以避免过度扩张不干涉原则的适用范围，造成不必要的法律冲突。

审慎原则的适用可以从以下几个方面入手。第一，在判断网络干涉行为是否具备“胁迫性”时，不仅要考虑行为的直接影响，还要综合评估其间接后果。各国在判断网络行为时，应当避免仅从技术角度或直接结果出发，而应全面考虑行为对国际社会的长远影响。第二，各国在应对低程度网络干涉行为时，应遵循适度原则，即在回应网络干涉时，应采取合理且不超出必要范围的反措施，避免造成过度损害，根据实际情况选择最适合的对策，以免冲突的升级。第三，审慎原则的适用还应通过加强国际合作与对话来实现，各国应共同讨论如何合理定义“胁迫”行为，并通过建立共同的应对机制来减少网络干涉带来的不确定性和风险。

五、加强对网络空间干涉行为进行现实规制的国际合作

加强国际合作是确保有效规制网络空间干涉行为的关键。通过推动全球法律框架的建设、加强信息共享与透明度、推动多边机制下的对话与协商、促进技术和能力建设、建立追责和制裁机制，国际社会可以有效应对网络空间中的跨国干涉行为，确保全球网络安全与秩序的稳定。这不仅有助于防范和应对现有的网络干涉威胁，也为未来网络空间的治理提供坚实的基础。

第一，应当加强跨国信息共享与透明度。网络干涉行为的隐蔽性和高技术性使得难以追踪到干涉行为源头，这更增加了归责和法律追诉的困难。因此，各国应加强信息共享和透明度，在网络监控、数据共享和证据收集方面开展更广泛的跨国合作。国家间应建立共同的技术平台，建立信息共享机制，有助于各国深入

⁵ See 孙一榕: The Future of Due Diligence in Cyberspace, New York University Journal of International Law and Politics Journal, Vol.54:2, p753(2022).

识别网络干涉行为的来源，以采取有效应对措施，加强对网络干涉行为的预防与规制。此外，针对不同层级的网络攻击，应制定适应性强、透明的评估标准，以确保各国在应对这些行为时能够协调一致。

第二，建立有效的追责和制裁机制。国际组织应当建立网络空间国际争端解决中心，负责调查、解决网络干涉行为的归责问题，技术专家与国际法学者参与其中，以确保裁决的科学性与合法性。国际社会应制定清晰的归责标准，明确网络干涉行为的定义，并设立跨国追责机制，通过相关国际机构，对涉嫌实施网络干涉的国家或行为体进行规制，对网络干涉行为进行有效制裁。此外，应各国也根据国际法以及国内法的规定，采取必要的反制措施，确保主权不受侵犯。

六、 结论

网络空间的隐蔽性使得传统国际法中的不干涉原则在网络空间的适用面临诸多挑战，法理上包括非国家行为体归因困难、对“内政”概念与范围界定不清、胁迫性认定标准不明等问题，现实中包括网络空间不干涉行为证明困难以及执行机制缺失。这些问题导致网络干涉行为的界定和责任追究变得复杂，低效的国际治理机制进一步加剧了困境。国际社会应当通过联合国框架凝聚共识，明确“胁迫性”标准与审慎原则，加强跨国信息共享与追责机制；推动技术赋能，如利用区块链存证等技术手段提升证据的可追溯性和透明度；倡导国际法的动态改革，以适应网络空间的快速发展。确定不干涉内政原则适用于网络干涉行为的具体规则，达成有效国际共识并最终形成国际条约的过程难免曲折。应当认识到，现有国际法为网络空间国际法治实现所提供的并非固定不变的模板，而更像是具有原则性和基础性的规则框架，各国应当在坚持国际法基本原则不变的前提下根据网络空间特点对具体规则进行变通和发展。

参考文献

- [1] 邵沙平. 国际法（第四版）[M],北京: 中国人民大学出版社,2020.
- [2] 北约网络合作卓越中心特邀国际专家组.网络行动国际法:塔林手册 2.0 版 [M].黄志雄等,译.北京:社会科学文献出版社,2018.
- [3] 刘晗.域名系统、网络主权与互联网治理历史反思及其当代启示 [J]. 中外法学, 2016, 28 (02).
- [4] 刘碧琦. 国际法在网络空间的适用问题研究[D]. 武汉大学, 2020.
- [5] 黄志雄.网络空间国际规则制定的新趋向——基于《塔林手册 2.0 版》的考察[J].厦门大学学报(哲学社会科学版), 2018(01).
- [6] 朱玲玲. 从《塔林手册 2.0 版》看网络攻击中国家责任归因的演绎与发展[J]. 当代法学, 2019, 33(1).
- [7] William Ossoff. Hacking the Domain Reserve: The Rule of Non-Intervention and Political Interference in Cyberspace[J]. Harvard International Law Journal,2021,62(1).
- [8] 孙一榕. The Future of Due Diligence in Cyberspace[J]. New York University Journal of International Law and Politics, 2022, 2(54).

智能船舶港口国管辖的内涵阐释、困境剖析及规制应对

崔明珠*

摘要：全球航运业正加速迈向智能化、无人化与数据化的范式转型，智能船舶技术深刻重塑海洋运输的底层逻辑与治理格局。港口国管辖作为国际海洋法体系中平衡船旗国监管缺陷、维护海洋权益的核心机制，其制度效能面临技术革命的系统性考验。在船旗国责任弱化与航运安全需求强化的双重背景下，如何重构智能时代的管辖权框架，既关乎国际海事公约体系的适应性演进，更涉及数字主权、技术伦理与国际公共利益的复杂博弈，成为全球海洋治理不可回避的战略议题。国际社会应以IMO为核心平台，推动公约条款的动态解释与专属规则创设，强化技术标准与法律义务的衔接耦合，构建智能船舶港口国管辖的国际监管体系。中国作为智能航运变革的引领者，需统筹国内立法创新与国际规则话语权建设，通过输出港口国监督检查制度规范、全过程监管进离港船舶和明确突发事件管辖执行依据等路径，实现从国际规则引领到本土治理效能提升的闭环演进，为智能时代海洋法治贡献兼具创新性与安全性的中国方案。

关键词：智能船舶；港口国管辖；港口准入；港口国监督；船舶引航

引言

全球航运业加速向智能化、绿色化转型，智能船舶技术的突破性发展正深刻重构国际海运生态。作为人工智能、物联网与大数据技术的集成载体，智能船舶通过无人化操控、自主决策系统及泛在化数据交互，实现了从传统“人—船”二元结构向“算法—机器—网络”多维范式的跃迁。国际海事组织（IMO）将其定义为“具备不同自主等级的海上自主水面船舶（MASS）”，其第四级全自主船舶已逐步从概念验证迈向商业化试航，标志着航海活动正式步入“后人类中心”时代。智能船舶与港口国管辖的制度耦合，已逐渐成为全球海洋秩序重构进程中亟待破解的核心议题。

探究智能船舶港口国管辖法律问题，不仅为智能航运产业提供理论支撑，更具重要的实践价值。科技的发展不仅塑造立法方向，更构成法律制定的基础要素。

* 崔明珠，哈尔滨工程大学人文社会科学学院，2022级法学硕士研究生，本论文是国家社科基金一般项目“基于国际海事组织立法进程的中国智能船舶法律规制研究”（23BFX198），黑龙江省哲学社会科学研究规划项目“《极地规则》生效背景下俄罗斯、加拿大北极通航法律政策演变及中国应对研究”（22GJB130），中央高校安全研究专项课题“北极航道航行安全法律问题研究”，黑龙江省法学会2024年重点课题《北极命运共同体视域下中国参与北极航运安全治理的路径研究》的阶段性成果。

智能航运技术的快速演进与现行法律体系的脱节,亟需系统性重构智能船舶管辖法律框架。现阶段国内外学者关于智能船舶港口国管辖的研究呈现明显的不均衡态势。在法律责任研究层面,国内外学者主要聚焦于碰撞事故后的民事与刑事责任分配,以及远程操作人员的法律定位问题;而在港口国管辖领域,现有成果多局限于公约适用性分析和传统管辖规则梳理,鲜少探讨智能船舶对国家管辖权的实质性影响,仅浅尝辄止地论及管辖权基础;受制于技术形态的不确定性,多数研究仅基于“无船员”这一基本特征展开,缺乏对技术变革引发的系统性影响分析;传统船舶管辖研究虽较丰富,但受各国立法差异影响,普遍欠缺理论一般性。为此,本文通过阐释港口国管辖的基本内涵,分析智能船舶港口国管辖的基本内容,剖析智能船舶港口国管辖的法律困境,提出智能船舶港口国管辖体系构建的路径及中国方案,为智能航运时代的法律治理提供新思路。

1. 智能船舶港口国管辖的内涵阐释

1.1 港口国管辖的概念引入

港口国管辖作为国际海洋法体系中的重要制度,是主权国家在全球化航运治理中行使管辖权的重要体现。管辖权的本质在于国家对特定领域或对象进行规范管理的能力,而船舶管辖作为国家主权的延伸,直接关系到海上航行安全、海洋环境保护以及国际航运秩序的维护。

在国际航运的广阔舞台上,港口及其附属的无数设施与功能构成了国际海洋运输系统不可或缺的基石。港口不仅是国际航运的便捷枢纽,为船舶的停泊与货物的转运提供便利,同时也为港口国提供了对外国船舶进行监管的契机与动因。自《联合国海洋法公约》(UNCLOS)明确界定港口国管辖权以来,这一概念便在国际海洋法领域崭露头角。¹港口国管辖(Port State Jurisdiction, PSJ)是国际海洋法体系中一项兼具主权属性与全球治理功能的特别规定,其核心在于港口国对自愿进入其港口的外籍船舶,针对违反国际公约或国内法的行为所行使的立法、执法及司法权力,即港口国在其港口或内水范围内,对外国船舶及其相关活动行使的管辖权。这种管辖权源于国际法中的“港口国主权”原则,即港口国在其领土范围内享有主权,有权制定和实施法律,以维护港口的安全、秩序和环境保护等公共利益。早在19世纪,一些国家就已经开始对进入其港口的外国船舶实施一定的管辖措施,但早期的管辖主要集中在船舶的进出港管理、关税征收等方面,目的是维护国家的主权和经济利益,随着国际贸易的发展和海洋活动的增加,港口国管辖的范围逐渐扩大到船舶安全和卫生等领域。在处理不合规船舶时,港口国权力可分为港口国控制(PSC)和港口国管辖权(PSJ),PSJ涉及港口国对违反国际规则的船舶提起诉讼并处以罚款的权力,而PSC仅限于采取行政管控措

¹ 宋晓:《域外管辖的体系构造:立法管辖与司法管辖之界分》,《法学研究》2021年第3期,第171-191页。

施，如扣留船舶直至整改完成或令其前往最近船厂维修。值得注意的是，PSJ 不仅限于对港口或沿海国海域内违法行为的追诉，根据 UNCLOS 第 218 条，PSJ 还涵盖对领海外违法行为的追诉。²

港口国管辖自规定以来经历了从补充性机制向强制性义务的深刻转型。早期国际航运治理高度依赖船旗国管辖，港口国仅在船舶自愿入港时实施有限监督，其功能多限于技术标准核查。然而，随着船舶开放登记制度的滥用导致“方便旗”现象泛滥，以及海洋污染、非法捕捞等跨国问题的加剧，国际社会开始重构治理模式。20 世纪 80 年代，区域合作机制（如巴黎备忘录、东京备忘录）的兴起，标志着港口国管辖从被动补充转向主动干预，通过联合检查、信息共享等方式强化对低标准船舶的制约。进入 21 世纪，《港口国措施协定》（PSMA）³等文件的生效，将港口国责任从选择性权利升级为必要义务，要求缔约国对船舶安全、污染防治、劳工权益等实施全链条监管，此时港口国不再仅是主权边界的守卫者，更被视为国际公共利益的“执法代理人”，其职能范围从传统的技术监督拓展至渔业管理、生物安全、反恐安保等多元领域，形成了以港口为支点的跨国治理网络。

1.2 港口国管辖适用智能船舶存在差距

航运技术革新正深刻重塑行业生态，船员培训机制、船舶制造规范、航行规则体系及安全管理标准均面临巨大变革。在这一转型进程中，各国海事监管部门传统的管理范式已难以应对新兴需求，亟待构建面向未来的海事治理能力体系。⁴一方面，检查手段的技术性失灵。以港口国监督（PSC）为例，传统港口国监督依赖目视检查、设备测试与文书审查，但这些方法无法穿透智能船舶的“技术黑箱”。在航行自由与海上交通安全相冲突时，船旗国对于沿海国交通安全权益的关注存在天然的局限。⁵在智能船舶配员减少的趋势下，PSC 检查程序面临着智能船舶无法配合港口国监督官员检查登船检查的挑战，智能船舶是否至少具有与传统船舶相同的安全水平难以判断。如果智能船舶远程操作中心位于船旗国，如何实现对其检查成为难题。虽然《MASS 规则》谈判中提及“无论何种操作模式或自主程度均应有一名人类船长负责 MASS”。⁶但是智能船舶船长是否需要在船上以及其职责仍需要进一步的讨论。即使未来 IMO 制定了智能船舶最低安全配员标准及规则，船上船员是否具备能满足履行 PSC 官员检查的相关能力也存疑。同时，智能船舶的监管高度依赖航行数据（如传感器日志、决策轨迹），当船舶在港口国水域发生系统故障时，需依赖船旗国技术支持，倘若船旗国主张权利而

² Bang H S. Port state jurisdiction and article 218 of the UN convention on the law of Sea. J. MaR. l. & com, Vol.40(2009), p. 291.

³ 参见《港口国措施协定》（PSMA），at <https://www.fao.org/port-state-measures/zh/>.

⁴ 高宇，黄玉龙：《海事管理数字化转型工作研究》，《水上安全》2022 年第 4 期，第 52 页。

⁵ 邢厚群：《论国际海洋法治视域下我国海上交通安全立法》，《社会科学》2021 年第 6 期，第 107 页。

⁶ 国际海事组织，海上安全委员会，《海上安全委员会第一百零七届会议报告》，MSC 107/20，2023 年 6 月 26 日，第 41 页。

拒绝提供支持，港口国既无技术能力跨境提取加密数据，也缺乏国际法授权突破主权壁垒，如此整个调查活动将陷入被动。此外，“方便旗”智能船舶的可能性存在也是限制 PSC 检查执行的潜在挑战。⁷

另一方面，区域性协调安排不足弱化监管效能。各国海事主管机关与 IMO 合作，就 PSC 措施的实施达成 9 个区域性港口国监督谅解备忘录及美国海岸警卫队（USCG）。覆盖了不同的地理区域，但由于经济和文化的差异以及海洋环境的复杂性导致不同的区域性协议规则之间存在冲突。例如，《巴黎港口国监督谅解备忘录》和《亚太地区港口国监督谅解备忘录》不会检查所有进入该地区的船舶，但会抽查一部分船舶。在计算船舶检查的数量和比例时，两份谅解备忘录的计算方法不同。不同区域性协议不禁止成员国身份重叠，也就是说一个港口国可以同时是多个区域性协议成员，规则的冲突可能使成员国管辖权行使产生混乱。⁸智能船舶的算法协同失误可能引发“责任真空”——船旗国、港口国、技术供应商均以“超出可控范围”推诿责任。这种系统性失效不仅消解管辖权的威慑力，更激励航运业将风险转嫁至法律灰色地带，危及全球海运安全秩序。

1.3 智能船舶港口国管辖的规范构造

随着智能船舶技术的发展，港口国管辖的范围和内容也在不断扩展和深化，以适应新技术带来的挑战和需求。对于智能船舶而言，港口国管辖不仅涉及传统的船舶安全、环境保护和船员权益保护等领域，还扩展到了网络安全、数据隐私、自主系统监管等新兴领域，智能船舶的港口国管辖包括港口国对智能船舶实施的港口国监督、进离港管辖和突发事故管辖。

1. 智能船舶的港口国监督

港口国监督（Port State Control, PSC）指港口国在其管辖水域对外籍船舶开展的合规性核查，旨在确认船舶技术状况与设备配置是否达到国际标准，同时审查船员资质及操作程序是否符合法律规范以弥补船旗国和沿海国管辖的不足。⁹涉及船舶安全检查和国内法律法规主要有《中华人民共和国海上交通安全法》（以下简称《海上交通安全法》）¹⁰和《中华人民共和国船舶安全监督规则》（以下简称《船舶安全监督规则》）¹¹。在传统检查模式下，港口国当局通常依赖与船员的

⁷ 袁雪，于博：《国际海事组织海上自主水面船舶的法律规制进程及中国因应》，《国际法研究》2023 年第 1 期，第 38-57 页。

⁸ Linlin Chen, Research on the Modern Maritime Supervision Mechanism from the Perspective of Global Maritime Governance. China Legal Science, Vol. 11, No. 3 (2023), p. 118.

⁹ See Port State Control, IMO, International Maritime Organization, <https://www.imo.org/en/OurWork/MSAS/Pages/PortStateControl.aspx>.

¹⁰ 《中华人民共和国海上交通安全法》第 88 条规定：“海事管理机构对在中华人民共和国管辖海域内从事航行、停泊、作业以及其他与海上交通安全相关的活动，依法实施监督检查。海事管理机构依照中华人民共和国法律、行政法规以及中华人民共和国缔结或者参加的国际条约对外国籍船舶实施港口国、沿岸国监督检查。”

¹¹ 《中华人民共和国船舶安全监督规则》第 2 条：“本规则适用于对中国籍船舶和水上设施以及航行、停泊、作业于我国管辖水域的外国籍船舶实施的安全监督工作。”《船舶安全监督规则》是目前我国海事管理机构开展船舶安全检查的主要依据。

直接互动以获取必要信息。然而，智能船舶船上可能无驻留船员可供问询，而船舶证书及操作记录均存储于远程控制中心，因此，智能船舶将促使港口国调整检查范围、评估标准及核查重点，尤其需要解决人员接触缺失和文件调取困难等新型监管难题。

智能船舶在结构设计层面与传统船舶存在本质差异。由于无需配置船员生活区域，其船体构造显著简化，这使得港口国监督的常规检查项目，如船员居住条件评估，失去适用基础。尽管智能船舶的航行操作仍依赖人工决策，但操作者可能位于港口国境外，并不实际随船抵港，这导致港口国当局在人员资质核查方面面临管辖权障碍，进而引发是否应当将远程操作中心人员纳入港口国监督范围，以及如何建立相应的跨境核查机制的问题。智能船舶依靠先进的传感器、通信技术和自动化控制系统运行，港口国需要检查这些系统是否运行正常，数据传输是否稳定准确。船员对智能系统的操作能力也是监督重点。虽然智能船舶自动化程度高，但船员依然是保障船舶安全的关键因素。港口国要评估船员是否经过专业培训，能够熟练操作和维护智能系统，在系统出现故障时能否及时采取有效的应急措施。此外，智能船舶产生的大量数据的管理和使用也在监督范围内。这些数据涉及船舶航行信息、货物信息等，港口国要确保数据的存储、传输符合相关安全和隐私规定，防止数据泄露对船舶安全和港口国利益造成威胁。

PSCOs 对传统船舶检查的基本程序包含到港前对船舶外观进行初步检查，到港后登船检查相关证书和设施的深度检查，随后形成检查报告，对可能存在严重缺陷的船舶现场提出处理意见。智能船舶的出现使 PSC 船舶检查的方式发生了重大改变。在船舶安全检查过程中，PSCOs 必须获得船员配合。登船后，官员通常会直接与船长及值班船员进行对话，询问了解船舶实际状况，船员依法应当如实提供所需信息。由于智能船舶船上无人值守，若检查发现船舶存在缺陷，或远程操作人员有违国际规则或该国国内法之举，PSCOs 除限制船舶离港外，似乎难以直接对相关责任人采取强制措施，且受此限制的船舶亦可能给港口运营增添不必要的麻烦。对智能船舶实施 PSC 安全检查时，由于缺乏船员配合，PSCOs 无法通过传统问询方式获取信息，仅能依靠人机交互界面从船舶智能终端提取有限数据。当发现船舶存在技术缺陷或远程操作违反国际公约及国内法规时，PSCOs 难以对实际责任主体采取直接强制措施，这严重限制港口国当局的执法效果，对港口正常运作造成不当干扰。¹²

2. 智能船舶的进离港管辖

船舶进离港管辖是对港口当局对进出港船舶的动态监管过程，也是船旗国、港口国管辖权发生转移过程。船舶进离港管辖是港口国基于主权和安全需求，对船舶进出港口全流程实施的监管体系，涵盖港口准入、船舶引航、口岸检查等核

¹² 潘虹：《智能船舶监管法律问题研究》，硕士学位论文，2020 年，大连海事大学，第 34 页。

心环节,各环节以人工操作为主,强调船舶适航性、船员资质、文书合规性等静态要素,并通过引航员现场操控、多部门分头登临检查等方式实现管理目标。涉及船舶进离港管辖的国内法律法规有《海上交通安全法》、《中华人民共和国港口法》(以下简称港口法)¹³、《船舶进出港报告管理办法》¹⁴、《船舶引航管理规定》¹⁵、《国际航行船舶进出中华人民共和国口岸检查办法》¹⁶等。智能船舶技术的成熟,其自主航行能力、数据交互特性及无人化特征对传统管辖框架提出了系统性变革需求,推动船舶进出港管辖向智能化、协同化、动态化转型。

UNCLOS 第 25 条、第 211 条及第 255 条赋予港口国对船舶进出港的自由裁量权,港口国可通过设置船舶安全与防污染标准行使准入控制权。¹⁷传统船舶的港口准入审查聚焦于船舶登记证书、船员适任证明、载货清单等基础文件,其核心是验证船舶的“合法性”与“适航性”,智能船舶的港口准入则需新增技术认证维度,包括但不限于自主航行系统的可靠性验证、远程控制与网络通信能力的认证、式智能船舶的分级认证等。电子证书查验是智能船舶准入的核心环节。IMO 修订的《国际便利海上运输公约》(FAL 公约)要求通过“单一窗口”系统实现电子证书的标准化传输与验证。智能船舶需实时提交电子化的适航证书、网络安全认证及远程操作合规文件。¹⁸此外,考虑到智能船舶的系统属性,对智能船舶的港口准入需要实现动态适配性,如当船舶通过系统升级优化航行算法时,需触发技术参数的重新核验,避免“一次认证、终身有效”的传统模式导致技术迭代与监管脱节。

船舶引航,作为兼具主权属性与安全价值的特殊海事活动,由具备资质的引航员借助拖带等技术手段,引导船舶自开放水域安全进入指定港口或航道。¹⁹船

¹³ 参见《中华人民共和国港口法》2019 年 1 月 7 日, http://www.npc.gov.cn/zgrdw/npc/xinwen/2019-01/07/content_2070259.htm, 2025 年 3 月 2 日。

¹⁴ 参见《交通运输部海事局关于印发<船舶进出港报告管理办法>的通知》, 2021 年 8 月 11 日, <https://www.msa.gov.cn/html/xxgk/tzgg/cbjd/20210811/A03639D3-34FA-40C9-8A3C-50F07DAA7B9F.html>, 2025 年 2 月 3 日。

¹⁵ 参见《交通运输部关于修改<船舶引航管理规定>的决定》, 2021 年 9 月 13 日, https://www.gov.cn/zhengce/zhengceku/2021-09/13/content_5637004.htm, 2025 年 2 月 3 日。

¹⁶ 参见《国际航行船舶进出中华人民共和国口岸检查办法》, 2019 年 3 月 2 日, https://www.gov.cn/gongbao/content/2019/content_5468i919.htm。

¹⁷ UNCLOS 第 25 条:“在船舶驶往内水或停靠内水外的港口设备的情形下,沿海国也有权采取必要的步骤,以防止对准许这种船舶驶往内水或停靠港口的条件的任何破坏。”UNCLOS 第 211 条:“各国如制订关于防止、减少和控制海洋环境污染的特别规定作为外国船只进入其港口或内水或在其岸外设施停靠的条件,应将这种规定妥为公布,并通知主管国际组织。如两个或两个以上的沿海国制订相同的规定以求协调政策,在通知时应说明哪些国家参加这种合作安排。每个国家应规定悬挂其旗帜或在其国内登记的船只的船长在参加这种合作安排的国家的领海内航行时,经该国要求应向其提送通知是否正驶往参加这种合作安排的同一区域的国家,如系驶往这种国家,应说明是否遵守该国关于进入港口的规定。本条不妨害船只继续行使其无害通过权,也不妨碍第二十五条第 2 款的适用。”UNCLOS 第 255 条:“各国应尽力制定合理的规则、规章和程序,促进和使得在其领海以外按照本公约进行的海洋科学研究,并于适当时在其法律和规章规定的限制下,便利遵守本部分有关规定的海洋科学研究船进入其港口,并促进对这些船只的协助。”

¹⁸ 韩佳霖,李香遐,章文俊:《便利运输框架下海上自主水面船舶立法》,《中国海事》2023 年第 7 期,第 57-60+65 页。

¹⁹ 陈长涛:《由船舶引航展望智能船舶发展》,《中国水运》2023 年第 10 期,第 5 页。

舶引航既体现国家对其管辖水域的主权行使,又直接关乎港口设施与船舶本身的航行安全。²⁰为保障港口国主权安全,《海上交通安全法》第31条第二款、第三款分别规定引航员须实际登船执行引航作业,引航期间不解除船长指挥和管理船舶责任,外籍船舶适用强制引航制度。SOLAS公约第V章规定,船舶须配备符合标准的引航员登离装置,并由甲板值班驾驶员确保设备就位,全程负责引航员的安全接送作业。由此可见,现行港口引航相关法律主要适用于引航员实际登船实施引航作业的情形。尽管智能船舶上没有船员,但对智能船舶实施引航仍有必要。²¹考虑到智能船舶的技术特性,智能船舶引航制度的构建需首先界定引航员在各控制层级的功能定位,进而探索与“人机等效替代”场景实现最优适配的新型引航操作范式。²²其次,可考虑对智能船舶实施强制引航豁免及远程引航作为强制引航豁免的替代方案的可行性。传统引航活动中引航员登船作业具有主权象征意义。若实施豁免需向远程操作端开放港口核心数据,此举虽便利船舶操控,但却可能导致敏感信息外泄威胁港口安全。而远程引航的有效实施需满足船舶与港口远程操控系统的技术兼容以及港口基础设施的智能化改造。当引航员远程接管智能船舶控制权时,事故责任划分标准亟待重构,这要求对现有责任认定机制进行全面评估与更新。²³

传统船舶口岸检查以海关、卫生检疫与移民检查为三大支柱,均高度依赖船员在场配合。例如,海关官员需登船核对纸质舱单与货物实物,卫生检疫依赖现场采样与船员健康申报,移民部门则通过查验船员护照确认人员合法性。船舶口岸检查是一国政府行使管辖权之关键途径。然而,智能船舶的无人化属性迫使检查机制向电子化、自动化转型。为此,智能船舶的口岸检查实则要加强信息共享与数据连通性。智能船舶需与港口管理系统、远程操作中心及他船建立结构化数据交互机制,涵盖航行计划同步、环境感知数据实时传输及协同避碰信息交换。

24

3. 智能船舶的突发事件管辖

船舶突发事件管辖制度以船旗国管辖为主导,结合属地管辖与国际习惯法原则构成。UNCLOS明确规定,船旗国对本国籍船舶享有行政、技术及社会事务的排他性管辖权²⁵,涵盖对船舶建造标准监管、适航性认证,以及环境污染事故的

²⁰ 《船舶引航管理规定》第3条:“船舶引航是指引领船舶航行、靠泊、离泊、移泊的活动。”

²¹ 郑世江:《远程控制无人船法律适用问题研究》,硕士学位论文,大连海事大学,2019年,第24页。

²² 周广华,胡建军,潘登等:《自主船舶与引航工作最佳耦合模式》,《世界海运》2023年第6期,第18页。

²³ 《中华人民共和国海商法》第39条:“船长管理船舶和驾驶船舶的责任,不因引航员引领船舶而解除。”

²⁴ 韩佳霖,李香遐,章文俊:《便利运输框架下海上自主水面船舶立法》,《中国海事》2023年第7期,第57-60+65页。

²⁵ UNCLOS第94条第1款:“每个国家应对悬挂该国旗帜的船舶有效地行使行政、技术及社会事项上的管辖和控制。”

责任追究。²⁶而港口国在传统框架下的管辖权限主要体现在对到港船舶的有限监管²⁷，例如依据 SOLAS 公约检查船舶安全证书，或依据 MARPOL 公约核查污染物处理记录。对于船舶碰撞事故，港口国通常仅能在事故发生于其领海或内水时行使属地管辖权，且需遵循船旗国优先原则。而在污染事故中，港口国可通过滞留船舶、强制整改等手段间接干预，但其监管重心仍限于传统船舶的机械缺陷或船员操作过失，例如检查油污应急计划是否符合国际标准。²⁸

智能船舶的兴起促使港口国对事故监管的拓展与升级。首先在技术合规层面，港口国海事当局需建立智能系统分级认证制度，根据船舶自主等级实施差异化监管。在数据监管方面，实时监管获取智能船舶的航行数据，例如，通过船舶自动识别系统（AIS）与电子航海日志（ELOG）的联动，追踪智能船舶在港口水域的航迹偏差，并对异常数据，如连续 10 分钟未执行避碰指令就启动预警。在应急响应机制上，港口国需针对智能船舶特性制定专项预案。一方面，建立远程控制系统失效时的强制干预程序，例如通过海事专用频段向失控船舶注入停航指令；另一方面，完善事故调查中的数据取证规则，明确船舶“黑匣子”（航行数据记录仪）的解码权限与证据效力。

二、智能船舶港口国管辖的国际规制动态

2.1 IMO 智能船舶的监管规划

随着信息和通讯技术的发展，智能化、数字化和绿色化技术成为海洋领域发展的中坚力量。MASS 作为新兴技术的集成体现，正在 IMO 海事安全领域进行如火如荼的讨论。由于其目标是完全无人运营船舶，因此带来了整个产业生态的变革，需要对现有 IMO 立法做系统性的调整。²⁹IMO 规制 MASS 的最终目标是制定一份基于目标型标准的强制性规则，为达到此目标，其立法规划大致可分为“三步走”：

第一步，在 IMO 所有文书框架下开展 MASS 的监管范围界定工作（RSE）并得出不同文书适用智能船舶存在的共性问题。在这一环节中，RSE 工作分头展开，主要以海上安全委员会（MSC）和法律委员会（LEG）为主导，FAL 最后加入智能船舶 RSE 工作，三个委员会在各自职权范围内开展 RSE 工作并逐步形成各委员会框架下的智能船舶工作路线图。

²⁶ 白佳玉：《〈联合国海洋法公约〉缔结背后的国家利益考察与中国实践》，《中国海商法研究》2022 年第 2 期，第 3-13 页。

²⁷ UNCLOS 第 218 条第 3 款：“当船只自愿位于一国港口或岸外设施时，该国应在实际可行范围内满足任何国家因认为第 1 款所指的违章排放行为已在其内水、领海或专属经济区内发生，对其内水、领海或专属经济区已造成损害或有损害的威胁而提出的进行调查的请求，并且应在实际可行范围内，满足船旗国对这一违反行为所提出的进行调查的请求，不论违反行为在何处发生。”

²⁸ 参见《船舶油污应急计划填报需要注意的事项》，2021 年 1 月 13 日，<https://www.xindemarinenews.com/data/haishifagui/2021/0113/26495.html>，2025 年 2 月 3 日。

²⁹ 冯书桓：《IMO 海上水面自主船舶立法进展与趋势》，《中国船检》2022 年第 9 期，第 56 页。

第二步，依据 RSE 工作结果，形成非强制性规则，现已形成初步的规则草案。³⁰为顺利推进 MASS 文书制定工作，MSC 第 105 次会议、LEG 第 109 次会议和 FAL 第 46 次会议批准成立 MASS 联合工作组（MASS-JWG），旨在解决三个委员会各自进行的 RSE 工作中确定的共同高度优先问题，但同时各委员会框架下 MASS 的规制工作仍在进行，结合委员会的职权范围分领域解决 MASS 适用问题。需要注意的是，联合工作组的设立不是为了监督或指导 MASS 的立法工作，而其是一种向不同委员会提供咨询意见的协作机制，并作为一个推动三个委员会就共同问题开展工作重要辅助工具。

第三步，完成从“非强制性 MASS 规则”向“强制性 MASS 规则”的过渡。2024 年 5 月，MSC 第 108 次会议上，委员会同意修改 MASS 规则的路线图，预计将于 2025 年 5 月最终确定并采用“非强制性 MASS 规则”；2026 年上半年制定经验构建阶段的框架（EPB）；2028 年开始在“非强制性 MASS 规则”的基础上制定“强制性 MASS 规则”，并考虑对 SOLAS 公约进行修订，以供规则的通过；2030 年 7 月 1 日前，通过“强制性 MASS 规则”，并于 2032 年 1 月 1 日生效。³¹

2.2 ISO对智能船舶管辖的标准赋能

国际标准化组织（ISO）扮演着连接 IMO 与船舶产业间的关键纽带角色，通过持续更新船舶设计建造国际标准，引领全球航运与造船行业发展。ISO 标准体系主要覆盖整船设计规范、船用设备标准、航运运营规程、内河船舶技术规范及大型休闲船舶制造标准五大领域。ISO/TC8 船舶与海洋技术委员会创立于 1947 年，系 ISO 最早成立的技术委员会之一，2016 年 9 月，ISO/TC 8 成立了智能航运工作组（ISO/TC8/WG10），工作组基于大数据的船舶运营分析、物联网在造船业的应用，以及传感技术在海洋工程中的创新实践，对接 IMO 立法需求整合全球资源进行智能化发展顶层规划。当前，工作组所推进的一项关键任务在于编制智能航运国际化路线图 2.0 版本，旨在为相关国家及国际组织在智能船舶国际化顶层顶层设计方面给予参照，标准化路线图整体上遵循明确重点技术领域、进行差距剖析、预判未来标准化行动的逻辑脉络。

2018 年，工作组发布智能航运国际化路线图 1.0 版，从信息通信基础设施、智能船舶本体技术、航运概念创新、船舶运营管理系统，以及智能化支持服务体系领域系统规划船舶和航运智能化发展。2021 年，工作组结合智能航运技术演进与 IMO 立法动态，对标准化重点领域进行了战略调整，在概念层面，除保留通用技术与分类标准外，新增教育培训模块以契合 IMO 最新要求；将“支持服务”重新定义为“港口服务”，突出自动靠离泊与智能装卸系统，反映了船-港协同发展的技术整合趋势。随着智能船舶技术的深化应用，网络安全已成为关键议

³⁰ See MSC 108/4, Annex 1, page 1, <https://docs.imo.org/Shared/Download.aspx?did=147939,2024-6-4>.

³¹ See Maritime Safety Committee - 108th session (MSC 108), 15-24 May 2024, <https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-108th-session.aspx,2024-5-28>.

题，其影响范围涵盖船舶内部系统、船-岸通信及船-船交互环节。基于甚高频数据交换系统（VDES）等新技术的船-船通信，对数据安全提出了更高标准的防护要求。

当前 ISO/TC8 的标准体系分为通用指南与功能规范、技术解决方案以及具体设备与应用标准三类，共同支撑航运业智能化转型。工作组未来将重点推进信息基础设施数据规范、服务器技术要求、网络安全评估框架，以及船岸通信协议的标准化工作。³²2024 年 3 月，ISO 正式批复成立国际标准化组织船舶与海洋技术委员会智能航运分技术委员会（ISO/TC8/SC15），由中国承担该委员会的秘书处工作。³³

2.3 IACS对智能船舶认证检验的基准化作用

国际船级社协会（IACS）作为全球航运业技术标准制定的核心机构，其主要职能涵盖制定和维护船舶建造、检验、入级等方面的统一标准，通过对船舶技术规范의 严格把控，保障船舶在安全、环保、质量等多方面达到既定标准，进而促进全球航运业的稳定与可持续发展。IACS 和 IMO 一直保持紧密合作。如今，IACS 已经成为 IMO 在船舶设计、建造与设备规范领域的首席技术顾问机构，为其政策制定与标准实施提供了不可或缺的专业支撑。2016 年 12 月，在中国船级社（CCS）担任 IACS 轮值主席期间，促成 IACS 与 IMO 签署具有里程碑意义的合作备忘录。作为 IMO 首次与非政府组织（NGO）签订的此类协议，备忘录不仅肯定了 IACS 在 IMO 技术咨询体系中的特殊地位，更标志着 IACS 发展进程中的重要转折点。³⁴

MASS 的出现，使得航运业对技术要求的发展有了新的需求。IACS 凭借自身的技術贡献，在各类讨论平台上展示专业能力，成为监管机构和航运业信赖的合作伙伴。在 IMO 的 RSE 工作中，IACS 参与了 SOLAS 第 II-2 章的相关技术讨论。根据 IMO 制定的 MASS 试验指南，IACS 期望借此获得与船舶控制和工程系统相关的重要安全信息，以及风险评估结果，这些信息将由 IMO 成员在相关委员会和小组委员会的后续会议上共享。此外，IACS 与 ISO 之间的工作也有联动。IACS 就船级分类的主题与 ISO 建立紧密联系，包括但不限于参与 ISO 关于 MASS 术语的项目并提出意见。同时 IACS 还参与了欧洲海上安全局（EMSA）关于 MASS 的工作组。³⁵

IACS 指出船舶的网络安全事件可能会对生命、财产和环境产生直接和有害的影响，因此，IACS 加强了针对船载计算机系统的可靠性和功能有效性的关注

³² 冯书桓，岳宏：《智能船舶信息联通国际标准发展动向》，《中国船检》2022 年第 6 期，第 53 页。

³³ 李天煜：《标准引领造船与航运智能化转型发展》，《质量与标准化》2024 年第 4 期，第 1-4 页。

³⁴ 王思佳：《IACS：在海事立法与行业需求间寻求平衡》，《中国船检》2020 年第 1 期，第 28 页。

³⁵ See IACS high level Position on MASS, https://www.maritimeuk.org/documents/528/George_Reilly.pdf; IACS POSITION PAPER MASS (Maritime Autonomous Surface Ships) (Revision 3), <https://iacs.s3.af-south-1.amazonaws.com/wp-content/uploads/2023/09/29165530/IACS-MASS-Position-paper-Rev3.pdf>.

力度。³⁶船舶智能化在提升航运操作效率的同时，也因网络系统的集成而衍生出新型海事网络安全威胁。2018年11月，IACS正式发布12份关于船舶网络安全的建议案(Rec.153-164)³⁷，从系统设计原理、运行鲁棒性及网络弹性，对数字化船舶通信网络提出全面规范要求，涵盖系统运维管理、应急响应机制、网络安全防护、架构设计标准、设备清单管理以及数据安全保障等关键领域。³⁸2022年5月，IACS为提升船舶网络风险抵御能力，颁布两项具有强制效力统一要求URE26与URE27。³⁹该标准适用于2024年1月1日后签约建造的新船，为MASS的网络安全体系构建提供了关键制度保障。⁴⁰

2.4 IALA对船舶助航服务的智能化改造

国际航标组织（IALA）作为非营利性国际非政府组织，其前身为国际航标协会，IALA以促进全球海上助航服务体系标准化与协调化为核心使命。2024年8月22日，IALA于法国宣布《国际航标组织公约》即日生效，国际航标组织正式成立。⁴¹中国是国际航标组织首批创始成员国之一，目前已有包括中国在内的35个国家加入《公约》。较于此前的非政府间组织，转制后的IALA在组织架构、运作模式及全球影响力方面发生了较大转变，这要求相关方及时开展制度适应性研究并制定配套应对策略。⁴²

IALA在智能航运领域的核心贡献集中体现于电子航海（E-navigation）及配套数据系统建设。IALA提出的E-navigation概念已获IMO采纳并定义为，运用电子化技术实现船岸航海数据的协同采集、智能整合、高效交换、可视化呈现与深度分析，全面提升从泊位到泊位的导航服务质量，最终实现航行安全强化、海上安保完善与环境保护优化，为全球航运业的高效安全运营提供技术支撑。IALA还特别强调E-Navigation系统架构必须遵循用户中心设计原则，其核心构成涵盖船载/岸基系统及导航通信基础设施两大模块，这使得IALA标准体系与智能船舶发展高度协同，当前已发布的标准主要规范海事信息服务领域，包括船舶动态

³⁶ 参见《重大变化：IACS最新要求，快来看看你的船舶是否满足》2022年4月27日，<https://www.xindemarinenews.com/data/haishifagui/2022/0427/38110.html>，2025年3月2日。

³⁷ 参见《智能船舶标准体系建设指南（征求意见稿）》，第1页
https://wap.miit.gov.cn/cms_files/filemanager/oldfile/miit/n1146285/n1146352/n3054355/n7697926/n7697960/c7897865/part/7897871.pdf。

³⁸ 参见《中国船级社签发首张IACS网络安全符合证明》2019年6月26日，https://www.eworldship.com/html/2019/classification_society_0626/150540.html，2025年2月4日。

³⁹ URE 26 主要涵盖设备识别、保护、攻击检测、响应和恢复；URE 27 旨在确保第三方设备供应商对系统完整性进行保护和加固，提供了船上系统和设备的网络韧性要求，并提供了与用户和船上基于计算机的系统之间的接口有关的附加要求，以及新设备在船上实施前的产品设计和开发要求。

⁴⁰ 张淑静：《海上自主水面船规则及航海保障应对探讨》，《航海》2023年第1期，第59页。

⁴¹ 参见《〈国际航标组织公约〉生效 IALA 正式转变为政府间组织》，2024年8月23日，<https://www.msa.gov.cn/page/article.do?articleId=FDBA937C-C5AE-43B2-9D54-993B9CED921C>，2025年2月3日。

⁴² 曾繁屏，肖小春，杨弘融等：《国际航标组织转制和我国对策探讨》，《中国海事》2025年第1期，第22页。

数据交换、航行预警信息发布等关键环节。⁴³同时,IALA 已着手编制《IALA MASS 指南》为各国推进智能航运提供规范性框架,推动各成员国间通信协议标准化、船舶交通管理(VTS)系统适配性改造,以及岸基智能服务平台建设。⁴⁴

2024 年,IALA 发布《海上自主水面舰艇(MASS)的未来场景》⁴⁵,探讨了未来二十年内逐步采用 MASS 的前景以及该技术的潜在驱动因素和挑战⁴⁶,文件提出未来 MASS 的四种应用情景即许多配备自动化功能的有人驾驶船舶、少数无船员的自主船舶、更多配备自动化功能的有人驾驶船舶和部分无船员的自主船舶,为行业提供了分阶段规划依据。同时,IALA 积极研究区域性安排对跨国航行的示范作用,倡导统一安全标准与责任界定。针对船员短缺问题,IALA 指出自动化可缓解人力压力,但也强调无人船舶需应对岸基维护成本上升等经济性瓶颈,并呼吁行业关注船员技能转型(如远程操作培训)与港口基础设施升级(如冗余通信网络)。在环境目标上,IALA 将 MASS 与减排战略结合,支持优化航线规划与清洁燃料应用,助力全球可持续发展目标。

三、智能船舶港口国管辖的困境剖析

3.1 远程操作中心的PSC检查面临技术性失灵

港口国监督(PSC)是港口国在其主权范围内对外国籍商船实施的强制性检查制度,港口国通过系统性的港口检查程序,重点核查船舶技术状况及其设备是否符合国际公约的技术规范以及船员配备与操作实践是否达到法定标准要求。大多数海事公约都规定在船舶抵达外国港口时必须进行检查,以核实其是否遵守国际海事规则和条例。虽然船旗国对维持船舶标准负有主要责任,但 PSC 可作为检测不达标船舶的最后“安全网”。

远程操作中心(Remote Operations Centers, ROC)是智能船舶发展衍生的新生事物,也是监控和控制智能船舶的核心枢纽。ROC 对智能船舶航行、作业活动实施远程控制,相当于传统有人驾驶船舶的驾驶室,不同的是该“驾驶室”不设在船上,而是建在岸上。IMO 语境下远程操作中心是“一个远离 MASS 但可以操控船舶部分或全部功能的地点”,而 CCS 在《自主货物运输船舶指南(2018)》第 1.3.1 条⁴⁷将远程操作中心定义为“船外某一能够提供足够的支持和保障能力并监视和控制船舶的集中处所”。⁴⁸智能船舶通过网络与 ROC 连接,两者共同组成

⁴³ 冯书桓,岳宏:《智能船舶信息联通国际标准发展动向》,《中国船检》2022 年第 6 期,第 53 页。

⁴⁴ 张宝晨,耿雄飞,李亚斌等:《船舶智能航行技术研发进展》,《科技导报》2022 年第 14 期,第 53 页。

⁴⁵ See THE FUTURE OF MARITIME AUTONOMOUS SURFACE SHIPS (MASS), at <https://www.iala.int/content/uploads/2024/02/The-Future-of-Mass-2024-Portrait-simples-pages-for-website-corrected.pdf>

⁴⁶ See at <https://www.iala.int/technical/mass/>.

⁴⁷ 参见《自主货物运输船舶指南(2018)》,

<https://www.ccs.org.cn/ccswz/file/download?fileid=201900004000006120>

⁴⁸ 郑睿:《海上自主水面船舶适航性的认定》,《浙江海洋大学学报(人文科学版)》2024 年第 6 期,第 10 页。

一个系统,智能船舶的安全航行不再像以前一样仅依赖于船舶自身的状况,而且还依赖于 ROC 的状况及其两者之间的网络连接状况。ROC 是保障智能船舶安全航行不可或缺的系统装置。因此,必须把 ROC 纳入安全检查的范围。然而,关键的问题是如何对 ROC 实施安全检查。

首先,ROC 检查主体认定困难。此问题涉及由谁负责 ROC 的安全检查,这牵扯到 ROC 的法律属性问题,即 ROC 是否属于船舶。很明显,按照有关船舶的定义,ROC 不属于船舶。关于 ROC 是否属于智能船舶的组成部分。对这个问题,可能存在“组成说”和“独立说”两种观点。“组成说”认为智能船舶需要 ROC 监控,当然属于智能船舶的组成部分。因此,港口国可依据公约授权对智能船舶及 ROC 实施安全检查,而船旗国检查则要依据 ROC 的位置问题分两种情况进行讨论。对于 ROC 位于智能船舶船旗国境内的,船旗国可直接对智能船舶及 ROC 进行管辖。而如果 ROC 的位置不在船旗国境内,那么船旗国管辖的法律依据将丧失,因为一国管辖下的 ROC 不可能是另一国国旗船舶的一部分。“独立说”认为智能船舶与 ROC 是相互独立的。此时,对港口国而言其只能基于公约授权对智能船舶进行安全检查而不能对 ROC 进行安全检查,而对船旗国来说同样需要讨论 ROC 的位置问题。ROC 处于船旗国境内时,船旗国可直接对其实施安全检查,倘若不在,则也无权管辖。通过以上分析不难看出,确定 ROC 的安全检查主体不是一个简单问题。⁴⁹

其次,对异地 ROC 实施 PSC 困难。PSC 由船舶所在的港口国实施,尽管未来可能出现移动式 ROC,现有绝大多数 ROC 与智能船舶所处位置不同,港口国没有机会对其进行 PSC 检查。然而,ROC 对智能船舶的运营起着重要作用。若无法对 ROC 实施有效监管,其运行合规性将完全取决于船旗国监管效能。⁵⁰因此,对 ROC 实施 PSC 检查将更加依赖船旗国的配合。然而,如何监督船旗国尚未明确,这直接影响到智能船舶的安全治理效能,所以未来需要深化与船旗国的执法协作,明确船旗国对智能船舶的强化监管责任。⁵¹此外,还需要解决如何对 ROC 实施安全检查、确定安全检查的内容、选择合适的安全检查方法和途径等问题。⁵²

第三,PSC 检查技术标准不一。部分智能船舶配备的远程控制功能在 PSC 检查中也引发了可靠性验证的难题。对于检查人员而言,确认远程控制在各种复杂环境下的可靠性是保障船舶安全的关键环节。然而在实际操作中,要模拟所有可能出现的复杂场景几乎是不可能完成的任务。不同品牌和型号的智能船舶,其远程控制系统在技术原理、操作方式和性能表现上存在显著差异,缺乏统一的测试

⁴⁹ 张清宝:《无人船海事监管法律问题研究》,《法学前沿》集刊 2023 年第 4 卷,第 6 页。

⁵⁰ Noma T. Existing conventions and unmanned ships-need for changes? 2016, p 54.

⁵¹ 李志伟,《无人船管辖法律问题研究》,硕士学位论文,大连海事大学,2020 年,第 39 页。

⁵² Zhang Q. A study on legal issues of unmanned ship maritime regulation, 2022, pp. 32-37.

标准和方法。这使得检查人员在面对各式各样的远程控制系统时，难以凭借现有的经验和手段判断其可靠性。一旦船舶在航行过程中遭遇紧急状况，若远程控制无法有效响应控制指令，船舶将陷入危险境地，对海上生命财产安全和海洋环境构成严重威胁。

3.2 自主航行模式与强制引航制度适配困难

UNCLOS 赋予港口国管辖权并非绝对权力，而是在公约规范下的有限裁量权。其中最具影响力的当属港口国对船舶强制引航制度的实施权限。港口国实施的强制引航制度既满足了航行安全的技术需求，又实现了国家主权的合法彰显，通过专业引航员的水域专业知识保障船舶在复杂航道的航行安全与通航效率，作为 UNCLOS 框架下沿海国行使内水及领海管辖权的具体体现，该制度本质上是国家通过航行管制实现主权宣示的法律机制。尽管智能船舶因无人在船特性显著降低了港口及狭水道航行中的人员伤亡风险，但引航制度的核心价值还在于预防船舶与港口基础设施的碰撞事故。鉴于智能船舶仍存在操纵系统故障、环境感知误差等技术风险，维持专业引航服务具有不可替代的安全保障作用。⁵³与 IMORSE 工作中所考虑的核心问题相比，智能船舶引航问题的关注度较低，而实际上这将成为限制智能船舶投入使用的重大障碍。

首先，引航相关实施标准不一。从引航实现方式来看，“引航员登船引航”⁵⁴、“引航装置安装”⁵⁵的相关规定在智能船舶场景下将难以适用，远程引航实施缺乏操作规范指引。从现有规范内容来看，目前国际层面并没有统一的关于船舶引航的规定，而更多为各国国内法规制，引航标准差异巨大，不存在能够适应智能船舶引航的统一的国际法律文书。例如，关于引航员的管理，我国引航员管理全权交由交通部海事局进行，由地方港口管理部门负责监管其辖区内沿海水域的海上交通安全，包括引航管理。⁵⁶同样，加拿大由国家对引航员能力标准和引航证书的颁发进行监管，但划定强制引航区域和颁发引航豁免证书等事务则委托给地方当局。⁵⁷而在其他国家，引航监管几乎完全委托给地方当局，由其控制引航员的证书发放、强制引航区域的划定和引航豁免证书的颁发。例如，英国设立了 90 余个“主管港口当局”监管与引航相关的所有事务⁵⁸，土耳其的引航服务由 38 个

⁵³ 郑世江：《远程控制无人船法律适用问题研究》，硕士学位论文，大连海事大学，2019 年，第 24 页。

⁵⁴ See IMO, SOLAS Consolidated Edition 2014 6th (IMO Publishing, London, 2014) 305, Regulation 23(2.2) of Chapter V of SOLAS provides that the rigging of the pilot transfer arrangements and the embarkation of the pilot “shall be supervised by a responsible officer having means of communication with the pilot bridge.”

⁵⁵ See Resolution A.1045(27) 30 November 2011, Pilot Transfer Arrangements. See also the revised poster in IMO Circular MSC.1/Circ.1428 of 28 May 2012.

⁵⁶ 参见《中华人民共和国引航员管理办法》第 3 条，

https://xxgk.mot.gov.cn/2020/gz/202501/t20250120_4162822.html；《船舶引航管理规定》第 4-7 条，

https://www.gov.cn/zhengce/zhengceku/2021-09/13/content_5637004.htm。

⁵⁷ See Transport Canada, Pilotage Act Review (2018) [Canada] ii, vii, 71-2. To take just one example of the inconsistencies, the Great Lakes Pilotage Authority has granted 318 pilotage exemption certificates, but the Pacific Pilotage Authority has issued none at all.

⁵⁸ Pilotage Act (1987) [United Kingdom] ss. 2(1)(a) (whether pilotage compulsory), 3 (licensing), 7 (designation of compulsory pilotage areas), 8 (exemptions).

不同引航区域的公共机构、私营港口公司和引航公司以合作的形式提供。⁵⁹

其次，引航豁免适用智能船舶存在难题。部分地区允许对特定的船舶或船舶类别给予引航豁免，然而多数地区只对符合一定条件的个人给予引航豁免，而不是给予特定船舶或船舶类别。对于智能船舶而言，其不存在符合条件的个人来申请引航豁免证书。此外，还有很多地区尚未规定引航豁免，这直接限制了智能船舶在这些地区的港口停靠和贸易活动，使得智能船舶的贸易范围受到极大制约。

60

最后，船长与引航员的角色关系界定模糊。在传统引航模式中，船长与引航员的职责和权力界限较为明确，引航员负责船舶操作，船长保留指挥权。但智能船舶的出现打破了这种传统模式。智能船舶场景下不存在传统意义上的船长，而关于岸基操作人员在船舶运行中的角色定位目前也尚不明确，其是否能被视为船长存在争议。如果岸基操作员被视为船长，那么他与远程引航员之间的法律关系，在船舶操作和决策过程中的权力分配以及在引航水域发生碰撞后的责任承担都需要重新进行界定和规范。这一系列问题的解决，不仅依赖于船舶技术的进一步发展，还需要相关法律法规的明确规范，目前尚处于不确定状态，给智能船舶的运营和管理带来了法律层面的挑战。

3.3 智能船舶污染损害的管辖权真空

在 IMO MASS Code 立法讨论中，阿根廷代表团曾在 IMOMSC 第 105 届会议上指出，UNCOLS 未涉及第三、四级自主船舶，因此，此类船舶进入沿海国管辖水域需得到沿海国的同意。⁶¹2018 年，MSC 同意邀请 MEPC 审查其职权范围内的文书。MEPC 同意在 MSC 就监管范围研究取得重大进展后，于未来审议该问题，以便该委员会能够跟进此类研究进程。⁶²尽管国际社会已经逐渐认识到智能船舶发展中的技术不确定性，并指出需要在海洋环境法的背景下考虑智能船舶，但仍没有进行全面深入的研究，以探讨其在国际环境法框架下的合法性和合规性，为此本文对此问题展开讨论。在预防船舶造成海洋污染方面，港口国管辖权的有效行使至关重要。UNCLOS 明确规定各国有权采取相应的法律和措施，以减少和控制其港口，内水和领海内船舶对海洋环境的污染。虽然智能船舶集成了诸多绿色环保设计理念，但由于其全新的操作模式和当前法律监管框架存在结构性错位，反而制约其环保效能的充分释放，甚至衍生新型船源污染风险，由此可能使环保技术创新陷入“制度性消解”的窘境⁶³，这种制度性障碍不仅削弱了港口国环

⁵⁹ Ö. Uğurlu, M. Kaptan, S. Kum, and S. Yildiz. Pilotage services in Turkey; key issues and ideal pilotage, *Journal of Marine Engineering & Technology*, Vol. 16, No. 2(2017), pp. 51-54.

⁶⁰ Davies M. Pilotage of autonomous and remotely-controlled ships, *Autonomous Ships and the Law*. Routledge, 2020, p. 287.

⁶¹ See Report of The Maritime Safety Committee on Its 105th Session, MSC 105/20, para. 7.19

⁶² See Report of The Marine Environment Protection Committee on Its 73rd Session, MEPC 73/19, para. 2.3

⁶³ Chang, C.H., Kontovas, C., Yu, Q., Yang, Z. Risk assessment of the operations of maritime autonomous surface ships. *Reliability Engineering & System Safety*, Vol. 207, Issue, 2021. Pp.107324

境执法效能，更导致污染事故处置陷入“有管辖权无执行手段”的尴尬境地，对海洋环境保护构成系统性风险。

首先，智能船舶船源污染管辖权存在竞合。智能船舶的全球航行特性易引发多重管辖冲突。根据 UNCLOS 第 94 条⁶⁴，船旗国对智能船舶负有首要管辖义务，但智能船舶的远程控制削弱了其实际监管能力⁶⁵，船旗国、沿海国与港口国的管辖权可能重叠。此外，网络攻击引发的污染事故可能同时触发网络安全法、环保法等多领域管辖，导致责任竞合与法律衔接缺失。港口国在行使监督权时，需重新界定船舶“适航性”标准，例如是否豁免“最低安全配员”要求，或如何评估远程操作系统的合规性。

其次，智能船舶防污责任主体不明。传统法律框架下，船长和船员构成污染防治的义务主体，其物理在场性为港口国行使管辖权提供锚点。例如，MARPOL 公约要求油类记录簿签署、船上海洋污染应急计划执行等义务，均预设人类决策者的直接参与。而智能船舶的“无人化”特性瓦解了这一基础，港口国既无法依据现行公约对“虚拟船长”实施责任追溯，也难以穿透至远程操作中心、算法供应商及服务商等主体⁶⁶，最终导致污染损害赔偿陷入多方推诿的僵局。此外，现有法律既未明确岸上操作员是否替代船员义务，也未规定自主技术能否履行传统船员的污染防控职能，可能导致责任归属真空。⁶⁷

最后，港口国对智能船舶污染应急处置规范缺失。船舶意外污染是船舶碰撞、管道破裂、倾覆或火灾等非人为意愿的海上事故导致的结果。⁶⁸为避免此类情况发生，MARPOL 公约要求特定船舶配备“船上海洋污染应急计划（SMPEP）”，这样的计划至少应包括：“船长或其他负责船舶的人员报告油类或有毒液体物质污染事故应遵循的程序”以及“船上人员在事故发生后立即采取的减少或控制油类或有毒液体物质排放的详细措施”。⁶⁹港口国监督官员（PSCOs）依据区域谅解备忘录实施的突击检查，以船员在场、证书可查、设备可视为基础操作逻辑。然而，智能船舶可能导致 PSCOs 的检查出现问题。一方面，ROC 可能位于第三国，导致 PSCOs 与远程操作人员的实时沟通面临时差、语言及司法管辖壁垒从而导致污染扩散；另一方面，电子证书的跨国认证标准缺失，使得 MARPOL 要求的记录簿、应急计划等核查演变为数据主权博弈。港口国当局如何实施上述污染事

⁶⁴ 《联合国海洋法公约》第 94 条：每个国家应对悬挂该国旗帜的船舶有效地行使行政、技术及社会事项上的管辖和控制。

⁶⁵ 赵骏，李婉贞：《人工智能对国际法的挑战及其应对》，《浙江大学学报(人文社会科学版)》2020 年第 2 期，第 10-25 页。

⁶⁶ Chougrani E H, Al Ali N A R, Chebotarev V. Legal regulation of the safety of autonomous ships: environmental challenges and approaches, *E3S Web of Conferences*. EDP Sciences, 2021, p. 296.

⁶⁷ 俞毅：《智能航运背景下智能航海保障发展初探》，《交通运输管理干部学报》2019 年第 2 期，第 12-15 页。

⁶⁸ Karahalios, H. The severity of shipboard communication failures in maritime emergencies: a risk management approach. *International Journal of Disaster Risk Reduction*, Vol. 28(2018), pp.1-9.

⁶⁹ Xing, Wang wang. Contemplating Maritime Autonomous Surface Ships (MASS) under the international law on ship-source pollution. *Marine Pollution Bulletin*. Vol.207, Issue. 2024. pp. 11688.

故的实体和程序性检查尚不明确。

四、智能船舶港口国管辖的规制应对

4.1 制定远程操作中心PSC检查的标准

当前PSC机制因智能船舶的无人化、远程化和数据依赖性面临技术性失灵，亟需通过国际协作实现标准统一与技术创新。首先，要将远程操作中心（ROC）纳入“船舶组成部分”的法律框架，并明确其合规性标准，修订现有PSC检查程序，允许通过远程接入方式对智能船舶的软件系统、数据记录及网络安全性进行核查，确保检查手段与技术发展同步。其次，建立统一的PSC技术标准与认证机制。智能船舶的异构性导致其远程控制系统、通信协议和数据格式差异显著，现有PSC检查工具难以适配。应推动IMO联合ISO、IEC等机构，制定智能船舶设备性能测试的技术性标准，涵盖传感器精度、算法决策可靠性、网络安全防护等级等关键指标。最后，强化区域PSC合作与数据共享机制。智能船舶的全球化运营要求突破传统PSC的区域限制。针对ROC的跨境监管难题，可探索“主检国+协检国”模式，当智能船舶的ROC位于第三国时，港口国可请求船旗国或ROC所在国提供远程检查协助，通过视频连线、数据实时传输等方式完成联合检查。此外，还需建立智能船舶PSC风险评级体系，结合船舶技术状态、历史违规记录及网络安全等级，实施差异化监管，优化资源配置，加强船旗国、港口国与沿海国之间的合作。

就国内而言，应当制定ROC安全检查规则。首先需要明确由谁来实施检查，对此有学者提出两种方案，一种认为应由ROC所在的国家具体负责安全检查并安全检查报告；另一种认为各船旗国的智能船舶只能由设置在船旗国境内的ROC进行控制，应由船旗国负责现场安全检查，出具检查报告并督促缺陷整改，且能提供电子版报告供港口国查验。⁷⁰本文认为，两个方案各有优势。倘若采取第一个方案，本文建议引入智能船舶第三方监督检查机制，港口国主管部门可以根据船舶的风险等级、船旗国信誉等因素，合理安排第三方监督检查的比例，对于高风险船舶或信誉不佳船旗国的船舶，可以增加第三方监督检查的频次，加强监管力度。⁷¹倘若采取第二个方案，港口国则要加强与船旗国的联系，严格船旗国的责任。其次，应在《船舶和海上设施检验条例》中增设“远程操作中心检验”条款⁷²，将智能船舶远程操作中心列为法定检验对象，规定远程操作中心检验的内容，包括法定检验的种类，即建造时，申请建造检验；使用时，申请定期检验；海上交通安全主管机关责成检验的，应当进行临时检验；智能船舶与远程操作中心连接时，申请临时检验等等，检验合格后检验机构应当签发相应的检验证书。

⁷⁰ 张清宝：《无人船海事监管法律问题研究》，《法学前沿》集刊2023年第4卷，第7页。

⁷¹ Tomotsugu Noma, Existing conventions and unmanned ships-need for changes? 2016, pp. 54-55

⁷² 参见《中华人民共和国船舶和海上设施检验条例》，
https://www.gov.cn/gongbao/content/2019/content_5468917.htm。

同时,海事局也应当制定专门的智能船舶与 ROC 检验规则作为法定检验标准⁷³,应明确 ROC 的检查内容,除了对系统进行检查外,重要的是对 ROC 的配员情况进行检查,如配员数量、人员资质、人员监管船舶数量等等。同时,还需要对 ROC 安全运营和日常管理情况进行检查,通过核查运营工作日志和日常管理记录以审查可能存在的风险漏洞。

4.2 明确强制引航适用智能船舶

智能船舶的进离港涉及主权宣示、技术适配与风险防控等多重问题,需通过国际规则协调构建全流程监管框架,平衡航行自由与港口国安全权益。传统引航制度以“船员在场”为前提,与智能船舶的远程控制模式严重冲突。为此,从国际层面来看,可通过修订 SOLAS 第 V 章,增设“远程引航”条款,允许港口国通过专用通信信道向智能船舶发送导航指令,并规定岸基操作员与引航员的权责边界。此外,需推动国际船舶引航规则的出台,将智能船舶引航豁免条件标准化。例如,对通过国际认证的自主航行系统,可授予“自动引航豁免证书”,允许其在不依赖人工引航的情况下进入特定水域,但需实时向港口国传输航行数据以供监督。

国内层面,探索制定智能船舶强制引航的数字化规则。首先,建立分级化引航模式。根据不同自主等级分别开展引航活动。对 L1-L2 级自主航行能力相对较弱的智能船舶或者在一些特殊情况下,如港口水域发生紧急事件、航道条件发生重大变化等,应强制实施引航。对于 L3-L4 级自主航行能力较高的智能船舶,可以在满足一定条件的情况下,适当放宽引航要求,如对通过中国认证的智能船舶授予“技术可信豁免”,允许其在指定航段免于传统引航。其次,实现远程引航法定化,可在《船舶引航管理条例》中增设“远程引航操作条款”,授权港口国海事部门通过 5G 专网对境外控制中心实施电子核查,要求运营商提供实时操作日志及视频流数据。同时,与主要航运国签署双边协议,建立跨境引航员资质互认机制。最后,为确保智能船舶引航工作的顺利开展,还需加强引航员与智能船舶之间的协同配合。

4.3 增设智能船舶污染损害管辖的规定

国际层面,应当建立跨国事故调查与应急响应机制。智能船舶事故数据常存储于跨国服务器,单一国家调查权受限。IMO 应推动海事事故调查国际合作协定的签署,赋予港口国“数据调取优先权”,并设立“国际海事事故数据库”,要求船旗国、制造商与运营商强制上传事故相关数据。同时,需成立“智能船舶应急响应中心”(MASS-ERC),配备跨国专家团队与标准化应急处置协议,确保在黑客攻击或系统失控时,可快速接管船舶控制权,避免次生灾害。此外,需要填补污染损害赔偿的国际规则真空。智能船舶的无人化导致传统污染责任主体船长、船

⁷³ 潘虹:《智能船舶监管法律问题研究》,硕士学位论文,大连海事大学,2020 年,第 41 页。

员消失，MARPOL 公约的“船东责任”条款需扩展至远程操作员与技术服务商。国际油污赔偿基金（IOPC）应修订赔偿范围，将算法错误引发的污染纳入承保范围，并通过“绿色债券”机制筹集应急资金⁷⁴，明确智能船舶污染应急国际标准，要求船舶配备“自主溢油回收系统”，并与港口国监测网络实时联动，实现污染源的快速定位与封堵。

国内层面，需要增设智能船舶污染损害的特别管辖条款。首先，增设“智能船舶污染特别责任”条款，将锂电池泄漏、自主排污系统故障等纳入严格责任范畴，取消赔偿责任限制。同步发布“智能船舶防污技术规范”，强制配备污染传感器与自动封堵装置。其次，强化跨境应急协作。依托区域谅解备忘录建立“智能船舶污染响应联盟”，制定统一电子证书互认标准与应急数据共享协议。当发生跨境污染时，港口国可一键触发联盟成员国的无人清理船协同作业。最后，建立资金保障机制。设立“智能船舶污染赔偿基金”，按船舶吨位收取年度保证金。基金用于垫付应急处理费用并向责任方追偿。同时，要求保险公司开发“算法缺陷险”，将承保范围扩展至自主系统故障导致的生态损害。

结语

智能船舶的技术兴起对港口国管辖制度构成系统性挑战，暴露了传统国际海事法律框架在技术适配性、责任界定及监管效能等方面的结构性滞后。尽管 IMO 通过制定基于目标的 MASS 规则等文件初步构建了智能船舶的国际监管框架，但无人化、自主化及数据化特征仍引发管辖权冲突、算法责任空白及 PSC 技术失灵等深层问题。现行公约体系对“人”的规制依赖与智能船舶的技术逻辑存在本质冲突，区域性协调不足与标准碎片化进一步加剧监管割裂。未来需强化国际公约的强制力层级，推动技术审查与责任机制的全球统一，构建“技术实践-法律治理”回应式监管范式体系，平衡航行自由、主权安全与全球航运秩序。智能船舶技术的迭代与港口国管辖制度的重构，本质上是海洋治理范式从物理疆域向数字主权延伸的深刻映射。为此，需要国际社会应以 IMO 为核心平台，推动公约条款的动态解释与专属规则创设，强化技术标准与法律义务的衔接耦合，构建智能船舶港口国管辖的国际监管体系。中国作为智能航运变革的引领者，需统筹国内立法创新与国际规则话语权建设，通过输出港口国监督检查制度规范、全过程监管进离港船舶和明确突发事件管辖执行依据等路径，实现从国际规则引领到本土治理效能提升的闭环演进，为智能时代海洋法治贡献兼具创新性与安全性的中国方案。

⁷⁴ 参见《什么是绿色债券，为什么其市场增长如此之快？》，2024 年 12 月 21 日，<https://cn.weforum.org/stories/2024/12/what-are-green-bonds-climate-change-cn/>，2025 年 4 月 5 日。

“荷花号”原则的澄清和再发展

——以网络间谍活动的国际法应对为例

文金金¹

摘要：国际常设法院在“荷花号”案中的论证逻辑表明，“荷花号”原则有着严格的适用前提和条件。领土主权是“荷花号”原则的适用前提，一国未经允许不得在他国领土上以任何形式行使其权力。“荷花号”原则的适用条件是出现国际法的“规制空白”，国家此时拥有任意行动的自由。“荷花号”原则可发展为一项法律适用规则，首先应寻求是否有具体国际法规则判断一项活动的非法性，如无则在相关国际法领域中审查是否存在国际法规范，涵盖国际法具体规则和国际法基本原则。对于网络间谍活动，其本身构成对他国领土主权的侵犯，不符合“荷花号”原则的适用前提，国际法对此也并未出现“规制空白”。针对包括网络间谍活动在内的网络行动，尽管国际法目前出现了“规则真空”，但尽管没有专门的国际法规则，但以主权原则为代表的国际法基本原则，能够在网络空间中发挥规制作用，应当在实践中丰富国际法基本原则的内涵。

关键词：“荷花号”原则；网络间谍；国际法基本原则

引言

“荷花号”原则产生于1927年“荷花号”案的判决中，国际常设法院在讨论“土耳其对他国公民行使刑事管辖权是否违反国际法”这一问题时，将国际法视为禁止性规则体系并提出“对国家独立自主的限制不能被推定”。该案中的这一论断被后来的国际法学者发展为一项法律原则，即“荷花号”原则。“荷花号”原则适用的部分意涵在于，国家行使主权的自由只受到有关国家同意的禁止性规则的限制，在没有禁止性规定的情况下，国家可自由采取行动而无需国际法规范的授权。由此，“荷花号”原则在国际法规范适用问题上带来了一定范围和条件下的“法不禁止即自由”。基于“荷花号”原则所包含的国家中心主义思想，该原则通常被一些学者用于论证国家有权自由行为而不构成对国际法的违反。²国际法作为不同于国内法的规则约束体系，由于其缺乏集中权威对规范作出统一制定和解释，对“荷花号”原则的讨论较为莫衷一是。一些错误主张在于，将“荷花号”原则认定为一种笼统的原则，即“国际法不禁止即允许”，但这种解释并没有准确把握到“荷花号”案中国际常设法院的推理逻辑。³

¹ 武汉大学法学院国际公法学研究生，18374030661

² See Marco Vohringer, State Jurisdiction and the Permissiveness of International Law: Is the Lotus Still Blooming 7(1) LSE Law Review 29-60 (2021).

³ Hugh Handeyside, The Lotus Principle in ICJ Jurisprudence: Was the Ship ever Afloat, 29(1) Michigan Journal of

在网络空间这一国际法的新疆域中，国际法的造法进程滞后于信息技术发展的背景下，在一些网络活动没有具体的专门性规则发挥规范作用时，是否意味着各国可依据“荷花号”原则主张自身的行为自由，是网络空间国际法在发展过程中需要澄清的重要问题。以网络间谍活动为例，由于目前尚不存在规制和平时期网络间谍活动的专门性规范，一些学者凭借对“荷花号”原则的片面理解，以证明网络间谍活动是国际法上的“法无禁止即自由”，不违反现行国际法。⁴但，“荷花号”原则存在于国际常设法院的判决中，被赋予了特定的适用条件和前提，其完整意涵要求审慎识别国际法规范的存在。

一、“荷花号”原则的内在法律逻辑

“荷花号”原则既产生于国际常设法院关于“荷花号”案的判决中，有必要重新审视“荷花号”案的相关论证，以便澄清“荷花号”原则的内在法律逻辑。

（一）“荷花号”原则的适用前提：领土主权原则

“荷花号”原则是国际常设法院在讨论“土耳其行使刑事管辖权是否违反国际法”这一问题时提出的，土耳其和法国作为争端方所签订的《洛桑公约》第15条涉及刑事管辖权的分配问题，而第15条规定土耳其与其他缔约国之间的所有司法管辖问题应依据“国际法原则”作出决定。⁵法院通过条约文本解释和上下文解释的方法，认为第15条文本所述“国际法原则”，仅指国际法一般原则中关于司法管辖的规定。⁶法院继而转向了对有关司法管辖的国际法规范的考察。

由于法国在诉讼中主张，“土耳其法院具有管辖权，必须能够指明国际法所认可的赋予土耳其管辖权的依据”，而土耳其政府则认为，只要管辖权的行使不与国际法原则相冲突，土耳其就可以主张管辖权。⁷法国和土耳其的不同主张，本质上是适用国际法规范判断行为违法性的两个不同角度。最终，国际常设法院考虑到两国提交诉讼所签订的特别协议，协议规定双方达成一致要求法院裁定的问题是，“土耳其是否违反了国际法原则，以及若有违反，具体涉及哪些原则”。⁸这直接决定了国际常设法院对于土耳其行为违法性的判断角度。

法院基于上述违法性判断角度，认为“问题的关键并非在于确认哪些原则允许土耳其进行刑事诉讼，而是界定是否存在土耳其可能违反的国际法原则”。法院进一步指出，“这种问题表述方式也符合国际法的本质及其现有状况”，并引出了“荷花号”原则的经典论述，即“国际法调整独立国家之间的关系。对各国具有拘束力的法律规范源于这些国家的自由意志，该自由意志体现于……国际公约

International Law 72 (2007).

⁴ See K. Browne, The paradox of peacetime espionage in international law: From state practice to first principles, 23 Australian international law journal 109-123 (2018).

⁵ Article 15 of the Convention of Lausanne of July 24th, 1923, provides that: “Subject to the provisions of Article 16, all questions of jurisdiction shall, as between Turkey and the other contracting Powers, be decided in accordance with the principles of international law.”

⁶ SS Lotus (France v Turkey) (Judgment) PCIJ, Series A No 10 (1927), page 17.

⁷ Ibid, page 18.

⁸ Ibid, page 16-17.

或……惯例中。因此，对国家独立自主的限制不能被推定。”

尽管“荷花号”案判决从“法无禁止即自由”的角度提出了“荷花号”原则，并以此判断土耳其行使行使管辖权的行为不违反国际法，但该案中的魏斯法官和芬利法官在反对意见中提出了不同的观点，认为问题应当在于在国际法中找到土耳其行使刑事管辖权的授权，而非寻求是否违反国际法规则。⁹上述法官的反对意见表明，“荷花号”原则所体现的违法性判断角度是存在争议的。有学者则认为，以“法无规定即禁止”为核心的“荷花号”原则，是对判决的不完整解读，因为多数法官意见实则是在规则不明确的情况下对违法性问题保持沉默。¹⁰布赖尔利(Brierly)也对“荷花号”原则所基于的极端实证主义思想表示反对，并主张不能仅从国家主权独立而推断出适用于特定事实状态的法律。¹¹一些学者更是将“荷花号”原则片面地理解为国际法上的“法无规定即自由”，并以此为“荷花号”原则的全部意涵。¹²这种理解这显然是对国际常设法院论述的阉割，忽视了后续论述中的必要前提。

笔者认为，基于“荷花号”原则对法律适用问题的填补作用，其不应被完全否定，当然，更不应被夸大成为国家行为合法性的粉饰工具，其真实意涵应当完整审视国际常设法院的论述，其中以领土主权原则为必要之前提。国际常设法院既主张“对国家独立自主的限制”不能被推定，随即补充了作为前提的“首要限制”，即“除非存在相反的允许性规定，一国不得以任何方式在他国领土内行使权力”。换言之，一国独立自主行为仍然受到国际法上领土主权原则的首要限制，而在首要限制之外，可能的限制还包括体现国家自由意志的国际公约或惯例，在没有上述限制的情况下，不能当然认为国家的自由行为是受到限制的。

国际常设法院之所以将领土主权原则定义为国家独立自主行为的“首要限制”，本质上是领土主权原则的国家中心色彩所决定的。胡伯法官(Judge Huber)在1928年“帕尔马斯岛案”中将主权定义为，“国家间关系中的主权意味着独立，指的是在该区域内排他性地行使国家职能的权利，不受任何其他国家干涉。”¹³国际法院作为国际常设法院的承继机构，在“科孚海峡案”、“尼加拉瓜案”、“以色列隔离墙案”中也无不强调领土主权原则，一国不得在未经他国同意的情况下在另一国领土上采取强制性措施，包括军事行动、干涉内政等行为。领土主权原则意味着主权国家在领土范围内有权独立自主行使权力，同时要求国家承担未经允许

⁹ SS Lotus (France v Turkey) (Judgment) PCIJ, Dissenting Opinion of Judge Weiss, page 42; Dissenting Opinion of Judge Finlay, page 52.

¹⁰ See An Hertogen, Letting Lotus Bloom, 26(4) The European Journal of International Law 904-908 (2015).

¹¹ J.L. Brierly, The 'Lotus' Case, 44 L.Q. REV. 154, 155 (1928), reprinted in THE BASIS OF OBLIGATION IN INTERNATIONAL LAW AND OTHER PAPERS 143-44 (Sir Hersch Lauterpacht ed., 1958).

¹² Yuval Shany, Toward a General Margin of Appreciation Doctrine in International Law?, 16 Europe. Journal of International Law. 907, 912 (2005). Anne Peters, 'Does Kosovo Lie in the Lotus-Land of Freedom' (2011) 24(1) Leiden Journal of International Law 100. See Katja Schöberl and Linus Mührel, Sunken Vessel or Blooming Flower? Lotus, Permissions and Restrictions within International Humanitarian Law, Björnstjern Baade/Linus Mührel/Anton O. Petrov (eds.) International Humanitarian Law in Areas of Limited Statehood, Adaptable and Legitimate or Rigid and Unreasonable? 1st Edition 2018, pp.66.

¹³ Island of Palmas Case United States v Netherlands (Award, Permanent Court of Arbitration, 829, 4 April 1928) [29].

不得侵入他国领土的义务。国家作为权利享有者和义务承担者的双重中心身份，决定了“荷花号”原则之下，国家独立自主行为的必要前提是符合领土主权原则。

（二）“荷花号”原则的适用条件：国际法的“规制空白”

在“荷花号”案的判决中，国际常设法院还指出了“荷花号”原则的适用条件，但这一必要条件同样却被一些学者所模糊，从而导致了“荷花号”原则被解释为“凡是国际法没有明确禁止的，就是被允许的”。¹⁴根据笔者前述对国际常设法院论述的分析，以领土主权原则为前提，国家自由行事的权利还受到国际法规范的限制。国际法规范作为一种限制，源于其是国家自由意志的体现，本质上是国家在国际社会中的“自我约束”。¹⁵从反面观之，当不存在国际法规范的限制，即国际法出现了“规制空白”现象时，国家的自由意志不能被推定为受国际法约束，国家在不违反领土主权原则的基础上具有行为自由。上述反面角度也正是以“荷花号”原则论证国家行为合法性的常见方法。

重新审视“荷花号”一案的判决可以发现，“荷花号”原则的本质是将国际法规范视为禁止性规则体系而构成的一项剩余规则，但该项“剩余规则”的适用受到严格限制。领土主权原则作为“荷花号”原则的适用前提，约束着国家独立自主的行为。国际法规范作为国家行为自由的必要条件，当出现“规制空白”的情形时，“荷花号”原则作为一项剩余规则为国家行为的自由提供了法律基础，国家在符合领土主权原则的前提下行事才属合乎国际法。

二、“荷花号”原则的再发展：国际法上的法律适用规则

“荷花号”原则若是囿于过分狭窄的涵义，显然无法满足全球化条件下高度复杂的社会生活。¹⁶“荷花号”原则所包含的国家中心主义的思想，存在一定的时代局限性，随着国际法律体系趋向复杂化，国家独立自主的行为越来越受到国际法律规范的约束。国际法院在后续的司法判例中，从未直接引用过“荷花号”原则，仅有三次在案情实质裁决中提及“荷花号”案，国际法院所持的模糊态度也表明了“荷花号”原则存在的狭隘性。¹⁷

无论是从“荷花号”原则的前提还是适用条件而言，都绕不开国际法的适用。国际法作为不同于国内法的分散化体系，如何从法律适用的角度判断某一行为是否违反国际法，国际法院在经典的“核武器咨询意见案”中作出过相关论证，该案也被视为是国际法院对“荷花号”原则的间接适用和发展。¹⁸因此，“荷花号”案后法律适用方法的发展，或能为“荷花号”原则在当今复杂的国际法体系中注入新的

¹⁴ See An Hertogen, Letting Lotus Bloom, 26(4) The European Journal of International Law 902 (2015).

¹⁵ Oliver Dörr, Kirsten Schmalenbach, Vienna Convention On The Law Of Treaty: A Commentary, Springer-Verlag GmbH Germany 2018, 2nd edition, pp.472

¹⁶ 参见陈一峰：《国际法不禁止即为允许吗？——“荷花号”原则的当代国际法反思》，载环球法律评论 2011 年第 3 期，第 132-141 页。

¹⁷ Máté Csérmus, Might contain traces of Lotus: The limits of exclusive flag state jurisdiction in the Norstar and the Enrica Lexie cases, Leiden Journal of International Law (2023), 36, 949-950.

¹⁸ See Hugh Handeyside, The Lotus Principle in ICJ Jurisprudence: Was the Ship ever Afloat, 29(1) Michigan Journal of International Law 71-94 (2007).

活力。

（一）从具体规则到相关规范的周延考察

鉴于国际法固有的缺陷，以及政治、经济和技术变革对传统领域的冲击，国际司法机构在案件审理过程中，无法直接适用法律规则的现象并不少见。在“烟雾案”、“巴塞罗那牵引公司案”、“科孚海峡案”、“北海大陆架案”、“东部延伸案”等案件中，国际法院都未能找到可直接适用的国际法规则。¹⁹但在这些案件中，国际法院并没有当然认定国家行为合乎国际法，而是广泛考虑了相关领域可能适用的国际法规范，并以规范为基础作出了国家行为合法或非法的结论。目前在国际社会的一些新型领域，国际立法的确出现了滞后的现象，但缺乏直接适用的具体规则并不当然意味着国家独立自主行为的合法性。²⁰“荷花号”原则中“规制空白”这一适用条件，不是指缺乏具体明确的法律规则，而是缺乏可供适用的国际法规范，这无疑需要对国际法规范从具体规则到相关规范进行周延的考察。国际法院在“核武器咨询意见”案中，正是对国际法规范进行了由点到面的审视。首先，法院审查了国际法中是否有具体规则规范诉诸核武器本身的非法性的问题，尽管尚不存在适用于核武器规制的具体规则，但国际法院也并没有给出合法性的确定结论或拒绝作出判决，进而，国际法院考察了适用于武装冲突领域的法律，包括适用于武装冲突的人道主义法原则和规则以及中立法，以审查诉诸核武器的非法性问题。²¹从上述实践可知，国际法院对国际法规范的适用采取了谨慎态度，尽管法院并未直接引用“荷花号”原则，但该案的法律适用方法体现了从特别法到普通法的法理逻辑，实为对“荷花号”原则下“规制空白”这一条件的全面审查。

（二）相关规范涵盖国际法规则与原则

在不存在具体规范可供直接适用时，国际司法机构将目光扩大至相关领域的规范，这些规范应涵盖国际法规则和国际法基本原则。联合国国际法委员会通过的《国家对国际不法行为的责任条款草案》（“ARSIWA”）第12条规定，“当一国的行为不符合一项国际义务对它的要求时，该国即违反了该义务，不论该义务的起源或性质如何”。²²该条是对国际不法行为客观要件的表述，“义务的起源”显然涉及到对国际法规范的考察，无论这种规范属于条约规则、习惯法规则还是一般法律原则。当一个国家的行为违背了对该国有约束力的国际义务时，国际不法行为成立，此处国际义务的来源可以是国际条约、国际习惯、国际法上的一般法律

¹⁹ Kiyotaka Morita, The Issue of Lacunae in International Law and Non Lique Revisited, Hitotsubashi Journal of Law and Politics 45 (2017), pp.33.

²⁰ Steven R. Ratner, Symposium: State Intelligence Gathering and International Law - Introduction (2007) 28(3) Michigan Journal of International Law 539. Yang Liu, A Meta-Critique of Frontier Scholarships on the Laws of Peacetime Espionage: Towards a Systemic Framework for Lex Specialis, (2022) 14(1) Case Western Reserve Journal of Law, Technology and the Internet 49-51. Kevin Jon Heller, In Defense of Pure Sovereignty in Cyberspace, (2021) 97 International Law Studies Series. US Naval War College 1453-1454.

²¹ Legality of the Threat or Use of Nuclear Weapons, ICJ, Advisory Opinion of 8 July 1996.

²² Article 12 of Draft Articles on Responsibility of States for Internationally Wrongful Acts: “There is a breach of an international obligation by a State when an act of that State is not in conformity with what is required of it by that obligation, regardless of its origin or character.”

原则，甚至可以是国家单方面的行为所产生的国际义务。²³因此，判断某一行为的客观违法性时，需要周延适用国际法规范，包括条约、习惯和一般法律原则中体现的规则及原则。在适用“荷花号”原则下“规制空白”这一条件判断国家行为的合法与否时，应当全面考察国际法规则和国际法基本原则。

综上所述，国际法院在“核武器咨询意见”案中的上述法律适用方法和国际法委员会的实践表明，法律适用方法应遵循从特别法到普通法的法律适用逻辑，以及原则和规则应当共同适用的法律适用逻辑。“荷花号”原则所包含的“规制空白”这一必要条件，应当在以上两种法律适用逻辑的轨道上进行审查。

三、适用“荷花号”原则以“合法化”网络间谍活动的谬误

由于现有国际法规范仅对武装冲突时期的间谍行为作了较为明确的规定，而平时时期的间谍问题则凸显空白。²⁴有的学者主张根据“荷花号”原则，国际法上对于平时时期的网络间谍活动没有法律依据可供适用，在出现法律空白的情况下应当被认为合法。²⁵实际上，平时时期的网络间谍活动不符合“荷花号”原则的适用前提，也并非属于“规制空白”的适用情形，不能以“荷花号”原则为由主张合法。

（一）网络间谍活动本身违背领土主权原则

国际常设法院在“荷花号”案中提出的，“任何国家未经允许不得在他国领土内行使权力”的原则，是国际法律秩序的基石之一。²⁶这一原则正是领土主权原则的精要表达。1970年《关于各国依联合国宪章建立友好关系及合作之国际法原则宣言》（“《国际法原则宣言》”）中提出了“各国主权平等”原则，其要素包括“各国享有充分主权之固有权利”和“各国均有义务尊重其他国家的人格”。领土主权和领土完整是国家政治独立的重要标志，是久经确认的最基本的国际关系准则和最重要的国际法基本原则。²⁷领土主权原则一方面具有权利的性质，即各国在领土范围内行使充分主权所固有的权利，具体表现为主权国家能够独立自主地进行内政事务的安排，对本国领土范围内的事务行使全面的管辖权，不受其他任何国家和非国家行为体的干涉。另一方面，一国领土主权的充分行使有赖于其他国家履行相应不作为义务，不对他国行使固有权利造成干扰，这也正构成了不干涉原则的内涵。主权不仅意味着权利，也意味着义务，一国应在遵守其国际法义务的基础上行使主权，例如根据其签署的条约或习惯法所产生的义务，肆无忌惮地行使主权将违反国际法。²⁸

²³ 参见朱文奇：《现代国际法》，商务印书馆2013年版，第432页。

²⁴ 国际人道法对战时间谍的地位和权利作出了较为详尽的规定。1977年《日内瓦第四公约关于保护国际性武装冲突受害者的附加议定书（第一议定书）》第46条即规定了间谍的身份界定与排除间谍待遇的情形。

²⁵ Steven R. Ratner, Symposium: State Intelligence Gathering and International Law --Introduction (2007) 28(3) Michigan Journal of International Law 539.

²⁶ See Armin von Bogdandy, Markus Rau, The Lotus, Max Planck Encyclopedia of Public International Law [MPEPIL] Oxford Public International Law (<http://opil.ouplaw.com>). (c) Oxford University Press, 2023.

²⁷ 参见梁西主编：《国际法》，曾令良修订主编，武汉大学出版社2011年版，第126页。

²⁸ 尼古拉斯·查古里亚斯：《网络空间主权原则的地位及其重要性》，载中国信息安全，2021年第11期，

主权原则具有权利义务的双重性质，在网络空间中亦然。网络主权是国家主权在网络空间的自然延伸，是一国基于国家主权对本国境内的网络设施、网络主体、网络行为及相关网络数据和信息等所享有的对内最高权和对外独立权。²⁹网络主权这一概念的诞生，意味着领土主权原则在网络空间中产生了延申。尽管网络空间没有确定的边界，但领土主权原则要求一国在实施网络行动时应当尊重他国的领土主权和安全，在享有网络空间自主权的同时，不侵犯他对领土范围内的通信技术和设施。网络间谍活动，由于其本身固有的技术属性，必然会涉及对他国领土上计算机系统的入侵。³⁰在领土主权原则的范围内，计算机系统并不会脱离领土而存在，一国实施网络间谍活动势必会产生侵入他国领土的后果，而未经一国允许侵入其领土显然构成对领土主权原则的违反。

结合领土主权原则的完整内涵，网络间谍行为利用跨国信息窃取、监控等手段对他国领土范围内的计算机系统进行攻击，显然违反了领土主权原则。而“荷花号”原则允许国家独立自主行为，是受限于领土主权原则这一前提的，不得以“荷花号”原则主张网络间谍活动的合法性。

（二）国际法针对网络间谍活动不存在“规制空白”

除开“领土主权”原则，“荷花号”原则的适用还要求出现国际法上的“规制空白”。对于网络间谍活动而言，“规制空白”意味着国际法上不存在可供适用的规范来规制网络间谍活动。美国国防部总法律顾问保罗·内伊在2020年的演讲中称，“反间谍条约的缺失以及许多国家开展间谍活动的具体实践表明，并不存在禁止间谍活动的习惯国际法规范”。³¹该观点存在两大谬误，首先，不存在反间谍条约和禁止间谍活动的习惯法，并不意味着出现“荷花号”原则下的“规制空白”，因为“规制空白”不仅要求考察规制间谍活动的特殊国际法，还应寻求可供适用的一般国际法规范。“没有专门的禁止”并不意味着当然适用“荷花号”原则。目前国际法上虽然没有禁止间谍活动的具体规则，但国家主权原则可以用于规制网络间谍活动，“荷花号”原则无法支持网络间谍活动的合法性。³²领土主权作为国家主权原则的核心内容，使得国家能够在领土范围内排他性行使权力。国家主权是一国对领土内最高权威和控制的主张，象征着基于领土的政治和法律共同体的连贯性、统一性和独立性。³³网络间谍活动所具有的跨领土入侵的特征，在法律层面已然

第 81 页。

²⁹ 武汉大学、中国现代国际关系研究院、上海社会科学院：世界互联网大会成果文件《网络主权：理论与实践》，https://cn.wicinternet.org/2019-10/21/content_36314787.htm。

³⁰ Kevin Jon Heller, In Defense of Pure Sovereignty in Cyberspace (2021) 97 International Law Studies Series. US Naval War College 1453-1454.

³¹ Paul Ney, “DOD General Counsel Remarks at U.S. Cyber Command Legal Conference”, <https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/>.

³² Yang Liu, *A Meta-Critique of Frontier Scholarships on the Laws of Peacetime Espionage: Towards a Systemic Framework for Lex Specialis* (2022) 14(1) Case Western Reserve Journal of Law, Technology and the Internet 49-51.

³³ Jean Cohen, *Globalization and Sovereignty: Rethinking Legality, Legitimacy and Constitutionalism* (Cambridge

违反领土主权原则，破坏了国家主权的统一性。

综上所述，对于网络间谍活动能否适用“荷花号”原则这一问题而言，西方学者存在的谬误在于忽视了“领土主权”原则这一必要前提，以及采取了片面的法律适用方法。在主权原则延申入网络空间的应有之义下，明确各国在领土范围内进行网络活动的权利与义务的统一性，以此明确网络间谍行为的违法性，有助于更好规制网络间谍问题。

四、网络空间“规则真空”的国际法应对：国际法基本原则的规制作用

网络空间国际法作为国际法发展的“新疆域”，现有国际法在网络空间的适用面临滞后性问题。一方面，网络空间的特殊性使得传统国际法规则难以直接适用，规则解释面临现实分歧。另一方面，网络空间缺乏专门性国际规则，导致了包括网络间谍活动在内的网络行动陷入是否构成国际法上“法无禁止即可为”的泥潭。网络空间“规则真空”状态会导致网络空间的国际治理面临法律不确定性，亟需在国际法层面探索可行的规制路径。

国际法仍然存在着很多缺漏和不足的现状下，基本原则的存在对于法律欠缺情况下的解释和适用具有重要的帮助作用。³⁴对包括网络间谍活动在内的网络行动而言，缺乏专门性规范并不代表其当然的合法性，网络空间仍然需要在一般国际法规范的规制之下，以主权原则为代表的国际法基本原则能够在理论和实践层面发挥有效规制作用。

（一）理论依据

在理论层面，国际社会所关注的空间中既没有绝对的自由，也没有绝对的封闭，囊括海洋、空气空间、外层空间和网络在内的所有具有国际意义的空域应当共同遵守国际法基本原则。³⁵积极利用国际法基本原则对缺乏专门性规则的网络活动进行规制，在目前尚缺乏新规则的背景下，具有一定现实性和必要性。在基本法律价值尚不明晰的国家和国际社会，完全指望规则来解决纷争既缺乏规则上的可能性，也缺乏理念上的公正性。³⁶以主权原则为代表，其作为国际法上的基本原则，是国家基本权利和义务的基础。³⁷从国际法发展的历史来看，尽管纳入国际法规范的对象不断扩张，从领土到海洋到外层空间，但国际法基本原则具有一般性和普适性，其本质上是调整国际关系和规范国家权利的规范。网络空间的形成主要基于互联网这一现代通讯技术，而从法律角度来说，人类通过这一新技术从事的一些活动，与通过传统手段从事的类似活动并没有本质上的区别。³⁸以

University Press, 2012) 26.

³⁴ 【日】村瀬信也：《国际法原理》，秦一禾译，中国人民公安大学出版社 2012 年版，第 30-31 页

³⁵ 参见何志鹏、都青：《从自由到治理：海洋法对国际网络规则的启示》，载厦门大学学报（哲学社会科学版），2018 年第 1 期，第 16 页。

³⁶ 何志鹏：《国际法基本原则的迷失、动因和出路》，载当代法学 2017 年第 2 期，第 43 页。

³⁷ 参见梁西主编：《国际法》，曾令良修订主编，武汉大学出版社 2011 年版，第 84 页。

³⁸ 黄志雄：《国际法在网络空间的适用：秩序构建中的规则博弈》，载环球法律评论，2016 年第 03 期，第 8 页。

主权原则为代表的国际法基本原则，能够规制那些缺乏专门性规范的网络活动，关键在于如何对国际法基本原则进行解释。以网络间谍行为为例，应当以领土主权为基础，明确主权原则在规制网络间谍活动中的核心地位。这是由于，领土主权这一概念相较于主权原则更为明晰，国际法院在科孚海峡案中明确指出，“阿尔巴尼亚作为科孚海峡沿岸国，对其领海享有主权，其他国家不得未经许可进入其领海。”³⁹领土主权是国家对其领土享有排他性控制权的体现，考虑到网络间谍活动必然会涉及对一国领土范围内计算机系统的攻击，以领土这一概念可将存在于虚拟空间中的网络间谍活动固定在受害国领土之上，具象化主权原则的内涵能够更好发挥主权原则的规制作用。就不干涉原则而言，以英国为代表的国家积极推动“控制自由”标准来解释传统国际法上不干涉原则的“强制”要件，以弥合其主张“主权原则论”而可能导致的不周延现象，防止一些行为无法被纳入违法性的判定之中。这种方式本质上是考虑到网络空间的虚拟性，尽可能地丰富国际法基本原则在网络领域的内涵，使得国际法基本原则在面临违法性判断问题上能够落到实处。因此，运用理论手段对国际法基本原则进行解释，能够将没有专门化规则以规制的网络行动纳入一般国际法规范的框架之内，从而有效发挥国际法基本原则在网络空间中的规制作用。

（二）实践依据

在实践层面，以主权原则为代表的国际法基本原则适用于网络空间，已经被国际实践所证明。联合国信息安全政府专家组分别在2013年、2015年和2021年通过的共识性报告，一再确认了“国家主权和在主权基础上衍生的国际规范及原则适用于国家进行的信息通信技术活动，以及国家在其领土内对信息通信技术基础设施的管辖权”，标志着国际社会已就主权原则适用于网络空间达成基本共识。⁴⁰各国的立场文件中也都在不同程度上涉及到主权原则、不干涉原则、禁止使用武力原则等国际法基本原则如何适用于网络空间的主张。以网络间谍活动为例，尽管各国针对网络间谍这一问题的对外立场表达方式虽各有不同，但已有一些国家利用主权原则指出网络间谍活动的非法性。波兰在国家立场文件通过提出违反主权原则的可能表现，反对以“干扰国家机构的运作，包括阻碍公共实体的信息通信技术网络、服务或系统的正常运行，或窃取、删除或公开披露属于此类实体的数据”为形式的网络间谍活动。⁴¹哥斯达黎加在其立场文件中也指出，“监控行动可能以违反国家主权或国际法的其他规则的方式进行”。⁴²棱镜门事件曝光后，

³⁹ Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania), Judgment (Merits), I.C.J. Reports 1949.

⁴⁰ 黄志雄、孙芸芸：《网络主权原则的法理宣示与实践运用——再论网络间谍活动的国际法规制》，载《云南社会科学》，2021年第6期，第76页。

⁴¹ 波兰立场文件：<https://www.gov.pl/web/diplomacy/the-republic-of-polands-position-on-the-application-of-international-law-in-cyberspace>.

⁴² 哥斯达黎加立场文件：https://docs-library.unoda.org/Opened_Working_Group_on_Information_and_Communication_Technologies_-_2021/Costa_Rica_-_Position_Paper_-_International_Law_in_Cyberspace.pdf

巴西等国对美国采取的网络间谍活动提出了强烈谴责,认为其行为违反了国家主权原则。⁴³ 上述国家的立场表明,主权原则能够被国家灵活运用,并以此为法律基础论证网络间谍活动的非法性。在国内立法层面上,美国的《计算机欺诈和滥用法》和中国的《网络安全法》均明确禁止未经授权的网络入侵和数据窃取,主权国家通过行使国内立法权对网络间谍活动进行规制,本质上仍是主权原则发挥规制作用的体现。由此可见,尽管在实践中网络空间缺乏专门性立法,但各国能够通过立场表达和国内立法等方式,将国际法基本原则运用到网络行动的规制之中。

（三）应对路径

从传统领域到网络空间领域,国家主权原则等国际法基本原则能否发挥填补规则空白的作用,仍面临着不确定性争议。⁴⁴如何更好发挥以主权原则为代表的国际法基本原则的规制作用,以弥合网络空间国际法上的“规制真空”,仍然任重道远,但鉴于国际法基本原则在理论上的可适用性和已有国际实践,可充分利用联合国这一平台展开更广泛的讨论,并充分发挥国际法学者在规范解释和适用上的引领作用。

主权国家作为国际法产生与发展过程中的主要角色,应当在国际实践中丰富国际法基本原则的内涵。以主权原则为代表的国际法基本原则均源于《联合国宪章》,并得到国际社会的广泛认可。在国际组织深度参与国际社会治理的背景下,其成员国具有接受和服从国际组织参与相关国际法实施的义务。⁴⁵联合国作为国际上最具普遍性和权威性的多边平台,理应成为各国表达立场、协调规则的核心舞台。首先,各国应以联合国为平台,积极释放对国际法基本原则的立场主张,结合网络空间中的具体行动和风险挑战,丰富国际法基本原则的内涵。如果各国局限于单纯承认一般性的原则适用于网络空间,只会导致国际法基本原则在网络空间中被束之高阁,无法发挥其规制具体网络行动的示范性作用。其次,在国际法学界参与网络空间国际治理的背景下,学者们积极探讨现有国际法基本原则在网络空间这一新领域的适用,对于国际法基本原则在网络空间中的落地起到了引领性的效果。例如,“网络主权”概念的提出,强调国家对其网络空间的独立权、管辖权等,积极推动不干涉原则和禁止使用武力原则在网络空间中的转型,并产出了以《塔林手册 2.0 版》、法国白皮书为代表的诸多理论成果。以《塔林手册 2.0 版》为代表,其推崇的“规模与效果”标准被吸纳在众多国家的立场文件之中,极大可能在潜移默化中影响未来国际法规范的走向,以填补网络空间的“规则真

⁴³ R. Rolle, Lawyers to Act in N.S.A. Spy Row, <https://www.tribune242.com/news/2014/jun/05/lawyers-act-ns-spy-row>.

⁴⁴ 黄志雄、孙芸芸:《以合作主权观应对主权在网络空间适用的挑战》,载《边界与海洋研究》2023 年第 5 期,第 22 页。

⁴⁵ 参见饶戈平主编:《国际组织与国际法实施机制的发展》,北京大学出版社 2013 年版,第 21 页。

空”。⁴⁶学者厘清国际法基本原则适用于网络空间的必要性，并掌握其如何科学适用于网络空间的方法论，有助于政府在国际舞台上运用国际法理论更好表达自身立场主张，也能使以主权原则为代表的国际法基本原则在不偏离其固有涵义的基础上，在网络空间这一新疆域更好发挥对各国网络活动的规制作用。

五、结论

“荷花号”原则表明，一国的自由行事应在其领土范围内，一国未经允许不得在他国领土上行使权力，并只有当国际法出现“规制空白”时，一国的自由行事才合乎国际法。“荷花号”原则可发展为一项法律适用规则，判断一行为是否合法，应对国际法规范进行周延考察，先寻求具体规则，再审查包括国际法基本原则在内的相关国际法规范。网络间谍活动本身违反领土主权原则，国际法对此也并不存在“规制空白”，不能适用“荷花号”原则以证明网络间谍行为的合法性。

长期以来，国际社会通过不断的立法活动与实践创设了大量调整与规范国际关系的习惯法与成文法，从无到有地构建起以规范为主体结构的国际法律体系。⁴⁷网络间谍行为所引发的相应国际法规制的争议，根源在于现有国际法规范难以明确调整网络空间所涉的国家权利义务要素，主权原则作为国际法基本原则，由于其固有性质具有宽泛的特征，则尤为会出现解释的难题与纷争。对于以网络间谍活动为代表的网络行动，虽然现行国际法存在“规则真空”，但即使缺乏专门的法律规范，以主权原则为核心的国际法基本原则仍可适用于网络空间并发挥约束作用，应在实践中进一步充实国际法基本原则的内涵。

⁴⁶ 大多数国家（如巴西、爱尔兰、爱沙尼亚、澳大利亚、德国、非盟、荷兰、加拿大、挪威、欧盟、瑞士、新西兰等）支持以网络行动的规模 and 效果来判断是否构成“使用武力”。

⁴⁷ 张文显著：《全球治理与国际法》，法律出版社 2020 年版，第 73 页。

人工智能的军事化利用及其法律规制

张昕怡¹

摘要：当前人工智能技术已全方位融入军事各领域全流程，形态多样，呈现出人机共生发展倾向，为军事领域带来巨大机遇同时也带来必须正视的风险挑战。人工智能技术的军事应用突破了传统战争形态，战争边界愈发模糊，国际战略信任受到威胁，对国际法构成了严峻挑战。人工智能的军事化利用应遵守国际法，但现有国际法规则体系难以全面应对其滥用风险。针对该议题的规制成果呈现为碎片化样态，国际社会仅达成部分原则性共识，理念基础尚存在分歧，高层级的具有约束力的国际条约尚未形成且面临一定障碍。因此，应充分发挥联合国作用，构建标准与流程明确的外部审查机制，确立“有意义的人类控制”原则，建立国家责任与个人刑事责任相结合的多元责任体系以化解风险。

关键词：人工智能的军事化利用 人工智能技术 国际法 滥用风险 规则体系

近年来，人工智能技术不断取得突破，广泛应用于各领域，但其迅猛发展同时带来各国政治意图的介入，其军事化利用尤需警惕。人工智能越来越成为推动新一轮军事革命的核心驱动力，当前战争形态已由机械化、信息化转向智能化。人工智能技术上的优势为军事领域带来巨大机遇同时也带来必须正视的风险挑战。观俄乌冲突战场可知人工智能技术已融入军事系统各环节各领域，自主武器系统成为人工智能军事化利用的一个子项，当今人工智能的军事化利用早已呈现出武器化态势。国际社会虽已认识到其对国际法构成了严峻挑战，但对其如何与国际法规则体系相衔接讨论并不充分，技术治理与规则治理存在脱节。当前针对该议题的治理成果呈现出碎片化、表层化甚至阵营化，产出高层级的有约束力的国际条约对发挥人工智能军事应用的正面价值至关重要。但人工智能技术特性以及军事国防领域的特殊性决定了构建对人工智能军事利用有约束力的规则体系尚有较长的路要走。

一、人工智能的军事化利用及其对国际法的挑战

人工智能的军事化利用是指智能科学理论与技术在军事领域的应用和发展，以及催生的军事力量编程和作战理论的持续变革，²是将人工智能相关技术、平台与服务应用到军事领域，使其成为赋能军事行动的重要驱动力量，进而提升军事行动的效率、精准度和自主性。³实践表明，人工智能的军事化利用对传统战事

¹ 张昕怡，北京理工大学法学院 2024 级国际法方向硕士研究生。

² 廉颖婷、王淑梅：《人工智能军事应用诸多法律问题待解》（人工智能军事应用法律问题学术研讨会新闻稿），载《法制日报》2019 年 10 月 24 日第 7 版。

³ 朱启超：《人工智能武器化的发展趋势及治理策略》，载《人民论坛》2025 年 2 月 5 日第 2 版。

形态的改变已对国际法构成严峻挑战。

（一）人工智能军事化利用形态

与以往大多数新兴军事技术不同，人工智能技术并不专属于某个作战空间或领域，而是能够应用于各领域并提升武器装备性能的一种技术，甚至可以改变现有战争形态。⁴ 也就是说，人工智能的军事化利用演变方向已体现为武器化且不局限于武器化发展。在大数据和人工智能技术赋能下，传统作战流程将实现智能化再造，即由“态势感知—指挥决策—攻防协同—综合保障”向“全域态势智能认知—人机一体混合决策—有人/无人自主协同—主动按需精准保障”转变。⁵当前战场实践与各国科研投入均表明人工智能已融入军事各领域全流程，应用形态集中体现为如下四个方面：

1、智能情报分析

人工智能技术在情报收集的广度、情报分析的深度与速度方面均体现出人类与其他信息技术无法比拟的优势，也正是此种优势推动人机关系由“人在回路中”向“人在回路外”转变。在情报收集与分析速度上，人工智能的应用使数人数年的工作量得以在数秒内完成。⁶在情报收集的广度上，人工智能的应用可同步整合多端口来源信息，将各作战单元进行扁平化统筹，收集信息类型包括图文、语言、电磁信号等。具备自然语言处理功能（NLP）的语言模型可以实时翻译和解释多种语言，语言与文化差异已失去作为情报传递介质的独特优势。在情报分析的深度上，人工智能可对收集数据自动进行深度挖掘，挖掘出各类碎片化情报信息中的深层脉络与因果联系，将海量碎片化数据快速转变为可以利用的情报，从而提升情报分析质效。⁷美国在情报领域着力提升自动化处理与分析能力，通过 AI 赋能，推进了阿拉丁视频（Aladdin Video）、隐喻（Metaphor）、更好地从文本中提取增强检索（BETTER）等多个项目。⁸美国雷神技术公司（Raytheon Technologies Corporation）研制的 ISTAR（情报、监视、目标识别和跟踪）系统，涵盖了情报采集、监视、目标识别及跟踪功能，可汇聚来自卫星、舰船、飞机及地面站等多元信息源的数据，并对其进行深度分析与处理。当前人工智能技术的情报分析能力逐渐得到肯定，但对其输入数据可信度大多仍持保留意见。

2、智能辅助决策

当今智能化战场条件呈现出高对抗性、高动态性、高复杂性，在此过程中，

⁴ 李享、罗天宇：《人工智能军事应用及其国际法问题》，载《信息安全与通信保密》2021年第1期。

⁵ 朱启超：《人工智能武器化的发展趋势及治理策略》，载《人民论坛》2025年2月5日第2版。

⁶ 量子计算与人工智能的结合加速密码破译进程。谷歌公司宣称，其发布的 Willow 量子芯片能够在 5 min 内完成一次“标准的基准计算”，而传统超级计算机完成同样的任务需要“10 的 25 次方年”。

⁷ 朱启超：《人工智能武器化的发展趋势及治理策略》，载《人民论坛》2025年2月5日第2版。

⁸ 项目内容各有侧重，均旨在提高获取与分析情报能力，包括自动化视频提取、音频提取、知识展示和搜索、自动识别、分析语言文化中的隐喻、从文本中自动化提取和检索个性化的多语种语义等。参见王静雅、申华、沈彦：《智能化时代美国情报分析转型工作研究》，载《情报杂志》2022年第11期。

人工智能的大规模军事化利用既是因也是果。基于数据、样本、知识与模型的输入，人工智能技术逐渐成为战场态势综合感知、作战筹划、作战计划与推演评估的中坚力量，有由“人机交互”向“人机共生”发展趋势。早在 2016 年年底，美军便启动“指挥官虚拟参谋”（Commander's Virtual Staff, CVS）混合智能辅助决策系统。该系统能基于态势感知数据，识别和推理态势演变并形成未来态势图及作战方案，但作用只停留于信息提供阶段。随着人工智能技术的飞速迭代，该类系统作用已由辅助数据处理、辅助信息管理，发展到辅助决策分析，进而又发展到了辅助决策支持，对做出决策的影响持续加大。⁹该类系统技术着力于实现态势感知全维全时、信息联通精准高效、指挥控制智能自主，甚至可能承担未来战场的“数字指挥官”角色。美国防部已与 OpenAI 等企业合作开发了军事专用大模型，可解析复杂战场情报、在几秒内生成战术建议，甚至模拟敌方决策逻辑。俄罗斯“战斗指挥信息系统”、英国“THEIA 计划”和法国的“The Forge”数字决策支持引擎也在不断升级。在此发展趋势下，人机关系亟须得到谨慎对待，责任主体亟待明确。

3、智能武器装备

武器装备搭载人工智能技术已成为常态，甚至有演变为军备竞赛加剧倾向。根据人的参与程度不同，当前智能武器装备可分为两种形态，分别是智能增强武器装备与无人智能武器装备。智能增强武器装备系统仍坚持人类主导地位，AI 技术提供辅助以增强装备的感知分析、协调行动等方面能力。而无人智能武器装备系统则可全部或部分与人类脱节，以 AI 技术为主导，人类仅设定任务边界或紧急干预。俄罗斯研发了可搭载常规弹头或核弹头的“波塞冬”核动力无人水下航行器，以色列研制出完全自主武器“哈比”（Harby）巡飞弹，乌克兰成立了首支无人机作战部队可独立承担作战任务，美国柯林斯航空航天公司与雷神公司研发的无人机可在与操控员失联的情况下自主协同执行威胁探测和打击任务。武器装备的无人化自主程度、精确程度、集群化程度不断提升引发了国际社会的担忧，各界均开始讨论其对国际人道法及国际责任法的冲击及应对策略。其中，自主武器系统最具代表性及争议性。联合国早在 2014 年便在《特定常规武器公约》框架下开启了“致命性自主武器系统”问题的讨论进程。但随着近年来人工智能军事化利用技术与形式的更新，传统议题成果已难以应对新兴挑战，针对“人工智能的军事化利用”议题亟须得到系统性展开。

4、智能军事后勤保障

人工智能技术已与无人设备相结合融入军事物流链、联合投送链、战场救治链，促进保障体系信息化、智能化，多域联合作战保障的精准性与时效性得到极大提升。搭载人工智能技术的无人装备可在战场自动识别伤员并利用机器人进行

⁹ 参见杨露菁编：《作战辅助决策理论与方法》，国防工业出版社 2023 年版，第 13 页。

转移、救助，全域武器装备弹药情况可被实时搜集并通过无人设备进行投放补充。俄乌冲突战场已体现出新质后勤保障力量在人工智能技术赋能背景下能够得到有效统筹协调、联合与综合运用，人道主义精神也经由此种方式得到一定程度地践行。

（二）人工智能军事化利用影响

作为一把“双刃剑”，人工智能武器一方面以态势感知智能化、决策快速化、作战自动化大幅提升打击效能，另一方面导致外交斡旋余地减少，使冲突全面升级甚至核战争爆发成为可能。¹⁰当前，人工智能技术的深度融合对局势产生如下方面影响：

1、人工智能技术迭代突破传统战争形态

在实体战场中，人工智能技术通过战场实时监测、实时情报分析、轨迹追踪、目标监控等方式极大提升态势感知质量与效率，与武器相结合打造“无人高速率战场”，物理打击范围涵盖水陆空，¹¹打击精准度极高。¹²此外，人工智能技术的应用还加剧了虚拟空间危机。认知战、网络战、舆论战、电子战颠覆了传统战场形态。网络空间、电磁空间与认知空间作战相互交织，军队应用人工智能通过网络、电磁空间制造错误、虚假信息、信号欺骗敌军变得可能。

当前战争已不是传统物理层面力量对抗，而是科技与战略的全方位对抗。人工智能技术的军事化应用或将改变现有战略力量的平衡，削弱核国家核威慑的能力，增加危机意外升级的可能性，鼓励军备升级和军备竞赛，冲击和挑战以核威慑力量为基础的战略稳定，进而动摇甚至可能破坏现有全球战略稳定的根基。¹³在当今多极地缘政治秩序与多极核环境下，人工智能技术的跨数量级突破愈发放大其潜在不稳定性。美国兰德公司认为，人工智能的军事应用会削弱构成核战略理论的主要根据，迫使依赖报复打击的国家不得不面临要么提前发动进攻，要么输掉战争的困境。¹⁴人工智能军事利用实力相当的交战主体之间的较量愈发演变为人工智能技术军事化利用水平的较量，而技术搭载水平悬殊国家间的冲突则加剧战争的不对称性。为应对此种不确定性，各国竞相发展军事领域的人工智能技术水平，部分原本在军事部署与作战策略领域采取防御型的国家也以此为“契机”转为进攻性，国际紧张局势升级，战争的暴力性加剧。

2、人工智能技术模糊战争边界

¹⁰ 刘国柱、尹楠楠：《美国国家安全认知的新视阈：人工智能与国家安全》，载《国际安全研究》2020年第2期。

¹¹ 美国海军已经研发出多款水下无人潜航器，如“刀鱼”和“蝠鲼”，并投入使用。它们能够自主完成侦察、电子战以及打击等多种任务，超大无人潜艇“虎鲸”能够在数月内自主航行。

¹² 2025年1月，乌克兰国防情报总局对外宣布，一艘海上无人艇发射的对空导弹，击落了一架俄罗斯军用直升机，这是现代战争史上首次有直升机被无人艇用导弹击落的实例。这意味着无人艇在与直升机之间的对抗中再度攻守异位。

¹³ 张东冬：《人工智能军事化与全球战略稳定》，载《国际展望》2022年第5期。

¹⁴ 转引自朱荣生：《【CISS研究报告】人工智能军事化的挑战及中美合作应对探讨》，载清华大学战略与安全研究中心，<https://ciss.tsinghua.edu.cn/info/OpinionsandInterviews/5608>。

智能化战争无论是在时间进程、空间进程，还是宏观领域、微观领域方面，都高度错综复杂，难以形成清晰明了的进度条。当前，人工智能的军事应用覆盖实体战场与虚拟战场，“多元战场”并行。新型技术空间和传统物理空间紧密融合，战争对抗由传统空间向“陆、海、空、天、电、网、心”多维一体拓展；作战域由单一物理域向“物理域、信息域、认知域、社会域”深度融合；制权争夺的重心由“信息、海洋、天空”向“智能、太空、网络”转移，形成“耦合紧密、互联互通”的一体化全域多维战场。以人工智能技术为支撑的现代智能化战争使得战争的边界被无限延展甚至有“诱导作战”情况发生，¹⁵战争的时间边界难以再清晰划分，战争进程模糊。此外，人工智能在军事领域的高强度多角度深度运行还模糊了战军兵种职能、杀伤效能等方面的边界。¹⁶

随着人工智能技术在军事领域应用愈发成熟，人工智能技术与武器的边界愈发模糊，人工智能技术内嵌于战争全流程，内嵌于各式军事武器。此种紧密结合性与算法“黑箱”使得难以明确认定人工智能技术何时开始发挥作用以及在何种程度上发挥作用，进一步增加了处理人机关系的难度，同时构成了对国际责任法的挑战。规则的缺漏加剧了人工智能的军事滥用风险。智能化技术手段愈发展，战争的潜能就愈被无限释放，原有战争边界就不断处于“打破—重构—再打破—再重构”的循环过程，而这就是“模糊”背后的本质。¹⁷

3、人工智能技术加剧信任危机

人工智能的军事利用使人工智能技术与国家政治利益相挂钩，某些掌握技术优势国家人为制造技术壁垒，泛化国家安全概念，争夺技术霸权，试图在战场上与战场外始终保持领导地位。此种政治目标的介入扩大了人工智能军事利用的潜在风险，科技力量在此背景下反而成为“潘多拉魔盒”。

部署军事人工智能可能会重塑不同国家之间以及国家与各类非国家行为体之间的力量平衡。数字鸿沟不仅没有成为国际合作的基础，反而成为霸权主义国家以“技术合作”为旗帜“胁迫”他国加入合作体系打造军事阵营的“筹码”。当前国际实践反映出人工智能的军事利用阵营化趋势主要通过对前端技术的控制实现，包括人工智能技术的软件核心即数据以及硬件核心半导体材料。2024年10月21日，美国司法部（DOJ）发布了一份拟议规则制定通知（NPRM）——有关防止受关注国家或涵盖人员访问美国敏感个人数据和政府相关数据的规定。该规定再一次将中国列为“外国敌手”，实行在经贸领域的数字单边保护主义。美国还在数字经济国际规则、网络空间治理等方面加大对东盟的引导，并联合东盟排查关键技术过度依赖中国的“风险”，严格对中国相关技术项目的投资审查，挤压中国数

¹⁵ 2025年2月，美国DARPA发布“金苹果”（Kallisti）项目招标书，寻求新的方法确保国家安全决策者能够优化威慑或引导对手行动的战略。

¹⁶ 参见湛力、武智：《透视智能化战争的“模糊性”》，载《解放军报》2025年1月14日。

¹⁷ 湛力、武智：《透视智能化战争的“模糊性”》，载《解放军报》2025年1月14日。

字空间。¹⁸人工智能技术成为霸权国家组建政治格局的工具，此种对技术前端的限制使紧张局势升级，加剧了信任危机。

（三）人工智能军事化利用对国际法构成的严峻挑战

首先，人工智能的军事利用对国际责任法构成挑战。人工智能技术特性造成因果关系与主观过错认定困难，从而导致责任认定存在困难。在战争进程中，人工智能技术尚未改变军事系统物理形态且军事系统可根据态势在 AI 驱动与非 AI 驱动模式进行切换，人工智能技术作为多领域科技全方位全链条融入军事进程，难以将人工智能技术与传统因素相剥离认定何种因素造成损害。¹⁹此外，人工智能技术难以规避算法“黑箱”，其难以回溯性与难以解释性也对因果关系判断抛出难题。

除了人工智能技术的算法特性，人工智能的军事利用对国际责任法构成挑战的另一原因是其参与主体身份多样，国家责任与个人刑事责任的适用难以匹配当前现状。技术研发人员个人违反国际人道法的战争罪行的刑事责任的实现面临严峻挑战。此外，主权国家已不是战争中的唯一主体。科技公司与国防部门的合作日益紧密，而商业公司在人工智能技术研发—配置—部署—军事利用的全链条中具有极大的自主性。除了研发主体众多，企图攻击人工智能技术系统的主体同样众多，人工智能的军事应用可能成为助长恐怖主义的工具。而这些对传统责任认定提出的难题尚未得到与技术进步相匹敌的规则回应，规则体系中责任板块的漏洞或将成为人工智能技术在军事领域滥用的“诱饵”。

其次，人工智能的军事利用对国际人道法构成挑战。智能化武器缺乏同理心与人道主义精神，难以作出与人类情感相结合的判断，同时还存在技术漏洞、误判攻击目标风险，可能直接或间接打击民用设施、误伤平民。2023 年，美军的一款人工智能将渔船甲板上的阴影当作导弹发射管，将渔船误判成了军舰。巴以冲突中，人工智能技术被用来肆意攻击敌国网络、电力等系统，医院、银行等几近瘫痪，区分原则受到严峻挑战。除了智能化武器自身的道德缺位，人工智能技术也影响了人类的价值判断。尽管当前尚未出现真正意义上的强人工智能，非完全自主武器系统的决策权仍在人类手中，但人工智能技术的应用对人类决策行为产生极强的引导性，人类作战思维被重塑。人工智能技术提供的数据信息与决策指引降低了指挥官及士兵出于人道主义的判断考量，人工智能技术压缩决策时间使人类的“即时反应”不可避免地受其影响。

此外，人工智能技术对武器管制及正常国际贸易构成威胁。人工智能的出现使各国开始积极探索军民两用的转化，²⁰军用民用界限在战争情境下由于技术的

¹⁸ 陈小宁：《美国发展数字合作，挤压中国空间》，载《世界知识》2022 年第 9 期。

¹⁹ Centre for International Governance Innovation, *International Norms Development and AI in the Military Domain*, The Hague Centre for Strategic Studies 2024.

²⁰ 2021 年，北约创新基金（NIF）启动，旨在投资有助于国防和安全的军民两用技术，与北大西洋国防创新加速器一脉相承。2024 年，北约更新了其《人工智能战略》，同样强调加快技术转化。

“隐蔽”而被模糊。为民用目的开发的人工智能可能被滥用，包括用于政治假情报、网络攻击、恐怖主义或其他恶意目的，这对两用技术和国际治理构成重大风险。

²¹由于人工智能技术的军民两用性质，可以作为插件存在的人工智能模块尚未受到条约的重视，某些人工智能技术甚至能够作为民用品出口，完全规避《武器贸易条约》对最终用途的约束，这有可能进一步强化相关武器的交易与滥用，导致此类武器或技术向恐怖主义或犯罪组织非法扩散。²²俄乌冲突伊始，民用改装无人机的大规模使用加剧战争紧张态势，人工智能技术的搭载使得民用物体在战场摇身转变为军事武器，此种军民用途的肆意切换同样对国际人道法构成挑战，也为国际贸易营造出紧张气氛。

二、人工智能军事化利用的现有法律规制及其面临的挑战

尽管国际社会已对“人工智能的军事化利用”议题予以高度重视，但针对该议题的具体国际条约尚未形成，高层级“硬法”缺失，已有成果呈现为“碎片化、表层化”样态。2024年7月，联合国秘书长所作“当前的科学技术发展及其可能对国际安全和裁军努力造成的影响”的报告指出“虽然有一些外部倡议，但目前在联合国主持下没有一个政府间进程来解决人工智能在军事领域的负责任的生命周期问题”。²³尽管该结论是通过整理自主武器系统议题发展而得出的，但人工智能的军事利用作为自主武器系统的上位且出现更晚的概念，该结论同样揭示了人工智能的军事化利用规制现状的关键问题。

（一）现有规制共识

人工智能技术虽已贯穿战争全流程，其附属性身份尚未被突破，仍需与实体工具相结合服务于实体战场，故人工智能的军事利用无疑需遵守适用于传统战争的国际法，包括《联合国宪章》、国际人道法等。对此，国际社会已形成普遍共识。

2024年9月，荷兰海牙战略研究中心（HCSS）和加拿大国际治理创新中心（CIGI）联合发布《军事领域的人工智能与国际规范制定》报告（下文简称海牙报告），通过对已有成果进行统计，该报告指出当前针对人工智能的军事利用议题的探讨共识集中在七个方面，分别是遵守国际法、可归责、可解释性与可回溯性、应对算法偏见、可信任、可控与人机交互问题，²⁴与北约人工智能战略中的六点原则高度契合。也就是说，国际社会已经对人工智能的军事利用潜在问题集中在法律领域、技术领域与伦理领域内部与这三个领域的交叉有充分认识，其中应首要遵守国际法的共识基础已经达成。

2024年10月，联合国发布了《军事领域的人工智能及其对国际和平与安全

²¹ A/79/224.

²² 朱启超：《人工智能武器化的发展趋势及治理策略》，载《人民论坛》2025年2月5日第2版。

²³ A/79/224.

²⁴ 参见 Centre for International Governance Innovation, *International Norms Development and AI in the Military Domain*, The Hague Centre for Strategic Studies 2024.

的影响》决议草案。²⁵该草案是首份完全聚焦于人工智能的军事利用的联合国成果，申明国际法，包括《联合国宪章》、国际人道法和国际人权法，适用于人工智能能力及其在军事领域启用的系统的整个生命周期中发生的受其管辖的事项，并决定在大会第八十届会议临时议程题为“全面彻底裁军”的项目下列入题为“军事领域的人工智能及其对国际和平与安全的影响”的分项。

此外，作为比“人工智能的军事化利用”议题发展更为成熟的下位议题，关于“致命性自主武器系统”的规制成果具有借鉴意义。自 2014 年起，联合国《特定常规武器公约》（CCW）框架下的政府专家组开始讨论“致命性自主武器系统”（LAWS），聚焦于自主武器的伦理和法律风险。2019 年《特定常规武器公约》缔约国会议一致通过了 11 项指导原则，明确了国际人道法与国际责任法的适用。²⁶该项成果的发布代表了针对致命性自主武器系统的高层级规制框架及共识已经形成。值得注意的是，从该 11 项指导原则的表述出发可以发现其并不局限于致命性自主武器系统。其中第一项原则指出国际人道法继续完全适用于所有武器系统；第四、第五项原则均使用“新武器”表述国际法的适用及国家的义务；第七项原则使用“任何武器系统新技术”表述风险评估的必要性。因此，该规制成果对人工智能的军事化利用议题同样具有适用可行性。

（二）现有国际法规则的适用

人工智能的军事化利用在作战手段和方法领域提出了需要应对的新问题。对此，除了在上述《特定常规武器公约》缔约国会议通过的 11 项指导原则有所涉及，国际法中还存在两个条款具有宏观规范意义。其一是 1949 年通过的《日内瓦公约第一附加议定书》第三十六条，该条规定“在研究、发展、取得或采用新的武器、作战手段或方法时，缔约一方有义务断定，在某些或所有情况下，该新的武器、作战手段或方法的使用是否为本议定书或适用于该缔约一方的任何其他国际法规则所禁止。”不同于倡导性口号，该条款可谓明确的义务来源，规定了对新式武器的审查义务，为国家责任的适用提供了依据。其二是著名的马尔顿斯条款，内容为“在颁布更完整的战争法规之前，缔约各国认为有必要声明，凡属他们通过的规章中所没有包括的情况，居民和交战者仍应受国际法原则的保护和管辖，因为这些原则是来源于文明国家间制定的惯例、人道法规和公众良知的要求”。该条款为新式作战手段和方法设定了框架，但原则框架性约束既是其适用优势同时也是难以准确适用的劣势所在。

海牙报告通过统计分析指出人工智能的军事利用应遵守国际法在国际上已达成广泛共识，但对于现有国际法是否完备尚未形成一致意见。因此，梳理现有相关国际法规则并考虑人工智能的军事化利用与其匹配度至关重要。

²⁵ A/C.1/79/L.43.

²⁶ 李驰江：《人工智能在军事领域的应用及全球治理》，载《人民论坛·学术前沿》2021 年第 10 期。

现行国际法对作战手段和方法均有相应成熟的国际法律规制。国际人道法限制两大类作战手段，分别是使用大规模杀伤性武器和使用野蛮及残酷的常规武器。1977 年通过的联大决议重申大规模杀伤性武器定义为原子爆炸武器、放射性物质武器、致命化学和生物武器，以及任何今后发展的、与原子弹或其他上述提到武器在毁灭性上类似的武器。²⁷同时认为由于存在处理具体类型的大规模毁灭性武器条约，便无必要单独制定一个处理该大类武器的普遍性条约。在此存在两个问题。首先，人工智能军事化利用当前面临最大的风险是核扩散风险，而现有相关规则并不充分，难以应对滥用风险。相比较而言，当前规制化学和生物武器的条约成熟且相对完善，而对于核武器的规制尚存在分歧，仅有 1968 年的《不扩散核武器条约》以及 1996 年的《全面禁止核试验条约》（当前尚未生效）这两个多边条约及有限数量的双边条约。其次，从大规模杀伤性武器的定义可推知其内涵是可适应实践发展的，而现有条约规制对象是明确具体的而难以将新型类似武器囊括其中。

相比之下，有关常规武器的规则更能适应人工智能军事化利用新兴发展但仍难以全面匹配。有关常规武器的规则主要体现在《禁止或限制使用某些可被认为具有过分伤害力或不分皂白效果的常规武器的公约》（《常规武器公约》）及其议定书²⁸中。该套规则体系既涉及传统武器也涉及激光等新型武器，并以效果作为判断标准，其核心在于国际人道法的基本原则。但该套规则侧重于对过分伤害加以判断，而当前人工智能军事化利用主要对“不分皂白”构成挑战，两者重点存在差池。以“激光”为例，现有规则规制激光致盲武器，而当前激光滥用风险在于对打击目标的误判且通常不以人为目标。对此，红十字国际委员会早在 1987 年便提出相关预测，即远距离遥控武器或者通过遥控器控制的战场武器，把战场变得自动化……而针对这一演进发展起来的反措施——特别是电子屏蔽（或干扰）——进一步恶化了冲突不分皂白的程度。²⁹

国际人道法限制的作战方法包括不分青红皂白的攻击方法、背信弃义的作战方法和改变环境的作战方法，人工智能在军事领域的应用主要对前两项作战方法构成挑战。《第一附加议定书》第五十一条对“不分皂白的攻击”的含义做了概括，规制无区别地打击军事目标和平民或民用物体的限制并列举“如果将平民或民用物体集中的城镇、乡村或其他地区内许多分散而独立的军事目标视为单一的军事目标而进行炮轰，那么这种攻击就是不分皂白的”。人工智能的应用使作战手段和方式的精准度不断提升，但此种特点在其失控或错误条件下便成为对国际人道法最大的挑战。此外，当前人工智能的应用可使部分常见民用物品随时转变为武

²⁷ 联合国大会 32/84B 号决议。

²⁸ 《第一议定书》关于无法检测的碎片，《第二议定书》关于地雷、饵雷和其他装置，《第三议定书》关于燃烧性武器，《第四议定书》关于激光致盲武器，《第五议定书》关于战争遗留爆炸物。

²⁹ Commentaries on APs, s.1476.

器形态，颠覆作战员对“武器”的传统认知与信任，有被解释为背信弃义的可能性。

综上所述，现有对作战手段和方法加以限制的具体规则与人工智能军事化利用风险规制重点存在错位，或是重点不同或是内涵有待更新。国际人道法基本原则、《日内瓦公约第一附加议定书》第三十六条与马尔顿斯条款存在适应人工智能军事化利用发展的空间，但仍需向产出具体的规则方向努力。

（三）达成规制共识面临的挑战

对现有成果进行梳理可以发现规制人工智能军事化利用的原则性共识已经达成，即人工智能的军事化利用完全受国际法规制，但对遵守国际法的强调仅停留在应然层面而难以进入实然层面，产出针对该议题的多边条约仍存在较大障碍。导致目前困境的核心原因在于规制理念与规制重点的分散，达成共识的理念基础尚未达成一致。

美国与欧盟对人工智能军事化利用议题的体系建设处于领先水平，在世界范围内具有示范作用，但二者规制重点全然不同。欧盟侧重于“标准化”，而美国侧重于“技术化”。与人工智能技术规则规制思路相一致，欧盟对“人工智能的军事利用”议题同样主张划分层次，根据风险等级进行分层级治理，致力于首先统一AI、AI 的军事应用定义。与欧盟思路相仿，OECD 于 2024 年发布《人工智能（AI）事件及相关术语定义》报告，并于 2025 年发布了《AI 事件报告共同框架》，提出了一个人工智能（AI）事件报告的通用框架，旨在为各司法管辖区和行业的利益相关者提供一个全球基准。³⁰上述做法均旨在为人工智能在军事领域的应用划定理论规则框架。而纵览美国人工智能军事化利用规范体系可知，美国规则战略的设定均旨在为技术发展创造便利条件，同时引导技术发展克服技术弊端特性，既重视战略牵引又重视配套法规的可操作性，³¹旨在为人工智能在军事领域的应用划定技术规则框架。规制重点的不同可能导致在紧急状态下作出不同的价值取舍，价值排序的不同导致理念基础的分化。

在规制理念上，美国与北约以“竞赢”为目标制定并输出战略规则，通过结盟、达成战略合作等方式为其目标服务。³²而中国以“人类命运共同体”理念为主线，主张共商共建共享。美国于 2019 年发布《保持美国在人工智能领域的领导地位》行政令，又于 2025 年发布《消除美国在人工智能领域领导地位的障碍》行政命

³⁰ OECD publishing, *Towards a Common Reporting Framework for AI Incidents*, OECD Artificial Intelligence Papers, February 2025, No. 34.

³¹ 参见司孟韩、龙坤、徐能武：《美国军事人工智能战略的调整：动因、影响与启示》，载《情报杂志》2024 年第 12 期。将 2023 年美国国防部发布的《数据、分析与人工智能采用战略》（*Data, Analytics, and Artificial Intelligence Adoption Strategy*）及采取的配套措施与前期的《国防部人工智能战略概要（2018）》（*Summary of the 2018 Department of Defense Artificial Intelligence Strategy*）、《国防部数字现代化战略》（*DoD Digital Modernization Strategy*）、《国防部数据战略》（*DoD Data Strategy*）等战略文件进行对比总结出其当前规制重点。

³² 参见孙成昊、王叶涓：《北约人工智能战略：内涵、动因与挑战》，载《国际论坛》2022 年第 5 期；*Summary of NATO's revised Artificial Intelligence (AI) strategy*, https://www.nato.int/cps/en/natohq/official_texts_227237.htm。

令,明确指出美国在人工智能领域的持续领导对维护美国经济和国家安全并以符合美国价值观、政策和优先事项的方式塑造人工智能的全球演变至关重要,³³公开表明其政治野心,将技术发展放在第一位。当技术与规则发生冲突时,规则需为技术让步。³⁴同美国在外空领域发展思路相一致,除了战略规则,美国大力发展基础设施并不断推进技术演变,提高自身防护能力同时意图通过技术引领掌握更大话语权。³⁵在规制理念上,北约虽受美国竞赢思维影响,但仍具备自身特点。北约人工智能新战略并不寻求改变盟国业已采用的标准,也不寻求影响、取代现有的国家或国际义务和承诺,而是通过六项原则的设定,划定盟国开发应用人工智能战略的“基线标准”,为之后更高层次的标准化合作奠定原则性基础。³⁶中国一贯认为人工智能发展和军事应用问题与全人类命运息息相关,需要各国团结应对以推动人工智能技术和平利用,相继推出《中国关于规范人工智能军事应用的立场文件》《全球安全倡议》《中国关于加强人工智能伦理治理的立场文件》《全球人工智能治理倡议》等系列文件,致力于促进人工智能领域的国际交流与合作,第 78 届联合国大会通过了中国主提的加强人工智能能力建设国际合作决议。³⁷可以看出,上述北约对盟国内部采取的措施理念与中国主张的共同体理念存在部分相似,但总体规制理念的分歧同样带来理念基础的分化。

理念基础的分化亟须联合国平台作用的发挥,国际社会针对该议题进行的自发研讨已遇瓶颈。2023 年,首届“军事领域负责任人工智能峰会”在荷兰海牙召开,包括中国、美国在内的 60 多个国家签署了一项“行动倡议”(REALM 2023 Call To Action),支持“军事领域负责任使用人工智能”的立场。³⁸同年,28 国在首届人工智能安全全球峰会签署《布莱奇利宣言》以就前沿人工智能带来的机遇和风险建立共识。2024 年,第二届“军事领域负责任人工智能峰会”在韩国首尔召开,但该会议成果“行动蓝图”难以同首届峰会一样达成广泛共识,原因在于“行动蓝图”涉及更为具体的规则,超越原则倡议层级。

联合国于 2023 年发布的政府简报《新和平纲领》中便强调应警惕“将新技术和新兴技术武器化的危险,必须确保为消除新技术和新兴技术武器化的危险所采取的步骤,不限制全球南方国家获得此类技术承诺为促进可持续发展目标带来的巨大惠益”;³⁹安理会举行了关于“人工智能:国际和平与安全的机遇与风险”的会

³³ Executive Order on Maintaining American Leadership in Artificial Intelligence.

³⁴ Executive Order on Removing Barriers to American Leadership in Artificial Intelligence.

³⁵ 美国于 2020 年发布《关键与新兴技术国家战略》,其中明确指出维护美国在全球关键和新兴技术领域领导地位。此外,美国国防部于 2023 年发布的《数据、分析与人工智能采用战略》提出人工智能需求层次“金字塔”,并明确将中国视为唯一战略竞争对手。

³⁶ 孙成昊、王叶滔:《北约人工智能战略:内涵、动因与挑战》,载《国际论坛》2022 年第 5 期。

³⁷ 司孟韩、龙坤、徐能武:《美国军事人工智能战略的调整:动因、影响与启示》,载《情报杂志》2024 年第 12 期。

³⁸ 参见 <https://www.government.nl/documents/publications/2023/02/16/realim-2023-call-to-action>;
<https://www.government.nl/ministries/ministry-of-foreign-affairs/activiteiten/realim>。

³⁹ A/77/CRP.1/Add.8. 该报告背景为在 2021 年联合国秘书长所作“我们的共同议程”的报告(A/75/982)所载提议的基础上,提出着重于行动的建议。

议。同年，联合国秘书长安东尼·古特雷斯宣布成立人工智能高级别咨询机构以使人工智能造福全人类并避免风险，该机构发布了《为人类治理人工智能》最终报告，首次为全球人工治理提出联合国框架内的具体建议，提出了 7 项其认为“敏捷、适应性强且有效”的措施。⁴⁰2024 年，联合国未来峰会通过《全球数字契约》，提出“加强人工智能国际治理，造福人类”的目标，授权联合国设立人工智能国际科学小组和人工智能治理全球对话等机制。2021—2024 年，联大一委三次高票通过“和平利用科技”相关决议，聚焦技术出口管制。上述成果均表明联合国虽已关注人工智能的军事化利用议题，但焦点尚集中于技术层面，对该议题与国际法规则衔接的关注亟待加强。

三、应对人工智能军事化利用潜在风险的国际法律制度的完善

当前人工智能的军事利用尚停留在概念性名词阶段，内涵与外延均不明确。因此，对人工智能军事利用的规制要统筹变量与不变量，选定一基点进行范围辐射，明确其定义。其中，人工智能技术发展日新月异，形式多样，宜定为变量；军事利用为判断基点，宜定为不变量；辐射范围涵盖海陆空实体战场与网络空间虚拟战场。当前技术规则与法律规则的衔接在人工智能军事化利用领域并不顺畅。因此，应加强利用法律手段为人工智能技术设置边界，价值伦理与技术规则并重，不以军事利用为理由放纵发展人工智能技术，规制重点由后端移至前端并贯彻全链条来填充国际规则框架。

（一）构建标准与流程明确的外部审查机制

将人工智能的军事利用有效限制在国际人道法范畴内需要设定人工智能军事利用技术红线，搭建标准与流程明确的外部审查机制。国际社会对人工智能军事利用的技术有更为充分的讨论且对其潜在风险有一致共识，故建立体系化的审查机制对技术加强审查具备理论与实践的可行性。

《日内瓦公约第一附加议定书》第三十六条要求缔约国对新武器、作战手段或方法进行法律审查。具体到该议题下，应先明确人工智能军事化利用的内涵与外延，梳理涉及技术种类、技术功能以及应用此类技术会造成的可预见性后果。明确上列事项后再对应构建审查机制。审查范围应包括软件（模型算法、数据、系统运行）、硬件（人工智能技术搭载武器）与参与人工智能军事利用的非国家主体资格及行为三方面；审查标准为国际人道法既有规则及原则。因此，模型算法与系统运行预先设定不得与人类共有伦理价值相悖，输入数据应注重可靠性与可信性，人工智能技术搭载武器不得突破国际人道法对作战手段和方法的限制，非国家行为体不得擅自嵌入有违正义考量算法。国际社会应寻求变量中的“底线”，

⁴⁰ 该七项举措为：设立国际科学顾问委员会以提供有关人工智能的公正且可靠的科学知识，开展政治对话以形成保障人权的监管共识，建立国际人工智能以确保人工智能系统的跨境技术互操作性，搭建全球网络建立人才支撑，设立全球基金负责解决能力和协作差距，搭建全球数据框架确保人工智能系统的透明度和问责制，在联合国秘书处内设立一个小型人工智能办公室以支持和协调上述措施的实施。

将价值伦理渗透入全流程规则体系，技术规则体系与伦理价值并重，通过法律手段的调和兼顾规则边界与技术红线，对《日内瓦公约第一附加议定书》第三十六条进行配套细化与落实。

（二）确立“有意义的人类控制”原则

人机关系是人工智能在各领域应用的重点，在军事领域尤需保有人类控制权。基于人工智能算法的难以解释性与难以预测性，人类应参与 OODA 循环⁴¹全链条，人工智能系统驱动与非人工智能系统驱动的可转化性应作为“应急按钮”予以保留。人工智能技术的大规模应用使国家为了获得军事行动中的时间优势，或者确保自身有足够的“快速反应”能力，而越发依赖于技术“操作”弥补人类即时反应的不足，其结果是危机以超脱人类反应的速度螺旋上升。⁴²在人工智能军事化利用技术飞速迭代背景下，人机关系的变化可能突破 OODA 循环而发展成为 OODA 螺旋。因此，“有意义的人类控制”不可单单保留人类形式上的控制权，而应关注人类实质控制权的保留，即发挥人类思维独有性的真实人类行为而非基于人工智能技术逻辑惯性作出的“被动受控”行为。

此外，除了技术层面，确立“有意义的人类控制”原则还与责任主体的确定息息相关。首先，该原则内涵表明责任不能转嫁给机器，系统不具备承担责任的能力。其次，人类控制员与系统运行各环节的明确匹配有利于因果关系的判断，从而确定是国家还是个人承担责任。如若个人承担责任，可以确定是系统研发者还是指挥官、战斗员、交易方承担责任。因此，“有意义的人类控制”原则的确立与国际责任法在人工智能军事化领域的细化同样密不可分。

（三）建立国家责任与个人责任相结合的多元责任体系

应对人工智能的军事利用对国际责任法的冲击，该议题下责任机制亟须明确，完善包括国家责任与个人责任的多元责任体系。

首先，国家应为其造成损害后果的作为或不作为负责，也应对人工智能系统运行行为负责。其中，作为包括国家故意研发、投放、使用有违国际人道法的人工智能技术系统及搭载人工智能技术的武器系统；不作为包括国家对人工智能军事利用链条中的主体及系统的审查监管失职；而基于系统运行的难以预测性与系统的工具属性，国家应对系统运行行为承担责任以避免“责任洼地”，技术算法系统不具备承担责任的资格。⁴³当前，北约已明确人工智能出现错误或造成损害属于人为责任。

除了国家，该链条中众多非国家主体对结果影响重大，包括技术研发人员、指挥官、战斗员、非法交易方等，其承担责任情形同样亟待明确。但比对战争法

⁴¹ OODA 循环：由美国空军上校约翰·博依德（John Boyd）提出，主要用于描述在武装冲突中，敌对双方如何通过“观察—判断—决策—行动”这一循环程序来进行较量。

⁴² 转引自朱荣生：《【CISS 研究报告】人工智能军事化的挑战及中美合作应对探讨》，载清华大学战略与安全研究中心，<https://ciss.tsinghua.edu.cn/info/OpinionsandInterviews/5608>。

⁴³ 当前，北约人工智能战略概要文件已明确人工智能出现错误或造成损害属于人为责任的原则。

构成要件，该议题下的个人“故意”认定存在障碍且同样由于人工智能技术特性，“故意”的限定难以有效应对人工智能军事利用的潜在风险。因此，当务之急不仅是要明确责任，更需探索与技术进步相匹配的责任体系完善方向，可以将结果与因果关系判断两方面相结合进行归责，明确主观主义与客观主义的适用情形及判断步骤。

（四）构建联合国框架下规范人工智能军事化利用的体系化机制

2023年12月，联合国人工智能高级别咨询机构发布《为人类治理人工智能》的临时报告，该报告指出虽然已有数百份关于AI治理的指南、框架和原则被各国政府、公司、联盟和国际组织采纳，但没有一个具有真正的全球影响力，这导致AI治理始终存在代表性、协调性和执行力方面的问题。当前在国际层面对“人工智能的军事化利用”这一独立议题进行深入研究已势不可挡，但人工智能的军事利用治理平台与治理规则呈现出极强的碎片化表层化，甚至出现“阵营化”趋势，故应在联合国框架下逐步构建针对该议题体系完善的治理机制。

第一，发挥联合国大会及联大裁军事务厅的影响力，推动人工智能的军事利用议题系统深入展开。2021年，联大一委首次通过“和平利用科技”决议。2024年，联大一委第三次高票通过同名决议，决议强调和平利用科技及相关国际合作对经济社会发展的重要性，敦促各国在履行防扩散等国际义务、维护国家安全利益的同时，取消对发展中国家和平利用科技的不合理限制，平衡推进防扩散出口管制目标与各国和平利用科技并开展相关国际合作的权利。中国裁军大使沈健强调各国尤其是大国对在军事领域研发和使用人工智能技术应该采取慎重负责的态度，不谋求绝对军事优势，避免损害全球战略平衡与稳定。在联合国框架下，对“自主武器系统”的讨论也更加深入并不断取得阶段性进展，展现了联合国平台的独特优势。但当前人工智能的军事利用这一上位议题大多是在人工智能技术议题中作为附带性表述论点或是作为独立议题仅展开表层性探讨，尚未进入独有研究轨道。因此，应发挥联合国大会及联大一委的特有优势作用，梳理现有架构，明确各职能部门分工，加强各机构互动，对人工智能的军事化利用这一议题进行深入系统讨论并推动形成原则与规则并具的规制体系，力求形成“软硬结合”的框架治理机制，促进形成具有广泛共识的全球人工智能军事利用治理框架和标准规范，推动构建人工智能时代的人类安全共同体。⁴⁴

第二，加强建设安理会的监督与核查机制。虽然《日内瓦公约第一议定书》明确规定缔约国需对新武器、作战手段或方法进行法律审查，但只有英美等少数国家真正践行。要规避人工智能军事利用的潜在风险，对人工智能技术、技术搭载军事系统、商业研发主体、技术应用操作主体、应用行为及效果进行外部监督

⁴⁴ 司孟韩、龙坤、徐能武：《美国军事人工智能战略的调整：动因、影响与启示》，载《情报杂志》2024年第12期。

与核查至关重要。此外，在恐怖主义泛滥的当下还应对武器贸易加强审查。尽管军事国防领域具有特殊性，但对人工智能技术“潘多拉魔盒”效应的畏惧是人类共识，故安理会应逐步建立针对人工智能军事利用的可行监督机制，防范人工智能军事化滥用及向恐怖主义扩散风险。

第三，发挥国际法院及国际法委员会作用，推动规制人工智能军事化利用的“软硬结合”国际法规则体系构建。国际法院可通过咨询管辖权明确国际责任法、国际人道法在人工智能军事化利用领域的具体适用。国际法委员会可整合软法治理成果，结合实践困境对人工智能的军事利用这一议题展开深入研究，提供国际立法完善建议。

四、结论

在世界和平与发展面临多元挑战的背景下，各国应秉持共同、综合、合作、可持续的全球安全观，通过对话与合作，就如何规范人工智能军事应用寻求共识，构建有效的治理机制，避免人工智能军事应用给人类带来重大损害甚至灾难。⁴⁵在此过程中，应充分考虑人工智能技术特性，将伦理价值与技术规则通过法律手段进行调和并贯穿于人工智能军事应用全链条，确保有意义的人类控制。人工智能技术及系统并不具有承担责任的身份，国家责任与个人刑事责任在人工智能军事化利用背景下的适用规则情形亟待明确。现有国际法规则体系难以与该议题充分衔接，应对原则性条款加以具体化。在此过程中应充分发挥联合国平台作用，谋求国际社会共同价值基础，加强国际合作，警惕以技术优势谋求规则话语权的意识形态输出，推动针对该议题的高层级的具有约束力的国际条约产出。总体而言，人工智能在军事领域的应用对国际法构成了挑战，但尚未造成冲击，故应积极应对以推动二者的充分衔接。

⁴⁵ 《中国关于规范人工智能军事应用的立场文件》，该文件是中国首次就规范人工智能军事应用问题提出的倡议，于2021年12月13日由中国特命全权裁军事务大使李松在日内瓦召开的联合国《特定常规武器公约》第六次审议大会上提交。

人工智能立法视野下无人船责任分配机制的规范进路

王枫予*

摘要：近年来人工智能技术正影响航运产业变革，无人船责任分配机制变革亟待法律层面的有效回应。在 L4 级完全自主船舶场景下，传统海事法律框架在“去船员化”转型中面临多重困境：多元主体介入影响责任边界划分，部分主体法律属性悬置、责任竞合等问题较为突出，技术监管存在整体缺位；此外，法律规制框架较为滞后分散，MASS Code 草案等效原则难以应对算法危机。作为航运系统重要组成部分，算法决策机制正逐渐颠覆传统过错认定规则。其不可逆推性或影响因果关系认定，动态演化特性亦冲击产品缺陷判断标准。同时，责任主体的算法权责结构存在失衡情形。未来，应通过国际标准本土化构建相适应的法律框架，建立分层责任体系并完善算法全周期风险防控机制，推进技术治理模式从被动响应向主动规制转型。可尝试以等效替代原则衔接传统责任规范路径，以平衡技术创新与风险分配。

关键词：人工智能立法 无人船 责任分配 算法决策 等效替代原则

第一章 问题的提出

人工智能迅猛发展正重塑航运产业，无人船凭借自主导航、远程操控和算法决策等技术特性逐步实现商业应用，催生了“去船员化”的新型航运模式。然而，传统海事法律框架以“船员过错”为核心构建的归责体系面临颠覆性挑战。

IMO 将智能船舶划分为 4 个等级，其中第 4 级别 MASS，即完全自主船舶的规范力度和远程控制船舶相比较为缺少。尽管当前技术难以完全支撑自主船舶航行，但仍有必要对其责任分配机制进行探讨，以为将来相关立法提供借鉴。在 L4 级船舶情景下，传统船长及船员一线控制权让渡于算法系统后，责任主体呈现出权重转移与边界模糊的双重特征。一方面，多元操作主体影响归责逻辑的实际效能。¹远程操作人员、算法开发商、船舶所有人等多元主体介入船舶管理，但现行法律尚未明确其责任顺位与连带关系；²另一方面，过错判定标准失衡。传统意义上的“合理注意义务”难以适用于基于大数据的算法决策系统，同时智能航行

* 王枫予，男，哈尔滨工程大学人文社会科学学院，硕士研究生。

本论文是国家社科基金一般项目“基于国际海事组织立法进程的中国智能船舶法律规制研究”

(23BFX198)，黑龙江省哲学社会科学研究规划项目“《极地规则》生效背景下俄罗斯、加拿大北极通航法律政策演变及中国应对研究”(22GJB130)，黑龙江省法学会 2024 年重点课题《北极命运共同体视域下中国参与北极航运安全治理的路径研究》，中央高校安全研究专项课题“北极航道航行安全法律问题研究”的阶段成果。

¹ 林涓民：《论人工智能立法的基本路径》，载《中国法学》2024 年第 5 期，第 96-99 页。

² 邢厚群：《智能船舶背景下的海事立法挑战与回应》，载《中国海商法研究》2024 年第 4 期，第 41 页。

系统的不可解释性进一步放大了归责障碍。³

IMO 于 2023 年末发布《海上自主水面船舶规则》(MASS Code) 草案, 其中主张等效替代原则, 即无人船责任水平“不低于传统船舶”, 但其自愿性规范属性及碎片化条款难以全面应对算法危机、远程控制中断等新型风险场景。⁴ 欧盟《人工智能法》将智能航运系统纳入高风险监管范畴, 要求开发商承担严格安全评估义务, 却未与相关海事法律规范形成衔接机制, 导致算法责任与航行责任法律衔接出现一定空白。⁵ 我国虽于 2023 年出台《船舶自主航行试验技术与检验暂行规则》, 但此类行政规范性文件效力层级难以突破现行海商法以“承运人过错”为中心的责任框架, 更无法覆盖算法解释权、数据主权等前沿研究问题。⁶

应当注意的是, 更深层次矛盾在于技术特征与法律价值的结构性冲突。一方面, 无人船依赖实时数据交互与跨境算法协作, 但其引发的网络攻击、管辖权冲突等问题凸显国际规则协同机制的缺位。⁷ 另一方面, 技术中立性原则与航运安全价值存在失衡性问题, 部分学者倾向于通过“沙盒监管”鼓励技术创新, 而相对少地关注算法歧视、数据垄断等技术滥用行为。⁸

在上述背景下, 构建无人船责任分配机制应当面临多重问题, 包括法律逻辑重构的技术契合水平、国际规范本土转化的实现效果, 即 IMO MASS 规则与国内立法体系的兼容程度, 以及风险分配的可接受性等。⁹ 面对全球性人工智能立法浪潮, 亟需探索构建从“被动”到“主动”的规范进路, 以回应航运智能化转型深层次法治需求。

第二章 无人船责任分配机制现存困境分析

2.1 责任主体划分不清

对于航运系统算法, 在 L4 MASS 操作情景中, 传统“过失”和“过失责任”概念面临解释性危机。算法可能引发的技术故障(如导航系统错误或软件缺陷)需重新明确责任归属, 尤其是当缺乏人类直接干预时, 如何界定系统开发者、远程操作中心(ROC)或网络服务提供商的过错。

当前, 现行立法与公约未能明确多元主体间的法律关系, 形成责任主体识别与边界不清的双重困境。首先, 远程控制人员的法律属性界定仍然较为模糊。陆基操作的远程控制人员取代船上船员履行船舶操控职责, 但其身份性质缺乏法律层面的明确。若其与承运人存在长期雇佣关系, 则可被认定为“受雇人”, 承运人

³ 陈祎锟:《无人船运输的转承担责任研究——兼论<海商法>的修改》, 载《北京理工大学学报(社会科学版)》2023 年第 6 期, 第 111-115 页。

⁴ I:\MSC\108\MSC 108-4-1.

⁵ 参见 Regulation (EU) 2024/1689 (《人工智能法》) 第 16 条。

⁶ 《船舶自主航行试验技术与检验暂行规则(2023)》第 5 条。

⁷ 吴熙辰:《新型“网络海盗”的安全风险及其法律规制》, 载《法律研究》集刊 2024 年第 1 卷, 第 264 页。

⁸ 邢厚群:《智能船舶背景下的海事立法挑战与回应》, 载《中国海商法研究》2024 年第 4 期, 第 45 页。

⁹ 详见 2022 年 1 月 10 日 IMO MSC《MASS 规则路线图草案》

需承担替代责任；¹⁰而若以独立合同人提供相关技术服务，其责任归属则将排除于承运人责任网络之外。实践中，远程操作人员工作场所虽在陆地，但其操作失误引发的航行事故与传统船员过错造成的损害具有同质风险，法律地位的不对等性直接导致责任豁免依据缺失。MASS Code 草案虽引入等效责任原则，但未就远程控制人员与船员的权责匹配细则作出规定，使得相同危害行为面临差异性归责。

另外，AI 航行系统开发者与承运人的责任竞合边界不明。AI 航行系统作为无人船重要控制单元，其算法缺陷可能导致路径规划错误、避碰决策失灵等系统性风险。虽然《民法典》第 1203 条将产品责任归属于生产者与销售者，但承运人基于“谨慎处理”义务仍可能因未及时更新算法或发现系统漏洞承担过失责任。技术中立性原则下，开发者主张其仅承担技术层面的设计责任，而承运人则被要求对算法应用后果承担兜底责任，二者责任竞合缺乏明确划分标准。另外，当 AI 系统产生超出预设场景的自主决策时，开发者能否援引技术不可预见性免责仍存争议。¹¹

此外，船舶所有人与实际控制人的连带责任缺失。在无人船“船岸分离”运营模式下，船舶所有人可能将船舶租赁予技术公司实施远程控制，形成所有权与控制权的实质分离。现行海商法强调承运人责任，但未规定船舶所有人与实际控制人（如算法服务商）的连带责任。若当算法服务商篡改航行数据导致船舶碰撞时，受害人需通过《产品质量法》单独起诉开发者，而无法依据海事特别法追究船舶所有人责任。¹²欧盟《人工智能法》将涉及船舶安全的关键算法归类为高风险 AI，需满足严格数据治理和透明度标准，相关条款明确要求船舶系统取得许可，并在危险时平等保护所有交通参与者，同时限制生物识别技术滥用。¹³然而，其中第 16 条中，有关高风险人工智能系统提供者安全评估义务仅能约束开发者，未能衔接船舶所有人监管责任。¹⁴

对于责任主体划分问题，IMO MSC109 会议中各国发表了不同观点。比利时提出通过远程操作管理（ROM）机制解决船旗国、ROC（远程操作中心）与船公司之间的责任衔接问题。中国明确提出 MASS 证书应区分“无船员”状态的特殊豁免条款，建议在港口文件中明确记录“责任交替条件”，并要求 ROC 和船公司分别对数据完整性及人机界面设计负责。法国、西班牙和 IMRF 联合主张 MASS 需具备与传统船舶同等的搜救义务能力，强调算法设计中应融入应急响应模块以

¹⁰ 陈祎锬：《无人船运输的转承责任研究——兼论〈海商法〉的修改》，载《北京理工大学学报(社会科学版)》2023 年第 6 期，第 110-112 页。

¹¹ 范晓波，陈怡洁：《船舶无人化趋势下 AI 航行系统的责任探析》，载《中国海商法研究》2021 年第 4 期，第 42 页。

¹² 孙誉清：《商用无人船法律地位的界定》，载《武大国际法评论》2019 年第 6 期，第 127-128 页。

¹³ 刘子婧：《欧盟〈人工智能法〉：演进、规则与启示》，载《德国研究》2024 年第 3 期，第 118-120 页。

¹⁴ 邢厚群：《智能船舶背景下的海事立法挑战与回应》，载《中国海商法研究》2024 年第 4 期，第 39 页。

实现主动援助。此外，俄罗斯要求界定“单一责任主体原则”，即特定航行阶段仅允许一名实体（如船长或远程操作员）承担最终责任。

2.2 技术监管结构性缺位

无人船技术的快速发展对现有海事监管体系形成系统性危机，网络攻击风险加剧与适航义务履行冲突、沙盒监管模式缺位构成技术监管的双重结构性缺陷，亟待通过立法更新与制度修正予以融合。

一方面，网络风险防御机制与传统适航义务存在适用冲突。传统适航义务聚焦船舶物理安全与船员资质，要求承运人确保船舶处于“适航状态”并配备适格船员。然而，无人船依赖远程控制与数据交互系统，网络攻击可通过操控导航参数、植入恶意代码等方式劫持船舶，形成远超传统适航义务覆盖范围的安全威胁。国际海事组织 2023 年调查显示，近半数航运企业认为现行海事法律对智能船舶网络安全防护要求存在空白。尽管《国际船舶安全运营规则》（ISM Code）要求船舶建立安全管理体系，但未明确数据加密、系统冗余等技术标准，导致承运人难以将网络安全成本纳入适航义务履行框架。根据欧盟《人工智能法》定稿版本内容，智能航运系统被列为高风险人工智能，但其第 16 条聚焦算法透明度，未与《海上人命安全公约》（SOLAS）的适航标准形成有效衔接。¹⁵

另一方面，技术测试活动面临法律保障空白。无人船需经过实船测试验证技术可行性，但现行海事立法缺乏概括性监管机制。以中国《船舶试航活动通航安全监督管理办法》为例，其要求测试方提前报备航行计划，并限制试航区域范围，但未规定试错容错机制。实际操作中，智能船舶测试若突发技术故障而违反《海上交通安全法》第 3 章“海上交通条件和航行保障”条款，测试企业面临行政处罚与民事责任双重风险。韩国《自主航行船舶法》引入沙盒监管机制，允许测试船舶在限定区域内豁免部分船舶配员与设备标准，并在测试事故中减免企业责任。

2.3 法律规制框架滞后分散

无人船责任分配机制所面临的一大困境在于现行法律框架呈现滞后及分散化特征。此类特征既体现在国际公约与国内立法衔接缺失，同时表现为技术标准与制度规范发展进程的差异性，从而致使责任认定与风险分配缺乏统一标准。

从国际法层面观察，现行海事公约体系尚未形成针对无人船的规范性框架。尽管国际海事组织（IMO）于 2018 年启动海上自主水面船舶（MASS）立法研究，但其制定《海上自主水面船舶规则》（MASS Code）草案目前仍处于探讨阶段，未形成事实上的国际法律约束效力。关键性公约如《国际海上避碰规则》（COLREGS）中的“合理瞭望”“良好船艺”条款与无人船的自主决策机制存在实质冲突；《海员培训、发证和值班标准公约》（STCW）设定的船员资质要求亦难

¹⁵ 皮勇：《欧盟〈人工智能法〉中的风险防控机制及对我国的镜鉴》，载《比较法研究》2024 年第 4 期，第 69-73 页。

以契合岸基控制人员的职能定位。此类立法滞后导致国际航运实践中存在规范空白，对无人船航行安全标准的解释差异逐渐扩大。

在国内法领域，《海商法》《海上交通安全法》等基础性海事法律规范仍以传统船舶为规制对象，对无人船法律地位、责任主体等关键问题缺乏明确定义。中国《海商法》将“船舶”界定为“海船和其他海上移动式装置”，但未对自主航行系统的法律属性作出直接乃至间接回应，导致无人船适航性判断标准缺失。尽管近年来中国船级社不断发布最新版《智能船舶规范》，但其技术性标准尚未转化为法律层面的强制性规则。

此外，专项责任规则的分散性进一步削弱法律体系的协调性。在碰撞责任领域，《1972 年国际海上避碰规则》要求船舶间通过视觉信号协商避让，而无人船 AI 决策系统可能无法完全模拟人类船员的互动逻辑，从而引发责任认定困境。在网络安全责任方面，国际海事委员会（CMI）问卷调研显示，75% 的成员国尚未建立针对无人船网络攻击的归责规则，现有责任体系难以覆盖远程控制中断等新型风险。¹⁶规则分散的状态使得航运主体面临法律适用不确定性和合规成本上升等多重压力。

第三章 算法特性与责任分配机制的法理冲突

3.1 算法决策机制对传统责任认定规则的影响

传统侵权责任制度以过错责任原则为核心，其归责基础建立在“理性人标准”之上。¹⁷在传统船舶碰撞法律框架中，责任认定需考察船员是否存在未尽到合理注意义务或违反避碰规则的操作行为。¹⁸然而，智能船舶的算法决策机制从根本上阻断了这一机制，即当航行控制权从人类船员让渡给 AI 系统时，“理性人标准”的适用面临多重困境。首先，算法决策缺乏人类主观过错范畴的归责基础，其决策行为仅为概率计算结果，无法用传统过失理论中的“预见可能性”或“结果回避可能性”进行评价；¹⁹其次，算法具有不可逆推性，导致决策过程形成技术性探索空白，即便系统输出了错误指令，也难以通过现有技术手段完整回溯决策路径；再者，理性人判断所依据的职业经验与伦理准则，在算法中被转化为数据参数，使得传统过错认定的规范性标准失去参照标准。²⁰

算法系统的技术特性还导致责任关联体系的断联。现行法律中的近因原则要求损害结果与行为之间存在直接、可预见的因果关系。但智能船舶事故往往由算法缺陷、数据污染、网络攻击等多重因素叠加引发。²¹面对复合型致损路径，传

¹⁶ 邢厚群：《智能船舶背景下的海事立法挑战与回应》，载《中国海商法研究》2024 年第 4 期，第 39 页。

¹⁷ 同上，第 41 页。

¹⁸ 严谨：《智能船舶的法律规制研究》，2020 年哈尔滨工程大学硕士学位论文，第 37 页。

¹⁹ 吴熙辰：《新型“网络海盗”的安全风险及其法律规制》，载《法律研究》集刊 2024 年第 1 卷，第 263 页。

²⁰ 徐晨艺：《智能船舶网络风险责任保险法律问题研究》，2022 年大连海事大学硕士学位论文，第 45 页。

²¹ 范晓波、陈怡洁：《船舶无人化趋势下 AI 航行系统的责任探析》，载《中国海商法研究》2021 年第 4 期，第 43 页。

统责任追溯机制无法有效区分各环节主体的过错程度与责任份额。此时，若严格适用“近因原则”，可能完全免除算法设计者的责任，因为直接致损的传感器故障可能被认定为偶然技术缺陷，而非系统性过错。

现行产品责任制度同样面临适应性危机。根据《民法典》第 1202 条，产品责任成立需以存在产品缺陷为要件，而智能船舶主要涉及设计缺陷问题。但算法系统的特殊性质导致设计缺陷判断面临双重障碍：一方面，算法运行模型具有动态演化特征，其缺陷可能产生于系统运行阶段的数据迭代，而非原始设计阶段；另一方面，等效替代原则的适用标准实效，当算法设计符合行业通行技术规范但无法达到人类船员的避碰效能时，是否构成“不合理危险”缺乏明确判定规则。²²这导致海事司法实践中索赔方可能难以举证设计缺陷存在，侵权责任与产品责任发生规范竞合却无法有效衔接。

面对算法带来的责任认定变革，法律亟需构建相应评价标准。基于伦理准则的可解释性要求和技术参数的可验证性标准应成为新型评价体系的基础。²³例如，可通过建立“最小安全冗余度”指标，要求算法系统的避碰决策能力不低于同期人类船员的平均水平；同时引入算法运行日志的法定存证制度，为决策过程追溯提供技术支撑。²⁴

3.2 风险分配模式的技术识别障碍

无人船运行中形成的“远程控制-算法决策”二元结构，引发了代理人法律关系的瓦解问题。远程控制人员虽维持名义上的控制权，但算法系统的自主决策已突破传统代理关系中“指示遵从”的法律要件。²⁵此类权责分离的状态是形式监管，但为实质自治，使得雇主责任中的“选任监督”要件面临适用困境。即便船东严格履行人员培训与设备管理义务，仍无法有效预防算法决策失误可能造成的损害后果。

深度学习算法的动态演化特性与产品责任制度产生冲突。对于产品缺陷判断，可以投放市场时科学技术水平为基准。然而，智能航行系统通过持续数据训练不断优化决策模型，其缺陷可能始于系统运行阶段的算法迭代，而非初始设计环节。这种不可逆的技术演进特征，导致既有的设计缺陷认定标准失去时间维度上的基本参照。国际海事组织（IMO）的研究报告显示，62%的智能船舶系统缺陷产生于部署后的在线学习过程，使得生产者可通过软件更新规避产品责任。

算法决策的可解释性缺陷加剧了责任豁免困境。现行法律体系尚未建立技术

²² 郭萍、武利海：《无人船适用领海无害通过制度的困境及对策》，载《武大国际法评论》2022 年第 1 期，第 50 页。

²³ 陈景辉：《算法之治：法治的另一种可能性》，载《法制与社会发展》2022 年第 4 期，第 124 页。

²⁴ 方阁、初北平：《海事网络安全风险保险的法律治理研究》，载《江西社会科学》2020 年第 5 期，第 183 页。

²⁵ 陈秭赜：《无人船运输的转承责任研究——兼论〈海商法〉的修改》，《武汉大学学报（哲学社会科学版）》2023 年第 5 期，第 103 页。

透明度评估的法定标准,使得算法系统的不透明特性无法转化为有效的责任抗辩事由。根据欧盟《人工智能法》的技术规范,高风险 AI 系统需达到 Level 3 以上的可解释性标准,但在海事领域缺乏对应的实施细则²⁶。这种制度空白可能导致两种情况:一、运营商承担超出实际控制能力的严格责任;二、完全免除技术开发方的责任后果。

国际标准体系的碎片化状态进一步放大了技术风险。尽管 IMO 于 2021 年发布《自主船舶试航暂行导则》,但各国在算法安全水平设定等方面仍存在显著差异。美国海岸警卫队采用“功能等效”原则,允许通过算法实现传统船员义务;而日本国土交通省《自动航行船舶安全技术指南》则坚持“性能卓越”标准,要求算法系统须超越人类平均水平。这种标准分野导致跨国航行船舶面临责任认定冲突,不同国家认证的算法系统在他国海域可能被认定为存在技术缺陷。MASS Code 草案明确要求将安全为核心的设计原则融入系统算法开发中,包括故障安全机制和应急协议设计以确保风险管理。韩国通过 KASS 项目提出了针对自主导航系统算法的海试验证程序,重点测试算法在不同技术模块之间的互联性表现(如避碰、路径规划等),并要求开发完整的验证与确认(V&V)计划以确保算法可靠性。IACS 强调数据质量对算法决策的直接影响,建议建立数据评估框架,并针对 MASS 不同自主等级实施定制化数据验证流程。

构建契合算法特性的风险分配机制,需从以下方面突破技术滞碍。首先,建立技术缺陷认定体系,将算法更新纳入持续性合规审查;其次,引入算法透明度分级制度,根据可解释性水平分配举证责任;最后,推动形成全球统一的算法安全基准,通过国际海事组织框架协调各国技术标准。同时,针对算法特性下的责任归属问题,中国在 LEG111 中建议通过制定实施指南,要求新兴责任主体(如 ROC)以“等效原则”履行原船员职责,同时在公约框架内明确适用条款与例外情形,同时建议通过合同条款明确 ROC 的法律地位和跨境管辖权问题,并提议在《联合国海洋法公约》(UNCLOS)框架下建立旗国与 ROC 所在国间的协同监管机制。²⁷

3.3 责任主体的算法权责结构失衡问题

对于完全自主船舶,人工智能自主决策系统深度参与船舶管理,算法成为责任分配机制的重要变量,岸基控制中心、算法开发者、导航服务商与网络安全维护方之间形成了多维度权责博弈。这既源于技术控制权的结构性分层,亦反映出传统归责原则在智能航运场景下的价值冲突。当前连带责任制度因欠缺精细化划分依据,难以有效平衡技术创新激励与风险防范的立法目标。²⁸

²⁶ 余成峰:《法律人工智能新范式:封闭与开放的二元兼容》,载《中外法学》2024 年第 3 期,第 601 页。

²⁷ 郑玉双:《计算正义:算法与法律之关系的法理建构》,《政治与法律》2021 年第 11 期,第 101 页。

²⁸ 方阁、初北平:《海事网络安全风险保险的法律治理研究》,载《江西社会科学》2020 年第 5 期,第 181 页。

一方面，算法开发者与船舶运营方的权责失衡凸显了技术和规范的冲突。从技术实现路径观察，算法开发方掌握逻辑框架的初始设定权，可通过深度学习模型持续修正系统运算规则，形成隐蔽的算法权力优势。²⁹这种现象导致产品缺陷认定标准趋于模糊，《产品质量法》第46条以“不合理的危险”或技术标准作为缺陷判断要件，但无人船算法设计的泛化特征使得设计缺陷的因果关系产生泛化性的影响。例如，远程控制中心人员因算法决策失误引发碰撞事故时，既需判定算法参数的初始设计合理性，又要考量操作人员的风险干预能力，现行司法解释尚无法对“注意义务”的算法化标准作出明确回应。³⁰俄罗斯关于无人船立法计划中计划要求开发商对智能航行系统缺陷承担连带责任，尝试将系统开发商纳入责任主体范围，这过度强调开发者严格责任可能压缩技术迭代空间，需要构建依技术参与度划分的梯度责任体系。

另一方面，运营主体间的协作治理缺陷可能引发责任体系衔接不畅。岸基控制中心兼具算法执行与实时干预双重功能，其“一对多”控制模式打破了传统船舶管理责任边界；导航服务商通过数据接口实时校准航行路径，网络维护方则承担防控系统性网络攻击的关键职责，三者形成递进式技术依赖关系。但多主体介入使得过错形态呈现混合性特征，《民法典》中产品责任制度难以应对算法动态优化引发的责任问题。国际海事组织提出的等效替代原则虽为责任配置提供了“不低于传统船舶安全水平”的判断基准，但未解决多主体责任体系的量化标准，在实践中易产生责任承担空白或责任竞合的相关情况。

解决此类问题需从权责结构与证明规则两个维度展开。首先，应建立以算法控制权为中心的权责体系，将算法开发者的框架设计权、运营方的实时干预权等予以分层解决，对应设置差异化归责要件。其次，可借鉴欧盟《人工智能法》第3章的安全协作机制，要求各方建立联合风险评估与履责记录系统，通过区块链技术实现过错行为的追溯性验证。³¹最后，在证明规则层面引入算法解释性标准，对具备可解释特征的算法适用过错推定，而对不透明性算法实施严格责任。

基于以上分析，基本揭示了无人船责任制度的深层矛盾：传统海事法律规范依赖线性因果关系，以较为单一的当事方作为责任分配机制的中心主体，这与智能航运系统算法支持下的“网状”责任结构形成重大冲突。

第四章 无人船责任分配机制的优化路径

4.1 推进国际标准国内转化的实现

国内需构建、优化具有可操作性的国内立法框架。首先，可以选择以体系化立法为主，辅助技术规范进行动态调整。优先通过修订现有《海上交通安全法》

²⁹ 余成峰：《法律人工智能新范式：封闭与开放的二元兼容》，载《中外法学》2024年第3期，第594-595页。

³⁰ 陈秭赜：《无人船运输的转承责任研究——兼论〈海商法〉的修改》，载《北京理工大学学报（社会科学版）》2023年第6期，第108-111页。

³¹ 参照程啸《侵权责任法》（第三版）

《海商法》，直接或者间接增设无人船相关条款，同时在人工智能立法框架下筹划跨领域的系统性立法。该思路借鉴法国等国的立法经验，将传统船舶法律体系与新技术的特殊性相衔接，避免重复性立法资源浪费。其次，建议实施地方试点机制。可以选取航运基础较好地区，制定区域性无人船管理规章，重点突破智能航行规则、远程监测等关键技术标准的立法验证。³²试点区域的实践数据可为国家层面立法提供技术安全水平和责任分配机制的实证支撑，如通过测试确定无人船引航权属、避碰规则等核心要素。³³

同时，建议参照 MASS 规则制定路径，将强制性技术规范（如通信协议、故障冗余系统要求）通过《船舶自主航行试验技术与检验暂行规则》等形式逐步转化为立法内容。应鼓励无人船研发机构参与立法研讨，形成技术安全共识，为法律责任界定提供专业依据。最后，应当继续对接国际立法动态。³⁴进行深入参与 IMO MASS Code 制定及相关立法进程，推动我国技术标准与国际公约框架紧密衔接。³⁵通过国内立法试验与国际规则输出双重保障，进一步提升我国海洋法治话语权，为国内企业参与全球智能航运竞争提供建设性条件。³⁶

4.2 基于技术特性构建责任分层体系

无人船责任分配机制需要根据技术发展水平与运营阶段差异进行层级化设计。首先，在测试期，无人船技术成熟度与算法可靠性尚未得到充分验证，采用严格责任制度有助于强化技术开发主体的风险防控意识。测试主体应当对无人船设备故障、算法偏差引发的损害承担无过错责任。³⁷依据我国《船舶自主航行试验技术与检验暂行规则》，测试期间需建立风险准备金与强制保险配套措施，为责任承担提供制度保障。³⁸待技术进入商用阶段后，系统可靠性经规模化应用验证，可调整适用过错责任原则，将过错判断基准从“人类注意义务”转向“等效替代原则”。若智能航行系统的安全效能达到或超过传统有人船水平，则技术使用者得以免责；若存在系统性缺陷或安全标准低于有人船，则认定其未尽等效替代义务。这种从严格责任向过错责任的梯度过渡，既可保障技术研发的积极性，又能引导商业化应用的风险治理。

算法决策的可解释性应当成为责任构成要件。无人船的责任判定需要突破传统侵权法中“理性人”标准的局限，建立以技术透明度为核心的前置评估机制。通

³² 严正浩：《技术算法缺陷对智能船舶适航的影响》，载《水运管理》2024年第3期，第29页。

³³ 孙武：《智能船舶规范法规发展趋势分析》，载《船舶工程》2024年第10期，第18-19页。

³⁴ 方阁、初北平：《海事网络安全风险保险的法律治理研究》，载《江西社会科学》2020年第5期，第181页。

³⁵ 袁雪、于博：《国际海事组织海上自主水面船舶的法律规制进程及中国因应》，载《国际法研究》2023年第1期，第48-51页。

³⁶ 杨文丛：《人工智能船舶对我国〈海商法〉及相关规定的影响》，载《水运管理》2022年第2期，第23-25页。

³⁷ 邢厚群：《智能船舶背景下的海事立法挑战与回应》，载《中国海商法研究》2024年第4期，第45-47页。

³⁸ 详见《船舶自主航行试验技术与检验暂行规则（2023）》第28条。

过引入“算法决策可解释性”分级评估制度，要求技术提供者对路径规划、避碰决策等关键算法进行形式验证与场景模拟测试。当算法决策过程无法追溯或存在无法合理解释的异常行为时，应推定技术缺陷与损害结果之间存在因果关系，由技术提供者承担举证责任倒置的证明义务。

2023 年 MASS Code 草案提出技术阶梯式责任框架，第四级全自主船舶的技术基准被设定为“行业最优实践水平”，要求责任判定需参照同期市场上同类无人船的技术性能指标。此类动态基准的确立，有效解决了等效替代原则的局限性，即当无人船整体安全效能超越传统有人船时，法律评估应从新技术对标旧标准转向同代技术横向比较。该类调整既防范了技术垄断导致的责任虚化风险，又避免了过高责任水平对技术迭代的抑制。此外，本文认为需引入分层留白规则。韩国《自主航行船舶法》采用框架立法与授权立规协同的模式，对责任限额、免责事项等争议问题仅作出原则性规定。通过立法留白，可维持法律体系稳定性，为无人船责任规则后续精细化调整预留制度空间。

4.3 建立算法风险防控机制

无人船责任分配需以风险预防为主，构建分层递进式的算法风险防控机制。一方面，可以尝试设立海事沙盒试验区，推行“分级豁免+动态观测”制度。针对智能算法的演进性与海上场景的复杂性，IMO《自主船舶试航临时指南》建议在可控范围内允许技术测试豁免部分现行法义务。可参考韩国《自主航行船舶法》，依据算法成熟度将测试分为多个阶段。封闭场地测试阶段允许完全豁免传统船舶配员要求，开放海域测试阶段仅豁免部分航行规则，试运营阶段可以采取“等效替代”原则动态调整豁免范围。同时，要求测试主体上传航行决策日志，对路径规划、避碰响应等核心算法进行实时跟踪与异常预警。³⁹

另一方面，可尝试构建技术合规前置条件。欧盟《人工智能法案》将高风险算法纳入强制认证范畴，要求开发方提交形式化验证报告与场景模拟测试数据。我国可建立“两阶段认证体系”。⁴⁰设计阶段需通过第三方机构对算法决策边界进行形式化验证，确保其决策逻辑符合《避碰规则》等强制性规范；商用前需完成类似于虚拟场景测试，验证极端海况下的系统稳定性。对于未通过认证的算法，限制或者禁止投入商业运营，并建立开发者“负面清单”以强化主体责任。⁴¹另外，应当完善强制责任保险制度，进行风险分散。《2013 年北欧海上保险计划 2023 年版》（The Nordic Marine Insurance Plan of 2013, Version 2023）设置了类似于“算法缺陷责任险”。我国可尝试借鉴该模式，明确须投险种，覆盖算法设计漏洞导致

³⁹ 桑家军、严忠伟：《智能船舶对船舶保险条款的适用性研究》，载《中国海事》2024 年第 5 期，第 22 页。

⁴⁰ 杜健、李宁、赵亮等：《智能船舶发展若干问题的思考与建议》，载《中国海事》2023 年第 4 期，第 46 页。

⁴¹ 邢厚群：《智能船舶背景下的海事立法挑战与回应》，载《中国海商法研究》2024 年第 4 期，第 39-41 页。

的直接损失；船舶运营方需对运营风险进行投保，覆盖动态观测中未能识别的残余风险。⁴²

4.4 加快技术治理模式转型

传统法律规制响应难以与无人船技术迭代速度同步，亟需调整技术治理模式。针对无人船技术快速更迭与法规滞后性的矛盾，建议建立技术标准动态清单机制。⁴³该制度可以未来出台的 MASS Code 为框架与标准，并结合中国船级社《智能船舶规范 2025》，将船舶自主避碰系统、远程操控协议等关键技术的安全水平纳入管理范围。通过立法授权动态增补技术指标并通过“等效替代原则”实现合规性审查，既保留现行船舶安全责任体系框架，又为新兴技术应用预留制度空间。⁴⁴

此外，应当建立算法开发、部署、运行的伦理审查机制，即在开发阶段引入第三方评估机构，对避碰决策算法进行偏见检测与可解释性验证；在航行周期内强制披露核心算法的功能边界与应急逻辑，保障事故追溯时的责任判定可行性。同时在岸基控制系统中融入审计性质的模块，确保远程操控指令的透明化记录。同时，可在部分具有实施条件的沿海城市先行构建“监管沙盒”机制，允许无人船在划定水域开展混合交通场景下的技术试航。依据《船舶自主航行试验技术与检验暂行规则》设立分级测试标准，针对遥控驾驶、自主靠泊等场景采取差异化的法律责任豁免政策。⁴⁵

结语

人工智能立法背景下的无人船责任分配机制面临传统海事法律框架与技术特性的结构性冲突。研究表明，L4 级无人船“去船员化”颠覆了以“人类过错”为中心的归责逻辑，多元主体权责交错、算法不可逆推性及国际规则碎片化构成主要规范困境。算法决策对过错认定规则的重构、技术中立性与航运安全的价值失衡，亟需通过法律框架的动态适应性调整予以回应。

未来相关研究可关注以下方面：首先，深化算法可解释性与司法判定标准的衔接机制，构建分层举证规则，明确算法缺陷的技术认定路径；其次，推动责任体系从个体归责向系统化风险分配转型，探索算法全周期责任与保险体系的协同模式，平衡技术创新激励与公共安全保障；另外，强化国际标准本土化与区域协作，针对跨境航行管辖权冲突与数据主权问题，构建“旗国-港口国-技术提供国”协同监管框架。此外，需关注动态场景下等效替代原则的适用边界，推动海事法律从被动容错机制转向基于技术验证的主动预防体系，为全球智能航运治理贡献

⁴² 郑志峰：《人工智能使用者的立法定位及其三维规制》，载《行政法学研究》2025 年第 1 期，第 26-27 页。

⁴³ 蔡莉妍：《智能船舶法律规制的困境与突破》，载《华南理工大学学报（社会科学版）》2020 年第 6 期，第 7 页。

⁴⁴ 崔佳璐：《IMO 框架下智能船舶适航的法律规制研究》，载《海洋开发与管理》2023 年第 7 期，第 26 页。

⁴⁵ 沈伟、赵尔雅：《数字经济背景下的人工智能国际法规制》，载《上海财经大学学报》2022 年第 5 期，第 125-127 页。

中国方案。

深度伪造技术法律规制的中欧对比

顾杰媛¹

摘要：Chat GPT、Deepseek 等一系列生成式人工智能系统的出现，使得每一个人可以轻而易举地使用人工智能生成内容。纵观各类深度伪造案件，可以从“生成-传播”两个阶段解构其可能引发的风险，并以深度伪造技术的活动范式为介质分析风险的产生。欧盟和中国均通过多维度的法律对深度伪造问题作出了规制。但通过比较欧盟与中国对深度伪造法律规制的不同规定，会发现二者对人工智能系统提供者所设定的义务强度不同、侧重保护的权益不同、对事前监管和事后救济所采取的态度也不尽相同。这两种规制模式一定程度上阻止和减轻了非法的深度伪造活动，但也都存在固有的局限性。且深度伪造作为一项新兴技术，除了注重监管以外，还需要在规制与创新之间取得平衡²，并需要各国加强国际协同与合作，共同应对深度伪造技术这一复合体所带来的时代挑战。

关键词：深度伪造 欧盟人工智能法 欧盟数字服务法案 监管与创新

一、“深度伪造”概念辨析

谈论深度伪造法律规制问题必须先从技术上和法律上解析深度伪造这一概念。技术上的深度伪造概念源于“深度学习”（Deep learning）和“伪造”（Fake）的结合，是指基于深度学习模型——生成式对抗网络（Generative Adversarial Networks, GAN），不断进行神经网络和技术迭代，以生成愈加逼真的深度伪造内容的过程。图像和音视频伪造是深度伪造技术应用最广泛的场景，例如大众广知的“AI 换声换脸”就是深度伪造技术的典型应用³。而从本质上讲，深度伪造技术的根本逻辑可概括为“学习-训练-生成”，即深度学习模型通过对大量数据的学习，最终生成深度伪造内容。

法律上的深度伪造概念则在不同国家有着不同的规定。《欧盟人工智能法案》将深度伪造定义为人工智能生成或操纵的图像、音频或视频内容，类似于现有的人物、物体、地点、实体或事件，并且会使人错误地认为其是真的或真实的⁴。欧

¹ 顾杰媛，女，上海对外经贸大学国际组织学院（贸易谈判学院）硕士研究生，联系方式：18274039919

² Guan Zheng, Jinchun Shu, Kaigeng Li, *Regulating deepfakes between Lex Lata and Lex ferenda—a comparative analysis of regulatory approaches in the U.S., the EU and China*, *Crime, Law and Social Change* (2025) 83:1, <https://doi.org/10.1007/s10611-024-10197-z>

³ 刘东亮,黄安,杨晓科:《深度伪造技术滥用及其法律规制》,载《延边大学学报(社会科学版)》2025年第1期。

⁴ 《AI Act》Article 3 (60): (60) ‘deep fake’ means AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful.该中文译文为丁晓东教授所译版本。

盟这一定义强调了深度伪造内容对人所造成的影响，并将其作为认定深度伪造的依据之一，这与欧盟一贯来强调人权的立场相契合，并定下了欧盟对生成式人工智能深度伪造全生命周期法律规制模式的基调。与欧盟强调公众对深度伪造看法的立场不同，中国从技术出发对深度伪造的定义作出了规定。中国的官方文件中并没有采取“深度伪造”这一字眼，而是使用了“深度合成”这一说法。《互联网信息服务深度合成管理规定》（以下简称《深度合成管理规定》）明确规定：深度合成技术，是指利用深度学习、虚拟现实等生成合成类算法制作文本、图像、音频、视频、虚拟场景等网络信息的技术。这一定义排除了“伪造”这一略带贬义的表达，更加强调深度合成的技术性而非负面性，更显中立。以技术为切口，与《深度合成管理规定》主要针对深度伪造技术提供者和技术支持者作出义务规定的模式相符合，也为后续既要求监管、也要求坚持发展创新的态度打下基础。欧盟和中国对深度伪造所做的不同定义已经在一定程度上凸显出二者不同的立法偏好。

二、“深度伪造”可能引发的法律风险

纵观深度伪造所带来的各类风险问题，可以将其总结为“生成-传播”两个阶段。以深度伪造最典型的案件为例，2017年12月，在美国知名社交平台红迪网 (Reddit) 上，一个署名为“Deepfakes”的用户发布了一系列换脸女明星的色情视频。这一案件先由深度伪造技术的使用者利用人工智能系统“生成”深度伪造视频，再通过社交媒体进行“传播”，引发了公众对隐私权和名誉权的担忧，也对网络环境社会秩序产生了不良影响。基于深度伪造“生成-传播”的范式，可以分阶段地评估其可能引发的风险。

（一）“生成”阶段可能引发的法律风险

基于前述对深度伪造技术本质的认识，对深度伪造的“生成”阶段可以进一步分解提炼为“收集数据、训练模型、生成内容”⁵三个步骤。三个步骤可能引发的风险也不尽相同。

1. 收集信息——数据隐私被侵犯

深度伪造系统依赖海量个人生物识别数据（如面部特征、声纹信息）进行模型训练，这些海量信息的获取活动成为了风险的来源。数据获取常涉及未经授权的抓取行为：例如通过社交媒体爬虫收集用户公开照片，或非法购买黑市数据。这种未经得个人同意抓取、收集公民个人信息的行为显然会侵犯公民的数据隐私。除未经得授权收集和使用数据外，还存在超出获得了一定授权但收集和使用数据超出授权范围的侵权情形，这种过度采集数据的行为同样是对公民个人隐私的侵犯。

2. 训练模型——数据偏见与算法歧视

⁵ 刘东亮,黄安,杨晓科:《深度伪造技术滥用及其法律规制》,载《延边大学学报(社会科学版)》2025年第1期。

“生成”阶段的数据偏见与算法歧视的风险，根植于海量数据当中，并在深度学习算法下放大。当模型从海量数据中进行提取时，其并非完全客观中立，而是被动继承了人类社会沉淀在数据中的结构性偏见。例如，在一项商业性别分类研究中，研究人员分析了微软等公司的面部识别系统，发现其训练数据中深肤色女性样本数据大量缺失（仅占 4.7%-7.7%）⁶，导致分类准确率严重下降，也加剧了对女性和少数族裔的刻板印象以及歧视。这种生成结果看似源于“数据分布不均”，实则折射出深度学习模型在数据采集阶段的选择性忽略。同时，更加值得注意的是，当我们试图补救这种风险时，结果往往会进一步加剧偏见固化。现有的模式大部分是通过添加偏见过滤词库、设置道德约束规则来遏制有害生成，但这种表层干预无法根除模型认知框架中的深层偏见，反而会带来一种“逆向歧视”。

3.生成内容——生成虚假信息和非法内容的风险

深度伪造技术所生成的内容并不必然违法，但这一技术所合成的内容不可避免地带有虚假性这一特征，存在着被用来制造虚假信息和非法内容的可能性，以其潜在的非法性和虚假性特征带来法律风险。深度伪造内容、虚假信息和非法内容这三个概念并不能等同，但三者互有交叉，一旦深度伪造内容落入虚假信息或非法内容的概念范畴中，则可能受到相应法律法规的规制。

在“生成”阶段，深度伪造技术基于深度学习的内容生成机制，易生成含有虚假内容的文本或图像，抑或带来非法内容。此类内容不仅可能直接违反《网络安全法》《数据安全法》对网络信息内容的规范性要求，更因生成过程的自动化特征，容易导致对违法内容和虚假信息对生成主体难以追溯，引发法律风险。例如生成具有暴力暗示的隐喻性文本或隐含极端意识形态的符号化图像。此类内容通常借助技术中立性外衣以规避现行的合法审核机制，影响法律事实的认定。

（二）“传播”阶段可能引发的法律风险

发达的社交网络为深度伪造的传播提供了温床，也可能导致一些不合法的深度伪造内容愈加广泛传播，最终导致严重后果的发生。需要明确的是，深度伪造内容传播引发风险的前提是深度伪造内容属于虚假信息或非法内容，合法的深度伪造的内容传播并不会带来消极影响。

1.侵犯公民名誉权和肖像权的风险

“深度伪造”在传播阶段对公民名誉权的侵犯，本质上是一种依托技术优势对人格权益的系统性破坏。在侵犯名誉权的情形中，“深度伪造”技术的应用已呈现高度场景化的特征，面对诸如职场诽谤、家境纠纷等不同领域时，生成式人工智能会自动选取最为贴合的传播手段，在算法的助推下形成全网级的传播，而受害者往往面临“自证清白”的悖论，传统诽谤行为的破坏力在“深度伪造”的帮助下被

⁶ Joy Buolamwini: *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018.<https://proceedings.mlr.press/v81/buolamwini18a.html>

放大了数倍。据统计，截至 2023 年，主流社交平台在线 95,820 次深度伪造中，98%是色情内容，其中 99%的视频针对女性⁷。这种名誉损害一旦形成，便具有极强的“数字顽固性”。即便最终通过技术手段证明内容虚假，受到辟谣和删帖滞后性的影响，传播过的平台仍会残留截图、录屏或文字讨论，受害者不得不面对一种“数字幽灵”的长期困扰。除伪造色情内容这一典型外，针对政治家、企业家或是社会名人生成的一些发表不当言论的深度伪造内容，也会对其名誉权造成损害。此外，深度伪造内容如果输出了未经授权的公民肖像，也存在侵犯公民肖像权的风险。

2. 损害国家安全的风险

生成式人工智能深度伪造技术对国家安全带来的法律风险，首要体现在对国家信息主权的系统性侵蚀。深度伪造通过算法驱动来生成用户所需要的内容，可大规模生成具有政治煽动性、军事误导性或社会分裂性的一系列虚假信息，此类信息会对国家安全带来不可忽视的危险。深度伪造技术若被境外敌对势力或恐怖组织操控，可能直接冲击国家的关键信息基础设施，干扰政府权威信息发布渠道的公信力，甚至引发国际社会对主权国家决策真实性的信任危机，构成对国家数字主权的实质性威胁。例如，2024 年美国的总统大选被称为“人工智能时代的首场美国大选”，选举时期中深度伪造所带来的乱象频发。布鲁金斯学会研究发现，美国右翼极端分子炮制了关于救灾、移民、犯罪、边境安全等的深度伪造内容，在塑造公众对候选人和竞选议题的看法方面产生了负面影响⁸。

其次，该技术通过破坏社会认知秩序来危及国家安全治理的根基。由于深度伪造生成的虚假音视频内容具有高度仿真性，能够精准模仿国家领导人、国家的军事指挥系统等，起到“以假乱真”的效果。当此类伪造内容突破传统内容审核机制，而以不可遏制的趋势在网络空间扩散时，不但可能歪曲国家一些重大政策的传播效力，更可能在国际社会制造虚假的军事动态或外交声明，导致国际社会对本国战略意图产生误判，进而进一步诱发地缘政治冲突或带来经济制裁风险。

三、“深度伪造”法律规制的中欧模式

对于深度伪造所可能引发的各种风险，中国和欧盟均采取了一定的措施进行规制。但通过其具体规则的设定会发现二者的具体侧重点并不一致，进而反映出二者对于深度伪造这一领域的立法倾向。

（一）对“生成”阶段法律风险的规制

1、对数据隐私的保护

⁷ Kayleigh Gibson, *Gender bias, AI, and deepfakes are promoting misogyny online*, Online Violence Against Women, <https://blogs.lse.ac.uk/wps/2025/01/09/gender-bias-ai-and-deepfakes-are-promoting-misogyny-online/>

⁸ Darrell M. West, *“How Disinformation Defined the 2024 Election Narrative”* The Brookings Institution November 7, 2024, <https://www.brookings.edu/articles/how-disinformation-defined-the-2024-election-narrative/>.

对公民个人隐私的保护是 AI 时代全球各国都需要回应的问题。欧盟的《通用信息保护条例》（General Data Protection Regulation，以下简称 GDPR）是欧盟保护公民个人数据的有力武器。GDPR 首先对数据实施了分类管控，将生物识别数据、遗传数据等敏感个人数据列为“特殊类别个人数据”，原则上禁止处理此类数据，但列举了十种的例外情形，如基于个人明确同意、公共利益等允许处理此类数据⁹。对于活跃于音视频领域、依赖海量公民生物识别数据的深度伪造技术而言，这一分类规定为公民的个人数据提供了有力保障。同时，GDPR 赋予公民的知情权（获知数据如何被收集、使用）、访问权（免费获取个人数据副本）、更正权（要求修正不准确数据）、删除权（也称被遗忘权，在数据非法处理、目的已达成等情况下，可要求删除数据）、限制处理权（在争议期间暂停数据处理）等权利¹⁰也为公民捍卫自己的合法权益提供了法律依据。

《中华人民共和国个人信息保护法》（以下简称《个人信息保护法》）同样对公民个人信息作出了分类，将与深度伪造相关的生物识别等信息列为敏感个人信息，但其并没有像欧盟一样采取原则上禁止并列举例外情形的方式规制该类数据，而是设定了一定的条件，满足条件即可处理该类数据——只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下，才允许处理敏感个人信息，并要求征得公民个人的单独同意。

欧盟原则上禁止+十种情形例外允许的规制方式给处理个人信息提供了明确稳定的框架，但其较为绝对的明确类型可能会增加企业的合规成本，因此抑制创新。中国满足具备必要性、采取保护措施、征得同意三个条件即可处理敏感个人信息的模式较为泛式，但相较欧盟的模式会更为灵活，一定程度上有利深度伪造技术的发展创新。

2、对数据偏见与算法歧视的规制

改善数据偏见与算法歧视，就需要做到数据公平和算法透明。欧盟的《人工智能法案》（EU AI Act，以下简称 AIA）第十条明确规定了涉及使用数据训练 AI 模型的高风险人工智能系统应当符合的质量标准，并进行了明确列举，该标准要求对于训练模型中所可能存在的违法的歧视与偏见，应当采取措施对该歧视与偏见进行检测、预防并减轻，并强调了数据公平¹¹。同时，AIA 关于技术文档的规定以及要求建立质量管理体系并系统记录的规定实现了算法透明，其中要求公布所使用的模型为数据偏见与算法歧视的溯源提供了依据¹²。这两份文件被要求于该人工智能系统上市或投入使用十年后交给国家主管部门。

中国在《生成式人工智能服务管理暂行办法》中也强调了训练数据的真实性、

⁹ 《General Data Protection Regulation》Article 9.

¹⁰ 《General Data Protection Regulation》Article 15-19.

¹¹ 《General Data Protection Regulation》Article 10.

¹² 《General Data Protection Regulation》Article 13、17.

准确性、客观性、多样性，但对质量标准的规定并未向欧盟一样逐个列举，仅概括性地提出了以上四性的要求¹³。《深度合成管理规定》也有关于加强技术管理、保障训练数据安全的要求，并要求企业自主展开安全评估¹⁴。在提高算法透明度方面，中国制定了事前备案机制。《互联网信息服务算法推荐管理规定》对具有舆论属性或者社会动员能力的服务提供者作出了备案要求，而《生成式人工智能服务管理暂行办法》援引了《互联网信息服务算法推荐管理规定》的备案规定¹⁵。

对于数据偏见和算法歧视这一与技术高度相关的议题，欧盟依然选择在法律中具体列举对于质量标准的要求，而其算法透明的要求高度依赖企业的自主记录，且上交文档的时间要求存在明显的滞后性。中国的规定又稍显笼统，可能会导致企业规避合规义务的情形，但其较为严格的事前备案制度一定程度上又加强了生成式人工智能系统的算法透明，稍微弥补了其模糊标准的缺陷。

3、对虚假信息和非法内容的规制

在深度伪造内容传播前，对深度伪造所生成内容通用的规制方式是设定显性标识义务。深度伪造内容构成虚假信息或非法内容，其中必要的要件之一是内容受众的错误认识。而显性标识义务是指为深度伪造内容附上显著提示，使深度伪造内容隐藏的虚假性外显，从而打破公众的错误认识状态，达到防范风险的目的。欧盟和中国都采取了这一做法。

欧盟的 AIA 法案要求人工智能系统的部署者披露深度伪造内容是人为生成或操纵的¹⁶。而《深度合成管理规定》则要求具有生成或者显著改变信息内容功能的深度伪造服务的提供者，在生成或者编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况¹⁷。

在对虚假信息和非法内容的规制路径上，欧盟与中国所采取的显性标识义务模式较为一致，其目的都是为了捅破深度伪造内容的虚假性，从而导致深度伪造构成虚假信息或非法内容的要件无法实现，进而实现管控风险的目的。

（二）对“传播”阶段法律风险的规制

1、对“传播”行为的规制

当深度伪造内容落入虚假信息或非法内容的范畴时，该内容的传播也必然需要法律规制以尽可能降低其可能造成的损害后果。由于传播需要依托网络媒体平台，那么规范平台主体的义务是监管传播阶段风险的有效手段。以通知删除规则为核心的“避风港”规则无论在中国还是欧盟都对网络间接侵权产生了重要影响¹⁸。

欧盟《数字服务法案》（Digital Services Act，以下简称 DSA）对服务提供商

¹³ 《生成式人工智能服务管理暂行办法》第七条第三款。

¹⁴ 《深度合成管理规定》第十四条、第十五条。

¹⁵ 《生成式人工智能服务管理暂行办法》第十九条。

¹⁶ 《AI Act》Article 50.4

¹⁷ 《深度合成管理规定》第十七条。

¹⁸ 丁晓东：《作为举报治理的通知删除：避风港规则反思》，载《法学论坛》2025年第2期。

设定的避风港规则为：在提供上不实际知情非法内容或在得知非法内容后及时删除的情形下，服务提供商对非法内容不承担责任¹⁹。同时要求服务提供商在服务条款中公开内部投诉处理系统的程序规则²⁰。其侧重点是对用户权利的保护与对平台透明度的关注。

《中华人民共和国民法典》（以下简称《民法典》）在侵权责任编中规定了我国的避风港规则。其首先构建了我国避风港规则的基本情形：网络用户实行侵权行为后，权利人有权通知网络服务提供者，网络服务提供者应当将通知转送相关的用户并及时采取必要措施。未及时采取必要措施的，对损害的扩大部分与用户一起承担连带责任²¹。同时明确了红旗规则，网络服务提供者知道或应当知道用户利用其网络服务侵犯公民合法权益的，应当采取必要措施，否则与用户承担连带责任。

相较于欧盟更加关注用户权利与平台透明度的重点，中国的规定显得更加围绕“平台责任”这一核心。红旗规则使得对平台的责任更加严格，加重了平台的审核义务，国内有一些大型平台（如哔哩哔哩）采取了事前审查的形式，在信息发布前对用户发布的信息进行审核，这其实承担了高于欧盟的事后风险评估的责任。

2、对公民名誉权和肖像权的保护

对公民名誉权和肖像权的这两项基本权利的保护是全球的共识。《欧洲人权公约》（European Convention on Human Rights,以下简称 ECHR）同样规定了公民私人得到尊重的权利。而欧盟的成员国也以其国内法对公民的名誉权和肖像权加以保护。例如德国《民法典》中明确规定故意或过失侵害他人名誉需承担赔偿责任。这些对名誉权和肖像权的专门性规定与 GDPR 和 DSA 中有关深度伪造的规定形成协同作用，一同为公民个人权利提供保护。

中国则以《民法典》为核心强化对公民名誉权和肖像权的直接保护。其中，第一千零一十九条明确规定“任何组织或者个人不得以利用信息技术手段伪造等方式侵害他人的肖像权”，明确禁止深度伪造技术非法使用公民生物特征。第一千零二十四条进一步将“侮辱、诽谤”等行为纳入名誉权侵害范畴，涵盖利用合成内容贬损他人社会评价的情形。针对深度伪造的传播特性，第九百九十七条创新性引入人格权侵害禁令制度，允许受害者在侵权行为持续期间向法院申请责令停止传播²²。同时，《民法典》与《生成式人工智能服务管理暂行办法》形成衔接，也更好地降低了深度伪造所带来的对公民权益的风险。

尽管对公民名誉权和肖像权的保护是国际共识，但欧盟与中国的保护路径呈现出了一定的思维差异。“传播”阶段后对公民名誉权和肖像权的保护其实可以视

¹⁹ 《Digital Services Act》Article 6.

²⁰ 《Digital Services Act》Article 17.

²¹ 《民法典》第一千一百九十五条。

²² 朱嘉珺：《生成式人工智能虚假有害信息规制的挑战与应对——以 ChatGPT 的应用为引》，载《比较法研究》2023 年第 5 期。

为一种事后救济。欧盟层面更倾向于直接针对以数据权利为基础，通过前述的 GDPR、DSA 和 AI Act 构建技术透明度规则，强调平台风险防控责任与个人知情权保障，在事前监管和预防上大做文章，对事后的救济稍显单薄。中国的《民法典》将人格权单独成编，以人格权保护为核心，侧重侵权行为的事后救济与传播阻断。这种差异折射出欧盟“个人赋权”与中国“秩序优先”的治理逻辑分野。

3、对国家安全的保护

国家安全是每个国家的根基，在深度伪造问题上，欧盟构建了以《人工智能法案》（AI Act）和《网络安全法案》（Cybersecurity Act）为核心的风险防控框架，以阻止、减少深度伪造技术对国家安全的冲击。AIA 的四级人工智能风险分类系统首先为国家安全提供了保障，通过分级管理排除了容易对国家安全造成动摇的人工智能系统的使用²³。《网络安全法案》则要求设立欧盟级网络安全认证体系，强制关键数字服务（如 AI 训练平台等）通过 ENISA 认证，确保技术供应链安全²⁴。此外，DSA 也对平台进行分类，要求超大型平台评估算法对选举安全、公共秩序的系统性风险，这一举措也对深度伪造技术可能对国家安全造成的风险进行了有效监管。

相较于欧盟较为分散的法律规制模式，中国则以《国家安全法》《数据安全法》《网络安全法》为基础，构建集中化的深度伪造技术安全治理体系。《国家安全法》明确将“网络与信息安全”纳入国家安全范畴，要求防范利用信息技术危害国家主权与制度安全的行为²⁵。《数据安全法》要求建立数据分类分级保护制度，规定涉及国家安全的核心数据实行严格管制，禁止境外机构未经批准获取重要数据²⁶。《网络安全法》要求关键信息基础设施运营者将境内收集的个人信息与重要数据本地化存储²⁷，深度技术研发机构需通过国家安全审查²⁸。《生成式人工智能服务管理暂行办法》进一步对生成式人工智能提出了细化要求，即要求生成式人工智能不得生成含危害国家安全内容。通过多部门法、从总体到细分领域为国家安全的提供稳定坚固的法律保障框架。

在国家安全这一重要问题上，欧盟选择以风险分级为核心，通过分级管理其认为可能产生国家安全风险的平台。而中国则延续一贯重视国家安全的立场，采取“安全优先”原则，通过立法明确技术研发与数据流动的禁区，依托行政审查与强制本地化措施集中管控风险。二者虽均聚焦技术供应链安全，但欧盟侧重市场主体的自我规制能力，中国更依赖国家主导的穿透式监管。

对上述两个不同阶段中欧的规制模式进行对比，不难看出二者对于深度伪造

²³ 《AI Act》 Article 5.

²⁴ 《Cybersecurity Act》 Article 6.

²⁵ 《国家安全法》第二十五条。

²⁶ 《数据安全法》第二十一条。

²⁷ 《网络安全法》第三十七条。

²⁸ 《网络安全法》第三十七条。

全生命流程的监管与规制均有自己的模式，且都布局到了相似的权利领域。但也可以从其中总结出二者在三个维度上的差异：一是欧盟相较于中国更加强调技术透明度，关注企业合规；二是欧盟更加关注公民个人（用户）的权利，与其强调人权的立场相一致；三是欧盟更加关注事前审查合规，中国事后救济的规定比欧盟完善。

四、中欧治理路径的困境与中国因应

中欧在深度伪造法律规制上的差异化路径，映射出数字时代技术治理的复杂性与多元性。尽管双方均致力于遏制深度伪造的滥用风险，但其规制逻辑的差异亦暴露出三方面亟待解决的深层问题：一是监管与创新的平衡困境，二是跨境规制冲突的协调难题，三是国际合作机制的缺位。

（一）打破监管与创新的动态平衡困境

在 AI 监管领域，监管与创新的平衡是无法逃避的课题。欧盟以高合规标准确保技术透明度，但其具体列举的各类情形由于其限制性会导致企业合规成本的显著增加，从而抑制技术迭代活力。但考虑到中国的监管实效和合规模糊化问题，具体的人工智能系统分级分类管理和数据分级分类管理是需要的，但在具体的规制领域可以避免采取欧盟较为绝对化的举例分类形式。对中国而言，维持监管与创新的动态平衡，可以对医疗、金融等高敏感领域实施严格准入与实时监测，而对艺术创作、教育工具等低风险场景引入“监管沙盒”，允许在可控范围内试错创新，从而在安全底线与创新激励间寻求动态平衡。

（二）化解跨境规制冲突的协调难题

中欧在数据主权与平台责任上的规则冲突，加剧了跨国企业的合规困境。欧盟的 DSA 通过域外效力²⁹强制全球平台遵守其透明度规则，而中国《数据安全法》并不绝对允许数据出境，而是要求满足一定条件³⁰。二者在数据跨境流动与平台责任分配上形成对立。以深度伪造检测技术为例，欧盟要求公开模型训练数据来源³¹，而中国《网络安全法》第三十七条的本地化存储要求可能限制数据共享，导致技术协同受阻。为此，中国可主动探索“等效互认+负面清单”机制：在数据分类标准、深度伪造显性标识技术等领域推动互认协议，同时以负面清单明确不可妥协的安全红线（如军事数据、选举干预内容等）。

（三）填补全球化治理与国际合作的缺位

深度伪造的跨国传播与暗网匿名性，使得单一司法辖区的规制效力显著受限。当前中欧均侧重内部治理，缺乏国际协同规则与执行机制。例如，欧盟的“人工合成”标识（与中国“AI 生成”标签的技术标准尚未统一，导致跨境平台需重复标注；针对暗网深度伪造犯罪，双方亦未建立联合溯源与证据交换机制。破解此困

²⁹ 《Digital Services Act》Article 3.

³⁰ 《数据安全法》第三十一条。

³¹ 《AI Act》Article 13.

局需依托多边平台（如联合国、G20）推动《全球人工智能伦理公约》的制定，明确深度伪造的“技术滥用清单”（如伪造领导人声明、军事指令），并共建开源检测工具库与跨境司法协作网络。同时，可成立国际深度伪造治理联盟，通过定期技术峰会共享风险案例与治理经验，例如欧盟 ENISA 与中国网信办联合发布跨境威胁预警，实现从“孤立防御”到“协同共治”的转型。

五、结语

中欧对深度伪造的法律规制路径，深刻体现了数字时代技术治理的复杂性与多元性。欧盟以《人工智能法案》与《数字服务法案》为核心，构建了“技术透明-权利保障”的治理框架，通过数据主权与算法可追溯性平衡技术风险与市场创新；中国则以《生成式人工智能服务管理暂行办法》、《网络安全法》、《深度合成管理规定》为基础，形成“穿透式监管-安全优先”的集中化模式，依托数据本地化与分类分级制度强化技术可控性。二者虽均聚焦“生成—传播”全链条风险防控，但在价值逻辑与工具选择上存在根本差异：欧盟以“个人赋权”为核心，通过透明度规则与平台问责制分散风险责任；中国以“秩序优先”为导向，通过行政审查与强制合规集中管控风险。

然而，面对深度伪造技术的快速迭代与全球化挑战，中欧模式均暴露了其局限性。欧盟的高合规成本可能抑制企业创新活力，而中国的概括性标准易导致合规虚置。更严峻的是，跨境规制冲突加剧了跨国企业的法律困境，而国际合作机制的缺位进一步削弱治理效能。

未来深度伪造治理需超越“安全管控”与“技术自由”的二元对立，转向“韧性治理”范式：其一，构建动态分类治理机制，对医疗、金融等高敏感领域实施严格准入，对低风险场景采用“监管沙盒”允许可控创新；其二，推动跨境规则互认，在数据分类、标识技术上建立等效协议，以负面清单明确安全红线；其三，依托联合国、G20 等平台制定《全球人工智能伦理公约》，明确“技术滥用清单”，并共建开源检测工具库与跨境司法协作网络。

中欧规制路径的互补性为全球治理提供了新思路：欧盟的算法可追溯性可增强中国事前和事中监管的精准度，而中国的穿透式审查能补足欧盟事后救济的滞后性。进一步探索“监管科技”在风险预警中的应用，或基于区块链构建不可篡改的内容溯源网络，或将成为平衡技术创新与法律价值的关键路径。唯有在核心权益保障基础上，实现敏捷立法、技术赋权与国际协同，方能应对深度伪造这一复合体带来的时代挑战。

网络空间国家行使预先性自卫权的国际法探析

——基于对各国国家立场文件的分析

张宇星³²

摘要：近年来，随着网络攻击的复杂性、隐秘性及破坏性显著增强，一些网络空间中的国家立场声明、《塔林手册 2.0》以及部分国际法学者提出，可以在网络空间中针对尚未实际发生但迫在眉睫的网络攻击行使预先性自卫权。然而，将传统武力攻击情境中本就存在争议的预先性自卫引入网络空间缺乏合理性和合法性。在技术层面上，由于网络攻击具有隐秘性、情报获取难度较大，难以收集准确清晰的证据，因此对尚未实际发生的网络攻击进行评估不具有现实性，难以进行实践，预先性自卫的理论仅仅是“纸上谈兵”。在法律层面上，网络空间缺乏边界和匿名性使得国家归因难以实现，并且判断网络攻击的诸如“迫在眉睫”“最后机会之窗”等紧迫性标准具有主观预测性、模糊性和政治性，极易产生自卫权的滥用风险。中国始终坚持《联合国宪章》中第 2 条第 4 款规定的“禁止使用武力”原则，反对在网络空间中行使预先性自卫，主张发挥联合国安理会在应对网络攻击方面的作用，以遏制网络空间的军事化。

关键词：网络空间；预先性自卫；迫在眉睫；武力攻击

近年来，随着技术的发展，网络攻击呈现出更强的破坏力、隐蔽性和复杂性，对各国的网络安全构成了前所未有的严峻挑战。2013 年，北约卓越合作网络防御中心（NATO Cooperative Cyber Defence Centre of Excellence）组织专家组发布了《国际法在网络战中的适用塔林手册》，其中规则第 15 条明确规定：“当网络武力攻击已经发生或迫近时，可使用武力行使自卫权。”2017 年，迈克尔·施密特教授主编的《塔林手册 2.0 版》沿用了上述条款（规则第 73 条），进一步强调即使《联合国宪章》第 51 条未明确授权对预期攻击采取自卫行动，一国也不应被动等待敌方攻击，而应在攻击迫在眉睫时即行使自卫权。³³为避免遭受网络攻击的国家在实施自卫行动之前就已经遭受到无法补救的损失，以美国、以色列等为代表的部分国家开始推动自卫权在网络空间的扩大适用，发布国家立场声明，明确支持国家在网络空间行使预先性自卫权（Anticipatory Self-Defense）。

值得思考的是，当前美西方国家针对网络空间行使“预先性自卫权”提出了许

³² 武汉大学法学院硕士研究生，15274836826 1404591149@qq.com

³³ 参见[美]迈克尔·施密特主编：《网络行动国际法塔林手册 2.0 版》，黄志雄等译，社会科学文献出版社 2017 年版，第 62-63 页。

多看似合理的法律依据，试图在网络空间规则博弈中，推动有利于其国家利益的规则构建。然而，预先性自卫权作为一种带有积极攻击性质的权利，并非仅仅是对网络攻击的消极防御，是否会造成网络空间军事化？网络空间中行使预先性自卫权的理论仍然存在较强的主观预测性、模糊性和政治性，是否极易导致武力的滥用？鉴于此，有必要对各国发布的国家立场、预先性自卫权的基本理论进行全面而深入的分析，揭示预先性自卫在网络空间中适用的理论缺陷与潜在风险，并进一步阐明和提出相应的中国立场与对策。

一、各国对于网络空间行使预先性自卫权的立场分析

自 2018 年以来，越来越多的主权国家开始在网络空间规则博弈的背景下争夺“话语权”，积极地通过书面文件、高官演讲等形式参与网络空间国际规则的制定。³⁴鉴于国际习惯法形成是国际造法的重要路径之一，各国立场文件中的自卫权主张尤其值得深入探讨。

（一）支持行使“预先性自卫权”的国家

“预先性自卫权”的概念在以美国、以色列、英国、澳大利亚、法国、德国等多个西方国家的官方立场文件或者声明中得到了支持。

作为在网络空间占主导地位的大国之一，美国持续推进诉诸武力的法律在网络空间中的适用，并试图扩大自卫权的适用前提。最早在 2012 年 9 月，时任美国国务院法律顾问高洪柱发表了题为《网络空间国际法》的演讲，³⁵明确表示了美国对“预先性自卫权”的支持立场。在回答“一国是否可以通过行使国家自卫权来应对计算机网络攻”的问题时，他主张：“《联合国宪章》第 51 条所承认的一国的国家自卫权，可以被相当于武力攻击或者迫在眉睫的武力攻击威胁的计算机网络活动所触发。”

在本次演讲之后，2016 年布莱恩·依根也在其演讲中表明国家立场：“在某些情况下，一国可以采取本来是违反国际法的行动来应对恶意网络活动。一个例子是使用武力进行自卫以应对实际或迫在眉睫的武力攻击。”³⁶之后在 2020 年，美国国防总法律顾问保罗·内伊发表演讲时³⁷表示“在行使其固有的自卫权时，一国可以使用必要和相称的武力来应对实际或迫在眉睫的武力攻击。这一点在网络背景下与在任何其他背景下是一样的。”从这三次演讲中可以看出，美国一直支持对“迫在眉睫”的网络攻击行使自卫权，并主张预先性自卫权的行使是出于维护国

³⁴ 参见黄志雄，罗旷怡，《各自为“辩”：网络空间新近国家立场声明的总体考察与中国因应》，载《云南社会科学》2023 年第 6 期，第 91-104 页。

³⁵ International Law in Cyberspace, 18 September 2012, available at <https://2009-2017.state.gov/s/l/releases/remarks/197924.htm>.

³⁶ 美国国务院法律顾问布莱恩·依根（Brian J. Egan）2016 年 11 月 10 日在其母校伯克利大学发表的题为《国际法与网络空间的稳定》的演讲。该演讲英文版见 <https://www.justsecurity.org/wp-content/uploads/2016/11/Brian-J.-Egan-International-Law-and-Stability-in-Cyberspace-Berkeley-Nov-2016.pdf>.

³⁷ 美国国防部总法律顾问保罗·内伊（Paul Ney）2020 年 3 月 2 日在美国网络司令部年度会议上发表的主旨演讲。该演讲英文版见 <https://www.defense.gov/News/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/>.

家安全的需要。这一立场与美国在网络空间一直实行的“先发制人的战略”相符合，呈现“扩张性”“转守为攻”的趋势。美国一直以来妄图以“预先性自卫”“先发制人”的自卫模式，对其他国家发起所谓的网络武力自卫攻击，实际上是为了维护美国在网络空间的领导性地位和绝对优势。

除了美国以外，澳大利亚和英国也相继对“预先性自卫权”的理论表示支持，并进行了较为详细的理论阐释。英国司法部长、王室法律顾问杰里米·赖特³⁸在其2018年首次公开表达英国对国际法适用于网络空间的立场，他指出“如果网络行动造成与武力攻击同等规模的死亡和破坏，或构成迫在眉睫的此类威胁，就会产生《联合国宪章》第51条所确认的采取自卫行动的固有权利。”他认为当代有关自卫权的法律体系亟需调整，并指出自己早在2017年于伦敦国际战略研究所的演讲中，已就“当代自卫权”的核心问题进行了详尽论述。³⁹

2017年4月11日，澳大利亚司法部长、参议员乔治·布兰迪斯在昆士兰大学就“国际法中对迫在眉睫的武装袭击的自卫权”发表了公开演讲：“当攻击方明确致力于发动武力攻击、而受害方不采取行动就会失去有效自卫的最后机会时，该国可对武力攻击采取先发制人的自卫行动。”⁴⁰这一观点也在澳大利亚2021年的官方国家立场文件中被重申。⁴¹澳大利亚司法部长在演讲中表示其发表的意见是对于杰里米·赖特2017年演讲的重述与补充，因此在对于“预先性自卫权”合理性的论证上，不难发现这两个国家都是基于以下几点：1、“卡罗琳号”事件已经逐渐发展成为习惯法。2、随着科技发展，新的威胁例如恐怖主义、极端主义不断涌现，国际法应当随着国家实践的发展和不断变化的情况而演变。3、采取自卫行动并不意味着一个国家被动等待攻击。4、均认可“伯利恒原则”所述因素可以作为一个评估预先性自卫权的框架。

法国国防部2019年报告指出：“在例外情形下，法国允许对‘尚未实施但即将实施的、紧迫且确实的、潜在影响足够严重的’网络攻击作出预先性自卫（*légitime défense préemptive*）。”⁴²德国也在2021年的国家立场文件中多次表明支持《塔林手册2.0》的观点：“《联合国宪章》第51条要求一个国家可以诉诸自卫的攻击是“迫在眉睫”的。这同样适用于针对恶意网络行动的自卫。”⁴³

³⁸ 英国司法部长、王室法律顾问杰里米·赖特（Jeremy Wright QC）议员于2018年5月23日发表的题为《21世纪的网络与国际法》的主旨演讲。该演讲英文版见

<https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>.

³⁹ The Modern Law of Self-Defence, 11 January 2017, available at <https://www.ejiltalk.org/the-modern-law-of-self-defence/>

⁴⁰ Developments in International Law: Self-Defence Against Imminent Armed Attack, 11 April 2017, available at <https://law.uq.edu.au/event/session/6404>.

⁴¹ 澳大利亚2021年立场文件。该文件是联合国政府专家组2021年报告（A/76/136）附件（主题为“国际法如何适用于各国使用信通技术”的国家自愿提交文件正式汇编）的一部分。

⁴² 法国使用了“preemptive self-defence”这一不太贴切的英文措辞，但从其表述内容来看，法国实质上是在阐释预先性自卫。See France’s Position Paper, para.1.2.2.

⁴³ On the Application of International Law in Cyberspace: Position Paper, March 2021, [https://www.auswaertiges-amt.de/blob/2446304/32e7b2498e10b74fb17204c54665bdf0/on-the-application-of-international-law-in-cyberspacedata.pdf\(hereinafter‘Germany’s Position Paper’\)](https://www.auswaertiges-amt.de/blob/2446304/32e7b2498e10b74fb17204c54665bdf0/on-the-application-of-international-law-in-cyberspacedata.pdf(hereinafter%20Germany%20s%20Position%20Paper)), p.16.

（二）未表态支持“预先性自卫权”的国家

对于是否能在网络空间中针对“迫在眉睫”的攻击进行自卫的问题，仍有欧盟、非洲、中国等大部分的地区或国家持消极反对的态度。部分国家虽然没有明确提及“预先性自卫权”的概念并进行批判，但会表示应严格按照《联合国宪章》第 51 条援引单独或集体自卫的权利，即仅能在武力攻击实际发生时行使自卫权。

例如，2019 年，爱沙尼亚总统 Kersti Kaljulaid 在网络冲突国际会议（International Conference on Cyber Conflict 2019）开幕式上发表演讲，表示：“国家有权对恶意网络行动作出回应，包括使用外交应对和反措施，必要时还可行使固有的自卫权。”⁴⁴2024 年 11 月，欧盟发布《欧洲联盟及其成员国关于将国际法应用于网络空间的共同谅解的宣言》，⁴⁵提出了有关自卫权的规定：“遭受构成武装攻击的网络行为的国家，可以根据《联合国宪章》第 51 条的规定，行使其固有的个体或集体自卫权。”

2024 年 1 月，非洲联盟在《关于在网络空间使用信息和通信技术适用国际法的非洲共同立场》（CAP）⁴⁶也提出“非盟成员国强调，从法律角度来看，《联合国宪章》第 51 条允许各国在针对联合国会员国‘发生武力攻击时’使用武力进行单独或集体自卫。”非洲联盟指出，各国针对迫在眉睫的武力威胁行使自卫权的观点尚存争议，司法判例和专家意见也未达成一致。非盟强调此问题需要进一步深入审议，尤其要考虑网络空间的特殊性及其可能对禁止使用武力原则完整性的影响。此外，非盟强调《联合国宪章》第 51 条明确规定自卫权仅适用于实际发生的武力攻击，并主张对自卫权的适用进行严格限制性解释，以有效维护国际和平与安全。因此可知，实际上有大部分多的国家仍然反对“预先性自卫权”的理论，认为会对。

2024 年，古巴也在其国家立场文件中表明，网络空间武力行使应采取谨慎的态度：“由于网络空间缺乏明确的监管和统一的概念，不能因此就将“网络攻击”定义为“武装攻击”，从而不合理地扩展这一概念，以此为侵略行为辩护，借口是所谓的合法自卫权。”⁴⁷

（三）各国立场比较分析

⁴⁴ 爱沙尼亚总统 Kersti Kaljulaid 于 2019 年 5 月 29 日在塔林举办的 2019 年网络冲突国际会议（CyCon 2019）开幕式上发表的演讲。该演讲英文版见 <https://ceipfiles.s3.amazonaws.com/pdf/CyberNorms/LawStatements/Remarks+by+the+President+of+the+Republic+of+Estonia+at+the+Opening+of+CyCon+2019.pdf>。

⁴⁵ Declaration by the European Union and its Member States on a Common Understanding of the Application of International Law to Cyberspace, available at <https://data.consilium.europa.eu/doc/document/ST-15833-2024-INIT/en/pdf>。

⁴⁶ The Common African Position on the Application of International Law to the Use of Information and Communication Technologies in the Cyberspace, available at <https://ssrn.com/abstract=4714756>。

⁴⁷ Position Paper of the Republic of Cuba on the Application of International Law to Information and Communication Technologies in Cyberspace, 28 June 2024, available at [https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-__\(2021\)/Documento_de_posici%C3%B3n_de_Cuba._Aplicaci%C3%B3n_del_Derecho_Internacional_a_las_TIC](https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-__(2021)/Documento_de_posici%C3%B3n_de_Cuba._Aplicaci%C3%B3n_del_Derecho_Internacional_a_las_TIC)

通过统计可以对比出各国对于网络空间行使“预先性自卫权”的立场表现出显著的分歧。支持“预先性自卫权”的国家，如美国、英国、澳大利亚等，更强调国际法需要适应网络空间中的新威胁，倾向于“向前防御”。与此相对，非洲和欧盟大部分国家则更强调在网络空间也需要严格遵循《联合国宪章》的规定，防止自卫权的滥用，避免网络空间成为武力冲突的新领域。总体而言，仅有小部分国家明确支持预先性自卫权，大部分国家/地区仍未表态或者持否定态度。

表 1 国家/地区对于网络空间行使“预先性自卫权”的立场文件统计
(截至 2025 年 3 月)

年份	支持行使预先性自卫权的国家立场文件		未表态支持预先性自卫权的国家立场文件	
	数量	国家/地区名称	数量	国家/地区名称
2012	1	美国	0	
2016	1	美国	0	
2018	1	英国	0	
2019	1	法国	1	爱沙尼亚
2020	3	美国、新西兰、以色列	3	芬兰、捷克、伊朗
2021	5	巴西、德国、英国、澳大利亚、美国	12	爱沙尼亚、俄罗斯、哈萨克斯坦、荷兰、肯尼亚、罗马尼亚、挪威、日本、瑞士、新加坡、意大利、中国
2022	1	英国	3	波兰、瑞典、加拿大
2023	0		4	爱尔兰、丹麦、巴基斯坦、哥斯达黎加
2024	0		5	欧盟 ⁴⁸ 、非盟、奥地利、古巴、捷克
2025	0		1	哥伦比亚

二、预先性自卫权的理论概述

作为《联合国宪章》第 2 条第 4 款所确立的禁止使用武力原则之例外，国家自卫权的行使构成国际法框架下合法使用武力的情形。《宪章》第 51 条规定：“联

⁴⁸ 欧盟在宣言附录中列出了一套非详尽的法律要素，称这些要素是对欧盟成员国当前和未来的国家立场的补充，并不影响国际法适用于网络空间的共识的未来发展。

联合国任何会员国受武力攻击时，在安全理事会采取必要办法，以维持国际和平及安全以前，本宪章不得认为禁止行使单独或集体自卫之自然权利。”⁴⁹自卫权作为国家固有的权利，是指国家在遭到外来侵犯时，单独或者与其他国家共同抵抗侵略的权利。国际法院在尼加拉瓜诉美国案中明确了自卫权已成为一项习惯国际法。

50

对于网络空间而言，在联合国信息安全政府专家组(UNGGE)2015年与2021年两份报告中，专家组都强调了《联合国宪章》内容将适用于网络空间，指出各国拥有采取符合国际法和得到《宪章》承认的措施的固有权利，即自卫权。⁵¹网络攻击如果导致大规模人员伤亡或物理破坏，构成了联合国宪章第51条所界定的“武装攻击”，那么将为国家在网络空间中行使自卫权提供法律依据。⁵²预先性自卫权是关于上文所述自卫权的延伸概念。

(一) 概念厘清

当前学界对“预先性自卫权”这一概念尚未形成统一的定义，例如，在 Malcom N. Shaw 的《国际法》一书中，将“先发制人的自卫”的概念与“预防性自卫”作为等同概念分析。⁵³在《武力攻击的紧迫性问题——国家援引自卫权应对紧迫性武力攻击的晚近实践》一文中，学者将“预先性自卫权”等同于“先发制人自卫”。⁵⁴因此本节将对预先性自卫权的概念进行厘清，以推动后续论证的展开。

根据对预先性自卫权扩大适用的程度差异，国际法学者提出了与预先性自卫权极为相近的几个概念，包括拦截性自卫权 (Interceptive Self-Defense)、预先性自卫权 (Anticipatory Self-Defense)、先发制人自卫权 (Preemptive Self-Defense) 与预防性自卫权 (Preventive Self-Defense)。其中，以色列的国际法学者约拉姆·丁斯坦 (Yoram Dinstein)⁵⁵ 提出的拦截性自卫权是指在武装攻击尚未完全实施、但已经启动且不可逆转的情况下，受害国提前采取军事行动，对即将到来的攻击进行拦截，以防止其达到最终目标，他认为拦截性自卫权是《在联合国宪章》第51条中自卫权的语境之下的。这一概念是自卫权中最小程度的扩大适用。

对于“预先性预先性自卫权”和“先发制人的自卫权”这两个极为相近的概念的区分，Geoffrey S. DeWeese 指出⁵⁶根据美国陆军《武装冲突法手册》(Law of

⁴⁹ 《联合国宪章》第51条。

⁵⁰ Nicaragua v. US, ICJ Reports, 1986.

⁵¹ See Open-ended Working Group on developments in the field of information and telecommunications in the context of international security: Final Substantive Report, 12 March 2021, UN Doc. A/AC.290/2021/CRP.2.

⁵² Hayward, R.J, Evaluating the "imminence" of a cyber attack for purposes of anticipatory self-defense, Columbia Law Review, Vol.177:399, p.399-433 (2017).

⁵³ Malcom N. Shaw, International Law, (9th Edition, Cambridge University Press, 2021), P. 2430.

⁵⁴ 黄瑶，陈虹霏，《武力攻击的紧迫性问题——国家援引自卫权应对紧迫性武力攻击的晚近实践》，2023年第8卷第1期。

⁵⁵ Yoram Dinstein, War Aggression and Self-Defence, Cambridge University Press, 2017.

⁵⁶ G. S. DeWeese, Anticipatory and preemptive self-defense in cyberspace: The challenge of imminence, 2015 7th International Conference on Cyber Conflict: Architectures in Cyberspace, Tallinn, Estonia, p. 81-92(2015).

Armed Conflict Deskbook) 的定义,⁵⁷预先性预先性自卫权是指“在迫在眉睫的武装攻击发生之前使用武力(using force in anticipation of an imminent armed attack)”,而先发制人的自卫权则被视为在预先性预先性自卫权这个宽泛概念下的子集(subset)。Gill 和 Ducheine 不赞成这种观点,⁵⁸认为这两个术语含义一致。David Sadoff 则认为,⁵⁹两者的主要区别体现在时间维度上,即预先性预先性自卫权适用于攻击已具备明确的能力和意图、即将发生的情况,而预防性自卫权更多地基于对未来可能受到攻击的恐惧,即攻击尚未迫在眉睫,但国家可能在不久的将来成为武装攻击的目标。这种观点与 Michael Reisman 的理论相似,⁶⁰他指出:“那些主张预先性自卫权的人能够指向一个清晰可见的、即将发生的威胁,而先发制人的自卫权则更多依赖于‘可能性’,即仅基于一种潜在的推测。”因此,两者的主要区别在于威胁的紧迫程度不同,即采取防御行动所依据的威胁是否足够具体和迫近。

多数国家在其立场文件中普遍表明不接受“预防性自卫权”理论。该概念被界定为针对非迫在眉睫的威胁采取的措施,并被视为国际法下的非法行为。相较于仅应对即时威胁的先发制人自卫权,预防性自卫权的适用边界显著扩大,其预设的防御时间轴可延伸至数年之久。该理论所针对的威胁形态具有高度不确定性——既包括尚未成形的潜在风险,也涵盖未来未知时间可能发生的假设性攻击。预防性自卫权的核心特征在于脱离实际威胁的实证基础,转而依据对未来风险的主观预判。正如迈克尔·沃尔泽(Michael Walzer)所揭示:“预防性战争的决策逻辑基于对危险的标准化评估,但这种标准既非源自客观现实,也不受领土安全紧迫性的制约。它本质上根植于决策者的主观认知,反映着权力制衡的思维定式。”⁶¹

总结而言,本文认为这四个概念各不相同,预先性自卫权的基本内涵是一国对尚未实际上开始、但可以合理地认为已迫在眼前的武力攻击,预先使用武力、行使自卫权的行为。⁶²

(二) 来源、实践与发展

预先性自卫权最初被提出是为了适应不断更新的现代化武器装备,很有可能让被攻击的国家在自卫行动开始之前就已经遭受到无法补救的损失。⁶³预先性自卫权理论的起源可以追溯至 19 世纪的卡罗琳号事件(Caroline case)。1837 年,英国军队为阻止美国支持加拿大叛乱者,越境袭击了停泊在尼亚加拉河的美国蒸

⁵⁷ Law of Armed Conflict Deskbook (William J. Johnson & David H. Lee, editors, 2014).

⁵⁸ Terry D. Gill and Paul A.L. Ducheine, Anticipatory Self-Defense in the Cyber Context, 89 International Law Studies 438, 452-53 (2013).

⁵⁹ David A. Sadoff, A Question of Determinacy: The Legal Status of Anticipatory Self-Defense, 40 GEO. J. INT'L L. 523, 530 (2009).

⁶⁰ W. Michael Reisman & Andrea Armstrong, The Past and Future of the Claim of Preemptive Self-Defense, 100 A.J.I.L., P. 526 (2006).

⁶¹ Michael Walzer, Just and Unjust Wars: A Moral Argument With Historical Illustrations (4th ed., 2006).

⁶² 参见[英] 詹宁斯, 瓦茨修订:《奥本海国际法》, 王铁崖, 陈公绰, 汤宗舜, 周仁译, 北京: 中国大百科全书出版社, 1995 年版, 第 310 页。

⁶³ 贾兵兵,《国际公法: 平时时期的解释与适用(第二版)》, 清华大学出版社。

汽船卡罗琳号。对于这一事件，美国抗议英国侵犯美国的领土主权，而英国辩护它的行为出于自卫的必要。美国国务卿丹尼尔·韦伯斯特（Daniel Webster）在此背景下提出了自卫权仅适用于“仅当自卫的必要性是迫在眉睫的、压倒一切的，且别无他法可选、无暇考虑时”的情况。⁶⁴

1967 年第三次中东战争期间，以色列以先发制人打击的方式袭击了埃及军队，因为埃及、叙利亚等国家已经在以色列边境集结了军队。以色列政府宣称，这场攻击是自卫性质的，是为防止阿拉伯军队入侵所必需的。安理会和联合国大会都没有谴责以色列“侵略”行为。1981 年 6 月 7 日，以色列空军对伊拉克境内正在建设的奥西拉克核反应堆发动空袭，以消除伊拉克未来可能拥有的核威胁。以色列在安全理事会声称：“采取消除这个危及以色列生存的可怕的核威胁的举措，以色列不过是在行使国际法以及《联合国宪章》中关于自卫的合法权利。”

2001 年 9·11 恐怖袭击发生后，布什总统在 2002 年《美国国家国家安全战略报告》中指出“我们必需调整迫在眉睫的威胁的概念，使之于今天对手的能力与目标相适应，美国长期保留采取预防性行动的权利，以消除对我们国家安全的足够威胁。”⁶⁵首次把“先发制人”战略正式纳入美国 21 世纪国家安全战略之中，强调美国将在威胁完全形成之前就采取行动，向恐怖分子和敌对国家发动主动进攻，消除威胁，确保美国绝对安全。此后的 2006 年《国家战略新版本》继续强调这一理论。⁶⁶

在 21 世纪，各类网络安全威胁不断增长，以电脑病毒和其他恶意软件为“武器”的网络攻击逐渐成为一种新的战争形态或作战手段，⁶⁷网络攻击往往具有隐蔽性、速度快等特点。对此，著名学者 Michael Schmitt 在文章中指出，“网络行动加剧了采取“最后机会窗口”策略的动机。这类行动通常隐蔽进行、瞬间发生，并可能引发灾难性后果；而且网络攻击手段相对容易获取且分布广泛。鉴于这一作战现实，遵循传统的时间标准既不合理，也极其危险。毕竟，从决定发动网络武装攻击到其执行及后果显现之间，可能仅以毫秒计；因此，为使自卫权真正发挥作用，国家必须在得知攻击即将发生时迅速采取有力措施，否则一旦犹豫，就可能失去有效防卫的机会。”⁶⁸2018 年，《塔林手册 2.0》的规则第 73 条规定：“当网络武力攻击已经发生或者迫近时，可使用武力行使自卫权”，国际专家组明确提出支持预先性自卫权，认为“即使《联合国宪章》第 51 条没有明确规定可以对即将发生的武力攻击采取自卫行为，一国也不必任凭敌方准备攻击而坐视不理。”

⁶⁴ Letter from Daniel Webster to Lord Ashburton (Aug. 6, 1842), in 2 INTERNATIONAL LAW DIGEST 412 (John Bassett Moore ed., 1906).

⁶⁵ The White House, The National Security Strategy of the United States of America 15 (Sept. 2002).

⁶⁶ The White House, The National Security Strategy of the United States of America 18 (March 2006).

⁶⁷ 黄志雄，《国际法视角下的“网络战”及中国的对策——以诉诸武力权为中心》，2015 年，第 37 卷第 5 期，第 145 页。

⁶⁸ Schmitt, Michael N., The Law of Cyber Conflict: Quo Vadis?, Stanford Law & Policy Review, Vol. 25:269, p. 286 (2022).

（三）紧迫性标准

在探讨预先性自卫权合法性问题时，一个核心争议在于如何解释“紧迫性”这一关键标准。国际法理论和实践中目前存在几种明显不同的观点。

一种观点认为，自卫行动的实施必须严格限于敌对方攻击即将实施的瞬间（the attack is about to be launched），即只有当攻击几乎已经启动、尚未实际发生的短暂时刻，才能合法采取预防性行为。⁶⁹若将这一严格标准应用至网络领域，意味着只有在攻击者即将发起网络攻击（如攻击指令即将被执行的时刻），受威胁国家方才具备合法行使自卫权的条件。然而，这种过于狭窄的解释可能使自卫权在网络攻击环境中难以实施。鉴于网络攻击通常发生迅速，从攻击者点击按钮到实际攻击造成后果之间，几乎不存在足够的反应空间。这种情况类似于等待袭击者手持炸弹、即将引爆装置之时才允许采取防卫行动，使自卫行动几乎无法有效展开。因此，《塔林手册》的专家组多数成员拒绝接受这种极为严格的解释方式。

另一种更具灵活性的观点则主张，应将“紧迫性”理解为“最后机会之窗”（the last possible window of opportunity）。⁷⁰在该理论下，自卫行动不需要等到攻击即将启动的最后时刻，而是可以更早地在攻击发生前的某一阶段实施——这一窗口可能非常靠近攻击启动前，也可能距实际攻击尚有相当一段时间，关键在于能否有效阻止攻击的发生。在网络冲突情境下，这意味着潜在受害国必须基于有限信息对威胁进行风险评估，包括：（1）攻击方实际发动攻击的可能性；（2）攻击造成严重损害后果的可能性；（3）阻止该攻击的有效时机何时关闭。这种观点虽然为网络环境中的预先性自卫权提供了一定法律空间，但其主观性和不确定性也增加了国际法适用的复杂性和被滥用的风险。

近年来，更进一步延伸的所谓“延长的紧迫性”（elongated imminence）观点也有所出现。⁷¹该观点认为，自卫行动的时机甚至可以提前至攻击计划的准备阶段，例如不必等到恐怖分子登机实施袭击，而是在发现恐怖分子开始制造自杀炸弹背心时即进行反制。然而，这种所谓“延长的紧迫性”本质上并未创造新的法律标准，而是对上述“最后机会之窗”理论的一种极端延伸。其争议之处在于，这种过度宽泛的标准易造成主观判断泛滥，显著降低了自卫行动的确定性和可控性，从而增加了国际社会对自卫权被滥用的担忧。

三、网络空间行使预先性自卫权的内在缺陷

本节将探讨网络空间行使预先性自卫权存在的缺陷：在技术层面，由于网络攻击具有高度隐秘性，且情报获取困难，往往难以收集到准确清晰的证据。因此，

⁶⁹ Brian T. O'Donnell, Michael N Schmitt, Computer Network Attack and International Law, Naval War College, 1999, at 930.

⁷⁰ Michael N. Schmitt, Cyber Operations and the *Jud Ad Bellum* Revisited, Villanova Law Review, 592 (2011).

⁷¹ Daniel Klaidman, Kill or Capture: The War on Terror and the Soul of the Obama Presidency, 219 (2013).

对尚未实际发生的网络攻击进行评估缺乏现实性，难以付诸实践，预先性自卫理论因此显得空洞无力，只是“纸上谈兵”。在法律层面，网络空间缺乏明确的边界和匿名性使得国家归因变得极为困难。此外，判断网络攻击是否具备“迫在眉睫”或“最后机会之窗”等紧迫性标准往往依赖于主观预测，具有模糊性且带有一定的政治色彩，这使得自卫权的行使容易存在滥用风险。

（一）技术缺陷

在现代科学技术发展的推动下，网络空间（Cyberspace）逐渐成为继陆地、海洋、天空、外空之外的第五空间，与人类活动的现实空间深度融合。互联网、电信网络、计算机系统等组成了网络空间中信息技术基础设施。在 21 世纪，各类网络安全威胁不断增长，以电脑病毒和其他恶意软件为“武器”的网络攻击逐渐成为一种新的战争形态或作战手段。⁷²网络攻击旨在改变或摧毁目标计算机中包含的信息，目的是使对方的指挥、控制和通信系统失效，对目标之外的事物造成损害。具体来说，网络攻击的目标可能包括通过制造网络流量过载或迫使网络中断，从而阻止用户的正常访问；或是渗入网络系统，监视其运行情况并窃取敏感数据；亦或入侵系统，篡改并操纵存储于其中的信息。⁷³本节对网络攻击的特征、使用情境进行分析，从网络攻击难以提前预测和难以溯源展开论述。

1. 网络攻击难以预测

网络攻击一般可以分为单独的网络攻击和与传统动能武力共同使用的网络攻击两种情形。对于单独的网络攻击，大多数都不太可能直接造成严重的物理伤亡，或对一个国家的关键基础设施造成长时间的显著破坏。进行如此大规模的网络武装攻击并不是轻易能够实现的，需要进行大量的努力，涉及相当长的时间、技术和训练有素的人力资源，可能还需要得到拥有先进网络能力的国家的支持，才能通过单独使用网络来取得毁灭性结果的能力。迄今为止，尚未发生单一的网络攻击直接构成武力攻击的情形。

相比于网络攻击单独发生，网络攻击与传统动能武力攻击结合使用的情景更加可能造成威胁。这类联合攻击已经有过实例。例如，2008 年俄罗斯与格鲁吉亚的冲突中，俄罗斯对格鲁吉亚进行了军事干预，并支持亲俄的分裂势力。除了使用传统武力外，俄罗斯国家机构和支持的黑客组织还实施了网络行动，虽然这些网络行动的影响较为有限，未构成武装攻击，但如果这些网络行动扩展到破坏关键军事通信系统或削弱敌方武器系统的程度，它们就会成为武装攻击的一部分。在这种情境下，网络行动与动能攻击相辅相成，网络技术被用作动能攻击的支持或准备工具，形成了整体的武装攻击。这种联合攻击方式使得网络攻击不仅是军

⁷² 黄志雄，《国际法视角下的“网络战”及中国的对策——以诉诸武力权为中心》，2015 年，第 37 卷第 5 期，第 145 页。

⁷³ Kittichaisaree Kriangsak, *Public International Law of Cyberspace*, Springer International Publishing Switzerland 2017, p. 153.

事行动的辅助手段，还可能在攻防中发挥关键作用，达到提升战争效能的目的。

74

丁斯坦（Dinstein）曾指出：“虽然对于预先性自卫权和相关其他几个术语的界限可能有所不同，但它们的共同特点是都源于一种对未来攻击的预测性。”⁷⁵无论是哪种情景下，都需要对网络攻击进行提前预测并做出及时反应，这完全取决于一国的网络技术能力。预先性自卫权的一个关键因素就是需要提前预测到攻击将会发生，并在攻击实际发生前进行自卫。然而，网络攻击从发起到到达的时间是很短暂的，如果一个国家在网络攻击的计划阶段未能察觉，它就无法提前评估是否符合“迫在眉睫”的条件。例如，2021 年针对欧盟的勒索病毒攻击大部分在 30 秒至 2 分钟内完成，受害国根本没有足够时间来判断并追踪病毒何时会进入本国领土。

同时，受害国在遭受网络攻击后，若未能及时确认攻击发生，也无法在短时间内判断是否需要行使预先性自卫权。Stuxnet 震网事件便是一个典型例子，这是一种高度复杂的恶意软件，由美国和以色列联合制造，最早于 2007 年秘密安装在伊朗纳坦兹的计算机系统中，针对伊朗的核计划并进行物理破坏。然而，直到 2010 年，计算机安全研究人员才发现这一攻击。因此，在这种情况下，受害国未能及时确认攻击行为，无法迅速做出对于预先性自卫权的判断。

2. 网络攻击难以溯源

关于网络攻击的溯源问题，有学者提出需要考虑源头、结构和目的三个方面。⁷⁶首先，从源头角度来看，技术相对落后的发展中国家由于技术条件的限制，网络空间的特殊性使得受害国往往对攻击毫不知情，或者即使发现后也难以进行有效的逆向工程，导致调查无法推进。即便是技术先进的国家能够破解攻击实施者隐藏的软件源头，恢复被删除的攻击痕迹，找到真实的源地址（如网络协议和 IP 地址），仍然可能遇到 IP 源地址被伪造并在多个物理位置被篡改的情况，甚至攻击的 IP 源地址可能来自一台已被破坏的计算机。此外，如果攻击是通过跨国合作组织实施的，那么识别攻击者就更为复杂，这使得锁定攻击发起人身份变得异常困难。

其次，从结构角度来看，开发恶意软件的成本远低于破解其结构的成本。恶意软件的设计往往是为了迷惑敌方，攻击者可以以极低的成本不断调整并组合欺骗手段。即便是通过简单的语言更改、语种变动或算法调整，敌方就会在应对过程中付出大量时间和资源。这使得恶意软件的破解变得更加困难。最后，从目的角度来看，虽然前两者容易被掩盖，但攻击的目的通常较难伪装，因此识别攻击

⁷⁴ Gill, Terry D. and Paul AL Ducheine, *Anticipatory self-defense in the cyber context*, Israel Yearbook on Human Rights, Volume 43 (2013).

⁷⁵ Yoram Dinstein, *War Aggression and Self-Defence*, Cambridge University Press, 2017.

⁷⁶ 何志鹏，王韵涵：《网络攻击困境与应对的国际法审思——以“武力”为视角》，载《哈尔滨工业大学学报（社会科学版）》2023 年 9 月，第 25 卷第 5 期。

者的目的往往是确认其身份的有效手段。然而，攻击者可以通过在源头实施欺骗，或者通过故意制造其他犯罪活动（如盗窃等常见网络犯罪）来掩饰其真正的政治目的。这种手段使得攻击者的真实意图更难被揭示出来。

（二）法理缺陷

1. 预先性自卫权的相关概念：具有模糊性

有学者就曾指出预先性自卫经常被混淆为先发制人式自卫，但两者存在本质不同。这种混淆就是网络空间引入预先性自卫权的首要法律风险。无论是国际法理论界和实务界，抑或是国际法律文件，往往会交替使用这两个概念，以至于这两种截然不同的自卫权经常被混为一谈。⁷⁷比较典型的是法国立场文件中使用“*légitime défense préemptive*”来指称“预先性自卫”，法语中“*préemptive*”对应的英文中的“*preemptive*”，实际应当是先发制人的意思。因此可以预料的是，在网络空间引入预先性自卫权，极有可能被某些国家利用为先发制人式军事行动的法律依据。

除了预先性自卫权概念的模糊性，其使用的具体紧迫性标准也在有较多争议。各国对“迫在眉睫”一词的含义缺乏共识。国际法并没有对该术语在武力攻击背景下做出权威性的法律定义。武力使用的法律规制本就存在一定的不确定性，如果不同国家依据对适用法律规则的不同理解采取军事行动，将会引发严重的担忧，其中包括可能滥用预先性自卫权和法律责任问题。对于盟国和联盟伙伴而言，如果各国对“迫在眉睫”的理解不同，还可能导致在防御性军事行动的协调上产生实际操作困难。

另一方面，有学者指出“迫在眉睫”的标准不仅仅只是涉及到时间上的要求。例如，美国国务卿丹尼尔·韦伯斯特（Daniel Webster）所提出的“仅当自卫的必要性是迫在眉睫的、压倒一切的，且别无他法可选、无暇考虑时”才能行使自卫权，除了时间上的接近和紧迫，“压倒一切的”“别无他法可选”等内容都涉及到了对必要性（Necessity）和相称性（Proportionality）要求。这种对迫在眉睫与必要性的混同，本质上使“迫在眉睫”的概念在武装攻击的语境下失去独立意义。⁷⁸

因此，近年来部分国家对“迫在眉睫”的概念的理解逐渐摆脱了传统上仅考虑未来武力攻击在时间上的临近性，而趋向于纳入更多的因素来评估采取预先性自卫行动的合法性。英国和澳大利亚在说明自身立场时均援引了伯利恒原则（Bethlehem Principles）中的第 8 项原则，⁷⁹强调应从多个维度和情形进行综合判断，包括威胁的性质和紧急程度、攻击发生的可能性、攻击是否构成持续武装

⁷⁷ 张华，《网络空间适用自卫权的法律不确定性与中国立场表达——基于新近各国立场文件的思考》，载《云南社会科学》，2021年第6卷，第81-92页。

⁷⁸ O'Meara, Chris, Necessity and Proportionality and the Right of Self-Defence in International Law, Oxford University Press (2021).

⁷⁹ Bethlehem, Daniel, The Right to Self-Defense Against Non-State Actors, American Journal of International Law 106, no. 4, p. 775-788(2012).

行为模式的一部分、预期攻击可能的规模和损害程度,以及是否存在其他更合适、更少损害的应对机会等。然而,这项标准也在使用中也具有较大的模糊性。需要指出的是,这些原则在实施过程中,需要涉及到对于多个评估因素之间如何相互作用、各因素权重如何确定,以及如何准确界定“迫在眉睫”攻击的临界点的问题,但没有获得各国的普遍共识。伯利恒原则本身在国际法领域具有高度争议性,其反映的并非现行的法律规则(*lex lata*),而是包含着明显的不确定性。

总体而言,“预先性自卫权”在理论上仍存在较大争议,主要表现为概念混杂和紧迫性标准的不确定性,这使得行使“预先性自卫”缺乏法理依据。

2. 条约解释视角:“预先性自卫权”不具合法性

在对预先性自卫权合法性的分析中,一些学者和国家立场文件采用了条约解释的路径。支持预先性自卫权的学者通过对《联合国宪章》第 51 条的文本进行法律解释,结合文件起草的过程分析,得出“预先性自卫权”是在《联合国宪章》第 51 条下自卫权的合理解释范围内,“武装攻击”也含有“武力威胁”的意思。以色列司法部副部长罗伊·舍恩多夫在 2020 年演讲中,也主张在处理条约条款是否适用于网络空间的问题时,应依照国际法条约解释的通行规则进行审慎分析,明确条款在网络环境中的适用性和相关性,最后主张:“当国家或非国家行为体在网络领域使用武力被视为构成实际或迫在眉睫的武力攻击时,受到攻击的国家可根据《联合国宪章》第 51 条规定的固有自卫权采取行动。”⁸⁰

反对行使预先性自卫权的学者刘仑通过对《联合国宪章》第 51 条进行条约解释来论证预先性自卫不具备合法性。他认为《宪章》已经清晰地规定,只有在“发生武力攻击时”(If an armed attack occurs),受害国才能行使自卫权。即便是支持预先性自卫的学者,也只是通过突出该条款中的“固有权利”或试图模糊第 51 条中的语义,以证明预先性自卫合法化的可能。然而,无论如何,第 51 条中“发生”(occurs)一词的含义都是非常明确的。他进一步指出《塔林手册》却将“迫在眉睫”与“发生”并列,视两者为网络攻击场景下行使自卫权的同等前提,这种做法一定程度上已经背离了《联合国宪章》第 51 条的原意,因此《塔林手册》在明确承认预防性自卫合法性的问题上显得突兀且武断的。⁸¹

对此,本节试图从法律解释的角度出发,根据 1969 年《维也纳条约法公约》第 31 条确立的条约解释框架,遵循“术语通常含义——上下文联系——条约宗旨”的分析路径,对《联合国宪章》第 51 条中“武装攻击”概念进行解释。⁸²

首先,就术语的通常含义而言,权威词典释义体系呈现出显著的语义趋同性。

《韦氏百科词典》将“攻击”定义为“旨在战胜敌人的进攻性军事行动”,强调其作

⁸⁰ See Israel's Perspective on Key Legal and Practical Issues Concerning the Application of International Law to Cyber Operations, 26 January 2021, available at <https://digital-commons.usnwc.edu/ils/vol97/iss1/21/>.

⁸¹ 刘仑,《从〈塔林手册〉看网络攻击中自卫权的行使》,载《法学研究》2015 年第 36 卷第 2 期,第 182 页。

⁸² Vienna Convention on the Law of Treaties (1969), Article 31.

为已实施之敌对行为的本质属性；《牛津词典》则界定为“暴力尝试伤害、征服、击败”，着重突出其具体实施的暴力特征。两定义均指向已实际发生的物理性军事行动，未涵盖单纯威胁性行为。此解释路径在《宪章》第 51 条文本结构中得到印证，该条款特别规定“武装攻击发生”（*occur*）作为自卫权行使前提，其动词时态明确排除未现实化的潜在威胁。

其次，从体系解释的角度考察，《宪章》条款间形成了严密的逻辑关联。第 2 条第 4 款对“武力威胁”与“武力使用”的双重禁止，⁸³与第 51 条仅规定“已发生攻击”的自卫权例外形成鲜明对照。此差异化的立法技术表明：当时制定宪章的专家们有意将“武力威胁”与“实际的武力攻击”作严格区分，前者仅构成第 39 条下安理会管辖的“和平之威胁”，后者方属国家单方自卫权的触发要件。这种规范配置在第七章制度设计中尤为显著——第 39 条明确将“对和平的威胁”（*threat to the peace*）、“和平之破坏”（*breach of the peace*）与“侵略行为”（*act of aggression*）三者并列，而第 51 条仅选取“实际发生的武装攻击”作为例外情形，其规范意图在于构建“安理会集体安全机制优先，国家自卫权严格受限”的武力使用控制体系。

84

最后，基于《宪章》的宗旨和目的进行解释。《宪章》序言明示“我联合国人民同兹决心：欲免后世再遭今代人类两度身历惨不堪言之战祸”的核心价值，第 1 条第 1 款⁸⁵确立联合国之宗旨为“维持国际和平及安全；并为此目的，采取有效集体办法，以防止且消除对于和平之威胁，制止侵略行为或其他和平之破坏；并以和平方法且依正义及国际法之原则，调整或解决足以破坏和平之国际争端或情势。”如果将“武装攻击”扩张解释至包含潜在威胁，将实质架空第 2(4)条的禁止性规范，导致自卫权滥用风险。此种解释既违背条约的“限制国家单方面使用武力”的宗旨，亦与《维也纳公约》第 31 条第 1 款要求的“善意解释”原则相违背。

综上，通过对术语的通常含义、上下文本、立法宗旨和目的分析，《宪章》第 51 条项下的自卫权行使严格限定于已实际发生的武装攻击情形。所谓“预先性自卫”因缺乏规范基础与价值正当性，构成对国际法上武力使用制度的实质性突破，应认定其不具备合法性。此结论不仅符合条约解释方法论要求，更是维护集体安全机制有效运作的必然选择。

3. 习惯法的演变：“预先性自卫权”缺乏国家实践

支持预先性自卫的学者常使用的理论依据为 1837 年的卡罗琳号案（*Caroline case*），主张已经形成的“卡罗琳号”标准，并且这一标准并非静态，而是在后续的实践中成为了一组指导性原则（*a set of guiding principles*），因此在具体适用时需

⁸³ 《联合国宪章》第 2 条第 4 款。

⁸⁴ Michael Bothe, *Terrorism and the Legality of Pre-emptive Force*, *European Journal of International Law*, Vol.14:227, P. 230 (2003).

⁸⁵ 《联合国宪章》第 1 条第 1 款。

依据当时的相关情境，以及可获得的信息来评估预期性自卫行动的正当性。⁸⁶还有学者解读卡罗琳案承认了预先性自卫为一项习惯国际法，《联合国宪章》51条并未限制预先性自卫的适用，因为预先性自卫是一项固有权利（the inherent right）。

87

但卡罗琳号案发生在《联合国宪章》体系建立之前，能够多大程度在《联合国宪章》没有明确允许的情况下扩张自卫权的使用仍然是存疑的。事实上，国际法院在该案中的判决并非针对预先性自卫的有效性。“卡罗琳号”标准必须其他事实和法律因素的背景下加以适用，而不能仅被视为抽象概念，也没有形成一般性的指导原则，任何超出“卡罗琳号”标准的先发制人自卫行动，都在法律上是不可行的。

通过前文对于网络空间立场文件的对比分析可知，在实践中，国际社会对于预先性自卫理论仍然存在巨大分歧，除了美国、以色列等国家持明确支持的立场，极少数国家用这一观点来正当化自己的武力使用行为。国际法院（ICJ）也已确认，《联合国宪章》第51条“不允许国家使用武力来保护其“被感知的安全利益”。无论是好战言论还是未来获得核武器的可能性，都不足以触发自卫权。

因此可知，实际上仍然有相当多的国家仍然反对这一预先性自卫的理论，这意味着该权利并未得到普遍接受，“预先性自卫”更没有被确立为一项现行的国际习惯法。

四、网络空间行使预先性自卫权的风险

预先性自卫不仅存在技术上和法理上的缺陷，其适用也有风险和挑战。

（一）归因错误与误判风险

若因情报错误或证据不足而错误归因，可能对无辜国家或实体实施武力反应，引发严重国际后果。对自卫权的行使施加时间上的要求，旨在防止国家在网络空间滥用武力。攻击的迫近程度始终是判断防御行动合法性与必要性的关键因素之一。理论上，攻击发生的时间越远，潜在威胁的预测准确性就越低。这种基于时间的限制能够有效抑制自卫权的滥用风险，因为随着时间间隔越远，对未来攻击发生的可能性及其严重性的评估难度会显著增加。更长的时间间隔意味着防御国在决策过程中需要考虑更多的不确定因素，包括攻击方可能调整或取消原定的攻击计划、通过外交或其他和平手段解决争端的可能性提高、联合国安理会（UNSC）可能采取有效措施使军事行动不再必要等。因此，攻击与防御行动之间的时间间隔越长，潜在受害国所面临通过和平方式解决问题的压力也越大，而非直接诉诸武力。

在网络空间中，网络攻击的特性使得其评估尤其复杂，网络行为通常具有高

⁸⁶ Terry D Gill, The Temporal Dimension of Self-Defence: Anticipation, Pre-emption, Prevention and Immediacy, *Journal of Conflict and Security Law*, Vol. 11, p. 361-369 (2006).

⁸⁷ Derek Bowett, *Self-Defence in International Law* (Manchester: Manchester University Press, 1958), 3-25, 269.

度的不确定性，潜在目标国家在很多情况下很难准确判断攻击的存在、性质以及可能带来的损害程度。例如，即便网络攻击成功入侵了水坝或核电站等敏感设施，也不必然意味着攻击会达到国际法上认定的“武力攻击”门槛。以 2010 年伊朗布什尔核电站和纳坦兹核设施遭受的 Stuxnet 蠕虫病毒攻击为例，虽然攻击对设施运行造成了一定的干扰，但并未造成任何人员伤亡或实质财产损失。按照国际法中效果标准的衡量，这类攻击未达到构成武力攻击的严重性门槛，因此受害国并不具备行使自卫权的法律依据。如果此时，国家对于这类攻击提前行使了自卫权，将会极易造成网络空间武力的滥用。

（二）破坏“禁止使用武力”机制

事实上，在过去 20 多年的反恐行动中，预先性自卫的理念在西方国家得到更高的推崇。北约国家在反恐战争中频繁动用武力，宣称是基于“先发制人”的自卫需求，但事实上，这些行动未能有效维护世界和平，反而加剧了地区和全球的不稳定。若允许这种倾向延伸至网络空间，将可能导致网络安全国际规则朝着有利于军事扩张的方向发展，这与维护全球网络空间和平的初衷相悖。

18 世纪和 19 世纪的国际实践不得不面对这样一个现实：在当时，禁止战争的理念并不可行。在具体案例中，人们难以确定哪一方的战争是正义的，哪一方并非如此。因此，原本旨在规范战争行为的“正义战争”原则（*bellum iustum*）最终适得其反，使国际法对战争的诉诸变得无动于衷。直至第一次世界大战结束，战争的合法性才在国际法框架下受到限制。

因此，如果试图确立一项无法有效界定什么是“合法使用武力”的规则，那么最终可能会破坏导致“禁止使用武力”原则和机制，会破坏 20 世纪最重要、最有益的文化与政治成就之一。这一风险是具有现实性的，因为禁止使用武力的规则本身就十分脆弱，该规则并非源自国家实践的自然演进，事实上，历史上从未形成过持续性的“避免使用武力”国际惯例。第一次世界大战后，国家行为体对武力使用的态度发生变化，一种新的法律信念（*opinio iuris*）逐渐形成，即战争本身是不合法的，各国官方对此的立场亦随之转变。这一转变根植于更深层次的社会变革，即社会价值判断的演进。第一次世界大战所带来的巨大人类苦难，使得战争不再被社会视为一种“正常现象”。在此之前，和平主义往往受到当时社会体系的轻视甚至嘲讽。然而，战争的惨烈代价促使人类重新审视其正当性，从而推动了国际法对武力使用的限制。

简而言之，禁止战争的规则以及随后对军事力量的全面限制，本质上是社会价值观变化的产物。而这一过程的可逆性决定了其脆弱性。尽管我们距离重新回到军事化社会的价值体系尚有距离，但如果希望国际法继续发挥约束武力使用的作用，我们必须警惕任何试图将军事力量描绘为“非邪恶”或“合理”事物的言论。历史的教训已然清晰：若再次接受诸如“正义战争”这样模糊而宽泛的概念，我们

便是在为武力使用的合理化铺平道路，这将使世界倒退至 19 世纪——那个战争仍未被视为“非法”的时代。

自 1945 年宪章生效以来，“禁止使用武力”原则一直是当代国际法体系的基石。尽管这一原则曾被法律学者作出不同解读，并且在实践中屡遭违反，但毫无疑问，它依然是一条几乎被普遍接受的国际法和国际关系基本准则，并被广泛承认为具有强行法（*jus cogens*）的地位。⁸⁸不仅是在现实世界，在网络空间，预先性自卫的出现非常有可能破坏当前的“禁止使用武力”原则。

（三）导致网络空间军事化

网络空间军事化是指国际社会中的各个行为体将网络空间技术和资源应用于军事领域，政府和军队不断增加在网络空间的军事投资，形成了一种在提升网络攻防能力和网络战取胜能力方面的国际竞争趋势。⁸⁹当前，各国尤其是美西方大国日益加大对网络空间的军事投入，网络空间的军事化趋势愈发明显。这一发展主要表现为以军事目的为导向、聚焦安全领域的网络对抗频繁发生，各国竞相强化网络作战力量建设，网络武器的研发与运用日趋激烈，网络军备竞赛呈现白热化态势。⁹⁰

推动将包括“预先性自卫”在内的自卫权适用于网络空间，实质上将进一步加剧网络空间的军事化趋势，其背后反映的是美西方大国依托技术优势，意图在网络空间确立主导地位的战略图谋。在 2016-2017 届联合国信息安全政府专家组（UNGGE）会议上，美欧国家基于此前 G7 峰会的《七国集团网络空间原则和行动》（*G7 Principles and Actions on Cyber*），试图将战争法和自卫权引入网络空间，将传统战争行为与网络战挂钩。然而，以中俄为代表的网络“后起国”一致认为，这种将网络空间视为传统军事行动领域的做法，无法有效缓解网络空间军事化的趋势，反而可能加剧网络军备竞赛，最终导致谈判失败。

作为《联合国宪章》所允许使用武力的唯一例外，自卫权的适用本应受到极为严格的限制，然而一些大国却通过对《宪章》条文的扩张性解释，企图赋予“预先性自卫”以“合法”地位，此举无疑背离了国际法的基本精神。

可以预见，一旦“预先性自卫”被广泛适用于网络空间，各国将在技术主导的预测能力竞赛中愈发依赖对网络武器的投入，并将其纳入传统军事战略体系之中。网络军队建设将持续加快，围绕军事目的的网络对抗日益频繁，网络空间被越来越多的国家明确视作未来战争的新型战场。在此背景下，网络军备朝着战场化、实战化方向迅速推进，网络武器的研发与应用也进入竞争白热化阶段，全球网络

⁸⁸ Yoram Dinstein, *War Aggression and Self-Defence*, Cambridge University Press, 2017, p. 101-116.

⁸⁹ Ronald Deibert, *Tracking the Emerging Arms Race in Cyberspace*, *Bulletin of the Atomic Scientists*, Vol. 67, No. 1, 2011, p. 2.

⁹⁰ 杜雁芸：《网络空间军事化发展态势及其应对》，载《太平洋学报》2021 年第 29 卷第 12 期，第 53-64 页。

安全形势面临更为复杂与严峻的挑战。⁹¹

五、中国关于网络空间行使预先性自卫权的立场表达

中国对网络空间适用自卫权问题的一直以来采取谨慎态度。将在动能武力攻击中本就存在较大争议的“预先性自卫”适用于网络空间，不仅不具备国际法上的依据，同时存在自卫权滥用等风险。因此，中国反对预先性自卫适用于网络空间，坚持自卫权的行使必须与《联合国宪章》相符合，遏制网络空间军事化。

（一）反对在网络空间行使预先性自卫权

目前网络空间的国际法治仍处于初步探索阶段，西方国家较早地意识到，全球化的网络空间可以成为传播和推广西方意识形态和价值观的重要渠道。⁹²美国、以色列、英国、澳大利亚等西方国家支持“预先性自卫”的立场，实际上反映了这些国家试图凭借其在网络技术和规则制定上的优势，从而在网络空间推动有利于本国国家利益的进攻性网络行动，以维持在网络空间的霸权地位。因此，在战略层面，中国应当揭示西方国家推动自卫权扩大适用的本质⁹³：以网络攻击的特殊性为借口，提倡引入本就在传统武力攻击下就富有争议预先性自卫，实际是倒逼国际法上诉诸武力的法律向军事化的方向发展，本质是一种“偷梁换柱”的做法。

美国新一轮“中国威胁论”兴起，大肆炒作所谓中国黑客攻击，声称“中国黑客”对美国电信公司和基础设施发动了大规模网络攻击。美国不断重复这些毫无根据的指控，试图将谎言包装成“真相”，借此散布对中国的不实指责和错误认知，以便为自身的政治利益服务。众所周知，美国是世界上最大的“黑客帝国”，利用其技术优势对其他国家政府甚至是本国公民进行长期监控和窃密，但一些对中国抱有偏见的美国政客却无视事实，指责中国是美国网络安全的最大威胁。

部分国家在网络安全问题上模糊《联合国宪章》的武力使用限制，试图以法律解释的方式消解国际法对于武力行为的约束。中国对此持反对态度，并基于条约解释原则，主张《联合国宪章》第51条中的“武力攻击”一词不包括“即将发生的攻击”。部分国家试图混淆二者概念，以便在网络空间中降低自卫权的行使门槛，这种做法既不符合条约解释原则，也不符合国际安全的现实需求。特别是在网络空间，网络攻击的手段、影响和性质均与传统武力攻击存在显著区别，若简单适用传统武力规则，容易导致误判和滥用。

（二）积极开展网络外交，推动制定网络空间国际规则

当前，网络空间的国际秩序构建中，美国通过“密集式”立法、先发制人的预防战略和“集体性自卫”的国际合作方式，在全球网络空间博弈中占据了先机。对

⁹¹ 杨楠：《网络空间军事化及其国际政治影响》，载《外交评论(外交学院学报)》2020年第37卷第3期，第69-93页。

⁹² 黄志雄：《网络空间国际法治：中国的立场、主张和对策》，载《云南民族大学学报（哲学社会科学版）》2015年第32卷第4期，第137页。

⁹³ 张华：《网络空间适用自卫权的法律不确定性与中国立场表达——基于新近各国立场文件的思考》，载《云南社会科学》2021年第6卷，第91页。

于包括“预先性自卫”在内的自卫权扩大适用,我国也应当通过官方立场声明和通过非官方的学者学说、开展互动式演讲等方式明确表示反对,积极参与网络空间国际规则的制定,提高我国在网络空间国际秩序构建中的话语权和影响力。⁹⁴

一方面,应当加强包括自卫权在内的网络空间法治领域的主要争议问题的学术研究,为我国的相关网络外交提供坚实的理论支撑。这些研究将有助于从制度和能力建设层面,帮助我国积极、深度参与网络空间国际规则的制定,切实提升我国在全球网络空间国际秩序构建中的话语权和影响力。⁹⁵

另一方面,应该采取“多轮并驱”的策略,积极参与网络空间国际规则的博弈,尤其是在包括《塔林手册》在内的具有广泛影响力的国际法规则的编写过程中发挥作用。尽管《塔林手册》并非官方立场文件,但它对后续各国发布立场文件产生了重要影响。例如,德国的国家立场文件多次引用了《塔林手册》的观点,特别是在论述“预先性自卫”合理性时,亦以《塔林手册》作为法理依据。

（三）充分发挥联合国安理会在网络空间中的职能

面对国际社会在使用武力和自卫权问题上所面临的现实问题,中方始终坚持应以多边主义的方式寻求解决方案。法律理论的传统智慧指出,仅凭实体法律规范并不足以保障充分的法律确定性,因此应借助程序性规则,来达成各方均能接受的社会与政治共识。这一智慧在当前国际社会所面临的问题中具有特别的现实意义。

中国认为,应坚定维护以联合国宪章为核心的国际法治体系,特别是安理会在维护国际和平与安全方面的核心作用。实践证明,联合国安理会的授权是解决国际安全困境的最有效、多边且合法的程序机制。联合国安理会作为国际集体安全体系的核心机构,体现了国际社会必要的权力平衡与政治智慧。尽管安理会在运作过程中可能出现困难,但这种权力制衡机制恰恰是《联合国宪章》设计的本意所在,有利于防止个别国家的单边主义和霸权行径。任何绕过安理会的单边行动,不论其动机如何,都无法在国际法框架下获得充分的合法性,也不利于国际和平与安全。

因此,中国一贯主张,各国应当严格遵守联合国宪章,尊重安理会在维护国际和平与安全中的核心作用。国际社会应积极推动多边合作,通过协商对话解决争端,避免以单边主义和预防性武力手段解决问题。唯有如此,才能真正维护国际社会的稳定与繁荣,实现持久和平与共同安全。

六、结语

网络空间的兴起深刻重塑了国家安全的边界和国际法治治理模式。尽管“预

⁹⁴ 黄志雄:《国际法在网络空间的适用:秩序构建中的规则博弈》,载《环球法律评论》2016年第38卷第3期,第5-18页。

⁹⁵ 黄志雄:《网络空间国际法治:中国的立场、主张和对策》,载《云南民族大学学报(哲学社会科学版)》2015年第32卷第4期,第141页。

先性自卫”理论看似为国家安全提供了新的工具，但其潜在的法律风险不容忽视。在网络空间的背景下，中国始终坚持构建以《联合国宪章》为核心的国际法体系，坚决反对自卫权的过度扩张，主张通过更加完善的程序设计，如通过安全理事会授权，寻求和平解决国际争端。正如古语所言：“天下兼相爱则治，交相恶则乱。”面对不断演进的网络攻击和挑战，各国只有秉持合作共赢、相互尊重的原则，才能真正实现网络空间的持久和平，携手构建网络空间命运共同体。

国际法视野下的网络战规制困境与中国因应

谢霖

摘要：网络战作为新型冲突形态，其国际法规制面临传统规则适配性与新兴机制碎片化的双重困境。传统国际法体系适用困难，新兴规则虽尝试填补空白，但难以形成有效共识。困境成因可溯至国际法造法机制滞后于技术迭代、主权概念受网络无界性消解，以及软法规范非约束性与治理权力失衡的交互作用等。中国应对策略应以“网络空间命运共同体”为核心理念，主张坚守网络主权、推动国际法适应性修订，倡导传统规则与技术解释标准革新，并通过强化国际合作与能力建设，弥合数字鸿沟，实现网络空间的持久和平与共同发展。

关键词：网络空间 网络战 国际法规制 中国应对

当前，网络战作为一种低成本、高回报的战争手段，正愈发频繁地登上国际冲突的舞台，成为国家间博弈的新领域，因其具有匿名性、远程性和快速传播等特点，使得网络空间充满潜在危险。近年来，爱沙尼亚遭受的网络瘫痪、俄罗斯对格鲁吉亚实施的网路攻击以及伊朗核工业和科研设施遭受的“震网”病毒攻击等，不仅揭示了网络战的巨大破坏力，也深刻影响了国际政治经济格局。这些事件暴露了网路攻击军事化的趋向，而其后果则进一步彰显了加强网络安全体系建设的紧迫性。

在此背景下，习近平总书记高度重视网络安全工作，多次作出重要指示批示，强调“没有网络安全就没有国家安全，就没有经济社会稳定运行，广大人民群众利益也难以得到保障”。¹ 习近平总书记的论断深刻阐明了网络安全在国家安全体系中的基础性地位，为我国网络安全事业的发展指明了方向。然而，面对日益复杂的国际网络环境和不断演变的网路战威胁，现有的网络安全规范体系仍面临诸多挑战，国际法在规制网路战方面的困境尤为突出。因此，为有效应对网路战带来的挑战，增强我国在网络空间的话语权，我们应首先理清现有规制网路战的国际法规则，在此基础上，进一步深入剖析国际法在网路战规制方面的困境，并对困境成因进行溯源分析，最后，结合中国实际情况，提出具有针对性和可操作性的因应方案，以期为我国乃至全球网络安全体系的完善贡献智慧和力量。

一、网络战：国际法规制之现状概览

（一）传统国际法体系的适应性探索

1. 传统国际法体系中的约束性规则

¹ 《习近平出席全国网络安全和信息化工作会议并发表重要讲话》，载 https://www.gov.cn/xinwen/2018-04/21/content_5284783.htm，2025 年 4 月 5 日访问。

传统国际法体系对于网络战的约束性规范主要体现在《联合国宪章》（下称“《宪章》”）和《日内瓦公约》及《海牙公约》中。

一是《宪章》中的约束性规范。传统军事攻击行为因具备攻击结果的严重性、侵入性等特点而受《宪章》规制，网络战若符合上述特征亦应纳入该框架。作为国际法核心文件，《宪章》虽未直接规定网络战，但其基本原则可为网络战规制提供法律依据。根据《宪章》第2条第4款确立的禁止使用武力原则，²若网络战的效果达到传统武力威胁或使用武力程度，可以视为构成对该原则的违反。³这种解释既符合国际法精神，亦是应对网络空间安全威胁的必然要求。第51条赋予受害国在遭受“武力攻击”时的自卫权，⁴若网络攻击达到传统武力攻击程度且可准确溯源，受害国在符合必要性及相称性原则前提下可行使自卫权。同时，《宪章》第六章和第七章赋予安理会采取集体强制行动的权力，当网络战产生等同于传统军事攻击后果时，同样适用该规制体系。

二是《日内瓦公约》及《海牙公约》中的约束性规范。国际法院在使用核武器合法性的咨询案中曾表示，适用于武装冲突的海牙公约体系以及日内瓦公约体系，两者紧密联系，已经逐渐成为一个复杂的整体，即“国际人道法”。⁵虽然关于国际人道法是否可以适用于网络战的问题仍存在争议，但红十字国际委员会从一开始就采取了明确和肯定的立场，表示武装冲突期间的网络行动，或者网络战，都要受到国际人道法的规范。网络行动依赖于不断发展的新技术这事实，并不能阻止武装冲突中使用这种技术作为作战手段和方法时国际人道法的适用。⁶国际人道法中涉及网络战的约束类规则主要包括以下内容：

一是区分原则，该原则集中体现在1949年《关于武装冲突中保护平民的日内瓦公约》（简称“日内瓦第四公约”）和1977年《日内瓦公约第一附加议定书》（简称“第一附加议定书”）中。⁷其中，《第一附加议定书》第48条规定，“为了保证对平民居民和民用物体的尊重和保护，冲突各方无论何时均应在平民居民和战斗员之间和在民用物体和军事目标之间加以区别，因此，冲突一方的军事行动仅应以军事目标为对象。”据此可知，若区分原则适用于网络战规制，则网络战武器的选择、网络战的目标选择上，都应避免影响平民生活。

² Article 2(4) of *United Nations Charter*: All Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations.

³ 参见蔡高强、焦园博：《“网络武装冲突”的国际法阐释》，载《湘潭大学学报（哲学社会科学版）》2017年第41期第3卷，第16-23页。

⁴ Article 51 of *United Nations Charter*: Nothing in the present Charter shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member of the United Nations, until the Security Council has taken measures necessary to maintain international peace and security.

⁵ 参见宋杰：《“国家在战争或武装冲突中使用核武器的合法性问题咨询意见案”述评》，载《武大国际法评论》2007年第2期。

⁶ 参见克努特·德尔曼、洛朗·吉塞勒、蒂尔曼·罗登霍伊塞尔、丁玉琼：《国际人道法对网络战的可适用性及其适用》载《国际法研究》2019年第4期，第3-17页。

⁷ 参见孙国顺：《谈国际人道法中的区分原则》，载

《<https://www.icrc.org/zh/doc/resources/documents/misc/6rc9fx.htm>，2025年4月5日访问。

二是比例原则，该原则在《第一附加议定书》第 57 条中有明确规定，其强调攻击手段与军事利益需保持合理平衡，禁止采用导致过度附带损害的作战方式。⁸例如，针对军事指挥系统的网络攻击若引发大规模民用设施瘫痪或环境灾难，则可能违反《第一附加议定书》第 54—56 条关于生存必需物体与危险装置的保护规定。

三是马尔顿条款，该条款是由著名国际法学家雷德里克·马尔顿（Friedrich Martens）于第一次海牙和平会议期间提出，最早记录于 1899 年和 1907 年的海牙公约的序言中，之后被引进于《第一附加议定书》第 1 条第 2 款。该条款要求在缺乏明确条约规制时，仍须遵循既定国际习惯、人道原则及公众良知，⁹若适用该条款，则可为规制网络战提供兜底性规范。

2. 传统国际法体系中的追责性规则

在传统国际法体系中，《国家对国际不法行为的责任条款草案》（Responsibility of States for Internationally Wrongful Acts）和《国际刑事法院罗马规约》（Rome Statute of the International Criminal Court）的相关条款为网络战追责提供了基础性框架，但其适用性需结合网络战的特殊性进行解释与拓展。就《国家对国际不法行为的责任条款草案》而言，根据该法第 4 条至第 11 条，当网络攻击行为可归因于国家机关或其授权实体时，国家即需承担国际不法行为的责任。¹⁰而对网络战而言，归因问题尤为关键。若网络战行为可归因于一国（如通过证明该行为由国家机关实施，或非国家行为体受国家“有效控制”），则受害国应有权通过援引《国家对国际不法行为的责任条款草案》要求行为国停止不法行为、赔偿损失，甚至采取反措施。

就《国际刑事法院罗马规约》（简称“罗马规约”）而言，其通过界定战争罪、危害人类罪等核心罪行的构成要件，从个人刑事责任角度为网络战追责提供可能。根据该规约，若网络攻击构成战争罪或危害人类罪（如故意攻击平民设施导致大规模伤亡），相关个人可能被追究刑事责任。此外，《罗马规约》第 25 条确立的个人刑事责任原则，以及第 28 条规定的指挥官责任制度，为追究网络战策划者、实施者及其上级指挥人员的刑事责任提供了法律路径。¹¹

（二）新兴国际法规则的碎片化发展

1. 以《塔林手册 2.0》为核心的软法规范

⁸ 参见《日内瓦公约第一附加议定书》第五十七条 攻击时预防措施，载 <https://ihl-databases.icrc.org/zh/ihl-treaties/api-1977/article-57?activeTab=>，2025 年 4 月 5 日访问。

⁹ 参见《日内瓦公约第一附加议定书》第 1 条第 2 款：在本议定书或其它国际协议所未包括的情形下，平民和战斗员仍受来源于既定习惯、人道原则和公众良心要求的国际法原则的保护和支配。载 <https://ihl-databases.icrc.org/zh/ihl-treaties/api-1977/article-1?activeTab=1949GCs-APs-and-commentaries>，2025 年 4 月 5 日访问。

¹⁰ United Nations, *Responsibility of States for Internationally Wrongful Acts*, at https://legal.un.org/ilc/texts/instruments/english/draft_articles/9_6_2001.pdf (Last visited on April 5, 2025).

¹¹ International Criminal Court, *Rome Statute of the International Criminal Court*, at <https://www.icc-cpi.int/sites/default/files/2024-05/Rome-Statute-eng.pdf> (Last visited on April 5, 2025).

从 2009 年起, 北约网络合作防御卓越中心牵头组织了一个由 20 名西方国家的战争法、军事法及网络技术专家构成的国际专家组, 他们历经努力, 完成了《塔林手册》第一版的编纂工作。¹² 至今, 该手册已迭代至第二版, 即《网络行动国际法塔林手册 2.0 版》(以下简称《塔林手册 2.0》), ¹³ 成为研究网络空间国际法适用问题的权威文献, 其全面性和系统性为国际法学界所公认。

《塔林手册 2.0》作为网络空间国际法适用的重要学术成果, 在网络战规则体系的扩展方面展现出一定突破。其内容分为四个部分、154 条规则, 几乎涵盖了所有可能适用于网络空间的国际法规则, 并对其提出适用意见。手册中和网络战较为相关的内容为第三部分的“使用武力”和“集体安全”(规则 68-79), 以及第四部分“网络武装冲突法”(规则 80-154)。该手册承认主权在网络空间的适用, 并且对在网络空间中使用武力或以武力进行威胁的行为进行了界定, 明确指出当网络攻击达到一定程度时, 其与传统攻击一样具有非法性。在这种情况下, 受攻击的国家有权行使自卫权, 但手册也对这种自卫权的行使条件进行了详细规范。在承认网络攻击可以被视为“使用武力”的基础上, 手册进一步探讨了网络攻击升级为“武装冲突”的标准。此外, 手册还特别强调了区分原则在网络战争中的应用, 详细讨论了如何在网络攻击中正确区分军事目标和平民目标。这些规定确保了即使在网络环境中, 国际人道法的基本原则也能得到遵守。

可以看出, 《塔林手册 2.0》在现有国际法框架内, 深入探讨了诸多关于网络战规制的问题。手册中提到的大多数原则都有相应的国际法基础作为支撑, 这表明尽管网络空间带来了新的挑战, 但许多传统的国际法原则仍然适用, 并且可以通过适当的调整来适应新的情况。尽管《塔林手册 2.0》本身不具有法律约束力, 但其仍然为网络战的国际法规制提供了一个理论框架, 也为处理涉及网络攻击的实际案例提供了实用指南。

2. 联合国框架下的“双轨制”模式

联合国信息安全政府专家组 (United Nations Group of Governmental Experts, 简称“UNGGE”) 由联合国大会第一委员会于 2004 年成立, 并分别在 2010 年、2013 年、2015 年、2021 年发布了四份共识报告, 重点在于为国家在网络空间的行为规范提供自愿且非约束性的指导。UNGGE 报告中最值得关注的是 2015 年公布的《政府专家组关于在国际安全背景下的信息与通信领域的发展报告》, 其中确立的 11 条网络空间国家行为规范也在 2021 年的最新报告中得到进一步重申。¹⁴ 该报告肯定了国家主权原则、战争法中的区分原则与比例原则等内容能够

¹² N. Simovic, Zivorad Rasevic & Vladimir M. Simovic, *Cyber Warfare and International Cyber Law: Whither?*, J. Crimin. & Crim. L., Vol.58:23, p.28 (2020).

¹³ 北约网络合作防御卓越中心特邀国际专家组编写, [美]迈克尔·施密特总主编, [爱沙尼亚]丽斯·维芙尔执行主编, 《网络行动国际法塔林手册 2.0 版》, 黄志雄等译, 社会科学文献出版社 2017 年版。

¹⁴ 参见左亦鲁:《国家安全视域下的网络安全——从攻守平衡的角度切入》, 载《华东政法大学学报》2018 年第 21 卷第 1 期, 第 148-157 页。

在网络空间得到适用，并细化了相关原则内涵。同时，该报告明确了国家在网络空间需履行国际法上的义务，力求信息通信技术的和平利用。¹⁵

联合国信息安全开放式工作组（Open-ended Working Group，简称“OEWG”）的诞生则与 2017 年 UNGGE 第五届会议的谈判破裂有关，当年由于各国就网络空间军事化、传统军事手段与网络攻击之间的关系存在根本分歧，导致未能形成共识报告，¹⁶这对 UNGGE 进程来说是一次严重的挫折，也体现出各国对网络空间军事化趋势问题存在截然不同的态度。为继续推进 UNGGE 的工作，俄罗斯和中国等国于 2018 年联合提出了建立一个开放的专家工作组的建议。但是，以美国为首的西方发达国家对此表示反对，并主张恢复原有的 UNGGE 机制。最终，两个提案均获得通过，标志着联合国在网络安全规则的制定上进入了“双轨制”阶段。

不同于 UNGGE 仅有部分主权国家参与的闭门会议形式，OEWG 允许包括联合国成员国、私营企业、民间社会组织、技术社区以及个人在内的多个主体参与到讨论中来。经过多轮商讨，OEWG 和 UNGGE 先后发布了最终共识性报告，并在 2021 年 9 月召开的第 76 届联合国大会上提交表决。两份报告表决通过后便成为正式的联大文件，在解决网络空间国际治理核心问题中起到促进国际社会共识形成的作用。

二、网络战规制的双重困境：传统规则局限与新兴机制缺陷

（一）传统国际法规制存在规则适配与追责困境

1. 约束性规则之适配困境

网络空间的迅猛发展使得传统国际法在网络军事行动中的适用面临前所未有的挑战，其中“武力”界定及相关原则的适用问题尤为突出。

其一，《联合国宪章》第 2 条第 4 款“使用武力”与第 51 条“武力攻击”的认定缺乏明确标准。网络攻击的物理破坏性、隐蔽性及技术复杂性导致其是否构成“武力”存在争议，以国际法院在“刚果案”中的意见为例，法院指出判断一国对另一国的行为是否构成使用武力时，应考虑该国行为的严重性及持续性。¹⁷然而，这一标准在网络战中便显现出使用局限。网络攻击行为的严重性及持续性往往难以即时评估，这种模糊性为各国在网络战中的行为提供了极大的操作空间，削弱了“禁止使用武力原则”的规制效果。

其二，国际人道法中的区分原则与比例原则在网络场景中操作性不足。就“区

¹⁵ General Assembly of the United Nations, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, at https://digitallibrary.un.org/record/799853/files/A_70_174-EN.pdf?ln=zh_CN (Last visited on April 5, 2025).

¹⁶ 参见徐培喜：《米歇尔 Vs. 米盖尔：谁导致了 UNGGE 全球网络安全谈判的破裂？》，载《信息安全与通信保密》2017 年第 10 期，第 10-12 页。

¹⁷ ICJ, *Judgement on the armed activities on the territory of the Congo (Democratic Republic of Congo v. Uganda)*, at <https://www.icj-cij.org/public/files/case-related/116/116-20051219-JUD-01-00-EN.pdf> (Last visited on April 5, 2025).

分原则”而言,《日内瓦公约》及其附加议定书规定的“区分原则”要求交战双方在战斗中需对民用设施与军用设施、战斗员与非战斗员进行区分。¹⁸然而,在网络战中,这一原则的适用却面临诸多困境。首先,由于信息基础设施普遍存在的军民融合属性,攻击方在实施网络作战行动时面临双重困境:既难以对目标属性进行精确识别,又需承担超出传统物理战场的法律审查义务;此外,依据《日内瓦公约》确立的作战身份公示原则,在虚拟空间遭遇执行障碍——相较于物理战场的显性身份标识,网络攻击者可通过技术手段实现多重匿名化,这不仅导致防御方在攻击主体识别上存在技术滞后性,更使得“战斗员”与“非战斗员”的区分标准面临适用困难。这种技术特性与法律规范的结构性矛盾,实质性地削弱了区分原则在网络空间的约束效能。

就“比例原则”而言,其核心要义在于要求冲突方对军事行动可能产生的附带平民损害与预期军事利益进行预先权衡,禁止二者间出现明显失衡。虽然《塔林手册 2.0》、UNGEE 报告等文件均认为其能够适用于网络战领域。然而,网络攻击的技术特性对传统比例原则的适用范式提出了结构性挑战。相较于传统物理攻击,网络攻击呈现出显著的“技术扩散性”与“后果不可控性”。以针对军事指挥系统的网络攻击为例,攻击者通过植入恶意代码虽能达成类似传统物理攻击的战术效果,但其技术载体具有自主传播、跨系统渗透等特性。2010 年“震网”病毒事件即为典型案例,该病毒在破坏伊朗核设施离心机系统后,仍持续感染全球包括印度尼西亚、印度、巴基斯坦、德国等国家的技术系统,印证了网络武器“溢出效应”的客观存在。¹⁹此种技术特征导致攻击者难以在事前准确评估军事目标与民用设施间的“数字边界”,更无法可靠预测攻击后果的时空范围。尽管马尔顿条款为法律空白提供了兜底性规范,但其难以转化为具体规则,极易导致司法实践中裁判尺度不一,进而引发争议。

2. 追责性规则之实施困境

传统国际法体系在规制网络战的追责实践中面临多重困境,其核心矛盾在于既有规则的制度设计与网络战的技术特性及实施模式之间存在结构性张力。首先,归因难题构成追责的首要障碍。《国家对国际不法行为的责任条款草案》虽确立了国家责任的两大归因路径——国家机关行为与非国家实体受“有效控制”下的行为,但在网络战中,攻击主体常通过分布式跳板服务器、虚拟私人网络等技术手段隐匿真实来源,加之国家与非国家行为体之间可能存在的“战略模糊”关系,如国家通过技术支援、资源提供等间接方式支持黑客组织,使得证明国家与攻击行为之间存在“完全依赖”或“有效控制”的实质性联系面临极高举证门槛,²⁰更何

¹⁸ 参见王孔祥:《区分原则在网络战中的适用》,载《国际安全研究》,2013 年第 31 卷第 1 期,第 142-153 页。

¹⁹ See Kerr, Paul K., John Rollins, and Catherine A. Theohary. *The stuxnet computer worm: Harbinger of an emerging warfare capability*, Congressional Research Service, p.1 (2010).

²⁰ Boudreaux B, Davis J S, McGovern G, et al., *Stateless Attribution: Toward International Accountability in*

况，国家要对利用本国领土范围内的网络基础设施从事的各种行为加以有效监控，其技术难度远远超过对传统私人行为（特别是军事行为）的监控难度。²¹此外，网络攻击的即时性与匿名性导致溯源技术依赖各国情报能力差异，客观上为行为国制造了责任规避空间。而倘若技术优势国家实行不透明、无直接证据的归因，将给网络空间留下更多不稳定因素。虽大部分国家坚持仅仅是指向性的、无直接证据归因不应被允许，然而此前美国与北约成员国建立的“联合归因”（collective attribution），对认定的网络攻击发动国进行“指出并羞辱”（Naming & Shaming），便极有可能是利用技术优势形成小团体的霸凌行为。

其次，国际法规范与网络战行为样态的适配性不足加剧了法律适用困境。《罗马规约》将战争罪与危害人类罪的构成要件锚定于物理性暴力或大规模人道灾难，而网络攻击可能表现为数据篡改、关键基础设施瘫痪等非动能损害，此类新型攻击手段是否落入传统罪行的语义射程尚存争议。即便攻击造成实质人员伤亡，网络空间的作战指挥链往往呈现去中心化特征，指挥官责任制度要求的“实际知悉与有效控制”在匿名化攻击中难以举证。此外，国际刑事法院的属人管辖权受制于缔约国数量及合作意愿，主要网络作战主体（如国家支持的黑客组织成员）常具有国籍国司法庇护，导致个人刑事责任追究机制陷入实效性困境。

（二）新兴国际法规则存在软法约束与模式建构困境

1. 软法性规范之约束困境

新兴国际法规则在网络战规制中的碎片化发展面临多重困境，其核心问题体现在规则制定主体的局限性、法律适用的模糊性以及国际共识的脆弱性。以《塔林手册 2.0》为代表的软法突破虽在理论上构建了网络空间国际法适用框架，但其编纂过程由北约网络合作防御卓越中心主导，国际专家组构成以西方国家军事及法律专家为主，发展中国家代表性不足，导致规则体系隐含对技术优势国家单边利益的倾斜。例如，手册将“审慎原则”引入网络空间治理，要求国家对其领土内发起的网络攻击承担预防义务，但该原则的适用边界模糊，可能被技术强国用作干涉他国内政的借口，而发展中国家因技术能力限制难以有效履行此类义务，加剧了国际法实施的不平等性。

2. “双轨制”模式之建构困境

联合国框架下的“双轨制”模式则暴露了国际治理机制的结构性矛盾。2017 年 UNGGE 谈判破裂直接反映出各国对网络空间军事化趋势的立场对立。有学者认为，所谓并行机制根本上是大国博弈的后果，两个方案根植于美俄两国对于网络空间治理话语权的长期分歧，在诞生和运行过程中就体现了各自诉求。²²

Cyberspace, at https://www.rand.org/pubs/research_reports/RR2081.html (Last visited on April 5, 2025).

²¹ 参见黄志雄：《论网络攻击在国际法上的归因》，载《环球法律评论》2014 年第 5 期，第 157-168 页。

²² Kevin Yu：《联合国并行机制 GGE 和 OEWS：管窥网络空间治理大国博弈》，载微信公众号“海国图智研究院”，2021 年 4 月 26 日上传。

“双轨制”模式本身存在的问题也使其在规制网络战方面难以取得迅速成效，其通过的决议并不具有明确、直接的法律拘束力。UNGGE 与 OEWG 在 2021 年的两份报告中明确其在网络和平利用方面制定的新规则仅是“自愿、非约束性”规范，参与国家可自由选择是否履行或者参考。²³UNGGE 与 OEWG 属于联合国框架下大会的特定专家组，大会通过的决议中只有针对联合国内部事务达成的决议对组织及会员国有拘束力，其他决议仅具有劝告性质。

总体而言，当前新兴国际法规则在网络空间的适用仍呈现“软法主导、硬法缺位”的特征，难以有效应对网络战技术迭代与战术隐蔽性带来的挑战。

三、规制困境的生成机理：规范供给与治理结构的双重失衡

（一）传统国际法体系的形成机制缺陷

1. 国际法造法机制的时代滞后性

国际法的造法机制本质上是基于主权国家间的共识和实践形成的，其发展速度往往难以跟上技术变革的步伐。在网络战领域，传统国际法规则的形成机制暴露出显著的时代滞后性。首先，国际法的核心原则和规则多是在工业时代背景下形成的，其适用范围和规范对象主要针对物理空间的军事行动。例如，《联合国宪章》第 2 条第 4 款关于“禁止使用武力”的规定，以及《日内瓦公约》中区分原则和比例原则的设定，均是基于传统战争形态和物理空间的军事行动，彼时的立法者无法预见网络技术对战争形态的革命性改造，导致既有规则在技术解释层面存在先天性缺陷。然而，网络战的虚拟性和技术复杂性使得这些原则在网络空间的适用面临挑战。例如“震网”病毒通过代码渗透导致伊朗核设施瘫痪，其物理破坏由离心机设备异常引发，但直接作用机制是数据篡改而非动能打击。传统国际法面对网络战后果的延时性（如病毒潜伏扩散）和系统性（如电网崩溃引发连锁反应）时，既缺乏即时评估的技术基准、又难以界定破坏的时空边界。这种规范与现象的错位，本质是工业时代战争逻辑对信息革命的失配。

其次，国际法的造法过程往往需要经过漫长的谈判和共识达成，而网络技术的快速发展使得这种机制难以及时回应现实需求。美国陆军参谋长前高级顾问 Maren Leed 曾提到，网络武器可以有許多适应环境的属性，从生命周期的成本角度看，它们比其他的武器系统更为优越。²⁴在网络武器飞速发展的背景下，法律更新周期与技术创新周期形成严重脱节，这种造法机制的滞后性导致传统国际法规则在网络战领域的适用存在规则空白和解释分歧，进而削弱了其规范效能。

2. 主权概念的技术性消解与规范重构困境

传统国际法体系以主权国家为核心，其规范设计和责任追究均建立在清晰的

²³ See General Assembly of the United Nations, *Report of the group of governmental experts on advancing responsible state behaviour in cyberspace in the context of international security*, at https://digitallibrary.un.org/record/3934214/files/A_76_135-EN.pdf (Last visited on April 5, 2025).

²⁴ Leed Maren, *Offensive Cyber Capabilities at the Operational Level*, Center for Strategic International Studies, p.1 (2013).

国家行为认定基础之上。然而，网络技术的匿名性和跨境性使得主权国家的行为边界在网络空间变得模糊，传统主权概念在网络战规制中的适用面临技术性消解的困境。

从表层来看，网络攻击的匿名性使得行为归因变得异常困难。如前所述，攻击者可以通过技术手段掩盖真实身份，甚至利用非国家行为体或第三方服务器实施攻击，这使得传统的国家责任认定机制难以有效运作。此外，非国家行为体的介入使得国家行为的认定变得更加复杂，而国家对网络攻击行为的监控和管理能力也难以达到传统物理空间的标准。

从深层来看，国际法追责机制的内在缺陷进一步加剧了规制困境。现行国际责任追究体系依赖于主权国家的自愿合作与情报共享，但网络战溯源的高度技术依赖性导致证据获取能力与司法管辖范围呈现不对称性。主要网络作战力量往往受特定国家司法庇护，且更为根本的是，国际法规则演进受制于国家间政治博弈，主要网络强国对规则解释权的主导与双重标准实践，使得国际社会难以就网络战规制形成实质性共识。这种制度性困境的本质，折射出主权国家有界性与网络空间无界性之间的深刻矛盾，最终导致传统国际法在网络战规制中陷入规范效力递减的困境。

（二）新兴国际法机制的治理结构矛盾

1. 软法生成路径的主体代表性缺陷

新兴国际法机制在网络空间的规制实践中，呈现出明显的软法主导特征。以《塔林手册 2.0》和 UNGGE 报告、OEWG 报告为代表的软法规则，试图在网络战领域填补传统国际法的规则空白。然而，这些软法规则的生成路径和主体代表性存在显著缺陷，导致其规范效力和普遍性受到质疑。

首先，《塔林手册 2.0》虽然通过系统性地将国际法原有规则通过解释引入网络战场景，并且也构建了相对完整的规则框架，但其本质仍为学术性重述，缺乏国际社会的普遍认可。但其“重述国际法”的本质决定了其无法突破物理战争范式的桎梏。网络武器的代际演进，如人工智能驱动的自主攻击系统，已超前于国际法发展，形成“技术领跑—规则跛行”的代际差。此外，手册的编纂主体以西方国家为主导，其规则设计未能充分考虑发展中国家在网络技术能力和军事实践上的差异性，内容和立场难免带有地缘政治倾向，削弱了其作为国际法共识载体的中立性。

其次，联合国框架下的“双轨制”机制在网络空间的规制实践中也暴露出代表性不足的问题。UNGGE 和 OEWG 的代表性主要集中在技术优势国家和西方盟友，而广大发展中国家在网络空间的意志和利益难以得到有效表达。美国国务院网络事务协调官米歇尔·马尔科夫（Michele Markoff）曾要求在 2021 年 UNGGE 最终报告文本中彻底删除本应出现在每一条款开头的“各国同意”一词，反映了普

遍的分歧。伊朗代表甚至曾因不满报告内容威胁退出 UNGGE 机制，反映出中小国家在双轨制中的边缘化困境，进而削弱了规则体系的普遍接受性。

2. 治理机制形成中的权力博弈失衡

新兴国际法机制在网络空间的治理实践中，不可避免地受到大国权力博弈的影响，因为发达国家与发展中国家在网络空间军事化、自卫权适用等核心议题上存在根本性立场冲突。在联合国框架下的网络空间国际规则博弈场中，UNGGE 被认为是代表西方网络发达国家由美国主导的进程，OEWG 被认为是代表网络发展中国家由中俄主导的进程。例如，以美国为首的西方国家依托技术优势，倾向于扩大《联合国宪章》第 51 条自卫权的解释范围，主张将网络攻击纳入“使用武力”范畴，并通过《塔林手册 2.0》等软法工具推动国际人道法向网络空间的延伸，从而为其网络军事行动提供合法性背书。而中国、俄罗斯等国家则主张网络空间应成为和平利用的领域，而非军事对抗的新战场，试图平衡技术霸权对国际法解释权的垄断。这种意识形态与战略利益的割裂，使得双轨制机制沦为大国争夺规则话语权的场域，而非凝聚共识的平台。

由此，“双轨制”机制陷入了“规制悖论”：既需要突破传统治理范式应对诸如网络战的新型网络威胁，又受制于国际体系固有的权力结构；既期待技术创新赋能治理变革，又面临技术政治化带来的制度解构风险。这种结构性矛盾的本质，是网络空间技术逻辑与政治逻辑尚未实现规范性耦合的必然产物。

四、中国应对网络战国际法规制困境的策略框架

（一）基本立场：坚持“网络空间命运共同体”核心理念

2015 年 12 月，习近平主席在第二届世界互联网大会开幕式上发表主旨演讲强调：“网络空间是人类共同的活动空间，网络空间前途命运应由世界各国共同掌握。各国应该加强沟通、扩大共识、深化合作，共同构建网络空间命运共同体。”²⁵这一论述为中国参与网络空间国际治理提供了根本遵循。2024 年 11 月，习近平主席在世界互联网大会乌镇峰会的致辞中指出，“中国愿同世界各国一道，把握信息革命发展的历史主动，携手构建网络空间命运共同体，让互联网更好造福人民、造福世界”，²⁶再次向国际社会发出了关于世界互联网治理与发展的中方愿景。网络空间的互联互通性打破了国家间物理边界，促使各国形成“一荣俱荣、一损俱损”的依存关系，这种依存关系不仅体现在经济发展层面，更深刻影响着国家安全、社会稳定与文化认同。因此，构建网络空间命运共同体绝非权宜之计，而是应对数字时代共同挑战、实现可持续发展的必然选择。

在全球网络空间治理机制构建过程中，我国应始终展现出负责任大国的担当，

²⁵ 《习近平在第二届世界互联网大会开幕式上的讲话》，载 https://www.gov.cn/xinwen/2015-12/16/content_5024712.htm，2025 年 4 月 5 日访问。

²⁶ 《习近平向 2024 年世界互联网大会乌镇峰会开幕视频致贺》，载 https://www.gov.cn/yaowen/liebiao/202411/content_6988342.htm，2025 年 4 月 5 日访问。

包括主张在网络空间坚持国际法的基本原则，包括主权原则、不干涉内政原则和禁止使用武力原则。网络空间不是法外之地，国际法应当在网络空间得到适用和遵循。其次，我国应当强调网络空间的和平利用，反对将网络空间军事化，主张各国应当共同努力，避免网络战对国际和平与安全造成的威胁，网络战的规制应当以预防和减少网络冲突为目标，通过国际合作与对话，构建网络空间的信任机制。最后，我国应主张通过国际合作，推动网络战国际法规制的完善。国际法规制的完善是应对网络战挑战的关键，而这一过程需要国际社会的共同努力。

（二）具体策略：多维度推进网络空间治理

1. 推动传统国际法规则的适应性修订

传统国际法规则在网络空间的适用面临诸多困境，特别是在网络攻击的阈值界定、国际人道法原则的适用以及归因机制方面。我国应当积极推动传统国际法规则在网络空间的适应性修订，以确保网络军事行动的法律界限清晰，避免各国在网络空间的行为失控。例如，我国可以在联合国框架下，推动《联合国宪章》第2条第4款和第51条中明确网络攻击的阈值界定，确保“使用武力”和“武力攻击”的认定标准在网络空间得到明确。通过明确网络攻击的法律定义和适用标准，各国在网络空间的行为将更加规范，网络战的风险将得到有效控制。

此外，强化平民与军事目标的区分原则是保护平民免受网络战伤害的重要保障。我国应始终秉持人道主义精神，积极参与国际人道主义救援行动。在网络战中，保护平民免受伤害是国际法的基本原则之一。国际社会应明确网络空间中平民与军事目标的区分标准，避免误伤无辜。这要求各国在网络攻击前进行详尽的目标审查，确保攻击行为严格限于军事目标。对于违反区分原则及比例原则的行为，我国应与国际社会一道予以谴责并采取相应措施，共同维护网络空间的和平与稳定。

最后，我国应当在国际社会中倡导建立科学合理的归因机制。网络攻击的匿名性和技术复杂性使得归因变得异常困难，这为网络攻击提供了便利条件。我国可以通过技术手段和国际合作，提升归因的准确性和可靠性，以确保网络军事行动的责任追究机制的有效性。例如，我国可以提出采用分布式账本技术（如区块链）等先进技术手段，记录并验证网络攻击的来源和行为轨迹，从而提高归因的透明度和公信力。通过建立科学合理的归因机制，我国可以为国际社会提供解决网络攻击归因困境的中国方案。

2. 推动软法治理的去政治化与共识凝聚

软法治理在网络空间国际法规制中具有重要作用，但其政治化倾向和非约束性特征使其在规则解释和实施中存在一定局限性。《塔林手册2.0》作为网络空间软法的代表性文献，在国际社会中具有一定的影响力。然而，其毕竟由西方主导编纂，其中代表西方立场的、可能引导“网络战合法化”结论的内容在国际社会上

仍具有争议。因此，我国应当推动软法治理的去政治化与共识凝聚，确保软法治理的公正性和中立性。例如，我国可以在 UNGGE 和 OEWG 中提倡软法规则的去政治化，推动各国在软法规则的制定和实施中达成共识。通过加强软法治理的共识塑造，我国可以为网络空间的和平与安全提供坚实的法律保障。

3. 强化网络战应对能力建设与风险预防

网络战能力建设是网络空间治理的重要组成部分，我国需要通过技术研发和人才培养，提升网络战的防御和响应能力。虽然反对网络空间军事化是我国应坚持的主张，然而由于近年“网络战”事件频发，我国应当有所警觉、加强网络战应对能力建设，例如，我国可以加大对网络安全技术的研发投入，提升网络攻击的检测、防御和溯源能力。通过强化网络战应对能力建设，我国可以为网络空间的安全与稳定提供坚实的技术保障。

此外，我国应当加强网络战风险预防机制的建设，通过风险评估和应急响应，确保网络空间的安全和稳定。例如，我国可以建立网络安全风险评估和预警系统，及时发现和修复潜在的安全漏洞，降低网络风险。同时，我国可以制定网络战争应急预案，确保在网络战争发生时，能够迅速和有效地进行应对。通过加强网络战风险预防机制的建设，我国可以为网络空间的安全与稳定提供坚实的保障。

4. 构建网络空间国际合作治理平台

超越“零和博弈”思维，我国需将人类命运共同体理念注入网络战规制体系。构建合作治理平台。中国代表团在 OEWG 发言中明确指出，“我们应团结而非分裂，对话而非对抗，对外释放维护网络空间和平与稳定的积极信号，携手构建网络空间命运共同体”。²⁷在“能力建设”议题下，中国代表团从宏观战略视野方面指出，“各国应推动国际网络安全合作与援助，实现公平合理普遍的互联网接入和 ICT 的普及化，弥合数字鸿沟，确保人人都能平等分享数字红利，实现全球共同、可持续发展。”²⁸具体而言，我国可加强“南南合作”，与发展中国家共同合作研究，从发展中国家视角阐释区分原则等规则的适用标准。此种策略既能保持规则灵活性，又可逐步改变西方主导的解释话语结构。并且，可通过“数字丝绸之路”建设，向发展中国家开放网络安全防护技术，降低其因技术脆弱性被迫卷入网络军备竞赛的风险。

五、结语

网络战作为一种新兴的战争形态，已经对国家间的安全与稳定构成了严峻挑战。国际法在面对网络战这一新型冲突形态时，存在着明显的滞后性与不完整性，导致现行国际法体系在规制网络战方面存在诸多困境。网络战的规制完善是一场

²⁷ 《中国代表团在联合国信息安全问题开放式工作组二期会上的发言》，载 http://un.china-mission.gov.cn/zgylhg/cjyjk/202204/t20220408_10665738.htm，2025 年 4 月 5 日访问。

²⁸ 《中国代表团在联合国信息安全开放式工作组首次会议关于能力建设的发言》，载 https://documents.unoda.org/wp-content/uploads/2021/12/Statement-of-China_ICT-OEWG-8th-plenary-meeting_capacity-building_DEC-16-PM_CHN.pdf，2025 年 4 月 5 日访问。

长期的革命，既需要坚守主权平等、人道主义等核心价值，也需包容技术创新带来的规则创新。

我国应对网络战国际法规制困境的策略选择，本质上是新兴大国推动全球治理体系变革的创造性实践。这种实践既要坚持维护网络空间主权安全的底线思维，又要展现推动构建网络空间命运共同体的责任担当；既要突破西方主导的规则制定垄断，又要避免陷入“另起炉灶”的制度对抗陷阱。因此，我国的应对策略应超越“被动适应”，以“规则共建者”的身份，在传统与新兴、安全与发展、单边与多边的动态平衡中寻求突破：通过解释性决议与技术标准，实现《联合国宪章》原则在网络空间的有机衔接；在规制网络战的同时，为人工智能等技术的和平利用保留空间；在强化自身能力的同时，推动联合国框架下的多边机制改革，防止“数字霸权”的滋生。唯有通过持续的规则创新、机制实践与价值传播，我国才能为网络空间治理贡献真正具有普遍意义的解决方案，推动国际法在数字时代的适应性演进与合法性更新，最终实现网络空间的持久和平与共同发展。

智能航运时代智能船舶远程操作中心的法律定位与治理规则重 构

朱艺辰*

摘要：伴随前沿技术的发展，智能船舶逐渐成为航运领域的焦点。智能船舶远程操作中心，作为智能船舶技术体系的核心组成部分，其重要性日益凸显。但由于远程操作中心的法律属性尚不明确，远程操作人员的责任界限模糊，并且其与船旗国实质联系被削弱，远程操作中心所在位置衍生出管辖冲突等问题，给智能船舶的法律治理带来了新的考验，亟待进一步加强重视。需明晰远程操作中心的法律地位以适用现有相关海事公约的适用、强化智能船舶与船旗国之间的实质性联系以明确船旗国对远程操作中心监督和检查的责任，并通过探索建立国际合作机制以解决管辖冲突等途径，以保障智能船舶远程操作的安全与规范，更好地适应智能航运时代的发展需求。强调智能船舶远程操作中心法律治理的重要性，以期为智能航运的发展提供法律保障。

关键词：智能船舶 远程操作中心 远程操作人员 法律治理 法律属性

伴随着大数据、移动物联网、云计算及人工智能等前沿信息技术的迅猛发展，无人驾驶技术亦迈入了一个日益成熟的新纪元。在这一背景下，智能船舶作为该技术与航运业深度融合的产物，预示着航运业未来发展的新方向。在全球海洋经济日益繁荣、国际贸易持续增长的今天，智能船舶作为航运业的新兴力量正迅速崛起，逐步成为推动航运业转型升级、实现高质量发展的关键力量，引领着海上运输与作业模式的深刻变革。智能船舶远程操作中心（Remote Operation Centre，简称 ROC）作为这一变革的关键枢纽，整合了先进的信息技术、通信手段和智能操控系统，使船舶得以突破传统的人工现场操控局限，实现远距离的精准监控与高效运行。其宛如智能船舶的“智慧大脑”，通过复杂的算法和数据传输，对船舶的航行方向、速度、动力系统、货物状态以及周边环境等进行实时分析和决策，极大地提升了航运的效率、安全性和经济效益，为全球贸易的顺畅流通开辟了新的航道。

然而，新生事物在成长过程中必然面临诸多挑战，在全球航运业加速迈向智

* 朱艺辰，哈尔滨工程大学人文社会科学学院 2023 级法律专业国际法方向硕士研究生。

本论文是国家社科基金一般项目“基于国际海事组织立法进程的中国智能船舶法律规制研究”

（23BFX198），黑龙江省哲学社会科学研究规划项目“《极地规则》生效背景下俄罗斯、加拿大北极通航法律政策演变及中国应对研究”（22GJB130），中央高校安全研究专项课题“北极航道航行安全法律问题研究”，黑龙江省法学会 2024 年重点课题“北极命运共同体视域下中国参与北极航运安全治理的路径研究”的阶段性成果。

能化的进程中，智能船舶远程操作中心作为实现船舶自主航行、岸基协同控制的核心枢纽，正推动海事法律体系经历从“物理空间治理”到“网络空间治理”的范式变革。智能船舶远程操作中心在法律治理方面陷入了重重困境。随着其应用范围的不断扩大和技术的持续迭代，现有的法律框架在面对这一全新的运营模式时，暴露出诸多的不适应性和空白之处。从远程操作中心的法律属性界定到远程操作人员的资质认证与责任划分，再到智能船舶运营过程中的监管机制，一系列复杂而棘手的问题接踵而至，给航运业的稳定发展带来了潜在的风险和不确定性。此外，智能船舶技术的快速发展与法律更新的相对迟缓之间的矛盾日益尖锐，进一步加剧了法律治理的难度。

因此，深入剖析智能船舶远程操作中心法律治理的现实困境，并积极探寻切实可行的纾解进路，已成为当下航运业乃至全球贸易领域亟待解决的关键课题。这不仅关乎航运企业的合法权益与运营安全，更关系到海洋环境的保护、海上生命财产的安全以及国际航运秩序的稳定与可持续发展。唯有通过构建健全、完善且与时俱进的法律治理体系，才能为智能船舶远程操作中心的蓬勃发展保驾护航，使其在法治的轨道上充分释放创新活力，助力航运业驶向更加智能化、高效化和安全化的未来。

一、问题的提出

（一）智能船舶“人的因素”的核心地位

智能船舶的蓬勃发展浪潮中，尽管先进技术不断革新并深度嵌入船舶的运行体系，但“人的因素”依然占据着关键地位。智能船舶的核心目标是实现安全、高效且精准的海上运输与作业，这一目标的达成无法脱离人的参与和把控。国际海事组织（International Maritime Organization，简称 IMO）其下设的法律委员会（Legal Committee，简称 LEG）在回顾关于海上水面自主船舶（Maritime Autonomous Surface Ships，简称 MASS）的不同公约监管范围界定结果时，提出的五个共性偏差，包括船长的作用和责任、远程操作人员的角色与职责、责任问题、智能船舶的相关定义和术语以及证书标准，都与“人的因素”密切相关¹。

在智能船舶技术的研发与应用进程中，人的因素是推动技术不断进步和优化的关键动力。在船舶系统的设计阶段，船员基于实际操作体验和需求，能够为技术开发者提供具有针对性的功能优化建议。例如在人机界面设计方面，船员的反馈能够引导开发者设计出布局简洁明了、操作流程便捷的界面，使船员能够快速上手并准确操作，减少因界面复杂或操作不便导致的人为错误。²在技术测试和

¹ Secretariat of Legal Committee of IMO, Report of the Legal Committee on the Work of its 109th Session, LEG 109-16-1, Retrieved from <[https://docs.imo.org/file:///D:/E6%A1%8C%E9%9D%A2/LEG%20109-16-1%20-%20Report%20Of%20The%20Legal%20Committee%20On%20The%20Work%20Of%20Its%20109Th%20Session%20\(Secretariat\).pdf](https://docs.imo.org/file:///D:/E6%A1%8C%E9%9D%A2/LEG%20109-16-1%20-%20Report%20Of%20The%20Legal%20Committee%20On%20The%20Work%20Of%20Its%20109Th%20Session%20(Secretariat).pdf)> (visited on 1 December 2024)

² McCabe, Faye. *Future human-system interaction techniques to influence perceived trust*. Diss. University of Birmingham, 2024.

验证过程中，人的专业知识和经验也不可或缺。船员能够参与测试方案的制定和实施，通过实际操作检验技术的可靠性和有效性，发现潜在问题并及时反馈给开发者进行改进，确保智能船舶技术能够满足实际航运需求。

在智能船舶的日常运营阶段，尽管自动化和智能技术不断发展，但人的因素对于保障船舶安全运行的仍至关重要。在船舶驾驶台和机舱的日常操作中，电子系统虽承担着重要的运行职责，但当这些先进系统出现故障或遭遇复杂情况时，船员凭借其专业技能和丰富经验，能够迅速接管操作，熟练地操控如阀门、轮机等关键设备，确保船舶维持正常运行状态，有效应对突发技术故障带来的安全风险。在复杂多变的海上环境中，如恶劣天气、船舶突发机械故障或面临潜在碰撞风险等紧急情况时，船员的经验和判断力更是成为保障船舶安全的核心力量。其能够迅速且准确地评估局势，果断采取诸如调整航向、控制航速、启动应急设备等一系列有效措施，及时化解危机，避免事故的发生与恶化。³

（二）远程操作中心的兴起

在智能船舶人机交互的场景中，人的因素同样是实现高效协同作业的核心要素。船员在与船上各类智能设备配合操作过程中，能够凭借自身的经验和判断力，根据实际航行情况灵活调整操作策略，确保任务顺利执行，在靠泊、离泊等操作难度较大的环节，船员能够充分发挥其操作技巧和经验，与智能设备紧密配合，完成精准操作。而随着智能技术在船舶领域的深度应用，智能船舶远程操作中心应运而生，传统的船上作业将逐渐被在远程操作中心进行的远程操作作业或智能航行系统所取代，传统船员的适航等基本义务由智能船舶远程操作人员代为履行，远程操作人员也就成为智能船舶的远程操作的主体。远程操作中心对于智能船舶而言至关重要，其主要作用在于监督智能船舶的运营状况，确保航行安全，同时优化航行路径和效率，在遇到如复杂海况、紧急避险等场景的必要情况时，远程操作人员能够迅速介入进行操作，通过预设的自动化程序或人工干预，对船舶进行远程控制，但这种控制权限的行使是有限的，在确保安全的前提下最小化人为干预，以充分发挥智能船舶的自主能力。无论是船员与船上各类智能设备的配合操作，还是远程操作中心人员与船舶自动化系统的协同工作，人类都承担着至关重要的任务。

二、智能船舶远程操作中心的法律定位辨析

（一）远程操作中心的法律实践

IMO 下设的海事安全委员会（Maritime Safety Committee，简称 MSC）在召开的第 99 次会议上，将智能船舶称为海上水面自主船舶，并按照自主等级将 MASS 细化为四个层次：“船员在船且辅以自动化系统与辅助决策支持的船舶”、“船员在船但实现远程操作的船舶”、“完全无船员在船但实施远程操作的船舶”，

³ Lützhöft, Margareta, and Jonathan Earchy. *Human-Centred Autonomous Shipping*. Taylor & Francis, 2024.

以及“实现完全自主航行的船舶”。⁴中国船级社发布的《智能船舶规范（2024）》将远程控制船舶定义为船舶能够被船舶之外的一个远程控制站或控制位置进行控制，以实现船舶的运行，并将远程控制船舶分为 R1 和 R2 两个等级，R1 船舶主要功能由远程控制站控制操作，船上船员对船舶状态进行监视，在应急情况或必要时接管船舶的操作，根据设计确定的船舶运行场景，对非远程控制的系统和设备进行操作；R2 船舶船上无船员，完全实现远程操作⁵。

IMO 对 MASS 的分级中，L2、L3、L4 级 MASS 均涉及到远程操作船舶，其中第二级属于有人在船的远程操作情形，L3、L4 级 MASS 属于无人在船的情形，L4 级 MASS 即使实现自主航行，但也需要由远程操作中心对其进行监控，以更好地应对紧急情况。当前在学术研究领域、行业规范、政策指导性文件等相关文献资料中，与“远程操作中心”相关的术语还包括“岸基操控中心”、“远程控制站”、“远程操作工作站”，虽然这些术语本质上指向类似的功能实体，但它们在各自的语境下所强调的侧重点各不相同。岸基操控中心（Shore Control Centre，简称 SCC）通常是指位于陆地上的操作控制中心，其包括监控中心与信息支持中心，对智能船舶进行遥测遥控并提供信息支持，在船岸之间传递信息和指令⁶，学界通常使用这一术语。远程控制站（Remote Control Station）则强调对单个无人船的进行实时、精确的控制，设备相对集中，我国海事局发布的《船舶自主航行试验技术与检验暂行规则》、《无人艇技术与检验暂行规则（征求意见稿）》，中国船级社发布的《智能船舶规范（2024）》中均使用这一术语，在 IMO《海上自主水面船舶规则》（International Code of Safety for Maritime Autonomous Surface Ships，简称 MASS 规则）草案制定过程中该术语也被提及，但由于在 SOLAS 公约中存在“控制站（Control Stations）”的概念，为避免混淆，后又更改为远程操作工作站（Remote Operations Workstation）。在 MSC 第 107 次会议上，考虑到多个远程控制站可以设立在一个远程操作中心内，因此为简化概念，原则上同意了远程操作中心的概念。⁷随后在出台的 MASS 规则草案中，即使用远程操作中心的概念，将其定义为一个可以操作智能船舶的某些或所有功能的远离智能船舶的地点。⁸远程操作中心的特点在于对船舶集中管理的能力，一个 ROC 不仅仅局限于对一艘船舶的管理，其能够同时监控并管理多艘船舶，并且使用 ROC 是强调远程功

⁴ International Maritime Organization (IMO), *IMO takes first steps to address autonomous ships*, at <https://www.imo.org/en/MediaCentre/PressBriefings/Pages/08-MSC-99-MASS-scoping.aspx> (Last visited on December 24, 2024)

⁵ 中国船级社《智能船舶规范（2024）》第 8.1.3 条。

⁶ Dybvik, H., Veitch, E., & Steinert, M., *Exploring challenges with designing and developing shore control centers (SCC) for autonomous ships*. in: Proceedings of the Design Society, DESIGN Conference, Cambridge University Press, 2020.

⁷ International Maritime Organization (IMO), *Report of the Maritime Safety Committee on its 107th Session*, MSC 107/20 (26 June 2023).

⁸ International Maritime Organization (IMO), *Report of the Correspondence Group on Maritime Autonomous Surface Ships*, MSC 108/4 (13 February 2024), Draft International Code of Safety for Maritime Autonomous Surface Ships (MASS Code).

能并降级控制功能，对自主船来说，ROC 的主要任务是监督船舶运营而不是控制船舶。⁹

由此可知，远程操作中心，是一个能够从集中位置对智能船舶进行操作的核心场所，其不仅仅是一个物理空间或技术平台，更是连接船舶、岸基、数据、人员与决策的综合体，远程操作人员通常使用先进的网络通信和计算技术来监控和操作远程系统的运行，以实现船舶的实时精准的远程监控、控制、优化与决策支持，提升航运效率与安全水平。

（二）远程操作中心的法律属性

在当前的智能航运领域，远程操作中心的法律属性尚未明确，但对智能船舶远程操作中心的法律属性进行探讨至关重要。首先，从责任界定角度来看，远程操作中心的法律定位模糊，明确其法律属性有助于在事故发生时，准确界定船舶所有人、运营人、远程操作中心及相关技术提供商的责任范围，保障各方合法权益，比如若远程操作中心被视为船舶的一部分，那么船舶所有人可能要承担主要责任；但若其被看作独立的服务实体，责任的划分就会更加复杂。其次，对于法律体系而言，智能船舶是新兴事物，传统的海事法律可能无法完全适应其发展需求，确定远程操作中心法律属性，可以为海事法律的修订和完善提供方向，这不仅有助于我国建立健全智能船舶相关法律体系，也能在国际海事法律制定中发挥积极作用。再次，明确远程操作中心的法律属性能为行业发展提供规范指引，确定其法律属性后，可制定相应的监管政策和标准，促使相关企业规范自身行为，促进智能船舶行业健康有序发展。最后，从安全保障角度考虑，明确远程操作中心的法律属性可以促使相关海事部门加强对其的监管，确保其在操作过程中严格遵守安全规范，降低事故风险，为海洋环境的安全提供有力保障。

从船舶法律定义来看，远程操作中心显然不属于传统意义上的船舶范畴之内；而至于其是否构成智能船舶的组成部分，则存在争议。在学术界，远程操作中心与智能船舶之间的关系存在着一体说和独立说，一体说认为，智能船舶若脱离远程操作中心则无法运行，二者不可分割，因此应将远程操作中心视为智能船舶的一部分，并以整体的方法对二者进行评估；独立说则强调，智能船舶与远程操作中心各自作为独立的实体存在，二者之间并非固定的一一对应关系，一个远程操作中心可能控制多艘智能船舶，而一艘智能船舶在运行过程中也可能被多个远程操作中心同时控制或进行交替控制，二者并非包含关系，只是功能上的相互配合关系。¹⁰比如在智能船舶安全检查方面，远程操作中心是否应该被纳入安全检查的范围以及如何对远程操作中心进行安全检查。就远程操作船而言，远程操作中心相当于传统有人驾驶船舶中驾驶室的角色；就自主船而言，在正常作业的情况

⁹ Lunde - Hanssen, L. S. (2020). *Identification of information requirements in ROC operations room*. IFE/E..

¹⁰ 参见张清宝：《无人船海事监管法律问题研究》，载《法学前沿》2023 年第 4 卷，第 8-17 页。

下，其能够依赖其内置的智能自主航行系统实现自主运行，但也并不意味着其能够完全脱离外部监控，远程操作中心持续监测其航行状态与作业活动，在突发应急事件时，迅速切换至远程操作模式，以确保船舶运行的安全与效率。由此可见，智能船舶的适航与安全航行保障已不再单纯依赖于船体自身的技术状况与性能，而是扩展到远程操作中心的稳定运行以及其与智能船舶之间的网络通信能力上，二者共同组成了一个系统。远程操作中心作为这一系统中不可或缺的关键部分，其安全性与稳定性直接关乎智能船舶的整体安全性能。因此，远程操作中心必须被纳入安全检查的范畴。然而，如何对远程操作中心实施有效的安全检查，就涉及到何方主体应承担起对远程操作中心进行安全检查的职责，由此也就需明确远程操作中心法律属性。若采纳一体说，港口国可国际公约的明确授权，对涵盖远程操作中心在内的整个无人船体系执行全面的安全检查，但对船旗国而言则需分情况讨论，若远程操作中心设立于其操控的智能船舶的同一船旗国领土内，即两者均处于船旗国的直接主权管辖之下，船旗国具备对其实施安全检查的合理性与合法性，但远程操作中心若设立于其操控的智能船舶所属的船旗国之外，则与智能船舶分属于不同国家的管辖范围，但在逻辑上一国管辖下的远程操作中心明显无法成为另一国注册船舶的构成部分。¹¹若采纳独立说，根据国际公约，港口国的安全检查权仅限于到港的外籍智能船舶本身，现行公约并未赋予港口国对远程操作设施进行安全审查的权力，由此港口国无权对远程操作中心进行任何形式的检查。船旗国同样需分情况探讨，若远程操作中心在船旗国境内，则可基于属地管辖权对其进行安全检查，但若位于船旗国境外，船旗国则无权对其进行安全检查，此时智能船舶的安全运行则难以得到保证。

智能船舶远程操作中心及其人员的出现，作为智能船舶运行体系的核心组成部分，其重要性日益凸显。但这一新兴领域在快速发展的同时，也面临着诸多监管、运营、安全等方面的挑战，明晰远程操作中心的法律属性，构建科学、合理、有效的法律治理体系，成为当前亟待解决的问题。

三、智能船舶远程操作中心治理的现实困境解构

（一）远程操作人员的责任界限模糊

海事法律体系的发展相对滞后于智能船舶技术的快速进步。传统的海事法规大多是基于有人驾驶船舶的运营模式制定的，对于远程操作人员这种新兴的职业角色和远程操作模式缺乏充分的考虑和相应的规定。虽然 IMO 等国际组织的智能船舶研究进程已取得显著成果，但目前尚未形成完善的、专门针对远程操作人员责任界定的国际法律规则，各国国内立法也处于探索和完善阶段，这使得远程操作人员在法律适用上存在诸多空白和模糊之处，在此情况下仍需为有关法律问

¹¹ 参见张清宝：《无人船海事监管法律问题研究》，载《法学前沿》2023 年第 4 卷，第 8-17 页。

题提供优化措施、减除阻碍因素。¹²

关于《1978 年海员培训、发证和值班标准国际公约》(International Convention on Standards of Training, Certification, and Watch keeping for Seafarers, 简称“STCW”)对智能船舶的法律适用, 公约明确了海员(船员)的“在船”性质, 表面上将远程操作人员认定为船员不具备可能性, 从而使得远程操作人员在船员资质认定及适任标准方面较难直接使用有关条款。但就立法目的而言, STCW 规定在船应当基于传统海运体系下一般船舶有人驾驶要求, 而未完全预料智能船舶无人化趋势及所致船员身份认定问题, 明确在船性质则要求船长及其他船员能够恪守航行准则, 规范职务行为, 以维护海上人命、财产安全。智能船舶操纵监控以及船舶安保等多由远程操作人员进行, 不仅限于船上船长或其他船员, 能够脱离传统物理空间限制而维持常规航行。

既然现行 STCW 难以直接适用远程操作人员, 则应当考虑以修订公约或其他形式构建远程操作人员操作规范体系, 在这种情况下, 确定远程操作人员的能力并明确其责任界限至关重要。¹³传统海事法律对船员在船履行职责期间的责任有较为明确的规定, 涵盖了船舶航行、货物管理、安全维护等多个方面。然而, 远程操作人员并不在船上工作, 其工作环境和方式与传统船员存在显著差异, 但在某些情况下, 又与传统船员的职责存在交叉和重叠。

一方面, 在责任认定的维度上, 相较于传统船员直接在船舶上执行任务的直观责任认定模式, 远程操作人员对船舶的操控行为跨越了物理空间的限制, 这一转变增加了责任界定的难度。¹⁴在事故发生时, 难以区分是船舶自身的机械故障、智能系统的程序漏洞, 还是远程操作人员的失误所导致, 当船舶出现偏离航线的情况, 可能是由于远程操作人员输入了错误的指令, 但也可能是船舶的定位系统出现故障或者其他原因, 若由此导致船舶发生碰撞事故, 界定远程操作人员应承担的责任份额成为了一项极具挑战性的任务, 不仅需要详尽的事故调查, 还需要依赖于先进的技术手段和明确的法律判断标准。此外, 智能船舶的操作流程涉及多个环节与系统的协同工作, 远程操作中的通信环节也可能出现问题, 信号中断、数据传输延迟等情况均可能影响远程操作人员的判断与操作, 但目前对于这些通信故障导致的责任归属尚未有清晰的界定规则, 使得远程操作人员在操作流程中的责任界限模糊难辨。

另一方面, 从责任划分的主体关系来看, 远程操作人员与船舶所有人、管理人以及技术支持方等关键主体之间的责任界定同样缺乏统一且明确的标准。在实

¹² 参见谭韵璇, 刘博明等:《智能船舶技术发展对船员职业的影响分析》, 载《中国海事》2024 年 1 期第 26-28 页。

¹³ Kuntasa, Tanan, and Taih-Cherng Lirn. *A conceptual model of autonomous ship remote operators' competency*. *The Journal of Navigation* (2024): p.1-22.

¹⁴ Choi, Junghwan, and Sangil Lee. *Legal status of the remote operator in maritime autonomous surface ships (mass) under maritime law*. *Ocean Development & International Law* 52.4 (2022): p.445-462.

践中,船舶所有人可能会倾向于将事故责任归咎于远程操作人员的不当操作,以此作为推卸自身在船舶日常维护与安全管理上应负责任的借口。同时,技术支持方也可能因远程操作人员的直接介入,而试图模糊其在系统稳定性保障、故障预警及应急响应等方面的责任界限。这种相互推诿责任的风险可能会加剧责任认定的难度。构建一套科学、合理且具备可操作性的责任划分体系,对于明确各方责任、保障远程操控船舶的安全运营具有至关重要的意义。

(二) 远程操作中心与船旗国实质联系削弱

《联合国海洋法公约》(United Nations Convention on the Law of the Sea, 简称 UNCLOS) 第 91 条赋予船旗国确定给与船舶其国籍、为船舶登记和提供悬挂该旗帜的权利的条件,对船舶授予登记的基本要求是船旗国与船舶之间必须建立“真正联系”,但“真正联系”并没有一个明确的界定。¹⁵1986 年通过的《联合国船舶登记条件公约》对于船旗国和船舶的真正联系原则解释为要求船旗国与船舶之间建立实质性联系,包括经济联系(如资金参与船舶所有权)、国籍联系(如配备本国船员)和管理联系(如船旗国内设立船公司代表处或者管理人),以确保船旗国能够有效地对船舶进行管辖和控制,解决开放登记制度带来的问题,但该公约至今未能生效。¹⁶在国际海洋法法庭的“塞加号”案(The M/V“SAIGA”(No. 2) Case)中,“真正联系”的要求被解释为船舶注册国需确保船旗国义务的有效履行,这种义务是由船舶登记的国家保证的,船旗国必须“对该船行使有效的管辖权和控制权,以确保该船按照公认的国际条例、程序和惯例运作”,法庭还强调,国家不能仅以船舶与船旗国之间没有“真正联系”为由拒绝承认船舶的国籍。¹⁷在弗吉尼亚 G 号商船(The M/V“Virginia G”Case)案中,国际海洋法庭进一步强调了船旗国在船舶监管中的核心作用,法庭认为一旦船舶注册,船旗国有义务根据 UNCLOS 第 94 条对其行使有效管辖和控制,以确保其遵守普遍接受的国际规定,这构成了“真正联系”的本质。¹⁸

就传统船舶而言,船旗国往往通常将其核心管辖权聚焦于海上劳动活动的监管上,通过对船舶上的船长及船员的监管,来实现对船舶整体运营的有效管辖与控制。具体而言,船旗国可以制定一系列的法律法规和规章制度,对船长及船员的资质、培训、工作条件等方面进行详细的规范和要求,同时,还可以通过定期的检查和监督,确保船舶上的劳动活动符合相关的标准和规范,从而保障船舶的安全运营和海洋环境的保护。但智能船舶作为新兴技术产物的出现,船旗国长期以来的监管模式其与船舶之间的“真正联系”原则就不得不被重新审视。就智能船

¹⁵ 联合国《联合国海洋法公约》第 91 条。

¹⁶ 联合国《联合国船舶登记条件公约》第 5-9 条。

¹⁷ International Tribunal for the Law of the Sea, Saint Vincent and the Grenadines v. Guinea, The M/V 'SAIGA' (No. 2) Case (1 July 1999), Reports of Judgments, Advisory Opinions and Orders.

¹⁸ International Tribunal for the Law of the Sea, Argentina v. Ghana, The M/V 'Virginia G' Case (14 April 2014), Reports of Judgments, Advisory Opinions and Orders.

船而言，其运营正逐步脱离传统的人工直接干预模式，智能船舶的大部分操作与联络工作不再依赖于船上的船长和船员，而是由远程操作中心及其人员承担，或是由预设的智能系统进行自主处理。但远程操作人员和智能系统的设计者并不需要在船上进行工作，也不需要具有船旗国国籍，远程操作中心也并不需要船旗国境内设立，其均可能位于世界各地，完全打破了传统船舶运营的地域限制。船旗国难以对智能船舶行使包括船舶登记、证书发放、检查、事故调查、执行法律法规等在行政、技术和社会事务等方面的有效管辖和控制¹⁹，这也就表明确立船旗国与智能船舶之间的“真正联系”将更可能成为一种形式意义上的存在。有观点认为需明确“真正联系”的真正内涵，以有效敦促船旗国对于船舶的监管；也有观点认为船舶一旦注册登记即实现了与船旗国所谓的“真正联系”²⁰；甚至有观点认为船旗国与智能船舶之间存在“真正联系”是虚拟的、不真实的，二者之间无法建立明确的联系，“真正联系”的概念无法适用。²¹

（三）远程操作中心所在位置衍生管辖冲突

控制智能船舶的远程操作中心及其人员所在的位置衍生了管辖冲突的问题，进而导致在智能船舶的检验与安全检查、航行安全的监管、事故责任的判定，以及远程操作人员的劳动权益与安全保障等多个关键领域的管辖权归属问题上产生争议，影响相关法律的适用性及执行效率。

首先，船旗国管辖权与其他国家对智能船舶远程操作人员的管辖权可能产生管辖冲突，当远程操作中心位于船旗国境内时，船旗国自然能够便利地行使其对远程操作人员的管辖权和控制权，确保操作的合规性与安全性。然而，若远程操作中心被设立在船旗国以外的国家领土上，情况则变得复杂。虽然船旗国基于属人管辖原则，对远程操作人员拥有专属管辖权，但远程操作中心所在国依据属地管辖原则，同样可主张对该中心及其人员的管辖权。这种双重管辖权的存在，极易导致远程操作中心所在国的属地管辖权与船旗国的管辖权之间产生管辖冲突。以远程操作人员在工作时受到损害为例，若将其视为智能船舶的船员，船旗国可能会基于船员保护的原则，主张对他们的管辖权。然而，由于远程操作人员可能身处其他国家，根据《2006 年海事劳工公约》（Maritime Labour Convention, 2006，简称 MLC）的相关规定，他们并不属于船旗国应直接保护的海员范畴。²²因此，船旗国仅在存在例外规定的情况下，才能行使有限的管辖权，难以对远程操作人员实施全面、有效的管辖和控制。而远程操作中心所在国，虽然具有属地管辖权，但根据 UNCLOS 的相关规定，却无需承担相应的管辖义务，²³进一步加剧了管辖

¹⁹ 联合国《联合国海洋法公约》第 94 条。

²⁰ Strati, Anastasia, Gavouneli, Maria & Skourtos, Nikolaos S. (eds.) (2006). *Unresolved issues and new challenges to the law of the sea: time before and time after*. Vol. 54. Martinus Nijhoff Publishers.

²¹ Van Hooydonk, Eric, *The law of unmanned merchant shipping—an exploration*, The Journal of International Maritime Law, Vol.20, No.3 (2014), p.403 -423.

²² 国际劳工组织《2006 年海事劳工公约》第 4-5 条。

²³ Secretariat of Legal Committee of IMO, Report of the Legal Committee on the Work of its 111th Session, LEG

权的模糊性。在此情境下，则难以明确判定是由船旗国行使管辖权，还是由远程操作中心所在国行使属地管辖权。远程操作人员虽然不直接身处海上环境，但他们的工作直接关系到智能船舶的安全与航行效率，他们的劳动权益、工作环境安全等问题同样不容忽视。然而，由于管辖冲突的存在，远程操作人员可能面临法律保护不足的风险，影响其权益保障。

其次，当控制智能船舶的远程操作中心位于另一艘具有不同国籍的船舶上时，情况更为复杂，两艘船舶分别属于不同的船旗国，²⁴各自受其船旗国的管辖。远程操作中心所控制的智能船舶的船旗国，可能会对远程操作中心及其人员主张管辖权，以确保其操作符合本国的法律法规，但远程操作中心所在船的船旗国，同样可能基于属地管辖原则，主张对远程操作中心及其人员的管辖权。在这种情况下，法律适用和责任判定将变得更加复杂。

此外，不同国家对于 UNCLOS 等国际法规在智能船舶领域内的具体应用与解释，可能因各自法律体系、技术发展水平及海洋战略利益的不同而存在差异，这种差异可能体现在对国际法规条文理解上的细微差别，也可能深刻影响到各国在制定相应的国内法律与规章制度时所采纳的标准与原则，尤其在智能船舶这一新兴术领域，各国法律制度的创新性与前瞻性将直接影响其对智能船舶的监管方式与力度。进而言之，由于各国对 UNCLOS 等国际法规理解上的差异，加之智能船舶技术本身的复杂性与跨界性，导致在涉及智能船舶的管辖权规定上也可能存在不同，包括在事故调查、责任追究以及损害赔偿等关键环节引发冲突，由此也就增加了国际间协调与合作的难度。如何在尊重各国法律制度差异的基础上，建立一种更加统一、高效且被广泛接受的国际法律框架，以有效应对智能船舶带来的新型法律挑战，也成为当前需解决的问题。

四、智能船舶远程操作中心治理规则的重构路径

（一）明晰远程操作中心的法律属性

目前，国际公约、各国立法以及国内法律法规对于智能船舶远程操作中心的法律属性尚未有明确的规定，需尽快明确远程操作中心的法律属性。笔者认为，在此问题上，可采纳“一体说”，将远程操作中心视为智能船舶的组成部分。从智能船舶的整体系统架构来看，远程操作中心与智能船舶之间存在着紧密的联系，其在智能船舶的运作体系中，扮演着类似于传统船舶驾驶台或控制中心的角色，是确保智能船舶正常运行不可或缺的关键要素。在传统船舶的航行过程中，船舶的操控完全依赖于船舶驾驶人员在船上专属的驾驶室内，执行一系列操作指令，以实现对船舶的全面驾驭与操控。而在智能航运领域这一传统模式得到了根本性

111 - 17, Retrieved from <<https://docs.imo.org/Category.aspx?cid=45&documentlang=en%2Czh&session=111>> (visited on 29 November 2024)

²⁴ Dong, Bing, Bautista, Luis & Zhu, Lei, *Navigating uncharted waters: Challenges and regulatory solutions for flag state jurisdiction of Maritime Autonomous Surface Ships under UNCLOS*, *Marine Policy*, Vol.161:1.

的转变，智能船舶的航行不再依赖于直接的人为操作于船舶之上，船舶的“驾驶室”概念在物理空间上发生了迁移，船舶运行转而依赖于远程操作人员在远离船舶的远程操作中心内，通过科技手段对船舶的航行状态、设备运行等进行实时掌控和调整，进行远程操控，从而确保智能船舶能够安全、高效地执行航行任务。从远程操作中心的主要功能来看，其包括持续监测船舶运行、导航和控制、建立通信链路、决策支持以及船上系统和设备的维护和诊断等²⁵，可看出远程操作中心的功能与船舶的运行、安全、管理等方面密切相关，通过远程的方式为船舶提供了全方位的支持和保障，由此可以认为远程操作中心是智能船舶系统不可或缺的关键组成部分。

因此，建议 IMO 在 MASS 规则中明确远程操作中心的法律属性，明确其与智能船舶之间的法律关系，赋予其相应的权利和义务，以确保其在航行过程中既能有效履行监控、指挥与控制等核心职能，又能承担相应的法律责任与风险，并进一步明确细化远程操作中心的位置、功能定位、技术要求等关键要素，以确保各国在立法与实践中有统一的参照标准。

（二）明确船旗国监督和检查的责任

尽管“真正联系”的概念、要求均未界定，但仍需要进一步明确智能船舶与船旗国之间“真正联系”的具体要求和标准，以确保船旗国与智能船舶之间更紧密的联系，明确智能船舶船旗国对远程操作中心监督和检查的责任。

IMO 的 MASS 联合工作组（The Joint MSC-LEG-FAL Working Group）在第 2 次会议中讨论船旗国对远程操作中心所在地的管辖权和责任问题时，确定可采用与《国际船舶安全营运和防止污染管理规则》结构（International Safety Management Code，简称 ISM 规则）类似的监督框架，由船旗国政府对位于其领土管辖范围之外、悬挂其国旗经营船舶的“公司”进行监督，对其旗下运营的船舶进行认证。关于 ROC 和 MASS 认证的监督机制应涵盖 ROC 和 MASS 两个级别，MASS 的船旗国应负责 MASS 和相关 ROC 的验证、调查和认证，运营 MASS 和 ROC 的公司应分别进行认证，然后在验证其兼容性后进行联合认证，每个船旗国应对其旗下旗帜的 MASS 与特定 ROC 的组合认证负责。对 ROC 及其设计的认证应确保与相关 MASS 的兼容性，ROC 的组织和管理认证应确保 MASS 特定任务和 ROC 特定任务的执行；对 MASS 的认证则需包含 MASS 船旗国对 MASS 和 ROC 之间关系进行认证。认证应独立于另一个 MASS 的管理、操作或技术要求。并应规定 MASS 和 ROC 由同一国际安全管理（ISM）公司运营，负责 MASS 的 ISM 公司应按照 ISM 规则维持责任并进行认证。

IMO 在 MASS 规则草案中明确，详细规定远程操作中心及其人员在智能船舶运营中的角色和责任，应要求远程操作中心定期报告智能船舶的运行情况，以

²⁵ ICMASS 2018, *Proceedings of the 1st International Conference on Maritime Autonomous Surface Ships*. Online. SINTEF, 2018. Available from: <https://hdl.handle.net/11250/2684110>.

便及时掌握船舶的状态，加强智能船舶与船旗国之间的实质联系，确保智能船舶的安全运营。可倡导船旗国与船级社进行合作²⁶，船旗国可以通过委托的船级社对智能船舶的关键方面进行监督和控制，以确保智能船舶保持良好的技术状态并符合环保要求，并督促船级社针对远程操作人员工作地点等方面制定和实施相应的技术标准。此外，也可考虑比利时、利比里亚和韩国通过引入远程操作管理（ROM）的概念²⁷，经认证的 ROM 公司持有由给定船旗国颁发的远程操作管理合规文件（ROM DoC），作为 ISM 规则的补充安全机制。

（三）推进国际监管合作机制的建立

船旗国可以制定具体的规定和标准，要求智能船舶的远程操作中心位于其领土内，以建立更明确的实质联系，这样虽然在一定程度上有助于加强船旗国对智能船舶的管辖和控制，但从长远来看，伴随着智能船舶技术的飞速发展与全球化合作的日益加深，单一强调船旗国对远程操作中心地理位置的限制，已不符合当前智能船舶产业的发展趋势。

为了促进智能船舶产业的健康发展，可由 IMO 应发挥主导作用，引领国际间建立灵活高效的监管合作机制。一方面，对于远程操作中心在船旗国领土外的情况，船旗国管辖权不应损害远程操作中心及其人员所在国家的管辖权，应由船旗国与远程操作中心所在国家进行合作，以确保海上航行安全、环境保护及船员权益的有效维护。如两国签署的合作协议，明确各自在监管方面的责任与权限，建立信息共享平台，允许船旗国与远程操作中心所在国家实时交换关于船舶动态、航行计划、安全警报、环境风险及船员健康状况等关键信息，以有效平衡船旗国与远程操作中心所在国家之间的管辖权。另一方面，船旗国应加强与港口国的合作，港口国应要求智能船舶及其远程操作中心建立完善的数据管理和安全体系，确保数据的真实性、完整性和安全性，²⁸并参考现有区域备忘录中船舶港口国监督程序方法制定检查程序与执行标准，通过跨区域港口国合作协调机制强化对智能船舶的监管效果。²⁹中国向 MASS 联合工作中提交的文件中建议智能船舶及其远程操作中心应持有 2023 年《港口国监督程序》（The Procedures for Port State Control, 2023）（A.1185(33)号决议）附录 12 中所列的适当证书和文件，以证明符合相关要求并支持安全操作；对于无船员的智能船舶，应考虑豁免仅与船上人类活动相关的记录、手册和指示；明确智能船舶对电子格式证书和文件（包括记录、

²⁶ Chen, Hao, *Challenges Raised by Unmanned Ships and Flag State's Possible Jurisdiction Cooperations with Other Stakeholders, Uncrewed Vessels and International Law*, p.206-236 (2024).

²⁷ Belgium, Liberia & Republic of Korea, *DEVELOPMENT OF A GOAL - BASED INSTRUMENT FOR MARITIME AUTONOMOUS SURFACE SHIPS (MASS)*, Concepts on the management of remote operations, Retrieved from <<https://docs.imo.org/Category.aspx?cid=49&session=108&documentlang=en%2Czh>> (visited on 11 November 2024)

²⁸ Demirci, Seyid Mahmud Esad & Cicek, Kadir, *Intelligent ship inspection analytics: Ship deficiency data mining for port state control*, Ocean Engineering (2023).

²⁹ 参见蔡莉妍：《智能船舶法律规制的困境与突破》，载《华南理工大学学报(社会科学版)》2020 年第 22 卷第 6 期，第 7 页。

手册和指示)的使用和接受。³⁰船旗国应积极配合港口国的检查工作,提供必要的智能船舶信息和支持,远程操作中心也应保存好必须配备的证书文件、操作记录并提供访问权限,以便港口国查验,在检查过程中发现的任何缺陷或违规行为,应及时通知船旗国,并共同商讨解决方案。同时,船旗国在保持必要监管的同时,应承认并接受其他国家或地区在智能船舶远程操作、数据管理、安全维护等方面所采取的有效措施和认证结果,减少不必要的重复监管和壁垒。

五、结语

作为全球海洋大国、航运大国,中国在智能船舶技术、海事监管等方面积累了丰富的经验。技术层面,中国在远程通信、实时数据分析及智能决策支持等方面取得了显著进展,为智能船舶的远程精准操控提供了坚实的技术支撑;海事监管层面,通过不断升级的现代化监测手段,如卫星遥感、雷达监测、AIS 系统等,对智能船舶的运行状态进行全方位、实时的监控,确保其在海上的航行安全有序。秉持开放合作的精神,中国始终积极参与 IMO 等全球性机构的各项活动,主动与世界各国分享智能船舶技术与海事监管经验,助力全球智能船舶产业的发展。当前智能船舶远程操作中心的法律治理问题尚属新兴领域,现行法律法规大多基于传统船舶制定,难以直接适用于智能船舶的远程操作模式,其不仅关系到智能航运技术的健康发展,更直接影响到海事安全及国际航运秩序的稳定。面对既存的法律治理空白,中国应陆续调整国内法规、完善海事监管,并适时向 IMO 提交反馈建议,在 MASS 规则制定中引领话语权,尽快明确远程操作中心的法律属性,加强智能船舶与船旗国的实质联系,解决船旗国与远程操作中心所在国家之间的管辖冲突。

随着 IMO 等国际组织对智能船舶法律治理框架的持续优化,以及世界各国在立法、司法层面的积极探索与深入实践,有望建立健全一套科学合理、切实可行智能船舶远程操作中心的法律治理规则体系,为智能船舶安全、稳定运行提供法律保障,以更好地迎接智能航运时代的到来,构建更加安全可靠、绿色环保、高效便捷的全球航运新生态。

³⁰ Secretariat of MASS-JWG, *DEVELOPMENT OF A GOAL-BASED INSTRUMENT FOR MARITIME AUTONOMOUS SURFACE SHIPS (MASS)* Report of the MSC-LEG-FAL Joint Working Group on Maritime Autonomous Surface Ships (MASS) on its third session. MSC 109/5/2(30 August 2024).

Reconstruction of the Legal Positioning and Governance Rules of the Remote Operation Center of Smart Ships in the Era of Smart Shipping

ZHU Yichen

Abstract: Along with the development of cutting-edge technology, intelligent ships have gradually become the focus of the shipping field. As the core component of the intelligent ship technology system, the importance of the remote operation center of intelligent ships is becoming more and more prominent. However, as the legal attributes of the remote operation center are still unclear, the responsibility boundaries of the remote operators are blurred, and the substantive connection with the flag state has been weakened, the location of the remote operation center has given rise to jurisdictional conflicts and other problems, which has brought a new test to the legal governance of the intelligent ships and urgently needs to be further strengthened and paid more attention to. It is necessary to clarify the legal status of remote operation centers for the application of existing relevant maritime conventions, strengthen the substantive connection between intelligent ships and flag states to clarify the responsibility of flag states for the supervision and inspection of remote operation centers, and explore the establishment of an international cooperation mechanism for the resolution of jurisdictional conflicts in order to safeguard the safety and standardization of the remote operation of intelligent ships and to better adapt to the needs of the development of the era of intelligent shipping. The importance of the legal governance of the remote operation center of smart ships is emphasized with a view to providing legal protection for the development of smart shipping.

Keywords: intelligent vessels ; remote operation center ; remote operators ; legal governance ; legal Attributes

网络犯罪的刑事管辖兼论《联合国打击网络犯罪公约》的管辖协调

欧宝盛 中南财经政法大学法学院

摘要：当下网络发展和数字化程度提高，引发了经典管辖权的积极冲突，但现有的管辖原则并非排斥新型的数字空间。网络空间固有的地域性、执行管辖与规定管辖的限缩与界分和国家主权的强性要求使经典管辖原则使其对网络空间及其活动仍具有支配力和解释力，但既有管辖规则仍面临网络和信息环境在规则硬度和适用扩张上的挑战。《公约》作为首个打击网络犯罪的普遍性公约，全面吸纳了经典管辖原则，但在管辖权协调方面存在规则缺失，亦同样陷入经典管辖原则的适用困境。对既有管辖管辖规则的改构无法真正实现冲突管辖权的分配，管辖规则层面的失灵推使管辖权分配寻求具体确定。通过归纳经典管辖原则的联系要求、解剖经典管辖核心的领土联系、追溯国际私法向联系原则的转变，发现国际私法与国际公法在国家规定管辖权构造上共享一套“联系”理论；并且经过考察网络环境对网络活动与有关因素的联系度的影响，主张《公约》将密切联系作为抽象的实质标准，作为国家协调管辖权冲突的自平衡机制。同时，依据“密切联系”原则的具体化，初步确定有限度的、优先的管辖权，为公约打击犯罪提供确定性、可预见性。

关键词：网络犯罪 《联合国打击网络犯罪公约》 经典管辖原则 联系 密切联系

引言

互联网并不是新兴的事物，经过三十多年的快速发展，人类已进入网络时代。特别是随着信息通讯技术和数字化的发展，互联网的应用生态和技术结构不断发生深刻迭代升级。在此过程中传统的刑事活动出现去现实空间的网络化趋势，网络空间和计算机系统越来越多出现网络犯罪，对国家安全和社会秩序构成重大威胁。由于互联网具有无边界性和虚拟性等特征，频繁的网络刑事活动实现不定向犯罪乃至犯罪的全球化，向国家打击网络犯罪和刑事管辖提出了诸多挑战，最为突出的是网络空间中超越地域与时间限制的犯罪活动导致的国家管辖权的重叠。进入二十一世纪后，为适应网络刑事行为区别于传统犯罪的新特点、新形势，各国学者针对网络犯罪刑事管辖先后提出了新网络主权¹、网络自治理论²、网址来

¹ See BARLOW J.P. A Declaration of the Independence of Cyberspace[EB/OL]. (1996-02-08).

² 参见李本灿. 《联合国打击网络犯罪公约》核心议题述评[J]. 法学论坛, 2025, 40(1): 92-103.

源地管辖理论、有限管辖原则及实害联系原则等新管辖权理论³试图重构或改变国际法上现行的管辖权规则以实现国家管辖的有序化。然而主张给予网络空间以独立地位的主权理论和自治理论实际上早已被国家实践所否定,国家不但重返网络空间而且正尝试建立“网络主权”及有关监管秩序安排。⁴至于试图寻找互联网与现实空间连接的物理层从而为国家提供管辖依据的网址管辖理论、由属地管辖原则衍生的有限管辖也难以为国家实践接受,在网络犯罪国际打击的立法层面上未能得到适用。⁵实害联系原则要求行为结果与行为目标国具有现实联系,强调损害结果的实质发生。从理论思想来看,实害联系原则仍然难离属地管辖原则的范畴支配,只在个别公约中有所体现。⁶总体上,这些理论主张尽管采取了既有管辖原则的改良方法,但未能适应网络时代管辖权冲突的普遍性。

尤为关键的是,在中国、俄罗斯等发展国家的倡议和长期努力下,作为全球首部关于打击网络犯罪的普遍性条约《联合国打击网络犯罪公约》(以下简称《公约》)面世。⁷《公约》在规定打击网络刑事活动的实体范围的同时,还就国家管辖权确定作出了具体分配。⁸然而,《公约》关于管辖权的规定基本上是对现行国际法管辖规则的复写,鉴于现行国际管辖规则在面对灵活、随机的网络犯罪上的不足,《公约》的管辖权部分对于缓解网络环境下潜在的管辖权积极冲突有何积极意义尚值得审视。甚至部分研究已经就《公约》管辖权规定提出管辖协调或改变管辖结构的主张。⁹总的来说,尽管全球性的《公约》已经通过,国家在网络犯罪刑事管辖方面的“竞争”仍将是贯行联合国《公约》、有效遏制网络犯罪的头等问题。为此,有必要认真梳理当下国际法管辖原则在适用于复杂的网络环境时存在的缺陷并重新检视其适用性;在此基础上检讨《公约》在管辖权冲突中发挥的效果,进而因应网络犯罪特点为《公约》管辖权分配提出建设性方案。

³ 参见张婷,陈立. 跨国网络犯罪中刑事管辖权的适用困境及解决[J/OL]. 中南民族大学学报(人文社会科学版), 2023, 43(10): 108-115, 185; 李晓明, 李文吉. 跨国网络犯罪刑事管辖权解析[J/OL]. 苏州大学学报(哲学社会科学版), 2018, 39(1): 89-100.

⁴ 在网络空间建立国家主权实际上是“国际法适用于网络空间”的命题。越来越多的国家发布了国际法适用于网络空间的立场文件。

⁵ 在主要的网络犯罪公约上,这些国际条约仍然遵循了现行国际法关于管辖权的原则。See Budapest Convention on Cybercrime - Cybercrime[EB/OL]. (2001). <https://www.coe.int/en/web/cybercrime/the-budapest-convention>; ARAB LEAGUE. Arab Convention on Combating Technology Offences[EB/OL]. (2010). <https://dig.watch/resource/arab-convention-combating-technology-offences>.

⁶ 在《阿拉伯国家联盟打击信息技术犯罪公约》关于管辖权的规定中,公约规定了基于安全和利益受到损害的管辖权。See ARAB LEAGUE. Arab Convention on Combating Information Technology Offences[EB/OL]. (2010). <https://www.asianlaws.org/gcld/cyberlawdb/GCC/Arab%20Convention%20on%20Combating%20Information%20Technology%20Offences.pdf>.

⁷ See UNITED NATIONS. 联合国大会通过具有里程碑意义的《打击网络犯罪公约》[EB/OL]. (2024-12-24).

⁸ See UN GA. United Nations Convention against Cybercrime: UN Doc.A/RES/79/243[R/OL]. (2024-12-31).

⁹ 参见张婷,陈立. 跨国网络犯罪中刑事管辖权的适用困境及解决[J/OL]. 中南民族大学学报(人文社会科学版), 2023, 43(10): 108-115, 185; 李晓明, 李文吉. 跨国网络犯罪刑事管辖权解析[J/OL]. 苏州大学学报(哲学社会科学版), 2018, 39(1): 89-100; 敬力嘉, 罗贯琛. 《联合国打击网络犯罪公约》的困境与优化进阶[J]. 警学研究, 2024(4): 7-21; 李本灿. 《联合国打击网络犯罪公约》核心议题述评[J]. 法学论坛, 2025, 40(1): 92-103.

一、现有管辖规则在网络刑事管辖中的困境

网络空间以其基础设施和终端设备与现实空间相联系，但信息技术和数据交互将世界连接成为一个虚拟空间，其无边界性和互联性脱离了传统的法律空间支配，对现有国际法经典管辖原则的领土支配和人身控制造成了现实威胁。

（一）现行国际法经典管辖权的僵化

在国际法上，国家对某些主体、事项享有管辖权是构成国家主权的基础，管辖权是实现国家主权的必要和基本方式。依据一般国际法，国家可以基于以下几项原则行使管辖权：属地原则、属人原则、保护原则和普遍性原则，并构成管辖权领域的习惯国际法。¹⁰ 尽管这些一般原则为国家行使管辖权提供了多种依据，但在领土主权理念的影响下，管辖权四大原则在国家实践中享有不同的管辖权地位：国家往往视领土为绝对性地域，同时领土被认为是“主权”的具体承载，因此属地管辖一般被置于最高位阶。在国际法历史上，自常设国际法院结束“*lotus*”案后¹¹，国际社会基本上确立了以属地性或领土性为首的经典管辖原则。在后来为解决跨国犯罪引渡问题的《哈佛犯罪公约草案》（《Draft Convention on Jurisdiction with Respect to Crime-1935》）中，国际法学者再次确认了以属地管辖权和属人管辖权为主的经典管辖权规则。¹² 《哈佛犯罪公约草案》是对这种以领土为核心的管辖权分配范式的不正式确认。¹³ 直到今天，国际法上的管辖原则仍未发生基本变化。但随着互联网和信息通讯技术的不断发展，网络犯罪正在以其巨大的不确定性、偶然性和无边界性冲击着这种以领土性为重的传统管辖权体系。着眼于国家在物理空间内对有关对象的实际控制的经典管辖权体系面对虚拟性、开放性的网络活动陷入了规则僵化的困境。¹⁴

（二）经典管辖权适用于网络刑事管辖的实践困境

属地管辖权、属人管辖权及保护性管辖权构成了国际法经典管辖权体系的主要部分。属地管辖原则以管辖对象与国家领土的实在联系为依据，在国家实践中逐渐发展出主观属地原则和客观属地原则。¹⁵ 属人管辖原则则以人的身份为连接点行使管辖权。在去领土化的网络空间中，属地管辖因其固有的地域性而广泛受到网络活动的潜在影响，而属人管辖则由于现代主体的流动性和组织性扩张其“长臂管辖”。保护性管辖更容易进一步借用信息通讯技术的高度交互性实现宽泛

¹⁰ See UN GA. Report of the International Law Commission - Fifty-eighth session: UN doc.A.61/10[R/OL]. (2006).

¹¹ 常设国际法院在“荷花号案”判决中否定了国家在域外行使权力，由此国家的域外管辖权只保留了立法层面上的权力。See PCIJ. The Case of the S.S. Lotus, France v. Turkey, Judgment, 7 September 1927, Permanent Court of International Justice (PCIJ)[EB/OL]. (1927)

¹² See Draft Convention on Jurisdiction with Respect to Crime[J/OL]. The American Journal of International Law, 1935, 29(1): 439-442.

¹³ See SVANTESSON D J B. A New Jurisprudential Framework for Jurisdiction: Beyond the Harvard Draft[J/OL]. American Journal of International Law, 2015, 109: 69-74.

¹⁴ See SVANTESSON D J B. A New Jurisprudential Framework for Jurisdiction: Beyond the Harvard Draft[J/OL]. American Journal of International Law, 2015, 109: 69-74.

¹⁵ See AMERICAN LAW INSTITUTE. Restatement of the law - the foreign relations law of the United States[M/OL]. Philadelphia, Pennsylvania: American Law Institute, 2018

解释。经典管辖原则既表现了作为确定性规则的硬化，又以其内涵的抽象性广泛地网络主体行为发生接触。

1.属地管辖原则的矛盾

基于属地管辖原则，网络犯罪刑事管辖权的“爆发”。依据主观属地原则和客观属地原则的区分，网络犯罪刑事管辖中的主观属地、客观属地可分别表示为网络犯罪行为（实施）地、网络犯罪结果（发生）地。由于网络空间的无国界性和传播的广泛性，网络行为者的行为及其产生的危害结果将随着数字信息的即时传达和交换而实际或潜在地延展到世界各地。换言之，网络犯罪下的行为地、结果地极易因互联网的信息通讯技术和监管的缺失而大量泛点分布，从而诱发大量的管辖权主张、激烈的管辖权冲突。网络空间的公共性和开放性催生了行为要素和结果要素的扩张，从而形成网络空间所到之处被纳入犯罪关联地的情况。¹⁶ 这就引起了对网络犯罪刑事管辖要件：行为地和结果地资格认定的复杂挑战。

网络物理层传导产生“异型”网络行为。在现实的物理世界中，刑法上的行为总是受到实在时空的严格限制，行为者的行为通过有形传导只能及于因果关系所能到达的地方或程度。而互联网在颠覆刑事行为的表现形态的同时，也衍生了异于物理行为的行为形态。受互联网物理构造的影响，行为转化得到的数字信号需要经由多个物理中介，如服务器、路由器等。而这些基础设施因对数字信号的处理程度差异可能会引致有关国家的管辖要求。相对于网络行为者在现实空间的实行行为，经由网络空间传导的信息数据可以冠称为“投射行为”或“映射行为”。在性质上，不论是经由物理措施导致的投射行为，还是网络行为者在设备终端的实行行为导致的网络结果，都只是互联网促进的“远距离行动”的结果。¹⁷ 诚然，这种投射行为仍然在网络犯罪中作为不同角色和地位。无疑，经典管辖原则可能有必要考虑该种行为形态及其多样化行为类型下的管辖地位。同时，正是由于识别数据处理地存在不确定性和易变性，行为地的管辖地位有所下降，客观属地原则在国家实践中以其结果发生的确定性占据适用优势。¹⁸

结果地的“阴假性”分布。从网络犯罪结果地的遍在分布看，规定网络信息交互的逻辑层决定了网络犯罪行为导致的危害结果将以潜在的形式分布。可见网络在制造广泛的危害结果的同时，“埋藏”了大量的潜在结果地。实际上，在行为危害烈度较低的民事侵权领域，欧洲国家的法院已就类似情形有过判例。¹⁹ 据此，基于网络数据获取的无疆域性和共享性，信息传播、获取、访问地的遍在分布要求对结果地应当采取审慎的严谨态度进行认定。网络传播在形成“访问地”“接受

¹⁶ 参见裴炜. 网络空间刑事司法域外管辖权的数字化转型[J/OL]. 法学杂志, 2022, 43(4): 109-128.

¹⁷ See URBAS G. Cybercrime, Jurisdiction and Extradition: The Extended Reach of Cross-Border Law Enforcement[J]. Journal of Internet Law, 2012, 16(1): 1,8-17.

¹⁸ See SINO-EUROPEAN EXPERT WORKING GROUP ON THE APPLICATION OF INTERNATIONAL LAW IN CYBERSPACE (EWG-IL). Jurisdiction in Cyberspace[EB/OL]. (2024-11-26).

¹⁹ 如法国法院在二十世纪初审理的“雅虎”案，法国法院就要求服务器位于美国的雅虎公司采取 ip 锁定等技术阻止法国境内获取、访问该网站。

地”的同时，可能会逾越司法管辖区域产生外溢的影响，导致部分国家声称行使基于影响的效果原则。不过，目前效果原则仍然集中于经济金融领域，将其适用于网络空间将承受有关国家的质疑。²⁰但是在法理上，效果原则发端于属地原则的领土性联系，还有待观察后续相关的国家实践。

2. 属人管辖原则的流动性困境

属人管辖原则行使管辖权的连接点较为具体、明确，网络活动者的国籍差异和主体确定是导致管辖权积极冲突的重要原因。首先，在网络环境下网络活动的归因存在技术上困难。行为者使用多种技术手段，例如使用VPN、构建darkweb以规避国家网络监管。因此活动者的身份具有隐蔽性，精通信息通讯技术的hacker更是国家网络安全部门的“心患”。其次，具有不同的国籍身份的人也将会导致管辖权重叠。更复杂的是，国家管辖对于网络组织犯罪的无力。有的国家穿透行为体的组织，行使基于控制的属人管辖权。²¹这实质上是对属人管辖原则进行层级扩张。最后，被害人的广泛分布也会导致基于被动属人管辖权的冲突。

此外属人管辖原则还可能由于行为人对属地管辖的规避导致管辖权消极冲突。而实践中属人管辖并不如属地管辖有效。换言之，属地管辖可能因为行为人的主观规避而陷入被动局面。

3. 保护性管辖原则的宽泛适用

保护性管辖原则允许国家以其安全、核心利益行使管辖权。在网络空间下，保护性管辖原则可能会循着宽泛解释的逻辑主张基于网络设施、网络安全和秩序等理由的管辖权。这也与国家晚近在其他领域如关税、出口管制等的实践路径一致。²²鉴于网络犯罪的破坏性和广泛性，保护性管辖原则可能被援引以实现管辖。

总的来说，经典管辖原则在适用于网络空间中的刑事管辖时既存在硬性规则不适应网络行为的灵活性和无边界性的困境，又存在被过度解释、推理而扩大国家管辖范围的可能性。经典管辖原则表现的碎片化，有必要考虑从借以国际私法的灵活标准重新调整其管辖标准。²³

二、既有规定管辖权规则的适用辩证

据上述分析可见，经典管辖原则确立的管辖权具有深厚的物理现实基础，以领土性为首的管辖逻辑强调国家对一切的有效调控。然而，网络时代和数字化时代的到来使网络应用与活动开始试图掘起这一根基。在数字化和互联网世界中，网络犯罪的增长趋势迫使我们有必要重新考虑管辖权的适用。²⁴这种现实状况

²⁰ See SINO-EUROPEAN EXPERT WORKING GROUP ON THE APPLICATION OF INTERNATIONAL LAW IN CYBERSPACE (EWG-IL). Jurisdiction in Cyberspace[EB/OL]. (2024-11-26).

²¹ See SINO-EUROPEAN EXPERT WORKING GROUP ON THE APPLICATION OF INTERNATIONAL LAW IN CYBERSPACE (EWG-IL). Jurisdiction in Cyberspace[EB/OL]. (2024-11-26).

²² 美国政府近来频繁援引1977年的《国际紧急经济权力法》实施单边制裁措施和关税加征。

²³ See SCHULTZ T. Carving up the Internet: Jurisdiction, Legal Orders, and the Private/Public International Law Interface[J/OL]. European Journal of International Law, 2008, 19(4): 799-839.

²⁴ See RYNGAERT C. An Urgent Suggestion to Pour Old Wine into New Bottles: Comment on “A New Jurisprudential Framework for Jurisdiction[J/OL]. American Journal of International Law, 2015, 109: 81-85.

诱导人们向改构或建构新的管辖权规则方向思考,但本文正想提醒的是,网络空间的虚拟性并非彻底。网络空间在物理维度上并不存在,正如人类社会在文明演进与发展中不自觉地围蔽、构造“社会空间”²⁵一样,数字化的互联网仍然倚赖物理空间。由此,反思经典管辖原则的适用性未必等于重新建构管辖权。

正如前文所言,过往提出的所谓新的管辖权理论在国家实践中并没有得到肯定性适用。这些新管辖权理论大致可分为两类,一类是主张网络自由的新主权理论和网络自治理论;另一类包括网络地址管辖理论、有限管辖原则和实害联系原则。这类管辖权理论实质上属于对现行管辖权规则的改进或细化,如因应互联网的物理结构而新增 IP 这一连接点。这表明,我们或许无法脱离既有管辖规则而虚构一套新的管辖权理论。

(一) 规定管辖权与执行管辖权的限缩与界分

从性质而言,跨国网络犯罪的刑事管辖一般属于国家域外管辖问题的范畴。而在国际法上,管辖权类型的区分随之使跨国网络犯罪刑事管辖分为不同层面的问题。传统上,国际法把国家的管辖权分为两大类:规定(立法)管辖权(jurisdiction to prescribe)及执行(执法)管辖权(jurisdiction to enforce)。²⁶ 在规规定管辖权方面,国际法一般允许国家通过法律的制定确立域外管辖权,只要这些域外管辖权符合国际法上的管辖原则即可。与规定管辖权可以将国家管辖权延长至本国领土之外相反,执行管辖权则受到严格的属地性限制。除非基于另一国授权或同意,一国不得在他国领土上行使属于主权的权力。规定管辖权与执行管辖权在权力范围上的结构性差异使得跨国网络犯罪的刑事管辖更为复杂化,因此本文有必要就跨国网络犯罪的规定管辖权和执行管辖权进行区分。

概括而言,一方面,立法管辖权与执法管辖权涵盖的问题域有所区别;另一方面,网络空间活动引起的诸多挑战均可回落至执法管辖层面。一般而言,立法管辖权是指国家对某些事项、行为确立的法律管辖权,而执法管辖权则指国家机关执行、实施并实现法律管辖权的权力。概念的内涵尚可区分,然而,在复杂的法律实践中二者的界限似乎变得模糊。例如,国家行政机关的具体行政行为一般认为属于执法管辖²⁷,但有立法权的行政机关实施的抽象行政行为,即行政立法,则在法理上陷入两面困境。从立法权来看,这种行政立法属于立法管辖²⁸,从法律体系上讲亦具有执法属性。特别是在数字化趋势下,国家管辖权的既有属地性受到极大挑战,进一步导致了网络空间领域与网络活动有关的立法管辖权和执法

²⁵ “社会空间”或“空间的生产”是列斐伏尔提出的重要思想,空间是社会的产物,社会空间是除自然空间和实在空间之外的独立空间。参见亨利·列斐伏尔. 空间的生产[M/OL]. 刘怀玉等,译. 商务印书馆, 2021.

²⁶ 关于管辖权的类型还存在“管辖权三分法”,即依照管辖权的权能表现分为立法管辖权、执法管辖权、司法管辖权。这里本文的执法管辖权采取广义的执法含义,还包括司法管辖权。

²⁷ See SINO-EUROPEAN EXPERT WORKING GROUP ON THE APPLICATION OF INTERNATIONAL LAW IN CYBERSPACE (EWG-IL). Jurisdiction in Cyberspace[EB/OL]. (2024-11-26)

²⁸ See SINO-EUROPEAN EXPERT WORKING GROUP ON THE APPLICATION OF INTERNATIONAL LAW IN CYBERSPACE (EWG-IL). Jurisdiction in Cyberspace[EB/OL]. (2024-11-26)

管辖权之间界限模糊。²⁹ 事实上, 尽管立法管辖权在法理上规定并决定着执法管辖权; 但在法律实施的个案上, 执法管辖亦具有相当的独立意义。换言之, 立法管辖与执法管辖在国家管辖功能分配下有不同的关切和目标。现实情况是, 随着立法管辖权不断地向“写入”执法管辖权, 立法管辖实现对执法管辖的规则“掌控”, 由此发生了立法管辖权对执法管辖权的规则“重述”。这就愈发形成立法管辖与执法管辖的交集。然而, “就立法管辖权而言, 其本质规制的是国家的抽象权力行为”。³⁰ 相对而言, 执法管辖权回应的是具体权力的行为准则和约束。国际机关实施立法管辖权, 是由抽象的规范层面进入直接及于管辖对象的操作层面。³¹ 因此, 若果对立法管辖权进行原旨主义的限缩, 立法管辖权将回归至抽象事项管辖权力的确定与分配, 而执法管辖权则落入实现立法管辖权的具体的权力实践。但需要承认的是, 经过抽象权力限缩的立法管辖权与现实的执法管辖权之间将存在巨大的规则真空。在法律实践中, 立法管辖权往往通过规则“写入”完成填补, “膨胀”的立法管辖权是法律有效运行的必然要求。

据此分析, 在跨国网络犯罪领域的立法管辖权经过国际法抽象层面的权力限缩后应限于网络罪案的管辖权确定, 而不论具体的执法管辖规则。跨国网络犯罪刑事管辖既包括刑事立法管辖, 也包括刑事执法管辖及司法管辖。具体而言就是, 规定管辖权与执行管辖权的二元界分: 立法管辖权确定网络行为的犯罪性质, 执法管辖权确定网络犯罪的具体管辖。³² 因此在规定管辖权与执行管辖权二元界分的视野下, 国家确定对具体网络活动罪案的管辖权与国家如何完成网络犯罪行为追诉是两个独立问题, 而国家确立网络犯罪这一抽象的管辖权力的依据是现行国际法管辖原则。换言之, 现行国际法管辖原则在某种意义上发挥国际法的“立法管辖权”功能。然而, 在打击跨国网络犯罪问题下, 执法管辖的地域性禁忌进一步使跨国网络犯罪的刑事侦查和司法审判问题独立化, 并成为完成全过程刑事管辖的重大法律“障碍”。这同时表明, 执法管辖权的绝对地域性与立法管辖权的国际共享性构成了跨国网络刑事管辖的结构性矛盾。无论国际管辖与国内管辖都属于地域管辖的范畴, 都高度依赖在某个地域内实现。³³ 网络空间特点下的跨国网络犯罪是去领土性、隐蔽性的, 但国家管辖根本上还是地域性的。就此而言, 跨国网络犯罪引起的管辖挑战是无法避免的难题。正如上述分析, 这种管辖挑战包括管辖权分配和执行管辖, 既属于立法管辖层面, 又共属于执行管辖的范畴。也许是鉴于绝对属地的执法管辖权为国家对网络空间的监管、规制带来现实阻碍,

²⁹ See RYNGAERT C. Extraterritorial Enforcement Jurisdiction in Cyberspace: Normative Shifts[J/OL]. German Law Journal, 2023, 24(3): 537-550.

³⁰ 参见裴炜. 论网络犯罪跨境数据取证中的执法管辖权[J]. 比较法研究, 2021(6): 30-45.

³¹ 参见裴炜. 论网络犯罪跨境数据取证中的执法管辖权[J]. 比较法研究, 2021(6): 30-45.

³² See O'KEEFE R. Universal Jurisdiction: Clarifying the Basic Concept[J/OL]. Journal of International Criminal Justice, 2004, 2(3): 735-760.

³³ 参见林洵. 跨境网络侵权诉讼的国际管辖权分配规则研究[J]. 国际法学期刊, 2023(3): 108-132, 156-157.

有的学者在研究中进一步区分规定管辖权与执行管辖权,提出了“调查管辖权”³⁴,并借此厘清数字化带来的规定管辖权与执行管辖权的混淆。

(二) 经典管辖原则对网络活动具有解释力

首先网络空间及其活动的地域性使现有属地管辖原则仍对网络活动具有解释力。如前文所分析,网络空间并非实在,具有所见即一切的特点,高度倚赖现实的物理支持。网络空间的内容层和物理层都是物理世界的物质构成,无法摆脱国家的领土范畴。因此网络空间的基础设施和信息技术都由于固有的领土性而落入国家的管辖范围,国家不但对网络空间的物理构成享有事实控制权,也对网络空间的内容展示有正当的管制权力。现代国家是一个有自然疆域的空间体,同时也是一个意识形态空间和法律空间,而网络空间在信息和技术设施方面的地域性使其成为国家控制下的法律场所。³⁵ 同时,网络空间的地域性为国家进入网络空间、进行内容和活动的监管提供了技术路径。依据网络构建的物理附属性,如 IP 地址、国家域名可能构成界定国家管辖的空间范围的属地连接点。³⁶ 对社会主体而言,网络空间就是呈现的全部。这就是说,需要区分抽象意义的网络空间与具体的网络空间。后者由行为主体的网络活动及其后果所决定。网络空间的开放性、无界性只为行为者的网络投射提供了技术路径,从根本上来讲,法律规制的对象没发生改变——主体的行为。一旦作出网络活动,现实的行为与数字化的网络传播便同时发生了。无疑的是,具体的网络空间的行为及其结果仍然与特定地域保存着地域性关联。在属地管辖原则下,行为的地域性联系足以满足管辖的初始依据要求。

其次,网络活动行为体的身份联系和组织性并未脱离既有属人管辖原则的涵摄。开放的网络环境并未改变网络行为者与国家的国籍关联,只是在身份发现和主体确定等方面陷入实践困境。现实世界人和物的广泛流动,引起了国家属人管辖的不便利。这种管辖困难同样发生在去领土化的网络空间。网络空间的匿名性和去中心化挑起了网络行为主体接受属人管辖的归因“不能”,非国家行为体如私人公司在组织上的复杂化和流动性造成了短臂的属人管辖。即便如此,明确的、稳固的属人联系仍然为国家的管辖提供连接点支撑。因此,属人管辖原则继续通过身份和组织关系的确定保有管辖的合法性。

最后,保护性管辖原则在根本上对网络空间有国家安全、维护文化价值等利益关切,实质上与过往在其他物理领域的适用逻辑一致。网络空间的地域性与属人性表明,网络活动与物理空间保持着密切的现实联系,现行的国际法管辖原则可以在解释的范围内继续适用于网络活动。因此在思考经典管辖原则适用于网络

³⁴ See RYNGAERT C. Extraterritorial Enforcement Jurisdiction in Cyberspace: Normative Shifts[J/OL]. German Law Journal, 2023, 24(3): 537-550.

³⁵ 参见刘艳红. 网络犯罪的刑法解释空间向度研究[J/OL]. 中国法学, 2019(6): 202-223.

³⁶ 参见吴琦. 网络空间中的司法管辖权冲突与解决方案[J]. 西南政法大学学报, 2021, 23(1): 47-64.

犯罪时，由网络空间向物理空间转化的理论思想³⁷应当得到坚持。

（三）国家主权作为国际法管辖权规则的权力基础

管辖权原则一向被视为国家主权的横向调节工具，发挥主权的核心作用。在国际视野下既有管辖原则是国家主权对一切有效控制的规则表达和法律概括，构成了实现国家主权的必然出路。因此，国家主权在地域及对象上的天然支配权决定了现有管辖权规则与国家主权不可剥离。在国家主权原则下，经典管辖原则发展出的连接点，如领土、国籍，与国家紧密联系，代表了主权的实质因素。因此，经典管辖原则蕴含着维护、尊重国家主权并调节它们相互关系的要求。正如全文所言，网络犯罪既可以是国内管辖的事项，又可以是国际管辖的问题。但从国内视角来看，网络犯罪是刑法的，而刑法根本上又是地域的。³⁸ 刑法在现代国家具有较高的法律地位，是实现国家权威、维护公共秩序的基本法律手段。刑法作为规定一国采取刑事惩罚的基本法律，在跨国犯罪问题上兼具实体性和程序性，即当刑法确立规定管辖权时使打击犯罪范围的实体内容与确立域外管辖权的程序问题等同。这表明，国家行使域外管辖权与国家刑法的基本利益深度捆绑。换言之，现行国际法的管辖权规则在很大程度上与国家主权深度捆绑。如果存在与一国有国际法管辖规则上的联系的某一事项确系维护国家利益的必要，那么这一事项的管辖权不仅能得到来自国家主权原则的合法性和正当性支持，还获得国家强制力的实现。跨国网络犯罪问题也莫不如此。就此而言，国家主权构成了施行经典管辖原则的权力基础。《公约》的通过和将来的签署、生效将使《公约》在规定的犯罪范围内替代了国内刑法，但《公约》在管辖权确定与实体内容上采取了体例形式上的分离。

在立法管辖与执法管辖的分离视角下，打击跨国网络犯罪的现实挑战由立法管辖权和执法管辖权共同承受。对立法管辖抽象权力的限缩，同时保留执法管辖绝对的地域性，可以发现经典管辖原则作为立法管辖规则的独立性，并旨在分配这类管辖权：抽象国家管辖权。继而在这一二元视角下，国际法经典管辖原则对网络空间及其活动仍然具有解释力。网络活动的监管必须诉诸某种管辖规则，因此现行国际法的管辖规则具有坚实的适用效力。最后，在国家主权的解释框架下，经典管辖原则作为实现国家主权的基本规则适用于网络空间。简而言之，网络空间并没有根本上改变网络犯罪的管辖权理论基础³⁹，国际法并不需要重新构建网络犯罪的管辖权理论，既有国际法管辖原则仍适用于网络空间。在这一结论基础上，鉴于经典管辖原则在网络犯罪中表现出的规则僵化、管辖权数量“爆发”及管辖权冲突不可避免。由此，经典管辖原则在网络条件的刑事犯罪中不得不表现为

³⁷ 参见孙潇琳. 我国网络犯罪管辖问题研究[J/OL]. 法学评论, 2018, 36(4): 186-196

³⁸ See BRONITT S, GANI M. Shifting boundaries of cybercrime: from computer hacking to cyber-terrorism[J]. Criminal Law Journal, 2003, 27(6): 303-322.

³⁹ See REGERINGSKANSLIET R och. Position Paper on the Application of International Law in Cyberspace[EB/OL]. Regeringen och Regeringskansliet, 2022.

理论的适用性与实践的不适应性的双重张力。在坚持经典管辖理论的根本适用性的前提下，应当从循适当的理论与实践路径改变管辖权冲突现状。因此，网络犯罪域外刑事管辖的核心议题应当是谋求恰当的方法实现管辖权协调。

三、《公约》刑事管辖权体系与管辖权协调设想的检视

《公约》是全球范围内首个针对网络犯罪的普遍性公约，《公约》的通过标志着世界各国共同打击网络犯罪迈出了重要的第一步，在国际法层面为国家法律的域外管辖提供了外部合法性。⁴⁰ 在联合国《公约》之前，国际上已存在较为成熟的《布达佩斯网络犯罪公约》。但该国际协定由发达国家主导，主要反映的是网络技术水平较高的发达国家在网络安全与治理方面的需求，忽略了发展中国家现实的网络国情。且由于《布达佩斯网络犯罪公约》制定年份较早，随着网络现实和信息技术的发展已难以适应当下治理网络犯罪的新要求，因而未能成为广泛接受的全球性公约。⁴¹ 尽管《公约》具有广泛的代表性，但由于发展中国家与发达国家两大阵营在犯罪实体内容方面的分歧较大，导致《公约》在犯罪条款设置上靠近《布达佩斯网络犯罪公约》。⁴² 然而本文无意就《公约》的实体方面展开评述。相比于在实体内容上的不足，《公约》在管辖权安排的程序问题上取得了宝贵的进步。

（一）《公约》管辖权体系评述

《公约》的管辖权安排需要结合其他网络与信息技术犯罪公约比较地看待，同时还应对具体规则予以现实和法理的考察。

1.完整的管辖权规定

从比较角度来看，《公约》采取的管辖权规则具有全面性。欧洲联盟的《布达佩斯网络犯罪公约》（以下简称《布达佩斯公约》）在其程序部分对国家的刑事管辖权作出了规定，共设置了五项条款。概括而言，《布达佩斯公约》在第二十二条第一款确立了基于领土和船舶、航空器的属地管辖权，基于国籍的积极属人管辖权。⁴³ 随后第三款就罪犯所在地国确立管辖权后将罪犯引渡至他国进行了限制，即不允许仅以国籍为由引渡至他国。⁴⁴ 该款的实质目的在于确立罪犯控制

⁴⁰ See REED C. Why judges need jurisprudence in cyberspace[J/OL]. Legal Studies, 2018, 38(2): 263-278.

⁴¹ 参见裴炜，黄志雄，冯俊伟，等。《联合国打击网络犯罪公约》：背景、内容与展望[J]. 数字法治，2024(5): 9-27.

⁴² 参见李哲，朱晓琴。《联合国打击网络犯罪公约》的中国立场与核心问题[J]. 北京师范大学学报（社会科学版），2024(5): 124-133.

⁴³ 《布达佩斯网络犯罪公约》“Article 22 – Jurisdiction”第一款：Each Party shall adopt such legislative and other measures as may be necessary to establish jurisdiction over any offence established in accordance with Articles 2 through 11 of this Convention, when the offence is committed: a in its territory; or b on board a ship flying the flag of that Party; or c on board an aircraft registered under the laws of that Party; or d by one of its nationals, if the offence is punishable under criminal law where it was committed or if the offence is committed outside the territorial jurisdiction of any State. See Budapest Convention on Cybercrime - Cybercrime[EB/OL]. <https://www.coe.int/en/web/cybercrime/the-budapest-convention>.

⁴⁴ 《布达佩斯网络犯罪公约》“Article 22 – Jurisdiction”第三款：Each Party shall adopt such measures as may be necessary to establish jurisdiction over the offences referred to in Article 24, paragraph 1, of this Convention, in cases where an alleged offender is present in its territory and it does not extradite him or her to another Party, solely on the basis of his or her nationality, after a request for extradition. See Budapest Convention on Cybercrime

国对于罪犯国籍国的优先管辖权，在这个意义上，该款具有管辖权协调性质。第四款表明公约确定的管辖权并不排除缔约国国内法的管辖权⁴⁵，该款旨在分割《布达佩斯公约》与国内法的关系及尊重国内法的管辖。第五款规定了多个缔约国就共同罪行声称管辖权时应当协商确定。⁴⁶ 因此该款对于管辖权积极冲突规定了协商确定方式。总体而言，《布达佩斯公约》的管辖权条款主要包括属地与属人管辖权、国籍引渡限制、尊重国内刑法管辖、管辖权冲突协商。

在欧盟主导建立的《布达佩斯公约》之外，一些国家联盟和地区组织在联合国《公约》之前也有就联合打击信息技术犯罪的条约尝试。这些信息技术犯罪条约主要包括阿拉伯国家联盟的《阿拉伯国家联盟打击信息技术犯罪公约》（以下简称《阿拉伯公约》）、非洲联盟的《非洲联盟网络安全和个人数据保护公约》（以下简称《非盟公约》）。《阿拉伯公约》在第三十条“权能”中对管辖权作出了规定。在管辖权资格方面，《阿拉伯公约》参照了《布达佩斯公约》基于属地管辖和属人管辖的管辖权规定，且新增了一项接近保护性管辖的规定。⁴⁷ 与《布达佩斯公约》相同，《阿拉伯公约》也限制罪犯控制国的国籍引渡。⁴⁸ 更为重要的是，作为通过年份较晚的信息网络犯罪公约⁴⁹，《阿拉伯公约》首次确立了具有优先次序的管辖权，即国家安全利益受害国处于第一顺位，犯罪行为实施地所在国处于第二顺位，犯罪嫌疑人国籍国居于第三顺位。⁵⁰ 并在此基础上，《阿拉伯公约》还

- Cybercrime[EB/OL]. <https://www.coe.int/en/web/cybercrime/the-budapest-convention>.

⁴⁵ 《布达佩斯网络犯罪公约》“Article 22 – Jurisdiction”第四款：This Convention does not exclude any criminal jurisdiction exercised by a Party in accordance with its domestic law. See Budapest Convention on Cybercrime - Cybercrime[EB/OL]. <https://www.coe.int/en/web/cybercrime/the-budapest-convention>.

⁴⁶ 《布达佩斯网络犯罪公约》“Article 22 – Jurisdiction”第五款：When more than one Party claims jurisdiction over an alleged offence established in accordance with this Convention, the Parties involved shall, where appropriate, consult with a view to determining the most appropriate jurisdiction for prosecution. See Budapest Convention on Cybercrime - Cybercrime[EB/OL]. <https://www.coe.int/en/web/cybercrime/the-budapest-convention>.

⁴⁷ 《阿拉伯国家联盟打击信息技术犯罪公约》第 30 条“Competence”第一款：Every State Party shall commit itself to adopting the procedures necessary to extend its competence to any of the offences set forth in Chapter II of this Convention, if the offence is committed, partly or totally, or was realized: a- in the territory of the State Party b- on board a ship raising the flag of the State Party. c- on board a plane registered under the law of the State Party. d- by a national of the State Party if the offence is punishable according to the domestic law in the location where it was committed, or if it was committed outside the jurisdiction of any State. e- if the offence affects an overriding interest of the State. See ARAB LEAGUE. Arab Convention on Combating Information Technology Offences[EB/OL]. (2010). <https://www.asianlaws.org/gclid/cyberlawdb/GCC/Arab%20Convention%20on%20Combating%20Information%20Technology%20Offences.pdf>.

⁴⁸ 《阿拉伯国家联盟打击信息技术犯罪公约》第 30 条“Competence”第二款：Every State Party shall commit itself to adopting the procedures necessary to extend the competence covering the offences set forth in Article 31, paragraph 1, of this Convention in the cases in which the alleged offender is present in the territory of that State Party and shall not extradite him to another Party according to his nationality following the extradition request. See ARAB LEAGUE. Arab Convention on Combating Information Technology Offences[EB/OL]. (2010). <https://www.asianlaws.org/gclid/cyberlawdb/GCC/Arab%20Convention%20on%20Combating%20Information%20Technology%20Offences.pdf>.

⁴⁹ 公开信息显示，《阿拉伯国家联盟打击信息技术犯罪公约》于 2010 年通过。详见：ARAB LEAGUE. Arab Convention on Combating Technology Offences[EB/OL]. <https://dig.watch/resource/arab-convention-combating-technology-offences>.

⁵⁰ 《阿拉伯国家联盟打击信息技术犯罪公约》第 30 条“Competence”第三款：If more than one State Party claim to have jurisdiction over an offence set forth in this Convention, priority shall be accorded to the request of the State whose security or interests were disrupted by the offence, followed by the State in whose territory the offence was committed, and then by the State of which the wanted person is a national. 参见 See ARAB LEAGUE.

规定了多个国家处于同一管辖权层级如两个国家同属受害国时，以最先正式提出引渡请求的国家为优先。⁵¹ 这一条款构成了管辖权优先层级的协调。概括而言，《阿拉伯公约》的管辖权安排包含了属地与属人的基本管辖权、国籍引渡限制和管辖权协调。《非盟公约》作为规制网络通讯技术、维护网络安全的重要国际协定，是治理跨国网络犯罪国际条约的重要组成。然而，遗憾的是《非盟公约》没有就管辖权作出规定，而是选择把确立管辖权的问题交回缔约国由其国内法确定。

52

联合国《公约》在其第二十二条就管辖权规定了包括上游犯罪管辖权确定、下游洗钱罪管辖权确定、关于引渡的管辖权、管辖权协商及允许国内法管辖的内容。具体而言，在管辖权确定方面，《公约》确认了国家基于犯罪行为发生地、犯罪受害国民、犯罪嫌疑人人身联系（国籍或惯常居所）及作为犯罪目标或具有保护性利益享有管辖权。⁵³ 对于下游的洗钱罪，《公约》也确立了洗钱目的地国的管辖权。这些管辖权反映在国际法上的经典管辖原则，包括了主观属地原则、积极国籍原则、被动国籍原则及客观属地原则或保护性管辖原则。⁵⁴ 在引渡方面，《公约》区分了被指控犯罪人的本国国籍与外国国籍，但《公约》都要求被指控犯罪人控制国遵行“或引渡或起诉”原则，实际上给予了缔约国选择性管辖权。在管辖权协调、《公约》的国际管辖与缔约国的国内管辖方面，《公约》与《布达佩斯公约》有着相似规定——由缔约国磋商、尊重缔约国国内法确立的管辖。

从上文可见，《公约》采用了完整的属地管辖原则和属人管辖原则。甚至在

Arab Convention on Combating Information Technology Offences[EB/OL]. (2010). <https://www.asianlaws.org/gclid/cyberlawdb/GCC/Arab%20Convention%20on%20Combating%20Information%20Technology%20Offences.pdf>.

⁵¹ 《阿拉伯国家联盟打击信息技术犯罪公约》第 30 条“Competence”第三款：In case of similar circumstances, priority shall be accorded to the first State that requests the extradition. 参见 See ARAB LEAGUE. Arab Convention on Combating Information Technology Offences[EB/OL]. (2010). <https://www.asianlaws.org/gclid/cyberlawdb/GCC/Arab%20Convention%20on%20Combating%20Information%20Technology%20Offences.pdf>.

⁵² 参见《非盟公约》在“Article 31 – Adapting certain sanctions to information and communication technologies”涉及刑事处罚的第一款及第二款。See AFRICAN UNION. African Union Convention on Cyber Security and Personal Data Protection[EB/OL]. (2014-06-27). https://africanlii.org/akn/aa-au/act/convention/2014/cyber-security-and-personal-data-protection/eng@2014-06-27#chp_IV_art_32.

⁵³ 《联合国打击网络犯罪公约》第 22 条有关款项：一、各缔约国均应在下列情况下采取必要措施，对根据本公约确立的犯罪确立管辖权：

- (一) 犯罪发生在该缔约国领土内；或者
- (二) 犯罪发生在犯罪时悬挂该缔约国国旗的船只上或已根据该缔约国法律注册的航空器内。
- 二、在不违反本公约第五条规定的情况下，缔约国还可在下列情况下对任何此类犯罪确立其管辖权：
 - (一) 犯罪系针对该缔约国的国民；或者
 - (二) 犯罪系由该缔约国国民或在该缔约国领土内有惯常居所的无国籍人实施；或者
 - (三) 犯罪系发生在该缔约国领土以外，且属于根据本公约第十七条第一款第(二)项第 2 目所确立的犯罪之一，其目的是在该缔约国领土内实施根据本公约第十七条第一款第(一)项第 1 目或第 2 目或第(二)项第 1 目确立的犯罪；（犯罪发生在外国，但意图在国内实施掩饰或隐瞒犯罪所得的犯罪）或
 - (四) 犯罪系针对该缔约国。See UN GA. United Nations Convention against Cybercrime: UN Doc.A/RES/79/243[R/OL]. (2024-12-31).

⁵⁴ 实际上，《公约》在文本上的表述为“犯罪系针对该缔约国”，这可能导致理解的分歧。一方面根据客观属地原则，犯罪所针对的国家是犯罪目标国，在一般预期下，该国将成为犯罪结果发生地。另一方面，鉴于网络犯罪的危害性、破坏性之大，犯罪结果可能损害不特定国家的重要安全利益。

属人管辖原则上，将个人的属人联系扩展至无国籍人的惯常居所。相较于《布达佩斯公约》和《阿拉伯公约》限于传统的行为地管辖权和国籍管辖权，《公约》采取的管辖原则更为全面，覆盖了大部分管辖利益。这种对国际法经典管辖原则的全面吸纳，不仅重申了既有管辖权规则在网络空间的适用性，也有利于限制国家在国内利用立法管辖权不当地改写管辖权规则从而扩张国家的跨国信息网络刑事管辖。

2.具体规则的考究

如前文所总结，联合国《公约》在确定缔约国管辖权问题上具有全面性，其采用的管辖权类型与先前的打击网络犯罪公约并无根本差异。⁵⁵ 接前文结论，《公约》体现的属地管辖原则和积极属人原则，其在网络环境下的适用性更无需再赘论。但亦无可避免，由于网络活动的遍在性和网络空间的去领土化特点，属地与属人的经典管辖方法将引致管辖权数量的增多和混乱。在属人管辖问题上，《公约》罕见地引入国际私法的“惯常居所”概念用作管辖连接点。一般而言，在既有管辖原则中，“惯常居所”具有较大的流变性，并不适宜应用于属地性严格的国内公法或国际公法的问题上。且这一概念时常夹杂民商事关系的平等性，关注的是维护民商事行为的稳定性与预见性，与刑事管辖所具备的严厉规制性相去甚远。当然，《公约》采取了无国籍的范围限缩。赋予无国籍人惯常居所地所在国以管辖权，填补了积极属人管辖原则对无国籍行为者的适用漏洞，行为《公约》的重要创新。还需特别注意的是《公约》吸纳了“被动国籍原则”。这一属人管辖原则仍然依据有关主体的国籍身份，但一直以来受到较大争议。⁵⁶ 无疑，在网络犯罪背景下，《公约》对这一管辖原则的适用将刺激受害国国民据此行使管辖权。另外，尽管受害者与相关国家存在国籍联系，但是不当的管辖主张可能涉嫌侵犯他国主权。因此“被动国籍原则”的适用还需待实践检验。《公约》还首次就上游网络犯罪的下游洗钱罪规定了目的地的管辖权，是对打击网络犯罪的重要完善和补充，但可能产生与上游犯罪有关管辖权协调和司法衔接等问题，未来《公约》可能需要对此作出细化规定。关于《公约》“犯罪系针对该缔约国”确定的管辖权问题，根据客观属地原则和保护管辖原则的解释，这一规定具有不确定性。无论《公约》指向客观属地原则抑或保护管辖原则的意图为何，该规定的用词“针对”将导致管辖权的不确定性和开放性。在这一规定之上，客观属地原则与保护管辖原则如同杠杆两端的滑动砝码，为缔约国创造了巨大的解释空间。该规定似乎勤勉地关注到了网络犯罪行为对各国造成的影响，但忽视了为合理、恰当确定管辖地应具有谦抑性和明确性。最后《公约》还在本国国民引渡问题上作出了重要

⁵⁵ 参见李本灿. 《联合国打击网络犯罪公约》核心议题述评[J]. 法学论坛, 2025, 40(1): 92-103.

⁵⁶ See Theory of Extraterritoriality of States and Jurisdiction In International law | Legal Service India - Law Articles - Legal Resources[EB/OL]. <http://www.legalserviceindia.com/legal/article-3648-theory-of-extraterritoriality-of-states-and-jurisdiction-in-international-law.html>.

变革，不再坚持过往“本国国民不予引渡”的国家实践。对于本国国民同等对待，要求犯罪嫌疑人国籍国执行“或引渡或起诉”有利于推进网络罪案终结并缓解管辖权冲突。在未来引渡实践中，在引渡问题上还可能出现被控诉的犯罪者所在地国与多个缔约国的引渡选择问题。如何使这一问题避免政治化及遵循公义和正当方法解决，是《公约》未来值得思考的问题。

（二）协调管辖权冲突的《公约》规则与理论设想

有效的管辖权分配是打击网络犯罪、维护世界人们网络权益的基本前提。《公约》采取了近乎全面的管辖权规则，在网络和通讯技术的开放背景下，多种管辖权规则相竞合将成必然。由此，应当检视《公约》在管辖权竞争上的规定效果。同时，基于《公约》管辖权规则与经典管辖原则在冲突问题的同质性，也有必要检讨学界近来提出的网络空间中有关刑事犯罪管辖权积极冲突的理论方法。

1. 《公约》在管辖权协调上的“隐身”

根据前文分析，《公约》仅规定了缔约国遇其他缔约国就同一行为诉诸实际的刑事管辖时得相互磋商、协调行动。因此面对管辖权冲突，《公约》实际上只明确规定了“协商”方式。同时，《公约》也未就协商的程序、结果作进一步规定。模糊的规定极可能使协商方式沦为国家完全自主裁量的程序形式，并缺乏实际的约束力，对网络空间犯罪引起的管辖权增加和重叠无法发挥有效的协调、支配作用。事关，“协商”一词本无规则的确定性，决定了缔约国自由意志的主导地位。然而还需注意的是，除了《公约》明示规定的管辖协商，有关引渡的管辖规定也能潜在地、在一定程度上梳理“拥挤”的管辖权要求。如上文所言，《公约》要求罪犯所在地国“或引渡或起诉”，实际上迫使罪犯所在地国将管辖权交由其他主张罪犯管辖的国家或保留在本国。在管辖权效果上，引渡管辖的规定对罪犯控制国与其他管辖权主张国进行了先后的区分。就此而言，可以合理地期望管辖权“要么交由外国要么保留本国”的操作向导在多个国家中实现管辖权分配。然而，更为复杂的情形将使这一明确的操作规则难以充当管辖权协调规则。例如，当犯罪嫌疑人分布于多个国家时，仍可能导致管辖权的积极冲突；当罪犯所在地国与其他犯罪管辖主张国就犯罪构成存在异议时，即发生管辖权的消极冲突，“或引渡或起诉”未必能够妥善解决。实际上，尽管被指控犯罪人身处何国的地理标准足够明确具体，但也具有一定的流动性和偶然性。一贯适用该标准，可能导致国家间的猜疑及网络罪案实际管辖的不合理性。据此可见，这种潜在的引渡管辖规则过于简单，难以满足分配网络犯罪管辖权的复杂要求。综而言之，《公约》在协调管辖权冲突方面存在两个“一明一暗”的规则。“协商”规则浮于程序，不足以为解决管辖权冲突提供规定硬度；“或引渡或管辖”则偏于简化，容易造成实践中的僵化局面。由此可见，《公约》在管辖权调控方面存在规则缺失，不能不称之为《公约》一大遗憾。

2. 管辖权规则改构或有序化

接上文关于“经典管辖原则在网络空间中的适用性”分析的基础上,可以确定的是,既有管辖权规则导致的管辖权冲突也将现实地发生在《公约》的实践中。实现管辖权的有效调配成为经典管辖原则适用于网络犯罪的基本问题。管辖权冲突的解决必然意味着寻求某种能够支配绝大部分国家的分配规则。换言之,当下国际法需要诉诸有关管辖权冲突规则,这也是《公约》的缺失。然而,目前国家实践未有冲突规则的直接指示⁵⁷,国际法上亦未提出一项规则解决合法而冲突的管辖权。⁵⁸

当下,越来越多的学者支持既有管辖规则仍然适用于网络环境,也意识到了网络空间犯罪活动的刑事管辖需要某种程度的管辖权层级分配规则,并提出了多种管辖权规则的改构模式或次序安排方案,下文就此进行类型化分析。

(1) 属地管辖原则的刑事化改构

这类管辖权模式主张由明确的犯罪目标国享有优先管辖权,或在没有明确的犯罪指向国时根据“实际受害国”确定优先管辖权;对于处于同一管辖权顺位的国家,则由最先控制犯罪人国实行“或起诉或引渡”原则。⁵⁹ 首先,这里的“犯罪目标国”借鉴了现实物理的犯罪构成理论中的主观犯意理论。同时,根据客观属地原则,如行为意图或实际在某一国家完成,该国便可行使管辖权。⁶⁰ 显然,这种主观的目标指向也是对客观属地原则的提炼。这种目的性原则在网络空间条件下具有确定性和限缩作用,有利于一定程度减少主张管辖权的国家数量。其次,这类管辖权模式引入了实害联系原则。由于网络犯罪产生的影响可能超越既定的犯罪目标国之外,并对其他国家构成潜在的危险,采取这一原则也有利于限制有资格的管辖权主张。但如前文分析,实害联系原则在本质上仍是属地管辖原则的具体化,某种程度上为限制管辖权数量适用于网络犯罪中是不言自明的。

综合而言,目标国、受害国体现的目的原则与实害原则都是对属地管辖原则的改写,实质上维持的是犯罪活动与特定国家不同程度的领土联系。尽管这一结构模式对限制并确立管辖权的优先性具有积极意义,但还可能重蹈经典属地管辖原则的既有困境。

(2) 经典管辖原则的位阶确认

较多的研究都表明,应当关注国家对属地管辖具有的天然利益和正义性,国民及其组织受到国家的正当支配居于属地性之下;根据既往的国家做法,有限的国家核心利益被限制于特定领域而处于辅助地位;普遍管辖原则由于涉及全人类

⁵⁷ See KRISCH N. Jurisdiction Unbound: (Extra)territorial Regulation as Global Governance[J/OL]. European Journal of International Law, 2022, 33(2): 481-514.

⁵⁸ See RYNGAERT C. The concept of jurisdiction in international law[M/OL]//ALEXANDER ORAKHELASHVILI. Research Handbook on Jurisdiction and Immunities in International Law. Northampton, MA, USA: Edward Elgar Publishing, 2015: 50-7.

⁵⁹ 参见李晓明,李文吉. 跨国网络犯罪刑事管辖权解析[J/OL]. 苏州大学学报(哲学社会科学版), 2018, 39(1): 89-100.

⁶⁰ 参见甘勇. 《塔林手册 2.0 版》网络活动国际管辖权规则评析[J/OL]. 武大国际法评论, 2019, 3(4): 117-135.

共同利益只应适用于特定情势。据此，以属地管辖为优（或以实害原则作为属地管辖原则的限制或体现），属人管辖次之，保护管辖为辅助，普遍管辖为补充的管辖权分配方案被提出了。⁶¹ 这一管辖权体例实质上是对既有管辖原则的重新确认，事实上领土性或属地性在国家实践中向来作为国家的首要考虑，既有管辖原则本就具有强烈的属地性。在这一位阶安排下，属地管辖权居于首位，尽管实害联系原则的适用一定程度限制了有资格的属地管辖权，但亦只存在单一的限制规则。这一规则涉嫌对涉及国家主权利益进行了位次的排序，实践情形中或导致个别国家的抗议。因此，这一位阶安排看似为管辖权冲突构建了有优先级别的分配规则，实际上仍难以回避同类型及不同类型管辖权冲突的难题。可以进一步认为，对经典管辖原则的“改编”或排序呈现的是理论和规则的状态，在复杂多样的网络犯罪形态前，采取机械、硬性的认定标准不是最有利的。

（3）单一管辖模式

上述分析表明，在多种管辖原则内周旋仍无可避免管辖权的激烈冲突。领土管辖和属人管辖是对国家将其控制和管治能力施于不特定人员、财产等的反映，它为国家提供了多样的域外管辖依据，但在网络空间条件下造成了管辖竞争的低效。鉴于互联网必然的管辖权广泛分布，有的观点主张网络犯罪的刑事管辖应选择单一的主观属地管辖模式。⁶² 根据主观属地原则，将只有网络犯罪行为实施地所在国享有管辖权。选择单一的主观属地原则为现行多样的管辖权进行了“瘦身”，从技术上而言只保留了一种管辖依据，由此改变管辖权的遍在分布和竞争的现状。然而，这一方法过于机械和简单，极大地忽略了其他有合理利益的关联国的管辖诉求，此外也不能协调行为地国家之间的管辖权对立。

从以上管辖权规则的改构和次序化安排来看，继续建基于经典管辖原则的各种管辖权分配方案仍然无法适应网络空间背景下的刑事管辖冲突。实际上，正如上文所分析，网络空间的开放性、去领土化和即时性等特性正在放大以领土性为首的经典管辖原则的规则硬度和解释边界。因此，停留于经典管辖原则的规则方法或将无法有效改善管辖冲突，有高度必要进一步深入至规则层掩盖下的被支配对象：具体的管辖权，亦即从规则层到具体层的转变。

四、因应《公约》管辖权协调的自平衡机制和规则构建

据上分析，经典管辖原则的抽象性和概括性无力应对公共性和非物质性的网络活动。网络犯罪现实促使我们考虑转向更为灵活合理的规则标准。有关最新研究指出，“国家可以根据‘利益平衡’分析来约束自身管辖权”，或者“有必要类似国

⁶¹ 目前而言，我国大部分研究大体上均参用这一管辖权协调体例。参见刘艳红. 网络犯罪的刑法解释空间向度研究[J/OL]. 中国法学, 2019(6): 202-223; 张婷, 陈立. 跨国网络犯罪中刑事管辖权的适用困境及解决[J/OL]. 中南民族大学学报(人文社会科学版), 2023, 43(10): 108-115, 185; 李晓明, 李文吉. 跨国网络犯罪刑事管辖权解析[J/OL]. 苏州大学学报(哲学社会科学版), 2018, 39(1): 89-100.

⁶² See ARNELL P, FATUROTU B. The prosecution of cybercrime – why transnational and extraterritorial jurisdiction should be resisted[J/OL]. Journal of Crime and Justice, 2022, 42(5).

际私法一样建立管辖权分配的一般性规则”。⁶³ 尽管利益考量高度依赖缔约国的自行判断,但至少表明了在该案中管辖权的具体化确定对管辖权协调具有积极效果。循着这一思路,更为实证和明确的“联系”原则/标准可以进一步启发《公约》的管辖权分配方法。

（一）“密切联系”原则作为协调管辖权冲突的自平衡机制

在网络和信息通讯环境中,网络活动按其烈度从行政监管到刑事规罚,无不充满国家权力的冲突。借用理性和双边的方法,如国际私法中的平衡方法——冲突规范与管辖构造会有所裨益。⁶⁴ 联系原则是国际私法上用以构造国内法院管辖跨国民商事关系的理论基础,与国际法施于国家的合法管辖原则在法理上同构。因此《公约》应当考虑借鉴国际民事管辖理论对联系要素的适用,并因应互联网和通讯技术的虚拟性确证“密切联系”原则,赋予缔约国在网络犯罪刑事管辖权确定上有限度的自平衡权力。

1.由经典管辖向“联系”原则的理论转向

“联系”原则代表作为权力施加者的国家与权力指向对象,如法律主体、事项等的实质性关联。因此,国际私法的管辖理论通常要求国家法院与案件存在真实性联系。⁶⁵ “联系”原则也代表了被调整对象接受特定司法管辖的合理性要求和国家司法管辖的权力正当性。但在国际法上的管辖原则中,“联系”要素并没有被列为一项独立的管辖权依据。实际上,随着人类社会逐步迈入发达工业阶段,复杂的现代生活和社会流动使人们难在领土范围内获得足够的结果。⁶⁶ 德国国际法学者曼恩在上世纪六十年代就对传统的领土原则适用于现代生活表示怀疑,转向关注“联系”,并主张基于“联系”的国家监管。⁶⁷ 近年来更有国际法学者对以领土为核心的管辖原则提出了范式转变,确立以“实质性联系”“利益平衡”为核心的管辖权原则。⁶⁸ 这些主张表明,随着社会生活的复杂化和数字化程度加深,经典管辖原则在授予国家管辖权中的合理性不断发生偏差,仅凭属地与属人联系也无法胜任管辖权确定任务。事实上在经济监管领域中,国家在出口管制和市场监管等具体领域正在行使越来越频繁和广泛的域外管辖权,法律的属地性也不再为国家与

⁶³ See SINO-EUROPEAN EXPERT WORKING GROUP ON THE APPLICATION OF INTERNATIONAL LAW IN CYBERSPACE (EWG-IL). Jurisdiction in Cyberspace[EB/OL]. (2024-11-26).

⁶⁴ See GINSBURG T. Comment on “A New Jurisprudential Framework for Jurisdiction”[J/OL]. American Journal of International Law, 2015, 109: 86-88.

⁶⁵ See AMERICAN LAW INSTITUTE. Restatement of the law - the foreign relations law of the United States[M/OL]. Philadelphia, Pennsylvania: American Law Institute, 2018

⁶⁶ See A. MANN F. The Doctrine of Jurisdiction in International Law (Volume 111)[M/OL]//Collected Courses of the Hague Academy of International Law. Brill, 1964.

⁶⁷ See A. MANN F. The Doctrine of Jurisdiction in International Law (Volume 111)[M/OL]//Collected Courses of the Hague Academy of International Law. Brill, 1964.

⁶⁸ 具体而言,该学者主张了国家行使管辖权的三项原则:(1) 该事项与寻求行使管辖权的国家之间存在实质性联系;(2) 寻求行使管辖权的国家在该事项中具有合法利益;(3) 考虑到国家合法利益和其他利益之间的平衡,行使管辖权是合理的。See SVANTESSON D J B. A New Jurisprudential Framework for Jurisdiction: Beyond the Harvard Draft[J/OL]. American Journal of International Law, 2015, 109: 69-74.

私主体间的关系提供确定的答案了。⁶⁹“联系”原则的法理基础就在于国家主权在被规整客体中的具体化，随着联系要素的增减，联系程度也随之成为调节国家间关系的指向针。因此，应当期待“联系”原则在指导国家行使规定管辖权过程中渐进地承担实质的规则功能。

2.“联系”原则在管辖权中的法理逻辑

“联系”原则不仅直接适用于国际法上的管辖权原则，也构成国家实现域外利益的法理基础。在“联系”要素上，国际私法和国际公法有不同的关注点，前者偏重个案中的具体化认定和私利保护，而国际公法与国内公法受限于地域禁忌和主权独立，但一致基于联系的管辖基础构造。在行为烈度较刑事犯罪较低的跨国侵权问题上，“合理联系”因素被引入结果地管辖权。⁷⁰

（1）经典管辖原则中的联系要素

“联系”原则既具有抽象内涵，亦可实证地表现为连接因素，从而具有可控制性、明确性并继而量化。在经典管辖原则中，每一原则都以具体的管辖基础授权国家行使管辖权，这些管辖基础构成了联系原则中的具体要素，并被认为证明了管辖事项与管辖国之间存在充分的联系。⁷¹ 具体而言，属人管辖原则以个人身份（如国籍、住所的居民）为“联系”，其联系要素包括当事人的国籍、住所或居所，这些联系要素又称连接点。源于国家对其国民或居民的控制权。属地管辖原则以可追踪的地理位置为“联系”，其联系要素主要表现为案件中的行为、事件或财产发生或存在于管辖国领土之内，例如行为发生地、结果发生地等。根据属地管辖原则的结果形态，国家实践发展出了“效果原则”。效果原则体现为事件的重大影响，是一种管辖范围的领土延伸，它与属地管辖原则一样基于这种领土性联系允许受影响国行使管辖权。效果原则的出现反映了人们对“联系”内涵认识的进一步发展。保护原则以国家的基本安全和核心利益为“联系”，实际上是对联系原则的最明确表达，⁷² 直接地揭示了“联系”所体现的实质关联的抽象性。至于普遍管辖原则，也表明了所涵盖的事项是与所有国家有着根本和密切联系的事项，这些事项事关国际社会的根本利益。⁷³

可见，透过经典管辖原则的管辖依据可以发现隐含的联系因素。换言之，国际法上的管辖原则在根本上建基于被调整对象与国家存在的“联系”，根据“联系”的实证化从而接受为国家共同承认的法律实践。当现行的经典管辖原则被国家援引时，实际越过了内在的联系原则而直接适用了其代理规则，如地域、行为、身

⁶⁹ See GINSBURG T. Comment on “A New Jurisprudential Framework for Jurisdiction”[J/OL]. American Journal of International Law, 2015, 109: 86-88.

⁷⁰ 参见张文亮. 跨国侵权结果地的认定和管辖权行使[J/OL]. 法商研究, 2023, 40(3): 131-143.

⁷¹ See O’KEEFE R. Universal Jurisdiction: Clarifying the Basic Concept[J/OL]. Journal of International Criminal Justice, 2004, 2(3): 735-760.

⁷² See O’KEEFE R. Universal Jurisdiction: Clarifying the Basic Concept[J/OL]. Journal of International Criminal Justice, 2004, 2(3): 735-760.

⁷³ See UN GA. Report of the International Law Commission - Fifty-eighth session: UN doc.A.61/10[R/OL]. (2006).

份等联系。⁷⁴ 《美国对外关系法重述（第三版）》在管辖权原则部分对经典管辖原则和实质性联系的条序分布一定程度上反映了这种逻辑关系，表明了对国家行使管辖权具有的联系的关注。⁷⁵

（2）联系作为领土“扩张”的秘密

领土构成国家的基本生存空间，自人类社会建立国家以后，边界和“领土”便产生了。进入现代社会后，国家领土基本趋于稳定。然而，与此相反的是国家越来越多地将国家权力延伸至领土外的他国人和物。除了国际民商事管辖中外国法院判决的承认与执行外，便是国家积极地行使域外管辖权。⁷⁶ 领土应避免片面地理解为国家的绝对限制和绝对权力，它是一个政治和历史的空間，它是国家权力的实在场域。由此，我们可以理解：现代国家管辖制度的扩张正是谋求控制更为广阔的权力空间。⁷⁷ 正如国际法上承认国家在其领土内的积极管辖权，但并未向国家要求行使消极的域外管辖权。国家在域外权力的延伸尽管受制于主权原则的领土性，但通过对本国主体的控制、利用国家经济实力等联系因素，国家的权力仍然延及于领土外。一方面本质上，在现代化和全球化中各国间的“联系”加深、多样化和复杂化不可避免；这种联系。另一方面，国家主体之间的密切联系为国家域外管辖制度的扩张奠定了实在基础，构成国家域外立法管辖的基本出发点。在多边国际秩序中，“联系”的抽象性意味着“联系”的跨国性。它不受国家领土作为边界的物理隔离，也不受国家领土作为概念的精神建构。联系超越了领土限制，将各国的法律制度、事实行为连接起来，为域外管辖权的扩张提供了应用和理论解释的进路。

（3）国际私法共享管辖权的联系理论

根据前文所述，国际法上的经典管辖原则来源于对事物联系的关注、发现和确认。由经典管辖原则向联系原则纵深转向，是适应现代国际社会管辖现状的必然要求。历史上，这一转变同样发生在国际私法的民事管辖中。美国冲突法重述在 1970 年代经历了由传统的机械的司法管辖向“政策”“利益”和合理性的灵活司法管辖转变。⁷⁸ 更早地，美国法院经历“国际鞋业公司”案后迈入了民事管辖快速扩张阶段，依靠“最低限度联系”长臂管辖外国主体。⁷⁹ 另外，除国际民事管辖中一般管辖体现了案件与管辖国存在的稳定和具体的联系外，各国允许法院行使的

⁷⁴ See RYNGAERT C. An Urgent Suggestion to Pour Old Wine into New Bottles: Comment on “A New Jurisprudential Framework for Jurisdiction”[J/OL]. American Journal of International Law, 2015, 109: 81-85.

⁷⁵ 在《美国对外关系法重述（第三版）》的管辖权部分，第 402 条规定了属地原则等现行国际法管辖原则，第 403 条合理性原则规定了考虑行为与国家领土的联系程度。

⁷⁶ 在域外管辖权的属事范围上，存在是否包括国际民商事司法管辖的不同看法。这里本文认为域外管辖并不包括平等主体之间的国际民事管辖。

⁷⁷ See KRISCH N. Jurisdiction Unbound: (Extra)territorial Regulation as Global Governance[J/OL]. European Journal of International Law, 2022, 33(2): 481-514.

⁷⁸ See WATT H M. A Private (International) Law Perspective Comment on “A New Jurisprudential Framework for Jurisdiction”[J/OL]. American Journal of International Law, 2015, 109: 75-80.

⁷⁹ See International Shoe Co. v. State of Wash: 326 U. S. 310[A/OL]. (1945-12-03).

特别管辖尤为关注法院地国和民商事争议之间的特别联系。⁸⁰ 我国《民事诉讼法》的涉外编在最新修订中也采用了“适当联系”原则⁸¹，构成了我国国际民事管辖的特别管辖权。联系原则以其对当事人利益的关注和国家行使管辖权力的合理性限制已成为国际私法调控国家司法管辖与民事主体的基本方法。传统上，在国际领域中对私人利益的高度关注构成国际私法的典型特征，将联系原则“纳入国际公法的管辖权合理性分析”表明了管辖权项上国际公法与国际私法的趋同。⁸² 在这个意义上，真实和实质性联系的要求构成了将国际公法和国际私法联合起来的共同纽带。⁸³ 由此，国际法上的立法管辖规则——经典管辖原则与国家在国际私法的民事管辖规则在法理上共享一套联系理论。

3.“密切联系”的联系原则相对化

经验上，正义和合理的法律原则都依赖于人的主观运用。网络犯罪行为与根据经典管辖原则确定的管辖国之间存在真实和实质的联系是联系原则的基本要求。事实上，在国际机制中这一联系要求的裁量权通常由国家自拥。而联系原则具有抽象内涵，且历史上更多由国内法院在司法实践中承担这一角色。换言之，有关国家判断联系的质量是联系原则适用的关键因素。⁸⁴ 在网络条件下，联系原则适用于网络犯罪刑事管辖权的优先分配仍然面临挑战。

首先，网络信息技术对传统现实联系的弱化。计算机设备等终端凭借信息技术的交互功能独立地开辟了新的虚拟空间，网络主体所作行为不再强烈地在物理规律下发生及于对象的效果，而是指向不被现实空间容纳的网络界面。例如，某一互联网访问者在国地存取色情数据，转而在 B 国利用获取的色情数据制造成不同形式的淫秽作品，后辗转 C 国通过侵入服务系统获取有关用户信息进行色情散播。在这一过程中，该名访问者仅在网络设备上实施信息技术行为，与人身所在国丧失了传统的现实关联。可见，在网络虚拟环境中法院地与当事人的行为——后果的传统关联被削弱了⁸⁵，现实空间仿若沦为容纳行为的场域。在传统法律适用理论中，行为地属地性联系往往构成法律规则有效适用的根本条件；一旦网络法律行为的行为后果脱离了领土限制，法律管辖的联系基础便被抹杀。然而，这一认识只流返于传统的联系。事实上这种传统联系是完全领土性、属地性的，人们的认知也容易碍于物理空间的有形体现。实际上“联系”是多样的、无形态的，也可以仅表现抽象独立的行为。如本文分析，经典管辖原则对网络活动仍

⁸⁰ 参见肖永平，“长臂管辖权”的法理分析与对策研究[J/OL]. 中国法学, 2019(6): 39-65.

⁸¹ 我国《民事诉讼法》第二百七十六条：除前款规定外，涉外民事纠纷与中华人民共和国存在其他适当联系的，可以由人民法院管辖。《中华人民共和国民事诉讼法》[A/OL]. (2023-09-01).

⁸² See RYNGAERT C. An Urgent Suggestion to Pour Old Wine into New Bottles: Comment on “A New Jurisprudential Framework for Jurisdiction”[J/OL]. American Journal of International Law, 2015, 109: 81-85.

⁸³ See SVANTESSON D J B. A New Jurisprudential Framework for Jurisdiction: Beyond the Harvard Draft[J/OL]. American Journal of International Law, 2015, 109: 69-74.

⁸⁴ See RYNGAERT C. An Urgent Suggestion to Pour Old Wine into New Bottles: Comment on “A New Jurisprudential Framework for Jurisdiction”[J/OL]. American Journal of International Law, 2015, 109: 81-85.

⁸⁵ 参见湖北省高院民事庭. 涉外民事诉讼管辖权问题研究[M/OL]. 1 版. 武汉大学出版社, 2008.

具有根本的解释向度，国家依据属地原则及其内在的主权保护权利对网络法律行为具有管辖利益。因此，这种属地联系并未被“抹掉”，而是被弱化了。

其次，网络活动对联系的泛化。如网络空间使网络行为的属地联系被弱化的逻辑一致，网络介质的分布性广的技术特点使网络行为和数字化结果可能留下广布的“痕迹”。这些技术连接如服务器位置，可能被指与国家存在领土联系，从而主张具有管辖权。⁸⁶某些技术应用的网络化，如云计算技术，甚至“使地域位置成为一种偶然现象”。⁸⁷ 这些网络在线活动在物理上与多个国家的领土发生现实接触（信息交互与传输），“而与这些国家没有任何真正的实质性联系”。⁸⁸ 却去这一类型的网络活动，个人和私人组织，尤其是具有全球影响力的跨国互联网公司和具有主导地位的服务处理者的网络活动仍可对不特定国家产生程度不等的影响。⁸⁹ 就前一种网络活动而言，根据国际法上的经典管辖原则，这些网络在线活动关联国家均可以行使属地管辖权。然而根据联系原则，纯粹的联系将劣后于受到网络活动的主要影响。但对于后一种更为主要的网络处理行为，在联系原则下，许多受到影响的国家都可以主张管辖权。⁹⁰ 可见，网络技术构造决定了网络活动的遍布性，同时使活动与多个国家发生关联。

综合地看，网络技术对经典管辖原则之下的联系原则也形成了现实威胁，应当对联系原则在网络犯罪管辖中的适用进行相对化限缩。回归联系原则的基本方法有助于寻求适用于网络犯罪的恰当方法。联系原则以对象与国家之间的联系为依据行使管辖权，当联系因素在多个国家分布的集中程度和这些要素与关联国家的紧密程度发生变化时，联系程度随之变化。这一理论方法也体现在国际私法中确定准据法的“最密切联系原则”，可以进一步类推至网络环境下的刑事犯罪。但与国际民事管辖中的最密切联系国不同，刑事管辖附有强烈的主权意义。如果在网络刑事打击上采用唯一性管辖标准，涉嫌存在管辖权的威权主义，并导致某些国家的管辖霸权。此外，国际民事管辖和刑事管辖存在相异的目标追求和法理逻辑。因此，在最密切联系原则之下和一般的联系原则之上，“密切联系”原则可作为《公约》的犯罪管辖权冲突规则。另外，鉴于前文分析，经典管辖原则在规则层的“不能”应当进一步下沉至具体层，并考虑到网络信息技术对联系的弱化与泛化，应当对密切联系作为具体层考虑管辖权优先性的核心标准。同时，避免联系的实证化停留在事实因素，对密切联系应当在抽象层面进行把握以适应复杂的联

⁸⁶ See RYNGAERT C. The concept of jurisdiction in international law[M/OL]//ALEXANDER ORAKHELASHVILI. Research Handbook on Jurisdiction and Immunities in International Law. Northampton, MA, USA: Edward Elgar Publishing, 2015: 50-75

⁸⁷ See RYNGAERT C. Extraterritorial Enforcement Jurisdiction in Cyberspace: Normative Shifts[J/OL]. German Law Journal, 2023, 24(3): 537-550.

⁸⁸ See SVANTESSON D J B. A New Jurisprudential Framework for Jurisdiction: Beyond the Harvard Draft[J/OL]. American Journal of International Law, 2015, 109: 69-74.

⁸⁹ See RYNGAERT C. An Urgent Suggestion to Pour Old Wine into New Bottles: Comment on “A New Jurisprudential Framework for Jurisdiction[J/OL]. American Journal of International Law, 2015, 109: 81-85.

⁹⁰ See MEETALI RAWAT. Transnational Cybercrime: Issue of Jurisdiction[J/OL]. International Journal of Law Management & Humanities, 2021, Volume 4(Issue 2): Page 253-266.

系现实。

密切联系原则包含了“足够和密切的实质联系”的基本含义。在这一原则指导下，经典管辖原则上的适格国家或基于案件方便原则、国际礼让原则等合理性规则禅让或以协商方式主张优先管辖权，密切联系的规则限度和规则共识有利于促进管辖权的有序行使。

（二）密切联系原则适用于管辖权分配的具体化

密切联系原则本质上倚赖裁决者的主观善用和智用，赋予了国家有限度的自平衡权力。从规则建构层面来看，单一的“比例性”原则⁹¹似乎并不足够。网络犯罪活动广泛的“数字痕迹”、实害后果辐射之广、影响地的潜在分布及流动的属人性将使现实情况更为混乱。因此，还应对密切联系原则进行部分规则细化，以确立比较情形下的优先管辖权或适格管辖权，进而便利密切联系原则在有限管辖权范围内确定优先性管辖权、提高分配效果。

1. 有限的适格管辖权

在以领土性为重的经典管辖原则中，属地管辖原则是网络犯罪中最为主要的管辖权依据。由于网络犯罪所造成的结果形态异于现实的物理犯罪，且网络空间具有的广泛接触地点，因此应当对属地管辖原则的结果要件进行限缩以确立了某些情形下有限的管辖权。

具体而言，《公约》应因应实体部分的具体犯罪对属地原则的主张作出结果要求。第一，对于抽象危险的行为犯确定主观属地原则。第二，对于具体危险犯和实害犯要求结果的实质发生，以客观属地原则为限。第三，对于预备的犯罪亦应采取限制立场，限于属地管辖原则。另外，关于“结果地”的认定，鉴于网络信息内容广泛分布并处于可获取状态的现实，对于某些依赖信息接受主体获知的犯罪类型可以借鉴公众知晓地、网络内容实际被访问地。⁹²

除了限缩属地原则的“结果”要件外，还可以“时间”限制国家声称的管辖权，即国家据称的管辖联系必须与网络犯罪行为发生时同时存在。⁹³ 例如无国籍人进行的网络犯罪仅与该无国籍人的惯常居所地所在国存在属人联系，犯罪行为实施后而取得别国国籍，该国不能主张对该罪犯的属人管辖权。

2. 优先的管辖权

在网络空间特性下，网络犯罪结果可能呈现大范围分布，导致属地管辖权的遍在冲突。即使经过上述适格管辖权的限制后，仍面临剩余同类属地管辖权的竞争。应当对密切联系原则进行方法论上的细化，期待在属地管辖中通过有预见性

⁹¹ 密切联系原则在规则形式上采取了管辖权与联系基础的平衡结构，“公法中的一种普遍反应是比例性（原则）”。See RYNGAERT C. An Urgent Suggestion to Pour Old Wine into New Bottles: Comment on “A New Jurisprudential Framework for Jurisdiction[J/OL]. American Journal of International Law, 2015, 109: 81-85.

⁹² 参见张文亮. 跨国侵权结果地的认定和管辖权行使[J/OL]. 法商研究, 2023, 40(3): 131-143.

⁹³ See O’KEEFE R. Universal Jurisdiction: Clarifying the Basic Concept[J/OL]. Journal of International Criminal Justice, 2004, 2(3): 735-760.

的规则分配管辖权。

在客观属地管辖中,联系要素的体现集中为结果要素与行为与结果要素的关联性。由此,对优先的客观属地管辖权应综合衡量“结果地”结果要素的集中程度及犯罪行为和犯罪结果之间的关联性。⁹⁴ 前者表现为网络犯罪行为的对法益的危险和损害程度,后者应通过“目标测试”原则予以考量。“在过于宽泛和过于狭窄的管辖范围之间”,必须找到一条“中间道路”——目标原则以管理网络犯罪的跨境外部性。⁹⁵ 目标测试原则表明了行为人与行为目标国之间更直接的真实联系,根据该原则,犯罪意图的真实指向国与犯罪活动具有更高管辖利益。对网络活动者而言,可以合理地预见被管辖国,一定程度体现了人权保护。⁹⁶ 通过目标测试,可以考察网络活动与结果地国的联系程度,从而确定客观属地原则的优先性。

至于属人管辖权、保护管辖权或主管属地管辖权与客观属地管辖权的优先层级。首先,属地原则体现的领土性联系强弱情况并不是恒定的,应避免在属地管辖、属人管辖及保护管辖中作出事前的机械安排。其次,在密切联系原则下,联系程度的紧密性是抽象标准,应由国家以协商方式具体化确定,以避免《公约》在规则层的硬化导致管辖权持续冲突。

结 语

《公约》在管辖权体例上基本参照了现行经典管辖原则,一方面具有广泛的国家意志基础,但另一方面未能够因应网络犯罪特点对管辖权安排作出限制或优先分配,某种程度上是对国家管辖利益的文本宣示。经典管辖原则属于规则层的权力安排,性质上作为国际法上的“立法管辖权”规则。但在更为灵活的网络活动刑事管辖中,为实现有效的国家资源分配和和谐国际关系,《公约》应当尝试在应用的具体层中发力,类比国际私法上的平衡和冲突思维确定更合理的管辖国家。未来,密切联系原则应当在管辖权冲突协调中发挥理论指导与具体规则适用的双重功能。

⁹⁴ 参见刘艳红. 网络犯罪的刑法解释空间向度研究[J/OL]. 中国法学, 2019(6): 202-223.

⁹⁵ See SCHULTZ T. Carving up the Internet: Jurisdiction, Legal Orders, and the Private/Public International Law Interface[J/OL]. European Journal of International Law, 2008, 19(4): 799-839.

⁹⁶ See RYNGAERT C. The concept of jurisdiction in international law[M/OL]//ALEXANDER ORAKHELASHVILI. Research Handbook on Jurisdiction and Immunities in International Law. Northampton, MA, USA: Edward Elgar Publishing, 2015: 50-75

人工智能自动武器的国际法回应

陈佳欣*

摘要：人工智能自动武器系统（AWS）的自主决策能力突破传统战争“人在回路”模式，引发责任归属模糊、国际人道法核心原则的技术性消解，以及军备竞赛与国际规则滞后的合规矛盾。现行国际法框架因人类中心主义预设与静态规则设计，难以应对算法黑箱、行为不可预测性等特性，暴露出《特定常规武器公约》（CCW）的扩展性局限、国际人道法弹性解释的边界模糊，以及国际人权法个案化评估能力的缺失，需通过技术治理嵌入法律框架、责任机制革新与军控范式转型进行突破，构建动态适配的全球协同的国际法治理体系：首要深化“有意义的人类控制”（MHC）原则进行适应性改革，短期应依托 CCW 强化合规审查机制，中期推动“自主武器管制议定书”，长期设立全球 AI 军控机构，整合中国“人类命运共同体”伦理框架与国际法院咨询意见的法理指引，以规避 AWS 沦为“国际法真空地带的杀戮工具”。

关键词：人工智能自动武器；有意义的人类控制（MHC）；特定常规武器公约（CCW）；国际人道法；人工智能国际治理

一、人工智能自动武器的技术特点与法律风险

人工智能自动武器的技术革新虽为军事行动带来效率革命，但其自主性与智能化的双重特性——尤其是突破“人在回路”的决策模式——正在颠覆传统国际法的规制逻辑。其引发法律风险绝非单纯的技术副作用，而是其技术基因与国际法体系的结构性冲突的必然产物。

（一）人工智能自动武器的定义、技术特点及应用现状

1. 人工智能自动武器的定义与技术特点

AI 自动武器（Autonomous Weapons Systems, AWS）是指通过人工智能算法和传感器实现自主感知环境、目标识别与跟踪、决策和执行攻击任务的武器系统，无需人工直接干预。这些系统能够独立完成作战任务，具备实时数据分析、快速反应和精确打击的能力，适用于多种作战场景，包括侦察、监视、精确打击和后勤支持等。

在诸多特征之中，人工智能自动武器区别于传统武器的根本之处在于“智能”和“自主”，是一种“能思考”的武器系统。¹ AWS 的核心特征——自主性，例如 2020 年利比亚冲突中，土耳其的 STM Kargu-2 无人机在没有人为监督的情况下自主

*中国海洋大学法学院硕士研究生。

¹ 参见侯嘉斌、李军：《人工智能武器：法律风险与规制路径》，载《中国信息安全》2019 年第 12 期，第 90 页。

攻击人类目标，成为首个 AI 杀戮案例，标志着 AWS 首次突破“人在回路”限制。

2.当前应用现状与案例

AI 自动武器早已不再停留于概念层面，而是逐渐转化为真实的战争实践。大量不同类型的 AI 自动武器已被广泛应用于现代战争中，包括无人机、无人地面车辆（UGVs）、无人水面舰艇（USVs）、智能导弹和机器人战斗车辆等，这些系统在战场上减少了对人类士兵的依赖，提高了作战效率和精确度。如美国海军的无人水面舰艇 Sea Hunter、以色列的“铁穹”系统和美国海军舰艇上的“宙斯盾”系统。美国国防部计划到 2035 年初步建成智能化作战体系，到 2050 年实现作战平台、信息系统、指挥控制的全面智能化。在中国，早在 2017 年就利用人工智能成功实现 119 架小型无人机飞行试验，名为“海翼 1000”的水下无人滑翔机也已投入使用。

随着技术的不断进步，AI 自动武器也正向模块化、轻量化和智能化方向发展，以适应不同作战需求。例如，智能瞄准系统和实时数据分析技术的应用显著提升了武器系统的灵活性和精确度。同时，通过机器学习和深度学习算法，这些系统能够不断优化性能，适应复杂环境和动态威胁。

（二）人工智能自动武器引发的主要法律风险

AI 自动武器代表了现代军事技术的重大突破，但其发展也伴随着巨大的伦理和安全挑战，如责任归属、武器失控风险以及国际人道法冲突等问题。

1.责任归属的双重困境

当 AI 自动武器系统在军事行动中犯错或造成伤害时，责任归属难以明确。例如，AWS 的完全自主性可能导致无法完全区分平民和非战斗人员，如果无人机面部识别系统错误地识别目标，导致误杀平民，责任应由制造商、程序员、操作员还是 AI 系统本身承担？²

这种责任归属模糊性引发了法律争议，争议根本原因在于 AWS 的“算法黑箱”特性导致了责任链断裂。³例如：以色列国防军使用“Habsora”AI 目标系统，能够通过分析无人机拍摄画面、社交媒体帖子、地面特工获取的信息、手机位置数据等多种信息，以识别和定位哈马斯武装人员。然而 AI 在处理复杂战场环境时可能存在误判，且大量数据的输入并不意味着目标识别的绝对准确性，系统学习过程中的错误累积可能导致不准确的袭击目标，从而导致加沙地带平民死亡人数持续增加。

在此情况下，国际法上的责任归属机制面临着双重困境：在个人刑事责任层面，AI 系统的自主性切断了《罗马规约》第 28 条要求的“有效指挥与控制”链条，

² See Klein, M. “We are on the brink of a third revolution in warfare, therefore we should forbid AWS” *AI & Society*, January 27, 2021. <https://ai-society.michelklein.nl/2021/we-are-on-the-brink-of-a-third-revolution-in-warfare-therefore-we-should-forbid-aws/>.

³ See Tiwari, G. “Artificial Intelligence and International Humanitarian Law” *A Contrario ICL*, June 11, 2018. <https://acontrarioicl.com/2018/06/11/artificial-intelligence-and-international-humanitarian-law/>.

AI 系统也无法承担个人刑事责任，由此导致了国际刑事法院（ICC）追责困境；在国家责任层面，自主武器引发的附带损害可能因缺乏“有效控制”证据而规避《国家责任条款》的追责。

2. 武器失控和合规难题

AWS 的自主决策过程缺乏透明度和可解释性，增加了法律和道德上的复杂性，特别是在实际操作中可能无法遵守国际人道法的要求，造成失控局面。⁴

首先，算法透明性要求与技术黑箱存在矛盾。AWS 的复杂算法和机器学习模型通常被视为“黑箱”，其内部运作机制难以被外部观察者或操作人员完全理解。这种不透明性使得在实际部署前难以进行全面的合规性评估，增加了其在战场环境中违反国际人道法的风险。

其次，战场环境的动态性与预先合规评估之间存在冲突。战场环境瞬息万变，AWS 需要在高度不确定和复杂的条件下做出快速决策。传统武器审查依赖静态测试，如《第一附加议定书》第 36 条通常假设武器在设计和测试阶段可以预见其所有可能的行为，这在 AWS 具有实时环境适应能力的情况下显然难以实现，使得预先合规评估失效。

同时，技术迭代下的军备竞赛与法律滞后性形成恶性循环。AWS 的开发可能引发大国之间的武器竞赛，导致军备竞赛加剧，从而催化了 AI 自动武器的应用演化。美国“复制者计划”拟在 2025 年前部署千架 AI 无人机，刺激中俄加速对应武器研发。AWS 技术迭代速度与法律修订周期的脱节，形成“监管时差黑洞”。如此不仅增加了 AWS 失控的风险，也使得国际社会在应对这些风险时面临更大的挑战，进一步威胁国际和平与安全。

3. 国际人道法冲突

2015 年，红十字国际委员会起草的《国际人道法和当前武装冲突中的挑战》研究报告中，专门分析了自主武器所带来的法律挑战，认为人工智能武器的运用是否符合国际人道法取决于武器类型、所执行的任务、适用环节、意图攻击目标等多方面因素。AWS 的自主性使其在选择目标和攻击时无需人类干预，这引发了对国际人道法中“比例原则”、“区分原则”和“预防性原则”等核心原则的挑战：

第一，区分原则的失效。AWS 复杂战场中可能无法区分平民与战斗员，可能导致对平民的伤害，这直接违反国际人道法的区分原则。例如，加沙冲突中以以色列使用的 AI 目标识别系统“Habsora”因数据偏差导致平民设施误判。联合国武器审查案例表明，现有 AI 系统在混杂场景（如城市战）中的误判率远超 IHL 容忍阈值（通常要求<5%）。⁵

⁴ See Ariel Conn. “How Can We Make Sure Autonomous Weapons Are Used Responsibly?” *IEEE Spectrum*, December 7, 2022. <https://spectrum.ieee.org/autonomous-weapons-trust>.

⁵ See Aulia Putri Yunanda. *Autonomous Weapon Systems dan Legalitas Penggunaannya Dalam Hukum Humaniter Internasional* *Jurist-Diction*: Vol. 2 No. 2, Maret 2019.

第二，比例原则的失衡。AWS 的自主攻击模块（如实时损伤评估系统）可能通过强化学习动态调整打击强度，但算法难以量化评估“军事利益”与“附带损害”的平衡（《第一附加议定书》第 51 条第 5 款）。传统武器系统在复杂环境中本就具有性能偏差，而 AWS 因依赖动态学习算法，其行为变异率可能成倍增加。

第三，预防性原则的技术悖论。AWS 的“黑箱”决策机制（如深度神经网络的不可解释性）与《第一附加议定书》第 57 条要求的“采取一切可行预防措施”直接冲突。更严峻的是，《日内瓦公约》第 36 条要求武器需通过“可预测性”审查，但 AWS 的动态学习和调整交战规则（如实时损伤评估模块）使其行为超越一般预设规则，从而难以满足这一要求。

最后，AWS 更是直接挑战了现行国际人道法中的马尔顿条款（《第一附加议定书》第 1 条第 2 款），可能突破“人道原则与公众良知”的伦理底线。⁶⁷

综上，AI 武器的自主决策能力突破了传统武器“人在回路”的控制模式。目标识别算法的误差率（尤其在复杂战场环境下）、深度学习系统的不可解释性，以及自主武器集群的涌现行为，均可能导致违反国际法规则和核心原则，引发一系列的国际法律风险。

二、现有国际法对人工智能自动武器的适用性分析

现有国际法针对非自主武器时代的传统武器制定，难以适配 AI 特性。《日内瓦公约》第 36 条要求武器具备“可预测性”，但 AWS 的动态算法调整能力（如实时损伤评估模块）使其行为超出预设规则。同时，国家责任受制于《国家责任条款》第 8 条“有效控制”标准，难以追究算法自主行为后果。《特定常规武器公约》（CCW）虽具扩展性，但其议定书修订进程缓慢，尚未形成对 AWS 的专门约束。国际法院指出，现行法律在归责时面临“心理要件缺失”难题——AI 无法具备犯罪意图，导致战争罪追责陷入真空。具体阐述如下：

（一）传统武器的国际法规制框架及核心原则

国际法对传统武器的规制框架及核心原则主要体现在以下几个方面：

一，国际人道法（IHL）基本原则。区分原则（区分战斗员与平民）、比例原则（攻击收益与附带损害平衡）、预防原则（攻击前验证目标合法性），规范了传统武器的使用。这些原则旨在防止不必要的伤害和无差别伤害，并确保武器不会被滥用或非法转移。

二，《特定常规武器公约》（CCW）。框架性公约，通过议定书扩展规制范围。该公约是国际社会对常规武器进行规制的重要法律框架，其核心内容包括禁止使用无法区分目标的武器、限制可能造成过分伤害或不必要痛苦的武器等。公约通过五个议定书进一步细化了具体限制措施，如禁止地雷、燃烧武器和盲眼激光武

⁶ 参见李强，谢丹：《人工智能武器的国际人道法规制：中国学者的视角》，载微信公众号“红十字国际委员会”，2019 年 6 月 6 日。

⁷ 王韵涵：《自主武器系统对人权与人道的挑战及法律应对》，载《人权》2023 年第 3 期，第 145 页。

器等。

三，《武器贸易条约》(ATT)。该条约旨在规范武器的国际贸易，防止武器非法流入市场并被用于恐怖主义等非法目的。其核心原则包括透明度、负责任的转让以及禁止向违反国际法的实体转让武器

四，国际公约与习惯法的结合。国际法中关于武器控制的规定既包括条约法，也涉及习惯国际法。例如，《日内瓦公约》及其附加议定书、《海牙公约》等均对传统武器的使用提出了明确限制。

五，禁止或限制某些传统武器的国际条约。如《禁止或限制使用燃烧武器议定书》、《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》等，这些条约通过具体条款限制了某些传统武器的使用，以减少战争中的人道主义灾难。

国际法通过一系列条约和习惯法，构建了一个多层次的规制框架，旨在限制武器的使用，保护平民和战斗人员的安全，并防止武器被滥用或非法转移。

(二) 现行国际法律规则的适用可能性

但传统法律能否覆盖 AI 武器？上述对于传统武器的国际法规制框架，是否适用于自主武器系统这一新兴技术的发展？核心争议在于国际法的“技术中立性”与“动态适应性”之间的张力。现行国际法规则对 AI 自动武器系统的适用性分析如下：

1. 《特定常规武器公约》(CCW) 的扩展性

自 2014 年《特定常规武器公约》(CCW) 框架下成立“致命性自主武器系统 (LAWS) 政府专家组”以来，国际社会围绕 AWS 的规制路径展开多轮博弈。截至 2023 年，该机制已召开九次正式会议，但实质性进展有限。各国立场分化明显：禁止派以奥地利、巴西、巴基斯坦为代表的 30 余国主张通过“禁止自主武器公约”，认为 AWS 的自主性必然违反国际人道法核心原则。规制派如美国、俄罗斯、以色列等国强调“现有法律足够”，主张通过技术标准补充而非禁止。美国提出“有意义的人类控制 (MHC)”原则，要求人类保留攻击决策的最终否决权，但未明确如响应时间阈值或地理部署范围等操作标准。⁸

尽管 2023 年 CCW 审查会议通过《负责任人类控制原则声明》，但其缺乏强制约束力，且未解决算法透明性验证、行为动态监测等关键问题，无法有效解决 AWS 的合规性问题，特别是无法覆盖对闭源系统的监管。例如，以色列“Habsora”系统因闭源特性逃避审查，其目标识别模型的数据偏差率未被纳入公约监管范围。

同时，以 CCW 为代表的现有军控机制基于“物质特性分类标准”（如毒性、爆炸当量）规制武器，而 AWS 的威胁源于算法行为模式的不可预测性。例如，

⁸ 参见梁正，宋雨鑫：《人工智能国际治理规则的发展》，载微信公众号“清华大学人工智能国际治理研究院”，2025 年 3 月 21 日。

CCW 第五号议定书对“激光致盲武器”的禁止仅针对其物理损伤机制，但 AWS 的伦理风险来自算法黑箱导致的决策不可逆性。2023 年联合国裁军研究所（UNIDIR）报告指出，传统武器分类体系无法覆盖“动态学习型 AWS”——这类系统在部署后可通过实时数据更新改变攻击逻辑，形成“后验不可控性”。⁹对此，2024 年联合国《全球数字契约》尝试将“算法武器化”（algorithmic weaponization）纳入军控范畴，要求成员国对 AWS 的机器学习模型实施“可解释性审查”（explainability audit）。然而，美国、俄罗斯等国以“技术中立性”原则反对该条款，认为算法本身无善恶属性，监管应聚焦于“使用意图”而非技术载体。总而言之，各国立场的分歧和自身标准的局限，导致了 CCW 对于人工智能自动武器的扩展规制十分有限。

2. 国际人道法（IHL）的弹性空间

国际人道法（IHL）通过日内瓦四公约及其附加议定书确立了战争行为的底线规则，但其能否适配 AWS 的自主性仍然有待考证：现行国际人道法的“马尔顿条款”（《第一附加议定书》第 1 条第 2 款）规定“缺乏具体规则时，冲突方仍需遵循人道原则与公众良知”。部分学者主张将其扩展解释为“禁止超出人类伦理底线的自主攻击”，例如禁止 AWS 使用面部识别实施定点清除。¹⁰然而，条款的模糊性可能导致部分国家以“符合军事必要性”为由实施自主攻击，凸显弹性解释的局限。同时，马尔顿条款虽为新技术规制提供弹性空间，但算法决策的动态性使得传统武器审查机制难以实施有效的事前合规评估。

另外，红十字国际委员会通过的《新武器、作战手段和方法法律 审查指南》虽然也为各国确立法律审查机制的实体和程序方面提供了框架性建议，但该《指南》仅是红十字国际委员会提供的一个建议性文件，对各缔约国如何审查新武器、作战手段和方法不具有国际法约束力。¹¹

以上，国际人道法的弹性解释空间虽带来了对 AWS 的规制可能，但也为相关责任主体提供了抗辩、出罪甚至不受约束的广阔出口。

3. 国际人权法（IHRL）的延伸适用

国际人权法（IHRL）通过“国家人权保护义务”对 AWS 形成间接约束，但其适用边界面临技术特性与法律原则的深层冲突。具体表现在以下三方面：

第一，比例原则的技术性消解。根据《公民权利与政治权利国际公约》（ICCPR）第 6 条，生命权保护要求武力使用必须符合严格的比例原则，即“军事利益”与“附带损害”需实现个案化平衡。然而，AWS 的算法设计往往以“军事效

⁹ See Evhen Tsybulenko, A. Kajander. “Customary International Humanitarian Law and Article 36 of Additional Protocol I to the Geneva Conventions: A Stopgap Regulator of Autonomous Weapons Systems?” *TalTech Journal of European Studies*(2022).

¹⁰ See Andi Faradilla Ayu Lestari. “AWS: A Threat to International Humanitarian Law or a Necessary Technological Evolution?” *Modern Diplomacy*, April 4, 2024. <https://moderndiplomacy.eu/2024/04/04/aws-a-threat-to-international-humanitarian-law-or-a-necessary-technological-evolution/>.

¹¹ 李寿平：《自主武器系统国际法律规制的完善和发展》，载《法学评论》2021 年第 1 期，第 170 页。

能最大化”为优化目标并进行自主决策，容易导致比例性框架被突破。以美军 Project Maven 系统为例，其强化学习模型在追求打击效率的同时，可能会忽视对平民伤亡概率的动态评估。这种算法偏向性直接违反 IHRL 对生命权的“最小伤害”要求，暴露出机器决策与比例原则的根本性矛盾。

第二，人案化评估能力缺失。HRL 要求武力使用必须基于“个案化必要性评估”（individualized necessity assessment），即针对具体情境中的威胁性质、对象身份及环境特征进行独立判断。然而，AWS 的算法逻辑往往基于群体数据特征进行概率判断，缺乏对具体场景中人权的动态响应能力。¹²例如，以色列军方通过 Habsora 系统识别和打击目标，导致了一定比例的误判案例，造成了平民伤亡。此类算法和模型将复杂人权场景简化为数据标签，其黑箱特性使攻击决策无法追溯至具体法律要件，形成“算法暴政下的权利剥夺”。¹³

第三，系统性人权风险的不可控性。AWS 的规模化部署可能引发系统性人权危机，例如，在埃塞俄比亚提格雷冲突中，政府军使用自主无人机，造成了平民伤亡，由于算法无法精准区分武装人员与逃难平民。此类系统性风险暴露出现行人权法的规制盲区——当国家以“系统自主决策”为借口规避责任时，ICCPR 第 2 条第 3 款规定的“有效救济权”形同虚设。再如，反恐行动中，AI 可能误将平民集会识别为“敌对行为模式”并自主开火，这直接违反了 ICCPR 第 6 条的禁止任意剥夺生命权义务。

（三）目前人工智能国际治理的结构性缺陷

现行国际法规则的适用局限性，根本在于其法律框架建立在人类中心主义预设之上，未能回应“机器代理人”的法律地位问题。而针对“机器代理人”武器的立法回应，全球国家却存在着分歧。这暴露了目前人工智能国际治理的结构性缺陷。

1.非人类主体的责任规制真空

当前的国际法律框架，包括《日内瓦公约》和《海牙公约》，主要针对传统战争中的武器使用规则，而 AI 自动武器系统（AWS）的特性（如自主决策、实时评估附带损害等）使其难以完全符合现有法律要求。

最直接的表现是，AI 无法满足《罗马规约》第 30 条“心理要件”，导致战争罪追责出现真空；而国家责任受制于《国家责任条款》第 8 条“有效控制”标准，难以追究算法自主行为后果。国际法院 2024 年咨询意见也明确指出，现有法律无法解决“非人类主体故意性”难题。当 AWS 因数据偏见误伤平民时，开发者可援引《罗马规约》第 32 条“事实错误”抗辩，指挥官则以“算法自主性”脱责，造成“责任蒸发”。

¹² See M. Brehm, “Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapon Systems Under International Humanitarian and Human Rights Law” *Social Science Research Network*(2017).

¹³ See Brehm, M. “Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapon Systems Under International Humanitarian and Human Rights Law” *Social Science Research Network*(2017).

2.全球治理阵营的分裂态势

联合国秘书长古特雷斯呼吁在 2026 年前禁止无人监督的致命自主武器系统，并强调 AWS 的使用必须符合国际人道法和人权法。多个国际组织和学者也提出了限制或禁止 AWS 的倡议。例如，SYRIZA-Progressive Alliance 提出禁止超出人类可控范围的进攻性自动武器，并建立问责机制。此外，联合国人权观察组织和哈佛大学国际人权法研究中心也呼吁制定相关国际法规，限制 AWS 的研发、管理和使用。¹⁴ 2024 年联合国大会以 161 票通过《人工智能安全倡议》，呼吁建立“人类最终控制权”的全球标准。

由此可见，国际社会对于保持人类对技术控制权具有共识。然而，这些倡议和决议往往缺乏强制约束力，这意味着它们的执行效果依赖于成员国的自觉遵守和国际合作。美国、俄罗斯、以色列等国坚持“技术优先”立场，认为现有国际法足以应对 AWS 风险。

美国在 2024 年联合国决议中推动“灵活治理”框架，强调技术创新与军事优势的结合，其“复制者计划”加速部署千架 AI 无人机，刺激全球军备竞赛。欧盟虽支持监管，但其《人工智能法案》侧重商业伦理，对军事 AI 采取模糊立场，暴露其战略自主性与国家军备诉求的矛盾。¹⁵

三、人工智能自动武器国际法治理体系的构建方案

现有国际法虽可通过弹性条款部分覆盖 AWS 风险，但其静态性、人类中心主义预设与技术的动态演化存在结构性矛盾。未来需通过技术治理嵌入法律框架、责任机制革新与军控范式转型进行突破突破，构建动态适配的治理体系。从国际法原则自身的适应性改革、人工智能武器的全球治理路径、中国发挥的关键作用和司法机构的协作力量，分别阐述如下：

（一）国际法适应性改革：深化“有意义的人类控制”（MHC）原则

2012 年，美国国防部颁布了 3000.09 指令，首次在国别层面提出自主武器系统的全生命周期必须得到人类控制并充分反映人类指挥官、操作员的意图。2014 年，红十字国际委员会向全球传达就 LAWS 的关键功能实施 MHC 的理念，形成了广泛共识。同年，联合国《特定常规武器公约》正式开始将 LAWS 遵循 MHC 的伦理议题纳入框架议程。¹⁶要使 MHC 原则在实践中能真正有效应用，必须落实“可读性”标准 and 责任原则。

1.设置人权法的“可读性”

IHRL 隐含“决策可追溯性义务”，应明确“有意义的人类控制”（MHC）原则

¹⁴ 杨通进：《寻求人工智能伦理的国际共识——以欧盟人工智能伦理建设为例》，载《阅江学刊》2021 年第 4 期，第 41 页。

¹⁵ 参见梁正，王净宇：《美、欧人工智能治理的分歧、影响及其对中国的启示》，载《湖北社会科学》2024 年 12 期，第 42-50 页。

¹⁶ 俞鼎：“有意义的人类控制”：一项新的人工智能伦理治理原则，载中国理论法学研究信息网，2023 年 2 月 10 日，<https://legal-theory.org/?mod=info&act=view&id=27029>。

的操作标准（如地理部署半径<5 公里、任务周期<72 小时）。同时，若国家无法解释 AWS 攻击决策的具体逻辑（如面部识别触发攻击的置信度阈值），则构成对生命权的程序性侵犯。¹⁷

2. 引入严格责任原则

此前应先明确使用 AWS 的国际责任，高级军事指挥官和文职官员应对严重违反国际人道法的行为（如杀害已经投降的战俘，或者不分青红皂白地攻击平民）承担个人刑事责任，同时取得并部署 AWS 国家应对其行为承担国家责任。同时还应涉及人工智能武器的设计者、生产商和销售商。¹⁸进一步地，在《国家责任条款》条款的基础上，对 AWS 引发的损害适用严格责任——即只要证明损害结果与算法缺陷存在因果关系，开发者与部署国即承担连带责任，无需举证主观过错。

3. 提出“预见性控制”

针对《罗马规约》第 28 条“有效控制”标准，在 AWS 场景下可提出“预见性控制”概念：若指挥官明知算法存在系统性风险（如训练数据偏差率超警戒值）仍部署使用，即构成战争罪。

（二）全球治理路径：分阶段建立 AI 武器技术标准与审查制度

AWS 对国际法的挑战本质上是技术自主性与人类控制权的博弈。为化解危机，需采取三阶路径：

1. 短期强化合规审查

短期内应强化 CCW 框架下的合规审查机制，将技术透明治理嵌入国际法律框架。对人工智能武器的合法性审查，需要贯穿从系统研发到采购及使用的整个过程。¹⁹合法性审查标准应包括以下内容：是否属于人道法严禁使用的武器；是否采用禁止使用的作战手段和方法；是否遵守国际人道法的区分原则、比例原则和预防原则等，并应充分结合 AI 技术特点进行细化，以确保《第一附加议定书》第 36 条审查机制的有效性。如：

一，限制 AWS 的攻击目标、操作时间和地理范围，如限制 AWS 部署半径（<5 公里）与任务周期（<72 小时），通过可验证设计降低法律风险。

二，强制 AWS 通过算法透明性和伦理测试，强制公开算法决策逻辑的关键阈值（如目标置信度、攻击触发条件）、关键参数（如 SHAP 值）以验证合规性。

三，要求自主武器系统必须配备“人类紧急干预接口”（如响应时间<0.8 秒）和算法决策日志记录，确保指挥官可终止攻击链。

¹⁷ See Berkant Akkuş, “POSITIVE HUMAN RIGHTS OBLIGATION OF STATES AND THE USE OF AUTONOMOUS WEAPON SYSTEMS DURING LAW ENFORCEMENT OPERATIONS.” *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*(2022).

¹⁸ 参见张卫华：《人工智能武器对国际人道法的新挑战》，载《政法论坛》2019 年第 4 期，第 153-154 页。

¹⁹ 参见付姝菊：《人工智能武器对国际人道法的冲击与回应》，载《华北水利水电大学学报(社会科学版)》2024 年第 5 期，第 79 页。

四，要求 AWS 部署国公开目标识别模型的数据偏差率，要求必须证明所部署的 AWS 具备遵守这些发国际法原则的能力（如通过模拟战斗测试验证能够区分平民和非平民）。

2.中期推动专项立法

中期应推动专项立法。一，可借鉴《集束弹药公约》模式，制定“自主武器管制议定书”，明确“人机协同”的操作边界——例如设定“必要性阈值”（如仅在敌方电子干扰导致通信中断时启动自主攻击），并规定指挥官负有“动态干预义务”（每秒至少一次人工确认指令），使军事审查机构具备实时监测算法演化的技术能力。二，可借鉴 IAEA 核查机制，对自主武器算法实施第三方“白盒测试”，特别是要求闭源武器必须接受第三方审计。三，可借鉴欧盟《人工智能法案》的“高风险系统”分类，建立 AWS 算法透明度分级制度。如将“动态攻击模式”纳入 CCW 的“不可接受伤害”清单。²⁰

3.长期构建全球军控

长期则应构建全球 AI 军控机构：将《关于负责任地军事使用人工智能和自主技术的政治宣言》等软法转化为硬法义务，并通过安理会决议建立 AWS 部署的事前报备制度，防止技术滥用；建立第三方主导的“算法黑箱验证制度”，要求武器系统接受独立技术联盟（如全球人工智能伦理委员会）的实时审计；同时，建立区域性 AWS 算法认证中心，平衡军事创新与人类安全。

（三）中国角色定位：发展人类命运共同体下的智能武器伦理

在上述进程中，作为 AI 技术大国、安理会常任理事国和关键利益攸关方中国的参与对治理格局具有关键意义：提出的“人类命运共同体”框架下“智能武器伦理准则”，将“人类道德可追溯性”设为研发红线，不仅为发展中国家争取规则话语权开辟新路径，更通过“数字丝绸之路”推动治理标准互认（如与东盟签订算法透明度双边协议）。应以“全球人工智能治理倡议”搭建平台，持续发展人类命运共同体下智能武器的治理倡导，重塑传统武器的治理范式，助力在全球范围内建立更加统一和协调的 AI 自动武器治理框架。²¹

（四）司法机构协作：发挥国际法院咨询意见的潜在作用

国际法院（ICJ）作为联合国主要司法机构，可通过发布咨询意见，厘清 AWS 适用国际人道法的具体标准，填补 AI 武器治理的法理真空。具体例如：针对《日内瓦公约》第 36 条“武器可预测性”要求，ICJ 可发布咨询意见明确其适用于 AI 武器的具体标准，如要求算法决策逻辑必须满足“全生命周期可审查性”；ICJ 可援引加沙冲突案例，界定 AI 武器区分战斗员与平民的技术底线，从而明确“区分

²⁰ 参见龙坤，徐能武：《致命性自主武器系统军控——困境、出路和参与策略》，载《国际展望》2020 年第 2 期，第 78-102+152 页。

²¹ 参见费艳颖，赵凯：《全球人工智能治理倡议的世界意义》，载中国社会科学网-中国社会科学报，2024 年 6 月 19 日，https://theory.gmw.cn/2024-06/19/content_37387897.htm。

原则”的操作指南，如规定目标识别模型必须通过“混淆矩阵测试”（平民误判率 $\leq 5\%$ ），否则视为违反《第一附加议定书》第 48 条。

四、结语

技术革命的浪潮正以不可逆的姿态重塑战争规则，利比亚战场上自主武器的血腥实践已昭示：我们并非在“迎接”未来，而是被迫直面技术超前性与法律滞后性共生的现实困局。应对这一挑战，国际法需构建阶梯式的治理框架：短期应依托《特定常规武器公约》强化技术透明机制，强制部署国公开算法决策的关键阈值，以遏制滥用风险；中期需推动“自主武器管制议定书”落地，明确人机协同的操作边界；长期则需成立全球性 AI 军控机构，建立第三方主导的“算法黑箱验证制度”。与此同时，国际法院可通过发布咨询意见，为立法提供法理支点。在此进程中，中国的角色至关重要，持续深化“人类命运共同体”框架下“智能武器伦理准则”的倡导与实践。

国际法对于 AI 自动武器的回应，必须超越传统“禁止/许可”的二元规制逻辑，转而构建“技术标准-法律责任-伦理约束”三位一体的治理范式。唯有通过技术透明化、责任链条固化与法律动态适配的协同治理，以预防性法律机制锚定技术狂飙，以多边共治重塑权力格局，才能避免 AWS 沦为“国际法真空地带的杀戮工具”，找到人工智能战争与人类文明存续之间动态平衡的支点，捍卫马尔顿条款所承载的人类文明底线。

国产开源人工智能兴起影响下数字丝绸之路的制度优化

何鑫龙¹

摘要：数字丝绸之路作为“一带一路”倡议在数字领域的延伸，围绕基础设施和制度两个方面推动“硬联通”和“软联通”。长期以来，数字丝绸之路的建设面临着结构性困境，包括技术应用困境、政策协调困境以及地缘政治困境。而以 DeepSeek 为代表的国产开源人工智能作为重要的数字基础设施，由于算法优化和开源策略两大特点，在国际上引起巨大反响，对于化解上述困境有着重要意义。然而“硬基建”应当和“软制度”两条腿走路，DeepSeek 的广泛应用可能阻碍数据流通、滥用误用引发风险以及加剧国际治理规则的割裂。因此，我国应当从加强推动到促进开源人工智能利用领域的相关规则标准建设，具体而言包括推动数字丝路数据流通等领域机制的完善、与开源人工智能同时推广安全标准、建立协助发展中国家提升开源人工智能能力的长效机制等等。

关键词：DeepSeek 开源人工智能 数字丝绸之路 “一带一路”倡议 数字治理

一、前言

“数字丝绸之路”是以深化数字经济国际合作为基本内容，通过与“一带一路”沿线国家和地区共建信息基础设施、推动信息共享、促进信息技术合作、推进互联网经贸服务和加强人文交流，共同发展更广范围、更宽领域、更深层次的信息经济，携手共建政治互信、经济融合、文化包容的网络空间“命运共同体”。²人工智能为代表的数字技术对于经济社会等领域产生了全方位的变革。习近平总书记强调，“人工智能是引领这一轮科技革命和产业变革的战略性技术，具有溢出带动性很强的‘头雁’效应”。³当前，人工智能在第四次科技革命之中处于核心地位。世界主要国家都将发展人工智能放在首要地位。⁴联合国在《加强人工智能能力建设方面的国际合作》决议中提到：“注意到开源软件、开放模式和开放数据等方法 and 业务模式在传播人工智能惠益方面的积极益处及潜在风险，并认识到发展中国家参与相关全球合作的重要性，”⁵

¹ 黑龙江大学学生，17721374260

² 《探讨“数字丝绸之路”发展之路》，载中国中央人民政府网 2019 年 08 月 27 日，https://www.gov.cn/xinwen/2019-08/27/content_5425070.htm。

³ 《加快形成以人工智能为引擎的新质生产力》，载求是网 2024 年 10 月 30 日，http://www.qstheory.cn/qshy/jx/2024-10/30/c_1130214673.htm。

⁴ 参见余乃忠：《人工智能时代的中国机遇：第四次科技革命的领导者》，载《重庆大学学报（社会科学版）》2020 年第 2 期，第 185-194 页。

⁵ 《加强人工智能能力建设方面的国际合作》，载联合国 2024 年 07 月 05 日，https://www.fmprc.gov.cn/wjb_673085/zzjg_673183/jks_674633/fywj_674643/202412/P020241121627250527218.pdf

在全球数字经济蓬勃发展和“一带一路”倡议进一步深入发展的复合背景下，数字丝绸之路正在面临着技术垄断、治理碎片化和地缘政治竞争等多方面因素带来的挑战。数字丝绸之路中许多发展中国家数字领域的发展举步维艰。然而，以 DeepSeek 为代表的国产开源人工智能作为重要的数字基础设施，为打破数字丝绸之路发展的结构性困境提供了突破口。然而“硬基建”的突破仍需“软制度”作为保障和支撑。为推广国产开源人工智能进行配套的制度升级，平衡开源生态与数字主权、技术发展与风险防控，成为数字丝绸之路制度优化的重点。

本文立足数字丝绸之路的技术应用、政策协调与地缘政治竞争的三重困境，分析 DeepSeek 等国产开源人工智能破局逻辑和潜在的挑战，从数据流通机制、安全标准建设与能力建设长效机制等维度提出制度优化方案。

二、数字丝绸之路当前的困境

作为“一带一路”倡议在数字领域的延伸，数字丝绸之路的建设同样要坚持以基础设施“硬联通”为重要方向，以规则标准“软联通”为重要支撑。⁶二者相辅相成、共同促进。同时，数字丝绸之路建设的目的并非在全球数字治理领域划定“势力范围”，而是要建立一个开放包容的数字治理体系，成为全球治理的重要组成部分。数字基础设施也在不断演进。当前，新一代人工智能基础设施的构成更加复杂，包括算力和 Maas（模型即服务）以及相关工具。大模型是新一代人工智能基础设施的重要组成部分。⁷人工智能技术的兴起，极大地推动了数字基础设施的变革。随着人工智能的广泛应用，与之相适应的规则标准也应运而生，全球各国国内以及国家之间涌现出了大量相关规则标准。人工智能带来的挑战是全球性的，所以必须通过全球治理应对相关挑战和问题。⁸这使得国家间的合作更加必要。在此背景下，推进数字丝绸之路建设的过程中，受多种因素影响，在“硬基建+软制度”的建设模式，以及开放包容治理体系建立的目标导向下，数字丝绸之路建设逐渐显现出了诸多结构性困境和系统性挑战。

（一）技术应用困境

人工智能开发和使用的成本极高，需要超算中心、数据中心、网络基站等数字基础设施和相关的技术人才作为支持，此外还可能支出使用相关技术的专利费用，广大发展中国家难以承担这些高昂的成本。⁹据国际能源机构的报告显示，到 2026 年，数据中心、加密货币以及人工智能的数字技术或服务的电力需求相较于 2022 年，将会增加 160~590 太瓦时，大约相当于增加一个瑞典到一个

⁶ 参见《共建“一带一路”：构建人类命运共同体的重大实践》，载中国中央人民政府 2023 年 10 月 10 日，https://www.gov.cn/zhengce/202310/content_6907994.htm

⁷ 参见中国信息通信研究院云计算与大数据研究所等：《新一代人工智能基础设施白皮书》，<https://www.sensecore.cn/whitepaper.pdf>。

⁸ See Jean Garcia Periche, *Artificial intelligence and the future of global governance*, AI Policy Exchange (23th March 2025), <https://aipolicyexchange.org/2020/08/07/artificial-intelligence-and-the-future-of-global-governance/>

⁹ 参见封帅：《高昂成本加剧全球人工智能“发展鸿沟”》，载新民网 2024 年 11 月 14 日，<https://paper.xinmin.cn/html/xmwb/2024-11-14/14/197077.html>。

德国的用电量。¹⁰除此之外，此前在国际上知名度较高、产品性能领先的 Open AI 采用的是闭源收费的模式，如果发展中国家想要使用高级功能或者应用于商业领域，需要额外缴纳商业授权费等一系列费用。而“一带一路”沿线大多是发展中国家和不发达国家，难以负担高昂的成本。而且发达国家和发展中国家的数字鸿沟还在进一步扩大。以网络为力量，发达国家在广泛采用 5G 网络的同时，一些低收入国家还不得不使用 2G 和 3G 网络。许多先进的人工智能技术可能在广大南方国家当中难以部署。

当前，人工智能技术聚集在少数国家手中。以基础模型为例，根据斯坦福《2025 人工智能指数》报告，世界上绝大多数基础模型来源于美国、中国和部分西欧国家。大部分国家扮演的是使用者的边缘角色，难以参与到人工智能的改进当中。因此，发展中国家即便能够部署人工智能，也会面临本土化不足带来的一系列问题：一方面，Chat GPT 等模型在训练时主要使用的是英文数据，在生成其他语言的内容时出现误解和偏见的概率更高，而且在实际应用过程中透露出对于西方价值观和利益偏好的维护，并不能做到价值中立；另一方面，广大南方国家的应用需求和西方国家并不相同，其独特的需求难以得到满足。¹¹不同地区国家国情不同，在发展过程中各有侧重，如果西方提供的人工智能不经过本地化定制，则难以起到理想的效果。¹²

（二）政策协调困境

数字丝绸之路的内容还包括促进规则标准方面的互联互通。现实中，“一带一路”倡议覆盖的范围极广，各国之间的法规、政策各不相同。从纵向发展水平上来看，数字丝绸之路沿线国家在基础设施和技术水平等方面的数字鸿沟也映射到了制度发展水平上。各国的法规政策的发展程度不尽相同，以中亚地区为例，总体上看，中亚地区国家的数字立法水平有待提升，同时也缺乏类似东盟的区域数字规则，而且中亚五国之间的立法水平同样存在差异，哈萨克斯坦相对较为完善，存在数据等领域的立法，但是塔吉克斯坦、土库曼斯坦的数字立法在很多领域仍存在空白。¹³在国际合作方面也是如此，在联合国《治理人工智能，助力造福人类》报告当中列举的七个非联合国国际人工智能治理倡议当中，有一百余个国家未参加任何治理举措，其中绝大部分为亚非拉国家。¹⁴基

¹⁰ See *Electricity 2024-Analysis and forecast to 2026*, IEA (January 2024), <https://iea.blob.core.windows.net/assets/18f3ed24-4b26-4c83-a3d2-8a1be51c8cc8/Electricity2024-Analysisandforecastto2026.pdf>.

¹¹ 参见郑恩、刘沫潇：《全球传播秩序的算法重构：DeepSeek 挑战西方中心主义的 AI 认知霸权》，载《学术探索》2025 年；韩永辉、周港隽、徐翠芬：《人工智能全球治理的现状、困境和中国路径》，载《特区实践与理论》2024 年第 6 期，第 94-102 页。

¹² 参见韩建伟：《中东数字鸿沟问题与“数字丝绸之路”的高质量发展》，载《国外社会科学》2022 年第 4 期；黄玉沛：《中非共建“数字丝绸之路”：机遇、挑战与路径选择》，载《国际问题研究》2019 年第 4 期；王海燕：《中国与中亚国家共建数字丝绸之路：基础、挑战与路径》，载《国际问题研究》2020 年第 2 期。

¹³ 同上。

¹⁴ 参见《治理人工智能，主力造福人类》，载联合国，

于物质基础，发展中国家很难和国际上的先进标准进行对接，例如欧盟的 GDPR 数据专属规则的应用对于相关技术、硬件设施、配套制度以及资金支持要求极高。发展中国家往往面临着一刀切式的艰难抉择。¹⁵

人工智能的应用是一把双刃剑，在带来机遇的同时也带来诸多安全隐患和挑战。若不加以规制，则可能对一国的政治、社会、经济带来重大风险。¹⁶因此，对于人工智能的安全治理受到各国广泛关注和重视。从横向上来看，数字丝绸之路沿线国家国情各异、核心利益诉求也有所不同。¹⁷新加坡等数字经济、数字技术发达的国家，往往关注增强自身核心竞争力，而更多的数字水平较为薄弱的发展中国家往往会采取防御性技术民族主义战略。¹⁸在总体政策上强调政府的重要作用，往往倾向于通过加强监管和限制以减少风险，在国际竞争方面则采取一定程度的封闭政策和保护政策以保护本国技术的发展，促进实现科技自主、自强。在这种情况下，各国在人工智能等数字领域法规政策的差异极大增加了数字丝绸之路建立规则标准“软联通”的难度，无论是协调各国政策还是建立统一规则，都将面临不同程度的阻碍。

（三）地缘政治困境

当前，在数字领域，国际社会围绕数字治理能力展开了激烈竞争。¹⁹中国、美国和欧洲都正采取新措施来应对新时代的机遇与挑战。²⁰长期以来，全球在数字领域的治理深受美国主导的“中心——外围”霸权体系的影响。²¹美国在人工智能领域占据优势地位，以 Open AI 出品的 ChatGPT 为代表的人工智能产品在全球范围内拥有远超其他同类产品的性能与影响力。同时，美国为了维持自身的霸权优势，还在算力等方面对我国以及部分“一带一路”倡议沿线国家进行限制或封锁。例如，美国推出《人工智能扩散框架》并在 2025 年 1 月生效，其中将全球国家分为“授权国家”“高风险国家”“一般国家”。²²除了“授权国家”之外，其余国家都会受到不同程度的限制或者封锁。还包括对 AMSL 等非美国企业对中国出口进行施压。²³除了出台法案对我国进口硬件进行影响之外，还对

https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_zh.pdf, 2025 年 03 月 24 日访问。

¹⁵ 参见李蕊：《发展中国家因应 GDPR 数据跨境传输规则的困境与启示——以印度为例》，载《武大国际法评论》2024 年第 4 期，第 76-93 页。

¹⁶ 参见于蓝：《人工智能安全隐患与挑战：新兴技术带来的风险》，载《新安全》2025 年第 3 期。

¹⁷ 参见黄日涵、柏瑞山：《从治理规则看“数字丝路”建设》载光明网，https://theory.gmw.cn/2022-08/13/content_35951121.htm。

¹⁸ 参见孙海泳：《进攻性技术民族主义与美国对华科技战》，载《国际展望》2020 年第 5 期。

¹⁹ 参见蒋万胜、刘玲霞：《国际数字主权领域内的多元竞合与应对》，载《中国特色社会主义研究》2023 年第 5 期。

²⁰ 参见黄黎洪：《中国建设数字丝绸之路与数字化时代全球治理的变革研究》，载《电子政务》2019 年第 10 期。

²¹ 参见张东冬：《数字外交强化与美国全球数字竞争的全新图景》，载《国际关系研究》2025 年第 1 期。

²² See Framework for Artificial Intelligence Diffusion, FEDERAL REGISTER, <https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion>

²³ 参见《美国“设卡阻挠”中国再造台积电、ASML》，载观察者网，https://www.guancha.cn/economy/2024_12_04_757660.shtml

我国的基建出口进行阻挠，例如对于华为的 5G 技术进行打压，企图在数字领域进一步“去中国化”。²⁴在意识形态领域，美国为首的西方国家以“数据威权主义”对数字丝绸之路进行污名化，是美国开展技术脱钩和“小院高墙”格局的延伸。²⁵

对于丝绸之路沿线的其他国家而言，大多发展中国家不同程度地受到美国为首的西方国家的数字殖民，在物理空间和网络空间受制于拥有技术优势的发达国家。²⁶在外交方面，美国以自身主导的数字技术产业链为基础，积极推动构建“民主科技联盟”，沿线其他发展中国家面临两难的困境。以美国推行的“清洁网络计划”为例，“清洁网络计划”剥夺其他国家选择数字服务供应商和服务的自由，迫使一些国家在中美在数字领域的竞争当中选边站。

三、开源人工智能对于困境突破的影响路径

联合国《全球数字契约》中提到：“我们承诺到 2030 年：（a）通过多利益攸关方合作，开发、传播和维护安全、可靠的开源软件、开放数据、开放人工智能模型和开放标准，造福全社会（目标 8、9 和 10）……”²⁷相对于此前国际上影响力较大的 Open AI 等公司推出的 ChatGPT 等闭源人工智能模型，以 DeepSeek 推出的 V3 和 R1 模型为代表的开源人工智能同时有着算法优化和开源策略两大特点，具有低成本、高性能、开放性等优势。对于沿线广大发展中国家而言，极大降低了其使用人工智能技术的成本。而且 DeepSeek 相较于 Meta 等公司的开源模型，开放程度更高，不仅便于本地部署，满足发展中国家对于安全的需要，而且也降低了发展中国家参与到技术创新和优化过程中的门槛。对于数字丝绸之路所追求的推动数字化转型和打造数字领域互联互通的目标以及推动联合国 SDGs 的实现有着重要推动作用。

（一）以本地部署强化主权安全保障，筑牢发展共识基础

数字主权安全具有技术依赖性，数字技术直接关系到国家的数字管辖权、参与国际治理的权力以及维护自身安全的能力。²⁸以 DeepSeek 为代表的开源人工智能的开放程度非常高，并且部署成本低，极大降低了本地部署的门槛，允许发展中国家自主独立地部署技术链条，为其构建自主可控的数字主权体系提供了可能的路径。《人工智能伦理问题建议书》中指出：“对数据的使用必须尊重国际法和国家主权。”²⁹在立法领域领先的欧盟在其数字治理当中也重视对于

²⁴ 参见王义桅：《美国打压华为背后的几点启示》，载环球网 2020 年 07 月 28 日，<https://opinion.huanqiu.com/article/3zEJyXE4ReY>

²⁵ 参见刘国柱：《“数字威权主义”论与数字时代的大国竞争》，载《美国研究》2022 年第 2 期。

²⁶ 参见蔡翠红、于大皓：《“帝国的新衣”：世界数字体系下的美国数字霸权》，载《当代世界与社会主义》2024 年第 3 期。

²⁷ 《全球数字契约》，载联合国 2024 年 09 月 22 日，<https://www.un.org/zh/documents/treaty/A-RES-79-1-Annex-1>

²⁸ 刘玲霞：《数字主权安全的理论内涵、现实挑战与应对路径》，载《陕西师范大学学报（哲学社会科学版）》2024 年第 1 期，第 122-133 页。

²⁹ 联合国教科文组织：《人工智能伦理问题建议书》，载联合国数字图书馆，

数据和技术的控制和本地化。³⁰在当前数字领域竞争激烈和逆全球化思潮抬头的近代背景下，通过构建自主可控的数字技术、加强对于数据的管控以维护国家主权和保障国家安全，在国际社会上已经得到了一定程度和范围的认可和实践。

然而，理论上的数字主权不等于实际上对于技术的主导权。长期以来，许多在数字领域处于劣势的南方国家需要依赖他国数字技术。长期以来，Open AI、X 平台、META 等来自美国的非国家行为体对其他国家的主权造成了冲击。³¹以数据领域为例，美国 Open AI 的闭源运营策略与 2018 年出台的《澄清境外合法使用数据法案》（下文简称“云法案”）可以形成一套“组合拳”。首先，使用闭源人工智能就意味着数据难以做到本地化。人工智能模型的训练和使用需要大量的数据作为支持，然而作为使用方的国家难以选择数据的储存位置与对数据进行监管。在 Open AI 等人工智能系统的部署和使用中，往往需要受到服务提供方的多重限制。而这种技术层面的限制机制将与美国相关制度形成衔接，为美国依据《云法案》进行长臂管辖提供了技术基础。《云法案》在数据管辖权扩张，削弱其他国家对于本国数据的自主控制权，并且在司法审查豁免等方面提供了制度“保护伞”。³²技术上的差距造成了国家在数字领域得不对等性，发展中国家的数据要流向美国，而美国的数据却很少以对等的方式流向发展中国家。以美国为首的数字霸权主义国家构建了“技术+法律”的双重霸权工具，在这个工具当中，技术先发优势为制度结构性压迫提供基础。虽然很多发展中国家也有数据本地化立法的规定，但是由于缺乏自主技术，仍然需要依靠美国为首的发达国家提供的数字技术。本国的法律不得不让步于技术缺乏的现实。联合国多项文件中提到的数字领域的主权原则也变得形同虚设。

因此，发展中国家保护自身主权安全的物质基础是自主可控的数字技术，其中就包括数字基础设施。开源人工智能作为数字基础设施之一，对于发展中国家的主权安全至关重要。2024 年世界政府峰会上，英伟达创始人兼 CEO 黄仁勋提出了“主权人工智能”的概念。虽然当前关于主权人工智能学界尚未存在相对一致的定义，但是显然，建立自主安全可控的人工智能基础设施是其中的重要环节。³³主权人工智能概念反映出主权概念在人工智能时代的延续，对于国家安全至关重要。³⁴DeepSeek 开源的特点降低了本地部署人工智能的门槛，

https://unesdoc.unesco.org/ark:/48223/pf0000381137_chi/PDF/381137chi.pdf.multi，最后访问时间 2025 年 03 月 21 日。

³⁰ 孙通、舒伟超：《从数据保护到数字主权构建：欧盟数字安全治理的历程、动力与限度》，载《世界经济与政治论坛》2025 年第 1 期。

³¹ See Timmers P., *AI challenging sovereignty and democracy*. Turkish Policy Quarterly, 45-55(2021).

³² See CLOUD Act § 2713; CLOUD Act § 2523.

³³ 参见付建龙：《加强“主权 AI”建设，提升竞争力》，载《人民周刊》2024 年 11 月 18 日，第 8 版。

³⁴ 谢新水：《智能跃迁、开源创新与主权 AI：DeepSeek 现象推动人工智能开源创新生态体系建设》，载《电子政务》2025 年。

可以有效地成为维护发展中国家维护主权安全的技术保障，全面开源的策略允许开发者基于模型进行二次开发。³⁵发展中国家可以自主掌控人工智能技术的全过程，打破闭源人工智能的“技术黑箱”垄断，更易于监管，并且可以在部署过程中插入监管系统，维护国家主权安全。本地化部署还有助于打造关键领域的闭环训练体系。例如在国防、医疗等重点领域实行全过程本地化训练，从数据采集到应用于模型训练和验证的全部环节均可在本地完成，避免高安全敏感性的数据流出。技术依赖导致本国法律政策的自主性受限，反之，技术自主性的增强将会增强本国法律政策的落实，从而促进数字主权的实现。

（二）依托低成本部署特性，助推弥合全球数字鸿沟

人工智能对于新质生产力的发展具有重要的推动作用，可以为实现经济转型注入强大动力。³⁶在 Open AI 于 2020 年发表的论文中提出了“规模法则”，指出模型规模、数据量和计算资源与模型性能之间存在正向关系。³⁷长期以来，人工智能发展以发展算力和加大投入为途径。根据斯坦福《2025 人工智能指数》报告显示，例如 GPT-4、Gemini Ultra 等先进模型的训练成本十分高昂。³⁸众多南方国家难以承担研发与部署人工智能的费用，只得被迫采用发达国家提供的人工智能方案，除了数据存储、运输以及系统维护等必要成本之外，还要支付额外的授权费用。进一步加大了南方国家的使用负担。算力是数字经济时代的核心生产力，然而，在清华大学等机构发布的《2022-2023 全球算力指数评估报告》中统计的 15 个国家中，巴西、南非等南方国家距离发达国家有一定差距。而且在算力领域同样存在着“马太效应”：算力先发国家优势会随着算力投资比重的增加进一步获得强化。³⁹

算力领域的差距同样映射到了全球人工智能产业链当中。美国等国家凭借自身技术优势占据产业链的高端位置、掌控着相关产业的关键环节，加剧了数字体系中的不平等。⁴⁰在技术上处于劣势的广大南方国家面临着产业发展的“低端锁定”。⁴¹即南方国家在缺乏产业基础和遭遇产业升级壁垒的双重挑战之下，难以独立实现技术追赶。除了人工智能技术的“马太效应”之外，数字经济本身还具有集聚效应，美国等国可以凭借技术优势获得更大的垄断利润。⁴²2003 年

³⁵ 魏钰明、贾开等：《DeepSeek 突破效应下的人工智能创新发展与治理变革》，载《电子政务》2025 年。

³⁶ 董志勇：《“人工智能+”赋能新质生产力发展》，载人民网 2025 年 01 月 13 日，

<http://theory.people.com.cn/n1/2025/0113/c40531-40400643.html>

³⁷ Kaplan J, McCandlish S, Henighan T, et al. *Scaling laws for neural language models*. Computer Science(2020).

³⁸ See *The 2025 AI Index Report*, Stanford University (11 April 2025), <https://hai.stanford.edu/ai-index/2025-ai-index-report>

³⁹ 参见《2022-2023 全球算力指数评估报告》，载浪潮信息 2023 年 07 月 12 日，

<https://www.ieisystem.com/global/file/2023-09-14/16946384123912c916ead8a3cb8ee62018a90521267362c.pdf>

⁴⁰ 参见蔡翠红、于大皓：《“帝国的新衣”：世界数字体系下的美国数字霸权》，载《当代世界与社会主义》2024 年第 3 期。

⁴¹ 参见秦北辰：《数字技术、权力失衡与全球南方产业发展的困境》，载《外交评论（外交学院学报）》2024 年第 3 期。

⁴² 参见孙志燕、郑江淮：《全球价值链数字化转型与“功能分工陷阱”的跨越》，载《改革》2020 年第 10 期，第 63-72 页。

至 2023 年间，仅有大约三十个国家拥有具备国际影响力的机器学习模型，美国以 61 个的数量位居榜首，远超位居第二的中国的 15 个，占全球总数近一半。而 2019 年到 2023 年间发布基础模型的机构当中，有五家机构贡献了十个以上的基础模型，其中前四名均为美国企业。⁴³除了在人工智能产业分工中居于下层之外，发展中国家还被迫承受连带的其他代价，例如数字化带来的环境影响。联合国《2024 年数字经济报告》中指出：发展中国家既无法充分享受数字化带来的惠益，又深受其负面环境影响的困扰。许多发展中国家一方面成为原材料和能源的供应方，另一方面还要承接数字化产业当中产生的废物。⁴⁴

DeepSeek 通过模型和算法创新极大降低了使用成本和训练成本，在性能相近的情况下极大降低了对算力的依赖。⁴⁵对于广大发展中国家而言，DeepSeek 可以从三方面对于破除困境起到作用：首先，DeepSeek 推出的模型通过算法优化，具备轻量化的特点，除了模型自身需要资源较少之外，还能够通过蒸馏进一步压缩成本，一定程度上可以弥补算力方面的差距；其次，DeepSeek 推出的模型均有较高的开源等级，发展中国家不仅可以免费本地部署，还可以以此为基础进行更贴合本国需要的应用开发；最后，DeepSeek 推出的开源模型也可以在一定程度上减少地缘政治竞争影响下的算力封锁的影响，为发展中国家提供了加大算力和其他资源投入以外的另一条人工智能发展路径。

（三）借助开源策略破局，有效对冲美国数字霸权垄断

当前，数字技术成为新时期地缘政治竞争的焦点。⁴⁶美国十分重视人工智能在军事国防和产业数字化方面的强大动力，将人工智能技术的领先与维护美国霸权紧密结合。⁴⁷美国构建了一套“小院高墙”战略，一方面，利用价值观构筑科技联盟，筛选盟友；另一方面，加强对于技术主导权和话语权的控制，防止技术流向其他国家。⁴⁸最终总体上构成一个以美国为圆心，盟友为外围，其他国家为边缘的技术霸权圈层体系。为了维护这个体系，美国搭建起涵盖硬件、软件和规则在内的“三维”支撑架构，其中三个要素相互支持。除了先进算力和人工智能模型之外，美国还在规则层面加强技术封锁。在治理规则层面，人工智能全球治理领域内规则博弈十分激烈，技术上占据优势地位的国家和地区通过技术标准和规则输出等方式增强自身的制度性权力。以欧盟为代表的技术优

⁴³ See *The 2025 AI Index Report*, Stanford University (11 April 2025), <https://hai.stanford.edu/ai-index/2025-ai-index-report>

⁴⁴ 参见《2024 年数字经济报告|概述》，载联合国，https://unctad.org/system/files/official-document/der2024_overview_ch.pdf，2025 年 03 月 27 日访问

⁴⁵ 参见陈智：《新一代生成式人工智能的价值定位、演进逻辑及前瞻路径——以 DeepSeek 为例》，载《新疆社会科学》2025 年。

⁴⁶ 刘国柱：《地缘政治视野下的新兴技术与美国技术民族主义》，载《当代世界》2020 年第 10 期，第 30-37 页。

⁴⁷ 戚凯：《ChatGPT 与数字时代的国际竞争》，载《国际论坛》2023 年第 4 期，第 3-23+155 页。

⁴⁸ 参见高奇琦：《论通用人工智能的新突破对世界政治经济的多重影响》，载《国际观察》2024 年第 3 期。

势方旨在在国际规则制定当中获得主导地位，通过“布鲁塞尔效应”将《通用数据保护条例》《人工智能法案》等欧盟内部法规的效力扩大化，将其上升为全球标准。

数字丝绸之路也覆盖了大量的南方国家，这些国家在国际数字治理当中处于弱势地位，倾向于通过法律等手段维护本国的数据主权，具体表现形式包括数据本地化和数据跨境管制。⁴⁹当前，部分数字霸权主义国家滥用长臂管辖，侵犯其他国家的数据主权。例如，美国以《云法案》为法律工具推行长臂管辖的举措，迫使其他数字技术水平处于劣势的国家不得不采取更加保守的数据法规以维护数据主权。⁵⁰对全球数据的正常流动造成了阻碍。而俄罗斯、印度等技术水平和话语权稍逊的国家则制定有利于维护数字主权和技术自主化的法规，表现为防御性的政策，以谋求在弱势领域的“战略对冲”。⁵¹DeepSeek 等国产人工智能为突破当前地缘政治困境提供了突破路径，开源架构和技术透明性有助于打破“小院高墙”体系下的技术垄断。使得发展中国家能够建设自主可控的数字化能力，减少对于霸权技术体系的依赖，为多元主体的包容性治理体系的建设提供物质基础。

四、国产开源人工智能助数字丝绸之路的法律挑战

以由中国企业自主研发的 DeepSeek V3、R1 等开源模型为例，其不仅成为数字丝绸之路建设中的新型数字基础设施，更因为其技术普惠的特征和开源策略，契合了《全球数字契约》当中数字公共产品的开放性、包容性以及可持续性等特征。⁵²但是也正是因为这些特性，开源人工智能是一把双刃剑。想要充分发挥开源人工智能的强大动力，就必须在其推广和利用的过程中关注可能出现的新问题和风险。要确保积极作用大于消极影响，并且将消极影响控制在可接受范围之内。在数字丝绸之路的建设中，这些消极影响很有可能影响跨国协作。数字丝绸之路的规则互联互通面临三重外部挑战：就“一带一路”范围内的国家而言，一方面，各国之间的国情和利益关注不同，在有关数据治理的规则上存在较大差异，可能为开源人工智能的推广造成一定的阻碍；另一方面，各国之间的发展水平不同，数字治理能力也不同，部分国家由于相关领域规则的缺乏可能成为数字丝绸之路规则标准互联互通建设当中的短板。而在“一带一路”倡议建设的外部，以美国为首的数字霸权国家将人工智能治理为代表的数字

⁴⁹ 参见姚天、王宇：《数据主权视阈下的数据跨境流动及中国因应》，载《国际商务（对外经济贸易大学学报）》2025年第2期，第141-156页。

⁵⁰ 参见廖明月、王佳宜、杨映雪：《美国 CLOUD 法案数据跨境执法中的安全风险与中国的应对》，载《图书馆论坛》2025年第1期，第128-137页。

⁵¹ 参见赵文丹、王婷婷：《简析俄罗斯互联网治理的理念及模式》，载《中国广播电视学刊》2023年第5期。

⁵² 参见中国信通院：《人工智能基础设施发展态势报告（2021年）》，载中国信通院2022年01月27日，<https://www.caict.ac.cn/kxyj/qwfb/ztbg/202201/P020220127577307796240.pdf>；《全球数字契约》，载联合国，<https://www.un.org/zh/documents/treaty/A-RES-79-1-Annex-I>，2025年03月30日访问。

治理议题政治化、意识形态化，围绕技术标准和治理规则的制定与中国开展竞争，打压中国推动建设的数字领域的“人类命运共同体”。

（一）跨境治理规则挑战：开源生态与维护数据主权政策的矛盾

数据是人工智能的“粮食”和“养料”。数据贯穿人工智能利用和发展的全过程。投入人工智能的数据的质量和数量与人工智能的性能呈现正相关。⁵³在国际层面，数据的流动性为人工智能在全球范围内的推广、运用和发展提供了充分条件。⁵⁴同时，数据的流动性也是围绕开源人工智能所开展的全球合作的基础。

然而，当前数字丝绸之路的国家之间差异较大，不同国家对于数据治理有着不同的态度。围绕国家间数据流动问题，国产开源人工智能的推广以及围绕开源人工智能开展的相关合作面临着两个层次的法律挑战分别是：区域内国家数据规制存在差异、同一国家加入的不同规则体系之间存在竞合。中国—东盟是全面战略伙伴，在数字经济和智能发展领域有着坚实的合作基础和广泛的前景。⁵⁵以此为例，国产开源人工智能在东南亚国家的推广正面临着一系列的法律挑战。

相较于欧盟，东盟在数字领域一体化程度不高，虽然东盟先后出台了《东盟个人数据保护框架》《东盟数字一体化框架》等跨境数据治理领域的法律法规，但是因为缺乏法律约束力，尚未在范围国家里提出一套一致的规则。⁵⁶在东盟国家内部，不同国家之间对于跨境数据流动的治理模式有着较大的差异。总体而言，新加坡和菲律宾对于数据流动并持较为开放的态度，新加坡《个人数据保护法》第15条规定了若干视为同意的情形，简化了收集、使用数据的程序。第17条和附表一、附表二规定了无需同意便可收集、使用的数据类型，涵盖个人重大利益、公共利益和广播服务。第二十六条规定了当接收者能够提供“等同的保护水平”时，便不对数据传输进行限制。⁵⁷菲律宾的《数据隐私法》第11~14条从原则、无需同意的情形、敏感信息利用的情况、信息分包等角度对于数据信息收集使用进行了规定，同时在第6条当中规定了域外适用，表现出对于跨境数据流动的认可。⁵⁸而越南《个人数据保护法》则十分重视数据本地化，对于数据流动的限制较为严格，例如第7条和第八条分别规定了对于个人

⁵³ 参见马治国、张楠：《通用人工智能的数据风险及法治应对路径》，载《北京工业大学学报（社会科学版）》2024年第5期，第131-142页。

⁵⁴ 参见谢新水：《智能跃迁、开源创新与主权 AI:DeepSeek 现象推动人工智能开源创新生态体系建设》，载《电子政务》2025年第3期。

⁵⁵ 参见《中国—东盟人工智能发展与治理合作：进展观察和推进建议》，载中国国际问题研究院，<https://www.ciis.org.cn/xwdt/202412/W020241208737956451968.pdf>，2025年04月02日访问。

⁵⁶ 林梓瀚：《东盟数据跨境流动制度研究：进程演进与规则构建》，载《世界科技研究与发展》2024年第3期，第306-317页。

⁵⁷ See Personal Data Protection Act, Singapore Statutes Online(03 April 2025), <https://sso.agc.gov.sg/Act/PDPA2012?WholeDoc=1#pr26->.

⁵⁸ See DATA PRIVACY ACT OF 2012, NATIONAL PRIVACY COMMISSION(03 April 2025), <https://privacy.gov.ph/data-privacy-act/#>.

数据的访问限制和对于个人数据处理的程序，相较于新加坡，越南对于“视为同意”的情况进行了较为严格的限制，第 10 条对于无需同意便可收集、使用个人数据的情况限制也较强，不包含广播服务等情形。第 21 条对于数据的跨境转移进行了严格的规定，除了要获得主体同意之外，还需要接受定期的评估并进行传输记录。⁵⁹除次出外，例如马来西亚、印度尼西亚等国也有着不同的规定。⁶⁰在数据政策差异显著的情况下，构建依托开源人工智能的互联互通面临着制度性障碍。区域内数据无法互联互通意味着难以构建规模庞大的训练数据集。而 DeepSeek 提供的开源人工智能又降低了本地部署的门槛，意味着国家可以在需要的情况下加强数据本地化程度，与开源社区的数据共享需求不符，割裂技术生态。

当前全球数据治理领域的国际规则存在体系庞杂，相互交叉的情况，许多国家同时成为不同相关领域国际条约的缔约方，存在因此导致规则重叠或者使用冲突的情况。在东盟国家当中，文莱、马来西亚、新加坡和越南同时是 RCEP 和 CPTPP 的成员。而 RCEP 和 CPTPP 对于数据治理的要求不尽相同。通过对比二者在“电子商务”章节中的规定，包括但不限于如下区别：1.对于跨境数据传输的规定，RCEP 第十五条规定了数据流动当中关于“合法的公共政策目标”的例外措施，并且给予了文莱、缅甸、柬埔寨和越南落实跨境传输条款的缓冲期，CPTPP 要求例外措施必须满足“无歧视”的要求，并且没有缓冲期；2.关于电子认证的规定上，RCEP 第六条规定，当事方可以自主选择认证技术和模式，而 CPTPP 第 14.6 条规定禁止缔约方对于认证技术的选择进行限制；3.就个人信息保护而言，RCEP 第八条规定“缔约方应当在可能的范围内合作”，CPTPP 第 14.11 条对于建立更具兼容性的数据流通机制提出了更加具体的要求。RCEP 和 CPTPP 之间的重叠和冲突仅仅是一个缩影，在众多不同类型、不同范围的法规之间，deepseek 等国产开源人工智能不仅在跨境数据获取和传输的过程中面临存在分歧甚至相互冲突规则的限制、难以构建统一的开源生态社区。在 DeepSeek 等模型的推广过程中面临着不同法规之间适用的不确定性，很可能面临需要做好应对多项数据治理规则的准备，将大幅度提升合规成本。

（二）技术安全规则挑战：跨境风险防控规则的缺失

《全球数字契约》提到：“我们确认，数字公共产品，包括遵守隐私和其他适用国际法律、标准和最佳做法且无害的开源软件、开放数据、开放人工智能模型、开放标准和开放内容，使社会和个人能够将数字技术用于其发展需求，并能促进数字合作和投资。”⁶¹然而，开源人工智能的应用存在诸多潜在的风

⁵⁹ See Nghị định số 13/2023/NĐ-CP của Chính phủ: Bảo vệ dữ liệu cá nhân, VIETNAM GOVERNMENT(04 April 2025), <https://chinhphu.vn/?pageid=27160&docid=207759&classid=1>

⁶⁰ 赵祺：《中国—东盟跨境数据流动治理合作：现实与路径选择》，载《南洋问题研究》2024 年第 3 期，第 85-98 页。

⁶¹ 《全球数字契约》，载联合国 2024 年 09 月 22 日，<https://www.un.org/zh/documents/treaty/A-RES-79-1->

险，既包括涉及知识产权侵权、不正当竞争等合规性风险，也存在危及一国政治、社会安全的安全性风险。不同种类的风险相互关联交叉，共同构成开源人工智能应用的风险体系。除了风险的复杂性之外，开源人工智能还有着技术迭代速度快、推广范围广的特点。⁶²同时，由于 DeepSeek 为代表的模型开放程度较高，还存在二次开发后的追责难题。开源人工智能风险与治理之间的差距愈发明显。鉴于开源人工智能治理必须面对平衡其创新和风险防控的需求，在国际层面，有必要构建包含广泛主题的多主体跨境协同治理机制。⁶³这一机制的建立需要对于分类监管模式的设计和主体之间的权责分配进行合理规定，对于治理主体的立法水平提出了极高的要求。

当前，追求人工智能的风险进行分级分类的精准管理已经成为当前国际社会的普遍共识。⁶⁴在人工智能治理领域的区域法规当中立法水平和一体化程度较高的是欧盟地区的法规。以《人工智能法》为例，虽然《人工智能法》给予了开源人工智能豁免，当时一旦开源人工智能属于第 5 条或者第 50 条当中规定的投入市场或者投入使用，或者被认定成高风险系统，则不会受到豁免。实现监管的路径是进行分级分类管理，除了分级内容外，53 条规定了监管沙盒机制，58 条规定了监管任务，104 条对于开源人工智能模型的透明度进行了规定。除此之外，《人工智能法》还对相关主题的风险报告义务。⁶⁵构成了一套全过程强监督的监管模式。然而，这种监管模式在规避风险的同时还意味着高昂的成本以及对产业发展的障碍。⁶⁶数字丝绸之路沿线国家相当一部分不具备构建“欧盟模式”的一体化监管规则体系。目前，欧盟所认定的“具有充分的数据保护水平”的国家当中不包含任何东盟国家。⁶⁷根据斯坦福《2025 人工智能指数》报告，从地理区域上看，东盟、中东、中亚以及非洲等地都存在同区域国家之间人工智能相关法规立法数量差异较大的情况。⁶⁸这反映出这些国家难以将已有的区域合作基础延伸到开源人工智能监管领域。⁶⁹这使得开源人工智能在这

Annex-I

⁶² 参见张亮、陈希聪：《生成式人工智能背景下的跨境数据安全规制——基于 DeepSeek、ChatGPT 等主流 AI 的思考》，载《湖北大学学报（哲学社会科学版）》2025 年第 2 期。

⁶³ See Cross-border collaboration in the digital environment, Digital Regulation Platform(30 September 2020), <https://digitalregulation.org/cross-border-collaboration-in-the-digital-environment-2/>.

⁶⁴ 参见吕德文：《治理技术如何适配国家机器——技术治理的运用场景及其限度》，载《探索与争鸣》2019 年第 6 期，第 61-63 页。

⁶⁵ See European Parliament, Artificial Intelligence Act(Jun. 13, 2024)(AI Act), Article 53, Article 58, Article 97, Article 104, Article 121.

⁶⁶ 参见郭江兰：《高风险人工智能监管的批判与超越》，载《北京工业大学学报（社会科学版）》2025 年。

⁶⁷ Adequacy decisions, European Commission(03 April 2025), https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en.

⁶⁸ See The 2025 AI Index Report, Stanford University (11 April 2025), <https://hai.stanford.edu/ai-index/2025-ai-index-report>

⁶⁹ 参见韩建伟：《中东数字鸿沟问题与“数字丝绸之路”的高质量发展》，载《国外社会科学》2022 年第 4 期；王海燕：《中国与中亚国家共建数字丝绸之路：基础、挑战与路径》，载《国际问题研究》2020 年第 2 期。

些地区推广时，可能面临因当地法规不完善、区域合作不足导致的法律保障和合规困难的问题。

同时，在人工智能全球治理领域，跨国企业、非政府组织等非国家行为体的相对影响力有所上升，在行业标准等跨国规则的制定中具有重要地位。⁷⁰例如美国的 Meta 公司于 2025 年发布的《前沿人工智能框架》（Frontier AI Framework）当中提供了对于开源人工智能系统及其衍生应用的评估框架，这将会随着其提供的 Llama 开源模型的扩散对全球产生广泛的影响。⁷¹在这些非国家行为体对相关规则 and 政策的制定产生积极影响的同时，也要注意，这些行为体主要集中在技术先进的国家和地区。从这个角度讲，广大南方国家的利益和需求可能被进一步边缘化。⁷²因此，国产开源人工智能模型的推广将会面临发展中国家实际需求与来自发达国家行为体的高标准规则之间的矛盾，面临着不同行为体的规则冲突，难以兼顾各方利益的困境。

（三）规则重构挑战：全球治理权竞争

各国在人工智能治理规则领域的竞争反映出科技竞争与地缘政治竞争的融合。人工智能治理规则存在地缘政治化的风险。在 DeepSeek-R1 为代表的国产开源人工智能模型推广的过程中引发了部分国家在治理规则方面的联合打压。在 2025 年初，DeepSeek APP 一度登上多国软件下载榜榜首后，部分国家以国内法规为依据对 DeepSeek 进行了不合理封禁。典型的打压形式包括“意识形态化”和“泛安全化”。美国、澳大利亚等国先后对 DeepSeek 以“安全”为由进行调查。⁷³在制度和话语权方面，美欧等国具有“先发优势”，渲染数字领域的“中国威胁论”将会对我国数字丝绸之路打造人工智能制度层面互联互通造成阻碍。美西方国家英美两国拒绝签署《关于发展包容、可持续的人工智能造福人类与地球的声明》（简称《巴黎人工智能宣言》），其中，英国以国家安全为由拒绝签署，美国虽未直接给出理由，但是表达出对于欧盟较高程度监管的反对，这进一步反映出全球人工智能治理中的分歧日益加剧。⁷⁴参照此前我国数字技术和数字基础设施出海遇到的法律困境，DeepSeek 等国产人工智能在推广过程中将可能面临两重源自国际竞争的法律风险。

⁷⁰ 参见薛澜、赵静：《人工智能国际治理：基于技术特性与议题属性的分析》，载《国际经济评论》2024 年第 3 期。

⁷¹ 参见《Meta 公布「Frontier AI Framework」白皮书，强调在人工智慧系统加上安全防护》，载 yahoo 新闻 2025 年 02 月 05 日，<https://tw.news.yahoo.com/meta-released-the-frontier-ai-framework-white-paper-emphasizing-the-importance-of-adding-security-protection-to-artificial-intelligence-systems-165719412.html>；Frontier AI Framework, Meta (03 April 2025), <https://ai.meta.com/static-resource/meta-frontier-ai-framework>.

⁷² 参见张勤、仝宇：《人工智能驱动全球经济治理变革的理论逻辑及政策路径》，载《亚太经济》2024 年第 3 期，第 1-11 页。

⁷³ See *OpenAI says Chinese rivals using its work for their AI apps*, BBC NEWS(29 January 2025), <https://www.bbc.co.uk/news/articles/c9vm1m8wpr9o>; *Australia bans DeepSeek on government devices over security risk*, BBC NEWS(4 February 2025), <https://www.bbc.co.uk/news/articles/c8d95v0nr1yo>.

⁷⁴ 《共促全球人工智能治理合作》，载人民网 2025 年 02 月 14 日，<http://world.people.com.cn/n1/2025/0214/c1002-40418256.html>.

其一来源于数字丝绸之路范围之外，主要包括美国的长臂管辖和封锁。美国《明确境外数据的合法使用法》《科学与芯片法案》等法规的影响将波及我国国产开源人工智能在数字丝绸之路范围内的推广。例如，DeepSeek 较强的性能表现引发了美国对于半导体出口管制政策是否存在漏洞的怀疑。因此可以预见美国将进一步加强对中国在半导体领域的封锁。⁷⁵除了直接进行制裁和封锁之外，美国还会通过二级制裁，敦促其盟友对中国相关科技企业和技术推广进行制裁和打压。⁷⁶其二来源于数字丝绸之路内部。主要体现欧盟通过严苛的法规对自身数字主权的保护，例如《一般数据保护条例》《人工智能法》等法律均对数据流动进行了严格的管控，《数字市场法》进一步扩大了欧盟的监督权。⁷⁷例如意大利较早地以不符合隐私政策为由封禁 DeepSeek。⁷⁸受美欧影响，韩国、澳大利亚、中国台湾等国家和地区也在不同程度上对 DeepSeek 进行调查和限制。⁷⁹

五、促进开源人工智能赋能数字丝路的制度建设

面对数字丝绸之路沿线国家与沿线国家在发展阶段、技术水平和制度环境方面存在差异的现状，以及国产开源人工智能面对的法律挑战，需要构建“能力提升——风险防控——合作促进”三位一体的制度体系。平衡开源人工智能的发展和​​安全。优质国产开源人工智能除了 deepseek 之外，还包括月之暗面推出的 kimi1.5 等等。⁸⁰我国应当就数据流通、技术安全、能力建设等领域推动相关法规的完善，并且以制度作为保障开展与数字丝绸之路各国的合作，充分发挥国产开源人工智能的技术红利，促进数字丝绸之路的“硬联通”和“软联通”，构建包容普惠的人工智能治理体系，实现人类命运共同体的伟大愿景，促进各国在数字领域共同发展。

（一）推动区域数据流通等法规的衔接

我国《全球数据跨境流动合作倡议》指出，当前全球数据跨境流动实践中存在涉及国家安全、公共利益、个人隐私以及知识产权等风险，我国主张尊重各国的政策法规，通过负面清单等方式构建数据流通体系，同时反对数据问题泛安全化和非法获取数据的行为。⁸¹当前，数字丝绸之路范围内存在 RCEP、

⁷⁵ See *DeepSeek, Huawei, Export Controls, and the Future of the U.S.-China AI Race*, CSIS(07March 2025), <https://www.csis.org/analysis/deepseek-huawei-export-controls-and-future-us-china-ai-race>

⁷⁶ 参见戴淑婷、宋黎磊：《欧美科技制裁协作的动力与分歧：基于欧盟的视角》，载《国际安全研究》2023 年第 5 期，第 81-107+159 页。

⁷⁷ 参见宫云牧：《数字时代主权概念的回归与欧盟数字治理》，载《欧洲研究》2022 年第 3 期。

⁷⁸ See *Italy's regulator blocks Chinese AI app DeepSeek on data protection*, Reuters(05 February 2025), <https://www.reuters.com/technology/artificial-intelligence/italys-privacy-watchdog-blocks-chinese-ai-app-deepseek-2025-01-30/>.

⁷⁹ 参见《从 DeepSeek 海外监管挑战看 AI 企业出海合规》，载四川省数字贸易网，<https://www.tradesichuan.com/jmzx/1281.html>，2025 年 04 月 09 日访问。

⁸⁰ 《2024 中国开源年度报告》，载 github，<https://kaiyuanshe.github.io/2024-China-Open-Source-Report/preface.html>，2025 年 03 月 28 日访问。

⁸¹ 《全球数据跨境流动合作倡议》，载中国网信网，https://www.cac.gov.cn/2024-11/20/c_1733706018163028.htm。

DEPA、CPTPP 等多套规则，总体上这些规则都是对数据跨境流动持鼓励态度的，但是在具体监管的程度和范围上不尽相同。相对而言，RCEP 的安全例外条款较为保守。我国可以在一定范围内推动 RCEP 等区域协定向 DEPA、CPTPP 等标准对接，将数据重新进行量化评估，在 RCEP 等区域协定当中推动对于数据敏感程度认定标准的建立。考虑到越南等国对于数据流动较为保守的国内政策，可以将不同领域的数据进行分别评估，对于国内公共数据采取谨慎保守的标准，对于个人数据则可以根据自愿原则进行适当放宽标准。除了数据流动之外，还可以在电信服务方面降低互联互通的成本，例如对于 RCEP 当中适当降低对于服务提供者与领土范围的绑定，以促进分布式架构等部署模式的推广。除了推动区域协定的升级之外，还应当及时更新 FTAs 等数字经贸条约。增添围绕数据流动、人工智能准入等方面的内容。⁸²为了保持与各国之间的标准规则对接顺利，可以主导制定合作规则框架。⁸³在数据流动等领域达成共识，在此基础上进行细化标准的制定。

（二）开源人工智能技术与安全标准同步推广

CSIS 的一篇报告指出，DeepSeek 具有较高的开源程度，并且与 Open AI 和 Anthropic 等公司推出的人工智能相比，缺乏安全防护措施。并且将此问题上升到政治安全层面，将 DeepSeek 意识形态化，渲染“中国威胁论”和“数字威权主义”。⁸⁴为了应对西方对于我国开源人工智能技术的恶意渲染，我国应当构建起数字丝绸之路中的中国国际话语体系，化解来自西方的话语攻势。⁸⁵在开源人工智能治理领域，中国应当以技术上的领先优势作为推动统一安全标准构建的基础。安全性的提升能够为国产开源人工智能带来更多话语权。⁸⁶DeepSeek 等开源人工智能更易于监管，在透明度合规方面具有巨大的优势。因此，应当将重点放在部署和应用层面。

具体做法上，一方面，可以从软法出发，推广我国《人工智能大模型私有化部署技术实施与评价指南》等国内行业标准国际化，为其他国家制定相关标准法规提供样本。另一方面，还应当推动建立对区域内开源社区的管理，推动多元共治局面的形成。

（三）建立以建设人工智能能力为内容的长效机制

人工智能能力分为技术能力和治理能力等多个方面。我国在《人工智能能力

⁸² 参见何志鹏、周萌：《共建“数字丝绸之路”的中国角色、挑战及应对》，载《东北亚论坛》2024 年第 6 期。

⁸³ 参见郭柚坊、陈咏梅：《论“数字丝绸之路”数字合作规则的构建》，载《湖北大学学报（哲学社会科学版）》2025 年第 2 期。

⁸⁴ See Delving into the Dangers of DeepSeek, CSIS(24 February 2025), <https://www.csis.org/analysis/delving-dangers-deepseek>.

⁸⁵ 参见张起铭：《数字丝绸之路的国际认知分歧及其根源、影响与对策》，载《世界社会主义研究》2024 年第 11 期。

⁸⁶ 参见段玉聪：《抢占 AI 话语权：DeepSeek 的技术优势、战略布局与未来生态图景》，载《新疆师范大学学报（哲学社会科学版）》2025 年第 4 期。

建设普惠计划》中提出未来我国的行动方向，通过对这些方向进行分析，长效机制仍然围绕“硬联通”和“软联通”两方面内容进行制定。⁸⁷

硬联通角度，我国可以依托开源人工智能构建围绕数字基础设施领域的合作平台，创建模块化的合作机制，以应对与数字丝绸之路不同国家开展合作的需求。例如面对中东地区存在较大地区鸿沟的情况，可以重点开展网络基础设施等人工智能相关配套设施和建设；对于中亚地区紧邻中国的地理位置，则可以开展关于跨境贸易相关人工智能应用的部署。⁸⁸同时在更新合作双边或者多边合作机制时，应当明确技术援助以及知识产权开放等保障开源人工智能顺利部署和使用的条款。

软联通角度，针对沿线国家治理水平差异较大的现状，为了弥合治理领域的“数字鸿沟”，可以明确化治理能力有待提升国家的过渡措施。⁸⁹给予这些国家与区域较高标准对接一定的缓冲期和“能力输入”到“体系自主”一定的转化空间。

而面对开源人工智能技术迭代速度快的情况，相关规制框架也要保持动态更新。可以在特定区域和领域建立“沙盒机制”。⁹⁰对于一些具有一定风险的领域可以建立“监管沙盒”，探索构建更加具体的监管制度，平衡技术创新与规制。在确保监管制度有效的情况下进行人工智能在更大领域和范围的推广。

六、结语

作为“一带一路”倡议在数字领域的延伸，数字丝绸之路无疑是实现数字治理的重要平台。国产开源人工智能通过降低技术成本和本地部署的门槛，为破解数字丝绸之路沿线国家发展和多领域互联互通的困境提供了全新路径。2025年4月，美国在全球范围内掀起新一轮的关税战，可以预测到全球贸易将会面临更加紧张的局面。在单边主义抬头的国际环境下，全球在人工智能等数字领域的合作必然也会受到影响。这既是风险也是机遇。数字丝绸之路的制度优化不仅担负着技术合作的功能，更承担着构建普惠包容的全球数字治理秩序的使命。中国应当继续秉持开放包容的态度，让数字技术发展的成果惠及更多国家，为超越零和博弈，为促进人类发展提供新的动能。

⁸⁷ 参见《人工智能能力建设普惠计划》，载外交部2024年09月27日，https://www.mfa.gov.cn/web/wjbx_673089/xghd_673097/202412/t20241218_11496414.shtml。

⁸⁸ 参见韩建伟：《中东数字鸿沟问题与“数字丝绸之路”的高质量发展》，载《国外社会科学》2022年第4期；王海燕：《中国与中亚国家共建数字丝绸之路：基础、挑战与路径》，载《国际问题研究》2020年第2期。

⁸⁹ 参见郭柚坊、陈咏梅：《论“数字丝绸之路”数字合作规则的构建》，载《湖北大学学报（哲学社会科学版）》2025年第2期。

⁹⁰ 参见杨丰一：《人工智能“监管沙盒”：理论溯源、实践考察与制度构建》，载《中国行政管理》2024年第12期。

全球人工智能治理的软法机制

王玉婷¹

摘要：人工智能技术的快速发展引发全球性治理挑战，传统国际法因滞后性与僵化性难以应对人工智能技术的动态风险。软法凭借灵活性、适应性和包容性，成为全球人工智能治理的核心机制。软法能通过形式上的非约束性与实质上的合法性、有效性，调和各国利益分歧，塑造共同价值。全球人工智能治理的软法模式体现为：以伦理原则为基础，通过联合国教科文组织《人工智能伦理建议书》等框架凝聚共识，平衡创新与安全；二是依托技术标准，借助市场激励与供应链机制推动合规与安全；三是加强国际合作与能力建设，弥合技术鸿沟。中国参与路径需注重发展共同利益、提升造法主体代表性和构建软硬法协同机制。当前软法治理面临碎片化、执行力弱化及大国博弈等挑战，未来需整合多层次规范体系，促进多元主体协作，平衡主权与全球公共利益。软法作为动态协商工具，为技术快速迭代中的秩序构建提供了可行路径，但其长效治理效能依赖于国际社会的持续对话与制度创新。

关键词：全球人工智能治理；软法；《全球人工智能治理倡议》；国际造法

一、引言

人工智能技术是新一轮科技革命和产业变革最具代表性的颠覆性技术，给人类社会带来巨大发展机遇的同时，也带来了诸多全球性挑战。²由于技术复杂性高、涉及议题领域广泛且行为体众多，若缺乏明确的规则和秩序，各方在人工智能伦理、规范及安全领域的诸多分歧难以化解，将引发更多、更复杂的冲突。³由此，人工智能治理成为全球治理的新兴议题，全球主要国家和地区纷纷将人工智能规制纳入正式立法议程。⁴

国际法是全球人工智能治理的重要依据，然而当前的国际造法难以跟进人工智能的带来的新风险。⁵一方面，国际法规则本身具有滞后性，难以迅速对技术发展作出回应；另一方面，人工智能迭代更新给现行国际法概念、原则、体

¹ 中国社会科学院大学学生。

² 《全球人工智能治理研究报告》，2024 年，
<https://cn.wicintenet.org/site395/20241122/f44d305ea1af28aa58d809.pdf>，访问时间：2025 年 1 月 27 日，第 1 页。

³ 鲁传颖：《全球人工智能治理的目标、挑战与中国方案》，《当代世界》2024 年第 5 期，第 25 页。

⁴ 如欧盟于 2024 年正式出台《人工智能法》，美国于 2023 年发布《人工智能风险管理框架》《关于安全、可靠和值得信赖地开发和人工智能的行政命令》，中国分别于 2023、2024 年出台《全球人工智能治理倡议》《生成式人工智能服务管理暂行办法》、《人工智能安全治理框架》1.0 版等。

⁵ 阎德学：《全球人工智能治理：联合国的角色》，《国际问题研究》2024 年第 4 期，第 13 页。

系造成不同程度的冲击。在对人脑结构和意识产生机制进行高度模拟的基础上,人工智能已具有自学习、自适应、自组织等特性,能够对周边的复杂环境和动态变化进行主动适应⁶,这意味着人工智能技术可能不断产生新的、多样的风险。由于软法的灵活性、适应性等特点契合人工智能敏捷性治理的需求,当前的全球人工智能治理总体上呈现“软法之治”的面貌。

人工智能治理中的软法议题已受到国内外学界的关注,既有研究呈现四种分析近路。一是,从实证法视角分析当前全球人工智能治理中的软法的形态、作用、及其存在的缺陷和完善方案⁷;二是,从技术发展与软法的关系入手,对软法在技术治理方面的实践经验进行研究,包括环境技术、纳米技术、信息通讯、生命科学等,并以此为人工智能治理提供借鉴⁸;三是对某一类型的软法工具在全球人工智能治理中的生成与发展进行分析,阐明其作用及实施机制。⁹既有研究勾勒出当前全球人工智能治理碎片化的治理模式,肯定了软法在人工智能治理中发挥的作用,并对全球人工智能治理的未来发展展开丰富的想象。但同时,既有研究一方面对“软法”这一基础概念的内涵莫衷一是,未能厘清软法在国际法层面的理论意义;另一方面对全球人工智能治理中的软法框架只作描述性分析,未能从体系或结构层面剖析其中的效力逻辑与运行机理,致使全球人工智能治理中的软法研究仍呈现模糊的面貌。鉴此,本文首先厘清软法的内涵与效力来源,肯定软法与全球人工智能治理的契合性,以更全面深入地进一步对全球人工智能治理中的软法工具进行类型化分析,阐明不同软法发挥实质效力的机制,最后从加强涉外法治建设的视角出发试图提出中国参与全球人工智能治理的有效路径。

二、全球人工智能治理中的软法

学界很早就注意到科技发展与国际法的关系。科技的发展深刻影响当代国际造法,其不仅加速国际习惯法的造法进程,而且催生了非约束性和约束性文件之间的过渡形式,同时创造了更多共同利益,促进国际共识与合作。¹⁰这些过渡形式被视为国际软法。传统上认为,软法作为国际造法的过渡形式,主要发挥“中

⁶ 张劲松:《人是机器的尺度——论人工智能与人类主体性》,载《自然辩证法研究》2017年第1期,第52页。

⁷ 叶淑兰 李孟婷:《全球人工智能治理:进展、困境与前景》,载《国际问题研究》2024年第5期; Mark D. Cole, AI Regulation and Governance on a Global Scale: An Overview of International, Regional and National Instruments; Olivia J. Erd'elyi & Judy Goldsmith, “Regulating Artificial Intelligence Proposal for a Global Solution”, (2022) 39 Government Information Quarterly 1.

⁸ Carlos Ignacio Gutierrez, Gary Marchant & Lucille Tournas, “Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance”, (2020) 61(1) Jurimetrics 133; Coglianese, C, “Environmental Soft Law as a Governance Astrategy”, (2020) 61(1) Jurimetrics 19; Bowman, Diana M, “The Role of Soft Law in Governance Nanotechnologies”, (2020) 61(1) Jurimetrics 57; Thierer, Adam, “Soft Law in U.S. ICT Sectors: Four Cases Studies”, (2020) 61(1) Jurimetrics 79; Stevens, Yvonne A, “Soft Law Governance: A Historical Perspective From Life-Science Technologies”, (2020) 61(1) Jurimetrics 121.

⁹ José-Miguel Bello y Villarino, “Global Standard-Setting for Artificial Intelligence:

Para-regulating International Law for AI?”, (2023) 41 The Australian Year Book of International Law 157.

¹⁰ Manfred Lachs, “Thoughts on Science, Technology and World Law”, (1992) 86 American Journal of International Law 673, p.694.

转”价值，如促进有约束力的条约的达成或提供国际习惯法形成的证据。¹¹但随着全球治理的兴起，软法在全球治理中发挥日益重要的作用，如在全球金融治理领域，软法并不当然是造法的“过程”，也有可能是造法的“结果”，能够凭借自身特点和优势长期稳定存在并独立发挥作用。¹²

尽管在理论上软法的内涵与效力不甚明确，软法在全球人工智能治理中的使用仍十分普遍。这一方面是因为软法的灵活性特点与人工智能技术快速迭代、自适应、不确定性相契合，能够有效平衡安全与创新两方面价值；另一方面是因为当前全球人工智能治理呈碎片化特点，各主权国家难以对在议题形成具体而深入的共识，软法能够在最低限度调和各国间利益冲突，促进共同理解，塑造共同利益与价值，进而推进全球人工智能治理朝着公正、合理的方向发展。

（一）软法的界定与内涵

相较于硬法，软法的概念更为模糊和复杂，学界至今未能形成对软法的统一的、明确的权威的定義。部分学者对软法的概念持批判态度，认为软法对国际规范体系的稳定性和明确性构成挑战，主张在国际法层面应坚持法与非法的二元区分。¹³软法这一概念起源于国际法，在国际法和国内法层面均得到广泛使用，但两个层面的用法亦有较大差别，如一些国内法学者认为国际法本身就属于软法。¹⁴本文主要在国际法层面探讨软法的内涵与效力。

关于软法的内涵，既有研究有三种代表性的观点。最常使用的软法概念是“不具有法律约束力，却能够产生法律效果的法”。¹⁵此种观点遵循的是一种二分法模式，即软法与硬法能够进行明确区分，相对而言，同时具有法律约束力和法律效果的法就是硬法。而另外两种观点认为软法与硬法并不是绝对的概念，而是相对的。第二种观点认为，硬法和软法之间并无清晰的分界线，“硬”和“软”是针对相关规范的法律化（legalization）的程度而言的。¹⁶而此种法律化程度由义务性（obligation）、精准性（precise）和授权性（delegation）三方面确定。¹⁷据此，提供具有约束力的义务，措辞精确，并在法律实施过程中提供某种形式的授权的规范是硬法，而在任一维度上被弱化的规范就是软法，且弱化的程度越高，法律就

¹¹ Alan Boyle & Christine Chinkin, *The Making of International Law*, Oxford University Press, 2007, p.218-219.

¹² 廖凡：《论软法在全球金融治理中的地位和作用》，《厦门大学学报（哲学社会科学版）》2016年第2期。

¹³ See Jan Klabbers, “the Redundancy of Soft Law” (1996) 65 *Nordic Journal of International Law* 167; Prosper Weil, “Towards Relative Normativity in International Law?” (1983) 77 *American Journal of International Law* 413.

¹⁴ 姜明安：《软法的兴起与软法之治》，载《中国法学》2006年第2期，第26页。

¹⁵ Francis Snyder, “Soft Law and Institutional Practice in the European Community”, In Stephen Martin (eds) *The Construction of Europe* (1994), Springer, Dordrecht.

¹⁶ 肯尼斯·艾尔伯特提出法律化（legalization）标准，据此，法律化的判定应考虑三个要素：义务性（obligation）、精准性（precise）和授权性（delegation）其中，义务性是指国家或其他行为体被特定的规则和承诺所约束，这意味着特定的规则或承诺具有法律约束力，即其行为要受到国际法原则和话语体系的规范；精准性是指规则清晰地定义其所要求、授权和规制的行为；授权性是指第三方被授予实施、解释和适用规则，解决争议的权力。

¹⁷ Chris Ingelse, “Soft Law”, (1993) 20 *Polish Yearbook of International Law*, 75.

越“软”。¹⁸第三种观点由联合国法律事务厅首席法律干事阿诺德·普龙托（Arnold N. Pronto）提出，他认为硬法和软法的界定应从形式和实质作区分，如此，软、硬法的界定最终导致 4 种类型的法律，即形式与实质上都“硬”的法、形式与实质上都“软”的法、形式“硬”实质“软”的法以及形式“软”实质“硬”的法。¹⁹这一界定模式中，形式上的标准是清晰的，即相关法律在形式上是否具有法律约束力，更具体而言，法律是否属于《国际法院规约》第 38 条规定的国际法渊源。但实质标准相对模糊，依据阿诺德·普龙托所列举的四种法律类型，大概可以推断实质标准包含两方面内容，一是条文的措施是否具体、详细，足以指导相关主体的行为，二是该法是否在国际司法机构中得到实质的适用或被在后续的发展中被采纳为正式的法。

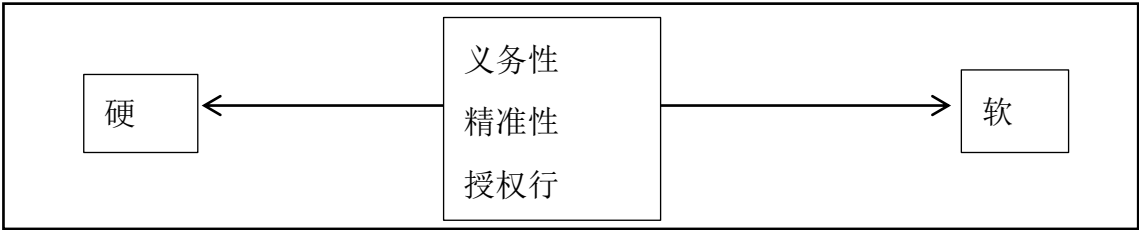


表 1:通过法律化程度界定法的“软”和“硬”程度

形式 实质	硬	软
硬	（硬，硬）	（软，硬）
软	（硬，软）	（软，软）

表 2:从实质和形式两方面界定硬法和软法

要厘清软法的真正内涵，需将此三种软法概念拆解，并将其中的核心要素相互联系起来理解。第一种观点的两个核心要素为法律约束力和实质效果，其中法律约束力与法律化中的义务性属同一要素，实质效果的性质与含义却具有模糊性，这一界定模式的缺陷也在于此。第二种观点反映了形式主义法律观，即只关注法律的形式要素，忽略了法律的实质要素，而软法的出现在很大程度上是为了弥补国际法发展中的实质正义赤字，因而此种界定方式是狭隘且片面的。此外，精确

¹⁸ Sylvia I. Karlsson-Vinkhuyzen & Antto Vihmap, “Comparing the legitimacy and effectiveness of global hard and soft law: An analytical framework”, (2009) 3 Regulation & Governance 400, p.402.
¹⁹ Arnold N. Pronto, “Understanding the Hard/Soft Distinction in International Law”, (2015) 48 Vanderbilt Journal of Transnational law 941.

性和授权性标准显得冗余且容易引起混乱,即使只看形式渊源,能够满足法律化标准的似乎只有表述清晰、建立相应争端解决机制的国际条约,而习惯、一般法律原则等国际法渊源在这一界定方式中仍属于软法,这无疑是有悖于常理。因此,准确性和授权性方面的弱化应被视为软法所具有的一般性特征,而不宜作为软法的界定标准。第三种观点的形式标准仍集中于义务性这一要素,但诡吊的是,究其实质要素的本质,似乎仍未摆脱形式主义的窠臼,条文的详细程度体现准确性要素,被司法机构实质性的适用体现授权性,后续被采纳为正式的法体现义务性,或具备义务性的潜质。

综合而言,应从形式和实质两方面理解软法的内涵。一是,软法在形式上不具有法律拘束力,即不产生直接的权利与义务,违反软法规范的行为不会产生法律责任,在此意义上软法的遵行具有体现一定程度的自愿性。二是软法具有一定的法律效力,这一效力一方面体现为软法反映实质正义,即实质合法性(legitimacy),另一方面体现为软法得到了一定的遵行,或产生了预期的效果有利于实现治理目的,即有效性(effectiveness)。同时,软法在形式上一般具有低准确性、低授权性特征。

(二) 软法的效力结构

前已基本述明软法的形式结构与特征,其在实质层面的效力结构仍需说明。法律效力涉及到法律的本原问题。对此,实证法学派强调主权国家的同意是国际法的唯一效力来源。国家同意作为国际造法的唯一合法性来源在全球治理背景开始面临压力,特别是像国际公共机构或非国家实体这样的其他行动者扮演着越来越重要、不可或缺的角色。²⁰尤其是二战后国际强行法的确立和发展表明高于个体国家意志的国际社会公共利益作为国际造法之基础的观点愈加得到重视,这被称为国际造法的“立法转向”。国际造法的“立法”模式认为国际法规则的形成无需每个国家的同意,而是需要国际社会的集体同意,集体同意的形式包括协商一致、多数表决。²¹

国际造法的“立法”模式同样重视国家主权和意志对国际法的基础性意义,但其同时认为具有约束力的法律需要一个更高层次的规范,即一种基础规范(Grundnorm),从中导出法律的约束力。现代实证主义理论认为这一基础规范即为由国家组成的国际社会(international community of states)。²²在现代国际法体系中,国际强行法的存在和发展是对传统的以国家同意为基础的国际造法模式构成了补充和修正。这意味着,在制定国际法规范时,除了考虑国家的同意外,还

²⁰ Matthias Goldmann, “Soft Law and Other Forms of International Public Authority – The View from Discourse Theory: A Reply to Jaye Ellis”, (2012) 25 Leiden Journal of International Law, 373.

²¹ 黄蓬北:《论国际造法的“立法”转向——基于普遍意志的契约论解释》,《清华大学学报(哲学社会科学版)》2025年第11期,第249页。

²² Matthias Goldmann, We Need to Cut Off the Head of the King: Past, Present, and Future Approaches to International Soft Law, (2012) 25 Leiden Journal of International Law, 335, p.343.

需兼顾国际社会的普遍利益和共同价值，从而确保国际法的公正性和有效性。国家是否遵守国际法并不总是取决于法律是否具有强制性，是否存在违法的制裁措施。有研究表明，法律化程度并不能解释规范的有效性。艾伦·卢茨（Ellen Lutz）和凯瑟琳·西金（Kathryn Sikkink）在人权领域的规范的法律化与有效性之间的关系，研究发现在“法律化”程度最高的领域，即酷刑，遵守程度最低；而在“法律化”程度最低的领域，即民主治理，履约程度最高。²³

软法的生成和发展是国际造法模式转变趋势下的产物。首先，软法至少代表了一定程度的国家同意，尽管与有约束力的硬法相比，这一同意是不成熟的。其次，软法是国际社会共同利益与价值的体现，并在其实施和发展过程中不断地塑造共同利益与共同价值。最后，软法向硬法的转化不是软法效力的来源，而是软法具有效力的证明，即因为软法反映了某种正当性因而获得了完全的国家同意。

自然法学派认为法律的效力来自其体现的正义与道德原则，即合法性；社会法学派则认为法律的效力本质上是法律实施的实际效果，即实效性或履约性。软法若不抵触硬法或硬法原则、精神，又大致上符合一定范围内社会对更好的“公共善”的认知和期待，就具有了“有效性”，就会对应地产生说服的约束力。²⁴软法的效力包含两层结构，一是合法性，二是有效性，二者相互联系和作用。

1. 合法性

合法性是一种普遍的认知，即一个实体的行为受正当且理性的理念、价值、规范的约束。²⁵法律性与合法性的区别在于，前者为实体规范的外在抽象载体，或者是规范得以产生的形式或程序，与程序正义相关，而后者为实体规范的具体内容，体现实质正义。²⁶可以说，软法是为了实现实质正义的一种特殊的立法模式。肯尼斯·艾尔伯特的法律化标准集中体现了新自由制度主义国际关系学者对国际法的认识，他们认为，法律的主要作用在于可以提供预测的规则和稳定的制度结构。²⁷该观点引发了建构主义学者的批评。批评者的质疑集中于两个方面，首先，此种法律观他们的法律观只体现了自由主义、实证主义的观点，是狭隘的、西方化的。²⁸该标准所定义的硬法，即通常意义上的法律，主要通过案例和法院，或者通过正式的条约谈判来构建，而法律适用的过程被普遍视为解决争议的过程，其大多是在正式的制度化背景下进行，法律的效力完全来自于条约所强加的明确义务。在这种观点中，法律仅具有约束作用，它在社会生活中没有创造或生成的

²³ Finnemore M, Toope SJ, Alternatives to “Legalization:” (2001) 55(3) *Richer Views of Law and Politics*. *International Organization* 743, p.755.

²⁴ 沈岩：《论软法的有效性与说服力》，《华东政法大学学报》2022年第4期，第104页。

²⁵ Mark C. Suchman, *Managing Legitimacy: Strategic and Institutional Approaches*, *The Academy of Management Review*, 1995,20(3), p.574.

²⁶ 江河：《从大国政治到国际法治：以国际软法为视角》，《政法论坛》2020年第1期，第55页。

²⁷ Jutta Brunnée & Stephen J. Toope, *Legitimacy and Legality in International Law: An Interactional Account*, Cambridge University Press, 2010, p.11.

²⁸ Finnemore M, Toope SJ, Alternatives to “Legalization:” (2001) 55(3) *Richer Views of Law and Politics*. *International Organization* 743, p.745–746.

力量,然而,现实生活世界中运行的法律既构成关系也界定可接受的行为。同时,建构主义学者们主张将法律置于更为广泛的社会背景,关注产生特定身份与行为的文化因素。²⁹

建构主义法律观认为,法律规范根植于社会群体的共同理解。只有社会发展产生了对规范的共同需求,且对用于指导行为的特定规范的需求有共同理解时,社会规范才会产生。³⁰规范发起者和知识共同体通常在推广规范方面具有影响力,但规范的有效创制并非单向的思想或价值观投射,而依赖于群体性的共同理解。此种共同理解本质上是互动性的,通过实践形成的,这些共同体塑造了国际社会各种行为者之间的共同参与治理的行为。这种参与由现有的规范和社会结构所框定,而行为者的积极参与反过来又维持和重塑了规范与结构,共同理解可能仅仅包括对社会中法律需求的基本接受,也可能具有实质内容且负载价值。

软法具有精密的道义结构,它是可以概括国家间共识的简洁公式,在一定程度上反映了国际社会的共同利益与价值。³¹Braithwaite 和 Drahos 提出以“对话网络(Webs of Dialogue)”为主的效力机制。对话网络发挥效力的机制为:首先设定各方关切的议题,营造各方互动和合作的环境,使得复杂的相互依赖和问题联系能够促进达成协定,其次形成规范性承诺并使遵守习惯制度化;最后对违约行为奖赏和谴责加以制度化。³²软法在共同利益与价值的基础之上提供了一个各国交流、对话、合作的平台,构建了相应谈判程序与法律框架,能在相互对话与实践互动中增进对某一议题的共同理解,通过构建并深化“对话网络”促进实现治理目标。

2. 有效性

有效性与合法性时常相伴相生。一般而言,规范的有效性从两个维度进行评价,一是解决问题的有效性,某个规范的设计与采纳是为了特定的治理目标,那么规范的有效性自然体现为其实现治理目的的程度;二是行为的有效性,即国家等相关主体在多大程度上遵守该规范。³³对于软法而言,解决问题的有效性更多体现在治理过程的维持与推进。与硬法以完备的形式确立稳固的合作关系不同,软法注重关系与过程,在对话与实践过程中谋求合作,类似于秦亚青教授提出的关系治理,关系治理以过程为取向,在这里,过程绝不仅仅是个体行为体互动和交往的背景或是平台,而是具有本体地位,是治理行为的内在动力。在某种意义上讲,维护了过程就是维护了治理,使过程在符合预期的轨

²⁹ Finnemore M, Toope SJ, Alternatives to “Legalization:” (2001) 55(3) *Richer Views of Law and Politics. International Organization* 743, p.755.

³⁰ Jutta Brunnée & Stephen J. Toope, *Legitimacy and Legality in International Law: An Interactional Account*, Cambridge University Press, 2010, p.56-87.

³¹ 曹伟:《国际环境法造法机制研究》,载《中国人民大学学报》2016年第5期,第115页。

³² Braithwaite J & Drahos P, *Global Business Regulation*. (2000) Cambridge University Press, p.556.

³³ Sylvia I. Karlsson-Vinkhuyzen & Antto Vihmap, “Comparing the legitimacy and effectiveness of global hard and soft law: An analytical framework”, (2009) 3 *Regulation & Governance* 400, p.405.

道上运行本身就是治理有效性的体现。³⁴

行为的有效性按主体划分,可分为国家的遵行和非国家主体的遵守。国家的遵守体现为国家在软法规定的原则、规则、建议等为指导行为,一般通过国内法律、政策施行。如巴塞尔委员会制定发布的巴塞尔协议等一系列文件的内容被包括中国在内的众多国家不同形式、不同程度地转化为国内立法,为各国银行业监管者所遵行。³⁵至于国家遵守软法的原因,通说认为国家利益与国际法所体现的道德观念是其得以遵守的主要原因。³⁶从建构主义视角来看,后者具有根本性的地位。因为尽管国家利益可以驱动国家行为,但国家利益并非一成不变,而是由国家身份决定,国家身份由国家共有的知识与价值建构。³⁷有效性其至少包含两个要素:一是决策程序的内生性问题,在国际法律框架内,决策过程本身是国际行为体开展重复博弈的过程,这一过程基本确立了理解与共识,因此规范的实施和遵守会得到基本的保证;二是对国际行为体需要在基本共识的基础上持续地学习和互动,以提升合作的广度和深度。³⁸

非国家行为体(私主体)的遵守需从两个层面来理解。一是,当国际软法被国家实质性地转化为国内法时,处于该国管辖范围内的相关私主体实际上具有遵守该法的法律义务。二是,当国际软法未被国家转化,或国家以政府或相关公共机构的名义发布相应的软法实施时,相关私主体不承担法律义务,在一定程度上,其拥有遵守或不遵守该软法的自由。以人工智能治理中最主要的主体——科技企业为例,企业落实软法的途径包括:人事架构、组织架构、员工培训、产品与服务设计、指标设计、绩效激励、资金分配等。³⁹

(三) 全球人工智能治理的规范需求与软法的契合性

全球治理是指在缺乏主导性政治权威却又相互依赖的全球化背景下,为了应对全球性问题和追求人类共同利益,包括非国家行为体在内的国际行为主体,通过各种强制性的正式管理或规制以及非正式的社会化倡议、公共程序或机制,来避免全球风险和追求可预见的、安全的社会秩序的行为。⁴⁰循此逻辑,全球人工智能治理可界定为,在缺乏统一国际权威的数字化时代背景下,为应对人工智能技术发展引发的伦理、安全等共同风险,由主权国家、国际组织、科技企业、公民社会等多元行为主体,通过各种正式或非正式的机制实现技术可控性、权利保

³⁴ 秦亚青:《世界政治的关系理论》,上海人民出版社,2021,第432页。

³⁵ 廖凡:《论软法在全球金融治理中的地位和作用》,《厦门大学学报(哲学社会科学版)》2016年第2期,第25页。

³⁶ 参见韩永红:《国际法何以得到遵守——国外研究述评与中国视角反思》,载《环球法律评论》2014年第4期。

³⁷ 参见秦亚青:《国际政治的社会建构——温特及其建构主义国际政治理论》,载《欧洲研究》2001年第3期。

³⁸ 参见任洪生:《国际关系与国际法研究:历史、理论与议程——兼论构建中国全球治理自主知识体系》,载《人民论坛·学术前沿》2025年第1期,第19页。

³⁹ Gary E. Marchant & Carlos Ignacio Gutierrez, "Soft Law 2.0: An Agile and Effective Governance Approach for Artificial Intelligence", (2023) 24(2) Minnesota Journal of Law, Science and Technology 375, p.395.

⁴⁰ 江河:《从大国政治到国际法治:以国际软法为视角》,《政法论坛》2020年第1期,第53页。

障和风险预防的持续性协调过程。

人工智能迫切需要全球治理是因为：一方面，人工智能产生的外部效应跨越国家边界，需要各国共同努力。作为新一轮科技与产业变革的重要力量，人工智能日趋成为各国竞争力角逐的主赛道，在监管与创新的矛盾关系中，监管宽松的国家在开发某些人工智能应用时可能会获得竞争优势，因而造成“逐底竞争”的风险。全球治理有益于为人工智能发展创造公平竞争的国际环境符合国际社会共同的利益。另一方面，人工智能的发展是一个跨国界的过程，跨国公司是主要的行为者，其可以在多个国家的司法管辖区内行为，只有引入共同规则，各国才能确保人工智能企业面临相似的监管环境，这能在促进人工智能跨国界发展的同时减少“监管套利”行为。⁴¹

沈岍教授将国内人工智能治理采取软法的原因总结为五点：一是软法可灵活应对人工智能技术的快速发展；二是软法适合解决人工智能带来的多样化挑战；三是软法能够为人工智能利益相关者的合作创造选择实验的空间；四是软法在事实上产生的压力使其具有一定的效力；五是软法更容易跨国适用。⁴²这五个原因更多是从技术维度阐明软法与人工智能治理的契合性，同样适用于全球人工智能治理。不过，全球人工智能治理涉及到不同国家间的竞争、博弈与合作，在国际层面采取软法有其特殊原因。

第一，针对人工智能治理的谈判主体的多元性以及巨大的利益分歧使得通过条约来实现国家的合意的成本与难度大幅度提高，尤其是人工智能技术具有高度复杂性、需要最高层次的协同行动，目前来看，各国不愿承担具有法律约束力的义务，使得迅速通过缔结条约来进行造法变得困难重重。第二，人工智能加剧技术鸿沟，容易产生技术霸权，而软法能够降低制度成本，容纳差异性，具有包容性、适应性特征，能一定程度上消解西方国家的技术霸权，推进全球人工智能治理当前，以美国为中心的西方发达国家仍然主导着全球治理进程，同时占据技术优势的美国利用人工智能实行技术垄断、技术封锁、脱钩断链，以保证其持续获得大量的财富和利益，巩固其在全球经济中的支配地位，侵蚀国际合作根基，威胁全球产业链、创新链、供应链的安全稳定，将给世界经济带来巨大的风险和不确定性。⁴³第三，国际组织、科技企业等非国家行为体在全球人工智能治理中发挥着重要作用，软法为非国家行为体参与治理的提供了切实有效的途径。总之，软法能有效回应全球人工智能治理的现实需求，通过一种全球协商机制稳步调节智能化发展的方式与进程。

⁴¹ Tallberg, Jonas et al., “The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research”, (2023) 25 International Studies Review 1, p.6.

⁴² 沈岍：《论软法的实施机制——以人工智能伦理规范为例》，《财经法学》2024年第6期，第116-117页。

⁴³ 李猛：《人类命运共同体视角下人工智能风险全球治理的国际法规制路径探究》，载《宁夏社会科学》2024年第2期，第117页。

三、全球人工智能治理的软法模式及其实施机制

全球善智应该包含两点内涵：一方面，人工智能的发展要进一步增强全球社会中的平等和正义。另一方面，善智的目的是发展良善的人工智能，这意味着人工智能的发展要推动全球性问题的解决，而不是加剧全球性问题。⁴⁴

为了尽可能全面有效和可信，软法必须直接面对其根本弱点。迄今为止，大多数采用和实施软法的努力都集中在识别负责任发展人工智能的关键实质性原则。原则很重要，因为它们设定了人工智能治理试图实现的实质性目标。如果没有关注这些目标，即软法原则所关注的，人工智能治理将会失去方向且效率低下。软法原则是必要的，但并不足够，所需的还有顺利实施原则的过程和机制，以解决软法的根本弱点。⁴⁵

（一）伦理原则：通过塑造共同价值提供治理基础

以伦理价值为导向的原则性治理是当前全球人工智能治理主要的治理机制。国际社会通过多边倡议、区域立法、行业准则和伦理框架等形式，构建起初步的治理体系。在多边层面，联合国教科文组织于 2021 年出台《人工智能伦理问题建议书》，是首个全球性人工智能伦理框架，确立了透明度、问责、隐私保护等 11 项治理原则，鼓励各国根据本国国情制定具体的人工智能伦理规范。在区域层面，经济合作与发展组织（OECD）于 2019 年提出的《关于人工智能的经合组织原则》以“可信人工智能”为基础，提出五大核心原则，包括包容性增长、可持续发展和增进人类福祉；尊重人权、民主与法治；透明性与可解释性；稳健性、安全性与可靠性；问责制。这些原则已被二十国集团（G20）采纳为国际共识，并成为全球 AI 治理的重要参考框架，成为美、欧、日等主要经济体的政策基础。此外，国际社会就全球人工智能治理开展广泛合作，“G20 人工智能原则”强调包容性、可持续性和以人为中心的人工智能发展。2020 年 6 月，由加拿大、法国等 G7 国家牵头成立全球人工智能伙伴关系（GPAI），截至 2023 年其成员扩展至 29 个。该平台的核心目标是：促进人工智能技术的负责任发展，平衡创新与伦理风险；推动人工智能国际合作，弥合发达国家与发展中国家的技术鸿沟；为政策制定者提供科学依据，支持全球人工智能治理规则构建。全球人工智能安全峰会聚焦人工智能安全风险与治理，以风险识别、探索国际监管规则，平衡创新与安全以及国际合作机制为核心议题。首届峰会的最重要的成果是达成《布莱切利宣言》，获中美英等 28 国签署，标志着国际社会首次就人工智能安全风险取得国际共识。

值得一提的是，尽管从形式上看，欧洲委员会制定、已获美国、英国和欧盟等签署的《人工智能与人权、民主和法治框架公约》具有法律约束力，但就实质

⁴⁴ 高奇琦：《全球善智与全球合智：人工智能全球治理的未来》，《世界经济与政治》第 39 页。

⁴⁵ Gary E. Marchant & Carlos Ignacio Gutierrez, “Soft Law 2.0: An Agile and Effective Governance Approach for Artificial Intelligence”, (2023) 24(2) Minnesota Journal of Law, Science and Technology 375, p.402.

而言，该公约依然具有软法的性质。不同于一般的公约，该框架公约旨在通过渐进式监管过程建立初始未明确的义务和实施机制，随后通过协议加以具体化。其包含具有广泛的承诺，但缺乏具体性。其主要内容包括三类，一是设定框架公约和整个条约制度的目标，即确保人工智能系统生命周期内的活动完全符合人权、民主和法治”，并鼓励各国密切关注人工智能对人类健康和环境、就业和劳动等社会经济方面的风险与影响。二是阐明人工智能应用领域的重大原则性问题，包括保护人权、公共部门和民主程序、个人隐私和自主权益、提升识别和预防人工智能风险的能力、开展国际合作。三是建立国家间的广泛承诺、确保条约制度的持续性，如举行缔约方会议进行定期磋商以促进公约的有效应用和实施，缔约方对公约履行情况进行报告的义务。

原则导向与规则导向的主要区别在于规则的使用方法，规则导向监管通过详尽的法规确保市场合规，但可能存在限制市场和监管机构的灵活性的弊端；原则导向治理侧重于规则的目的和根本义务，采用高层次原则而非具体规则，不具体规定达成目标的方式，为市场提供发展空间。⁴⁶原则导向治理方式发挥作用的逻辑是，通过设定制度目标和预期结果，确立行为体的基本义务，构建了一个具有预见性和前瞻性的治理框架，通过灵活和包容的方式在人工智能技术不断进步和规则的稳定性之间寻求平衡。

原则导向的治理框架有三个核心特征。第一，抽象性与开放性。各类软法文件不预设具体技术场景（如生成式人工智能或自主武器），而是通过“人权保护”“透明度”“可问责性”等抽象表述，为各国提供适应性解释空间。例如，联合国《人工智能伦理框架》提出“增进人类福祉”原则，既涵盖医疗领域人工智能应用的公平性，也延伸至军事领域自主武器的合法性判断。第二，具有价值兼容性。通过提炼不同文化传统的最大公约数（如中国倡导的“人类命运共同体”与欧盟“以人为本的人工智能”），原则导向淡化意识形态差异。例如，G20《人工智能原则》将“包容性增长”与“可持续发展”并列，兼顾发达国家与发展中国家的诉求。第三，强调风险预防。人工智能伦理规范并不是针对人工智能本身的，而是对研究、开发与应用人工智能的人类提出要求，其目标是要求研究者、开发者与应用者遵循一定的规范，以使人工智能带来的伦理风险降到最低。⁴⁷各类文件从不同方面要求对人工智能的潜在系统性风险（如算法歧视、深度伪造等）采取前瞻性约束。例如，《布莱切利宣言》（2023）将“前沿人工智能风险”纳入治理范围，指出即使科学证据不充分也需提前管控。

伦理原则具有国家驱动性，反映各国基本机制取向。经济与合作发展组织（OECD）、法国、德国、加拿大等 15 国建立的“人工智能全球合作伙伴关系”及

⁴⁶ 徐玖玖：《人工智能的道德性何以实现？——基于原则导向治理的法治进路》，《现代法学》2021 年第 3 期，第 31 页。

⁴⁷ 沈岷：《论软法的实施机制——以人工智能伦理规范为例》，《财经法学》2024 年第 6 期，第 116 页。

欧盟提出的《人工智能与人权、民主和法治框架公约》都反映了西方自由主义价值观。⁴⁸国家之间的战略竞争增加了达成国际共识的难度，导致在全球层面缺乏统一的人工智能治理规则。各国针对某一治理事项的态度越是多样化，国际法就越需要软性法律（second degree law）的存在，这种法律并不旨在解决深层次的价值冲突，而是为了使各种不同的价值体系相互间共存成为可能。⁴⁹因此，国家间的战略竞争已成为全球人工智能治理规则不平衡的核心动因。技术主导权争夺、地缘政治博弈与安全考量，使得主要大国在规则制定中优先维护自身利益，而非寻求全球共同价值。如中、美两国在人工智能伦理框架上持有不同立场，中国2023年《全球人工智能治理倡议》强调“主权平等”和“发展权优先”，主张各国自主制定人工智能治理规则，反对以人权为借口干涉他国内政。其《生成式人工智能服务管理暂行办法》要求内容安全审查，确保“社会主义核心价值观”融入人工智能系统。而美国于2022年发布《人工智能权利法案蓝图》将“算法公平性”“隐私保护”作为核心原则，联合盟友在经合组织框架下推动“民主价值观导向”的人工智能伦理标准，试图将中国排除在规则制定体系外。这一立场对立也体现在国际造法实践中，美国主导的“人工智能全球合作伙伴组织”（GPAI）拒绝中国加入，指责中国人工智能技术“缺乏透明度”，同时中国在联合国教科文组织推动《人工智能伦理建议书》时，弱化“人权”条款，强调“文化多样性”。软法在塑造和分享价值观以及创设国家行为预期方面发挥着重要作用。⁵⁰软法的出现客观反映了国际法发展的需求，允许包含相互冲突的标准和目标，并为国家在提出和反驳主张时提供灵活性。

（二）技术标准：以市场激励机制为主的遵守机制

全球人工智能治理高度依赖技术标准的制定与实施。标准是一种协调机制，可以确保世界各地生产的产品实现共享。标准提供了开发新技术所需的制度基础，并提供了以可控方式进行开发的安全流程。2022年7月国际电工委员会（IEC）与国际标准组织（ISO）联合发布两项人工智能新的基础标准，旨在为数字化转型提供重要的基石。技术标准作为标准的一类，兼具技术性、社会性和可通约性，能够更为有效地创建、演绎和引导人工智能技术的有序稳健发展。⁵¹一方面，技术标准通过设定行业规范、算法透明度要求、风险管理标准等，确保人工智能技术的负责任使用；另一方面，共同的技术标准通常是国际合作和创新的基础，不同国家和国际组织制定的独特治理方法可能导致技术标准的碎片化，使得跨国合

⁴⁸ José-Miguel Bello y Villarino, “Global Standard-Setting for Artificial Intelligence: Para-regulating International Law for AI?”, (2023) 41 The Australian Year Book of International Law 157, p.164.

⁴⁹ J. H. H. Weiler & Andreas L Paulus, “The Structure of Change in International Law or Is There a Hierarchy of Norms in International Law?” (1997) 8 European Journal of International Law 545, p.565.

⁵⁰ C. M. Chinkin, “The Challenge of Soft Law: Development and Change in International Law”, (1989) 38 International and Comparative Law Quarterly 850, 865.

⁵¹ 王平、梁正：《标准和标准化概念的多学科观点（之五）——Brunsson 和 Busch 的社会学观点》，载《标准科学》2020年第5期。

作和技术交流变得更加困难。⁵²

作为“建构主义”实践和理性渐进的混合“制标”机构，国际标准组织（ISO）最大的建构性贡献是，在标准制定过程中，将主权国家行为的诸多标签、特定制标行为和市场主体行为的创新、竞争及逐利，经由技术组织融为一体，实现合作、多赢、技术共享、消费最大化的市场格局；其理性渐进主要表达为坚守标准的技术问题归技术创新和技术专家解决，即通过技术组织这个平台进行，使 ISO 及其产出结果即标准文件在严格程序约束的形式中，呈现出或成为公平、合理和实质性平等的本色及其保证体系。⁵³

自发性的标准化之所以能够形成的关键在于网络效应。网络效应是经济学上的一个概念，它解释了在缺乏统一、协调性的安排之下，不同个体之间如何在无意中形成一个普遍标准。网络效应是指随着使用同一产品或服务的用户的数量增加，产品或服务的隐含价值也随之增加，这反过来又吸引了更多的用户，如此循环往复，造成越来越多的用户涌向该产品的“滚雪球”效应。网络外部性源于对兼容性的需求，例如随着更多消费者购买电话，电话对用户价值会增加，仅仅因为用户可以拨打更多人的电话，如果只有一个人拥有电话，它将毫无价值。这是因为物品的效用取决于它帮助用户进行互动的能力，即用户从特定物品中获得的效用依赖于同一“网络”中其他用户的数量。⁵⁴

尽管 WTO/TBT 要求各个国家都要采用国际标准，但其并未对国际标准给出一个明确而清晰的界定。跨国标准和标准制定联盟在某些方面可能会相互展开激烈的竞争，但在其他情况下它们又是互补的。原则上，它们也可以简单地中长期共存——构建不同标准的“俱乐部”或“社区”。随着时间的推移，它们还可以互动并创建协调机制。⁵⁵正式标准系统中的标准产生模式都大致相同，即所谓的“技术民主”，而且都要按照不同的领域建立技术委员会，标准制定流程采用严格的“开放透明、协商一致”的基本原则。⁵⁶ISO 在市场上通过认证体系提供市场治理工具，提升自身在国际贸易中的话语权。⁵⁷

技术标准发挥效力的核心在于市场激励，即遵守技术标准可以获得有价值的市场资源，例如提供认证机制，对于人工智能领域的表示遵守透明度或隐私实践

⁵² 马光、王丽雯：《国际软法视角下人工智能全球治理的现状与建议》，载《海关与经贸研究》2024 年 5 月第 45 卷，第 3 期。

⁵³ 王艳林、叶雅婷：《论国际标准化组织技术组织治理的建构》，《经贸法律评论》2024 年第 6 期，第 138 页。

⁵⁴ Bryan H. Druzin, “Spontaneous Standardization and the New Lex Maritima”, in Miriam GOLBY, ed., Oxford International Arbitration Series: The Continued Development of Shipping Law (Oxford: Oxford University Press, 2016), 63, p.64-65.

⁵⁵ 王艳林、叶雅婷：《论国际标准化组织技术组织治理的建构》，《经贸法律评论》2024 年第 6 期，第 137 页。

⁵⁶ 王平、房庆：《正式标准化组织的标准系统研究》，《标准科学》2020 年第 11 期，第 10 页。

⁵⁷ 王艳林、叶雅婷：《论国际标准化组织技术组织治理的建构》，《经贸法律评论》2024 年第 6 期，第 125 页。

的标签或认证就扮演了这样的角色。⁵⁸另一个重要的动机是保护利益。私营部门实体的目标是增加股东价值，软法可以成为保护这些利益的工具。许多从事开发人工智能方法和应用的组织已经制定了实践原则，以保护他们的声誉，进而保护他们的底线。⁵⁹技术标准的实施同样依赖供应链机制，其中，上游供应商和下游采购商都必须遵守其供应链伙伴提供的规格要求，以满足公司希望实施的产品管理责任和预防责任措施。这种供应链管理是基于这样的认识演变而来的：如果一家公司购买并使用了供应商的缺陷产品，或者向下游客户提供自己的产品，而这些客户可能不负责任地使用它们，特别是如果该公司资金雄厚，那么该公司可能会被追究责任和承担赔偿责任。⁶⁰

（三）国际合作：通过能力建设机制促进治理效能

如果说对话是国家通过它学习新问题并建立起足够关注以成为发展国际规范的一部分，并使之得以遵守的机制。国家还可以通过能力建设机制遵守规范。具言之，国家某一遵守规范的前提是拥有实施该规范的能力和知识，例如，当某一条约的实施需要影响非国家行为体的行为时，就需要相关国家制定“详细的行政规定和有力的执行措施”，这反过来需要“科学和技术判断、行政能力和财政资源”，而许多国家并不具备相应的能力。这可能超出了许多国家的能力。国际条约通常包括与实现特定目的有关的技术和人力资源建设措施等内容，或者由各种国际组织和为实现具体目标设计的双边协议事后创建的，这些措施会促进遵守，并专门针对那些人力资源和经济资源有限的国家。⁶¹

2024年7月1日，第78届联合国大会上，中国提出的加强人工智能能力建设国际合作决议获得协商一致通过，标志着国际社会在推进人工智能（AI）领域合作与发展上迈出了重要一步。此次决议得到了140多个国家的支持，凸显了全球范围内对于加强AI能力建设、促进技术普惠的广泛共识。该决议核心在于强调以人为本、智能向善的发展原则，鼓励国际合作以助力各国，尤其是发展中国家，提升其在人工智能领域的实力。

四、中国参与全球人工智能治理的路径

碎片化的治理体系可能导致国家以本国国内法代替国际法，为保护人工智能产业的发展而设置相关的贸易壁垒或放宽国内监管，从而使得人工智能产业创新和治理体系效率显著降低。⁶²软法规范作为“效力来源”可以塑造国家利益和国家

⁵⁸ Carlos Ignacio Gutierrez, Gary Marchant & Lucille Tournas, Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance, (2020) 61 Jurimetrics 133, p.148.

⁵⁹ Carlos Ignacio Gutierrez, Gary Marchant & Lucille Tournas, Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance, (2020) 61 Jurimetrics 133, p.148.

⁶⁰ Carlos Ignacio Gutierrez, Gary Marchant & Lucille Tournas, Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance, (2020) 61 Jurimetrics 133, p.

⁶¹ Sylvia I. Karlsson-Vinkhuyzen & Antto Vihmap, “Comparing the legitimacy and effectiveness of global hard and soft law: An analytical framework”, (2009) 3 Regulation & Governance 400, p.408.

⁶² Robert F. Trager, & Ben Harackp et al., International Governance of Civilian AI: A Jurisdictional Certification Approach, <https://arxiv.org/abs/2308.15514>, p.12.

的观念与价值。通过强调参与性与互相信任，国际共同体成员能够针对全球人工智能治理达成一定程度的共识从而开始初步的行动，增强国际社会对人工智能治理合作意愿与合作效能。

（一）在国家同意的基础上加强国际合作发展共同利益

尽管国家同意作为国际造法的唯一合法性基础存在一定的局限性，但不可否认的是，国家依然是最重要的国际法主体，国家同意依然是最重要的国际造法依据。国际法中的“国家同意”原则强调，任何国际规则的制定与实施均需基于主权国家的自愿参与。在因为国家利益和价值等规范性力量是国家同意的根本动因，那么国家对统一性的国际法规则的同意应建立在共同利益与共同价值上。

通过国际合作推动技术普惠将有助于发展共同利益与共同价值。在人工智能治理领域，当前发达国家在资金、技术以及信息上占据绝对优势，而发展中国家处于相对劣势。因此即便制定权利义务较为明确的硬法，很多发展中国家事实上不具备履行与发达国家相当义务的能力。这一问题无法通过硬法规范下的法律责任追究和制裁来解决，而软法也只是对规范法律约束力的妥协，并不能从实质上提高国家履行法律义务的能力。⁶³因此，基于实现全球人工智能善治、维护全球共同利益的目标，人工智能技术处于领先地位的国家应当向发展中国家提供资金、技术和能力方面的支持，并进行信息上的沟通与交流，缩短国家间在话语和力量上的差距，能够推动形成统一的治理规范，通过技术、价值、规范的相互间建构推动全球人工智能治理走向善治。

（二）增加造法主体的代表性和包容性

每个国家参与一般国际造法的过程，既是站在自身的角度代表自己的意志，同时也是作为国际社会的一部分、代表普遍意志发声。国家经过商谈和沟通形成的意志，经历了理性化的过程，因而不同于各国个体意志的集合。在国际造法过程中，参与国越多，对国际社会的代表性就越高，得出的国际法规则就越具有合法性，也更可能符合理性。⁶⁴

人工智能的溢出风险加剧了国际社会的“结构性不平等”，致使发展中国家在人工智能领域一直处于接受者的地位，这无疑会影响全球人工智能治理国际造法的合法性。长期来看，全球人工智能鸿沟既是联合国开展全球人工智能治理面临的重要挑战，也是其治理进程中的重要治理对象。⁶⁵增加造法主体的代表性和包容性是使全球人工智能治理体系更加公正、合理的应然路径。

增加造法主体的代表性和包容性同时意味着全球人工智能治理体系需要容纳多元化的文化传统、治理理念和价值趋向。以现有的软法框架为基础，以多边

⁶³ 曹炜：《国际环境法造法机制研究》，《中国人民大学学报》2016年第5期，第119页。

⁶⁴ 黄蓬北：《论国际造法的“立法”转向——基于普遍意志的契约论解释》，《清华大学学报(哲学社会科学版)》2025年第11期，第259页。

⁶⁵ 阎德学：《全球人工智能治理：联合国的角色》，《国际问题研究》2024年第4期，第136页。

机制为主要平台，推动不同文明间的相互交流，增进相互理解。以应对共同的挑战为目标，确保全球人工智能治理的规则既符合国际社会的普遍利益，也兼顾各国的特殊需求。此外，还应积极邀请非国家行为体，如国际组织、跨国公司、非政府组织等，参与全球人工智能治理的造法过程，这些非国家行为体往往拥有独特的专业知识和实践经验，能够为全球人工智能治理提供新的视角和思路，从而推动造法过程的创新和发展。

（三）发展软硬结合的综合造法程序

原则导向的治理规范极为重要，各类软法文件在关键的人工智能伦理原则上达成一致，是技术治理向前迈出的重要一步。但仅有原则是不够的，利益相关者必须超越原则，转向过程，以识别和实施具体项目——在开发或使用人工智能的实体内部和外部，将伦理原则付诸实践。⁶⁶软法与硬法深度融合，形成动态适应的规则体系，灵活应对技术快速迭代带来的治理挑战。⁶⁷

一方面，需通过硬法明确禁止性原则，硬法是对人工智能技术发展设定“不可逾越红线”的核心机制，这一法律工具的构建是为了回应技术失控风险。据统计，生成式人工智能的参数量年均增长 10 倍，若无禁止性条款，可能触发“技术奇点失控”。同时，某些不可逆的损害必须通过硬法预防，如深度伪造技术导致虚假信息传播速度超事实核查能力 300%，需通过硬法禁止恶意使用场景。

另一方面，需注重软法和硬法的互动、协调及相互转化。实践证明软法是平衡创新空间、塑造共同价值、共同利益的重要工具。硬法能够为软法提供合法性源泉，国际标准化组织将三种结构性、规范性权力的约束——瑞士民法为主要的国内法、组织章程为核心的自治法和 WTO/TBT 协定、WTO/SPS 协定为原则的国际法，凝练为以公开、透明、协商一致为价值导向的标准化活动守则和以多元、正当、实践为市场导向的标准中立原则。⁶⁸同时准可能会成为国内法的一部分，因为它们是规范制定者的（有意或无意的）参考点，具有“潜意识扩散”作用。⁶⁹国际法标准化组织制定的软法在国家实践中产生实际的法律效果，间接性地发挥效力。标国际标准化组织通过与国际电工委员会联合组建的专门技术委员会，系统化推进人工智能治理相关标准的制定，逐步构建起覆盖人工智能全生命周期的标准体系，已出台的成果包括管理体系标准（ISO/IEC 42001）、风险管理指南（ISO/IEC 23894）、可信赖人工智能框架（ISO/IEC TR 24028）。欧盟《人工智能法》中的 74 项技术规范直接引用 ISO 标准，中国《生成式人工智能服务管理办

⁶⁶ Carlos Ignacio Gutierrez, Gary Marchant & Lucille Tournas, “Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance”, (2020) 61(1) Jurimetrics 133, p.134.

⁶⁷ 张欣：《人工智能治理的全球变革与中国路径》，《华东政法大学学报》2025 年第 1 期，第 21 页。

⁶⁸ 王艳林、叶雅婷：《论国际标准化组织技术组织治理的建构》，《经贸法律评论》2024 年第 6 期。

⁶⁹ José-Miguel Bello y Villarino, “Global Standard-Setting for Artificial Intelligence: Para-regulating International Law for AI?”, (2023) 41 The Australian Year Book of International Law 157, p.171-172.

法》兼容 ISO/IEC 42001 体系要求。

五、结语

作为传统硬法的补充与过渡，软法凭借其灵活性、适应性和包容性，有效回应了人工智能技术快速迭代与治理需求之间的张力。法律的效力不依赖于立法者和守法者之间的等级制度，而是依赖于所有有关的参与者之间的相互作用，法律不是一条“单行道”，只有当参与者合作建立共同的理解并维护合法性实践时，法律才能存在。⁷⁰

国际造法的“立法”转向表明，软法并非仅是硬法的前奏，而是逐渐形成了一套独立的规范生成机制，其核心在于通过动态协商塑造共同利益，而非单纯依赖国家同意的传统模式。软法既承载了各国对共同利益与价值的初步共识，又通过原则导向的治理框架为多元主体提供了合作空间。然而，软法主导的治理体系也面临碎片化、代表性不足及执行力弱化的挑战，凸显了国际社会在协调安全与创新、主权与全球公共利益之间的深层矛盾。

全球人工智能治理亟需构建更具包容性与协同性的国际法框架。如何在技术鸿沟与地缘博弈的背景下，推动发达国家与发展中国家的实质性合作，平衡“人权导向”与“主权导向”的价值分歧？如何将碎片化的软法规范整合为层次分明、互操作性强的治理体系，并探索软硬法协同的综合性路径？此外，非国家行为体在造法中的角色如何进一步制度化，以增强治理机制的合法性与实效性？这些问题不仅关乎人工智能技术的可持续发展，更是对全球治理体系韧性与创新能力的重大考验。唯有通过持续对话与制度创新，方能在技术变革中实现秩序与公正的共融。

⁷⁰ Jutta Brunnée & Stephen J. Toope, *Legitimacy and Legality in International Law: An Interactional Account*, Cambridge Studies in International and Comparative Law, Cambridge University Press, 2010, p.7.

人工智能数据训练的合理使用规则研究——中美欧的比较分析

吕丹丹¹

摘要：ChatGPT、Deepseek 等生成式人工智能大模型在训练过程中对海量数据的使用引发合理使用规则适用争议。欧盟、美国与中国在相关法律制度与司法实践中呈现出差异化路径。在合理使用规则上，欧盟《人工智能法案》《数字单一市场版权指令》的“默示许可+选择退出+文本与数据挖掘例外”、美国判例法中的“转换性使用”规则，中国采取“半开放式”的立法模式。未来，中国合理使用制度设计应在促进技术创新与保护著作权之间寻找合理平衡的路径，注重数据来源合法性审查、使用目的公益性考量以及权利补偿机制，完善训练端的合理使用规范和输出端的信息披露机制，推动大模型数据训练与输出管理的双轮驱动治理，从而实现对合理使用边界的动态平衡。

关键词：生成式人工智能 Deepseek 数据训练 合理使用 版权

一、问题的提出：大模型数据训练原理与争议

近年来，以大语言模型为核心的系统逐步实现了对自然语言的理解、生成与交互。自 Transformer 结构提出以来，模型的参数规模、数据规模和训练算力持续扩展，成为推动生成式人工智能能力跃迁的三大要素。其中，数据训练作为模型学习语言规律的基础环节，在整个模型生命周期中占据核心地位。数据训练已成为决定模型竞争力的关键环节之一。从 ChatGPT 的数据爬取到 DeepSeek 的数据蒸馏，人工智能在采集、训练和使用中对数据的处理存在较大争议，但由于各国缺乏明确且匹配技术发展的监管规范，需要学术界及时回应。²

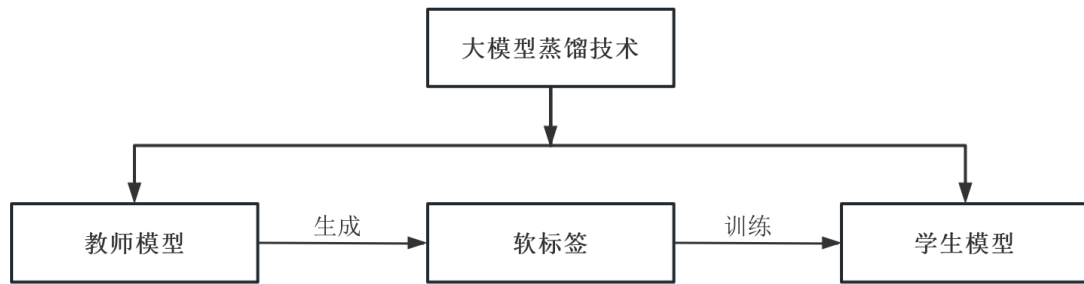
ChatGPT 代表了主流大语言模型的数据训练路径，其预训练阶段基于互联网公开数据，涵盖维基百科、网络文档、开源书籍等语料。Deepseek 作为中国本土大模型代表，其通过模型蒸馏的创新方法³，DeepSeek-V3 的训练只需要不到 600 万美元的 Nvidia H800 芯片的计算能力引起了全球人工智能界的关注⁴。大模型蒸馏技术的核心原理是知识传递机制，学生模型（如 DeepSeek）通过模仿教师模型（如 ChatGPT）的输入与输出，结合软标签知识不断优化自身性能，从而在计算资源受限的环境中实现接近教师模型的性能。

¹ 吕丹丹，华东理工大学法学院硕士研究生，研究方向：人工智能法，国际经济法。

² 顾男飞：《从 ChatGPT 到 DeepSeek：人工智能蒸馏数据的风险治理》，载《图书馆论坛》2025 年。

³ 白云怡，陈青青：《“美政界应认清：中国崛起有利于美国”》，载《环球时报》2025 年 3 月 28 日第 5 版。

⁴ Eduardo Baptista, What is DeepSeek and why is it disrupting the AI sector?, Jan 29, 2025 <https://www.reuters.com/technology/artificial-intelligence/what-is-deepseek-why-is-it-disrupting-ai-sector-2025-01-27/>.



2023 年 12 月 27 日,《纽约时报》等原告指控 OpenAI 和微软通过抓取网络内容(包括受版权保护的新闻文章)训练 ChatGPT,2025 年 3 月 26 日,美国纽约南区联邦地区法院法官 Sidney H. Stein 就《纽约时报》等新闻机构起诉 OpenAI 及微软的版权侵权案发布命令,驳回了 OpenAI 要求直接驳回案件核心指控的动议,允许案件的主要版权侵权主张进入审理阶段。⁵ 2025 年 1 月,OpenAI 指控 DeepSeek 在数据训练过程中通过蒸馏技术不恰当地从其模型获取数据,OpenAI 本身也被指控在开发 ChatGPT 时不当访问了其无权访问的内容,其已成为多起包括来自《纽约时报》的此类诉讼的对象。⁶ 欧盟、中国、美国以及其他国家,大模型训练过程中往往使用大量包含受版权保护的内容,但极少实际取得权利人的明示或默示授权,引发数据训练是否符合版权合理使用或版权例外制度的争议。2025 年 2 月 9 日,经济合作与发展组织(OECD)发布的《人工智能基于抓取数据的知识产权问题》(Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data)专题报告指出,数据抓取是开发许多大型语言模型所需的训练数据的主要获取方法,但频繁涉及版权等法律争议,亟需全球协调的政策框架与技术创新应对挑战。⁷ 本文通过比较分析欧盟、美国对合理使用规则与司法裁判,为中国合理使用规则提供借鉴,厘清数据训练合理使用规则的边界。

二、欧美中规则比较观察：数据训练合理使用规则差异

合理使用制度作为各国著作权法中平衡权利排他与表达自由的重要工具,是指在特定情况下,可以不经著作权人许可使用其已经发表的作品,且无需向著作权人支付费用。

(一) 欧盟：默示许可+选择退出+例外规定

在《欧盟人工智能法案》中,针对通用人工智能模型(General Purpose AI Model,以下简称 GPAI 模型)的数据训练规则主要规定在第 53 条第 1 款第(c)项与第(d)

⁵ The New York Times Company v. Microsoft Corporation (1:23-CV-11195). District Court, S.D. New York. <https://www.courtlistener.com/docket/68117049/the-new-york-times-company-v-microsoft-corporation/?page=4>

⁶ Kevin Collier and Jasmine Cui, OpenAI says DeepSeek may have 'inappropriately' used its data, Jan. 30, 2025, <https://www.nbcnews.com/tech/tech-news/openai-says-deepseek-may-inappropriately-used-data-rcna189872>.

⁷ OECD. Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data https://www.oecd.org/en/publications/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data_d5241a23-en.html.

项。⁸该规定确立了 GPAI 模型提供商在处理受版权保护内容时所应遵循的合规义务，确立了版权人选择退出机制与要求公开训练数据摘要。依据第 53 条第 1 款第(c)项，GPAI 模型提供商在数据训练阶段应当为权利人提供明确、可操作的选择退出机制。该机制以 2019 年欧盟《数字单一市场版权指令》(以下简称《DSM 指令》)第 4 条第 3 款为依据，要求提供商采取有效手段确保权利人得以表达其拒绝将作品用于模型训练的意愿。根据第 53 条第 1 款第(d)项，GPAI 模型提供商还应编制并公开训练数据摘要。摘要应依照欧盟人工智能办公室 (AI Office) 统一制定的模板进行撰写，内容需涵盖训练数据的来源、类型及代表性内容等，使得权利人能够了解其作品是否被使用。此外，《DSM 指令》规定，在满足数据合法来源的前提下，科研机构和文化遗产机构以科研为目的“文本和数据挖掘”行为属于可被许可行为，且在“科学研究型数据挖掘”情形中不允许权利人通过协议或单方声明等方式予以保留。⁹《DSM 指令》第 7 (2) 条¹⁰以及 2001 年《信息社会指令》(以下简称《InfoSoc 指令》)第 5 (5) 条“文本和数据挖掘的例外和限制仅适用于某些特殊情况，这些特殊情况不得与作品或其他主题的正常利用相冲突，也不得不合理地损害权利人的合法利益”¹¹。2024 年 12 月，欧盟成员国在回复“生成式人工智能与版权关系的政策调查问卷”时，会员国普遍支持为行使退出权引入共同标准，以便在欧盟内部为选择退出制度的运作提供更多的法律确定性。

12

《欧盟人工智能法案》	第五十三条 通用人工智能模型提供者的义务 1. 通用人工智能模型提供者应当： (c) 制定一项政策，遵守联盟有关版权和相关权利的法律，特别是通过最先进的技术，识别和遵守根据第 2019/790 号指令 (EU) 第 4(3) 条表达的权利保留； (d) 根据人工智能办公室提供的模板，制定并公开有关通用人工智能模型训练内容的足够详细摘要。
《DSM 指令》	第三条 用于科学研究目的的文本和数据挖掘 1. 成员国应对第 96/9/EC 号指令第 5(a) 条和第 7(1) 条、第 2001/29/EC 号指令第 2 条以及本指令第 15(1) 条规定的权利作出例外规定，允许研究组织和文化遗产机构为科学研究、

⁸ EU Artificial Intelligence Act. Article 53: Obligations for Providers of General-Purpose AI Models, <https://artificialintelligenceact.eu/article/53/>.

⁹ Directives on Copyright in the Digital Single Market, <https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng>.

¹⁰ Directives on Copyright in the Digital Single Market. Article 7, <https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng>.

¹¹ InfoSoc Directives. Article 5, <https://eur-lex.europa.eu/eli/dir/2001/29/oj/eng>.

¹² Council of the European Union. (2024). Policy questionnaire on the relationship between generative Artificial Intelligence and copyright and related rights – Revised Presidency summary of the Member States contributions (16710/1/24 REV 1). <https://data.consilium.europa.eu/doc/document/ST-16710-2024-REV-1/en/pdf>.

	文本和数据挖掘目的而对其合法接触的作品或其他内容进行复制和摘录。 2. 按照第 1 款制作的作品或其他内容的复制件应当以适当的安全级别保存，并可以保留用于科学研究的目的，包括用于验证研究成果。 3. 应允许权利人采取措施确保作品或其他内容所在的网络和数据库的安全性和完整性。此类措施不得超出实现该目标所必需的范围。 4. 成员国应鼓励权利人、研究组织和文化遗产机构就第 2 款和第 3 款所述义务和措施的应用确定共同商定的最佳做法。
《DSM 指令》	第四条 文本和数据挖掘的例外或限制 1. 成员国应对第 96/9/EC 号指令第 5(a) 条和第 7(1) 条、第 2001/29/EC 号指令第 2 条、第 2009/24/EC 号指令第 4(1)(a) 和 (b) 条以及本指令第 15(1) 条规定的权利，规定为文本和数据挖掘目的复制和提取合法可访问的作品和其他主题的例外或限制。 2. 根据第 1 款进行的复制和提取，可在文本和数据挖掘所需的时间内保留。 3. 第 1 款规定的例外或限制适用于以下条件：该款所述作品或其他内容的使用尚未得到其权利人以适当方式明确保留，例如，对于在线公开提供的内容，则以机器可读的方式保留。
《DSM 指令》	第七条 2. 2001/29/EC 号指令第 5(5) 条适用于本标题规定的例外与限制。
《InfoSoc 指令》	第五条 例外和限制 5. 第 1、2、3 和 4 款规定的例外与限制仅应在某些特殊情况下适用，这些情况不与作品或其他内容的正常利用相冲突，也不会不合理地损害权利人的合法利益。

表 1 欧盟数据训练合理使用相关规定

欧盟在 GPAI 模型数据训练的治理上，采取“默示许可+选择退出+例外规定”的制度框架。在作品合法获取的前提下，如权利人未明确表示反对，推定其同意其作品用于模型训练。同时给予权利人通过技术机制行使排除权的可能，从而平衡人工智能技术发展的数据可得性与维护著作权人的合法权益。

（二）美国：“四要素”和“转换性使用规则”

美国版权法中的合理使用原则¹³（Fair Use）为在未经版权持有人许可的情况下使用受版权保护的材料提供合法性依据。根据美国普通法的先例，合理使用被编入现行法规，它没有分类或具体的定义，而是根据对四个因素的裁决评估来决定。¹⁴其一，使用的目的和性质，考察使用行为是否具有商业性质以及是否具备转换性。商业性质的使用可能对合理使用认定不利，但非决定性因素；而转换性使用，即对原作品增加新的表达、意义或信息，如评论、批评、戏仿、新闻报道或学术研究等，更易被认定为合理使用。其二，版权作品的性质，事实性作品相较于创造性作品更易被认定为合理使用对象。其三，使用的数量和实质性，使用的数量越少且非原作品核心或最具价值部分，越可能被认定为合理使用。其四，使用对原作品市场价值的影响，若使用行为未对原作品市场造成显著负面影响，甚至可能促进其销售，则更可能被认定为合理使用。合理使用的认定是法院对上述四个因素进行综合权衡的结果，而非单一依赖某一因素，每个案件的具体情况都会影响最终的判决。美国多次运用合理使用原则以应对新技术所带来的挑战，合理使用的先例经验使得美国技术开发市场更具有活力，促使技术公司和资本家对新技术进行投资，而无需担心陷入著作权纠纷的风险。¹⁵

（三）中国：“半开放式”的合理使用立法模式

对于合理使用制度，《中华人民共和国著作权法》（2020年修正）引入“三步检验法”来进行使用作品的“合理性”界定，要求作品使用中“不得影响该作品的正常使用，也不得不合理地损害著作权人的合法权益”。¹⁶在促进技术创新和商业发展确有必要的特殊情形下，如果该使用行为既不与作品的正常使用相冲突，也不至于不合理地损害作者的正当利益，可以认定为合理使用。¹⁷中国《著作权法》第24条第1款列举了12种具体合理使用情形，第1款第13项规定了兜底情形。立法者严格限定了兜底情形的成立要件和合理使用的具体类型，该合理使用条款也仅属于“半开放式规定”。¹⁸《生成式人工智能服务管理暂行办法》（以下简称《暂行办法》）第7条对数据来源合法性要求进行了规定，生成式人工智能服务提供者应当依法对生成式人工智能产品的预训练数据、优化训练数据来源的合法性负

¹³ 17 U.S. Code § 107 - Limitations on exclusive rights: Fair use.

¹⁴ Dan L.Burk, Algorithmic Fair Use, 86 U.CHI.L.REV.283(2019).

¹⁵ 熊琦，陈子懿：《美国人工智能模型训练合理使用认定的成案经验研究》，载《科技与法律(中英文)》2024年第6版。

¹⁶ 吴汉东，刘鑫：《〈著作权法〉第三次修改的立法安排与实施展望》，载《法律适用》2022年第4期。

¹⁷ 最高院2011年《关于充分发挥知识产权审判职能作用推动社会主义文化大发展大繁荣和促进经济自主协调发展若干问题的意见》第8条 妥当运用著作权的限制和例外规定，正确判定被诉侵权行为的合法性，促进商业和技术创新，充分保障人民基本文化权益。正确认定合理使用和法定许可行为，依法保护作品的正当利用和传播。在促进技术创新和商业发展确有必要的特殊情形下，考虑作品使用行为的性质和目的、被使用作品的性质、被使用部分的数量和质量、使用对作品潜在市场或价值的影响等因素，如果该使用行为既不与作品的正常使用相冲突，也不至于不合理地损害作者的正当利益，可以认定为合理使用。对设置或者陈列在室外社会公共场所的艺术作品进行临摹、绘画、摄影或者录像，并对其成果以合理的方式和范围再行使用，无论该使用行为是否具有商业目的，均可认定为合理使用。

¹⁸ 刘云开：《人工智能训练作品的著作权合理使用进路》，载《东北大学学报(社会科学版)》2025年第1期。

责，确保数据来源合法，不侵害他人知识产权等合法权益。¹⁹推荐性国家标准《信息安全技术生成式人工智能预训练和优化训练数据安全规范》（以下简称《数据安全规范》）和《生成式人工智能服务安全基本要求》（以下简称《基本要求》）进一步规范了数据训练的合法性和安全性。

三、欧美中典型司法裁判：数据训练合理使用边界的实践检验

欧盟、美国、中国在司法实践中探索合理使用规则的边界，欧盟德国 Laion 案中法院判定非商业性人工智能训练数据符合例外规定，美国“汤森路透诉罗斯案”中法院判定不构成合理使用，中国杭互奥特曼案中法院判定构成合理使用。

（一）欧盟：德国 Laion 案非商业性人工智能训练数据符合例外规定

2024 年 9 月 27 日，德国汉堡地方法院在 Robert Kneschke 诉 LAION 案中认定人工智能训练数据挖掘行为构成著作权合理使用，²⁰这是全球首例就人工智能训练是否构成版权侵权问题予以明确的裁决。本案原告 Robert Kneschke 系知名摄影师，他将其涉案摄影作品授权予图片代理商在线展示并许可使用，且代理商网站通过明确的权利保留声明（以自然语言表述）禁止任何形式的下载、抓取或缓存行为。LAION 为一家非营利组织，其主要业务是对 Common Crawl 抓取的公开数据集进行再加工，形成包含图片链接及 ALT 文本的 LAION-5B 数据集。该数据集通过下载原图并利用专门软件分析 ALT 文本与图片内容的对应性，LAION 声称分析完成后删除已下载数据。LAION-5B 数据集目前被广泛应用于人工智能大模型训练，包括业界知名的 Stable Diffusion 模型。原告其摄影作品被收录于由被告 LAION 组织加工形成的 LAION-5B 数据集中，原告主张被告在未经许可的情况下对原告摄影作品进行抓取、分析及存储，侵犯了其对作品享有的著作权。

欧盟针对人工智能训练数据的著作权利限制与例外制度主要建立在《DSM 指令》第 3 条规定了以科学研究为目的的文本与数据挖掘（TDM）例外，第 4 条则确立了普通 TDM 行为的版权例外。德国通过《德国版权法》（以下简称 UrhG）²¹对该指令进行国内法转化，具体体现为 UrhG 第 44b 条的一般性 TDM 例外规

¹⁹ 《生成式人工智能服务管理暂行办法》第 7 条 生成式人工智能服务提供者（以下称提供者）应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定：（一）使用具有合法来源的数据和基础模型；（二）涉及知识产权的，不得侵害他人依法享有的知识产权；（三）涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形；（四）采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性；（五）《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》等法律、行政法规的其他有关规定和有关主管部门的相关监管要求。

²⁰ Hamburg Regional Court, Germany 2024: Robert Kneschke v. LAION e.V., Case No. 310 O 227/23.

²¹ UrhG. https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html

定²²，以及第 60d 条专门针对科研目的的 TDM 特殊例外条款²³。在本案审理过程中，原被告双方就 Laion 未经授权使用涉案照片进行人工智能训练的行为是否构成 UrhG 第 44a 条规定的临时复制²⁴、第 44b 条规定的文本与数据挖掘例外以及第 60d 条规定的科研目的文本与数据挖掘例外等法律适用问题展开争议。法院经审理认定，Laion 对涉案照片的复制行为虽构成对原告开发权（exploitation rights）的侵害，但可援引 UrhG 第 60d 条关于科研目的的文本与数据挖掘的例外规定获得免责。法院认定 LAION 援引 UrhG 第 60d 条具有正当性，其核心裁判要旨在于：首先，构建人工智能训练数据集的行为本质上属于科学研究范畴；其次，判断科研目的的商业属性应基于活动实质而非机构组织形式或资金来源，数据集免费开放的特征佐证了其非商业性质。该判决确立的“非营利机构构建-数据采集处理-公开共享”模式，为人工智能训练数据的著作权合规使用提供了潜在路径，但需注意该路径的适用严格受限于科研目的要件，商业性人工智能开发仍需寻求其他合法性基础。

LAION 案判决虽非终局裁决，但作为全球首例涉及人工智能训练数据版权侵权的司法判例，其裁判将对欧盟成员国实施 TDM 例外条款产生示范效应，具有重要借鉴价值。该案核心争议聚焦于 UrhG 第 60d 条（对应欧盟《DSM 指令》第 3 条）的适用条件，法院认定非营利机构基于科研目的的数据集构建行为可豁免侵权责任。判决刻意回避了对更具普适性的第 44b 条（对应指令第 4 条）适用要件的认定，特别是未就权利人在 robots.txt 中的技术措施是否构成权利保留这一关键问题作出回应。法院裁判态度一方面为非商业性人工智能研究保留发展空间，另一方面则为商业性人工智能训练行为的合法性判断预留解释余地。从法解释学角度，判决将数据集构建行为直接等同于“科学研究目的”的论证逻辑存在商榷空间，未能充分回应 TDM 例外条款体系内部（第 44b 条与第 60d 条）的规范

²² UrhG. 44b 文本和数据挖掘：（1）“文本和数据挖掘”是指对单个或多个数字或数字化作品进行自动分析，以收集信息，特别是有关模式、趋势和相关性的信息。（2）为进行文本和数据挖掘，允许复制合法可访问的作品。当不再需要复制品进行文本和数据挖掘时，应删除复制品。（3）仅当权利人未保留使用权时，才允许根据第（2）款第 1 句进行使用。对于可在线获取的作品，使用保留仅以机器可读的格式进行，方有效。

²³ UrhG. 60d 用于科学研究目的的文本和数据挖掘：（1）允许按照以下规定，为了科学研究的目的，进行复制以进行文本和数据挖掘（第 44b 条第（1）款和第（2）款第 1 句）（2）研究机构有权进行复制。“研究机构”是指大学、研究机构和其他从事科学研究的机构，只要它们①追求非商业目的，②将所有利润再投资于科学研究或③根据国家批准的授权采取符合公众利益的行动。第一句规定的授权不适用于与私营企业合作的研究机构，该私营企业对研究机构具有一定程度的影响力，并可优先获取其科学研究成果。（3）下列人员还有权进行复制：①向公众开放的图书馆和博物馆，以及档案馆或影视或音频遗产领域的机构（文化遗产机构），②个人研究人员，只要他们追求非商业目的。（4）根据第（2）款和第（3）款获得授权，并且为了非商业目的，可以将根据第（1）款制作的复制品提供给下列人员：①明确界定的人员范围，以进行联合科学研究；②为监控科学研究的质量而设立的第三人。联合科学研究或者科学研究质量监控一旦结束，必须终止向公众提供。（5）根据第（2）款和第（3）款第 1 项获得授权的人员可以保留根据第（1）款制作的复制品，从而采取适当的安全措施，防止未经授权的使用，只要它们是出于科学研究的目的或监测科学成果质量的需要。

²⁴ UrhG 44a 临时复制行为：允许进行临时复制行为，这些行为是短暂或偶然的，并且构成技术过程的必要组成部分，其唯一目的是实现第三方之间的网络传输或合法使用作品，并且没有独立的经济意义。

竞合问题。这种规范适用的模糊性，可能催生“形式非营利实体+实质商业利用”的规避行为。

（二）美国：“汤森路透诉罗斯案”不构成合理使用

ROSS Intelligence 公司为开发基于人工智能的法律检索系统，试图获取 Westlaw 数据库中的 headnotes 等编辑性法律信息用于人工智能模型训练。由于 Westlaw 所有者汤森路透公司（Thomson Reuters，以下简称 TR）拒绝授权，ROSS 遂通过第三方 LegalEase 获取 25000 份以 headnotes 为基础制作的“Bulk Memos”，间接使用了 TR 编辑内容。TR 提起诉讼，主张其 headnotes 构成独立的版权作品，ROSS 未经授权在训练阶段复制该内容构成直接侵权，且不构成合理使用。特拉华联邦地区法院最终认定 ROSS 对 2243 条 headnotes 构成侵权，驳回其合理使用等抗辩主张。

美国《版权法》第 107 条²⁵确立了合理使用抗辩的四项因素，即使用的目的与性质、作品的性质、使用部分的数量与实质性、对原作品市场的影响。在本案中，法院强调了第一项“使用目的与性质”和第四项“市场影响”的优先权重，并据此否定了 ROSS 的合理使用主张。法院裁判认为：其一，数据商业性使用与“转化性”不足的判定。法院认定 ROSS 的数据使用具有明确商业目的，其训练目标是开发一款与 Westlaw 构成直接市场竞争关系的法律人工智能检索工具。同时，尽管 ROSS 主张其使用为“中间性复制”（intermediate copying），人工智能模型并不输出 TR 的表达内容，法院仍强调该使用并未具备足够“转化性”（transformative use），即未对原作品“添加新的表达、意义或信息”。其行为核心仍在于复制并抽取 TR 编辑成果用于同类市场，目的高度重合。因此，第一要素明确不利于 ROSS。其二，未对外展示亦不能否定“市场替代性”。第四因素即“市场替代性”成为法院否定合理使用的关键。法院认为，即便 ROSS 并未将 TR 内容直接展示于最终用户，其将 headnotes 作为训练数据使用，实质上可能侵蚀 TR 向第三方授权人工智能训练数据的潜在市场。这种人工智能语义向量化重构的行为，在市场影响层面与传统复制并无本质差异，构成对 TR 合理预期市场的侵蚀。尤其值得注意的是，法院明确指出合理使用不以是否“展示于公众”为标准，而应关注使用行为是否替代了著作权人原有或潜在市场。这一认定对当前普遍存在的“训练阶段非公开复制”辩护路径构成实质限制。其三，创作程度虽低，但仍不构成免责事由。虽然 TR 的 headnotes 为编辑性内容，创作自由度相对有限，但法院认为该等内容仍满足 Feist 案所确立的“最低创作性”标准，属于受版权保护的表达。其非纯粹事实信息，不属于“思想-表达合并”（merger）或“场景设定”（scènes à faire）下

²⁵ 第 107 条专有权的限制：合理使用 在确定特定情况下对作品的使用是否构成合理使用时，应考虑以下因素：（1）使用的目的和性质，包括此类使用是否属于商业性质或用于非营利性教育目的；（2）版权作品的性质；（3）所使用部分相对于整个版权作品的数量和实质性；以及（4）使用对版权作品的潜在市场或价值的影响。如果在考虑了上述所有因素后做出合理使用认定，则作品未发表这一事实本身并不妨碍该认定。

的不可保护内容。

本案确认即便是人工智能训练阶段的间接使用行为,也必须接受四因素审查,不能因“中间性”“非展示性”而免于侵权责任。法院尤其强调了转化性与潜在市场的核心地位,警惕人工智能使用行为以“训练”为名规避许可制度,从而构成“隐蔽性替代”。

（三）中国：杭互奥特曼案构成合理使用

2024年12月,杭州中院对上海某文化发展有限公司(以下简称某文化公司)与杭州某智能科技有限公司(以下简称某智能公司)著作权侵权及不正当竞争案作出二审判决,维持了一审中数据训练使用他人作品构成合理使用的认定。²⁶本案中,原告上海某文化公司经授权取得了奥特曼系列形象的著作权独占权利。被告杭州某智能科技有限公司运营的人工智能平台通过提供图生图、文生图服务及LoRA模型训练功能,允许用户上传奥特曼图片并基于平台提供的基础模型进行训练,生成奥特曼LoRA模型。训练完成后的奥特曼LoRA模型可被用户反复使用并在此基础上生成其他奥特曼图片并发布在平台上。某文化公司主张,某智能公司通过输入图片等数据进行训练后生成的方式将侵权图片和侵权模型置于信息网络中,侵害了某文化公司的信息网络传播权并构成不正当竞争。

司法裁判指出,人工智能对作品的使用核心目的在于将他人作品作为分析样本,用于提升模型创作能力的训练过程,而非再现作品本身具有独创性的表达形式。现有证据不足以证明该类使用行为对权利作品的正常市场利用构成实质性影响,亦未对著作权人合法权益造成不当损害,因而可以认定此类使用为合理使用。

四、中国人工智能数据训练合理使用方案

著作权法既需要平衡和协调好著作权人、人工智能研发方及社会公众之间的利益,促进社会文化繁荣,也需要通过合理使用等规则积极回应科技的创新发展,防止著作权法不恰当地阻碍技术进步。²⁷

（一）构建技术创新与版权保护的平衡机制

1.原则：不宜将数据训练活动统一纳入合理使用

人类创造者的未来面临生成式人工智能的风险,政策制定者必须保障创作者的权利并实施透明度规则。²⁸一方面,不宜将人工智能模型训练或预训练活动纳入著作权合理使用范畴。中国是包括《伯尔尼公约》《世界版权条约》(WCT)、《与贸易有关的知识产权协定》(TRIPS)等多个国际版权条约的缔约国,全面接受“国民待遇”与“最低保护标准”等国际版权规则。中国对本国著作权人作品的保

²⁶ 杭州互联网法院(2024)浙0192民初1587号民事一审判决书,杭州市中级人民法院(2024)浙01民终10332号民事二审判决书。

²⁷ 张吉豫,汪赛飞:《大模型数据训练中的著作权合理使用研究》,载《华东政法大学学报》2024年第4期。

²⁸ CISAC. Study on the economic impact of Generative AI in the Music and Audiovisual industries, <https://www.cisac.org/Newsroom/news-releases/global-economic-study-shows-human-creators-future-risk-generative-ai>.

护也必须延伸至其他缔约国权利人，其所作出的版权政策选择不仅具有国内适用效力，也可能通过条约机制对外发生法律影响。在此框架下，若中国在法律上明确将生成式人工智能训练纳入合理使用范围，允许人工智能开发者在未事先取得许可的情况下使用受保护的中文作品数据，理论上也将适用于《伯尔尼公约》的其他 179 个缔约方对中文作品的使用。一旦合理使用被滥用，或缺乏跨境数字内容使用的追踪机制，可能对中国本土的版权产业带来复杂影响。

2.建立对话协商机制

生成式人工智能企业与著作权集体管理组织的对话与协商机制，不仅有助于人工智能企业依法合规地获取训练数据，更为权利人提供了集体维权和收益分配的制度平台。大模型的训练依赖于海量的数据资源，其中包括大量受著作权法保护的类型作品。此类作品的使用不仅涉及复杂多样的著作权归属与许可问题，还面临难以逐一取得权利人授权的现实障碍。基于此，集体管理组织在数据许可与使用环节中发挥协调功能，能够有效缓解人工智能开发主体与权利人之间的信息不对称，降低交易成本，提升数据获取的合规性与效率。

（二）训练阶段数据训练合理使用边界

1.商业目的大模型数据训练：法定许可

针对人工智能开发企业在训练模型和商业应用过程中对受版权保护作品的使用，构建事前授权与事后补偿相结合机制。人工智能企业基于商业目的使用受版权保护作品前，应依法获得权利人或其代理的著作权集体管理组织的许可，并依据使用范围、方式和商业价值支付合理的版权使用报酬。探索建立人工智能著作权补偿金制度，人工智能企业可通过统一缴纳补偿金的方式获得作品使用权，具体金额可根据模型训练的规模、类型及用途，结合著作权集体管理组织进行协商与核定，并定期分配给作品权利人。

与此同时，应建立专门面向人工智能场景的版权纠纷调解机制，由行政机关牵头设立由法学专家、行业代表、权利人代表组成的第三方中立机构，开展调解、仲裁及专业意见审查，以缓解海量人工智能训练行为所带来的版权争议压力，强化预防性治理与程序性保障。

从国际经验来看，多个国家已对人工智能训练与著作权之间的冲突进行了制度性回应。在制度类型选择上，部分国家提出可将人工智能训练过程中的作品使用行为纳入法定许可制度的适用范围，即法律规定在满足特定条件下，使用者无需取得个别许可，但需支付报酬并履行登记、报告等义务。此类制度兼具规范性与灵活性，被认为是在人工智能产业发展与著作权保护之间实现平衡的次优解。法国通过强化集体管理组织作用、推动平台谈判和设定版权使用补偿机制，为生成式人工智能企业提供授权通道；在音乐与新闻作品领域，已通过集体许可协议解决数据使用争议，并探索合理报酬标准的立法设定。我国可在现行《著作

权法》第 24 条合理使用与第 23 条法定许可条款基础上进行功能性扩展，推动形成中国特色的人工智能著作权协调机制。

2. 非商业目的大模型数据训练：合理使用

为实现人工智能基础能力的公共性建设与创新生态的可持续发展，有必要在现有著作权制度框架内，为非商业性大模型训练行为探索建立适度的“合理使用”规则与开源豁免机制。在技术进步与公益性目标之间寻求制度性张力的动态平衡，既有助于保障教育科研机构、初创社区等主体对知识资源的合法可及，也可通过引导合规行为、防止滥用行为，为后续商业化开发留下制度空间。2018 年日本《著作权法》进行了修改，对于利用物联网、大数据、人工智能等技术进行创新，且对版权作品市场影响不大的创新作品的使用需求，设立了“灵活的权利限制规定”，以方便版权作品的使用。²⁹在其 2018 年修订的《著作权法》中确立了“信息分析使用”例外³⁰，允许包括企业在内的使用者在不向公众传播作品的前提下，对作品进行分析性使用，如模型训练。

其次，可借鉴开源软件治理经验，推动构建以开源人工智能内容协议为核心的训练数据共享机制。该机制应明确数据用途范围（如仅限研究用途）、作品著作权状态、衍生模型可否商业化等条款，从而为开源大模型或公益大模型训练提供版权层面的合规路径。例如，当前已有如 LAION、OpenSLR 等公益数据集，在明确版权状态与使用限制的基础上，允许在非商业前提下进行自由使用，其模式可为我国建立相关制度提供参考。中国在完善版权制度以适应人工智能发展的过程中，应考虑建立用途导向+主体限制+比例控制+退出机制相结合的非商业使用豁免规则。可先由政策或司法解释层面对“科研训练”的合理使用边界进行明确，再通过专门立法或授权机制，逐步探索人工智能领域的开源内容使用规范，推动我国人工智能技术生态的开源化、规范化与可持续发展。

（三）输出阶段信息披露与透明机制

在生成式人工智能完成模型训练并进入输出应用阶段后，模型生成的内容已开始面向社会传播，输出阶段的信息披露与透明机制，既是防范生成式人工智能黑箱化风险的关键，也是实现用户知情、自主判断与权利人维权的必要条件。人工智能生成内容的可识别性与可追溯性，直接关系到用户的知情权、权利人的维权可能性以及监管机构的治理能力。《暂行办法》第 12 条规定了生成式人工智能服务提供者的标识义务。³¹《人工智能生成合成内容标识办法》（以下简称《标识

²⁹ 日本文化厅：人工智能与版权，<https://www.bunka.go.jp/seisaku/chosakuken/aiandcopyright.html> 访问日期：2024 年 4 月 12 日。

³⁰ 日本《著作权法》第三十条之四（非以享受作品所表达的思想或情感为目的的使用）在以下情形中，任何人均可在被认为必要的范围内，以任何方式使用受著作权保护的作品，且不构成对著作权的侵害：当使用作品是为了电子计算机的信息处理，或为其他非以享受作品中所表达的思想或情感为目的，且该使用是以人类无法通过感知认知的方式进行的。但若该使用方式将对著作权人之利益造成不当损害，则不适用本条之规定。

³¹ 《生成式人工智能服务管理暂行办法》第 12 条 提供者应当按照《互联网信息服务深度合成管理规

办法》)第4条³²规定了提供生成合成服务的企业应首先确保对人工智能生成内容添加醒目的显式标识。《标识办法》第5条³³要求服务提供者按照深度合成管理规定第16条,在生成内容的文件元数据中添加隐式标识,鼓励服务提供者在生成合成内容中添加数字水印等形式的隐式标识。输出端标识机制强调模型输出阶段的信息透明与责任可追溯,构成对数据训练阶段合法性的重要支撑。通过强化生成内容的可识别性与可追溯性,反向推动训练阶段的数据使用趋于规范,推动大模型数据训练与输出管理“训练合规—输出可控”的双轮驱动治理,从而实现对合理使用边界的动态平衡。

五、结语

生成式人工智能大模型的数据训练机制,正在深刻改变人类创作方式,也引发数据训练合理使用规则争议。欧盟、美国、中国等相关诉讼不断涌现,司法裁判在实践中试图厘清合理使用的边界。未来,中国亟需构建具有前瞻性的制度回应机制。一方面,应完善训练端的合理使用规范,以制度性授权的形式明确可使用数据类型及条件;另一方面,也应建立输出端的信息披露机制,增强人工智能生成内容来源的透明度与可追溯性。未来的数据训练合理使用规则的边界应在尊重著作权人合法权益的同时,保障人工智能发展的空间,以实现创新与规范的共生共赢。

定》对图片、视频等生成内容进行标识。

³² 《人工智能生成合成内容标识办法》第四条 服务提供者提供的生成合成服务属于《互联网信息服务深度合成管理规定》第17条第1款情形的,应当按照下列要求对生成合成内容添加显式标识:(一)在文本的起始、末尾或者中间适当位置添加文字提示或者通用符号提示等标识,或者在交互场景界面、文字周边添加显著的提示标识;(二)在音频的起始、末尾或者中间适当位置添加语音提示或者音频节奏提示等标识,或者在交互场景界面中添加显著的提示标识;(三)在图片的适当位置添加显著的提示标识;(四)在视频起始画面和视频播放周边的适当位置添加显著的提示标识,可以在视频末尾和中间适当位置添加显著的提示标识;(五)呈现虚拟场景时,在起始画面的适当位置添加显著的提示标识,可以在虚拟场景持续服务过程中的适当位置添加显著的提示标识;(六)其他生成合成服务场景根据自身应用特点添加显著的提示标识。服务提供者提供生成合成内容下载、复制、导出等功能时,应当确保文件中含有满足要求的显式标识。

³³ 《人工智能生成合成内容标识办法》第5条 服务提供者应当按照《互联网信息服务深度合成管理规定》第16条的规定,在生成合成内容的文件元数据中添加隐式标识,隐式标识包含生成合成内容属性信息、服务提供者名称或者编码、内容编号等制作要素信息。

鼓励服务提供者在生成合成内容中添加数字水印等形式的隐式标识。

文件元数据是指按照特定编码格式嵌入到文件头部的描述性信息,用于记录文件来源、属性、用途等信息内容。

Research on reasonable use rules for artificial intelligence data training: a comparative analysis of China, the United States and Europe

Abstract: The use of massive amounts of data in the training process of large generative AI models such as ChatGPT and Deepseek has sparked copyright disputes. The EU's Artificial Intelligence Act and the Digital Single Market Copyright Directive's 'implied licence + opt-out + text and data mining exception', the 'transformative use' principle in US case law, and the normative differences in the reasonable use provisions of China's Copyright Law. The EU, the United States and China have taken different paths in their relevant legal systems and judicial practices. China's future institutional design should seek a reasonable balance between promoting technological innovation and protecting copyright. It should focus on the review of the legality of data sources, consideration of the public welfare of the purpose of use, and a rights compensation mechanism. It should also improve the reasonable use norms at the training end and the information disclosure mechanism at the output end, and promote the two-wheel drive governance of 'training compliance-output control' for large model data training and output management, so as to achieve a dynamic balance of the boundaries of reasonable use.

Keywords: generative artificial intelligence; Deepseek; data training; fair use; copyright

WTO 法下人工智能产品属性认定研究

吴梓杰 赵汉卿¹

摘要：人工智能模型的涌现为国际法领域带来了“人工智能中的国际法”这一议题。在国际贸易领域，人工智能首先对传统的“商品-服务”二分法提出了挑战。一方面，人工智能系统本身作为一种“数字产品”，其“数字产品”的属性长期以来在 WTO 法领域内的归类问题饱受争议，无论是学界、还是 WTO 业界对其均无普遍之共识。另一方面，人工智能产品嵌入至硬件设备中，使得这类传统的“硬件产品”同时具有了潜在的“服务属性”。本文以此为基本背景，归纳了过往学界、实务界的相关理论、判例，试图以一种“类型化”的方法，对人工智能属性划分以及 WTO 法中规则适用问题探寻一条可行进路。

关键词：人工智能 WTO 法 AIoT 数字服务

一、问题的提出

人工智能这一数字经济的新业态，不同于过往的电子商务、大数据等技术手段，其对人类生活的影响将深入到日常生活的方方面面，对国际贸易的影响也更为复杂。世界贸易组织（WTO）在其最新发布的《ChatWTO: 生成式人工智能与国际贸易分析白皮书》中便指出生成式人工智能在国际层面在传播和伦理、就业和环境等五个领域造成一定影响。²基于上述挑战，各国政府都在尝试通过监管手段调整人工智能造成的潜在风险。

不同的国家试图通过不同的角度、手段对不同领域内的人工智能实行治理，由此导致了人工智能治理规范“碎片化”的现象。³WTO 在一份 2024 年发布的报告中特别指出了人工智能治理的“碎片化”是一个需要关注的问题。⁴所谓“制度碎片化”（fragmentation）系指以某一特定问题领域为重点，但在范围、成员和通常规则上各不相同的国际机构的拼凑而成。⁵对应到国际贸易领域，当前，与人工智能产品贸易相关的规则，如人工智能伦理问题、数据跨境问题、隐私与知识产权

¹ 作者简介：吴梓杰，上海对外经贸大学国际组织学院（贸易谈判学院）法与经济学专业硕士研究生，邮政编码：201620，通信地址：上海市松江区文翔路 1900 号，联系电话：18876599357，电子信箱：24358002@suibe.edu.cn、florian492@outlook.com。赵汉卿，上海对外经贸大学国际组织学院（贸易谈判学院）国际经贸规则专业本科生，邮政编码：201620，通信地址：上海市松江区文翔路 1900 号，联系电话：13391255331，电子信箱：13391255331@163.com。

² ChatWTO: An Analysis of Generative Artificial Intelligence and International Trade 2024, World Economic Forum, <https://www.weforum.org/publications/chatwto-an-analysis-of-generative-artificial-intelligence-and-international-trade/> (last visited Dec. 2, 2024).

³ Peter Cihon, Matthijs M. Maas & Luke Kemp, *Fragmentation and the Future: Investigating Architectures for International AI Governance*, 11 Global Policy 545 (2020).

⁴ *Trading with Intelligence: How AI Shapes and Is Shaped by International Trade*, World Trade Organisation, 2024, p. 46.

⁵ Frank Biermann, Philipp Pattberg, et al., *The Fragmentation of Global Governance Architectures: A Framework for Analysis*, 9 Global Environmental Politics 14 (2009).

的保护等散落在 CPTPP、DEPA、欧盟颁布的《人工智能法案》与《通用数据保护条例》等各个细碎的法案当中。因此,就“人工智能贸易”这个议题,存在多个与之相关的治理规则、治理结构,且参与到某个规则当中的国家同时也可能是其他规则的参与方,故人工智能贸易制度“碎片化”的问题尤为凸显。

制度碎片化可能导致不同的制度规则之间在治理目标上可能产生冲突。就国内立法而言,欧盟通过《人工智能法案》将人工智能产品进行分类,并针对不同的类别施加了不同的透明度义务。而美国则希望推动其国内的人工智能产业发展,因此相比较于欧盟,对其国内的人工智能服务提供者施加了较轻的义务。不同治理思路的背后反映了不同国家和地区的自身利益和社会价值观,而这些不同的框架之间的区别可能导致人工智能国际治理的冲突:对国家而言,另一个国家实施的人工智能政策可能会被视为不正当的贸易壁垒;对企业而言,不同的监管政策对跨国人工智能服务提供者带来了高昂的合规成本。⁶这些成本则可能最终转嫁到消费者身上,造成消费者福利的损失。以一言以蔽之,制度碎片化导致重复监管、监管脱逃等问题,提高了治理成本。考虑到区域间的贸易安排应当与承认这些区域安排的多边协议规则相兼容。⁷因此,需要建立一个全球范围内的、统一的人工智能国际贸易治理机制,调和不同国家之间在规制思路、目标上的冲突,以中心化的形式,降低人工智能的治理成本,实现人工智能国际贸易的有效治理。

但中心化治理结构的缺陷表现在“创建这个组织机构本身是一个缓慢的任务”。⁸然而,国际规则、组织的形成具有路径依赖的特性。有鉴于此,面对人工智能对国际贸易的挑战,诉诸于既存的国际间贸易机制,而非创立一个新的体制,是成本最小化思想下的一个理性选择。作为当前世界范围内最大的国际间贸易组织,世界贸易组织因其具有的“准立法、准司法与准行政”功能为人工智能时代的跨国贸易争议解决提供了一个现存的框架。⁹但人工智能产品模糊了传统“商品-服务”二分法下的分类规则。¹⁰考虑到“商品-服务”二分法是整个 WTO 规则制定的基础,在限制人工智能的贸易措施出现时,如无法确认其违反的是《货物贸易总协定》(下称 GATT)还是《服务贸易总协定》(下称 GATS)项下的义务,则依托于 WTO 考虑人工智能的国际贸易问题则无从谈起。本文针对人工智能产品

⁶ Emily Jones, *Digital Disruption: Artificial Intelligence and International Trade Policy*, 39 Oxford Review of Economic Policy 70 (2023).

⁷ Giorgio Sacerdoti, *WTO Law and the “Fragmentation” of International Law: Specificity, Integration, Conflicts*, in Merit E. Janow, Victoria Donaldson, & Alan Yanovich eds., *The WTO: Governance, Dispute Settlement & Developing Countries*, Juris Publishing, Inc., 2008, p. 2.

⁸ Supra note 2, Peter Cihon, Matthijs M. Maas & Luke Kemp.

⁹ 郑蕴:《国际贸易法体系的碎片化结构:历史性变局、主要矛盾与中国对策》,载《国际经济法学刊》2023年第3期。

¹⁰ 例如,人工智能模型的训练需要海量的数据来源,这要求数据在一定程度上是可自由访问的,从这个角度看人工智能的贸易具有“服务”属性。但如果我们希望人工智能去完成一些更为复杂的工作,例如设计、制造机械产品,在生物医学领域分析蛋白质结构等等,则人工智能模型必须被部署到硬件设备中,这使得人工智能的时代“服务”和“商品”之间的区别逐渐变得模糊。这一特性是随着“制造服务化”

(Servitization of Manufacturing) 的商业模式特点所决定的。see Shin-yi Peng, *A New Trade Regime for the Servitization of Manufacturing: Rethinking the Goods-Services Dichotomy*, 54 Journal of World Trade (2020).

对 WTO“商品-服务”二分法的挑战，通过对人工智能产品类型化区分的尝试，提出了一条认定人工智能产品属性的可行进路。

二、人工智能产品的基本类型

（一）“嵌入式”的人工智能产品

依照欧盟《人工智能法案》对“人工智能系统”的定义：“AI 系统是指一种基于机器的设计系统，可以实现不同程度的自主运行，并且在部署后可能表现出适应性。该系统根据接收到的输入，推断出如何生成预测、内容、推荐或决策等输出，这些输出能够影响物理或虚拟环境。”¹¹如果一件 AI 产品能够对“物理环境”产生影响，其必须具备“足以影响物理世界的基本能力”。换言之，这种 AI 系统不可能是部署于云端的一种“智能问答助手”，他必须具备属于其自身的，可以和外界实现交互的硬件设施，才有可能对“物理世界”造成影响。意即：这是一种部署于硬件设备中的 AI 系统，我们将其称之为是“嵌入式”的 AI 系统。

嵌入式的 AI 系统在日常生活中十分常见，例如特斯拉公司生产的电动汽车具备自动驾驶的功能，能够实时地识别路况，并及时作出相关的回应，这就是一种典型的“嵌入式”AI 系统。除此以外，各种智能手表可以实时检测人类的心跳、血糖等生物信息，并提供相应的反馈或者医疗建议，这也属于“嵌入式”AI 系统的范畴。

（二）非“嵌入式”的人工智能产品

除了“嵌入式”的 AI 系统以外，AI 产品还可能部署于云端，通过互联网技术实现传播，我们将其称之为是“非嵌入式”的 AI 产品。¹²非嵌入式的产品在日常生活中也十分常见。最典型的例子是由 OpenAI 公司提供的 ChatGPT。ChatGPT 有两种访问形式：其一是消费者可以直接在其官网中与之对话，实现交互；其二是消费者可以购买 OpenAI 公司提供的互联网接口（API），通过 API 在消费者自身设计的软件程序中，调用 ChatGPT 与之实现交互。无论是何种形式，都不涉及“实体”的传输，二者在本质上是数据通过互联网的传播。故将其认定为是“非嵌入式”的产品。

欧盟在立法中，支持了此归类模式。作为全球数字立法的先行者，欧盟于 2019 年通过了《关于提供数字内容和数字服务的合同特定方面的指令》（下称《数字合同指令》）以及《关于货物买卖的合同特定方面指令》（下称《货物买卖合同指令》），聚焦提供数字内容和数字服务两类交易内容。¹³欧盟各国随即展开相应

¹¹ Article 3(I), European Union AI Act.

¹² 需要特别指明的是：嵌入式与非嵌入式的区别，是针对消费者而言进行划分的。如果某项 AI 系统部署于消费者所有的，传统意义上的“物”上，我们更倾向于认为这是一种“嵌入式”的 AI 产品。如果该 AI 部署于非消费者所有的“物”上，通过互联网技术使得其自身得以满足消费者的效用，则我们倾向于认为这是一种“非嵌入式”的 AI 产品。

¹³ 金晶：《数字交易的欧盟路径与德国模式——《德国民法典》数字产品合同规则对债法的“补丁式”更新》，载《欧洲研究》2023 年第 3 期。

立法工作，例如德国于 2022 年修订《德国民法典》，在债之关系法中新增第 327 条“数字产品合同”（Verträge über digitale Produkte）。¹⁴此轮立法值得镜鉴之处在于：欧盟、德国立法中针对“数字内容”和“数字服务”设立了统一的上位概念，即“数字产品”（digitale Produkte），此规范结构与 CPTPP 等世界上主要经贸规则在形式上较为类似。其次，德国针对《数字合同指令》的转化中，充分接纳了“技术中立”的原则，其背后之法理依据与现行 WTO 判例法相吻合。¹⁵

其一，数字内容（Digitale Inhalte）。《德国民法典》327 条第 1 句将“数字内容”界定为：以数字形式创建和提供的数据。¹⁶数字内容的关键之处在于其以“数据形式出现”的属性。¹⁷因此，在认定一件产品是否属于“数字内容”时，关键之处在于考虑其是否以“数据”的形式被提供，至于其内容如何，则不在考虑的范畴内。¹⁸其二，数字服务（Digitale Dienstleistung）。327 条第 2 句对“数字服务”做了列举式的说明，简而言之系指“使消费者能够使用数字内容的服务。”¹⁹

依照该归类方法，可以很好契合当前非嵌入式的 AI 产品的主要交易方法。还是以 ChatGPT 为例：（1）当消费者购买 ChatGPT 会员，通过互联网技术在线访问 ChatGPT，实现交互时。此时，在本质上是消费者通过互联网以及 OpenAI 的平台作为中介，访问了电子形式的大模型，属于“数字服务”的交易。（2）消费者通过 API 技术访问大模型，或将大模型部署在自己开发的软件中，此时属于通过 API 作为中介，又或者消费者可以直接将大模型下载至本地。此时，消费者拥有的是“以数据形式存在的模型本身”，属于数字内容的交易。

三、人工智能产品性质认定的规则现状

非嵌入式的人工智能产品的跨境贸易与数字产品具有高度的相似性，毋宁说人工智能在一定程度上是“数字产品”的交集之一。因此，参考当前学界对“数字产品”属性认定的探讨对研究“非‘嵌入式’人工智能产品”的认定颇有裨益。当前国际经贸立法与实践中，对于数字产品的属性认定主要存在服务说²⁰、货物说²¹以及两种不同的观点。对于嵌入式的人工智能产品，各国普遍将其视为货物，不存在认识上的重大分歧。但由于该货物具有“服务”属性，因此学者主张将其视为“混合产品”²²。在此，需要考虑的问题是：一项限制“嵌入式人工智能产品”的措施，应当受到 GATT 还是 GATS 的约束。

（一）人工智能（数字）产品的货物属性

¹⁴ §327 ff. BGB

¹⁵ 同前注〔12〕，金晶书。

¹⁶ §327 I BGB.

¹⁷ Ulber und Bernzen/Specht-Riemenschneider, in: Erman Kommentar, 17. Aufl. 2023, S. 1539 Rn 16.

¹⁸ 同前注〔12〕，金晶书。

¹⁹ Helmut Köhler, BGB Allgemeiner Teil, 46. Aufl. 2022, S. 314 Rn 1.

²⁰ Sam Fleuter, *The Role of Digital Products Under the WTO: A New Framework for GATT and GATS Classification*, 17 Chicago Journal of International Law (2016).

²¹ Farrokh Farrokhnia & Cameron Richards, *E-Commerce Products Under the World Trade Organization Agreements: Goods, Services, Both or Neither?*, 50 Journal of World Trade (2016).

²² Anupam Chander, *The Internet of Things: Both Goods and Services*, 18 World Trade Review S9 (2019).

美国等部分国家将人工智能或数字产品认定为货物的理由在于：GATT 并未将其自身的适用范围限制在“有形”的产品当中，因此不能由于数字产品的“无形”性，就排除 GATT 的适用。²³其次，依照“技术中立”的相关理论、判例：既然在服务贸易领域 WTO 判例法支持了“服务贸易之提供与其提供形式无涉”的原则，²⁴那么在货物贸易领域，似乎没有理由推翻这一原则。第三，从学理看：首先，货物作为法律意义上的“物”的本质属性，乃其背后所凝结之一定的人类劳动成果。作为数字内容的人工智能产品的背后，亦凝结有一定的人类劳动。那么认定其为“货物”是较为妥当的一种做法。其次，根据 Peter Hill 的观点：“货物”的本质是作为一个“实体”，权利人可以在“货物”上建立所有权，以供权利人重复使用；而对于一项服务，在给付完成后权利人对于相对人的给付请求权随之消灭。²⁵在这个意义上，如果某项人工智能产品是可以被下载并重复使用的（例如各大人工智能公司都提供了可部署在本地的预训练模型），因而应当归类为“货物”。

然而，仅从 WTO 判例法对 GATT 的适用范围看，当数字产品（人工智能产品）是以无形的互联网手段提供时，将数字产品认定为货物的理由似乎不能成立。从专家组在 DS236 认为“货物”一词的普通含义是指“除货币外的有形或可移动的个人财产”。²⁶因此，对于以无形方式提供的数字产品或人工智能产品，主张 GATT 并未将其自身的适用范围限制在“有形产品”中的抗辩不能够得到 WTO 判例法的支持。其次，针对“技术中立”这一原则而言，无论是 GATT 还是 GATS 都未曾将该原则以制定法的方式予以明确，因此，“技术中立”仅仅是 WTO 裁判过程中予以说理的依据，恐怕不是真正意义上的 WTO 法律渊源。²⁷综上所述，恐怕不能仅凭以上论据就将人工智能（数字）产品归类为货物。

美国主张人工智能（数字）产品贸易应当适用 GATT 背后的经济根源在于：美国自身的数字产业在世界范围内极具优势。如采用 GATT 模式，则各国难以对美国数字产品进入该国市场施以限制性措施。因此，GATT 模式将更有助于美国发挥自身的产业优势，通过数字产业获得更多的经济利润。²⁸然而，这对于自身数字产业本就脆弱的第三世界国家而言，GATT 下自由贸易的模式将导致自身数字产品市场受到外部产品的冲击，将不利于提升本国的脆弱产业，恐怕将使得这部分国家陷入数字贸易的“比较优势陷阱”中。

（二）人工智能（数字）产品的服务属性

中国、欧盟以及其他一些第三世界国家普遍主张在“服务贸易”也即 GATS 的

²³ 陆一戈：《数字产品的非歧视待遇问题研究》，载《武大国际法评论》2024 年第 1 期。

²⁴ Anupam Chander, *Artificial Intelligence and Trade*, in M. Burri ed., *Big Data and Global Trade Law*, Cambridge University Press, 2021, p. 120.

²⁵ Peter Hill, *Tangibles, Intangibles and Services: A New Taxonomy for the Classification of Output*, 32 *Canadian Journal of Economics* 426 (1999).

²⁶ Report, United States — Preliminary Determinations with Respect to Certain Softwood Lumber from Canada, WT/DS236/R para 2(a).

²⁷ PR, United States-Measures Affecting the Cross-Border Supply of Gambling and Betting Services, para.6.671..

²⁸ 罗云开：《货物抑或服务：数字产品属性探究》，载《社会科学研究》2023 年第 6 期。

语境下讨论人工智能（数字产品）的贸易问题。其理由在于：第一，GATS 对贸易自由化的约定相比较于 GATT 而言较为宽松，成员国可以自行决定是否开放相应服务。将人工智能（数字产品）的贸易问题纳入 GATS 的范畴下更有利于减轻市场开放对成员国内部数字市场的外生冲击。其次，人工智能产品的贸易可能伴随隐私保护、数据安全等公共政策问题，而 GATS 第 14 条一般例外条款允许成员方采取其认为必要的任何措施来实现某些公共政策目标，因此在将人工智能产品归类为服务，将更有利于促进此类目标的实现。²⁹

在学理上，当前学界通过“技术中性”和“可接触性”普遍将数字产品贸易纳入到服务贸易的框架下。一方面，当前数字产品贸易的主要形式是通过互联网实现产品的在线提供，因而符合了“跨境服务”的基本特征。进一步的，数字产品普遍不具有物理性质的实体属性，因而不符合 WTO 专家组就 GATT 中“产品”须具有实体属性的定义。综上所述，将数字产品认定服务是符合 WTO 的基本框架的。

（三）人工智能（数字）产品的混合属性

在当前的国际贸易问题的相关文献中，由于人工智能产品自身与数字经济的高度融合性，因此主要在数字贸易的相关范畴下进行讨论。然而，世界当前主要经贸规则中对于数字经济的内涵普遍定义不清，甚至有时采取回避的态度：以 CPTPP 为例，其将与人工智能贸易相近的“数字贸易”定义为：“以数字编码形式制作、用于商业销售或分发的计算机程序、文本、视频、图像、音频记录或其他产品，并可以通过电子方式传输。”³⁰然而，除“通过网络传播”外，人工智能产品的贸易还可能以有型的方式开展：例如一台部署至硬件设备中的“法律服务机器人”从 A 国出口至 B 国，从性质上看该产品具有“有形性”，因此应当适用 GATT 的有关规定，但该产品进入市场后在事实上承担的是类似于人类律师的法律服务提供职能，因此也存在 GATS 的解释空间。³¹应当认识到，国际贸易实务与学界对“数字贸易”的认识也是在贸易数字化的进程中不断改善、进步的。因此，当数字贸易进入人工智能时代，笔者以为：过往针对于“数字产品”的贸易法规、判例，恐怕难以因应人工智能产品的贸易纠纷。³²

不宜将人工智能产品贸易与产品贸易混为一谈的理由首先在于：人工智能产品贸易的概念显然要远远超过数字产品贸易的内涵。以上述 CPTPP 为例，若采

²⁹ 同前注〔22〕，陆一戈书。

³⁰ CPTPP, Chapter 14. Art 14.1

³¹ 实际上，对“货服二分”的挑战并非人工智能时代新产生的议题。早在上世纪 80 年代便有针对“搭载了电子芯片的卡西欧手表是否因其提供了‘日历查看’服务因而需要受到 GATS 的规制”的疑问。然而细看此类问题不难发现其荒谬之处：如果采用此种标准构造货物和服务之间的界限，那么任何产品的背后都蕴含了一定的人类劳动，此劳动得满足一定人之效用；由此，“货物-服务”之间的界限将不复存在。由是观之，人工智能产品在货服二分下的归类模糊在一方面是人工智能产品对传统国际贸易法体系的挑战；另一方面则是以 WTO 法为代表的国际贸易法在创制伊始便对货物、服务之内涵、界限等关键性问题语焉不详，由此引致技术进步对自身体系的冲击。

³² Han-Wei Liu & Ching-Fu Lin, *Artificial Intelligence and Global Trade Governance: A Pluralist Agenda*, 61 *Harvard International Law Journal* (2020).

CPTPP 就数字产品贸易之概念对作为人工智能产品贸易的范围，则例如智能机器设备等搭载了人工智能系统等硬件设施，将不属于“数字贸易”的范畴。人工智能律师、投资顾问师等亦是如此。正是由于人工智能产品自身的特殊属性，使得“货服二分”下的传统国际经贸规则出现了适用上的模糊空间，因此，有必要审视人工智能对“货服二分”体系的挑战，从贸易的基本形式上对不同类型的人工智能产品贸易进行划分，不宜一概而论，将人工智能贸易统摄为货物或服务的一种。

四、WTO 法下人工智能产品归类问题的解决进路

（一）前提：WTO 法对人工智能产品跨国贸易的可适用性

讨论人工智能产品在 WTO 规则下的分类问题的前提在于：影响人工智能自身或其衍生物贸易的措施确受到 WTO 规则的约束。有学者的观点认为，由于 WTO 规则在创制时起草者并未认识到人工智能产品的存在，且 WTO 文本中也并未提到人工智能的语词，因此 WTO 规则不能够适用于人工智能产品的国际间贸易。³³但这种观点并不能够成立——本文认为：结合人工智能的技术路径和 WTO 规则的文本，尤其是 GATT 和 GATS 的文本，以及过相关规则的立法目的等，可以得出的结论是：WTO 规则的确可以适用于影响人工智能产品贸易的措施。

人工智能产品的组成成分可以被划分为硬件、数据以及算法（软件）三大部类。³⁴倘若影响硬件、数据、算法的措施都可以受到 WTO 规则的约束，那么没有理由认为其最终组成的人工智能产品不能为 WTO 规则所约束。就硬件而言，如果某个人工智能算法已经被部署到硬件设备上（AIoT），我们当然可以直接将其视为是一种商品，进而通过 GATT 规制影响了人工智能产品的贸易措施。当然这种方法也可能产生一定争议，例如人工智能产品因其自身之特性及其用途的差异，究竟应当归属于“商品”还是“服务”在部分学者看来是一个尚无定论的问题。³⁵但无论我们将人工智能认做是“商品”还是“服务”，其结果都必然使得影响人工智能的贸易措施可以纳入到 GATT 或 GATS 的语境下。因此，对于影响了搭载了人工智能系统的硬件设施的措施，适用当前的 WTO 规则在解释论上不存在障碍。

在数据、软件的跨境交易层面，制定于前互联网时代的 WTO 规则时常遭受

³³ see Supra note 23, Anupam Chander, p. 120.

³⁴ see Giovanni Briganti Dini, *The EU's Response to the Fragmented Emergence of Artificial Intelligence*, in O. Costa, E. Soler i Lecha & M. C. Vlaskamp eds., *EU Foreign Policy in a Fragmenting International Order*, Springer Nature Switzerland, 2025. 具体而言，依照《欧盟人工智能法案》2.5g(1)条对人工智能产品的定义，人工智能产品应当是一种“基于机器”的系统，在部署后具有一定的自主运营能力，并且可以进行推理、预测，并足以对现实世界造成一定影响。基于上述定义，不难归纳出人工智能产品应当具有这三个组成成分。其中，硬件设备是人工智能得以与现实世界交互的物理基础，算法是人工智能决策的逻辑基础，数据则是沟通交互和决策的桥梁。持相同观点者，又见 Ching-Fu Lin, Shin-yi Peng & Thomas Streinz eds., *Artificial Intelligence and International Economic Law Disruption, Regulation, and Reconfiguration*, Cambridge University Press, 2021, p. 4.

³⁵ 参见戴艺晗：《国际贸易法视域下的人工智能规制——以 WTO 规则为视角》，载《上海财经大学学报》2023 年第 2 期；Dr Abid Shah & Ximei Wu, *Towards AI-Driven Trade Regulation: Regulatory Challenges and Global Commerce Within WTO Framework*, Social Science Research Network, 2024.

到无法跟上快速发展的数据交易环境的质疑。³⁶但 WTO 规则在当前新业态跨境交易中的滞后性不等同于其规则本身无法被适用于数据跨境交易当中。人工智能以无形的方式的跨境部署与电信服务存在一定的相似性，本质上都是“数据”的跨国流通。从 GATS 规则的文本看，除了正文规定的一般规则和义务外，还两个附件涉及电信服务，其中一个附件是 GATS 附件 5。该附件第 2 条（a）款明确了该附件的适用范围包括“所有影响成员国进入和使用公共电信运输网络和服务的措施”。³⁷从“利用公共电信运输网络和服务”自身的含义看，附件 5 第 3 条（b）款进一步明确到“此类服务除其他外可包括电报、电话、电传和数据传输”。³⁸因此，数据交易问题可以纳入到 GATS 附件 5 当中。更进一步地，根据 GATS 第 XXIX 条，GATS 所包含的附件都是 GATS 的组成部分，³⁹因此附件中的规则应与 GATS 的其他部分一并解读。综上所述，数据的跨境流动应当被视为一种服务，故数据跨境贸易问题可以纳入到 GATS 中进行讨论。⁴⁰

其次，人工智能产品可以不依赖于实体形式而以“软件”的形式跨境部署，并完成原本应当由人类完成的工作。因此，当 GATS 中已经为成员国所承诺的服务，基于数据技术，以某种全新的形式出现，对这些服务的限制性措施是否违反 WTO 义务？GATS 将服务划分为：跨境服务、跨境消费、商业存在和自然人流动。⁴¹从解释论的角度看，能够直接适用于数据要素跨境流动的，当属第一类服务，也即从某成员境内向其他成员境内提供服务。当然在第二类、第三类服务类型中也可能涉及数据流动的问题。⁴²在 *US-Gambling* 一案中，安提瓜和巴布达（安提瓜）认为美国联邦和州法律的许多条款，特别是《电线法》、《旅行法》和《非法赌博业务法案》法律在事实上对互联网赌博实施了有效禁令——这实际上排斥了安提瓜的服务提供商，而美国在其承诺表中对此作出了承诺——因此违反了 GATS 第 XVI 条。⁴³专家组指出：“GATS 模式 1 包括从一个世贸组织成员的领土向另一个世贸组织成员的领土提供服务的所有可能方式。因此，模式 1 的市场准入承诺意味着其他成员的供应商有权通过所有交付方式提供服务，无论是邮寄、电话、互联网等，除非成员的附表另有规定。我们注意到，这符合‘技术中立’原则，而

³⁶ see Andrew D. Mitchell & Neha Mishra, *WTO Law and Cross-Border Data Flows: An Unfinished Agenda*, in M. Burri ed., *Big Data and Global Trade Law*, Cambridge University Press, 2021, p. 83.

³⁷ General Agreement on Trade in Services, Annex 5, 2(a)

³⁸ General Agreement on Trade in Services, Annex 5, 3(b)

³⁹ General Agreement on Trade in Services, Art XXIV.

⁴⁰ see Rüdiger Wolfrum, Peter-Tobias Stoll & Clemens Feinäugle, in *WTO - Trade in Services*(Max Planck Commentaries on World Trade Law), Brill Nijhoff, 2008, p. 687, Rn. 1.

⁴¹ General Agreement on Trade in Services, Art I.2.

⁴² see Andrew D. Mitchell & Jarrod Hepburn, *Don't Fence Me In: Reforming Trade and Investment Law to Better Facilitate Cross-Border Data Transfer*, *Yale Journal of Law and Technology* (2018).对此也有学者持不同观点，例如 Maria Diory F. Rabajante 认为人工智能贸易无法为 GATS 当前所规定的 4 种服务类型所涵摄。see Maria Diory F. Rabajante, *Artificial Intelligence and International Trade Law*, in A. Poorhashemi ed., *Artificial Intelligence and the Future of International Law: Bridging Rights, Trade, and Arbitration*, Springer Nature Switzerland, 2024.

⁴³ see Douglas A. Irwin & Joseph Weiler, *Measures Affecting the Cross-Border Supply of Gambling and Betting Services (DS 285)*, 7 *World Trade Review* 71 (2008).

这一原则似乎在很大程度上得到了世贸组织成员的认同。”⁴⁴ 究其深层原因，在于“以不同方式提供的服务，不足以使得成员国逃脱其在 WTO 中所应遵守的义务”。⁴⁵毋宁说 GATS 创立之初的目的，在于规制“限制服务贸易的措施”，⁴⁶而不考虑为该措施所限制之服务贸易以何种方式提供。⁴⁷这一观点也得到了 WTO 判例法的承认：在 *US-Gambling* 案中，上诉机构虽未能回答有关 GATS 准确适用于电子服务提供以及各种数字产品分类的问题，但承认了该案中通过电子方式提供的服务属于 GATS 所规定的模式 1 项下的服务类型。⁴⁸该案明确了 WTO 规则的确可以适用于电子商务和数字产品的跨境贸易中，这一裁决进一步地为中国影音案所支持：在中国影音制品案中，上诉机构裁定，“如果一部电影的内容是通过实物交付材料传播的，则中国的《电影管理条例》将不可避免地规定谁有权进口这些商品，因为电影的内容是通过实物表达和嵌入的。”⁴⁹该裁决实际上表明了，无论“电子产品”本身是否是一件产品，只要电子产品以实物的形式出现，那么其就在事实上“包含在了一件产品”中。因此，结合 WTO 文本和专家组、上诉机构做出的裁决，本文认为：当前人工智能本身及人工智能产品的跨境交易可以被纳入到 GATS 或 GATT 的规则当中。

（二）人工智能产品属性判定的基本路径

（1）嵌入式的人工智能产品应当属于一种“具有服务属性”的货物

就“嵌入式”的 AI 系统而言，将其视为“货物”的核心逻辑根植于其物理实体的本质属性与国际贸易规则体系的交互作用。从物质层面看，嵌入式人工智能系统作为具备物理形态的工业制成品，其硬件构成（芯片、机械结构、电子元件等）完全符合 WTO《关税与贸易总协定》（GATT）对“货物”的界定标准，必须通过海关申报、缴纳关税并受技术性贸易壁垒约束，这一特性在特斯拉电动汽车、苹果智能手表等产品的全球供应链管理中体现得尤为显著。从 WTO 的相关判例看：在“加拿大期刊案”中明确以纸质载体的物质性优先于广告服务内容作为判定依据，类似原则延伸至人工智能系统领域时，即便设备附带数据服务，其作为工业制品的实体属性仍不容抹去。故将其视为一种“货物”，即符合当今国际贸易的基本情况，也契合 WTO 的既有判例。

（2）非嵌入式的人工智能产品属于服务

首先，无论是将其视为“货物”还是“服务”，在既有的讨论中，都是一种“基于 WTO 规则”的讨论。然而，数字产品、人工智能的出现，对制定于前互联网时代的 GATT 和 GATS 提出了根本性的挑战。无论将其视为货物还是服务，仅仅基于

⁴⁴ PR, United States - Measures Affecting the Cross-Border Supply of Gambling and Betting Services, para 6.285.

⁴⁵ Ibid, para. 3.150.

⁴⁶ General Agreement on Trade in Services, Art I.

⁴⁷ see Supra note 23, Anupam Chander, p. 120, Fn. 15.

⁴⁸ Supra note 41, Maria Diory F. Rabajante, p. 34.

⁴⁹ ABR, China - Publications and Audiovisual Products, para 188.

“法律规则”而讨论，恐怕难以得到具有最终意义上的结论。依照法理学家德沃金的观点，法律规则是以一种“全有或全无”的方法适用的。⁵⁰在这种分析思路下，将“非嵌入式”的 AI 系统视为是“货物或服务”，必然要给出具有绝对说服力的理由，排除另一者的适用，然而在当前的学理与判例的范围内，任何一方都无法给出这个理由。因此，我们主张一种“基于 WTO 根本原则”的分析方法，因为法律原则是以一种“优先性”的方法适用的。⁵¹质言之，在这种分析思路下，将“嵌入式的 AI 系统”视为服务，并不是因为其在绝对意义上不能被视为货物，而是将其视为服务，进而受到 GATS 纪律的约束，相比较于将其视为“货物”，将更有利于促进 WTO 基本原则、目的的实现。

GATT 与 GATS 中，均包含了“促进贸易自由化”的语词或与之类似的描述，但 GATT 的前言指出：国际贸易应“旨在提高生活水平，确保充分就业和大量且稳步增长的实际收入和有效需求，充分利用世界资源，并扩大商品的生产和交换。”⁵²而这些目的是通过“缔结互惠互利的安排，旨在实质性地减少关税和其他贸易壁垒，并消除国际贸易中的歧视性待遇，从而为这些目标做出贡献。”⁵³因此，应当认识到，贸易自由化应当有助于“提高生活水平、促进充分就业”等，而后者才是 WTO 的根本目的与规则制定的一般性原则。⁵⁴进一步地，在 GATS 中与之类似，其序言中指出在实现服务自由化的同时，也要保护发展中国家的国家利益。若将“非嵌入式的 AI 产品”视为一种“货物”，则各成员国必须开放自身市场，允许美国等人工智能技术产业发达的国家的产品进入自身市场，从短期看，进口国

⁵⁰ Ronald Dworkin, *Taking Rights Seriously*, Harvard University Press, 1977, p. 24.转引自舒国滢：《法理学导论》，北京大学出版社 2019 年版，第 113 页。

⁵¹ 同上注，第 114 页。

⁵² Preamble, GATT.

⁵³ Preamble, GATT.

⁵⁴ 实际上，无论是 GATT 还是 GATS，从未将“自由贸易”定为其唯一或首要目标。细看 GATT 和 GATS 的序言便可发觉此重要特点。GATT 在第二段中说明其目的在于“提升生活质量、确保充分就业、稳步增长的实际收入和有效需求、充分发挥世界资源的潜力，扩大商品的生产和交换”：

„The Government of ... Recognizing that their relations in the field of trade and economic endeavour should be conducted with a view to raising standards of living, ensuring full employment and a large and steadily growing volume of real income and effective demand, developing the full use of the resources of the world and expanding the production and exchange of goods,“

紧接着，第三段中，才出现了“贸易自由、互惠”等表述（为了实现这一目的，应通过互惠、贸易自由化等手段进行）。同样，在 GATS 序言中，也有类似结构：GATS 序言的第二段明确表示其目的在于促进成员国和发展中国家的经济发展：

„Members... Wishing to establish a multilateral framework of principles and rules for trade in services with a view to the expansion of such trade under conditions of transparency and progressive liberalization and as a means of **promoting the economic growth of all trading partners and the development of developing countries**;

在第三段中，才提及“贸易自由”等术语。综上所述，应当认识到：WTO 的基本目的是和实现基本目的所依赖的手段，是不同的概念，不能混为一谈。

那么，依照《维也纳条约法解释公约》第 31 条，也就不应当将序言中从未提及的目的，纳入到法条解释的过程当中。所以对 WTO 法条进行目的解释的过程中，应当清楚地认识到：WTO 法的基本目的是：“提升生活质量、确保充分就业、稳步增长的实际收入和有效需求、充分发挥世界资源的潜力，扩大商品的生产和交换（GATT）以及促进所有贸易伙伴的经济增长和发展中国家的发展（GATS）”。

的消费者福利当然得到了提升，但从长期来看，这必然使得进口自身的人工智能产业受到发展，不符合此类国家的发展权。

若将“非嵌入式 AI 产品”视为“服务”，由于 GATS 相比较于 GATT 具有较低程度的贸易自由化要求，成员国可以不选择在自身的减让表中对相应服务作出承诺，这在一定程度上可以保护成员国自身的产业不受外部贸易的冲击。在另一方面，人工智能产品对贸易以外的隐私、国家安全等也带来了相应的影响，将其视为“服务”是一种更有利于维护各成员国自身利益的手段。综上所述，本文主张将“非嵌入式的 AI 系统”视为一种服务。

五、结语

人工智能的产生对主要制定于前互联网时代的 WTO 规则提出了严峻的挑战。因此，依托于 WTO 解决人工智能时代可能出现的贸易纠纷，必须依赖于 WTO 创建之初所预设的基本目的，在平等协商、互利共赢的前提下，解决贸易争端。而非动辄以“脱钩断联”、“无故制裁”的手段，挑战世界贸易秩序。

欧盟《AI 框架公约》的规则输出、理论基础与中国因应

杨乐天*

摘要：人工智能技术的迅猛发展促使全球治理步入规则构建的关键时期。欧盟率先通过《AI 框架公约》提出以风险管理、人权保护与发展安全平衡为核心的治理框架，意在确立其在全球 AI 治理中的领导地位。本文运用国际关系霸权稳定论、博弈论、新自由制度主义三大理论，探讨欧盟的战略逻辑，揭示其在规则输出过程中遭遇的霸权能力不足的问题，通过“制度先行”策略争夺规则话语权建立 AI 框架公约的动力，维持 AI 框架公约制度所需的新自由制度主义理论和在未来维持 AI 框架公约所需解决的问题。最终，基于对 AI 框架公约及欧盟 AI 国际战略的分析，提出以技术发展为先导、构建统一市场、依托 WTO 改革推动全球 AI 治理合作等策略建议，旨在为中国参与全球 AI 治理合作提供参考。

关键词：人工智能治理 欧盟 AI 法案 AI 框架公约 国际机制 中国因应

一、问题的提出

人工智能技术的革命性突破与跨领域应用，如 Chatgpt、Deepseek 等，引发了通用人工智能新一轮的发展热潮，推动全球人工智能治理进入规则构建的关键阶段。同时，随着人工智能领域发展不平衡、规则不完善、秩序不合理等问题已经溢出网络空间，直接威胁现实社会。¹如何治理人工智能，成为世界各国共同面对的挑战和难题。

目前，欧盟、美国、中国等国均将人工智能立法提上议事日程。我国十三届、十四届全国人大常委会立法规划均将人工智能立法列入其中。国务院 2023 年、2024 年也连续两年将人工智能法草案列入年度立法计划。2025 年 2 月 12 日，瑞士联邦通信管理局（OFCOM）发布的一份《世界各国和地区人工智能法规分析报告》显示，世界大多数国家有一个或以上的国家人工智能战略或行动计划，但很少有国家拥有生效的人工智能特定法律约束性文件，人工智能的国际法治理则更是立法的空白区域。²

2024 年 3 月 13 日，欧盟率先通过全球首部 AI 法案——《欧盟 AI 法案》³，

*杨乐天，男，2002 年生，上海对外经贸大学国际组织学院（贸易谈判学院）2024 级法与经济学专业硕士研究生。研究方向为欧盟法、数据法、人工智能法。联系方式：电话 19832590976，邮箱 yanglt2024@163.com。通讯地址：上海市松江区文翔路 1900 号，邮编 201620。

¹ 《人工智能发展报告（2024 年）》，中国信息通信研究院，2024 年 12 月。

² Swiss Federal Office of Communications (OFCOM). Analysis Report on Artificial Intelligence Regulations in Countries and Regions Worldwide[R]. 2025.

³ 《人工智能法案（Artificial Intelligence Act）》，以下简称“AI 法案”，（EU）2024/1689，欧洲议会与欧盟理事会 2024 年 6 月 13 日通过，载 EUR-Lex: <https://eur-lex.europa.eu>。

并于同年5月17日通过《AI框架公约》⁴，来具体落实《欧盟AI法案》的域外监管。两者共同构成了欧盟AI领域的治理规范。欧盟的AI治理，立足于“不让任何人掉队，人人享有自由、保护和公平”的欧洲价值观，以风险管理为基础，以人权民主为基石，以促进AI的发展与安全的平衡的三大理念作为AI框架公约的理念。本文第一部分，基于AI框架公约以风险分级为方法，以人权保护为基石、以平衡发展与安全为发展理念三大核心内容分别进行辩证分析，第二部分，引入国际关系理论中的霸权稳定论的基础力量模型、博弈论的智猪模型和罗伯特·基欧汉的新自由主义制度理论对欧盟AI框架公约的战略意图和实施逻辑进行阐释，在第三部分，并对中国未来如何推动AI的国际合作从发展理念，参与基础和实践路径三方面提出建议。

二、《AI框架公约》的核心治理架构

本章围绕《AI框架公约》的核心治理架构展开分析，分析《AI框架公约》基于风险的分级管理机制、以数字人权为基石的保护方式，以及平衡发展与安全的理念。首先，基于风险的分级管理机制通过风险评估体系实现分级分类治理，结合全生命周期监管和透明度原则，构建治理框架，以应对人工智能技术的快速发展和潜在风险。其次，以人权为基石的保护方式，解析“数字人权”所应保护的个人信息权的双重属性（人格权与财产价值），但现行国际法包括AI框架公约在内，仍存在数字时代人权保护方面的立法空白。最后，平衡发展与安全的理念通过引入“监管沙盒”机制，促进人工智能的创新与安全协调发展，同时强调各国需在发展与安全的光谱上寻找适合自身的折中点。本章通过对上述核心机制的分析，揭示《AI框架公约》在AI治理中的制度设计逻辑及其面临的挑战。

（一）基于风险的分级管理机制：分级分类、全生命周期监管与透明度原则

《AI框架公约》第1条第2款规定，各缔约方须基于风险危害程度与发生概率评估体系，实施“风险分级分类治理机制”。结合AI框架公约其他章节的配套条款可知，该监管方法是构建以分级分类为核心，全生命周期监管为监管范围、辅以透明度原则的预防与救济框架。⁵而基于风险的分级管理的思维逻辑在于，“如果行政规制机构能识别不同领域、不同环节中人工智能应用风险的程度差异，设定和实施相应的规制措施，可使规制措施与被规制的风险级别相称”⁶，避免给低风险企业造成过重合规负担，实现监管资源的最大化利用，以下就风险分级、全生命周期、透明度三方面进行讨论。

以风险分级为核心。根据AI框架公约要求，各国自行制定国内的风险分级

⁴ 欧洲委员会，《欧洲委员会人工智能、人权、民主和法治AI框架公约》（Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law），以下简称“AI框架公约”（AI Framework Convention），2024年5月17日通过，2024年9月5日开放签署，访问日期2025年3月26日，<https://www.coe.int/web/conventions/full-list?module=treaty-detail&treaty-num=225>。

⁵ AI Framework Convention, Article 1(2).

⁶ 宋华琳：《人工智能立法中的规制结构设计》，载《华东政法大学学报》2024年第5期，第6-20页。

分类管理办法，当前世界，欧盟作为制度先行者，基于已有 AI 法案中风险的分级管理——四分法⁷，将风险整体横向分为不可接受的风险、高风险、有限风险和极低风险四种。⁸其评估方法、评估主体和评估标准依风险等级而不同，见于表 1，为各国提供了风险分级管理的模板。但风险管理方式的设计究其本质仍为各国国内法内容，允许各缔约方根据自身情况制定具体的国内立法。又由于各国对于 AI 的风险分级方式尚在探索之中，并无明确纪律，欧盟的四分法是否为各国所采纳，也须待未来讨论。

表 1：欧盟 AI 法案——风险分级管理模式

风险等级	评估方法	评估主体	评估标准
不可接受风险	法律直接禁止	欧盟委员会通过立法明确禁止行为	是否属于第 5 条禁止的 AI 实践
高风险	强制性合规评估	自评+第三方机构审核（部分行业）	符合附件Ⅲ高风险清单并满足第 9—15 条合规要求
有限风险	透明度检查	市场监管机构抽查，用户可起诉	是否符合第 52 条透明度义务
极低风险	自愿性评估	无指定主体，企业自主决定	无强制要求，鼓励企业遵循伦理指南

资料来源：根据欧盟 AI 法案整理

以全生命周期监管为范围，为“基于风险分级的管理”构建配套体系，AI 框架公约的规制范围涵盖人工智能系统的全生命周期。⁹根据 AI 框架公约条款的规范性解释，全生命周期监管机制旨在通过系统性地规范 AI 的各发展阶段。该机制要求将监管义务贯穿于技术研发、部署应用退役及迭代更新的全过程，以此确保 AI 框架公约规制范畴既涵盖既存风险，亦能前瞻性地应对潜在风险。该制度设计体现了 AI 框架公约制定者构建适应性治理框架的立法智慧——通过将技术快速发展与不可预测性特征纳入制度考量，使 AI 框架公约在动态演进中保持规范效力。简言之，立法者通过采用前瞻性（preventive）与预防性（precautionary）双重考量，将人工智能技术可能引发的最大范围风险纳入 AI 框架公约的规范化治理体系。¹⁰这一点中美欧三方面在操作的事实上已达成了共识，欧盟在其 AI 法案中要求建立全生命周期监管并且要求需要定期进行系统性的审查和更新¹¹，美

⁷ AI Framework Convention, Article 16.

⁸ Artificial Intelligence Act, High-level Summary, 27 February 2024, available at <https://artificialintelligenceact.eu/high-level-summary/>.

⁹ AI Framework Convention, Article 16.

¹⁰ AI Framework Convention Explanatory Report, para. 14.

¹¹ AI 法案第 9 条第 2 款

国 NIST 框架中的全生命周期包括设计、收集和处理数据、建立、使用和修改模型、开发和使用、操作和监控、使用或被影响的群体和生态等环节。中国信息通信研究院“可信人工智能总体框架”、“人工智能风险管理体系”均涵盖规划设计、数据处理、模型建设、测试验证、部署上线等阶段。¹²AI 框架公约通过国际法的方式，将这一“全生命周期监管”的理念，输出到 AI 框架公约各缔约国。

以透明度原则为辅助措施。¹³AI 框架公约将 AI 透明性要求分为“外部可理解性要求”¹⁴和“内部可解释性要求”¹⁵两个方面。一方面，外部可理解性旨在保障在不同风险等级项下用户的知情权与救济权¹⁶，通过提供输出结果的自然语言或可视化解释，使非技术用户（如患者、消费者）能够理解人工智能系统的决策过程。违反外部可理解性可能导致用户救济权受限，并依据第 14 条进行追责。2022 年，欧盟委员会通过了《人工智能责任指令》¹⁷和新《产品责任指令》¹⁸以指令为依据进行救济。当前中国学界对于人工智能应用责任方式问题尚缺乏体系化思考，相关争议非常大，在此不做讨论。¹⁹但外部的可理解性是用用户“知道其受到侵害”的前提，换言之，“外部可理解性”的要求在于向用户“打开黑盒”知道其是否受到侵害。

另一方面，为实现风险分级的准确性，内部可解释性则侧重于确保系统的合规性与安全性²⁰，通过实现模型内部逻辑的代码级或数学可解析性，使技术专家和监管机构能够深入理解系统的运行机制用于合规审查。²¹即“内部可解释性”要求可使专家从内部分析“黑盒”。最后，为平衡国家利益，国家可依据国家安全例外来豁免透明度义务。但 AI 框架公约对于国家安全的豁免条款的范围语焉不详，给予了各国充分的自由解释权。

总之，由于人工智能技术本身，其功能和类型具有多样性，风险程度具有可塑性、变化性，除极少数应用于高风险领域的场景以外，不适宜采用像对待药品、核设施那样的严格意义上的风险规制原则。²²而就欧盟提供的模板 AI 法案来讲，其采取的是“横向分级”的方法而非“纵向”的依据具体场景进行规制。²³至于 AI 规

¹² 中国信息通信研究院. 人工智能治理蓝皮书（2024 年）[EB/OL]. (2024-12). <https://www.caict.ac.cn/kxyj/qwfb/bps/202412/P020241227660032159191.pdf>.

¹³ AI Framework Convention, Article 8

¹⁴ 外部可解释性（Explainability）要求 AI 系统能够向用户和受影响者提供其决策或输出的清晰解释，确保决策过程可理解且透明。

¹⁵ 内部可理解性（Interpretability）要求 AI 系统的内部工作机制可追溯且能被人类理解，确保技术透明和可审计。

¹⁶ AI Framework Convention, Article 14.

¹⁷ 《关于使非合同性民事责任规则适应人工智能的指令（人工智能责任指令）的提案》

¹⁸ 《关于缺陷产品责任的指令的提案》

¹⁹ 参见杨立新：《人工智能产品责任的功能及规则调整》，载《数字法治》2023 年第 4 期；张安毅：《人工智能侵权：产品责任制度介入的权宜性及立法改造》，载《深圳大学学报（人文社会科学版）》2020 年第 4 期；王乐兵：《自动驾驶汽车的缺陷及其产品责任》，载《清华法学》2020 年第 2 期等。

²⁰ AI Framework Convention, Article 8,16.

²¹ AI Framework Convention, Article 16.

²² 周学峰：《论人工智能的风险规制》，载《比较法研究》2024 年第 6 期，第 42-56 页。

²³ 所谓横向规制，是指立法不区分行业、领域而就人工智能风险的共同性问题作出一般性规定。所谓纵

制采用“横向分级”抑或“纵向场景”何种为宜，则留待各国去探索，但其关于全生命周期和透明度的制度设计可资赞同。

（二）以数字人权为基石：人权保护的新维度与挑战

以人权为基石，《AI 框架公约》以维护人权、民主与法治为价值基础，其中人权保护是核心。²⁴人权是指作为“以人或者人性作为其来源的，一个人应该拥有的权利”。²⁵由于两次世界大战给世界人权带来的沉重苦难，1948 年后联合国大会通过的《世界人权宣言》《公民权利和政治权利国际 AI 框架公约》和《经济、社会及文化权利国际 AI 框架公约》共同构成国际人权宪章体系的基本框架。²⁶但如果说人与自然分离是生物意义上第一次“人成之为人”的话，那么，“以网络化、数字化和智能化的方式变革工业生产方式和农业生产方式，并再一次以新的联合方式颠覆工业文明时代的制度安排、思想观念和思维方式等，重塑符合信息文明的概念范畴和政治、经济、文化、法律等社会体制”，这就导致了信息意义上第二次“人成之为人”。²⁷因此，AI 框架公约所保护的，应当为“数字人权”。

1. 数字人权的内涵界定：以个人信息权为基础

AI 框架公约第 4 条要求各缔约方确保人工智能系统生命周期的活动符合保护人权的国际法和国内法义务。²⁸根据解释报告引用的 25 个人权文件，现有的国际规范大多由于订立时间较早，大多数没有在数字时代及时更新其数据人权的内容而不适于 AI 时代的人权保护要求，存在立法空白。²⁹以其中与数字人权最相关的《欧盟 108 号 AI 框架公约》为例，其数字人权的内容主要包括基于知情权和透明权、数据控制权、数据更正和删除权。³⁰其本质仍属隐私权在数据空间的延伸适用，天然带有“被动性”。³¹但数字时代人权的保护基础在于“个人信息权”，这与传统人权中基于隐私权逻辑的“数据权利”保护存在根本性的区别。但 AI 框架公约对此并未进行界定。

因此，为厘清数字人权内涵，首先应当区分传统隐私权与数字时代个人信息权之间具有本质差异的法律属性。“个人信息权”具有人格属性与财产价值双重性质³²，一方面，人格权部分以防御性为主，在人工智能语境下，其规制的方向应

向规制，是指区分不同的行业、领域而就每一行业、领域中的人工智能风险问题分别予以立法规制。

²⁴ 根据 AI 框架公约的解释文本，AI 框架公约所涵盖的人权分布于联合国 AI 框架公约、欧盟 AI 框架公约和美洲共 25 个法律文件。参见 AI Framework Convention Explanatory Report, para. 39.

²⁵ [美] 杰克·唐纳利：《普遍人权的理论与实践》，王浦劬译，中国社会科学出版社 2001 年版，第 12 页。

²⁶ 刘志强：《论“数字人权”不构成第四代人权》，载《法学研究》2021 年第 1 期，第 20-34 页。

²⁷ [英] 卢恰诺·弗洛里迪：《信息伦理学》，薛平译，上海译文出版社 2018 年版，第 351 页。转引自马长山：《智慧社会背景下的“第四代人权”及其保障》，载《中国法学》2019 年第 5 期，第 5-24 页。

²⁸ AI Framework Convention, Article 4.

²⁹ AI Framework Convention Explanatory Report, para. 39.

³⁰ Convention for the Protection of Individuals with Regard to the Automatic Processing of Personal Data (Convention 108), opened for signature January 28, 1981, Council of Europe Treaty Series No. 108, ETS 108.

³¹ Convention for the Protection of Individuals with Regard to the Automatic Processing of Personal Data (Convention 108), opened for signature January 28, 1981, Council of Europe Treaty Series No. 108, ETS 108.

³² 个人信息也具有效用性、稀缺性、可控性、流通性具有天然的财产基因。参见彭诚信：《论个人信息的双

当为算法下的人权保护，因为即便是匿名的信息，通过算法的计算仍然可以计算出个人隐私用于侵犯人权。如著名的塔吉特百货孕妇营销案，百货公司利用顾客的购买数据推测出该少女怀孕，并向其推荐一系列孕妇产品。³³另一方面，数字人权所保护的个人信息也具有效用性、稀缺性、可控性、流通性³⁴，具有天然的财产基因，换言之其财产性为数据时代的“发展权”，保护个人与数字时代开发者如 AI 的开发者共享数字福利也应当是数字人权下人权保护的应有之义。

2. 数字人权保护场景：以隐性侵犯为特征

除数字人权的内涵未加界定之外，AI 框架公约对于人权的保护方式也不符合时代变化。其主张保护的依据主要在于人类尊严和个人自主权³⁵、平等与不歧视³⁶、隐私和个人数据保护³⁷三方面，并要求各缔约方提供有效救济途径（如起诉机制、信息透明），用于因人工智能活动导致的人权侵犯。³⁸AI 框架公约所保护的“人权”与现行国际法中的人权内容，属于制度上的“显性人权保护”。但在人工智能语境下，对于人权的侵犯可以是“隐性虚拟的侵犯”，给予起诉权利、信息透明的方式未必能够满足人权保护的需要。

对人权的“隐性侵犯”，典型如著名的“剑桥分析”案，本案对于个人隐私权、政治权利的严重侵犯，在美国引起了轩然大波。本案源于 2016 年唐纳德·特朗普的意外当选和其后披露的“剑桥分析”³⁹丑闻，即通过问卷调查的形式欺骗性地从数千万 Facebook 用户那里收集个人信息⁴⁰，用于选民分析和定位。通过获取大量 Facebook 信息，形成用户行为数据库⁴¹，再运用大数据分析心理学分析技术，构建民众性格档案，识别出具有摇摆政治态度的选民群体。基于其性格特征及认知模式，在适宜的时机、适当的场合，采用最能影响目标选民政治倾向的信息传播方式，实现对相关政治理念的精准推送与传播。最终影响其投票决策。最终，鉴于剑桥分析事件，美国联邦贸易委员会（FTC）对 Facebook 处以 50 亿美元的罚款。Facebook 被责令成立独立的隐私委员会，并指派“合规官员”，以保障用户

重法律属性》，载《清华法学》2021 年第 6 期，第 78-97 页。

³³ 吴军：《智能时代——大数据与智能革命重新定义未来》，中信出版集团 2016 年版，第 152—155 页。

³⁴ 彭诚信：《论个人信息的双重法律属性》，载《清华法学》2021 年第 6 期，第 78-97 页。

³⁵ AI Framework Convention, Article 7.

³⁶ AI Framework Convention, Article 10.

³⁷ AI Framework Convention, Article 11.

³⁸ AI Framework Convention, Article 14.

³⁹ 剑桥分析将精准营销与心理学分析用在政治宣传中，利用问卷调查的方式，获取读取用户脸书数据授权并收集大量社交网络数据后，与其在多种渠道购买的民众土地登记、汽车信息、购物习惯、俱乐部会员资格等个人信息相整合，以最易影响目标选民政治倾向的信息呈现方式，向其精准推送和传播相关政治理念，最终影响其投票抉择。

⁴⁰ 根据联邦贸易委员会指控：剑桥分析公司、尼克斯和科根欺骗了消费者，谎称他们没有从被要求回答调查问题并分享一些脸书个人资料数据的 Facebook 用户那里收集任何个人身份信息。参见：FTC Sues Cambridge Analytica, Settles with Former CEO and App Developer, July 2019, <https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-sues-cambridge-analytica-settles-former-ceo-app-developer> (last visited Apr. 7, 2025)。

⁴¹ 前剑桥分析公司首席执行官尼克斯曾公开表示，该公司存档了几乎每位美国成年人的性格档案（约 2.2 亿人）。

隐私权益。与此同时，剑桥分析公司于 2018 年 5 月提交了破产申请。⁴² 本案通过 Facebook 开放的 API 接口，将调查问卷嵌入 Facebook 平台。用户在参与问卷互动时，不仅其社交图谱、行为偏好等核心隐私数据在毫无察觉的情况下被不当采集，更在完成问卷后，遭到专业数据公司“剑桥分析”基于心理测量学模型的精准操控——该机构运用大数据挖掘与行为心理学技术，针对用户政治倾向实施潜意识层面的定向干预来改变个人想法，影响个人政治权利的行使。⁴³

在人工智能时代背景下，即便经过匿名化处理，个人隐私和数据依然有可能被人工智能算法所推算。并且个人信息的功能性发生了转变，商业实体日益依赖于用户的网络行为轨迹及数据进行预测性分析或精准营销。商业实体拥有强烈的经济激励去搜集个人信息，用户的浏览历史、点击行为以及购买记录均在其监控范围之内。这种现象导致了人们在数字空间中的透明度前所未有地提高，通常将此现象称为“数据监控”。⁴⁴ 通过“数据监控”获取大量数据之后，再通过人工智能，使信息自由流通受阻，信息市场被高度扭曲，每个人都生活在算法编织的信息“滤泡”（Filter bubbles）当中，此时，即便人权形式上依旧存在，但业已实质上被无形侵犯。⁴⁵ 因此，人工智能项下的人权保护方式也应当有所变化，而非将眼光局限于提供透明性和起诉救济机制，但 AI 框架公约并未对此进行界定。

（三）平衡发展和安全的理念：AI 发展与风险控制的协调

AI 框架公约各缔约国对于 AI 的治理进行的立法，本质上都是在发展与安全的光谱上选择最适合的折中点，这一点由缔约各国进行国内立法寻找，属于国内法内容。对于 AI 发展的收益预期，高盛研究部的经济学家 Joseph Briggs 和 Devesh Kodnani 在《2025 年全球人工智能投资有望增至 2000 亿美元》研究报告中指出，生成式 AI 具有巨大的经济潜力，在广泛使用后十年内每年可提高全球劳动生产率超过 1 个百分点以上。⁴⁶ 为抓住这一发展机会，全球各国纷纷推出各自 AI 发展战略。⁴⁷ 如图 1 所示，2023 年，美国投资的 672 亿美元大约是中国（78 亿美元）的 8.7 倍，是第三英国（38 亿美元）的 17.8 倍（图 1 全球人工智能私人投资情况）。可见欧盟在 AI 的私人资金投入相比于美国和中国完全不在一个数量级。因此，欧盟委员会也推出了自己的 InvestAI 计划，将调动 2000 亿欧元用于人工智能投资，并为人工智能千兆工厂再设立一个 200 亿欧元的新基金，以期使

⁴² FTC Sues Cambridge Analytica, Settles with Former CEO and App Developer, July 2019, <https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-sues-cambridge-analytica-settles-former-ceo-app-developer> (last visited Apr. 7, 2025)。

⁴³ 穆琳：《“剑桥分析”事件“算法黑箱”问题浅析》，载《中国信息安全》2018 年第 4 期，第 92-94 页。

⁴⁴ 彭诚信：《论个人信息的双重法律属性》，载《清华法学》2021 年第 6 期，第 78-97 页。

⁴⁵ Eli Pariser, *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*, New York: Penguin, 2012.

⁴⁶ Briggs, J., & Kodnani, D. (2025). 高盛观点 | 全球人工智能投资规模 2025 年或将接近 2000 亿美元. 高盛. <https://www.goldmansachs.com/worldwide/greater-china/insights/AI-investments-to-approach-200-billions>.

⁴⁷ Swiss Federal Office of Communications (OFCOM). (2025, February 12). Analysis of artificial intelligence regulation in countries and regions around the world. Federal Office of Communications (OFCOM).

欧洲成为人工智能大陆并追赶与美国和中国的差距。⁴⁸

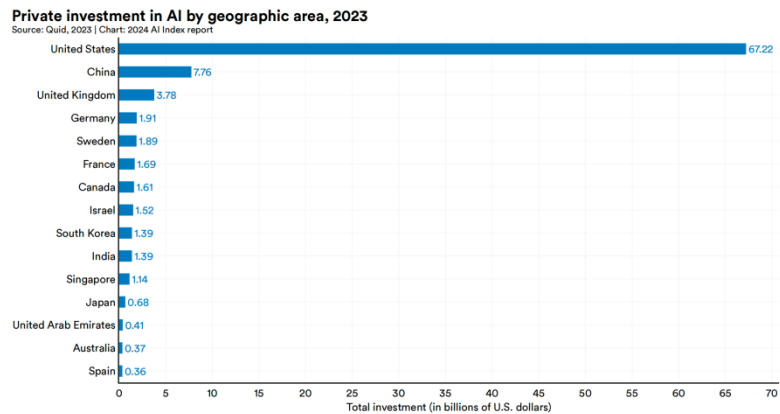


图 1：全球人工智能私人投资情况（按地区划分，单位为亿美元）⁴⁹

资料来源：Stanford HAI, 2024, P.247

但发展 AI 就意味着要能面对 AI 在不同情境下表现出不同程度的复杂性、不透明性和自主。首先，就 AI 的复杂性而言，人工智能技术作为一项新兴技术，其技术特征、商业模式尚未定型，目前仍处于不断发展和演变的过程，未来走向尚存在许多不确定性，依据场景进行风险分级未免为时尚早。⁵⁰当前，诸如通用大型语言模型 ChatGPT 等主流发展模型的应用场景极为广泛。欧盟所提出的四分法，基于人工智能的应用领域，将面临 AI 大模型应用的无边界性与四分法应用领域的局限性，这构成了四分法规制手段难以应对的难题。换句话说，欧盟这种对风险进行静态分级的做法，存在一定的机械性和分类上的疏漏。⁵¹

其次，就 AI 的不透明性而言，但人工智能依赖于代码和算法模型，而绝大多数的模型都是不透明、高度专业化的，这意味着人工智能可以轻易规避审查。⁵²并且规制者相较科技企业，明显处于信息与技术劣势，若想要实现有效的全过程监管，提升分析、评估、控制人工智能风险的能力，也需耗费巨大的人力、物力、财力资源。⁵³

⁴⁸Commission Launches New InvestAI Initiative to Mobilise €200 Billion of Investment in AI, European Commission: Digital Skills and Jobs, <https://digital-skills-jobs.europa.eu/en/latest/news/commission-launches-new-investai-initiative-mobilise-eu200-billion-investment-ai> (last visited Apr. 5, 2025)。

⁴⁹ Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, et al. Artificial Intelligence Index Report 2023. Stanford Institute for Human-Centered Artificial Intelligence, 2023. <https://aiindex.stanford.edu/report/>。

⁵⁰ 周学峰：《论人工智能的风险规制》，载《比较法研究》2024 年第 6 期，第 42-56 页。

⁵¹ 解志勇. 高风险人工智能的法律界定及规制[J/OL]. 中外法学, 1-21[2025-04-11].

<http://kns.cnki.net/kcms/detail/11.2447.D.20250326.1346.004.html> .

⁵² 约翰·帕克：《全民监控——大数据时代的安全与隐私困境》，关立深译，金城出版社 2015 年版，第 299 页。

⁵³ 解志勇. 高风险人工智能的法律界定及规制[J/OL]. 中外法学, 1-21[2025-04-11].

<http://kns.cnki.net/kcms/detail/11.2447.D.20250326.1346.004.html> .

再次，就 AI 的自主性而言，人工智能“基于概率的文本生成方式”⁵⁴即使对开发人员来说，大型模型在预训练阶段如何处理个人信息数据和揭示词汇概率分布规律也是未知的。因此，包括埃隆·马斯克（Elon Musk）在内的 1000 余名专家曾共同签署公开信，呼吁至少在六个月的时间内暂停开发超越 GPT-4 的大型模型。

55

总之，AI 既可能会增进人权，也可能会侵蚀人权。因此，AI 框架公约要求各缔约方两方面促进 AI 的发展与安全的平衡。首先，从 AI 本身角度，要求各缔约方采取适当措施促进人工智能系统的可靠性及其输出的信任度，实现 AI 的高质量发展，以提升 AI 本身的可靠性。⁵⁶另一方面，要求各方引入“监管沙盒”，建立受控环境，用于开发、实验和测试人工智能系统。监管沙盒滥觞于英国对于金融科技设计的“监管沙盒”。⁵⁷监管沙盒的理念在于促进私人实体、监管机构和其他利益相关者在沟通的情况下进行测试金融科技，以在早期识别潜在风险和问题，以达成“敏捷治理”。⁵⁸就欧盟 AI 法案而言，其要求欧盟每个成员国必须确保其国家主管机构设立至少一个 AI 监管沙盒，并于 2026 年 8 月 2 日前投入运营。至于各国 AI 沙盒如何制定与实施，则由各国自行选择。⁵⁹因此，当下世界尚未有真正成功的“AI 监管沙盒”供各国借鉴，AI 的治理如何实现发展与安全的平衡，是一个不断发展的课题。

三、《AI 框架公约》的理论基础与实施逻辑

本章基于霸权稳定论、智猪模型与新自由制度主义视角，系统剖析欧盟推动《AI 框架公约》的理论基础与实施逻辑。首先，引入霸权稳定论揭示欧盟在 AI 原料、资本的来源、市场以及在高附加值产品生产上的劣势，难以通过传统霸权路径主导国际机制；其次，引入智猪模型阐释欧盟作为“次强国”主动承担规则制定成本，以抢占全球话语权的博弈动机；最后，引入新自由制度主义指出 AI 框架公约通过降低交易成本、强化透明度与预期稳定性，试图构建无霸权依赖的国际合作框架。

（一）基于霸权稳定论的分析路径：欧盟在 AI 治理中的霸权能力评估

“AI 框架公约”旨在建立一项关于人工智能国际合作的治理机制。⁶⁰国际机制，

⁵⁴ 人工智能大模型的主要技术原理是在预训练阶段挖掘训练数据中包含的词元（token）间概率分布规律，并以参数形式保存于大模型之中，然后依据这种词元间概率分布规律推导出用户需求的信息。

参见解志勇. 高风险人工智能的法律界定及规制[J/OL]. 中外法学, 1-21[2025-04-11].

<http://kns.cnki.net/kcms/detail/11.2447.D.20250326.1346.004.html>.

⁵⁵ “没有人——甚至他们的创造者——都无法理解，预测，或可靠地控制（no one—not even their creators—can understand, predict, or reliably control）”。参见：赵子昂. 美国观察 | 暂停大型语言模型实验的请愿书及其批判：跨越人工智能的“卢比孔河”[EB/OL]. 复旦中美友好互信合作计划, [2025-04-11].

<https://fdi.fudan.edu.cn/8a/68/c21253a494184/page.htm>.

⁵⁶ AI Framework Convention, Article 12.

⁵⁷ 廖凡. 论金融科技的包容审慎监管[J]. 中外法学, 2019, 31(03): 797-816.

⁵⁸ Financial Conduct Authority. Regulatory sandbox [EB/OL]. (November 2015). Financial Conduct Authority, [2025-04-11]. <https://www.fca.org.uk/publication/research/regulatory-sandbox.pdf>.

⁵⁹ EU AI Act, Arts 57, 58.

⁶⁰ AI Framework Convention, Preamble.

即“一系列围绕行为体的预期所汇聚到一个既定国际关系领域而形成的隐含的明确的原则、规范、规则和决策程序。原则是指对事实、因果关系和诚实的信仰；规范是指以权利和义务方式确立的行为标准；规则是指对行动的专门规定和禁止；决策程序是指流行的决定和执行集体选择政策的习惯”。⁶¹AI 框架公约符合该国际机制的定义，那么如何维护好该国际机制并使其扩大影响，此处拟引入国际关系的霸权稳定论进行分析。

霸权稳定论由经济学家查尔斯·金德尔伯格⁶²提出，后经罗伯特·吉尔平⁶³、斯蒂芬·克拉斯纳⁶⁴等学者进一步发展，其中心命题是：第一，国际政治中的秩序是由一个主导国家创立的。第二，国际秩序的维持需要霸权国家的持续存在。正如金德尔伯格所言“要使世界经济稳定，需要一个稳定者且只能由一个稳定者”。⁶⁵

根据霸权稳定论基础力量模型四要素，霸权国家必须控制原料、资本的来源、市场以及在高附加值产品的生产上具有竞争优势。那么欧盟是否具备这四要素呢？首先 AI 的原料众多，以最重要的模型和算力为例，现在几乎所有 AI 大模型训练时采用的 Transformer 网络结构，由美国谷歌公司在 2017 年提出。论及算力，NVIDIA⁶⁶、AMD⁶⁷、Intel⁶⁸三家显卡厂商均为美国企业，而美国一手云大厂，一手英伟达，手握着全球 AI 的核心资源，欧盟对此并无优势。⁶⁹其次，资本的来源而言，根据欧盟统计局数据，2024 年第三季度，欧盟资本账户显示-17 亿欧元的赤字，相较于 2023 年同期的-4 亿欧元赤字有所增加。⁷⁰这意味着欧盟目前扮演的是资本输出国的角色，资本流动呈现外流趋势，这在资本控制力上并不占据优势。在市场领域，根据欧盟委员会发布的统计数据，2024 年，欧盟的国内生产总值（GDP）已达到 24.7 万亿美元，占全球 GDP 的 14.7%，为世界第三大经济体，故其市场吸引力足够。最后，在技术领域，从技术端而言，2023 年，61 个知名 AI 模型来自美国的机构，远远超过中国大陆的 15 个，法国仅 8 个。⁷¹以 AI 独角

⁶¹ Krasner, S. D. (1983). *International Regimes*. Ithaca, NY: Cornell University Press, p. 2, cited in Keohane, R. O. (1984). *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton: Princeton University Press.

⁶² Charles Kindleberger.

⁶³ Robert Gilpin

⁶⁴ Stephen Krasner

⁶⁵ Kindleberger, Charles P. *The World in Depression, 1929-1939*. Berkeley: University of California Press, 1973.

⁶⁶ NVIDIA：成立于 1993 年，总部位于美国加利福尼亚州圣克拉拉市，是全球知名的显卡设计和制造商，其产品广泛应用于游戏、专业图形处理以及人工智能等领域，拥有 GeForce、Quadro、Tesla 等多个系列的 GPU 产品。

⁶⁷ AMD：创建于 1969 年，总部位于美国加利福尼亚州桑尼维尔市，是高性能与自适应计算领域的领先企业，其 GPU 产品系列涵盖 Radeon RX 系列、Radeon Pro 系列、Radeon Instinct 系列等，主要面向游戏、图形设计、专业图形市场以及高性能计算和深度学习市场。

⁶⁸ Intel：创建于 1968 年，总部位于美国加利福尼亚州圣克拉拉市，虽然主要以 CPU 生产闻名，但近年来也推出了独立显卡产品，如 Iris Xe 图形、Iris Xe MAX 图形、Arc A 系列 GPU 等。

⁶⁹ 全球 AI 大模型一览：中美之外还有谁？. 虎嗅网. 2023 年 3 月 28 日. 2025 年 10 月 21 日. <https://www.huxiu.com/article/957075.html>.

⁷⁰ European Commission. *Balance of payments statistics quarterly data*[R]. 2024.

⁷¹ Stanford Institute for Human-Centered Artificial Intelligence (HAI). *Artificial Intelligence Index Report 2024*[R]. Stanford University, 2024.

兽企业为基础，据 PR Newswire 报道，美国和中国在现存 AI 独角兽企业数量上占据领先地位，引领全球。全球现存 AI 独角兽企业分布于 54 个国家或地区，其中美国拥有 838 家（占比 44.9%），中国有 510 家（占比 27.3%），中美两国的总和占全球总量的 72.2%⁷²，显然欧盟相对于美中，在 AI 技术和独角兽企业数量方面并无优势。综合上述基础力量模型，欧盟并不具备“霸权”标准。

根据“霸权稳定论”，霸权既是合作的必要条件，也是合作的充分条件。既然欧盟并不具备绝对意义上的霸权，则自然不具备稳定合作的能力。那么欧盟力推《AI 框架公约》的动力和理论依据从何处取得？因此需要引入博弈论和另外的国际关系理论——新自由制度主义理论进行分析，以期取得对《AI 框架公约》的逻辑分析。

（二）基于智猪模型的动力分析：欧盟作为“次强国”的规则制定动机

从历史视角审视，国际机制的构建往往并非源于公正无私的理想主义者对共同善的追求，而是主要由政府官员主导，其核心动机在于追求所谓的国家利益最大化，同时也不可避免地掺杂了个人或集团利益的考量。尽管其言辞可能冠以全球福祉或构建安全相互依存世界的名义，但实质上，此类机制更多地反映了现实主义的权力博弈与利益权衡。⁷³欧盟设计《AI 框架公约》也是基于对技术主权、伦理标准主导权及全球市场竞争力的多重考量。⁷⁴通过“智猪模型”，可以理解为何欧盟会在即便没有“霸权能力”的前提下，在美国和中国这两个 AI 领域的领先国家之前，率先通过 AI 框架公约建立 AI 的国际治理框架。

智猪模型（Boxed Pig Game）是博弈论中解释公共物品供给问题的经典模型，其核心逻辑在于：在资源有限的情境下，规模较大的参与者（“大猪”）因行动收益更高，倾向于主动承担供给成本；而规模较小的参与者（“小猪”）则选择“搭便车”策略，以最小成本获取公共物品收益，此处的大猪、小猪在表 2 中指代大国、小国。⁷⁵该模型的关键变量有三，

一是成本—收益比：大猪的个体收益覆盖行动成本的概率更高；二是策略依赖性：小猪的收益完全依赖大猪的行动选择；三是公共物品的非排他性：规则一旦建立，所有参与者均可共享其红利。

在国际 AI 治理领域，规则制定权可视为一种“全球公共物品”。收益则象征着 AI 时代到来后的潜在利益。大国主动承担规则制定的成本，以期获得长期的收益，博弈中可选的策略组合如表 2 所示。欧盟通过《人工智能 AI 框架公约》

⁷² YZone Research Center, "2024 Global Unicorn Landscape: AI Reshapes the Ecosystem," PR Newswire, April 1, 2025, <https://www.morningstar.com/news/pr-newswire/20250401cn54533/2024-global-unicorn-landscape-ai-reshapes-the-ecosystem> (accessed April 5, 2025)。

⁷³ 罗伯特·基欧汉. 霸权之后-世界政治经济中的纷争与合作. 上海：上海人民出版社，2006。

⁷⁴ European Commission, WHITE PAPER On Artificial Intelligence: A European approach to excellence and trust, Brussels, 19.2.2020, COM(2020) 65 final, available at https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf。

⁷⁵ 王则柯，《平话博弈论》，北京：中信出版社，2011 年 4 月第 1 版，第 71～73 页。

将人权、民主等价值观融入技术治理之中，填补中美技术竞争中的“治理空白”，借此加强其作为“全球规范塑造者”的地位，避免成为“等待的小猪”。因此，尽管与美中相比，欧盟在 AI 领域实力上属于“小国”，但相较于其他国家，欧盟无疑是“大国”，通过承担规则制定的成本，可以换取全球话语权和长期战略利益。

表 2：智猪模型博弈策略表

策略组合	大国收益	小国收益	纳什均衡	经济学解释
大国制定规则，小国不制定	5	4	均衡点	大国承担成本，小国搭便车；总收益最大化
大国等待，小国制定规则	9	-1	不稳定	小国成本高但收益低，策略不可持续
双方都制定规则	4	1	低效	双方均摊成本，但收益被竞争稀释
双方都不制定规则	0	0	零和	无行动导致资源未被利用，集体损失

资料来源：作者自制

尽管上述博弈策略表明，作为全球 AI 领域第三大实体，欧盟通过制定 AI 规则来提供公共产品确实是其最佳策略选择，然而，公共产品具有显著的外部性。这意味着那些未承担公共产品供应责任的个体能够“免费搭车”（free rider），自动享受他人提供的公共物品带来的好处。换句话说，当个体之间存在共同的类似公共产品的需求时，这些个体可能不会采取有效的集体行动来实现这些共同利益，这正是奥尔森所阐述的“集体行动的逻辑”。⁷⁶如何解决这一问题，对于欧盟能否成功增强“布鲁塞尔效应”而言也属关键。现欧盟已将提供《AI 框架公约》作为国际合作的公共产品，其 AI 法案也作为全球 AI 治理模板为众多国家参考，未来如果其使用与 GDPR 跨境数据流动的相似策略——充分性认证。⁷⁷则可以通过将合规成本转嫁给非 AI 框架公约国家，从而提高收益，并通过内部 AI 框架公约的相互监督与合作，增加内部收益，减少搭便车现象的发生解决“集体行动的逻辑”，但其机制的实施的理论依据仍有待分析。

（三）基欧汉新自由制度主义的应然适用：国际制度建设与合作的可能性

基欧汉认为，无霸权状态下的合作是可能的，通过国际制度建设和国际制度安排，世界政治可以朝着进步的方向演化，人类的和平和福利是可以得到保障的。

⁷⁶ 奥尔森，《集体行动的逻辑：公共物品与团体理论》，北京：商务印书馆，2009 年。

⁷⁷ 《通用数据保护条例》(GDPR)第 45 条:当欧盟委员会作出认定，认为相关的第三国、第三国中的某区域或一个或多个特定部门、或国际组织只有在提供充足保护时,才可以将个人 数据转移只到第三国或国际组织。此类转移不需要特定的授权。

⁷⁸当然其前提是，国际制度能降低交易成本、提供信息透明度和稳定预期来促进合作，则国际机制可以通过各国合作的形式存在。

首先，关于降低交易成本，从人工智能的定义而言，欧盟在 AI 框架公约中采用的定义与 OECD 定义⁷⁹完全一致，这一点值得肯定。因为各国的 AI 立法和监管需要定义作为基础，该定义由经合组织的 38 个成员国共同修订，采用 OECD 定义，有助于 AI 框架公约与 OECD 相衔接，未来也能增加跨司法管辖区的合作可能性。

其次，关于透明度，公约通过初始声明文件⁸⁰和定期实施报告制度⁸¹来降低信息的不对称，提高透明性，但根据 AI 框架公约第 3 条，所有关于国防相关的事项均不在 AI 框架公约规制范围之内。⁸²因此具体关于 AI 监管合作的透明度执行情况，也有待于各国合作的效果来验证。

最后，稳定性而言，就文本的稳定性而言，修改公约须经过修正案的提出、沟通、审议、批准、转发和生效等步骤，若某国未接受修正案，则该修正案对其不产生约束力，因此，AI 框架公约的具有文本上的稳定性。⁸³就成员的稳定性而言，根据 AI 框架公约第 31 条，加入需经现有成员国一致同意（而非多数决）⁸⁴，表明 AI 公约的成员制度权限来自成员国的共同授权，而非某一强权的单边意志，其成员的稳定性的预期可以得到保障，公约可以提供在成员和文本上的稳定预期，这一点从稳定性上可资赞同。

（四）AI 框架公约未来实施所面临的挑战：交易成本、透明度与预期稳定性

即便 AI 框架公约能够作为国际机制持续存在，但其提供如上三种功能的可能性仍然存在疑问，在此就 AI 框架公约提供降低交易成本提升透明度和稳定预期的实际能力进行分析。

首先，国际制度需通过清晰规则降低交易成本、增强预期稳定性，但欧盟 AI 框架公约规定的分级分类管理制度本质为各国国内法，由于各国国情不同，风险分级的立法和执行必然会存在碎片化问题。就分级分类标准而言，欧盟提供的是根据使用场景进行的“四分法”划分标准。但人工智能具有科技含量高、技术迭代

⁷⁸ 苏长和：《解读〈霸权之后〉——基欧汉与国际关系理论中的新自由制度主义》，载《美国研究》2001 年第 1 期，第 138-146 页。

⁷⁹ A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments. Different artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment.

⁸⁰ 根据公约第 3 条，各国在签署或批准时需提交“实施方式声明”说明如何履行对私人行为者活动的监管义务（选择适用公约原则或采取其他措施）参见 AI Framework Convention, Article 3(1)(b).

⁸¹ 根据公约第 24 条，各缔约国成为缔约国后前两年内需提交定期实施报告，详细说明为落实公约第 3 条第 1 款 a 项和 b 项（公共机构及私人行为者活动监管）所采取的措施，并持续更新实施公约的进展及措施：AI Framework Convention, Article 24.

⁸² AI Framework Convention, Article 3.

⁸³ AI Framework Convention, Article 28.

⁸⁴ AI Framework Convention, Article 31.

快、情景急速变化等特征，在人工智能开发者、提供者等主体与行政管理部门之间始终存在信息不对称，人工智能开发者、提供者等主体相对占据信息优势。⁸⁵且鉴于未来的发展趋势是“通用大模型”，即 AI 能够适应各种场景和领域，此时过早地依据应用场景来划分风险标准，其分级归类会对监管部门造成归类困难，也会提高通用大模型的合规成本。

并且 AI 框架公约以安全例外豁免的方式可使各国免于执行透明度义务。各国对于安全的标准不同，对于将“由各国自由裁量的与国防有关的事项”排除在《AI 框架公约》范围之外，那么如果通用 AI 作为国防标准之内，如何规制？⁸⁶最后，预期的稳定性而言，首先成员的稳定性，制度设计为“非排他性”的开放型制度，但以一致同意为加入原则，后续的成员加入需要取得全体成员的同意，因此成员的稳定性可得保障。其次，文本的稳定性，AI 框架公约规定了修订程序，并且不同意修订的成员可以选择不适用修订版本，文本上的稳定性也可以得到保障。⁸⁷然而，也正因为文本修订的严格程序，尽管通过 AI 框架公约可以深化 AI 的全球治理，但要形成一个实质性的 AI 治理系统，所面临的再谈判成本无法通过 AI 框架公约来大幅降低。

总之，欧盟 AI 框架公约由于欧盟不具备霸权能力，但依据智猪模型的分析，欧盟又具备主动制定规则的动力，因此只能在新自由制度主义理论的基础上通过提供降低交易成本、提供信息透明度和稳定预期来使 AI 框架公约真正成为协调机制稳定存在，同时这三方面也是 AI 框架公约及其他国际机制实现其价值的必备条件。但 AI 框架公约作为并无实质性争端解决机制的“软法”⁸⁸，其能否达成以上要求，成为 AI 国际治理的制度平台，尚无定论，有待观察。

四、我国参与全球 AI 治理的可行建议

欧盟采取的策略是“制度先行”，通过预先立法确保掌握规则的制定权，并创建 AI 法案模板。随后，利用签订 AI 框架公约确立“条约必守义务”⁸⁹并通过以“市场换规则”的方式，推广 AI 法案内容，同时投资发展本土的 AI 大模型技术。欧盟正通过“先制定规则，再掌握技术”的战略，进行全球 AI 布局。因此，随着欧盟 AI 立法的不断完善和 AI 战略的深入实施，我国感受到的紧迫感也日益增强。我国必须加快吸收欧盟在 AI 立法方面的经验，并构建一个适应我国国情的 AI 战略。否则，一旦落后，将会步步落后。在全球 AI 规则竞争中失去先机，意味着丧失 AI 规则的主导权，这将在未来很长一段时间内对我国参与国际战略竞争产生极其不利的影响。然而，在制定规则的同时，也不能忽视技术发展的客观规律，

⁸⁵ 宋华琳：《人工智能立法中的规制结构设计》，载《华东政法大学学报》2024 年第 5 期，第 6-20 页。

⁸⁶ AI Framework Convention Explanatory Report, para. 36.

⁸⁷ AI Framework Convention Article 28.

⁸⁸ AI Framework Convention, Article 29.

⁸⁹ Vienna Convention on the Law of Treaties, opened for signature 23 May 1969, 1155 UNTS 331 (entered into force 27 January 1980).

以免过于严格的监管阻碍 AI 技术的健康发展。⁹⁰本章拟结合对于欧盟 AI 国际合作的战略，提出未来中国推动 AI 国际合作的路径。

（一）发展理念：以技术竞争力的提升，作为参与全球 AI 治理实力基础

根据霸权稳定论，推动国际合作的霸权源于原料、资本、市场和生产能力的四方面优势。从原料的角度，尽管中国在算力方面相比美国存在劣势，但 Deepseek 的发展证明，通过模型的创新仍然能够缩短在算力的劣势。从资本的角度，中国一方面通过“一带一路”、亚投行提升对外投资力度，另一方面通过制度型开放和《2025 年稳外资行动方案》来提升外资引入力度。⁹¹市场方面，中国是全球第二大消费市场，拥有全球最大规模中等收入群体，蕴含着巨大投资和消费潜力。⁹²在技术层面，中国在人工智能（AI）技术领域扮演着追赶者的角色。中国的大型模型技术主要源自产业实践，例如阿里巴巴、百度、腾讯等公司的 AI 大模型已被广泛应用于广告推送、社交平台图片识别以及内容分发等场景。在应用广度和深度方面，中国已超越欧盟。中国有能力通过四个方面的基础力量模型提升其 AI 技术竞争力，但应以“人类命运共同体”的理念来增进全人类的共同福祉，而非采取霸权逻辑。

（二）参与基础：以构建统一的大市场，作为参与全球 AI 治理的利益基础。

根据新自由制度主义，合作需要共同的利益为前提，中国可以通过“市场换市场，共同利益换规则”的方式进行国际合作。欧洲的推动 AI 国际治理的问题在于，其本身缺乏一个统一的大市场。在 GDP 总量方面，欧盟在 2022 年为 16.65 万亿美元，与中国持平；在人口规模上，欧盟的 2022 年人口为 4.46 亿，甚至超过了美国的 3.32 亿人口。然而，由于欧盟由 28 个国家组成，拥有 23 种官方语言，加之与美国的紧密联系，这些因素共同作用导致欧盟在互联网时代未能孕育出一家大型互联网企业。因此，在数据量、云计算、推理训练等与人工智能大模型相关的基础设施领域，欧盟正逐渐落后。并且 FLI⁹³认为，欧洲对其他国家的技术依赖可能阻碍欧盟参与制定人工智能全球标准的努力。⁹⁴而中国由于其政治体制和政治传统的原因，中国是一个高度中央集权的国家，由一个极有组织的党中央所领导，中央指挥地方如身之使臂，极少发生抗命的事情，因此我国的政策反应速度也更加迅速，这一点相比于欧盟和美国都有更大的优势。⁹⁵

⁹⁰ 谈晓文：《欧盟数据共享治理模式研究及中国镜鉴》，载《德国研究》2024 第 5 期，第 113-133 页。

⁹¹ 国务院办公厅关于转发商务部、国家发展改革委《2025 年稳外资行动方案》的通知。中国政府网. 2025 年 2 月 17 日. 2025 年 10 月 21 日. https://www.gov.cn/zhengce/content/202502/content_7004409.htm。

⁹² 习近平会见国际工商界代表时的讲话（全文）。共产党员网. 2025 年 3 月 29 日. 2025 年 10 月 21 日. <https://www.12371.cn/2025/03/29/ARTI1743210307804350.shtml>。

⁹³ FLI（Future of Life Institute）成立于 2014 年，是一家非营利性机构，机构宗旨在于引导变革性技术朝着造福人类的方向发展，远离极端大规模风险。

⁹⁴ 全球 AI 大模型一览：中美之外还有谁？. 虎嗅网. 2023 年 3 月 28 日. 2025 年 10 月 21 日. <https://www.huxiu.com/article/957075.html>

⁹⁵ 黄仁宇，《万历十五年》（中华书局，1982 年版）。

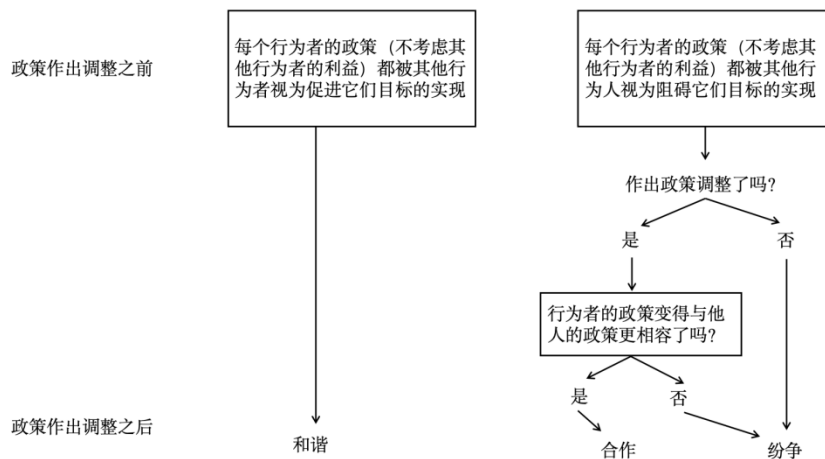


图 2：和谐、合作、纷争与政策调整之间关系图⁹⁶

资料来源：基欧汉，2006

中国凭借 14 亿人口和庞大的统一市场优势，已成为全球第二大消费市场，拥有世界上规模最大的中等收入群体，预示着巨大的投资和消费潜力。中国正致力于高质量的发展，加速推进绿色化、数字化、智能化的转型，产业配套能力强大，成为新一轮科技革命和产业变革的理想试验场。结合“一带一路”倡议和亚洲基础设施投资银行等投资项目，如果未来合作国家认为与中国合作有助于实现其人工智能发展目标，则最终可以促成和谐的共同人工智能发展环境，即便不是相同的目标，中国有较高的行政效率，也可以迅速调整政策以协调政策目标达成合作。⁹⁷

（三）实践路径：以现有国际机制的改革，作为参与全球 AI 治理的治理方式

首先，通过对现有国际机制的改革和发展，来降低交易成本。以国家完全自利为假设来讲，各国需要这样的一个国际机制，来降低交易成本，提供相对对称的信息，稳定各国预期，并使特定协议能够相对容易地作出，即在现有国际机制框架下达成特定协议的专门努力，比之没有国际机制框架下达成专门协议更加容易。⁹⁸例如世界贸易组织谈判功能重振中的“联合声明倡议”的开放式新诸边模式。⁹⁹欧盟的策略是通过另开一套“AI 框架公约”的形式对 AI 的国际治理重新构建，则其所面临的边际成本相对于利用 WTO 等已成熟的国际机制而言较大。其次，就提升透明度而言，欧盟对于透明度的规定方式基于技术中立，但会导致技术中

⁹⁶ 罗伯特·基欧汉. 霸权之后-世界政治经济中的纷争与合作. 上海：上海人民出版社，2006，第 52 页。

⁹⁷ 习近平会见国际工商界代表时的讲话（全文）. 共产党员网. 2025 年 3 月 29 日. 2025 年 10 月 21 日. <https://www.12371.cn/2025/03/29/ARTI1743210307804350.shtml>。

⁹⁸ 罗伯特·基欧汉. 霸权之后：世界政治经济中的合作与纷争. 上海：上海人民出版社，2006。

⁹⁹ 石静霞. 世界贸易组织谈判功能重振中的“联合声明倡议”开放式新诸边模式. 法商研究. 2022 年第 5 期, 第 3-17 页. 访问日期：2025 年 10 月 21 日. <https://www.rdfybk.com/qw/DownPdf?id=812637>。

立理念的理解不同。中国可以通过与各国共建透明度标准，通过技术共享与透明度规则共建的方式，提升共同透明度从而打通国内外市场。其三，稳定预期，中国可以在《全球人工智能治理倡议》的基础上建立国际人工智能技术共享平台，推动负责任技术流动，避免数据壁垒和技术垄断限制发展中国家的技术升级，确保发展中国家能公平利用数据资源。此外，全球人工智能合作也需要加强国内外企业和学术机构的协作，推动跨国企业共建人工智能产业生态，促进全球人工智能人才交流，避免“人才保护主义”阻碍技术进步，通过交流来促进理解和合作，进而通过交流稳定各方预期。¹⁰⁰

利己的政府能够在共享利益的基础上，理性地去建立国际机制，政府会遵守机制中的规则，即使这样做可能不符合它们的短视利益。在当下的逆全球化背景下，由于世界某些特殊成员国家内部存在的政策不稳定性，而导致部分国家并不遵守国际机制中的规则。WTO 及其他国际组织固然存在本身的问题，但由于构造国际机制本身的困难，应当尽可能地调整现有机制，而不是推倒重来，才是真正理性的行为。¹⁰¹因此，机制倾向于不断演化，现存的 WTO 作为国际组织是有价值的，它能够降低合法交易的成本，其已成熟的规则可以减少行为的不稳定性，其最惠国待遇、国民待遇、特差待遇等基本原则已经为大多数国家所接受，如果重新开发一套崭新的国际 AI 治理机制，则接受新国际机制的各国政府所承担的义务和成本，无法重新进行估算。

五、展望：全球 AI 治理的未来图景

人工智能作为新技术，给人类带来了机遇和挑战，欧盟以其制度先行的理念，对 AI 的国际治理进行了探索，其结合 OECD 定义以扩张规制范围的立法技术确实值得赞同。¹⁰²但欧盟和美国对于 AI 国际治理空白占领得越深入，中国的规则话语权也将愈来愈少。但即便如此，中国仍然应当根据 AI 发展的客观规律进行整理。现阶段 AI 正处于发展的起步阶段，发展才是核心需求，其安全保障应当以“伦理指导前置，公私合作治理”的方式，而非欧盟式的“横向四分法治理”。¹⁰³人类从工业革命以来，一直在面对全新的科技和规制难题，新技术会带来新问题，出现问题，强监管灭掉确实干净，禁止发展确实是一种高效的手段，但后续问题仍然会出现，结合历史思考可得出一个悖论，人本身就是问题的根源，要不要灭掉？AI 治理追求的是创新与安全，而招致治理失败的是极端。

总之，中国在未来无论是从全人类的共同福利角度还是从战略发展的角度抑

¹⁰⁰ 加强人工智能国际治理和合作（深入学习贯彻习近平新时代中国特色社会主义思想）--理论-中国共产党新闻网. 中国共产党新闻网. 2025 年 3 月 3 日. 2025 年 10 月 21 日.

<http://theory.people.com.cn/n1/2025/0303/c40531-40429403.html>.

¹⁰¹ 罗伯特·基欧汉. 霸权之后：世界政治经济中的合作与纷争. 上海：上海人民出版社，2006。

¹⁰² 欧盟 AI 框架公约对于 AI 的定义与 OCED2023 年修订的 AI 定义保持一致。

¹⁰³ 中国信息通信研究院. 人工智能治理蓝皮书（2024 年）[EB/OL].
<https://www.zgcsswdx.cn/uploads/file/20241229/1735473522180483.pdf>, 2024-12-29

或从国际安全的角度，都无可避免地要进入到 AI 国际治理领域。欧盟通过 AI 框架公约进行了尝试，但 AI 的发展是一条长期的路径，欧盟预先的国际立法未必需要担心，随着技术的发展和技術标准和伦理标准等的愈加成熟并获得广泛认可，彼时也许是真正意义上达成 AI 治理国际合作平台的契机，无论是 WTO 还是其他国际组织，如能实现降低交易成本、提高治理透明度并稳定各国预期，则其成功便具备可能性。

网络空间治理下的中国国际话语权：基本内涵、面临困境及提升路径

*蒋莉

摘要：作为人类社会的第五空间，网络空间已成为各国争夺国际话语权的关键领域。网络空间治理下国际话语权的争夺，本质上是在网络空间治理规则秩序建构中主导力和塑造力的争夺。然而，当前国际话语权在网络空间治理中分布不均，加之网络空间治理的国际规则适用范围受限以及中国国际议程设置能力不足，制约了中国国际话语权在该领域内的提升。作为互联网领域的新兴大国，中国应坚持网络主权原则，倡导构建网络空间人类命运共同体、以联合国为主要平台并借助其他多边机制，推动新的国际公约的制定，同时强化在网络空间治理领域的议程设置能力，从而提升在网络空间治理下中国国际话语权。

关键词：网络空间治理 中国国际话语权 网络主权 人类命运共同体

随着网络技术更迭加快，网络攻击行为海量增长，网络威胁也逐步从虚拟空间向物理空间蔓延，给国际网络空间治理带来了重大挑战。网络空间作为人类社会的第五空间，需要相应的规则来维持其正常秩序。目前，众多国家已经相继颁布了与网络治理相关的国内法律法规。然而，鉴于网络空间本质上是一个无物理边界且高度虚拟化的空间，单纯依靠各国的国内法律制度管理，显得捉襟见肘。因此，这一领域迫切需要国际社会共同协作，制定出一套具有普遍约束力的国际规则体系，以实现对网络空间的全面治理。不过，目前，国际社会尚未就网络空间治理形成统一且全面的国际规范体系，网络空间的国际秩序仍处于形成的初级阶段。而国际规则的构建过程，本质上是一个涉及各国利益博弈和争夺国际话语权的复杂过程。这意味着，网络空间国际规则体系逐步成型的关键时期，对于意在增强自身在网络空间治理领域影响力的国家而言，无疑是一个不可错失的历史性机遇。为了确保网络空间这一新兴领域中的话语权分配能够更加均衡，不被单一或少数国家（如西方国家）所主导，中国作为世界第二大经济体，应当敏锐地把握当前的历史契机，并积极参与其中。在此过程中，一个核心而关键的任务在于，中国如何有效提升自身在网络空间领域中的国际话语权。这不仅关乎中国在全球网络治理格局的地位和影响力，也是维护国家网络主权、促进网络空间健康有序发展的重要保障。

当前，学界对于中国国际话语权的研究，在新闻传播学、政治学、哲学、文

* 蒋莉，福州大学法学院 2023 级硕士研究生。

化学领域等多个学科内均有较为广泛的涉及。具体而言,新闻传播学领域的研究侧重于增强中国的国际传播效能,尤其是提升舆论引导能力和传播能力,旨在破除西方对中国舆论构陷和偏见。¹在政治学领域,有观点认为国际话语权本质上体现的是国家利益之间的争夺以及国家间权力的博弈。中国作为新兴大国,伴随着综合国力的显著提升,提升国际话语权以争取应有的国家权益显得尤为重要。²为此,中国应从大国外交、理念引领、国际传播以及人才培养等多个维度出发,构建一个与自身大国地位相匹配、且富有中国特色的国际话语权体系,从而增强在国际政治舞台上的影响力。³此外,我国还应高度重视“全球南方”国家话语权的发展,积极发掘和扩大共同的“朋友圈”,通过凝聚广泛共识,共同推动人类命运共同体的构建。⁴哲学和文化学领域的研究都共同呼吁深化国内理论文化的研究,尤其是需增强对中国式现代化的理论自觉。⁵与此同时,我国应深入挖掘优秀传统文化的精髓,在知识生产和理论生产上赢得主动权,并在此基础上,通过创新话语表达和叙事策略,扩大中国话语的国际影响范围。⁶相比之下,针对网络治理领域国际话语权的研究尚显不足,相关成果也聚焦于网络传播能力的话语权构建,以及如何破解西方舆论困境的策略。前述研究现状表明,一方面,尽管这些研究涵盖了不同的学科领域,但跨学科的比较与分析揭示出,这些成果在差异之中蕴含着相通之处,在于均致力于探索提升中国国际话语权的多元路径,这对于网络空间治理同样具有重要的启示与借鉴意义。另一方面,网络空间治理领域国际话语权的系统研究亟待加强,方能全面回应全球化背景下网络空间治理的复杂挑战。

本文通过解读国际话语权的内涵,分析当前国际社会的网络空间的治理现状以及中国在网络空间治理规则建构中所面临的挑战,并在此基础上,探讨并提出提升中国在网络空间治理框架内国际话语权的路径。

一、网络空间治理话语权的基本概念

(一) 国际话语权的概念

在传统语言学的角度,人们借助话语进行社会交流,“话语”作为一种交流工具,具有社会性和平等性。随着后殖民主义和后现代主义思潮的兴起,话语也超出了传统语言学的范畴与权力之间产生了密切的联系。一方面,后殖民主义主张,语言不仅是交流工具,更是一种权力结构,通过对一切渗透着人类活动的文本资料进行分析,人们可以研究蕴于语言中的权力支配关系。另一方面,后现代主义

1 孟晨、祁峰:《提升中国网络空间国际话语权:价值意蕴、战略方向与多维路径》,载《学术探索》2022年第11期。

2 陈正良、周婕、李包庚:《国际话语权本质析论——兼论中国在提升国际话语权上的应有作为》,载《浙江社会科学》2014年第7期。

3 叶淑兰:《中国国际话语权建设:成就、挑战与深化路径》,载《国际问题研究》2021年第4期。

4 孙吉胜:《“全球南方”国际话语权的构建及前景》,载《当代世界》2024年第7期。

5 付高生、龚剑飞:《中国式现代化国际话语权的丰富内涵、建构境遇与提升路径》,载《江西社会科学》2024年第4期。

6 张志洲:《增强中国国际话语权:理论自觉、话语创新与问题应对》,载《当代世界与社会主义》2024年第2期。

认识到话语在权力组成中的作用,指出所有权力政治的现实都是通过话语得以构建。⁷其中,法国哲学家米歇尔·福柯提出的“话语权力论”十分具有代表性。该理论主张,“人通过话语赋予自己权力”,认为话语不仅仅是思维符号和交际工具,而且是人们斗争的手段和目的。⁸此外,从语义的层面深入剖析,“权”这一概念有双重意蕴,一为“权利”,二为“权力”。话语权首先应具备发声的“权利”,这是话语权得以施展其效能的前提与基础,构成了话语得以表达的先决资格,确保了主体拥有发声的可能性。缺失了“权利”层面,话语权便失去了行使的正当资格。而“权力”则影响着话语的实际效力与影响力范围,行为主体借助话语在某一领域争夺更多的权力,以此拓展自身影响力,这是话语权的核心要义。同时,“权力”也决定了行为主体的话语能否得到广泛认可并最终有效施行。换言之,话语权是指个体或者群体通过话语媒介传播理念与观点,并据此提升自身在某一特定领域中影响力和说服力的能力。

随着国际竞争态势的加剧,国际话语权已逐渐演变国际社会利益博弈的核心焦点。关于国际话语权这一概念,学界也存在诸多观点。梁凯音认为,国际话语权指以国家利益为核心,就国际社会事物和社会发展发表意见的权利。⁹张康乐认为,国际话语权是涉及媒体传播、国际政治、国际法律秩序等多个领域内国际影响力的体现。¹⁰赵磊认为,国际话语权以国家实力和国际道义为基础,目的是一国价值观在国际社会实现共振。¹¹龙钰认为,国际话语权是指国家行为体在国际社会中,为争取国家权益与提升国际地位,所具备的话语表达资格及其实际影响力的体现。它不仅彰显了一国话语在国际话语体系中的优势地位,更体现了一国话语对国际话语体系的影响力、主导力、支配力。¹²可以看出,学者们普遍认同,国际话语权体现了一国在国际社会中的综合影响力,是其叙述自身故事、输出与传播价值观的资格与能力。进一步而言,国际话语权对于国际规则和标准的塑造力和支配力,不仅彰显了其内在的权力构成,也是国家在国际社会中国际地位的重要体现。综上所述,国际话语权可以理解为以国家利益为核心,涉及国际传播、国际政治和国际法律秩序等多个领域,以争取国家权益和国际地位为目的,兼具权利和权力特性的能力,它不仅赋予了一国对国际事务发表意见的权利,而且也是一国话语对国际规则的塑造力和支配力的综合体现。

（二）网络空间治理的界定

7 阮建平:《话语权与国际秩序的建构》,载《现代国际关系》2003年第5期,第31页。

8 陈正良、周婕、李包庚:《国际话语权本质析论——兼论中国在提升国际话语权上的应有作为》,载《浙江社会科学》2014年第7期,第78页。

9 梁凯音:《论中国拓展国际话语权的新思路》,载《国际论坛》2009年第11卷第3期,第43页。

10 张康乐:《增强国际话语权的国际法路径研究》,载《中国社会科学院大学报》2022年第8期,第111页。

11 赵磊:《增强中国国际话语权的现实挑战与应对》,载《人民论坛·学术前沿》2024年第5期,第77页。

12 龙钰:《叙事·制度·学术——中国国际话语权提升理路》,载《中南大学学报(社会科学版)》2022年第28卷第6期,第141页。

为厘清“网络空间治理”的概念，首先在于明确网络空间的本质。随着互联网技术的飞速发展，网络空间应运而生。其不仅是第三次科技革命（其标志性技术成就涵盖核能、电子计算机与自动化技术）的重要产物，还广泛吸纳了目前蓬勃发展的以互联网为依托的大数据和物联网以及人工智能技术的第四次科技革命的核心内容。因而，网络空间是一个不断动态发展的概念范畴。¹³那么，何谓网络空间治理？以往，“互联网治理”更为经常使用。根据联合国互联网治理工作（Working Group on Internet Governance）的定义，“互联网治理是政府、私营公司和社会，根据各自的角色、共享的原则、规范、规则、决策过程以及程序，来规范互联网的发展和使用”。¹⁴不过，“互联网治理”的议题主要聚焦于技术维度和社会层面的交互影响，ICANN（Internet Corporation for Assigned Names and Numbers）的成立开启了互联网应如何进行有效控制和管理的初步探索。¹⁵但随着互联网技术的进一步发展，互联网治理的议题已不再局限于技术与社会范畴，而逐渐扩展至政治、经济、文化、社会乃至军事战略等多个维度。在此背景下，互联网治理的内涵和外延不断得到丰富和拓展，逐渐转向“网络空间治理”。网络空间治理旨在确保网络空间稳定、安全及有序发展，并通过政府、企业、个人等多方主体的协同合作，遵循共同原则与规则，管理网络事务。¹⁶综上所述，无论是互联网治理还是网络空间治理，均体现了多元主体协商共治的过程，只是网络空间治理的涵盖范围更为广泛，不仅涉及到利用技术和法律手段规范网络，解决技术、政治、经济、文化等方面的问题，还涵盖网络空间资源分配和权力划分的问题。¹⁷本文所探讨的是在网络领域中如何提升中国国际话语权的问题，其不仅与国家的政治、经济以及互联网技术等层面息息相关，而且涉及到不同国家间对国际规则与秩序构建的主导权争夺。鉴于此，使用网络空间治理这一表述更为恰当。

（三）网络空间治理下的国际话语权的界定

网络空间治理下的国际话语权可以被视为是国际话语权在网络空间的拓展与延伸。相较于其他领域的全球治理，网络空间治理领域内缺少全面且统一的国际规范，尚未形成稳定的国际秩序。鉴于此，各国在网络空间治理进程中，积极通过话语权的竞争与博弈，试图获得更多的权力资源和影响力。如前所述，话语权不仅局限于“发声”的基本权利，更是话语的威力与有效性，它既是权力行使的一种工具，也是权力分配与运作的结果体现。一国能否在网络空间治理中占据话语权及其话语权的强弱，将直接影响其在该领域内的影响力大小。然而话语权

¹³ 蔡翠红：《全球大变局时代的网络空间治理》，载《探索与争鸣》2019年第1期，第24页。

¹⁴ WGIG. Report of the working group on internet government[EB/OL].<https://www.doczj.com/doc/4610730888.html>, 2024年12月11日最后访问。

¹⁵ 郎平：《概念变迁视角下的网络空间国际治理》，载《信息安全与通信保密》2024年第4期，第9页。

¹⁶ 闫晓丽：《网络治理的概念及构成要素》，载《网络空间安全》2018年第9卷第5期，第28页。

¹⁷ 杨晓茹：《中国共产党网络空间治理思想研究》，华南理工大学2022年博士学位论文，第25页。

不仅体现为一种影响力,在更深层次上,它是一种具备塑造社会认知、构建话语框架与导向舆论走向的塑造力。互联网时代,话语权更是规则制定与未来秩序主导权的关键所在。¹⁸换句话说,拥有更多的话语权的一方,往往能够主导网络空间规则的制定,进而引导并重塑网络空间治理领域的国际秩序。因此,网络空间治理下的国际话语权可以界定为一国在网络空间中所展现的话语影响力,以及在网络空间规则制定过程中的主导能力,这种能力不仅决定了其在当前治理格局中的地位,更能引导未来网络空间治理领域国际秩序的重塑方向。

二、国际网络空间治理现状

(一) 网络空间治理规则和秩序主导权博弈激烈

随着网络技术的迅速发展,网络安全威胁的态势日益严峻,这对国际社会带来了前所未有的挑战,亟需建立新的网络空间治理规则,以有效应对当前的和未来可能涌现的各种风险。因此,各国纷纷开展对话与交流,旨在推动新的网络空间治理国际规则的制定。与此同时,围绕网络空间治理规则和秩序的构建,不同国家及阵营之间竞争与博弈愈发激烈,各方均期望通过该领域的规则制定和秩序塑造,争取更多国家利益,并以此提升自身的国际影响力和国际话语权。

以惩治网络犯罪为例,目前最具有代表性的国际公约主要是《布达佩斯网络犯罪公约》和《联合国打击网络犯罪公约》。《布达佩斯网络犯罪公约》是由欧洲委员会牵头,美国、加拿大、日本等国参与制定的世界上第一部网络犯罪公约。该公约界定了各种网络犯罪行为,明确规定了各国在应对网络犯罪应采取的法律措施,并着重强调开展跨国合作的重要性,目的是在全球范围内提升打击网络犯罪的能力。¹⁹然而,该公约只是一个区域性的公约,欧美国家却试图将其打造成为打击网络犯罪的全球性的法律标准,并且反对谈判制定新的全球性的打击网络犯罪的公约。²⁰并且,该公约在实际应用中却显现出了若干局限性。一方面,该公约内容重点针对的是计算机相关犯罪,对于网络恐怖主义和利用互联网实施的传统犯罪等领域并未涉及。另一方面,从利益平衡的角度分析,该公约由欧美国家牵头起草,主要体现了发达国家的利益诉求,未能充分兼顾发展中国家与发达国家利益平衡。这种利益分配不均在一定程度上限制了该公约在全球范围内的广泛接受度和影响力。此外,该公约需要现有缔约国一致同意的加入模式不具备国际公约所具有的开放性,也削弱了该公约在全球范围内的实施效果。²¹因此,国际社会迫切需要一个更具有广泛代表性和适应当前网络犯罪状况的新公约。

2024年8月8日,根据联合国决议成立的特设委员会,在其第7次总结会议续会上,经表决正式通过了《联合国打击网络犯罪公约》草案。该草案的顺利通

¹⁸ 王向阳:《网络空间治理的国际规范研究》,载《情报杂志》2021年第40卷第7期,第82页。

¹⁹ 李哲、朱晓琴:《〈联合国打击网络犯罪公约〉的中国立场与核心问题》,载《北京师范大学学报(社会科学版)》2024年第5期,第125页。

²⁰ 徐峰:《网络空间国际法体系的新发展》,载《信息安全与通信保密》2017年第1期,第77页。

²¹ 江溯:《打击网络犯罪的国际法新机制》,载《法学》2022年第11期,第49页。

过,是中国、俄罗斯等发展中国家阵营多年不懈努力与积极推动的成果,标志着联合国框架下,打击网络犯罪国际新秩序逐步形成。²²《联合国打击网络犯罪公约》的重要目标之一在于致力于消除数字鸿沟,减轻技术发展不平衡所带来的负面影响。同时,该公约也旨在为打击网络犯罪的国际合作提供一个在全球范围内具有普遍约束力的法律框架,对于网络空间的专门性立法而言,具有重要的示范意义与引领作用。²³尽管相较于《布达佩斯网络犯罪公约》,该公约在诸多方面取得了显著性进步,例如在国家间互相提供技术援助和能力建设的问题上,增加了考虑发展中国家利益和需求的内容。但是,在该公约草案的谈判过程中,发展中国家还是不得不作出了一定的妥协和让步,尤其是在刑事定罪的范围和与数据主权相关的跨境取证这两个核心问题上,发展中国家与欧美国家之间存在较大分歧,双方仍在持续进行博弈与协商。未来,如何进一步平衡各方利益,推动公约的完善与实施,将是国际社会面临的重要课题。

在网络安全国际规则制定领域,最具影响力的国际组织文件是《塔林手册》和《信息安全国际行为准则》(以下简称《行为准则》)。2013年,北约下设的研究与培训机构“网络防务合作卓越中心”推出了《关于网络战国际法适用的塔林手册》(以下简称《塔林手册》1.0版),主要对武装冲突法适用于网络空间进行了具体解释和讨论。该手册将网络安全明确为维护国家主权的重要内容,将境外网络攻击由犯罪行为上升为战争行为,并规定国家可以依据国际法采取军事行动。²⁴2017年,北约进一步推进相关工作,组织撰写了包括平时法在内的《塔林手册》2.0版本,将原先用于处理网络战争的法律拓展到和平时期网络行动的国际法规则,实现了战争时期与和平时期网络空间国际规则的全覆盖,这一新版本的手册也更名为《可适用于网络行动的国际法的塔林手册》。从某种程度上讲,《塔林手册》填补了国际网络空间法律规则的空白,为网络战和网络行动等相关问题提供了重要的规则指引。然而,值得注意的是,该手册只是一个示范性文件,尚不具备法律拘束力。而且,《塔林手册》无论是在专家组的构成上,还是在2.0版本新增专题的甄选以及若干规则和评注的内容的形成上,首先体现了西方国家的立场和关切。²⁵例如,西方国家认为现行国际法应适用于网络空间治理,主张既存的且具有国际习惯法地位的“实然法”,而不是倡导新的“应然法”。²⁶从本质上审视,该手册仍是基于西方国家主导规则体系下的产物。而《行为准则》则是中俄等广大发展中国家针对网络空间信息安全威胁制定的规则实践,2015年,中

²² 同前注 18

²³ 武汉大学:《网络空间国际规则前沿与态势简报第四十六期》,载微信公众号“网络空间国际法前沿”2024年9月10日, <https://documents.un.org/doc/undoc/gen/v24/036/33/pdf/v2403633.pdf>.

²⁴ 司文文:《塔林手册》,载《中国投资》2018年第13期。

²⁵ 黄志雄:《网络空间国际规则制定的新趋向——基于〈塔林手册2.0版〉的考察》载《厦门大学学报(哲学社会科学版)》,2018年第1期,第7页。

²⁶ 黄志雄:《〈塔林手册2.0版〉:影响与启示》,载《中国信息安全》2018年第3期,第85页。

俄等上合组织成员国共同向联合国大会提交了经过修订的《行为准则》。该准则详尽地列出了十三条网络空间国际行为的规范,广泛涵盖了国际和平与安全、人权和基本自由、打击网络犯罪和恐怖主义、互联网国际治理、能力建设和信任措施建设等多个方面,是目前国际社会在信息安全领域最为全面且系统的文件。该准则的核心宗旨旨在倡导构建一个和平、安全、开放、合作的信息空间环境,确保信息和网络技术能够促进人类的全面发展、保障民众福祉,并有力维护国际和平与安全。²⁷然而,遗憾的是,后续由联合国专家组就此展开的讨论未能促成新共识的形成,最终未能实现对该《行为准则》的预期进展。

尽管《布达佩斯网络犯罪公约》和《联合国打击网络犯罪公约》已经为打击网络犯罪提供了相对全面的规则框架,然而它们的适用范围仅限于网络犯罪领域。而且,《塔林手册》在初期同样着重探讨了网络犯罪的相关议题,这说明了国际社会最初聚焦于犯罪这一极端严重的行为。此外,虽然《行为准则》在当前国际社会的信息安全领域中是较为全面且系统的文件,但它也只是提出了较为原则性的规定,在法律条文上缺乏明确性和具体性。例如,《行为准则》并未明确规定如何界定信息安全威胁和应采取的措施,各国对准则的理解和适用上可能存在差异,从而影响准则的普遍性和有效性。而且,它与《塔林手册》均不属于国际法的范畴,亦不具备法律约束力。这些现象反映出当前网络空间规则的适用范围尚显不足,仍有待进一步拓展和完善。

(二) 网络空间治理模式之争

在探讨国际网络空间治理模式的议题中,各国意见分歧显著,尤以“多边”与“多利益攸关方”两种治理模式之争为甚。西方国家秉持“全球公域”理念,力推“多利益攸关方”模式,视网络空间为全球不可分割的一个整体,反对任何国家单独行使主权与管辖权,主张应由政府、私营部门及社会公众等多元主体共同制定网络运行规范。²⁸其中,具有代表性的是美国主导的 ICANN 的“多利益攸关方”治理模式。然而,“棱镜门”事件后,美国虽迫于舆论压力,有意放弃对 ICANN 的直接管理,但坚持管理权转移须维持“多利益攸关方”模式。²⁹由政府、私营部门及社会公众等多利益攸关方主导网络空间政策和标准制定,是一种自下而上的网络空间治理模式。在该模式下,其他国家政府必须在政府咨询委员会达成普遍共识后,方能向 ICANN 董事会提出建议。³⁰尽管政府也是多元共治的主体,但其并不能在网络空间治理中发挥实质性的影响,主权国家的角色逐渐被边缘化了。

²⁷ 赵宏瑞、李树明:《网络空间国际治理:现状、预判、应对》,载《广西社会科学》2021年第11期,第109页。

²⁸ 郭炯、洪永红:《全球网络治理的法律困境与出路》,载《湘潭大学学报(哲学社会科学版)》2017年第41卷第3期,第26页。

²⁹ 龙坤、朱启超:《网络空间国际规则制定——共识与分歧》,载《国际展望》2019年第11卷第3期,第48页。

³⁰ 李彦:《中国参与互联网治理国际制度建构的议程设置策略》,载《传媒观察》2020年第3期,第72页。

因为该模式从本质上来讲,是美国为了防止其“敌对国家”参与国际网络空间治理的手段。换言之,“多利益攸关方”模式表面上是为非国家行为体主导管理互联网创造条件,但实际上仍是网络强权主导下的超主权模式。³¹

相对而言,中俄倡导“多边”治理模式,强调政府在网络空间治理中的核心作用。《行为准则》第八条明确指出:“各国政府应平等参与、履行职责,推动构建多边、透明、民主的互联网国际管理体系,确保资源公平分配、普及接入,以及互联网稳定安全运行。”此条款反映了“多边”治理模式核心理念。该模式实质上是中俄等发展中国家在网络空间治理中坚持主权原则的体现,即旨在以政府主导、多边参与为指导原则,推动网络空间国际规则朝着更加公正合理的方向发展。这两种治理模式的争论本质是不同国家间的治理理念的分歧,更体现了对网络空间治理下国际话语权的核心内容界定权的角逐。其根本目的在于影响网络空间治理领域国际规则的构建和塑造,从而引导规则及其制定过程朝着最符合自身利益的方向发展。

（三）中国的网络空间治理现状

针对网络空间治理,中国分别从国内和国际层面用实际行动积极回应,以便更好地应对未来可能面临的挑战。在国内层面,2014年,中央网络安全和信息化小组的正式成立,标志着我国网络发展的一个重要转折点,即从单纯的信息技术基础设施建设阶段,跨越至从网络空间安全的这一视角来审视全球态势的阶段。这一举措不仅预示着我国正从网络大国稳步迈向网络强国,也彰显了网络安全已被提升至国家战略的核心地位。随后,在法制建设方面,我国于2015年7月通过了《国家安全法》,同年12月又颁布了《反恐怖主义法》。这些法律的出台为构建国家安全法律体系奠定了坚实的基础。2016年11月颁布的《网络安全法》更是对网络信息安全、网络运行安全、网络风险的检测预警等方面都作出了明确规定,进一步强化了网络空间安全防护体系。与此同时,为了细化并补充网络安全法律体系,一系列的行政法规、部门规章和规范性文件相继出台,针对特定领域的网络安全问题提供了具体的指导和规范,如《计算机信息系统安全保护条例》《计算机信息网络国际联网安全保护管理办法》《计算机软件保护条例》等。³²在个人信息保护和数据安全上,我国更是加大了立法力度,先后出台了《个人信息保护法》和《数据安全法》,为公民个人信息安全和数据安全提供了坚实的法律保障。在刑法领域,2009年通过的《刑法修正案(七)》增加了“出售、非法提供公民个人信息罪”“非法获取公民个人信息罪”“非法控制计算机信息系统罪”“提供侵入、非法控制计算机信息系统程序、工具罪”,进一步加大对侵犯公民个人

³¹ 赵宏瑞、李树明:《网络空间国际治理:现状、预判、应对》,载《广西社会科学》2021年第11期,第110页。

³² 郭少青、陈家喜:《中国互联网立法发展二十年:回顾、成就与反思》,载《社会科学战线》2017年第6期,第217页。

信息和计算机系统安全的惩处力度。而 2015 年通过的《刑法修正案(九)》则明确规定了网络服务提供者负有安全管理公民个人信息的义务,进一步强化了网络服务商的责任。这一系列网络空间治理相关的法律、行政法规和政策文件的出台,为中国更好地参与网络空间治理提供了法律指引和行动指南。

在国际层面,中国提出构建网络命运共同体的理念,并主动融入联合国网络安全的进程中,坚决拥护联合国在国际网络空间治理中发挥核心领导地位。例如,中国和俄罗斯一起向联合国提交新的《行为准则》,积极推动《联合国打击网络犯罪公约》谈判进程等。³³而且,中国还积极倡导网络空间的国际合作,于 2017 年 3 月发布首份《网络空间国际合作战略》,首次全面系统地提出了中国在推动网络空间国际交流合作方面的主张,以期为全球网络空间的和平、安全与发展贡献了中国智慧与中国方案。中国还举办了世界互联网大会,进而成立了世界互联网大会国际组织,旨在为全球互联网领域的资源共享与协同治理提供一个高端对话与合作的平台。此外,外交部条约法律司作为我国网络国际法领域的主管部门,近年来积极开展网络空间国际法问题实务和调研,包括牵头网络空间全球大会(即“伦敦进程”)、中国—东盟网络工作小组、亚洲—非洲法律协商组织网络空间国际法工作组、联合国网络犯罪政府专家组等多个重要国际平台的工作,同时,参与其他部门涉网络外交和法律事务的协作。³⁴从举办世界互联网大会到成立世界互联网大会国际组织,再到举办中非互联网发展与合作论坛与中国—东盟信息港论坛等活动的成功举办,中国始终致力于打造一个和平、安全、开放、合作且有序的网络空间环境,推动网络空间命运共同体的构建进程。³⁵这一系列的举措不仅体现了中国在网络空间治理领域的责任担当,也彰显了其对全球互联网健康发展的深远考量与积极贡献。

三、网络空间治理中国国际话语权提升的困境

(一) 网络空间治理下国际话语权不平衡

权力结构是影响国家发展的基础性因素,个人或者利益集团通过权力资源博弈塑造制度和规则,从而影响国家未来的发展格局。³⁶网络空间治理下的国际话语权是权力分配和运作结果的体现,其力量大小不仅决定了利益相关活动中一国的地位、支配权力以及施加影响的范围,而且深刻影响着网络空间国际规则的制定、解释与适用。换言之,网络空间治理下国际话语权的分配,不仅影响着网络空间中国际规则的制定和国际秩序的重塑,更是衡量各国在网络空间治理领域国际影响力的大小和话语权强弱的重要标尺。

当前**网络空间治理国际秩序**正处于初期形成阶段,中国作为互联网领域的核

³³ 《携手构建人类命运共同体:中国的倡议与行动》,载《人民日报》2023 年 9 月 2 日,第 6 版。

³⁴ 徐峰:《网络空间国际法体系的新发展》,载《信息安全与通信保密》2017 年第 1 期,第 78 页。

³⁵ 林子涵:《中国为全球网络空间治理贡献力量》,载《人民日报海外版》2023 年 8 月 14 日,第 10 版。

³⁶ 高波:《权力结构视角下的发展陷阱——基于对委内瑞拉“蓬托菲霍体制”的分析》,载《中国国际政治研究》2020 年第 1 期,第 15 页。

心国家，是参与**国际网络空间治理**的不可或缺的主体力量。对于中国而言，这正是中国为**网络空间治理国际秩序**构建建言献策的重要机遇。但是，中国参与国际网络空间治理仍面临多重挑战。譬如，目前欧美等西方国家已占据了网络空间治理先机，导致网络空间治理国际话语权不平衡，西方国家在该领域的影响力仍更为显著。如前所述，不论是《布达佩斯网络犯罪公约》，还是《塔林手册》都是由欧美国家牵头制定，体现其利益诉求。同时，由于网络空间发展的不均衡性导致数字鸿沟扩大，³⁷给国际网络空间治理增添了复杂性。另外，作为互联网和域名系统的核心基础资源，根服务器的运行和管理影响着全球互联网发展格局。而全球 13 个根服务器中，美国掌握了主服务器在内的 10 个根服务器。因此，长期以来，美国凭借其自身拥有的信息技术、人才、资本等显著优势，以及在各类互联网组织中的主导地位，牢牢把握了互联网资源的分配，并垄断了技术标准的制定，从而推行网络霸权，³⁸导致了网络空间治理的权力结构失衡。网络空间权力结构作为物理空间权力结构在网络空间的映射，其当前形态和配置主要反映了物理空间的政治权力结构与配置的特征。³⁹这种权力结构的失衡，实质上揭示了网络空间治理下国际话语权的失衡状态，即网络空间治理国际话语权仍掌握在欧美国家手中。

（二）网络空间治理的国际规则适用范围具有局限性

国际规则的制定以各国之间达成的共识为基础，即所谓的“国家同意”。但是，在网络空间治理国际规则制定的问题上，各国之间的意见分歧较大。发展中国家认为，现有的国际法规则难以应对新兴网络空间治理的复杂性，亟需新的国际法规则规范网络空间活动。而欧美发达国家则主张，直接将现有的国际法规则适用于网络空间治理足以应对网络空间治理中的挑战。以《联合国打击网络犯罪公约》为例，在是否应将使用计算机技术而触犯的传统犯罪纳入公约所打击的网络犯罪范畴问题上，发展中国家普遍赞成将其纳入公约的定罪范围，但欧美等发达国家仍然持反对态度，坚持主张直接适用传统国际法。

而在跨境取证相关的数据主权问题上，发达国家与发展中国际也存在很大分歧。例如，发展中国家坚持国家主权原则和数据主权原则，反对“数据控制者模式”，主张“数据储存地模式”，即支持跨境取证时，国内主管机关不能强制要求本国境内的个人或网络服务商提交其控制或拥有的数据，而是应该请求数据储存地有关机关提供数据，这实质上是为了应对美国的“长臂管辖”。⁴⁰然而，这一主张

³⁷ 翁臻：《网络空间命运共同体视角下网络空间国际治理》，载《海峡法学》2023 年第 25 卷第 4 期，第 32 页。

³⁸ 方芳、杨剑：《网络空间国际规则：问题、态势与中国角色》，载《厦门大学学报(哲学社会科学版)》2018 年第 1 期，第 23 页。

³⁹ 王向阳：《网络空间治理的国际规范研究》，载《情报杂志》2021 年第 40 卷第 7 期，第 81 页。

⁴⁰ 李哲、朱晓琴：《〈联合国打击网络犯罪公约〉的中国立场与核心问题》，载《北京师范大学学报(社会科学版)》2024 年第 5 期，第 131-132 页。

遭到了欧美等发达国家的强烈反对。此外，作为目前国际社会具有代表性的网络空间治理规则，《布达佩斯网络犯罪公约》和《联合国打击网络犯罪公约》只是一个针对网络犯罪领域的规范性文件，对于亟需关注的关键信息基础设施安全、跨境数据保护、网络攻击溯源等问题，目前仍缺乏统一的国际法规范，导致网络空间治理国际规则适用范围具有局限性。同时，《塔林手册》和《行为准则》也仅仅是在网络安全领域具有影响力的国际组织文件，并不具备法律拘束力，这也反映了国际社会目前大多数规则以不具有法律拘束力的软法形式存在，从而在适用性和确定性上易引发诸多问题。

（三）中国在网络空间治理中议程设置能力不足

成功的议程设置是提升国际话语权的重要渠道。⁴¹然而，当前中国在网络空间治理体系中的国际话语权与其日益增强的综合实力之间处于不对称的状态。导致这种不对称的状态的重要原因之一在于中国在网络空间治理议程设置方面的能力不足，主要体现在以下几个关键维度：

其一，中国在形成国际重大网络治理议题方面的能力有待加强。重要性议题的选择和界定是获取国际关注的前提。相较于其他全球治理议题，我国目前对于网络空间治理议题的研究关注相对较少，但欧美国家已将网络空间议题视为关键的研究方向，从而导致国际重大网络议程的主导权往往集中于西方国家手中。这些国家凭借其深厚的历史积累、技术优势及强大的国际影响力，在国际话语权的竞争中抢占了先机，⁴²从而在网络空间治理领域国际话语权的争夺中取得了领先地位。

其二，中国在传播议题方面的能力有所欠缺。一方面，西方国家凭借其网络技术优势和网络语言优势引导国际议题的发展，将不符合本国利益和价值观的议题重新加工整合，甚至时常肆意歪曲事实抹黑中国，塑造了不利于我国的舆论环境，以至于我国的议题难以寻求广泛认同。另一方面，我国网络话语国际传播能力不足。我国不仅面临着世界主流媒体缺失的问题，而且各种非媒体形式的国际传播途径，如国际会议、国际组织以及民间外交等网络传播途径发育不全，且两者之间协调程度不够。⁴³因此，我国的网络空间治理议题难以得到有效传播。

其三，网络空间治理议程设置的制度平台受限。议题的选定和利益动员只是为议题进入议程做铺垫，而“进入渠道”则是国际议程设置成功与否的关键。⁴⁴“进入渠道”是指行为体构建议题的场所，国际议程的设置是通过控制“进入渠道”

⁴¹ 杨娜、程弘毅：《理解国际议程设置：议题传播与框架互动》，载《当代亚太》2022年第5期，第136页。

⁴² 胡文秀、刘振霞：《全球治理视域下中国国际话语权的提升》，载《中州学刊》2020年第8期，第57页。

⁴³ 许开轶：《论中国网络空间国际话语权的建构》，载《南京师大学报(哲学社会科学版)》2022年第4期，第121-123页。

⁴⁴ John A. Vasquez & Richard W. Mansbach, "The Issue Cycle: Conceptualizing Long-Term Global Political Change," pp.260-261(1983).

来完成的。⁴⁵换言之,“进入渠道”即议程“在哪说”。⁴⁶在国际网络空间治理领域,议程设置的平台与场所主要是指与网络空间治理相关的国际组织和国际会议,议题发起者利用这些平台开展利益动员,旨在推动本国议题上升为国际议程,⁴⁷进而提升该国的国际话语权。然而,中国的议程设置面临着制度平台受限的挑战。例如,ICANN大会、WSIS(World Summit on Information Society)和IGF(Internet Governance Forum)均是网络空间治理议程设置的重要平台,但欧美等网络强国凭借其自身的软硬实力,频繁主导网络空间治理国际议题的界定,不断巩固自身在网络空间治理体系中的话语权,导致中国难以有效地推进网络空间治理国际议程。因此,在网络空间治理国际规则的制定角逐中,我国对国际网络空间治理议程设置的影响相对有限,这无疑也制约了我国网络空间治理国际话语权的提升。

四、提升中国网络空间治理的国际话语权的途径

(一) 坚持网络主权原则,推动构建网络空间人类命运共同体

国家主权作为《联合国宪章》的基本原则,强调主权平等是国际合作与交往的基石。主权平等是合作的基础,而公平是构建国际秩序的内在要求。网络主权原则旨在构建合理有效的网络空间秩序规则,要求各国在网络空间中也应平等地参与国际决策与合作,确保网络主权得到尊重。《联合国宪章》所确立的主权原则及其精神,是当代国际关系的基石,这一原则同样适用于网络空间的治理。习近平总书记也早在首届互联网大会上便提出了“尊重网络主权”的观点,并在随后的第二届互联网大会讲话中,将其列为全球互联网治理体系变革的四大原则之首,进一步凸显了网络主权原则的重要性。⁴⁸此外,《行为准则》也明确指出,与网络相关的政策决策属于各国主权范畴,这不仅代表了中国一贯坚持网络主权原则的立场,也倡导将该原则确立为网络空间治理的基本原则,体现了中国在网络空间治理领域的积极态度与贡献。

然而,鉴于网络空间所具有全球性和虚拟性,其边界的界定相较于海域、空域等自然空间显得尤为复杂,因而,一些以美国为首的西方国家试图借助“全球公域”理论模糊网络空间的边界,主张网络空间不存在主权,任何个体或组织均可对其进行自由利用和管理,进而拓展自身的权利范畴。⁴⁹这些国家表面上是宣扬网络自由主义,实际上则是利用自身在网络领域的强国地位和技术优势,以巩固和扩大其在网络空间的霸权。根据国际电信联盟(ITU)2024年的统计数据,全球约80亿人口中,仍有约三分之一处于无法接入互联网,且这部分人群主要

⁴⁵ Steven G. Livingston, The Politics of International Agenda—Setting: Reagan and North—South. *International Studies Quarterly*, (3), P314—326 (1992).

⁴⁶ 李益斌、王昊语:《国际议程设置与中国倡导信息安全规范的扩散》,载《外交评论》,2023年第2期,第28页。

⁴⁷ 韦宗友:《国际议程设置:一种初步分析框架》,载《国际关系研究》2011年第10期,第48页。

⁴⁸ 黄志雄:《筑牢网络空间治理的主权基石》,载《中国信息安全》2021年第11期,第66页。

⁴⁹ 罗楚湘:《网络空间国际治理中国方案的形成、遵循与路径》,载《广西社会科学》2024年第1期,第115页。

分布在发展中国家，凸显了发达国家与发展中国家之间存在巨大的数字鸿沟。⁵⁰作为网络空间治理的重要参与者，中国代表着广大发展中国家的利益，有责任在网络空间治理国际规则制定过程中，坚持以网络主权为基本原则，即应在尊重网络主权的前提下发展真正“多边主义模式”的发展，坚决反对任何以制定网络空间国际规则为由，干涉其他国家网络空间治理事务的“霸权主义”和“单边主义”行为。同时，中国也应坚持倡导各国应在尊重互联网自身的发展规律基础上，充分考虑各国国情差异，允许各国自主选择互联网发展的道路和管理模式，确保各国能够真正平等地参与国际网络空间的治理。⁵¹此外，中国可以依托人类命运共同体的理念，及共建共商共享的原则，积极倡导互联网基础设施互联互通，以缩小国家之间数字发展鸿沟，推动经济全球化的深入发展，增进各国人民福祉。尤其是要重视“全球南方”作用，中国与东盟、非洲国家以及拉美国家间长期开展网络空间安全国际合作，增强了网络空间治理共识，该举措是构建网络空间人类命运共同体的伟大实践，促使“全球南方”国际话语权的影响力不断增强。⁵²我国应继续深化与“全球南方”国家之间网络空间治理合作，从而进一步提升中国网络空间治理中的国际话语权。

（二）以联合国为主要平台，辅以其他多边机制推动制定新的多边公约

中国应积极有效地参与进由联合国主导的，以政府专家组和开放式工作组为核心的“双轨制”框架下针对网络空间国际法建构的各项活动中。该“双轨制”的目标就是形成一个基于共识的、全面的网络空间法律框架，有助于扩大多边立法路径，有利于非成员国，尤其是发展中国家，能够根据本国法律实践，参与到网络空间国际规则建构过程中，⁵³充分体现了多边利益的代表性，与我国倡导的多边主义理念相契合。中国应当充分利用联合国“双轨制”提供的平台，在数据治理、人工智能治理等新兴领域积极推动新议题的研讨与探索，力求在这些领域掌握规则的创制权和话语权。⁵⁴通过将本国利益融入创制的国际规则制定的过程中，中国才将有机会影响对网络空间国际规则的走向产生实质性影响，⁵⁵从而由规范倡导者逐渐变为规范的领导者。

此外，中国可以积极借助上海合作组织（上合组织）及世界互联网大会等区域性平台的力量，以推动网络空间治理相关多边公约的制定。上合组织作为一个

⁵⁰ ITU, Internet use, <https://www.itu.int/itu-d/reports/statistics/2024/11/10/ff24-internet-use/>

⁵¹ 芮必峰、张冰清：《建立国际网络空间新秩序》，载《国际新闻界》2017年第39卷第6期，第13-14页。

⁵² 安怡宁、陈兆源：《携手“全球南方”构建网络空间命运共同体》，载《人民论坛》2024年第11期，第82页。

⁵³ 田立：《国际安全视角下的中国参与网络空间国际法建构的路径选择》，载《云南社会科学》2021年第6期，第95页。

⁵⁴ 刘金河、杨乐：《联合国框架下双轨制进程的博弈焦点与发展趋势》，载《中国信息安全》2021年第9期，第84页。

⁵⁵ 宋杰：《从英美实践来看我国参与国际法律事务的有效性》，载《比较法研究》2015年第2期，第158页。

以中国为主导的区域性政府间合作组织，虽然在包容性方面存在一定局限性，但中国仍可通过在成员国间达成的合作共识，努力将其转化为具有普遍约束力的规范性共识。同时，世界互联网大会作为中国主导的新兴国际组织，已成长为国际网络空间治理中不可或缺的重要力量，为各国政府、企业及学者搭建了交流、对话与合作的桥梁，并为国际网络空间治理做出了积极贡献。为充分发挥这些区域性平台的作用，中国应在这些平台上积极发表官方声明与建议，利用多边机制来促进网络空间治理相关多边公约的形成。鉴于国际法本身是一个在实践中不断演进与发展的规则体系，明确的官方主张和政府实践对于国际规则的塑造具有举足轻重的作用。⁵⁶因此，我国需在该平台举办的各项活动中积极提出我国有关网络空间治理的主张，挖掘共同“朋友圈”以形成广泛共识，从而推动其上升为制度化和规范化的多边公约，这不仅是中国在网络空间治理领域发出声音、表达立场的重要途径，更是在推动相关规范形成的过程中，提升中国国家话语权的有效方式。

（三）强化中国网络空间治理议程设置能力

中国作为全球互联网使用人口最多的国家，应该充分利用其网络大国的独特优势，在网络空间治理领域深化理论贡献，强化理论供给的广度和深度。在此过程中，中国应坚定不移地秉持网络多边主义，积极推动网络空间人类命运共同体的构建，并发起更多旨在提升发展中国家在网络空间治理下国际话语权的议题。另一方面，中国需充分发挥政府在治理过程中的主导作用。具体而言，政府应定期发布一些较为全面的、深入的网络空间治理的官方声明文件，运用严谨的法律的逻辑和话语体系，反映我国在网络空间国际规则核心问题上的立场和诉求，以实质性地引领国际议题的讨论方向，促进网络空间国际规则的形成。⁵⁷

其次，我国应加强官方国际主流媒体的建设，发挥政府力量将网络媒体资源进行整合，打造出综合性多语种的国际新闻网站。与此同时，政府通过政策、资金和技术支持推动民间媒体国际化发展，发挥民间媒体灵活、多样化的传播优势，弥补中国在网络空间治理议题在国际传播方面的不足。⁵⁸此外，我国应借助举办各种研讨会和互联网大会的有利时机，同各国代表和知名权威专家积极交流，旨在塑造有利于我国的舆论环境，从而在国际网络空间治理中获得更大的影响力。

最后，我国还可以充分利用联合国、上合组织、二十国集团（G20）和世界互联网大会等多个具有国际影响力的平台，作为宣示中国网络空间治理理念和呼声的议程重要进入渠道。我国需通过积极设置代表我国治理策略的网络空间国际议程，同时，广泛进行利益动员，寻求更多国家的支持，以形成议题的集体认同，

⁵⁶ 黄志雄：《网络空间国际规则制定的新趋向——基于〈塔林手册 2.0 版〉的考察》载《厦门大学学报（哲学社会科学版）》，2018 年第 1 期，第 8-9 页。

⁵⁷ 黄志雄：《积极参与网络空间国际规则博弈加快建设网络空间国际法强国》，载《中国信息安全》2021 年第 1 期，第 135 页。

⁵⁸ 许开轶：《论中国网络空间国际话语权的建构》，载《南京师大学报（哲学社会科学版）》2022 年第 4 期，第 124-126 页。

促使更多的中国理念融入国际网络空间治理实践。⁵⁹

五、结语

网络空间国际话语权不仅代表着一国在国际网络空间治理中的话语影响力，更是一国对网络空间治理国际秩序引导力和支配力的综合体现。然而，我国在网络空间治理中的国际话语权提升与综合实力增长处于不对称的状态，主要原因是西方发达国家凭借历史经验和技術优势在网络空间治理中占据优势地位，尤其是美国等网络强国凭借其网络霸权在全球范围内横行。并且，目前网络空间治理国际规则以不具法律约束力的软法居多，导致规则适用范围具有局限性。同时，我国网络空间治理议程设置能力也有待增强。因此，我国必须坚定不移地维护网络主权原则，以此作为抵御网络霸权、推动构建网络空间命运共同体的坚实基础。此外，我国应高度重视联合国及多边机制的造法功能，积极贡献中国智慧与中国方案，从而推动网络空间治理多边公约的制定与出台。与此同时，中国应强化网络空间治理议程设置能力，加强议题引领和议题传播能力的建设。通过选择恰当且合适的平台开展广泛的利益动员，推动我国议程上升为国际议程。这一系列的举措不仅有助于提升网络空间治理下中国国际话语权，更是将我国价值理念融入新国际规则制定和国际秩序构建中，从而获取更多国家利益的有效策略。

未来，如何推进网络空间治理无疑将是一个持续性的重要课题。各国将继续围绕网络空间治理规则和秩序的构建不断博弈。鉴于此，中国需牢牢掌握网络空间治理国际规则和秩序构建的主导权，我国才有望在网络空间国际治理体系的变革中发挥更加积极的作用，引领网络空间治理的国际秩序朝着更加公正、合理的方向发展。这不仅符合全人类的共同利益，也将确保网络空间的发展成果能够由全人类共建、共商、共享，真正实现网络空间的共治、共享、共赢。

⁵⁹ 郭志奔：《中国智库参与全球互联网治理的国际议程设置路径研究》，载《智库理论与实践》2022年第7卷第6期，第43页。

《联合国打击网络犯罪公约》电子数据分类的国内法转化研究

——基于《刑事诉讼法》第四次修改之背景

杨宇轩 中南民族大学

摘要：《联合国打击网络犯罪公约》（简称《公约》）作为全球首部网络犯罪治理综合性国际公约，将电子数据作“流量数据”“内容数据”“订阅用户信息”三分并设置了差异化的取证措施，体现了国际共识。相较于《布达佩斯公约》经验，该公约强化“网络主权”原则，回应发展中国家关切，成为平衡数据主权与跨境协作的关键规范。作为《公约》主要推动者，中国需通过国内法转化履行国际义务，尤其需借《刑事诉讼法》第四次修改契机，弥合现行规范与《公约》的差距。当前，《刑事诉讼法》对电子数据的分类尚显粗疏，虽通过司法解释对电子数据采取“四分法”或“七分法”，但囿于立法层级低、规则碎片化、执行不统一等问题，难以有效对接《公约》要求。应借修法之机引入《公约》“三分法”，并通过明确电子数据分类对应关系、优化侦查措施适用、完善非法证据排除规则等路径整合并优化规则，提升取证法治化水平，彰显中国对“网络主权”的坚守与国际合作的诚意。

关键词：《联合国打击网络犯罪公约》；电子数据；《刑事诉讼法》第四次修改；网络主权

2024年12月24日，在第七十九届联合国大会上，全体联合国会员国以协商一致的方式正式通过了具有历史性意义的《联合国打击网络犯罪公约》（文件全称为《联合国打击网络犯罪以及为打击利用信息通信技术系统实施的某些犯罪并共享严重犯罪电子证据而加强国际合作公约》，后称《公约》）。这是继《联合国反腐败公约》之后时隔二十年面世的又一部新型国际立法，也是迄今为止在网络和数字领域第一部由联合国主导制定的全球性国际公约。^[1]

从《公约》的谈判历程及最终案文的通过情况来看，目前对于《公约》在形式上不存在明确拒绝签署的国家反对意见；同时，与刑事实体法相比，各个缔约国对打击网络犯罪程序法方面的意见更趋统一，《公约》的篇章结构及主要条款，尤其是电子数据定义即类型以及取证方面的条款，与2001年欧洲理事会制定的《布达佩斯网络犯罪公约》（Budapest Convention on Cybercrime，后文简称《布约》）及2021年11月《布达佩斯公约第二附加议定书》（后文简称《布约第二议定书》）保持了较大相似性，可见在打击网络犯罪中，电子数据规则的国际共识更为成熟，甚至可以说，《公约》的出台是迈向“全球电子证据法”的重要一步。本

^[1] 参见叶伟、姜博谦：《〈联合国打击网络犯罪公约〉开启网络空间国际规则新章》，载《中国信息安全》2024年第8期。

文围绕刑事诉讼电子数据分类理论和意义展开,梳理中国刑事诉讼法律规范和《公约》有关规定,尝试找寻两者之间存在的潜在冲突,以期能为中国刑事诉讼领域必要的制度和规则调整,尤其是为《刑事诉讼法》的进一步修改提供一些参考,促进《公约》批准后更好地转化与执行。

一、《公约》对于电子数据分类的规定及其评价

在《公约》以前,《布约》曾是设计跨境电子数据获取的最为广泛接受区域性公约,截止目前有 78 个缔约国和 17 个观察员国,被誉为是跨境电子数据“获取指南、清单或示范法”。^[2]但随着网络犯罪形式的变化,《布约》已经表现出来了一定的落后性,如过分注重“人权保护”,致使其在防治网络犯罪的格局受到限缩;同时,以法国为代表《布约》缔约国所倡导的“网络自由”,也与其他国家主张的“网络主权”相冲突,致使《布约》民主性不足。^[3]

许多发展中国家也意识到《布约》体系下欧美国家立法存在一系列不公正性问题,^[4]也明白,在各国刑事司法合作、电子数据规则碎片化背景下,亟待生成的是惩治网络犯罪的全球性的国际规则。因此,中国、俄罗斯等国没有选择加入《布约》,而是致力以“网络主权”为出发点,推进《公约》的落成。

(一)《公约》对于电子数据分类的规定

《公约》电子数据分级分类的规定,分散于《公约》总则第 2 条“术语使用”及第 4 章、第 5 章当中。^[5]《公约》将电子数据细分为“流量数据(traffic data)”“内容数据(content data)”“订阅用户信息(subscriber information)”三类,并按照此分类标准,于第四章规定了针对电子数据各缔约国所应采取的措施及程序,为不同类型电子数据作出了共通但又有所区分的安排,又于第五章中规定了跨境取证国际合作的有关事宜。

可以认为,《公约》首先于总则对电子数据的作出了“三分法”的分类,一定程度上是对《布约》的继承与发展;将电子取证条款区分为国内措施与国际合作两个部分,其中第四章所规定的,正是电子取证国内措施、手段,旨在强化缔约国的电子取证能力,较少涉及国家间的协调,对《布约》有所继承;第五章由于涉及国际合作具体事项,基于“尊重主权”,给予了各缔约方自主采取措施的空间,一定程度上较《布约》更为宽松。

具体而言,《公约》针对以上三类电子数据,在第四章设置了快速保全和披露、提交令、搜查和扣押、实时收集和拦截等六项取证措施。其中公约第 25 条

^[2] See Parties/Observers to the Budapest Convention and Observer Organisations to the T-CY - Cybercrime, <https://www.coe.int/en/web/cybercrime/parties-observers>, 最后访问时间: 2025 年 3 月 29 日。

^[3] 包括中国在内的东亚很多国家和地区对其持保留意见。

^[4] 欧美国家利用国内法,“对外”扩展执法管辖权,“对内”却采取严格的管控措施限制本国数据出境,而且组建基于地缘政治立场的跨境数据取证“小圈子”将我国和广大发展中国家排除在规则体系之外,加剧了全球跨境数据取证格局的分裂,参见魏光禧,刘树想:《跨境数据取证中的公私合作——以〈联合国打击网络犯罪公约〉为视角》,载《河北法学》2024 年第 8 期,第 80 页。

^[5] 需要说明的是,与电子数据的有关术语是在确定公约的主要实质性条款之后才确定的。

“快速保全（expedited preservation）”规则，规定各缔约国在涉案电子数据极易丢失或被修改的情况下，可以责令有关主体在必要时限内，保全并维护电子数据完整性，并对有关程序保守秘密。《公约》第 28 条“搜查和扣押（search and seizure）”规则与中国《刑事诉讼法》表述较为一致，同时对三种电子数据都适用。

依据《公约》的定义，“流量数据”是指显示通信的起点、终点、路径、时间、日期、数据量大小、持续时间或基础服务类型的电子数据，具有易灭失、易篡改的特点。^[6]因此，针对“流量数据”，《公约》第 4 章第 26 条特别规定了“部分披露（partial disclosure）”措施，其旨在确保缔约国主管机关或该机关所指定人员识别有关服务提供者（不论一个还是多个）和所涉通信及其传输路径。第 29 条则规定了针对“流量数据”的“实时收集（real-time collection）”规则。^[7]

依据《公约》的定义，“内容数据”指与使用信息通信技术系统传输的数据的实质内容有关，且不属于订阅用户信息或流量数据的任何电子数据，主要包括图像、文本消息、语音消息、录音和录像等。“内容数据”一般涉及沟通交流中传递的信息实质内容，通常具有较高的侵入性。针对“内容数据”，《公约》第 4 章第 30 条对“内容数据”规定了“拦截（interception）”措施，类似于对载有内容信息的电子数据进行实时侦听的措施。

依据《公约》对“订阅用户信息”的界定，它是指服务提供者持有的与其服务订阅用户有关的、除流量数据或内容数据以外的信息。其具体又可以进一步分为三类：（一）所使用的通信服务种类、与之相关的技术条款以及服务期限；（二）根据服务协议或安排提供的订阅用户身份、邮政地址或地理地址、电话或其他接入号码、账单信息或付款信息；（三）根据服务协议或安排提供的关于通信设备安装地点的任何其他信息。需要特别指出的是，《公约》对“订阅用户信息”的定义包括“IP 地址”，与《布约》对此概念的定义有所差异。

针对“订阅用户信息数据”，《公约》第 4 章第 27 条设置了“提交令（production order）”条款。“提交令”制度最早来源于《布约》第 18 条；而 2017 年俄罗斯向联合国大会提交的《公约（草案）》中亦设置了此条款，只是针对情形有所区别，但共同点在于认为部分电子数据有跨境调取的必要性，区别在于《公约（草案）》将数据类型限制在“订阅用户信息”。即一定程度上反映出各缔约国跨境调取“订

^[6] “流量数据”系指与使用信息通信技术系统进行的通信有关的任何电子数据，由构成通信链一部分的信息通信技术系统生成，显示通信的起点、终点、路径、时间、日期、数据量大小、持续时间或基础服务类型；

^[7] 其内容为：“各缔约国均应采取必要的立法措施和其他措施，授权其主管机关采取下列行动：（1）在本国领土内应用技术手段实时收集或记录；（2）强令服务提供者在其现有的技术能力范围内：在本国领土内应用技术手段实时收集或记录；或者配合并协助主管机关实时收集或记录；

与其领土内使用信息通信技术系统传输的特定通信有关的流量数据，因其本国法律制度原则而无法采取本条第一款第（一）项所述措施的，可采取其他必要的立法措施和其他措施，确保在其领土内应用技术手段实时收集或记录与在其领土内传输的特定通信有关的流量数据。

各缔约国均应采取必要的立法措施和其他措施，责成服务提供者对本条规定的任何权力的行使情况以及与之相关的任何信息保守秘密。”

阅用户信息数据”的认同与支持，并反映了公私合作模式的运用。

（二）透过《公约》规定，评价中国对电子数据分类的立场

经历多轮谈判并最终妥善落成的《公约》文本中，第5条就确立了维护主权的基本原则；“尊重缔约国主权”之措辞，也在规定如何开展合作以及具体合作事项的条款中被多次重申。

就维护主权这一原则而言，过去中国坚守“数据存储地模式”，顾虑即在于网络空间主权（或者说数据主权）安全。在此模式下，我国采取的是通过“一案一议”或外交渠道开展刑事司法协助活动或警务合作的传统模式。结合第五章“国际合作”相关条文来看，《公约》充分尊重和保障各缔约国的主权利益，肯认中国“主权原则在网络空间中的延伸”的立场，也使得《公约》成为首部将我国积极推动的“网络主权”理念转化为具有法律约束力的明文规定的国际协定。^[8]

而在贯彻这一原则之下，值得注意的就是基于《公约》中电子数据的分级分类而严格限定适用范围的“提交令”条款的通过。设置与“提交令”类似条款的国家在《公约》缔结前既已达到相当规模，其更大的价值在于追求跨国打击犯罪的效率。相反的，传统刑事司法合作中“一案一议”模式的取证效率极低，^[9]与电子数据不稳定、易灭失、需快速保存和提取的特征与要求相抵牾，而面对“暗网”、云计算应用等场景时，甚至可能无法确定电子数据的物理存储位置而导致取证困难。固守“数据存储地模式”看似维护了境内数据安全，但也很大程度堵塞了获取他国电子数据的路径，不利于与他国开展合作，有效打击日趋严峻的跨国网络犯罪；长远来看，也可能与我国所缔结或所欲参与的区域贸易协定中数字贸易以及（尤其是数据跨境流动）的规定相冲突，有碍于我国数字经济的发展。^[10]因此，我国对“提交令”条款的接纳态度，体现了我国在“数据存储本地化模式”和适度开放数据流动之间谋求最优平衡点的积极尝试，也体现了我国对促进各方合力打击网络犯罪的报以较高诚意。

与此同时，《公约》第6条规定的是尊重人权原则。就尊重人权这一原则而言，由于实时收集流量数据和拦截内容数据两种取证措施对隐私和个人数据安全带来的风险更高，因此《公约》基于人权保护的相称性原则，对有关取证措施作出了针对“犯罪”与“严重犯罪”区分适用的规定。^[11]我国虽然在公约谈判中持续表示反对对人权保护的“片面注重”，^[12]但在《公约》中仍肯认了《布约》

^[8] 参见吴沈括：《联合国打击网络犯罪公约》的治理体系与中国因应，载《中国法律评论》2025年第1期，第220页。

^[9] 有数据显示，传统双边司法协助进行的证据调取平均时效9个月以上，证据效力尽失，对于犯罪的打击效果会非常微弱，载刘品新，谢鹏程主编《电子证据与网络犯罪治理20讲》，清华大学出版社2024年版，第121页。

^[10] 采取“数据本地存储模式”的保守态度无可厚非，但笔者认为，要想参与并引领国际数据规则的制定，分享全球数字经济的红利，必然要着眼于面向开放的自由流通，参见陆栋：《首都大数据跨境交易之刑事风险防范与化解——以刑事司法协助为视角》，载《中国刑事司法》，2024年第2期，第171页。

^[11] 根据《公约》规定，“严重犯罪”是指可能判处4年以上有期徒刑的犯罪。

^[12] 参见王渊洁：《联合国打击网络犯罪公约》特委会会议综述，载《中国国际法年刊（2022）》，法律出版

缔约国有关主张的积极意义，在电子数据传递的类型上，明确对“内容数据”进行了限制和保留，并设计了对权利保护的一致性与相似性审查；在电子数据跨境的合作上限定为涉及预防、侦查、调查或起诉严重犯罪的需要，而不能用于民事、行政或者商业查询，这实则为“人权保障”价值（后文也称“权利保障”价值）进一步在刑事诉讼中贯彻作了铺垫。

总的来说，中国无论是基于“保护数据主权”立场，还是基于契合刑事诉讼“人权保障”的精神基石，都需要于刑事诉讼领域将电子数据分类予以明确、落实，来满足《公约》的需要。

二、中国刑事诉讼规范对电子数据分类的规定及其评价

（一）中国《刑事诉讼法》对电子数据分类的规定

2012年，中国刑事诉讼法首次将“电子数据”作为一类独立的法定证据种类。^[13]其进步意义在于赋予了电子数据的法定证据身份，一定程度上避免电子数据在证据转化方面耗费过多人力与时间，继而提高司法诉讼效率。^[14]

但是，2012年刑事诉讼法并没有根据电子数据特点进一步规范界定其概念内涵、证据形态、取证行为方式以及与权利保障相适应的侦查取证措施等内容。尤其在取证方式上，没有将电子数据作为勘验、检查、搜查等侦查行为的实施对象予以程序性规制，仅在“查封、扣押”章节中有所提到相关程序性规定。换言之，2012年的《刑事诉讼法》规定过于原则化，也并未对如何保护电子数据进行规定。而2018年《刑事诉讼法》沿袭了2012年《刑事诉讼法》有关电子数据法定证据种类的规定，但也未在其他条款涉及电子数据取证、审查、认定等问题。

总体来看，《刑事诉讼法》对于电子数据的规定，是“宜粗不宜细”立法理念的秉持。

（二）中国刑事诉讼司法解释对电子数据分类的规定

在2012年《刑事诉讼法》出台以后，中国最高人民检察院、最高人民法院、公安部陆续地针对性地出台了相应规范，这些规范“成果”依据取证措施的强制性程度，以“是否涉及公民隐私信息”为标准对其展开了区分，^[15]来对刑事诉讼活动

社2023年版，第478页。

^[13] 此前，电子数据的概念解释主要依靠最高人民检察院、最高人民法院和公安部在办理涉及相关证据的案件时出台的规定和规范，如2010年两高三部颁布的《关于死刑案件审查判断证据若干问题的规定》中，在第29条就对电子数据作出“电子邮件、电子数据交换、网上聊天记录、网络博客、手机短信、电子签名、域名等”的列举。由于电子技术发展较快，这些规定中的列举已经不能明确地覆盖此后电子数据实际所涉范围，故在此不作详细讨论。

^[14] 此前由于欠缺法律的明确规定，加上各部门存在认识分歧，往往需要将QQ聊天记录等电子数据转换成勘验笔录等其他证据形式才能在刑事诉讼中予以使用。

^[15] 如2012年11月《人民检察院刑事诉讼规则（试行）》，于第9章第6节明确规定了电子数据能够作为调取、查封以及扣押的对象，并规定了执行及审批的主体；2012年12月，《关于适用〈中华人民共和国刑事诉讼法〉的解释》第93条，对于通过勘验、检查、搜查等手段收集到的电子数据（包括电子邮件、电子数据交换、网上聊天记录等），就如何进行合法性与规范性方面的审查进行了阐释，包括审查是否随原始存储介质移送、所收集到的电子数据与案件事实有无关联等。同年12月《公安机关办理刑事案件程序规定》，于第59条和第227条分别针对电子数据的调取与扣押进行了规定；并对可能过分干预公民个人隐私的技术侦查作出“仅适用于严重刑事犯罪”的限定。

中所涉电子数据展开分级保护，但这种分级较粗疏，对电子数据仅限简单列举，不涉及对电子数据类型化区分。质言之，这些非专涉电子数据的刑事规范，既无电子数据分类之名，亦对电子数据分级不足。

直到 2016 年，《刑事电子数据规定》的出台，才对电子数据作了具有针对性、统合性的规定。首先，其将电子数据概括地解释为“案件发生过程中形成的，以数字化形式存储、处理、传输的，能够证明案件事实的数据”；其次，《刑事电子数据规定》第 1 条第 2 款对电子数据采用了“示例+特征”的列举方式，并将电子数据的类型作了“网络平台发布的信息、网络应用服务的通信信息、用户身份及行为信息、电子文件”的“四分法”划分。

有学者认为，《刑事电子数据规定》这种分类方式主要以电子数据所承载信息的公开程度为标准，在网络空间中的网页、微博等电子数据多属于“公开发布”，朋友圈、贴吧等属于“半公开性”的电子数据，而电子邮件、短信微信等通讯则属于“高私密”电子数据。同时以“秘密性”为标准，以所涉及的国家秘密、商业秘密以及个人隐私（后文简称“三密”）来对电子数据进行类型化处理。可见，分类列举的依据是电子数据的存在形式与隐私期待，已经初步体现了电子数据分类的理念，这有利于办案机关对各种信息材料分门别类、有所取舍，提高办案效率；并且考虑到了今后电子数据发展的前景，用“包括但不限于”的表达方式为电子数据未来可能展现出新的形态留出了一定的解释空间；^[16]其后出台的 2023 年通过的《海警机构办理刑事案件程序规定》以及 2024 年通过的《国家安全机关办理刑事案件程序规定》也延续了对电子数据的“四分法”。^[17]

对于不同类型的电子数据，在取证程序要求上也存在差异化的设置，如对于通信信息的收集提取可能涉及技术侦查措施的，要经过严格的批准手续；^[18]与此同时，在电子数据取证措施类别中补充规定了“冻结”措施的条件与方法；^[19]而在电子数据跨境取证措施方面，16 年《刑事电子数据规定》由原来的远程勘验，^[20]扩展至网络在线提取、网络远程勘验和技术侦查三种。

^[16] 参见陈易臻：《侦查中的数据分类分级保护制度研究》，西南政法大学博士学位论文，第 44～45 页。

^[17] 《海警机构办理刑事案件程序规定》第 55 条第 2 款在“四分法”基础之上特别增加了“电子海图、船舶航行轨迹等信息”这一类别，但总体上依然是对“四分法”的承继。

^[18] 参见喻海松：《网络犯罪二十讲（第二版）》，法律出版社 2022 年版，第 217 页。

^[19] 《电子证据规定》第 11 条规定了冻结电子数据的条件：“具有下列情形之一的，经县级以上公安机关负责人或者检察长批准可以对电子数据进行冻结：（一）数据量大，无法或者不便提取的；（二）提取时长，可能造成电子数据被篡改或者灭失的；（三）通过网络应用可以更为直观地展示电子数据的；（四）其他需要冻结的情形”；第 12 条第 1 款规定了冻结电子数据以及解除冻结的手续：“应当制作协助冻结通知书，注明冻结电子数据的网络应用账号等信息，送交电子数据持有人、网络服务提供者或者有关部门协助办理。解除冻结的，应当在 3 日内制作协助解除冻结通知书，送交电子数据持有人、网络服务提供者或者有关部门协助办理”；第 12 条第 2 款规定了冻结电子数据的方法“冻结电子数据，应当采取以下一种或者几种方法：（一）计算电子数据的完整性校验值；（二）锁定网络应用账号；（三）其他防止增加、删除、修改电子数据的措施。”

^[20] 远程勘验，于 2005 年公安部出台的《计算机犯罪现场勘验与电子证据检查规则》就已规定：“是指通过网络对远程目标系统实施勘验，以提取、固定远程目标系统的状态和存留的电子数据。”

2019年1月，公安部颁布《公安机关办理刑事案件电子数据取证规则》（后文简称《电子数据取证规则》），首先对“电子数据取证”做出了开放列举式规定；^[21]在吸收《刑事电子数据规定》保护“三密”的同时，还设置了对“警务工作秘密”的保护，并规定了行政执法过程中的电子数据转化问题。其次，规定了扣押、封存原始存储介质、现场提取、网络在线提取、冻结、调取电子数据等收集提取电子数据的措施。

2021年最高人民检察院发布的《人民检察院办理网络犯罪案件规定》第27条对电子数据进行分类可谓最为详尽，^[22]它在前述规范的基础上，将电子数据“四分法”中“用户注册信息、身份认证信息、电子交易记录、通信记录、登录日志等信息”进一步细分为“用户身份信息”“用户行为信息”“行为工具信息”“系统运行信息”；同时对原“四分法”下第四类“电子文件”特别添加了“及其时间、大小等文件附属信息”。可以认为，上述文件主要是对实务界经验教训的总结提炼。

（三）对既有规范中电子数据分类规定的分析与评价

《公约》言及“各国可以采取多种立法措施”来履行其所规定的义务，换言之，中国对《公约》批准生效时，无论是立足于《刑事诉讼法》，还是立足于实务部门出台司法解释等规范性文件，都需要实现《公约》关于刑事诉讼电子数据分类的要求。需要思考的是，既有规范是否有足够的灵活性来对接《公约》术语体系？

正如一些学者指出的那样，不论是16年《刑事电子数据规定》、19年《电子数据取证规则》，还是23年《海警机构办理刑事案件程序规定》以及24年《国家安全机关办理刑事案件程序规定》等规范中的分类标准，主要围绕证据形式展开，更毋庸提及18年《刑事诉讼法》；^[23]前述规范大多由侦查机关参与制发，其结果是刑事侦查中的技术性经验积累，顺利地转化为规范性文件，获取执法依据，目的在于便于取证机关准确把握刑事司法实践中电子数据的范围，克服涉电子数据刑案的惯常复杂性和长周期性，并实现取证的高效性。但反过来看，由国家专门机关通过制定司法解释、规范性文件来进行电子数据侦查取证中的自我授权，这些机关就同时承担了“规则创设者”和“规则执行者”的双重角色，有悖正当程序的基本精神，远没有达到实现电子数据侦查取证措施法治的目标要求。

就国内法治层面来看，现行规范也“力有不逮”。电子数据冻结的具体规范欠

^[21] 《电子数据取证规则》第3条：“电子数据取证包括但不限于：（一）收集、提取电子数据；（二）电子数据检查和侦查实验；（三）电子数据检验与鉴定。”

^[22] 《人民检察院办理网络犯罪案件规定》第27条规定“（一）网页、社交平台、论坛等网络平台发布的信息；（二）手机短信、电子邮件、即时通信、通讯群组等网络通讯信息；（三）用户注册信息、身份认证信息、数字签名、生物识别信息等用户身份信息；（四）电子交易记录、通信记录、浏览记录、操作记录、程序安装、运行、删除记录等用户行为信息；（五）恶意程序、工具软件、网站源代码、运行脚本等行为工具信息；（六）系统日志、应用程序日志、安全日志、数据库日志等系统运行信息；（七）文档、图片、音频、视频、数字证书、数据库文件等电子文件及其创建时间、访问时间、修改时间、大小等文件附属信息。”

^[23] 参见裴炜、黄志雄等：《〈联合国打击网络犯罪公约〉：背景、内容与展望》，载《数字法学》2024年第5期，第20～21页。

缺，对相应证据予以排除或者对权利受侵害主体予以救济的规定尚付阙如等等；有关规范的执行不力，体现在电子数据封存扣押中唯一性特征缺失、记录不全面和封存不完全，电子数据勘验检查规范要求执行性差，侦查机关的电子数据提交与展示方式老旧，^[24]侦、检机关在取证、审证过程中过分重视内容数据而忽视附属信息等。^[25]究其原因，电子数据案的具体办案人员会依赖自己既往的办案经验和行为惯性，体现出“上有政策、下有对策”。此外，中国全国范围内规范执行情况也不具备统一性，一些地区面对跨地域涉电子数据案件，侦办意愿严重不足；一些地区则出现利用相关权利保护条款等规范漏洞，与一些配侦公司开展“远洋捕捞”式取证，出现异地趋利性执法，使得这些机关“在收集电子数据信息方面几乎处于‘放飞自我’的状态”。^[26]

从涉外法治层面来看，16年《刑事电子数据规定》有关电子数据侦查取证措施的拓展，就曾遭到外国的质疑与批评，认为中国的规定有违他国网络主权之嫌。^[27]在《公约》转化执行之前，如果不对跨境电子数据取证的恣意作进一步规制，不仅会给传统《布约》阵营进一步批判《公约》留下空间，更会带来域外对中国积极构建的“网络空间主权”主张的质疑。

三、《刑事诉讼法》修改背景下引入《公约》电子数据分类规定的必要性及可行性

基于前文述原因，笔者认为不能仅仅靠司法解释不断扩充相应措施，而有必要通过修改《刑事诉讼法》来落实《公约》的规定。笔者尝试基于以下三个方面，对于刑事诉讼法继受《公约》“三分法”的必要性加以论证。

（一）基于《刑事诉讼法》第四次修改的机遇恰适性

首先，2023年9月，中国十四届全国人大常委会也将《刑事诉讼法》修改列为一类立法规划项目，并且属于全国人大常委会立法规划三类立法项目中的第一类，即属于条件比较成熟、任期内拟提请修改的立法项目，旨在通过这一刑事诉讼领域基本法，回应刑事诉讼跨国化、数字化的需求与挑战。本次修法机会难得，而电子数据规则的完善无疑是对跨国化、数字化的需求需求的最直接、有效的回应，则其首要立足点即在于电子数据的分级分类。

其次，从中国《刑事诉讼法》证据制度的整体上看，本身就存在严重的规范

^[24] 参见尹鹤晓：《电子数据取证：法律规制与侦查实务》，中国政法大学出版社2024年版，第106～118页。

^[25] 虽然2021年的《人民检察院办理网络犯罪案件规定》第27条对电子数据作出了最为详尽的分类，但不管是公安机关在调取证据的过程当中，还是检察机关在审查证据的过程当中，都更重视电子邮件的内容、聊天记录等内容数据，对于文件、文档本身的属性、制作人、发件人、传递的路径等附属信息往往会忽视。参见刘坤：《海量电子数据审查常见问题及应对措施》，转引自刘品新、谢鹏程编《电子证据与网络犯罪治理20讲》，清华大学出版社2025年版，第79页。

^[26] 参见陈永生：《电子数据收集应当被纳入本次〈刑事诉讼法〉修改》，载《法治研究》2024年第5期，第107页。

^[27] 参见刘品新：《信息网络犯罪中证据的取证规则》，第二届“证据法学论坛”，https://www.bilibili.com/video/BV1KM411V7Sw/?spm_id_from=333.1387.homepage.video_card.click&vd_source=bf464031c270fb60c3bffe2030e44af9，最后访问时间：2025年3月29日。

供给不足问题。从证据规范数量上看,涉及具体证据认定工作的条文仅 16 条,具体到电子数据更是寥寥。^[28]无论是基于一些学者所提,刑事诉讼法四修应基于“大改”还是“中改”,“加强证据制度建设”都是应有之举;^[29]更有学者着重强调电子证据之于证据制度完善的紧迫性与优先性。^[30]就此,本文认为,借助《公约》通过并预期将得到批准的“东风”,倒逼《刑事诉讼法》纳入新的电子数据分类,不仅可以在不改动“电子数据”这一立法用语的同时,增强其概念的周延性、准确性与统合力,也可以对既有司法解释、部门规范的拨冗、整合,有效克服我国证据规范来源的碎片化、片面化,实现“再法典化”。^[31]从立法内容上,落实《公约》更多是对确有实效的司法解释、成熟的实务经验在法律层面的提炼式呈现,在对《刑事诉讼法》修订的工作没有造成过大压力的同时,还能增强有关取证、审证规范要求的贯彻时效性和全国统一性。

再次,由于《刑事诉讼法》缺乏对电子数据的进行提纲挈领式的分类,没有统一有序的设置取证门槛和审查程序,即使是基于《公约》“打击网络犯罪”之意旨、口号,自上而下部署任务,也难以预见实务中侦查机关内部对于电子数据收集形成统一的认识。为便利本部门完成刑事案件的办理任务,执业惯性会驱使侦查部门采取较为恣意地启动电子数据收集措施,导致权力的滥用与膨胀。举例来说,如果要求侦查机关积极打击网络犯罪又由于缺乏法律明确的监督和制约,其会寻求将大量数据提前应用于“预测警务”;如果加之各部门间共享数据的监管体系和制约机制存在疏漏,可以预见,大量的“黑数据”将难以剔除,这会提高了公民个人信息的泄露风险,反而会影响公民对于犯罪预测的接受程度,影响刑事治理的实效。^[32]

(二) 基于国内法与国际法应有的协调一致性

一般认为,中国对于国际法的适用采取转化适用,在民商事等特定部门法领域,才会特别规定“国际法的优先适用”。在刑事诉讼领域,不乏可供参考的先例。如 2005 年 10 月 27 日中国批准《联合国反腐败公约》(后文简称《反腐败公约》)之后,在立法方面就是以优先修订《刑事诉讼法》的方式落实履约义务的,随后

^[28] 刑事诉讼法中涉及电子数据的规定,仅有第 50 条“证据种类”以及第 54 条有关“行政、刑事证据衔接”的规定,参见陈永生:《电子数据收集应当被纳入本次〈刑事诉讼法〉修改》,载《法治研究》2024 年第 5 期,第 107 页。

^[29] 主张“大改”的代表学者如陈卫东:《〈刑事诉讼法〉第四次修改前瞻》,载《政法论坛》2024 年第 1 期;主张“中改”的代表学者如龙宗智:《〈刑事诉讼法〉修改若干重点难点问题研究》,载《中国法律评论》2025 年第 1 期,第 20~22 页。但“大改”、“中改”者无一例外都重视对电子数据问题的研判。

^[30] 如樊崇义教授、姚莉教授认为,完善电子证据的相关规范,是本次《刑事诉讼法》修改的当务之急,参见苏菡乔:《刑事证据立法的中国图景——〈刑事诉讼法〉再修改与证据制度建设研讨会综述》,载《证据科学》2024 年第 5 期,第 576~586 页。

^[31] 参见聂友伦:《刑事诉讼法的解法典化与再法典化》,载《中外法学》2024 年第 5 期,第 1142~1148 页。

^[32] 参见胡铭、严敏姬:《大数据视野下犯罪预测的机遇、风险与规制——以英美“预测警务”为例》,转引自《中国犯罪治理蓝皮书犯罪态势与研究报告(2022)》第 140~142 页。

的步骤才是制定专门立法、出台有关的司法解释。^[33]因此,应当在中国接受《联合国打击网络犯罪公约》履约审议之前,优先修改刑事诉讼法,才符合以往的刑事诉讼立法规律。

与《反腐败公约》不同的是,在联合国平台谈判磋商、草案拟定和最终通过《公约》的整个过程,中国可谓首倡者,与俄罗斯等国是为《公约》最积极的推动方。经过调研访谈,笔者认为,《公约》中文版约文的最终落成,尤其是其中对各类电子数据术语体系的表述及界定,经历了中方各谈判部门的深思熟虑,会对于我国具有不同以往的羁束意义。如果在《刑事诉讼法》中未予体现,未来中国对《公约》的贯彻执行,以及《公约》的必要性,难免受其他国家质疑与非难。如果不采取一致性表述,能够构成“同一解释规则”的合理例外是值得警惕的。^[34]

其次,在刑事诉讼法中有效的落实《公约》中文版的语词,便于理论界与实务界在问题讨论时话语指涉对象的准确、统一。^[35]刑事诉讼法学界在讨论数字时代证据制度有关问题时,仍然混杂使用“数字证据”、“电子证据”等概念。^[36]需要指出的是,《公约》中文版在总则第2条第2款,已经明确使用了“电子数据”(《公约》英文版中称“electronic data”)这一术语并进行了界定,可见,代表立法者进行《公约》谈判的有关部门同志对于证据有关术语整体持严格态度,不仅仅考虑到了国内防范冤假错案的需要,还考虑到了国内刑法及其他部门法与《公约》衔接的问题。而学术交流的高效只是其附随效果,用语一致性的重点在于能够减少证明网络犯罪事实的过程中的障碍,促进跨国合作,应对当前网络犯罪“语言不通而犯意易通”的问题,有效打击跨国网络犯罪;此外,术语的沿用,也有利于处理在我国境内对外籍犯罪嫌疑人、被告人适用我国刑事诉讼法时的“汉译英”等翻译问题,能够更畅通地实现外国人的诉讼权利,避免有关争议。

再次,鉴于《公约》最终版本吸收了《布约》“7天24小时全天候网络的快速响应机制”,并要求各国建立专门的联络点以提供即时协助,表明了中国对此的认可。为回应《公约》这一要旨,中国需要确定在处理电子数据跨境请求时有关主管机关以及相关职责。而优先通过《刑事诉讼法》,解决公检法三机关在处

[33] 我国2012年修订《刑事诉讼法》增加犯罪嫌疑人、被告人逃匿、死亡案件违法所得的没收程序,并于2013年7月起接受《公约》第一周期履约审议;2017年出台司法解释,2018年颁布《监察法》,并积极推进制定《国际刑事司法协助法》。参见梅芬:《联合国反腐败公约》走过15年——合作应对腐败中国步履坚实,载《中国纪检监察》2018年第21期,第60~61页。

[34] “同一概念在统一法律部门内部,甚至在不同法律部门之间,都应当尽可能保持含义的一致”,参见王利明:《法律解释学导论——以民法为视角(第三版)》,法律出版社2021年版,第271页。

[35] 如,过往一些学者将“订阅用户信息”称其为“用户数据”(subscriber data)、“注册人数据”;一些学者对“流量数据”也称“电信资料”、“往来数据”,参见汪振林:《电子证据学》,中国政法大学出版社2016版,第71~73页。

[36] 胡铭、杨继文、张中等学者提出用“数字证据”的概念来包纳电子数据、人工智能数据在内的其他新证据样态;何家弘、汪振林、刘品新等学者则仍用“电子证据”表达与之相关的概念。

理不同类型电子数据跨区调取、审查权限的问题，完善各执法司法机关之间关于不同类型电子数据审查、处理的职权与分工，才能进一步考虑涉外情况，完善《国际刑事司法协助法》，搭建三机关应对跨境取证请求的协调机制。这也一定程度契合一些学者“在刑事诉讼法第四次修改中增设涉外刑事诉讼专章”的建议。^[37]

（三）基于纵向的法秩序统一性和横向的立法规律性

早前就有中国学者讨论，对于技术性较强的电子数据，是以司法解释、纪要性文件还是单行法规来规范电子数据。如前文所述，现阶段中国有关电子数据的主体规范主要散落在两高一部有关文件和司法解释之中，法律效力位阶不同，不可避免地呈现出散乱、非系统化状态。电子数据的“分散式立法”，势必影响实务部门执法的规范性和对外刑事司法合作的有序性。因此，笔者认为，对于电子数据问题，应对《刑事诉讼法》“宜粗不宜细”的立法惯性适时作出调整。我国不应再通过司法解释等方式来扩充我国在刑事司法领域的电子数据分类方式，而应当基于“法律保留”，有待《刑事诉讼法》在数据分级分类方面进行统摄性规范和细化，才能实现对《公约》中电子数据收集、提取等有关规定的转化衔接。

有学者认为，不应再对电子数据规定在三大诉讼法中进行分散式立法，而宜优先考虑对电子数据（电子证据）以单行法规的形式进行专门立法；^[38]须知，2025年全国人大重点领域立法规划中，尚没有提出《电子数据法》等单行法的制定计划，因此仍需要聚焦《刑事诉讼法》修改层面，讨论电子数据制度的完善。

同时，有学者从比较法角度，认为许多同时是《布约》缔约国的欧美国家秉持“不限定证据的形式，注重定案的实质性”的开放式立法体例，其本国刑事诉讼法也没有完全吸收《布约》分类的规定，以此反对“对于《公约》所规定的电子数据分级分类制度优先从《刑事诉讼法》的层面予以落实”的观点。诚然，一些缔结《布约》的欧洲国家到目前为止还没有出台单独的电子数据立法，甚至鲜见单独的证据法，这其中以法国和德国最为典型，其有关证据的规范呈现出在各诉讼法中散落分布、表述模糊的现象；^[39]但需澄清的是，这些国家在刑事诉讼领域近年来也陆续展开了对“不同强度获取电子数据的手段”的区分和限制，即这些国家可能无电子数据分类之名，但有基于基本权利保护而对电子数据分类处置之实。^[40]最重要的是，这些国家刑事诉讼法暂时没有电子数据分级分类的规定，并不代表两国的电子数据取证实务中缺乏“三分法”的法律依据。作为欧盟成员国，其所依据的2023年出台的《欧洲提交令和欧洲保全令条例》（后简称《条例》）就采

^[37] 参见陈光中：《完善涉外刑事诉讼若干问题探讨》，载《法学》2025年第2期，第98～113页。

^[38] 参见许晓彤：《组织机构电子文件证据效力保障研究》，中国社会科学出版社2023年版，第261～285页。

^[39] 参见尹鹤晓：《电子数据取证：法律规制与侦查实务》，中国政法大学出版社2024年版，第31-33页。

^[40] 尤其以德国2017年刑事诉讼法改革为典型，其通过修订第100条a、第100条b第100条d第100条e的内容，对源端电信监控和在线搜查等电子数据收集措施进行了规定，参见。施鹏鹏、褚侨：《德国刑事诉讼与证据制度专论（第一卷）》，法律出版社2023年版，第154～155页。

取了“三分法”的规定。^[41]质言之，这些欧盟国家通过直接履行《条例》，即可实现在欧洲范围内贯彻《公约》要求。同时须知，《条例》将于 2026 年 8 月正式生效。笔者认为，比较法上的经验反而能够更进一步增加理论与实务界关于“中国刑事诉讼法亟待回应新的电子数据分类要求”的确信。

四、藉由《刑事诉讼法》引入《公约》电子数据分类规定的可行性建议

笔者认为，藉由《刑事诉讼法》贯彻《公约》电子数据“三分法”有其必要性，亦存在直接迁移《公约》电子数据“三分法”的可行性基础。回顾前文所述公、检等部门的规范文件，电子数据的分类相较《公约》只多不少，甚至可以期待，未来《公约》召开缔约国会议、讨论附加议定书问题时，作为“数据大国”的中国，可以将本国电子数据分类取证的工作经验上升为国际标准。进一步需要思考的是，中国既有的“四分法”“七分法”如何与《公约》“三分法”对应？既有的取证措施规则、机制是否完备，与《公约》项下的规则与要求如何对应？笔者认为，至少可以立足于以下三个方面，在《刑事诉讼法》层面对《公约》电子数据分类及其要求予以落实。

（一）明确《公约》对电子数据分类与中国刑事规范中分类的对应关系

2021 年《人民检察院办理网络犯罪案件规定》（后文简称《规定》）第 27 条对电子数据的分类最全，是因该《规定》除了需要面向办案人员，为办案审查提出规范指引外，还考虑了要面向技术人员，对于技术支持办案提出基本要求，因此规定较为充实。^[42]笔者认为，由于《刑事诉讼法》主要面向公检法等具体办案职能部门，其分类及表述以《公约》“三分法”为准；同时，可以考虑将《规定》“七分法”根据《公约》“三分法”的规定进行逐项对应。笔者建议，《规定》第 3 项除“生物识别信息”中的“人脸、指纹、声纹、虹膜、DNA 等”以外，包括“用户注册信息、身份认证信息、数字签名”以及“生物识别信息”中的“行为特征（步态、击键习惯等）”在内的“用户身份信息”，还包括《规定》第 17 条、当中有重点提及但没有划归至“七分法”下的“域名、IP 地址（仅限用于揭露用户登录行为）、终端 MAC 地址、通信基站”等网络标识信息，应当视为《公约》中的“订阅用户信息”；这种分类的理由在于这些电子数据一般是调查嫌疑人身份时的第一线索，目的是能够反映电子设备为犯罪嫌疑人所使用，查证网络犯罪行为主体与现实主体的“人机同一性”，相较其他电子数据类型不太敏感，与“订阅用户信息”的设定宗旨一致。笔者需要特别解释“用户身份信息”不包括“人脸、指纹、声纹、虹膜、DNA”的原因。理论上讲，在刑事诉讼中，对于犯罪嫌疑人指纹、人像等的收

^[41] See Regulation (EU) 2023/1543 of the European Parliament and of the Council of 12 July 2023 on European Production Orders and European Preservation Orders for electronic evidence in criminal proceedings and for the execution of custodial sentences following criminal proceeding, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023R1543&qid=1740232208220>

^[42] 最高人民检察院副检察长孙谦就《人民检察院办理网络犯罪案件规定》答记者：
https://www.spp.gov.cn/xwfbh/wsfbt/202102/t20210226_509964.shtml#2，最后访问时间：3 月 30 日。

集都是必需的，相对来说保护层级的强烈性有所削弱，但基于《公约》第 36 条“保护个人信息”以及我国《个保法》上“敏感信息”的规定，再结合欧洲的《提交保存令条例》对生物识别隐私期待过高的立法实践，笔者认为，将《规定》附则中第 62 条第 6 项，作两分法，将“人体所固有的生理特征”这类生物识别信息排除出“订阅用户信息”为妥。

《公约》中的“流量数据”，除了包括《规定》第 27 条第 4 项用户行为信息、第 5 项行为工具信息、第 6 项系统运行信息以及第 7 项中的附属信息，还包括《规定》第 18 条所提到的用以反映犯罪人客观行为的轨迹的“IP 地址、WiFi 信息、地理位置信息”。这里需要特别说明一下 IP。如果要求提供 IP 地址、访问号码等相关信息的唯一目的不是为了识别用户，则通常会要求它们获取更多侵犯隐私的信息，例如用户的联系方式和下落，同时它们以结构化和标准化的格式呈现，比内容数据更容易分析和处理，则必须将 IP 地址、访问号码等视为“流量数据”。其分类的理由在于，“流量数据”相对敏感，对公民个人信息权的干预程度更高，因此基于“流量数据”的分类宗旨，应当严格限制授权。

而《规定》第 27 条第 2 项网络通讯信息以及第 7 项电子文件属于“内容信息”。由于“内容数据”关涉公民的通信自由和通信秘密权，应遵守更严格的要求，以反映此类数据更敏感的性质；同时基于法益保护和相称性原则，对这类电子数据获取限制在更严重的罪行上，不会过度影响“服务提供商”等从业者；至于“网页、社交平台、论坛等网络平台发布”的公开性质的电子数据以及基于法律法规在网络平台公开的 IP 标识（国内网站上的 IP 标识不同于前述 IP 地址，一般只具体到国家、省份），属于《公约》一般意义上的表述的“个人数据”（与已识别或可识别身份的自然人有关的任何信息）。当然，当出现“个人数据”海量收集时候也应给予规制，收集的表明必要性与紧迫性，但这种情况通常涉及《数据安全法》《个人信息保护法》中电子数据的分级问题，限于篇幅，需要另文论述。

（二）明确有关侦查取证措施对《公约》分类下电子数据的对应适用

《公约》第 28 条所规定的“搜查和扣押”，同时针对三种电子数据，不仅要求适用于可能的“存储介质”，其第 1 款第 1 项还要求直接适用对象是“电子数据”及其系统。对电子数据的搜查、扣押措施，学理上一般认为，电子数据搜查分为“直接模式”和“间接模式”，其中“直接模式”是指侦查机关直接对电子数据进行搜查，而不依赖于对存储介质的搜查。因此可以认为，《公约》的规定有电子数据“直接搜查”优先的意味。《刑事诉讼法》要想改变中国既往实践中一惯的“间接搜查”模式的适用范围，也就意味着通过现场提取、网络在线提取、网络远程勘验等侦查措施将被纳入“搜查和扣押”制度中进行规制。^[43]因此，无论从对接《公约》

^[43] 本身也有学者认为，《公约》第 28 条的规定，明确中国实务谈判部门对于“电子数据本身可以作为搜查的客体”的态度，从实质角度来理解法律文件中的电子数据远程勘验、现场提取、在线提取，完全可以将之整合进“搜查”措施的框架下，参见冯俊伟：《数字时代背景下刑事取证制度的困境与回应》，载《北方法

的角度讲，还是保障“数字时代基本权利”的角度来说，中国《刑事诉讼法》都有必要对电子数据的直接搜查进行规定。基于刑事诉讼改革的渐进性，笔者认为可以在“间接搜查”模式中，嵌入适用电子数据“直接搜查”模式。^[44]但此前有学者统计，现阶段刑事诉讼中的主要是警察、检察官等取证主体不擅长于数字取证。^[45]因此笔者认为，中国《刑事诉讼法》可以参照“勘验、检查”的规定，搜查中可以指派、聘请有专门知识的人，以解决部分电子数据自身技术性强的问题，便于有效搜查，更高效率地打击网络犯罪。

对于《公约》规定的第 29 条针对“流量数据”的“实时收集”规则，和第 30 条针对“内容数据”的“拦截”规则，都采用了“应用技术手段实时收集或记录”的表述，并规定了“服务提供者”收集、配合及保密义务；区别在于针对“内容数据”这种措施，仅限定于“最高刑不少于四年”的“严重犯罪”。基于刑事诉讼法学理论中的“强制性侦查措施”和“任意性侦查措施”之区分，可以认为，对于“流量数据”的“收集”，通常并不会干预公民的通信自由和通信秘密权，但是会干预公民个人信息权，因此可归于“任意性侦查措施”；对于承载着公民的通信自由和通信秘密权的“内容数据”进行通常会持续较长时间的“拦截”，则需要严格遵循《公约》重罪原则，应当认为是“强制性侦查措施”，与中国电子数据“技术侦查”中的“数据监控”类似。因此，对于“内容数据”“拦截”规制进行国内法转化时，应当适用对“技术侦查”的规制。如对微信、QQ 聊天记录的提取和后续聊天记录的监控、对手机定位或 IP 地址的提取等行为，由于需要使用远程数据提取、在线监控、私人数据提取等技术，应当都被归于技术措施的应用范围。当然，在电子数据的技术侦查过程中，由于其具有时效性和易被篡改的特征，取证工作需要争分夺秒来尽可能赶在证据被更改删除前完成。有学者认为，如果适用严格审批程序，就会因为相关手续的繁琐导致侦查人员工作起来束手束脚，甚至延误最佳取证时机，如果“不仅不利于侦破案件，还会造成事实性的违法”。^[46]这种观点有一定道理，如果不适用事前严格审批，而以发布通知书等书面令状调取电子数据，对特别情况在事前说明理由，也面临在涉外层面因技术侦查扩张受域外非难而承担国家责任的风险。因此需要审慎考虑电子数据技术侦查的规则表述。

笔者认为，当向网络服务提供者调取网络通讯信息等“内容数据”扣押时，首先应当适用“邮件检查扣押”，即经过县级以上公安机关负责人的审批后才可以适用；对于承载公民个人信息的“订阅用户信息”，调取可能会干预或者侵犯其信息

学》2024 年第 1 期。

^[44] 参见谢登科：《电子数据搜查的两种模式》，载《交大法学》2025 年第 2 期，129～133 页。

^[45] 据实务部门统计，刑事电子数据提取率不足 2%，参见刘浩阳等：《公安机关办理刑事案件电子数据取证规则释义与实务指南》，中国人民公安大学出版社 2020 年版，第 158～159 页。

^[46] 参见裴炜：《“爬虫第一案”的反思》，转引自刘品新、谢鹏程主编《电子证据与网络犯罪治理 20 讲》，清华大学出版社 2024 年版，第 181 页。

自决权的，可以转化为适用“调取”，但需要将此类电子数据的调取界定为强制性侦查措施，可以设置审批程序；对于“网页、社交平台、论坛等网络平台发布”的曾公开过的电子数据进行调取，可以转化为适用“调取”，但对此类电子数据的调取需要将其界定为“任意性侦查措施”，可不设置相应审批程序，但在调取时需要出具相应的说明文件。

而关于《刑事电子证据规定》第 12 条首创的对电子数据的“证据型”冻结措施，笔者认为在语词上需要与《公约》针对犯罪财产所得而设置的“财产型”冻结相区别。基于对于电子数据冻结实务的考察，笔者认为，由于指导意见目前尚付阙如，公、检察机关等司法部门“应如何冻结、针对不同类别电子数据的调取需要哪些手续、冻结通知书应如何制作和下达”等操作也是一头雾水；而涉及电子数据的刑事案件往往案情复杂，办案周期长，大型的电子数据运营商、提供商等第三方平台又限制电子数据保存期限，从公共环境中调取或冻结数据不及时，导致关键证据灭失的情况时有发生。为了避免电子数据冻结措施沦为空转条款，也为避免《公约》的适用出现错误，笔者认为，在保留《刑事诉讼法》第七章附带民事诉讼章节第 102 条中“财产型”冻结同时，不宜将《刑事电子证据规定》中的“冻结”措施吸纳至《刑事诉讼法》第五章证据制度中，而应当明确规定“电子数据保全”措施，并采用“快速保全”的表述。

（三）明确“三分法”下电子数据“三性”标准和非法证据排除问题

本文的写作，对于证据属性的讨论不可避免。此前学界存在各种学说，但幸运的是目前有关学者“证据有‘三性’”上共识较强，主要分歧在于“‘三性’之中何为第一性”。

针对电子数据，通说认为客观性（包括电子数据真实性、完整性）是电子数据被采纳的前提，2016 年以后，中国的司法实践对于电子数据完整性的重视日渐突出，成为各机关审查的核心。但是基于警务预防的需要，存在大量的“黑数据”，进而产生了大量有关电子数据取证侵权问题的讨论。笔者认为，与《刑事诉讼法》第 50 条所确立的一般刑事证据审查框架相比，2016《刑事电子证据规则》至 2021 年《高法解释》所承继下来的电子数据规范框架不太强调证据的关联性，然而与网络犯罪相伴的海量数据问题及其过度调取问题又使我们不得不面对关联性问题。因此，笔者认为，对于电子数据规则的丰富，不仅要从区分电子数据类型以规制非法调取入手，还有注重对冗余电子数据的剔除与还原，也即将与案件事实无关但调取时又限于数据集形式获取的电子数据“不作为证据”。

同时，多数学者建议在《刑事诉讼法》修改背景下，增加针对合法性的排除规则，将电子数据取证侵害公民权利的情形纳入排除范围，表达了对电子数据的合法性关切。^[47]现行《刑事诉讼法》第 56 条并未将电子数据纳入非法证据排除

规则的适用对象，也没有回答电子数据的存储介质是否可以按照物证进行排非，电子数据的打印件是否可以按照书证进行排非的问题。而《刑事电子数据规则》所规定的“电子证据的真实性难以查实的，不能作为定案的根据”，也不属于针对电子数据非法排除规则。

可以肯定的是，应当根据《公约》最新的电子数据分类，对应地拓展非法证据排除规则的适用，对针对不同电子数据的取证加以规制。^[48]在考虑《公约》下电子数据非法证据排除的规则细化时，可以运用普通法的方法，关注欧美等《布约》缔约国的经典案例、素材作为参考，促进审判人员对侦查取证行为合法性的经验判断。同时，也要考虑通过进一步丰富指导性案例尤其是涉外跨境取证的指导性案例，逐步解决这一难题。

余论

（一）《公约》语境下电子数据的分类与国际数字治理语境下数据分类分级的关系

国际数字治理语境下“作为数据要素”的电子数据的所涉及的数据分类分级问题，和《公约》语境下“作为证据”的电子数据分类问题，是两个内涵不同但密切相关的问题。

国际数据治理这一宏观意义上的数据的分类分级，涉及行政、司法、国家主权与安全等各个领域数据的管控。尽管常被笼统讨论，但实质上包括数据分类与数据分级两个层面。数据分级通常是指按照数据所承载的法益大小，或者说处理数据的后果为标准而对数据采取的层级化处理；数据分类，主要是指按照数据所承载的内容、数据的来源等特征而对数据进行相对应的划分和处理。数据的分级的评判标准，一般认为，由于数据具有天然的流动性，应当基于国家实行安全制度的前提之下加以讨论，同时需要兼顾数据自由开放的价值实现；而数据的分类的评判标准，是“粗细得宜”，既具有足够的涵盖性，又不能因过度精细化而对数据属性的把握更加混淆和困难。

《公约》语境下“作为证据”的电子数据分类，涉及对证据的分类，限于司法领域的运用，旨在使刑事司法部门“认识条理化...促进处置的目的性与有效性”。^[49]作为证据的电子数据，一度被一些学者认为是将“各种传统证据部分地剥离出来而泛称的一种新证据形式”，^[50]虽然这种观念在实证法上得到一定程度的否定，但正如学者所说，“传统证据类型，它的规则是一个有限的集合”，而“任何一个规则，都很难将它（电子数据）描述严谨”，^[51]更有学者指出，电子数据审查应当

^[48] 参见郑曦：《数字司法实践的刑事程序规制——以〈刑事诉讼法〉再修改为背景》，载《法学家》2025年第1期，第62页。

^[49] 参见龙宗智：《证据分类制度及其改革》，载《法学研究》2005年第5期。

^[50] 参见何家弘、刘品新：《证据法学》（第4版），法律出版社，第185页。

^[51] 参见戴士剑讲座：《电子证据基础属性的理解与应用》，https://mp.weixin.qq.com/s/BliLAdThJM-tY_QZXfJuOQ，最后访问时间：2025年3月15日。

重视面向实践的类型化思维，以溯源审查、动态审查、多向审查等具体方法，充分发挥电子数据的证明作用。因此，现阶段中国确实需要对电子数据更加精确的分类，以对这种证据类型施加不同的规范。

笼统来论刑事诉讼中电子数据的分类分级，是指根据一定标准，对刑事诉讼过程中收集、提取的电子数据进行类型，并针对性地制定适宜的证据规则和程序性要求的制度设计。目前，《网络安全法》正处于修改议程中，因此囿于对中国最新立法动向的研究进程和文章篇幅的限制，本文仅就《公约》背景下刑事诉讼领域电子数据的分类进行讨论。然而即使仅从分类的意义上对电子数据进行认识，其积极意义也在于，在尽可能打击国内国际涉网犯罪的同时，也加强了对各种类型电子数据所涉及国家主权与安全、社会公共利益及公民权利等多个方面重要利益的考量，本质上，也是对“分级授权、分类应用、权责一致”的数字法治精神的进一步落实。^[52]经过笔者对与电子数据分级有关的规范进行梳理，发现《网络安全法》第 21 条明确规定了国家“实行网络安全等级保护”、网络运营者“采取数据分类、重要数据备份和加密”的措施要求，只是未涉及数据分级制度的构建；而《数据安全法》第 21 条同时兼顾了数据的分级与分类，但对数据分级分类仍不精细。目前，有关下位法存在“核心数据”“重要数据”“一般数据”以及“三级”“二级”“一级”等划分形式。^[53]这其划分标准需要紧密结合国家最新的数据安全审查政策，才能实现研究成果的实效性和制度设计的精确性；与此同时，立足企业的数据分类分级治理视角，在对《公约》所规定“订阅用户信息”、“流量数据”和“内容数据”的流动的操作实现过程中，必然需要我国数字企业的数据治理体系能够为自身提供完整、清晰的数据资产检索图谱以及敏捷、高效的数据识别调取能力，并且要能够同时满足不同国家关于数据跨境流动的合规要求。这些问题尚需时日，留待进一步研究深化。

（二）《刑事诉讼法》修改之外，应配套地优化电子证据取证的体制机制，加强能力建设

《公约》下的数据分类，既体现网络主权理念，又体现刑事诉讼领域人权保护的关切。本文的论述重点在于讨论《公约》第四章针对电子数据应采取的措施及程序在中国《刑事诉讼法》中的转化及其可能性。需要进一步讨论的是，基于《公约》电子数据分类的要求，第五章中有关跨境取证国际合作、公私合作有关事宜应当如何落实转化的问题，如国内有关部门法对于执法部门与服务提供商的

^[52] 参见郑文阳《数据安全法治化研究》，中国政法大学出版社 2024 年版，第 21 页。

^[53] 笔者认为，现已生效实施的《网络数据安全条例》系对《网络安全法》、《数据安全法》、《个人信息保护法》的细化规定，其实质体现了“核心数据”“重要数据”“一般数据”三分，https://www.gov.cn/zhengce/zhengceku/202409/content_6977767.htm，最后访问时间，2025 年 4 月 12 日；而《工业互联网安全分类分级管理办法》第 6 条则根据工业互联网企业由高到低分为三级、二级、一级，暗含了“将有关‘重要数据’作了三级划分之意”，https://wap.miit.gov.cn/jgsj/waj/wjfb/art/2025/art_72d3dab251474245908611263f50b096.html，最后访问时间，2025 年 4 月 12 日。

对接机制应当如何设置，同其他缔约国的电子数据证据互助应当如何规定。除了前文所述，在《刑事诉讼法》中对电子数据分类加以明确外，还需研判如何在《国际刑事司法协助法》《数据安全法》《个人信息保护法》等法规中增设专门的数据、信息保护和相应保障措施条款，明确有哪些国家标准得以参照、哪些国家标准尚付阙如。^[54]

除此规范层面的考量之外，实践层面还需要加强公检法等部门应对跨境取证合作中的能力建设。无论是电子数据取证还是审证，无论是域内还是域外，无论是民事诉讼还是刑事诉讼，想要提高数字时代涉电子数据案件办案质量与效率，都必须要求司法队伍对于新兴的电子数据没有抗拒心理且较为熟稔。而对于所谓人工智能证据、大数据证据，由于其与算法黑箱紧密结合，运行机理不被检察、审判人员所知。需要承认，如果没有具备相关专业知识的参与解读，检察官、法官们往往只能望洋兴叹但退一步讲，实务中，即使聘请有关技术专家进行评估论证，是否被采纳也是不必然的。“神证”到“物证”的历史已然昭示我们“朝前看”，唯有提高有关办案主体的理念与认识才可破局。

特别从侦办案件的取证主体角度讲，数据提权、网络监控与数据分析、网络远程勘验、暗网侦查以及加密追踪等技术手段需要及时强化。在技术、手段层面，中国其实并非弱国，但思维和习惯方面，需要公安等执法队伍走出舒适区，适时突破，对于欧美先进的执法经验也要敢于承认与借鉴。

笔者有注意到，有关学者提出“数字证据”和加强实质审查的提法，以破解电子数据制度阙如、空转及证据法定主义的积弊。^[55]但无论理论主张为何，要想让电子化、数字化证据“活起来”，须以数字侦查、数字检察、数字司法等能力建构为基础。只有将刑事诉讼领域的各项数字化工程系统性构建起来，形成“数字”诉讼体系，才能在刑事诉讼法中落实开放的数字证据制度并明确有关规范。

^[54] 以《个人信息保护法》为例，其中第 62 条也明确网信部门，针对小型个人信息处理者制定专门的个人信息保护规则、标准。这类国家标准，也被纳入 2025 年定制计划之中，期待早日落地。

^[55] 参见郑飞：《证据种类法定主义的反思与重构》，载《中国法学》2024 年第 1 期，第 122 页。



中国国际法学会
Chinese Society of International Law



大连海事大学
DALIAN MARITIME UNIVERSITY

中国国际法学会2025年学术年会
2025 Annual Academic Conference of the Chinese Society of International Law
暨世界反法西斯战争胜利及联合国成立80周年与国际法研讨会
and the International Law Symposium Commemorating the 80th Anniversary of the Victory of the World Anti-Fascist War and the Founding of the United Nations

加强涉外法治建设与推进中国式现代化：国际法的地位和作用
Strengthening the Role of Law in Foreign Affairs and Promoting Chinese Modernization: The Status and Role of International Law

主办单位: 中国国际法学会 大连海事大学
Co-sponsored by Chinese Society of International Law and Dalian Maritime University
承办单位: 大连海事大学法学院
Organized by Law School of Dalian Maritime University

2025年4月25日-27日, 大连
April 25-27, 2025, Dalian