2026年人工智能与国际安全研究

中美人工智能发展、应用与治理

2026 年 6 月

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

引 言

2026 年 5 月，中美两国元首在北京会晤期间，就人工智能问题进行了建设性交流，同意开展人工智能政府间对话。作为两个人工智能大国，中美双方应该携手促进人工智能发展和治理，推动人工智能更好服务人类文明进步和国际社会共同福祉。自 2019 年 10 月以来，清华大学战略与安全研究中心（CISS）与美国布鲁金斯学会（Brookings）共同举办中美人工智能与国际安全二轨对话，围绕人工智能领域相关议题共举办了 14 次对话。双方发表了项目阶段性研究报告，并各自发布了三份人工智能相关术语研究报告，不断增进彼此在关键议题上的认知和理解。在中美两国宣布重新开展人工智能政府间对话的背景下，CISS 人工智能项目组成员结合前期对话交流经验，分别围绕人工智能安全与合作、通用人工智能发展、军事领域应用以及虚假信息治理等议题，探讨中美人工智能领域的发展、应用与治理，为两国人工智能领域良性互动与共同发展提供相关思考。

目 录

人工智能安全——中美关系不应忽视的合作议题 01
◎肖　茜

中美如何做大人工智能合作的蛋糕 04
◎孙成昊

中美通向AGI的不同道路比较 07
◎鲁传颖

人工智能军事应用的中美竞合 10
◎祁昊天

AI虚假信息领域的中美合作可能 16
◎董　汀

人工智能安全
——中美关系不应忽视的合作议题

肖　茜

随着外界关于美国总统特朗普即将访华的讨论不断升温，当前舆论的关注点大多集中在关税、贸易以及半导体等议题上。也有观点认为，两国领导人在会谈中将提及人工智能相关议题。但如果人工智能仅仅被放在技术竞争的框架下讨论，那么中美两国很可能会错失当前为数不多、仍然存在合作空间、且迫切需要合作的重要领域之一。

近年来，美国对人工智能的政策逻辑正在逐渐从“创新治理”转向“战略竞争”。“赢得人工智能竞赛”的叙事正在主导美国国内人工智能政策讨论。出口管制、投资限制以及技术联盟也越来越多地被纳入国家安全的框架之下加以推进。

的确，人工智能正在迅速成为一种具有重大经济、军事和地缘政治意义的基础性技术。任何国家都不愿在这一领域落后。但问题在于，人工智能不仅仅是一项竞争性技术，它同时也是一种可能带来系统性风险的技术。

当前最严重的人工智能风险并不会止步于国界之内，也不可能由任何一个国家独立应对。无论是 AI 赋能的网络攻击、军事自主系统，还是大规模虚假信息传播，一个国家内部产生的不稳定因素都可能迅速外溢至整个国际体系。也正因此，人工智能安全应当成为未来中美高层接触中的重要议题之一。

中国已经多次展现出其在人工智能治理与安全问题上开展国际对话的开放态度。近年来，中国发布了一系列政策文件与治理倡议，强调在推动人工智能发展与保障安全之间实现平衡。同时，中国的科研机构、科学家以及政策专家也在国际上积极参与关于人工智能安全的讨论。

然而，在当前的地缘政治氛围下，这一点往往被外界忽视。今天国际社会关于人工智能的讨论大多聚焦于芯片、算力和基础设施。但在这些热点议题背后，还存在另一个现实：中美专家实际上在持续、认真地就人工智能风险问题开展交流。尽管中美关系近年来持续紧张，但过去几年间，由两国科学家、研究人员与政策专家参与的二轨对话仍在低调推进。这些讨论涵盖了从“失控风险”、AI 赋能网

络攻击、AI 与生物安全风险，到术语解释、信任建立措施等多个方面。

我个人在参与这些交流过程中观察到，尽管双方在政治立场上存在差异，但许多中美专家对于人工智能未来发展方向的担忧其实高度相似。越来越多的人意识到，一些风险已经大到任何国家都无法独自应对。

中美两国都有责任共同降低人工智能相关风险。作为全球最重要的两个人工智能国家，北京和华盛顿所做出的决策，不仅将影响技术发展的未来，也将深刻塑造全球人工智能风险环境。

由图灵奖得主 Yoshua Bengio 牵头、100 多位 AI 专家共同撰写的最新《国际人工智能安全报告》指出了两个尤其值得关注的风险领域。

第一个是人工智能与网络攻击的结合。先进人工智能系统正在迅速改变网络空间：它降低了复杂攻击的门槛，加速了自动化漏洞发现，并扩大了信息操控的规模。在当前大国之间本就高度缺乏信任的环境中，AI 赋能的网络事件极易被误判为“蓄意升级”。

第二个是人工智能与生物风险。人工智能已经在医学与生命科学领域显著加速科研突破。但与此同时，越来越多科学家担心，能力不断增强的 AI 系统可能会降低危险生物行为的技术门槛。这种担忧如今已经不再局限于政府层面，更涉及非国家行为体获得过去仅限国家级实验室掌握能力的可能性。

无论是中国还是美国，都不希望看到上述风险的发生。这也正是为什么中美维持人工智能安全沟通渠道至关重要。人工智能时代有一个被严重低估的风险：系统故障越来越可能被视为“敌意行为”。随着 AI 系统变得更加自主、更加复杂、更加“黑箱化”，各国政府越来越难以区分一次事故究竟是技术失灵、意外升级，还是蓄意攻击。在危机情境中，这种模糊性本身就可能成为不稳定因素。

历史已经反复证明，过度“安全化”技术只会加深不信任，并导致全球创新生态的碎片化。今天，人工智能治理本身正在被全面卷入地缘政治对抗之中。如果每一次人工智能能力突破都被置于零和竞争框架下解读，那么真正需要的安全合作，反而会在最关键的时候变得政治上不可行。

更加务实的路径是承认人工智能安全合作本身符合中美双方利益。中美可以扩大围绕前沿 AI 风险的专家对话，加强重大 AI 事件的沟通机制，推动人工智能

评估与安全测试的联合讨论，并探索军事 AI 领域的信任建立措施。

事实上，中美之间存在一些合作基础。过去几年，两国专家已经通过多个围绕人工智能与国际安全的二轨对话开展合作。科学家与研究人员持续就风险治理和新兴技术威胁交换看法。这些交流或许不会登上新闻头条，但它们的价值在于在一个越来越不确定的技术时代帮助减少误判与误解。在当前许多传统双边互信渠道已经明显削弱的背景下，人工智能安全是少数仍然存在建设性接触空间的领域之一。

但对华盛顿而言，这同样意味着需要进行某种思维转变。将中国完全视为人工智能领域的战略对手，或许能够形成国内政治共识，但也可能压缩双方在共同风险问题上的务实合作空间。并非所有人工智能治理议题，都应当被放在技术遏制或意识形态竞争框架下加以理解。

真正摆在中美两国面前的问题，并不仅仅是谁将在人工智能能力竞争中领先。更重要的问题是：世界上最大的两个大国，是否能够以足够负责任的方式行动，防止人工智能相关风险升级为影响所有人的全球性危机。也正因此，人工智能安全并非中美关系中的边缘议题。在这个充满不确定性的时代，它或许将成为检验中美之间是否仍然存在合作可能性的最重要议题之一。

作者：肖茜，清华大学战略与安全研究中心副主任

中美如何做大人工智能合作的蛋糕

孙成昊

回望过去一段时间，中美技术竞争如火如荼，然而双方也在人工智能合作方面取得了部分积极进展。2024 年 5 月中旬，双方在瑞士日内瓦举行首轮政府间人工智能对话，就人工智能科技风险、全球治理机制以及彼此关切的议题深入交流，标志着中美正式将人工智能议题纳入政府间对话。2024 年 11 月，中国国家主席习近平与美国前总统拜登会晤，就确保核武器始终处于人类控制之下达成重要共识，为人工智能军事化问题划定一条重要的“红线”，也为双边乃至全球层面技术风险管控与战略沟通奠定了基础。

特朗普执政后，尽管中美在人工智能领域激烈竞争，但在 2025 年釜山元首会晤后，两国均表示要在人工智能领域继续推进互利合作，显示出无论是民主党还是共和党政府，都认为有必要在人工智能这一议题上与中国保持接触与协作。

不可否认，当前中美政治氛围下，双方在人工智能与核融合（AI-nuclear nexus）方面实现更深层次、更具体合作依然面临诸多现实制约。核武器与其他关键战略军备具有高度敏感性，导致双方在短期内难以做到相互公开透明，而军备领域的核实（verification）困难与信任缺失进一步增加对话难度，但中美仍有必要最大程度探索在非敏感领域的合作。在涉军事问题上，中美可以尝试将重点放在如何推广、深化已有原则性共识上，而非立即开展深度军事对话或技术性合作。在非军事领域，中美则有必要在技术风险研判、伦理问题探讨与技术“向善”方面展开对话合作。

在军事领域，尤其是管控人工智能军事化应用风险方面，中美元首达成的原则性共识为下一步合作奠定重要基础。当前特朗普也正着力打造其“和平总统”形象，为中美在横向、纵向两个方面深化共识提供了契机。

在横向扩展上，中美可考虑往两个方向努力。一是把这一共识向其他核大国推广，包括英国、法国、俄罗斯等，逐步将双边共识转化为更具包容性和约束力的多边准则。同时，中美还可以考虑将这一议题纳入联合国等多边框架进行讨论，并在联合国体系内推动发表共同倡议或原则性声明，重申保持人类对于使用核武

器决定的这一核心原则，同时强调必须负责任部署和使用其他军用人工智能。二是中美可以考虑将这一原则进一步向更广泛的战略武器体系延伸，尝试将具备威慑力和巨大破坏力的常规战略武器纳入其中。但关于相关战略武器的定义和适用这一原则的范围需要双方进一步探讨，也可以先从二轨对话开始。

在纵向深化上，中美可在现有共识基础上探讨深化合作的具体路径。例如，在针对核武器系统的风险评估、管控机制上进一步细化，尝试将核武器系统和可能冲突场景拆分为具体部署、决策环节并评估潜在风险点，推动形成更细化、可操作的风险评估与管控框架。还需注意的是，全面禁止人工智能进入核指挥、控制和通信（NC3）系统并不现实，而且双方也需考虑人工智能与核武器体系结合可能带来的积极意义，而不是全面禁止二者融合。因此，如何确定更多彼此都能接受的“红线”变得尤为重要。

在此背景下，二轨外交可作为维系中美沟通的重要补充机制。双方智库、学术界及退役军方人士可以在非正式但制度化的对话平台上，就军用人工智能的风险评估、伦理准则以及危机情境中的决策等议题展开持续讨论。此类交流既能够在一定程度上摒除政治干扰，保持沟通延续性，也有助于为重启政府间磋商积累政策储备与专家共识。

在非军事领域，中美也可以共同推进人工智能全球治理。首先，中美双方可以聚焦人工智能带来的跨国界风险，并推动形成被更多国家接受的风险分级、分类与评估框架，并共同探索应对之策。无论是技术层面的智能体失控、大模型不可解释性和“幻觉”输出问题，还是应用层面的人工智能生物安全风险、人工智能“赋能”恐怖组织等非国家行为体、深度伪造对社会信任体系的侵蚀，这些都是中美共同面对的治理挑战。此外，从风险切入也有助于将中美从彼此防范的“技术竞争者”再定位为“风险共担者”，使中美即使处于“战略竞争”背景下，也能够在人工智能风险评估与管控领域找到共同关切。

其次，中美可以围绕人工智能对经济与社会结构的深刻影响开展对话。人工智能正在重塑劳资关系并改变传统生产模式，越来越多的行业呈现“无人化”和高度自动化倾向，这既可能提升生产效率，也可能造成结构性失业风险。如何确保人工智能提升人民生活，而不是加大社会资源分配差距，如何通过教育改革、

技能转型和数字基础设施普惠化等政策安排提升公众对技术变革的适应能力，是双方共同关注的问题。中美率先推进此类对话也能为全球其他国家提供思路参考。

再次，中美可以进一步探讨如何利用人工智能为全球公共治理提供新的解决方案。在气候治理领域，人工智能可用于灾害预测、极端天气模拟和应急响应；在国际和平与安全领域，人工智能技术有望被用于辅助冲突调解、维和部署优化与危机预警；在全球发展领域，通过推动开源技术和基础算法共享，也将有助于进一步缩小南北技术鸿沟。因此，中美可以依托现有联合国框架，将人工智能公共产品合作嵌入多边议程中，推动在气候监测、灾害管理、公共卫生预警系统、农业智能化、跨境传染病预测、基础教育数字化等领域的技术合作试点，使人工智能真正成为服务全球公共利益的工具。在这些领域开展合作不仅能够展现人工智能技术的正面价值，也有助于在两国“竞争激烈”的技术领域注入合作因素，改善两国关系。

最后，中美还应当共同应对人工智能带来的伦理与法律难题。人工智能技术的快速发展使得如何定义人类身份、保持人类主体性与价值感等基础性伦理问题变得尤为凸显。如何在技术高度自主化的时代保持人类对关键决策过程的有效掌控，为人工智能的研发与部署确立基本伦理原则，并在出现损害后明确责任主体并建立可追责机制，已成为全球共同关注的课题。围绕这些问题展开对话，有助于中美进一步增强互信，并探索、强化国际社会对技术发展的规范性共识。

中美在人工智能全球治理领域的合作并非意在构建某种“G2”模式，而是基于两国在技术创新能力、产业体系布局以及国际影响力而提出的要求。任何有效的全球层面技术治理都离不开中美参与。而人工智能作为典型的颠覆性技术，其发展不仅会深刻改变经济、社会与安全格局，也必然塑造未来国际秩序的规则基础和治理体系。如果中美能率先在风险认知、伦理原则和公共应用等领域迈出合作步伐，将具有大国带头的示范意义与引领作用。

作者：孙成昊，清华大学战略与安全研究中心副研究员

中美通向 AGI 的不同道路比较

鲁传颖

通用人工智能（Artificial General Intelligence，AGI）近年来重新成为人工智能研究与科技政策讨论中的核心议题。不同于以往围绕特定任务或应用场景展开的“弱人工智能”，AGI 通常被理解为具备跨领域学习、推理和适应能力的通用智能系统，其潜在影响被认为将超越单一产业或技术范畴，深刻重塑经济结构、社会治理乃至国际权力格局。正因如此，AGI 不再只是一个工程问题或学术问题，而逐渐演变为一个具有高度战略意义的政治经济学议题。

当前关于 AGI 的讨论，主要被两种主导性叙事所塑造。第一种是由技术社群主导的“技术可行性”叙事，强调从认知建模、学习机制等基础层面重新理解智能的本质；第二种是由产业界推动的“工程可行性”叙事，主张通过不断扩大模型规模、算力和数据，依托既有深度学习范式逐步逼近 AGI。这两种叙事在学界与产业界均拥有重要支持者，例如围绕“大模型是否能够通向 AGI”，杨立昆与戴密斯·哈萨比斯之间的分歧，便生动体现了这一争论。

然而，这两种看似对立的叙事实际上共享一个隐含前提：AGI 的发展主要是一个由技术内在逻辑与产业激励所驱动的过程，国家的作用更多体现在“支持”或“监管”层面。这一假设在弱人工智能阶段或许尚能成立，但在 AGI 这一高度不确定、投入巨大且外部性极强的领域中，却显得日益不足。AGI 面临的并非单一技术瓶颈，而是技术路径不确定、实现时间不可预测、社会影响难以评估等多重不确定性叠加。在此条件下，单纯依赖市场机制，不仅容易导致技术路径锁定和资源配置失衡，也难以应对由此产生的社会风险与政治压力。

正是在这一背景下，AGI 的发展必须被重新理解为一个战略问题。AGI 具有典型的公共品属性和战略属性，其研发过程需要长期、高风险、跨部门的投入协调，涉及算力、数据、人才、资本以及治理规则的系统整合。国家的关键作用，并不在于替代市场进行技术选择，而在于通过制度安排，在根本不确定性条件下，使技术探索在经济上可持续、在政治上可接受、在社会层面风险可控。

从这一视角出发，中美两国在 AGI 领域所呈现的路径差异，便具有了重要的

分析意义。作为当前全球人工智能发展的两大引领者，中美不仅在技术能力和投入规模上处于前列，更在制度结构、国家—市场关系以及技术治理理念上存在显著差异。这些差异并未停留在政策表述层面，而是逐步内化为不同的AGI推进方式，形成了两种具有代表性的制度化发展道路。

美方早在2016年奥巴马政府时期发布的《为人工智能的未来做好准备》就开始系统性布局AGI。特朗普第一任期《保持美国在人工智能领域的领导地位》行政令和《国家人工智能研发战略计划2019更新版》进一步强化了在AGI领域的技术领先战略。2025年1月《推进美国在人工智能基础设施领域领导地位的行政命令》以“星际之门”计划为抓手，集中推进超大规模数据中心和算力枢纽建设，意在为下一代通用模型提供持续扩展的算力底座。2025年7月密集发布的四项行政令构成了一个完整的政策组合，从《赢得竞赛：人工智能行动计划》到阻止“觉醒式”AI、加速数据中心建设、推动技术出口，既要确保前沿模型的突破性进展，又要通过基础设施投资和出口管制维持技术代差优势。11月发布的《启动创世纪任务》将前沿模型直接嵌入能源、材料、气候与生命科学等国家级科研工程，意图通过跨学科算力集成和模型协同应用，加快通用能力向科学发现与工程突破转化。这种布局高度依赖少数头部企业的技术能力，期待通过技术领先建立全球标准。

美国当前呈现出的，可以概括为一种“模型中心道路”。在这一模式下，AGI被视为技术创新的终极目标，制度性协调的核心在于集中资源推动模型能力的持续突破，并通过技术领先塑造事实性全球标准。政府在其中主要通过战略规划、基础研究资助和安全监管发挥作用，而具体的技术探索和工程实现，则高度依赖少数大型科技企业。这种模式在短期内具有较强的突破潜力和国际竞争优势，但其内在风险也同样突出：技术路径高度集中，容易形成锁定；算力与数据资源高度集中，放大伦理与安全争议；一旦关键技术路线遭遇瓶颈，其调整空间相对有限。

中国的战略文件没有明确提出AGI这个概念，并在很多做法上与美国有异同之处。相同之处，都是重视人工智能的技术和产业发展，不同之处在于中国实现AGI之路更多是通过应用创新来引领发展，而非完全押注前沿模型能力提升抵达AGI这条路。2017年《新一代人工智能发展规划》奠定了技术、产业、治理三位一体的基本框架，此后的政策演进始终围绕这一框架展开。2022至2025年间发布的一系列应用导向政策显示出明确的基础设施化取向，通过场景开放、算力网

络建设和跨部门协调，推动智能能力向实体经济和社会治理领域的系统性渗透。2025 年 7 月发布的《人工智能全球治理行动计划》和 2026 年 1 月的《人工智能 + 制造专项行动实施意见》，表明中国正在将 AGI 能力的培育嵌入到产业升级和全球治理议程之中。

这过程中，中国逐渐显现出一种“通用基础设施道路”。在这一模式下，AGI 并非被理解为单一的技术终点，而是被视为一种需要嵌入国家发展体系的基础性能力。这一模式不在于押注某一条技术路线的极限突破，而在于通过算力网络、数据体系、应用场景和组织机制的系统整合，推动智能能力在更广泛的经济和社会系统中扩散与积累。国家在其中扮演着更为主动的统筹者角色，通过基础设施建设、场景开放和跨部门协调，为多样化技术路径提供生存空间。

这两种道路并非简单的“国家主导”与“市场主导”之分，而是对同一问题的不同回应方式：在技术高度不确定、市场容易失灵的条件下，如何通过制度安排维持长期探索的可行性。美国的模型中心道路，更强调通过市场化竞争和技术集中来加速突破，其优势在于创新效率和全球影响力，但风险集中于技术与治理层面的不稳定性；中国的通用基础设施道路，则更强调通过系统整合和能力扩散来降低整体风险，其优势在于稳定性和可控性，但也面临协调成本上升和突破速度不确定的问题。

重要的是，这两种道路并不存在简单的优劣之分。它们反映的是不同国家在制度条件、发展阶段和战略目标约束下，对 AGI 这一高度不确定技术所作出的理性选择。AGI 的实现路径并非唯一，其社会嵌入方式、风险分布结构以及对国家能力的反向塑造，都将随着制度性协调方式的不同而显著分化。从更广阔的视角看，中美在 AGI 领域的分化，不仅关乎技术竞争本身，也预示着未来通用技术如何被纳入国家治理体系的不同可能性。AGI 最终是否实现，固然仍有巨大的不确定性，但可以确定的是，其探索过程已经成为检验国家制度协调能力的重要场域。理解中美通向 AGI 的不同道路，不仅有助于把握人工智能发展的现实走向，也为思考前沿技术与国家制度之间的关系提供了一个关键窗口。

作者：鲁传颖，同济大学政治与国际关系学院教授、副院长，上海市人工智能与社会发展研究会副会长

人工智能军事应用的中美竞合

祁昊天

新竞技场

人工智能（AI）正迅速成为中美军事关系中不可或缺的一环，被两国视为塑造未来军事实力格局的关键因素。美国强调加速 AI 技术在国防军事领域的应用和部署，以在未来战场保持优势地位。中国同样高度重视人工智能在国防中的作用。习主席多次强调加速推进“智能化”在军队建设中的应用，明确要求人民军队在 2035 年基本实现现代化、本世纪中叶建成世界一流军队，而 AI 等前沿技术被置于这一战略目标的重要地位。人工智能被两国纳入各自军事现代化战略的核心，事实上已构成中美军事博弈的全新竞技场。

然而，AI 在中美军事关系中扮演的角色并非单一的“冲突催化剂”。在两国都面临国防能力现代化需求的背景下，AI 对两国军事关系的影响呈现出多重面向。AI 军事应用充满复杂性，既能够在局部推动革命性的军事效能和作战方式的调整，也在兑现赋能预期方面充满挑战、不足和不确定性，并连带产生足以破坏作战效果、国家安全和战略稳定的风险与威胁。AI 应用和部署于军事领域，既是推动作战效能提升的利器，成为双方竞争领域，也能够促使双方正视共同面临的安全风险。因此，我们有必要在竞争与合作的双重视角下，对中美军事 AI 关系进行审视。此外，作为全球 AI 与军事科技举足轻重的两个大国，中美有必要共同承担责任，应对 AI 军事化所带来的结构性安全挑战，在相关治理领域寻求对话协调机制，防止技术及应用失控损害国家安全、扰乱地区乃至全球稳定。

不同层级的赋能与风险

AI 具有提升军事效能和引发多重风险这两方面的影响，如同硬币的两面，贯穿于上述各个层级、不同领域。在赋能方面，AI 可以帮助后勤、训练、指挥等领域实现自动化与效能优化，在直接作战和战场管理方面，AI 结合无人平台的出现有望至少在局部改变未来战争的范式。但与此同时，AI 引入军事领域会带来多重

潜在风险。其一，AI 可能压缩决策时间、放大冲突升级概率。其二，当前以神经网络和数据为基础的 AI 系统因其算法不透明性以及数据的不可靠性，可能诱发人为误判，导致地缘冲突或人道悲剧。更隐蔽的是，AI 被利用于信息战和认知战，通过生成海量虚假音视频或新闻，扰乱对方国内舆论和决策判断。在中美高度警惕彼此意图的背景下，这种信息层面的误导可能对战略稳定造成额外冲击。其三，军备智能化引发的竞争螺旋升级也会加剧，AI 竞赛无疑已成为当前两国军事安全竞争的新推手和新形态，双方在缺乏互信和有效沟通渠道的情况下，都担心在军事智能化上落后于对方，于是加速投入军备和防御措施。虽然美军高层试图释放克制信号，避免陷入失控军备竞赛，中国方面也反复强调“不与其他国家开展人工智能军备竞赛”。但是在客观效果上，两国显然都不愿也不会在 AI 军事应用上示弱，一方的新部署往往被另一方视作压力，从而推动其投入对等或更多资源追赶，这种局面可能令两国陷入“赢了竞赛却输了安全”的困局。这种困局并非“安全困境”，潜在的负面影响则大于安全困境，因为双方的出发点均非“满足现状”而是寻求通过新的技术可能性寻求有利于己方的改变。

具体而言，两国的 AI 军事应用在由战术到战略层面的部署存在不同的影响。在技战术层面，AI 已广泛融入两国武器装备和支援平台与系统。AI 的优势直接体现在战场管理微观层级的敏捷与精确上，以及战场之外的建设与管理，例如更加实时的情报处理、更自主的无人系统以及战场感知与火力打击链的加速闭合，能够提高感知、决策和打击效率，以及后勤物资管理和人员演训效率。这构成了中美竞争最直接也最激烈的领域之一，双方都希望通过 AI 赋能使己方士兵和武器比对手更“聪明”、更迅速、更高效。不过这一层级应用部署的影响范围也是相对有限的，无论是赋能还是失控风险方面，虽然存在出现连带效应和升级的可能。

在作战行动和战区层面，AI的影响开始放大，对中美军事态势的塑造更加显著。双方都在探索以 AI 优化指挥控制和战场管理，例如通过智能算法整合多源情报、辅助指挥官决策，从而在较大规模作战行动和战区层级形成具有“联合”和“全域”特征的感知与快速响应能力，实现跨军兵种、跨地域、跨作战域的信息融合与火力分配。这一层面的 AI 应用对中美军事关系影响的风险性也在提升，AI 赋能的快速决策和联合打击能力意味着双方在潜在对抗中的攻防博弈更趋激烈，敏感领域

和地域出现局面失控与升级的可能性提高，如果缺乏有效的沟通管控机制，高速运转的智能化战场可能超出人类指挥员的掌控，或“诱导”人类指挥员，令偶发事件迅速升级为难以遏制的军事对抗。AI 不仅是力量“倍增器”，更可能成为风险“放大器”。

在战略层面，人工智能对中美军事关系的影响更趋深远和复杂，既关系到战略稳定、核威慑稳定，也涉及全球安全格局的走向。首先，AI 可能冲击战略威慑与稳定。例如，智能化侦察与打击技术的发展，使得强者探测对方战略威慑力量并实施有效打击的可能性增大，削弱对方二次核打击能力成为一种可能和诱惑，而相对弱的一方如判断自身二次打击能力的生存性存疑，至少在理论上面对“use it or lose it”的两难，无论哪种机制发挥作用，都可能削弱双方的威慑平衡，“需要类似“不首先使用核武器”这样的原则、“互不首先使用”倡议之类的机制对稳定提供进一步保障。此外，AI 介入战略决策可能引发误判升级。在高度紧张的战略博弈中，如果决策者过度依赖 AI 提供的信息与建议，或 AI 支持的机构或个人，而这些算法存在偏见或不透明性，所依靠的训练数据存在缺陷，又可能出现恶意操纵导致的“深度伪造”、网络攻击等问题，那么战略层面的误判风险将大大上升。因此，在战略层次上，中美都对 AI 保持着警惕，包括强调确保核武器始终置于绝对安全和可靠、由人控制的状态。不过，对中美而言，最复杂的课题在于，AI 在战略层面的介入并非简单的“掌握按钮”或“替代决策”，而是 AI 在复杂的人机环中以各种方式、在各个环节实现了广泛的“嵌套”。像中美这样的大国所要管理的并非简单的某些 AI 赋能战略工具，而是庞大的复杂系统。总之，在战略领域，AI 既被两国视作确保未来战略优势的必争高地，同时也因其可能撼动战略稳定而应当引发双方对失控以及更复杂情形负面后果的担忧，前者推动竞争，而后者则构成双方合作管控的前提。

竞合交织态势

通过以上讨论，可以发现当前中美军事 AI 关系呈现出竞合并存交织的特征。一方面，竞争态势激烈。无论是在前沿技术研发、战术应用、作战概念创新还是体系性的现代化适应上，中美均视对方为主要竞逐对象。双方都将 AI 视为赢得未

来军事优势的关键，投入巨资、聚集力量，力求在 AI 技术突破和军事转化上跑在前列。在争夺 AI 军事优势的问题上，中美谁都不愿也不会主动让步，形成了典型的零和博弈状态。

另一方面，合作与风险管控的需求同样突出。特别是随着 AI 技术不断渗入战略领域，两国都有动机避免陷入安全闸门缺位或失效的危险状态。近年来，中美在一轨和二轨层面都开始探索就 AI 风险管控进行对话沟通的可能，两国也分别作出了管控高风险应用的动作。

在美国方面，相关约束已通过《美国国防部第 3000.09 号指令：武器系统自主性》实现一定程度的制度化。该指令要求自主和半自主武器系统在设计上必须确保指挥员和操作员能够对武力使用作出适当程度的人类判断，并要求某些自主武器在正式立项研发和部署列装之前接受高层审查。华盛顿发布的《负责任军事使用人工智能与自主性的政治宣言》则呼吁以负责任的方式使用军事人工智能，遵守国际法，并设置相应护栏，以减少非预期偏见、事故和非预期升级。

中国近来的相关努力采取了类似但在一定程度上更具明确规制色彩的形式。中方一再主张，各国尤其是大国应以审慎和负责任的态度对待人工智能军事开发与使用，确保此类系统安全、可靠、可控并始终处于人类控制之下，避免误用、滥用和无序扩散。中国还在多边场合积极推动这些理念，倡导在联合国框架下制定广泛接受的军事人工智能治理规则。中方已发布《全球人工智能治理倡议》和《关于规范人工智能军事应用的立场文件》，明确主张各国尤其是大国应以审慎和负责任的态度对待军事人工智能的研发和使用。中国强调，军事人工智能治理应以人类命运共同体理念为引领，坚持发展与安全并重，并防止军备竞赛。这些立场在精神上并不与美国倡导的“负责任人工智能”原则根本冲突，二者都体现了对人工智能安全与伦理问题的关注。习主席与拜登总统在 2024 年 12 月亚太经合组织领导人会议期间两国元首声明中明确强调，必须由人而不是人工智能或机器保持对核武器的控制，这给未来相关管控合作推进奠定了重要的政治基础。

中美在某些风险管控议题上存在进一步的协作与合作空间，例如建立沟通热线和预防机制以防 AI 引发误判危机，甚至在未来实现局部有限分享测试验证和算法安全相关信息，加强局部双向透明度，以减少重大误判风险。随着 AI 技术继续

发展，中美在管理和治理军事领域失控风险、人为误判和冒进风险、伦理规范合作等方面都存在协调甚至合作的必要。不过也需要注意的是，对共同风险的理性认知并不足以完全对冲或消除因互疑而互塑的消极行动循环。

竞合并存的态势给两国军事安全关系和国际安全治理都提出了新的课题——如何在竞争必要且必然存在的前提下，避免完全失控和非必要对抗的出现，并寻求将有限的合作共识转化为实质的风险管控。对于中美而言，这不仅是两军之间的问题，更是覆盖 AI 技术发展全生命周期、全应用部署网络的多领域和多部门协调问题，不仅是高政治和一线政策问题，也是两国二轨外交和学术科研交流协作的问题。

共同责任与治理展望

作为全球 AI 和军事科技的领先国家，中美在军事 AI 领域的竞合关系将对 21 世纪的国际安全产生深远影响。这不仅事关两国自身的战略利益，也攸关地区乃至世界的和平稳定。历史经验表明，新技术的军事化往往早于相应规则的建立，因此大国有责任超前思考并引领治理框架的塑造。随着 AI 以前所未有的速度、广度和深度融入军事领域，中美都无法独善其身、也无法通过单边行动有效应对 AI 带来的结构性安全挑战。正如中国军控白皮书中指出的，外空、网络、人工智能等新兴领域是全球治理的新疆域，需要各国特别是主要大国共同参与制定广泛共识的治理框架。

这意味着，中美应超越零和博弈，开展必要的对话、协调乃至合作。首先，两国应建立高层对话与沟通机制，借鉴美苏在核军控领域建立热线、签署协议的经验，中美可以探讨在 AI 军事应用领域设立常态化的沟通渠道，讨论签订行为守则或建立危机联络机制，在 AI 相关事故或异常情况发生时及时沟通，防止局势升级。值得注意的是，这类机制如果有可能建立并具备可行性，其任务应尽量收窄：不是“解决危机”，而是“核实某一与人工智能相关的危险异常究竟是真实存在、局部发生、功能降级、遭到伪装，还是仍在扩散”。例如，这类渠道应当主要服务于异常澄清和事件范围界定，而不是在实时状态下裁定责任归属，其目的在于防止因不确定性而引发螺旋式升级。它应当包括一套可在人工智能关联事件发生

时迅速交换的最低限度信息模板，例如：事件发生的时间窗口、受影响部门、对归因的初步置信度、潜在伪装迹象，以及为遏制扩散已采取的措施等。

第二，中美应在多边场合发挥建设性作用，支持联合国、五常、五常 + 或其他有效平台与框架，推动国际规则与标准的讨论甚至制定。这种工作的起点往往很低，政治阻力巨大，协调成本很高，但即便是诸如共同支持诸如“某些 AI 赋能武器或军事系统只有在特定条件下以特定方式才能够使用和部署”这类最低原则，即便缺乏完整的落地对标性，鉴于 AI 竞赛的全球扩散风险，中美仍然有责任携手推动相关议程的起步。

第三，中美应鼓励在官方层面之外持续加强二轨交流与合作，允许和推动两国学术、智库和广大的认知共同体开展联合研究、提出行动倡议、影响政策闭环，就 AI 对军事安全和战略稳定的影响建立基础知识体系、拿出可行政策建议。事实上，双方专家已通过不同平台进行了大量相关的二轨对话，并产生了实际政策影响。如何进一步有效打通部门壁垒，将相关成果纳入具体的军事安全决策和能力建设环路，应是下一步目标。

中美在 AI 军事应用领域既是竞争对手，也是面对共同风险、承担共同责任的利益攸关方。双方既要避免陷入螺旋竞争，也应在共同关切的问题上开展对话和协调，甚至将军事 AI 视为与核武器和战略稳定同等重要的双边安全对话核心议题。如何在追求军事效能、确保国防能力建设、军事安全利益与防范安全风险之间取得平衡，已成为中美两国两军共同面临的战略课题。在推动 AI 军事创新、维护自身军事安全利益的同时，中美两国需要拿出大国担当，共同塑造安全框架与可落地、因地制宜的规范规则体系，为相关地区乃至全球的长治久安做出贡献。这不仅符合中美双方的长远利益，也是对人类命运负责的必然要求。

作者：祁昊天，北京大学国际关系学院长聘副教授、国际安全与和平研究中心副主任

AI虚假信息领域的中美合作可能

董　汀

观察中美 AI 对话久了，会发现一个现象：围绕 AI 安全的议题这两年其实一直在扩展，从军用 AI 到前沿模型风险议题，从生物安全到网络安全，谈的并不算少。但是虚假信息几乎没有进入讨论的范畴，更谈不上合作。在既有的国际讨论中，虚假信息往往以指责的方式出现，我想讨论的是，它是否可以从“指责的对象”变成“可以共同处理的问题”。

今天的虚假信息有一个容易被忽略的特点。它的伤害往往发生在真相大白之前。一段来源不明的视频、一张出处不明的截图，短短几个小时就能在舆论场传播开来。等到事实终于被核实清楚，外交指责、舆论愤怒和平台处置都早已发生。

这类压力对新闻业是家常便饭。如果等上几个小时把事实理清楚，热点可能早已过去；如果不等，又可能造成错误传播。但是如今，这种困境正越来越多地外溢到外交回应和决策决断的环节。

中美当前如此高度互不信任的关系，会把“等不起”的状态进一步放大。低信任环境的一个特点是，任何一件事情弄清楚之前，不惮以最坏的恶意来作出解释。例如一段来源不明的视频，几个小时之内就可能被解读为是对方政府授意的有组织行为。人工智能时代，谣言的传播速度和误判合理化的速度都被极大放大。

值得欣慰的是，这并不意味着合作就毫无可能。合作完全可以基于相对朴素的共同认知。既然合成内容能在数小时内制造出一场危机叙事，任何一方都可能成为受害者，双方需要处理的，首先是如何稳定，而不是解释真相。

冷战时期美苏之间的“热线”，可以作为一个有用的参照。古巴导弹危机期间，美苏之间外交电报的传递动辄要花几个小时，赫鲁晓夫给肯尼迪的一份关键信件被接收后用了将近十二个小时才完成解码。双方都认为这种通讯延迟构成了重大风险。1963 年 6 月，两国在危机过后签署了谅解备忘录，建立了直接通信链路。1967 年中东六日战争期间，这条热线被首次正式启用。1971 年印巴战争和 1973 年赎罪日战争中也都曾使用过。事实上，热线并未消解冷战中的任何根本矛盾，但是它在多次紧要关头，为双方领导人提供了几个小时不必立即反应的时间，从而避免了因误判而导致的不可逆后果。这种“为核实争取时间”的思路，也许

正是今天 AI 虚假信息领域合作可以借鉴的。

除了争取时间以外，还有两个常被忽视的议题也同样可以合作。一是如果缩短信息生成与核实事实之间的时间差。如今，生成一段足以乱真的视频只需要分钟级，专业取证却需要数天甚至数周。在汹涌的舆论压力下，要求一国领导人向公众说“我们再等等调查结果”，政治上几乎不可行，仅靠提高公众的媒介素养也无济于事。

另一个议题更为隐蔽。未来的舆论影响活动，可能主要不再针对人，而是针对机器。当全球公众越来越习惯通过大模型和智能助手去理解中美关系、台海局势和贸易摩擦时，今天互联网上散布的低质量内容，就可能以“权威背景资料”的形式，出现在若干年后 AI 助手输出的答案中。

处理虚假信息的传统办法主要是删除和标记，但是生成的速度会随着模型能力提升而越来越快，删不胜删。因此更有效的方法或许是开展基础公共信息建设，让一个国家的权威资料和一手记录等，以机器可读的形式留在公共网络空间，让人和机器都容易搜索到。

具体可以开始合作的事情不需要宏大，或许有三个层面能够尝试。第一，在现有双边危机沟通机制中，增设一项专门针对时效性敏感信息事件的处置安排，让双方在涉及 AI 生成内容的重大事件中，能够先有几个小时的相互沟通核实的时间，再做出政策反应。第二，推动两国在内容来源标识技术规范上的互通，使一段生成影像所携带的来源标识在跨越国境后依然能被对方读取，如此一来，无论由哪一方的 AI 工具生成，对方都能查清它从何而来，以及是否被修改过。第三，双方各自加大向国际公共信息空间提供本国权威资料的力度，让主流大模型在回答中美关键议题时至少有可信的原始来源可引用。这三件事都不需要任何一方放下既有分歧，因为它们回应的是一个非常现实的问题，即当信息环境中缺少基本的核实能力时，最先失去反应时间的会是政府本身。

中美之间的结构性分歧不会消失，AI 领域的竞争也或将长期存在。然而即便如此，双方至少仍有一项共同利益，那就是不能让全球公共信息环境变成一个人人都来不及建立独立判断，政府也来不及修正的空间。

作者：董汀，清华大学战略与安全研究中心副研究员

CISS 人工智能项目简介

清华大学战略与安全研究中心（CISS）成立于2018年11月7日，是聚焦国际战略与安全领域研究的高校智库。自2019年以来，CISS聚焦人工智能技术发展前沿与国际安全治理问题，专门设立人工智能项目专家组，扎实推进人工智能与国际安全相关研究。同时，CISS持续与美国布鲁金斯学会、瑞士人道主义对话中心开展中美、中欧人工智能与国际安全二轨对话，不断拓展同联合国裁军研究所、红十字国际委员会等国际组织和智库机构在人工智能全球治理方面的项目合作，并通过联合研究和政策交流，形成一系列重要的研究成果和政策报告，为推动人工智能国际安全领域的交流合作积累国际共识。此外，CISS还积极承接来自外交部、科技部、财政部等国家部委委托的研究项目，持续深化人工智能治理的区域与国别研究，为相关政策制定与国际合作贡献力量。

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

ciss@tsinghua.edu.cn
010-62771388
清华大学明理楼 428A 室
http://ciss.tsinghua.edu.cn

微信公众号　官方网站　联系我们