《国际政治研究》(双月刊)2025 年 第 3 期

超越二元论：跨国人工智能竞合关系的风险理论

祁昊天

内容提要 跨国人工智能关系存在复杂的竞争与合作交织状态，无法仅用简单的“竞争—合作”二元两分框架来理解。引入“风险水平”和“风险可识别度”这两种风险特征属性，有助于对跨国人工智能关系进行更加细致地分类与分析。其中，高风险往往涉及军事应用部署、关键基础设施和敏感数据处理，容易带来重大和直接的损失；而较低风险则主要指商业应用、社会服务等领域，风险相对可控。风险可识别度高的情况通常与技术问题相关，而低可识别度则涉及复杂交互和长期间接影响。基于比较不同技术范畴和应用场景下的人工智能风险特征，跨国人工智能关系可分为监管防控、协同发展、竞赛对抗和谨慎隔离四种类型。此外，不同行为主体对风险水平的认知差异又会导致合作意愿与安全措施的不对称，并加大治理难度

* 祁昊天：北京大学国际关系学院副教授。(邮编：100871)

** 感谢《国际政治研究》匿名评审专家的意见和建议，文责自负。

和摩擦概率。通过对军事现代化竞争、战略核领域风控、绿色产业合作、大语言模型竞赛和算力能耗治理等典型案例的讨论,该风险理论框架的有效性得到了支撑。

关键词 非传统安全 人工智能 跨国人工智能关系 竞争与合作 风险理论 风险水平 风险可识别度

目前,人工智能(Artificial Intelligence, AI)在技术和应用层面仍存在很大不确定性,但它已在不同程度上影响着政治、军事、经济、外交等各个领域,跨国人工智能关系成为世界政治的新维度。本文旨在探讨跨国人工智能关系为何存在不同样貌,揭示不同技术范畴和互动场景中的竞合交织状态,并为跨国关系治理提供结合技术与政治逻辑的参考依据。

本文将"跨国人工智能关系"定义为国家与非国家行为体(如大型科技公司)、超国家行为体(如欧盟)等具有人工智能能力与政策基础的行为主体之间,在涉及人工智能技术发展与应用部署时的关系状态。人工智能发展水平往往体现了综合国力或地区整体实力。根据不同机构关于全球人工智能发展情况的梳理总结,目前,在国家和非国家行为体方面均总体呈现美国整体领先、中国局部占优、多国多地区跟随的格局。[①] 针对这一格局,传统上对于跨国人工智能关系的讨论通常局限于"竞争—合作"这种二元光谱。一方面,人工智能的发展本身构成了跨国竞争的新领域,并引发战略互疑,诱发人工智能"军备竞赛",进而影响地区甚至全球力量对比与战略稳定;另一方面,人工智能的快速发展也能够开辟新的合作领域,如科技研发、场景应用、产业链合作

① 中国科学技术信息研究所、北京大学:《2023 全球人工智能创新指数报告》,2024 世界人工智能大会科学前沿主论坛,2024 年 7 月 4 日; Human-Centered Artificial Intelligence of Stanford University, "Artificial Intelligence Index Report 2024," https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf, 2024-11-20; Human-Centered Artificial Intelligence of Stanford University, "Artificial Intelligence Index Report 2025," https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf, 2025-06-12.

与风险治理等。

这种二元两分的视角为理解跨国人工智能关系提供了基本参照系，却无法立体地解释竞合差异及竞合并存的复杂状态。例如，盟友间存在合作与矛盾分布不均衡的状态。[①] 又如，中美在人工智能发展及其应用部署上存在激烈竞争，美国频频对华设限并试图强化单向优势[②]，但同时，两国高层也对人工智能的安全与治理问题表达了基本共识。[③] 单一维度上的竞争或合作两分法很难对这种立体性与复杂性给出简洁有力的解释。基于此，本文将尝试突破二元视角来回答：跨国人工智能关系为什么及在怎样的条件下存在何种具体的竞争与合作差异？

考虑到目前应用的人工智能主要以人工神经网络（Artificial Neural Network，ANN）和数据驱动（Data-Driven）为基本技术路径，本文将从其固有的风险特征属性切入对跨国人工智能关系进行探讨。通过区分和比较风险水平高低程度与风险可识别度这两种属性在具体技术范畴、应用领域、部署阶段和状态下的影响，结合行为主体的风险认知倾向，本文提出一种解析跨国人工智能关系的分析框架。该框架有助于超越简单化的“竞争—合作”两分法，对跨国人工智能关系进行更为系统、立体和动态的解释。

本文采用分类法构建分析框架，通过划分风险水平与可识别度概括出四种关系类型，即消极合作性的监管防控、积极合作性的协同发展、冲突性的竞赛对抗以及具有“脱钩”倾向的谨慎隔离。在此基础上，本文针对每一种类型

① 例如，美国在人工智能军事应用方面盟伴合作中的“小圈子”倾向非常明显，与“五眼联盟”国家之间的合作紧密程度远高于其他盟友。而一些盟伴国家之间在总体不存在显著矛盾冲突的前提下，依然暗流涌动，如美与日、荷、韩在光刻机与芯片问题上。

② 如拜登政府卸任前于 2025 年 1 月 13 日发布的关于限制人工智能与芯片等出口的进一步举措，参见“Biden-Harris Administration Announces Regulatory Framework for the Responsible Diffusion of Advanced Artificial Intelligence Technology，” https://www.bis.gov/sites/default/files/press-release-uploads/2025-01/AI%20embargoed%20press%20release.pdf，2025-02-10。

③ 2023 年 11 月 15 日，习近平主席与拜登总统在旧金山会晤，同意建立人工智能政府间对话，参见《习近平同美国总统拜登举行中美元首会晤》，《人民日报》2023 年 11 月 17 日，第 1 版；当地时间 2024 年 11 月 16 日，习近平主席与拜登总统在秘鲁利马会晤时再次达成人工智能风险与安全的相关共识，参见《习近平同美国总统拜登在利马举行会晤》，《人民日报》2024 年 11 月 18 日，第 1 版；当地时间 2024 年 5 月 14 日，中美在日内瓦举行首次一轨人工智能对话，参见“中美举行人工智能政府间对话首次会议”，https://www.chinanews.com.cn/gn/2024/05-15/10216990.shtml，2025-01-11。

的风险属性特征,选取典型案例进行说明和分析,以验证分析框架的有效性并拓展政策参考的实用性。

跨国人工智能关系涉及复杂的政治、地缘、安全、决策者判断等因素,本文并不认为风险视角可以独自完整地呈现这种关系全貌,这也并非本文目的所在。引入风险视角是为了提供一种可能,使我们在结合技术逻辑与政治逻辑的基础上,刻画跨国人工智能关系的立体和多样性,并更加清晰地界定技术风险与政策行为之间的关系。

一、跨国人工智能关系的复杂性

跨国人工智能关系中的竞争与合作存在交织和混杂的状态,要比二元论的竞合更加复杂。但目前,学术和政策界尚未提出一个对不同人工智能关系状态进行体系性差异比较和分析的框架。为了在分析简洁性和现实复杂性之间取得平衡,本文尝试从风险特征切入,对关系类型进行区分,使之更具有学理上的可验证性与政策上的参考性。

(一) 人工智能竞争与合作

二元论光谱的一端是人工智能竞争的视角。人工智能被认为是未来跨国互动中影响竞争力的决定性因素,对竞争地位、相对优势与技术主权的争取和维护必然导致竞争的加剧,在技术研发、数据资源、人才储备和市场占有等方面的竞争趋势又会导致跨国"脱钩"倾向。[①] 例如,美国强调自身领先地位,警

① Michael Horowitz, et al., "Strategic Competition in an Era of Artificial Intelligence," Center for a New American Security, July 25, 2018, https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence, 2024-11-15; Ryan Sullivan, "The US, China, and Artificial Intelligence Competition Factors," *China Aerospace Studies Institute*, 2021, https://www.airuniversity.af.edu/Portals/10/CASI/documents/Research/Cyber/2021-10-04%20US%20China%20AI%20Competition%20Factors.pdf, 2024-11-15; Nurmukhammad Y. Samijonov, "The Race of Artificial Intelligence for Supremacy," *International Journal of History and Political Sciences*, Vol.3, No.9, 2023, pp.16-21; Maria Papageorgiou, et al., "China as a Threat and Balancing Behavior in the Realm of Emerging Technologies," *Chinese Political Science Review*, Vol.9, No.4, 2024, pp.441-482.

惕其他国家在经济和军事等领域通过人工智能技术的发展和应用危害美国国家安全[①];欧盟也要求加强自身人工智能发展独立性,避免产生技术依赖并保障欧盟的数字主权。[②]

由于兼具通用技术和军民两用技术的特征,人工智能不仅自身发展成为竞争领域,其应用和部署也会在多种场景(特别是安全关系)产生潜在的负面刺激。联合国很早便对人工智能的武器化、伴随而来的军备竞赛以及这种竞赛对战略稳定的影响发出了警告。[③] 而即便是在针对人工智能发展和应用的治理领域,由于利益、身份认同、意识形态、技术主导权等方面的分歧及相关国际机制的不足,同样存在全球性竞争。[④] 美国拒绝在 2025 年 2 月巴黎“人工智能行动峰会”联合宣言中签字及副总统万斯(J. D. Vance)的发言,便体现了狭隘政治立

① 参见美国白宫科技政策办公室 2025 年 2 月发出关于人工智能行动计划建议的信息请求(Request for Information, RFI)以及业界典型回应,The White House, “Public Comment Invited on Artificial Intelligence Action Plan,” https://www.whitehouse.gov/briefings-statements/2025/02/public-comment-invited-on-artificial-intelligence-action-plan/, 2025-02-27; OpenAI, “OpenAI's Proposals for the U.S. AI Action Plan,” https://openai.com/global-affairs/openai-proposals-for-the-us-ai-action-plan/, 2025-03-15; 白宫发出关于维护美国人工智能领导地位并移除现任政府认为无益政策的行政令,The White House, “Removing Barriers to American Leadership in Artificial Intelligence,” https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/, 2025-02-03; 美国人工智能国家安全委员会给前一届美国政府的意见,U.S. National Security Commission on Artificial Intelligence, “Final Report,” March 2021, https://reports.nscai.gov/final-report/, 2024-11-20.

② European Commission, “White Paper on Artificial Intelligence: A European Approach to Excellence and Trust,” February 19, 2020, https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf, 2025-10-11.

③ United Nations Institute for Disarmament Research (UNIDIR), “The Militarization of Artificial Intelligence and Its Impact on Strategic Stability,” August 2019, https://digitallibrary.un.org/record/3972613/files/Militarization-ArtificialIntelligence.pdf, 2024-11-23.

④ Huw Roberts, et al., “Global AI Governance: Barriers and Pathways Forward,” *International Affairs*, Vol.100, No.3, 2024, pp.1275-1286; Lewin Schimitt, “Mapping Global AI Governance: A Nascent Regime in a Fragmented Landscape,” *AI and Ethics*, Vol.2, No.2, 2022, pp.303-314; Sabine Mokry and Julia Gurol, “Competing Ambitions Regarding the Global Governance of Artificial Intelligence: China the US, and the EU,” *Global Policy*, Vol.15, No.5, 2024, pp.955-968; Emmie Hine and Luciano Floridi, “Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI policies,” *AI & Society*, Vol.39, No.1, 2024, pp.257-278.

场、意识形态阵营化和管制偏好差异对人工智能治理合作带来的阻力。①

“竞争—合作”光谱的另一端是人工智能发展与治理合作的视角,主要集中于推动技术全生命周期标准制定、伦理框架完善以及应对全球性挑战等。例如,在2024年博鳌亚洲论坛“AIGC改变世界”分论坛上,与会专家和企业高管均认为,人工智能发展不是零和博弈,任何国家都无法独自引领其发展,开展全球合作才能确保人工智能安全和造福人类,非赢即输的观念是错误的。②

在这种视角下,人工智能的快速发展为跨国治理合作提供了新的契机。虽然不同国家和组织在治理模式与目标上可能存在分歧,但关于人工智能安全治理和弥合全球数字鸿沟的跨国多边合作工作均在稳步推进中。例如,中国始终坚持人工智能技术以人为本和向善的原则,积极推进广泛共识基础上的全球人工智能治理框架和标准规范。2023年10月,中国提出《全球人工智能治理倡议》,强调发展人工智能应遵守适用的国际法,各国尤其是大国对重点领域如军事领域研发和使用人工智能应该采取慎重负责的态度,要确保人工智能始终处于人类控制之下。③ 2024年3月,联合国大会一致通过美国牵头提出的人工智能决议,就安全、可靠和值得信赖的人工智能系统达成了原则共识。④ 2024年7月,联大通过中国提交的加强人工智能能力建设的国际合作决议。⑤这些行动不仅是国际社会推进人工智能发展和治理合作的积极信号,也

---

① 美英两国拒绝签字,相关报道参见 https://www.theguardian.com/technology/2025/feb/11/us-uk-paris-ai-summit-artificial-intelligence-declaration,2025-3-20;万斯发言参见 https://www.presidency.ucsb.edu/documents/remarks-the-vice-president-the-artificial-intelligence-action-summit-paris-france,2025-3-20。

② 《博鳌亚洲论坛“AIGC改变世界”分论坛举行》,新华网2024年3月27日,http://www.xinhuanet.com/photo/20240327/974822835a2c4bdba35665c123a43f12/c.html,2025-01-05。

③ 2023年10月18日,习近平主席在第三届“一带一路”国际合作高峰论坛开幕式主旨演讲中提出该倡议,参见《习近平出席第三届“一带一路”国际合作高峰论坛开幕式并发表主旨演讲》,《人民日报》2023年10月19日,第1版;《全球人工智能治理倡议》全文参见 https://www.mfa.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml,2025-02-05。

④ Antony J. Blinken, “Consensus Adoption of U.S.-Led Resolution on Artificial Intelligence by the United Nations General Assembly,” March 21, 2024, https://www.state.gov/consensus-adoption-of-u-s-led-resolution-on-artificial-intelligence-by-the-united-nations-general-assembly/, 2024-12-01.

⑤ 《联大通过中国提出的加强人工智能能力建设国际合作决议》,新华社2024年7月2日,http://www.xinhuanet.com/20240702/7e7b6668eadf4738be49cb9b1769a654/c.html,2024-12-01。

有利于国际战略稳定。

（二）人工智能关系的风险视角

“竞争—合作”光谱为理解跨国人工智能关系提供了基准参照，但这种单一维度、二元化视角却很难呈现这种关系的复杂性、差异性与动态性。例如，中美之间在高敏感的人工智能军事应用领域存在直接竞争，但同时两国又表现出积极的合作信号。习近平主席与拜登总统在 2023 和 2024 年的两次亚太经合组织会晤中均强调了人工智能的治理合作问题，他们在 2024 年的会晤中更是强调了军事层面、与核武器有关的人工智能部署安全问题。

本文将通过风险视角构建跨国人工智能关系的比较与分析框架。从风险切入存在两方面的合理性和必要性：一方面，基于当前人工智能主要的发展和应用方向，风险在其技术范畴和部署方面都是基本性特征；另一方面，不同行为体均将风险作为人工智能发展战略相关路线规划的核心关切内容。

当前被广为应用的人工智能主要基于神经网络并由数据驱动。[①] 它依赖多层神经元模拟与海量数据为输入，通过统计学习方法训练模型，尤其是在深度学习领域。此类人工智能具有算法黑箱（Black Box）、解释性较差、复杂性高、输出非线性和应用情景多样等特征。这些特征使得人工智能在不同应用场景均表现出不同种类和程度的风险性。2025 年 4 月 25 日，习近平总书记在中共中央政治局第二十次集体学习时强调，“人工智能带来前所未有发展机遇，也带来前所未遇风险挑战”。[②]

既有风险研究为分析跨国人工智能关系提供了基础。体系化的风险研究

---

① 相关基本介绍参见邱锡鹏：《神经网络与深度学习》，北京：机械工业出版社 2021 年版；Mohamad H. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, 1995；以及国际组织在 AI 发展与治理领域强调的相关问题，如 UN Chief Executives Board for Coordination, “United Nations System White Paper on Artificial Intelligence Governance,” August 2024, https://unsceb.org/sites/default/files/2024-11/UNSystemWhitePaperAIGovernance.pdf, 2024-12-13。

② 《习近平在中共中央政治局第二十次集体学习强调 坚持自立自强 突出应用导向 推动人工智能健康有序发展》，《人民日报》2025 年 4 月 27 日，第 1 版。

大体始于20世纪70年代,被政府和企业用于评估、衡量威胁,作为判断行为合理性的路径。随着冷战后国际关系实践和研究愈加关注如气候变化、恐怖主义、跨国犯罪等超越国界的安全与治理议题,风险视角开始受到国际关系和安全研究领域的关注。①

国家的安全认知与行动往往套嵌于风险视角当中。② 相较于更加强调能力、意图与态势可见性的传统威胁视角,风险视角意味着更大的多样性和不确定性。③ 相关社会科学研究主要关注几个方面议题:如何通过治理框架识别、评估与管理风险,特别是在技术快速发展的背景下④;技术与政策的相互影响与塑造⑤;复杂技术的广泛社会影响及包括跨国与跨学科的风险治理等。⑥ 本文所考察的跨国人工智能关系在不同的技术应用场景下与这些既有议题和研究发生联系。

技术发展的风险往往难以被个人甚至国家轻易预防和控制,人工智能技术更是如此。人工智能相关风险既涉及技术发展本身所带来的不确定性与威胁,也牵涉应用场景中的安全隐患。在人工智能的设计、研发、训练、测试、部

---

① Karen Peterson, "Risk Analysis: A Field Within Security Studies?" *European Journal of International Relations*, Vol.18, No.4, 2011, pp. 693-717.

② Christopher Coker, *Globalisation and Insecurity in the Twenty-first Century: NATO and the Management of Risk*, London: International Institute for Strategic Studies, 2002; M. J. Williams, "(In) Security Studies, Reflexive Modernization and the Risk Society," *Cooperation and Conflict*, Vol.43, No.1, 2008, pp. 57-79.

③ Yee-Kuang Heng, *War as Risk Management: Strategic and Conflict in an Age of Globalized Risks*, New York: Routledge, 2006; Mikkel Vedby Rasmussen, *The Risk Society at War: Terror, Technology and Strategy in the Twenty-First Century*, Cambridge: Cambridge University Press, 2006.

④ Ortwin Renn, *Risk Governance: Coping with Uncertainty in a Complex World*, Routledge, 2008; Christopher Hood, et al., *The Government of Risk: Understanding Risk Regulation Regimes*, Oxford University Press, 2001.

⑤ David Collingridge, *The Social Control of Technology*, St. Martin's Press, 1980; Sheila Jasanoff, *The Fifth Branch: Science Advisers as Policymakers*, Harvard University Press, 1990; Wiebe E. Bijker, et al., *The Social Construction of Technological Systems*, MIT Press, 1987; Gary E. Marchant, et al., *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight*, Springer, 2011.

⑥ Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, Princeton University Press, 1984; Andrew R. Hale and Jan Hovden, "Management and Culture: The Third Age of Safety. A Review of Approaches to Organizational Aspects of Safety, Health and Environment," in *Occupational Injury: Risk, Prevention and Intervention*, 1998; Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable*, Random House, 2007.

署、使用和维护全生命周期中，均存在不同类型和程度的安全风险。[①] 这也意味着，跨国人工智能关系必然与风险属性紧密挂钩，不仅涉及一国内部经济和安全结构，也会在跨国层面刺激不同的冲突与合作倾向，甚至对全球格局产生影响。[②]

二、超越二元论的人工智能关系风险理论

通过风险特征切入来讨论人工智能的影响已渐成新的学术探索方向。相关研究包括对人工智能自主决策风险的认知[③]，人工智能的失控风险及应对[④]，伦理风险与治理[⑤]，长期风险与冲击[⑥]，军事和安全领域风险[⑦]，数据隐私和信

① 全国网络安全标准化技术委员会：《人工智能安全治理框架》1.0 版，https://www.cac.gov.cn/cms/pub/interact/downloadfile.jsp?filepath=ZBWvETi1XzcBKtOIkqelkCQWkQ6GkH4AOG55yBcnFKlQTPuoKqJJjq0JI1Zc9nIr2n5IdA/Bi～vuxCc4OOgnQTEjTcORl73kmCTdYOvdSjs=&fText=人工智能安全治理框架—中文版，2025-06-13。

② Richard A. Clarke and Robert K. Knake, *The Fifth Domain*, Penguin Random House, 2019; Erik Brynjolfsson and Andrew McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, W. W. Norton & Company, 2014.

③ Theo Araujo, et al., "In AI We Trust? Perceptions About Automated Decision-Making by Artificial Intelligence," *AI & Society*, Vol.35, No.6, 2020, pp. 611-623; 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; Uwe Klein, et al., "Application of Artificial Intelligence: Risk Perception and Trust in the Work Context with Different Impact Levels and Task Types," *AI & Society*, Vol.39, No.5, 2023, pp. 2445-2456.

④ Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014; Miles Brundage, et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, 2018; Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, 2019.

⑤ Corinne Cath, "Governing Artificial Intelligence: Ethical, Legal, and Technical Opportunities and Challenges," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol.376, No.2133, 2018, https://royalsocietypublishing.org/doi/epdf/10.1098/rsta.2018.0080, 2025-02-03; Luciano Floridi and Josh Cowls, "A Unified Framework of Five Principles for AI in Society," *Harvard Data Science Review*, Vol.1, No.1, 2019, https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8, 2025-02-03.

⑥ Daron Acemoglu and Pascual Restrepo, "Artificial Intelligence, Automation, and Work," National Bureau of Economic Research Working Paper Series, January 2018, https://www.nber.org/system/files/working_papers/w24196/w24196.pdf, 2024-11-20; Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W. W. Norton & Company, 2014.

⑦ Greg Allen and Taniel Chan, "Artificial Intelligence and National Security," Belfer Center for Science and International Affairs, Harvard Kennedy School, July 2017, https://www.belfercenter.org/sites/default/files/2024-10/Artificial%20Intelligence%20and%20National%20Security.pdf, 2024-12-23; Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, W. W. Norton & Company, 2018.

息安全[①]等。这些研究从不同角度探索了技术与社会的复杂关系,为更加立体地探讨跨国人工智能关系奠定了基础。但同时,既有研究也尚未系统建立针对人工智能关系的整体性分析框架。

本文将基于人工智能的风险可识别度与风险水平这两个属性构建分析框架。这样有助于将风险视角中的抽象因素转化为可测量和可比较的实证变量,从而增强分析框架的可靠性和实用性。

(一)风险可识别度

本文将风险可识别度定义为识别特定风险的可能性。总体而言,较低的可识别度会加剧不确定性和安全焦虑,不利于互动关系的合作性。高可识别度风险通常与相对"单纯"的技术要素相关,如算法错误、数据泄露和系统运行风险,较之低可识别度风险更有可能通过技术测试和评估进行识别和预防。[②]在算法方面,人工智能系统在执行任务时可能由于算法缺陷导致不正确的行为或决策,相对更容易在系统测试阶段被识别和修正。在数据方面,人工智能系统需要处理大量数据,管理不当会导致数据泄露,这类风险可通过安全审计和数据加密等手段部分规避。在系统方面,硬件故障、软件缺陷或外部攻击都可能导致错误的发生,这种风险可以通过测试和监控进行识别。以上这些风险通过技术测试、安全审查、性能评估和稳定性监控等方法进行识别的可能性相对较高,虽不能确保完全或及时预测甄别风险,但在原则上较容易实现,也

---

① Omer Tene and Jules Polonetsky, "Big Data for All: Privacy and User Control in the Age of Analytics," *Northwestern Journal of Technology and Intellectual Property*, Vol.11, No.5, 2013, pp. 239-273; Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Public Affairs, 2019.

② 参见全国网络安全标准化技术委员会:《人工智能安全治理框架》1.0 版;UN AI Advisory Body, "Governing AI for Humanity," September 2024, https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf, 2024-11-01; "How to Identify and Mitigate AI Risks," *PwC Australia*, May 21, 2024, https://www.pwc.com.au/services/artificial-intelligence/accelerate-responsibly-how-to-identify-and-mitigate-ai-risks.html, 2024-11-20; Kevin Buehler, et al., "Getting to Know—and Manage—Your Biggest AI Risks," McKinsey, May 3, 2021, https://www.mckinsey.com/capabilities/quantumblack/our-insights/getting-to-know-and-manage-your-biggest-ai-risks, 2024-11-20。

较容易在人工智能生命周期的早期发现。总体而言，高可识别度风险相对具有直接性、早期性、单纯性和短期性的特点。

可识别度较低的风险主要涉及人工智能技术与特定环境的交互，以及长期或复杂应用部署中产生的间接、延时与难测后果，具有间接性、互动性、复杂性和长期性的特点，常表现为长期影响、环境适应和复杂互动，如社会伦理影响、复杂军事应用、决策辅助系统、多系统集成等。在长期影响方面，这类风险相对更难在人工智能生命周期的初期通过标准测试或监控被识别。人工智能技术对社会、伦理和文化等领域的影响通常是微妙、渐进和累积的，或亦会在特定情境下突然爆发，需要通过长期监控、往复反馈或复杂模拟来逐渐识别。在环境适应方面，人工智能应用在不同条件下会出现性能波动，通常需要在多种环境下进行广泛测试才能发现风险所在。在复杂非预期互动方面，人工智能系统之间（或与其他系统）的交互可能引发难以预测的复杂效应，这种交互的具体影响在应用部署早期很难被精准预测。

（二）风险水平及认知差异

风险水平是指人工智能技术及其应用可能带来的负面影响程度。总体而言，风险水平的高低会影响合作或竞争倾向的程度。人工智能风险水平评估在各国和国际组织的法律法规与标准制定中有不同的方法和具体分类，但在原则上存在共通之处。有些分类主要考虑后果的严重程度，另一些则通过加入风险概率进行预期值测量，考虑到后者与可识别度存在一定重叠，本文将主要参考前一种路径。考虑到不同规范体系建设的成熟度差异，本文主要借鉴中国、美国、欧盟和联合国的相关工作进行风险水平划定。

不同行为体在风险评估和监管方面各有特点。例如，中国重视人工智能风险并同时兼顾技术和市场发展，欧盟以全面立法为基础并强调严格的风险

规制,美国则倾向于自愿性质的框架和行业特定监管。[①] 在综合借鉴这些体系的基础上,本文将人工智能风险分为较高、较低两类,并会考虑人工智能风险认知差异的影响。

较高风险涵盖对重要场景具有潜在重大影响的人工智能应用部署。[②] 这些风险通常涉及国家安全、关键基础设施、医疗健康和数据处理等领域,可能导致严重损失或灾难性后果。例如,在军事和国防领域,错误的决策或系统故障可能导致重大冲突风险或生命财产损失;在关键基础设施方面,如电网、水务系统或交通管理,故障或恶意攻击可能带来重大损失甚至社会经济停摆;在个人数据和隐私保护方面,敏感的个人健康信息、金融记录或行为数据泄露或不当使用会导致政治、经济和社会安全冲击;在医疗健康领域,诊断决策、治疗计划或仪器错误所引发的医疗事故会直接影响个体甚至群体健康状况和生命安全;在交通领域,如在条件不成熟前提下广泛部署自动驾驶系统,可能出现安全风险以及基础设施与法规兼容等问题;在传播领域,如自动内容生成、新闻撰写或图像视频编辑出现重大深度伪造(deep fake)灾害,对个人、社会乃至跨国关系都可能产生深远影响。

较低风险包括对社会、经济或个人生活有显著影响但通常可控的场景,主要涉及具体环境、行业流程或特定群体,风险后果较易通过监管手段、技术措施或应急响应来减轻或控制,对个人权利、自由或安全的威胁较小,一般不存在重大国家和跨国安全影响。例如,消费服务领域的个性化推荐系统,供应链

---

① EU Artificial Intelligence Act, "High-level Summary of the AI Act," February 27, 2024, https://artificialintelligenceact.eu/high-level-summary/, 2025-02-01; Russell T. Vought, "Guidance for Regulation of Artificial Intelligence Applications," *Office of Science and Technology Policy*, January, 2020, https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf, 2025-01-20; Gina M. Raimondo, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, January, 2023, https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf, 2024-12-01;《生成式人工智能服务管理暂行办法》,2023 年 7 月 13 日,https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm,2024-11-05;全国网络安全标准化技术委员会:《人工智能安全治理框架》1.0 版。

② 参见 UN AI Advisory Body, "*Governing AI for Humanity*;"全国网络安全标准化技术委员会:《人工智能安全治理框架》1.0 版。

管理、库存控制或财务分析等流程自动化中的应用，虽涉及数据风险但影响相对有限；在教育领域，尽管存在数据隐私和算法偏见风险，但较易通过适当的监管和技术手段减轻和局部化影响；在人力资源管理和就业平等方面，算法偏见会破坏公平，但可以通过提高技术和流程透明度降低负面影响程度和范围；在游戏、音乐推荐和其他娱乐形式中，影响通常限于用户体验层面；在日常生活的自动和自主化系统方面，故障、理解错误或错误信息提供可能导致不便，但一般不涉及深层次的个人或群体安全。

相比较而言，风险可识别度更接近于客观标准，这是由其核心技术特征和应用场景差异所决定的。而风险水平高低除了客观差异之外，还可能与行为主体的认知判断有关。

对风险水平的认识兼具客观、主观和社会建构性，可能受到科学、文化、价值、制度和权力等不同因素的影响。[①] 这些不同维度的差异会导致行为主体在技术接纳、监管偏好、约束水平等方面采取不同策略。例如，在中、美、欧三个体系之间，中、欧相对更倾向于将数据和伦理类风险视为高风险领域，而美国则倾向于采用自愿合规框架和行业自我监管，对同类领域的风险感知相对较低。欧盟和中国都在推动人工智能不同应用领域的立法工作，以建立安全和道德标准，但在风险评估的具体方法和监管框架上也存在差异。欧盟倾向于设立统一、高标、严格的标准和监管体系，而中国则更加重视在技术快速发展和风险监测管理之间寻求平衡。不同的风险认知不仅影响人工智能在不同技术生态环境中的发展，也会影响跨国人工智能关系。

（三）人工智能关系类型

风险可识别度与风险水平之间不存在明显内生关联，可以作为相互独立维度上的变量构成类型组合。总体而言，较高的风险可识别度有助于建立信

① 唐士其、庞珣：《综合安全论：风险的反向界定和政治逻辑》，《国际政治研究》2022 年第 6 期，第 9—25 页。

任或共同的威胁预期,跨国人工智能关系相对更倾向于合作。而风险可识别度低时,误解和不信任预期更容易上升,关系相对倾向于疏远或对抗。风险水平则总体上影响着合作和疏远的程度。基于此,本文将跨国人工智能关系分为四种类型:监管防控、协同发展、竞赛对抗和谨慎隔离。随着技术发展和应用部署变化,关系状态当然也可能在这些类型之间动态调整。

(1) 监管防控。当风险水平较高且较易识别时,即便其他条件不利于合作(如缺乏战略互信),各行为体之间通常会出于共同安全需求采取措施防控风险,并在重要领域进行监管合作,如伦理准则、数据治理、算法透明度等。这是一种消极性合作,其目的为降低误判或不安全操作的可能性,而非寻求共同发展。具体的协作应对策略可能包括加强高层沟通、信息共享、共同制定风险应急机制,或确保危机在萌芽阶段得到控制。这类协作往往出现在对国家和社会产生重大安全影响的领域。(2) 协同发展。当风险较低且易于识别时,跨国人工智能关系相对更侧重协同发展、积极合作,如经济发展、气候变化分析和疫情预测等。作为一种积极性合作类型,开放性的共赢发展有着更大可能,特别是在相对低敏感领域。例如,不同行为主体之间可以寻求协同创新,推动双边或多边共同研究基金和技术交流,开放相关领域市场与投资合作,共同推动有利于技术、知识和资本流动的措施。(3) 竞赛对抗。当风险较高且较难被识别时,由于担忧对方借人工智能获取优势地位并对己方安全造成威胁,相关行为体关系存在冲突倾向,可能陷入竞赛甚至对抗,尤其在对国家安全具有战略意义的领域表现突出,如涉及指挥控制和决策环节的人工智能军事应用。因双方对彼此的意图或风险管理能力存在疑虑,会倾向于选择强化自身防卫、加强技术限制,并可能在跨国舞台争取资源和伙伴,寻求集团式的竞赛与对抗。(4) 谨慎隔离。当风险水平较低但难以被识别时,人工智能关系陷入恶性竞争的可能性低于前一类型,但彼此接触会较为谨慎,存在一定脱钩倾向。相对于协同发展类型,这时的接触限于低水平有限互动,或伴随数据和技术转移方面的保护主义措施,以确保不陷入技术依赖并规避风险扩散。在这种类型中,跨

国人工智能关系会出现发展碎片化和治理协调困难等情况。

表　人工智能关系类型

	风险水平较高	风险水平较低
可识别度较高	(1) 监管防控(消极合作)	(2) 协同发展(积极合作)
可识别度较低	(3) 竞赛对抗(冲突倾向)	(4) 谨慎隔离(脱钩倾向)

图表来源:笔者自制。

除了这四种类型外,对风险水平的认知差异还可能进一步导致两种不对称关系。高风险可识别度总体有利于合作,但如果风险水平认知存在差异,合作意愿便会存在不对称性。此时,判定风险水平较低的一方会倾向于寻求合作,判断较高的一方则倾向于采取防范措施,双方可能因此在制定标准、共享信息和协同监管等方面缺乏一致性,进而导致治理合作难度上升。当风险可识别度较低时,跨国人工智能关系总体倾向将更强调安全保障。此时,关于风险水平的认知差异会导致不对称和不同级别的安全保障意识与行动,进而可能加剧人工智能关系的疏离和紧张,激化误解与误判,加深不信任并导致摩擦概率的上升。

至此,我们便建立了一个相对系统的风险分析框架,有助于超越竞争与合作两分法,对立体和动态的人工智能关系进行类型区分和探讨。下文将通过具体案例对安全、发展与治理等不同方面的人工智能关系进行讨论。

三、人工智能关系类型的案例讨论

本节将通过典型案例的讨论对前述四种关系类型和两种不对称关系倾向进行阐述。案例选择主要基于相关案例在对应关系类型中的典型性和代表性,涵盖了安全、防务、经济、治理等不同领域。每个案例对应于前述分析框架的一种类型,体现了不同风险水平和可识别度组合下人工智能关系的特征。

(一) 四种关系类型案例分析

1. 监管防控。军事领域的人工智能应用一般具有较高风险,但可识别度的高低取决于具体应用场景。当可识别度与风险水平均较高时,不同行为主体间即便存在竞争甚至对抗性关系,也有动机保持相对开放的沟通渠道,并会寻求通过外交手段降低误判和意外冲突的风险。例如,虽然人工智能的军事应用在中美之间总体强化了竞争甚至竞赛倾向,但是某些领域(如战略力量、核武相关问题)由于人工智能的部署可能出现极为严重的危机和失控风险,消极性的管控合作已成为重要议题。两国元首在2024年利马会晤中谈到人工智能的风险和管理时,特别突出强调了人类控制核武器使用的必要性。[①] 两国以及更大范围内的学术与政策界已开始讨论借鉴美苏热线机制[②]、加强人工智能突发事件通报和安全风险预防等具体工作安排。[③] 在人工智能与核武器的结合方面,全球政策界和学界在一定意义上对于管控和约束的重要性已形成原则性共识。[④]

---

① 参见《习近平同美国总统拜登在利马举行会晤》,《人民日报》2024年11月18日,第1版。

② Christian Ruhl, "The U.S. and China Need an AI Incidents Hotline," *Lawfare*, June 3, 2024, https://www.lawfaremedia.org/article/the-u.s.-and-china-need-an-ai-incidents-hotline, 2024-12-06.

③ Marcus Holmes and Nicholas J. Wheeler, "The Role of Artificial Intelligence in Nuclear Crisis Decision Making: A Complement, Not a Substitute," *Australian Journal of International Affairs*, Vol.78, No.2, 2024, pp.164-174.

④ 智库、学界日愈重视该议题的研究与国际合作,参见瑞士日内瓦安全政策中心的"五常国家"智库、学界圆桌会议项目,Geneva Center for Security Policy, "Co-Convenor's Summary: High-Level Roundtable on the Interface Between Artificial Intelligence and Nuclear Command and Control," February 7, 2025, https://www.gcsp.ch/global-insights/high-level-roundtable-interface-between-artificial-intelligence-and-nuclear-command, 2025-01-15; 瑞典斯德哥尔摩国际和平研究所同样长期关注这一问题,可参见 Vladislav Chernavskikh, "Nuclear Weapons and Artificial Intelligence: Technological Promises and Practical Realities," SIPRI Background Paper, September 2024, https://www.sipri.org/sites/default/files/2024-09/bp_2409_ai-nuclear.pdf, 2025-01-15; Vincent Boulanin, et al., "The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk," SIPRI, 2019, https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf, 2025-01-20; Graham Webster and Ryan Hass, "A Roadmap for a US-China AI Dialogue," The Brookings Institution, January 10, 2024, https://www.brookings.edu/articles/a-roadmap-for-a-us-china-ai-dialogue/, 2025-01-15; Zachary Kallenborn, "Giving an AI Control of Nuclear Weapons: What Could Possibly Go Wrong?" *Bulletin of the Atomic Scientists*, February 1, 2022, https://thebulletin.org/2022/02/giving-anontrol-of-nuclear-weapons-what-could-possibly-go-wrong/, 2024-12-15。关于这些研究机构在人工智能军事应用治理合作中的作用,参见 Haotian Qi and Wanmingkhwan Kamnerdrat, "How Epistemic Community Shapes Global Governance of AI in Military Domain," *The Chinese Journal of International Politics*, Vol.18, No.1, 2025, pp.85-122。

在民事方面，一些领域的人工智能部署存在可识别度较高且风险较高的特征，并存在较好的跨国监管合作。例如在跨境数据流动、隐私保护和安全标准方面，欧盟和日本近年来积极开展了监管合作。日本自 2019 年起启用《通用数据保护条例》(GDPR)同等水平协议，双方关于数据流动的协议于 2023 年秋纳入《欧盟—日本经济伙伴协议》(EPA)并在 2024 年夏生效。双方的数据处理标准趋同，允许在符合《通用数据保护条例》和日本《个人信息保护法》(APPI)的情况下，提升数据跨境流动的效率、透明度和安全性，减少隐私风险。[①] 又如在医疗和交通领域，由于涉及数据管理和隐私问题，一旦风险爆发可能导致大范围信息安全事故。一些国家和地区已在这方面开展积极监管与风险防控合作，针对相关系统中的人工智能应用推动数据保护标准化，确保隐私保护和数据使用的合规性。例如，欧盟与加拿大在相关标准的认定方面开展了持续有效的监管合作。[②]

2. 协同发展。当人工智能带来的风险较低且较易识别时，相关行为体之间更倾向于开放与合作。一些领域虽然可能涉及算法可靠性和数据管理问题，但相对而言，依赖非敏感数据且出现故障等问题的后果较为有限，并不直接影响大范围的群体或关键基础设施安全，属于低风险范围且较易识别。在智能制造领域，跨国商企合作便较为通畅。如中德开展的数字化工厂项目，双方通过共享机器学习和大数据分析等领域的技术标准、最佳实践和优化生产

① European Commission, "EU and Japan Conclude Landmark Deal on Cross-Border Data Flows and High-Level Economic Dialogue," October 8, 2023, https://ec.europa.eu/commission/presscorner/detail/en/ip_23_5378, 2024-11-15; "EU-Japan Deal on Data Flow Enters into Force," https://policy.trade.ec.europa.eu/news/eu-japan-deal-data-flows-enters-force-2024-07-01_en, 2024-11-15; 协议内容参见"Protocol Amending EPA—Data Flows and Personal Data Protection," https://circabc.europa.eu/ui/group/09242a36-a438-40fd-a7af-fe32e36cbd0e/library/f9c7b4f0-ea0f-467a-bb9e-208013b07312/details?download=true, 2024-11-15。

② Lisa R. Lifshitz and Roland Hung, "European Commission Approves of Canada's Data Protection Regime (Again)," January 22, 2024, https://www.torkin.com/insights/publication/european-commission-approves-of-canada-s-data-protection-regime-(again)#:~:text=The%20Commission%20found%20that%20Canada,Act%20("PIPEDA"), 2024-12-17; "Canada and the European Union Sign Agreement to Enhance Border Security," October 4, 2024, https://www.canada.ca/en/border-services-agency/news/2024/10/canada-and-the-european-union-sign-agreement-to-enhance-border-security.html, 2024-11-20; European Commission, "EU and Canada Launch Digital Partnership to Strengthen Strategic Cooperation," November 24, 2023, https://ec.europa.eu/commission/presscorner/detail/en/ip_23_5953, 2024-11-20.

线流程、故障预测和维护等工作,提升制造业的智能化水平。[①]在智能绿色农业领域,经由国际组织协调所建立的发达国家与欠发达地区合作也体现了这一关系类型。该领域的人工智能应用包括作物生长监测、精确灌溉、病虫害预测,以及从整体上提高气变应对和适应能力进而提高产量、生产效率并减少资源浪费。通过共享数据、算法模型和标准,智能农业系统的稳定性和可靠性得以提高。[②] 在发展赋能方面,发达国家和地区之间的合作已发展多年,如欧盟和日本以数据为核心所进行的智慧城市发展合作[③],双方在能源管理、网络、半导体、通讯、技术管理标准等涉及人工智能赋能发展的领域进行了持续的合作。[④] 中欧之间也在如智能交通这样的新兴领域稳步应对共同挑战,推进技术创新和探索标准制定。[⑤]

3. 竞赛对抗。人工智能在跨国关系中伴生的风险也可能存在长期、复杂和互动性,带来相对较低的可识别度,并助推竞争甚至对抗叙事的主导地位。例如,美国有观点认为,中美之间赢得人工智能竞赛的一方将主导未来全球发展,而出于国家安全、全球影响力及意识形态竞争等考虑,美国必须采取积极

---

① Siemens Numerical Control Ltd. (SNC), "The First Digital Native Factory—in Nanjing," https://www.siemens.com/global/en/company/stories/industry/2022/electronics-motors-digital-enterprise-digital-twin-siemens-numerical-control-nanjing-china.html, 2025-01-16.

② OECD, "Empowering Marginalized Communities: Korea's Climate-Smart Agriculture Programme in Kenya," in *Development Co-operation Report 2024: Tackling Poverty and Inequalities through the Green Transition*, July 7, 2024, https://www.oecd.org/en/publications/development-co-operation-report-2024_357b63f7-en/full-report.html, 2025-02-04.

③ 如日欧间 CpaaS.io 项目的成果,参见 2018 年 3 月项目终止成果报告 https://cpaas.bfh.ch/wp-content/uploads/2019/03/CPaaS.io_D8.4_Final_Report_UPDATE.pdf, 2024-12-16。

④ European Commission, "EU and Japan Advance Joint Work on Digital Identity, Semiconductors, Artificial Intelligence," April 30, 2024, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2371, 2024-11-23; 作为日本国民经济极重要领域的电力问题也是欧盟积极介入的领域,如欧盟委员会与日本政府联合支持的合作机构欧盟日本中心(EU-Japan Center)便将其视为重要的智慧城市合作增长点,https://www.eu-japan.eu/eubusinessinjapan/sectors/energy/smart-grid-smart-city, 2024-11-23。

⑤ 2024 年 9 月,中国智能交通产业联盟与欧洲职能交通协会(ERTICO)续签合作备忘录,继续推进双方合作,参见 ERTICO, "ERTICO and ITS China Renew Partnership at ITS World Congress 2024," https://erticonetwork.com/ertico-and-its-china-renew-partnership-at-its-world-congress-2024/, 2025-01-20。

行动赢得这一竞赛。[①] 美国政府与产业界领袖甚至直接对人工智能进行意识形态划线。[②] 在行动方面，美国大范围采取限制交流、打压中国公司，以及出口管制等措施遏制中国的人工智能发展，并以中美之间选边站队来确定第三方能否与美进行合作。[③] 2025 年初，深度求索（DeepSeek）发布了高性能且低成本的大模型[④]，这不仅引发全球科技界、产业界、人工智能与金融市场的冲击，也受到美国政府高层的关注，强化了对华人工智能竞争和技术管控的倾向。[⑤]

就具体领域而言，部分人工智能军事应用具有高风险和低可识别特征。例

① Doug Kelly, "Five Reasons Why America Needs to Win the Race for Artificial Intelligence," *American Edge Project*, April 13, 2023, https://americanedgeproject.org/five-reasons-why-american-needs-to-win-the-race-for-artificial-intelligence/, 2024-12-20; Alfred D. Hull, et al., "Why the United States Must Win the Artificial Intelligence (AI) Race," *The Cyber Defense Review*, Vol.7, No.4, 2022, pp.143-158; Tim Hwang and Emily S. Weinstein, "Decoupling in Strategic Technologies: From Satellites to Artificial Intelligence," Center for Security and Emerging Technology, July 2022.

② 2021 年 7 月，美国拜登政府商务部长雷蒙多（Gina Raimondo）在华盛顿举行的"全球新兴技术峰会"上表示，美国及其盟友必须紧密合作以保证人工智能的进步合乎"民主价值"，而"不能让中国决定围绕 AI 的规则"，参见 Jacob Fromer and Jodi Xu Klein, "US and Allies Must Set 'Democratic' Rules for Artificial Intelligence, Biden Administration Officials Say," July 14, 2021, *South China Morning Post*, https://www.scmp.com/news/china/diplomacy/article/3140997/us-and-allies-must-set-democratic-rules-artificial, 2024-12-16; Sam Altman, "Who Will Control the Future of AI," *The Washington Post*, July 25, 2024, https://www.washingtonpost.com/opinions/2024/07/25/sam-altman-ai-democracy-authoritarianism-future/, 2024-12-16。

③ 例如，2024 年 12 月、2025 年 1 月拜登政府卸任前推出的芯片禁令和人工智能扩散框架，参见 U.S. Department of Commerce, Bureau of Industry and Security, "Commerce Strengthens Export Controls to Restrict China's Capability to Produce Advanced Semiconductors for Military Applications," https://www.bis.gov/press-release/commerce-strengthens-export-controls-restrict-chinas-capability-produce-advanced, 2025-01-17; "Framework for Artificial Intelligence Diffusion," https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion, 2025-02-01; "Implementation of Additional Due Diligence Measureas for Advanced Computing Integrated Circuits; Amendments and Clarifications; and Extension of Comment Period" https://www.federalregister.gov/documents/2025/01/16/2025-00711/implementation-of-additional-due-diligence-measures-for-advanced-computing-integrated-circuits; 2025-02-01。

④ 深度求索：《DeepSeek-R1 发布，性能对标 OpenAI o1 正式版》，2025 年 1 月 20 日，https://api-docs.deepseek.com/zh-cn/news/news250120，2025-03-02。

⑤ David Ingram, "Trump Says China's DeepSeek AI 'Should Be a Wake-up Call' for American Tech Companies," January 28, 2025, https://www.nbcnews.com/tech/innovation/trump-china-deepseek-ai-wake-call-rcna189526, 2025-2-2; J. P. Morgan, "Is DeepSeek Drama a Gamechanger for the AI Trade?" January 31, 2025, https://www.jpmorgan.com/insights/markets/top-market-takeaways/tmt-is-the-deepseek-drama-a-gamechanger-for-the-ai-trade, 2025-3-5.

如,智能化网络攻防和致命性自主武器系统(LAWS)不仅可能有催化机器幻觉(hallucination)和数据泄漏等技术风险,还可能由于系统失控造成连带的冲突爆发与升级。[①] 但在这种风险特征的组合下,大国容易陷入新的竞赛诱惑。为了准备大国竞争背景下的"高端战争",美国政府和军方认为掌握"智能化战争"的战略主动权是抢占未来军事竞争制高点的关键,希望借助人工智能对先进武器平台的赋能实现"智胜"目标。由此,美国试图利用机器学习、计算机视觉、生成式人工智能、一体化和全域性的指挥控制等能力维持其军事优势和领先地位。[②]

4. 谨慎隔离。低风险通常意味着缺少紧迫的共同威胁,而低可识别度又同时放大了合作可能带来的不确定性,这种情况会刺激相对独立的技术发展和应用部署。例如,电力消耗和能源安全正在成为人工智能部署的连带问题。2024 年初,国际能源署(IEA)指出,人工智能在大型应用中的广泛采用可能极大地增加电力消耗,并对全球能源储备带来巨大影响。[③] 然而,由于相关风险

---

① 如 2023 年、2024 年美国海军、陆军先后发布《生成式人工智能和大语言模型指南》,均强调这类人工智能工具目前缺乏足够的可靠性,机器幻觉问题意味着会向用户提供看似真实但其实是完全错误和虚构的反馈内容,参见 U. S. Department of Defense, Department of Navy, "Guidance on the Use of Generative Artificial Intelligence and Large Language Models," September 6, 2023, https://www.doncio.navy.mil/ContentView.aspx? ID=16442, 2024-12-16; U. S. Department of Defense, Department of the Army, "Guidance on Generative Artificial Intelligence and Large Language Models," June 27, 2024, https://www.dau.edu/sites/default/files/webform/documents/27066/Army%20CIO%20Guidance%20on%20Gen%20AI%20and%20LLM_20240627%20%28003%29.pdf, 2024-12-16。

② 体现在美国近年在预算、项目和装备等方面的投入与发展,如 Cheryl Pellerin, "Project Maven Industry Day Pursues Artificial Intelligence for DoD Challenges," October 27, 2017, https://www.defense.gov/News/News-Stories/Article/Article/1356172/project-maven-industry-day-pursues-artificial-intelligence-for-dod-challenges/, 2024-11-20; U. S. Department of Defense, "DOD Announces Establishment of Generative AI Task Force," August 10, 2023, https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/, 2024-11-20; United States House of Representatives, Armed Services Committee, "Summary of the *Fiscal Year 2023 National Defense Authorization Act*," https://www.armedservices.senate.gov/imo/media/doc/fy23_ndaa_agreement_summary.pdf, 2024-11-20; 又如,试图融合人工智能、网络和无人平台的"复制者"计划,参见国防创新小组(Defense Innovation Unit)对项目的相关介绍 https://www.diu.mil/replicator,2024-11-20; 在不同军种以及与盟国之间开始试点探索的联合全域指挥与控制,U. S. Department of Defense, "DoD Announces Release of JADC2 Implementation Plan," March 17, 2022, https://www.defense.gov/News/Releases/Release/article/2970094/dod-announces-release-of-jadc2-implementation-plan/, 2024-11-20。

③ International Energy Agency, "Electricity 2024 Analysis and Forecast to 2026," January, 2024, https://www.iea.org/reports/electricity-2024, 2024-11-23.

不直接威胁重大国家安全，各国在合作与监管上基本各行其是，缺乏合作开发算力基础设施及预防能源危机的动机。深度求索的成功可能部分改变当前人工智能技术路径的能耗及成本，但是否会带来根本性变化仍需观察。局部成本下降以及相应而来更为广泛的部署，反而可能最终拉高整体能耗即“杰文斯悖论”(Jevons Paradox)，这种情形目前来看仍不能排除。①

在娱乐和创意产业中，人工智能应用广泛，其风险往往和长期交互有关，不易识别但总体风险水平相对较低。这会引发本土技术开发与推广的倾向，既保护和增强自身文化影响力，也避免对外部技术的过度依赖。例如，欧盟通过政策支持和资金资助，推动如“创意欧洲”(Creative Europe)等计划②，并通过《数字服务法案》(Digital Services Act，DSA)严格限制跨国公司在内容推荐和算法透明度方面的责任，鼓励符合欧洲价值观的算法。③

(二) 不对称关系及动态变化

除了以上四种典型关系类型外，如前所述，不同的风险水平认知还会带来两种关系的不对称性：一是合作意愿不对称及其对治理难度的影响，二是安全措施不对称及其对摩擦前景的冲击。

1. 合作意愿不对称，治理难度上升。对于数据安全这类可识别度较高的风险，不同行为主体对其风险水平存在认知差异。例如，中国与欧盟认为数据安全的风险较高，而美国则相对较低。欧盟采取了严格的监管措施，对企业数据使用、存储、共享提出很高的法律要求。中国在数据安全方面也设置了高标准，强调国家主权和数据保护，尤其是对个人隐私和敏感数据的保护。相比之

① 相关介绍参见 John M. Polimeni, et al., eds., *The Myth of Resource Efficiency: The Jevons Paradox*, Routledge, 2009。

② European Commission, Creative Europe Programme, https://culture.ec.europa.eu/creative-europe, 2025-01-03.

③ European Commission, *Digital Services Act: Ensuring a Safe and Accountable Online Environment*, October 27, 2022, https://commission.europa.eu/publications/legal-documents-digital-services-act_en, 2024-11-23.

下,美国对数据安全的监管相对宽松,更多依赖行业自律和市场导向的管理方式,只有部分州的措施相对严格。

这样的风险水平认知差异,使得中欧与美国在数据治理方面难以形成一致标准,增加了合作难度,也导致跨国企业在三地间面临不同的法律合规压力。2020年,作为美国与欧盟之间数据流动协作机制的"隐私盾牌"协议被欧盟法院裁定无效,其背后便是双方对于数据治理认知和法律要求的差异。[①] 直至美国提升了保护措施之后,欧美才形成新的机制即"欧美数据隐私框架"。[②]

2. 安全措施不对称,摩擦概率上升。在人工智能军事应用方面,不同国家的风险水平认知也存在差异。以致命性自主武器系统为例,国际军控界的普遍观点认为失控风险和伦理争议意味着这类部署需成为管控合作的重要领域。但是,由于对自主性、自主武器等核心概念的理解、风险预期及相关政治立场存在较大差异,管控合作的难度很大,进展有限。[③]如美俄两国便对相关风险及必要应对存在不同判断。美国相对强调确保致命性自主武器系统的可控性,防止复杂战场环境中的意外后果。[④] 而俄罗斯则更加侧重这类人工智能应用作为"游戏规则改变者"所带来的优势,强调国家有权自主决定相关系统的

① Court of Justice of the European Union, "The Court of Justice Invalidates Decision 2016/1250 on the Adequacy of the Protection Provided by the EU-US Data Protection Shield," July 16, 2020, https://curia.europa.eu/jcms/upload/docs/application/pdf/2020-07/cp200091en.pdf, 2024-12-26.

② European Commission, "Data Protection: European Commission Adopts New Adequacy Decision for Safe and Trusted EU-US Data Flow," July 10, 2023, https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3721, 2024-11-29; European Data Protection Board, "EU-U.S. Data Privacy Framework—F.A.Q for European Individuals," July 16, 2024, https://www.edpb.europa.eu/system/files/2024-07/edpb_dpf_faq-for-individuals_en_0.pdf, 2024-11-29.

③ Ingvild Bode, et al., "Prospects for the Global Governance of Autonomous Weapons: Comparing Chinese, Russian and US Practices," *Ethics and Information Technology*, Vol.25, No.1, 2023, Article 5; Alexander Blanchard, et. al., "Dilemmas in the Policy Debate on Autonomous Weapon Systems," February 6, 2025, https://www.sipri.org/commentary/topical-backgrounder/2025/dilemmas-policy-debate-autonomous-weapon-systems, 2025-02-09; Michael T. Klare, "Diplomatic Debate Over Autonomous Weapons Heats Up," April 2024, https://www.armscontrol.org/act/2024-04/news/diplomatic-debate-over-autonomous-weapons-heats, 2024-12-29.

④ U.S. Department of Defense, "DoD Directive 3000.09: Autonomy in Weapon Systems," January 25, 2003, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf, 2024-11-06.

开发和使用方式，而不应受到严格的国际限制。在联合国《特定常规武器公约》(CCW)框架下所进行的致命性自主武器系统多边谈判和具体政策实践中，俄罗斯始终对限制相关武器开发持保留态度。[①] 这种风险认知的差异便可能导致美俄在该领域的潜在摩擦，加剧双方未来由于相关部署而引发的战略互疑和互动不确定，增加意外冲突的风险，并给地区甚至全球安全带来隐患。

3. 关系类型的动态变化。跨国人工智能关系并非静态锁定在前述四种类型和两种不对称关系上。人工智能技术和应用的快速发展，意味着不同行为主体之间的关系可能存在动态调整。重大技术创新和应用迭代不仅可能重塑跨国人工智能的发展格局，也可能在风险水平及认知与可识别度方面带来改变。例如，原本难以识别的风险，也可能随着技术提升和数据积累，变得更易识别和防控管理。

如是，则新的技术和应用突破便可能打破原有平衡并促进合作。例如，原本在某些领域倾向于谨慎隔离的国家，可能会发现其所处环境的风险可识别度提高，进而转向更加积极的协同发展；而曾在高风险但难以识别场景下陷入竞赛对抗的国家，也会随着风险可识别度的提高，逐渐探索更趋消极合作化的防控合作。

当然，新的冲击也可能加剧跨国认知分歧与摩擦，当新的技术特征因迭代或涌现而发生时，以往被认定为低风险或可控风险的领域，可能转变为高风险。同一种技术在不同主体的视角中又可能被解读为截然不同的风险属性，从而使得潜在合作领域变得敏感和棘手，合作难度上升甚至带来新的对抗。

总之，本文关于风险分析框架及相关典型案例的讨论，其意义并不局限于提供对既有关系状态的刻画，更在于应对动态变化进行可能的前瞻性与适应性调整。在不同关系类型之间发生的迁移和重构，正体现了人工智能技术发展和应用部署存在的深层多样性和不确定性。

---

① Anna Nadibaidze, "Great Power Identity in Russia's Position on Autonomous Weapons Systems," *Contemporary Security Policy*, Vol.43, No.3, 2022, pp. 407-435.

结　语

竞争与合作是跨国人工智能关系的基本参照系,但是这一关系的复杂性与动态性无法由单一维度、二元化的光谱所呈现。人工智能技术的不确定性和应用场景的多样性注定了不同类型人工智能关系的长期并存。

在人工智能赋能的竞争和竞赛中胜出,或许是一些行为主体必然的优先选择,但对于单向优势的过度追求并不利于长远、稳定和可持续的发展。竞争即便不可避免,也应在竞争中寻求共识,推动共同的跨国标准和监管框架,降低技术孤立、标准割裂及对抗竞赛所带来的风险蔓延与失控。不同行为主体在推动技术发展的同时,只有正视技术本身的不确定性和突变性、应用中的系统性与复杂性,以及有限的技术可控性,才能对风险的冲击进行适当的提前干预部署。

在未来追求善用与善治人工智能的道路上,需要不同跨国行为体展现出责任感、协作精神与对内对外的协调能力,建立健全沟通、释疑与防控机制,建立定期的技术与政策对话渠道,制订跨国应急响应协议。这些不仅有助于管理偏冲突性的关系类型,也可为跨国合作奠定信任和行动基础。

从这一点来说,本文基于风险特征属性所进行的分析,较为立体地展现了技术逻辑与政治逻辑之间的动态关联,也为未来的技术和跨国关系风险治理提供了新的视角和思考基础,具有学理和实践两方面的意义。在学理层面,跨国人工智能关系的风险理论具有一定的适应性和灵活性,允许针对不同技术和应用领域的风险特征分析更多样场景下的关系类型。在实践层面,本文分析框架有助于揭示不同的风险和机会,为跨国人工智能相关风险管理和机遇把握提供依据。学术研究人员或可在此基础上进一步明晰状态解释和趋势判断的依据,而政策制定者则可以更好地定位响应策略并合理处理资源分配。结合这些学理与实践价值,本研究的未来拓展仍具有较大空间,特别是不同关系类型之间的动态转化与跨国人工智能关系的“帕累托优化”(Pareto Improvement)。