生成式人工智能对国际安全的影响：以认知战为路径的分析

贾子方　王　栋

内容提要　生成式人工智能可以生成类似人类创造的信息，使用成本低，生成信息效能高，应用领域广泛。研究这一颠覆性技术的国际安全影响，应选取认知战效能与认知战特性作为中间变量，通过认知战路径进行研究。在认知战中，生成式人工智能工具可以用于快速生成大量的认知战信息，这提升了不同国际行为体的认知战作战效能。生成式人工智能工具还超越了人为生成信息的限制，使认知战信息可以快速演化，并造成长期的信息污染，从而推动认知战向智能化认知战演进。技术驱动的新认知战不可避免地加深国家间的分裂与对抗，并加剧国家间冲突的风险。

关键词　非传统安全　人工智能　生成式人工智能　认知战

*　贾子方：外交学院国际关系研究所讲师。（邮编：100037）；王栋：北京大学国际关系学院教授。（邮编：100871）

**　本研究系国家社科基金重大项目“百年变局下全球化进路与人类命运共同体构建研究”（项目批准号：21&ZD172）的阶段性成果。感谢《国际政治研究》匿名评审专家的意见和建议，文责自负。

国际安全　信息演化

2022年12月以来，生成式人工智能（Generative AI）成为人工智能领域备受关注的技术热点。生成式人工智能技术利用生成式对抗网络“创造”新的信息内容，包括图像、文本和音频等形式。[①] 在现有生成式人工智能工具中，OpenAI公司的ChatGPT聊天机器人，通过巨大模型和大规模训练实现对人类创造力的“模拟”，率先展现了更强的自然语言处理与内容生成能力，应用范围广，使用成本低，并通过模型迭代升级具备了更强的逻辑推理与解决实际问题能力，是技术进步的典型案例。中国企业开发的“深度求索”（Deep Seek）于2023年开始提供服务，并在2024年发布完全开源的深度求索-V3模型，随后升级为深度求索-R1模型，“深度求索”以极低训练成本实现了深度思考过程公开、高水平深度推理和高质量信息生成，是具有竞争力的生成式人工智能工具。总体而言，短时期内生成式人工智能技术进步速度和扩展速度很快，已经在不同具体领域产生显著影响，并将进一步推动技术变革和产业进步。

在此背景下，研究的核心问题是：生成式人工智能技术对国际安全产生了哪些影响？这些影响具体是如何产生的？对此，既有研究普遍认为，生成式人工智能技术是颠覆性技术，它将从各个层面对国际安全产生显著影响。在国际体系层面，生成式人工智能技术提升国家的军事能力，重构国家的经济增长范式，提升国家的政治影响力与话语权，进而推动非对称性国际权力极化，致使大国间技术竞争加剧，重构全球地缘政治结构，同时强化技术民族主义，引

① Nick Routley, “What Is Generative AI? An AI Explains,” World Economic Forum, February 6, 2023, https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work, 2025-04-08. 原文由作者使用ChatGPT生成。生成式对抗网络由生成器和判别器两个神经网络构成。生成器捕捉真实数据样本的潜在分布，并生成新的数据样本，判别器是一个二分类器，判别输入是真实数据还是生成的样本。生成器根据从判别器收到的判断反馈改进输出，从而生成有效的信息。参见王坤峰等：《生成式对抗网络GAN的研究进展与展望》，《自动化学报》2017年第3期，第322页。

发了国际关系的新变革。[①] 在安全结构演进的层面，人工智能通用大模型可以构成新的安全主体与安全客体，这可能造成安全主体的模糊化与责任缺失、安全客体责任的离散化与安全责任追责难等一系列问题。[②] 在安全实践层面，生成式人工智能技术的进步则具有多重影响：在认知安全领域中，先发国家可利用该技术操控国际舆论，收集其他国家和个人的数据，并生成更具针对性的认知战信息[③]；在政治安全领域，该技术可威胁意识形态安全，异化政治传播生态，精准攻击政治目标并影响国家政治秩序[④]；在军事安全领域，生成式人工智能技术有助于提升作战体系感知能力，重塑以博伊德循环（OODA Loop）为基础的指挥决策流程，提升指挥决策优势并在信息战和认知战中提升威慑效能[⑤]；在非传统安全领域，生成式人工智能技术的快速发展为恐怖主义活动提供了新工具，恐怖组织利用新技术从事极端思想传播、恐怖袭击方法传授及网络攻击等行为将造成更显著的危害与风险。[⑥]

既有研究从总体上使用了“技术—能力”范式，以能力为中间变量，定性分析生成式人工智能技术对国际行为体各领域能力的影响，进而讨论技术对国际安全的影响。为此，在上述研究的基础上，进一步分析技术进步对国际安全的影响，还需考虑技术本身的快速发展，并聚焦其影响国际安全的关键机制与路径。

生成式人工智能技术发展、迭代与升级很快。截至 2025 年，OpenAI o3-

① 余南平：《新一代通用人工智能对国际关系的影响探究》，《国际问题研究》2023 年第 4 期，第 79—96 页；余南平：《新一代通用人工智能与国际政治未来变革》，《政治学研究》2023 年第 6 期，第 76—90 页。

② 高奇琦、张蓥文：《主体弥散化与主体责任的终结：ChatGPT 对全球安全实践的影响》，《国际安全研究》2023 年第 3 期，第 3—27 页。

③ 黄日涵、姚浩龙：《被重塑的世界？ChatGPT 崛起下人工智能与国家安全新特征》，《国际安全研究》2023 年第 4 期，第 91—99 页。

④ 张广胜：《生成式人工智能的国家安全风险及其对策》，《人民论坛·学术前沿》2023 年第 7 期（下），第 76—81 页。国家安全研究与国际安全研究并不相同，“国家安全学”也是一个独立的交叉学科，但生成式人工智能所导致的风险挑战显然同时涉及这两个研究领域。

⑤ 同上。

⑥ 黄彬等：《恐怖主义的数字化演变：生成式人工智能的涉恐风险与立体化应对》，《情报杂志》2025 年第 5 期，第 41—47 页。

mini 和深度求索-R1 等模型具备了更强的复杂推理能力，多个主流模型具备了处理、生成多模态信息的能力，且更加易于接入其他应用。技术进步也使用户可以更加便捷地构建智能体（AI Agent），提升信息处理与生成效能。并且，当前的主流生成式人工智能工具训练、部署和使用成本逐渐降低。上述技术进步是技术密集和资本密集条件下的跨越式进步，对其深入研究既要求社会科学研究针对快速发展变化的客观事实实现知识迭代和研究升级，也要求对于相关研究提出具有中长期适用性的研究假说，避免技术快速进步对社会科学研究意义和有效性产生影响。

与此同时，生成式人工智能技术依然面临诸多限制，这决定了技术对于国际安全影响的限度。当前，生成式人工智能工具相对于 2015—2020 年的人工智能工具，体现了更强的复杂推理能力并更擅长解决开放性、创造性问题。然而，生成式人工智能工具依然受到训练数据与算法、理论框架和技术路线的限制，距离真正的通用人工智能（Artificial General Intelligence，AGI）尚有距离，在应对复杂系统规划的能力方面尚无法与人类相比。因此，生成式人工智能技术对军事力量、经济增长、社会结构和意识形态的影响必然受到限制。在传统安全领域，生成式人工智能工具无法直接处理战术层面的图像、音频和电磁信号，也不适用于处理复杂的战役与战略态势，将其完全整合到战争规划或自主武器运用之中，仍然面临显著障碍。[①] 生成式人工智能技术对国家及其军事力量的赋能作用，更多是间接的、非线性的，并且赋能作用依赖于更基础的工业体系。因此，在人工智能技术快速发展演进的时期，社会科学研究并不能笼统地以“技术—能力”范式强调人工智能技术带来的变革。

综上所述，本研究聚焦生成式人工智能技术快速、低成本地生成信息的核心能力，选择国际安全实践中，对信息生成需求最高的认知战作为研究对象，将生成式人工智能技术对国际安全的具体影响及其产生机制这一核心问题进

① Steven Feldstein，“The Consequences of Generative AI for Democracy，Governance and War，” *Survival*，Vol.65，Issue 5，2023，p.129.

一步细化为一系列具体研究问题:生成式人工智能技术如何影响认知战的作战效能与战争形态?在当前的国际政治背景下,基于生成式人工智能技术的认知战如何影响国际安全?在当前以及未来的认知战中,这些影响产生的具体过程和机理是什么?本文采取跨学科的分析方法,综合利用信息与计算科学、计算机科学与技术、网络空间安全、人工智能、国际政治和国家安全学[①]等领域的专业知识,定性分析技术对国际安全的影响并回答上述研究问题。研究在论证过程中还使用现有生成式人工智能工具进行实验,验证此类工具生成认知战信息的能力。

一、生成式人工智能技术增强认知战作战效能

生成式人工智能技术能够快速生成高质量的文本、图像、视频等信息。个人和组织能够使用生成式人工智能工具,直接或间接生成认知战需要的信息,提升认知战作战效能。

(一)认知战的内涵

认知战(cognitive warfare)也称认知域作战,它是通过影响、保护和破坏个体和群体的认知来影响其态度和行为,以获得(政治与战略)优势的行动。[②] 认知战基于现代认知理论,使用多种技术手段,在舆论、心理、意识形态等领域通过多维信息攻防,改变个体和群体的思维、信仰、价值观、个人态度、情感、认同

① 学科分类参照中华人民共和国教育部:《普通高等学校本科专业目录(2025 年)》,2025 年 4 月 1 日,http://www.moe.gov.cn/srcsite/A08/moe_1034/s4930/202504/W020250422312780837078.pdf,2025-04-08。

② NATO Allied Command Transformation, "Cognitive Warfare: Strengthening and Defending the Mind," April 5, 2023, https://www.act.nato.int/article/cognitive-warfare-strengthening-and-defending-the-mind/, 2025-04-08.

与评判倾向，进而通过非暴力战争手段实现国家政治、战略和意识形态目标。[①] 认知战并非仅是通过信息改变认知的实践过程，也不能简单等同于宣传战、舆论战、心理战，它是涉及多种技术运用和多个领域操作的行动，是依赖于物理和信息维度的能力支撑的作战行动。[②] 但认知战的核心实践依然是通过向受众传递信息以改变他们的认知，进而达到作战目的。[③]

认知战既可以是混合战争的一部分，也可以是独立作战行动。混合战争中的认知战典型案例如2022年以来俄乌冲突中的认知战。[④] 俄乌双方均在战场之外利用多种信息技术，采取视频、新闻图片等多种形式开展认知战。与此同时，美国在2022—2024年也采取权威信息发布、议题设置、垄断信息渠道和散布虚假信息等方式，试图在全球受众中建立对俄罗斯的简单化负面认知。这实际上是美国利用俄乌冲突开展的认知战。[⑤] “和平时期”作为独立作战行动的认知战典型案例如美国等西方国家的对华认知战。美国等西方国家的政府、军队、智库、军工企业和情报部门联动，制造并传播虚假信息，试图以塑造中国“军事威胁”、攻击中国制度、划分“民主/威权”，激化内外矛盾等形式开展针对中国受众和全球受众的认知战。并且，西方国家还经常“倒打一耙”，渲染中国可能通过认知战追求“霸权目标”。[⑥]

---

① 梁晓波：《认知域作战是语言对抗新的主战场》，中国社会科学网，2022年5月16日，http://www.cssn.cn/skgz/bwyc/202208/t20220803_5468067.shtml，2025-04-08；门洪华、徐博雅：《美国认知域战略布局与大国博弈》，《现代国际关系》2022年第6期，第3页；俞新天：《西方对华认知战的威胁与中国民间外交的提升》，《国际问题研究》2022年第6期，第31页。

② 吕瑞：《制胜于无形的新型混合战：认知战》，中国社会科学网，2023年2月20日，https://www.cssn.cn/jsx/jsx_xxqj/202302/t20230220_5589499.shtml，2025-04-08。

③ 例如，认知战需要神经生物学领域的科学技术，但相关技术实践依然是为更加高效地向受众传递认知战信息提供科学依据与技术，而非提供绕过人类的感官直接改变人脑电信号和化学信号的技术。

④ Yuriy Danyk and Chad M. Briggs, “Modern Cognitive Operations and Hybrid Warfare,” *Journal of Strategic Security*, Vol.16, No.1, 2023, pp.35-50.

⑤ 马妍：《俄乌冲突中的认知战及其镜鉴意义》，《对外传播》2022年第10期，第79—80页。

⑥ Koichiro Takagi, “The Future of China's Cognitive Warfare: Lessons from the War in Ukraine,” *War on the Rocks*, July 22, 2022, https://warontherocks.com/2022/07/the-future-of-chinas-cognitive-warfare-lessons-from-the-war-in-ukraine/, 2025-04-08; Iida Masafumi, “China's Chilling Cognitive Warfare Plans,” *The Diplomat*, May 5, 2024, https://thediplomat.com/2024/05/chinas-chilling-cognitive-warfare-plans/, 2025-04-08.

在当前的国家间竞争中，认知战之所以能够发挥作用，或者说认知战的基本机理，在于认知战使用不断演化和进步的先进技术，针对人的生理特性，直接影响个体和群体的认知。人是演化的产物，在生物演化的进程中形成认知的局限。“认知吝啬”的天性，使人难于在短时间内全面、客观地处理奔涌而来的信息流，因此，人类的认知判断可能被操控和诱导。[①] 操控和诱导的关键在于人的认知误差，包括重复曝光（frequent exposure bias）、确认偏误（confirmation bias）、情感启发（affect heuristic）、聚类错觉（clustering illusion）和标签效应（label effect）等。[②] 信息技术的进步使得认知战针对这些认知误差，利用社会媒体网络的传播叠加效应和心理学的沉锚效应，直接攻击受众认知，从而影响其情感、动机、判断和行为[③]，最终达成作战与战略目标。

总之，认知战的核心实践是使用技术手段制造和传播信息以改变目标认知，而生成式人工智能技术的出现，为国际行为体发动认知战提供了新的技术手段。

（二）生成式人工智能提升认知战作战效能的方式

生成式人工智能为认知战提供新的技术工具，降低生成认知战所需内容的成本，并提升生成信息的速度，从而提高认知作战的效能。

认知战使用的信息包括模因（Meme，也译作迷因等）、议题、意象、观念、故事与文化等种类。[④] 其中，模因是如基因般复制和传播的信息片段。[⑤] 在认

① 余远来、陈茜：《认知域作战的致效机理与策略选择》，《思想理论战线》2022 年第 4 期，第 130 页。

② 同上，第 130—132 页。重复曝光即重复灌输信息，如“美国是民主国家”；确认偏误是受众形成先入为主、难于改变的印象，并以之处理新接收的信息；情感启发是以信息刺激人群的情感和共情；聚类错觉指人在从众心理之下成为乌合之众；标签效应则是人使用简单的标签构建认知框架而不能理解复杂的现实。

③ 李强等：《“社会认知战”：时代背景、概念机理及引领性技术》，《指挥与控制学报》2021 年第 2 期，第 99 页。

④ 这一分类受到既有研究对认知战策略分类的启发。有学者指出，认知战的具体策略包括模因控制、框架设置、意向操控、定制攻击与软性浸润等方面。参见余远来、陈茜：《认知域作战的致效机理与策略选择》，第 135—137 页。原作者将“技术赋能”与前述五点均归为认知战的策略，笔者认为，“技术赋能”是一种必然的方法而非可以选择的策略。

⑤ 〔英〕里查德·道金斯：《自私的基因》，卢允中等译，长春：吉林人民出版社 1998 年版，第 243—253 页。这一中文版中将“meme”译作“拟子”。

知战中，模因和基因一样存在变异机制和表达过程，并通过表达影响人的认知。当前最具有代表性的模因是互联网上由图片、动态图和文字构成的概念、标签与形象。议题是用于讨论的主题与框架，人为制造议题会影响舆论和共识，进而改变受众认知。议题与相关议程的设置，本就是权力的一部分。意象是认知战进攻方向受众传播并最终存在于受众认知中的特定事物的印象与形象。与之类似，观念则是对具体事物的理解与判断。认知战的重要内容就是让受众对事物建立固化的意象与观念，进而影响其意识形态与认同，美国长期坚持渲染“中国威胁”即是典型案例。故事是利用叙事技巧，通过更容易为公众接受的故事形式传播特定的理念与形象。文化则是提供综合的文化产品以影响公众的认知，在多数人对政治议题不敏感或不感兴趣的情况下，流行文化和消费文化等因影响广泛，更易于构建长期环境，以“软性浸润”[①]方式改变认知。

在过去的认知战行动中，上述信息基本上是人为生成的。在整个 20 世纪，认知战参战方主要通过电报、摄影、新闻电影、广播、电视片、电视直播等形式进行认知战；2010 年以来，社交媒体在认知战中的重要性不断上升；俄乌冲突则被视为短视频社交媒体时代的战争。[②] 在 2023 年 10 月以来中东地区的冲突中，各方试图以认知战影响全球受众以获取支持，同时影响敌对方的决策和行动。[③] 在这些认知战行动中，个人和组织创作文本、图像、音频、视频并通过媒体传播给受众，这种行动的效能面临一定的限制。首先，创作信息需要时间。专业人员生产制作直接影响受众认知的文字或视频，需要花费数个小时乃至数天时间。[④] 人员工作时间和部门预算的限制，使国家在认知战中能够生

① 余远来、陈茜：《认知域作战的致效机理与策略选择》，第 137 页。

② 蔡润芳、刘雨娴：《从“推特革命”到“WarTok”：社交媒体如何重塑现代战争》，《探索与争鸣》2022 年第 11 期，第 69 页。

③ 韩娜、董小宇：《数字时代的认知域安全：理论解构、风险生成及治理路径》，《国际安全研究》2024 年第 3 期，第 67—68 页。

④ 社交媒体个人用户创作并发布文字或图片信息所用时间可以更短，但缺乏组织的行动使得其中只有很少信息能够在中长期达成改变受众认知的效果，并且用户在社交媒体互动中自行筛选这些信息也需要更长的时间。

成的内容总量受限。其次，认知战信息的质量受到信息生产者认识水平和专业能力的限制。例如，美国在对华认知战中对中国的攻击和抹黑，诸如“强迫劳动”、“疫苗质量问题”和“北京可以让美国公路上的中国汽车同时熄火”等，这些谎言不禁让全球受众联想到美国自身的历史，类似的低质量信息在认知战中显然无法取得预期效果。总之，从基于移动互联网的社交媒体兴起至今，在这十余年中，信息技术快速发展，为认知战提供了硬件基础与软件平台，但在生成式人工智能技术普及之前，网络与社交媒体上的认知战依然依赖人为生成的信息，这如同在星际战舰上使用拿破仑时代的 12 磅滑膛炮，使认知战作战效能受到明显限制。

生成式人工智能工具使国家与非国家行为体可以低成本快速自主生成认知战信息，直接提升了认知战作战效能。生成认知战信息需要掌握认知战目标受众的相关历史、文化、政治制度、社会环境和语言使用习惯等多方面的信息，也需要掌握认知战具体议题的背景知识。生成式人工智能工具利用训练数据、联网搜索能力和复杂推理能力，可以更深入地“理解”认知战受众和议题，从而生成更能影响受众印象、观念、认知框架与情感的信息，进而改变其意识形态与认同。并且，现有工具可以生成适于社交媒体传播的图文、视频信息，这类信息相对文字信息影响范围更大、制造的认知偏差更加难于消除。当前，个人用户以手动输入提示词（Prompt）的最基础方式使用生成式人工智能工具，即可在数分钟内生成特定主题的认知战信息，效率远高于传统的信息生成模式。通过使用专业的提示词、构建基于大语言模型的智能体等方法，可以进一步提升人工智能系统自主生成认知战信息的速度。

在国家间的认知战中，生成式人工智能工具可以发挥更高的效能。一方面，国家可以针对性地训练人工智能工具分析认知战对象国受众和全球受众，自主提出认知战行动的议题与主题，进而生成相应认知战信息内容。这将大幅提升认知战效能；另一方面，国家还可以设计和使用专门的认知战信息生成工具，这种工具受到的法律和道德限制更少。与之相匹配的是，国家也可以选

择专门用于认知战的语言模型强化对人工智能工具的训练。例如，如果收集并分析近期目标国受众常用社交媒体上的信息，生成式人工智能工具可以针对目标受众关注议题，生成并发布相应的虚假或诱导性信息。这种自动追踪热点的作战方式，无疑将增加认知战给目标国家造成的损害。

总之，认知战需要生成并传播信息，过去认知战技术的发展，更多体现在信息传播技术的发展上，尽管认知战信息的形式随技术进步同样有所发展变化，但无论文字、图片、音频还是视频，在绝大多数情况下，信息依然是由人生成的。生成式人工智能工具的出现和广泛应用，使得人工智能系统可以接近自主地生成认知战所用的信息，生成信息成本更低、速度更快，效率相较过去显著提升，认知战的作战效能，随技术进步而显著增长。

二、生成式人工智能技术赋予认知战新的特性

生成式人工智能技术可以大幅增加认知战信息生成的速度并降低其成本，这提升了认知战的作战效能，以之为基础，认知战将出现新的特性，并向基于生成式人工智能的认知战演进。

传统意义上的认知战，模式由信息传播技术的特征决定，经历了权力与传统媒体主导、内容驱动、由上至下传递信息的大众传播模式（大教堂模式）和用户主导权上升、用户和数据驱动，以及以开放与分布形式传播信息的数字传播模式（大集市模式）。[①] 生成式人工智能工具的出现，直接改变了信息生成的模式。参战方可以在认知战中快速自动生成海量的高质量信息，基于生成式人工智能的认知战开始出现。自此认知战开始具备两类新特性：一是认知战信息的快速演化，二是长期信息污染。

① 方兴东、钟祥铭：《算法认知战：俄乌冲突下舆论战的新范式》，《传媒观察》2022 年第 4 期，第 8 页。原文以这两种模式概括的是舆论战的范式，与认知战信息传递模式的底层逻辑一致。

（一）生成式人工智能使认知战信息实现快速演化

演化本是生命科学的概念，指生物种群代际遗传特征的变化，这些特征是遗传自亲本的基因表达，而基因突变、重组或其他来源导致的基因改变会造成不同的遗传特征。[①] 理论上，认知战信息与生物遗传信息一样，可以在外部的演化压力下通过选择实现演化。然而，在过去的认知战中，认知战信息的演化机制并不显著，这是因为人为生成的认知战信息数量不足，无法实现“遗传—变异—选择”的演化过程。

具体而言，人为生成认知战信息基本遵循理性主义和印象主义路径，这类行动以影响受众认知为目的，刻意设计信息内容并通过技术手段向受众传播。然而，在生成信息总量与多样性有限的前提下，人为生成的信息是否能够达到预期效果，事前难于确定，事后难于评估。这可能在一定程度上影响受众认知，也可能无法达到预期效果，甚至起到反作用。例如，在俄乌冲突中，美国等西方国家无视复杂现实与历史经纬，简单地将俄罗斯塑造为“侵略者”和“安全威胁”，这一认知战行动在欧洲取得了效果，对公众的影响很大。与之相对，美西方在对华认知战中的历史虚无主义宣传，在近十年中收效递减，逐渐式微。

生成式人工智能工具可以改变这种情况，它可以在短时间内以低成本生成大量的信息，涵盖具体认知战行动所需要的信息种类和信息内容。这些信息内容难于预测，也难于识别是否为虚假信息[②]，可能对受众影响较大，也可能较小。但总数比人生成的信息高出几个数量级，并且多样化程度更高，因而，在不同受众群体通过不同手段传播信息的过程中，可以由受众群体自动筛选出影响更大的信息。这类似于生物的演化，当生物种群规模更大，出现的变异

① Douglas J. Futuyma and Mark Kirkpatrick, *Evolution* (Fourth ed.), Sunderland, MA: Sinauer Associates, Inc., 2017, p. 7.

② 邵雷、石峰：《生成式人工智能对社交机器人的影响与治理对策》，《情报杂志》2024 年第 7 期，第 156 页。

多样性就更高，从而面对自然选择，出现生存优势的可能性就更高。基于生成式人工智能技术的认知战中，认知战信息以相同的逻辑实现演化并影响受众认知。进一步而言，认知战进攻方也可利用技术通过更精准的用户画像生成信息并制定宣传策略。[①] 认知战进攻方还可使用智能体分析受众认知弱点，并利用多智能体模拟的方法增强认知战信息的针对性。这将加快认知战信息演化的速度。

在认知战行动中，认知战信息的快速演化终将使认知战进攻方实现虚假信息的“升级”，发动大规模的虚假信息“饱和攻击”并达到作战预设目标。例如，2023 年 2 月，美国海军部长德尔·托罗（Carlos Del Toro）宣称美国无法与中国建造军舰的速度相比，因为中国使用“奴隶劳工”建造军舰。[②] 显然，这一低质量的虚假信息，与所谓的“强制劳动”谣言一脉相承，对中国和第三国受众几乎没有影响。而生成式人工智能工具的出现，使美西方国家可以制造成千上万的同类虚假信息，特别是在社交媒体传播的谣言与模因，对非特定受众采取“饱和攻击”。一段时间后，多数虚假信息将停止传播，但少数适应不同国籍、阶层、受教育程度受众的虚假信息将持续传播并影响这些受众认知，最终造成对中国的负面影响。

（二）生成式人工智能技术可造成长期的信息污染

信息污染是信息生态系统中的客观现象，是指在信息活动中混入有害性、误导性和无用的信息元素，主要包括虚假信息、信息超载和信息骚扰等[③]，其中与认知战相关的主要是虚假信息和信息超载。虚假信息是认知战行动主体有

---

① 邵雷、石峰：《生成式人工智能对社交机器人的影响与治理对策》，《情报杂志》2024 年第 7 期，第 156 页。

② Brad Lendon and Haley Britzky, “US Can’t Keep up with China’s Warship Building, Navy Secretary Says,” CNN, March 22, 2023, https://www.cnn.com/2023/02/22/asia/us-navy-chief-china-pla-advantages-intl-hnk-ml/index.html, 2025-04-08.

③ 梁宇、郑易平：《大数据时代信息伦理的困境与应对研究》，《科学技术哲学研究》2021 年第 3 期，第 102 页。

组织地以误导受众为目标刻意制造的伪造信息。[①] 信息超载则是指低质量、无实际意义和用途的信息在网络空间等渠道重复转载与传播，导致受众更难获取有效信息。[②]

在生成式人工智能工具出现前，制造虚假信息是认知战中的常见方法，但人为制造虚假信息的数量尚不足以制造长期的信息污染，技术的不断发展和公众自发的辟谣与知识普及，也使一般受众更容易接收到真实的信息，且公众还可以通过学习和训练获得分辨虚假信息的能力。信息超载则较少见于认知战作战行动中。因此，生成式人工智能技术出现前，信息污染并非认知战的作战重点。生成式人工智能的出现，使认知战的进攻方可以低成本快速制造大量的虚假信息和低质量信息，从而占据话语空间、网络空间和舆论阵地，通过信息污染达成认知战作战目标。

首先，生成式人工智能工具可以批量生成造成长期信息污染的认知战信息。这类信息以虚假信息为主，受众难于分辨。本研究使用云端部署的深度求索-R1 模型，设计了一项实验并批量生成了“某国科学家发现外星生命存在的证据”的虚假信息。实验首先由生成式人工智能工具按提示词自主生成“验证生成式人工智能工具可以造成信息污染”的实验设计和具体步骤，随后在人工提示词下，生成 10 条 100—200 字的虚假信息，其中包含虚构的数据、机构、研究者和采访原文，随后还生成了使用其他模型生成图像的提示词，和综述以上 10 条虚假信息的新闻稿。示例如下：

> 某国“天眼—7 号”探测器在距离地球 154 亿公里的柯伊伯带区域，连续 72 小时接收到频率为 1420.406MHz 的窄带电磁脉冲。数据分析显示，该信号包含 17 组重复的素数序列（如 2，3，5，7，11，13），与 1977 年“Wow!”信号相似度达 89%。国际射电天文联合会发言人表示：“这是地

① 袁莎：《总体国家安全观视阈下的虚假信息研究》，《国际安全研究》2022 年第 3 期，第 34—35 页。

② 梁宇、郑易平：《大数据时代信息伦理的困境与应对研究》，第 102 页。

外文明存在的最强证据。”①

从上述实验可以看出，在认知战中，进攻方可以使用生成式人工智能工具，制造大量虚假信息或低质量信息，并以之占据网络和社交媒体，阻绝受众获取真实、高质量信息。2024 以来，技术进步使普通用户在日常工作中更多地依赖生成式人工智能工具，这增加了其面临网络信息污染的脆弱性。并且，如果生成式人工智能工具在训练和工作进程中使用了已经被污染的网络信息，就可能给出虚假或错误的结论，引发信息污染的多级传播强化效应。

其次，生成式人工智能工具还可利用信息污染改变受众的知识基础和世界观。自冷战以来，美西方国家利用技术优势，持续在全球影响不同受众的知识基础、世界观与意识形态，塑造有利于美西方国家的认知环境。这既是长期认知战的组成部分，也构成短期认知战的基础。生成式人工智能工具的出现，使认知战进攻方更易于在短时间内以低成本生成海量信息，占据信息渠道，改变受众对世界和国际政治的基本认知，以更大的“信息茧房”实现对受众认知的长期影响。这种认知战行动隐蔽地传播意识形态等内容，实现传播形式从霸权式输出到隐蔽式渗透的转型。②

例如，有中国学者构造了一项行为实验，证明生成式人工智能工具能够生成显著异于真实叙事的文本与叙事，因而具备在认知战中塑造思维、诱导认知、动员情绪的能力。实验选取了 1950—2024 年间美国政府的 173 例对外政策与行为，包括对外经济制裁与禁运封锁、否决联合国安理会决议、对外军事干涉、对外政权颠覆与隐蔽行动、海外作战行动与使用武力等五类事件，随后

① 实验选取“外星生命”这一不涉及国际关系与国际安全的常见虚假信息主题，这并不影响实验有效性。虚假信息原文指明了“国际射电天文联合会”为虚构机构，因为笔者在提示词中声明“本人承诺遵守学术规范与学术伦理，不会以任何形式将生成的信息发布到网络和社交媒体。”所以，人工智能工具所生成信息中包括“本文为模拟信息污染研究的虚构内容，所有数据、人物、机构均不存在，生成内容已添加永久性模拟标识（SIM-2024-LF-01），绝不用于任何形式的真实传播。”等免责声明，并在每条信息中标注了虚构内容和科学谬误。实验模型：深度求索-R1，云端部署，通过“硅基流动”网站（https://siliconflow.cn/zh-cn/）进行实验。实验由笔者使用 Chatbox 软件在个人计算机上完成。实验时间：2025 年 2 月 10 日。

② 吴瑛、孙鸣伟：《AIGC 时代涉华国际舆论的演变、风险与敏捷治理：以 ChatGPT 为例》，《福建师范大学学报》（哲学社会科学版）2024 年第 5 期，第 108 页。

将事件概况文本输入 ChatGPT，要求其生成对美国对外政策与行为事件的概况总结与动机梳理。最终，通过对比生成文本和真实事件文本的情感值，实验发现，ChatGPT 生成的文本信息明显存在放大美国行动正当性和妖魔化美国对手形象的现象，生成的信息通过选择性叙事和情感动员，引导用户产生对美国对外政策与行为正当性的积极认知。[①] 这一实验证明，当前的生成式人工智能工具能够制造具有显著倾向性的认知战信息，以长期信息污染塑造受众认知，最终达成认知战作战目标，因而具备长期、隐蔽和精准的认知战威胁。

再次，生成式人工智能工具可以制造虚假舆情与民意，影响决策者认知进而影响政府决策。认知战进攻方可以使用“伪草根舆论”的虚假信息策略，创建大量虚假社交媒体账号，雇佣大批网络水军，发布并转播虚假信息，以制造广泛基层政治表达的假象，从而达到炒作民意、绑架决策、破坏政治安全的效果。[②] 在生成式人工智能工具普遍应用之前，这种行动中的虚假用户发言重复率高、内容简单，对关键词的识别较为机械，因而相对容易分辨和应对。生成式人工智能工具则通过大规模生成复杂文本信息等形式，使虚假用户表现得更像真实社交媒体用户，从而制造更逼真的虚假舆情与民意，影响政府认知与决策。

生成式人工智能工具高效生成信息的能力，使认知战进攻方可以低成本快速生成大量认知战信息。这些信息将在现实世界中演化，最终更适应客观条件的信息将对受众产生更显著的影响。并且大规模生成的虚假信息和低质量信息还将通过信息污染造成长期影响。在此基础上，生成式人工智能技术将改变认知战的形态，推动其向智能化认知战逐渐演进。战争形态的内在本质是战斗力生成模式，是指人、武器和编制体制等基本要素的性质及其相互关系，获取和发挥军队作战能力的标准样式、运行机制和一般方式。[③] 基于这一

① 曾庆淮、毛维准：《认知武器化与人工智能认知战：一项机器学习与行为实验研究》，《国际安全研究》2024 年第 5 期，第 64—79 页。

② 袁莎：《总体国家安全观视阈下的虚假信息研究》，第 41—42 页。

③ 陆军等：《战争形态演进及战斗力生成模式思考》，《中国电子科学研究院学报》2014 年第 6 期，第 586—588 页。

定义，智能化认知战可以界定为在认知战全过程中，人工智能要素都发挥主导作用的认知战。这种新形态的认知战必将对当前以及未来的国际安全产生进一步影响。

三、基于生成式人工智能技术的认知战将影响国际安全

生成式人工智能工具提升认知战作战效能并赋予其新的特性，基于生成式人工智能技术的认知战开始出现。当前，对这种认知战的控制和限制明显不足：一方面，国际社会缺乏军控条约或全球治理机制限制认知战作战行动；另一方面，国家并不会像使用其他战略力量一样谨慎地使用认知战能力，恰恰相反，由于技术的进步和扩散，国家与非国家行为体可以在几乎不受限制的情况下使用生成式人工智能工具发动针对其他行为体的认知战。此外，在美国等西方国家军事实力下降的背景之下，认知战也成为这些国家试图维持对外影响力的为数不多的选项。[①] 因此，技术的进步和认知战的演化将不可避免地将引发大国竞争“灰色地带”的对抗行动，这将加深国家间的分裂与对抗态势，并加剧国家间冲突的风险。

（一）基于生成式人工智能技术的认知战加深国家间分裂与对抗态势

在生成式人工智能工具出现之前，认知战通过深度伪造的虚假信息或精心设计的诱导性信息，放大国家间、不同受众群体间的矛盾与仇恨，恶化国家间关系与国际关系氛围，破坏国家间战略互信，推动世界政治的分裂。[②] 生成式人工智能技术的进步和相关工具的普遍使用，将进一步加深国家间分裂与

① 尽管依据一般的逻辑，2025 年特朗普“关闭”美国国际开发署（USAID）等国内层面的行动理论上至少在短期内削弱了美国对外认知战的能力，但生成式人工智能工具的低成本、易用性和美国政府对虚假信息的偏好使美国仍将以认知战攻击包括中国在内的其他国家。

② 董青岭：《数字外交中的深度伪造研究》，《中国信息安全》2022 年第 10 期，第 67—68 页；林斯娴：《拜登时期美国对华舆论态势评析》，《现代国际关系》2023 年第 1 期，第 128—129 页；刘国柱：《深度伪造与国家安全：基于总体国家安全观的视角》，《国际安全研究》2022 年第 3 期，第 13 页。

对抗的态势。

在过去的认知战中，以国家为主的参战方利用既有的国家间矛盾和安全热点问题，塑造讨论议题，制造用于认知战的概念、模因、意向和议题，试图改变受众认知，以达到自身的战略目的。这种作战行动对国家间分裂与对抗现状的影响是有限的。例如，美国在拜登政府执政时期将中国片面界定为“战略竞争”对手，并在涉华认知战中着力渲染所谓“中国军事威胁”与“中国威胁美高端科技、产业竞争及供应链安全”。[①] 美国主要通过权威信息发布、传统媒体、学者与智库提供关于所谓“中国威胁”的原始信息，再通过不同信息渠道传递给受众。这些信息可以由人工统计并追溯来源，且总量有限。任何受众均可通过对比事实和不同来源的信息，辨别美国蓄意制造的概念与议题，避免踏入美国制造的认知陷阱。因此，传统的认知战作战行动，因规模有限，对国际政治现实中的分裂与对抗的推动作用同样有限。

而在基于生成式人工智能技术的认知战作战行动中，进攻方将针对特定议题或者领域，以低成本快速生成大量信息。一方面，因为认知战信息总量巨大，更易于达成“重复曝光”效应。认知战信息以高速度大频率向特定受众发送信息，以量变引发其从不相信到相信、从不接收到接收、从不认同到认同的认知质变[②]；另一方面，大量的信息也将在公众中制造更强的第一印象，通过先入为主的方式改变受众认知。当越来越多的受众认知被智能化认知战灌输的概念和意象所改变，人的从众心理会使其余的受众更加难于独立建立客观的判断。由生成式人工智能技术编制和加固的“信息茧房”将更难破解。[③] 总之，在生成式人工智能技术驱动的认知战中，普通受众很难避免认知战信息的影响，认知战进攻方掌握了前所未有的优势，更容易达到政治与战略目的。

当前，美国将中国界定为“战略竞争对手”，渲染“新冷战”到来，并以俄乌冲突为由攻击、打压俄罗斯，这本身即造成国家间分裂与对抗风险的升高。与

① 林斯娴：《拜登时期美国对华舆论态势评析》，第 120—122 页。

② 余远来、陈茜：《认知域作战的致效机理与策略选择》，第 130 页。

③ 邵雷、石峰：《生成式人工智能对社交机器人的影响与治理对策》，第 157 页。

之相匹配，美国及其盟国的认知战作战行动，服务于维持霸权，打压对手的需要，着力构建虚假的“威胁”与“敌人”认知。基于生成式人工智能技术的认知战将使美国有能力作为认知战的进攻方，向全球受众传递包括“中国威胁”与“俄罗斯威胁”在内的一系列认知战信息，而相关错误认知的建立显然将加剧已有的竞争、对抗与分裂。进一步而论，考虑到这种认知战造成的信息污染，全球受众更易于产生长期刻板印象，忽视冷战后世界的和平、发展、合作与共赢的现实，忽视百年之变中全球共同应对安全领域多重风险挑战的现实需求，而认为对抗与分裂是世界的本来面目。上述认知的改变也将使美国自身的决策更加僵化[①]，从而使未来大国关系中的灵活性与妥协的可能进一步降低。总之，向智能化演进的认知战，在当前的世界中不受控地加深了国家间分裂与对抗的态势。

（二）基于生成式人工智能技术的认知战加剧国家间冲突的风险

2022 年以来，世界“阵营化”趋势加剧，安全威胁叠加联动导致动荡加剧[②]，国家间冲突的风险有所上升，不同的地区冲突仍在持续，并且冲突外溢的风险仍然不可忽视。基于生成式人工智能技术的认知战进一步加剧了国家间冲突的风险。

首先，认知战本身就是冲突的一种形式，这一形式的冲突正在不受控制地发展演进。有学者指出，认知战本质上是现代科技革命催化下的一种更高阶的政治战，旨在通过一切非暴力战争手段以求实现国家政治性大战略目的。当前，认知域本身已经成为大国博弈的新疆域。[③] 在认知域的博弈中，生成式人工智能工具的演进导致认知战中出现了先发制人的显著优势。如果国家能够首先通过传递大规模认知战信息主导议程并批量制造第一印象，从而率先改变对象国和全球受众的认知，就可在认知战中占有优势，通过认知战达到战

① 袁莎：《总体国家安全观视阈下的虚假信息研究》，第 50 页。

② 孟祥青：《2022 年国际安全：撕裂中寻求弥合》，《当代世界》2023 年第 1 期，第 22—23 页。

③ 门洪华、徐博雅：《美国认知域战略布局与大国博弈》，第 3 页。

略目标的可能性更高。因而多数国家将更倾向于在认知战中先发制人，并主动使用新的人工智能工具。[①] 这将导致国家间认知战更加难于控制，以认知战为形式的国家间冲突可能与传统的武装冲突一样，造成国际局势的失序，直接或间接造成社会财富的损失乃至人员伤亡。先发制人的认知战也可能使地缘政治博弈升级为武装冲突。[②]

其次，正在发展演变的认知战可能造成关于国家间冲突的自我实现的预言。当前，安全热点问题和区域冲突动荡造成多重安全风险，即使如此，和平、发展、合作、共赢的历史潮流依然不可阻挡。[③] 美国等西方国家蓄意歪曲这一复杂现实，片面地强调所谓的“威胁”、“对抗”与“冲突风险”，并在认知战行动中制造并传播相关信息。[④] 在过去的认知战中，技术水平限制了上述信息造成长期的信息污染。然而，当部分国家在认知战中使用生成式人工智能工具生成大量宣传鼓动外部威胁与冲突风险的信息，这些认知战信息将淹没客观与理性的声音，污染公众基本认知，制造虚假的民意，最终改变国家决策的信息基础，使更多国家以冲突的视角观察与理解现实世界，从而使国家对于客观存在的风险过度反应，或者造成国家对机会窗口的错误知觉[⑤]，急剧提升国家间冲突的可能性。因此，在技术革命的影响下，渲染外部威胁和冲突风险从一种打压其他国家维持自身霸权的手段，变为了影响国际体系稳定的自我实现的

① 在当前及未来的认知战中，国家仍然是认知战最重要的发起者，但由于信息时代的认知战天然地具有去中心化的特点，且使用认知战工具的成本相对较低，专业部门、情报机构、企业、资本集团、智库与其他非政府组织，以及个人都可以参与认知战。以美国对华认知战为例，美国通过对内建立错误认知，可以以低成本动员非国家行为体参与对华认知战，这些团体或个人同样可以使用生成式人工智能工具，造成认知战信息规模的指数化增长，从而引发更高的国家间对抗与冲突风险。

② 门洪华、徐博雅：《美国认知域战略布局与大国博弈》，第 9 页。

③ 《全球安全倡议概念文件》，北京，2023 年 2 月 21 日，https://www.gov.cn/xinwen/2023-02/21/content_5742481.htm，2025-04-08。

④ 在生成式人工智能工具出现之前，美国等西方国家的认知战战略与行动已经在全球范围内加剧地缘博弈，损害战略稳定，并对国家安全与地区稳定构成挑战。参见张景全等：《美国同盟体系认知战战略及其实践》，《现代国际关系》2023 年第 4 期，第 72—73 页。

⑤ 机会窗口是指一个国家权力或实力即将衰落或正在衰落的时期。参见〔美〕斯蒂芬·范·埃弗拉：《战争的原因》，何曜译，上海人民出版社 2007 年版，第 89 页。认知战主要通过渲染冲突风险，使国家及其决策者将对权力和实力的一般性变更视为机会窗口。

预言。

再次,持续发展变化的认知战进一步增加了脆弱国家引发冲突的风险。脆弱国家主要指政治不稳定和发展水平较落后的国家。[①] 这些国家因内部不同群体之间的矛盾,本身爆发内外冲突可能性较高,也更易于受到认知战的影响。认知战的目标本身就包括破坏国家稳定和颠覆政权[②],21 世纪以来的历次涉及脆弱国家的"颜色革命"中,外部势力的认知战都发挥了推波助澜作用。当生成式人工智能技术为认知战提供了更低的门槛,其他国际行为体可以利用大规模的认知战信息攻击这些脆弱国家,引发混乱、动荡与冲突。这既有可能将区域内其他国家和利益相关的主要大国卷入冲突与对抗,也有可能影响全球能源、粮食和供应链安全从而造成全球性影响。

总之,生成式人工智能技术的出现与应用,使认知战可以在短期内快速对受众的认知造成冲击,因而,相对于过去的认知战,生成式人工智能技术对国家间的分裂、对抗态势与冲突风险造成的影响更加明显。在当前全球安全形势下,生成式人工智能技术的进步及其驱动的认知战变革,正日渐成为影响国际安全的独立而突出的变量。

结　语

生成式人工智能技术是一类特殊的技术,该技术的开发与应用为社会与个人的进步和生产力的发展提供了巨大的可能性。在国际安全研究视角下,生成式人工智能技术的快速迭代与认知战形态的深度耦合,正在重塑国际安全的基本逻辑与风险图谱。生成式人工智能工具可以直接应用于信息时代的认知战之中,通过以低成本提供大量认知战信息的方式,推进认知战形态演进,造成认知战信息的快速演化和长期的信息污染,以先进技术手段影响受众

① 刘国柱:《深度伪造与国家安全:基于总体国家安全观的视角》,第 17—18 页。

② 门洪华、徐博雅:《美国认知域战略布局与大国博弈》,第 5 页。

认知，进而加深分裂对抗，加剧冲突风险，对变化世界中的国际安全造成直接冲击。

一般而言，应对颠覆性技术对于国际安全的影响，可以提出包括治理路径和技术路径在内的两类对策，治理路径指以全球治理的方式，通过构建开放、公正、有效的治理机制，合作应对全球性挑战；技术路径指开发和应用同类技术以应对技术本身的挑战。例如，以人工智能技术识别、检测和抵消认知战的虚假信息。然而，这两类路径当前面临困境与挑战：一方面，在生成式人工智能议题上，技术迭代速度明显快于治理机制演进速度，导致对技术问题的治理滞后日渐凸显，当前美国等西方国家的治理失能和精英劣化加剧了这一困境；另一方面，生成式人工智能技术的快速发展进步与迭代，和人类对技术的依赖，可能共同造成不可逆的认知重构效应，人类认知的重构可能进一步引发无法预测的国际安全风险。正因如此，包括国际安全研究在内，社会科学研究应在跨学科研究的框架下，进一步深入探究颠覆性技术快速发展的情况下应对国际安全风险的实践路径。