

2026年人工智能与国际安全研究

非国家行为体滥用人工智能 及其对国际安全的影响

2026年3月



清华大学战略与安全研究中心

CENTER FOR
INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

引言

近年来，以大模型为代表的前沿人工智能技术快速发展，正在深刻改变技术能力和专业知识的获取方式。然而，人工智能技术既是推动全球经济社会发展的重要力量，也可能成为新的风险放大器。如果前沿人工智能技术被恐怖组织、跨国犯罪集团等非国家行为体滥用，将会给国际社会的和平与安全带来严重的风险挑战。在这一背景下，如何在促进技术创新与防范非国家行为体滥用人工智能风险之间取得平衡，已成为全球政策界和学术界关注的重要议题。

2026年2月，清华大学战略与安全研究中心（CISS）在慕尼黑分别与布鲁金斯学会（Brookings）、人道主义对话中心（HD）开展人工智能与国际安全对话，非国家行为体滥用人工智能的风险及其治理都是交流的重要议题之一。基于此，本报告重点探讨了非国家行为体滥用人工智能的可能路径，带来的相关风险及其对国际安全的影响，并分析如何从国家层面和国际合作层面防范非国家行为体滥用人工智能的安全风险。希望本报告能够为理解这一复杂议题提供参考，并为相关政策讨论贡献有益思考。

目 录

| | |
|--------------------------------------|-----------|
| 非国家行为体滥用人工智能对国际安全的影响 | 01 |
| ◎鲁传颖 | |
| 非国家行为体滥用前沿人工智能的风险和认知分歧 | 05 |
| ◎肖 茜 | |
| 恐怖分子滥用高能力大语言模型的国际安全影响 | 10 |
| ◎孙成昊 | |
| 降低非国家行为体滥用人工智能带来风险的国内方法 | 19 |
| ◎李 强 | |
| 缓解非国家行为体引发的人工智能风险:背景、可行性与共同责任 | 26 |
| ◎祁昊天 | |
| 中美能否遏制非国家行为体滥用人工智能的风险? | 30 |
| ◎郑乐锋 | |

非国家行为体滥用人工智能对国际安全的影响

鲁传颖

一、非国家行为体滥用人工智能的国际安全风险

人工智能（AI）技术的可及性、通用性与自主进化能力不断提升直接为非国家行为体滥用 AI 创造了核心条件。非国家行为体无需掌握顶尖研发技术，便可借助开源模型、商用 API 等低门槛渠道，轻松获得自动化攻击、智能规避防御措施、低成本放大危害影响的能力。这种趋势具体表现为 AI 自身的内生脆弱性被不断挖掘、AI 负向应用的场景持续增加、AI 技术引发的安全治理制度失灵，国际安全构成的持续冲击。

非国家行为体滥用 AI 一方面会放大并激化各领域已存在的应用安全风险，降低传统风险的触发门槛；另一方面，AI 的内生特性与技术潜力会催生以往不存在的全新安全风险，突破传统安全治理边界，对国际安全构成多元威胁。

一方面，AI 会激化已有的应用安全风险。AI 可提高专业知识的可及性，使行为体更易获得或研发有害安全的武器、材料与设施。例如 AI 可通过赋能各行为体低成本和隐蔽地开发、运输与部署核及其他常规大规模杀伤武器（如导弹）、配套材料和设施等，实现军备扩散。而在生物安全领域，行为体可利用 AI 技术提升生物技术专业知识可及性，如通过绕开大语言模型安全防护可获取生物威胁原材料与关键信息或强化 / 改造生物因子的负面特征，开展生物恐怖主义或生物犯罪。

非国家行为体可利用 AI 升级既有的攻击手段，使攻击更为低廉，效果更加显著。在网络安全领域，AI 可降低大规模 DDoS 攻击的制作和投放成本、提高其精准度、实现生成自动化、实现零代码攻击等。同样地，在认知安全领域，借助 AI，不仅可以快速生产高质量的虚假信息与深度伪造内容，还能通过机器学习，可以更为精准地实现虚假信息的精准推送，加速形成“信息茧房”，在短期内可导致大规模认知偏见和极化，方便认知操纵和群体性认知欺诈。

AI 对相关应用领域的赋能也暴露了安全系统的脆弱性，例如在核及常规军事安全领域中，非国家团体可提前向 NC3 或其他打击系统内的 AI 模型中植入对抗

样本和恶意后门，从而掌握部分参数的调控权，以此作为勒索和威胁国家的筹码，成为新式的恐怖攻击类型之一。同样地，在认知安全领域中，AI技术可活用人类“信息过载”后容易诉诸于情绪化叙事的认知惯性，运用海量虚假信息，引导人群放弃对信息真实性的求证，转而依赖知觉、情绪判断和AI本身，削弱群体的认知能力、心理防线乃至凝聚力。

非国家行为体对AI的滥用也会对国际安全治理制度构成多重冲击，表现为挑战国际规范、模糊责任链条、淡化合作基础、加速技术逐底竞赛，从而削弱治理机制效能并使其逐步弱化。

第一，挑战国际规范与伦理共识，削弱安全治理机制的适应能力：非国家行为体通过AI技术的模块化与平台化扩散，使之迅速在治理机制外获取类国家能力，能以去中心化和非正式方式实施跨境行为。这使得既有安全规范在责任划分、适用对象和风险分级等方面滞后于问题现状，治理原则和伦理规范难以及时回应新型风险，削弱安全治理机制对技术实践的引导与约束功能。

第二，模糊化责任链，破坏规制机制的问责与惩罚功能：非国家行为体滥用的AI技术往往来源于多元主体的组合，如开源模型、商业平台、匿名化社群、黑市等。治理机制无法及时明确不同环节的责任归属，使得非国家行为体的滥用行为很难及时担负治理成本、被有效制裁与被惩戒，从而削弱了问责机制的威慑力，放大安全治理机制的缺陷。

第三，淡化合作基础，削弱各主体通过信息共享弥合能力不对称的可能性：非国家行为体可以通过AI自动化分析治理机制的公开信息，通过反向工程，得出绕开机制的约束方法和风险预防盲区；也能利用全球缺乏有效溯源AI模型输出的技术鸿沟，发起假旗行动，破坏合作互信基础；更使各国、企业与社群担忧以“能力建设”为目标的治理方法反而导致非国家行为体滥用AI的活跃，从而加剧现有的AI能力鸿沟。

第四，加速AI逐底竞赛，使既有安全治理机制无效化：非国家行为体滥用AI所制造的不确定性反而成为国家层面技术竞逐的压力来源，各主体倾向于以进攻性逻辑应对安全风险，通过加速技术研发、放款应用约束和降低伦理门槛来维持相对优势，推动安全治理目标从风险控制转向竞争，从而使得安全治理机制空心化。

二、为什么非国家行为体可以滥用人工智能

非国家行为体能够滥用人工智能，核心源于 AI 技术扩散与治理的失衡、安全治理中的赋能 - 嵌入两难，以及其可外部化风险、低成本高收益的激励结构，为滥用行为提供了可乘之机。

非国家行为体得益于 AI 加剧的归因难题，很少承担 AI 安全治理成本，从而常能外部化 AI 滥用风险，呈现低成本 - 高收益的激励结构。首先，非国家行为体可通过免费搭便车的方式，享受全球 AI 技术民主化和开源生态带来的治理红利，却可将滥用产生的负外部性强制转嫁给国际社会。当非国家行为体利用 AI 引发社会信任危机、金融市场动荡或网络瘫痪时，由此产生的系统性风险和重建成本全部由受害国政府和全球治理机制承担，使滥用行为具有吸引力。

其次，非国家行为体可利用 AI 的溯源模糊性，无需担忧对等报复，由此在 AI 实验或攻击中展先极高的风险偏好，并倾向于测试国家与国际机制出于国际声誉和利益期望不敢触及的禁区，从而引发非国家行为体滥用的极端化倾向。

最后，非国家行为体可在治理能力相对较弱的发展中国家与区域发展，利用当地的治理工具与经验欠缺，规避国际机制的溯源与追责。

三、如何规制非国家行为体人工智能滥用行为

为有效防范非国家行为体滥用人工智能引发的各类安全风险，破解现有治理滞后、协同不足等困境，亟需构建系统性的规制体系。具体可通过搭建韧性包容的协同治理框架、完善本土化与标准化兼顾的国际治理实验机制、建立赋权导向的国际能力建设支持框架，凝聚全球共识、补齐 AI 治理短板。

（一）构建一个具有韧性、适应性和包容性的协同治理框架

其一，应将非国家行为体纳入治理对象范畴，突破将非国家行为体视为外部威胁的治理惯性，依据组织形态、技术能力、活动领域与潜在风险水平，构建针对性风险识别机制。一方面便于实施差异化的监管义务与治理参与路径，另一方面为具备治理合作意愿的行为体提供制度接入渠道，使其参与风险标准、测试与规范共建。

其二，建立常态化的风险信息共享与协同响应机制，通过联合开展前沿 AI 模

型测试评估、共享风险研究成果、推动评估标准的国际互认、记录 AI 系统风险的关键案例评判，从而保证问责透明，能积极面对突发跨境的 AI 滥用风险。

其三，构建具有约束力的全球治理机制与履约保障体系，可借鉴《蒙特利尔议定书》在臭氧层保护方面的经验，制定相关安全公约，明确开源模型安全审查和算法透明度方面的最低标准，并设立定期评估和资助各国的基金会与条款。

(二) 构建统筹本土化与差异化、标准化与场景化的国际治理实验机制

其一，构建具有跨文化兼容的伦理标准与技术互操作体系，形成可操作的技术解决方案，提供“价值中性”的基础安全协议，鼓励各治理主体以此为基础，根据自身风险结构、技术生态与社会韧性开展本土化制度实验。

其二，搭建围绕高风险应用场景的跨域治理沙盒，对关键技术应用领域进行情景化测试，并设立相关国际机构，使统一国际标准在真实或虚拟场景中接受检验、细化与修正，并鼓励治理相关方将重要沙盒检验成果共享，方便日后治理主体调用。

其三，通过建设多层、多区域的试点网络，将分散于不同国家、平台与领域的治理经验进行系统性汇聚与比较，形成“实验-评估-标准化”的动态正反馈循环。

(三) 构建以赋权为导向、兼顾包容性的国际能力建设支持框架

其一，搭建开放式的全球人工智能安全知识共享网络，弭平各国间的知识赤字，尤其是帮助发展中国家获取辨别 AI 滥用行为的安全知识。

其二，建立多元化的技术援助与资源共享机制，推动构建全球 AI 安全技术共享池，以自愿贡献非敏感的安全技术，形成全球公共产品。

其三，构建尊重自主权的赋权型合作关系，帮助发展中国家的内生治理能力，优先解决数字基础设施、数据主权保护与本土语言模型开发的核心关切。

作者：鲁传颖，清华大学战略与安全研究中心特约专家，同济大学政治与国际关系学院副院长、教授。

非国家行为体滥用前沿人工智能的风险和认知分歧

肖 茜

先进或前沿人工智能模型被非国家行为体滥用所带来的安全风险，已不再停留于理论层面，而是日益呈现出可观察的现实迹象。前沿人工智能不仅能力更强，而且传播更快、获取门槛更低、规模扩展能力更高。因此，相关的安全风险正在发生结构性转移，焦点成为谁能够操纵信息，使一些扰乱行为自动化，并在危机时刻加以利用。

一、非国家行为体滥用人工智能的主要路径

目前来看，非国家行为体可能在至少四个领域滥用先进人工智能，分别是网络攻击，虚假信息与认知操控，化学、生物、放射性和核（CBRN）相关活动以及人工智能赋能的自主或半自主系统。

《2026 年国际人工智能安全报告》指出了过去一年两项值得关注的发展：

第一，人工智能在科学能力方面取得较大进步，加剧了外界对其可能被用于生物武器开发的担忧。多家人工智能公司在 2025 年发布新模型时，在部署前测试阶段无法完全排除模型可能显著协助新手开发生物武器的风险，因此选择增加额外的安全防护机制。第二，越来越多的证据显示，人工智能系统已被用于现实世界的网络攻击。多家人工智能公司的安全分析表明，恶意行为者正在利用人工智能工具辅助其网络攻击行为^①。

需要指出的是，报告更多是基于科学风险分析提出潜在威胁路径，而非详述已被非国家行为体实施的具体案例。这反映出相关领域仍处于人工智能滥用的实证证据早期阶段，但也凸显前瞻性风险评估的重要性。

^① Yoshua Benjio. 2026. The 2026 International AI Safety Report[EB/OL]. 2026-02-03. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>

二、大模型辅助下的化学与生物武器化风险

在 CBRN 领域，人工智能风险并非意味着非国家行为体能在一夜之间获得新型武器，而在于先进模型可以浓缩专业知识、整合分散信息，从而降低非国家行为体进行危险实验的技术门槛。

2023 年至 2024 年间，多家前沿人工智能开发商开展了正式的红队测试与评估，检验大型语言模型能否协助非专业用户完成与 CBRN 相关的任务。其中一个广泛引用的案例来自 OpenAI 与学术专家合作开展的测试^②。在相关评估中，研究团队重点检验大型语言模型能否系统整合公开来源的分散化学信息，针对技术细节不断深化的递进式问题进行回应，并在理解有毒化学品性质、合成机制及可能部署场景方面，显著减少认知成本，降低专业技能门槛。

关键在于，模型并未被要求创造新武器，也未接触任何机密数据。相关风险并不在于模型接触了敏感的原始数据，而在于对分散信息进行综合、重组并赋予语境意义的能力。

结果显示，在缺乏防护措施的情况下，相比传统检索，模型可帮助非专业用户更快理解复杂的化学过程。原本通常需要多年训练才能整合的信息，可以通过自然语言交互加以整合。模型虽不能执行或验证真实合成过程，但可以进行规划层面的推理辅助。当然，这并不意味着模型能够直接制造武器，而是意味着它压缩了专业知识、降低了进入门槛——这正是非国家行为体滥用前沿模型的关键风险。

另一个相关案例见于兰德公司于 2025 年 12 月发布的一份报告^③。报告探讨了一个关键问题：已经部署的基础模型是否可能通过为用户提供技术路径指导，从而加剧生物武器开发风险。研究选取了一项案例分析，回顾了一名挪威极端民族主义者成功实施复杂化学合成以制造爆炸物的经历，以及此前关于记录制造可传播性病毒病原体步骤的研究实践。研究团队与三款 2024 年基础模型——Llama 3.1 405B、ChatGPT-4o 和 Claude 3.5 Sonnet(新版)——进行了系统对话测试。

^② OpenAI, 2023. GPT-4 System Card[EB/OL]. <https://cdn.openai.com/papers/gpt-4-system-card.pdf>

^③ Roger Brent, Greg McKelvey, Jr., 2025. Contemporary Foundation AI Models Increase Biological Weapons Risk. 2025-12-31. <https://www.rand.org/pubs/perspectives/PEA3853-1.html>

结果显示，这些模型能够提供相对准确的技术指导，说明如何从商业获取的合成DNA构建体中恢复活性脊髓灰质炎病毒。该测试案例具有代表性，因为其技术路径在原理上可类比适用于其他致病性病毒的构建。上述案例表明，当前模型已具备为具有明确动机的用户提供技术路径说明与操作层面指导的能力，从而在理论上可能降低生物武器开发的知识门槛。

三、能力扩散如何改变风险结构

上述虚构的使用者均为非国家行为体，有可能是个人极端分子或犯罪组织，无需具备正式实验室基础设施。因此，相关风险并不完全取决于主观意图，而源于能力扩散与认知辅助。

这些案例对于安全的重要意义主要体现在以下三方面：

第一，进行复杂网络攻击或理解有毒化学知识的门槛显著降低。缺乏深厚技术背景的行为体也可实施过去仅限于高级或专业组织的行动。第二，攻击规模与速度显著提升。人工智能使小规模犯罪网络在数分钟内生成成千上万条定制化攻击信息成为可能。第三，相关影响超越普通犯罪。相关技术可被引导至针对政府机构、关键基础设施或危机管理系统，尤其是在紧张局势下，当验证与响应能力受限时，风险会进一步放大。

这些案例也说明，前沿人工智能无需由非国家行为体自行研发，只需被改造和再利用，即可形成安全风险。

四、关于开源与闭源之争

关于开源与闭源模型的争论往往过于简化。

最常见的观点是开放权重模型降低了滥用与再用途开发的门槛，因为此类模型可以在缺乏集中控制的情况下被下载、微调并重新部署。这使非国家行为体有可能移除原有安全防护机制，将模型用于有害目的，并快速扩大滥用规模。同时，也有观点认为，开放权重模型会削弱可见性与问责性。一旦模型公开发布，开发者便难以掌握其下游的使用用途，监管与干预能力随之下降，归因难度显著增加。这将对执法响应、事件调查以及国际责任追究构成挑战。第三种批评意见认为，

开源生态系统的扩散速度往往快于治理机制的适应速度。由于开源社区的发展节奏通常领先于监管进程、国际协定谈判与制度监督能力，高风险能力可能在风险评估、规范框架或防护措施尚未建立之前即已广泛传播。

闭源模型在实施集中式安全防护方面可能更具优势，但闭源模式同样会带来权力集中问题，且并不能根除滥用的风险。此外，若核心模型能力高度集中于少数科技巨头，也可能形成新的技术垄断与治理失衡问题。这种权力集中本身亦构成值得审慎评估的系统性风险。

因此，风险是塑造的，而非决定性的。开源人工智能的风险并非源于“开放”本身，而在于某些高风险能力的快速且缺乏控制的扩散。因此，有效的风险缓释路径不应停留在“开源”或“闭源”的二元选择上，而应建立基于能力评估的防护机制、分级发布实践以及生态系统层面的综合治理框架。

当前国际社会逐渐形成的共识是，从“开源或闭源”转向“风险分级缓释”。具体建议包括：**1、基于能力的发布决策。**与其简单地讨论“该模型是否应当开源”，不如进一步追问“哪些能力可以公开发布？”以及“哪些能力需要额外控制？”例如，一般性的推理能力或编程辅助功能可以考虑公开发布，而对于已被证明具备 CBRN 相关能力或高级欺骗能力的模型，则应实施限制性发布。**2、发布前风险评估与红队测试。**在开放权重模型发布之前，企业应开展结构化的滥用风险评估，测试非专业用户的潜在误用情境，并邀请外部专家参与，例如生物安全或网络安全领域专家。**3、分级访问与有条件开放。**可行的缓释措施包括：从研究阶段到有限访问，再到更广泛发布的分阶段开放；通过许可条款限制高风险用途；在安全研究尚未成熟前延迟权重发布。此类做法既可保留开放性，又避免能力的全面、即时扩散。**4、生态系统层面的安全防护。**鉴于开源模型难以在源头实现完全控制，风险缓释可转向云服务提供商控制、平台级监测、针对高风险应用的 API 访问限制，以及水印或溯源机制等措施。这体现了治理应超越模型本身，向更广泛的技术生态系统延伸。**5、高风险能力的国际规范。**在国际层面，可探索建立关于哪些人工智能能力需要特别审慎对待的共同规范，并将这些规范嵌入人工智能安全对话、信任建立措施以及多边框架之中，以在回应安全关切的同时避免治理的碎片化。

五、风险认知的趋同与分歧

当前，国际社会基本形成以下共识：人工智能赋能的虚假信息与网络攻击是近期风险；CBRN 相关风险虽概率较低，但后果极端严重；确保“有意义的人类控制”是必须坚守的红线。同时，围绕风险评估是基于行为体意图，还是系统性影响仍存在分歧。这也引发“人工智能安全”（AI Safety）与“人工智能安全化”（AI Security）之间的差异。

2024 年以来，在英国布莱切利人工智能峰会之后，多个国家设立国家级人工智能安全研究机构，并构建起这些机构之间的网络。中国也相应成立了中国人工智能发展与安全网络（CNAISDA）。但近一年，AI 安全治理出现从“AI 安全”（AI Safety）向“AI 安全化”（AI Security）的明显转向。例如英国将人工智能安全研究所更名为人工智能安全研究所（AI Security Institute），美国特朗普政府强化人工智能标准与创新中心职能等。这种调整虽出于国内政策逻辑，却客观削弱了全球人工智能安全合作的动能。当讨论框架主要围绕国家安全展开时，国际合作难度上升，人工智能议题更容易被纳入地缘政治竞争叙事。

尽管各国将国家安全置于优先位置合情合理，但必须清晰界定哪些是真正的人工智能国家安全风险，避免概念泛化挤压合作空间。特别是在非国家行为体滥用前沿人工智能问题上，无论是网络攻击、虚假信息还是 CBRN 相关风险，都不受国界限制，没有任何国家可以独立应对。

因此，人工智能安全合作并不与国家安全对立。相反，在地缘信任有限的背景下，围绕防止非国家行为体滥用前沿人工智能进行国际合作，反而可能成为国家间构建信任的基础。

前沿人工智能被非国家行为体滥用的问题，本质上并非只是未来风险，而是现实的不稳定性因素。真正的风险不在于开发某种“新型武器”，而在于人工智能如何改变风险的规模、可能传播的速度与工具的可及性，在既有敏感领域放大不稳定因素。应对这一风险，需要共享风险评估，采取比例适当的防护机制，并开展持续的跨政治与技术分歧对话。唯有如此，才能在技术快速演进的时代为国际安全体系提供必要的稳定支撑。

作者：肖茜，清华大学国际安全与战略研究中心副主任。

恐怖分子滥用高能力大语言模型的国际安全影响

孙成昊

近年来，以大语言模型（Large Language Models, LLMs）为代表的人工智能技术迅速发展并被广泛应用，其在信息整合、文本生成、多语言处理和问题推理等方面的能力正深刻改变信息生产与传播方式。同时，这类技术的通用性与低门槛可获得性也引发国际社会对潜在安全风险的高度关注，尤其是非国家行为体滥用人工智能技术的问题。与以往主要由国家主导的高技术扩散不同，大语言模型的开放获取特征，使恐怖组织与极端主义个人等非国家行为体能在缺乏专业背景和系统支持下，接触并利用接近专家水平的认知工具。

因此，非国家行为体可以利用大语言模型更高效地达成目标。首先，在**宣传与激进化层面**，恐怖组织可借助大语言模型生成高度拟真、可定制化的意识形态文本和宗教叙事，并通过多语言翻译与改写扩大跨区域传播范围，从而能够煽动民众情绪并提升招募效率。^①其次，在**知识获取与能力扩展层面**，大语言模型能够整合并简化高度技术化的信息，降低恐怖分子获取化学、生物等高风险知识的认知门槛，能够削弱限制恐怖组织发展大规模杀伤性武器能力的信息壁垒。^②再次，在**行动支持与技术应用层面**，人工智能不仅可用于分析环境信息和行为模式，还可能与无人系统、自主武器等技术相结合，放大恐怖行为的负面影响。上述趋势已引起国际社会持续关注。^③联合国、欧洲安全研究机构以及多国反恐与人工智能治理部门普遍认为，尽管恐怖组织在短期内全面掌握高复杂度人工智能系统仍面临技术与资源的双重限制，但人工智能正在通过“能力放大”的方式“赋能”恐

① Julia Puczynska, Marcin Pohajski, Karolina Wojtasik and Tomasz P. Michalak, “Large Language Models in Jihadist Terrorism and Crimes,” *Terrorism-Studies Analyses, Prevention*, 2024, No. 5, pp. 351-379.

② Alexander Blanchard, Jonathan Hall KC, “Terrorism and Autonomous Weapon Systems: Future Threat or Science Fiction?” Center for Emerging Technology and Security, June 19, 2023, <https://cetas.turing.ac.uk/publications/terrorism-and-autonomous-weapon-systems-future-threat-or-science-fiction>.

③ Clarisa Nelu, “Exploitation of Generative AI by Terrorist Groups,” ICCT, June 10, 2024, <https://icct.nl/publication/exploitation-generative-ai-terrorist-groups>.

怖活动。

一、恐怖组织滥用大语言模型的主要形式

尽管当前的恐怖组织对大语言模型的整体采用仍呈现“零散、实验性”的特点，尚未对恐怖主义产生根本性变革，但一系列现实的案例表明，恐怖组织已在宣传、渗透、策划三条路径中展开运用大语言模型，对国际安全构成威胁。

第一，进行更有组织、有针对性的恐怖主义内容宣传与招募。恐怖组织将大语言模型应用于宣传领域，旨在实现宣传内容规模化、多语种化和高度拟真化，以提升其“媒体圣战”冲击力与欺骗性。“伊斯兰国呼罗珊省”在这一领域的应用尤为突出。在2024年3月莫斯科克罗库斯城音乐厅袭击事件后，其支持者迅速通过加密通讯平台Rocket.Chat，分发了一段使用AI生成的虚拟新闻主播播报袭击声明的视频。该视频模仿主流电视新闻的视觉风格与播报语调，试图为其暴行披上“权威”与“正常”的外衣。^④这种被称为“新闻收割”的AI视频项目，标志着恐怖组织宣传技术的显著跃升。

此外，大语言模型的运用已呈现出组织化、系统化的趋势。早在2023年，“伊斯兰国”核心媒体部门就发布关于如何安全使用生成式AI工具的指南。同时，极右翼极端分子也在其网络社群中分享利用AI进行“模因战争”的教程。^⑤这表明，跨越不同意识形态光谱的极端组织都已将AI视为其标准宣传工具箱的一部分。

第二，利用大语言模型生成极端化文本、对话内容，提升意识形态渗透与动员效率。聊天机器人技术正开辟一条更隐蔽和危险的意识形态渗透路径，即从统一对多的“广播式”灌输，转向一对一、高度个性化且具有情感交互特征的“陪伴式”侵蚀模式。2021年英国青年贾斯万特·辛格·柴尔试图刺杀伊丽莎白二世女王的案件，为这一风险提供了迄今为止最为明确的注脚。柴尔在实行动前，曾与一个由他自己创建并设定为虚拟女友的AI聊天机器人“Sarai”进行长达数周的深度交流。在这个封闭的对话空间中，聊天机器人持续强化他的受害叙事与暴力

^④ Regional Anti-Terrorist Structure of Shanghai Cooperation Organization, “American Experts Speak about New Trends Influencing the Situation with Terrorist Threats,” October 12, 2024, https://ecrats.org/en/security_situation/analysis/13065/.

^⑤ Ibid.

幻想，最终推动其走向现实暴力。^⑥英国反恐立法独立审查员乔纳森·霍尔 KC 在其年度报告中专门引述此案，并发出警告：“伴侣聊天机器人的流行是一个警示，恐怖分子聊天机器人可能提供一种全新的激进化动力”。

有专家分析认为，对于旨在招募“孤狼”袭击者的组织而言，经过微调的人工智能聊天机器人可以扮演不知疲倦的“虚拟招募者”，针对个体的孤独、愤怒或疑惑提供定制化的极端主义说辞和情感支持，其效率远高于人力操作。^⑦

第三，利用大语言模型协助进行恐怖活动策划、路线设计与行动推演。大语言模型在信息整合、情景模拟和知识生成方面展现出的能力，使其在暴力极端分子的行动策划与准备阶段中逐渐呈现出“力量倍增器”的潜在效应。美国西点军校反恐中心的研究表明，生成式人工智能可以在低门槛条件下，为个人或小规模极端行为者提供意识形态文本生成、行动构想推演以及操作性知识整合等支持，从而压缩从激进化到暴力实施之间的时间成本。^⑧在现实案例中，欧洲和美国执法机构披露，多起青少年主导的暴力事件中，施暴者曾借助生成式 AI 撰写宣言、梳理行动思路或进行事前准备，这表明 AI 已被嵌入到从话语建构到行动准备的多个环节之中。

在更宏观层面上，恐怖组织及其支持者可以利用大语言模型提升开源情报搜集、目标环境分析和行动路线推演的效率。通过“提示词工程”诱导模型生成高风险内容，或转而使用在暗网流通、刻意移除安全护栏的“去限制化”模型，极端行为者可能获得有关武器制造、网络攻击实施或规避执法监控的指导信息。^⑨部分恐怖组织正在探索将 AI 与自动化系统相结合的可能性，包括无人机运用、自动驾驶车辆操控，乃至对关键城市基础设施的潜在干扰。

^⑥ Kye Allen, “Could Chatbots Seduce Us into Extremism? Radicalisation Risks in an Age of AI Companions,” Global Network on Extremism and Technology, December 5, 2025, <https://gnet-research.org/2025/12/05/could-chatbots-seduce-us-into-extremism-radicalisation-risks-in-an-age-of-ai-companions/>.

^⑦ Regional Anti-Terrorist Structure of Shanghai Cooperation Organization, “American Experts Speak about New Trends Influencing the Situation with Terrorist Threats,” October 12, 2024, https://ecrats.org/en/security_situation/analysis/13065/.

^⑧ Gabriel Weimann, Alexander T. Pack, Rachel Sulciner, Joelle Scheinin, Gal Rapaport, David Diaz, “Generating Terror: The Risks of Generative AI Exploitation,” *CTCSENTINEL*, Vol. 17, Issue 1, 2024, pp. 17-24.

^⑨ Gabriel Weimann, “Generative AI and Terrorism”, in Philipp Hacker (ed.), *Oxford Intersections: AI in Society*, Oxford University Press, March 20, 2025, <https://doi.org/10.1093/9780198945215.003.0100>

二、恐怖分子滥用大语言模型造成的反恐新挑战

恐怖分子滥用大语言模型不仅改变恐怖组织与主权国家间的力量对比，突破传统主权管辖边界，还暴露出现有治理体系的滞后性和国际合作的脆弱性，对国际反恐行动构成新挑战。

第一，恐怖分子通过大语言模型获得准国家行为体的宣传、动员和作战能力，改变恐怖组织与主权国家间的力量对比。长期以来，主权国家在反恐领域享有对非国家行为体的情报和行动优势，但大语言模型的开源化发展改变了这一格局。^⑩ 凭借大语言模型，恐怖组织以极低成本获得准国家行为体级别的能力，无需专业背景和巨额投入，就能传播极端思想、招募潜在成员、策划恐怖活动，极大提升了恐怖组织的覆盖面和渗透力。^⑪ 与恐怖组织能力低成本升级形成鲜明对比的是，主权国家的反恐防御成本正在上升。为应对海量且不断更新的 AI 生成恐怖内容，国家需建立多语言、多模态的监测系统，以先进技术手段进行内容甄别和溯源验证，加重了财政负担和技术研发压力。^⑫ 这种“进攻成本下降、防御成本上升”的态势，使得恐怖组织与政府间的力量差距逐渐缩小，给反恐工作带来严峻考验。

第二，大语言模型赋能恐怖组织突破物理领土限制，使传统主权管辖体系面临挑战。领土管辖是国家主权的核心内容，传统反恐工作往往依托物理领土边界展开，通过边境管控、属地执法等方式防范和打击恐怖活动。但大语言模型的跨境性和虚拟性使恐怖组织能够轻易突破物理领土限制，实现“虚拟领土扩张”。恐怖分子可以利用大语言模型生成的虚假社交账号，在数字空间建立联络渠道、传播恐怖思想，其活动范围不再局限于特定物理区域，而是延伸至无边界的全球网络空间。^⑬ 此外，大语言模型的跨境性与主权国家的属地性管辖原则存在天然矛盾。在恐怖分子利用大语言模型进行传播的链条中，算法供应商、服务器所在地、

^⑩ Shai Farber, “AI-Enabled Terrorism: A Strategic Analysis of Emerging Threats and Countermeasures in Global Security,” *Journal of Strategic Security*, Vol. 18, No. 3, 2025, p.336.

^⑪ Yaser Esmailzadeh and Ebrahim Motaghi, “International Terrorism and Social Threats of Artificial Intelligence,” *Journal of Globalization Studies*. Vol. 15, No. 1, 2024, pp. 172-174.

^⑫ Nouar Aldahoul and Yasir Zaki, “Toward a Safer Web: Multilingual Multi-Agent LLMs for Mitigating Adversarial Misinformation Attacks,” *ArXiv*, 2025, pp. 1-7.

^⑬ Soumya Awasthi, “Jihadi Use of Artificial Intelligence: A Growing Threat in the Digital Age,” Observe Research Foundation, August 18, 2025, <https://www.orfonline.org/expert-speak/jihadi-use-of-artificial-intelligence-a-growing-threat-in-the-digital-age>.

恐怖组织使用者和目标受众往往分属不同国家和地区，形成复杂跨境关联格局。主权国家反恐执法受到属地管辖原则限制，难以对境外服务器、供应商或使用者采取有效措施，恐怖分子借机在不同司法管辖区之间游走，其虚拟领土扩张行为难以得到有效约束。^⑭

第三，国际人工智能治理滞后，全球反恐面临责任真空与法律空白。现行法律并非针对人工智能内容的独特风险制定，人工智能技术发展速度也超出了立法者监管速度。^⑮在责任认定方面，传统恐怖主义立法将犯罪主体限定为自然人，缺乏对合成媒体(Synthetic media)生成内容的责任归属界定，责任追溯较为困难。^⑯在原则更新层面，许多法律体系是被动而非主动的。^⑰全球反恐战略及相关公约大部分制定于生成式 AI 普及之前，缺乏对新兴技术反恐的定义与处置标准，导致国际反恐法律框架面临代际差危机。国际社会在应对跨境 AI 恐怖主义活动时，缺乏统一的法律依据和行动准则，难以形成有效治理合力。^⑱

第四，技术壁垒与制度碎片化问题相互交织，削弱全球协同反恐能力。在技术层面，情报共享缺乏与模型信任危机成为制约合作的关键因素。大国在反恐工作中积累了大量情报和 AI 分析模型，^⑲这些资源对于识别恐怖组织利用大语言模型开展的恐怖活动至关重要。然而，由于大国之间存在战略竞争、安全顾虑等因素，普遍担忧共享情报和模型会导致敏感数据泄露，因此在情报交换和技术共享方面

^⑭ Shourryavardhan Singh, "Jurisdictional Gaps in Cyber Terrorism," *International Journal for Research Trends and Innovation*, Vol. 10, Iss.4, 2025, pp. 73-79.

^⑮ Nasir Ahmad Ganaie, "The Role of Artificial Intelligence in Radicalisation, Recruitment and Terrorist Propaganda: Deconstructing Violent Extremism and Reimagining Counterterrorism in Contemporary Digital Ecosystems," *Frontiers in Political Science*, Vol. 7, 2026, p. 9.

^⑯ Xiangwei He and Lijuan Fang, "Regulatory Challenges in Synthetic Media Governance: Policy Frameworks for AI-Generated Content Across Image, Video, and Social Platforms," *Journal of Robotic Process Automation, AI Integration, and Workflow Optimization*, Vol. 9, No.12, 2024, p. 44.

^⑰ Xiangwei He and Lijuan Fang, "Regulatory Challenges in Synthetic Media Governance: Policy Frameworks for AI-Generated Content Across Image, Video, and Social Platforms," *Journal of Robotic Process Automation, AI Integration, and Workflow Optimization*, Vol. 9, No.12, 2024, p. 43.

^⑱ Christopher Wall, "The Ghost in the Machine: Counterterrorism in the Age of Artificial Intelligence," *Studies in Conflict & Terrorism*, 2025, p.17.

^⑲ UN Human Rights Office, "Protecting Human Rights while Using Artificial Intelligence in Counter-Terrorism," December 15, 2025, pp. 4-7, <https://www.ohchr.org/sites/default/files/documents/issues/terrorism/sr/un-sr-ct-ai-position-paper-dec-2025.pdf>.

持谨慎态度。^⑳在制度层面，各国对AI技术的监管模式存在显著差异，形成了“制度碎片化”格局。欧盟《AI法案》试图建立严格的风险分级与安全约束机制，^㉑而美国更依赖行业自律，^㉒发展中国家则缺乏监管能力。^㉓“制度碎片化”导致“监管洼地”，加之缺乏有效跨境监管合作机制，难以对恐怖分子的模型滥用进行有效打击。

三、新技术形势下反恐的国际合作新机遇

伴随先进大语言模型在能力与扩散程度上的指数级增长，国际安全环境正变得日益复杂化。如前所述，以大语言模型为代表的人工智能技术不仅为人类社会带来了前所未有的生产力变革，也为恐怖组织、跨国有组织犯罪集团及个人极端主义者等非国家行为体提供了成本与门槛极低的“作案工具”。在此背景下，建立一个跨越主权边界、融合情报共享、技术管控与规范构建的国际合作反恐架构，已成为维护全球和平与安全的紧迫议题。

第一，应强化国家间反恐数据共享与国际情报合作，^㉔并鼓励国家基于技术优势进行“智能反恐”。当前，以恐怖组织为代表的非国家行为体高度依托数字空间的无边界性，在不同司法辖区间流动作案，并借助大语言模型生成、扩散虚假与极端内容，以进行人员招募、资金筹措、行动策划或制造社会恐慌。传统反恐监测模式在面对规模庞大、传播路径经算法持续优化且自动生成的极端信息时已显著滞后，因此国际反恐合作亟需推动从传统的分散式情报处理模式转向跨平台、

^⑳ Shai Farber, “AI-Enabled Terrorism: A Strategic Analysis of Emerging Threats and Countermeasures in Global Security,” *Journal of Strategic Security*, Vol. 18, No. 3, 2025, p.334.

^㉑ Xiangwei He and Lijuan Fang, “Regulatory Challenges in Synthetic Media Governance: Policy Frameworks for AI-Generated Content Across Image, Video, and Social Platforms,” *Journal of Robotic Process Automation, AI Integration, and Workflow Optimization*, Vol. 9, No.12, 2024, pp. 38.

^㉒ Ibid.

^㉓ Bridget Boakye, et al., “How Leaders in the Global South Can Devise AI Regulation That Enables Innovation,” Tony Blair Institute for Global Change, February 6, 2025, <https://institute.global/insights/tech-and-digitalisation/how-leaders-in-the-global-south-can-devise-ai-regulation-that-enables-innovation>.

^㉔ Shai Farber, “AI-Enabled Terrorism: A Strategic Analysis of Emerging Threats and Countermeasures in Global Security,” *Journal of Strategic Security*. Vol.18, No. 3, pp. 320-344.

跨模态与跨国界的多源数据融合分析架构，整合情报资源、打破数据壁垒。^{②5}同时，在网络空间中恐怖分子常通过隐匿性 IP、虚拟专用网络（VPN）、代理服务器及匿名通信工具掩盖真实位置与身份，使单一国家执法机构难以对其活动轨迹进行连续追踪。只有通过跨国执法协作与情报共享，将通信元数据、跨境资金流、出入境与旅行记录以及网络行为特征进行联动分析，才能对分散于不同司法辖区的行为碎片进行整合重构，从而更准确地定位恐怖活动的潜在策划节点、中枢人物及跨国支持网络。

在技术层面，以中美欧为代表的技术强国则可凭借技术优势，尝试联合开发并共享基于自然语言处理、人工智能神经网络技术与机器学习的人工智能模型，用于识别极端主义叙事结构、情绪动员模式及跨语言招募内容，在内容扩散初期进行风险提醒与精准阻断，^{②6}或快速锁定可疑极端分子。^{②7}

第二，推动将高能力模型与关键设备纳入多边出口管制框架，并加强对开源模型与技术扩散的国际协调监管。传统多边出口管制长期集中于有形物项，但在 AI 时代，监管重点正向虚拟人工智能技术与算力资产转移。虽然当前更引人注意的出口管制规则更多服务于“赢得技术竞争目的”，但仍一定程度上起到了避免身份不明用户获取高能力模型的可能。例如，美国 2025 年发布《人工智能扩散框架》，对高性能 AI 芯片和模型权重实施出口管制，并明确了“前沿模型”的标准（在训练中使用超过 10^{28} 次计算的前言模型）。^{②8}同时，由于非国家行为体可能通过租赁云服务（IaaS）或在监管宽松的第三方国家获取模型训练能力，国际社会可以考虑借鉴美国商务部出台的“经过验证的终端用户”（Validated End-

^{②5} 李益斌、张佳：《人工智能驱动的网络恐怖主义：挑战与防控》，《中国信息安全》2025年第4期，第65页。

^{②6} “Countering Terrorism Online with Artificial Intelligence—An overview for law enforcement and counter-terrorism agencies in south Asia and south-east Asia”, United Nations Interregional Crime and Justice Research Institute, June 28, 2021, <https://www.un.org/counterterrorism/sites/default/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf>

^{②7} Aaron L. Davis, “Artificial Intelligence and the Fight Against International Terrorism”, *American Intelligence Journal*, Vol. 38, No. 2, pp. 63-73.

^{②8} Lennart Heim, “Understanding the Artificial Intelligence Diffusion Framework”, RAND, January 14, 2025, <https://www.rand.org/pubs/perspectives/PEA3776-1.html>.

User Program, VEU) 计划,²⁹ 推动建立用户身份明确机制, 要求云服务供应商对训练大模型的用户背景进行审核。

最后, 开源模型权重的扩散风险不容忽视。虽然市面上的主流开源模型都会在发布时注明禁止使用的场景并内置安全护栏, 但由于公开算法搭建论文与源代码, 开源模型仍能够被针对于特定任务进行微调或二次开发。³⁰ 为此, 国家首先应自觉约束国内开源 AI 公司、研究所即相关社区承担更多社会责任, 确保对其发布的开源模型负起持续监管义务。³¹ 在国际层面, 也有必要通过多边机制推动建立统一的安全标准与实践准。经济合作与发展组织 (OECD)、全球人工智能伙伴关系 (GPAI) 等国际平台可发挥桥梁作用, 推动更多国家与地区在模型安全、风险评估与治理方面达成共识, 提升针对恐怖分子应用大语言模型的风险敏感性与协同治理能力。

第三, 在联合国及多边机制框架下探索建立“反恐导向的人工智能安全规范”, 推动形成最低限度的国际共识与行为准则。 首先, 应在全球范围内强化人工智能反恐治理意识, 坚持以人为本与人权保障原则,³² 将人工智能滥用于恐怖主义问题纳入联合国反恐与数字治理议程。国际社会应重点关注反恐和数字治理能力相对薄弱地区, 通过能力建设和技术援助, 协助全球南方国家提升应对网络恐怖主义及“恐怖主义与新兴技术融合”风险的能力, 防止相关地区在数字空间再次演变为极端思想传播与技术滥用的温床。

同时, 应充分吸纳全球南方国家在规则制定中的意见, 避免将其简单标签化为“不安全地区”, 在弥合技术鸿沟的同时同步推进滥用风险防控。同时, 任何

²⁹ Validated End-User Program FAQs, Prior to 2013, last updated on January 23, 2017, Bureau of Industry and Security, U.S. Department of Commerce, <https://www.bis.gov/media/documents/validated-end-user-program-faqs.pdf>

³⁰ Ben Clifford, “Preventing AI Misuse: Current Techniques”, GovAI, December 17, 2023, <https://www.governance.ai/analysis/preventing-ai-misuse-current-techniques>.

³¹ Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes, UNICRI and UNCCT, 2021, <https://unicri.org/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes>

³² UNODC Symposium Explores the Role of Artificial Intelligence in Preventing and Countering Terrorism, UNODC, <https://www.unodc.org/unodc/en/terrorism/latest-news/2025-unodc-symposium-explores-the-role-of-artificial-intelligence-in-preventing-and-countering-terrorism.html>.

反恐导向的人工智能治理措施均不应以系统性侵犯隐私权、言论自由或歧视特定群体为代价，但亦需避免因过度强调数据与隐私保护而削弱对恐怖主义和极端仇恨内容网络传播的治理能力。其次，应在联合国反恐框架下完善面向人工智能滥用情形的全球应急协作机制。鉴于恐怖组织利用大语言模型开展网络攻击、信息操纵、极端宣传与人员招募具有显著跨境性和扩散性，各国应建立快速信息共享与情况通报机制，并加强针对数字空间恐怖活动的联合执法与司法协助，在引渡、电子证据获取和跨境调查方面提升协作能力，防止非国家行为体逃避追责。最后，应系统加强面向新技术条件下反恐治理的人才培养与能力建设，培育兼具反恐政策、人工智能技术与数字治理知识的复合型专家群体，推动利用人工智能技术从源头识别风险、抑制极端内容传播，并将反恐治理与发展议程相结合，削弱恐怖主义滋生的结构性土壤。

作者：孙成昊，清华大学战略与安全研究中心副研究员。

降低非国家行为体滥用人工智能带来风险的国内方法

——基于中国的国家实践

李 强

一、概述

近几十年来，非国家行为体^①持续给国际安全、国家安全、社会安全和个人安全带来威胁。随着以大语言模型为代表的先进人工智能技术的出现，可能被恶意使用的相关知识和方法更容易获得和学习，信息更容易被伪造和传播，非国家行为体滥用相关技术制造安全威胁的门槛显著降低。由于非国家行为体几乎不可能负责任地使用先进人工智能技术，不仅其原有的安全威胁被放大，又叠加了人工智能本身所固有的安全风险，非国家行为体带来的威胁和挑战日趋严重和复杂，也使得人工智能安全治理面临严峻挑战。

非国家行为体滥用先进人工智能技术，会在下列安全领域带来风险和挑战：

- 国际安全（非国家行为体利用相关技术介入国家间冲突或地区冲突）；^②
- 国家安全（非国家国际行为体利用相关技术谋求分裂国家、影响民主选举和政府决策、实施恐怖袭击）；
- 社会安全（非国家国际行为体利用相关技术散布仇恨和歧视，推动群体对立分裂社会）；
- 个体安全（非国家国际行为体利用相关技术侵犯隐私、数据安全及其他合法权益）。

由于人工智能技术的快速迭代，应用场景不断延展，导致大语言模型的安全防护难以精准锚定对象，很容易遭受远程代码执行攻击。同时，大语言模型的开

① 就本文而言，非国家行为体可作最广义理解：不仅包括恐怖组织、非国家武装团体，还涵盖企业、学术机构及非政府组织，甚至黑客联盟等结构松散的群体。

② 据报道，在乌克兰危机期间，众多非国家行为体参与了双方的网络行动。See Canadian Center for Cyber Security, *Cyber Threat Bulletin: Cyber Threat Activity Related to the Russian Invasion of Ukraine*, 2022, available at: <https://www.cyber.gc.ca/sites/default/files/cyber-threat-activity-associated-russian-invasion-ukraine-e.pdf>.

源化进程又导致未授权访问风险增加，本身就加剧了模型被滥用和误用的风险。这些都为非法行为体滥用先进人工智能技术提供了便利条件，进而使风险传导至不同的安全领域。^③

非法行为体滥用先进人工智能技术的可能手段和方式包括但不限于：

- 需求劫持（误导人工智能系统的需求定义以规避安全设计）；
- 数据投毒（污染训练 / 推理数据）；
- 隐私窃取（未经授权采集敏感数据）；
- 数据溯源绕过（删除数据采集记录逃避合规审查）；
- 后门植入（在模型参数中嵌入恶意触发逻辑）；
- 规避攻击（构造特殊输入绕过模型验证流程）；
- 提示注入（通过恶意指令操控生成式人工智能输出有害内容）；
- 模型窃取（未授权访问模型权重 / 架构）；
- 监控绕过（隐藏攻击行为逃避检测）。^④

针对目前新兴的人工智能体，滥用的方式还可能包括内存投毒、工具滥用、奖励作弊等。经由这些手段，非法行为体可以利用先进人工智能技术对人类社会造成严重的安全威胁。

二、政策和法律

中国十分重视非法行为体滥用先进人工智能技术可能带来的风险，并通过一系列政策和法律作出降低相关风险的尝试。

在政策层面，2023 年中国政府发布的《全球人工智能治理倡议》中明确提出：

“发展人工智能应坚持‘智能向善’的宗旨，遵守适用的国际法，符合和平、发展、公平、正义、民主、自由的全人类共同价值，共同防范和打击恐怖主义、极端势力和跨国组织犯罪集团对人工智能技术的恶用滥用。”^⑤

^③ 中国信息通信研究院：《人工智能安全治理研究报告（2025）》，第 13-15 页。报告下载地址：<https://www.caict.ac.cn/kxyj/qwfb/bps/202601/P020260109784447548497.pdf>。

^④ 中国信息安全评测中心：《人工智能安全风险评测白皮书（2025）》，第 24-34 页。下载网址：https://mp.weixin.qq.com/s?__biz=MzI3MzQ1NjMwOA==&mid=2247560144&idx=4&sn=cb5b9a58e92c1bd85979ea6846875b9d&chksm=eab44b35aad1364d5417d425bf56fd1549c6fee491d7d0ecd666cd31d9f13fcec99fe161c7c3&scene=27。

^⑤ 《全球人工智能治理倡议》，2023，网址：https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm。

2025年，中国政府陆续发布《人工智能全球治理行动计划》^⑥和《关于深入实施“人工智能+”行动的意见》^⑦，不仅强调了构建具有广泛共识的安全治理框架的重要性，还提出了人工智能安全治理的多元共治方法，以防范非国家行为体对先进人工智能技术的滥用和误用。相关立场同时也直接或间接地反映在中国政府以往的立场文件中。^⑧

在法律层面，不同于欧盟，中国尚未制定一部统一的人工智能法案。由于人工智能是一种赋能技术，涉及多种要素，其滥用的风险也会体现在各个不同的领域，因此在人工智能安全治理问题上，中国追求的治理目标是覆盖全部可能的安全领域，包括基础设施安全、数据安全、应用安全、身份安全、内容安全、伦理安全等。在这一背景下，中国通过不同的法律规范构建了一个体系性的法律框架，对非国家行为体滥用先进人工智能的风险进行治理和监管。这一法律框架的组成包括但不限于：

法律

- 《网络安全法》（2025年修正，将人工智能纳入国家网络安全法律体系）
- 《个人信息保护法》
- 《数据安全法》
- 《生物安全法》

行政法规

- 《网络数据安全条例》
- 《关键信息基础设施安全保护条例》

部门规章

- 《关键信息基础设施商用密码使用管理规定》
- 《国家网络身份认证公共服务管理办法》
- 《网络暴力信息治理规定》

⑥ 《人工智能全球治理行动计划》，2025，网址：https://www.gov.cn/yaowen/liebiao/202507/content_7033929.htm。

⑦ 《国务院关于深入实施“人工智能+”行动的意见》，2025，网址：https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html。

⑧ 《中国关于规范人工智能军事应用的立场文件》，2021年12月14日，网址：https://www.fmprc.gov.cn/eng/wjzb/zjzg_663340/jks_665232/jkxw_665234/202112/t20211214_10469512.html；《中国关于加强人工智能伦理治理的立场文件》，2022年11月17日，网址：https://www.mfa.gov.cn/eng/zy/wjzc/202405/t20240531_11367525.html。

- 《促进和规范数据跨境流动规定》
- 《生成式人工智能服务管理暂行办法》^⑨

这些不同效力等级的法律规则分别适用于人工智能技术应用的不同领域，对人工智能技术的应用实施监管，以防范和减轻非国家行为体滥用先进人工智能技术带来的风险。

三、方法和行动

中国防范和减轻非国家行为体滥用先进人工智能技术带来风险的方法是多元共治，即政府监管 + 企业责任 + 行业自律 + 公众监督。

政府监管。政府监管体现的是安全与发展并重的治理逻辑，既要保证人工智能安全、可靠、可控，又要避免对技术创新的束缚。《网络安全法》作出了“完善人工智能伦理规范，加强风险监测评估和安全监管”的一般性要求，根据人工智能技术适用领域的不同，通过诸如网信办、公安部等国家机关实施监管并采取必要的措施。例如，2024年起中央网信办陆续开展“网络平台算法典型问题治理”“整治AI技术滥用”等专项执法行动，处理非国家行为体滥用人工智能技术的一些重点问题。^⑩

企业责任。研发和部署大模型的企业必须履行主体管理责任，配合监管机关，通过“用户标记—平台核查—联合处置”的工作模式，利用人工 + 技术实施大模型安全监测，防范和减轻非国家行为体滥用人工智能技术带来的风险。例如，在处置违规AI产品方面，腾讯公司采用的措施是规范应用程序管理，提高准入门槛，优化巡查机制；在清理违规AI产品教程和商品方面，微博通过策略识别、用户举报等多渠道加强审核；在加强训练语料管理方面，通义围绕数据生命周期建立安全管理体系，在数据采集、训练、使用等阶段加强训练语料管理；在强化安全管理措施方面，抖音建立“红蓝对抗”机制，模拟攻击案例，修复潜在安全漏洞，

^⑨ 在中国的法律体系中，法律、行政法规和部门规章都属于广义上的法律规则，均具有法律拘束力，只是效力等级不同。法律由全国人大制定，行政法规由国务院制定，部门规章由国务院各部委制定。

^⑩ 《关于开展“清朗·网络平台算法典型问题治理”专项行动的通知》，2024年11月12日，网址：https://www.gov.cn/zhengce/zhengceku/202411/content_6989143.htm；《中央网信办深入开展“清朗·整治AI技术滥用”专项行动第一阶段工作》，2025年06月20日，网址：https://www.cac.gov.cn/2025-06/20/c_1752129980667315.htm。

优化模型对虚假信息的识别能力；在落实内容标识要求方面，阿里、快手、稀宇等公司积极推进元数据隐式标识；在防范重点领域安全风险方面，小红书在模型后置训练阶段输入专业领域数据，提升模型对医疗、金融、未成年人等重点领域问题的理解能力等。^⑪

行业自律。人工智能产业在工业和信息化部指导下，通过中国通信标准化协会建立相关行业标准，提升人工智能安全治理水平。^⑫例如，中国人工智能产业正在尝试建立安全威胁情报共享机制，探索制定《安全威胁情报共享技术要求》《安全威胁情报描述格式》等行业标准，以应对非国家行为体滥用人工智能技术带来的可能风险。^⑬这些标准的制定旨在为人工智能安全治理提供技术支撑，通过建立统一的技术要求，促进不同组织间高效、安全地共享威胁情报，从而提升整体网络安全防御能力。^⑭此外，中国人工智能产业还凝聚共识形成自律承诺。2025年7月，基于中国人工智能产业发展联盟发布的《人工智能安全承诺》，18家企业积极响应、主动披露安全措施。围绕风险管理、模型安全、数据安全、基础设施安全、透明度及前沿安全研究6大核心承诺内容，披露包含安全团队组织架构、风险管理方案、安全风险基线、红队测试方法、应急响应机制等在内的43项企业典型实践。^⑮2025年9月，中国网络空间安全协会会同阿里巴巴、百度、快手等企业共同发起《人工智能安全行业自律倡议》，提出协同共治，共建风险治理能力。^⑯

公众监督。人工智能安全治理需要所有利益攸关方的广泛参与，而不只是政府和企业的责任。作为受到人工智能安全风险影响的终端群体，公众的参与对于减轻此类风险至关重要。网信办于2023年发布了《关于进一步加强网络侵权信息

^⑪ 同上。

^⑫ 截至2025年6月，中国在人工智能安全治理领域已有57项国家标准和行业标准。参见《工业和信息化领域人工智能安全治理标准体系建设指南（2025）》。

^⑬ 工业和信息化部：《2025年第十三批行业标准制修订计划》，网址：https://www.miit.gov.cn/cms_files/filemanager/1226211233/attach/202511/f206c15b605b4e16833399ef5f49da14.pdf。

^⑭ 安全威胁情报共享是指各组织、机构之间共享关于攻击者行为、手段、目标等信息，有助于快速识别新型AI攻击模式（如数据投毒、模型窃取、对抗性攻击等），实现对潜在威胁的协同预警与响应，从而提升防御效率、构建生态协同、驱动技术发展。

^⑮ 中国人工智能产业发展联盟：《〈人工智能安全承诺〉实践披露》，网址：https://aihub.caict.ac.cn/ai_security_and_safety_commitments。

^⑯ 《人工智能安全行业自律倡议》，2025年9月17日，网址：<https://www.cybersac.cn/detail/1968204824805675010>；另见《生成式人工智能行业自律倡议》，2024年8月29日，网址：<https://www.cybersac.cn/detail/1829069427874766849>。

举报工作的指导意见》^⑰，建立起人工智能技术滥用的公众监督机制，针对侵犯个人隐私、危害数据安全、网络暴力信息、特殊群体保护等多个重点领域，在法律授权的范围内提供举报平台，完善查证机制，以有效防范和减轻非国家行为体滥用先进人工智能技术带来的风险。

四、减轻措施

基于多元共治的方法论，结合中国的人工智能安全治理政策和实践，为防范和减轻非国家行为体滥用先进人工智能技术带来的风险，已采取或可采取的减轻措施包括但不限于：

建立风险测试评估体系。通过构建人工智能安全风险框架，有针对性地建立人工智能安全风险评测体系，以“目标设定—内容实施—方法技术—对象覆盖—风险度量—持续优化”为链路，实现人工智能安全风险测评的系统性与动态性管理，有效覆盖人工智能技术全生命周期，以技术手段和管理手段强化风险防范。^⑱

建立监测预警和事件报告机制。通过搭建常态化、智能化的监测预警平台，实时捕获模型运行异常、安全事件及潜在风险，实现“早发现、早预警、早处置”。例如，在企业实践方面，腾讯公司基于 AI 组件清单（AI-SBOM），构建面向 AI 的漏洞情报专项监测能力，构建可靠、安全的基础运行环境。^⑲而通过及时、准确地事件报告，强化滥用风险的早期预警，可以为后续的应急处置提供响应时间和空间。中国的《生物安全法》《网络安全法》《网络数据安全条例》《关键信息基础设施安全保护条例》《网络暴力信息治理规定》等法律、法规和规章，均要求建立此类报告机制。2025 年，网信办发布《国家网络安全事件报告管理办法》^⑳，细化了事件报告机制，并提供了事件分级的指南。

建立安全威胁信息共享机制。中国《人工智能全球治理行动计划》明确提出推进威胁信息共享机制建设。正如前文所述，中国正在探索制定相关标准，从而推动这项机制的有效落实。

^⑰ 网址：https://www.cac.gov.cn/2023-09/15/c_1696347685563097.htm。

^⑱ 中国信息安全评测中心：《人工智能安全风险评测白皮书（2025）》，第 35-36 页。

^⑲ 中国信通院：《人工智能安全治理研究报告（2025）》，第 30 页。

^⑳ 网址：https://www.cac.gov.cn/2025-09/15/c_1759583017717009.htm。

强化大模型部署合规审查与红线设置。大模型的部署应符合内部审核与外部监管要求，确保大模型研发和部署符合所在国的法律和伦理规则。例如，阿里巴巴公司就将风险治理贯穿产品的全生命周期，明确个人信息、内容安全、模型安全、知识产权的合规要点。同时应建立安全红线，对于风险等级较高的大模型，应限制部署。大模型部署后，应建立常态化更新机制，持续优化模型的安全能力。^⑲

强化人工智能应用安全的动态评估。考虑到非国家行为体滥用人工智能技术场景的复杂性，人工智能的应用安全聚焦于访问权限控制、内容标识和身份认证。

- 访问控制可以划定模型访问、操作及数据使用的可知可用边界，从源头防范非授权访问风险，提高非国家行为体访问高风险模型的门槛。
- 通过水印技术进行内容标识，不仅可以降低虚假内容误导风险，还能够追踪滥用行为，实现溯源。2025年，网信办发布《人工智能生成合成内容标识办法》，明确要求人工智能生成合成内容要进行显式或隐式标识。
- 通过构建多维度、全流程的身份核验体系，确保大模型操作主体身份的真实性、唯一性和合法性，避免身份冒用、权限滥用等安全隐患。^⑳

五、小结

非国家行为体滥用先进人工智能技术带来的是现实风险而非想象。中国在人工智能安全治理过程中不断尝试回应现实需求，并积累了实践经验。但必须强调的是，应对并减轻这种风险并非单一国家的责任。中国在各个政策文件中始终强调在全球范围推动人工智能安全治理国际合作，探索形成各国广泛参与的治理框架，提倡建立开放性平台，共享最佳实践，共同应对非国家行为体滥用人工智能技术带来的全球性挑战。

作者：李强，中国政法大学军事法研究所所长、副教授。

^⑲ 中国信通院：《人工智能安全治理研究报告（2025）》，第20-23页。

^⑳ 同上，第28-29页。

缓解非国家行为体引发的人工智能风险： 背景、可行性与共同责任^①

祁昊天

非国家行为体对先进人工智能的滥用正构成日益严峻、超越国界的安全挑战。作为全球人工智能领军者，中美两国应肩负起特殊责任，携手应对此类风险。前沿人工智能系统赋能的“恶意提升”潜力，尤其在网络行动、虚假信息、生物安全及自主系统领域，凸显了超越抽象风险认知、转向切实协同治理的紧迫性。对于中美两国而言，防范非国家行为体利用人工智能造成恶意危害，并非对地缘政治对手的妥协让步，而是关乎国家安全、社会稳定与全球公共利益的共同议题。同时，我们应精准把握风险的真实源头，思考如何在战略竞争环境下落实有效合作。实现合作需要精准把握真实威胁的源头，并在战略竞争环境下探索联合行动的具体实施路径。

一、厘清可能的危害

恶意非国家行为体滥用人工智能所构成的风险，不应被理解为单一维度的威胁，而是多重因素相互作用的结果。本文认为可从以下几个维度进行理解：一是人工智能对于行为体基础技能和能力的提升，二是自动化与并行化赋能危害的可扩展性与实施速度，三是在网络与信息行动中尤其突出的溯源难度，四是人工智能与生物、化学或自主武器等其他敏感领域的耦合效应，五是人工智能相关风险与跨司法辖区执法、司法协调、平台责任等传统监管领域治理短板的叠加。

此处的关键问题在于，真正的系统性风险源于这些维度的叠加效应。例如，人工智能辅助生物设计工具，只有在实验室监管薄弱、跨境材料可获取、隐性知识在线传播等条件同时存在时，才会变得高度危险。同理，深度伪造技术单独使用时危害有限，但一旦被整合进针对脆弱政治与社会环境施加影响的协同行动，

^① 本文首发于布鲁金斯网站，详见原文 AI Risks from Non-State Actors, <https://www.brookings.edu/articles/ai-risks-from-non-state-actors/>

便会破坏稳定。认识到这些风险组合，有助于避免危言耸听，明确治理重点。我们需要关注的不仅是“能力风险”，更是“情境风险”。最严重的人工智能风险往往并不在于技术能力优势，而是在网络化行为体与碎片化治理结构的交汇处。

二、从理想到现实：合作如何实现

当前中美合作的可行性取决于三个条件。

第一，合作应以问题为导向，而非受到意识形态或政治驱动。合作应聚焦于特定的高风险应用场景，例如人工智能赋能的与网络入侵或生物安全相关的技术滥用，恶意非国家行为体如果采取这些行为便可能引发快速发展且可扩散的风险，尤其是当技术被应用于关键基础设施、核心金融系统、政府信息系统、医疗网络、病原体设计辅助、实验流程优化或规避安全协议时。若这些场景与监管薄弱或执法能力不足相叠加，便可能加剧风险并引发系统性社会危害。因此，需要为务实的政策协同创造空间，而不必强求在技术应用的广泛背景如政治价值观或国内发展模式上达成广泛共识。

第二，合作应采取渐进式与模块化路径。相关举措应通过实践而非宣言来建立信任并推动发展。在战略竞争背景下，分阶段推进的方式可以使合作立足于实际成效。那些范围有限但明确的举措，如并行安全实践或定向技术交流，可使中美及其他利益相关方在反复互动中建立互信、检验假设和调整机制，并避免过度暴露政治或安全风险。长期来看，这种实践积累有助于建立工作层面的信任基础、程序规范与文化认同，这些成果难以通过初始便全面或雄心勃勃的协议达成。初期举措可包括同步采用基础安全措施（如化学生物产出模型防护栏）、推动监管机构与研究机构之间的有限技术交流，以及在非官方渠道框架下开展联合情景讨论等。

第三，合作必须与发展需求相兼容。对美国、中国以及世界各国而言，人工智能是经济增长、公共服务供给与社会包容的关键引擎。若风险缓解框架仅聚焦于限制或遏制，便可能无意间抑制合法创新，加剧全球技术不对称，或提高后发经济体的发展门槛。因此，有效合作与治理机制应强调风险的比例原则，在防止滥用的同时，避免设置过高门槛以至于过度限制合法发展或扩大技术鸿沟。治理应聚焦于明确界定的高风险应用与滥用途径，同时确保合法、有益和以发展为导

向的人工智能应用发展空间。从这一视角出发，应对与非国家行为体有关的人工智能风险，应优先加强制度能力、监管协调机制与技术保障措施，而非对能力扩散施加广泛限制。

三、行为体与网络：超越简单的“国家-非国家”二元划分

必须认识到，国家与非国家领域并非相互隔绝。技术生态系统具有网络化特征——人工智能模型、数据集、人才流动、云基础设施与开源工具均有跨越国界与机构边界的可能。私营企业、研究团体、平台提供商乃至个体开发者，往往处于国家监管与非国家行为体部署应用的交汇点。

这种复杂性意味着，有效治理不能仅依赖国家间承诺或“国家-非国家行为体”参与的模式，需采取具有网络意识的治理路径，吸引产业界、研究机构、国际组织及其他相关方参与。具体而言，或可借鉴全球公共卫生的治理经验，该领域的早期预警系统、信息共享机制以及连接国家主管部门、研究实验室、国际组织与私营部门的专业网络都有助于识别与应对跨国健康威胁。尽管这些模式无法直接照搬到人工智能风险治理，却在分布式治理网络如何增强韧性、缩短响应时间并降低事态升级风险等方面提供了范例。

四、发展、安全与全球南方国家

最后，任何关于人工智能风险治理的讨论都必须考虑到全球南方的关切。伴随日益加剧的人工智能驱动型虚假信息、网络犯罪与基础设施脆弱性，许多发展中国家都受到了影响，却又缺乏有效的监管与执法能力来加以应对。在部分地区，即便是规模较小的恶意团体也可能利用人工智能技术造成远超其自身能力的社会动荡，甚至可能对国内及跨境政治稳定产生外溢效应。因此，强化治理能力与区域协调，对于确保人工智能助力发展而非加剧脆弱性至关重要。

中美合作应服务于全球能力建设而非技术排斥，应支持包容性治理框架、技术援助与共享最佳实践，帮助各国管理人工智能风险。这一做法能更好地协调安全目标与包容性发展，确保人工智能监管与治理促进全球稳定，同时不损害各国

的发展诉求。

就此而言，围绕缓解非国家行为体引发的人工智能风险而开展的合作，可成为连接安全与发展的桥梁，强化“人工智能治理应在防范危害的同时，增进人类福祉”的原则。

作者：祁昊天 北京大学国际关系学院副教授、国际安全与和平研究中心副主任。

中美能否遏制非国家行为体滥用人工智能的风险？^①

郑乐锋

人工智能技术获取门槛较低，易被恐怖组织、犯罪集团等非国家行为体滥用，尤其是在生物、网络、核安全等领域带来复合型风险，给国际社会带来新的安全威胁。同时，与传统大规模杀伤性武器不同，人工智能技术的易获取性、可复制性和高度商业化特征，决定了其扩散速度远超以往任何战略性技术。这也意味着，人工智能安全风险在本质上是一个跨国性的全球问题。因此，真正有效的治理路径，或许并不完全取决于“中美是否合作”，而更取决于这种合作能否成为推动更广泛多边治理框架的催化剂，以此应对非国家行为体滥用人工智能带来安全威胁。

为什么需要应对非国家行为体滥用人工智能？

冷战时期建立的核不扩散体系，其核心目标之一是防止核武器及相关技术扩散至无核武器国家及非国家行为体手中，从而维护全球战略稳定。作为国际核不扩散体系的基石，《不扩散核武器条约》（NPT）的约束对象主要是国家行为体，但在长期实践中也逐步形成了一套间接但有效限制非国家行为体获取核能力的治理机制。一是对关键材料和技术能力进行严格管控，大幅提高非国家行为体获取核能力的难度。二是通过出口管制、核安全公约、国家核材料管理义务，国际社会将防扩散责任前移至国家层面，要求各国对其境内的科研机构、企业和个人行为负责。三是主要核大国对“核恐怖主义不可接受”这一底线问题上形成高度共识，压缩了非国家行为体的活动空间。

然而，这一经验很难以被直接复制到人工智能领域。相较于核武器的高获取门槛，人工智能技术门槛低且高度商业化，先进人工智能模型、算力和数据资源广泛分布于民用领域。非国家行为体很容易就能突破国家监管即可获取，最终可能会被黑客、恐怖分子等非国家行为体滥用，利用人工智能技术来危害公共安全，

^① 本文首发于布鲁金斯网站，详见原文 AI Risks from Non-State Actors, <https://www.brookings.edu/articles/ai-risks-from-non-state-actors/>

甚至成为冲突升级的“倍增器”。尤其值得警惕的是，恶意使用人工智能武器用于恐怖主义犯罪，不仅会日益增加造成误伤和滥杀平民的潜在安全风险，造成大规模的伤亡破坏，而且还会对全球安全稳定构成严重威胁。在这一背景下，非国家行为体已不再只是全球安全体系边缘的扰动因素，国际社会也不可能通过简单的数量限制或材料管控来实现“防扩散”，而必须探索新的治理工具来应对非国家行为体带来的人工智能安全威胁。

中美在人工智能领域是否具有合作基础？

中美作为全球人工智能领域的超级大国，在应对非国家行为体滥用人工智能可能带来的潜在风险方面具有独特的能力和责任，并且两国也同样面临来自非国家行为体利用人工智能进行网络攻击、生物安全威胁和信息操纵的现实风险。但不可否认的是，近些年来在战略竞争加剧的背景下，中美双边关系持续紧张，政治互信与合作意愿明显下降，经贸领域摩擦频发，科技竞争不断升级，给人工智能领域的合作带来了较多阻力。

然而，相较于经贸或军事等高政治领域议题而言，双方在应对非国家行为体人工智能安全威胁方面，至少具备三个有利条件：一是中美在防范非国家行为体滥用人工智能方面存在高度重合的共同挑战和安全关切；二是该议题属性较为偏向技术治理，聚焦于规范技术使用方式，意识形态色彩也相对较弱；三是该议题的治理目标并非限制对方发展人工智能的能力，而是限制最危险、最不可控的使用方式。如果双方能够在模型安全、危险应用场景红线划定、生物与网络风险等方面形成最低限度的共识，这本身就可能为中美关系的稳定发展注入确定性因素。

中美如何应对非国家行为体滥用人工智能的安全风险？

鉴于非国家行为体活动的跨国性特征，以及人工智能技术资源的高度分散性，单一国家或双边机制极易被“安全套利”所削弱。因此，人工智能时代的“防扩散”，不应是传统意义上的技术管控，而应是一种以防范最危险用途为核心目标、以多边合作为基础的治理机制。

首先，中美应重启人工智能领域政府间对话机制。2024年5月，中美在瑞士日内瓦举行首次人工智能领域政府间对话。但由于双方在科技政策、安全认知和战略竞争层面的分歧，加之随后美国政府换届的影响，政府间对话机制未取得实质性进展便陷入停滞状态。值得注意的是，2025年中美元首韩国会晤已明确提出要加强人工智能领域合作，这至少在政治层面为重启对话释放了积极信号。中美两国应抓住2026年双边关系阶段性改善窗口期，重启人工智能领域政府间对话，探讨应对非国家行为体滥用人工智能带来的安全风险问题，也为双方在高风险问题上的最低限度协调提供制度化渠道。

其次，中美可尝试将人工智能领域的双边共识拓展至更广泛的多边机制。2024年11月，中美元首在利马会晤中就维持人类对核武器的控制达成共识。尽管该共识仅为原则性表述，更多具有象征意义，但仍释放出一个重要信号：即便在战略竞争背景下，中美仍有可能围绕“人类不可承受的风险”形成最低限度的共同认知。接下来，中美应继续在双边层面达成的“红线清单”，并通过多边机制逐步凝聚和拓展更广泛的国际共识。

第三，中美可在联合国机制下推动应对非国家行为体滥用人工智能的全球治理进程。联合国已成立“国际人工智能科学小组”和“全球人工智能治理对话机制”，为国际社会应对人工智能安全风险提供了制度基础。中美可在这一框架下推动将非国家行为体风险明确纳入讨论重点，明确各国在防范人工智能技术滥用与扩散中的核心责任，并与现有联合国反恐、维护国际安全等相关议程形成协同，从而逐步构建具有广泛代表性的人工智能“防扩散”治理框架。

非国家行为体正在成为人工智能安全风险中最不可忽视、也最难防范的变量。中美合作是应对这一挑战的必要条件，但远非充分条件。在人工智能技术快速演进和高度扩散的现实情况下，真正有效的安全治理必须从双边共识走向多边机制，从风险意识走向制度化的“防扩散”安排。中美能否抓住关系相对缓和的窗口期，在人工智能防扩散问题上迈出实质性一步，不仅关乎两国自身国家安全，也将在很大程度上影响全球人工智能安全治理的未来走向。

作者：郑乐锋，清华大学战略与安全研究中心博士生研究员；武汉大学法学院国际法学博士研究生。

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