2023 年第 2 期（总第 35 期）

# 国际战略与安全研究报告

# INTERNATIONAL SECURITY AND STRATEGY STUDIES REPORT

## 人工智能全球安全治理的发展态势与新动向

清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

# 人工智能全球安全治理的发展态势与新动向

朱荣生、苏适[①]

2022年底，ChatGPT凭借强大的信息收集和处理能力而爆火“出圈”，社会上掀起了对人工智能技术发展和应用风险的热烈讨论。这其中不乏对ChatGPT在军事上对作战行动影响的讨论。2023年3月14日，俄罗斯的一架苏-27战斗机与一架美军地MQ-9“死神”无人机发生碰撞，美军无人机受损并最终坠入黑海海域。该事件似乎验证了关于未来战场有人和无人装备将更加频繁互动的观点。这在一定程度上强化了国际社会对自主武器和人工智能军事化开展全球安全治理的关切和推动意愿。有鉴于人工智能军事化风险及其全球安全治理的讨论已成为一项热点话题，本文将重点分析人工智能全球安全治理的发展趋势，并结合近期荷兰主办全球军事领域负责任的人工智能峰会的主要观点进行评论。

## 一、人工智能全球安全治理的演进与挑战

随着人们对以致命性自主武器为代表的无人装备的安全担忧日甚，国际社会对管控相关技术的意愿愈发迫切。从当前的全球安全治理主张来看，有一些人极力推崇达成一项全面禁止致命性自主武器的国际协议；另一些人则认为应当确立武器研发、部署和使用的“良好

① 朱荣生，清华大学战略与安全研究中心特约专家；苏适，长安大学建筑工程学院本科生。

本文是国家社会科学基金“人工智能的国际安全挑战与规范演进研究（项目编号：21CGJ012）”的阶段性成果。

实践”。在全球安全治理路径争论不休的情况下，美欧等发达国家却把“软规则”建构视作建立自身发展优势的工具，利用其在治理机制中的主导地位建构排华“小圈子”。与此同时，推动人工智能安全治理的知识生产面临着供给不均衡、不充足的挑战。

### （一）取向多元化：对致命性自主武器进行管控争论不休

国际社会在如何管控致命性自主武器的议题上出现了“全面禁止”和“有限发展”的尖锐分歧。一方面，较为激进的观点主张全面禁止致命性自主武器的研发、部署和应用。人权观察组织、红十字国际委员会等国际组织不断发布研究报告，期望在国际社会形成禁止致命性自主武器的舆论压力。[②]它们呼吁能够倡导达成一项具有约束力的国际条约，尽早对致命性自主武器开启严格的军备控制。在2022年7月的《特定常规武器公约》政府专家会议上，阿根廷等十个人工技术弱国坚定地反对在国际场合讨论军事无人化的好处，联合发布了禁止致命性自主武器的路线图，这预示着中小国家将形成强化全面禁武主张的核心阵营。它们强调要尽早采取的预防措施是禁止发展、交易、部署和使用致命自主武器，从而彻底排除自主武器系统对人类造成巨大伤害的可能性。[③]

另一方面，美国拉拢盟友在2022年7月的《特定常规武器公约》政府专家会议上发布《致命自主武器系统领域新兴技术的原则和良好实践》，提出武器发展的全生命周期只要符合相关规制便是所谓的“良好实践”。发达国家联盟强调“良好实践”的原因是，它们认为完

② Peter Asaro, “On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making,” *International Review of The Red Cross*, Nov. 2012;Peter Asaro, “On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making,” *International Review of The Red Cross*, Nov. 2012

③ 朱荣生、乔光宇、苏适：《致命性自主武器全球治理中“人类控制”的联盟政治》，清华大学战略与安全研究中心，2023年第1期。

全自主的致命性自主武器还没有出现，当前的致命性自主武器处在“有意义的人类控制”之下，就算致命性自主武器还存在一些技术风险，这些风险也会随着科技的进步而得到缓解。④上述针锋相对的立场呈现出国际社会对规制致命性自主武器不同的路径和最终指向，而这种分歧恐怕在短期内是难以化解的。

### （二）机制集团化：全球安全治理结构正趋向西方主导的政治“中心化”

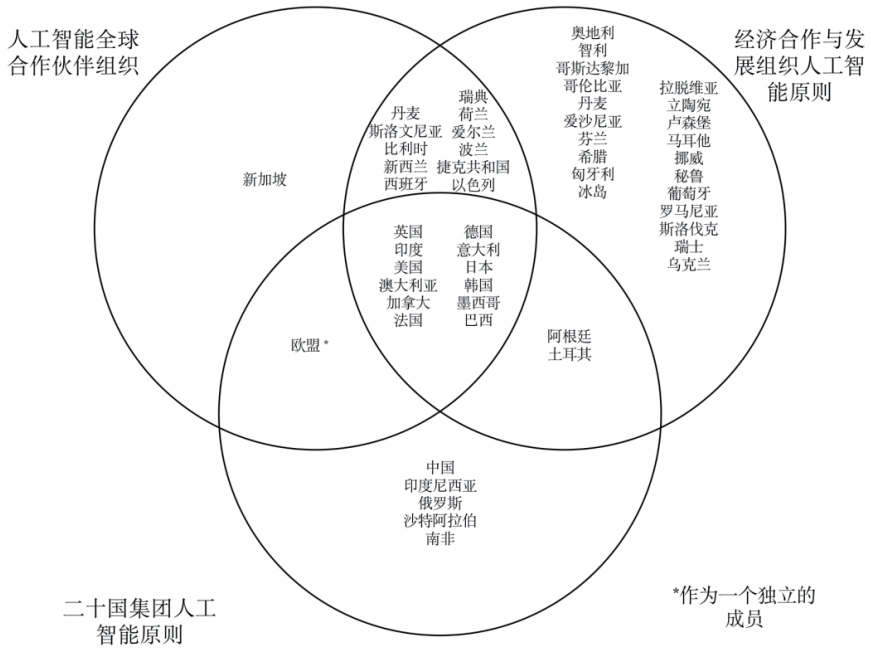



图1：人工智能全球治理主要机制

经济与合作发展组织、二十国集团以及人工智能全球伙伴组织是具有一定代表性的人工智能安全治理多边机制。从图1可看出，七国集团成员、澳大利亚、印度等等是这三个国际论坛的核心成员，中国与俄罗斯两个大国只是“二十国集团人工智能原则”的参与国。尽管

④ 朱荣生、乔光宇、苏适：《致命性自主武器全球治理中“人类控制”的联盟政治》，清华大学战略与安全研究中心，2023年第1期。

“中心化”的国际治理结构更有利于反应某些国家的关切、凝聚中心参与者的共识，但是这难免会让其他成员的治理主张边缘化。尤其是处于核心地位的美国正利用其在国际平台中的“核心地位”打造民主/威权集团对抗的治理结构，增加了不同成员推动人工智能全球安全治理合作的困难。

一是把美国的价值观嵌入“二十国集团人工智能原则”的文本中。2019年5月，经济合作与发展组织公布了经多个国家和国际组织研讨后的人工智能原则建议。在该原则通过后，特朗普政府的白宫首席技术官迈克尔·科雷特西奥斯（Michael Kratsios）指出，“历史上第一次，美国和世界上志同道合的民主国家将致力于共同的人工智能原则，来反映我们共同的价值观和优先事项。”[⑤]从结果来看，经济合作与发展组织人工智能原则的最终文本基本与《美国人工智能倡议》行政令的基调一致。[⑥]该原则建议后被二十国集团正式采纳为其人工智能原则。

二是美国加强与北约就人工智能军事应用的协调，逐步吸收新兴经济体并完善五眼联盟、四边安全对话、“人工智能国防伙伴关系”等机制，不断将人工智能议题纳入到遏华的亚太“小多边”安全体系中。这些行动的官方基调是西方需要共同捍卫芯片供应链安全，确保标准制定权不落入中国手中，以及协调集体安全军事安全行动。北约人工智能战略着力推动人工智能互操作性、技术标准对接、人力资本发展等；五眼联盟成员将制定解决人工智能应用和互操作性的方法，包括与北约一起搭建测试和应用人工智能的平台；四边安全对话将在美国、日本、印度、澳大利亚四方合作框架的基础上深化人工智能合作，并在印度–太平洋地区谈判达成正式的人工智能合作协议。由此

⑤ https://www.whitehouse.gov/ai/ai-american-values/.

⑥ 李括：《美国科技霸权中的人工智能优势及对全球价值链的重塑》，《国际关系研究》2020年第1期，第41页。

来看，美国正将人工智能治理议题安全化，将盟友推上新一轮军事竞争和规则制定权争夺的轨道。

三是美国聚拢西方国家搞排斥中国的新制度，拉拢盟友力推“人工智能全球合作伙伴”，肆意制造“中国滥用人工智能技术”的论调和阵营对抗。七国集团在2016年4月的信息通信技术（ICT）部长会议讨论了人工智能技术及应用监管议题。从2016年到2019年，七国集团（G7）以人工智能负责任应用及监管为核心的议题逐渐成熟。法国和加拿大倡议组建工作小组，并希望在2020年初正式启动人工智能全球合作伙伴（GPAI）。[⑦]白宫官员则对此方案持保留意见，强调该计划过度谨慎的监管规则会威胁和妨碍技术发展，并且认为经济合作与发展组织已经建立专家组并向成员国提供政策建议，所以人工智能全球合作伙伴是在重复经济合作与发展组织的工作。[⑧]法国数字事务部长塞德里克·奥（Cédric O)试图说服迈克尔·克拉西奥斯，表示，“如果你不希望西方国家采用中国模式，比如用人工智能来控制人口，你就需要建立一些普遍的规则。但这只针对一个国家”。考虑到继续反对该计划恐将让美国在西方集团内失去人工智能治理的规则主导权，而加入其中则可以借此扭曲中国的人工智能应用，美国最终决定支持启动人工智能全球合作伙伴。

### （三）知识供给匮乏：人工智能安全治理知识供给不均衡、不充足

先进人工智能技术的创新和大规模应用集中于少数大国。全球人工智能的研究在2020–2021年间共出版17万本期刊论文和七万部会议论文，中国、美国、欧盟和英国的出版物占据了其中的绝大多数。[⑨]

---

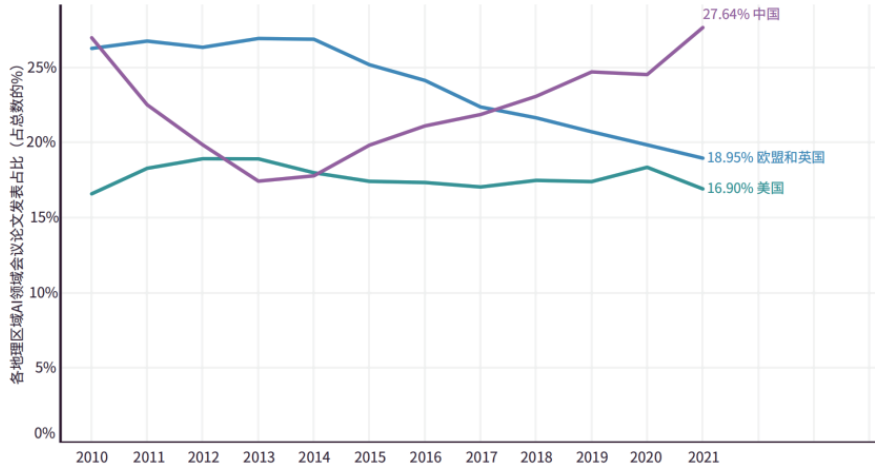
⑦ “The World Has a Plan to Rein in AI—But the US Doesn’t Like It,” *Wired*, Jan. 6 2020, https://www.wired.com/story/world-plan-rein-ai-us-doesnt-like/.

⑧ Ibid.

⑨ 杰克·克拉克等：《人工智能指数2022年度报告》，2022年，https://aiindex.stanford.edu/wp-content/uploads/2022/06/2022-AI-Index-Report_Chinese-Edition.pdf

纵观全球军用人工智能发展格局，有能力大规模研发、制造、部署和使用先进军用人工智能也只有大国。或许，从人工智能技术在俄乌冲突中的应用、伊朗核科学家和苏莱曼尼将军被“定点清除”、无人机助力阿塞拜疆取得“纳卡冲突”制空权等战例来看，相关武器的购买成本不是高不可攀。但是，大国在开发和评估新武器系统方面有更强大的能力，而弱国采购进行评估和创新的能力较弱。仅从目标识别环节来看，军用人工智能系统需要一个国家有强大的情报侦察能力以提供规模大和质量好的数据。而拥有更多的研发资金、科研人力和实战化训练的大国能够比中小国家获取更多、更好的战场数据。因此，大国可以开发出更加先进的算法和硬件。这意味着中小国家推动军备智能化升级和技术评估的天花板要低于大国，它们的人工智能安全治理的知识积累要弱于强国。[⑩]

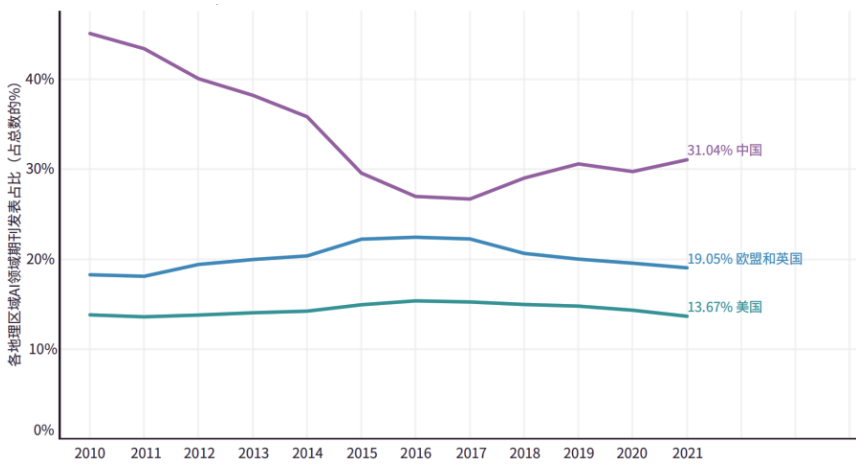
表1：2010–21年人工智能期刊出版量（占世界总量的%）[⑪]




⑩ 朱荣生、陈嘉澍、冯紫雯、陈琪：《人工智能军事目标识别的赋能效度和合规风险初探》，《计算机工程与应用》，2022年8月15日，第58卷。

⑪ 杰克·克拉克等：《人工智能指数2022年度报告》，2022年，https://aiindex.stanford.edu/wp-content/uploads/2022/06/2022-AI-Index-Report_Chinese-Edition.pdf

表2：2010–21年人工智能会议出版物（占世界总数的%）情况[12]




## 二、峰会关于人工智能军事化及其治理的主要观点

2023年2月15–16日，荷兰主办全球军事领域负责任的人工智能峰会（REAIM Summit）（以下简称“峰会”）。围绕人工智能的技术特性，负责任地开发和使用人工智能，以及探索人工智能治理框架等三个主题，峰会吸引了来自100个国家的2000余名参会者，以及80名政府代表的参加。60多个国家赞同了峰会发布的“行动倡议”。在这场后疫情时代关于军用人工智能治理的线下峰会中，与会代表普遍对人工智能的军事应用抱有较高期待，认为它将引发新一轮的战场革命，但也担忧对技术的滥用恶用将加剧国际安全的不确定性。总体看，会议讨论的内容反映出当前人工智能军事化及其治理的一些新动向。

### （一）人工智能将对军事领域产生颠覆性影响

在峰会的高端会议上，美国、芬兰、澳大利亚、立陶宛、马来

⑫ 杰克·克拉克等：《人工智能指数2022年度报告》，2022年，https://aiindex.stanford.edu/wp-content/uploads/2022/06/2022-AI-Index-Report_Chinese-Edition.pdf

西亚等国与会者强调，人工智能是一项强大的赋能技术，将在军事领域掀起新一轮变革。荷兰外交部部长沃普克·霍克斯特拉（Wopke Hoekstra）断言，人工智能有潜力彻底改变取得战争胜利的方式。韩国外交部长朴振表示，军事智能化有利于提高作战反应能力和减少误伤平民的可能性。他强调，韩国国防部推出“国防改革4.0”的一项重要目标是，将人工智能深度融入到军备建设。鉴于对人工智能技术增强国防预期的固化，世界主要国家将加速军事智能战略规划的制定和落实。

部分产业界代表也认为，人工智能军事化会提高国家安全。前谷歌执行董事长埃里克·施密特（Eric Schmidt）表示，美军推动军事智能化的脚步曾经较为缓慢，军工复合体的技术开发速度不如私营部门。但是，他认为美国在加速推动人工智能军事化方面的努力有利于增强国防实力。空中客车公司“未来空战系统”(FCAS)项目负责人布鲁诺·菲舍弗（Bruno Fichefeux）强调，人工智能在国防建设和武力使用中将扮演重要角色，它可以让指挥官以更快的速度做出更加合理的军事决策。

尽管峰会上的发言者总体上对人工智能的军事效能有较高的期待，但是一些军事安全研究者则态度保守，将人工智能视为既有武器的“放大器”。[13]此类分析认为，人工智能的军事应用和效能还没有全面铺开，目前缺少足够的有力证据支持它将对战场产生颠覆性效应。还有研究对军事智能和核武器进行对比后认为，核武器在使用之初就展现出了惊人的破坏力，刺激了大国间战略稳定理论的创新，而人工

⑬ James Johnson, “Artificial Intelligence and Future Warfare: Implications for International Security,” *Defense & Security Analysis,* Vol. 35, No. 2, 2019, pp. 147-169; Greg Allen and Taniel Chan, “Artificial intelligence and national security,” *Cambridge Belfer Center for Science and International Affairs*, July 2017.

智能则是一项有军事潜力的赋能技术。[14]而更为悲观的观点则指出，人工智能技术若得不到足够的信任就很难得到军事部门的认可。[15]

### （二）军事智能化加剧人道主义风险和军备竞争压力

在峰会的开幕式论坛中，“大赦国际”秘书长阿涅斯·卡拉马德（Agnès Callamard）极力主张尽早对人工智能军事应用开展严格的军备控制。她强调战争是肮脏的，时常伴随着人道主义灾难的出现。荷兰的恩诺·艾切尔斯海姆（Onno Eichelsheim）中将认同“战争是肮脏”的观点。不过，他反驳称，在竞争对手加速军事智能化迭代的情景下，荷兰将不得不做出相同的选择来保障国家安全。其中的政策含义是，技术发达国家没有很强的意愿支持率先禁武的军控主张。澳大利亚外长黄英贤（Penny Wong）赞同人工智能具有较强的军事变革潜力。但是，她担忧地表示印太地区正面临不断上升的军备竞争压力，呼吁国际社会推动对这项新兴技术进行规制，避免加剧区域安全的不稳定性。上述争论的核心关切是，国家面临生存压力而开启军备竞争和恶用滥用军事智能技术对平民造成死伤。

在峰会上，有代表希望将达成国际禁雷公约和国际禁止集束炸弹公约的成功经验借鉴到推动人工智能军备控制的进程中。地雷对平民和排雷人员造成的不幸伤亡已经被大量披露，形成了反对地雷的国际舆论压力，但是禁止人工智能军事化似乎并没有形成充足的民意基础。[16]当前，以致命性自主武器为代表的人工智能军事应用受到了广泛关注。根据益普索（IPSOS）在2021年对28个国家、地区关于致命

---

⑭ John Lewis, “The Case for Regulating Fully Autonomous Weapons,” *Yale Law Journal*, Vol.124, 2015.

⑮ “Artificial Intelligence and Life in 2030: the One Hundred Year Study on Artificial Intelligence,” *Stanford University*, Sep. 16.

⑯ Michael C. Horowitz, Julia M. Macdonald, “Will Killer Robots Be Banned? Lessons from Past Civil Society Campaigns,” *Lawfare*, Nov. 5, 2017, https://www.lawfareblog.com/will-killer-robots-be-banned-lessons-past-civil-society-campaigns.

性自主武器的态度的调查，有61.3%的受访者持有强烈反对的态度，有21.5%的受访者持支持态度；瑞典（76%）、土耳其（73%）和匈牙利（70%）的反对意见最为强烈；唯一获得多数支持的国家是印度（56%）；中国和美国的反对者分别占比42%和55%。[17]针对人工智能军事化风险的一项焦点议题是，它是否会在赋能军事平台后将非作战单位判定为攻击目标，进而实施打击对平民造成杀伤。谢菲尔德大学的诺尔·夏基（Noel Sharkey）认为，识别技术缺陷使人工智能赋能的作战平台不能区分合法和非法的攻击目标，这种缺陷有可能放大自主武器杀死平民的可能。相反的观点则认为，机器不会受到情绪的裹挟而对非战斗人员进行攻击，其打击精准度也高于人类操作员。[18]

### （三）人工智能安全治理方式的关注点趋向于规范军事行为而非规范军事装备

参与峰会讨论的官员和学者都认同有必要对人工智能军事化予以规制，避免其产生的潜在风险。现任联合国副秘书长兼裁军事务高级代表中满泉（Nakamitsu Izumi）在闭幕式发言上表示："传统的军备控制非常关注对能力或技术本身的规制。最近的军控讨论出现了新趋势。更加侧重于明确哪些是负责任的行为，哪些是不负责任的行为"。她声称这种军控方式是必要的，并且认可"负责任的人工智能"的框架。美国官方代表的政策宣介的核心主旨也是"反对不负责任的武力"，并且强调美国所发展的军事智能装备是合规的。这些观点的共通之处是更加强调对军事行为而非军事装备的规范。其原因可能有三方面。一是人工智能军备控制的知识供给不够充分，相关术语和需要被限制的技术界定不清，因而难以明确到底要控制哪些军备。二是技

⑰ https://www.ipsos.com/en-us/global-survey-highlights-continued-opposition-fully-autonomous-weapons.

⑱ "Task Force Report: The Role of Autonomy in DoD System," Department of Defense, *Defense Science Board*, July 2012.

术强国有可能希望在达成对军事装备进行控制的共识之前就确保自身的技术优势。三是打造“负责任的人工智能”的话语和技术体系有助于率先获得主导国际规范和规则的高地。

### （四）美国着力打造“负责任的人工智能”形象，争夺国际话语权

美军首席数字和人工智能办公室主任戴安·斯塔赫利（Diane Staheli）宣称，美国防部已通过六份建构“负责任的人工智能”的指导性文件。她认为，美军形成了包含技术研发、武器部署、系统测试、评估工具、人员教育、现实应用、法律责任等方面较为完善的监管体系。针对美国防部今年1月更新的3000.09指令，美国防部新兴能力政策办公室主任迈克·霍洛维茨（Michael Horowitz）声称，指令更新是为了建立人工智能和自主技术的军事行为规范，并且展现美国是一个有足够透明度的世界领导者。美方在峰会论坛的政策宣介中表明，美国想将自己打造成“负责任的人工智能”样板，进而提高在国际规则制定中的威望。

美国以意识形态划线对中国人工智能发展进行污名化。在峰会举办期间，一位美国防部高级官员向美媒“Defense One”表示：“我们现在有机会在某种程度上取得进展，建立负责任的行为规范……这对所有国家都有帮助，符合我们对国际人道主义法和战争法所作的承诺。中国和俄罗斯都没有公开声明它们所实施的程序，确保其军用人工智能系统能够安全可靠地运行。”客观而言，美国拥有领先的技术积累、丰富的管理经验、更多的实战化应用。所以，美军监管政策的制定会快于包括其盟友在内的其他国家。美军高官却刻意忽视现实情况，指责中国缺乏政策透明度和潜藏技术安全问题。这增加了中美在人工智能安全治理领域的合作难度。

发表日期：2023年3月23日

审编：肖茜　　签发：达巍

扫码关注我们

**清华大学战略与安全研究中心编印**

办公地点：北京市海淀区清华大学明理楼428房间

联系电话：010-62771388

http://ciss.tsinghua.edu.cn　邮箱：ciss@tsinghua.edu.cn