2023 年第 1 期（总第 34 期）

国际战略与安全研究报告

INTERNATIONAL SECURITY AND STRATEGY STUDIES REPORT

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
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

致命性自主武器全球治理中“人类控制”的联盟政治

朱荣生、乔光宇、苏适[①]

随着人工智能技术加速融入到武器系统和军事作战体系，人们愈发关注它将对国际和平与安全产生的重大影响。自2014年以来，国际社会在联合国《特定常规武器公约》框架下针对以致命性自主武器为代表的人工智能军事化的挑战和治理进行了激烈的辩论。在2016年，《特定常规武器公约》致命性自主武器政府专家组正式成立，并成为相关讨论的重要国际机制。尽管参与谈判的缔约国对这项新兴全球治理议题的理解越来越深入，但是在如何定义致命性自主武器以及人类对武力使用控制的方式和必要程度方面有不小的分歧。缔约国对于人类履行相关控制责任所涉及到的技术、操作、决策等方面有着不同的理解。[②]这在某种程度上可以归结为对致命性自主武器中“人类控制”议题的持久争论。围绕这一议题，国际社会形成了多个政治联盟。其中政治立场和治理分歧较为明显的是以美国为代表的发达国家集团和以十三国政治联盟为代表的发展中国家集团。[③]

---

① 朱荣生，启元实验室战略规划部青年研究员；乔光宇（Guangyu Qiao-Franco），拉德堡德大学国际关系助理教授、南丹麦大学AutoNorms项目高级研究员；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。

② “Areas of Alignment. Common Visions for a Killer Robots Treaty,” pp.12-17, https://www.hrw.org/sites/default/files/media_2021/07/07.2021%20Areas%20of%20Alignment.pdf.

③ 十三国集团的成员是阿根廷、哥斯达黎加、厄瓜多尔、危地马拉、哈萨克斯坦、尼日利亚、巴拿马、秘鲁、菲律宾、巴勒斯坦、塞拉利昂、委内瑞拉和乌拉圭。

基于政府专家组会议成员发布的文件、21次对不同国家代表团成员的访谈以及六次旁听政府专家组会议的记录，我们梳理了这两个具有巨大价值观差异和鲜明对立观点的联盟的安全关切和道德主张。我们发现，这种对“人类控制”概念的理解差异和不同期待深受国防研发能力、国际谈判资源、武力使用的历史经验等因素的影响。

一、“人类控制”议题下的政治联盟

自2016年政府专家组成立以来，相关讨论曾一度深陷对自主化、自动化、智能化等基本概念的循环争论。而“人类控制”概念的提出将国际争论的焦点引向了新的方向。尽管这一概念至今还存在模糊性，但这在某种程度上为各方提供了一个阐释不同立场的弹性空间。[④]在2019年政府专家组会议达成的指导性原则中就提出，要探索在不同的武器生命周期里融入人类控制的可能性。联合国裁军研究所（United Nations Institute for Disarmament Research）在2020年出版了一份报告，讨论在政治、战略、战术、作战等各个阶段加强人类的监管控制。同年6月，瑞典斯德哥尔摩和平研究所和红十字国际委员会从武器参数的控制、使用环境的约束、适当的决策监督等三方面提出了人类对致命性自主武器系统的监管建议。[⑤]这两份报告的内容折射出，相比于对技术概念的无尽争论而难以达成政治共识，人们对什么是“适当”的人类控制以及如何落实这一概念有着更大的探索动力。

---

④ Maya Brehm ,“Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapons Systems under International Humanitarian and Human Rights Law,” Geneva Academy of International Humanitarian Law and Human Rights, No. 9, 2017.

⑤ Boulanin, Vincent, Neil Davison, Netta Goussac, and Moa Peldán Carlsson, “Limits of Autonomy in Weapon Systems. Identifying Practical Elements of Human Control,” *Stockholm International Peace Research Institute*, June, 2020.

（一）以美国为代表的发达国家主张建立非约束性原则

在《特定常规武器公约》政府专家会议上，美国较早使用的概念是“人类判断”(human judgement)，而非“人类控制”。在美国2018年4月提交的立场文件中，“适当”的“人类判断”在某种程度上被宽泛地理解为不存在一个固定的、适用于所有情况的人类判断标准。⑥而美国认为，自主武器在“适当”的“人类判断”下可以起到更好的提高军事防御能力的作用。例如，“密集阵”近防武器系统、“宙斯盾”武器系统和“爱国者”防空导弹防御系统可以协助人类瞄准来袭炮弹。⑦由于公民社会组织和一些发展中国家愈发强烈地“禁武”呼声带来的武器发展压力，以美国为代表的发达国家逐渐接受“人类控制”这一概念。2021年6月，美国、英国、澳大利亚、加拿大和日本五国首次提交了联合立场文件，强调“人类判断”对确保致命性自主武器符合国际人道主义法至关重要。这标志着该联盟的初步形成。⑧

美国等发达国家组成的联盟主张对致命性自主武器建立一个相对宽松的、自愿性的国际规制清单。2022年3月7日，美国牵头澳大利亚、加拿大、日本、韩国、英国制定了《致命自主武器系统领域新兴

⑥ The United States. (2018, April 9–13). Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems. *U.S. Delegation to the CCW*. https://ogc.osd.mil/Portals/99/Law%20of%20War/Practice%20Documents/US%20Working%20Paper%20-%20Humanitarian%20benefits%20of%20emerging%20technologies%20in%20the%20area%20of%20LAWS%20-%20CCW_GGE.1_2018_WP.4_E.pdf?ver=O0lg6BIxsFt57nrOuz3xHA%3D%3D, para. 9-11.

⑦ Ibid.

⑧ Building on Chile’s Proposed Four Elements of Further Work for the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS), https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2021/gge/documents/Australia-et-al.pdf.

技术的原则和良好实践》。[⑨]该份文件提出了发展致命性自主武器的全生命周期只要符合相关规制便是所谓的“良好实践”。这些“良好实践”要求确保武器从研发到测试和运行都要有人类的监管，这些监管要符合国际人道主义和安全政策要求。而且，武器的使用者也必须接受相应的培训。发达国家联盟强调“良好实践”的原因是，他们认为完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。

（二）十三国集团力推达成有约束力的国际公约

较为悲观的观点认为，致命性自主武器的现实应用很难保证不会冲击交战规则，因而应尽早达成一项有约束力的国际公约以管制自主武器系统的研发、部署和应用。相关实践方面，人权观察组织、红十字国际委员会等国际组织加入“禁止杀手机器人”运动，并且极力倡导达成一项具有约束力的国际禁武条约。2012年11月9日，人权观察组织发布报告认为，自主武器从根本上违反国际人道主义法和战争法的精神，建议国际社会不仅要达成有国际法约束力的文书，各国还要制定法律在其大规模应用前进行全面禁止。[⑩]一些国际组织还不断细化禁武标准。2022年6月29日，“禁止杀手机器人运动”发布报告强调，不只是国家要停止生产、占有、获得、部署和转让致命性自主武器，研发人员、工程师、制造商、供应商也要被纳入到禁武条约的监管范围内。[⑪]

---

⑨ Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States, “Principles and Good Practices on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems,” March 7, 2022, https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2022/gge/documents/USgroup_March2022.pdf.

⑩ Bonnie Docherty, “Losing humanity: The case against killer robots,” *Human Rights Watch*, 2012, p. 5.

⑪ “Negotiating a Treaty on Autonomous Weapons Systems - The Way Forward,” June 29, 2022, https://www.stopkillerrobots.org/wp-content/uploads/2022/06/Stop-Killer-Robots-Negotiating-a-Treaty-on-Autonomous-Weapons-Systems-The-Way-Forward.pdf.

另一方面，技术弱国与公民社会组织有相似的主张。随着对致命性自主武器的认知程度逐渐提高，它们更加坚定地反对在国际谈判场合讨论军事无人化的好处，并且联合发布禁止致命性自主武器的路线图。古巴认为，《特定常规武器公约》不是考虑无人武器系统好处或军用人工智能优点的地方，而是要禁止和管理某些常规武器；巴基斯坦也认为，《特定常规武器公约》旨在结束军备竞赛、实现裁军以及编纂和发展相关国际法。⑫它们的核心主张是，《特定常规武器公约》是裁军平台，而非探讨武器发展带来战争优势的地方。

部分发展中国家强调，不论机器有多么完善，都不能将致命权交给机器。一些发展中国家自2017年以来依托不结盟运动推动致命性自主武器的治理进程，但代表成员的复杂性让众多成员难以发出统一的声音。然而，哥斯达黎加、巴拿马、秘鲁、菲律宾、塞拉利昂共和国和乌拉圭不满于不结盟运动在推动禁止致命性自主武器上的低效，在2021年6月的非正式谈判会议中以联合递交立场文件的方式组成了更为激进的“禁武”联盟。在2022年7月的《特定常规武器公约》政府专家会议上，阿根廷、哥斯达黎加、危地马拉、哈萨克斯坦、尼日利亚、巴拿马、菲律宾、塞拉利昂、巴勒斯坦、乌拉圭等十个国家联合递交了《自主武器系统新协议路线图》和《特定常规武器公约第六议定书》。⑬到了2022年12月，该联盟的数量达到了13个，正式形成

⑫ Ray Acheson, “CCW Report”, Vol. 10, No. 6, https://reachingcriticalwill.org/disarmament-fora/ccw/2022/laws/ccwreport/16301-ccw-report-vol-10-no-6.

⑬《特定常规武器公约》已有五个公约书分别是《关于无法检测的碎片的议定书》《禁止或限制使用地雷(水雷)、饵雷和其他装置的议定书》及其《技术附件》《禁止或限制使用燃烧武器议定书》《激光致盲武器议定书》《战争遗留爆炸物议定书》。“Roadmap Towards New Protocol on Autonomous Weapons Systems,” https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2022/gge/documents/G13_March2022.pdf; “Protocol VI,”https://documents.unoda.org/wp-content/uploads/2022/07/WP-Argentina_Costa-Rica_Ecuador_Nigeria_Panama_Philippines_Sierra-Leone_Uruguay.pdf.

了十三国集团。在“人类控制方面”，十三国集团认为要通过禁止研发和使用致命性自主武器的方式来避免武器瞄准人类的风险；必须要有明确的法律规定来保障自主武器有足够的可预测性、可靠性、可解释性，从而落实有意义的人类控制，避免出现法律漏洞和技术安全不足。总之，这些国家强调要尽早采取禁止发展、交易、部署和使用致命自主武器的预防性措施，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。⑭

二、“人类控制”治理观念分歧的原因

针对致命性自主武器“人类控制”的讨论，为何会产生两种截然相反的治理立场？如果仅从“技术决定论”的视角去理解这种分歧，难免忽视了政治、经济、文化、制度等诸多社会变革要素对技术创新和应用的影响。⑮基于对一线谈判人员的访谈和参与政府专家组会议的记录，我们发现不同的国防研发能力、国际谈判的可用资源、武器试用和使用的历史经验是分歧产生的重要原因。

（一）国防研发能力差距拉大加剧治理立场分歧

发达国家有推动军用技术创新和维持强大军力的利益。自上世纪70年代以来，美国及其军事盟友就对军用技术的自动化和人工智能军

⑭ Joint Working Paper Submitted by the Republic of Costa Rica, the Republic of Panama, the Republic of Peru, the Republic of the Philippines, the Republic of Sierra Leone and the Eastern Republic of Uruguay, June 2021, https://documents.unoda.org/wp-content/uploads/2021/06/Costa-Rica-Panama-Peru-the-Philippines-Sierra-Leone-and-Uruguay.pdf.

⑮ Jasanoff, Sheila, *Science at the Bar: Law, Science, and Technology in America*, Cambridge, MA: Harvard University Press, 1995; Collins, H. M, and Robert Evans, *Rethinking Expertise*. Chicago: University of Chicago Press, 2017; Bijker, Wiebe E., and John Law, eds., *Shaping Technology/Building Society: Studies in Sociotechnical*, Change. Cambridge, MA: MIT Press, 2010.

事化投入了大量资金。[16]在承认先进武器具有某种程度不可预测性的同时，发达国家认为它们已经部署的自主武器和正在研发的智能武器可以确保有效的人类监管。美国相信，致命性自主武器在人工智能的赋能下可以加速情报收集和分析、开展更加精准和快速的打击行动，并且减少作战人员的伤亡。[17]同时，对于人口可能越来越短缺的发达社会来说，设立严苛的限制自主武器相关技术发展的国际公约显然不利于解决军事作战人员成本高的问题。

对于技术实力不足的国家而言，它们恐将越发难以追赶发达国家的技术能力，因而会有更强的动力推动有约束力的条约来限制致命性自主武器的开发。先进且可靠的军事智能技术需要特定的训练数据、大量的研发资金以及与之相匹配的作战平台。这些“门槛”可能会限制技术弱国的国防研发能力，也会阻碍国际军控知识的创造和供给。处于弱势的发展中国家会很自然地选择与发达国家对立的立场。

（二）国际谈判资源的可用性越低，主张禁武的意愿就越强

国际谈判资源的不足限制了技术弱国在致命性自主武器“人类控制”议题上的论辩能力和意愿。在我们的访谈中，受访者频繁提到致命性自主武器的相关讨论涉及复杂的伦理、法律、技术问题。这意味着缺少谈判人员、技术知识和财政支撑的欠发达国家在回答什么是“适度的”“人类控制”的问题上很难给出有效方案。非政府组织第36

⑯ Haner, Justin, and Denise Garcia,“The Artificial Intelligence Arms Race: Trends and World Leaders in Autonomous Weapons Development,”*Global Policy*, Vol. 10, No.3, 2019,pp.331-337; Bode, Ingvild, and Hendrik Huelss, “Autonomous Weapons Systems and Changing Norms in International Relations,” Review of International Studies Vol.44, No.3, 2018, pp.393-413.

⑰ Mission of the United States,“CCW: U.S. Opening Statement at the Group of Governmental Experts Meeting on Lethal Autonomous Weapons Systems,”April 9, 2018, https://geneva.usmission.gov/2018/04/09/ccw-u-s-opening-statement-at-the-group-of-governmental-experts-meeting-on-lethal-autonomous-weapons-systems/.Mission of the United States,“3rd Meeting - 1st Session Group of Governmental Experts on LAWS 2021,”UN Web TV. August 4, 2021. https://media.un.org/en/asset/k1k/k1klog2whq.

条在国际裁军领域收集的数据显示："一个国家的收入越低，他们…出席、发言或者在会议上担任正式角色的可能性就越小。"⑱在我们参与的致命性自主武器专家会议中发现，一些欠发达国家在会议上的表现并不积极，有的国家派一名代表或者一名实习生代表，甚至根本不出席某些讨论。一位发展中国家的受访官员表示："对于许多国家来说，治理致命性自主武器不是优先项"。⑲这些发现表明，国际谈判资源越少的国家可能更无意参与到致命性自主武器"人类控制"的国际争论中。

为何技术弱国更倾向于全面禁止致命性自主武器？有受访官员认为，技术弱国对于发达国家肆意发展和扩散致命性自主武器有一种不满和失望的情绪，因而选择站在其对立面。⑳在这样的情绪和希望影响国际规则的愿望下，一些发展中国家不仅产生了结盟的偏好，也乐于与公民社会组织一道捍卫全面禁止致命性自主武器的主张。一些技术实力较弱的国家会主动接触公民社会组织，并委托这些组织人员担任其代表团的顾问。㉑由此来看，双方在国际禁武的道德政治上有着共同利益，而公民社会组织则成为了技术弱国的一种国际谈判资源。

（三）武器试验和使用的历史经验差别难以弥和

大国可以获取更多、更好的战场数据，以及开发出更加先进的算法和硬件。技术弱势国为了获取先进的军事技术会不得不从发达国家手中求购相关武器，进而越发依赖于发达国家的技术供应。阿根廷在《特定常规武器公约》会议上就指出，致命性自主武器需要强大的科学和技术能力，武器的生产国与进口国之间存在差距，前者具有评估

---

⑱ "Killer Robots: UK Government Policy on Fully Autonomous Weapons,"2013, http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

⑲ 对某国外交官员的访谈，日内瓦，2017年4月27日。

⑳ 对某国外交官员的访谈，日内瓦，2017年4月25日。

㉑ Ingvild Bode, "Norm-Making and the Global South: Attempts to Regulate Lethal Autonomous Weapons Systems," *Global Policy*, Vol.10, No.3, 2019.

新武器系统的强大能力，而进口国评估其采购后果的能力较弱。[22]同时，技术弱势国家还可能成为先进的致命性自主武器武器扩散的暴力“试验场”。

这种关切可以在灰暗的武器测试历史中寻得例证。1946 年的美国、1957 年的英国，到1966 年的法国都在南太平洋岛国在进行了核武器试验。这对生活在这些区域的公众健康和环境造成了严重后果。[23]2017年8月25日，时任俄罗斯武装部队军事科学委员会主席兼总参谋部副总参谋长伊戈尔·马库舍夫中将表示，俄罗斯在叙利亚测试的 200 多种武器在完成制定任务方面显示出高效能。[24]而那些被测试的武器包括较为先进的人工智能技术和自主武器。[25]这在一定程度上证实了巴勒斯坦代表团在《特定常规武器公约》政府专家会议上提出的担忧，即自主武器将由发达国家开发和部署，并在其他国家进行测试和使用。[26]

三、对致命性自主武器“人类控制”的思考

在地缘竞争摩擦加剧和对抗不断升温的境况下，凝聚军备控制的

㉒ Ray Acheson, CCW Report, Vol. 10, No. 3, 28 April 2022, https://reachingcriticalwill.org/disarmament-fora/ccw/2022/laws/ccwreport/16223-ccw-report-vol-10-no-3.

㉓ Fry, Greg, “The South Pacific nuclear-free zone: Significance and implications,” *Bulletin of Concerned Asian Scholars*, Vol. 18, No.2, 1986, pp. 61-72.

㉔ “Russia/Syria: More than 200 Weapons Tested in Syria Prove to Be Highly Effective,”*Russian News Agency*，August 28, 2017.

㉕ “Advanced military technology in Russia Capabilities, limitations and challenges,”Sep. 23, 2021, https://www.chathamhouse.org/2021/09/advanced-military-technology-russia/06-military-applications-artificial-intelligence

㉖ Ray Acheson, “CCW Report,” Vol. 10, No. 2, March 14, 2022, https://reachingcriticalwill.org/disarmament-fora/ccw/2022/laws/ccwreport/16025-ccw-report-vol-10-no-2.

政治共识将越发困难，达成一项有约束力的国际条约任重而道远。《特定常规武器公约》的《第四议定书》要求禁止使用以致人眼永久性失明为作战目的的激光武器。国际社会达成了这一有限禁武的国际条约可被看做是道德政治对权力政治的“胜利”。但是，这一限制范围有限，不能将这一成功案例作为推定致命性自主武器会被全面禁止的依据。例如，高能激光武器仍然有可能被用于攻击卫星。在十三国集团与公民社会组织强烈的道德呼吁下，一项对致命性自主武器的新议定书或有可能形成，但该协议恐怕难以实现对“人类控制”的严格要求。

建构人工智能及其赋能自主武器平台的国际治理规范应当倡导“伦理先行”。从上述两个政治联盟对加强“人类控制”的核心主张来看，它们对人工智能军事化和致命性自主武器快速发展都有着深刻的道德和安全关切。近年来，主要大国对确保人工智能中“人类控制”表现出一定的宣示姿态。2021年10月21日，北约防长会通过了首份《人工智能战略》，提出对人工智能程序的开发和使用要有人类的判断，明确对人追责的原则。2022年6月21日，美国防部发布《负责任的人工智能战略和实施路径》认为，“负责任的人工智能”包括人类在测试标准、问责审查、技术安全等方面的责任。2021年12月13日，中国向《特定常规武器公约》第六次审议大会提交的《中国关于规范人工智能军事应用的立场文件》强调，建立人工智能问责机制，对操作人员进行必要的培训。上述政策行动表明，主要大国和军事组织正在以软性的伦理规则强化对技术的监管。

“人类控制”概念可能在防止发生军事意外方面遭遇挑战。政策研究界愈发担忧军事智能化会加速战场速度，最终出现超出人类反应速度的“极速战”。[27]而大国为了维持军事优势有可能把更多的开火权

[27] John Allen and Darrell West, “Op-ed: Hyperwar Is Coming. America Needs to Bring AI into the Fight to Win — With Caution,” July 12, 2020, https://www.cnbc.com/2020/07/12/why-america-needs-to-bring-ai-into-the-upcoming-hyperwar-to-win.html.

下放给机器。[28]在这种悲观的预期下，人类可能在三方面愈发难以防范“极速战”下的军事意外。一是在军事竞争的压力下，匆忙部署不够安全和稳定的人工智能系统有可能使其在复杂的战场上失灵，进而引发难以预料的事故。二是在“极速战”的长期效应下，人类越发依赖机器给出的判断和建议，基于机器的建议而做出错误的军事决策，并最终导致意外发生。三是大国尚未就这项新兴技术对某些敏感的目标进行打击达成协议，它赋能的武器平台可能出现识别错误打击大坝、电站、医院等“禁区目标”，对平民造成不必要的附带伤害。

发表日期：2023年1月4日

[28] Scharre Paul, “Killer Apps,” *Foreign Affairs*, Vol. 98, No. 3, 2019, p.144.

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