人工智能的国际安全挑战及其治理

朱荣生[1]，冯紫雯[2]，陈　琪[1]，陈　劲[3]
（1. 清华大学战略与安全研究中心，北京　100080；2. 澳门城市大学葡语国家研究院，
澳门　999078；3. 清华大学经济管理学院，北京　100080）

摘　要：人工智能技术近年来的快速发展及其在军事领域的应用，引发了公众、政策界和学界的广泛关注。本文对人工智能军事化冲击国际稳定的理论和政策争论、人工智能时代国际安全治理的规范演化进行了总结和分析。在国际社会探索人工智能全球安全治理的进程中，国际社会出现了借鉴国际军控的经验教训、全面禁止致命性自主武器以及规范技术发展路径的多元主张。本文认为，多条路径主导和多元网络化发展的国际规范趋势导致不同的行为体之间形成了复杂的相互影响关系，这决定了没有一种治理主张可以在短期内主导规范演进的过程。

关键词：人工智能；国际安全；国际规范；战略稳定；中美合作
中图分类号：D815.5　**文献标识码**：A

International Security Challenges and Governance of Artificial Intelligence

Zhu Rongsheng[1], Feng Ziwen[2], Chen Qi[1], Chen Jin[3]
(1. Center for International Security and Strategy, Tsinghua University, Beijing 100080, China; 2. Institute for Research on Portuguese-Speaking Countries, City University of Macau, Macau 999078, China; 3. School of Economics and Management, Tsinghua University, Beijing 100080, China)

Abstract: In recent years, the rapid development of artificial intelligence technology and its application in the military field have attracted extensive attention from the public, policy and academic circles. This paper summarizes and analyzes the theoretical and policy debates on the impact of artificial intelligence militarization on international stability, and the norm evolution of international security governance in the era of artificial intelligence. In the process of exploring artificial intelligence global security governance, the international community has come up with multiple views on drawing lessons from international arms control, comprehensively banning lethal autonomous weapons and regulating the path of technological development. The study argues that the trend of multi-path and multi-network development of international norms leads to the formation of complex interaction among different actors. It decides that no single governance proposition can play an absolute leading role in short term.

Key words: Artificial intelligence; International security; International norm; Strategic stability; China-US Cooperation

基金项目：国家社会科学基金“人工智能的国际安全挑战与规范演进研究”（21CGJ012）。
收稿日期：2022-04-29
作者简介：朱荣生（1991—），男，北京人，博士后，研究方向为人工智能与国际安全。

0 引言

人工智能武器化对国际安全的挑战到底几何？一些战略学者对人工智能的军事效能有很高期待，认为它如同核武器横空出世一般引发军事革命，改变了人们对国际冲突和战场的基本认知[1]。一些军事安全研究者则态度保守，将之视为既有武器的“放大器”，认为它不足以改变国际安全体系，并质疑其战场应用效果能否满足预期[2]。在学者争论和政策辩论热闹非凡之时，中美对人工智能前沿技术的追赶竞争，使这项具有颠覆性潜能的技术成为“权力游戏”的新宠。

在大国战略竞争和军备竞赛升温的背景下，以自主武器为代表的人工智能赋能的武器备受推崇，人们的关注已从该技术在军事领域的发展机遇转向随之而来的严峻挑战，以及以何种理念、方式、工具开展有效的安全治理。参与到这场论辩的各方纷纷提出开展军备控制的倡议，但在如何实现维护国际稳定的方式上存在巨大差异。通常而言，当某一种巨大且急迫的危机降临，引起精英和大众的广泛关注和行动意愿，经历一段相对漫长的时间将催生出新的国际规范。在人工智能的军事应用还没有展现出全球性恶果之前，该如何平衡技术发展红利和潜在风险是一项难题。

基于上述理论思考的目标，本文结合国际安全的现有理论视角，从人工智能引发的国际安全挑战，比如军备竞赛、安全困境、危机稳定、中美科技领导权竞争等方面进行文献梳理，而后梳理多元行为体对于正在演进的国际安全规范过程的主张。有鉴于管控人工智能军事化的国际规范还处于争论发展阶段，本研究的目标不是提出规范建构的理论框架，而是对现有纷杂的各种治理观点及其逻辑假设进行梳理，为探索人工智能规范演进的路径提供新的思考方向和视角。

1 人工智能对国际安全的挑战

弗兰克·斯利珀（Frank Slijper）等学者对美国、中国、俄罗斯、法国、以色列和韩国的人工智能军事化行为的研究发现，这些国家正加大研发投资，联合企业和科研机构落实研究方案，都竞争性地坚持绝不能落后于人的人工智能战略[3]。人工智能属于军民两用技术，易于扩散的特性意味着斥巨资带来的技术优势很可能不会长期保持。但是，一些国家对人工智能军事化的执念不仅会挤兑提升国民福利的财政储备，还会引发军备竞赛，破坏战略稳定，零和的人工智能军备竞赛似乎成为自我实现的预言。

1.1 刺激军备竞赛论：引发大国军备争优，加深危机升级风险

人工智能有可能掀起一场军事革命。与核武器时代追求弹头数量优势相比，算法战时代追求的远非武器数量优势，而是信息技术领导地位，以获得军事实力“质”的突破[2]。面对智能技术迭代速度加快的趋势，军事技术差距放大了国家的生存恐惧，由此军备竞赛似乎不可避免。令人担心的不只是军备竞赛本身，还有尖端武器部署和反制技术研发的竞争，恶化了国家安全预期并陷入不断升级的安全困境。

国际军控学者认为，优势争夺引发军备竞赛，加大国际安全风险。2020 年 9 月 27 日，阿塞拜疆和亚美尼亚两国为争夺纳卡地区控制权爆发冲突。在地面进攻受阻的情况下，阿塞拜疆动用 TB-2 察打一体无人机、“赫尔墨斯”长程侦察无人机、“哈洛普”自杀无人机、“搜索者”中程侦察无人机、“人造卫星”短程侦察无人机和阿塞拜疆改装的安-2 无人机等取得纳卡冲突主导权。有观点认为，无人机展现出来的战场颠覆性效果恐怕会让更多国家购买先进无人机，导致武装无人机扩散问题[4]。另一方面，国家可能迫于军备竞赛压力而部署危险的人工智能系统。人工智能技术虽然强大却不一定可靠。很多系统易受攻击，或者在非训练环境中失灵[5]。从技术安全角度看，人工智能武器的固有安全风险有助于降低决策者的部署意愿，进而在一定程度上延缓军备竞赛。但是，国家渴求权力的现实会打破这种理想状态。有的国家迫于军备竞赛压力，会不断放宽对智能武器的安全限制，甚至会接受不合格的以及未经检测的武器系统[6]。

更为谨慎的分析认为，人工智能武器化使国家陷入安全困境。国家既难以明晰对方的战略意图，又恐惧于别国终将获得比自己更强大的军事力量，为了提高自身安全，应对他国的潜在进攻，从而选择增加军备。有观点认为，中美已经被拖入一场安全困境中，即美国将中国的人工智能发展看作一种安全威胁，而中国认为必须加快发展

速度以反制美国的各种压制政策[7]。由于人工智能技术具有军民两用的特性，可通过公开渠道获取相关技术要素并辅助训练军用系统，它的赋能效果带来何种程度的军事优势具有不确定性，这进一步放大了国家的生存恐惧。错误认知驱使美国更加怀疑中国维持两国关系稳定的意图，对意图的曲解将中美推向一场更加严峻的“科技战”，或者陷入赢得领导力却减少安全的自我挫败中。

1.2 破坏危机稳定论：增加战略决策不确定性，冲击危机稳定性

战略博弈终归是人和人的较量，引入非人类智能将加大战略误判的可能性。由人定义的算法和数据训练集会无意间将偏见带入训练系统，使人工智能给出错误建议。敌国可能污染己方的训练数据集，导致人工智能系统提供错误的侦查信息，诱骗己方落入敌人设置的陷阱。虽然人工智能可提出令人意想不到的建议，使实施建议的一方在博弈中以奇致胜，但可能增加对战略意图的判断难度，导致危机持续升级。除了理性决策的风险外，人工智能技术还会影响决策者的信念体系，让他们对机器抱有错误期待，相信机器总会正常运作并发挥奇效，因而有恃无恐地炫耀武力，从而造成危机不断升级[2]。

有研究认为，人工智能军事化造成的潜在战略风险是削弱核打击报复可信度，进而造成冲击危机稳定性的恶果。2017年5月，美国兰德公司发布报告表示，人工智能可以提高对竞争对手的核报复力量的侦查、跟踪、瞄准和打击的能力，这会削弱核报复的能力和可信度[8]。其可能结果将是加大依赖二次打击的国家应对危机的决策压力，迫使他们要么做出率先发动核打击的决定，要么提前落入失败的境地。另外，他们还可选择攻击敌方的探查器件、生成对抗性网络或者采取战略欺骗来防止对方打击自己的报复力量[9]。虽然这在一定程度上有助于保护自己的核报复设施，但加大了国家行为的不可预测性[10]。

上述参与讨论的部分专家认为，国际战略稳定不必然被人工智能轻易打破。人工智能可辅助决策者提高对情报的收集和分析能力，使威慑、保证和再保证更可信，起到巩固战略稳定的作用。这个过程会形成良性循环，最终降低战争风险。从博弈信息沟通而言，拥核国家互信不足或者难以揣测对方意图，更高效的侦察能力给决策者提供了更多可信的信息。此外，人工智能可更有效发挥对核设施的监督和核查作用，提高定位一国的导弹发射井的能力[11]，被威慑国可提升导弹发射井承受打击的防御力，保留第二次打击的战略威慑实力。

上述理论推演能否得到验证恐怕需要更多的时间，不过美国近年来的军备智能化迭代给维护国际稳定注入了更大的不确定性。2019年9月25日，时任美国联合人工智能中心主任杰克·沙纳汉（Jack Shanahan）中将公开表示，美国将确保继续由人类控制核发射决策，人工智能技术应用于武器平台应该扩大我们的优势来威慑战争的发生[12]。2019年11月1日，美国国防创新委员会发布的《人工智能原则：国防部人工智能应用伦理的若干建议》报告指出，当人工智能系统用于核武器等高危武器结合必须经过严格的测试和验证，就授权、安全和可靠性做更多研究项目，以避免出现技术事故[13]。美国国防部反复声明不会将危险的技术部署于核武器指挥控制系统，或者强调对大型武器平台的部署必须经过严格安全审查，都不是维护战略稳定可信度很高的信号，因为其刻意回避了美军对可搭载核武器的平台进行智能化升级。例如，美国空军已经将人工智能技术部署于F-35和可搭载核弹的B-2轰炸机上，以改进他们的攻击目标瞄准系统和忠诚僚机的协同[14]。对于技术安全和维护战略稳定的问题，美军的逻辑是不断升级智能化作战能力。相比于新兴技术在安全方面的不确定性，更好的办法或许不是加速智能化迭代，而是更加注重人类的控制和技术安全的可信度。

1.3 加深大国竞争论：人工智能竞争的意识形态化加剧了国际体系的分裂

政策界有一种观点认为，中美科技领导权的战略竞争正危险地滑向一场“大国政治的悲剧”。人工智能将从经济和军事上改变国家间的力量对比，甚至引发新一轮大国兴衰[15]。在人工智能军事应用方面处于劣势的国家面临更大的国际竞争压力[16]。2021年3月，美国人工智能国家安全委员会发布《最终报告》，提出创建技术咨询委员会，统筹制定一项同中国竞争的人工智能战略。同年7月，拜登政府的商务部长吉娜·雷蒙（Gina Raimo-

ndo）、国务卿安东尼·布林肯（Antony Blinken）、美国国家安全顾问杰克·沙利文（Jake Sullivan）等高官在美国人工智能国家安全委员会举办的讨论会上表示，美国必须确保自身的技术全球领先，要加紧制定新政府的人工智能战略，加强与盟友的合作来稳固所谓的“民主价值”以及主导规则制定权。这种零和竞争的思维会将世界推向一个分裂对立、激烈对抗的格局。联合国秘书长安东尼奥·古特雷斯（AntónioGuterres）曾警言道：世界上最大的两个经济体（中国和美国）创造了两个独立的、相互竞争的世界，各自拥有自己的主导货币、贸易和金融规则、自己的互联网和人工智能能力，以及自己的零和地缘政治和军事战略[17]。

相反的观点则认为，虽然权力争夺加深了中美长久以来的战略互疑和科技“脱钩”，但解决之道不是升级地缘政治竞争，而是要理解深厚的中美相互依赖现实，看到绝对收益对技术安全的重要性[18]。在人工智能技术发展方面，中国和美国已然形成不可分割的共同体。有研究通过对科学和工程领域的研究进行文献统计发现，如果美国贸然切断同中国的科研合作关系，将导致美国的发表成果数量大幅下降，承受更大的科研损失[19]。断绝中美科研合作的负面冲击是双向的。前外交部副部长傅莹认为，中美不应该任由传统的地缘政治、零和竞争思维主导两国关系，因为其结果将是自毁性质的[20]。

1.4　风险高估论：军事应用受阻，安全挑战被高估

除了上述强调人工智能时代的巨大挑战外，还有观点认为人工智能若得不到人类的信任就难有影响现实的机会[21]，它对国际安全的挑战自然就大打折扣。国家部署人工智能武器的动力是欲求它带来的军事优势，但人工智能的革命性效能如何还没有定论。当革命性的新技术嵌入到战略决策环节时，决策者和武器操作员都需要得到培训，才能适应新技术带来的变化。不同的军事部门在接纳和使用新技术的程度和速度不同，带来的潜在问题是军事部署和作战效率反而因协调障碍而降低。加之人工智能的判别能力还未达到人类智能的程度，经过智能系统提供的情报信息和分析有可能引发人类的错误决策。如果人工智能在战场上错误识别敌方目标，并影响指挥员的军事部署，显然会造成战场上的巨大损失。考虑到这些现实安全问题，人工智能何时能得到人类的信任，展现其全部军事潜力就不得而知了。

较为乐观的研究者认为，军事部门保守的制度和文化会阻碍人工智能的部署，部署人工智能系统要面临长时间的安全审查。以美国为例，递交申请实际上只需两周，却要花六个月分享必要的数据以获得组织的批准许可，之后再花六个月或一年时间让其他相关部门同意在平台上运行。在完成艰难的审批程序后，申请者还要再等几个月进行软件更新。冗长的审批过程反映出和平时期的精神状态和最小化风险的组织文化，严重阻碍人工智能的应用[22]。同时，新技术的应用还受到组织文化的限制。美国的 Maven 项目成员曾表示军事组织排斥这种改变，因为这项技术的融入可能带来颠覆性影响，但是这些好处不会马上就显现出来[23]。

2　国际安全规范演化路径的多元视角

人工智能不是单一技术，而是一种技术集群。人工智能全球治理天然地涉及多学科、多领域、多元行为和多元治理议题的特征，这些特征决定单一治理主体和治理方式无法解决所有治理关切。尽管参与论辩的各方提出了不同的治理主张，某些构想的政策可操作性不高，但多条治理路径必将长期影响国际规范演进。

2.1　军备控制路径：大国借鉴国际军控的经验教训

较为乐观的政策研究者认为，国家会采取自我军事克制行为，不开发和使用违规的人工智能系统。生化武器和人工智能技术有相似之处，均属于军民两用技术。格雷·阿伦（Greg Allen）对两种技术进行比较研究后提出，美国应该在规范军用人工智能研发政策中加入自我克制的内容。他以 1969 年美国结束研发攻击性生物武器为例声称，美国强大的核力量可以对其他国家的生化攻击实行有效威慑，满足安全预期，故此没有必要开发和部署可能威胁本土安全的人工智能技术[24]。该观点背后的政策含义是，决策者担忧人工智能武器系统出错，因此不能放松安全审查，武力使用需要审慎克制。还有观点认为，人工智能研发到使用的过程应严格遵守技术检测和法律审核程序，

美国要释放明确的军事克制信号，与包括敌对国家在内的世界各国就人工智能系统安全开展监察合作，避免出现不必要的危机[6]。

一些军控学者认为，人工智能国际安全治理需要借鉴核军控经验，建立大国军控机制。美苏经历漫长的军控谈判，建立了多种信任机制、对话渠道、克制措施和共同准则，这些是长年累月基于减少武器引发灾难性后果取得的成就，是军控工具箱不断升级的产物[25]。因此，人工智能的国际军控必然是一个漫长过程，很难就某些原则快速达成共识。尤尔根·阿尔特曼（JurgenAltmann）和弗兰克·绍尔（Frank Sauer）不完全认同这种观点，认为大国之间应尽快达成军事克制底线原则，为此可吸收核军控经验维持战略稳定。他们总结了三种方案：第一种是签订国际条约，限制发展智能武器；第二种是基于国际人道法，就军事“行为准则”签订不具约束力的协议；第三种是主要大国之间达成协议，限制某种人工智能武器。他们认为这三种方案各有优劣，可同时施行，互补不足。

从武器攻防角度看，国家应开发和部署防御性自主武器。防御性自主武器不仅能提高国家的防卫能力，还将提高进攻方的军事成本，从而降低攻击意愿。制造无人武器比维护传统大型武器成本更低，国家领导人可削减军事预算，鼓励减少其他国家军备开支，这在一定程度上有助于缓解安全困境。虽然没有纯粹的“攻击性”或“防御性”武器，但是可在攻击范围、弹药数量和自主化程度三个方面设定必要标准。防御性武器的攻击范围应该不广，携带弹药数量更少，也不需要具备进攻性智能武器所需的能力。相反，进攻性武器则需要较远的进攻距离，携带大量弹药以及先进的人工智能系统。这类先进武器的科技水平和生产成本很高，因而达成国际共识并实现禁止将更为容易。自我进化和自我复制的自主武器脱离了人类掌控，这种不具可预测性的超级人工智能尤其不能被用于军事[26]。

2.2 全面“禁武”的手段：国际组织倡导禁止完全自主的致命性自主武器

国际组织正致力于将禁止地雷的成功战略用于禁止致命性自主武器。2012年10月19日，7个非政府组织为了实现禁止自主武器的共同目标在纽约聚首，正式成立禁止杀手机器人运动。参与禁止杀手机器人运动的非政府组织曾通力合作，成功地推动建立全球禁止地雷、集束弹药和致盲激光武器的国际公约。共同创始人史蒂夫·古斯（Steve Goose）表示，希望禁止杀手机器人运动能够延续国际反地雷组织和集束弹联盟，推广达成国际禁雷公约和国际禁止集束炸弹公约的成功经验[27]。该运动的倡议协调人和主要推动者玛丽·韦勒姆（Mary Wareham）和周迪·威廉姆斯（Jody Williams）是禁止地雷和集束炸弹的国际活动的倡导者和实践者，在她们的努力协调下，禁止杀手机器人运动已壮大成为容纳来自63个国家的151个国际、区域和国家非政府组织的公民社会组织。

联合国《特定常规武器公约》会谈机制是致命性自主武器国际军控的核心机制。在2013年的《特定常规武器公约》缔约国会议上，各国激烈讨论是否在次年设立讨论致命性自主武器的非正式专家会议。至2022年8月，《特定常规武器公约》缔约国会议已就致命性自主武器问题举办6次政府专家组会议。各代表团认为，《特定常规武器公约》缔约国会议是讨论和解决致命性自主武器的核心平台，必须从武器开发到使用的过程确保人类的控制和监管[28]。

除了《特定常规武器公约》会谈机制外，联合国人权理事会和联合国大会第一委员会也起到禁止自主武器的推动作用。2013年5月30日，联合国人权理事会“法外处决、即审即决或任意处决问题”特别报告员克里斯托夫·海恩斯（Christof Heyns）向联合国大会强调，应加紧考虑机器人技术的发展和对法律、伦理和道德的影响[29]。2016年2月4日，克里斯托夫·海恩斯与和平集会和结社自由权利特别报告员麦纳·凯（MainaKiai）向联合国人权理事会提交调查报告，建议禁止无需人类有效控制的自主武器系统[30]。

2.3 技术发展规制的道路：知识共同体规范技术发展的专业治理

“人工智能向善”的共有观念已经形成，正发挥着延缓智能武器开发的作用。2015年1月12日，埃隆·马斯克和史蒂芬·霍金以及其他人工智能专家发表公开信强调，相关技术和产业研究不应只着眼于扩展人工智能的能力，还要考虑增加人类福祉[31]。这种共识影响企业研发的一个显著案例是，谷歌公司在压力下终止与美国国防部

的 Maven 项目合同。2018 年 3 月，谷歌公司被曝为美军研发军用无人机视频分析的 Maven 项目，此事引发部分谷歌公司员工不满，3000 多名谷歌公司员工联合发表公开信，敦促遵循谷歌座右铭“技术不能作恶”，要求尽快取消该项目并不再为美国防部提供开发人工智能武器的服务[32]。2018 年 6 月 4 日，谷歌公司云部门负责人黛安妮·格林(Diane Greene)迫于压力宣布，谷歌与美国防部的合约到期后将不签订新的合作计划。

共有观念起到规范国家发展人工智能武器的作用。技术民族主义者认为，企业有责任为国家研发高新技术武器，使本国有免于遭受威胁的自由。美国亚马逊公司的首席执行官杰夫·贝佐斯(Jeff Bezos)认为，如果美国的科技巨头都拒绝美国国防部，那么这个国家就会出安全问题[33]。从美国的尖端武器研发历史看，政府和私营部门合作制造了核武器、太空武器、网络武器。在人工智能研发方面，私营部门是最大的技术推进者。正因为此，美国防部的人工智能战略比以往任何时候更加依赖于与私营部门的合作关系。2020 年 2 月 25 日，美国防部正式采纳国防创新委员会建议的 5 项人工智能道德原则，为国防部弥合同私有部门的价值观分歧奠定了道德基础。美国军方和愿意为其开发武器的私营部门的合作关系不是上下隶属关系，而是了解和影响彼此观念体系从而互相妥协的共生过程。这一互动过程建立了对人工智能武器开发的共有观念，并规范着技术安全发展路径。

2.4　多路径演化的国际规范

在塑造人工智能的国际治理规范时，多元行为体如何获取和利用权力，权力政治和道德政治经历怎样的博弈过程，以及这些权力竞争如何影响国际规范演进，是下一步研究的方向。这些问题超出了本文的研究范围，只能留给未来的学术讨论。不过，要回答上述问题首先要明确存在怎样的规范演化路径，以及它们发挥影响面临的挑战是什么。

国家间战略互动有助于确立军事克制底线，维护全球战略稳定。鉴于人工智能军备竞赛已在某些领域展开，权力扩张的野望必然使大国手握进攻性无人武器。国家之间尽早采取军事克制措施，防止军备竞赛升级的重要性日益迫切。人工智能大国应该利用好当前的窗口机遇期，尽快划出战略竞争底线，达成军事克制协议或行为准则，启动信心建立措施。武器的先进程度不是保证国家安全的唯一来源，如果人工智能技术安全性提高，国家偏执部署自主武器则必然陷入某种安全困局，即政策制定者更有信心使用这些武器，进而刺激国家间军备竞赛。所以，如何平衡潜在的军事收益和可能成本，是政策制定者要思考的重要问题。

非政府组织主导推动一项禁武国际公约，约束国家使用武力。作为“伞”状的国际联盟，禁止杀手机器人运动提高了来自世界各地非政府组织对自主武器引发灾难性后果的认知，并促进各国政府在联合国主持下禁止致命性自主武器的努力。这些非政府组织借助媒体，向大众宣传禁武的法律和道德必要性，推动国际规范扩散。这种模式虽然在禁止地雷武器上有成功先例，但能否适用于致命性自主武器则要谨慎判断。地雷、集束炸弹、致盲激光武器等有明确的禁止对象，但致命性自主武器的定义存在广泛争论，难以明确禁止具体技术。致命性自主武器比地雷有更强的军事效能，对国家的军事安全诱惑力更大。所以，国家反对禁止致命性自主武器的声音要高于禁止地雷。地雷在世界各地造成大量人员死伤和致残的数据，有力地证明其使用有违人道，相关照片触动大众的内心，形成巨大的国内和国际禁雷压力。与之相比，缺乏自主武器导致人道主义灾难的明确证据，削弱了大众对该议题的关切程度和对政府的道义压力。

知识共同体规范技术发展表现出自生秩序的特点，将在国际治理中发挥两个重要作用：规避危险的技术路径，防止出现完全自主的致命性自主武器；明确被广泛接受的人工智能研发和应用原则，使国家在研发环节遵守相关法律和恪守国际道义。技术社群基于人类对生存的美好期待和道德伦理的坚守，本能地拒绝人工智能武器化，具有参与禁止自主武器的国际活动的动机。在大国争相部署自主武器的现实面前，民族主义和资本的结合必定打破知识共同体的统一阵线，使部分研发人员支持政府的军事项目。军事组织若要获得开发者的理解和支持，就必须正当化自己的行为。所以，政府和知识共同体的互动会影响彼

此的安全观念，共同规范技术的发展路径。知识共同体的观念多元性反映和建构了自生秩序的路径弹性，这种弹性让它可以推动不同甚至是相互矛盾的治理议题。

当下，人工智能国际规范的演进表现出了多条路径主导和多元网络化的特点，这项新兴技术的国际规范将沿着何种路径发展并不确定。国家最终接纳或内化何种规范，很大程度上由权力政治和道德政治的博弈结果决定。非政府组织禁止完全自主的致命性自主武器的设想和军控主张有相通之处，即希望避免人工智能武器化动摇战略稳定。这说明不同的治理路径之间并不必然相互矛盾，它们既存在相互协调合作的空间，也有相互对立的主张。既有研发人员出于民族主义感情参与制定国家武器研发计划，也有技术人员与国际组织合作反对人工智能武器化。持有不同信念的主体坚信自己的主张并付诸实践，导致不同的行为体之间形成了复杂的相互影响关系，没有一种治理主张可以在现实中起到绝对的主导作用。

3　总结及启示

核武器在被创造和使用之初就展现了惊人的破坏力，对战略稳定产生深远影响，改变了国际安全范式。与之相比，人工智能是一项正在发展的赋能技术，必须与网络技术、机器人、导弹等结合才能掀起智能战场革命。这场军事变革能否出现，要多久才能向世人昭示其全貌尚不得而知。有研究基于人工智能扩散的趋势，认为它将破坏军备稳定性。还有研究利用战争推演断定，它使安全危机以超越人类掌控的速度升级，令国家陷入一场不可避免的战争。这些研究丰富了安全困境、军备竞赛、威慑、错误认知等安全理论的解释力，为维护战略稳定的现实政策提供了新知识。但是，军事部门顾虑技术不成熟而出现武器失灵的情况，保守的战略文化以及和平时期冗长的制度审查阻碍人工智能武器系统的部署，削弱了人工智能的现实影响，这无疑使人工智能冲击战略稳定的解释力变得薄弱。目前来看，持有不同观点的各方恐怕很难相互妥协，因为他们大多从安全理论角度进行探讨，缺乏经验性研究。随着人工智能技术发展和军事应用普及化，学术界需要进行更多的经验性研究，讨论人工智能对战略稳定的影响。

达成人工智能治理规范的共识恐怕需要多方漫长且艰苦的努力，人工智能的国际安全治理进程可能遭遇知识供给不足的挑战。以国际社会讨论最为激烈的自主武器军控问题为例，自主武器缺乏被广泛认可的定义，因而国际组织无法清楚地要求国家禁止发展哪些技术，国际组织主张禁武也很难有政策操作性。如果贸然采纳国际组织全面禁止自主武器的建议，则必然会阻碍人工智能民用技术的发展以及其刺激数字经济增长的效果。国家可以主张技术发展会让自主武器更好地遵守国际法，或者将自主武器部署在远离人类的深海，削弱道德政治的压力。国家还能够发布人工智能治理原则，并对外做出军事克制承诺，减少国际道德的指责。这些政策选择的潜在含义是，大国对国际规则主导权的争夺不仅是对国际组织推动国际禁武议程的反应，也是国际竞争战略的一部分，这一点尤为明显地反映在中美竞争人工智能科技领导权方面。由此来看，道德政治与权力政治的博弈过程，以及大国的规则制定权之争是接下来国际规范研究的方向。

人工智能是一项不断发展的技术，与其忧虑不断变化的未来，不如根据武器的使用场景探讨治理方法。人工智能武器会否使攻击禁区目标成为可能，且攻击的方式不会从本质上违反比例原则和区分原则？根据战争法，武装攻击应该排除会对平民造成巨大伤害的水电站、大坝等禁区目标。但是，人工智能武器有可能在攻击基础设施附近的驻军时造成大坝崩塌的后果，致使处于下游的平民遭受巨大的财产和生命损失。关于人工智能安全治理的政策讨论，一种切实的方向可能是根据这些现实中可能发生的场景达成国际协议。战争本身充满欺骗，国家可以在不被发现的前提下攻击对方的传感器，污染敌方数据，把自己的军用设施标注为民用设施避免打击。如果这种行为被国际协议或国际法认定为非法，如何采取验证措施？对于这一难题，恐怕不仅需要学术界予以更多的关注找到破解之法，也需要决策者以卓越的智慧和“人类命运共同体”的理念推动维护国际稳定的全球议程。在百年未有大变局的当下，中美两国政府在未来有必要开展人工智能对话，

一个可能的议题是，双方都认可不将核武器指挥控制系统交给人工智能。但是，如何使中美相互之间共享基本的安全承诺信心，正遭遇两国负面战略互动持续加剧的重大挑战。

参考文献：

[1]AYOUB K,PAYNE K. Strategy in the age of artificial intelligence[J].Journal of strategic studies,2016,39 (5-6):794.

[2]JOHNSON J. Artificial intelligence and future warfare:implications for international security[J].Defense and security analysis,2019,35 (2):150-157.

[3]SLIJPER F,BECK A,KAYSER D. State of AI:artificial intelligence,the military and increasingly autonomous weapons[R].PAX,2019:4-5.

[4]朱启超,陈曦,龙坤．无人机作战与纳卡冲突[J].中国国际战略评论,2020 (2):167-183.

[5]HUNTER A P,SHEPPARD L R,KARLEN R,et al. Artificial intelligence and national security:the importance of the AI ecosystem [R].United States:Center for Strategic and International Studies,2018:45.

[6]SCHARRE P. Killer apps[J].Foreign affairs,2019,98 (3):135-144.

[7]LI Zheng.How to relieve the security dilemma of Sino-US technology competition[EB/OL].[2019-10-30].https://www.chinausfocus.com/peace-security/how-to-relieve-the-security-dilemma-of-sino-us-technology-competition.

[8]吉斯特·爱德华,安德鲁·洛恩．人工智能对核战争风险意味几何[R].兰德公司,2018 年.

[9]陈琪,朱荣生．为何担心人工智能冲击国际安全[J].人民论坛,2020 (8):124-127.

[10]ALTMANN J,SAUER F. Autonomous weapons and strategic stability[J].Survival,2017,59 (5):121-127.

[11]MARCUM R. Rapid broad area search and detection of Chinese surface-to-air missile sites using deep convolutional neural networks[J].Journal of applied remote sensing,2017,11 (4):1.

[12]FREEDBERG S. No AI for nuclear command & control:JAIC′s Shanahan[EB/OL].[2019-09-25].https://breakingdefense.com/2019/09/no-ai-for-nuclear-command-control-jaics-shanahan/.

[13]AI principles:recommendations on the ethical use of artificial intelligence by the department of defense[R].United States:Defense Innovation Center,2019:7-10.

[14]OSBORN K. Artificial intelligence is going to make America′s F-35 and B-2 even stronger[EB/OL].[2019-12-15].https://nationalinterest.org/blog/buzz/artificial-intelligence-going-make-americas-f-35-and-b-2-even-stronger-104967.

[15]傅莹．人工智能对国际关系的影响初析[J].国际政治科学,2019,4 (1):7.

[16]UN secretary-general′s high-level panel on digital cooperation. The age of digital interdependence.[EB/OL].[2019-06-15].https://www.un.org/en/pdfs/DigitalCooperation-report-for%20web.pdf.

[17]GUTERRES A. Address to the 74th session of the UN general assembly[EB/OL].[2019-09-24].https://www.un.org/sg/en/content/sg/speeches/2019-09-24/address-74th-general-assembly.

[18]ZHU Q,LONG K. How will artificial intelligence impact Sino-US relations[J].China international strategy review,2019,1 (1):139-151.

[19]LEE J J,HAUPT J P. Winners and losers in US-China scientific research collaborations[J].Higher education,2020,80 (1):57.

[20]傅莹．人工智能给我们带来的挑战．[EB/OL].[2019-12-17].https://www.un.org/en/pdfs/DigitalCooperation-report-for%20web.pdf.

[21]STONE P,BROOKS R,BRYNJOLFSSON E,et al. Artificial intelligence and life in 2030:the one hundred year study on artificial intelligence[R].United States:Standford University,2016.

[22]SCHMIDT E,WORK B,CATZ S,et al. National security commission on artificial intelligence interim report[R].United States:National Security Commission on Artificial Intellegence,2019:32.

[23]SLAYER K M. Artificial intelligence and national security[R].United States:Congressional Research Service,2019:18.

[24]ALLEN G,CHAN T. Artificial intelligence and national security[R].Cambridge:Belfer Center for Science and International Affairs,2017:56.

[25]GILL A S. Artificial intelligence and international security:the long view[J].Ethics and international affairs,2019,33 (2):174.

[26]KRISHNAN A. Automating war:the need for regulation[J].Contemporary security policy,2009,30 (1):187-188.

（下转第 179 页）