人工智能军事应用的国际安全风险与治理路径 *

龙　坤　　徐能武

【内容摘要】 人工智能的快速发展和军事应用潜藏着诸多风险，给国际安全带来了新挑战，主要体现在近期的战争门槛降低和自主武器扩散，中期的战略稳定性下降，远期的超级人工智能威胁人类整体安全。这些安全风险跨越国界，影响全球，无法由单一国家独自解决。在此背景下，人工智能全球安全治理的概念与实践应运而生。当前的人工智能全球安全治理领域呈现出多元行为体共同参与但协同不足，各类平台和规范不断涌现但效果有限，以致命性自主武器系统为主要治理对象但进展缓慢的特点。有效应对人工智能带来的国际安全风险，需秉持人类命运共同体理念，发挥相关行为体的重要作用，采取风险分级的全球治理模式，通过建立信任措施、分级管控自主武器系统、加强出口控制等措施，逐步推进人工智能全球安全治理进程，维护人工智能时代的国际安全。

【关键词】 人工智能　自主武器　安全风险　战略稳定　全球安全治理

【作者简介】 龙坤，国防科技大学军政基础教育学院博士研究生，国防科技战略研究智库实习研究员（长沙　邮编：410073）；徐能武，国防科技大学军政基础教育学院教授（长沙　邮编：410073）

【中图分类号】 D5　　**【文献标识码】** A

【文章编号】 1006-1568-(2022)05-0123-19

【DOI 编号】 10.13851/j.cnki.gjzw.202205007

* 本文系湖南省研究生科研创新项目“新兴颠覆性技术对国家安全的影响及治理策略研究”（CX2020004）的阶段性成果和第十三届“金仲华国研杯”获奖论文。调研过程中得到了国防科技大学朱启超研究员、谢海斌副教授及暨南大学陈定定教授的帮助，国防科技大学陈曦博士辅助搜集了部分资料，在此一并致谢。文中错漏概由笔者负责。

作为战略性新兴技术，人工智能在世界范围内快速发展和应用，其日益显现的算法偏见、安全风险等问题也越来越受到世界的关注。尤其是人工智能不断走向军事应用，给国际安全带来了诸多严峻挑战。在人工智能强大的军事应用潜力驱动下，主要军事国家围绕人工智能领域的军备竞赛日趋激烈，[①] 致命性自主武器系统（Lethal Autonomous Weapons Systems, LAWS）议题的凸显更是在国际范围内引起了越来越多的安全关切、伦理争议乃至对于“生存威胁”（existential threat）的强烈担忧。[②] 在此背景下，人工智能全球安全治理已经成为当今世界的一个重大议题。中国也高度重视人工智能全球安全治理，2017 年发布《新一代人工智能发展战略规划》，特别指出要“积极参与人工智能全球治理……共同应对全球性挑战”[③]。

人工智能全球安全治理属于人工智能全球治理的安全维度，目前已有学者对相关领域进行了初步探讨。[④] 但关于人工智能给国际安全领域带来的主要风险及全球治理方式，还缺乏相对清晰的思路和框架。本文首先基于时间序列的分析框架对人工智能国际安全风险进行考察，揭示人工智能全球安全治理兴起的动因。随后分析目前国际社会在人工智能全球安全治理领域的实际进展以及面临的主要问题。最后尝试提出一种基于风险分级的人工智能全球安全治理方案，探讨在当前形势下进一步推动人工智能全球安全治理的现实路径。

① Ariel Conn, “Is an AI Arms Race Inevitable?” Future of Life Institute, March 9, 2017, https://futureoflife.org/2017/03/09/ai-arms-race-principle/; and Paul Scharre, “Debunking the AI Arms Race Theory,” *Texas National Security Review*, June 29, 2021, https://repositories.lib.utexas.edu/bitstream/handle/2152/87035/TNSRVol4Issue3Scharre.pdf?sequence=2&isAllowed=y.

② Office for Disarmament Affairs, “Pathways to Banning Fully Autonomous Weapons,” United Nations, October 23, 2017, https://www.un.org/disarmament/update/pathways-to-banning-fully-autonomous-weapons/; Miriam Kramer, “Elon Musk: Artificial Intelligence Is Humanity’s Biggest Existential Threat, Live Science, October 28, 2014, https://www.livescience.com/48481-elon-musk-artificial-intelligence-threat.html.

③ 《新一代人工智能发展规划》，中国政府网，2017 年 7 月 8 日，http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm。

④ 代表文献主要有傅莹：《人工智能与国际安全治理》，观察者网，2020 年 12 月 19 日，https://www.guancha.cn/fuying/2020_12_19_575099.shtml。傅莹：《人工智能与国际安全治理路径探讨》，《人民论坛》2020 年 12 月下，第 2—7 页。阙天舒、张纪腾：《人工智能时代背景下的国家安全治理：应用范式、风险识别与路径选择》，《国际安全研究》2020 年第 1 期，第 4—38 页。龙坤、徐能武：《致命性自主武器系统军控：困境、出路和参与策略》，《国际展望》2020 年第 3 期，第 78—102 页。

一、人工智能发展应用带来的国际安全风险

正如德国学者乌尔里希·贝克（Ulrich Beck）所言，当今世界已经进入了“全球风险社会”。[①] 人工智能技术的快速发展应用也产生了系列国际安全风险。按照时间序列，可以将这些风险进行大致分类。从近期来看，主要安全风险包括战争门槛降低和自主武器扩散。从中期来看，人工智能可能降低战略稳定性，甚至引发核战争。从远期来看，其风险在于超级人工智能的崛起可能对人类整体带来生存威胁。[②]

（一）近期：战争门槛降低与自主武器扩散

战争门槛降低是当前人工智能带来的最为显著和紧迫的传统安全威胁。人工智能军事应用所带来的武器装备无人化和智能化可以有效降低己方军事人员在武装冲突中伤亡的风险，因而能够在一定程度上减少决策者面临的舆论压力，进而降低发动战争的门槛。诸多事件表明，以无人机为代表的自主武器系统正在为国家的军事行动提供更便捷、低成本、高收益的手段，使危险性较高、政治风险较大以及难以实现的打击成为可能，从而极大降低了战争门槛，容易引发国家间的军事冲突。[③] 此外，由于网络攻击的高速、隐蔽、低成本、难追溯等特征，各国有很大动力将人工智能运用于这一领域，从而可能诱发网络“闪战”（flash war），即国家间机器算法在极短的时间内进行无数次的自动网络攻防交互，导致冲突升级。[④] 鉴于网络攻击目前已具备引发物理伤害的能力，自然成为核威慑下成本较低和风险较小的作战形式之一，而人工智能会强化这一趋势，使战争门槛进一步降低。

① [德]乌尔里希·贝克、王武龙：《“9·11”事件后的全球风险社会》，《马克思主义与现实》2004 年第 2 期，第 70—83 页。

② 由于人工智能技术发展前景具有不确定性，超级人工智能究竟能否实现的争议较大且紧迫度不高，目前的担忧很大程度上还属于科幻层面的反乌托邦恐惧，因而本文的讨论重点在于近期和中期的安全威胁及其治理。

③ 如 2020 年发生的伊朗高级将领苏莱曼尼遭美军“捕食者”无人机暗杀及伊朗核科学家遭无人机刺杀等事件。

④ [美]保罗·沙瑞尔著：《无人军队：自主武器与未来战争》，朱启超、王姝、龙坤译，世界知识出版社 2019 年版，第 234—257 页。

自主武器扩散则是当前人工智能发展带来的比较紧迫的非传统国际安全威胁。[①] 特别是自主武器落入恐怖组织的风险正在增大。自“9·11”恐怖袭击事件以来，恐怖主义已成为威胁国际安全的一个突出因素，而自主武器对恐怖组织、跨国犯罪集团等非国家行为体具有巨大吸引力，能使其武力成倍增强，节省自身战斗人员，更高效地发动恐怖袭击。[②] 更为严重的是，这类非国家行为体通常不会考虑国际规范的限制，从而可能会增加其发起恐怖袭击的频次。在现实中，已经发生多起非国家行为体利用无人机发动袭击的案例。[③] 虽然目前的无人机大多数还依赖人的操控，自主程度还不高，但随着人工智能的不断发展和应用，恐怖分子可能制作或购买自主性越来越高的武器系统，从而使国际安全面临新威胁。

（二）中期：战略稳定性与核战争风险

从中期来看，人工智能将可能对战略稳定性带来较大的负面影响，甚至有引发核战争的风险。[④] 战略稳定性是冷战时期战略理论家发明的重要概念，旨在描述当潜在对手认识到如果与对方发生冲突难以得利的情况下双方就不会轻举妄动的情形，主要包括危机稳定性和军备竞赛稳定性。其中，危机稳定性是指先发制人与后发制人所造成的后果之间的差别。军备竞赛稳定性则表示一种军备行为是否会引起对手反应并导致军备竞赛状态。[⑤]

---

① 截至 2017 年 6 月，有 30 多个国家已经拥有或正在开发武装无人机，至少有 90 个国家及一些非国家行为体拥有非武装无人机。参见 Elisa Catalano Ewers, Lauren Fish, Michael C. Horowitz, Alexandra Sander, and Paul Scharre, “Drone Proliferation: Policy Choices for the Trump Administration,” CNAS, June 2017, http://drones.cnas.org/wp-content/uploads/2017/06/CNASReport-DroneProliferation-Final.pdf。到了 2020 年 3 月，已经有 102 个国家有积极的军事无人机计划。参见 Dan Gettinger, “Drone Databook Update,” Drone Center, March 2020, https://dronecenter.bard.edu/files/2020/03/CSD-Databook-Update-March-2020.pdf.

② Jacob Ware, “Terrorist Groups, Artificial Intelligence, and Killer Drones,” *Homeland Security News Wire,* September 2019, https://www.homelandsecuritynewswire.com/dr20190924-terrorist-groups-artificial-intelligence-and-killer-drones.

③ 如 2018 年 8 月委内瑞拉总统马杜罗（Nicolas Maduro）在演讲时遭遇一架载有 C4 炸弹的无人机爆炸袭击，险遭暗杀。Peter Bergen, Melissa Salyk-Virk, and David Sterman, “Non-state Actors with Drone Capabilities,” *New America Foundation*, July 2020, https://www.newamerica.org/international-security/reports/world-drones/non-state-actors-with-drone-capabilities/。

④ Edward Geist, and Andrew J. Lohn, “How Might Artificial Intelligence Affect the Risk of Nuclear War?” Rand Corporation, April 24, 2018, https://www.rand.org/pubs/perspectives/PE296.html.

⑤ 李彬：《军备控制理论与分析》，国防工业出版社 2006 年版，第 79—83 页。

第一，从危机稳定性来看，人工智能对于战略稳定的威胁主要包括三个方面。首先，攻防平衡的破坏诱发先发制人打击风险。根据攻防平衡理论，当防御占优时，战争发生的可能性会降低，战略稳定性会上升；而当进攻占优，发生战争的可能性上升，战略稳定性会下降。[①] 基于人工智能的兼备速度和隐身性的自主武器（如无人蜂群）为突破对手的防御体系提供了新的利器，可能会促进攻防平衡朝着进攻占优转变，[②] 刺激各国运用自主武器进行先发制人打击获取战略优势，进而降低首攻稳定性。其次，人工智能所强化的“遥感能力”会增大报复性二次核打击力量的脆弱性，从而削弱核威慑和核平衡。不断发展和进步的人工智能和传感器技术使核导弹发射基地、核潜艇等报复性核力量设施更容易被发现、定位和摧毁，从而威胁其生存能力，增大其脆弱性，破坏首攻稳定性和危机稳定性。[③] 这类运用人工智能威胁他国核力量生存能力的举动，将会引发相互猜疑和戒备，冲击最低核威慑战略，从而可能引发战略层面的不稳定。再次，一些国家对于人工智能时代“不对称战争”的担忧，导致将人工智能用于核指挥控制系统或提升核戒备等级而引发核风险。一般来说，拥核国家是否考虑使用自主系统，很大程度上取决于它们对于自身二次打击能力的认知。如果它们认为二次核打击力量较敌方更脆弱、更透明，就有可能在核武器系统中加快嵌入自主能力，特别是可以加速决策进程甚至是使“人在回路外”的系统。当前很大的风险在于，冷战

① Robert Jervis, “Cooperation Under the Security Dilemma,” *World Politics*, Vol. 30, No. 2, 1978, pp. 167-214; Jack Levy, “The Offensive/Defensive Balance of Military Technology: A Theoretical and Historical Analysis,” *International Studies Quarterly*, Vol. 28, No. 2, 1984, pp. 219-238; Stephen Van Evera, “Offense, Defense, and the Causes of War,” International Security, Vol. 22, No. 4, 1998, pp. 5-43.

② Kris Osborn, “Swarming Mini-Drones: Inside the Pentagon’s Plan to Overwhelm Russian and Chinese Air Defenses,” *National Interest*, May 10, 2016, https://nworeport.me/2016/05/12/swarming-mini-drones-inside-the-pentagons-plan-to-overwhelm-russian-and-chinese-air-defenses/.

③ Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, and Moa Peldán Carlsson, “Artificial Intelligence, Strategic Stability and Nuclear Risk,” Stockholm International Pe*ace Research Institute*, June 2020, https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk; 据称，美国防部已在资助一个项目，旨在运用人工智能帮助识别和预测核导弹发射，并可追踪和瞄准朝鲜等国的移动导弹发射平台，参见 Phil Stewart, “Deep in the Pentagon, a Secret AI Program to Find Hidden Nuclear Missiles,” June 5, 2018, https://www.reuters.com/article/us-usa-pentagon-missiles-ai-insight-idUKKCN1J114J。

时期俄罗斯的“死手系统”[①] 和“基于预警发射”[②]（launch-on-warning）等危险策略可能会在人工智能军事应用中逐步显现。[③] 但这些举措本身就潜藏着不稳定的风险，因为人工智能与人具有天然的差别，例如，存在“算法偏见”，在危急关头没有基于人类常识和恐惧心理进行干预和“终止”的机制等。从这些角度来看，人工智能军事应用将对危机稳定性带来极大冲击，甚至有引发核战争的风险。

第二，从军备竞赛稳定性来看，各国围绕人工智能军事应用领域的竞赛正在激烈展开。在无政府状态下，各国军队往往会竭尽所能地谋求针对竞争对手的军事优势，因此通常会追求发展那些能使其作战能力实现跃升的新兴技术。而人工智能所具有的颠覆性潜力让军事大国不断加大其军事应用力度，唯恐在这一领域落后而丧失战略机遇。美国、俄罗斯、英国、印度、北约等纷纷出台人工智能战略，[④] 加快推进人工智能的军事应用，竞相推进自主武器的试验和部署，这将不利于战略稳定。一些国家还开始发展具有“人工神经网络”的高超声速飞行器，[⑤] 尤其是美国军方高调推出了包括基于人工智能的“算法战”“马赛克战”“联合全域作战”等新型作战概念，企图在人工智能军事应用领域形成对中、俄等国的先发优势。[⑥] 从这些迹象不难

① [美]保罗·沙瑞尔著：《无人军队：自主武器与未来战争》,第 351—352 页。

② 李彬：《军备控制理论与分析》，第 84 页。

③ 例如，2015 年，俄罗斯就透露其正在研发名为“斯塔图斯-6”（Status-6）的终极核动力海底无人潜航器，旨在作为防止美国先发核打击的“末日武器”。该无人潜航器具有 1000 米的最大潜深，最大作战距离达 1 万公里，可以运载大型热核弹头。如美国突袭克里姆林宫，该武器将从俄罗斯北极地区的潜艇发射，以约 100 公里/小时（56 节）的航速在海底自动行进，最终向美国海岸线投下核弹头。Edward Geist and Andrew J. Lohn, “How Might Artificial Intelligence Affect the Risk of Nuclear War?” https://www.rand.org/pubs/perspectives/PE296.html.

④ U.S. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, February 12, 2021, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF; Newsroom, “NATO Releases First-Ever Strategy for Artificial Intelligence,” NATO, October 21, 2021, https://www.nato.int/cps/en/natohq/news_187934.htm?selectedLocale=en.

⑤ Andrey Ostroukh, “Russia Leads the World in Hypersonic Missiles Tech, Putin Says,” U.S. News and World Report, December 12, 2021, https://www.usnews.com/news/world/articles/2021-12-12/russia-leads-the-world-in-hypersonic-missiles-tech-putin-says.

⑥ 马斯克对于当前的人工智能军事应用态势表达了深深的忧虑，警告各国围绕人工智能的军备竞赛可能会导致第三次世界大战的发生。Ryan Browne, “Elon Musk Says Global Race for A.I. Will Be the Most Likely Cause of World War III,” CNBC, September 4, 2017, https://www.cnbc.com/2017/09/04/elon-musk-says-global-race-for-ai-will-be-most-likely-cause-of-

看出，一国在人工智能军事应用领域的军备行为容易引发其他国家竞相发展，并试图获取超越对方的军备和战略优势，导致这一领域的军备竞赛愈演愈烈，从而削弱战略稳定性。

（三）远期：超级人工智能威胁人类整体安全

人工智能技术会不会延续当前快速发展的趋势，抑或在现有技术发展过程中遭遇瓶颈？目前，人工智能理论界关于人工智能的本质和未来发展前景存在很大争议。一些学者认为人工智能不管如何发展，始终都是人类手中的工具，能够超越人类的人工智能永远不会出现。[①] 而以雷·库兹韦尔（Ray Kuzwill）和尼克·博斯特罗姆（Nick Bostrom）为代表的专家认为，根据加速回归定律下的技术进化理论，人工智能在未来某个时候将会突破“技术奇点”[②]，甚至智能爆发后会进化出某种高级智能形态，形成“超级智能”（super intelligence），具备人类无法理解或控制的能力，把人类远远甩在后面。这一派观点认为，超级人工智能迟早会出现，只是时间问题。但关于“奇点”出现时间的讨论，依然众说纷纭。总体来看，预测的大致时间可以分为几个阶段，一是 21 世纪中期，二是 21 世纪末期。一项包括人工智能领域的开创者之一尼尔斯·约翰·尼尔森（Nils John Nilsson）在内的 24 名人工智能专家的访谈结果显示，大部分专家认为有 50%的可能在 21 世纪中叶实现人类水平的机器智能，而 90%的可能性是在 21 世纪末，而在此基础上实现超越人类水平的超级人工智能还将需要 30 年左右。[③] 此外，学术界围绕超级人工智能的性质及其对人类的影响也存在不同意见。倘若超级人工智能出现，其对于人类的整体生存是否可能带来毁灭性的“存在性威胁”还有争议。有一些专家认为，超级人工智能会像《终结者》等科幻电影所预警的那样，在

---

ww3.html.

① “Potential Opportunities and Limitations of Military Uses of Lethal Autonomous Weapons Systems Submitted by the Russian Federation,” United Nations document CCW/GGE.1/2019/WP.1, March 15, 2019, https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/Documents/GGE.2-WP1.pdf.

② [美]雷·库兹韦尔著：《奇点临近》，李庆诚等译，机械工业出版社 2011 年版，第 3-4 页。

③ [英]尼克·波斯特洛姆著：《超级智能：路线图、危险性与应对策略》，张体伟、张玉青译，中信出版社 2015 年版，第 24—25 页。

有意或无意中毁灭人类。[①] 鉴于超级人工智能能否实现的不确定性较大，时间周期也很长，因此本文对远期超级人工智能带来的国际安全风险及其治理不进行过多探讨。

二、人工智能全球安全治理的进程与问题

在人工智能军事应用深刻影响国家安全，甚至可能危及全球战略稳定的背景下，各方逐步认识到必须高度重视其潜在风险，对人工智能造成的国际安全问题进行全球层面的治理。当前人工智能全球安全治理的基本结构正逐步生成，但仍面临许多问题。

（一）治理主体：多元行为体共同参与但协同不足

人工智能技术是一个类似于电力的“通用技术”和“使能技术”，能够赋能各行各业，并影响全球的政治经济发展。因而，关于其造成的国际安全问题的治理也必然涉及多个利益攸关方。在人工智能全球安全治理这一议题上，存在着十分多元的治理主体，它们通过磋商谈判、积极呼吁、发表宣言等形式推进这一领域的治理进程。

第一，主权国家是当前国际体系中最基本的单元，也在全球安全治理中扮演着关键角色。目前，越来越多的国家意识到人工智能发展应用所带来的安全威胁，开启了这一领域的国家治理，并积极参与到人工智能全球安全治理进程中来。中国向联合国提交关于人工智能军事应用的立场文件，倡导“智能向善”，强调“人工智能安全治理是人类面临的共同课题”，并呼吁各国政府、国际组织和其他行为主体秉持共商共建共享的理念，协力共同促进人工智能安全治理。[②] 美国国防部出台了专门的人工智能伦理准则和所谓负责任人工智能原则。[③] 其他很多国家也在联合国等多边场合积极参与人工智能

① 如“曲别针灾难”“黎曼猜想灾难”等人工智能通过反常目标实现方式在非故意的情况下毁灭人类。[英]尼克·波斯特洛姆著：《超级智能：路线图、危险性与应对策略》，第 150—153 页。

②《中国关于规范人工智能军事应用的立场文件》，外交部网站，2021 年 12 月 13 日，http://infogate.fmprc.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/jksxwlb_674635/202112/t20211214_10469511.shtml。

③ U.S. Department of Defense, *Implementing Responsible Artificial Intelligence in the*

全球安全治理的进程。2013 年以来，共有 96 个国家在日内瓦《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》（简称《特定常规武器公约》）相关会议和联合国大会等多边论坛公开阐述对致命性自主武器系统的看法，且有二十多个国家公开发表了声明。[①]

第二，联合国、欧盟、北约等国际组织对人工智能全球安全治理这一新兴领域也十分重视。联合国秘书长古特雷斯（António Guterres）自 2018 年以来一直呼吁禁止致命性自主武器系统，他认为这类武器在道德上令人厌恶，在政治上更是不可接受的。[②] 2021 年 12 月，在举行《特定常规武器公约》第六次审议大会上，他进一步呼吁各国尽快制定限制这类自主武器的计划。[③] 2019 年 4 月，欧盟委员会发布《可信赖的人工智能道德准则》（Ethics Guidelines For Trustworthy AI），旨在促进可信赖的人工智能发展，其中特别提及了对人工智能伦理风险和军备竞赛的担忧。[④] 2021 年 10 月 21 日，北约国防部长峰会通过北约首个人工智能战略，强调要以符合伦理道德的方式将人工智能应用于国防安全领域。[⑤]

第三，非政府组织、技术社群等次国家行为体在积极推动基于“人的安全”的人工智能全球安全治理中扮演着重要角色。例如，2012 年 10 月成立的“禁止杀手机器人运动”（Campaign to Stop Killer Robots）致力于禁止使

Department of Defense, May 27, 2021, https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPOSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF; and“DOD Adopts Ethical Principles for Artificial Intelligence,” *U.S. Department of Defense*, February 24, 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

① “Country Views on Killer Robots,” Campaign to Stop Killer Robots, March 11, 2020, https://www.stopkillerrobots.org/wp-content/uploads/2020/03/KRC_CountryViews_11Mar2020.pdf.

② “Convention on Conventional Weapons Policy Brief,” Campaign to Stop Killer Robots, November 2020, https://www.stopkillerrobots.org/wp-content/uploads/2021/11/CCW-Policy-Brief-November-2020.pdf; and António Guterres, “Remarks to the General Assembly on the Secretary-General’s Priorities for 2020,” United Nations, January 22, 2020, https://www.un.org/sg/en/content/sg/speeches/2020-01-22/remarks-general-assembly-priorities-for-2020.

③ “UN Chief Urges Plan to ‘Restrict’ Killer Robots,” RT, December 13, 2021, https://www.rt.com/news/543074-un-killer-robots-restrictions/.

④ European Commission, “The High-Level Expert Group on AI Presented Ethics Guidelines for Trustworthy Artificial Intelligence. Ethics Guidelines for Trustworthy AI,” April 8, 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines -trustworthy-ai.

⑤ Newsroom, “NATO Releases First-Ever Strategy for Artificial Intelligence,” NATO, https://www.nato.int/cps/en/natohq/news_187934.htm?selectedLocale=en.

用完全自主武器，从而保持人类对使用武力的实际控制，维护人类整体的安全。2014 年，以诺贝尔和平奖获得者、南非前大主教德斯蒙德·图图（Desmond Tutu）为代表的 70 多名宗教领袖、代表等，共同签署了一项跨宗教宣言，呼吁各国推进在全球禁止完全自主武器。[①] 技术社群也是人工智能全球安全治理的重要主体。2017 年 9 月，埃隆·马斯克（Elon Musk）联合 26 国共 116 名科学家和企业家致信联合国，呼吁禁止发展和使用致命性自主武器。[②]

（二）治理机制：各类平台和规范不断涌现但效果有限

第一，官方治理机制的主要代表是联合国体系内的《特定常规武器公约》会谈机制、联合国安理会、联合国人权理事会、联合国大会第一委员会（裁军与国际安全委员会）以及国际电信联盟（ITU）等。其中，最为突出的是《特定常规武器公约》机制。自 2014 年以来，《特定常规武器公约》缔约国、非缔约国和全球公民社会代表围绕致命性自主武器系统议题在这一平台上召开了多次非正式和正式会议。[③] 其他一些国际组织也就此议题取得了一些共识。[④] 2019 年 7 月，二十国集团（G20）达成《G20 人工智能原则》，强调人工智能开发应以人类为中心、负责任管理可信赖的人工智能为目标原则。[⑤] 另一特别值得注意的事件是，2021 年 5 月，中国在担任联合国安理会轮值主席期间，主持召开了“新兴科技对国际和平与安全的影响”阿里亚模式会议。这是安理会首次聚焦新兴科技问题，成为国际社会探讨人工智能等新兴科技国际安全治理的新机制。[⑥]

① “Religious Leaders Call for a Ban on Killer Robots,” PAX, November 12, 2014, https://www.paxforpeace.nl/stay-informed/news/religious-leaders-call-for-a-ban-on-killer-robots.

② Ryan Browne, “Elon Musk Says Global Race for A.I. Will be the Most Likely Cause of World War III, ” CNBC, September 4, 2017, https://www.cnbc.com/2017/09/04/elon-musk-says-global-race-for-ai-will-be-most-likely-cause-of-ww3.html.

③ 徐能武、龙坤：《联合国 CCW 框架下致命性自主武器系统军控辩争的焦点与趋势》，《国际安全研究》2019 年第 2 期，第 110—113 页。

④ 例如，2017 年 6 月，国际电信联盟在日内瓦召开“人工智能造福人类”（AI for Good）峰会，提出了人工智能的发展应当如何符合联合国可持续发展目标要求，造福而非戕害人类。ITU, *AI for Good Global Summit Report*, June 9, 2017, https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf。

⑤ “G20 AI Principles,” G20 Insights, July 9, 2019, https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf.

⑥ 钟声：《推动人工智能向善发展才是正道》，光明网，2021 年 12 月 15 日，https://m.gmw.cn/baijia/2021-12/15/1302721203.html。

第二，非官方的治理机制则包括非政府组织发起的国际运动和学术界的论坛等。截至 2021 年 9 月，已有超过来自 65 个国家共 180 个国际、区域和国家非政府组织加入了“禁止杀手机器人运动”，成为国际上推动禁止致命性自主武器系统最为活跃和有影响力的非官方组织力量。[①] 2017 年 1 月初，以人工智能开发者和公民组织为代表的技术社群举行了“受益人工智能”（Beneficial AI）会议，确立了“阿西洛马人工智能原则”[②]。在 2018 年举行的限制“完全自主武器”的正式谈判前夕，一些组织发起了一场宣誓行动，来自 36 个国家的多达 160 家与人工智能相关的公司、覆盖 90 个国家的 2 400 名 AI 领域的知名科学家和企业家共同表示，“绝不参与自主武器的开发”[③]。此外，国际电气和电子工程师协会（IEEE）、谷歌公司等技术组织也通过发起人工智能全球倡议或推出人工智能原则等方式发出各自在人工智能全球安全治理领域的声音。[④]

目前，针对人工智能国际安全的治理机制总体上具有多样化特征。主权国家行为体与各类非国家、超国家、跨国家行为体建立的不同机制相互交织，形成了人工智能全球安全治理体系的复合结构，共同推进这一领域的治理进程。但与此同时，围绕人工智能全球安全治理的治理机制还比较松散，碎片化特征明显，相关国际规范仍处在生成阶段。[⑤] 一方面，由于人工智能安全问题的复杂性、不确定性和地缘政治影响，各方围绕人工智能安全治理的意见和方案存在较大分歧，使得这种新规范的成型存在很大的难度。另一方面，现有治理机制明显效力不足，难以在全球层面形成共识并有效推进相关治理进程落地。

① “禁止杀手机器人运动”官网：https://www.stopkillerrobots.org/members/。

② “Asilomar AI Principles,” Future of Life Institute, January 2017, https://futureoflife.org/ai-principles/.

③ “About The Lethal Autonomous Weapons Systems (LAWS) Pledge,” *Future of Life Institute*, July 18, 2018, https://futureoflife.org/laws-pledge/.

④ “IEEE Ethically Aligned Design Document Elevates the Importance of Ethics in the Development of Artificial Intelligence (AI) and Autonomous Systems (AS),” *IEEE Standard Association*, December 13, 2016, https://standards.ieee.org/news/2016/ethically_aligned_design/; and “Artificial Intelligence at Google: Our Principles,” Google AI, May 26, 2018, https://ai.google/principles.

⑤ Martha Finnemore and Kathryn Sikkink, “International Norm Dynamics and Political Change,” *International Organization*, Vol. 52, No. 4, 1998, pp. 887-917.

（三）治理对象：以致命性自主武器系统为主要议题但总体进展缓慢

从理论上看，人工智能全球安全治理的对象包括前述人工智能带来的所有国际安全问题。而在当下的现实发展中，人工智能全球安全治理以对致命性自主武器系统的探讨和确立规制为主，究其原因，有以下几个方面。

第一，致命性自主武器系统一定程度上涉及了前述人工智能国际安全风险的大部分内容。这类武器如果全面用于战场，就可能引发冲突升级、威胁战略稳定，其扩散到恐怖组织中也会带来很大的人道主义风险，而完全自主的致命性自主武器系统也在一定程度上符合未来的超级智能给人类造成灾难的场景预设。关于这一问题，当前关键的分歧点在于，究竟什么是致命性自主武器系统？致命而非自主的武器系统不是问题，因为目前军队中几乎所有武器系统都具有致命性，已经有较为完善的法律进行规范。仅是自主而非致命的武器系统也不是问题，由于它不涉及人员伤亡的风险，因而也不会给国际社会带来安全挑战。但是，致命而自主的武器系统就会存在很大的安全风险。这类武器由于具备或高或低的自主性，可能与传统武器有着根本的不同。传统的武装冲突法只能对人进行问责，而无法对冰冷的机器进行惩罚。因此，目前在人工智能全球安全治理中最为紧迫的任务之一就是对这类武器进行管控。2013 年以来，致命性自主武器系统逐渐成为继核、太空、网络等国际军控领域之后的新热点。2016 年，联合国《特定常规武器公约》成立专门针对致命性自主武器系统问题的政府专家组，围绕这一问题举行了多轮磋商。2019 年，《特定常规武器公约》缔约国就自主武器系统规范和行动框架的多方面问题达成共识，同时通过了 11 项指导原则（guiding principles），这是国际社会迄今为止达成的基本共识。[①]

第二，从治理效果来看，这一领域的进展显得缓慢而艰难。以主权国家为代表的治理主体之间的理念分歧在很大程度上阻碍了这一进程的推进。由于美俄等大国的反对，《特定常规武器公约》机制下呼吁预防性禁止致命性自主武器系统的提议屡遭挫败。2018 年，以不结盟运动成员国为代表的发

① UN Convention on Certain Conventional Weapons, “Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System,” *United Nations*, September 2019, https://ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf.

展中国家建议尽早制定一项具有法律约束力的国际文书，但由于美、英、俄等国反对，这一提议无果而终。[①] 在 2021 年 12 月举行的《特定常规武器公约》第六次审议大会上，125 个国家表示有必要针对致命性自主武器系统进行新的规制，但同样由于美、俄、英等少数几个国家的强烈反对而再次折戟，无法向前推进。这次会议被一些观察家视为阻止杀手机器人的历史性机遇，而结果以失败告终。[②] 相比致命性自主武器系统，人工智能全球安全治理领域其他议题的治理进程则更显滞后，甚至很多进程尚未开启。例如，目前国际上关于人工智能军事应用引发的战略不稳定风险，只停留在学术探讨层次，核国家之间缺乏相应的风险管控和建立信任机制。

总之，当前国际社会已在致命性自主武器系统议题方面达成了 11 项不具有法律约束力的指导原则，但在战略稳定、危机管控、自主武器扩散等其他较为紧迫的人工智能全球安全治理领域，相关机制和方案仍暂付阙如。人工智能的全球安全治理的实质性进展有限，仍存在很多需要解决的问题。

三、推动人工智能全球安全治理的路径思考

要有效应对人工智能带来的多重国际安全风险和现实困境，不能僵化地用一种方法去应对所有问题，而应该根据不同风险和紧迫度制定相应的治理方案，通过全球层面的各行为体合作协同，推进治理进程。[③] 下文将围绕当

① United Nations Office for Disarmament Affairs, “Pathways to Banning Fully Autonomous Weapons,” https://www.un.org/disarmament/update/pathways-to-banning-fully-autonomous-weapons/; and Mattha Busby and Anthony Cuthbertson, “U.S., Russia Block Consensus at Killer Robots Meeting, Group Says,” CTV News, September 3, 2018, https://www.ctvnews.ca/world/u-s-russia-block-consensus-at-killer-robots-meeting-group-says-1.4077947#:~:text=GENEVA%20--%20A%20key%20opponent%20of%20high-tech%2C%20automated,humans%20stay%20at%20the%20controls%20of%20lethal%20machines.

② Sam Shead, “UN Talks to Ban ‘Slaughterbots’ Collapsed - Here’s Why That Matters,” CNBC, December 22, 2021, https://www.cnbc.com/2021/12/22/un-talks-to-ban-slaughterbots-collapsed-heres-why-that-matters.html.

③ 对于风险高、时间紧迫的安全威胁，需要尽快采取行动，运用预防性禁止、建立信任措施和出口管控等方式进行优先治理；对于风险低、时间紧迫的安全威胁，可以采取适当限制的治理措施；对于风险高、时间不紧迫的安全威胁，可以进行一定的预防性治理，但在其风险切实显露之前，不必投入过多的精力，如超级人工智能；而对于其他的风险低、时间也不紧迫的人工智能安全威胁，则可以暂时不用过多考虑。

前风险高、迫在眉睫的几大人工智能全球安全治理对象进行重点分析。

（一）加快建立信任措施保持战略稳定和管控战争风险

历史表明，具有重大战略价值的新兴技术（如核技术、太空技术、网络技术等）几乎必然走向军事领域。而从当前各大国加速推进智能化建设的发展趋势看，人工智能技术也同样不可避免地走向军事应用。在此背景下，更现实和可行的做法是共同构建人工智能全球安全治理的相关机制和规范，对人工智能赋能的武器装备进行有效规制。而现阶段，正是国家合作构建人工智能国际安全规范的关键窗口期。[①] 针对人工智能给战略稳定性、战争门槛及冲突升级等领域带来的安全挑战，目前是风险极高且最为紧迫的时期，需要国际社会对此高度重视并加强合作，加快建立人工智能军事应用的相关国际规范，并重点采取建立信任措施（CBMs）等方式进行管控。

第一，在核领域，拥核国家须强化合作，尽快签订相关双边和多边军控协定，禁止将人工智能技术用于核指挥控制系统中，确保人类控制所有核武器发射平台。核领域对国际安全的影响是最具毁灭性的，攸关全世界的生死存亡，因而管控人工智能在这一领域的风险尤为紧迫。如前所述，将人工智能应用到核武器的预警、控制以及运载系统中，可能会导致意外核风险的升级，对国际战略稳定造成灾难性后果。因此，核大国之间要尽快推进对话，优先对将人工智能应用于核武器系统的行为设定强有力的限制措施，尤其是有核国家间达成由人类严格控制核发射决定权的协议，确保人类控制所有核武器发射平台。这样才能降低意外核战争的风险，避免在没有人参与的情况下由于机器故障或误判直接导致核战争爆发等灾难性场景的出现。由于核武器的极端破坏性，各国必须确保人类严格控制核发射决策，同时明确人工智能赋能的无人武器装备不应用作核运载平台，避免人类失去对核武器的控制而发生核事故或意外核战争。

第二，在常规战争门槛和冲突升级领域，国家间需要加强合作，建立相关行为准则和协定，限制容易诱发冲突的场景和对象，以防止自主武器意外导致的冲突升级。在行为准则中，应当确保负责使用自主武器系统的指挥官

① 傅莹：《人工智能与国际安全治理》。

或战斗员要采取一切必要的预防措施，限制对军事目标和战斗人员的攻击，避免或尽量减少平民的意外伤亡和对平民财产的损害。与此同时，还需要确保有关使用人工智能军事系统的决定由人类军事指挥官负责，并且指挥官具有理解、解释和使用人工智能军事系统的技术能力。此外，还需要规定，在因部署启用人工智能军事系统而导致的任何非法、不道德或意外事件后，需要及时进行调查，确保责任分配的清晰、准确。[①] 拥有自主武器的国家间还可借鉴美苏《关于防止公海及其上空意外事件的协定》，谈判并签订关于防止自主武器系统事故的协定，并在具有争议的地区设立自主武器使用禁区，以防止自主武器引发的意外事故和冲突升级。

第三，从实现路径来看，可以先通过二轨对话的方式凝聚共识、增进互信，逐步推进到政府间的正式合作。认识共同体是推动全球安全治理和国际合作的重要途径，二轨对话是增进和扩大认识共同体的有效方法。[②] 从历史上来看，一项国际条约的诞生往往需要较长时间，会面临很多困难和障碍。在当前面临前述多重问题的现实境况下，通过二轨对话方式在各国专家层面逐步达成共识，再尝试推动国家间最后达成有约束力的条约和文书，这样循序渐进的推进方式更具可行性。特别是要加快双轨交流机制的构建，通过多边或双边方式推进二轨对话交流。在此过程中，也要积极发挥全球公民社会的积极作用，例如，吸纳“禁止杀手机器人”、人工智能跨国公司等领域的代表参与协商讨论，充分汇集各方诉求和智慧。目前，已有一些由来自不同国家的学者和专家就人工智能的军事用途进行探讨的二轨对话机制，如人道主义对话中心（Centre for Humanitarian Dialogue）组织的“人工智能赋能军事系统行为准则”国际对话，以及清华大学战略与安全研究中心与美国布鲁金斯学会组织的“中美人工智能安全治理”的二轨对话。[③] 二轨对话有助于

① “Code of Conduct on Artificial Intelligence in Military Systems,” Centre for Humanitarian Dialogue, August 2021, https://www.hdcentre.org/updates/release-of-draft-code-of-conduct-on-ai-enabled-military-systems/.

② Emanuel Adler, “The Emergence of Cooperation, National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control,” *International Organization*, Vol. 46, No. 1, 1992, pp. 101-145.

③ 笔者参加了其中的部分研讨对话活动。袁微雨：《CISS 举办第五轮“中美人工智能与国际安全对话”》，清华大学战略与安全研究中心，2021 年 11 月 9 日，http://ciss.tsinghua.edu.cn/info/yw/4185。

加深交流、凝聚共识，推动人工智能全球安全治理向前发展，为制定正式的国际行为准则乃至国际条约奠定良好基础。

（二）分级治理不同自主程度的武器系统

针对不同自主程度的武器系统及其安全风险，需要采取相应的治理举措。完全自主的武器系统，应当在国际社会进行预防性禁止，而半自主或其他低自主的武器系统，则应当采取限制措施、测试评估等方式进行风险管控。

如前所述，虽然超级智能的出现及其危害目前还存在较大不确定性，但“人在回路外”的完全自主的致命性自主武器系统在未来却是完全可能实现的，因而也是人工智能带来的重大而较为紧迫的国际安全风险。在国际社会中，预防性禁止完全自主武器是主流观点。[①] 根据“禁止杀手机器人运动”组织的数据统计，截至 2022 年 8 月，已有 30 个国家、110 个以上的非政府组织、4 500 多位人工智能专家、26 位诺贝尔和平奖获得者以及 61%的全球公众赞同禁止完全自主武器系统。[②] 完全自主的致命性武器系统具有明显的反人类特征，对国际安全可能造成颠覆性威胁和失控风险，因此合理的办法是对其进行预防性禁止。鉴于联合国《特定常规武器公约》机制在历史上对杀伤人员地雷、激光致盲武器等特定常规武器进行规制取得良好的效果，[③] 目前可以在这一框架下签订关于全自主致命性武器系统的议定书，进行预防性禁止。当然，前提是禁止的对象必须是完全自主的致命性武器系统，而不是半自主或有人监督的致命性武器系统。在这一问题的区分上，可以将中国于 2018 年 4 月在联合国《特定常规武器公约》关于致命性自主武器系统问题的政府专家组会议中提出的五项标准作为基本参照，尤其是从不可接受的致命性自主武器系统角度去界定其特征，具体包括致命性、完全自主性、不可控性、滥杀性和进化性。[④] 这一方案的优点是既可以规避完全自主的致命

① 根据益普索（Ipsos）公司针对 28 个全球主要国家近两万人进行的一项民调结果，超过五分之三（62%）的人反对使用致命性自主武器系统。“Opposition to Killer Robots Remains Strong Poll,” Big News Network, February 2021, https://www.bignewsnetwork.com/news/267723372/opposition-to-killer-robots-remains-strong-poll

② 参见禁止杀手机器人官网：“A Shared Movement” August, 2022, https://www.stopkillerrobots.org/a-global-push/a-shared-movement/。

③ Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, New York: W.W. Norton & Company, 2018, pp. 333-339.

④ UN Convention on Certain Conventional Weapons, “Group of Governmental Experts of

性武器给国际社会带来的安全和道德风险，又能将其与现有的自动化和半自主武器系统相区分，既不影响国家推进军事智能化进程，又能在一定程度上管控风险，从而容易在大国间达成一致。虽然目前美、俄等大国仍对全面禁止致命性自主武器系统持反对意见，但部分原因在于没有对需要禁止的致命性自主武器系统的范围进行清晰的界定。如果将致命性自主武器系统界定为具有上述五个特点的武器系统并进行预防性禁止，则符合各国的共同利益。

此外，针对半自主或低自主致命性武器系统，也需要对其进行一定约束，尤其是要对此类武器的可预测性、攻击目标类型、使用时长和范围、使用情形以及人类监督方面进行明确的限制，以预防和减少其用于战场引发的国家安全风险。具体而言，对不可预测和旨在对人类使用武力的自主武器系统也应当予以禁止。例如，可以参考《渥太华禁雷公约》中对杀伤人员地雷的禁止性规定，对旨在攻击人类的自主武器系统进行禁止。对于未被禁止的自主武器系统，也要对其设计和使用进行一定程度的规制，如限制其目标类型为仅攻击属于军事目标性质的物体，限制其使用时长、地理范围和规模，限制其使用情形须为不涉及平民或民用物体，限制其人机互动为需要确保有效的人类监督和及时干预等。[①] 针对可以部署使用的半自主或低自主武器系统，国际上需要共同制定一套通用的性能指标和评估标准来评估其有效性和安全性。在部署这类武器系统之前，各国应运用这套标准对其进行严格的测试和评估，确保系统达到商定的最低性能，从而避免自主武器系统以违背指挥官意图的方式造成非预期事故和冲突升级。[②]

（三）强化自主武器及相关技术出口控制

---

the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects Position Paper Submitted by China," United Nations, April 11, 2018, https://www.unog.ch/80256EDD006B8954/(httpAssets)/E42AE83BDB3525D0C125826C0040B262/$file/CCW_GGE.1_2018_WP.7.pdf.

① "ICRC Position on Autonomous Weapon Systems," International Committee of the Red Cross, May 12, 2021, https://www.icrc.org/en/publication/4550-icrc-position-autonomous-weapon-systems.

② Michael Horowitz and Paul Scharre, "AI and International Stability: Risks and Confidence-Building Measures," Center for a New American Security, January 12, 2021, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

国际不扩散机制是阻止和延缓特定军备和技术扩散的重要手段。国际社会经过长期共同努力，已构建了相对完善的国际不扩散机制，在防止大规模杀伤性武器及其运载工具扩散和维护国际安全上发挥了重要作用。在防止自主武器扩散问题上，也可通过相关国家的出口控制措施进行风险管控。拥有自主武器的主要国家需要加强出口控制，限制自主武器及相关技术向没有足够风险管控措施的其他国家行为体和非国家行为体的水平扩散。目前，国际社会已对自主武器扩散风险表达了关切，并在防止自主武器扩散到恐怖组织等非国家行为体问题上达成了基本共识。[①] 自主武器的扩散，尤其是向恐怖组织扩散的风险，不符合国际社会大多数成员的安全利益，对这类风险的治理具有良好的共有观念基础。对此，从拥有自主武器的国家源头进行出口控制尤为重要，尤其要限制人工智能赋能的轻小武器在世界范围内广泛传播，因为这些武器容易扩散到恐怖组织等非国家行为体中。

具体而言，可以借鉴历史上的“导弹及其技术控制制度”（Missile Technology Control Regime, MTCR），建立类似的“自主武器及其技术控制制度”。鉴于人工智能技术主导国大部分都是拥核国家，中、美、俄等国家可以就共同限制自主武器系统及其关键技术的出口进行协商，避免自主武器在世界范围内的泛滥，甚至流入恐怖分子手中。此外，还需在此框架内建立定期核查制度，设立联合审查机构或雇佣第三方机构定期对缔约国履约情况进行核实。同时，对自主武器的出口管控不应阻碍民用领域人工智能的正常交流，二者之间应当寻求适当的平衡。人工智能具有军民两用性，需要限制的是自主武器扩散，而非这类技术的正常交流和贸易。当然，如何既防止武器扩散而又不妨碍民用领域的技术交流存在挑战。目前已有的《禁止化学武器公约》和《禁止生物武器公约》通过聚焦使用目的而非具体的物质或技术对军民两用技术和产品的出口进行管制，取得了较好效果，这可以为自主武

① 联合国《特定常规武器公约》框架下针对致命性自主武器系统问题达成的 11 项指导原则中的第六条明确指出，在发展或取得基于致命性自主武器系统领域新技术的新武器系统时，应考虑“落入恐怖主义团体手中的风险和扩散的风险。”UN Convention on Certain Conventional Weapons, “Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System,” https://www.ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf.

器系统出口管制提供较好的经验借鉴。[①]

结 束 语

人工智能作为科技发展的前沿领域，本质上是全球性的通用泛在技术和先进生产力，其在全球化时代产生的安全风险也必然会波及全球，可能会严重威胁战略稳定，降低战争门槛，提高冲突频率，并给国际恐怖主义大开方便之门，甚至可能诱发超级人工智能的崛起，威胁人类整体安全。要应对人工智能带来的多重国际安全风险，需要全世界的主权国家、国际组织、跨国公司、全球公民社会等国际关系行为体加强合作，共同参与和贡献智慧，推动人工智能全球安全治理的进程，使得制度建设和伦理规范跟上人工智能发展演进的速度。作为引领人工智能发展应用的“领头羊”国家和世界前两大经济体，中、美两国积极承担这一领域的大国责任尤为重要。当前一段时期，正是管控人工智能国际安全风险的关键窗口期，两国需要强化合作，共同引领人工智能的全球安全治理进程，通过基于风险排序和时间紧迫度的分层治理方式化解安全风险，维护全球战略稳定。同时，也需充分发挥国际组织和全球公民社会的推动作用，强化主权国家、国际组织、全球公民社会等治理主体的良性互动和相互合作，形成松散耦合的人工智能全球安全治理机制，合力推进这一进程的有序前行。

[责任编辑：杨 立]

① 徐能武、龙坤：《联合国 CCW 框架下致命性自主武器系统军控辩争的焦点与趋势》，第 116 页。