2019 年第 8 期（总第 8 期）

国际战略与安全研究报告

INTERNATIONAL SECURITY AND STRATEGY STUDIES REPORT

不确定性：为何担心人工智能冲击国际安全？

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

不确定性：为何担心人工智能冲击国际安全？

陈 琪（清华大学战略与安全研究中心秘书长 / 教授）

朱荣生（清华大学战略与安全研究中心助理研究员）

科学家普遍认为，当前人工智能处于“弱人工智能”阶段，只能完成某一项特定的任务或解决某一特定问题。但也一些专家断言，人工智能军事化就如同核武器一般，将给国际战略带来范式性的冲击。然而，这场革命何时才能向世人呈现其全貌目前尚不可知。我们已经看到，一些人工智能企业迫于社会伦理的压力而放弃价值不菲的政府军事采购订单。但纵使智能系统被成功开发，国家也要花费数月甚至数年的时间来审批其使用。[①] 就武器产生的变革效应和方式而言，核武器与人工智能在性质上颇为不同。核武器在诞生之初就展示出了惊人的破坏力，进而对国际安全产生了深远的心理和物理影响。人工智能是赋能技术，必须与其他武器技术结合才有可能引发战场革命。而一些比较悲观论的专家则进一步预言，现有理论研究不足将限制人工智能的发展，人工智能在未来可能会陷入“寒冬”。既然人工智能带来的革命性影响是一个未知数，为何国际社会开始急于寻找关于人工智能的国际安全治理的答案呢?

我们在近期众多的国内会议和国际会议中发现，包括政策制定者、科学家、工程师、政治学者等在内的很多人都对人工智能带来的国际安全冲击抱有显著的焦虑和不安。这些焦虑和不安并非来自于人类对长期风险的警惕，而是因为人工智能技术以及其应用中产生的各种不确定性可能正在对国际安全形成巨大冲击，以及人们对这些不确定性

① *National Security Commission on Artificial Intelligence Interim Report*, Nov. 11th 2019, https://drive.google.com/file/d/153OrxnuGEjsUvlxWsFYauslwNeCEkvUb/view.

的认知动态变化。本报告主要探讨人工智能对国际安全带来的挑战，核心观点是：人工智能技术给国际安全带来冲击的同时也潜藏着成为稳定性因素的可能性。置于我们眼前的应该是慎种而明智的抉择，而不是不确定性带来的恐惧和绝望。人工智能最终将化身为“终结者”而给人类带来灭顶之灾，还是推动人类文明发展的“蒸汽机”，取决于人类如何认识和利用它。

一、人工智能冲击国际安全

人工智能是越来越重要的快速发展的基础技术。它可以通过赋能众多产业，提高国家经济竞争力，并且引发新一轮的产业革命。[②] 历史上的工业革命不仅改变了国家之间的权势争夺，还伴随权力转移以及接踵而来的大国兴衰。率先成功开发出新技术并在市场获得成功的国家将获得巨大的经济收益。根据麦肯锡的预测，考虑到转型成本和竞争效应，到2030年人工智能带来的经济增长将高达约13万亿美元，并使全球GDP每年上升约1.2%。这种经济冲击堪比19世纪的蒸汽动力、20世纪的工业制造和21世纪的信息技术对世界经济的影响。[③] 然而，这种数字经济增长并非公平分享。具有技术、资金、数据、人才等要素优势的发达经济体更可能收获更大的经济效益。广泛使用可信赖的人工智能产品将营造出更包容的环境，提高社会对新技术应用的容纳程度。数字基础设施不足、资本和技术基础薄弱以及基础更差的经济体则愈加难以追赶有先发优势的国家。数字鸿沟不可避免地越拉越大。

---

② 关于人工智能及其他数字技术赋能经济带来的发展差距，可参考 “The Age of Digital Interdependence,” *Report of the UN Secretary-General's High-level Panel on Digital Cooperation,* June 2019; “Digital economics How AI and robotics are changing our work and our lives,” Deutsche Bank Research, May 14th 2018。

③ “Notes From The AI Frontier Modeling The Impact of AI On The World Economy,” McKinsey Global Institute, Sep. 2018.

在人工智能竞赛中，有赢家，也有输家。没能成功开发人工智能技术并抢占市场的国家可能遭受经济损失。同样，在军事应用的投资和开发不足的国家将面临更大的安全威胁，其地缘影响力将被削弱。④2016 年以来，世界上多国纷纷公布人工智能发展战略，将自己定位为人工智能领导者或者表达出扮演关键角色的强烈意愿。以此观之，技术竞争绝非是中性的。随着通信技术的不断发展，有先发优势的国家想要确保在全球数字转型中获取更大的权势，或者防止自己屈居人下。故此，围绕着科技制高点的争夺不可避免地引发一场“大国政治的悲剧”。⑤

中美之间的贸易战和科技领导权之争印证了一论断。政策制定者很可能会高估技术突破带来的权势，低估国家间合作稳定国际安全的可能。按照现实主义的逻辑，大国怯于在这场影响国运的竞争中放松神经，它们既不确定自己会否成为赢家，也难以猜想在新的技术秩序变革中自身地位出现何种地位。于是，中美之间的大国竞争油然而起，正导致世界分裂成两个对立的体系。“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。⑥

令人担忧的不只是国际秩序向对抗性演进。科技的不断发展正在侵蚀战略威慑的基础，动摇大国间的战略稳定及其预期。拥核国家必然部署强大的第二次打击能力，确保对手不敢冒着同样被摧毁的风险而发动第一次打击。因此，确保报复可信度成为了大国核战略博弈的

---

④ Daniel Castro, Michael McLaughlin,Eline Chivot, “Who Is Winning the AI Race: China, the EU or the United States?”,*Center for Data Innovation,* August 2019.

⑤ John Mearsheimer, *The tragedy of great power politics*, W.W.Norton & Company, 2001.

⑥ António Guterres, “Address to the 74th Session of the UN General Assembly,” *United Nations Secretary General Speech*,24th September 2019, https://www.un.org/sg/en/content/sg/speeches/2019-09-24/address-74th-general-assembly.

重要基础。⑦ 可是，人工智能技术并不需要完全破坏大国战略博弈的基础，只需削弱核打击报复可信度就足够了。

人工智能已经具备了帮助人类从大量信息中筛选出导弹发射平台的能力。⑧ 强大的侦察力让中国和俄罗斯越来越来担忧美国发展跟踪和锁定移动导弹发射器技术发展成熟后会威胁到它们的报复能力。⑨ 如果无人武器设备的隐蔽性和突破性足够强大，国家就有风险更低、打击效能更高的攻击选择。⑩ 这使得进攻方有更大的战略优势。虽然这种战术不能保证自己能够免遭第二次打击，但这种可能性本身就已经十分可怕。面对核威慑，决策者不得不在极其有限的时间内做出决策，因而要承受更大的发动第一次打击的压力。或者，国家发展更加危险的武器来平衡威慑不足。由此引发的军备竞赛迫使国家部署不安全的人工智能系统，并进一步加大战略不稳定。对于防守方来说还有另一种选择，那就是攻击敌方的探查器件，生成对抗性网络或者采取战略欺骗的方式来防止对方打击自己的报复力量。这些努力将在一定程度上确保己方核报复设施的隐蔽性和安全性。但是，这种安全困境的加剧反而可能导致意外升级和加大战略部署判断的复杂性和误判。⑪

---

⑦ Keir A Lieber and Daryl G Press,"The new era of counterforce: Technological change and the future of nuclear deterrence," *International Security,* Vol. 41, No. 4, 2017, p. 9.

⑧ Richard A Marcum, et al, "Rapid broad area search and detection of Chinese surface-to-air missile sites using deep convolutional neural networks," *Journal of Applied Remote Sensing,* Vol. 11, No. 4, Nov. 13th, 2017.

⑨ 吉斯特·爱德华和安德鲁·洛恩：《人工智能对核战争风险意味几何？》，兰德公司，2018 年。

⑩ 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.

⑪ Jurgen Altmann and Frank Sauer, "Autonomous Weapons and Strategic Stability," *Survival*, Vol. 59, No. 5, 2017, pp. 121-127; Vincent Boulanin, "The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk", Vol. I; "Euro-Atlantic Perspectives," *Stockholm International Peace Research Institute,* 2019.

最终，国家不得不面对要么提前发动进攻，要么输掉战争的两难选择。

除了冲击战略稳定，人工智能还会改变国家间均势，加剧大国间冲突。算法战时代是以收集数据和训练算法实现军事实力中“质”的提升。[12] 这种“质”的飞跃在战场上将起到明显的效果。有技术优势的国家将在战场上部署更加先进的武器装备或者建立新的作战概念。技术落后的国家则因缺少应对策略而陷入相对劣势。按照现实主义的安全模式，大国可能寻求相对优势或者恐惧他国的军事优势威胁自己的安全，或者在其他地方寻找抵消对手优势的方法。这不仅激发新一轮的军备竞赛，还会加大国家间的战略互疑，引发预期之外的国际冲突。比如，一段“深度伪造”篡改的美军士兵被俄罗斯毒气杀死的虚假视频就可能导致两个大国动用核力量威慑对方。[13] 在军备竞赛的螺线轨迹上，自主武器越发先进以至于战争速度快到超出人类的反应，或者国家争相部署不安全的人工智能武器，都将给决策者带来巨大的心理压力，扭曲人类对战略应对的理性判断。

此外，自主武器扩散将降低战争门槛，加大国家间战争风险。军事行动往往伴随着人员伤亡的高风险。国家领导人要顾及到民众对战争伤亡的敏感性而不敢轻易对外军事行动。但是，自主武器减少伤亡率预期的特点会改变战略决策过程。决策者能够说服民众只需要较小的代价就能获得更高的战争回报。大国会进一步摆脱使用武力的国内限制，更有利于它们对外投射自己的军事影响力。不过，自主武器的扩散的最大受益人或许并不是大国而是技术基础较好的中等国家。它们可以借此弥补资源和人口上的劣势，从而改变国际常规武力的分配，提高自己在国际均势体系中的地位。在致命性自主杀伤武器缺乏有效

---

⑫ James Johnson,“Artificial intelligence & future warfare: implications for international security,” *Defense & Security Analysis,* Vol. 35, No. 2, 2019, p. 157.

⑬ Mark Fitzpatrick, “Artificial Intelligence and Nuclear Command and Control,” *Survival,* Vol.61, No.3 ,2019,pp.81-92.

国际规范的当今，大国和这些中等国家的均势变动将给国际安全体系带来怎样的冲击令人担忧。加之人工智能技术本身的脆弱性，更加大了突发事件发生的可能性。

二、探索维护国际安全治理

对技术发展抱有极强信心的观察家对人类事务却可能是极度悲观的。他们将这项技术视作“启蒙的终结”⑭、第三次世界大战的导火索、⑮人类的中介⑯等等。持相反态度的观察家则对人类避免灾难的能力更加乐观。目前来看，人工智能是一项不断发展的赋能技术，很难找到一个完美的定义。它带来的变革效应正在全球显现，完全禁止已然不可能，任由其肆意发展也与国际关切和人类共同命运不符。那么，如何对其进行国际治理呢？对政策制定者来说，所需要关注的并非只是技术在国际安全上引发的众多不确定性，还有他们自己对这项技术的认识不断丰富，进而改变了对当下以及未来的治理选择。摆在我们面前的并不是一个注定的悲剧，而是一个承担人类发展重任的抉择。

在联合国框架下，有多个机制都对限制自主武器发展进行了国际探讨。其中，联合国《特定常规武器公约》谈判机制自从 2014 年起已经召开了三次非正式专家会议和三次正式政府专家组会议。尽管在国际机制和国际规范的建立上取得了一定的突破，但是参与讨论的各方对于致命性自主武器的可行定义分歧严重，均认为其指涉对象十

⑭ Henry Kissinger, “How the Enlightenment Ends,” *The Atlantic*, June 2018, https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

⑮《埃隆·马斯克谈人工智能：人类可能在召唤恶魔》，2017 年 10 月 30 日，http://tech.ifeng.com/a/20171030/44736042_0.shtml。

⑯《霍金再抛人工智能威胁论：或招致人类灭亡》，2017 年 4 月 8 日，http://www.xinhuanet.com/tech/2017-04/28/c_1120889914.htm。

分模糊。现在来看，不论各方讨论的是怎样的定义，都在一定程度上与既有使用武器的规则存在矛盾。根据瑞典斯德哥尔摩和平研究所对 154 种武器系统的统计分析发现，只有 49 种武器可以在人类监管但不介入的情况下进行交战。它们主要用于对己方设施进行防御，比如保护军舰或基地、应对来袭导弹等等。[17] 致命性自主武器带来的战略红利牵动着各国的安全利益，给后续军控行动增加了困难和挑战。

尽管如此，这并不意味着我们无计可施或者注定陷入无尽的无谓争论之中。在前述 2014 年的第一次非正式专家会议中，各方分歧明显。有的国家代表认为，致命性自主武器易于扩散且极其危险，终将给人类带来巨大威胁，需要全面禁止开发和使用。与之相反的观点则认为，应该发展致命性自主武器，因为它未来将足够“聪明”，能够理解人类战争中的道德规范，甚至可为人类提供一种更加人道的战争选择。两种观点都预设了人工智能最终将足够强大，却给出了完全对立的应对方案。时至今日，他们预想还没出现，但看似不可调和的争论中日渐出现共识。参与讨论的各方同意，致命性自主武器的军控行为不应阻碍民用技术的开发和促进经济发展，即使定义不清也不妨碍其他问题取得进展。[18] 这表明，决策者的认知不是一成不变的。相关军控谈判不是一次性达成具有约束力的国际条约，而是制定出致命性自主武器的底线后采取多种军控手段结合的渐进方式建立“软法”。这需要国家采取自我限制，达成国家之间的非约束性协议，比如符合现行国际法和国际准则的“行为准则”。

在维护全球战略稳定上，国家并未进入你死我活的死胡同。人工智能技术可以作为维护战略稳定的工具。情报收集和分析的准确性提

---

⑰ Vincent Boulanin, and Maaike Verbruggen, “Mapping the development of autonomy in weapon systems,” *SIPRI Report*, Nov. 14th, 2018, p.26.

⑱《2017 年致命性自主武器系统问题政府专家组的报告》，《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》缔约方政府专家小组，2017 年 12 月 22 日。

高，也可能使威慑、保证和再保证更可信。在理想的情况下，如果获取更全面的情报和分析，则能加强对敌方的再保证。这个过程将促进形成良性循环，最终极大降低战争风险。在拥核国家互信不足或难以揣测对方意图的情况下，更高效的侦察能力可以给决策者提供更多可信的信息。[⑲] 即使人工智能可以定位一国的导弹发射井，国家也可以加大导弹发射井承受第一次打击的能力，从而保留第二次打击的战略性力量，形成战略威慑。换言之，在技术层面，人工智能引发的战略不稳定并非完全不可破解。此外，人工智能还可以在核裁军和去核化中起到对核设施的监督和核查的作用。

“大分裂”并不是中美两国唯一的选择。我们应乐观地相信，人工智能竞赛不是零和游戏。它应当是一项促进人类福祉和推动国家合作的工具。虽然特朗普政府的美国对外政策表现出领域“脱钩”的趋势，中美两国的科研合作反而呈现出更加紧密的情况。根据科睿唯安（Clarivate Analytics）提供的人工智能领域科技文献数据显示，从2013 年到 2017 年，中美两国国际合作论文数量增长最快，互为过去5 年开展国际合作最多的对象，合作论文 4000 多篇。[⑳] 通过对科学和工程领域的著作进行文献统计，有研究发现中美两国学者在 2014 年到 2018 年间共同发表的论文数量增长了 55.7%，并得出如果美国贸然切断同中国的科研合作关系会导致美国的发表成果数量大幅下降的结论。[㉑] 科学技术的发展和在世界范围的传播会让各国受益。一国的

---

⑲ 吉斯特·爱德华、安德鲁洛：《人工智能对核战争风险意味几何？》，兰德公司，2018 年；Vincent Boulanin,“The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk”, Volume I; “Euro-Atlantic Perspectives,” Stockholm International Peace Research Institute, 2019。

⑳ 科睿唯安信息服务：《人工智能领域科技文献中高产国家 / 地区的竞争力分析》，2018 年 12 月，第 13 页。转引自傅莹：《人工智能对国际关系的影响初析》，《国际政治科学》2019 年第 1 期，第 16 页。

㉑ Jenny Lee and John Haupt, “Winners and losers in US-China scientific research collaborations.” Higher Education, Nov. 7th, 2019.

技术突破也可能有利于另一国的社会发展。比如，位于北京的微软亚洲研究院曾经支持四名中国年轻学者发表了一篇关于深度残差学习的论文成为近年来该领域的重要引用文献。在论文发表后，四位作者中有一位前往美国脸书公司任职，其余三位则加入到中国国内的人工智能创业行列之中。[22] 他们的成功无疑让中美两国企业和整个人工智能研发界受益。

在国际安全事务中，不确定性是常态，确定性反而是稀缺的。随着人工智能技术的发展，对致命性自主武器的定义将不再只限制在技术层面而在于自主能力范围的问题上。例如，武器的智能化程度成为判定其性质和限制的标准。政策制定者需要担心的并不是那些不确定性导致的潜在权势流失，而是这种认知的固化带来的长远风险。他们不应该期待以“技术治理技术”的手段获得稳健安全的人工智能系统来解决面临的国际安全困境。安全的军事系统会在技术层面降低国家之间突发事件发生的可能性，但这意味着一国获得了只有偏利于它的利器。这将不可避免地加剧军备竞赛，甚至动摇战略稳定而产生灾难性后果。同时，国家也不应该为了赢得军备竞赛的胜利而争相部署不安全的人工智能系统。政策制定者既不能低估不同的战略文化对使用新技术的影响，也不应高估人工智能带来的安全冲击。目前的人工智能技术还不成熟，也没有足够的证据表明它可以替代现有的武器系统。政策制定者应该做的是承认不确定性决定人工智能的安全治理是一个动态的过程，任何基于在未来出现最好或者最坏结果的计划可能都很难实现。面对不断变动的国际环境，政策制定者需要在不断发展的技术和变化的人性中去寻找国际合作的空间。

---

[22] Matt Sheehan, “Who Benefits From American AI Research in China?” Oct. 21th, 2019, https://macropolo.org/china-ai-research-resnet/.

本期责编：周武华

清华大学战略与安全研究中心

办公地点：清华大学明斋 217

联系电话：010-62771388

电子邮箱：ciss_thu@163.com