跨越核技术类比陷阱：人工智能的国际安全风险及管控

苏若林　贾　开

内容提要　历史类比是人类最常用的理解世界进而进行决策的方法之一。随着人工智能技术的跨越式发展，学界和政策界试图通过与核技术的类比来分析人工智能带来的国际安全风险，并认为应参考核技术的全球管控模式来指导人工智能全球治理机制或平台的建设与改革。但这种简单类比可能存在认知误区，具体表现为颠覆性技术误区和冷战误区，前者是指在颠覆性技术框架下强调人工智能与核技术的相似性，却忽略了两者从技术属性到影响机制的实质差异；而后者是指在美苏冷战框架下思考中美人工智能竞争与合作，可能过分简化国际环境而无视其他行为体的作用。基于两类误区的分析，人工智能全球治理的利益相关方应该超越简单的核技术类比

* 苏若林：上海交通大学国际与公共事务学院副教授、上海市创新政策评估研究中心研究员。(邮编：200030)；贾开：上海交通大学国际与公共事务学院副教授。(邮编：200030)

** 本文是国家社会科学基金青年项目(项目编号：24CGJ007)的阶段性成果。感谢《国际政治研究》匿名评审专家的意见和建议，文责自负。

陷阱，结合人工智能自身特点及国际格局变化，采用更加匹配的分析框架，在准确把握中美人工智能竞争态势的前提下，发展出切实可行的治理模式，进而推动人工智能安全治理的全球合作。

关键词 非传统安全 人工智能 核技术 历史类比 国际安全风险

近年来，伴随人工智能技术创新及其应用步伐的加速，人工智能对于人类社会的影响力日益增加，这也使之成为国际社会普遍关注的重要议题。尤其是在以大语言模型为代表的通用人工智能技术出现重要技术突破的背景下，围绕前沿人工智能可能带来重大风险的担忧，正在推动世界主要国家积极涉入人工智能全球治理进程。2023 年英国“全球人工智能安全峰会”、2024 年中国“2024 世界人工智能大会暨人工智能全球治理高级别会议”、2025 年法国“人工智能行动峰会”都是该进程中的关键节点，而它们的共性是试图推动并建立人工智能全球治理的有效机制或平台，在管控风险的同时促进技术的健康发展。

但作为前沿新兴技术，如何理解人工智能带来的安全风险、通过何种机制或平台来管控人工智能安全风险，仍然是摆在决策者面前的重要挑战。在围绕该问题的持续讨论中，相当部分研究者和决策者将人工智能安全风险类比于核安全风险，并提出应参考核安全管控方式来应对人工智能安全风险。[①] 例

① See Matthijs M. Maas, “Two Lessons from Nuclear Arms Control for the Responsible Governance of Military Artificial Intelligence,” in Mark Coeckelbergh, et al., eds., *Envisioning Robots in Society-Power, Politics, and Public Space*, Vol.311, IOS Press, 2018, pp. 347-356; Eoin Micheál Mcnamara, “Nuclear Arms Control Policies and Safety in Artificial Intelligence: Transferable Lessons or False Equivalence?” *Finnish Institute of International Affairs*, January 2024, https://fiia.fi/en/publication/nuclear-arms-control-policies-and-safety-in-artificial-intelligence, 2025-03-15; Seokki Cha, “Towards an International Regulatory Framework for AI Safety: Lessons from the IAEA's Nuclear Safety Regulations,” *Humanities and Social Sciences Communications*, Vol.11, No.1, 2024, pp. 1-13; Andrew Jones, “From Nuclear Stability to AI Safety: Why Nuclear Policy Experts Must Help Shape AI's Future,” *European Leadership Network*, 25 April 2025, https://europeanleadershipnetwork.org/commentary/from-nuclear-stability-to-ai-safety-why-nuclear-policy-experts-must-help-shape-ais-future/, 2025-04-30.

如,蒙特利尔大学教授、“图灵奖”获得者约书亚·本吉奥(Yoshua Bengio)就旗帜鲜明地要求限制前沿人工智能技术的开放与扩散,因人工智能的超强能力使之类似于核技术,如果被误用滥用将给社会带来重大风险,所以,本吉奥认为,应该像管控核技术那样来管控人工智能技术的研发与应用。[①] 在国际关系领域,尤其是考虑到大国竞争的时代背景,诸多国际关系学者纷纷将目光回溯至冷战时期的核技术时代,认为历史上对于核技术及其风险的讨论为当前理解人工智能时代的国际安全风险及其管控提供了极强的借鉴意义,并提出应参考核技术的全球管控模式来指导人工智能全球治理机制或平台的建设与改革。[②] 其中,最具代表性的是亨利·基辛格(Henry A. Kissinger)与哈佛大学肯尼迪政府学院创始院长格雷厄姆·艾利森(Graham Allison)的观点,他们指出应参考之前的核技术管控机制来阻止当前的人工智能军备竞赛。[③]基辛格与艾利森的观点进一步被联合国秘书长古特雷斯所借鉴,从而转变为实际上的政策建议。在2023年7月联合国安理会首次举行的“人工智能与安全问题高级别公开会议”上,古特雷斯呼吁成立类似于国际原子能机构的国际人工智能监管机构,引起了彼时舆论的普遍关注。

① FAQ on Catastrophic AI Risks, https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/, 2025-02-15.

② Jürgen Altmann and Frank Sauer, “Autonomous Weapon Systems and Strategic Stability,” *Survival*, Vol.59, No.5, 2017, p.117; Matthijs M. Maas, “How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons,” *Contemporary Security Policy*, Vol.40, No.3, 2019, pp.285-311; Amandeep Singh Gill, “Artificial Intelligence and International Security: The Long View,” *Ethics & International Affairs*, Vol.33, No.2, 2019, pp.169-179.

③ Henry A. Kissinger and Graham Allison, “The Path to AI Arms Control,” *Foreign Affairs*, October 13, 2023, https://www.foreignaffairs.com/united-states/path-ai-arms-control, 2025-02-15.

历史类比是人类最常用的理解世界进而进行决策的方法之一[①]，是指将当前事件、情境或问题与历史上类似的事件进行比较，以发现两者的相似点和差异，从而帮助理解当前情境、指导决策或预测可能的结果。[②] 尽管参考核技术全球治理来指导人工智能全球治理改革一定程度上具有广泛的理论基础乃至正在形成政策共识，但是，这种简单类比本身可能存在认知误区，原因在于这一类比忽略了人工智能技术的独特性，不利于理解人工智能的技术逻辑、风险逻辑和管控逻辑，可能为有效应对人工智能时代的国际安全风险带来负面影响。事实上，如果以核安全管控方式来应对人工智能安全风险，不仅可能制约人工智能提升人类社会福祉的潜力，也可能因违背人工智能技术创新应用规律而难以有效应对安全风险的发生与扩散。一方面，核安全管控的严苛性势必影响人工智能技术的开放水平，只能在有限范围内被获取的技术资源与知识将严重影响人工智能作为一般通用技术在各个领域的普及，这不仅可能扩大“智能鸿沟”并进一步加速权力分化格局，也会反过来影响人工智能技术本身的可持续创新；另一方面，人工智能安全风险与核安全风险产生逻辑的差异性，使得以核安全管控模式处置人工智能安全风险存在不适配问题，而这可能

① Ernest R. May, “*Lessons*” *of the Past*: *The Use and Misuse of History in American Foreign Policy*, New York: Oxford University Press, 1973; Robert Jervis, *Perception and Misperception in International Politics*, Princeton: Princeton University Press, 1976, pp. 218-220, 274-275; Robert Jervis, “Political Psychology: Some Challenges and Opportunities,” *Political Psychology*, Vol.10, No.3, 1989, pp. 481-493; Howard Schuman and Cheryl Rieger, “Historical Analogies, Generational Effects, and Attitudes toward War,” *American Sociological Review*, Vol.57, No.3, 1991, pp. 315-326; Rose McDermott, *Political Psychology in International Relations*, Ann Arbor: The University of Michigan Press, 2004, pp. 62-63; Benjamin E. Goldsmith, *Imitation in International Relations*: *Observational Learning*, *Analogies*, *and Foreign Policy in Russia and Ukraine*, New York: Palgrave Macmillan, 2005; Giovanni Gavetti, et al., “Strategy Making in Novel and Complex Worlds: The Power of Analogy,” *Strategic Management Journal*, Vol.26, No.8, 2005, pp. 691-712; Keith J Holyoak, “Analogy and Relational Reasoning,” in Keith J Holyoak and Robert G Morrison, eds., *Oxford Handbook of Thinking and Reasoning*, Oxford University Press, 2012, pp. 234-259; 廖颖等：《不确定环境下的战略决策：类比推理的作用》，《外国经济与管理》2018 年第 8 期。

② David Patrick Houghton, “The Role of Analogical Reasoning in Novel Foreign-Policy Situations,” *British Journal of Political Science*, Vol.26, No.4, 1996, pp. 523-552; David Patrick Houghton, “Analogical Reasoning and Policymaking: Where and When Is It Used?” *Policy Sciences*, Vol.31, No.3, 1998, pp. 151-176.

会导致真正严重的人工智能安全风险不能得到有效管控,对社会带来实质性伤害。此外,也有些研究者意识到将中美之间的人工智能竞争类比美苏冷战期间围绕核武器的竞争对双边关系乃至国际格局稳定的负面影响,但却未对这种类比的谬误本质、深层次原因及其具体影响进行深入分析。①

正是因为存在上述问题,深入探讨人工智能的核类比逻辑的认知误区,廓清核技术安全和人工智能安全及其管控方式的差异,并在此基础上为未来人工智能安全治理提供扎实理论支撑,在当前便具有重要价值。为此,本文将详细描述人工智能的核类比逻辑的具体表现,分析其背后的根源及其会产生的误导性后果,进而提出知识界、政策制定者和大众应如何协作共同跨越这一类比陷阱。

一、人工智能与核技术:相似性及差异性

在国际关系领域,人工智能正逐渐成为影响全球政治、经济和安全格局的重要力量。② 学者和政策分析家们常常借用核技术作为类比对象,将人工智能置于类似核技术的认知和叙事框架中,以阐明人工智能的深远影响、潜在风险以及可行的管控路径。现有研究着重强调了人工智能与核技术在三方面的相似性:第一,颠覆性潜能。利益相关方之所以将核技术与人工智能技术相提并论,在很大程度上源于它们都被视为具有战略意义的变革性技术。具体来说,核技术与人工智能技术都具有民用和军用的双重用途属性,是对人类社会能产生深远影响的变革性力量,能够改变国家间力量对比、重塑国际秩序;③第二,灾难性后果风险。核技术和人工智能都伴随巨大风险,包括安全风险、伦

① James Johnson, "Inadvertent Escalation in the Age of Intelligence Machines: A New Model for Nuclear Risk in the Digital Age," *European Journal of International Security*, Vol.7, No.3, 2022, p. 351.

② 傅莹:《人工智能对国际关系的影响初析》,《国际政治科学》2019 年第 1 期。

③ Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review*, Vol.1, No.3, 2018, pp. 37-57; 贾子方、王栋:《人工智能技术对战争形态的影响及其战略意义》,《国际政治研究》2020 年第 6 期,第 38—39 页。

理问题和治理挑战，而核军备竞赛的历史经验常被用来警示人工智能领域的潜在军备竞赛[①]；第三，迫切性治理需求。灾难性后果和严峻的治理挑战推动着围绕两者的迫切性治理需求，而大国竞争背景的时代相似性讨论也在明示或暗示核风险管控路径对人工智能风险管控具有极强的借鉴意义。[②]

这些关于两者相似性的探讨在逻辑上是紧密相连的。核技术和人工智能都被视为具有颠覆性的关键技术，即体现出相比于之前任何技术创新都更大的变革潜能。因此，二者都被认为可能产生重大且不可逆转的灾难性后果，核扩散对人及自然的影响，以及前沿人工智能技术扩散在虚假新闻、舆论操纵等方面的风险，都成为其所处时代的迫切治理需求。但二者相似性类比并不简单局限于技术层面，大国竞争的时代背景是不可回避的另一重要政治经济因素。正是因为这两类技术的创新应用“恰好”处于国际形势走向对抗的历史进程中，从而放大了技术变革之于主权国家及国际格局变迁的重要性。

上述认知逻辑在理论上可被总结为国际关系理论中的“领先技术”(leading sector technology)视角。传统国际关系理论认为，技术是影响国家间权力结构的重要变量，而其中关键又在于所谓的“领先技术”，即能够引发颠覆性变革、极大提高生产率的主要技术创新。[③] 在技术演化史上，纺织业的蒸汽革命之于英国，以及化学工业之于德国都是支撑该论断的典型案例。[④] 此时，影响国家间实力对比的关键即在于一国能否相比于他国更早掌握“领先技术”。在“领先技术”理论视角下，核技术与人工智能都可被视为所处时代的“领先技术”，并因此会从根本上影响国家实力，进而影响国家间权力结构。

---

① Kenneth Payne, “Artificial Intelligence: A Revolution in Strategic Affairs?” *Survival*, Vol.60, No.5, 2018, p.15; 董青岭：《新战争伦理：规范和约束致命性自主武器系统》，《国际观察》2018 年第 4 期。

② Edward Moore Geist, “It's Already Too Late to Stop the AI Arms Race—We Must Manage It Instead,” *Bulletin of the Atomic Scientists*, Vol.72, No.5, 2016, pp.318-321; Matthijs M. Maas, “How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons,” pp.285-311.

③ 黄琪轩：《地缘政治的复兴与大国的技术竞争》，《国际政治研究》2023 年第 6 期，第 55—59 页。

④ Jeffrey Ding, *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition*, Princeton: Princeton University Press, 2024, p.3.

尽管此种类比具有一定说服力,但在理论演绎上仍然存在两个漏洞:第一,“领先技术”假说是否存在竞争性假说,即影响国家间权力结构变迁的关键因素是否必须是“领先技术”?第二,即便承认“领先技术”假说,核技术与人工智能是否都可被归类于此,并因此为了避免国家间的“逐底竞争”而要求建立类似的风险管控措施?从随后的理论讨论与实践探索来看,这两个问题的答案都有其复杂性,而这也说明简单类比的思考视角并不足以支撑“以核技术安全风险管控的传统思路来指导人工智能安全风险管控制度改革”的合理性。

事实上,尽管“领先技术”假说一直以来都是讨论技术革命影响大国权力变迁的主流观点之一,但近年来,一些相关研究已对此提出挑战,指出“领先技术”是否扩散至各个领域并引发普遍性变革是需要考虑的另一个关键维度。[①]例如,电气化技术革命从出现到真正促进全要素生产率增长迟滞了 50 年[②],如果仅以“领先技术”是否出现作为大国权力兴起的判断标准明显不够充分,因而,“技术扩散性”应该成为人们理解技术革命影响国际权力结构转移的另一重要技术特征——但在这一点上,核技术与人工智能技术存在较大差异,并表现出不同的安全风险,进而需要差异化的安全风险管控措施。总的来说,核技术与人工智能技术的差异性主要表现在以下几点:

第一,相比于核技术,人工智能技术的扩散性更强。1970 年首届世界地球日的组织者、环保活动家丹尼士·海耶斯(Denis Hayes)曾指出,核技术因其自由扩散将带来不可挽回的伤害而被要求置于集中控制的权威组织体系之下,与之相比,太阳能技术则更适合松散的自组织管理模式,并表现出更强的扩散性。[③] 在此意义上,人工智能技术更类似于太阳能而非核技术。一方面,从技术生产与再生产角度来看,不同于核技术主要源自高度管控、高度保密的实验

① Jeffrey Ding, *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition*, p. 5.

② Warren D. Devine, *An Historical Perspective on the Value of Electricity in American Manufacturing*, Institute for Energy Analysis, Oak Ridge Associated Universities, 1982, pp. 347-372.

③ Denis Hayes, *Rays of Hope: The Transition to a Post-Petroleum World*, New York: W Norton &. Co Inc., 1977, p. 159.

室科研，人工智能技术更多依赖于开源模式与开放创新。即使是以大语言模型为代表的通用人工智能技术是否应该坚持开源开放在当前引发了一定争议，但开源开放对人工智能技术可持续创新的重要性仍然不可否认[①]；另一方面，从技术价值实现角度来看，不同于核技术具有较明确限定的价值实现方式与过程，人工智能技术价值的实现与否在很大程度上取决于应用场景，而这也意味着人工智能技术不能被提前限定应用方式和应用范围，其价值释放的前提在于需要得到更为广泛的传播与扩散。

第二，相比于核技术，人工智能技术所引发的安全风险更为分散和不确定，这将带来新的大国竞争风险。考虑到核技术形成与应用过程的集中性，核技术引发的安全风险是明确而集中的；相比之下，人工智能技术安全风险体现出新的特点：一方面，开源开放与场景依赖意味着人工智能技术安全风险的分散与不确定，安全风险并不局限于特定区域、特定范围，而且在其产生之前不能准确预判风险类型及其程度。如果说核技术安全风险的本质是“信息不对称”（即风险监管者虽知晓风险内容但并不掌握风险分布），那么人工智能技术安全风险的本质就是“共同无知”，即风险监管者与其他利益相关方都无法准确预知风险内容和风险分布。[②] 另一方面，安全风险的分散和不确定性意味着大国竞争将面临新的风险格局。对核技术而言，大国竞争的对象和底线是清晰而明确的，核大国之间的技术竞争并不影响其防范核扩散的合作动机；与之相比，人工智能安全风险的分散和不确定使得大国竞争可能表现在多领域、多维度，而且也缺乏具有共识性的“风险底线”。

第三，人工智能技术安全风险的特殊性，表明对其管控措施不能简单沿用核技术安全风险管控的既有经验。正如海耶斯指出的，核技术的安全风险特征要求用集中控制的权威组织体系作为管控手段，这在国际层面意味着形成以防止核扩散为基本目标的强监管合作体系，但这一逻辑与人工智能技术的

① 傅宏宇等：《人工智能开源的价值、风险与生态治理研究》，《电子政务》2025 年第 3 期。

② 贾开等：《应对不确定性挑战：算法敏捷治理的理论界定》，《图书情报知识》2023 年第 2 期。

扩散性、人工智能技术安全风险的分散与不确定性特征并不匹配。从实践来看,尽管国际上存在要求成立类似核技术管控体系的改革建议,但人工智能安全管控在很大程度上体现为“机制复合体”特征,与核技术管控既有制度有较大差异。[①] 一方面,欧盟、美国、英国、中国等代表性国家或区域组织都在探索具有明显异质性特征的人工智能安全风险管控制度,而且人们很难有充分理由预期这种异质性会被削弱乃至消除;另一方面,从具体制度内涵来讲,围绕人工智能安全风险管控主题,当前已经形成分类分级、风险评估、风险审计、监管沙盒、安全公共产品等系列制度探索,同时,也形成了敏捷治理、实验主义治理、规制治理等指导未来进一步改革的系列理念创新,当前制度的丰富性、复杂性远超核技术管控的单一制度逻辑。

表　核技术与人工智能的差异性

	核技术	人工智能
技术生产层面	核技术形成与应用过程集中,应用方式和应用范围限定	人工智能依赖开源开放,松散的自组织管理模式体现出更强扩散性
风险影响层面	核安全风险明确,体现出“信息不对称”特征	人工智能安全风险分散且不确定,体现出“共同无知”特征
管控治理层面	核技术要求以集中控制的权威组织体系作为管控手段	人工智能安全管控体现出“机制复合体”特征,各国异质性明显

图表来源:笔者自制。

综上,核技术与人工智能技术在多个维度上体现出显著差异,二者的异质性甚至远大于“领先技术”假说下的共性特征。那么,为什么人们仍会做此类比?这样类比又会产生什么后果呢?

二、人工智能的核技术类比陷阱及其成因

类比推理是在理解复杂世界时人们无意识中使用的认知捷径[②],可以减

① 鲁传颖、〔美〕约翰·马勒里:《体制复合体理论视角下的人工智能全球治理进程》,《国际观察》2018年第4期。

② Jack S. Levy and William R. Thompson, *Causes of War*, John Wiley & Sons, 2011, p. 146.

少因不确定性导致的焦虑。正确类比强调的不只是比较对象的表面相似性或是属性相似性，而更多是其深层次结构的相似性。例如，结构映射理论（structure-mapping theory）就指出类比遵循系统性原则是复杂关系（或关系系统）的映射。[①] 同时，历史类比也不是只为寻求两者相似性，更重要的是试图通过类比来推断因果。[②] 更进一步来说，人们通过历史类比可以将过去的经验和教训应用于新的情境，为复杂问题提供一种易于理解的解释框架。[③] 在国际关系领域，对历史类比的相关讨论常见于战争、危机决策和不确定情境的分析。[④]

当前，人工智能技术实现跨越式发展为人类社会带来极大不确定性，人们很难避免地会使用类比推理来助力这一复杂现象的解释与应对。而通过核类比来理解人工智能的国际安全风险和管控就是这样一种尝试。对于国际关系学者来说，产生这种类比更是件很自然的事，其背后体现的是学界和战略界的认知惯性。首先，将新兴技术或者由此产生的新安全情境进行核类比在国际安全的讨论中并不是新现象，人工智能之前还存在不少有关网络武器以及无

① Dedre Gentner, "Structure-Mapping: A Theoretical Framework for Analogy," *Cognitive Science*, Vol.7, No.2, 1983, pp. 155-170; Dedre Gentner and Arthur B. Markman, "Structure Mapping in Analogy and Similarity," *American Psychologist*, Vol.52, No.1, 1997, pp. 45-56.

② Patrick H. Winston, "Learning and Reasoning by Analogy," *Communications of the ACM*, Vol.23, No.12, 1980, pp. 689-703.

③ Mary L. Gick and Keith J. Holyoak, "Analogical Problem Solving," *Cognitive Psychology*, Vol.12, No.3, 1980, pp. 306-355; Dominic D. P. Johnson and Dominic Tierney, *Failing to Win: Perceptions of Victory and Defeat in International Politics*, Cambridge: Harvard University Press, 2006, p. 54.

④ Alexander L. George, *Presidential Decisionmaking in Foreign Policy: On the Effective Use of Information and Advice*, Boulder: Westview Press, 1980; Yuen Foong Khong, *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of 1965*, Princeton: Princeton University Press, 1992; Abraham F. Lowenthal, *The Dominican Intervention*, Baltimore: Johns Hopkins University Press, 1994; Donald L. Horowitz, *The Deadly Ethnic Riot*, Berkeley and Los Angeles: University of California Press, 2001, p. 550; Jack S. Levy and William R. Thompson, *Causes of War*, p. 146; 张清敏、潘丽君：《类比、认知与毛泽东的对外政策》，《世界经济与政治》2010 年第 11 期；钱文荣：《一战最重要的教训是什么？：驳安倍把今日中日关系与一战时德英关系类比的谬论》，《和平与发展》2014 年第 1 期。

人机技术扩散的核类比讨论。[①] 其次,从学者研究脉络来看,关注人工智能的国际关系学者大都关心普遍意义上的新兴技术的系统影响,其中,许多学者如宾夕法尼亚大学教授迈克尔·霍洛维茨(Michael Horowitz)、斯坦福大学国际安全与合作中心的研究员爱德华·摩尔·吉斯特(Edward Moore Geist)、伦敦大学国王学院教授肯尼斯·佩恩(Kenneth Payne)等都曾长期研究核问题,因此,他们易于潜意识地将两者进行比较。再次,对于人工智能国际安全风险的讨论自然与核问题相关。当前,学界和政策界关于人工智能技术的一个很重要担忧就是其可能削弱核打击报复的可信度,从而冲击全球范围内的核战略稳定。[②]

然而,这种认知惯性下的类比并不意味着比较的准确性和可靠性。既有文献也认识到人工智能技术的核类比是不完美的,但仍试图强调这一类比的积极意义。[③] 人们常常会基于事物、事件相似性来进行预判和决策[④],因此,不正确类比不仅不会增进对未知事物和情境的理解,还会误导人们的判断与相关决策,甚至造成灾难性的后果。而当前将人工智能简单类比核技术并由此

---

① Peter D. Feaver, "Blowback: Information Warfare and the Dynamics of Coercion," *Security Studies*, Vol.7, No.4, 1998, pp. 88-120; David E. Sanger, "Path Set by US and China to Limit Security Breaches May Be Impossible to Follow," *The New York Times*, September 26, 2015, *Gale Academic OneFile*, https://go.gale.com/ps/i.do?id=GALE%7CA429845499&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=03624331&p=AONE&sw=w&userGroupName=anon%7E3eb10e78&aty=open-web-entry, 2025-02-09; Sarah Kreps and Jacquelyn Schneider, "Escalation Firebreaks in The Cyber, Conventional, And Nuclear Domains: Moving Beyond Effects-Based Logics," *Journal of Cybersecurity*, Vol.5, No.1, 2019, pp. 1-11; Michael C. Horowitz, "Do Emerging Military Technologies Matter for International Politics?" *Annual Review of Political Science*, Vol.23, No.1, 2020, p. 390.

② Keir A. Lieber and Daryl G. Press, "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security*, Vol.41, No.4, 2017, pp. 9-49; 蔡翠红、戴丽婷:《人工智能影响复合战略稳定的作用路径:基于模型的考察》,《国际安全研究》2022 年第 3 期;朱荣生等:《人工智能的国际安全挑战及其治理》,《中国科技论坛》2023 年第 3 期,第 162 页;Steve Fetter and Jaganath Sankaran, "Emerging Technologies and Challenges to Nuclear Stability," *Journal of Strategic Studies*, 2024, pp. 1-45.

③ Kareem Ayoub and Kenneth Payne, "Strategy in the Age of Artificial Intelligence," *Journal of Strategic Studies*, Vol.39, No.5-6, 2015, p. 810; Matthijs M. Maas, "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons," pp. 285-311.

④ Stephen J. Read, "Similarity and Causality in the Use of Social Analogies," *Journal of Experimental Social Psychology*, Vol.23, No.3, 1987, pp. 189-207.

进行政策建议的尝试就容易陷入错误类比陷阱，并导致错误认知和应对，其中最值得警惕的两个认识误区是颠覆性技术误区与冷战误区，前者主要涉及技术特性的理解，后者则更多围绕技术发展和应用所处情境的判断。

（一）颠覆性技术误区

人工智能的核类比很大程度上源自人们将人工智能技术和核技术都视为领先技术或是颠覆性技术这一认知。具体到国际安全领域，以自主武器为代表的人工智能技术的军事应用被广泛认为是继火药和核武器之后的“第三次战争革命”。[①] 对颠覆性技术判定本身就体现的是类比逻辑，顾名思义，颠覆性技术着重强调这些技术对人类社会将产生或已经产生的巨大影响，同时相应地暗示它们在议程设置中应该享有绝对优先级，需要整个社会更积极且谨慎地应对。自 2022 年底生成式人工智能技术的突破和大规模应用引爆了新一轮人工智能热潮后，世界范围内对人工智能及其安全风险的高度关注和激烈讨论已经充分印证此点。鉴于近年形成的对人工智能革命性影响的共识，有观点认为，如同核技术一样，人工智能会带来严重的伦理问题和安全风险，如人工智能的军事应用会进一步引发军备竞赛、增加战争风险，因此，需要严格管控技术发展以避免技术的无序扩散和应用。[②] 而反对者则认为，人工智能技术与核技术一样已经成为国家间实力竞争的重要一环，特别是在大国战略竞争背景下，任何强加的管控措施都会使国家错失发展机会乃至输掉国际竞争。[③]

虽然这些比较及思考有一定道理，但是，单纯在颠覆性技术框架下将人工

① Future of Life Institute, “Autonomous Weapons: An Open Letter from AI & Robotics Researchers,” February 9, 2016, https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/, 2025-02-09; James Johnson, “Artificial Intelligence & Future Warfare: Implications for International Security,” *Defense & Security Analysis*, Vol.35, No.2, 2019, p.150.

② 封帅等：《全球人工智能治理：多元化进程与竞争性图景》，《战略决策研究》2024 年第 3 期，第 100—102 页。

③ Michael C. Horowitz, et al., “A Force for the Future: A High-Reward, Low-Risk Approach to AI Military Innovation,” *Foreign Affairs*, Vol.101, No.3, 2022, pp.157-164.

智能与核技术类比,不考虑其他因素和维度,可能会极具误导性。因为这种类比并不是基于相似性的类比,而是基于差异相似性的类比。对大多数国际关系研究者来说,核武器和核技术是个例,是"几乎无关乎国家使用策略,技术本身就会对国际政治产生深远影响的小概率事件"。[①] 而人工智能与核技术同被认为是颠覆性技术并予以比较,看似提取的共性特征是两者都会对人类社会产生革命性的影响,本质体现的是两者相较各自时代其他技术的差异性,即两者皆是所处时代中能够产生广泛变革性影响技术,都是同时代技术中的"异类"。这种基于差异相似性的类比可能是所有类比中最具误导性一类,一方面,这种类比会使我们过分强调人工智能与其同时代其他技术的差异;另一方面,这一类比又试图从不同时代的"异类"中提取共性,最终造成错误判断、误导决策,甚至恶化国际形势。

第一,人工智能技术的核类比会加强人工智能是一种颠覆性技术的认知,赋予其新的政治意涵和社会意义。当前,世界范围默认人工智能具有革命性影响,并成功推动人工智能成为各国发展的重心,却少有人反思为什么人工智能会成为这个时代可能影响国际秩序的"颠覆性科技"。回溯"人工智能是颠覆性技术"这一叙事的形成过程,与核技术的类比功不可没。新技术本身并不具有任何意涵,是人们围绕技术进行的叙事构建赋予技术的社会意义。[②] 通过将人工智能与核技术类比,向公众传递的信息是人工智能具有同等的变革性,会对国际秩序产生同样巨大的不确定性和挑战。在这种以类比为代表的叙事构建中,人工智能作为颠覆性技术的共识逐渐形成,并暗含更多政治意义。

第二,人工智能技术的核类比会导致人工智能议题迅速安全化,从而不利于技术的长远发展。此前围绕 Open AI 和深度求索(DeepSeek)是否应该开源

---

① Michael C. Horowitz, "Do Emerging Military Technologies Matter for International Politics?" p. 386; Robert Jervis, *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon*, Ithaca: Cornell University Press, 1989.

② Nike Retzmann, "'Winning the Technology Competition': Narratives, Power Comparisons and the US-China AI Race," in Thomas Müller, et al., eds, *Comparisons in Global Security Politics: Representing and Ordering the World*, Bristol: Bristol University Press, 2024, pp. 238-239.

的辩论便体现了这一点。对于技术本身，开源代码和模型允许全球开发者共同参与改进和优化，从而加速技术的迭代和创新。但是，正如核武器的研发历程所表明，越是涉及国家安全的技术越会走向有限参与者的秘密研发。[①] 因此，开源反对者认为技术具有极强的国家安全属性，开源可能会使对手更容易获取关键技术，威胁自身安全。[②] 此外，人工智能不像核技术具有特定的应用方式，其应用场景非常广泛。且人工智能技术也并非最终呈现的实体，而是更像可用于广泛品类的产品、服务和系统中的重要组成部分和要素。因此，使用防核扩散这种视角来指导人工智能的管控是不合适的。人工智能管控的对象不应该是其最终形态和功能，而应该是可造成技术扩散的重要组成部分和要素。这样才能既保留人工智能创新的动力，又对其无序发展形成有力监管。[③]

第三，人工智能技术的核类比会带来对人工智能技术的压倒性关注，影响国家长远的发展布局。压倒性关注意味着与其他技术领域相比，人工智能获得了更多的资源配置，尽管当前相较人工智能技术的重要性来说无可厚非，但是，也要警惕技术领域的虹吸效应，避免错过下一轮技术革新的战略机遇期。人工智能技术的兴起可以追溯至 20 世纪 50、60 年代，但对人工智能颠覆性技术的判定却是在最近几年才形成的社会广泛共识。在 2022 年前，人们对于人工智能的变革性影响仍是持怀疑态度的，甚至有学者表示，“人工智能武器系统并不能如核武器那样具有影响全球权力均衡的可能”。[④] 正如无法提前预判人工智能的重要性一样，人类也无法准确预判下一次颠覆性技术，若一味只关注当前，可能就会错过下一个颠覆性技术的发展，毕竟很多技术的突破需要相

① Michael C. Horowitz, “AI and the Diffusion of Global Power,” in Centre for International Governance Innovation, ed., *Modern Conflict and Artificial Intelligence*, 2020, p. 35.

② Greg Allen and Taniel Chan, *Artificial Intelligence and National Security*, Vol.132, Cambridge: Belfer Center for Science and International Affairs, 2017.

③ Daniel Araya and Rodrigo Nieto-Gómez, “Renewing Multilateral Governance in the Age of AI,” in Centre for International Governance Innovation, ed., *Modern Conflict and Artificial Intelligence*, 2020, p. 10.

④ Amandeep Singh Gill, “A New Arms Race and Global Stability,” in Centre for International Governance Innovation, ed., *Modern Conflict and Artificial Intelligence*, 2020, p. 17.

当长的时间的积累和投入。

(二) 冷战误区

除了错误提取相似特征以外,对类比推理的另一大质疑就在于类比过程中对比较对象的过度简化。作为一种认知捷径,历史类比倾向于将当前的复杂问题简化为类似的历史问题,或是用已有概念框架来思考新现象,而提取相似性的过程却极易忽视新事件的多维性和新因素,造成误导性理解。[①] 重要的历史事件往往具有极为独特的背景和影响因素,很难复制与类比。[②] 此外,历史并不总是具有明确的因果关系,且事件的结果可能是多因素交互的结果,错误类比还可能源自对这些复杂性的忽略。[③]

具体到人工智能的核技术类比问题上,其中一个重要的支撑性观点是认为两种技术迅猛发展时所处的国际局势具有高度相似性,都是在两个大国正在进行战略竞争的特殊时期。因此,可以通过美苏核竞赛的历史经验来理解中美在人工智能领域的竞争动态,同时,中美对于人工智能安全风险的管控也可借鉴美苏两国在冷战期间为防止核扩散所做出的协调与努力。[④] 然而,这种类比将中美战略竞争简化为美苏两极对抗,既无视前后两对双边关系的差异,也忽略了其他行为体在当前所起到的重要作用,很容易产生误导性判断和政策建议。

现有研究表明,将中美战略竞争类比冷战美苏对抗是错误的:冷战呈现的是在两极格局下美苏对全球霸权的争夺。因此,美苏之间的竞争更多是零和博弈,两国及其领导的阵营也几乎没有经济上的相互依赖;而当今世界呈现多

① Robert Jervis, *Perception and Misperception in International Politics*.

② Margaret MacMillan, *The Uses and Abuses of History*, Profile Books, 2010.

③ Yuen Foong Khong, *Analogies at War: Korea, Munich, Dien Bien Phu, and the Vietnam Decisions of 1965*.

④ 朱荣生等:《人工智能的国际安全挑战及其治理》,《中国科技论坛》2023 年第 3 期,第 164 页。

极格局，中美在经济、科技和供应链上高度相互依赖，且双方竞争也并非零和。[①] 因此，即便两个大国在颠覆性技术上形成竞争，中美之间围绕人工智能的竞争也不能简单类比美苏之间围绕核技术的博弈。但通过将人工智能类比核技术，逐步营造中美必须“赢得科技竞争”的叙事框架[②]，会进一步强化中美在人工智能上的对抗性，从技术发展之争到国际规则制定权之争都可能被视为决定两国未来乃至全球秩序的关键，这不仅不利于推进中美两国在人工智能领域进行合作，而且还有可能推动中美走向人工智能军备竞赛。

除中美竞争动态不同于冷战期间的美苏博弈外，人工智能和核技术治理模式差异也增强这种类比的误导性，换言之，美苏围绕核扩散进行的合作并不能提供更多的借鉴意义。[③] 首先，人们学习历史并非只为了复刻过去的成功政策，更在于避免过去的错误。[④] 回顾历史，冷战时期各国围绕核风险的治理和管控其实受到“古巴导弹危机”等历史事件的推动。但是，当前人工智能技术发展并未遭遇大危机事件，却仍将人工智能类比核技术（特别是核武器）是试图以史为鉴，警示国际社会如若不对其加以管控，会导致技术扩散并产生危及国际秩序的负面效应。但值得注意的是，不同于核武器相对有限的使用场景（主要是战争和威慑），人工智能的应用场景极为广泛，涵盖经济、社会、军事等多个领域。因此，将人工智能与核武器类比可能会过度关注人工智能的军事应用及其风险，造成群体性恐慌，忽略其在其他领域的作用和影响。

其次，即便对人工智能技术扩散担忧是正确的，两者的技术特性也导致其

---

① Yuen Foong Khong, “The US, China, and the Cold War Analogy,” *China International Strategy Review*, Vol.1, No.2, 2019, pp. 223-237; Jiwu Yin, “The Cold War Analogy’s Misrepresentation of Essence of US-China Strategic Competition,” *China International Strategy Review*, Vol.2, No.2, 2020, pp. 257-269; 左希迎：《百年变局下的大国战争风险及其管控》，《外交评论》（外交学院学报）2024 年第 5 期，第 66—67 页。

② Nike Retzmann, “‘Winning the Technology Competition’: Narratives, Power Comparisons and the US-China AI Race,” pp. 237-256.

③ Daniel Araya and Rodrigo Nieto-Gómez, “Renewing Multilateral Governance in the Age of AI,” pp. 7-12.

④ Rose McDermott, *Political Psychology in International Relations*, p. 63.

管控路径并不具可比性。核技术治理主要集中于对原材料和设施的控制,而人工智能治理需要涵盖从算法设计、数据管理到集成电路等更加多样化的领域。[①] 以核武器为例,其开发需要稀缺的原材料(如铀、钚)和复杂的基础设施,技术门槛高且易于监控。特别是在冷战两大阵营对立且相对独立的情况下,核扩散的防控比较有效。但人工智能的开发依赖于算法、数据和计算能力,技术门槛相对较低且易于扩散,难以通过传统手段(如出口管制)完全控制。更加上当前全球化时代各国高度互通互联,人工智能技术的管控难度明显增大。因此,冷战时期对于核扩散的管控策略很难有效迁移至人工智能的管控,强行借鉴只会产生错误的政策建议。例如,2025 年 1 月,美国发布全球首个《人工智能扩散出口管制框架》,将中国、俄罗斯和朝鲜等国家列为"全面禁运国",即将所有高端人工智能芯片、模型权重和云计算服务全面封锁,禁止对这些国家出口。这份文件明显带有冷战色彩,试图通过阵营化的传统军控手段来建立"科技封锁圈",孤立核心竞争对手。如上所述,这一政策既忽略了人工智能技术本身的特性,也无视当前国际形势的变化,而科技封锁历史也表明,对手在封锁压力下会加速国产替代并实现技术自主,使该政策产生适得其反的结果。"深度求索"即是典型案例,它虽不一定代表中国突破了美国的技术封锁,但是"深度求索"的开源模式本身就意味着封锁的无效性,也再次证明人工智能领域存在丰富且庞大的国际合作空间。

再次,核技术治理主要依赖美苏两国主导国家间的条约和协议,如《不扩散核武器条约》,体现的是传统的基于权力和利益的风险治理合作。但是,人工智能的治理更需要多元主体参与的更灵活、多层次的框架,并要涵盖技术标准、伦理规范等多个方面,国际合作维度显著增高,体现了所需的国际治理体制的复杂性。[②] 因此,在此情况下再进行核类比,讨论人工智能的扩散风险和

① Allan Dafoe, "AI Governance: A Research Agenda," Governance of AI Program, Future of Humanity Institute, University of Oxford, 2018, pp. 48-49.

② 庞珣:《反思与重构:全球安全和全球治理的风险政治理论及方法》,《国际政治研究》2024 年第 2 期,第 31—35 页。

监管机制，未来人工智能风险管控将会强化冷战的对抗色彩和大国主导观念，严重忽视国际协作及国际社会其他国家和组织在此过程中的作用。同时，还容易引发有关《不扩散核武器条约》对后发国家核试验限制的联想，增强各国利用当前监管机制尚未成熟的窗口期抢先发展的动机，加剧资源浪费和军备竞赛。

正如现有研究所指出的，连接历史类比与政策偏好的因果路径有两条：一条是人们真正相信历史事件与当前情境之间的相似性，相信自己从历史中学到的经验和教训；另一条则是人们有意识地选择类比的历史来支持并推进自己的政策偏好，这也被称为历史的策略性使用。[①] 两者体现的恰恰是相反的因果逻辑，但却同时发生在人工智能与核技术的类比上。因此，除了无意识的错误类比导致的认知误区外，利益攸关方的主动性选择也造就了人工智能的核类比陷阱。具体来说，历史类比具有主观选择性，选择哪个历史事件和已有概念作为类比的对象通常是主观选择的结果。决策者倾向于选择支持自己观点的类比对象，而刻意忽略其他可能性。而技术发展本身就是围绕技术形态和随之而来的社会形态所展开的政治斗争和博弈的结果[②]，相应的历史类比就是推进自身政治议程的重要手段。例如，美国鹰派就试图拼命鼓吹冷战类比，以此进行国内政治动员并推动对华强硬。

三、认知重构、多元治理与走出类比误区

无论是出于认知局限导致的无意识错误类比，还是因为主观偏好导致的有意识选择性类比，都会促使全社会在核技术类比框架下形成对人工智能及其风险的错误认知，甚至产生不恰当的政策。为了更准确地理解人工智能的国际安全风险和管控，知识界应该达成对人工智能更科学认识的共识，并向大

① Jack S. Levy and William R. Thompson, *Causes of War*, p. 147.

② Daniel R. McCarthy, *Power, Information Technology, and International Relations Theory: The Power and Politics of US Foreign Policy and the Internet*, Basingstoke: Palgrave Macmillan, 2015, p. 33.

众普及正确的风险认知,而政策制定者则应通过公开、透明的决策过程确保相关政策的合理性,共同携手走出人工智能的核技术类比误区。

首先,人工智能的核技术类比构建了人工智能风险的特定错误认知,需要通过回归技术本身进行解构,并重建新的风险共识。建构主义学者认为技术本身是价值无涉的,其社会意义和重要性并非其技术属性决定的,而是政治斗争和叙事构建的结果。[①] 而风险也"并非关乎危险的事实,而是关于风险如何被政治化"。[②] "风险问题是政治问题而非技术问题,围绕风险展开的活动不仅是科技或官僚管理活动,更重要的是政治活动"。[③] 这一系列观点很好地总结了现有关于人工智能的国际安全风险的讨论。通过与核技术类比,人工智能技术逐渐被刻画为一种国家实力的来源,成为中美竞争的核心场域。作为一种新技术,人工智能对国际安全的冲击尚不明确,与核技术类比来理解这种不确定性,可能进一步引发从政策制定者、学者到普通民众的显著焦虑和不安。[④] 对于人工智能风险的恐惧既是该议题政治化的动力,也是其政治化的结果。面对日益增长的群体焦虑与恐惧,学者们反而需要回归技术本身,重新思考人工智能技术的真实特征及风险,形成知识界新的共识。

具体来说,解构人工智能核技术类比的过程其实是去安全化、去政治化的过程,也是重新理解不确定性的过程。人工智能技术的一个核心特点就是在技术、应用、环境等多个维度具有较高程度的不确定性,而这也是人工智能区别于其他技术的关键。在技术层面上,人工智能技术存在多个层次的不确定性:第一,实现人工智能的科学原理是不清晰的,无论是大脑的工作机制,还是神经网络的学习机制,都存在较多未知领域并因此带来"不可解释"的黑箱性

① Nike Retzmann, "'Winning the Technology Competition': Narratives, Power Comparisons and the US-China AI Race," p. 238.

② Mary Douglas, *Risk and Blame: Essays in Cultural Theory*, London: Routledge, 1992, p. 29.

③ 庞珣:《反思与重构:全球安全和全球治理的风险政治理论及方法》,《国际政治研究》2024 年第 2 期,第 14 页。

④ 陈琪、朱荣生:《不确定性:为何担心人工智能冲击国际安全》,《国际战略与安全研究报告》,2019 年第 8 期,第 1 页。

质;第二,影响人工智能技术性能的因素或机制是不清楚的,模型或参数权重的调整更多基于经验而非明确的因果关系;第三,实现人工智能的技术路径是多元化的,并不存在唯一或最优的技术路线,不同技术路线之间的竞争演化是常态。在应用层面,人工智能技术具有一般通用技术特性,即应用范围广、应用类型多。人工智能技术可以广泛应用于多领域、多场景,而在这些领域、场景中,其功能也是多样化的。例如,生成式人工智能可以广泛生成文字、图片、视频等多种媒体类型,也可以被广泛应用于客服机器人、个人助手、创意生产、代码生成、个性化教育等。同时,正因为这种功能上的多样性和范围上的广泛性,人工智能的应用场景或价值往往难以提前预知,这也同样带来应用层面的不确定性。在环境层面,人工智能研发和应用环境要求仍然存在不确定性,这不仅体现为在特定时间节点环境条件的不明确,也体现为伴随时间演化人工智能研发应用环境要求的动态变化方面。例如,在大型语言模型兴起之前,算力并非人工智能技术研发和应用的瓶颈资源,但大型语言模型作为人工智能一种技术路线的兴起则改变了这一环境条件,并且这种改变往往难以提前预知。

人们对于人工智能风险的担忧表面上是核技术类比为代表的技术政治意涵构建的结果,更深层次是因为这种构建符合进化心理:不确定性自然导致恐惧。[①] 然而,人工智能本身具有较高程度的不确定性并不意味着它对国际安全就会产生绝对意义上的负面冲击,“在国际安全事务中,不确定性是常态,确定性反而是稀缺”。[②] 更何况人工智能技术在其发展过程中也蕴含着稳定性因素,可以抵消可见的安全风险。例如,人工智能是一种“基础设施性技术”,而不仅仅是“武器性技术”,它具有在经济、社会和文化领域等多维度影响。因此,未来人工智能风险应对的关键不是如何在不断强化的恐慌中弥补所谓的风险漏洞,而是学会如何与这种不确定性共存。知识界对于人工智能技术风

① Shiping Tang, “Fear in International Politics: Two Positions,” *International Studies Review*, Vol.10, No.3, 2008, pp.451-471.

② 陈琪、朱荣生:《不确定性:为何担心人工智能冲击国际安全》,第 9 页。

险的再思考会进一步推动新的风险认知乃至共识的出现,扭转人工智能核技术类比造成的风险夸大与过度恐惧。

其次,在明确人工智能核技术类比的误区后,及时的风险沟通是必要的。知识界对人工智能风险的再认识需要及时向大众普及,消除大众因错误类比而导致的过度担忧和错误的风险认知。现有研究已经证明专家学者在推动民众风险认知中起到巨大的作用。例如,在人工智能领域,真正引发民众对人工智能风险担忧的并非现实生活中真实的灾难性事件,而是专家传递的信息。[①] 虽然安抚民众使其减少焦虑要远比利用恐惧心理引发公众焦虑困难,但这却刻不容缓。若放任核技术类比形成的对抗性和焦虑性的民众心理,其可能会逐渐产生极大的观众成本,推动更加激进的外交政策。

为消除大众的过分忧虑,在与其风险沟通过程中,选取更恰当的、本身不产生恐慌的类比,可能是一种有效方法。人工智能技术具有分散性和多样性的特点,在此维度上与互联网或生物技术等去中心化技术更为接近。此外,人工智能是赋能技术,更像可以助力其他技术发展的核心要素,这一点上又与阿斯巴甜和聚乙烯相似。[②] 而在人工智能的通用性上,与电[③]和蒸汽机[④]的类比更加形象。鉴于治理标准的多样性和复杂性,人工智能治理也可以参考食品安全治理。[⑤] 值得注意的是,选取不同的类比并不试图人为制造更多的类比陷阱,这一举动恰恰表明人工智能本身的独特性,我们无法完全依赖历史上的某单一技术来完全类比人工智能。此外,细化风险传递框架也是一种有效手段,在风险沟通中区分人工智能的短期和长期风险,避免将所有讨论都集中于“超

① Hugo Neri and Fabio Cozman, “The Role of Experts in the Public Perception of Risk of Artificial Intelligence,” *AI & Society*, Vol.35, No.3, 2020, pp. 663-673.

② Daniel Araya and Rodrigo Nieto-Gómez, “Renewing Multilateral Governance in the Age of AI,” p. 10.

③ 傅莹:《人工智能对国际关系的影响初析》,第 13 页。

④ Michael C. Horowitz, “AI and the Diffusion of Global Power,” p. 35.

⑤ Daniel Araya and Rodrigo Nieto-Gómez, “Renewing Multilateral Governance in the Age of AI,” p. 10.

级智能”或末日情景，而忽视了当下更现实的问题。[①]

最后，跨越人工智能的核技术类比陷阱还需要政策制定者的共同努力，人工智能不确定性程度高等独特特征意味着无法线性地使用历史思维来理解其变化中的风险，也无法通过照搬过去的经验来实现人工智能的有效治理。在治理层面走出核技术类比的误区最重要的就是要避免恐惧驱动的政策制定。[②]一方面，可以通过公开、透明的决策过程，以及平衡风险与机会的叙述，确保政策制定基于理性而非过度的情绪化反应；另一方面，在人工智能的全球治理方面，可以鼓励众多利益相关方一起参与政策协商和制定，推动跨学科、多利益相关者的协作，确保人工智能治理能够反映技术的多样性和复杂性。[③]

结　　语

人类对于新技术的理解常常伴随着认知悖论，一方面，新技术是打破常规的新现象；另一方面，人们又不能完全脱离现有的知识体系去描绘和形成对它的认知。[④] 因此，当人工智能技术的跨越式发展为人类社会带来了巨大的不确定性，需要更多地对技术本身及其风险进行正确理解时，以核技术类比为代表的历史类比成为学界和政策界广泛采用的认知框架。在国际关系领域，人工智能与核技术的类比虽有助于引发对人工智能安全风险及管控路径的深刻反思，但其简单化和误导性大大限制了这一类比的实际意义。其中最具代表性的错误认识是颠覆性技术误区和冷战误区，前者在颠覆性技术框架下强调人工智能与核技术的相似性，实则是基于差异相似性的类比，忽略了两者从技术

① Stephen Cave and Seán S. ÓhÉigeartaigh, “Bridging Near-and Long-Term Concerns about AI,” *Nature Machine Intelligence*, Vol.1, No.1, 2019, pp. 5-6.

② 苏若林：《外交决策中的风险偏好：概念与形成》，《国际政治科学》2021 年第 4 期，第 96—99 页。

③ Stephen Cave and Seán S. ÓhÉigeartaigh, “Bridging Near-and Long-Term Concerns about AI,” *Nature Machine Intelligence*, Vol.1, No.1, 2019, p. 6.

④ Nike Retzmann, “‘Winning the Technology Competition’: Narratives, Power Comparisons and the US-China AI Race,” p. 242；叶淑兰、李孟婷：《全球人工智能治理：进展、困境与前景》，《国际问题研究》2024 年第 5 期，第 116—117 页。

特性到实际影响的差异;而后者在冷战框架下思考中美人工智能竞争与合作,过分简化了国际环境,无视其他行为体的作用,不利于准确把握中美围绕人工智能的竞争动态,也不利于推动产生更行之有效的治理路径。

因此,为更有效地理解和管控人工智能带来的国际安全风险,社会各界需要集体协作超越这一类比陷阱,采用更加科学而现实的分析框架,结合其独特的分散性、多样性和全球性影响,发展出切实可行的治理模式。事实上,针对人工智能不确定性程度高的特征的治理改革已经在各国普遍展开,敏捷治理、适应性治理、实验主义治理等新的治理理论也在不断丰富人们对于安全监管制度的传统理解。在此意义上,部分领域的人工智能监管可被认为已经走出了核技术管控单一视角的政策局限。未来,需要将当前的丰富实践通过利益相关方多层次对话转化为共识性的全球治理机制,这也正是人工智能国际安全风险管控的希望所在。