意图孪生:人工智能军事化的国际安全挑战与治理路径

甘钧先
(浙江大学 马克思主义学院,浙江 杭州 310058)

摘要:人工智能对人类的威胁,并不是它在智能上的不断升级并最终进化出"类人"意识,而是它所承载的人类意图。神经网络技术的进步显示,人类正在将其意图复制到人工智能载体上。当前,人工智能军事化的浪潮承载着人类的消极意图,它将颠覆大国之间的军事实力平衡,更重要的是,它将导致国家之间的敌意孪生到人工智能系统中,进而给人类整体带来重大威胁。国际社会对人工智能军事化的管控,应该以人类共同安全为指导原则,以人类命运共同体为价值引领,构建基于人工智能的知识共同体,从科学层面判断人工智能意图孪生的技术拐点,并通过国际军控机制和国家之间军控协议缓解安全困境,多轨并行推进以精准管控人工智能军事化的安全挑战。

关键词:人工智能;意图孪生;自主武器;国际安全;全球治理;人类共同安全;人类命运共同体
[**中图分类号**]D5;E3/7 [**文献标识码**]A [**文章编号**]1672-934X(2024)01-0106-11
DOI:10.16573/j.cnki.1672-934x.2024.01.011

Twin Intention: International Security Challenges and Governance Paths for AI's Militarization

Gan Junxian
(School of Marxism, Zhejiang University, Hangzhou, Zhejiang 310058, China)

Abstract: AI's threat to human beings lies not in its continuous upgrading in intelligence and the eventual evolution of "human-like" consciousness, but in human's intentions it bears. Advances in neural network technology have showed that humans are replicating their intentions onto AI vehicles. The current wave of AI militarization embodies humanity's negative intentions, which will upend the military balance among great powers, and more importantly, will lead to the twinning of enmity among nations into AI systems, thus posing a serious threat to mankind as a whole. To response, the international community should take a common security as guiding principle, take a community with a shared future as value direction, build a knowledge community based on AI, judge AI's technological inflection point from scientific level, and alleviate the security dilemma through international arms control mechanism and inter-state agreements, in an effort to accurately manage and control the security challenges of AI militarization in parallel multi-tracks.

Keywords: artificial intelligence (AI); twin intention; autonomous weapons; international security; global governance; human's shared safety; human community with a shared future

收稿日期:2023-11-21
基金项目:国家社会科学基金重大项目(23ZDA127);浙江省哲社规划课题一般项目(22NDJC128YB)
作者简介:甘钧先(1978—),男,副教授,主要从事非传统安全问题研究。

一、人工智能的意图孪生现象及其安全内涵

意图是理解人类威胁的关键。根据意图的强弱，人类安全威胁可以分为：无意图威胁、弱意图威胁和强意图威胁。无意图威胁主要包括环境灾难，如地震、海啸、飓风等，它们具有破坏性，但本身没有表现出任何的意图。弱意图威胁主要包括特定动物和病菌带来的威胁，它们具有主动攻击人类的特性，但其意图是一种生物学设定，表现为生存本能。强意图威胁主要指人类自身的威胁，即人类个体或群体之间的相互威胁，比如不同民族、国家或文明之间的安全挑战。

人类彼此的安全威胁是一种强意图威胁。人类内部敌意不仅与民族或国家生存有关，也与权力、利益争夺和观念分歧相关。人类内部敌意主要表现为主动的敌意，并且不像某些生物攻击人类的威胁一样无法避免。人类可以心存善意，选择合作、协调和包容，但某些族群或国家为了更大的利益，选择排斥、打压其他群体，从而构成彼此的威胁。人类大多数安全困境，都是因为在交往过程中产生敌意而引发的。

技术与敌意的结合强化了人类安全困境。技术是中性的，其本身并不具有威胁属性，只有当它与人类意图结合在一起时，才会表现出安全内涵、表现出善恶。当人类使用技术作为攻击性武器时，技术的恶就会展现出来。当人类使用技术增进人类福祉时，技术的善就会表达出来。技术进步是人类文明的重要成果，它提高了人类面对自然环境威胁的能力，但也会引起人类内部对安全的忧虑，比如核能技术的运用。技术本身并没有任何意图，它只是人类意图的承载者。人工智能出现之前，人类所有技术都遵循这个逻辑。

人工智能是一项划时代的技术，它的出现改变了技术载体与意图的关系。根据历史经验，当人类停止使用技术载体作为武器时，技术载体就同时失去了意图。只有人类和技术载体结合时，技术载体才会显示出意图，但是人工智能改变了技术的中性特征。人工智能的目标是让机器拥有智能，也就是赋予机器以意图。当机器拥有类似于人类的意图之后，人类即使改变了自己的意图，或者终止了对机器的使用，人类意图依然可以继续存在于机器之中。人工智能可以将人类意图延续下去，相当于人类意图在机器上实现了孪生，它并不需要机器进化出“类人”生命意识就能够实现。本文将此现象视为人工智能的意图孪生。

按照智能化水平，人工智能可以分为三个阶段。第一阶段是初级人工智能，机器依然被人类操控，被动执行人类给出的任务。第二阶段是中级人工智能，机器被设定为脱离了人类控制之后依然可以自主执行任务的自动系统。中级人工智能具有一定程度上的环境适应性，而不再是基于人类的程序设定，基本上实现了意图孪生。第三阶段是高级人工智能，拥有类似于人类的生命意识，能够摆脱人类的设定自主开展行动。当前，人工智能的发展水平仍然处于第一阶段，它们是可以承载人类的意图，但无法自主思考、自主选择的机器，当人类终止意图或断开电源时，它们就不再执行人类意图。

人工智能在每一个发展阶段都具有不同等级的风险。第一阶段人工智能的风险在于，由于算法的不成熟或系统缺陷，它们可能错误地执行人类意图从而带来不可预期的灾难。第二阶段人工智能可以基于神经网络在环境互动中自主作出选择，人类意图通过智能机器得到持续执行。它的风险在于，人类意图复制或孪生到机器上，这意味着人类可能失去对机器的控制。第三阶段人工智能的风险在于，它最具威胁性，也最难实现对其的控制。它们拥有一定的自我意识，也就是拥有了专属机器的自我意图。

当前，神经网络技术的进展显示，人工智能也许很难获得生命意识，但很可能实现意图孪生。人类生命意识的基础是对世界的感受，而

感受来自神经元。人类感受是独特的、相对的,它以每个人的神经元系统为中介,集中表现为各类情感。人类情感(比如,饥饿、恐惧、喜悦、好奇等)是人类意图或生命活动的原始动力。但与人类情感不同,人类思维并不直接建立在感受性之上,它是理性和逻辑的表达,大脑的计算和推理在很大程度上并不依赖于情感。由于缺乏神经元对世界的感受,意识和情感几乎不可能被复制到机器系统上。但人类意图跟人类情感不同,它的实现更多地受制于人类理性计算。人类意图跟人类感受和理性的关系是:人类感受激发或驱动人类的意图,而人类理性策划实现人类意图的程序。因此,人类意图的本质是程序,实现人类意图的过程是算法,它可以被复制或孪生到机器上。让机器产生意识,如同让机器获得对世界的感受,目前看来几乎很难实现,但让机器产生并实现意图,如同让机器进行运算活动,这在技术层面上是能够实现的。人类依靠自身意识对机器进行原始的推动,赋予机器以人类意图,即程序设计,然后让机器自动执行程序。这个过程就是人类意图在机器上的孪生过程。

通过让机器完全自主行动,人类意图实现了延展。对第一阶段人工智能来说,它们不具有或具有很弱的环境适应性,无法对环境进行即时回应,必须依靠人类进行判断,因而不具有主动性。如,在军事博弈中,它们无法自我选择防守还是进攻。对第二阶段拥有孪生意图的人工智能来说,它们可以依据内置神经网络自动识别、追踪目标,自动发起攻击。一旦人工智能拥有孪生意图之后,就意味着它们在一定程度上脱离了人类控制。意图被赋予到机器上,即便人类放弃了意图,但机器仍然可以继续执行先前的意图。当人工智能发展到这个阶段之时,就达到了库兹韦尔和博斯特罗姆提出的“技术奇点”[1]。

任何一种实现了意图孪生的武器系统,当它们自主激活、自主进攻时,被进攻方都可能无法判断威胁来源,从而导致混乱的攻击和人类外交的失败[2]。从人工智能的技术发展来看,尽管第三阶段人工智能还不构成现实威胁,但第二阶段人工智能的威胁却近在咫尺,尤其在军事领域,无人武器的智能化程度正在得到不断提升。正如非政府组织“禁止杀手机器人运动”所宣称的,人类“正在逐步接近由机器决定杀死谁”。被誉为人工智能之父的杰弗里·辛顿认为,人类未来需要担心的不是机器的智能化提升,而是杀手机器人[3]。辛顿的话表明,威胁人类的不是机器不断提升的意识,而是它们的意图,尤其是人类孪生在它们身上的可实现的意图。

人类意图和孪生在机器上的意图存在着本质区别。人类意图,无论是善意的还是敌意的,都基于人类对社会和自我的理解,它们会在社会互动中不断改变。比如,在世界近现代史中,英法德之间的善意和敌意经过数次转换。但人工智能并不是社会存在物,它们没在社会环境之中,它们不理解善意与敌意的根源和意义,它们的意图是程序的设定。无论是人类意识还是意图,都指向外部实在的万物,人类意识是对万物的意识,人类意图也是对万物的意图。它们的根基是人类基于神经元之上的对世界万物的感受性。人类的价值观、交互理解、意义领悟、生命觉醒、共情等都建立在人类对万物的感受基础之上。因此,人类智能是一种生命感受型智能,而人工智能本质上是一种数据处理型智能。人工智能处理的是符号世界,即万物的表征,而非外部实在的万物。

人类意图之所以能够实现转变,是因为人类能够感受到对方的意图转变或对方的情感,如伤痛或恐惧等,也因为人类能够从生命觉醒中重新审视原有意图的价值。而孪生在机器上的意图则没有任何感受性,也就无法在万物变化的过程中重置意图。人类在经历了伤痛之后

会改变意图，比如，两次世界大战之后的欧洲改变了互为敌人的冲突状态。人类意图可以在对社会价值和生命意义的学习过程中得到动态调整，但人工智能的学习只是数据处理、算力增强、算法提升等，无法实现价值和意义的重塑，因而无法实现从敌意到善意的意图转换。人类的意图转换是人类获得和平的关键。尽管民用人工智能可以为人类社会带来解放性的生产力，但军用人工智能却天然地具有恶意，孪生在武器上的人类意图帮助人工智能武器维持了人类敌意。从这个角度来看，人工智能武器具有设计程序上的原罪。

人工智能的意图孪生现象正在发生。一方面，从技术转型来看，智能化是人类科技的未来趋势。人类技术经历了石器、青铜器、铁器和机器时代。人类科技进步的核心逻辑是：人类繁琐的生产活动不断地由工具辅助或交给工具来完成，提高生产效率的同时，把人类自身从生产活动中解放出来。迄今为止，人类的交通工具、通信工具和日常用具等大多实现了机械化。智能化是人类科技进步的趋势，即人类让机器直接从事人类不愿意从事的生产活动，尤其是危险性系数很高的活动。技术的演进逻辑决定了人类必然发展全自主机器来代替人工劳动。另一方面，从技术的迭代进程来看，只有赋予机器以人类意图，才能在更大程度上解放人类，因而全自主机器成为当前人工智能的主要努力方向。人类无穷尽的创新渴望，使得技术本身看起来仿佛是有生命的。人工智能最初看起来并非革命性的技术，但它具有进化性[4]，它能进化出让人惊叹的模样。当前，生成式人工智能已经可以自主生成图像和文本，可以进行人机对话，而“生产的深处指向创造”[5]，即未来人工智能可以根据环境自主作出判断和选择。2023年8月，马斯克开发的FSD Beta V12自动驾驶系统可以依据神经网络对交通环境独立作出创造性回应，摆脱了以前的程序设定和固定场景模拟。人工智能的孪生意图隐藏在机器系统表达的自主性之中。随着机器神经网络适应性和自主性的增强，人工智能的意图孪生将会变成现实。

军事活动是人类意图表现最为明确的领域，也是人类恶意最为集中的领域。人工智能军事化，尤其是致命性全自主武器的开发，恰恰是把最危险的人类意图孪生到机器上的人类实践，因而给人类带来的风险也最为紧迫。随着Chat-GPT的问世，越来越先进的自然语言处理技术、图像处理技术和人机互动模型，以及越来越强大的算力平台和算法优化，使得人工智能从人机交互不断向机器自主发展。当机器获得了一定程度的环境适应性，人工智能就获得了类似于人类的意图或意向性。尽管这种意图是人类构建的，但它可以在无人的情况下继续发挥主动性，自主激发攻击程序。一旦到达此临界点，即便人工智能永远无法获得生命意识，但马斯克所预言的“人工智能魔鬼”就会被唤醒。

二、当前主要大国人工智能军事化策略与意图孪生趋向

人类重大技术进步常常被用于增进军事实力。获取先进技术及技术军事化运用是人类竞争的重要动力和表现形态。如，军舰是轮船的军事化运用、战车是汽车的军事化运用、战斗机是飞机的军事化运用。同样，增进人类福利、把人类从繁重劳动中解放出来的人工智能也会经历相应的军事化过程。人工智能正在为全球主要大国的海陆空多个作战领域的武器赋能。这是否会带来重大风险，取决于国家之间的意图，以及人工智能武器的意图孪生程度。

近年来，全球主要大国都开始规划本国的人工智能发展战略，并迅速推进其军事化应用。美国是推动人工智能军事化最早也最激进的国家，同时也是在该领域拥有最强实力的国家。2015年，美国国防部实施“第三次抵消战略”，将人工智能与军事创新相结合，以此继续维持美国的全球霸权地位[6]。2016年，美国国防部

发布《为人工智能的未来做好准备》和《国家人工智能研发战略规划》,积极推动自主武器和半自主武器的研发。2017年,美国国防部提出“算法战”,并成立了“算法战跨职能小组”[7]。2018年,美国国防部发布《无人系统综合路线图2017—2042》,聚焦人机协作、互操作性及自主性,大力提升无人作战平台能力[8]。2019年2月,美国国防部发布《2018国防部人工智能战略概要:利用人工智能促进安全与繁荣》,提出通过产学研结合、与国际伙伴共同技术研发等方式提高美国人工智能军事化水平[9]。2021年6月,美国国防部副部长希克斯宣布推动人工智能和数据加速计划,推动实现以人工智能为基础的作战模式“联合全域指挥和控制”[10]。2018年以来,美国海军大力推动“海上幽灵舰队霸主”计划,多次在太平洋军事演习中使用无人舰艇。2023年9月,美国两艘“幽灵舰队”的无人舰艇“水手”号、“游骑兵”号出现在日本的横须贺基地。美国使用无人武器系统的目标是贯彻美军的分布式作战理念,试图在海上以最小代价损害其他国家的海洋权益。

俄罗斯也是推动人工智能军事化的主要大国。针对美国的“第三次抵消战略”,俄罗斯国防部制定了《2025年先进军用机器人技术装备研发专项综合计划》,提出了人工智能与军事运用的路线图[11]。2016年,俄罗斯政府发布《2025年前发展军事科学综合体构想》,将无人自主技术作为俄罗斯军事创新的重点。2019年,俄罗斯政府制定《俄罗斯2030年前国家人工智能发展战略》,将人工智能武器装备作为优先发展事项之一[12]。俄罗斯的军用人工智能战略非常重视战斗机器人。普京曾指出,“战斗机器人系统可以从根本上改变俄罗斯军队的运作方式。”[13]当前,俄罗斯正大力研制“阿尔戈”“铀—9”等陆战机器人部队和纳米机器人系统,以及“维塔兹”无人舰艇和“波塞冬”无人潜艇等[14]。俄乌冲突期间,俄罗斯更为重视无人机的使用,“牵牛星”“猎户座”等无人机系统在战场上得到了大规模的实践和运用。

英法日等大国也在积极推动人工智能的军事化运用。2019年1月,英国国防部发布《国防领域信息技术变革计划》,开始推动人工智能在军事领域的运用。2022年6月,英国国防部发布《国防人工智能战略》,为推动该战略的实施,英国国防科学技术实验室与人工智能研究院艾伦·图灵研究所等机构联合成立国防人工智能研究中心。2018年3月,法国国防部成立创新防务实验室,发布《战斗机人机编组计划》,开始利用人工智能集成更新武器系统,并在其人工智能规划中增加1亿欧元用于未来武器的创新研发[15]。2016年,日本发布《防卫技术战略》,明确提出将人工智能技术与军事创新结合起来,表达了其借助智能科技实现军事力量跨越的野心[16]。

当前,各国人工智能军事化的主要目标是发展各类无人系统,如无人机、无人舰船、无人潜艇、无人航母等打击平台。一些代表性的无人作战系统包括美国的RQ-4“全球鹰”战略无人机和无人舰艇“海上猎手”、俄罗斯的“阿尔戈”陆战机器人和“波塞冬”无人潜艇等。主要大国竞相推动的人工智能军事化战略,正在构成一场人工智能的军备竞赛。美国国防部海军部副部长罗伯特·沃克指出,一旦美国的竞争对手将自主性赋予机器系统,那么美国也将不得不采用对等的发展策略[17]。但是,一旦美国越来越激进地发展人工智能武器,强化军事霸权,其他大国也不得不参与这场竞赛。人工智能军备竞赛使得意图孪生加速实现,进而给人类带来真正的重大风险。

第一,人工智能军事化将打破当前的军事平衡,不仅激发当前大国之间的人工智能军备竞赛,而且刺激小国将人工智能与大规模杀伤性武器研发相结合,以期实现非对称性制衡。这两种情况都将恶化当前的全球安全局势。人工智能建立在数学、物理学、计算机科学等基础学科和强大的信息技术产业与芯片产业

基础之上，发展中国家的经济和教育基础都相对薄弱，短期内无法在科研领域取得较大进展。当前，积极推动人工智能军事化的大国在军事力量上已经远远超过那些无法获得顶尖人工智能技术的国家。人工智能军事化将拉大发达国家和发展中国家之间的经济差距和军事差距，使得发展中国家在国际秩序中处于更加不利的地位。霸权国家对发展中国家的打压将有增无减，进而刺激发展中国家追求获取非对称性的军事能力。对那些感到不安全的小国家来说，它们不仅存在着开发核武器的安全需求，也存在着开发人工智能武器系统的需求，还存在着将人工智能与核武器、生化武器等大规模杀伤性武器结合的需求。毫无疑问，无论何种武器领域的军备竞赛，都将弱化国家之间的战略信任，强化国际社会的敌意，从而给人类造成无法预期的后果。

第二，对于实现意图孪生的人工智能武器来说，它可能脱离人类的控制。当前，人工智能武器大体上属于半自主武器，仍然需要人类控制，但发展全自主武器是军用人工智能开发的必然逻辑。全自主武器就是无需人类控制的自主激活程序、自动寻找和跟踪目标、自动发动攻击的武器系统。人类对武器的控制主要体现在程序设计、程序激活、机器能源供应、自动攻击等领域，其中，自动寻找和跟踪目标在一定程度上已经可以实现。全自主武器的关键是让机器自动获取驱动能源，如通过太阳能自动激活机器，同时唤醒机器内置的识别、跟踪和攻击程序。一旦机器实现了全自主行动，也就变成了意图孪生的武器，就能够脱离人类的控制。能够自动唤醒攻击程序的全自主武器承载着人类的恶意，它将脱离人类控制，变成人类的梦魇。此外，由于地理上的隐蔽性和不可及性，当前人工智能军事化的热点区域在海洋，未来毫无疑问将深入到太空。深空和深海领域的军用人工智能系统可能会给人类带来重大风险。深空和深海本来就是人类难以控制的地理区域，但其恰恰又是人工智能实验的理想场所，从而可能变成最容易失控的区域。当前，美国研制的无人武器系统主要聚焦于海洋，如无人航母、无人潜艇等。那些可以自动激活攻击程序的人工智能武器隐藏在深海、深空、虚拟空间和微型空间中，很容易脱离人类的控制。这些武器一旦脱离人类控制，人类如何搜寻这类具有隐蔽性和攻击性的武器，以及如何再度控制或销毁它们，将变成国际社会的巨大挑战。

第三，实现意图孪生的人工智能一旦被恐怖主义和极端势力所掌握，他们将付出更小的成本发动更大规模的恐怖袭击。恐怖分子如果获得了无人机、自主武器或杀手机器人等人工智能武器系统，他们就会利用这些武器进行大规模的自主杀戮[18]。如，2023年10月5日，叙利亚霍姆斯军事学院在举行毕业典礼时，遭到无人机的恐怖袭击，造成重大伤亡。至今，没有任何组织宣布为此负责。恐怖分子发动恐怖袭击的意图是在民众中制造恐慌，同时又试图逃避追踪。人工智能武器系统是恐怖分子实现邪恶意图的“理想”工具。恐怖主义袭击的人工智能化，避免了袭击者与袭击目标的近距离接触，增加了其隐蔽性和灵活性，也将使得国际社会追踪及打击恐怖主义变得更为艰难[19]。此外，恐怖势力和极端势力的威胁不只在于使用自主武器发动袭击，还在于它们将人工智能与致命性传染病毒、生化武器等系统相结合，制造出更大的恐怖效应。

三、人工智能意图孪生的全球治理路径

传统研究较为关注人工智能是否可以进化出“类人”意识，担忧其取代人类，但人类安全的核心挑战是人工智能的意图选择，而不是意识进化。即使人工智能将来进化出了超级意识，如果它选择向善的意图，那么人工智能不是人类的威胁，而是人类的最佳伙伴。真正威胁人类的是人工智能意图孪生带来的不确定性。人类面临的真正难题是被延伸到机器上的人类恶

意能否被发现和阻止以及如何阻止。恶意是人类绝大多数威胁的根源,因此,全球人工智能治理需要聚焦人类的意图管理和军用人工智能的意图孪生。

对军用人工智能来说,它需要在两个平行层级进行意图管理。一是国家之间的敌对意图管理;二是意图孪生之后的人工智能管理。这是人工智能治理与其他领域全球治理存在的差异。针对国家层面的敌意来说,它的治理思维类似于核武器治理,需要达成国际协议和国家之间协议,避免部署人工智能武器或全面削减人工智能武器,控制人工智能军事武器的规模化使用等。针对意图孪生的人工智能武器来说,它根源于科学问题,类似于全球气候治理,需要更广泛的科学发现作为证据支撑。气候治理和人工智能治理,都要求科学介入政治。但人工智能治理的特殊性,意味着它与核军控和气候治理的路径都将有所不同。

全球治理的前提是安全化,即该问题变成一个受到国际社会关注的安全议题。一个没有得到合理安全化的领域,就不会生成有效的治理思路,也不会开展具体的行动。特定问题的安全化,必须经过安全话语传播和安全机制建立的阶段。安全话语的传播可以让民众认识到风险,安全机制的建立则可以启动安全治理的实践。当前,全球人工智能治理在很大程度上围绕智能技术带来的社会影响而展开,多聚焦在算法伦理、数据使用和污染、信息安全、劳工规范等领域,很少触及人工智能军事化的管控。在军用人工智能领域,安全话语供给明显不足,民众对人工智能军事化和意图孪生的威胁仍限于科幻层面,缺乏专业话语。同时,国际层面的安全机制也没有发挥应有的作用,国家之间的安全机制仍然没有得到构建,针对人工智能军事化的安全规范仍在讨论之中。基于以上存在的问题,当前推动人工智能军事化的治理需要遵循从话语到机制、从科学到政治、从联合国到主权国家的安全化路径。

(一)全球科学共同体合作路径

全球科学或知识共同体的合作,可以有效供给人工智能意图孪生的安全话语。与气候变暖一样,人工智能军事化的意图孪生也处在一个模糊地带,只有专业话语才能够将其科学地安全化。专业话语比政治话语或常人话语更具有权威性和说服力,能够将议题迅速安全化。例如,全球变暖需要气候专家发表专业的调查报告,解释气候变暖的形势以及可能的危害。人工智能军事化遵循同样的逻辑。人工智能发展到意图孪生的何种程度及其风险,需要强大的专业解释和预测。它属于科学范畴,必须开展相关的科学辩论,需要全球顶尖科学家组成的知识共同体作出判断。科学家的集体认知是人工智能军事化治理的话语基础。

鉴于安全话语的供给不足,国际社会应该推动在联合国层面建立监管人工智能技术前景的科学机构,其主要成员是全球人工智能领域的科学家。它的功能和职责是发表人工智能技术应用的风险评估报告,尤其是人工智能意图孪生的进展情况和前景预测。人工智能技术拐点的态势评估对其治理来说非常关键。一旦出现意图孪生的情况,就意味着人工智能即将出现失控态势,也意味着需要更紧密的全球合作来应对人工智能危机。此外,权威的态势评估也是人工智能分级管控的前提。自主武器系统的分级管控有利于精准控制军用人工智能,鼓励民用人工智能的发展,但人工智能的风险分级只有基于知识共同体的科学评估才具有合理性。

全球科学合作路径的必要性还体现在科学家和技术人员可以及时发现人工智能意图孪生的进展情况及其风险,及时采取行动。人工智能的开发者基本上都是科学家和技术人员,他们的专业知识和职业道德在极大程度上影响着人工智能军事化的治理进程。他们对技术风险的敏感性和行动力使其处于人工智能治理的最前沿。2015年10月,人类未来研究所发布了一

封《迈向强大而有益的人工智能研究》的公开信，呼吁将人工智能研究用于增进人类福祉。此公开信获得了霍金、马斯克、黄仁勋等人的支持，至今已有11 251份签名，其中大部分签名者为人工智能领域的教授、专家或从业者。2017年9月，马斯克联合116名科学家致信联合国，呼吁禁止使用致命性自主武器[20]。2018年3月，1 179名从事人工智能科学的教授和专家，联合3 100多名谷歌公司员工发表公开信，反对谷歌公司继续为美军国防部Maven项目提供服务，同时承诺不开发、不生产自主武器，积极支持国际社会在禁止自主武器方面所作的努力[21]。自主武器本质上就是实现了意图孪生的武器。人工智能专家将目标瞄准自主武器，正是为了预防意图孪生型武器的出现。

非政府组织等民间知识团体也是推动人工智能全球治理的重要组成部分，它们的积极活动也能够增加和传播安全话语。它们的职能是唤醒民众的安全意识，进而呼吁政府控制意图孪生型武器的开发。2012年10月，在纽约成立的非政府组织“禁止杀手机器人运动”，其宗旨是“更少的自主，更多的人性”。该组织呼吁制定新的国际法禁止或约束自主武器的开发和使用。该组织目前在全球拥有180多个分支机构，其目标是宣传研发和部署自主武器的危害性。他们的行动正在全球层面唤醒民众的安全意识和对自主武器的警惕[22]。

（二）联合国军控路径

《特定常规武器公约》缔约国会议是当前联合国层面针对军用人工智能的主要治理机制。2016年，缔约国会议设立了专家委员会，迄今为止，举行了6次专家咨询会议。缔约国会议讨论的焦点是自主武器，也就是实现了意图孪生的武器。根据历次会议的讨论，各成员国达成的基本共识是：从自主武器的开发到使用，必须保证人类对自主武器的全过程控制和监管[23]。联合国人权理事会在自主武器控制中起到了推动作用。理事会特别报告员迈纳·凯提交报告，建议全面禁止无需人类控制的自主武器系统[24]。上述事实表明，联合国在阻止自主武器上达成了越来越多的共识。但是截至目前，由于美国等国家对全面禁止人工智能武器持犹豫或反对态度，国际社会暂时无法就军用人工智能的研发和使用达成协议，也没有制定相应的国际法来约束军用人工智能的开发。

鉴于当前人工智能军控机制的治理效能不足，而《特定常规武器公约》缔约国会议、联合国人权理事会和红十字会等机构都不是专门的人工智能军控机构，因此，国际社会需要推动在联合国建立专门管控军用人工智能的机构，协商制定监管人工智能军事化的纲领性文件。这种机制类似于核武器裁减机制。它必须区分人工智能民用和军用的界线，鼓励推动民用人工智能，但约束其军事化使用。当前，各国正在积极探索人工智能的军事化运用，要想阻止这种趋势非常艰难，但有必要让各国的人工智能军事化应用处在相对透明公开的环境中，尽量将人工智能竞赛限制在民用领域，尽最大努力阻止自主武器的开发与部署。2023年10月18日，中国政府发布《全球人工智能治理倡议》指出，人工智能治理攸关全人类命运，国际社会必须建立专门的国际机构来监管人工智能的军事化使用[25]。

除了建立和完善针对军用人工智能的治理机制，联合国还需要更深层次地考虑人工智能军备竞赛的意图，防止军事实力失衡引发的冲突和危机。一方面，人工智能军备竞赛的根源与核武器开发一样，都源于国家的安全防备。在不安全的国际社会中，国家对待彼此的意图也会更多地趋于敌意，进而增强自主武器的研发动力。因此，人工智能的军控机制需要综合考虑地区安全环境，只有国家处于相对安全的环境中，才会弱化它们开发自主武器的动力。对小国家来说，自主武器跟核武器一样，都属于非对称型武器。如果处于不安全的状态中，它们就会选择开发自主武器以获得军事力量的平

衡。基于此,联合国军控的思路是通过弱化国家之间的敌对意图来阻止自主武器的开发和部署。联合国军控机制既要控制自主武器的研发,也要控制核武器、导弹防御系统、太空武器等高精尖武器系统的部署,以防止地区军事失衡带来的人工智能武器开发浪潮。另一方面,联合国军控机构需要与科学知识团体紧密合作,严格评估自主武器的智能化程度,控制意图孪生现象在武器系统上的实现。据此,联合国军控机构需要严密监控人工智能武器的资金链、产业链、供应链和技术专家信息库,防范自主武器被恐怖分子控制,同时建立对等的、针对所有国家的可持续核查机制,以及类似于气候治理中的国家自主评估报告和行动机制,以便联合国在最大程度上掌握自主武器的发展动向。

(三)国家之间的协调路径

大国协商是国际治理的重要组成部分,尤其是在科学技术领域。顶尖科学技术一般掌握在大国手中,它们是科技进步的主要推动者,也是科技危险的发现者。在核武器管控领域,美俄核军控协议是国际核武器治理的重要支柱,人工智能的军事管控亦是如此。当前,全球监管机制尚未建立,大国之间基于彼此意图的相互沟通、自我监督和相互约束极其重要。中美两国协商达成人工智能协议,可以成为全球人工智能治理的基础,为其他大国加入治理机制发挥示范作用。正如前外交部副部长傅莹所言,中美两国是在人工智能技术研究和应用领域发展最快的国家,两国需要加强协调与合作,尤其是讨论通过国际法和规范开展对人工智能武器的监管,鼓励采取克制态度以限制对人工智能数据的军事化滥用[26]。

2019年2月,中国科技部成立了国家新一代人工智能治理专业委员会,专门负责人工智能的管理。2021年12月,中国向《特定常规武器公约》审议大会提交了《中国关于规范人工智能军事应用的立场文件》。该文件旗帜鲜明地指出,人工智能的军事运用应该遵循以人为本、"智能向善"的伦理原则;呼吁各国不以谋求军事优势为由开发人工智能技术,确保人工智能武器系统永远处在人类控制之下;实施分级、分类管理,避免使用可能产生严重消极后果的不成熟技术[27]。2018年,美国成立"人工智能特别委员会""人工智能国家安全委员会",负责监管人工智能的开发与使用。2019年,美国国防创新委员会发布的《人工智能原则:国防部人工智能应用伦理的若干建议》指出,人工智能应用必须遵守"负责、公平、可追踪、可靠、可控"五条原则,同时必须严格测试和验证人工智能和核武器等高危武器的结合[28]。2018年,欧盟成立人工智能监管机构"人工智能高级别专家组",推动出台《人工智能法》[29],负责欧洲范围内人工智能治理政策的制定和实施。2019年4月,欧盟发布《可信赖的人工智能道德准则》,讨论全球人工智能军事化运用的可能风险[30]。目前,在主要西方大国中,美英等国家对克制使用自主武器仍显犹豫,但法德两国倾向于严厉禁止自主武器系统的开发[31]。

以上事实表明,全球主要大国在一定程度上愿意管控人工智能的军事化使用。人工智能向善的本质是人类意图向善,各国对此具有基本的共识,这意味着推动制定军用人工智能监管协议具有现实基础。中美等主要大国需要讨论人工智能军备竞赛的危害性,推动签署人工智能武器控制协议,尤其应该深入探讨深海、深空的自主武器部署,不仅要控制自主武器的研发进度,及时评估自主武器的危险程度,还应及时预防人工智能与大规模杀伤性武器的结合。基于大国掌握军用人工智能核心技术的事实,大国之间应该建立严密的防扩散机制,防止人工智能技术的全球扩散,阻止自主武器落入恐怖势力和极端势力手中。

当前,军用人工智能监管还面临霸权政治和冷战思维的强大阻碍。美国为了维持自身霸权地位,对人工智能的军事化持有极高热情。

美国前防长埃斯珀明确指出，人工智能将改变未来战争形态，美国必须赢得与中国和俄罗斯在人工智能领域的竞争[32]。正是这种维持全球霸权地位的错误意图，使得美国很难真心诚意地跟中国协商削减人工智能武器的部署。基于这样的安全困境，中国需要与其他发展中国家一起，以全球安全倡议为价值引领，遵循安全不可分割的原则，既通过中美军事对话来缓解彼此的意图和信任困境，也通过全球安全对话来抵制美国的霸权意图，从人类整体的角度出发，缓解地区安全困境和中美安全困境，推动构建中美人工智能军事化的监管框架，并逐步接纳其他大国共同商讨制定针对人工智能军事化的全球安全协议。

四、结语

当前，人工智能技术不断更新迭代，全球主要大国围绕人工智能的军事化趋势愈加明显，人工智能意图孪生变得更加可能，给人类未来带来巨大的安全风险。据此，国际社会既要针对人工智能导致的军事实力失衡展开谈判和控制，也需要在科学层面上紧密关注人工智能意图孪生的临界点。人工智能军事化的治理必须以人类共同安全为基本准则，以人类命运共同体为价值依归，坚持安全话语与安全机制结合、科学评估与政治合作结合、联合国监管与国家自律结合的多轨并行路径，以核军控和气候治理的历史经验来指导大国之间的谈判与协商，引领国际社会对人工智能军事化的协同治理。

[参考文献]

[1] [美]雷·库兹韦尔.奇点临近[M].李庆诚，董振华，田源，译.北京：机械工业出版社，2011：3-4.

[2] Dvorsky G. Henry kissinger warns that AI will fundamentally alter human consciousness[EB/OL]. Gizmodo, https://gizmodo. com/henry-kissinger-warns-that-ai-will-fundamentallyalter-1839642809.

[3] Joe Shute. The "Godfather of AI" on making machines clever and whether robots really will learn to kill us all [EB/OL]. https://www.telegraph.co.uk/technology/2017/08/26/godfather-ai-making-machines-clever-whether-robots.

[4] Johnson J. Artificial intelligence & future warfare: implications for international security[J]. Defense & Security Analysis,2019,35(02):147-169.

[5] 潘云鹤.人和人工智能(AI)共同进化[J].研究与发展管理，2023(4)：1-4.

[6] Fiott D.Europe and the pentagon′s third offset strategy[J]. The RUSI Journal,2016,161(01):26-31.

[7] Freedberg S."Algorithmic Warfare": DSD work unleashes AI on intel data[EB/OL]. Utah Standard News, https://www.utahstandardnews.com/algorithmic-warfare-dsd-work-unleashes-ai-intel-data/.

[8] Unmanned systems integrated roadmap 2017-2042[EB/OL].U.S. Department of Defense,https://www.defensedaily.com/wp-content/uploads/post_attachment/206477.pdf.

[9] Summary of the 2018 department of defense artificial intelligence strategy:harnessing AI to advance our security and prosperity[EB/OL]. https://media. defense. gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

[10] Cronk T M. Hicks announces new artificial intelligence initiative[EB/OL].U.S. Department of Defense,https://www.defense. gov/Explore/News/Article/Article/2667212/hicks-announcesnew-artificial-intelligence-initiative/.

[11] 武坤琳，葛悦涛.俄罗斯《2030年前国家人工智能发展战略》浅析[J].无人系统技术，2020(2)：63-66.

[12] Samuel Bendett. Red robots rising: behind the rapid development of Russian unmanned military systems[EB/OL]. https://thestrategybridge. org/the-bridge/2017/12/12/red-robots-rising-behind-the-rapid-development-of-russian-unmanned-military-systems.

[13] Bendett S.The rise of Russia′s Hi-Tech military[EB/OL]. American Foreign Policy Council, https://www. afpc. org/publications/articles/the-rise-of-russias-hi-tech-military#_ftnref1.

[14] Zysk K. Defense innovation and the 4th industrial revolution in Russia[J]. Journal of Strategic Studies, 2021, 44(04):543-571.

[15] 雷鸿竹，曾志敏，熊帅.人工智能武器的全球发展、治理风险及对中国的启示[J].电子政务，2019(11)：112-120.

[16] 邓美薇.日本人工智能的战略演进和发展愿景及其启示[J].日本问题研究，2022(2)：11-21.

[17] Johnson J. Artificial intelligence, drone swarming and

escalation risks in future warfare[J]. The RUSI Journal, 2020,165(02):26-36.

[18] Terrorist groups, artificial intelligence, and killer drones [EB/OL]. Homeland Security News Wire, https://www.homelandsecuritynewswire.com/dr20190924-terrorist-groups-artificial-intelligence-and-killer-drones.

[19] Haner J, Garcia D. The artificial intelligence arms race: trends and world leaders in autonomous weapons development[J].Global Policy,2019,10(03):331-337.

[20] Browne R. Elon musk says global race for AI will be the most likely cause of world war III[EB/OL]. CNBC, https://www.cnbc.com/2017/09/04/elon-musk-says-global-race-for-ai-will-be-most-likely-cause-of-ww3.html.

[21] Open letter in support of Google employee and tech-workers[EB/OL]. International Committee for Robot Arms Control, https://www.icrac.net/open-letter-in-support-of-google-employees-and-tech-workers/.

[22] Garcia D. Lethal artificial intelligence and change: the future of international peace and security[J]. International Studies Review,2018,20(02):334-341.

[23] 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[J].国际安全研究,2019(5):108-132.

[24] 朱荣生,冯紫雯,陈琪,等.人工智能的国际安全挑战及其治理[J].中国科技论坛,2023(3):160-167.

[25] 全球人工智能治理倡议[EB/OL].中华人民共和国国家互联网信息办公室,中央网络安全和信息化委员会办公室, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

[26] 傅莹.人工智能与国际安全治理路径探讨[J].人民论坛,2020(36):6-7.

[27] 中国关于规范人工智能军事应用的立场文件[EB/OL].外交部,http://infogate.fmprc.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/jksxwlb_674635/202112/t20211214_10469511.shtml.

[28] Defense Innovation Board. AI principles: recommendations on the ethical use of artificial intelligence by the department of defense[R]. United States: Defense Innovation Center,2019:4.

[29] 王秋蓉,李艳芳.抢占未来制高点:世界主要国家人工智能发展与治理政策扫描[J].可持续发展经济导刊,2019(7):19-22.

[30] European Commission. The high-level expert group on AI presented ethics guidelines for trustworthy artificial intelligence[EB/OL]. Ethics guidelines for trustworthy AI, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[31] 董建敏.未来战争将打响夺智权[N].中国国防报,2019-02-13.

[32] Esper M T. Winning the future with artificial intelligence [EB/OL]. Modern War Institute, https://mwi.usma.edu/winning-the-future-with-artificial-intelligence/.