自主武器如何变革未来战争——《无人军队：自主武器与未来战争》评介

朱启超　龙　坤

内容提要：保罗·沙瑞尔的近著《无人军队：自主武器与未来战争》全面考察了自主武器的发展历史、当前进展与未来趋势，探讨了人工智能对未来战争的多重影响。贯穿全书的核心问题是战场上生死攸关的决策权是否应该让渡给机器？沙瑞尔结合自身知识背景和丰富阅历，通过大量实地考察和专家访谈，对这一问题进行了多维度探讨。沙瑞尔界定了自主武器的相关概念，梳理了自主武器的优势与风险，并考察了自主武器对战略稳定性、战争法则以及战争伦理的影响。沙瑞尔还对自主武器国际军备控制进程的现状、问题及原因进行了深入分析，并提出了相关解决方案。虽然作者的立场和观点有其局限性，但该书为推动人工智能与未来军事变革的深入讨论，提供了有益参考。

关键词：人工智能　自主武器　战略稳定　战争伦理　国际人道法

自主武器对于现代战争样式、战争伦理、军队建设模式和国际战略稳定的冲击，正受到联合国军控与裁军会议、各国政要、军队将领和防务智库专家广泛关注，也成为新加坡香格里拉对话、北京香山论坛、世界和平论坛等研讨的热门话题，关于人工智能、自主武器与军事变革方面的学术研究成果也不断出

朱启超　国防科技大学前沿交叉学科学院国家安全与军事战略研究所所长、国防科技战略研究智库研究员。龙坤　国防科技大学文理学院硕士研究生、国防科技战略研究智库实习研究员。

现，[1] 美国智库新美国安全中心专家兼科技与国家安全研究项目主管保罗·沙瑞尔（Paul Scharre）的近著《无人军队：自主武器与未来战争》（*Army of None: Autonomous Weapons and the Future of War*）就是其中有代表性的一部。[2] 沙瑞尔在书中全面考察了自主武器的发展历史、当前进展与未来趋势，探讨了人工智能对未来战争的多重影响。该书一经出版，便迅速登上亚马逊军事类图书畅销榜，成为比尔·盖茨（Bill Gates）推荐的年度五佳图书之一，[3] 并荣获2019年度“威廉·E. 科尔比奖”。[4] 贯穿全书的主线是战场上生死攸关的决策权（life-and-death decisions）能否让渡给机器？这样做是合法和正确的吗？如何理解自主武器的优势和劣势？自主武器对于国际法、战争伦理和战略稳定性会带来哪些冲击？针对自主武器的军备控制能否成功？本文将对书中讨论的上述问题做简要评述。

一、自主武器的概念、历史与正负效应

关于人工智能与自主性的概念，现实中常常出现模糊不清的情况，以致陷于不同语境下的无端争执。在部分人的印象里，人工智能是《终结者》[5] 等科幻作品中向大众灌输的杀人机器形象，而在另一些人看来，人工智能只不过是“鲁姆

1 代表性成果有：徐能武、葛鸿昌：《致命性自主武器系统及军控思考》，《现代国际关系》，2018年第7期；封帅、鲁传颖：《人工智能时代的国家安全：风险与治理》，《信息安全与通信保密》，2018年第10期；Nathan Leys, “Autonomous Weapon Systems and International Crises,” *Strategic Studies Quarterly*, Vol. 12, No. 1, Spring 2018, pp. 48-73；Frank Sauer, “Autonomous Weapon Systems and Strategic Stability,” *Survival*, Vol. 59, No.5, 2017, pp.117-142; Patrick Lin, *Autonomous Military Robotics: Risk, Ethics, and Design*, US Department of Navy, Office of Naval Research, December 20, 2008, http://ethics.calpoly.edu/ONR_report.pdf, 2019-03-01; Andrew P. Williams, Paul Scharre, “ Autonomous Systems: Issues for Defense Policymakers,” https://www.researchgate.net/publication/282338125_Autonomous_Systems_Issues_for_Defence_Policymakers, 2019-03-01; Paul Scharre, “Autonomous Weapons and Operational Risk, ” Ethical Autonomy Project, Center for a New American Security, February 2016, https://www.files.ethz.ch/isn/196288/CNAS_Autonomous-weapons-operational-risk.pdf, 2019-03-01。

2 参见［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，朱启超、王姝、龙坤译，北京：世界知识出版社，2019年版。

3 Bill Gates, “When Ballistic Missiles Can See,” *Gates Notes*, December 3, 2018, https://www.gatesnotes.com/Books/Army-of-None, 2019-03-01.

4 “Paul Scharre Wins Colby Award for Book *Army of None* ,” April 23, 2019, https://www.cnas.org/press/press-release/paul-scharre-wins-colby-award-for-book-army-of-none, 2019-03-02. 威廉·E. 科尔比奖（William E. Colby Award），全称为威廉·E. 科尔比军事作家奖，该奖项于1999年由佛蒙特州诺威治大学的威廉·E. 科尔比军事作家研讨会成立，旨在表彰本年度为促进了解军事历史、情报行动或国际事务作出重大贡献的作品，具有较大的国际影响力。

5《终结者》（*The Terminator*）是美国著名科幻电影系列，著名电影杂志《电影周刊》在评选20世纪最值得收藏的一部电影时，此片以最高票数位居第一。目前，该系列已经出品包括《终结者1》《终结者2：审判日》《终结者3》《终结者2018》《终结者：创世纪》等电影。

巴”[1] 扫地机器人等生活产品。在一些人看来，自主性代表着机器具备“灵魂”或自我觉醒的意识，而在另一些人看来，自主性只不过是自动化的高级形式。大多数人对这些概念的认识尚处于一种朦胧的状态，而论述人工智能的作品又对此多语焉不详。对于这一问题，沙瑞尔在本书开始部分就给出了明确界定。他指出，自主性是赋能机器人的认知引擎。没有自主性，机器不过是任由人类操控的冰冷器械。在现实生活中，自主性并不意味着机器拥有灵魂或自由意志，而是一种机器自主执行某项任务的能力。理解自主性（autonomy）主要有三个维度。第一个维度是机器承担的任务类型，这些任务类型的重要性、复杂程度和风险不一，依据具体的任务可以界定其是否为自主系统。理解自主性的第二个维度是人机关系（human-machine relationship），据此可以将自主系统分为半自主（semi-autonomous）、有监督的自主（human-supervised autonomous）以及完全自主（fully autonomous）的系统。在半自主系统中，人处在 OODA 回路之中（in the loop），[2] 即机器执行一部分任务后，会停下来等待人类的指令授权再采取下一步行动。在有监督的自主系统中，人处在 OODA 回路之上（on the loop），机器能自主进行观察、判断、决策和行动，但人可以监督系统的运行，并在必要时进行干预。而完全自主系统则能够独立完成 OODA 回路，人处在回路之外（out of the loop）。理解自主性的第三个维度是智能程度（intelligence）。根据这一标准，可以将自主系统划分为自动的（automatic）、自动化的（automated）和自主的（autonomous）。自动系统是指简单的、基于阈值的系统，很少涉及决策过程，例如老式的恒温器。自动化系统是指一种相比自动系统更为复杂的基于规则的系统，需要考虑更多的输入条件并权衡变量，例如现代数字化可编程恒温器。自主系统则是指内部机制难以被用户掌握的复杂系统，它是目标导向的和自我指导的，例如无人驾驶汽车，只需要给予它一个目的地，过程全由机器自主规划和行驶。系统的复杂性与用户对于系统行为的可预测性是成反比的。系统越复杂，用户对其理解和预测的难度就越大。在此基础上，沙瑞尔将本书的核心概念——自主武器（autonomous weapons）界定为能够独立完成搜索目标、决定打击目标以及打击目标等全部作战任务周期的武器系统。[3] 根据前述自主性的分类，自主武器可以分为半自主武器、有监督的自主武器以及完全自主武器。通过概念的辨析与界定，沙瑞尔为人们理解人工智能与自主武器提供了一种更为清晰的框架，拨开了人工智能相关概念的迷雾。尤其是其按照人与 OODA 决策回路关系划分

1 鲁姆巴（Roomba) 是美国 iRobot 公司出品的自动清洁机器人的名字，广受消费者喜爱。

2 OODA 决策回路理论是约翰·博伊德（John Boyd，1927—1997）提出的。OODA 意指 Observation，Orientation，Decision，Action，即观察、判断、决策、行动。美军认为，规划军事行动和相关战略都可以用这一理论进行分析。博伊德当过战斗机飞行员，曾供职于美国空军参谋部装备战术需求分部，以美国空军上校军衔退役，还提出了战斗机设计影响空中格斗水平的“能量机动理论”。

3［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第57页。

自主系统程度的方法，简洁明了，为思考未来人机关系与战争形态提供了清晰思路。

基于这些概念，沙瑞尔回溯了自主武器的发展历史，并介绍了目前世界上业已存在的自主武器以及正在发展的相关项目。沙瑞尔指出，军事领域的机器人革命并非有意促成之举，而是在美军将大量机器人投入反恐战场的同时，不经意间走进了这场革命。从加特林机枪（自动武器），到精确制导武器（半自主武器系统），到“宙斯盾”（Aegis）作战系统（有监督的自主武器系统），再到“哈比”（Harpy）无人机（完全自主武器系统），沙瑞尔敏锐洞察到武器系统自动化程度不断提升的趋势。接着，沙瑞尔审视了世界上以美军为代表的正在进行的自主武器项目，包括远程反舰导弹（LRASM）、快速轻量化自主（FLA）、拒止环境下的协同作战（CODE）、对抗环境中的目标识别与匹配（TRACE）等。沙瑞尔注意到一场围绕机器人武器的军备竞赛正在悄然展开，并重点考察了韩国的SGR-A1型哨兵机器人、英国的“硫磺石导弹”（Brimstone）、英国的雷神无人机、俄罗斯的“平台–M”作战机器人等案例。他指出，自主武器军备竞赛最大的危险在于，由于担心他国率先研制出自主武器而使自身处于劣势，各国因而竞相开发自主武器。使这个问题变得更复杂的是，自主武器背后的驱动力量——人工智能是一种极其强大的技术，且大部分是软件，这意味着它可以被免费复制，并在转瞬之间跨越国界。此外，这类技术的门槛并不高，相关的技术知识很容易获取，高中生都可以制作出机器人。[1]

与所有技术和武器系统一样，自主武器也会带来正反两方面效应。自主武器的正面效应，或者说军事价值，主要体现在三个方面。一是能显著降低人力成本。以目前远程控制的无人机为例，一架美军“捕食者”无人机背后需要数十人进行操控和数据分析。如果提升无人机的自主性，使其能够自主执行相关任务，无疑将大大降低人力成本。[2] 二是拥有远高于人类的反应速度。根据约翰·博伊德（John Boyd）的OODA回路理论，作战双方谁能更快而准确地完成“观察、判断、决策、行动”这一回路，谁就更容易获得胜利。因此，利用机器远超人类的反应和决策速度将带来决定性的军事优势。美军认为，自主武器有望将这一回路的时间压缩至微秒乃至纳秒级，从而改变战争的游戏规则，尤其是在饱和攻击的情况下更为重要。[3] 三是能够在通信拒止的环境中实现自主作战。目前，依靠人类控制的武器装备均存在一个明显短板，那就是严重依赖通信，而电磁领域的对抗异常激烈。一旦通信被切断或干扰，武器系统就会失去控制，而提升武器系

1［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第142—146页。

2 United States Office of the Under Secretary of Defense, “Unmanned Systems Integrated Roadmap, FY2011–2036,” October 2011, p.50, https://info.publicintelligence.net/DoD-UAS-2011-2036.pdf, 2019-09-02.

3 United States Air Force, “Unmanned Aircraft Systems Flight Plan, 2009-2047,” May 18, 2009, p.41, https://fas.org/irp/program/collect/uas_2009. pdf, 2019-05-02.

统的自主性则有望解决这一棘手问题。

另一方面，自主武器也可能带来一些负面的风险和挑战。首先，自主武器的使用存在误判和误伤的风险。沙瑞尔指出，激活自主系统的本质在于人类将信任授予机器，而造成系统误伤的原因主要有两方面，要么是人类过于信任自主系统，对其有着“毫无保留和不加批判的信任”；[1] 另一种是对于自主系统缺乏信任，在需要发挥自主系统优势的时候没有使用自主系统。在第一种信任过度的情况下，自主系统可能以操作员预想之外的方式执行任务，就可能造成事故。为了阐述这一发现，沙瑞尔选取了两个代表性案例。第一个案例是发生在2003年伊拉克战场上的美国“爱国者”导弹部队误伤友军事件。在这一事件中，“爱国者”（Patriot）防空系统误将“龙卷风”（Tornado）友机误认为是反辐射导弹而击落。第二个案例是发生在1988年波斯湾的“文森”号驱逐舰（Vincennes）击落伊航655号客机事件。“文森”号驱逐舰战斗情报中心的人员不够信任自动化，而误将这一民用客机判断为伊朗的F–14战机，将其击落，造成290人全部遇难。[2] 这两个案例分别从信任过度和信任不足正反两方面表明，只有在恰当的场景恰到好处地信任和使用自主武器，才能避免误伤事故的发生。自主武器的第二个风险在于，推动网络空间的自主武器化可能导致“闪战”（flash war）的降临。沙瑞尔从美国股市应用自动交易算法出现故障，而引发了“骑士资本噩梦”（Knightmare）和“闪电崩盘”（Flash Crash）事件，自然联想到未来如果军事领域大量运用了自主系统，就可能会引发“闪战”。尤其是在网络空间，因为与物理空间的交互需要漫长时间不同，这一空间与股市极为相似，系统与系统之间的交互常常是以微秒来计算的。可能人类还没意识到，事故的灾难性后果就已经造成。最后，自主武器背后的支撑技术——人工智能本身存在脆弱性。引入更多自动化的元素意味着增加系统中的软件，而软件是由代码组成的，代码越多，存在漏洞和错误的概率以及遭受黑客攻击的风险就越大。同时，系统的自主性越高，复杂性就越高，用户对其理解和预测故障也就变得更加困难。目前，推动人工智能迅速发展的深度神经网络就存在容易被“欺骗”的脆弱性和不稳定性，且这类弱点往往与人类的常识和直觉不符，使得人们很难理解。针对这些风险，沙瑞尔在这里引入了查尔斯·培洛（Charles Perrow）的“正常事故理论”（normal accident），并考察了军事领域接近高可靠性的两个案例——美国海军“潜艇安全”项目（SUBSAFE）和“宙斯盾”作战系统，指出发挥人的干预作用才是实现高可靠性的关键。沙瑞尔强调，信任不等于盲目的信仰，需要在信任的基础上加强验证，并保持人在自主武器系统中的有效干预，作为“失

1 John K. Hawley, “Not by Widgets Alone: The Human Challenge of Technology-intensive Military Systems,” *Armed Forces Journal*, February 1, 2011, http://www.armedforcesjournal.com/not-by-widgets-alone/, 2019-04-02.

2［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第186—187页。

效保护装置”或“断路器”，才能有效管控自主武器带来的风险。

要在信任的基础上加强验证，并保持人在自主武器系统中的有效干预，才能有效管控自主武器带来的风险。

二、自主武器如何影响战略稳定性？

战略稳定性是20世纪的战略家针对核武器所创造的理论概念，指的是一种维持现有和平的状态，而不稳定性是指可能爆发战争的危险状态。用这一概念来分析自主武器，会得到一些有意思的结论。

战略稳定性主要有两种，一种是“先发打击稳定性”，另一种是“危机稳定性”。那么，自主武器究竟对战略稳定性会产生哪些影响呢？沙瑞尔分别从半自主武器和完全自主武器两个不同层面进行了考察。首先，自主武器如何影响先发打击稳定性？双方由于担心对方的报复自身难以承受，而自己又无法破坏对方的反击能力，因此不敢发动打击，这就形成了先发打击稳定状态。一些学者认为，无人机蜂群等半自主武器会更有利于进攻，降低武力使用门槛，因而会破坏先发打击稳定性。但沙瑞尔指出，这种论断忽略了防御方也拥有无人蜂群技术的情况。假如攻击方和防御方都拥有无人蜂群技术，那么攻防平衡并不一定会被打破。

其次，自主武器如何影响危机稳定性？从完全自主武器的层面来看，这类武器摆脱了对于通信的依赖，自然降低了攻击卫星系统等通信设备的必要性，进而减少先发打击的冲动。从这一角度来说，自主武器有助于提高先发打击稳定性。但另一方面，完全自主武器使武器系统完全脱离了人的控制，使得人无法控制危机自动升级，也无力终止战争。从这一角度来说，完全自主武器又降低了危机稳定性。关于这一问题，沙瑞尔举了一个很有意思的例子，1812年发生的新奥尔良战役正是在双方终止战争决议的消息没有到达战场的情况下发生的，造成了2000名士兵无辜阵亡。[1] 国家领袖和军事主官想要终止战争，但是却无力掌控战争进程。自主武器充斥战场的时代，是否也会发生类似的悲剧？这一点值得深思。同时，自主武器在速度方面的优势将加快战争节奏，缩短人类的决策时间，容易导致危机中的仓促应对和不必要的危机升级，这也会降低危机稳定性。

领导人心理也是影响危机稳定性的一大重要因素。[2] 在本书中，沙瑞尔特别关注了心理层面的影响。他指出，将自主武器引入战争中相当于引入了除冲突双方领导人之外的又一大变量，而这一变量在目前看来是难以预测和解释的。但与

1［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第343页。

2 罗伯特·杰维斯（Robert Jervis）的著作《国际政治中的知觉与错误知觉》是这一领域的代表作。参见：［美］罗伯特·杰维斯：《国际政治中的知觉与错误知觉》，秦亚青译，上海：上海人民出版社，2002年版。

我们的直觉不一致的是，沙瑞尔指出，正是由于在多数人的认知中，自主武器都是不可预测的，战略稳定性反而有可能得到提升。这里面的逻辑可以称为“疯狂机器理论”（mad robot theory），[1] 即自主武器行为的不可预测性反而引发了双方领导人的警惕，增强双方的相互威慑效力，进而提升稳定性。但同时沙瑞尔也强调，机器难以理解其行为可能带来的后果以及领导人的真实意图，没有办法像人类一样具有同理心且考虑全局，从而可能在危机中“将绳结越打越紧”，[2] 无法从战争的边缘悬崖勒马。

总体来看，沙瑞尔对于自主武器与战略稳定性这一议题的讨论是开放性的，人工智能技术正在快速发展和应用，很多影响目前还只是初露端倪。因此，自主武器究竟会增强战略稳定性还是破坏战略稳定性，目前还难有定论，需要进一步观察和研究。

三、自主武器如何挑战传统战争伦理？

自主武器的战场应用会带来很多伦理挑战，而沙瑞尔在这一部分的阐述可谓全书最精彩的部分，引人深思。沙瑞尔将自主武器形象地描述为“没有灵魂的杀手”（soulless killer）。[3] 没有人类的道德判断，机器在射杀帮助塔利班游击队侦察的小女孩时就会毫不手软，而不会像人类战士一样心存怜悯。这一点切中了人类与机器的一大本质区别——基于同理心与怜悯的道德判断，而这种道德判断是人之为人的重要表征。

当前，结果主义（consequentialism）与道德主义（deontological ethics）是讨论自主武器与伦理关系的两种主流框架。结果主义流派认为，对一件事情进行道德判断主要看它所造成的结果。而道德主义流派宣称，一件事情的对错是由其本身的规则决定的，与结果无关。[4] 结果主义在分析自主武器的伦理问题时，关注点主要放在自主武器对于战场伤亡尤其是平民伤亡这一结果的影响上。对此，沙瑞尔总结了自主武器可能会带来更多杀戮和伤亡的几点可能性。一是没有怜悯之心的机器会消除因心软而放过对手的现象。沙瑞尔援引了哲学家迈克尔·沃尔泽（Michael Walzer）对战争中的同理心与怜悯问题的研究成果，认为人类战争中存

1［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第354页。

2 这一隐喻源自古巴导弹危机期间赫鲁晓夫写给肯尼迪的信件，赫鲁晓夫奉劝肯尼迪不要采取冒险性行动将两国推向全面核战争的深渊，乃至两国领导人都无法停止（即将绳结越拉越紧，乃至拉绳子的人都无法解开，最终只能剪掉绳子）。参见 Department of State Telegram Transmitting Letter from Chairman Khrushchev to President Kennedy, October 26，1962，http://microsites.jfklibrary.org/cmc/oct26/doc4.html，2019年5月3日登录。

3［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第306页。

4 同上书，第307页。

在很多因怜悯而放对手一条生路的现象，这一现象也被称为“赤裸士兵”时刻（“naked soldier” moments）。[1] 而在自主武器的面前，这一现象很可能会不复存在，因为机器和算法只知道僵硬地执行程序任务，而不会存在任何所谓的怜悯之心。从这一角度而言，自主武器可能会比以往带来更多的战场伤亡。二是自主武器将人从瞄准和击杀的决策链条中移除，就会减轻甚至消除人在进行杀戮时所产生的道德责任和心理负疚感，从而导致更多的杀戮和伤亡。这一逻辑在于，人类不是一种冷血动物，对于自我犯下的杀戮行为或多或少都会存在心理负疚感。无论是冷兵器时代的短兵相接、手刃对手，还是当今以远程操控导弹来击毁敌方基地，看到自己的行为造成了与自己同为人类的对手或平民死亡，都会不可避免地带来武力使用者的心灵冲击感和负疚感。问题在于，将杀戮的决定让渡给冰冷的机器，或赋予机器以道德主体地位，无疑会增加杀戮的心理距离，弱化人类在杀戮问题上的道德负疚感，从而减少对于暴力使用的克制，甚至没有人感到有必要阻止自主武器的使用，这才是最令人担忧的结果。

另一方面，沙瑞尔也总结出了自主武器帮助减少战争中伤亡和暴行的可能性。首先，机器的冷酷性特征使得它们不会像人类一样因为人性恶念而对平民或俘虏施加酷刑、谋杀或强奸等暴行。在历史上的战争中，往往存在着人类的种种战争罪行，手握强大武器的士兵在某一刻萌生了人性邪恶的欲念，而无法抑制住内心的堕落，往往会将战争法则抛之脑后，从而使无辜的平民或俘虏惨遭暴行。而机器是冷酷的，没有人类的七情六欲，自然也就排除了这些基于人性欲念的暴行。其次，自主武器的精确性也有助于减少附带损伤。正如精确制导弹药能够有效减少人员伤亡一样，自主武器将成为下一代精确制导武器，比人类更加精确可靠，不会感到恐惧、嫉妒、愤怒，不会寻求报复，更不会背叛、逃跑和自杀。理论上，只要能对它进行有效编程，使之遵守战争法，就能在必要时进行杀戮，在行为非法时立刻终止，成为遵规守纪而不知疲倦的“完美士兵”。最后，随着技术的进步，理论上可以制造出遵守战争法则的自主武器，即为自主武器设计一个“伦理调节器”，阻止自主武器进行非法或不道德的行为，从而减少平民的无辜伤亡。自主武器系统还可以充当士兵的道德顾问，从而改善人类在战争中基于欲念的暴力行为。[2]

道德主义是分析自主武器影响战争伦理的另一种理念。这种理论认为，自主武器在本质上就是“毫无人性的、反人类的和不道德的”，[3] 即便它可以挽救更多的生命，也是错误的、根本不可取的。在这一问题上，沙瑞尔援引了哲学家彼

1 Michael Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, 4th Ed., New York: Basic Books, 1977, pp.138–142.

2［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第320页。

3 同上书，第322页。

得·阿萨罗（Peter Asaro）的观点。阿萨罗认为，讨论自主武器与伦理，最根本的问题就是其对于人的尊严和人权的影响。自主武器杀害人类，从根本上就侵犯了人的尊严。基于算法的杀人决定是武断的，没有经过人深思熟虑的生杀决策，是对人权和人类尊严的根本侵犯。[1] 但沙瑞尔对这一论断提出了尖锐质疑。在人的尊严方面，沙瑞尔认为，从来没有任何法律、道德或历史传统规定，战斗人员需要为敌人提供在战争中有尊严死亡的权利。相比自主武器杀人，人在战争中被敌人操控的机枪射杀、炸弹粉碎或被基于仇恨的人类间种族清洗和屠杀，都不存在任何尊严，也没有理由让我们感觉更好。因为，暴力和残酷向来是战争的本质，也是使对手屈服的必要手段。沙瑞尔指出，在分析自主武器与战争伦理问题时，我们需要分清楚哪些问题是关于战争本身的，哪些是关于自主武器的，而不能将二者混为一谈。例如，批评自主武器侵犯人的尊严问题就明显超出了自主武器的边界，而讨论到了战争的残酷性，而这一点是亘古未变的，与自主武器本身关系不大，因此意义也不大。

自主武器的战场运用还可能会剥夺军队专业人士做出生死攸关决策的权利，从而切断人类与武力使用之间长期存在的必然联系。很长时间以来，在波诡云谲的战争迷雾和相互矛盾的价值观中做出判断和决策，就是军队专业人员的职责所在。然而，自主武器的发展和应用可能会对这一传统造成冲击。在自主武器时代，军队专业人员的传统职能可能会被自主武器系统取而代之，武力使用将不再是人类的特权，生杀大权也不再是人类的专属，其本身存在的必要性也会遭受质疑。完全自主武器甚至会将感情因素彻底从战争中移除，使杀戮变得毫无人性和人道。

四、自主武器如何冲击国际人道法原则？

现有的国际武装冲突法（也称为国际人道主义法，International Humanitarian Law）主要有三个核心原则——区分性原则（the principle of distinction）、相称性原则（the principle of proportionality）和避免不必要痛苦原则（the principle of avoiding unnecessary suffering）。[2] 自主武器是否受到国际人道法的规制，主要就看能否符合这三条核心原则。对此，沙瑞尔进行了详细考察，得出了一些富有启发性的结论。

区分性原则是指，在战争中发动攻击时必须区分军事目标和非军事目标，不能故意攻击对方平民和其他民事目标。深度神经网络在目标识别领域的准确性，使得自主武器在区分军事物体和非军事物体方面有一些优势，但是这仅限于合作

1［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第325页。

2 同上书，第283页。

目标，在非合作目标和杂物等信号混杂进来之后，便仍然十分困难，且本身容易受到“欺骗攻击”。自主武器在区分军人和平民方面则存在更大的难度。因为，现实战斗中往往牵涉各色各样身份的人，军人、平民、警察、叛乱分子等混杂不清。在战争的乱局中，连拥有情境分析和判断能力的专业军人都很难做出区分，遑论冰冷而僵硬的机器。相称性原则是指，军事打击的必要性必须超过预期的民事附带损伤。在太空、深海等没有平民存在的环境中，自主武器符合这一原则并不困难，因为这里几乎不涉及对平民的附带伤害。但在人口稠密的地区，这一问题就变得十分棘手。问题的关键在于，自主武器很难判断和权衡军事必要性与附带损伤之间的比例关系。尤其是涉及平民伤亡时，需要复杂的道德推理能力，权衡多种行动方案的利弊及预期影响，而目前人工智能尚未实现这种能力，因此将自主武器部署战场会对这一原则带来很大冲击。避免不必要痛苦原则禁止使用会造成超出其军事价值的多余痛苦的武器，例如爆炸子弹、化学武器、激光致盲武器等。但问题是，这一原则很难对自主武器形成规制，因为自主武器主要涉及的是决策过程而非伤害机制。总的来看，目前的人工智能技术还很难使自主武器符合这三大颠扑不破的原则。但当前技术发展日新月异，人工智能已经在德州扑克、智力问答、围棋等人类智慧领域一路攻城略地，未来技术能否发展到符合这三大原则的程度，或者是否需要为自主武器另外制定国际人道法原则，还有待进一步观察。

除了这三大核心原则外，还有一些国际人道法规则也值得参考。例如，“攻击时的预防措施”（Precautions in Attack）规则要求准备或决定攻击的一方应采取一切可行的预防措施来避免伤害平民。但这一原则需要视具体的作战环境和可用的作战手段而定。此外，“退出战斗”（hors de combat）[1] 法则强调禁止伤害已经投降或无法战斗的参战人员。退出战斗包括被俘获、明确表示投降意愿以及无意识或因伤病而丧失作战能力的人员。第一种类型比较容易做到，但后两种类型就比较困难。问题的关键在于，机器人不能识别假投降、假受伤等涉及辨别人类意图的情形，容易受到敌方的欺骗。

另一大挑战是自主武器可能带来“问责空白”（accountability gap），即没有人能够为自主武器造成的后果负责。自主武器可能会误杀平民，这时产生的责任归属问题就变得十分棘手。战争中的问责制使得受害者或其家属能够通过“报应性司法”（retributive justice）惩罚犯罪者，以此阻止未来的非法行动。问责空白更大的危害是它可能引发一场道德危机，使得战场上的杀戮变得更加随意，因为没有人需要为此负责。此外，公众良知（public conscience）也是一个重要概念。但作者认为这一概念只不过是一根“脆弱的芦苇”，很难在推动自主武器军备控制上发挥实际效力。因为，这一概念并没有写进国际法，什么才算是公众的良知

1 法语，意指“out of the fight”或“no longer able to fight”，即主观上愿意退出战斗或丧失继续战斗的能力。

并没有一个固定的标准，很难进行测量，也容易受到“预设因素”的影响。[1]

总体而言，目前围绕自主武器对现有国际人道法冲击的辩论还在继续，很多问题仍悬而未决。例如，现有国际法是否足以规制自主武器的战场运用？是否需要为自主武器建立新的国际法规范？自主武器究竟能减少还是加剧平民伤亡？尽管这些争论目前并没有结果，但沙瑞尔一针见血地指出，在战争法则这一问题上，有一点必须要明确。那就是，制定于人类主导战场行动时代的传统战争法则，约束的对象不是机器，而是人类本身。因此，要想使传统国际人道法继续适用和有效约束自主武器的战场运用，就必须保持人类对攸关生死决策的控制权。毋庸置疑，如果让机器掌握完全的生死决策权，无异于使战争行为脱离战争法约束的缰绳。

要想使传统国际人道法继续适用和有效约束自主武器的战场运用，就必须保持人类对攸关生死决策的控制权。

五、针对自主武器的军备控制能否成功？

自主武器的军备控制问题已成为继核武器、生化武器、网络武器等议题之后的最新话题，也是各大国际安全论坛的焦点。在这一问题上，沙瑞尔秉持现实主义观点，认为制造更先进的军事人工智能系统是一种难以阻挡的趋势，希望各国不将人工智能这一强大而泛在的技术军事化，无异于期望各国不将电力运用至军事领域，是一种幼稚到了天真的想法。但同时，他又对致命性自主武器军备控制怀有一些希望。

沙瑞尔将人类在武器系统中扮演的角色分为武器操作者（essential operator）、失效保护者（fail-safe）以及道德判断者（moral agent）。他认为，随着武器系统自主性的提升，第一种角色的作用会越来越被淡化，后两种角色的重要性则会日益突出。从国际象棋世界冠军加里·卡斯帕罗夫（Gary Kasparov）被“深蓝”（Deep Blue）系统击败后开发出“半人马象棋”（centaur chess）的故事中获得启发，沙瑞尔预测，未来最好的作战系统将会是混合人机认知系统，即“半人马战士”（centaur warfighters），这是一种既发挥人工智能在信息处理精准性和可靠性方面的优势，也发挥人类决策鲁棒性和灵活性的复杂系统。但是，“半人马战士”这种理想化的人机编组模式，在有些情况下未必适用。当自主武器系统要求反应速度快于人类时，就需要采取有监督的自主武器系统。例如，当来袭导弹呈现饱和攻击态势时，实现自主防御就变得很必要。因为，在这种情况下，人类的反应速度可能根本无法有效应对，而这种代价可能是致命的。此外，当人类操作员与武器系统之间的通信受阻时，完全自主武器系统的重要性就凸显了。[2] 因为，完

1［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第296—299页。

2 同上书，第364—371页。

全自主武器不需要与人类操作员进行通信就能自主做出相关的作战行动，因而排除了对通信的需求。

在全书后半部分，沙瑞尔系统梳理了人类历史上主要的军备控制案例，总结出了影响军备控制成败的三个主要因素：对武器使用的可怕后果的认知、对武器军事价值的认知以及参与军备控制行为体之间的合作。[1] 沙瑞尔发现，成功的军备控制主要有两个共同特征，一是人道主义原因，即这类武器的使用造成的不必要痛苦或平民伤亡，超过了其军事应用价值，比如集束炸弹和地雷。另一个特征是这类武器威胁到了战略稳定性，比如核武器和化学武器。[2] 近年来，国际上在推动致命性武器军备控制上已经进行了很多努力，出现了多样化的平台和主体。联合国《特定常规武器公约》（Convention on Certain Conventional Weapons，CCW）就是官方多边会谈机制的代表。近年来，该机制已经就致命性自主武器军备控制议题召开了多次正式会议，并成立了专门的政府专家组（Group of Governmental Experts on Lethal Autonomous Weapons Systems，简称GGE on LAWS），吸引了近百个国家、大学科研机构和非政府组织参与进来。[3] 此外，国际红十字委员会、“阻止杀人机器人运动”（Campaign to Stop Killer Robots）[4] 等非政府组织也在积极地推动致命性自主武器系统的军控进程。尽管如此，自主武器军备控制的总体进展依然缓慢，并没有取得实质成果。沙瑞尔认为，这一困境的形成主要有三方面原因。一是定义问题。各国对于自主武器的定义并没有统一的认识，概念边界无法划定。二是技术问题。目前，自主武器的使用方式、使用环境以及对未来战争的影响仍不明晰，尚处在争论之中。三是政治原因。各国都希望从自主武器这一新兴领域最大限度地维护本国国家利益。更为关键的是，当前推动自主武器军备控制的主体是非政府组织而非主权国家，且参与的国家中并没有军事大国，而历史上成功的军控禁令大多由军事大国主导。正是由于这一现状，目前致命性自主武器军备控制的前景并不乐观。

尽管困难重重，但目前在致命性自主武器系统这一问题上，各国就保持必要的人类干预这条底线已达成了基本共识。沙瑞尔指出，要真正限制自主武器，需要明确关键概念，从理论上阐释某些自主武器的危害大于作战效用，并保证透明度。为此，作者提出了四种方案：一是禁止使用完全自主武器；二是禁止以人为

1［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第373—374页。

2 同上书，第386—387页。

3 关于这一机制近年讨论LAWS问题的主要观点和进展，可参见笔者发表的论文：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5期。也可参见联合国CCW官网：https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument，2019年6月5日登录。

4 “阻止杀人机器人运动”成立于2012年，目前已经发展成为近百个国际、区域和国家非政府组织组成的全球联盟，旨在预先禁止使用完全自主武器，保持人对武力使用的有意义的控制。参见其官网：https://www.stopkillerrobots.org/，2019年6月5日登录。

目标的自主武器；三是为自主武器建立行为准则；四是创建保持人类判断在战争中作用的通用准则。[1] 沙瑞尔强调，权衡人类决策对于自主武器干预的必要性，需要凝聚社会各界的智慧进行讨论，而不仅仅是学者、律师和军人的事情。为了避免出现一个各国武装力量皆无法控制的未来，世界各国需要进行积极合作。不难预见，这本书的问世也将进一步推动这一领域的讨论，并为国际社会探讨应对自主武器扩散的“药方”提供一些思路和启示。

六、结语

文面如人面。《无人军队：自主武器与未来战争》一书呈现的丰富内容，折射出本书作者保罗·沙瑞尔的丰富阅历。沙瑞尔1997年至2001年就读于美国圣路易斯华盛顿大学（Washington University in St. Louis），获得物理学学士学位和“优秀毕业生”称号。从华盛顿大学毕业后，沙瑞尔加入了美军特种作战司令部第75游骑兵团三营特种作战监视小组，担任狙击手和指挥官。在此期间，他三次被派驻阿富汗战场执行任务。2005年，他回到圣路易斯华盛顿大学攻读政治经济与公共政策硕士学位，并在2006年顺利毕业。随后，他被派往伊拉克战场，在伊拉克迪亚拉省第486民政营民政阿尔法工作队担任民政事务专员。2008年，沙瑞尔被调入美国防部负责政策的副部长办公室担任军事力量规划师（force planner）。在这里，他领导国防部工作小组起草了《国防部3000.09指令》，[2] 并制定了美军在武器系统自主性、情报监视侦察、定向能等技术领域的防务政策。[3] 2012年，他又升任负责国防政策的助理国防部长的特别顾问（Special Assistant to the Under Secretary of Defense for Policy），为制定无人自主系统、新兴武器技术等领域的防务政策发挥了关键作用。2013年，沙瑞尔离开五角大楼，进入了美国知名智库新美国安全中心（Center for a New American Security，CNAS）担任高级研究员和“科技与国家安全”（Technology and National Security）项目主任，领导该中心在人工智能、自主武器、未来战争等领域的研究。沙瑞尔著述颇丰，出版了《超级士兵》《自主武器与人类控制》《自主

1［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第395—401页。

2 U.S. Department of Defense DIRECTIVE 3000.09, *Autonomy in Weapon Systems*, November 21, 2012, https://cryptome.org/dodi/dodd-3000-09.pdf, 2019-06-07.

3 例如，沙瑞尔参与撰写了《2012国防战略指南》（*2012 Defense Strategic Guidance*）、《2010年四年防务评论》（*2010 Quadrennial Defense Review*），以及《部长级战略规划指南》（*Secretary-level Planning Guidance*）。参见U.S. Department of Defense: Sustaining U.S. Global Leadership: Priorities For 21st Century Defense, January 2010, https://archive.defense.gov/news/Defense_Strategic_Guidance.pdf, 2019年6月7日登录；Quadrennial Defense Review Report，February 2010，https://dod.defense.gov/Portals/1/features/defenseReviews/QDR/QDR_as_of_29JAN10_1600.pdf，2019年6月7日登录。

武器与作战风险》《20YY：机器人时代的战争》等多份高影响力的报告，[1] 并在《纽约时报》《华尔街日报》《时代周刊》《外交政策》《外交事务》《国家利益》等媒体和学术刊物上发表多篇文章和评论。他还担任美国外交关系委员会（CFR）的见习会员，多次在美国国会参众两院的武装力量委员会提供证词，并受邀在联合国、北约、五角大楼、中情局等机构举办的会议上发表演讲。[2]

从上述经历可看出，沙瑞尔既有基层的锤炼，又有高层战略规划的经验，还有着文理结合的教育背景，因此被比尔·盖茨称为“实战经历与高层战略思维兼备的思想家和作家”。[3] 文理结合的知识背景和丰富的工作阅历，使得沙瑞尔在处理自主武器与战争这一严肃主题上具有灵活的驾驭能力，在字里行间不时闪现出独有的洞察力，既体现出作者对人工智能的热情和期待，也流露出对自主武器可能破坏稳定性、挑战伦理道德、威胁战争法则的深深担忧。更难能可贵的是，沙瑞尔并没有仅仅站在美国的角度去审视自主武器与未来战争，而是某种程度上有着一种对全人类的终极关怀。沙瑞尔在全书的最后引用了科幻电影《终结者》中的一句经典台词，“未来并没有被设定。没有天定的命运，而是人类主宰自己的未来”。作为一个整体，人类可以决定将人工智能技术用于何种目的，也可以决定是否赋予机器足够自主性，使其决定自身行动而不受人类判断和决策影响。一言以蔽之，授权的基础是信任，人类对机器信任的多寡，一定程度上决定着机器将如何运转及其结果。人类可以合理地运用人工智能，建立一个更加安全、更多同情和关爱的世界，一个更少痛苦、事故和暴行的世界，一个运行良好但保留人类在必要时施加干预能力的人类主导的智能世界。或者，人类会被机器的快速运算速度、近乎完美的精准性等种种优点所诱惑，从而毫无保留地完全信赖机器，最后可能给世界带来灾难性后果。沙瑞尔强调，针对自主武器的军备控制，需要社会各界的参与和讨论，才能做出真正符合人类共同利益的明智选择。[4]

当然，这本书也难免存在一些局限性。例如，沙瑞尔指出，短期来看，自主武器将使得军事大国变得更加强大。但长期来看，随着这类技术和武器装备的扩散，天平会逐渐朝着相对弱小的国家倾斜。但他并没有详细分析自主武器对世界权力动态的影响机制及其背后的深层原因。对于致命性自主武器军备控制这一日益凸显的全球问题，沙瑞尔也只是简单地提出了四种方案，并没有对此进行深入阐述。此外，沙瑞尔在书中的一些观点和立场，也明显带有美式霸权色彩以及

1 参见保罗·沙瑞尔的个人主页：https://www.paulscharre.com/analysis，2019年6月10日登录。

2 参见新美国安全中心对沙瑞尔的介绍，https://www.cnas.org/people/paul-scharre，2019年6月10日登录。关于沙瑞尔的个人简历，可参阅 https://s3.amazonaws.com/files.cnas.org/documents/Scharre_Paul_Bio_050819.pdf?mtime=20190508103555，2019年6月10日登录。

3 Bill Gates, “When Ballistic Missiles Can See,” *Gates Notes*, December 3, 2018, https://www.gatesnotes.com/Books/Army-of-None, 2019-06-11.

4［美］保罗·沙瑞尔：《无人军队：自主武器与未来战争》，第405—406页。

面对其他国家的某种优越感。例如，在引言部分，他将叙利亚、伊拉克等国视为“流氓国家”（rogue states），认为这些最残暴的政权会无视自主武器军备控制协议，从而可能使人类面临最为黑暗的噩梦。[1] 在正文中，沙瑞尔使用“尖阁列岛”指代“钓鱼岛”，[2] 且没有做出特别说明，无疑有偏袒日本的倾向。在后记中，他点名批评了中俄两个大国，指出“俄罗斯正在利用机器人传播虚假信息，企图破坏西方民主国家的政治制度”，而中国则“正在朝着一个技术反乌托邦的监视国家发展，试图通过人脸识别技术和社会信用体系加强对公民的管理”。[3] 这些论断无疑都有失客观，存在着美式思维常见的偏见，需要读者在阅读过程中仔细甄别。

1［美］保罗·沙瑞尔:《无人军队：自主武器与未来战争》，第8页。

2 同上书，第229页。

3 同上书，第462—463页。