特色专题

进攻性人工智能的道德考量与全球治理

王英明,石超*

摘要 进攻性人工智能指人为地滥用人工智能(AI)来完成信息搜集并执行恶意进攻指令的智能体,其源于早期的棋盘博弈,现已被广泛应用于国家安全部署、情报搜集与意识形态等领域。然而,其道德问题不容忽视,具体讲,技术“黑匣子”致使伦理责任模糊,非法情报搜集威胁全球公众的隐私安全,复杂场景下机器判断具有不可控性,这都将国际局势推向新的政治不稳定境遇。有鉴于此,必须建立人类之于机器的“道德主权”,在全球范围内构建“复合网络”的伦理考量体系,并优化风险分级预案,从而为机器的“道德嵌入”、道德分级,以及“不可接受性风险”的伦理审查提供支持。

关键词 人工智能;伦理道德;风险治理;军事

科技创新成为国际战略博弈的主要战场,围绕科技制高点的竞争空前激烈[1]。进攻性AI(offensive artificial intelligence)主要指具备自主搜集、识别情报、决策并执行具有明显恶意攻击性任务的智能系统,可以被攻击者用来发动复杂且精准的网络攻击,目前已应用在目标侦察、精准打击、网络对抗、情报搜集及深度造假等多个场景,并呈现出向安全战、舆论战、信息战与认知战等领域全面扩散的趋势。然而,由于进攻性AI算法机制的不透明性、决策“黑匣子”等原因,其制定的解决方案可能会出现多场景应用的伦理道德失范,人类难以实现对AI决策过程的约束。因此,AI治理必须达成有效的国际合作与共识,明晰进攻性AI存在的场景、类型及特点,增强其在设计、决策与任务执行中的道德考量,构建其对科技伦理挑战的具象化图景,这对制定有效的AI伦理约束策略至关重要,或许可以为进攻性AI的伦理规范与道德框架优化提供理论基础。

中国石油大学(华东)马克思主义学院,青岛 266580

1 进攻性AI的演进历程与应用场景

进攻性AI的前身来源于计算机跳棋程序及其基础理论,并在实践中逐渐演变为具有进攻性的智能体。1956年,国际商业机器公司(IBM)的塞缪尔(Arthur Samuel)设计的西洋跳棋程序,可以通过观察棋子的走位构建计算机模型,并提供最佳的博弈策略。1994—2017年,AI机器人先后击败人类国际跳棋冠军、国际跳棋大师、国际象棋世界冠军、围棋世界冠军。AI在兵棋推演方面显示出强大的推算能力,其基于海量数据库做出的智能决策的速度和质量远超人类。凭借出色的表现,AI被逐渐应用于军事领域的兵棋推演与战略决策。

1.1 进攻性AI的定义及其理论流变

进攻性AI的概念及其理论流变源于早期的AI攻防技术与网络安全技术,这一概念没有达成共识前,常见于军事、安全、认知战与意识形态渗透等领域。2019年,由于网络安全与AI技术的深度关联,恶意利用AI进行网络进攻被称为进攻性网络武器[2]

收稿日期:2024-06-04;修回日期:2025-01-13
基金项目:教育部人文社会科学规划基金项目(24YJAZH094)
作者简介:王英明,硕士研究生,研究方向为技术伦理,电子信箱:yingmingwang@s.upc.edu.cn;石超(通信作者),副教授,研究方向为技术伦理、思想政治教育,电子信箱:superock@upc.edu.cn
引用格式:王英明, 石超. 进攻性人工智能的道德考量与全球治理[J]. 科技导报, 2025, 43(4): 37-45; doi: 10.3981/j.issn.1000-7857.2024.06.00622

(offensive cyber weapons),此可以视为进攻性AI的前身。詹姆斯·约翰逊(James Martin)认为由新一代AI增强的网络能力将促进核武器与战略非核武器的混合使用,并且极大加速战争进程并升级意外风险,呈现出进攻性色彩[3]。在此基础上,布莱辛·姆巴拉卡(Mbalaka Blessing)认为AI的进攻色彩并非仅仅体现在军事领域,在认知领域仍然会充满暴力与偏见[4]。伊斯洛尔·米尔斯基等于2022年正式提出"进攻性AI"这一概念,即"使用或滥用AI来完成恶意任务",在进攻工作量的基础上实现进攻精度、强度和广度全面提升与最大回报[5]。进攻性AI的理论体系随着在军事与安全领域的深入应用而逐渐完善,主要特征有2个:一是高度自主性的智能决策体,二是具有进攻能力的意识形态体。2024年,进攻性AI的使用主体逐渐由具备军事能力的集团转为数字资本主义操控的网络犯罪分子,也使新技术和新的商业模式(如黑客服务)被滥用于网络犯罪[6]。

1.2 进攻性AI与国家安全战略

AI技术的深度发展使得其逐渐在国家安全战略领域崭露头角。20世纪80年代,美国提出"空地一体战""非对称作战"等一系列作战概念,这些概念初步引入智能化元素,标志着AI技术被纳入军事战略考量之中。美国和其他一些国家通过发布战略规划文件,积极引导军事、安全领域与AI技术的深度融合,美国国务活动家兹比格涅夫·布热津斯基(Zbigniew Brzezinski)在《大棋局:美国的首要地位及其地缘战略》中提到,"一个集体安全体系,包括一体化的机会机构和部队"是美国成为"第一个也是唯一的全球性大国"的首要条件[7]。这也预示着进攻性AI的崛起。

进攻性AI已被应用于国家层面的战争发起与安全策略部署之中,用于执行各种带有战略意图的演习、军备及安全决策。早在2014年,美国时任国防部长哈格尔(Hagel)就曾提出[8]"第三次抵消战略"(third offset strategy),计划研发具有网络攻击和电子环境作战的网络赋能型自主武器。"我们必须发展和部署新的作战理论,包括多域作战和指挥控制,不仅要准备在空中、陆地和海上作战,还要在太空和网络空间作战。"2018年,《国防战略》进一步指出:"我们必须做出艰难的抉择……以打造一支致命、有韧性和快速适应能力强的联合部队"[9]。2019年,特朗普签署行政令《维护美国AI领导地位的行政命令》,这也是在政府层面上首次公开研究、推广、部署AI在军事领域的应用,标志着进攻性AI研发的组织化与"合法化"。2022年,美国《国家安全战略》报告指出,美国及其盟友、合作伙伴"正在投资一系列先进技术,包括在网络和空间领域的新技术应用、导弹挫败能力、可信赖的AI和量子系统,同时及时向战场部署新的军事力量……以保障我们的军事优势"[10]。加拿大军方认为,"国防和安全组织需要依靠科学技术来满足作战需求、预测和应对威胁,并满足现代战争中日益复杂的需求"[11]。在一定程度上,进攻性AI的研发、应用范围由美国向其周边盟友国家扩散,形成了以美国为核心的进攻性AI国家开发集团。

由宏观层面的进攻性AI国家战略部署转向已覆盖的微观实验环境与真实战场,则能进一步透视其威慑力量与本质。2018年,麻省理工学院研发的AI"诺曼"(Norman)在墨迹测验(Rorschach Inkblots)中发现,对进攻性AI持续性地灌输带有血腥、暴力倾向或属性的数据信息,就会强化机器的反理性、反人类的潜在威胁面。进攻性AI还被应用于战场,为大规模的突袭战提供有力的军备算力支撑(表1[12])。2024年6月,根据俄罗斯国防部报道,俄罗斯联邦武装部队各集团军的战术航空兵、无人驾驶飞行器和导弹部队,同时袭击了乌克兰无人驾驶攻击机群"马扎尔鸟"(Птахи Мадяра)、外国雇佣兵所在地,以及131个地区的阿扎瓦德军事装备集结地[13-14]。

表1 进攻性AI在现代化战场和军事武器中的应用

进攻性AI系统	现代军事武器/技术应用场景
目标驱动型系统	自主无人机(如MQ-9收割者)
自主系统	自动驾驶军用车辆(如KF51豹)
对话/人际互动	用于军事通信的聊天机器人(如美国陆军的Sgt. Star)
预测分析和决策	军事装备(如F-35)的预测性维护
超个性化	用于个性化士兵训练的GAN(生成对抗网络)
决策支持	军事行动中的人工智能辅助决策(如SAGE(半自动地面防空系统))
模式和异常识别	军事监视中的物体检测(如乌鸦无人机)

1.3 进攻性AI与意识形态渗透

相比于军事领域内进攻性AI的显性价值观特征,非战争形态的意识形态操控和情报战争则更为隐蔽。进攻性AI具备深度伪造信息的功能,正在全球信息环境中被非道德、非法地使用着,成为全球诈骗、假新闻泛滥、隐蔽的心理操控的得力工具。第一,进攻性AI加剧假新闻、假情报与虚假认知的传播和扩散。例如,假新闻、谣言、阴谋论和虚假信息正在Facebook、YouTube、Twitter、Instagram和WhatsApp等国外各大主流媒体上被广泛传播,而这一传播进程随着数字化生活场景的普及,较以往信息传播手段更为快速、轻松,且能深刻形塑受众的认知。Open AI公司开发的智能聊天机器人(如ChatGPT)能够生成类似人类的语言,并执行干扰公众认知、搜集民众情报的进攻性任务[15]。其技术的底层逻辑在于,“通过策略性的重复和强化,使其成为更容易感知、接受和记忆的主流解释”[16]。进攻性AI为何能够助长假新闻、假情报的快速散布,已成为心理学所关注的对象,“个人和集体在面对假新闻时具有脆弱性,他们容易接受看似有能力和可信的阴谋论”[17]。第二,全球网络安全系统面临着更为严峻的挑战。2021年3月,以色列国防部加沙分部成立代号为“南方向日葵”(Sunflower of the South)的情报团队,通过空地一体化的情报侦察系统从加沙地带搜集海量军事数据,实时传递给军事单位,并协同开发“福音”(Gospel)“智慧深度”(Depth of Wisdom)“炼金术士”(Alchemist)等进攻性AI,配合完成情报分析,使以色列对加沙地带房屋、建筑物、隧道、工业区、发电厂、港口,以及哈马斯在该地区的各类情况动态了如指掌,这在一定程度上反映了网络安全受到严重威胁的事实[18]。AI可为恶意程序提供动力,HYAS研究所的网络安全专家研发了一款名为“黑曼巴”(Black-mamba)的多模态恶意软件,它能够充当键盘记录器来窃听敏感数据,ChatGPT则为它们提供技术支持[19]。第三,进攻性AI为破坏网络安全和全球性诈骗活动提供便利。深度伪造技术涉及仿真性图片、语音和视频,“并且通过AI的改进,创建独特身份场景的图像”[20]。奥黛丽·雷蒙德(Audrey Raymond)在一项研究中指出,语音生成软件与其他欺诈方案(如商业电子邮件泄露)相结合,可为欺诈行为的实施提供便利,且具有极强的隐蔽性,在已经查明的2例案件中,预估已使公众分别损失24.3万美元和3500万美元[21]。不道德和非法使用进攻性AI,会大大增加深度伪造和金融犯罪的概率。此外,进攻性AI还可以通过分析、理解人类的情感(如恐惧、忧伤、失望等),生成针对性的文字、图片与视频,使全球公众陷入恐慌之中。

2 进攻性AI的道德伦理考量及风险肇因

从本质上讲,AI作为当前人类文明发展水平的先进生产力代表之一,本应致力于“缓解当今世界极为突出的财富不均和全球科技发展不平衡两大问题,并促进世界和平”[22]。也就是说,和平性、合法性、透明性等本应是其重要特征,尽管行为主义政治学一再强调价值“祛除”,但实践证明,进攻性AI难以做到以“和平”为真正的实然目的。完全定量化的进攻性AI可以为人类社会道德决策带来绝对客观的认识是不现实的,必须高度重视其潜在的伦理失范及其衍生的次生灾害。理由之一在于进攻性AI在构想之初的目的便是立于“人的反面”——针对和伤害人类。在战略博弈中,战争伦理也规定“有限的道德合法性”的先验立场,这构成进攻性AI伦理失范的根本性难题。二是由于信息情报采集数据的不完全真实性、算法设计人员的道德偏见、算法计算的发展限制、区域特色与差异等因素,进攻性AI可能带来算法偏见(algorithmic bias),而国际战略博弈往往事关全局,不像经济生产、人脸识别等容许一定的容错率,一旦战略决策出现偏差或误伤人类,其伦理风险可能导致的道德成本及政治代价都是极其高昂的。三是,定量化的进攻性AI难以满足政治的非中立、价值多元性,其不透明、难以道德审查等特征必然遭到质疑,当追求最大的利益冲动超越政治责任与社会责任,或将对全球公民权利、国际生态造成极大伤害。近年来,这一类伦理问题一直被悬置,“基于强AI技术的‘自主性’已成为现代武器系统的一个重要进化趋势,全面禁止致命性自主武器系统在现实中已不太可能,即便实施了禁令,也很可能不会产生效果”[23]。可见,仅仅将进攻性AI简单视为技术运算工具,可能会击破全人类的共同价值和道德底线。“技术上

最伟大的胜利与最大的灾难几乎并列”[24],进攻性AI的伦理道德问题亟须正视。

2.1 技术不透明性与“黑匣子”带来的伦理责任风险

进攻性AI的本质是“把物质粒子看作信息模式,把物理规律看作算法……即宇宙可以类比于一台可计算的量子计算机,于是一种信息论式的科学范式自然形成”[25]。进攻性AI能够最大限度地排除人为因素并建立其独特的数字范式,这种数据训练模式所形成的机器价值导向并非透明的,其决策过程也被置于“黑匣子”之中。进攻性AI在探究某一现象时,若试图涵盖所有相关细节,则难以取得实质性进展,算法设定者需要按照有意识地放大或缩小现象中的某些要素的方式,在舍弃那些具体、次要及偶发属性的基础上,最大限度地捕捉并展现事物的根本特性。“本质上就是关于时间与算力的复杂交互导致认识客体——机器本身无法成为一个完备的被认识状态”[26]。这样就意味着,进攻性AI在执行击杀任务的同时无法绝对秉承“价值中立”的原则,人类也无法确切考量进攻性AI对伦理、价值、生命、道德、权利等的追求与底线。

进一步来说,进攻性AI目前无法理解人类生命的价值,缺乏对自己或他人死亡事件之意义的感知能力。“机器人缺乏诸如痛苦、快乐一类的感知能力,这就意味着任何惩罚方式都无意义……任何对其不道德行为的事后指责都无意义”[27]。因此,进攻性AI“科学化”“客观化”“专业化”的幻觉也加剧了现代战争中的伦理失范。进攻性AI无需通讯员操作便可自动识别并集中打击目标群体,这种侦察范围已经实现超级扩散,士兵几乎不可能找到隐蔽之处。例如,乌克兰“尖帽”(Sharp Caps)无人机群反复轰炸有红外线感应的区域,直到该区域完全消失生命迹象,一名俄罗斯士兵在走投无路的情况下,只能试图通过向无人机投降获得生存的机会。尽管人类向机器人投降显得十分滑稽,但这的确是发生在战场上的真实情景[28]。

总之,进攻性AI的算法逻辑无法完全复现和遵循基于人类情感和经验的道德判断,其对任务执行的忠诚度显然要高于对人类生命意义与价值的感知,它并不能对其所执行的针对人类个体的伤害、虐待,甚至“斩首行动”的攻击性指令担负实际的责任。

2.2 情报无差别搜集的信息安全风险

进攻性AI非法从全球搜集军事情报与公众日常数据,其伦理失范问题可能引发全球道德恐慌。其一,进攻性AI的分析数据来源很可能依赖于国家间的不当监听、窃取信息的手段。早在2013年,英国《卫报》报道的“棱镜门”事件就曾指控美国白宫不仅监听本国民众的信息,还监听法国、德国等欧洲国家的政要与民众。近年来,为使算法深度搜集全球公众数据并应用于情报分析成为可能,美国认为,“对情报界获取、处理、评估和分析庞大的数据量,以及快速为国家安全机构提供有用情报的能力提出了挑战,美国情报界必须在日益不确定的国际环境中适应不断变化的任务需求”[29]。其二,现代战场的智能体为非法搜集情报提供了更多的便利。2023年,以色列国家网络局(INCD)、以色列国家安全总局“辛贝特”(Shin Bet)与以色列情报和特勤局“摩萨德”(Mossad)通过进攻性AI武器执行情报侦察任务,用于执行敏感数据窃取与关键信息基础设施破坏的任务,“果园行动”就是例子,在以色列空军空袭叙利亚核设施之前,以方8200部队向伊朗提供了信号情报[30]。不仅如此,随着现代通信技术的民用化,全球公众行为轨迹、行动路线和战略资源的分布情况均已暴露在智能体的分析框架之下,美国构建的“监听地图”呈现出新的数字霸权形态,肆意搜集公众数据用于各类国际战略竞争决策,并呈现出泄露隐私信息的可能性,为全球公众隐私带来未知风险。其三,非法的情报搜集与不明的政治意图杂糅,使AI伦理责任主体晦暗不明。进攻性AI的情报搜集“是否构成侵略、是否形成开战正义……要结合当时的国际背景、攻击的恶意程度,以及对国家主权和安全的威胁程度等因素来判断。”[31]同样,进攻性AI肆意搜集信息并与某种国家形式的开战权相结合,则会形成一种“绝对不对称”的战争技术凌驾于现实环境之上的局面,让国家层面的“道德腐败”有了可滋生的土壤。这意味着机器的道德判断演化为恐怖主义挑战人类文明的作恶形式,“它突破了所有的道德限制,以至于根本就不再有限制”[32]。由此推论,面临着已经发生的错误决策行为,并没有任何主体可以实际地、公平地承担灾难性后果的道德责任,这就进一步使算法决策陷入禁用难、审查难、问责难的尴尬境地。

2.3 对智能体进行人为控制的不确定性风险

多元化、复杂化的战争情境使进攻性AI的道德判断与区分原则背道而驰,人类预设的进攻程序缺乏满足复杂、海量的现代战争场景的道德应对能力,可能会诱导智能体做出不确定性控制下的极端行为,容易导致道德合法性危机。第一,进攻性AI的数据感知不确定性源于战场环境的不确定所引发的智能体行为的不确定。进攻性AI必须接受大量的战场图像、各种类型的传感器数据才能建立起对战场的感知系统,这种基于实时动态传播的战场感知,实际是以滞后变量作为解释变量的,各种小事都会改变AI模型的解释方式。美国陆军战争学院研究教授安东尼·普法夫(Anthony Pfaff)就对此表示过担忧,一旦使用进攻性AI控制核系统打击,AI机器的判断可能会超过人类的实际需要,这种不确定性控制无法避免[33]。第二,进攻性AI缺乏对攻击任务背后的人类及复杂社会关系的应然理解。“AI驱动的攻击可以引入新的威胁,扩展现有的攻击,改变威胁的典型特征,并且无法检测和追踪”[34]。例如,尽管以色列方面出动进攻性AI系统进行目标打击显示出明显效果,但是美国方面依然担心哈马斯可能会利用深度伪造来干扰其效能发挥。这包括对士兵图像、坦克造型、服装标志、士兵行为等识别的深度造假和干扰。一个战友因某种原因可能会被攻击系统误认为是敌人,这既增加了战场环境的复杂性,也为进攻性AI的战场行为带来更为严峻的道德挑战。

2.4 进攻性AI无节制发展的政治不确定性风险

进攻性AI的无节制发展可能会进一步加剧各国之间已然存在的认知偏差,进而加剧国际战略博弈的政治不确定性与力量失衡问题。近年来,美国多次以国防安全的名义采用进攻性AI技术开展其所谓的“定点清除”斩首行动,这进一步加剧了国际间实力不对称和权力相对性的威胁感知,“崛起国实力的增长引发守成国的恐惧……基于恐惧,守成国会采取包括预防性战争在内的‘破坏性’政策[35]”。约翰·米尔斯海默(John Mearsheimer)谈及认知误差对国际冲突所可能产生的影响时表示,各政治实体永远无法准确判定其他政治实体的真实意图,从而真正有效地采取有利于自身的欺骗性策略[36]。因此,基于非对称性的分析框架,不断增强进攻性AI的实力、采取富有攻击性的战略策略似乎成为唯一的选择,这将国际政治安全推向新的不稳定性范畴。进攻性AI引发的道德恐慌(moral panic)成为次生风险。道德恐慌是一种过度的社会反应现象,媒体、公众、利益集团和官方共同营造的道德恐慌对形塑越轨行为起到了推波助澜的作用。其一,动态竞争压力、不确定性意图、进攻性AI的动能与道德恐慌相互叠加。公众无法预测进攻性AI的真实意图,并且进攻性AI在决策方面的不透明性、去人格化,都会在一定程度上带来社会恐慌。当社会公众源于集体恐慌所产生的整体意志上升为国家意图之时,不确定性的加剧就会成为更加棘手的问题。“由于不确定性意图广泛存在,各国从维护本国安全利益出发,出于对潜在竞争对手军备行为的恐惧,推动AI军备竞赛持续升级”[37]。进攻性AI的不透明性还可能导致潜在对手的战略猜忌,做出更为强劲的、糟糕的假设,引发军备竞赛,造成国际范围内的恐慌。其二,道德恐慌所假定威胁的规模或严重程度往往是对真实情况的夸大。道德恐慌具有易变性和非对称性,前者意味着对进攻性AI的道德恐慌可能面临的政治不确定性将AI政治及其政治环境推向不明确或难以预测的状态,这种不确定性会加剧社会风险,这种易变性可能会随着AI的伦理约束而逐渐平稳;非对称性则使社会公众在道德判断与行为评价中对进攻性AI产生不公平、不一致的现象,情感状态可以显著影响人的道德判断,这会导致相同的主体在相似的情况下,因不同情感状态而做出不同的道德决策,进而形成道德对立。

3 价值对齐:建立全球机器伦理治理共同体

技术的进步总是伴随着新的伦理问题。进攻性AI最具伦理挑战性之处,莫过于道德推理能力的算法化。正是如此,才使进攻性AI成为“纯粹的工具”与“道德主体”之间的某种具有过渡性质的智能体,并由此使其具备某种人类“德性”之雏形。要真正使进攻性AI技术的伦理学后效满足现有的人类价值观体系,其主要举措并不在于野蛮地禁止此类技术的研发,而恰恰在于如何使此类技术趋于具备人类意义上的“道德推理能力”。

3.1 设立人对机器的“道德主权”原则

通过确立人类“道德主权”的价值共识来重塑和创新AI治理理念是推动全球AI治理体系完善并使国际秩序趋于稳定的逻辑起点。在传统的人机关系中,道德和伦理决策被视为人类的专属领域,但是,进攻性AI被赋予了越来越多的道德责任。什么是人类独有的道德感?机器是否能够拥有类似于同情心、正义感这样的道德情感?这些问题触及了人类的本质和未来,迫使人类重新思考人类之于机器的道德主权地位。首先,避免技术将人类“他者化”。技术“他者化”将人置于技术客体的治理方案。“他者”(the other)是具有悠久历史且影响深远的西方哲学概念,虚拟“客我”只是技术性在场而非真实肉身在场,虚拟“客我”甚至可以作为我的“他者”,将个体虚拟化[38]。技术“他者化”以数字模拟的方式区分人类的各种属性,人类的情感、价值与存亡均被“数字化”,其本质是将人与人置于对立的关系框架之中。因此,必须建立在战场环境中的新型人机关系,预防极端伦理问题的产生。其次,警惕虚幻的“技术共同体”。所谓的“技术共同体”是一种乌托邦式的幻想,进攻性AI设立的目的是在军事战争、战略博弈中取得技术性优势。技术治理同样存在国际竞争,很多专家预测,欧盟试图通过制定GPDR(《一般数据保护法》)一类的法案来建立“范式”,将AI法案的“布鲁塞尔效应”延伸到全球AI治理,从而使全球执行其AI治理标准[39]。虚假的“技术共同体”实际上是以一种隐蔽的方式继续着国际战略博弈,因此,必须及时发现并纠偏。最后,确保人类在战略博弈中的主体地位。马克思指出“占统治地位的是劳动力的结合(具有相同的劳动方式)和科学力量的应用,在这里,劳动的结合和所谓劳动的共同精神都转移到机器等上面去了”[40]。因此,若使人类这一有思想、有意识的生命体不至于沦为机器的附庸,使人的肉体和精神免于机器的统治和操纵,需要确立人类之于机器的“道德主权”,才能为设置一整套进攻性AI的治理框架提供价值原则。

3.2 凝聚全球范围的道德风险治理共识

国家层面应当积极推动AI风险治理由“线性”到“复合”、由“一元”向“多元”、由“人工”向“智能”方式的转变。近年来,各国在AI领域加快了探索和发展步伐。2020年,美国国防部(DoD)发布的《AI五大基本原则》,以指导美国军方在AI道德和合法使用AI系统方面处于领先地位[41]。2021年6月,美国政府问责局(GAO)发布了一份报告,明确了AI系统开发者和使用者的责任。2021年11月,美国国防创新部门发布了《负责任的AI指南》,为军用AI的构建和应用提供了明确方向。土耳其也于2021年8月24日发布了该国的首个《国家AI战略》,将技术发展与经济、文明实践经验相结合。英国则通过多个部门联合发布了《国家AI战略》,旨在将其打造成为全球AI安全领域的重要力量。2022年初,欧洲议会通过了《数字服务法》(《Digital Services Act》),并在上半年达成一致意见,规定了对大型互联网公司数据滥用的限制,并禁止在未获得用户同意的情况下定向推送广告。针对日益复杂的AI风险,2023年,美国国家标准与技术研究院(NIST)发布《AI风险管理框架》,用于更好地管理与AI相关的个人、组织和社会风险[42]。我国在网络、数据、算法、应用、算力等领域取得了显著进展。全国人大常委会已通过《网络安全法》《个人信息保护法》《电子签名法》《反不正当竞争法》等法律,并正在修订《电信法》和起草《网络犯罪防治法》等新法律。同时,国务院相关部门发布了《新一代AI伦理规范》《科技伦理审查办法(试行)》及《生成式AI服务管理暂行办法》等文件。

传统的AI风险治理是基于被动防御式的“链条”治理,其基本逻辑遵循“风险产生—认知—识别—分析预判—人工决策”的线性信息处理模型。实际上,各国对AI技术的全面介入已经为“复合网络”的综合道德考量模式提供了可行性基础。第一,构建纵深化、一体化、协同化的全球AI风险治理网络中心。尽管以美国为首的西方国家恪守“技术单边主义”,但是必须积极构建以国家利益为核心的“技术多边主义”来倒逼全球合作。推动AI技术发展搭建“去中心化”的治理平台,方能促使全球分散化、碎片化的信息技术要素的整合与流动。第二,构建风险治理主体在话语权上的“技术平衡”。针对不平等、不均衡的技术能力势差,进一步拓宽各类国家主体对技术风险治理标准的话语空间,构建以“风险均势”为导向的价值体系,从而打破大国对小国的“技术非对称”格局。第三,各国之间必须积极主动地寻求合作

与共治之道。“全球问题的解决,需要建立一种共同参与的治理方式,它应当是一种网络式的……各行为体在网络互动过程中,以妥协、信任、合作为基础,以平行互动来确认共同的目标、方式和途径,实施对公共事务的管理”[43]。

3.3 完善进攻性AI的风险评估体系

进攻性AI亟待建立统一的、完善的伦理风险评估体系与监管措施条例。关于AI的风险评级尚未在全球建立统一标准,2021年,欧盟委员会发布《AI法案》,将AI的风险分为最小风险、有限风险、高风险和不可接受风险。建立全球关于进攻性AI的风险评估体系需要满足3个条件。首先,应当建立完善的进攻性AI道德风险评估方案与参照细则。针对AI已经出现的风险预警,划分不同层次可视化、可量化数据标准的风险等级,形成对进攻性AI的全方位道德风险、伦理风险的评估,增强进攻性AI在全球范围内伦理风险透明度与公众的监管便捷性。其次,成立分级体系明确、责任划分清晰、实时动态评估的监管机构。“成立AI伦理委员会,进行伦理审查和推行负责任的创新理念”[44]。通过定期与不定期相结合的行业人员伦理培训,不仅能推动AI设计环节的各个周期性系统漏洞的修补与新系统开发,而且能将特定的道德、伦理风险概念灌输到开发者的头脑,从而对其所研发的进攻性AI负责。最后,确保进攻性AI投入战略博弈符合相应的风险等级。测试进攻性AI具体应用领域和场景中的表现,对其决策行为进行风险评估,合格后方能投入国家的战略博弈计划。

3.4 不可控场景的“道德嵌入”

AI的能力会随着时间的推移和使用方式的改变而改变,必须从设计之初就要考虑未来可能面临的所有风险类型与其道德规范框架,如此才不至于产生对进攻性AI发展走向的过度非理性恐惧。第一,增强“价值敏感”设计的伦理审查与可解释性。通过建立“机器对抗”的进攻性AI治理体系,设定能够自主基于时空条件处理进攻性AI风险数据的AI监管体,对多维数据、信息流进行基本的人类道德感知和伦理处置预判,协助治理主体对大量非结构化的数据进行清晰、聚合和序列化处理。由此形成进攻性AI数据分析、信息挖掘、情报检索等信息来源的道德性、合法性考量。第二,将进攻性AI的决策置于具备道德责任感的人类设计理念之中,进攻性AI不应受到非法的控制或歪曲,也不应受撒谎、操纵、胁迫或扭曲的影响。人类设计者必须考虑“其所有决策及其基本假设和道德框架,选择哪种伦理理论、包括哪些刺激、如何权衡所涉及的各个元素、如何进行计算以及通过哪种方式确定正确的行动得出结论”[45]。第三,应当置于宏大的全球叙事下考察进攻性AI伦理风险治理的意义。当进攻性AI作为一种工作模式,被置于社会、政治、组织动态之中时,应当审慎考虑其伦理风险,“将伦理思考视为其专业身份的一部分,并通过他们对这些缺失空间(技术、社会、法律或政治空间)的理解来看待AI的相关工作”[46]。

4 结论

综上,不同于AI在经济生产实践中的作用,进攻性AI应用于国际战略博弈或战场环境会构成更深层次的伦理风险挑战。需要注意的是,发展AI的目的本应是为人类福祉服务,而不是以所谓的“智能文明”和“机器文明”来取代“人类文明”。进攻性AI的运用直接关联到人的生命安全和尊严问题,其设计和运行必须具有极高的可靠性和安全性,任何技术缺陷或设计上的不完善都可能会对人类社会产生潜在的毁灭性打击。在深度考察进攻性AI伦理属性、伦理责任及伦理走向后,如何应对国际人道法和武装冲突法对于AI武器系统的适用性的争议问题,依然是国际社会当前所面临的共同且迫切的课题。进攻性AI的伦理风险治理,绝不仅限于建立限制进攻性AI发展的国际合作框架和规则体系,而是要更加迅速地推进进攻性AI的透明度改造和人类情感中具有伦理、道德、责任属性的因素嵌入,并确保可以追溯和审计,支持开展多领域、跨学科的对话以寻求规范之道。让进攻性AI技术充满人文(伦理)关怀,是在不确定风险下依然可以使技术保护全人类的价值基础,并在此基础上构建“智能威慑”以规范人类世界的技术运用方式。在此过程中,基于文明互鉴进行AI风险的全球治理并贡献中国智慧、中国方案、中国力量,才能确保这项强大的技术能力被谨慎地运用,服务于全人类的和平与安全。

参考文献(References)

[1] 习近平. 在中国科学院第二十次院士大会、中国工程院第十五次院士大会、中国科协第十次全国代表大会上的讲话[N]. 人民日报, 2021-05-29(002).

[2] Artificial intelligence and offensive cyber weapons[J]. Strategic Comments, 2019, 25(10): 10-12.

[3] James Martin Center for Nonproliferation Studies (CNS), Middlebury Institute of International Studies, Monterey. The AI-cyber nexus: Implications for military escalation, deterrence and strategic stability[J]. Journal of Cyber Policy, 2019(3): 442-460.

[4] Mbalaka B. Epistemically violent biases in artificial intelligence design: The case of DALLE-E 2 and Starry AI[J]. Digital Transformation and Society, 2023(4): 376-402.

[5] Mirsky Y, Demontis A, Kotak J, et al. The threat of offensive ai to organizations[J]. Computers & Security, 2023, doi: 10.1016/j.cose.2022.103006.

[6] Eider I, Oscar L V, Angel R, et al. Unleashing offensive artificial intelligence: Automated attack technique code generation[J]. Computers & Security, 2024(147): 104077.

[7] 布热津斯基. 大棋局: 美国的首要地位及其地缘战略[M]. 上海: 上海人民出版社, 2015: 25.

[8] Chuck H. Reagan national defense forum keynote[EB/OL]. [2024-02-20]. https://www.defense.gov/News/Speeches/Speech/Article/606635/.

[9] 2018 National Defense Strategy of the United States of America[A]. U.S. Department of Defense, 2018: 1.

[10] 2022 National Defense Strategy of the United States of America[A]. U.S. Department of Defense, 2022: 21.

[11] Wasilow S, Joelle B T. Artificial intelligence, robotics, ethics, and the military: A Canadian perspective[J]. AI Magazine, 2019(1): 37-48.

[12] Adib B R. Artificial intelligence in the military: An overview of the capabilities, applications, and challenges[J]. International Journal of Intelligent Systems, 2023(4): 1-31.

[13] Александр Полегенько. МО РФ заявило о поражении пунктов дислокации операторов ударных БПЛА ВСУ [EB/OL]. (2024-06-09) [2025-01-12]. https://tass.ru/armiya-i-opk/21052757.

[14] 俄罗斯卫星通讯社. 俄军对乌克兰保障武器生产的能源设施实施集群打击[EB/OL]. (2024-06-20)[2025-01-12]. https://sputniknews.cn/20240620/1059972414.html.

[15] Brady D L, Ting W. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? [J]. Library Hi Tech News, 2023(3): 26-29.

[16] Vivek K S, Isha G, Darshan S. Detecting fake news stories via multimodal analysis[J]. Journal of the Association for Information Science and Technology, 2020(1): 3-17.

[17] Bragarnich R. Observations on psychic vulnerability to media dissemination of false political - ideological messages (fake news)[J]. Journal of Analytical Psychology, 2022(2): 455-467.

[18] Yaakov K. Israel's Gaza War is like no other military operation in history[N]. Jerusalem Post, 2021-05-21.

[19] Ahmed D. Researcher create polymorphic Blackmamba malware with ChatGPT[EB/OL]. (2023-03-19) [2025-01-12]. https://www. hackread. com/chatgpt-blackmamba-malware-keylogger/.

[20] Cassandra C. Using artificial intelligence (AI) and deepfakes to deceive victims: The need to rethink current romance fraud prevention messaging[J]. Crime Prevention and Community Safety, 2022(1): 1-12.

[21] de Rancourt-Raymond A, Smaili N. The unethical use of deepfakes[J]. Journal of Financial Crime, 2023(4): 1066-1077.

[22] 艾伦·亨特, 刘舒羽. AI与和平学[J]. 史学月刊, 2016(7): 13-17.

[23] 董青岭. 新战争伦理:规范和约束致命性自主武器系统[J]. 国际观察, 2018(4): 51-66.

[24] 汉斯·昆. 世界伦理构想[M]. 周艺, 译. 北京: 生活·读书·新知三联书店, 2002: 16.

[25] 李建会, 符征, 张江. 计算主义:一种新的世界观[M]. 北京: 中国社会科学出版社, 2012: 238-244.

[26] 董春雨. 从机器认识的不透明性看AI的本质及其限度[J]. 中国社会科学, 2023(5): 148-166.

[27] 张亦佳. 强AI设备面临的伦理困境与风险[J]. 江苏社会科学, 2023(2): 58-67.

[28] Caitlin M. Battlefield demands spark AI race in Ukraine as war with Russia rages on[EB/OL]. (2024-03-31) [2025-01-12]. https://www. foxnews. com/world/battlefield-demands-spark-ai-race-ukraine-war-russia-rages.

[29] 刘胜湘, 李志豪. 美国军事战略的AI化趋势及其影响[J]. 国际展望, 2024, 16(3): 51-73.

[30] Jasper F. Israel's national cybersecurity and cyberdefense posture[R]. Zurich: CSS, 2020.

[31] 熊光清, 邓思敏. 网络战争的伦理约束问题:从正义的角度出发[J]. 学习与探索, 2020(7): 33-41.

[32] 迈克尔·沃尔泽. 正义与非正义战争[M]. 任辉献, 译. 北京: 社会科学文献出版社, 2015: 186.

[33] CBS News. The ethical challenges of AI on the battlefield [EB/OL]. (2024-05-09) [2025-01-12]. https://www. cbsnews. com/video/ethical-challenges-of-ai-on-battlefield/?intcid=CNM-00-10abd1h.

[34] Malatji M. Offensive artificial intelligence: Current state of the art and future directions[C]// 2023 International Conference on Digital Applications, Transformation & Economy (ICDATE). Sarawak, Malaysia: Miri, 2023: 1-6.
[35] 潘蓉, 肖河. 尚未触发的“修昔底德陷阱”与美国对华政策[J]. 国际论坛, 2020, 22 (2): 93-109.
[36] 约翰·米尔斯海默. 大国政治的悲剧[M]. 王义桅, 唐小松, 译. 上海: 上海人民出版, 2008.
[37] 张煌, 杜雁芸. AI军事化发展态势及其安全影响[J]. 外交评论(外交学院学报), 2022, 39(3): 99-130.
[38] 徐强, 单明娟. 我的“他者”与他者的“我”:数字主体的建构和确认[J]. 南京社会科学, 2022(7): 32-40.
[39] Marc R, Merve H. Artificial Intelligence and Democratic Values: Next Steps for the United States[EB/OL]. (2022-09-05) [2025-01-12]. https://www. cfr. org/blog/artificial-intelligence-and-democratic-values-next-steps-united-states.
[40] 马克思恩格斯全集(第30卷)[M]. 北京: 人民出版社, 1995: 588.
[41] DOD adopts ethical principles for artificial intelligence[EB/OL]. (2020-02-24) [2025-01-12]. https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.
[42] NIST AI risk management framework (AI RMF 1.0) launch [EB/OL]. (2023-01-16) [2025-01-12]. https://www. nist. gov/news-events/events/2023/01/nist-ai-risk-management-framework-ai-rmf-10-launch.
[43] 刘清才, 张农寿. 非政府组织在全球治理中的角色分析[J]. 国际问题研究, 2006(1): 48-52.
[44] 程新宇, 杨佳. AI时代人权的伦理风险及其治理路径[J]. 湖北大学学报(哲学社会科学版), 2024, 51(3): 159-166.
[45] Bellaby R W. Can AI weapons make ethical decisions?[J]. Criminal Justice Ethics, 2021(2): 86-107.
[46] Slota S C, Sherri R G, Nitin V. et al. Locating the work of artificial intelligence ethics[J]. Journal of the Association for Information Science and Technology, 2022(3): 311-322.

Ethical considerations and global governance over offensive AI

WANG Yingming, SHI Chao*

School of Marxism Studies, China University of Petroleum, Qingdao 266580, China

Abstract Offensive AI refers to intelligent agents that artificially abuse AI to complete information gathering and execute malicious offensive commands. Originated from board games, it has been widely used in many fields, such as national security deployment, intelligence gathering, and ideological infiltration, and the relevant ethical issues cannot be ignored. Specifically, the obscurity of ethical responsibilities due to the "black box" technology, the threat to the privacy and security of global public during illegal intelligence gathering, and the uncontrollable judgment of machines at complex scenarios will add more instability to the international politics. In view of the above situation, it is necessary to establish "moral sovereignty" of human beings over machines; meanwhile an ethical consideration system of "composite network" need to be built on a global scale, to optimize the risk classification plan and support the "moral embeddedness" of machines, moral classification and ethical review of "unacceptability risk".

Keywords AI; ethics; risk governance; military

（责任编辑　卫夏雯）