

# 全球安全治理视域下的 自主武器军备控制\*

刘杨钺

【内容提要】 随着人工智能的飞速进展，不断智能化的自主武器日益显现出伦理和安全风险，使得限制或禁止自主武器成为全球安全治理领域的新兴议题。相比其他军控对象，自主武器军控进程在过去几年中获得较快推进，其中《特定常规武器公约》会议已决定设立政府专家组专门探讨自主武器问题。文章系统梳理了自主武器军备控制的概念、伦理和安全争议，旨在通过这种梳理更好地理解自主武器军控进程发展的动因，并对这一进程的未来走向做出预测。当前，自主武器军备控制的主要动因集中在道德层面，特别是让机器自主决策杀伤引发的伦理忧虑。而从安全层面看，自主武器蕴涵的安全风险在其他新兴技术领域同样存在，而发展和使用自主武器带来的战略红利依然显著，这使得主要国家推动自主武器军控的意愿并不强烈。在权力政治与道德政治的博弈下，自主武器军控在可预见的时期内将难以形成实质性成果，稍有可能的是通过“软法”等非约束性方式塑造一定的国际规范。在这个过程中，中国可以发挥更加积极主动的作用，在确保战略利益的同时营造有利的大国形象。

【关键词】 人工智能；军备控制；自主武器；国际安全；安全治理

【作者简介】 刘杨钺，国防科技大学文理学院副教授（长沙 邮编：410074）。

【DOI】 10.14093/j.cnki.cn10-1132/d.2018.02.003

【中图分类号】 D815.5 【文献标识码】 A 【文章编号】 2095-574X (2018) 02-0049-23

---

\* 本文系国家社科基金青年项目“国际网络冲突态势变化及应对策略研究”（项目批准号：17CGJ004）的阶段性成果。感谢《国际安全研究》期刊匿名审稿专家提出的修改意见和建议，文责自负。

技术变革是国际政治最具结构性意义的宏观变量，而 21 世纪以来人工智能的飞速发展已经吹响下一场重要变革的前奏。近年来，美国、英国、日本等发达国家纷纷推出人工智能国家发展战略或宏观政策报告，显示出对这一技术领域潜在战略价值的高度重视。<sup>①</sup> 从国际政治博弈来看，人工智能最首要的价值在于对国家间战略能力分配的潜在改变。事实上，在过去几十年里各国均致力于提高武器系统自主能力和智能化程度，以此克服人类处理信息能力的局限性，提升决策速度和提高打击精度。相关技术已经在各类防空反导预警、精确制导等系统中得到广泛应用。随着大数据和机器深度学习的不断发展，技术的智能化水平出现了根本性飞跃，武器系统独立完成复杂任务的能力得到显著增强，人工智能的军事化应用似乎已经驶入快车道。

与此同时，人工智能军事化蕴藏的风险也开始引发广泛关注。如斯蒂芬·威廉·霍金（Stephen William Hawking）所言，人工智能的发展“要么是人类历史上最好的事，要么是最糟的”。<sup>②</sup> 一批知名科学家和技术专家多次发出呼吁，要求国际社会采取实质性举措限制这一危险进程，特别是限制那些致命性自主武器系统（Lethal Autonomous Weapons Systems, LAWS）的发展和运用。特斯拉创始人埃隆·马斯克（Elon Musk）甚至警告，国家间人工智能的军备竞赛，有可能成为第三次世界大战的起因。目前，针对自主武器军备控制的动议已经提上联合国若干军控机制的正式议程，并得到部分国家的明确支持。然而，自主武器军控进程究竟有着怎样的前景？当前军控进程取得的较快发展是否意味着国际社会能够形成实质性的治理举措？本文从概念、价值和安全的三个层面分析自主武器军备控制面临的争议和问题，并对这一进程的发展前景做出推断。

### 一 全球安全治理背景下的自主武器军控动议

基于对话、协调与合作的安全治理逐渐成为冷战后各国应对全球安全威胁的重要方式，不同于传统集体安全模式，全球安全治理呈现出许多新的特征。第一，对人类安全的终极关怀。由于跨国性、全球性安全问题不断涌现，安全问题的威

---

① 赵刚：《人工智能大国战略》，载《环球》2017年第6期，第26-27页。

② 霍金：《让人工智能造福人类及其赖以生存的家园》，中国广播网，2017年4月28日，<http://tech.china.com.cn/it/20170428/296510.shtml>。

胁对象逐渐从单一国家拓展到人类共同体。国际安全也开始更多地聚焦“人的安全”这一根本性问题。<sup>①</sup> 关注焦点的“位移”使得那些触及人道主义准则和人类普遍安全的议题获得高度重视。第二，受科学技术发展影响显著。许多新兴安全问题的出现，本质上都源于社会技术体系变化的内在矛盾。网络、外层空间、生物、深海探测与开发利用等领域的技术突破，均衍生出新的亟待治理的安全难题。第三，参与治理的行为主体多元化。全球安全治理往往涉及不同的利益攸关方和参与者，除政府外，国际组织、非政府组织、企业、社会团体等行为主体也对治理过程、议题和话语产生影响。<sup>②</sup> 多元主体也使得安全治理的方式不再局限于传统的国家间合作，而是更多地表现为对话交流和综合协调等形式，自下而上的社会治理有时成为安全治理的主要推动力量。军备控制领域也日益呈现类似趋势，“国际反地雷组织”和“国际废除核武器运动”等案例都或多或少反映出上述特征的影响。

价值关怀、技术推动和多元治理等因素同样影响着自主武器军备控制进程。就第一点而言，涉及人类根本安全和尊严的道德关切已经成为自主武器军控的主要动因，后文将进一步对此进行阐述。从技术发展上看，国际社会对自主武器的关注与武器系统的自动化和智能化发展密不可分。<sup>③</sup> 这些技术进展使得越来越多的武器装备在部分甚至多数功能环节——如侦查、跟踪、目标识别甚至打击——实现了自主控制。<sup>④</sup>

军备控制进程所体现的特征则更为明显。关于限制或禁止致命性自主武器发展的探讨，显现出典型的自下而上的推动过程。对自主武器风险的忧虑首先来自学术

---

① Barry Buzan and Lene Hansen, *The Evolution of International Security Studies*, Cambridge: Cambridge University Press, 2009.

② 蔡拓、杨雪冬、吴志成主编：《全球治理概论》，北京：北京大学出版社 2016 年版，第 10-12 页。

③ 人工智能领域的一些标志性技术进步，在很大程度上增强了国际社会对相关技术的担忧，如 1997 年 IBM 计算机“深蓝”战胜国际象棋冠军卡斯帕罗夫，谷歌 Alpha Go 多次横扫人类围棋冠军等。由于人工智能技术的高度军民两用性，这些事件极易转化为人们对武器智能化进展的延伸想象。

④ 通过对公开武器贸易数据进行整理，希瑟·罗夫（Heather Roff）编纂了可被归入自主武器范畴的武器装备数据库，记录了多达 284 种各类武器，包括大量导弹、鱼雷、无人机和无人作战车辆等。参见 Heather Roff, “Dataset: Survey of Autonomous Weapons Systems,” <https://globalsecurity.asu.edu/robotics-autonomy>。

界。<sup>①</sup> 英国科学家诺埃尔·夏基 (Noel Sharkey) 2007 年便撰文提出警告, 称“我们正如梦游般走入一个新的世界, 在何时何地杀伤何人皆由机器定夺……为自主机器人建立国际规则和道德规范势在必行, 否则为时将晚”。<sup>②</sup> 2009 年夏基等学者共同组建了“机器人军备控制国际委员会”(International Committee for Robot Arms Control, ICRAC), 宗旨是推动国际社会形成具有法律约束力的协定, 对自主武器系统的研发和使用加以禁止。2012 年“人权观察”组织 (Human Rights Watch) 发布了一份题为《反对机器人杀手》的报告, 认为武器系统的自主化趋势将会深刻挑战法律和伦理规范, 因而必须及早订立未雨绸缪的禁令。<sup>③</sup>

在此基础上, “阻止机器人杀手运动”(Campaign to Stop Killer Robots) 于 2013 年在英国伦敦成立。作为一个联盟组织, 该运动的成员包括人权观察、大赦国际、ICRAC、帕格沃什科学和世界事务会议 (Pugwash Conferences on Science and World Affairs)、国际和平局 (International Peace Bureau) 等 63 个非政府组织。自此, 该运动成为国际社会中推动自主武器军备控制最为积极、也最为显要的社会力量, 通过普及宣传、学术交流、组织活动等多种方式, 深度参与到各个国际组织关于自主武器的磋商会谈之中。2015 年逾千名国际知名科学家 (包括霍金和马斯克等人) 更是联合发表公开信, 警告可能出现的军事人工智能的军备竞赛, 并呼吁禁止进攻性自主武器的发展。公开信中强调, “人工智能技术已来到临界点……其风险异乎寻常……今天攸关人类 (生存) 的问题在于, 究竟是开启人工智能的全球军备竞赛, 还是从起点上对其加以防范。”<sup>④</sup>

在全球安全治理中, 多元行为体不仅仅是议题和话语的设置者, 也是治理进程的推动者甚至塑造者。来自学术界和公民社会的广泛呼吁, 直接推动了自主武器军

---

① 一些具有代表性的学术思考可参见 Laurie Calhoun, “The Strange Case of Summary Execution by a Predator Drone,” *Peace Review*, Vol. 15, No. 2 (May 2003), pp. 209-214; Robert Sparrow, “Killer Robots,” *Journal of Applied Philosophy*, Vol. 24, No. 1 (February 2007), pp. 62-77; Peter Asaro, “How Just Could a Robot War Be?” in P. Brey, A. Briggle, and K. Waelbers, eds., *Current Issues in Computing and Philosophy*, Amsterdam The Netherlands: IOS Press, 2008, pp. 50-64; Jürgen Altmann, “Preventive Arms Control for Uninhabited Military Vehicles,” in R. Capurro and M. Nagenborg, eds., *Ethics for Robotics*, Heidelberg: AKA Verlag, 2009.

② Noel Sharkey, “Robot Wars are a Reality,” *The Guardian*, August 18, 2007.

③ *Losing Humanity: the Case against Killer Robots*, 2012, Human Rights Watch, [http://www.hrw.org/sites/default/files/reports/arms1112\\_ForUpload.pdf](http://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf).

④ Samuel Gibbs, “Musk, Wozniak and Hawking Urge Ban on Warfare AI and Autonomous Weapons,” *The Guardian*, July 27, 2015.

控议题进入政府层面。在联合国框架下，多个机制都对限制自主武器发展展开了探讨。一是人权理事会，其中法外处决、即决处决或任意处决问题特别报告员（Special Rapporteur on extrajudicial, summary or arbitrary executions）多次向联合国大会提交报告，分析自主化和人工智能技术的军事应用可能给人权和国际人道主义法带来的冲击。2010 年的临时报告中就已提出“不应仅强调这种技术进步带来的挑战，还应强调积极主动地采取措施和办法，以确保这种技术促进更加有效地遵守国际人权和人道主义法规的能力得到优化”。<sup>①</sup>二是联合国大会第一委员会（裁军与国际安全委员会）。从 2013 年起，自主武器系统开始成为该委员会每年会议的议题，就自主武器表达关切或忧虑的国家逐年增多。<sup>②</sup>

除了人权理事会和第一委员会，《特定常规武器公约》（Convention on Certain Conventional Weapons, CCW）会谈机制实际上成为自主武器军备控制的核心平台。从功能上看，《特定常规武器公约》的目的在于对可能造成不必要人员伤亡，或者不加区分地威胁平民安全的特定武器进行限制或禁止，这与国际社会对自主武器的伦理关切比较吻合。在 2013 年的缔约国会议上，与会各方同意于次年设立非正式的专家会议，专门就致命性自主武器系统进行讨论。从 2014 年至 2016 年，非正式专家会议连续召开了三次，参与者既包括《特定常规武器公约》缔约国代表，也有作为观察员的非缔约国代表以及各类国际组织、非政府组织和学术机构的相关专家。讨论的议题包括自主武器的技术发展趋势、概念界定和这些武器所引发的道德、法律和安全等各种问题。<sup>③</sup>2016 年底这一进程再次获得新的突破：12 月召开的《特定常规武器公约》第五次审议大会决定成立政府专家组（Group of Governmental Experts），以便将致命性自主武器的讨论提升到更为正式的层面。<sup>④</sup>对于一向进展

① Philip Alston, “Interim report of the Special Rapporteur on extrajudicial, summary or arbitrary executions,” A/65/321, United Nations, August 2010; 另参见 Christof Heyns, “Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions,” A/HRC/23/47, United Nations, April 2013.

② 根据“阻止机器人杀手运动”的统计，在发言中提及自主武器问题的国家从 2013 年的 16 个增长到 2016 年的 36 个。参见 <http://www.stopkillerrobots.org/chronology/>。

③ 相关信息可以参见 2016 年非正式专家会议的报告：*Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, 2016, [http://www.unog.ch/80256EDD006B8954/\(httpAssets\)/DDC13B243BA863E6C1257FDB00380A88/\\$file/ReportLAWS\\_2016\\_AdvancedVersion.pdf](http://www.unog.ch/80256EDD006B8954/(httpAssets)/DDC13B243BA863E6C1257FDB00380A88/$file/ReportLAWS_2016_AdvancedVersion.pdf)。

④ 根据时间安排，政府专家组本应于 2017 年 8 月和 11 月举行两次会议。目前，8 月份的会议因经费不足而被取消，11 月份的会议于 13-17 日如期举行。联合国官方网站的公告：[http://www.unog.ch/80256EE600585943/\(httpPages\)/3CFCEEEF52D553D5C1257B0300473B77?OpenDocument](http://www.unog.ch/80256EE600585943/(httpPages)/3CFCEEEF52D553D5C1257B0300473B77?OpenDocument)。

曲折的《特定常规武器公约》框架而言，自主武器军控从获得关注到向前推进的速度可谓不同寻常。<sup>①</sup>

据统计，截至 2017 年已有 19 个国家正式表示应采取措施禁止致命性自主武器的发展，包括巴基斯坦、埃及、阿根廷等。<sup>②</sup> 然而，这些明确的反对者仍然以中小国家为主，主要大国在限制自主武器问题上存在不少分歧。例如，美国国防部虽于 2012 年出台条令，规定自主和半自主武器的使用必须以“适当水平的人工判断”为前提，<sup>③</sup> 但其又将人工智能和机器人技术视为“第三次抵消战略”的重要基石。<sup>④</sup> 一边是限制其使用，另一边则是鼓励相关技术发展。二者之间的张力必然影响到美国在自主武器军备控制上的政策主张。那么，应当如何评判自主武器军控进程取得的进展？本文认为，从概念、伦理和安全三个方面来看，自主武器军备控制进程要进一步取得实质性成果，仍然面临多方面争议的困扰，其前景并不容乐观。

## 二 概念层面：自主武器军控指涉对象模糊

“军备控制指的是通过限制军备的发展和通过使用方式来控制军备发展水平”，<sup>⑤</sup> 特别是对特定武器系统的数量、类型、性能和使用加以限制。对于自主武器系统军备控制进程而言，首要问题就在于对军控对象进行明确界定，<sup>⑥</sup> 因为这直接关系到

---

① Frank Sauer, “Stopping ‘Killer Robots’: Why Now Is the Time to Ban Autonomous Weapons Systems,” *Arms Control Today*, October 2016, [http://www.isodarco.it/courses/andalo18/doc/sauer\\_Stopping-Killer-Robots.pdf](http://www.isodarco.it/courses/andalo18/doc/sauer_Stopping-Killer-Robots.pdf).

② Campaign to Stop Killer Robots, “Country Views on Killer Robots,” May 23, 2017, [http://www.stopkillerrobots.org/wp-content/uploads/2013/03/KRC\\_CountryViews\\_May2017.pdf](http://www.stopkillerrobots.org/wp-content/uploads/2013/03/KRC_CountryViews_May2017.pdf).

③ Department of Defense Directive, “Autonomy in Weapon Systems,” November 21, 2012, [https://fas.org/irp/doddir/dod/d3000\\_09.pdf](https://fas.org/irp/doddir/dod/d3000_09.pdf).

④ Robert Work, “The Third U.S. Offset Strategy and Its Implications for Partners and Allies,” January 28, 2015, <http://www.defense.gov/News/Speeches/Speech-View/Article/606641/the-third-us-offset-strategy-and-its-implications-for-partners-and-allies>.

⑤ 李彬：《军备控制理论与分析》，北京：国防工业出版社 2006 年版。

⑥ 例如，中国代表团在 2016 年参加《特定常规武器公约》第五次审议大会时，递交的关于自主武器的立场文件中明确指出，“确定致命性自主武器的概念和范围是进行其他方面探讨的前提条件”。参见 “The position paper submitted by the Chinese delegation to CCW 5th Review Conference,” [http://www.unog.ch/80256EDD006B8954/\(httpAssets\)/DD1551E60648CEBBC125808A005954FA/\\$file/China's+Position+Paper.pdf](http://www.unog.ch/80256EDD006B8954/(httpAssets)/DD1551E60648CEBBC125808A005954FA/$file/China's+Position+Paper.pdf).

哪些武器会被禁止而哪些可以继续使用。

然而，“自主武器系统”并不是一个清晰的概念。“人权观察”在前述《反对机器人杀手》报告中，将完全自主的武器系统界定为“能够在无人参与或干预的情况下选择目标并施加武力”的机器。<sup>①</sup>许多自主武器系统的批评者和组织使用的定义与上述基本相近。例如，彼得·阿萨罗（Peter Asaro）认为，自主武器系统就是那些“不依靠直接的人工监督和杀伤决策，而能够（自主）定位并发起潜在致命攻击”的武器系统。<sup>②</sup>联合国法外处决、即决处决或任意处决问题特别报告员克里斯托夫·海恩斯（Christof Heyns）则将“自主机器人杀手”（LARs）定义为“一种机器人武器系统，一经启动，即可在无需人类操作员进一步干预的情况下选择和打击目标”。在界定这类武器系统时，“一个重要因素是机器人在挑选目标和使用杀伤力时可以作出自主‘选择’”。<sup>③</sup>进一步说，这类武器系统应当运行于无人监管的开放空间，并依赖传感器实现与任务环境的实时互动。<sup>④</sup>

上述定义大体勾勒出致命性自主武器系统的两项基本特征。一是高度自主，即能够在无人参与的情况下凭借自身的信息处理能力，独立完成目标搜寻、识别和发动攻击的全部过程。二是具有攻击性的杀伤力，而不仅仅是用于防御性的侦查、监控、分析等任务。这两项特征实际上构建起“机器人杀手”的核心意象——不受人控制而能够自主决定杀戮的机器。因此，不论是联合国<sup>⑤</sup>还是“阻止机器人杀手运动”，在探讨致命性自主武器系统时的指涉对象是基本一致的。但对于军备控制而言，上述定义和特征仍然过于宽泛。

首先，自主性和杀伤性并没有充分反映出武器系统的智能化程度，而后者实际上才是国际社会的真正忧虑所在。按照上述定义，许多现行的（甚至部分过时的）

---

① *Losing Humanity: the Case against Killer Robots*, 2012, p. 2, Human Rights Watch, [http://www.hrw.org/sites/default/files/reports/arms1112\\_ForUpload.pdf](http://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf).

② Peter Asaro, “On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making,” *International Review of the Red Cross*, Vol. 94, No. 886 (June 2012), p. 690.

③ Christof Heyns, “Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions,” A/HRC/23/47, United Nations, April 2013.

④ Noel Sharkey, “Automating Warfare: Lessons Learned from the Drones,” *Journal of Law, Information and Science*, Vol. 21, No. 2 (December 2012), pp. 140-154.

⑤ 联合国对致命性自主武器系统所下定义为“能够在无人干预情况下识别和攻击目标”（的武器系统）。可参见联合国对致命性自主武器系统的背景介绍：[http://www.unog.ch/80256EE600585943/\(httpPages\)/8FA3C2562A60FF81C1257CE600393DF6](http://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6)。

武器装备都会被纳入“致命性自主武器”之列。例如，美军普遍列装的“密集阵”（Phalanx）近防系统能够自动搜索、跟踪、评估和打击来犯的反舰导弹和飞机，全部作战流程都是依靠高性能计算机自动完成；以色列 20 世纪 90 年代研制的“哈比”（Harpy）无人机则属于典型的“即发即弃”（fire and forget）的装备，地面平台将无人机发射后便不再实施操控，由无人机自主探测并摧毁敌方雷达；韩国在朝韩非军事区边境部署的超级庇护 II（Super aEgis II）自动炮塔能够根据热能和运动来识别目标，这一武器系统原本同样具有自动发射功能，只是在客户普遍要求下增加了人工干预环节。<sup>①</sup> 诸如此类的武器均具备自主发现目标和实施打击的能力，因而在一定程度上都符合自主武器的定义范畴。但这样一来，自主武器军备控制的目标对象就不仅仅包括尚未出现的高度智能化的武器装备，也包括大量早已广泛使用和不断更新的平台系统，这显然难以在各国形成普遍共识。

基于此，一些国家和学者在界定自主武器系统时，更加强调这些武器的人工智能色彩。例如，希瑟·罗夫（Heather Roff）便认为，需要加以限制的自主武器系统不仅仅是能够自动寻找目标并进行攻击的武器平台，还必须具备自主学习的能力。<sup>②</sup> 也就是说，这类武器不是依靠事先给定的（程序化的）固定特征来识别目标，而是通过学习来寻找那些并不知悉先验特点的潜在杀伤对象。而按照英国国防部的说法，自主系统应能够“理解更高层次的意图和趋势。通过这种理解以及对环境的认知，此类系统能够采取合适的行为来实现预期的效果……自主系统将具有自我意识，其对外部输入做出的反应应当无异于甚至优于人工系统的反应”。<sup>③</sup> 问题是，这样的定义虽然凸显了自主武器的智能特性，但也使得概念变得过于模糊。武器的学习能力究竟需要达到何种程度才能真正算得上“自主”？上述定义显然无法为军备控制提供切实有效的标准。在罗伯特·斯帕罗（Robert Sparrow）看来，自主武器系统应被理解为一个智能化程度由低到高的连续体，较低的一端是能够探测重量并据此决定是否引爆的人员杀伤型地雷，另一端则是具备高度人工智能（甚至类似人类智慧）的武器系统。<sup>④</sup> 军

---

① Simon Parkin, “Killer Robots: the Soldiers that Never Sleep,” BBC, July 16, 2015, <http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-that-never-sleep>.

② Heather Roff, “The Strategic Robot Problem: Lethal Autonomous Weapons in War,” *Journal of Military Ethics*, Vol. 13, No. 3 (November 2014), pp. 211-227.

③ UK Ministry of Defence, *The UK Approach to Unmanned Aircraft Systems*, 2011, <https://www.gov.uk/government/publications/jdn-2-11-the-uk-approach-to-unmanned-aircraft-systems>.

④ Robert Sparrow, “Robots and Respect: Assessing the Case against Autonomous Weapon Systems,” *Ethics and International Affairs*, Vol. 30, No. 1 (March 2016), pp. 93-116.

备控制不可能也没有必要针对这个连续体中的所有武器，但如何划定（以及是否可能划定）红线可能成为军控向前推进的重要障碍。

其次，致命性自主武器概念中的自主性和致命性特征都具有变动性，很容易受到技术发展变化带来的影响。<sup>①</sup> 以无人机为例，现有大多数军用无人机装备并不属于联合国关于致命性自主武器军控的探讨范围，因为这些无人机往往由人员直接远程遥控，或者至少其关键环节（特别是发动攻击的决定）需要人的指令性输入，也就是所谓的“人在环内”（in-the-loop）的控制模式。但随着战场环境日益复杂化和战争节奏加速以及前述武器自主化带来的军事优势，越来越少的人工干预将成为主流。美国空军在 2009 年起草的一份规划中就已提出，“人的作用将逐渐从‘在环内’变为‘在环上’（on-the-loop），即仅对特定决定的执行进行监督。同时，人工智能技术的进步将使系统能够在法律和政策限定之下完成战斗决策和行为，而不必要求人为介入”。<sup>②</sup> 随着人工干预逐渐减少，机器在做出自主决策时如果产生错误或偏差，有可能难以被人工监管环节及时发现和纠正。对于军备控制探讨来说，如何准确区分“人在环上”还是“人在环外”（out-the-loop，即事实上无人干预机器决策流程），如何判定对武器系统进行的间接人工干预是否充足，以及如何对武器在实际使用过程中的自主化程度进行核查，似乎都成为颇为棘手的难题。因此，由技术发展引致概念内涵本身的变化，可能阻碍自主武器军控形成有效共识和可行措施。类似情况也适用于“致命性”特征。考虑到人工智能技术的军民两用性，非攻击性的武器平台往往只需稍加改装或增加组件，就能具备攻击性和杀伤力。对某种技术应用的简单禁止，很难确保其不用于其他不适当的用途。<sup>③</sup>

最后，更为重要的是，自主性和致命性特征并不能反映自主武器系统蕴涵的全部风险。如同下文将要进一步分析的那样，自主武器系统可能深刻改变战争的成本收益，极大降低战争门槛，引发国家间的军备竞赛，对传统的战略稳定性带来冲击，

---

① 在 2016 年《特定常规武器公约》关于致命性自主武器系统的非正式专家会议上，部分代表便提出，真正意义上的“致命性自主武器”尚不存在，且相关技术还在不断演化之中，因此要达成关于此类武器的定义极其困难。参见 *Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, 2016, [http://www.unog.ch/80256EDD006B8954/\(httpAssets\)/DDC13B243BA863E6C1257FDB00380A88/\\$file/ReportLAWS\\_2016\\_AdvancedVersion.pdf](http://www.unog.ch/80256EDD006B8954/(httpAssets)/DDC13B243BA863E6C1257FDB00380A88/$file/ReportLAWS_2016_AdvancedVersion.pdf)。

② U.S. Air Force, *Unmanned Aircraft Systems Flight Plan, 2009–2047*, 2009, <http://www.govexec.com/pdfs/072309kp1.pdf>。

③ Wendell Wallach and Colin Allen, “Framing Robot Arms Control,” *Ethics and Information Technology*, Vol. 15, No. 2 (June 2013), p. 132.

并为各类非国家行为体提供更为便捷但也更加危险的工具。这些风险并不是武器系统自主性和致命性的直接产物。例如，导致发起战争成本下降主要是来自武器的无人化状态，特别是无人作战对避免己方人员伤亡起到重要的作用。不论武器决策流程是否受到人工干预，这种作用同样存在。也就是说，仅仅针对自主性和致命性来推动的军备控制议程，只能管控致命性自主武器系统带来的部分伦理和安全问题，这将在某种程度上降低对军控谈判必要性和有效性的认识。对主要大国而言，在自主武器的概念上难以达成一致，既是维护各自战略利益最大化的需要所致，也给通过操纵概念议题进一步巩固在相关领域的技术优势提供了可能。<sup>①</sup>

### 三 价值层面：自主武器军控的主要动力

致命性自主武器之所以受到国际社会关注，很大程度上来自其对国际规范的潜在冲击。事实上，对自主武器最为有力的批评意见几乎都集中在这一层面。具体来看，涉及自主武器伦理价值规范的讨论呈现出两条不同的逻辑。

第一种逻辑遵循结果论（consequentialist）的推断方式，关注的是行为的后果是否符合普遍的道德关切。就自主武器而言，这意味着使用这类武器产生的后果可能有违道德规范。在这里，首要的批评意见认为自主武器无法有效区分平民和战斗人员。“冲突各方无论何时均应在平民居民和战斗员之间和在民用物体和军事目标之间加以区别”，<sup>②</sup>这一区分原则已成为战争法基础原则之一。自主武器的批评者认为，完全自主的武器系统难以有效区分平民和战斗人员，很可能增加平民伤亡风险。要有效进行人员区分，自主武器不仅需要对特定区域内战斗人员的可识别特征进行判断（例如判断可疑人员是否携带枪支），还必须对非战斗人员的特征也加以识别，否则即使对战斗人员的攻击也可能导致不必要的平民伤亡。<sup>③</sup>此外，区分原则还可能受到其他挑战：如何区别战斗人员和失去战斗力的受保护人员；如何区别

---

<sup>①</sup> Edward Moore Geist, “It’s Already Too Late to Stop the AI Arms Race – We must Manage it Instead,” *Bulletin of the Atomic Scientists*, Vol. 72, No. 5 (August 2016), pp. 318-321.

<sup>②</sup> 《一九四九年八月十二日日内瓦四公约关于保护国际性武装冲突受害者的附加议定书（第一议定书）》，第四十八条，《议定书》全文可在以下地址获取：[http://www.icrc.org/chi/assets/files/other/mt\\_070116\\_prot1\\_c.pdf](http://www.icrc.org/chi/assets/files/other/mt_070116_prot1_c.pdf)。

<sup>③</sup> Marcello Guarini and Paul Bello, “Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters,” in Patrick Lin, Keith Abney and George A. Bekey, eds., *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge, Mass.: MIT Press, 2012.

战斗人员和其他持有武器但非敌对方人员（例如维和人员或者单纯持有武器进行自卫的平民），等等。这些区分都需要武器系统对极其复杂环境中的诸多可预见和不可预见因素进行准确认知、分析和预测，以现有的人工智能水平似乎还难以企及。

结果论方面的第二个问题在于，自主武器系统可能深刻改变了战争决策规则并降低了战争门槛。通常情况下，民众对战争伤亡的敏感性极大影响着领导者做出的战争决断及其进程。<sup>①</sup> 例如，越南战争后期美国国内反战情绪的不断上涨，与美军在“春节攻势”等战役中伤亡日趋增多不无关系。美国在索马里、伊拉克、阿富汗等地的军事行动中反复呈现类似情形。而自主武器系统（包括无人机作战）显著降低了对己方伤亡的预期，这使得领导者的战争决定更容易获得民众支持，战争开启后的政治风险也相对较低。实证研究也发现，相较于动用地面部队或是常规空中打击，民众更愿意支持无人机作战。<sup>②</sup> 因此，武器系统自主化带来的低伤亡甚至零伤亡，可能会极大削弱限制战争的民意因素，使得战争成为更加难以管束的危险工具。联合国秘书长在 1998 年的报告中就对科学技术发展的伦理风险提出了警告，认为自主平台能力的增长“打开了一种可能性，那就是各国可以进行战争而不受到其人民反对牺牲人命的限制。”<sup>③</sup> 按此趋势发展，国家在战争方面的决策将越来越宽松，使用武力逐渐沦为仅靠财政和外交的考量，导致武装冲突走向“常态化”。<sup>④</sup>

此外，自主武器系统的实际效用有时也受到质疑。例如，美军在推动武器系统自动化发展中的一个重要依据是自主武器能够降低人力成本，甚至降低对战斗人员专业技能的要求。但有的学者指出，这种对于自主武器的期待不过是“技术主义者的盲目热忱”。从“爱国者”防御系统和“捕食者”无人机系统的实际使用情况看，自主武器并不能减少人力和训练成本开支，反而对武器设计人员和操作人员之间的

---

① Louis Klarevas, “The ‘Essential Domino’ of Military Operations: American Public Opinion and the Use of Force,” *International Studies Perspectives*, Vol. 3, No. 4 (November 2002), pp. 417-437.

② James Walsh and Marcus Schulzke, *The Ethics of Drone Strikes: Does Reducing the Cost of Conflict Encourage War?* U.S. Army War College Press, 2015; Michael Horowitz, Paul Scharre and Ben FitzGerald, “Drone Proliferation and the Use of Force: An Experimental Approach,” Center for a New American Security report, 2017, <http://drones.cnas.org/reports/drone-proliferation-use-force/>.

③ United Nations Report of the Secretary-General, “Role of Science and Technology in the Context of International Security and Disarmament,” A/53/202, July 1998, <https://www.un.org/disarmament/topics/scienceandtechnology>.

④ Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*, London and New York: Routledge, 2009.

融合提出了更高要求。<sup>①</sup> 如果自主武器系统在节约战争开支方面不能产生预期的效果，那么进一步提升武器自动化和智能化水平的重要前提条件便会受到削弱，这对那些主张用效用对冲风险的人而言显然不是好消息。

第二种论证逻辑沿着义务论（Deontological）展开，其关注的是行为本身是否体现道义准则。从义务论逻辑出发对自主武器的批评更为尖锐。这一视角认为让机器自主抉择对人类的杀伤，其行为本身（而不是行为的结果）就有违道德。至少三方面因素支撑着上述批评。

一是责任缺失。自主武器系统的核心特征在于自主。特别是随着智能化和学习能力的进一步提升，自主武器的发展趋势将朝着更少人工干预、更多独立决断和更难以预测的作战后果方向演进。这样一来，假如高度自主的武器系统由于信息不全、计算失误、软硬件失灵等诸多不可预知的原因，导致误伤（杀）平民或违反了相称性等基本战争法则的错误，如何归责就变成了颇为难解的问题。有学者将这一困境称为“三难”，即无论是自主武器系统的制造者、执行军事任务的指挥者，还是武器系统本身，都无法作为机器自主杀伤行为的责任主体。<sup>②</sup> 尤其是三者中的后者，承担责任所意味着的指责、惩罚或奖励，对武器系统本身均没有实际意义。这种“责任鸿沟”<sup>③</sup> 显然会使攻击行为从一开始便不具备合法性。<sup>④</sup>

二是风险不对称。责任问题主要涉及高度智能化和自主化的武器系统，而风险问题则普遍存在于各个程度的自主武器当中。随着空中打击、远程进攻逐渐成为现代战争的主导样式，风险的天平越来越多地开始倒向平民和技术落后的一方，以美

---

① Robert Hoffman, Timothy Cullen and John Hawley, “The Myths and Costs of Autonomous Weapon Systems,” *Bulletin of the Atomic Scientists*, Vol. 72, No. 4 (June 2016), pp. 247-255.

② Robert Sparrow, “Robotic Weapons and the Future of War,” in Paolo Tripodi and Jessica Wolfendale, eds., *New Wars and New Soldiers: Military Ethics in the Contemporary World*, Surrey: Ashgate, 2011.

③ Andreas Matthias, “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata,” *Ethics and Information Technology*, Vol. 6, No. 3 (September 2004), pp. 175-183.

④ 也有人自主武器的“责任鸿沟”问题提出了辩解，认为行为与责任之间的联系并不必须是因果关联的，也可以是人为建立起来的。因此，指挥官可以通过授予自主武器系统“自由权力”（blank check），同时对其行为后果承担责任。参见 Marc Champagne and Ryan Tonkens, “Bridging the Responsibility Gap in Automated Warfare,” *Philosophy and Technology*, Vol. 28, No. 1 (March 2015), pp. 125-137. 但本文作者认为这一点并不具有充足的说服力。战场上的指挥官虽然时常需要为其下属的失当行为负责，但这种责任必然是有限度的，而且实施失当行为的主体本身也不能免除责任。同理，自主武器系统的部分自主行为结果可以追责到某一层级的指挥人员，但指挥人员难以承担自主武器系统可能造成的所有后果。

国为代表的西方国家对于战争中己方的人员伤亡愈发敏感，而自主武器系统的发展无疑使伤亡风险分配进一步有利于无人化战争的发起者。也就是说，自主武器所带来的己方安全风险的最小化，是以对方安全风险的升高为代价的。即使对受攻击方没有本质影响，己方伤亡概率的显著降低（甚至消失）也会导致战争更加有利可图，而不具备无人作战技术能力的国家和人口则承受着战争的主要代价。这种不对称的风险分布与传统战争法的基本精神背道而驰，后者要求将非战斗人员的安全考虑置于作战人员之前而不是相反。<sup>①</sup> 美国在阿富汗、巴基斯坦等国实施的无人化军事行动虽仍然受到远程操控指挥，但其“通过规避美军士兵在战场上的伤亡，既使美国国内民众远离了战争效应，又使战略家们不被技术进步带来的逻辑和伦理缺陷所牵绊”。<sup>②</sup> 实际上，人工智能的基础架构看似是客观公允的机器算法，但背后隐藏的却是不平等的权力关系的进一步固化。正如英国政府首席科学顾问马克·沃尔波特（Mark Walport）所说，“机器学习可能会内部化在量刑或医疗历史中存在的所有隐性偏见，并通过它们的算法外部化。”<sup>③</sup> 同样，自主武器和无人作战也使不同行为体之间不平等的技术能力转化为强加在弱势行为体之上的风险“偏见”。

三是从根本上否定人的尊严。战争本质上是人与人之间的互动行为，而战争法即是由人的互动所产生的社会规范。也就是说，人际互动关系构成了战争伦理各项原则（不管是区别原则还是相称性原则）的基本前提。交战双方在步入战场时，先决条件就是清楚地认识到自己是“选择”成为战斗人员并承担可能被杀伤的后果。这种认识乃是交战正义（或许乃至开战正义）的基础。而当高度自主化和智能化的武器系统做出攻击决定时，这种人际互动关系荡然无存。<sup>④</sup> 从某种意义上说，“已经有人在决定攻击对象的生死了”。<sup>⑤</sup> 将人的主观意念排除在战争这一人际互动过程之外，会造成对人的尊严的彻底挑战。

---

① Jeff McMahan, “The Just Distribution of Harm between Combatants and Noncombatants,” *Philosophy and Public Affairs*, Vol. 38, No. 4 (Fall 2010), pp. 342-379.

② Sarah Kreps and John Kaag, “The Use of Unmanned Aerial Vehicles in Contemporary Conflict: A Legal and Ethical Analysis,” *Polity*, Vol. 44, No. 2 (March 2012), pp. 260-285.

③ Mark Walport, “Rise of the Machines: Are Algorithms Sprawling Out of Our Control?” *Wired*, April 1, 2017.

④ Peter Asaro, “On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making,” *International Review of the Red Cross*, Vol. 94, No. 886 (Summer 2012), pp. 687-709.

⑤ Robert Sparrow, “Robots and Respect: Assessing the Case against Autonomous Weapon Systems,” *Ethics and International Affairs*, Vol. 30, No. 1 (March 2016), pp. 93-116.

需要指出的是，一些支持者也针对上述批评提出了反驳理由。例如，对于自主武器无法区分平民，有人则认为机器不会像人那样受到恐惧、愤怒、报复心等非理性因素影响，因而随着技术不断进步，自主武器在区别目标对象时或许更能避免种种人为错误。<sup>①</sup> 但这类反驳理由似乎并不足以平息人们对自主武器价值风险的忧虑，因为这些理由的一个重要前提条件是自主武器也能遵循一定的道德准则，特别是能够掌握战争法和其他的交战规则。<sup>②</sup> 且不论这一点本身是否能在技术上实现，掌握这些规则是否等同于道德其实都具有较大争议，<sup>③</sup> 而以战争法作为智能化自主武器的全部道德来源也未必能使其足够应付战争中的所有情境。<sup>④</sup> 如果不能预期人工智能将逐渐发展出近乎于人类的道德能力，那么对自主武器系统的自主程度和杀伤性能进行一定限制应当是更为合理的选择。<sup>⑤</sup>

总的来看，虽然支持者从使用效能等方面出发为自主武器系统进行了辩护，但这类武器的进一步发展可能挑战传统战争法中的区分原则、相称原则等基本准则，这些价值风险很难在可预见的未来通过技术手段加以解决，这成为推动致命性自主武器军备控制进程的主要动因。

### 四 安全层面：自主武器军控受制于预期收益

安全问题同样是军备控制探讨的重要议题。特别是考虑到军备控制的结果直接

---

① Ronald Arkin, "The Case for Ethical Autonomy in Unmanned Systems," *Journal of Military Ethics*, Vol. 9, No. 4 (December 2010), pp. 332-341.

② Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall/CRC, 2009, <http://www.doc88.com/p-2146925595970.html>.

③ 有学者依据康德的哲学理念指出，“先验伦理的主体必须是人，必须具有自我意识，必须归属于人类共同体之中”。在这样的标准下，机器即便能够习得规则，也达不到人的道德状态。参见 Bernd Carsten Stahl, "Can a Computer Adhere to the Categorical Imperative? A Contemplation of the Limits of Transcendental Ethics in IT," Paper Presented at the International Conference on Systems Research, Informatics and Cybernetics, Baden, Germany, 2002.

④ 马蒂亚斯认为，仅仅依靠算法形成的道德伦理解决不了真实世界中普遍存在的道德张力，自然也无法应对战场上的道德难题，后者通常包括“涉及生死、正当理由、惩罚和报复以及特定文化背景中的伦理问题等矛盾抉择”。参见 Andreas Matthias, "Algorithmic Moral Control of War Robots: Philosophical Questions," *Law, Innovation and Technology*, Vol. 3, No. 2 (December 2011), pp. 279-301.

⑤ Wendell Wallach and Colin Allen, "Framing Robot Arms Control," *Ethics and Information Technology*, Vol. 15, No. 2 (June 2013), pp. 125-135.

影响到国际体系权力分配，安全关切对军控进程可能起到更为关键的作用。要求对自主武器发展加以限制的人认为，这类武器极易引发国家间军备竞赛、颠覆现有力量对比、导致冲突升级和不稳定性，并为非国家行为体提供易于获取的危险工具。

首先，自主武器或是人工智能的军备竞赛似乎已是进行时。以无人机发展为例，如前所述，超过 90 个国家已拥有军用无人机装备或相关技术能力，其中美军无人机占其军用飞机的比例在 2005-2013 年间从 1/20 迅速攀升至 1/3。有预测甚至认为到 2021 年全球无人机市场规模将达到 940 亿美元。<sup>①</sup> 随着军用无人装备的急剧增多，无人机的军事应用将不会局限于反恐和特种任务需要，而一旦这些武器应用于国家间军事安全互动，现有战争样式和交战规则都可能重新改写。更重要的是，没有人能够准确预估无人化、智能化战争究竟会产生何种图景。降低的战争门槛、膨胀的武器库规模、不确定的技术演进路径，这些因素使得自主武器军备竞赛可能成为国家间（尤其是区域国家间）战略互疑的新来源。<sup>②</sup> 与此同时，由于人工智能发展及军事应用早在冷战时期就已开始，自主武器的扩散并不仅限于传统视域中的无人机等系统，而是可能在各个军事领域全面铺开。有学者对此提出警告：“如果自主武器获得发展和部署，它们将最终在每个领域都安家落户——（不管是）空中、太空、海洋、陆地或是网络空间。它们将成群捕猎，编织成无人武器系统的复杂网络”。<sup>③</sup>

其次，自主武器对国际体系力量对比的影响至少表现在两个方面。第一种影响是破坏战略稳定格局。从冷战以来，全球战略稳定的重要基石之一在于大国间相对平衡的核威慑。这种相互威慑局面的前提条件是一方在一定程度上允许另一方拥有适度的核报复能力。而伴随无人技术的不断发展，无人武器装备特性将逐渐从目前强调持续性（续航能力、低能耗）为主，转向侧重速度和隐身性能的突防性为主。日益兴起的无人机蜂群战术同样为突破防御体系提供了新途径。<sup>④</sup> 这些变化将使技术发达国家拥有风险更低、打击效能更大的攻击工具，从而对对手的战略威慑能力

① Scott Shane, "Coming Soon: the Drones Arms Race," *New York Times*, October 8, 2011.

② Michael Boyle, "The Race for Drones," *Foreign Policy Research Institute E-Notes*, January 26, 2015.

③ Heather Roff, "To Ban or Regulate Autonomous Weapons," *Bulletin of the Atomic Scientists*, Vol. 72, No. 2 (March 2016), pp. 122-124.

④ Kris Osborn, "Swarming Mini-Drones: Inside the Pentagon's Plan to Overwhelm Russian and Chinese Air Defenses," *The National Interest*, May 10, 2016.

构成严重挑战。新型无人武器的发展“不仅意味着高度可信的威胁，因为使用这种武器对（己方）人员毫无损伤，而且提供了可升级的远程精确打击能力，能够逃避完善的空中防御体系”。<sup>①</sup> 对于那些原本具有拒止能力的国家来说，下一代无人装备的机动性、隐蔽性和自主性可能使其基于报复能力的威慑战略趋于无效。基于同样理由，成本相对低廉并可长时间蛰伏的水下无人系统，将致使水下作战环境趋向透明，导致潜艇这一核威慑重要手段面临较大威胁。<sup>②</sup> 第二种影响则来自常规力量方面。发展自主武器所需的资源禀赋或许与传统武器有所不同，对技术水平的要求显然高于对人口、能源等要素的要求。在这种情况下，有观点认为技术发达的中等国家可能是自主武器赋权的最大受益者，因为自主武器系统的快速发展意味着“军事力量将逐渐与人口基础脱钩，而后者传统上是衡量军事实力的重要指标”。<sup>③</sup> 总之，不论是削弱战略威慑还是改写常规力量分配，自主武器都可能为国际体系注入更多不确定性和不稳定性。

再次，自主武器还可能引发冲突升级。由于使用自主武器代价较低，国家行为体可能越来越倾向于使用这类武器来探查对手的能力、决心和回应策略。例如，在2012年的一起事件里，黎巴嫩真主党使用伊朗制造的无人机，意图对以色列核设施情况进行侦查，但遭后者空中力量击落。<sup>④</sup> 而在2016年底，美国无人潜航器在南海海域侦察时也引发了外交风波。问题在于，一方面，国家具有强烈动机来利用这种武器获取情报、增强态势感知、宣示主张甚至发动有限进攻；另一方面，这些行动的对象国却并不容易准确判断行为的真实意图，例如可能将抵近侦察误认为先发制人的打击。自主武器是否能在执行任务时释放明确无误的信号来表明自身意图，可能将成为困扰国家间安全互动的难题。更何况，对于本身处于敌对或紧张关系的国家而言，自主武器的侵犯行为即便本身并不严重，也可能被解读为对安全利益的严重挑衅和更大范围攻击的序曲，导致对象国采取更加严厉的回应措施而引

---

<sup>①</sup> Michael Mayer, “The New Killer Drones: Understanding the Strategic Implications of Next-Generation Unmanned Combat Aerial Vehicles,” *International Affairs*, Vol. 91, No. 4 (July 2015), pp. 765-780.

<sup>②</sup> David Connett, “Trident: Nuclear Deterrent under Threat from Underwater Drones, Expert Warns,” *Independent*, December 26, 2015.

<sup>③</sup> Robert Work and Shawn Brimley, *20YY: Preparing for War in the Robotic Age*, Washington DC: Center for a New American Security, 2014, p. 33.

<sup>④</sup> BBC News, “Hezbollah Admits Launching Drone over Israel,” October 11, 2012, <http://www.bbc.com/news/world-middle-east-19914441>.

起不必要的冲突升级。<sup>①</sup>此外，自主武器高度依赖对外部环境的感知和信息交换，由此产生意外事故和人为恶意干预的可能性也会升高。例如，如果无人机在执行侦察任务时，遭到黑客入侵或其他形式的电磁干扰而出现坠毁、撞击、爆炸等反常行为，<sup>②</sup>同样可能引起目标对象的误判和升级回应。

最后，非国家行为体（尤其是恐怖主义组织）可能掌握并使用此类武器，构成了对自主武器安全忧虑的另一重要理由。自主武器技术及其装备具有显著的军民两用性，而且一些小型无人装备的成本逐渐降低，非国家行为体从公开渠道获取相关装备并进行加工、改造的难度并不大。同时，无人装备可以采取远程操控甚至自主操控的方式完成打击任务，这似乎与极端分子青睐的暴力行为模式相当匹配。美国联邦调查局在 2011 年侦破并预防了一起针对五角大楼的恐怖袭击，策划者所考虑使用的手段便是通过无人机装载爆炸物进行空袭。英国反恐专家则提出警告，认为恐怖分子有可能利用无人机对民航客机进行攻击。<sup>③</sup>使用自主武器发动恐怖袭击可能还具有一定的象征意义。戴维·邓恩（David Hastings Dunn）便指出，“考虑到美英在伊斯兰世界大量使用无人机带来的争议，有能力使用这种武器对其本土（美国和英国）发动攻击，以一种非对称战争手段对称地做出回应，对许多恐怖组织而言颇具吸引力”。<sup>④</sup>因此，在一些人看来，更加智能的自主杀伤技术落入非国家行为体手中，只会助长恐怖主义和其他危险行为，增加国际体系面临的安全威胁。<sup>⑤</sup>

假设这些安全风险真实存在，但它们是否已经足够强烈，以至于必须通过对自主武器加以限制才能避免呢？在这一点上存在着许多尖锐的反对意见。这些质疑集中反映在以下几个方面：

首先，上述安全风险在其他技术领域同样存在，而这些领域的军备控制有的甚

---

① Michael Boyle, “The Costs and Consequences of Drone Warfare,” *International Affairs*, Vol. 89, No. 1 (January 2013), pp. 1-29.

② 许多试验已经验证了军用和民用无人机都容易遭受黑客攻击和劫持，参见 Mary-Ann Russon, “Wondering How to Hack a Military Drone? It’s All on Google,” *International Business Times*, May 8, 2015, <http://www.ibtimes.co.uk/wondering-how-hack-military-drone-its-all-google-1500326>.

③ Doug Bolton, “Terrorists Could Use Drones to Attack Planes and Spread Propaganda, Government Security Adviser Warns,” *Independent*, December 6, 2015.

④ David Hastings Dunn, “Drones: Disembodied Aerial Warfare and the Unarticulated Threat,” *International Affairs*, Vol. 89, No. 5 (September 2013), p. 1243.

⑤ Frank Sauer, “Stopping ‘Killer Robots’: Why Now Is the Time to Ban Autonomous Weapons Systems,” *Arms Control Today*, October 2016, [http://www.isodarco.it/courses/andalo18/doc/sauer\\_Stopping-Killer-Robots.pdf](http://www.isodarco.it/courses/andalo18/doc/sauer_Stopping-Killer-Robots.pdf).

至尚未在国际社会形成粗浅共识。在这一点上，网络武器提供了很好的例证。<sup>①</sup> 网络武器本质上是数据代码，这使其获取、传播和扩散都较为简单；由于网络攻击在物理层面的破坏效应日益突出，特别是在伊朗核设施遭受病毒攻击后，网络武器对于战略稳定性的负面影响开始受到广泛关注；恐怖分子利用网络攻击造成关键基础设施瘫痪和社会混乱的场景假定也屡被提及。总之，网络武器同样可能为国际体系带来严重安全风险。<sup>②</sup> 然而，虽然一些国家已经掌握了较为发达的网络攻击能力，并积极利用网络技术进行信息窃取和情报活动，但高度破坏性的网络攻击（特别是针对民用基础设施的攻击）却极少出现。<sup>③</sup> 与之相应，网络空间军备控制的可行性和必要性一直处于争议之中。<sup>④</sup> 国际层面虽有关于网络空间行为规范的若干动议，但网络军控进程并未获得实质性推动。既然网络、全球快速打击、反卫星等新技术手段带来的安全威胁更加现实和深刻，而针对这些技术装备的军控进程始终踟躇不前，那么以（相对较弱的）安全风险为理由推动自主武器军备控制就必然面临更大阻力和反对。网络空间并没有出现大规模的灾难性冲突，对此一种可能的解释是，对关键基础设施（特别是民生设施）的攻击容易触碰和违背国际法的基本原则，从而将国家置于国际道义的巨大压力之下。也就是说，国家行为体在使用新兴武器技术时有可能会进行自我克制，以避免过度暴力。如果这一逻辑成立，那么给自主武器系统设限实属多此一举，因为国际人道主义法已经为武器使用提供了总体规则，同样能够对自主武器系统起到充分的约束作用。<sup>⑤</sup>

---

① 相对而言，“网络武器”（cyber weapons）这一概念更容易引起争议，因为其本质上是一种数据代码，并不符合人们关于武器装备的惯常认知。但随着网络技术快速融入各国军事战略体系，网络攻击能够引发的物理损伤越来越具有现实性，关于网络武器及其管控的探讨也开始逐渐增多。参见 Tim Stevens, “Cyberweapons: An Emerging Global Governance Architecture,” *Palgrave Communications*, Vol. 3 (January 2017), pp. 1-6。

② Lucas Kello, “The Meaning of the Cyber Revolution: Perils to Theory and Statecraft,” *International Security*, Vol. 38, No. 2 (Fall 2013), pp. 7-40.

③ Erik Gartzke, “The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth,” *International Security*, Vol. 38, No. 2 (Fall 2013), pp. 41-73; Thomas Rid, *Cyber War Will Not Take Place*, Oxford: Oxford University Press, 2013.

④ Christopher Ford, “The Trouble with Cyber Arms Control,” *The New Atlantis*, Fall 2010, pp. 52-67; Thomas Rid and Peter McBurney, “Cyber-Weapons,” *RUSI Journal*, Vol. 157, No. 1 (February 2012), pp. 6-13.

⑤ 英国对限制自主武器系统的发展表达了反对的声音，重要理由便是认为现有的国际人道主义原则仍然能够有效管控自主武器系统。参见 Owen Bowcott, “UK Opposes International Ban on Developing ‘Killer Robots,’” *The Guardian*, April 13, 2015。

其次，自主武器为国家行为体带来的战略收益可能对冲上述安全风险。虽然掌握和使用自主武器的低门槛蕴涵着诸多风险，但恰恰是这些特点使得自主武器较之常规武器而言，潜在的应用范围更广且灵活性更强。对于国家行为体来说，自主武器能够减轻其发起和参与对外军事行动时可能面临的国内舆论压力，增加执行任务的工具选项。特别是在必须做出某种姿态但又需避免局势过度恶化的情况下，自主武器可以降低做出姿态的潜在成本。美国在过去十多年里陆续在巴基斯坦、也门等地开展无人机反恐行动，年均进行空袭有数十次之多，<sup>①</sup>但国内反对声音相对于伊拉克和阿富汗战争来说并不显著。而从空袭有效性来看，有研究发现美国的无人机反恐确实降低了恐怖活动的频率和规模。<sup>②</sup>无人机空袭行动在奥巴马任内达到顶峰，或多或少也是从小布什政府的“有人”战争“虽胜尤败”吸取了教训。与此相类似，德国、以色列等国家研究人员也认为，几乎可以忽略的己方伤亡是无人机等自主武器最突出的战略价值。<sup>③</sup>因此，相比其潜藏的安全风险，自主武器能够提供的红利却是触手可及、实实在在的。成本收益分析影响着国家（特别是技术优势国家）推动军备控制的意愿，这一点在自主武器领域显然也不例外。

最后，上述主张自主武器存在严重安全风险的观点，部分也被认为言过其实。例如，针对自主武器技术会引发军备竞赛和技术扩散的看法，有学者认为这种扩散效应不会特别强烈，因为高新技术的扩散和军事应用需要强有力的组织能力和基础设施支撑，即便美国、英国、德国这些技术发达国家，在推广无人作战模式时也面临重重困难。<sup>④</sup>而在极端分子可能利用自主武器这一点上，相反的观点则认为与现有无人装备相比，恐怖分子可能更加青睐他们早已熟悉的攻击装备和袭击模式，对

---

① The Bureau of Investigative Journalism, “Drone Warfare,” <https://www.thebureauinvestigates.com/projects/drone-war>.

② Patrick Johnston and Anoop Sarbahi, “The Impact of US Drone Strikes on Terrorism in Pakistan and Afghanistan,” *International Studies Quarterly*, Vol. 60, No. 2 (June 2016), pp. 203-219.

③ Kelley Saylor, et al., “Global Perspectives: A Drone Saturated Future,” Center for a New American Security, <http://drones.cnas.org/reports/global-perspectives/>.

④ Andrea Gilli and Mauro Gilli, “The Diffusion of Drone Warfare? Industrial, Organizational, and Infrastructural Constraints: Military Innovations and the Ecosystem Challenge,” *Security Studies*, Vol. 25, No. 1 (February 2016), pp. 50-84. 关于自主武器扩散制约因素的探讨，另可参见 Micah Zenko and Sarah Kreps, “Limiting Armed Drone Proliferation,” Council on Foreign Relations Special Report, No.69, June 2014.

许多恐怖活动目标来说自主武器或许并不是更优的选择。<sup>①</sup> 换种思路,假如恐怖分子迟早会掌握自主武器技术,国家行为体更不应当否定这种武器,最好的做法应该是继续推动其发展,以提高防御能力。<sup>②</sup>

综上所述,自主武器及相关技术发展或许会给国际安全互动带来新的挑战。但就现实性而言,自主武器的安全风险并不如网络武器、太空武器等那般明显,前述提及的一些风险(例如武器扩散和可能导致冲突升级)都被认为是高新技术武器化的共有特征,而且这些风险的真实严重程度尚有待实证检验。进一步削弱国家推动自主武器军控意愿的重要因素,在于使用这类武器仍然有利可图。不论是避免己方人员伤亡,还是执行反恐等特殊作战任务,以无人机为代表的自主武器已经愈发成为相关国家强化安全战略的有力工具。如果自主武器作战样式的有效性一再得到肯定,这将会鼓励其他国家仿效这一做法,在更大范围、更深程度上推动自主武器发展和应用。有意思的是,虽然这一发展轨迹印证了前述关于自主武器扩散的担忧,但在现实上也使自主武器军控道路愈加布满荆棘。

### 五 自主武器军备控制前景不容乐观

自主武器军备控制问题从获得关注到提上议程,进展过程可谓比较顺利。但从这一问题正式成为国际军控讨论议题开始,进一步迈向实质化发展的进程将变得十分坎坷。从目前来看,要形成正式的具有约束力的军控结果,不论其内容是禁止自主武器还是限制其发展,都将极为困难。

根据前述分析可以看出,管控自主武器之所以成为国际社会探讨议题,内在缘由是对技术发展不确定性的忧虑,而恰恰是这种不确定性,也成为阻碍自主武器军控取得实质性成果的重要因素。对于自主武器军备控制的支持者来说,最有力的主张在于,让武器自主决定杀伤对象有违伦理。但这一主张暗含的前提假设是,人工智能将必然发展到某一高度,以至于军事决策者认为完全废除人工干预,让武器系统自行决定从目标识别到发起攻击的完整链条,对实

---

<sup>①</sup> Brian Jackson and David Frelinger, "Emerging Threats and Security Planning: How Should We Decide What Hypothetical Threats to Worry About?" Occasional Paper, Santa Monica, CA: Rand Corporation, 2009.

<sup>②</sup> Irving Lachow, "The Upside and Downside of Swarming Drones," *Bulletin of the Atomic Scientists*, Vol. 73, No. 2 (February 2017), pp. 96-101.

现武器作战效能最大化而言最为有利。反对控制自主武器发展的人则认为，自主武器终将足够“聪明”，能够习得并严格遵守人类的战争规范，甚至比人类做得更好。无论哪种观点，都是基于对人工智能及其相关技术发展和后果的预估，而这种预估（尤其是对技术发展后果的评估）本身就极具争议，<sup>①</sup>使得不同立场难以实现有效弥合。同样，前述关于自主武器战略安全影响的分歧，也是围绕对不确定的技术发展后果进行预测来展开的。与针对核武器、化学武器等进行的军备控制不同，自主武器军备控制并没有明确的特定对象，或者说其对象仍处于快速地不断演变之中。这意味着自主武器军备控制只能是一种预防性的控制，控制对象是技术发展尚未完全展现的潜在后果。<sup>②</sup>而这要在国际社会获得普遍共识并由此形成军控的切实行动，恐怕在可预见的时期内并不现实。

从本质上看，权力政治与道德政治的博弈将决定自主武器军备控制究竟能否实现。自主武器军备控制之所以成为显要话题，主要还是来自道德层面的考量。对人工智能潜在社会效能的深层次忧虑，加上对这种颠覆性技术军事应用的伦理担忧，汇集成国际社会对管控自主武器发展的呼吁。目前自主武器军控进程的推动者是以社会力量为主，例如非政府组织和专家学者等。在自主武器的伦理争议中，最为核心的是不应将杀人决定权让予机器，应当说这一主张仅就伦理而言还是颇具说服力的。问题在于，仅有伦理方面的动因，仍不足以推动自主武器军备控制走向现实。军备控制从根本上说首先是权力政治的产物，军控结果往往体现为主要国家现实利益计算和讨价还价的结果。因此，自主武器军备控制要获得进一步推进，不仅依赖于道德层面汇聚更强烈的社会力量，更重要的则是在安全议题上获得足够的动力。简单来说，自主武器军备控制需具备至少三项前提条件：一是伦理争议突出，特别是在自主武器是否能真正实现绝对自主这一问题上达成明确共识；二是各主要国家均认为自主武器并非军事上必不可少；三是自主武器的安全风险能够真实呈现在国

---

<sup>①</sup> 布拉德·阿伦比（Brad Allenby）认为，当前关于新兴技术安全风险的评估混淆了三个不同层次的效应，第一个层次是技术本身的工具效应，第二个层次是技术在系统层面的效应，第三个层次则是技术对人类社会、经济、政治产生的宏观效应。第三个层次效应的预测尤为困难，许多技术产生的长远影响并不是设计者一开始所能预想的。参见 Brad Allenby, “Emerging Technologies and the Future of Humanity,” *Bulletin of the Atomic Scientists*, Vol. 71, No. 6 (November 2015), pp. 29-38.

<sup>②</sup> Denise Garcia, “Future Arms, Technologies, and International Law: Preventive Security Governance,” *European Journal of International Security*, Vol. 1, No. 1 (February 2016), pp. 94-111.

际社会面前。<sup>①</sup> 显然，后两项条件要实现极为困难，这意味着自主武器军备控制进程将继续呈现道德政治独角戏的局面，前景并不乐观。有学者坦言：“特别是考虑到当今的全球化文化以及新兴技术能够提供的战略和军事优势，无论是基于文化的、竞争的或是宗教的理由，对技术发展施加有意义的限制似乎难以成功”。<sup>②</sup>

应当指出，自主武器军备控制的可能形式并不是单一的而是多元的。第一种形式是军备控制中通常使用的国际条约，即通过正式的有约束力的协定，对特定自主武器的发展加以禁止，或对其使用进行限定。这种条约将不可避免地涉及自主武器界定、履约核查、违约惩罚等一系列具体问题，这些问题直接牵动着各国战略利益，因而要达成协议的可能性最低。第二种形式是各类不具有现实约束力的“软法”，如非强制的技术标准、指导原则、行为准则等。<sup>③</sup> 比方说，人工智能研究共同体可以自行制定建议性的指导原则，对如何控制自主武器相关技术研发过程中的各类风险进行说明。或者，国家间可以推广发展和使用自主武器的行为准则，对自主武器军事应用如何与国际法相协调加以阐释。但即便是非强制性约定，如何判定行为是否合乎规范仍然相当困难，这种约定究竟能产生多大效果也令人怀疑。第三种形式则是国家单方面做出声明，承诺不使用完全自主的“机器人杀手”。<sup>④</sup> 这种方式实际上对国家约束力最低，因为绝对意义上的自主武器还颇为遥远。但即使这种单方面声明也不容易，因为这牵涉国家对声明可能对其未来军备发展带来的非预见性制约的判断。在国际层面上，这种声明对减缓自主武器军备竞赛的实质意义也十分有限。总的来看，后两种形式的军控努力可能性远远大于正式的约束性协定，但要实现这种较低层次的军控“成果”，也仍需在前述军控前提条件上取得重要进展，更

---

① 与此类似，尼古拉斯·马什（Nicholas Marsh）列举了自主武器军控向前推进的两项条件，一是形成关于“自主武器”的清晰概念，二是说服各国发展和使用自主武器并非有利可图。本文认为，概念争议的产生实际上是权力政治与道德政治博弈踟躇不前所致，主要国家对自主武器有利可图且安全风险不显著的认知，使得概念分歧成为在军控进程中讨价还价和刻意拖延的工具。参见 Nicholas Marsh, “Defining the Scope of Autonomy,” *Peace Research Institute Oslo (PRIO) Policy Brief*, No. 2, 2014.

② Brad Allenby, “Emerging Technologies and The Future of Humanity,” *Bulletin of the Atomic Scientists*, Vol. 71, No. 6 (November 2015), pp. 29-38.

③ Gary Marchant and Brad Allenby, “Soft Law: New Tools for Governing Emerging Technologies,” *Bulletin of the Atomic Scientists*, Vol. 73, No. 2 (March 2017), pp. 108-114.

④ Frank Sauer, “Stopping ‘Killer Robots’: Why Now Is the Time to Ban Autonomous Weapons Systems,” *Arms Control Today*, October 2016, [http://www.isodarco.it/courses/andalo18/doc/sauer\\_Stopping-Killer-Robots.pdf](http://www.isodarco.it/courses/andalo18/doc/sauer_Stopping-Killer-Robots.pdf).

离不开主要国家的政治决心和魄力。

总之，自主武器军备控制进程体现了全球安全治理的一般特征，即价值关怀、技术驱动和多元治理，但这些因素同样也是全球安全治理时常停滞不前的重要原因。自主武器可能冲击传统国际法涉及人的根本安全的价值规范，但又不足在国家行为体（尤其是大国）之间产生足够的安全激励以推动实质性的管控措施。对自主武器的担忧与科技的快速发展变革息息相关，但同样，技术发展的不确定性也使得人们在如何应对其负面影响上难以形成共识。

## 六 结论

管控“机器人杀手”的动议虽然已经吸引了国际社会的广泛关注，但就目前而言，推动形成有效军控产出的动力尚不充分。主要原因在于，自主武器军备控制的道德呼吁尽管较为有力，但在安全领域却还难以触发主要大国的集体行动。目前相对最好或最具现实性的军控结果，是通过非强制性的方式塑造一种反对绝对自主的“机器人杀手”的国际规范。既然实质性的军备控制很难实现，对中国来说，可以考虑积极推动和引领自主武器军备控制进程，提出关于禁止绝对自主武器的基本原则甚至单方面声明，同时提出自主武器军控的应然标准，将自主武器与现有无人装备加以区别，以便在维护自身战略利益的同时，积极抢占国际道义制高点和议程设置能力，进一步巩固负责任的大国形象。

此外，高质量的知识供给也是提升军控议程设置力和话语权的重要途径，应当大力鼓励战略研究和工程技术的深度对话，厘清人工智能和自主武器可能产生的社会效应，并通过多样化方式对自主武器相关的假定和逻辑关系加以科学验证。

总之，“将人工智能可以做出的积极贡献最大化，同时将其有害后果降至最低，将是我们这个时代最艰巨的公共政策挑战之一。”<sup>①</sup>

【收稿日期：2017-07-13】

【修回日期：2017-10-22】

【责任编辑：苏娟】

---

① 约翰·桑希尔：《用人类智慧应对人工智能挑战》，[英]《金融时报》中文网，2017年4月24日，<http://www.ftchinese.com/story/001072308#adchannelID=2100>。