中国新一代人工智能发展战略研究院 CINGAI

安远 AI CONCORDIA AI

I-AIIG 清华大学人工智能国际治理研究院 INSTITUTE FOR AI INTERNATIONAL GOVERNANCE TSINGHUA UNIVERSITY

前沿AI风险与应急响应研究报告

（征求意见稿）

2026年7月

执行摘要

研究背景

本报告聚焦前沿人工智能（前沿AI），即能力处于最前沿的人工智能模型和系统。其不断的能力跃升和应用推广正深刻重塑生产生活与社会治理范式：一方面，前沿AI正在深度赋能科学研究、医疗健康、网络安全防御等重要战略领域；**另一方面，其可能引发的灾难性风险，也已成为全球AI治理关注的核心议题之一。**联合国人工智能独立国际科学小组近日指出，“科学界目前无法保证，随着能力持续增强，人工智能不会——无论是自主行为还是因恶意用户所致——造成灾难性危害”，这意味着全球政策制定者必须采取有效行动。[1]多个国家和地区开始针对前沿AI采取专门治理措施，我国网络安全标准化技术委员会（下文简称“TC260”）发布的《人工智能安全治理框架》2.0版也指出，需要积极研究应对人工智能灾难性风险的共识性准则，推动其健康发展和规范应用。[2]

当前各国针对前沿AI风险的治理工具与制度安排，大多集中于事前预防管控环节，如大模型安全评测、算法备案登记、透明度披露约束等事前预防性机制。**与之相对，当前全球在前沿AI引发灾难性事件后的应急响应范式与能力方面，普遍仍有建设空间，**[3]包括前沿AI风险事件分类分级标准、监测预警方法、跨国与跨部门协同机制等方面。[4]一旦前沿AI风险在现实世界中造成灾难性事件，可能造成无法逆转的严重影响。如何应对前沿AI风险引发的突发事件或危机，已是一个全球共同面临的难题与考验。

在我国，构建适应人工智能风险的应急响应体系正在进入治理视野。顶层战略层面，2025年4月，习近平总书记在中共中央政治局第二十次集体学习时作出部署，提出要“构建技术监测、风险预警、**应急响应**体系，确保人工智能安全、可靠、可控”。[5]2026年3月印发的

---

1 联合国：《“科学依据已然存在”：联合国秘书长欢迎首份全球人工智能评估报告》，2026年7月2日，https://unsdg.un.org/zh/latest/stories/“科学依据已然存在”：联合国秘书长欢迎首份全球人工智能评估报告。另见联合国人工智能问题独立国际科学小组：《联合国人工智能问题独立国际科学小组的初步报告：对人工智能带来的机遇、风险和影响的询证评估》，2026年7月，https://www.un.org/independent-international-scientific-panel-ai/sites/default/files/2026-07/zh_preliminary_report.pdf。

2 《人工智能安全治理框架2.0》，全国网络安全标准化技术委员会（TC260），第2页，2025年9月。

3 同注1 (联合国人工智能问题独立国际科学小组，2026)；另参见 Oxford Martin AI Governance Initiative, "Open Problems in Frontier AI Risk Management", Feb 2026, p. 16, https://aigi.ox.ac.uk/wp-content/uploads/2026/02/Open-Problems-in-Frontier-AI-Risk-Management-Final.pdf.

4 Rand, "Examining risks and response for AI loss of control incidents", July 30, 2025, https://www.rand.org/pubs/research_reports/RRA3847-1.html.

5 新华社：《习近平在中共中央政治局第二十次集体学习时强调：坚持自立自强、突出应用导向、推动人工智能健康有序发展》，2025年4月26日，https://www.news.cn/politics/leaders/20250426/63f5cde8f4b54e22ba35aa7ec7884b3a/c.html。

《中华人民共和国国民经济和社会发展第十五个五年规划纲要》（《“十五五”规划纲要》）也提出“推动建立人工智能全生命周期风险管理制度，健全覆盖安全监测、风险预警、**应急响应**的风险防控体系”。[6]**制度规范层面**，中共中央、国务院2025年2月印发《国家突发事件总体应急预案》，首次将“人工智能安全”纳入突发事件的综合监测体系。[7]而与部分可能演变为“黑天鹅”或极端事件的前沿AI风险相关的是，新修订的《突发事件应急预案管理办法》提出可结合行业或地区风险评估实际，编制“小概率、高风险、超常规极端情形下”的巨灾应急预案。**标准指南层面**，TC260发布的《人工智能安全治理框架2.0》及《生成式人工智能服务安全应急响应指南》也提供了用软性规范实现“敏捷治理”的有益探索。

然而，健全适应前沿AI风险的应急响应体系，不仅需要完善事件发生后的响应处置措施，更需要统筹规划前端治理体系，目前仍有诸多基础性问题有待研究，例如：

- 在**风险格局**层面，前沿AI可能引发什么类型的突发事件或危机？相关风险的演变可能存在何种特殊性，从而提高突发事件的发生概率、传播速度和损害结果？
- 在**标准体系与能力建设**层面，如何对与前沿AI风险相关的突发事件进行分类分级并制定应急预案？与前沿AI风险相适应的应急响应体系中，关键的能力要素有哪些？如何对相关技术与标准进行有效性评估并动态更迭？
- 在**制度定位**层面，前沿AI风险可能给既有制度带来何种挑战？与之相适应的应急响应体系建设应主要依托哪些现行制度与治理框架？属于哪些监管部门管辖？

本报告面向人工智能、网络安全、公共卫生等领域的风险防控与应急处置实务工作者及相关研究人员，从第一个问题切入，分析前沿AI可能如何重塑突发事件的风险格局，并挑战不同领域的既有应急治理措施；以此为基础，围绕第二、三个研究问题探索对策建议。

报告所针对的前沿AI风险与应急响应体系研究尚处起步阶段，且覆盖多个领域。受研究视角与调研广度所限，本报告分析框架、对策建议及相关细节仍有待完善。我们衷心期望这份阶段性研究成果能够引发各界关注与深度研讨，同时为跨领域交流创造契机、搭建平台。故此，报告以征求意见稿形式发布，恳切欢迎各界实务工作者与专家学者提出宝贵意见和反馈。

研究范围

风险类别：在应急体系中，与前沿AI相关的风险可区分为两类。第一类是前沿AI本身作为风险来源或风险放大因素，在网络、公共卫生、金融等领域提高突发事件的发生概率、传播速

[6] 《中华人民共和国国民经济和社会发展第十五个五年规划纲要》，全国人民代表大会，2026 年3月12日，https://www.nda.gov.cn/sjj/swdt/xwfb/0305/20260305164304775031044_pc.html。

[7] 《国家突发事件总体应急预案》，中共中央、国务院，2025年2月25日，3.2.1“监测”，https://www.gov.cn/zhengce/202502/content_7005635.htm。

度与损害结果。第二类是前沿AI作为应急管理的应用工具，在赋能应急管理的过程中，由于人为操作失误或故障等原因产生的负面影响。**本报告重点关注第一类风险**，[8]并在此基础上聚焦突发事件（如流行病）的应对，而非针对远期、缓释风险（如大规模失业）的结构性应对。

突发事件筛选逻辑：前沿AI可能以不同形式、不同程度参与或驱动突发事件的形成。本报告通过两个维度进行筛选，[9]重点关注那些既可能造成严重社会危害、触发高级别应急响应，又因前沿AI能力特征而重塑风险格局，对当前应急响应能力提出新挑战的突发事件。

报告框架

第一部分：围绕前沿AI进行介绍，包括其范围、正向赋能、风险趋势与应急治理现状。

第二部分：从“源头—演化—后果”三个风险演进阶段，分析前沿AI风险如何重塑突发事件风险格局。源头阶段，风险更加“看不清”：前沿AI使恶意行为门槛降低，威胁主体泛化，攻击面显著扩大；此外，模型内生风险增加，模型自身成为独立风险源。**演化阶段，应对可能“来不及”**：前沿AI操作流程高度压缩与任务自动化，可能导致风险规模化并发，并让演化呈非线性与突变性特征。**后果阶段，危害边界更容易“管不住”**：单领域突发事件的危害上限提高以及跨领域级联失效情形，导致损害后果超过社会承载能力。

第三部分：针对具体风险领域进行分析。基于风险可能性、后果严重性和治理可干预性三个维度，报告选取**网络安全、公共卫生和金融安全三个领域**，依托前述“源头—演化—后果”的框架，分析前沿AI在该领域的具体风险趋势及其潜在治理挑战。此外，报告还简要介绍了前沿AI可能驱动的**跨领域复合型突发事件**示例，包括“数字—物理空间”型、“认知—行动干预”型，以及“基础设施同质化导致的系统性故障”型事件。

第四部分：提出对策建议。前沿AI风险与突发事件的应急响应往往超出单一市场主体、单行业监管部门的处置能力与管辖边界，因此需要从整体性治理的视角出发，统筹强化应急响应能力，提升应对韧性。本报告从顶层设计、应急预案、能力建设、运行机制及协同治理五个维度提出十条对策建议，以期为前沿AI灾难性风险防控与危机管理的制度研究与实践探索补充视角、提供思路。

[8] 对于后者，在应急管理体系中引入AI技术，虽可提升效率，但也可能带来决策依赖、信息误判及人机协同失效等问题。该类人工智能风险主要影响应急管理能力本身，而非在社会整体层面重塑突发事件的生成机制和风险格局。它与第一类风险在影响路径和应对机制等方面存在较大不同，本报告不做详述，仅在必要时予以提及。

[9] 详细分析见本报告第二部分：前沿AI如何重塑突发事件风险格局。

贡献与致谢

本报告的主要贡献者

安远AI：陈欣怡（主要撰写人）、方亮、谢旻希、张静昕、李心宁
中国新一代人工智能发展战略研究院：王芳、唐楚雄、张馨月、周思蒙
清华大学人工智能国际治理研究院：肖茜
浙江大学智能安全与智慧应急团队：杜跃进

致谢

衷心感谢清华大学人工智能国际治理研究院院长薛澜在本报告选题上给予的早期启发与关键指引，以及中国新一代人工智能发展战略研究院执行院长龚克对本报告给予的全面战略性指导。

感谢以下专家对具体风险领域提出的宝贵建议：

- **应急管理：**清华大学公共管理学院应急管理研究基地副主任吕孝礼、北京师范大学风险治理创新研究中心主任张强
- **网络安全**：国家计算机病毒应急处理中心研究员张洋、奇安信人工智能公司副总裁刘岩
- **公共卫生**：教育部战略基地“天津大学生物安全战略研究中心”创始主任张卫文、国家卫生健康委前沿生物技术政策研究重点实验室常务副主任薛杨
- **金融安全**：中国人民财产保险股份有限公司风险研究院总经理李晓翾、北京国家金融科技风险监控中心有限公司技术专家郭晓兵

同时感谢孟庆一、梁家铭、安远AI其他同事等对报告的贡献与支持。

目录

第一部分 前沿AI范围及发展趋势

一、前沿AI的范围

本报告所研究的“前沿AI（Frontier AI）”，是指当前处于全球技术前沿、具备较强通用能力或特定领域高能力的人工智能模型与系统，既包括能够在多领域环境中执行多样化任务的通用型人工智能系统（General-Purpose AI，GPAI），也包括在网络安全、生物设计等特定高风险两用（dual-use）领域具备突出能力的专用模型与系统。前沿AI的典型应用形态包括但不限于生成式人工智能模型、智能体、多智能体协作系统等。

前沿AI正在加速与网络空间、关键基础设施、科学研究、金融市场等社会生产和生活领域深度融合。与传统或能力边界较窄的人工智能工具相比，前沿AI通常具有更强的知识泛化、复杂任务分解、多模态生成、环境感知、工具调用、自主规划与持续迭代能力，[10]并已经受到多个地区和国家的监管关注。[11]

二、前沿AI的正向赋能

前沿AI正成为推动社会进步的重要技术力量。其首先拓展了人类的专业知识与能力边界，在复杂推理、代码生成、多模态理解和科学求解等高阶认知任务中，已接近或超越人类基线水平。[12]在此基础上，其自主规划与执行能力也在逐步上升：英国人工智能安全研究所（UK AISI）的一项研究指出，前沿AI在无人类指导下自主执行任务的时长大约每8个月翻一倍，完成长达一小时复杂软件任务的成功率也从2023年末的不足5%激增至2025年的40%以上。[13]这些能力发展让前沿AI工具在复杂系统理解和干预中展现出巨大潜力，并能够在科学研究、医疗健康、网络安全防御、产业智能化升级等场景，赋能经济社会提质增效与治理能力提升。[14]

以应急与危机管理领域为例，面对高度不确定的情境，前沿AI工具可帮助决策者完成多源异构信息统一理解、情景推演、方案生成、指挥调度等，[15]提升决策的科学性与响应效率，已

---

[10] Bengio, Yoshua, et al., "International AI Safety Report 2026" , February 2026, p. 25, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.

[11] See Regulation (EU) 2024/1689 (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024; Transparency in Frontier Artificial Intelligence Act, S.B. 53, 2025-2026 Reg. Sess. (Cal. 2025); Responsible AI Safety and Education Act, S. 6953-B, 2025-2026 Reg. Sess. (N.Y. 2025).

[12] 同注10 (Bengio, Yoshua, et al., 2026)。

[13] UK AI Safety Institute, "Frontier AI Trends Report" (2025), https://www.aisi.gov.uk/frontier-ai-trends-report.

[14] Stanford HAI, "Artificial Intelligence Index Report 2026" (2026), https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf.

[15] 许欢：《以人工智能应急大模型赋能现代化应急管理》，人民政协网，2026年6月26日，https://www.rmzxw.com.cn/c/2026-06-26/3938477.shtml。

成为我国人工智能赋能治理的重要战略发展方向。[16]例如，在自然灾害监测预警方面，近年来的华为盘古气象大模型[17]、上海人工智能实验室风乌大模型[18]、GenCast[19]等新一代天气预测模型，可显著提升极端天气的预测精度、时效和空间分辨率。在突发公共卫生事件的应对方面，国内外多个机构联合共建“中国—葡萄牙人工智能与公共卫生技术‘一带一路’联合实验室”，研发大规模流行病时空预测模型，支撑跨国跨境协同监测与早期预警。[20] 此外，流行病防范创新联盟（CEPI）发起“百天使命”倡议（The 100-Day Imperative），以期在识别出具有大流行潜力的新型病原体后100天内完成疫苗、诊断试剂和治疗药物的研发与应急部署。该倡议一方面将前沿AI作为缩短研发周期、提升全球监测预警能力的重要技术支撑，另一方面将前沿AI带来的新型生物安全风险纳入治理框架，获得G7和G20领导人共识，为构建适应前沿AI时代的全球公共卫生应急体系提供了重要探索。

三、前沿AI的风险趋势

在前沿AI提供多元正向价值的同时，技术迭代伴随的安全风险与隐患也已经显现，并已得到多方关注。近年来，前沿AI模型在应用场景中的风险和威胁案例已被广泛记录。[21]2026年6月，麻省理工学院（MIT）发布了一份专项调研报告。该研究对全球30余个国家的270余位人工智能领域专家开展访谈调研，一定程度上体现了对风险可能性的共识判断。报告识别了24类风险，并指出：若前沿AI技术不加干预地照常发展，其中有18类风险在未来五年内有至少10%的概率引发灾难性后果。按该报告定义，灾难性后果指“造成100万人以上死亡，或是带来超1000亿美元经济损失”。其中滥用、危险能力涌现等风险尤为受到关注。[22]2025年，部分前沿AI模型开发商已主动上调其模型在网络攻击、生化安全等领域的安全防护等级。[23]2026

---

16 国务院《关于印发〈现代化应急体系建设 “十五五” 规划〉的通知》，2026 年 6 月 8 日，https://www.gov.cn/zhengce/zhengceku/202606/content_7071452.htm。

17 Bi K, Xie L, Zhang H, et al., "Accurate medium-range global weather forecasting with 3D neural networks", Nature, 2023,619: 532-538. https://doi.org/10.1038/s41586-023-06185-3.

18 上海人工智能实验室：《“风乌”大模型发布，全球气象有效预报时间首破10天》，2023年4月7日，https://www.shlab.org.cn/news/5443382。

19 Price I, Sanchez-Gonzalez A, Aaltonen J, et al., "Probabilistic weather forecasting with machine learning", Nature , 2024. https://doi.org/10.1038/s41586-024-08252-9.

20 广州医科大学：《中国—葡萄牙人工智能与公共卫生问题“一带一路”联合实验室》，https://www.gzhmu.edu.cn/info/1124/52601.htm。

21 AI Incident Database, "Entities Index," accessed July 9, 2026, https://incidentdatabase.ai/entities/.

22 MIT FutureTech, "Prioritization of Risks from Artificial Intelligence: A Delphi Study of 272 International Experts", June 2026, https://cdn.prod.website-files.com/669550d38372f33552d2516e/6a172558bd2947234379749f_a8684052fd49a64374c9a9d3e4e5ab59_Prioritizing%20the%20risks%20from%20Artificial%20Intelligence.pdf.

23 Anthropic, "Activating AI Safety Level 3 protections", May 22, 2025, https://www.anthropic.com/news/activating-asl3-protections; Jakob Steinschaden, "GPT-5 classified as high risk in terms of biological and chemical weapons", August 8, 2025, https://www.trendingtopics.eu/gpt-5-classified-as-high-risk-in-terms-of-biological-and-chemical-weapons/.

年4月以来，一些具备高风险能力的前沿AI模型或系统转向更为谨慎的发布策略[24]：例如，Anthropic的Claude Mythos Preview因展现出较强的漏洞发现与利用能力，初期仅向少数经过筛选的组织提供受限访问；[25]微软的MDASH系统同样因其强大的漏洞自动化发现、验证和利用能力，仅限内部使用和少数客户有限私有预览；[26]OpenAI的GPT-Rosalind则因具备较强的生命科学推理分析与药物研发等能力，采取了需经审核的可信访问机制。[27]

未来，前沿AI开源生态的深化可能让风险态势更加复杂。开源模型大幅降低了前沿AI技术的开发与部署门槛，在推动技术普惠的同时，也提升了模型滥用、恶意微调、违规部署的潜在风险。[28]2026年4月起，中央网信办部署开展“清朗·整治AI应用乱象”专项行动，提出将重点整治“开源模型安全管理不到位”的问题，并明确指出“目前开源社区缺乏身份认证和安全管理机制，对社区上传的数据集、模型代码等内容未建立有效审核和应急处置机制，未及时清理存在严重风险隐患的数据集或开源模型。”[29]随着前沿AI能力持续通过开源生态扩散，其风险能力也同步扩散，风险暴露面和潜在影响范围也将进一步扩大。

与此同时，智能体研发与落地的加速，[30]也带来不容忽视的治理挑战。[31]研究指出，由于智能体具备自主规划、工具调用、跨环境执行等能力，其风险敞口相应扩大；当事件发生时，人类往往缺乏有效的实时干预窗口。[32]今年5月印发的《智能体规范应用与创新发展实施意见》明确提出需要防范智能体不当行为引发重大风险，具体的安全风险包括数据投毒、隐私泄露、算法篡改、系统漏洞、运行失控等。[33]未来，随着多智能体系统的进一步发展，其能力上限可能显著高于单一智能体，安全防护也可能更具挑战性。[34]

---

[24] Nikita Ostrovsky, "'Too Dangerous to Release' Is Becoming AI's New Normal," Time, April 24, 2026, https://time.com/article/2026/04/24/claude-mythos-chatgpt-rosalind-release-dangerous/.

[25] "Anthropic Claims Its New A.I. Model, Mythos, Is a Cybersecurity 'Reckoning'", The New York Times, April 7, 2026, https://www.nytimes.com/2026/04/07/technology/anthropic-claims-its-new-ai-model-mythos-is-a-cybersecurity-reckoning.html?eafs_enabled=false.

[26] Taesoo Kim, "Defense at AI speed: Microsofts new multi-model agentic security system tops leading industry benchmark", May 12, 2026, https://www.microsoft.com/en-us/security/blog/2026/05/12/defense-at-ai-speed-microsofts-new-multi-model-agentic-security-system-tops-leading-industry-benchmark/ .

[27] Introducing GPT-Rosalind for life sciences research, Open AI, April 16, 2026, https://openai.com/index/introducing-gpt-rosalind/.

[28] 安远AI、北京大学人工智能研究院、北京大学武汉人工智能研究院、北京通用人工智能研究院：《基础模型的负责任开源》，2024年4月，https://concordia-ai.com/wp-content/uploads/2024/04/基础模型的负责任开源.pdf。

[29] 《智能体规范应用与创新发展实施意见》，国家网信办、国家发展改革委、工业和信息化部，2026年5月8日，https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm。

[30] Massachusetts Institute of Technology (MIT), "The AI Agent Index", 2025 Edition, https://aiagentindex.mit.edu.

[31] Sophia Wang：《智能体安全：2026年AI落地过程中最容易被低估的治理挑战》，2026年4月27日，International Data Corporation，https://www.idc.com/resource-center/blog/智能体安全：2026年ai落地过程中最容易被低估的治理/。

[32] 同注10 (Bengio, Yoshua, et al., 2026)。

[33] 《智能体规范应用与创新发展实施意见》，国家网信办、国家发展改革委、工业和信息化部，2026年5月8日，https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm。

[34] 同注1 (联合国人工智能问题独立国际科学小组，2026)，第35页。

近期“OpenClaw”风险事件即是智能体与开源风险叠加的典型案例。2026年3月，国家互联网应急中心、国家网络与信息安全信息通报中心等多家机构先后发布该工具的安全风险预警，指出其存在大量高危漏洞，并存在配置缺陷、提示词注入与指令劫持、自主决策偏差与不可逆操作、第三方生态恶意投毒等多种风险。[35]因部署门槛低、扩展能力强，该开源智能体在安全防护缺失的条件下快速扩散。大量用户在未做安全加固的情况下自行部署，已引发凭证泄露、权限提升攻击、持久化后门植入等实际安全事件，[36]成为前沿AI风险逐步向现实网络安全危害传导的典型样本。

四、前沿AI风险的应急治理现状

1、前沿AI的潜在灾难性风险已进入全球治理视野

国际方面。联合国减少灾害风险办公室（UNDRR）的研究报告指出，人工智能正日益成为推动存在性风险的重要驱动因素。在全球高度相互依赖的结构中，局部事件更易迅速外溢并演变为极端性全球灾难。[37]2026年6月，联合国人工智能独立国际科学小组警告称，人工智能“可能自行或因恶意用户而造成灾难性危害”，同时该技术“发展速度已超过科学界的理解能力以及各国政府的适应能力”。[38]欧盟、美国、韩国等地已经针对高风险、高影响力的人工智能模型和/或系统出台专门的监管措施。[39]其中，美国加州《SB-53法案》以前沿AI模型的灾难性风险为核心监管对象，并设定了可量化的损害阈值，专门围绕前沿AI引发50人以上伤亡或10亿美元以上重大财产损失的风险可能性采取监管措施，建立安全事件报告等制度。

国内方面。2025年2月发布的《国家突发事件总体应急预案》正式将人工智能安全纳入监测预警环节。作为应急预案中的纲领性文件，该预案“适用于党中央、国务院应对特别重大突发事件工作”，[40]而“特别重大”是我国突发事件分级中最为严重的一级，其在不同领域的阈

---

[35] 国家互联网应急中心：《关于OpenClaw安全应用的风险提示》，2026年3月12日，https://www.cert.org.cn/publish/main/11/2026/20260312144519429724511/20260312144519429724511_.html；工信部：《关于防范OpenClaw（“龙虾”）开源智能体安全风险的“六要六不要”建议》，2026年3月11日，https://nvdb.org.cn/publicAnnouncement/2031684972835299329。

[36] Cloud Security Alliance, "Claw Chain: Four CVEs Enable Full AI Agent Compromise", May 17, 2026, https://labs.cloudsecurityalliance.org/wp-content/uploads/2026/05/CSA_research_note_openclaw-claw-chain-cve_20260517-csa-styled.pdf.

[37] UNDRR, "Existential Risk and Rapid Technological Change : Advancing risk-informed development" (2023), p. 14， https://www.undrr.org/media/86500/download?startDownload=20260426; UNDRR, "Hazards with Escalation Potential : Governing the Drivers of Global and Existential Catastrophes " (2023), p. 26, https://www.undrr.org/media/89456/download?startDownload=20260426.

[38] 联合国：《从人工智能到“杀人机器人”：联合国秘书长发出紧急治理呼吁》，2026年7月6日，https://news.un.org/zh/story/2026/07/1142428。

[39] Act on the Development of Artificial Intelligence and Establishment of Trust [AI Basic Act], ch. 1, art. 2 (Republic of Korea, 2026); Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024; Transparency in Frontier Artificial Intelligence Act, Cal. S.B. 53 (2025); Responsible AI Safety and Education (RAISE) Act, N.Y. Laws of 2025.

[40] 《国家突发事件总体应急预案》，1.2 “适用范围”。

值不同（参见本节附表：国家突发事件分级分类标准），仅以人员伤亡和经济损失的相关阈值为参考，其通常对应30人以上死亡、1亿元以上直接经济损失。此外，相较于其1.0版本，TC260《人工智能安全治理框架》2.0版新增了“灾难性风险”这一表述，并设置了“特别重大”这一安全风险级别，特指“具有灾难性和系统性威胁特征，对国家安全、社会秩序和公民权益造成颠覆性或不可逆转的特别严重的影响。”[41]这些文件显示，我国已初步认识到前沿AI具备驱动严重突发事件甚至危机的现实风险。

2、构建适配前沿AI风险的应急响应体系正在进入我国治理议程

顶层战略层面，2025年4月，习近平总书记在中共中央政治局第二十次集体学习时，提出要构建人工智能应急响应体系，确保其安全、可靠、可控。[42]《“十五五”规划纲要》则进一步提出推动建立人工智能全生命周期风险管理制度，明确指出需要健全覆盖“应急响应”的人工智能风险防控体系。

法律制度与规范性文件层面，我国应急体系的基本法律依据为《中华人民共和国突发事件应对法》（以下简称“《突发事件应对法》”），[43]其确立了突发事件的基本分类框架（参见本节附表：国家突发事件分级分类标准）。尽管《国家突发事件总体应急预案》已将其纳入监测环节，但“人工智能安全”并未直接体现在突发事件分类框架之中。然而，制度层面的两个动向可能与前沿AI风险应急响应相关，值得进一步关注：

第一，针对巨灾应急。国务院办公厅2024年1月印发《突发事件应急预案管理办法》，补充了“小概率、高风险、超常规极端情形下”的巨灾应急预案编制规定，提出可结合行业或地区风险评估实际，制定的巨灾应急预案。尽管该规定并非专为前沿AI风险所设，但前沿AI引发的部分安全事件确实与巨灾的特征相契合。[44]故此，目前巨灾应急的相关制度框架，可能为构建适配前沿AI风险的应急响应体系提供良好基础。

第二，部分行业已经开始关注前沿AI应用风险及配套应急能力建设。例如，国家金融监督管理总局出台了《关于银行业保险业人工智能安全开发应用的指导意见》，要求针对金融人工智能应用“全面提升安全防护和应急处置能力”，具体要求如“将人工智能应用纳入业务连续性管理体系，开展业务影响分析，制定应急预案，加强安全运行管理、事件处置和容灾能力建设。”[45]

---

41 《人工智能安全治理框架2.0》，第25页。

42 《习近平在中共中央政治局第二十次集体学习时强调：坚持自立自强、突出应用导向、推动人工智能健康有序发展》，共产党员网，2025年4月26日，https://www.12371.cn/2025/04/26/ARTI1745635107727513.shtml。

43 《中华人民共和国突发事件应对法》，中华人民共和国主席令第二十五号，2024年11月11日，第一章第二条。

44 《人工智能安全治理框架2.0》，第2页、第4页、第25页。

45 《关于银行业保险业人工智能安全开发应用的指导意见》，2026 年 6 月 18 日，https://www.nfra.gov.cn/cn/view/pages/governmentDetail.html?docId=1261784&itemId=&generaltype=1。

在标准指南层面，TC260在其发布的《人工智能安全治理框架》2.0版中提出“建立人工智能应用运行监测能力和安全事件应急预案”“定期开展应急演练”和“及时优化应急策略”等，[46]为探索构建适应前沿AI的应急响应体系提供框架性认同。此外，TC260还出台了《生成式人工智能服务安全应急响应指南》。[47]这份指南提供了人工智能风险应急响应方面的软性治理探索。但是，这份指南的适用对象是生成式人工智能的应用服务安全，适用主体则聚焦相关服务的提供者。故此，就构建适应前沿AI风险的国家应急体系而言，目前标准指南层面的探索还较为有限。

[46] TC260：《人工智能安全治理框架2.0》，第20页。

[47] TC260：《网络安全标准实践指南—生成式人工智能服务安全应急响应指南》，2025年9月。

本节附表：国家突发事件分级分类标准

1、突发事件分类

我国《突发事件应对法》将突发事件划分为四大类型，《国家突发事件总体应急预案》则在各类型下列举了更多类别。[48]

类型	示例
自然灾害	水旱、气象、地震、地质、海洋、生物灾害和森林草原火灾等
事故灾难	工矿商贸等生产经营单位的各类生产安全事故，交通运输、海上溢油、公共设施和设备、核事故，火灾和生态环境、网络安全、网络数据安全、信息安全事件等
公共卫生事件	传染病疫情、群体性不明原因疾病、群体性中毒，食品安全事故、药品安全事件、动物疫情，以及其他严重影响公众生命安全和身体健康的事件
社会安全事件	刑事案件和恐怖、群体性、民族宗教事件，金融、涉外和其他影响市场、社会稳定的突发事件

表1.1 突发事件分类

2、特别重大突发事件分级标准（示例）

我国根据社会危害程度、影响范围等因素，将自然灾害、事故灾难和公共卫生事件划分为特别重大、重大、较大与一般四级，社会安全事件的分级标准另行规定。[49]突发事件的级别将直接决定预警标准，并影响响应程度和处置资源投入等关键决策。[50]

其中，不同领域“特别重大”突发事件的分级标准（见下表）与本报告的分析较为相关。作为突发事件分级体系中的最高一级，其阈值构成了我国识别极端冲击的关键基准。该阈值不仅是事件定级依据，也可以作为评估前沿AI是否显著改变突发事件风险结构的重要参照。如果前沿AI的介入显著降低了相关事件跨越这一高等级阈值的难度，则意味着前沿AI的影响已超出一般技术风险范畴，并可能导致既有应急响应措施的失效。

[48] 《国家突发事件总体应急预案》，1.3 “突发事件分类分级”。

[49] 《突发事件应对法》，第一章第三条。

[50] 闪淳昌、薛澜主编：《应急管理概论：理论与实践》（第二版），第67页。

指标纬度	“特别重大”突发事件阈值（示例）	相关突发事件类型/种类
人员伤亡	30人以上死亡，或100人以上重伤	事故灾难类事件，含生产安全、交通运输
	在一起突发中毒事件中，中毒人数在100人以上且死亡10人以上，或者死亡30人以上	公共卫生类事件
经济损失	1亿元以上直接经济损失	事故灾难类事件，含网络安全、生产安全、交通运输
传染病疫情	发生新传染病或我国尚未发现的传染病发生或传入，并有扩散趋势，或发现我国已消灭的传染病重新流行	公共卫生类事件
关键系统中断	关键信息基础设施整体中断运行 6 小时以上或主要功能中断运行 24 小时以上	事故灾难类事件，如网络安全；社会安全类事件，如金融市场稳定性等
	影响一个或多个省级行政区 50% 以上人口，或者 1000 万人以上用水、用电、用气、用油、取暖、交通出行、就医、购物等工作、生活	事故灾难类事件，如网络安全；社会安全类事件，如金融市场稳定、社会稳定等
数据安全	泄露 1 亿人以上公民个人信息	事故灾难类事件，如网络安全 社会安全类事件，如金融安全

表1.2 特别重大突发事件分级标准（示例）

第二部分 前沿AI如何重塑突发事件风险格局

突发事件筛选逻辑

应急体系的主要治理对象是“突发事件”，而前沿AI风险可能以不同形式、不同程度参与突发事件的形成。本报告从以下两个维度进行筛选，重点关注既具有严重影响后果，又因前沿AI能力特征而呈现出新的风险格局，对当前应急响应能力可能形成挑战的突发事件。

第一，从**风险结构**来看，本报告重点关注那些因前沿AI介入而重塑风险格局的突发事件。相较于传统人工智能或其他技术工具，前沿AI可能显著改变风险形成与演化路径，从而提高突发事件发生概率、传播速度或损害结果，对当前应急体系与响应能力造成挑战。

第二，从**影响后果**来看，本报告重点关注可能造成严重社会危害、触发高级别应急响应的突发事件。现行各应急预案中“特别重大”突发事件的标准，可作为研判前沿AI风险程度的参照系（相关标准参加第一部分附表：国家突发事件分级分类标准）。

分析框架

不同类型突发事件由于致灾机理、传播路径等存在差异，其具体演化机制并不相同，当前学界尚未形成适用于各类突发事件的统一演化框架。[51]但部分理论研究认为，突发事件通常涉及从孕育、发生、发展到终结的动力学过程。[52]与此同时，风险管理等理论则关注事件发生之前及其后的风险传导过程，通常围绕风险源、潜在事件、事件后果及其发生可能性等要素对“风险”进行描述。[53]

两类研究虽然关注对象不同，但共同揭示了风险从源头形成、触发事件到产生后果的动态演化规律。其中，事件可被理解为风险在特定条件下的具体呈现，既承接风险源的持续作用，也是风险后果形成的重要中介。结合这两种视角，本报告对风险演进过程适度抽象，将其概括为**“源头（Source）—演化（Evolution）—后果（Consequence）”**，其中的“演化”阶段不仅包括风险的扩散与传导，也涵盖风险触发具体突发事件及事件持续发展的过程，即将微观层面的突发事件置于整体风险演化链条之中，从更全局性的视角分析前沿AI如何改变风险的形成、传播及影响路径，重塑突发事件风险格局。

---

51 张海涛、李佳玮、周红磊等：《重大突发事件演变机制：认知框架与理论方法》，载《情报学报》，2021年第9期。

52 钟开斌、薛澜：《中国特色应急管理理论研究的基本图谱》，载《中国社会科学》，2025年第1期；马建华、陈安：《突发事件的演化模式分析》，载《安全》，2009年第12期。

53 GB/T 23694-2024《风险管理 术语》，3.1.1“风险”。

阶段	前沿AI风险趋势	潜在治理挑战
源头	• 恶意行为门槛降低，威胁主体泛化，攻击面显著扩大 • 模型内生风险增加，模型自身成为独立风险源	**“看不清”：** 源头风险识别、风险评估和监测预警复杂性提升
演化	• 流程高度压缩与任务自动化 • 多点并发导致规模化影响增强 • 风险演化的非线性与突变特性	**“来不及”：** 监测预警、风险评估和应急响应的时间窗口进一步缩短
后果	• 单领域危害上限提高 • 跨领域级联失效与系统性风险加剧	**“管不住”：** 已知风险抬升及未知风险引入，导致应急预案与遏制措施储备不足、治理协同决策复杂性提升

表2.1 “阶段—前沿AI风险趋势—潜在治理挑战”对照表

一、源头阶段“看不清”

阶段含义

风险源通常指可能单独或共同引发风险的各种要素。[54]在前沿AI的语境下，风险源可能包括模型危险能力（如网络攻击）、模型特性（如幻觉）、系统配置、部署环境、人机交互以及恶意滥用等因素。[55]在源头阶段，应急体系的核心治理目标是对风险源进行识别，并通过风险监测、早期预警与及时干预等手段，从源头降低风险向有害事件转化的可能性。

前沿AI风险趋势

- **恶意行为门槛降低，威胁主体泛化，攻击面显著扩大。** 前沿AI凭借知识生成、代码编写、任务规划及自动化执行等能力，将原本依赖长期训练、专业经验与高资源投入的复杂操作扩散至中低能力主体，显著降低了恶意行为体实施网络攻击、高精度社会工程学攻击、有害生物设计的门槛。此技术扩散效应导致潜在威胁主体数量大幅增加。同时，前沿AI基于海量数据分析、跨领域知识关联与仿真推演，能更高效地识别系统脆弱点，使网络供应链、关键岗位及生物实验流程等环节中的隐蔽风险点暴露更为彻底，攻击面随之急剧扩大。[56]

[54] GB/T 23694-2024《风险管理 术语》，3.3.10“风险源”。

[55] Oxford Martin AI Governance Initiative, "Open Problems in Frontier AI Risk Management"， page 14, Feb 2026, https://aigi.ox.ac.uk/wp-content/uploads/2026/02/Open-Problems-in-Frontier-AI-Risk-Management-Final.pdf.

[56] 具体分析见本报告第三部分 领域分析：前沿AI风险格局与治理挑战。

- **模型内生风险增加，模型自身成为独立风险源。** 前沿AI模型内部决策过程具有高度复杂性与黑箱特征，可解释性受限。在某些场景中，模型可能出现目标偏离、异常输出或系统功能紊乱，极端情况下甚至产生超出预期的自主失控行为。这使得模型自身成为潜在的独立风险源，显著加剧了风险识别与评估的复杂性。

潜在治理挑战

风险源不确定性的提升，对源头风险识别、评估及监测预警提出了严峻挑战。传统的监测和响应范式以高危能力主体有限为关键假设，主要聚焦重点场景中的高风险主体与高危行为，可能难以覆盖随前沿AI能力演进而出现的大量威胁主体与新型威胁源，从而阻碍风险的主动防控、动态监测与早期预警。

二、演化阶段“来不及”

阶段含义

风险演化是指风险源由潜在状态逐步发展为现实突发事件，并持续扩散、升级直至造成实际损害的动态过程。在演化阶段，应急体系的核心治理目标是及时发现风险演化迹象，强化实时监测、早期预警和及时响应能力，从而在风险演化初期切断其升级链条，遏制事件的扩散与规模扩大。

前沿AI风险趋势

- **流程高度压缩与任务自动化。**前沿AI具备对复杂任务进行拆解与端到端执行的能力，可能将原本依赖人工协同、耗时冗长的风险活动演化为半自动或自动化流程，显著缩短风险行为的实施周期。例如，在网络攻击中，前沿AI可高效串联漏洞扫描、攻击路径构建、代码编写及投放优化，将原本需要多阶段人工干预的攻击链条压缩为连续化的自动化过程，极大提升了从漏洞发现到攻击实施的时效性。
- **多点并发导致规模化影响增强。**在部分场景下，前沿AI可能导致风险行为多路径并行，促使风险活动从单点、低频的定向实施，转向多点并发、批量化的规模扩张，显著提升了单位时间内潜在风险行为的覆盖密度与破坏规模。例如，在网络空间，这可能表现为恶意行为体在多个节点的同步渗透攻击。在金融领域，则可能表现为多机构同质化交易与风控模型同步触发一致性操作，引发跨市场多品类资产共振波动与流动性收缩。在认知干预层面，则可能出现恶意主体利用前沿AI批量生成虚假的医疗或财经市场信息，迅速引发挤兑，影响社会秩序与稳定。

- **三是风险演化的非线性与突变特性。**前沿AI使风险演化呈现出“非线性突变”，而非传统的“渐进式加重”特征。[57]自动化决策、高速传播和大规模执行能力等前沿AI风险能力，可能使得风险在更短时间内完成积聚并突破关键阈值，使事件更快由局部风险演变为严重危机，压缩监测预警和应急响应窗口。

潜在治理挑战

风险演化速度的压缩，对传统依赖人工研判、层级决策和渐进式响应的应急响应模式提出挑战。一方面，危机升级速度显著加快，人工分级审批和决策流程可能难以及时响应。另一方面，风险演化路径更加复杂且不确定，超出历史经验范畴，导致既有应急预案和响应范式可能难以有效适配新型风险情景。[58]

三、后果阶段“管不住”

阶段含义

后果通常指事件对风险管理目标产生的影响。[59]在前沿AI的语境下，后果主要表现为前沿AI驱动的突发事件对人身财产安全、关键基础设施、经济与社会秩序、国家安全等造成的实际损害，以及由此引发的跨领域、跨区域级联影响。在后果阶段，应急体系的核心治理目标是控制事件影响范围和损害程度，防止风险进一步扩散和级联演化，并推动受影响系统尽快恢复正常运行。

前沿AI风险趋势

- **单领域危害上限提高。**前沿AI显著增强复杂系统的设计、推演与功能优化能力，极大地提升了高风险技术路径的产出效率，显著增加了突破既有安全阈值的成功概率。例如，在生物安全领域，前沿AI在病原体结构优化与功能增强方面的赋能，可能被用于增强已知病毒甚至创造未知病毒，导致病毒致病性或传染性增强，提升单次公共卫生突发事件的潜在损害上限。
- **跨领域级联失效与系统性风险加剧。**随着网络空间算力资源与关键业务向云计算平台、大型数据中心等数字基础设施高度集中，一旦底层算力平台或核心云基础设施受到攻击，或是因故障、人为操作失误等因素停摆，历史事件表明其安全边界的溃决将不再局限于单一实体，而是会沿着高度耦合的上下游产业链进行级联式传导。[60]在前沿

[57] 同注55 (Oxford Martin AI Governance Initiative, 2026), page 16。

[58] 同注55 (Oxford Martin AI Governance Initiative, 2026), page 16。

[59] GB/T 23694-2024《风险管理 术语》，3.3.10“风险源”。

[60] Microsoft, "Azure status history", July 2024, https://azure.status.microsoft/status/history/?trackingId=KTY1-HW8&utm.

AI的语境下，基础模型、训练数据及软件组件的广泛复用，使不同系统和行业之间形成更强的技术同源性。一旦基础模型存在共性缺陷、共享组件出现漏洞或关键节点发生失效，相关风险可能在多个系统中同步暴露。例如，同一基础模型若存在共性缺陷，可能在金融、交通、公共服务等多个领域同时触发异常，进而引发跨领域连锁反应，将单点事故演化为社会层面的系统性冲击。

潜在治理挑战

在前沿AI风险趋势下，损害后果可能突破既有应急处置能力边界，事件遏制与恢复难度显著增加。损害后果上限可能被突破，加剧事态遏制与恢复难度。一方面，前沿AI可能推动既有风险突破传统危害上限，使单次突发事件造成更大范围、更高烈度的人员伤亡、经济损失或关键功能瘫痪，现有应急资源配置、救援能力和恢复能力可能面临更大压力。另一方面，前沿AI还可能催生超出既有认知的新型风险，使现有应急预案、处置流程和技术手段难以覆盖未知情景，事件初期容易出现风险识别不足、处置策略缺失和决策依据有限等问题，导致事态迅速突破治理体系的承载边界。与此同时，同源模型漏洞、共享基础设施失效等因素可能导致风险在多个行业和领域同步扩散，传统以单部门、单领域为主的应急指挥和协同机制难以及时统筹跨领域资源，增加了阻断级联传播、恢复关键功能和维护社会稳定的难度。

第三部分 领域分析：前沿AI风险格局与治理挑战

本报告选择网络、公共卫生和金融领域作为风险领域示例进行详细分析。尽管这三个领域在治理逻辑、制度结构与风险形态上存在显著差异，然而，这三个领域**在风险可能性、后果严重性与治理可干预性三个维度**上呈现出高度共性，从而构成当前分析前沿AI风险演化及其对应急响应机制挑战的关键切入点，以下详细分析：

风险可能性

前沿AI可能在网络安全、公共卫生与金融安全三大关键领域引发的灾难性事件或造成的系统性风险，已经为国际研究与政策讨论所高度关注：

- **网络安全领域**，随着Mythos等高风险模型的出现，前沿AI在网络攻防领域的能力跃迁已促使网络安全专家和监管机构重新评估传统防御体系的有效性。2025年，人工智能安全国际对话伦敦会议（IDAIS-London）将网络安全识别为当前AI风险最为集中的重点领域，认为现有治理体系已出现明显能力缺口，需要加强国际协同应对。[61]同时，MIT面向30余个国家、270余名AI领域专家开展的调研显示，网络攻击能力提升被普遍视为未来前沿AI最值得关注的几个风险方向。[62]
- **公共卫生领域，**前沿AI对病原体设计、实验方案优化、实验故障排查等生物能力的赋能，使其潜在风险逐渐由学术界讨论转向现实治理议题。生物安全是上述IDAIS-London会议识别的另一前沿AI重点风险领域。此外，核威胁倡议组织（NTI）与慕尼黑安全会议围绕前沿AI与生物技术融合开展桌面推演，模拟前沿AI加剧全球灾难性生物事件的风险演化过程，并据此发布《防范全球性灾难：人工智能与生物学交叉领域的风险、机遇与治理方案》，[63]共同反映出国际社会已开始将相关风险作为现实安全挑战纳入应急准备和治理体系。
- **金融安全领域**，前沿AI的网络攻击能力的发展同样引发了监管部门和国际金融组织的高度重视。MIT上述调研将金融业识别为受前沿AI冲击最为脆弱的行业之一，反映出金融体系在高度数字化、自动化和算法依赖背景下面临更大的风险敞口。与此同时，国际清算银行（BIS）、金融稳定理事会（FSB）等国际机构近年来持续提示，前沿AI可能通过算法同质化、自动化交易、第三方模型集中依赖以及网络攻击等途径放大系统

[61] International Dialogues on AI Safety, "IDAIS-London Statement," April 2026, https://idais.ai/dialogue/idais-london/.

[62] 同注22 (MIT FutureTech, 2026)。

[63] Nuclear Threat Initiative (NTI), "Safeguarding Against Global Catastrophe: Risks, Opportunities, and Governance Options at the Intersection of Artificial Intelligence & Biology", December 2025, https://www.nti.org/wp-content/uploads/2025/12/2025-NTI-BIO-TTX-Report-FINAL.pdf.

性金融风险；OpenClaw等开源智能体框架能力的快速演进，则进一步强化了各国对于前沿AI在核心金融业务中安全应用的关注。

后果严重性

《新时代的中国国家安全》白皮书明确将网络安全、生物安全与系统性金融风险防范纳入国家安全体系[64]。本报告选取的网络、公共卫生与金融三个领域，分别对应国家关键信息基础设施、人民生命健康与社会经济稳定的底线，具有高度的代表性和现实紧迫性。这些领域一旦发生重大突发事件，可能造成严重甚至具有长期性与结构性影响的危害后果，显著冲击国家安全与社会运行秩序。

治理可干预性

三个领域均具备相对成熟或已有基础的应急响应体系，但前沿AI的介入可能使既有的响应体系出现能力缺口。需要特别指出的是，前沿AI的通用性失控风险目前在全球学界和业界仍存在较大争议与不确定性，而应急治理工具的有效运用高度依赖于具体的风险领域、触发场景和作用链条；若缺乏具象化的风险锚点，应急预案的编制、监测预警与响应机制均难以落地。因此，选择网络、公共卫生与金融这三个具备相对成熟应急框架与明确场景边界的领域，为检验和升级传统应急治理工具提供了最具现实意义的试验场——既有基础框架可依托，又有明确的补强需求。

需要说明的是，这三个领域并不意在、也难以穷尽前沿AI相关的全部突发事件类型，而是意在通过这些示例领域的分析，引发对前沿AI风险与其应急响应的关注与讨论，推动相关研究与制度落地。未来，随着技术演进和实践深化，更多领域可能被纳入进一步研究。

本部分前三节的分析在统一框架下递进式展开，依次从（1）既有制度抓手；（2）现行制度能够发挥有效性的关键底层假设；（3）前沿AI新增风险点，来论证（4）治理挑战。

第四节则跳出单一突发事件领域，分析跨领域复合型的突发事件及其风险与挑战。《国家突发事件总体应急预案》明确指出四大类型的突发事件“往往交叉关联、可能同时发生，或者引发次生、衍生事件，应当具体分析，统筹应对”。[65]在现实情境中，各类突发事件往往相互关联和影响，呈现跨类型耦合的常态特征。[66]在前沿AI时代，复合型事件的演化将更加复杂，因此有必要加强关注、识别关键的制度漏洞并进行预防性治理。

---

64 国务院新闻办公室：《新时代的中国国家安全》白皮书，2025 年 5 月，http://www.scio.gov.cn/zfbps/zfbps_2279/202505/t20250512_894771.html。

65 《国家突发事件总体应急预案》1.3“突发事件分类分级”。

66 参见薛澜、钟开斌：《突发公共事件分类、分级与分期：应急体制的管理基础》，载《中国行政管理》，2005年第2期；闪淳昌、薛澜主编：《应急管理概论：理论与实践》（第二版），第20页。

一、网络安全

前沿AI使得网络空间成为新技术安全风险传导的首要载体，网络安全也因此成为当前受前沿AI技术冲击最为直接的领域。近期Claude Mythos Preview、开源智能体OpenClaw等代表性事件接连出现，分别反映出闭源高能力模型攻防能力跃迁与开源智能体风险快速扩散两类趋势，标志着前沿AI驱动的网络安全风险已从技术预判进入现实应急处置视野。随着各行业数字化、智能化程度持续提升，业务与系统间的耦合依赖关系日益复杂，单点网络安全事件极易引发跨行业连锁反应，进而可能触发社会或国家层面的安全危机。

《国家网络安全事件报告管理办法》对网络安全事件进行了分类、分级，其中特别重大网络安全事件的量化标准（参见本节附表：国家网络安全事件分级分类标准）构成了该领域的应急红线，可以作为研判前沿AI风险的参照系。

（一）源头阶段：异常行为监测更加困难

现有制度抓手

当前的网络安全应急体系高度强调源头防范，在**攻击方监测**方面，现有体系通常以异常行为识别与持续态势监测为主要技术路径，通过漏洞全生命周期管理[67]、安全日志审计与关联分析[68]、入侵行为特征检测等手段，构建常态化威胁感知能力。在**防护体系维度**，我国采用层级化、差异化的资源配置策略[69]，将核心防护力量与监管资源向关键信息基础设施及核心业务系统倾斜。

关键假设

上述制度抓手主要建立在两项基本假设之上：

高危攻击能力相对稀缺。在传统资源约束下，能够触发特别重大网络安全事件的攻击主体通常需要具备接近国家级行为体或高级持续性威胁（APT）组织的综合能力，包括全链路渗透攻击技术、长期情报侦察、大规模资源投入以及复杂攻击组织能力，攻击准备和实施周期往往长达数月甚至数年；其攻击行为具备一定可观测性，因此能够围绕重点目标开展差异化防护和持续监测。

---

67 《网络产品安全漏洞管理规定》，2021年9月1日；GB/T 30276—2020《信息安全技术 网络安全漏洞管理规范》。

68 《中华人民共和国网络安全法》，2017年6月1日，第二十三条第（三）款。

69 《关键信息基础设施安全保护条例》、GB/T 39204—2022《信息安全技术 关键信息基础设施安全保护要求》、网络安全等级保护2.0标准体系；此外，根据GB/T43269—2023《信息安全技术 网络安全应急能力评估准则》，针对不同等级的关键信息基础设施，其在人员、制度、技术、措施等方面的应急能力要求显著不同。

攻击面总体相对可控。在传统资源约束下，攻击者难以同时识别、利用并维持大规模攻击路径，网络攻击通常集中于关键信息基础设施、核心业务系统等少数高价值目标，边缘系统、普通终端及关键岗位人员由于攻击成本较高，通常难以成为有效的攻击突破口。

前沿AI风险趋势

当前沿AI被作为网络攻击工具时，可能通过降低高危攻击能力门槛和扩大可利用攻击面，削弱现有网络安全防控措施中“攻击能力相对稀缺、攻击面总体可控”的基本假设。

首先，网络攻击能力规模化扩散。基于跨域知识生成、代码自动化生成、复杂任务推理及流程协同能力等技术特征，前沿AI可以在网络攻击链路中的多个环节提供关键能力支撑[70]，包括零日漏洞挖掘[71]、攻击面情报采集与解析[72]、攻击策略动态自适应与监测规避[73]等核心环节。尽管部分能力仍处于学术研究和受控环境测试阶段，但在漏洞挖掘和武器化等网络杀伤链的前中期阶段，前沿AI已经开始提供实质帮助。例如，Mythos、MDASH[74]等前沿AI模型和系统在批量、自动化挖掘漏洞方面展现出强大能力。在此基础上，部分模型研发机构的安全团队已观察到疑似AI辅助漏洞开发与攻击工具生成、甚至自主代理形成完整攻击链的实际案例[75]，表明前沿AI降低攻击门槛、辅助甚至自动化推进网络攻击的现实可行性正在提升。

与此同时，人员、供应链、非核心组件等维度形成的攻击面持续扩大，各类环节暴露的安全脆弱性显著加剧。前沿AI模型显著增强了恶意行为体发现并利用暴露面和脆弱点的能力，可以针对安全运维与决策人员、非核心资产及供应链等防护薄弱环节实施更加精准和有效的攻击。在针对人的社会工程学攻击中，前沿AI能够自动化分析社交画像、公开数据等开源情报，显著提升攻击的精准性与规模化水平。[76]此类工具已在地下黑市流通[77]，并已经在一起欺诈案

---

[70] Yujin Potter et al., "Frontier AI’s Impact on the Cybersecurity Landscape", last modified November 27, 2025, https://arxiv.org/pdf/2504.05408.

[71] Anthropic, "Assessing Claude Mythos Preview’s cybersecurity capabilities," April 7, 2026, https://red.anthropic.com/2026/mythos-preview/.

[72] Mateusz Kazimierczak et al., "Impact of AI on the Cyber Kill Chain: A Systematic Review", December 30, 2024, https://doi.org/10.1016/j.heliyon.2024.e40699.

[73] 同注70 (Yujin Potter et al., 2025)。

[74] MDASH在CyberGym已经超越Mythos，见 https://www.microsoft.com/en-us/security/blog/2026/05/12/defense-at-ai-speed-microsofts-new-multi-model-agentic-security-system-tops-leading-industry-benchmark/。

[75] Google Threat Intelligence Group, "GTIG AI Threat Tracker: Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access", May 12, 2026, https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access; Anthropic, “Disrupting the first reported AI-orchestrated cyber espionage campaign”, Nov 13, 2025, https://www.anthropic.com/news/disrupting-AI-espionage; Michael Clark, "JADEPUFFER: Agentic ransomware for automated database extortion", Sysdig, July 1, 2026, https://www.sysdig.com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion.

[76] 同注10 (Bengio, Yoshua, et al., 2026)。

[77] 同注70 (Yujin Potter et al., 2025)。

件中造成高达2500万美元的经济损失。[78]在供应链与资产管理领域，组织对第三方服务的深度依赖及智能体向非技术部门的扩散，正导致“影子资产”激增。[79]数据显示，去年涉及第三方的入侵事件占比已达48%，且同比增长60%。[80]大量未被纳管的新增暴露面与复杂的供应链依赖关系，可能稀释传统物理隔离与边界防御的有效性。部分地区政府部门已经针对前沿AI可能对关键信息基础设施带来的挑战发布了警示。[81]

治理挑战

攻击能力扩散导致异常行为监测更加困难。现有源头防控体系以异常行为识别、日志关联审计、漏洞周期管控为核心抓手，其效能依托于攻击特征可沉淀、行为基线可参照、威胁主体可聚焦的传统条件。然而，网络攻击能力的扩散可能大幅抬升监测难度：一方面，APT组织等高能力威胁主体，可以借助前沿AI让其实战手段更加高效、隐蔽和灵活；另一方面，它使得原本缺乏资源、仅具备普通技术的恶意行为体也能制造高破坏性的网络安全事件，导致威胁来源复杂且多点高发，增加研判难度与检测噪声。

攻击面扩大使得层级防护体系资源配置逻辑面临重构。前沿AI正在改变攻击面分布的相对稳定性与可控性，使原本相对清晰的防护边界呈现动态扩展与外溢特征。其一，攻击路径自动化与跨域渗透增强。前沿AI提升攻击路径发现与组合能力，使攻击能够更快跨越非核心资产、影子资产及供应链节点，实现向核心系统的快速渗透，边缘节点原有的风险缓冲作用明显减弱。其二，针对“人”的攻击门槛下降与规模化增强。前沿AI降低社会工程学攻击的定制成本，使针对关键岗位人员的信息操控与欺骗行为更易实施，突破既往人员安全防线的成本下降、确定性提高。其三，攻击面外延扩展与边界模糊化。随着第三方数字服务与智能体应用扩散，未纳入统一管理的数字资产持续增长，系统边界由静态防护边界转向动态网络结构，导致传统基于边界清晰性的防护资源配置模式适配性下降。

（二）演化阶段：人工研判与决策速度承压

现有制度抓手

根据《国家网络安全事件报告管理办法》等规定，网络安全事件应急响应高度依赖人工研判和决策来划分突发事件等级并进行事件报告。从各地各行业主管部门到中央应急主管机构，

---

[78] Barbara Kollmeyer：《AI深伪视频会议诱骗香港公司职员转账2,500万美元》，华尔街日报，2024年2月6日，https://cn.wsj.com/articles/ai深伪视频会议诱骗香港公司职员转账2-500万美元-30cffa25?eafs_enabled=false。

[79] Cloud Security Alliance, "The AI Vulnerability Storm: Building a Mythos-ready Security Program", April 18, 2026, https://labs.cloudsecurityalliance.org/wp-content/uploads/2026/04/mythosreadyv92.pdf.

[80] Verizon, "2026 Data Breach Investigations Report", May 2025, https://www.verizon.com/business/resources/reports/dbir/.

[81] Matthew Mohan, "CSA tasks critical information infrastructure leaders to review cyber risks due to AI-enabled threats", May 5, 2026, https://www.channelnewsasia.com/singapore/csa-cyberattacks-ai-mythos-cii-owners-parliament-6100061.

突发事件的响应和处置均要依托人工审核研判，分级别后，在相应的规定时间内推进 “监测—研判定级—事件上报—预警发布/启动响应” 的决策流程（参见下表）。[82]

报告主体分类	事件分级	时限	首次报告对象	重大 / 特别重大事件后续流转要求
涉及关键信息基础设施的网络运营者	所有事件	1 小时内	保护工作部门、公安机关	属于重大、特别重大事件的，保护工作部门收报后第一时间向国家网信部门、国务院公安部门报告，最迟不得超过半小时
中央和国家机关各部门及其直属单位的网络运营者	一般	2 小时内	本部门网信工作机构	无
	重大、特别重大	2 小时内	本部门网信工作机构	各部门网信工作机构收报后第一时间向国家网信部门报告，最迟不得超过 1 小时；国家网信部门收报后及时向有关部门通报
其他网络运营者	一般	4 小时内	属地省级网信部门	无
	重大、特别重大	4 小时内	属地省级网信部门	省级网信部门收报后第一时间向国家网信部门报告，最迟不得超过1小时，同时向同级有关部门通报

表3.1.1 网络安全事件报告流程

关键假设

风险演化具有渐进性，给应急响应留出了时间窗口。传统网络攻击通常需要经历情报侦察、漏洞挖掘、工具开发、渗透利用和横向移动等多个阶段，攻击准备周期往往以月甚至年计。这一时间跨度使得监测预警、人工研判和层级决策具备时间窗口，制度流程能够在风险大范围扩散前完成预警发布和应急响应。

人工研判的准确性优先于单纯时效性。由于网络安全事件的技术复杂性和影响外溢性，制度设计强调“先研判、后定级、再响应”，通过多层人工审核确保事件性质、危害程度和影响范围的判断准确，避免因过早或过晚预警导致资源错配或社会恐慌。

[82] 《国家网络安全事件应急预案》，3.3“预警研判和发布”。

前沿AI风险趋势

多个机构发布的报告共同指出，以Mythos为代表的前沿AI模型已展现出极强的漏洞挖掘能力，传统的防御窗口期可能从数月被压缩至数小时。人类需要在此时间内完成漏洞补丁开发、发布和部署，[83]从而引发攻防时间在技术层面的结构性失衡。[84]

此外，当前的失衡并非单纯源于技术本身，而同时来源于能力转化速度的不对称。攻击侧可以更快、更少制度约束地将前沿AI的技术加成转化为实战能力；而防御侧则必须履行合规审计、预算采购、安全审批及跨部门协同等多个需要人工研判的内部流程。相关研究指出，技术与制度的多种因素叠加，使得当前阶段前沿AI带来的技术红利正率先向攻击侧倾斜。[85]

治理挑战

前沿AI的快速发展可能对风险演化阶段中应急响应的制度假设造成挑战，攻防两侧的能力演化在结构上出现了不平等的倾向。

前沿AI驱动下，网络攻击的隐匿性、自动化和扩散速度持续提升，而既有应急响应范式仍主要依赖人工研判和层级决策机制，二者之间的时间尺度差异不断扩大，监测预警、事件研判和应急响应等关键环节面临更大的时效性压力。

此外，现行分级响应机制也面临新的适配挑战。前沿AI驱动的网络攻击在事件初期往往仅表现为局部异常或低风险事件，但一旦突破关键节点，可能在短时间内快速演化为跨系统、跨供应链的重大网络安全事件。同时，事件定性与溯源难度进一步增加，风险特征往往滞后于事件实际演化过程，容易导致预警等级调整和应急响应启动滞后，压缩有效处置窗口。

（三）后果阶段：针对关键信息基础设施战略目标的攻击威胁升高

现有制度抓手

我国网络安全治理体系以关键信息基础设施（CII）保护制度为核心，通过等级保护制度、分级防护体系、态势感知与应急响应机制等手段，对高价值系统实施重点防护。同时，依托入侵检测、漏洞管理、威胁情报共享及攻防演练等机制，构建覆盖事前防护、事中监测与事后处置的全周期安全治理框架。

---

[83] Cloud Security Alliance, "The AI Vulnerability Storm: Building a Mythos-ready Security Program", April 18, 2026, https://labs.cloudsecurityalliance.org/wp-content/uploads/2026/04/mythosreadyv92.pdf.

[84] Anthropic, "Project Glasswing: An initial update", May 22, 2026, https://www.anthropic.com/research/glasswing-initial-update.

[85] 同注83 (Cloud Security Alliance, 2026)；注70（Yujin Potter et al.,2025）。

关键假设

上述制度抓手的基本假设包括：高等级网络攻击通常由少数高能力行为体实施，其攻击能力具有稀缺性与高成本特征；二是关键信息基础设施在常态对抗条件下具备较强防护裕度，可造成重大破坏的攻击通常需要较长准备周期与复杂资源投入。在上述假设下，网络安全治理总体建立在持续攻防对抗、分层防御和事件响应基础上，更强调通过纵深防御、持续监测和应急处置控制风险演化，而非应对短时间内急剧发生与扩散的系统性风险。

前沿AI风险趋势

如前所述，前沿AI显著提升了复杂网络攻击的能力上限，高强度攻击的技术门槛显著下降。在此背景下，关键信息基础设施在面对高强度、自动化与快速适应性攻击时，其既有防御能力所预留的缓冲空间可能被压缩，潜在漏洞也可能被迅速挖掘和武器化，部分传统依赖人工响应与时间窗口的防御机制可能难以有效约束突发性高强度攻击扩散。对此，部分政府部门已向各行业关键信息基础设施运营者发出了前沿AI风险警示。[86]金融等具体行业的监管机构也对前沿AI滥用攻击支付系统、窃取金融数据的潜在威胁作出风险提示。[87]

治理挑战

前沿AI推动网络攻击能力上限显著提升，使单次事件可能造成的损害后果显著增加，从而改变既有“可控强度对抗”假设。在极端情境下，关键信息基础设施可能面临更高强度、更快扩散速度与更强跨系统影响的复合冲击，使传统分级防护体系的“梯度缓冲”功能被削弱。

与此同时，攻击能力提升并不必然受制于资源约束，而更多受到行为体策略选择与风险博弈约束，使得原本以“能力稀缺性”为基础的风险控制逻辑面临重估。在这一背景下，当前治理体系不仅需要应对更高强度攻击风险，还需重新评估关键基础设施在极端条件下的系统韧性与失效边界。

（四）防御赋能与治理机遇

尽管在战术层面上，网络攻击往往呈现出一定的“先手优势”与非对称特征，但从国家应急管理与安全治理的战略视角来看，网络安全体系有望逐步走向以动态对抗与持续适应为核心的“新均衡范式”。

---

[86] Matthew Mohan, "CSA tasks critical information infrastructure leaders to review cyber risks due to AI-enabled threats", May 5, 2026, https://www.channelnewsasia.com/singapore/csa-cyberattacks-ai-mythos-cii-owners-parliament-6100061.

[87] Financial Stability Board，"The Financial Stability Implications of Artificial Intelligence", November 14, 2024；中国人民银行：《中国金融稳定报告（2025）》，2025年12月。

技术层面，前沿AI亦可显著增强网络防御能力，推动防御体系由规则驱动向智能驱动升级。具体而言，基于大模型与智能体的安全系统能够在威胁检测、漏洞挖掘、异常行为识别、攻击路径推演与自动化响应等环节实现全流程赋能。一方面，AI可通过对海量日志、流量与系统行为的实时分析，提升对复杂、隐蔽与变种攻击的识别能力；另一方面，具备一定自主规划能力的智能体系统，可在攻击发生早期即完成分级响应、隔离与修复建议生成，从而显著压缩攻击窗口期并降低人为处置延迟。

在此基础上，防御体系具备“累积叠加效应”。与攻击者通常需要针对不同目标与场景重复构建攻击链条不同，防御方在持续对抗过程中能够积累更完整的系统运行上下文与攻击模式知识，并将每一次处置经验转化为模型更新、策略优化与规则增强的输入。这种基于反馈闭环的学习机制，使防御能力随时间演化呈现边际递增特征，从而逐步累计安全优势，而非线性消耗型投入。

与此同时，治理层面也在进行资源再配置。随着前沿AI能力逐步嵌入关键基础设施安全治理框架，全球范围内的高端安全人才、算力资源与技术研发投入有望向防御侧集聚。通过将大规模算力与模型能力注入安全运营与应急响应体系，可以在整体上提升单位防御资源的覆盖密度与响应效率，从而在结构上对冲攻击方的分散式、低成本试探策略。这一趋势也推动安全治理从传统“事后修补”模式，向“预测—拦截—自适应修复”的主动防御模式演进。

从长期演化来看，上述由人工智能驱动的防御能力重构，有望逐步压缩网络攻击的结构性非对称优势。在持续强化关键基础设施安全韧性的同时，也为构建适应人工智能时代复杂威胁环境的系统性韧性底座提供支撑。

本节附表：国家网络安全事件分级分类标准

根据《国家网络安全事件报告管理办法》等制度文件，网络安全事件指由于人为原因、网络遭受攻击、网络存在漏洞隐患、软硬件缺陷或故障、不可抗力等因素，对网络和信息系统或其中的数据和业务应用造成危害，对国家、社会、经济造成负面影响的事件。[88]

网络安全事件按发生后的危害程度、影响范围等因素[89]分为四级；针对最为严重的三个级别，相关规定和标准以非穷尽枚举的方式提供了分级定量指标。[90]其中，“特别重大网络安全事件”（详见下表）作为制度认定的最高红线，其量化阈值代表了当前网络安全体系的风险极值，可以用作基准来研判前沿AI等新技术所带来的风险。

类型[91]	阈值描述
运行中断	关键信息基础设施整体中断运行 6 小时以上或主要功能中断运行 24 小时以上
	省级以上党政机关门户网站、中央重点新闻网站因攻击、故障，导致 24 小时以上不能访问
数据安全	1亿人以上公民个人信息泄漏
	核心数据、重要数据泄露或被窃取、篡改、假冒，对国家安全和社会稳定构成特别严重威胁
有害内容传播	省级以上党政机关门户网站、中央重点新闻网站、超大型网络平台等被攻击篡改，导致违法有害信息特大范围传播
社会影响	影响一个或多个省级行政区 50% 以上人口，或者 1000 万人以上用水、用电、用气、用油、取暖、交通出行、就医、购物等工作、生活
经济损失	1亿元直接经济损失

表3.1.2 特别重大网络安全事件分级标准（示例）

[88] 参见《国家网络安全事件报告管理办法》，2025年11月1日，第十二条；相同定义见《信息安全技术 网络安全事件分类分级指南》（GB/T20986-2023）；更早于2017年11月1日颁布的网信办《国家网络安全事件应急预案》，其中定义略有不同：“本预案所指网络安全事件是指由于人为原因、软硬件缺陷或故障、自然灾害等，对网络和信息系统或者其中的数据造成危害，对社会造成负面影响的事件。”

[89] 《中华人民共和国网络安全法》，第五十五条；另参见《信息安全技术 网络安全事件分类分级指南》（GB/T20986-2023），6.1“分级方法”，该标准后于前述网络安全法制定，对分级的表述有所不同——按照事件影响对象的重要程度、业务损失及社会危害的严重程度这三个要素进行分级。

[90] 《国家网络安全事件报告管理办法》，第十二条。

[91] 为便于分析和理解，本报告将《国家网络安全事件报告管理办法》及相关标准所规定的分级量化阈值归纳为以下四类。

二、公共卫生

近年来的研究表明，前沿AI模型在生物化学知识获取与推理、实验设计和复杂问题求解等方面已达到甚至超过部分专业研究人员水平[92]，显著提升了生物和化学领域技术扩散的效率。在赋能生命科学研究的同时，作为一项典型两用技术，前沿AI在生物和化学领域的能力一旦被滥用，可能导致病毒、细菌、危险化合物等有害生物与化学制剂加速开发与投放。2025年，慕尼黑安全会议与核威胁倡议组织（Nuclear Threat Initiative, NTI）联合进行桌面推演，认为利用AI的生物技术能力发动疫情并非遥不可及，而是一种现实中完全可能发生的、后果严重的威胁；更值得警惕的是，现有的安全机制可能不足以应对AI加持下的生物威胁。[93]

突发公共卫生事件指包括重大传染病疫情、群体性不明原因疾病、群体性中毒等，[94]我国针对特别重大公共卫生事件的现行分类分级标准（见本节附表：国家公共卫生事件分类分级标准）构成了该领域的应急红线，可以作为研判前沿AI风险的参照系。本节以我国公共卫生领域传染病防治的应急体系为例，该领域构建于高危物质与实验室源头管控、传染病监测预警与应急响应、特异性医学干预与救治能力储备等核心治理抓手之上。[95]然而，前沿AI技术的发展，通过降低非自然突变改造技术门槛、压缩爆发周期、削弱既有干预手段有效性等方式，正改变公共卫生领域传染病防治相关的风险演化逻辑，冲击了现有治理抓手。

（一）源头阶段：高危物质和实验室源头管控难度增加

现有制度抓手

现行生化安全前置防控依托高危基因序列筛查、实验室分级准入等核心管控手段，以此约束高危致病因子、有毒化学品的合成活动。通过实验室审批、实验伦理审查、研究过程监管及对核酸合成订单进行严格的生物安全筛查，实现对高危物质、病原体及危险遗传材料的源头封堵。

关键假设

这一治理抓手的有效性主要建立在“资源壁垒”假设之上。过去，高危生物威胁主要来自拥有专业科研团队的国家级行为体：科研研发门槛高、资源密集，只有具备特定资质与资源的主体方能开展危化品物资采购和高危生物研究，且现有核酸筛查协议足以识别并拦截已知的致病性基因序列。

---

92 同注10 (Bengio, Yoshua, et al., 2026)。

93 同注63 (Nuclear Threat Initiative, 2025)。

94 《中华人民共和国突发公共卫生事件应对法》，2025年11月1日，第二条。

95 《中华人民共和国生物安全法》，2021年4月15日；国家疾病预防控制局《传染病疫情信息公布管理办法》，2026年5月9日；《全国疾病预防控制行动方案（2024—2025年）》，2024年5月21日。

前沿AI风险趋势

前沿AI可赋能有害生物制剂的采购、合成、实验故障排查等多个环节，提供技术支持、操作指导。前沿AI大幅压缩了生物知识与实验技能的学习周期，不仅拉低了专业门槛，更将实验能力下沉至低资源场景。前沿AI在众多衡量生物武器研发相关知识的基准测试中，其表现已达到甚至超过专家水平。例如，OpenAI的o3模型在病毒学实验室操作规程故障排除方面的评测，表现优于94%的领域专家，[96]已有前沿AI模型能够提供从商业合成DNA构建体中恢复活性脊髓灰质炎病毒的详细技术指导。[97]

前沿AI在DNA合成筛查规避方面的辅助尤为值得关注。[98]作为生物技术供应链的核心环节，全球多数核酸合成供应商均对订单序列执行生物安全筛查，阻断有害序列合成交付。然而，微软联合多家机构开展的红队演练表明，前沿AI可借助开源蛋白设计工具改写危险蛋白序列、规避筛查[99]；其针对72种高危蛋白生成的7.6万余个合成同源物变体，在4款主流生物安全筛查系统中漏检率达30%到70%。[100]

治理挑战

滥用防控从有组织、可溯源的机构实体，转向离散化、难预判的个体行为主体。既有防控体系的制度设计，内嵌着对科研活动集中化、建制化的基本预设——高风险生物实验需要专业团队、高级别实验室和持续资金投入，这使得监管可以通过许可证审批、机构备案与人员背景审查等机制实现有效覆盖。然而，前沿AI通过将专家级的实验设计、故障排查与方案优化能力封装为易用的工具，大幅压缩了专业知识的获取周期，使原本受限于资源与技能壁垒的实验活动下沉至低资源场景。此时，威胁源不再指向少数大型科研机构，而可能泛化为任何能够调用前沿AI模型并接触基础实验设备的个体或小型团队。使得传统基于“机构准入”和“关键人员资质”的监管模式难以覆盖分散的威胁源。

由基于已知序列对比的核酸筛查防线面临挑战。核酸合成订单的生物安全筛查，其技术有效性建立在威胁序列已知的前提之上——即通过维护已知危险序列的黑名单，对合成订单进行

---

96 同注10 (Bengio, Yoshua, et al., 2026)。

97 Roger Brent and T. Greg McKelvey Jr, "Contemporary AI foundation models increase biological weapons risk," arXiv:2506.13798, June 2025, https://ui.adsabs.harvard.edu/abs/2025arXiv250613798B/abstract.

98 利用AI生成功能高危但序列全新的DNA/RNA蛋白片段，通过序列伪装规避现有核酸筛查系统，MIT实验显示可轻易突破主流供应商检测，见安远AI，天津大学生物安全战略研究中心，《人工智能 x 生命科学的负责任创新》，2025，https://concordia-ai.com/wp-content/uploads/2025/07/Responsible-Innovation-in-AI-x-Life-Sciences-cn.pdf.

99 Antonio Regalado, “Microsoft says AI can create ‘zero day’ threats in biology,” MIT Technology Review, October 2, 2025, https://www.technologyreview.com/2025/10/02/1124767/microsoft-says-ai-can-create-zero-day-threats-in-biology/?trk=article-ssr-frontend-pulse_publishing-image-block.

100 利用开源AI蛋白设计工具对72种已知危险蛋白（如蓖麻毒素、肉毒毒素等）进行“同义改写”，生成超过7.6万个“合成同源物”变体，并投入4种主流生物安全筛查系统中进行测试，结果发现漏检率高达30%—70%，见：Bruce J. Wittmann et al., “Strengthening nucleic acid biosecurity screening against generative protein design tools“. Science, 2025, 390: 82-87. DOI:10.1126/science.adu8578.

比对拦截。但前沿AI可能实现危险基因进行同义重组、片段拆分等，使最终交付的DNA序列在局部匹配度上与已知危险条目脱钩，却保留完整的有害功能。这意味着，筛查所依赖的已知数据库，可能被AI生成的未知但等效变体所绕过，供应链入口处的技术屏障出现缺口。

（二）演化阶段：传染病监测预警与应急响应体系面临效能挑战

现有制度抓手

现行传染病监测预警与应急响应体系依托多层次的制度与技术安排，力求在疫情出现初期实现快速锁定与有效遏制。在监测预警层面，已建立传染病网络直报系统、不明原因肺炎监测网络、口岸检疫与发热筛查机制，辅以PCR核酸检测、宏基因组测序等实验室诊断技术，以实现对新发突发传染病的早期识别与病原体鉴定。在应急响应层面，已构建分级响应的应急预案体系、区域性联防联控机制、疫苗与药物战略储备，以及基于流行病学模型的传播风险评估与隔离封锁措施。上述制度安排构成从发现到阻断的完整响应链条。

关键假设

这一体系的有效运行建立在若干核心前提之上：其一，病原体的基因序列相对稳定或在已知变异范围内演化，检测探针与引物能够有效匹配目标基因特征，使实验室诊断具备足够的敏感性与特异性；其二，传染病的基本传播动力学参数处于过往经验可覆盖的范围之内，使得基于既有模型的传播风险评估与干预策略设计具有可操作性；其三，从病原体出现到形成大规模社区传播之间存在足够的防御窗口期，使监测系统有充裕时间完成发现、鉴定、上报与响应部署。三个前提共同支撑着感知—研判—阻断的闭环有效性。

前沿AI风险趋势

前沿AI赋能生成式设计，加速病原体改造。前沿AI赋能生成式设计具体表现为AI在蛋白质结构预测、功能突变预判及序列从头生成等领域的能力跃升。包括以下两个层面：在在蛋白质结构和功能预测层面，AlphaFold 3等工具能高精度预测蛋白质与配体、核酸及蛋白质的相互作用。[101]此外，前沿AI模型能高效预测单点或多点突变对蛋白质功能的影响，[102]这意味着恶意行为者无需进行大规模实验筛选，可利用AI快速探索突变空间，识别增强毒力、逃避免疫识别或提高传播能力的关键变异位点。在序列生成层面，前沿AI已被证明可设计出与天然序列差异

[101] 《上海交大超算平台用户手册》，AlphaFold 3词条，https://docs.hpc.sjtu.edu.cn/app/bioinformatics/alphafold3.html。

[102] Li, Xinning, et al. "Proteins Need Extra Attention: Improving the Predictive Power of Protein Language Models on Mutational Datasets with Hint Tokens." NAR Genomics and Bioinformatics 7.3 (2025): lqaf128.

显著却保留功能的新型蛋白质。[103] 这意味着恶意行为者无需从现有毒株出发，可直接设计具有预期毒力特征的新序列。

前沿AI赋能多源数据融合与传播策略优化。在生物制剂的投放与传播环节，前沿AI同样展现出显著的能力跃升。兰德公司2025年发布的报告明确指出，前沿基础模型除可帮助恶意行为者生成有害生物制剂，还可为投放提供支持，[104]包括如何在公共场所部署、如何选择最佳天气条件和袭击目标，以及探讨如何逃避检测。[105]已有研究展示了前沿AI可整合多源数据，包括社交媒体人口密度地图、空气质量监测数据、[106]气象条件（风向、温度、湿度对气溶胶扩散的影响）、关键交通枢纽客流及人口流动轨迹等，通过融合多模态异构数据的深度学习方法（如CNN）与传播动力学模型，模拟病原体在不同环境条件下的扩散规律和人群暴露程度。[107]虽然这些技术可用于疫情预警和公共卫生风险评估，但相同的能力若被恶意利用，意味着攻击方可以科学选择高效传播的投放地点、优化投放时机以利用有利的气象条件，并通过模型反演计算所需的最小有效剂量。这将使攻击方能够以更精准的投放策略和更少的制剂来实现更大范围的人群暴露和更高效的扩散。

治理挑战

监测预警方面，早期发现能力从技术底层遭到削弱。前沿AI引入的人为非自然突变可能使基于已知序列设计的PCR检测引物与探针无法有效结合，宏基因组测序依赖的同源性比对难以注释AI从头生成的新型序列。一旦疫情暴发，病原体鉴定与溯源在第一时间即面临“测不出、判不明”的困境，预警系统的响应起点被人为延迟，监测端“早发现”的核心功能面临失效风险。

应急响应方面，时效的压缩与决策信息不确定性的叠加。响应时效层面，AI驱动的多源数据融合与传播策略优化能力，使恶意行为者能够以多点同步投放策略实现跨行政边界的快速扩散，病原体扩散速度超越响应体系的空间调配与时间部署能力，传统的密接追踪、交通管制、区域隔离等阻断手段难以形成有效包围圈，防御方从发现病例到实施有效阻断的可操作窗口被收窄。决策信息层面，AI设计引入的非自然突变使得病原体的毒力、传播力、潜伏期等关键流

[103] Jeffrey A. Ruffolo, et al., "Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences" bioRxiv, April 22, 2024, https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1.

[104] Rand, "The Weapons of Mass Destruction AI Security Gap", Feb 12, 2026, https://www.rand.org/pubs/commentary/2026/02/the-weapons-of-mass-destruction-ai-security-gap.html.

[105] Gabriel J.X. Dance, “A.I. Bots Told Scientists How to Make Biological Weapons,” The New York Times, April 29, 2026, https://www.nytimes.com/2026/04/29/us/ai-chatbots-biological-weapons.html.

[106] Romain Molinas, et al., “A novel approach for predicting epidemiological forecasting parameters based on real-time signals and Data Assimilation,“ arXiv:2307.01157, July 3, 2023, https://ar5iv.labs.arxiv.org/html/2307.01157.

[107] Abhinav Sharma, et al., “Artificial Intelligence in Epidemic Modeling and Pandemic Preparedness: A Comprehensive Review,“ Archives of Computational Methods in Engineering, vol. 33, pp. 7085–7115, February 23, 2026, https://link.springer.com/article/10.1007/s11831-026-10525-7.

行病学参数难以基于历史经验与既有知识库快速研判，应急决策者在资源调配方案、封锁范围划定、预警级别提升等关键节点上面临比传统疫情更大的信息不确定性与判断难度，进一步加剧响应迟滞或干预力度失当的风险。

（三）后果阶段：生物威胁的潜在危害边界扩展

现有制度抓手

现行对生物威胁后果的控制能力，建立在特异性医学干预储备的基础之上。已建立针对特定高威胁病原体的疫苗战略储备、抗血清及抗病毒/抗菌药物储备体系。其核心逻辑是针对已知威胁储备已知手段，以在疫情发生后通过特异性医学干预手段降低重症率与死亡率，将健康损害控制在可接受范围之内。

关键假设

这一制度的有效性建立在病原体抗原表位与药物靶点相对稳定的前提之上。即传统储备的特异性医学干预手段，包括针对特定病原体的疫苗、抗血清及抗病毒药物，能有效覆盖疫情中实际流行的病原体变种，其保护效力与治疗效果不会因病原体的自然变异而出现根本性削弱。

前沿AI风险趋势

前沿AI对生物威胁后果控制环节的冲击，集中体现为AI赋能的多重功能修饰使病原体的抗原表位与药物靶点被人为定向改造，进而使特异性医学干预储备的有效性面临深层次侵蚀。

AI技术赋予恶意行为者设计具有复杂致病机制的新型病原体的能力，使病原体可同时叠加多种危险特性。现有自然病原体虽部分具有高致死率，但往往在传播效率、免疫逃逸或环境稳定性等方面存在自然演化的内在制衡——例如高致病性禽流感病毒对人类致死率极高，但目前在人群中不具备高效传播能力，因而其大流行威胁相对有限。[108]然而，兰德公司2026年发布的风险评估报告系统性地识别出五种可通过AI工具修饰的生物功能：宿主范围或趋向性改变、基因组复制增强、免疫或医疗对策逃避、环境稳定性增加、传播动力学增强。[109]其中，免疫或医疗对策逃避直接指向对疫苗与药物储备有效性的侵蚀。AI可使恶意行为者在单一病原体中有意叠加免疫逃逸与高传播性等多种风险特性，使新型病原体的抗原特征远比自然演化出的病原体更为复杂。

[108] Jane Merrick, “Biowarfare could be tailored to target race and sex – and UK ‘must better prepare‘ “, iNews, January 1, 2026, https://inews.co.uk/news/politics/biowarfare-tailored-target-race-sex-uk-better-prepare-4094036.

[109] RAND, “Developing a Risk-Scoring Tool for Artificial Intelligence–Enabled Biological Design,“ February 11, 2026, https://www.rand.org/pubs/research_reports/RRA4490-1.html.

已有的研究证明，在免疫逃避设计方面，AI模型已被用于预测病毒的免疫逃逸路径，以哈佛大学团队的EVEscape模型为代表，利用AI预测了83种新冠病毒刺突蛋白变体，这些变体足以逃避接种过疫苗或自然感染人群所产生的抗体[110]，展示了AI在预判与设计免疫逃逸路径方面的高效能力。在从头病毒设计层面，斯坦福大学团队已利用AI成功设计了16种此前不存在的功能性噬菌体，验证了AI生成的新型基因组不仅能够模仿自然序列，甚至可在特定功能上超越自然演化的限制——此即全球首个AI设计病毒的研究，为AI设计具有复杂致病特征的新型病原体提供了技术可行性的初步证据。[111]

治理挑战

上述AI技术路径对风险后果控制体系构成了新的挑战，并主要体现在**储备体系的靶标失效。**现行储备体系针对的是已知病原体的特定抗原表位与药物靶点，储备相应的疫苗、抗血清及抗病毒药物。然而，AI对病毒表面抗原蛋白关键表位与药物靶点的定向改造能力，使储备疫苗诱导的中和抗体无法有效识别变异毒株，抗病毒药物的分子结合靶点亦可能被改变而失去抑制作用。更为根本的是，AI的序列从头生成能力可直接设计出与任何已知病原体均无抗原相似性的新型蛋白质序列，使得储备针对特定病原体的干预手段这一前提条件本身即不成立——防御方无法在疫情暴发前预知需要储备何种疫苗或药物。由此，后果控制体系在最为核心的降低重症率与死亡率功能上，面临干预工具箱的空心化风险。同时由于新型疫苗与药物从研发到完成临床安全性与有效性验证，其刚性时间要求仍以年为单位计算，这意味着，面对AI设计的新型或深度改造病原体，后果控制体系在可预见的未来时间内，可能长期处于无针对性手段可用的状态。

（四）防御赋能与治理机遇

需指出的是，当前关于前沿AI生物风险的评估，主要依赖知识问答、故障排查与代码生成等基准测试或红队测评结果。尽管部分模型在相关基准上表现出较高能力，但科学界对由模型在知识型任务或红队场景中的高分，是否可外推至真实世界中的端到端有害能力仍存在显著分歧。关键约束在于，从信息生成到实际危害之间仍存在多重非数字化瓶颈，包括高度依赖经验的隐性知识（tacit knowledge)、湿实验操作能力、材料获取与供应链约束、以及跨环节整合能力缺失等。[112]因此，将基准测试或局部红队结果直接等同于现实世界中的完整威胁能力，可能存在对风险的高估。而在这一情况下，生物安全领域的防御方，仍然具有技术和治理方面的防御突破口。

---

[110] Noor Youssef, et al., “Computationally designed proteins mimic antibody immune evasion in viral evolution“, Immunity, 2025, 58(6): 1411-1421.e6, https://www.sciencedirect.com/science/article/pii/S1074761325001785.

[111] https://www.ithome.com.tw/pr/172602.

[112] Epoch AI, "Do the biorisk evaluations of AI labs actually measure the risk of developing bioweapons?", June 13, 2025.

1、技术机遇：前沿AI对公共卫生安全能力的正向赋能

与风险扩散并行的是，前沿AI技术同样显著增强了公共卫生与生物安全防御能力，并已在多个环节实际应用。一方面，在病原监测与早期预警领域，基于机器学习的流行病信号分析系统已被用于从新闻、旅行数据与临床报告中识别潜在疫情风险。[113]另一方面，在药物与疫苗研发环节，AI驱动的分子设计与蛋白结构预测显著压缩研发周期，例如基于结构预测与生成式模型的靶点筛选、[114]抗原设计与候选分子优化已被多家生物技术公司用于加速研发流程。[115]而在生物安全防护方面，DNA合成筛查系统与序列风险检测算法也逐步用于识别潜在高风险序列，从源头降低滥用风险。[116]这些进展表明，前沿AI在提升风险能力的同时，也在同步强化防御体系的“感知—设计—响应”能力。

2、治理机遇：制度层面的潜在结构性优势

公共卫生与生物安全领域本身已具备较为成熟的多层级治理体系，包括跨国监测网络、国家疾病控制机构、以及与科研机构之间的协同机制，这为前沿AI能力的嵌入提供了制度基础。同时，相较于分散化、低可见度的风险生成路径，防御体系在资源动员、标准制定与信息汇聚方面具有更强的集中性与协调能力，使其更易通过标准化工具（如安全评估框架、红队机制、模型审计与合规要求）被增强。此外，在应急响应链条中，防御方还可以通过“前移能力建设”的方式，将AI能力嵌入早期预警、快速研发与资源调度环节，从而在一定程度上抵消能力扩散带来的不确定性冲击。

[113] BlueDot, "ILI Pulse: What do we know about emerging SARS-CoV-2 variants and global COVID-19 activity?", December 6, 2023, https://bluedot.global/understanding-sars-cov-2-variants-and-global-covid-19-activity/.

[114] Zhang, K., Yang, X., Wang, Y. et al., "Artificial intelligence in drug development", Nat Med 31, 45–59 (2025). https://doi.org/10.1038/s41591-024-03434-4.

[115] Lucy Vost, Yael Ziv, Charlotte M. Deane, "Incorporating targeted protein structure in deep learning methods for molecule generation in computational drug design", Chemical Science, 2025, 16 (44): 20677–20693. https://doi.org/10.1039/d5sc05748e.

[116] Gretton, D., Wang, B., Edison, R., et al., "Random adversarial threshold search enables automated DNA screening", bioRxiv, 2024, doi: https://doi.org/10.1101/2024.03.20.585782.

本节附表：国家公共卫生事件分类分级标准

以下表格基于我国现行《国家突发公共卫生事件应急预案》等文件，提供我国公共卫生领域突发事件的分类与分级体系，以供参考：

类别	阈值示例[117]
已知传染病疫情	• 肺鼠疫、肺炭疽在大、中城市发生并有扩散趋势，或肺鼠疫、肺炭疽疫情波及2个以上的省份，并有进一步扩散趋势 • 发生传染性非典型肺炎、人感染高致病性禽流感病例，并有扩散趋势。 • 发现国内已消灭的传染病重新流行
不明原因传染病疫情和/或疾病扩散	• 发生新传染病或我国尚未发现的传染病发生或传入，并有扩散趋势，或发现我国已消灭的传染病重新流行 • 出现群体性不明原因疾病波及多个行政区，并有扩散趋势
危险致病因子丢失	• 鼠疫菌强毒株丢失、被盗 • 高致病性病原微生物、国内尚未发现或已经宣布消灭的病原微生物、未列入《人间传染的病原微生物名录》的高致病性病原微生物或疑似高致病性病原微生物菌（毒）种或样本丢失、被盗 • （其他）烈性病菌株、毒株、致病因子等丢失事件
群体性中毒	• 在一起突发中毒事件中，中毒人数在100人以上且死亡10人以上，或者死亡30人以上 • 在一个行政区内，24小时内出现2起以上可能存在关联的同类中毒事件时，累计中毒人数在100人以上且死亡10人以上，或者累计死亡30人以上 • 两个以上行政区内发生同类重大突发中毒事件（II级），且事件原因存在明确关联

表3.2 特别重大公共卫生事件分级标准（示例）

[117] 参照《国家突发公共卫生事件应急预案》、《北京市突发公共卫生事件应急预案》（2021年修订）等。

三、金融安全

在我国突发事件分类体系之中，金融类突发事件隶属于社会安全事件，[118]其可能表现为货币危机、金融危机、债务危机、宏观经济周期波动等[119]。作为国家安全的重要组成部分，系统性金融风险的防范得到党中央和多个政府部门的高度重视。[120]习近平总书记多次强调，应有效防范化解重大经济金融风险，牢牢守住不发生系统性金融风险的底线。[121]

前沿AI可能诱发系统性金融风险的可能性正受到越来越多关注。MIT一项覆盖30余个国家、270余名AI领域专家的调研指出，金融业是面对前沿AI风险冲击最脆弱的行业之一。[122]国际货币基金组织原副总裁、中国人民银行原副行长朱民指出，金融智能体等前沿AI的“技术迭代、速度、规模已远超现有监管体系，亟需前瞻性监管与全球合作”。[123]近年来，国际货币基金组织（IMF）、金融稳定理事会（FSB）、国际清算银行（BIS）等国际金融机构以及多个国家的监管部门也已针对金融领域的前沿AI风险发出警示。[124]

以下结合国内外最新研判，分析前沿AI在金融领域的典型风险机制与现实案例，并剖析其对我国金融安全应急体系带来的前沿挑战。本节主要关注前沿AI应用于金融领域的风险，即前沿AI模型或系统运行中因其内生技术特性等引发的风险。前沿AI滥用所导致的对金融信息基础设施的网络攻击、金融虚假信息大规模扩散目前也受到高度关注，鉴于威胁模型不同，不在本节详细介绍。

---

118 同注7（《国家突发事件总体应急预案》，2025）。

119 闪淳昌、薛澜主编：《应急管理概论：理论与实践》（第二版）。

120 国家金融监督管理总局党委理论学习中心组：《坚持把防控风险作为金融工作的永恒主题——学习<习近平关于金融工作论述摘编>》，共产党员网，2024年4月19日，https://www.12371.cn/2024/04/19/ARTI1713482516816275.shtml；习近平：《走好中国特色金融发展之路，建设金融强国》，求是网，2026年1月31日，https://www.qstheory.cn/20260131/487aa5b5e0804f7ea968118e541b4e91/c.html。

121 习近平：《当前经济工作的几个重大问题》，求是网，2023年2月15日，https://www.qstheory.cn/dukan/qs/2023-02/15/c_1129362874.htm。

122 同注22 (MIT FutureTech, 2026)。

123 韩忠楠、余世鹏：《朱民：应重视金融智能体背后的宏观风险 中国经济质效稳步提升》，证券时报网，2026年5月19日，https://www.stcn.com/article/detail/3916340.html。

124 参见中国人民银行：《中国金融稳定报告（2025）》；Federal Reserve, "Artificial Intelligence in the Financial System"，May 1, 2026, https://www.federalreserve.gov/newsevents/speech/bowman20260501a.htm；Canadian Centre for Cyber Security, "Statement from the Canadian Centre for Cyber Security on frontier artificial intelligence models and their impact on cyber security", June 24, 2026, https://www.canada.ca/en/communications-security/news/2026/06/statement-from-the-canadian-centre-for-cyber-security-on-frontier-ai-models-and-their-impact-on-cyber-security.html.

（一）源头阶段：基于历史经验和风险可预测性的风险防范工具效果下降

现有制度抓手

我国金融风险防范体系总体坚持风险预防为主，通过风险监测预警、风险评估、压力测试、风险评级、资本与流动性监管等制度工具[125]，对金融机构风险进行持续识别、评估和防控。《中华人民共和国金融稳定法（草案）》提出建立健全金融风险监测、评估、预警和早期纠正机制；《商业银行资本管理办法》要求商业银行建立覆盖各类风险的资本管理和风险评估机制；中国人民银行、国家金融监督管理总局等监管部门也持续组织压力测试、机构风险评级和现场、非现场监管，提前识别和防范系统性金融风险。

关键假设

上述制度工具的运行建立在一个重要前提之上，即金融风险来源总体可识别、风险形成机制具有一定稳定性，监管机构能够依据历史数据、既有风险模式和合理情景假设，对风险来源、风险暴露及潜在影响进行识别、评估和预演，并据此实施资本约束、风险预警和事前防范。

前沿AI风险趋势

模型黑箱。前沿AI模型存在一定程度黑箱，其内部推理过程和决策依据难以被完整理解、解释和验证，可能降低金融机构和监管部门对模型行为、能力边界及潜在失效模式的认知能力，使风险源头的不确定性明显增强，影响其在金融领域的可靠性[126]。国际清算银行报告指出，黑箱特征模糊了模型输出背后的计算逻辑，导致金融机构的结果无法被理解、解释或复现，从而威胁金融稳定性。[127]英格兰银行[128]、金融稳定理事会[129]也提出了相关顾虑，强调必须将增强模型可解释性和透明度纳入金融机构对前沿AI的全生命周期管理。

模型能力边界和失效模式的不确定性。 前沿AI大模型通常依赖大规模数据驱动学习，其能力边界和失效模式难以事先准确识别，金融机构可能难以判断模型将在何种市场环境、输入条件或业务场景下产生偏离预期的输出。这种不确定性并不意味着模型在日常运行中不可靠，而是在训练数据之外或极端情景下的行为缺乏充分验证。使风险源头更难预测，也削弱了监管

---

125 《商业银行资本管理办法》，国家金融监督管理总局令2023年第4号，2024 年1月1日施行，第一百二十六条。

126 上海财经大学、蚂蚁集团、国家金融科技测评中心：《大模型在金融领域的应用技术与安全白皮书》，2024年3月，；https://sime.sufe.edu.cn/43/76/c10224a213878/page.htm。

127 Bank for International Settlements, "Managing explanations: How regulators can address AI explainability," FSI Occasional Paper No. 24 , September 2025, https://www.bis.org/fsi/fsipapers24.pdf.

128 Bank of England, "Financial Stability in Focus: Artificial intelligence in the financial system", April 9, 2025, https://www.bankofengland.co.uk/financial-stability-in-focus/2025/april-2025.

129 Financial Stability Board, "Sound Practices for Responsible Adoption of Artificial Intelligence (AI)"， June 10, 2026, https://www.fsb.org/uploads/P100626.pdf.

机构和金融机构基于历史经验识别潜在风险的能力。英国政府的报告指出，前沿AI运行于开放环境，其安全性难以穷尽验证，在超出训练数据分布的场景中可能出现非鲁棒性和意外失效。[130]

治理挑战

前沿AI模型的黑箱特征、复杂决策机制以及能力边界的不确定性，使部分风险源头难以及时发现和准确刻画，也增加了识别模型失效、验证模型输出以及构建合理风险情景的难度。监管机构和金融机构既有的压力测试、风险评估和风险评级等制度工具可能无法覆盖关键风险情景，导致风险预警和事前防范能力下降。

（二）演化阶段：基于渐进演化和人工介入能力的响应工具适配度降低

现有制度抓手

在金融风险演化阶段，主要依托风险评估、压力测试等工具进行风险监测预警和宏观审慎监管[131]，对风险态势进行动态监测。而当市场出现剧烈波动、系统性金融风险上升时，监管部门可依托涨跌停制度[132]、临时停牌[133]、中央银行流动性支持等常态化市场稳定机制，及时开展风险研判与跨部门政策协调，为市场功能恢复提供缓冲窗口。

关键假设

上述制度工具隐含的核心假设是：

风险暴露与风险传导具有相对渐进性。从风险积累、风险暴露到市场恐慌形成，通常仍存在一定时间窗口。无论是信用风险、流动性风险还是市场风险，其传导往往需要经历信息披露、市场定价调整、机构行为变化等过程。监管部门能够在风险扩大前完成监测、研判、决策和干预。

市场参与者之间存在认知差异和行为异质性。金融体系中存在银行、证券公司、基金、保险机构、散户等不同类型投资者，不同主体的信息来源、风险偏好和决策逻辑存在差异。因此即使面对同一冲击，市场通常不会完全同步行动，而是形成一定程度的行为分散和相互对冲，从而减缓风险扩散速度。当市场出现剧烈波动时，暂停交易或提供缓冲时间能够促使市场重新评估信息、恢复价格发现功能，并重新形成流动性供给。

---

130 UK Department for Science, Innovation and Technology, "Frontier AI: capabilities and risks – discussion paper", page 16, October 2023, https://assets.publishing.service.gov.uk/media/6a3c02f9d52550a19950f3cf/frontier-ai-capabilities-risks-report_l.pdf.

131 《宏观审慎政策指引（试行）》，中国人民银行公告〔2021〕第25号，2021年12月31日。

132 《中华人民共和国期货和衍生品法》，中华人民共和国主席令第一一一号，2022年8月1日。

133 《中华人民共和国证券法》，中华人民共和国主席令第三十七号，2020年3月1日，第111条、第113条。

前沿AI风险趋势

模型集中与同质化放大羊群效应，引发市场系统性共振。当多个金融机构在市场分析、量化决策中使用相似模型、相同数据源和相同服务提供商时，其输出的高度相关性将强化机构间的风险互联性，进而加剧市场羊群行为与顺周期性。[134]美联储在其研究报告中指出，当前整个行业的资金融通和决策对生成式模型展现出极高的依赖性。[135]这一风险也受到我国金融领域监管机构和专家的关注。[136]例如，中国人民银行金融研究所副所长莫万贵指出，这种依赖可能导致模型趋同，并引发系统性金融风险。[137]国际上，金融稳定理事会[138]、英格兰银行[139]、美国财政部[140]等机构也在其政策报告中均表达了这一顾虑。

前沿AI可能因不同情形的模型趋同风险导致流动性踩踏和市场闪崩：当模型输出错误，该错误认知可能被放大，导致大规模非理性决策。若模型产生幻觉或被干扰，其给出的“错误建议”会被趋同的算法架构快速放大。与传统人工操作的随机失误不同，算法同质化使得错误决策具备了“高并发”特征，将单一模型的错误转变为全市场的瞬时抛售或误判。然而，当模型输出正确时，一致性的集体避险行为也可能引发流动性踩踏。在极端市场突发事件中，AI基于市场规律给出的“避险建议”是正确且理性的。但若全市场模型趋同，所有机构同时执行相同买卖动作，会导致市场流动性瞬间枯竭。

自动化执行导致风险爆发速度压缩。近年来，金融机构开始探索将前沿AI进一步嵌入交易执行、风险管理和资金管理等核心业务流程。例如，相关研究指出，随着金融智能体技术不断成熟，前沿AI正由信息分析和决策支持工具逐步演变为能够自主执行交易、资金划转等关键业务流程的自动化执行主体。[141]

若缺乏有效干预措施，上述前沿AI风险因素的叠加可能在未来诱发严重的宏观金融风险。2010年“闪崩”事件表明，高度自动化交易系统之间的相互作用可能在缺乏充分人工干预的情况下，导致市场流动性在极短时间内快速收缩，并放大资产价格波动。未来，随着前沿AI系

---

134 Liang, Nellie., "Remarks by Under Secretary for Domestic Finance Nellie Liang on Artificial Intelligence in Finance", Speech delivered at the OECD–FSB Roundtable on Artificial Intelligence in Finance, May 22, 2024, https://home.treasury.gov/news/press-releases/jy2383.

135 Financial Stability Implications of Generative AI: Taming the Animal Spirits，https://www.federalreserve.gov/econres/feds/files/2025090pap.pdf

136 《关于银行业保险业人工智能安全开发应用的指导意见》，金发〔2026〕8号，2026年6月18日。

137 莫万贵等：《人工智能在金融领域应用的风险与应对》，载《新金融》，2025年第11期。

138 同注87 (Financial Stability Board, 2024)。

139 Bank of England, "Artificial Intelligence and Machine Learning", October 26, 2023, https://www.bankofengland.co.uk/prudential-regulation/publication/2023/october/artificial-intelligence-and-machine-learning; Bank of England, "Financial Stability in Focus: Artificial intelligence in the financial system", April 9, 2025, https://www.bankofengland.co.uk/financial-stability-in-focus/2025/april-2025.

140 同注134 (Liang, Nellie., 2024)。

141 同注13 (UK AI Safety Institute, 2025)。

统广泛应用于金融决策、风险管理和自动化交易，这种全自主的交易闭环将使得风险在数秒内即可跨机构、跨市场完成扩散，从而彻底打破传统监管机制所依赖的响应时间窗口，使市场风险的应急处置面临前所未有的时效性挑战。

治理挑战

前沿AI风险形成与扩散速度显著压缩，挑战现有基于渐进演化和人工介入的响应流程。当前的治理范式高度依赖“监测—研判—决策—干预”的链式流程，决策周期通常以分钟或小时计；而AI驱动的决策闭环能在瞬时完成资产配置与资金划转，使得风险在监管机构做出反应之前便已完成扩散。这种“行政决策时滞”与“机器执行速度”的结构性错位，导致监管陷入滞后陷阱：即便监管部门具备完善的技术监测能力，也因缺乏与机器同频的自动化干预抓手，导致风险化解策略在执行端陷入实质性真空，既有工具难以匹配市场瞬息万变的风险节奏。

模型集中与同质化强化市场共振，基于行为异质性和市场自我修复能力的响应工具被削弱。当金融机构在模型架构与决策逻辑上高度一致时，市场在遭遇极端冲击时不再触发多元博弈下的自发平衡，而是演变为算法共振诱发的集体性同步调整。对于监管而言，这种异质性的丧失构成了深层的治理难题：监管工具的效能基础在于利用不同主体间的偏差进行边际调节，但当所有智能体均在同一时间做出相同响应（如同步撤资或避险）时，市场就失去了有效的对冲力量，流动性恢复的内生修复能力被彻底阻断。这意味着，传统依赖“监测市场边际变化—通过政策引导博弈”的调控手段可能失去效用，监管部门试图通过市场化手段维持稳定的难度显著增大，迫使政策干预被迫从“常规边际调节”转向“非常态全系统阻断”，既有政策工具箱的调控精度与效力面临严峻挑战。

（三）后果阶段：风险影响更易突破危机边界，挑战国家金融安全韧性

前述源头和演化阶段的分析，主要聚焦于前沿AI如何改变金融风险的形成机制和传播路径，即为何部分金融风险更容易发生、更快扩散。而在后果阶段，本报告进一步关注：当前沿AI驱动的金融风险突破传统金融体系的承载边界后，是否可能演化为影响国家金融稳定和经济安全的宏观风险，其治理逻辑也可能相应超出传统金融风险管理和应急管理范畴，进入国家安全层面。

随着前沿AI风险趋势的进一步发展，其可能并非取代传统金融风险，而是在既有风险基础上引入新的风险生成、放大与跨域传导机制，使部分金融风险突破传统经济周期和市场波动的影响范围，与数字基础设施、数据资源、关键技术和国际竞争格局等因素深度耦合，从而产生过去较少出现的新型国家金融安全风险。

现有制度抓手

宏观金融稳定治理层面，我国已初步构建以中央金融委员会统筹，[142]金融监管部门和中国人民银行等机构协同履职的金融稳定治理体系，通过宏观审慎管理、风险监测研判和金融风险处置机制，对重大金融风险进行统筹协调，并依托存款保险制度、金融稳定保障基金、中央银行最后贷款人职能等工具，为极端情形下的风险化解和金融体系稳定提供支撑。

金融基础设施安全保障层面，我国依托《中华人民共和国网络安全法》《中华人民共和国数据安全法》以及关键信息基础设施保护制度，强化金融领域网络安全、数据安全和技术风险治理，由网信、金融监管等多个部门协同开展关键系统保护和风险防范，保障支付清算、金融信息系统等关键基础设施安全稳定运行。

在跨境金融风险防范与国际协调层面，我国通过跨境资本流动管理、国际金融合作机制以及与国际金融组织和其他经济体的政策协调，防范外部金融风险冲击和跨境风险传导，维护金融体系稳定。

关键假设

上述制度工具隐含的核心假设包括：

风险影响范围和传导路径具有可隔离性。现有金融稳定治理体系建立在风险传播路径总体可识别、关键风险节点可定位的基础之上。即使发生重大金融冲击，监管部门仍可通过识别重要机构、关键市场和基础设施节点，采取流动性支持、风险隔离和政策协调等措施阻断风险扩散。

风险来源与处置资源具有可控性。现有金融安全体系建立在国家能够通过监管能力、技术保障和政策工具，对金融运行关键环节实施有效管控与治理的基础之上。即使面临跨境风险输入，也能够依托国内治理体系和国际协调机制维护金融稳定。

前沿AI风险趋势

金融前沿AI应用成为新型战略博弈工具。前沿AI在跨境金融交易与国际资产定价中的非对称应用，正成为地缘政治博弈的新型工具。国际货币基金组织（IMF）工作人员讨论纪要指出，高度集中的AI模型可能被武器化，用于对特定脆弱经济体实施定向的金融市场施压，从而对国家主权构成新型威胁。[143]金融稳定理事会在关于金融科技跨境风险的报告中明确指出，前沿AI的跨国界应用可能加剧地缘政治紧张局势在金融体系中的传导，并放大跨境资本流动的波动

[142] 2023年中共中央、国务院印发《党和国家机构改革方案》，2023年3月。

[143] Frank Adelmann et al., "Cyber Risk and Financial Stability: It’s a Small World After All", IMF Staff Discussion Note, December 2020, https://www.imf.org/-/media/files/publications/sdn/2020/english/sdnea2020007.pdf.

性。[144]此外，国际清算银行关于大型科技公司在金融领域渗透的报告同样强调，如果核心大模型技术被少数地缘政治利益体垄断，将削弱新兴市场国家的宏观经济调控能力与货币主权。这一风险趋势可能导致未来非国家行为体可训练具备强对抗策略的金融智能体，捕捉特定国家跨境支付清算基础设施（如CIPS）或离岸货币市场的流动性漏洞，发起毫秒级的智能化算法闪击战。这种攻击隐蔽性较强，常伪装于正常合规的资本流动中，旨在瞬间剥夺核心资产的定价权并侵蚀国家金融主权。

前沿AI强化金融系统关联性，推动风险向跨市场、跨领域级联演化。前沿AI驱动的金融风险在后果阶段具有极强的非线性跨域传播性，可能迅速从虚拟数字空间传导至物理空间与经济社会全局。当自主智能体因底层大模型趋同或极端市场信号发生非理性共振时，其带来的损害不仅局限于传统证券市场的市值暴跌，更会引发系统重要性金融机构的非典型性瘫痪与信用链条断裂，导致金融风险向公共财政、社会稳定等非金融领域无序蔓延。美国金融稳定监督委员会（FSOC）在其年度报告中已将前沿AI列为系统性金融安全的一大关键脆弱性，强调其可能引发跨市场的连锁反应和不可预测的级联失效。[145]世界经济论坛（WEF）《全球风险报告》亦指出，AI系统与关键金融基础设施的深度耦合，极易在极端黑天鹅事件中催生出超越传统金融危机边界的跨域复合型灾难。[146]（注：关于前沿AI与深度伪造融合引致的金融领域“认知—干预型”突发事件，将在本部分第四节：跨领域复合型突发事件进行介绍。）

治理挑战

风险影响跨市场、跨领域扩散，挑战传统风险隔离机制。前沿AI的风险发展趋势可能导致金融危机跨领域、跨境传导风险强化。经济全球化深入发展的今天，金融危机外溢性突显，国际金融风险点对国内各产业也可能造成冲击。随着前沿AI更深度地嵌入关键金融基础设施，不同机构和不同市场之间的数据共享、模型调用和系统联动程度持续提高。在此背景下，局部模型故障、错误风险信号或异常决策结果可能沿着金融市场复杂的依赖关系快速扩散。原本局限于单一机构或单一市场的风险，可能进一步向跨市场、跨机构乃至跨行业方向传导，更易诱发金融危机，增加系统性金融风险形成和扩散的可能性。

开放对抗格局下，主权应急“防火墙”构建与国际竞争力保持的两难。从完全抵御外部金融风险冲击的视角来看，最安全的做法是建立完全隔离的金融科技与数据防火墙。然而，过度严苛的隔离和应急关断机制，会使得我国金融体系失去利用全球前沿AI技术提升效率、扩大国际竞争力的机会，导致在国际金融科技竞争中落后；但若为了扩大开放与国际竞争而盲目接入

[144] 同注87 (Financial Stability Board, 2024).

[145] Financial Stability Oversight Council, "2025 Annual Report", U.S. Department of the Treasury, December 11, 2025, https://home.treasury.gov/system/files/261/FSOC2025AnnualReport.pdf.

[146] World Economic Forum, "The Global Risks Report 2026", January 14, 2026, https://reports.weforum.org/docs/WEF_Global_Risks_Report_2026.pdf.

全球算法网络，又等于将国家的宏观金融稳定暴露在外来前沿AI的跨国输入性闪击风险之中。如何在防住极端突发打击与构建领先的国际竞争能力之间找到应急治理的最佳战略平衡点，是践行总体国家安全观指引在人工智能与金融深入融合领域必须破解的时代命题。

（四）防御赋能与治理机遇

1、 推动金融风险感知升维与主动防御体系构建

前沿AI技术通过扩展非结构化数据的处理维度与多模态信息的实时分析能力，可以帮助推动金融风险管理从“事后被动响应”向“事前高维预测”跨越。利用大模型的涌现能力与深度强化学习，金融机构能够构建涵盖宏观经济指标、地缘政治舆情、跨境资本流动等多维度的全景风险看板。这有助于金融体系在微观层面精准捕捉异常交易行为，在宏观层面提早研判系统性风险的传播路径，及时阻断风险形成或传导，显著提升金融体系的韧性。

针对前沿AI可能引发的“毫秒级算法闪击战”，传统的“人工介入+熔断”机制已无法有效应对，倒逼金融领域重构基于AI对抗的主动防御体系。通过“以AI制AI”的思路，防御方可训练具备强鲁棒性的内生防御前沿AI应用，开展常态化的模拟算法红蓝对抗。 这种智能免疫系统能够动态识别恶意算法对跨境支付结算设施或离岸市场的流动性试探，实施毫秒级的自适应反制与流动性对冲，从而筑牢抵御外部恶意智能化攻击的“数字防火墙”。

2、 推进主权人工智能基础设施建设，赋能监管科技的范式转型

应对地缘政治背景下的技术垄断与主权侵蚀，其核心战略机遇在于加速构建自主可控的“主权金融大模型”及其中央算力基础设施，从根本上摆脱对外部垄断技术的依赖。以此为契机，国家金融监管部门可推动监管科技的范式转型。利用主权AI工具，监管者能够实施对全网自主智能体网络的穿透式监管，对算法共振、合谋操纵等内生结构性脆弱进行实时仿真与压力测试，确保技术红利牢牢掌握在维护国家金融主权的手中。

3、促进跨域复合型风险的全球多边治理与标准共建

前沿AI风险的非线性跨域传播特征，客观上催生了全球金融治理机制重塑的窗口期。各国在防范AI引发的“级联失效”和“黑天鹅事件”上存在共同利益。这为推动构建前沿AI金融应用的安全全球标准与合规框架提供了契机。通过探索建立跨境跨域的AI威胁情报共享机制、联合沙盒测试机制以及危机跨境处置协议，从而逐步在全球金融层面率先达成前沿AI的治理共识，化技术反噬风险为全球协同治理的制度机遇。

四、跨领域复合型突发事件

跨领域复合型突发事件是指风险在不同领域之间相互传导、叠加演化，并形成跨行业、跨区域、跨系统连锁冲击。《国家突发事件总体应急预案》指出，我国规定的四类突发事件在现实中“往往交叉关联、可能同时发生，或者引发次生、衍生事件”。[147]由于跨领域复合型突发事件的演化机理复杂多样且具有高度不确定性，[148]对跨部门协同要求很高，已成为应急工作中的重点和难点。[149]这一挑战也与国际多灾种防灾减灾的最新趋势高度一致。《2015—2030年仙台减少灾害风险框架》中明确强调，现代社会的风险具有高度的系统性与网络化特征，应采用多灾种和跨部门的方法开展灾害风险治理，加强对不同风险因素及其相互作用的识别、评估和管理，以应对自然、技术、生物等多种风险交织可能带来的复合性影响。[150]

需要说明的是，以下分析主要基于前沿AI能力与风险发展趋势开展未来情景推演。当前，部分关键技术能力及风险要素已在网络攻击、蛋白质预测等具体方面得到验证，但不同风险能力在现实中的实际伤害、组合方式、影响范围及最终后果仍具有较大不确定性。因此，本节旨在结合前文提及的已有技术证据和风险演化规律，分析前沿AI在未来可能引发的典型复合型突发事件，为应急准备和能力建设提供参考。下文选取三类前沿AI时代的复合型突发事件进行简要分析。

（一）数字—物理空间的复合型突发事件

前沿AI不仅可以影响数字空间，也可能通过网络攻击等方式影响物理空间。这类复合型突发事件的**核心传导机制在于**：恶意主体滥用前沿AI强大的自动化代码生成、零日漏洞挖掘及自动渗透能力，发动高强度的网络攻击，导致关键基础设施、工业控制系统或高风险实验室瘫痪，最终在现实世界酿成重大人员伤亡与财产损失。潜在突发事件示例包括：

既有高风险设施的恶意利用。攻击现有基础设施以造成生化危险品泄漏，其领域知识门槛远低于“从零合成”新毒剂。通过前沿AI对既有生化设施与高危物质存量进行恶意利用，能够直接跳过生物化学危险品极其繁琐的“前体获取、配方研发、安全生产、物流运输”等高门槛环节。攻击者可通过AI优化的网络工具，入侵制药企业、危险化学品仓库或高级别生物实验室的控制系统，篡改温控数据、加压阀门参数或通风系统指令，即可远程制造灾难性的危险品泄漏事件。

---

147 《国家突发事件总体应急预案》，1.3 突发事件分类分级。

148 薛澜：《学习四中全会〈决定〉精神，推进国家应急管理体系和能力现代化》，载《公共管理评论》2020年第1期。

149 樊博：《面向多灾种联动的跨部门应急网络研究》，2021年8月6日，https://ccmr.sppm.tsinghua.edu.cn/8thflt/1150.jhtml。

150 联合国大会：《2015—2030年仙台减少灾害风险框架》（Sendai Framework for Disaster Risk Reduction 2015–2030），A/RES/69/283，2015年6月3日，https://www.undrr.org/media/16179/download?startDownload=20260713。

复杂物理系统的链式瘫痪。除了生化设施，数字—物理空间的跨域破坏还可能发生在高度智能化的城市交通与物流网络中。例如，当包含成千上万辆智能汽车的“自动驾驶车队”遭遇前沿AI主导的集群式网络勒索或系统性劫持时，攻击者可瞬间制造大规模城市交通瘫痪、关键运输通道受阻，甚至引发群体性物理碰撞事故。这种数字空间的恶意指令向物理空间机械运动的瞬间传导，构成了智能时代跨域复合突发事件的典型图景。

（二）认知—行动干预的复合型突发事件

前沿AI不仅可能重塑物理世界的安全格局，也可能深度介入人类的认知、决策与行为过程。这类复合型突发事件的**核心传导机制在于**：利用前沿AI生成具有深度误导性、高度拟真性的大规模虚假信息，在认知层面进行高精准的有害操纵，进而诱发群体性事件风险[151]；在宏观层面，甚至可能干扰国家战略决策。其可能的风险事件示例包括：

社会微观层面的舆情事件。依托生成式AI等工具，恶意主体能够在极短时间内以极低成本定制出海量的“深度伪造”音视频与叙事文本。当这些虚假信息流通过社交媒体算法定向推送给特定人群时，可能触发社会层面的集体非理性行为，例如TC260《生成式人工智能服务安全应急响应指南》明确提出了前沿AI可能被用于辅助生成虚假新闻信息、虚假财经信息和虚假医疗健康信息，[152]从而引发侵权欺诈、公共卫生恐慌和金融市场挤兑等事件。[153]

国家宏观层面的决策干扰。相比于社会层面的群体恐慌，认知干预在国家战略决策层面的复合危害性可能更加致命。例如，在生化放核威胁领域，以往的风险模型常聚焦于“人工智能系统直接剥夺人类权限、自动化发射核武器”等假设场景，但兰德研究院组织多位国家安全专家进行的研讨认为，在现实政治与地缘对抗中，前沿AI可通过干扰决策支持系统进而引发高层误判，是更具现实紧迫性的威胁。[154]当这种认知操纵沿决策链条逐级传导至最高决策层时，其引发的战略误判和冲突升级风险可能显著上升，并成为前沿AI时代的重要战略安全挑战。

（三）基础模型同质化引发系统性故障的复合型突发事件

如果说前两类事件侧重于外部恶意主体的定向攻击，那么第三类复合型突发事件则根植于现代智能社会技术底座的结构性脆弱。这类复合型突发事件的**核心传导机制在于：**当全球AI生

[151] 刘清民、王芳：《框架理论视角下人工智能风险的社会建构：基于事件新闻文本挖掘的分析》，载《图书与情报》，2025年第3期。

[152] 同注47 (TC260, 2025)。

[153] 高福：《加强科普，抵御“信息流行病”》，载 《光明日报》，2026年6月6日，https://news.gmw.cn/2026-06/06/content_38814409.htm；Deloitte Center for Financial Services, "Generative AI is expected to magnify the risk of deepfakes and other fraud in banking", May 29, 2024, https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html.

[154] Rand, "AI Implications for Chemical, Biological, Radiological, and Nuclear Defense Policy and Programs", Apr 16, 2026, https://www.rand.org/pubs/perspectives/PEA4611-1.html.

态高度依赖极少数头部企业提供的基础模型或云服务，基础设施的高度同质化可能引发系统性脆弱与漏洞。一旦底层模型发生未知的技术故障、遭遇恶意诱导或陷入失控状态，其风险将瞬间向下游使用相同基础设施的广大行业传导，引发全社会多领域的系统性共振瘫痪。其可能的风险事件示例包括：

风险的“级联放大”。前沿AI产业已逐步呈现出较为明显的分层生态结构：底层由少数通用基础模型提供核心能力，上层则发展出大量基于基础模型开发、微调的行业应用。这一模式显著提升了技术创新和应用效率，但也意味着一旦基础模型出现系统性缺陷或固有偏差，其影响可能沿产业链向下游应用扩散，并在广泛部署过程中被放大。例如，在金融领域，中国人民银行金融研究所副所长莫万贵在其文章中指出，如果多数金融机构在AI底层技术上过度依赖少数技术服务商，则后者实质上就成为具有系统重要性的金融基础设施，形成“大而不能倒”的风险隐患。一旦这类技术服务商发生技术故障或服务中断，可能引发连锁反应，危及整个金融体系的稳定。[155]

非对称攻防格局的改变。在传统的网络安全语境下，恐怖组织等非国家行为体若想同时瘫痪金融、交通和医疗三大行业，必须分别针对这三个行业的专属网络发动三次复杂的渗透攻击，这是一项极度困难的任务。然而，在基础模型同质化的时代，非国家行为体不再需要逐一攻破下游的垂直系统，而只需集中全部技术力量，针对一个共享的底层开源模型、某个通用的AI供应链组件，亦或是核心的AI算力云服务实施单点突破。[156]这种针对“单一源头”的成功攻击或恶意诱导，能够通过供应链的网络共振，瞬间转化为跨行业的系统性级联故障，给社会运转带来严峻挑战。

[155] 同注137 (莫万贵等, 2025)。

[156] 国家网络安全通报中心：《大模型工具Ollama存在安全风险》，澎湃新闻，2025年3月3日，https://www.thepaper.cn/newsDetail_forward_30289619。

第四部分 对策建议

如前文分析，前沿AI风险具备跨领域、跨部门、跨系统的传导特征，传统的单一主体治理难以有效应对。当前核心挑战在于，如何将前沿AI风险纳入国家总体安全观与应急管理体系，并明晰其制度定位。

为此，本部分立足“大安全大应急”视角，从宏观层面的顶层战略设计，到微观层面的应急预案、能力建设、运行机制及协同治理，提出十条对策建议。需说明的是，构建适配前沿AI风险的应急响应体系，相关研究仍处于起步阶段，本报告旨在为相关领域研讨提供补充视角与有益思路。后续对策建议的落地与优化，需与各方主体深入交流，协同打磨。

一、完善顶层战略设计

建议1：强化国家战略统筹，构建适应前沿AI风险的应急响应体系

明确前沿AI风险治理的制度定位，研究如何将其纳入现有安全与应急治理体系。《“十五五”规划纲要》已提出健全人工智能风险防控应急响应机制的部署，[157]但如何将相关风险纳入当前安全与应急的制度框架与治理范畴，仍需进一步探索。建议研究确立前沿AI驱动的突发事件在当前突发事件分类体系中的制度定位，探索增设“人工智能安全事件”独立类型，或将其有机整合至现有突发事件分类体系，完善相关事件分类分级标准与响应要求，[158]提升各领域应急体系应对新型技术风险的承载能力。

探索构建有法可依、权责明晰的前沿AI风险应急指挥体系与工作体制。当前应急体系以分类管理、分级负责、属地管理为主，但前沿AI风险具有传导路径复杂、跨灾种耦合性强等特征，可能导致管辖主体模糊、责任边界不清等问题。建议开展治理试点，明确前沿AI驱动突发事件的类别归属、响应层级与处置机制，逐步消除监管盲区，推动治理体系向法治化、规范化有序过渡，确保未来应对此类突发事件时具备明确的制度依据和处置能力。

[157] 《中华人民共和国国民经济和社会发展第十五个五年规划纲要》，2026年3月。

[158] 宋劲松，《构建适应“人工智能+”的应急响应体系》，载《学习时报》，2025年8月13日，https://cbgc.scol.com.cn/news/6625492。

二、健全应急预案体系

建议2：建立前沿AI风险分析框架，推动应急预案与演练体系动态优化

构建以前沿AI风险识别为基础的关键领域应急预案体系。在公共通信和信息服务、金融、公共卫生、交通等关键领域，现有应急预案制定过程中可能尚未充分考虑前沿AI技术发展带来的增量风险。建议研究建立面向前沿AI驱动突发事件的风险识别与分析框架，推动开展常态化风险评估，将前沿AI技术应用带来的风险变化纳入相关行业的应急体系。重点研判前沿AI模型或系统部署后可能导致的攻击面扩展、关键业务依赖关系变化、人机协同失效以及跨系统风险级联等复杂风险，并据此动态调整和完善相关应急预案，[159]确保前沿AI驱动的突发事件风险在重点领域和行业能够被及时识别、有效介入。

加强前沿AI风险应急演练，牵引相关预案动态迭代。前沿AI可能引发的突发事件，可能具有风险因素耦合复杂、快速演化等特征。建议定期组织重点风险场景下的压力测试、模拟推演和跨部门协同演练，检验现有预案的适用性与响应能力。通过对演练过程中的响应、协同和能力短板进行复盘分析，反向推动预案修订和能力建设，避免预案停留于文本层面的制度安排，推动应急体系能够随着前沿AI技术发展和风险形态变化持续优化，实现“风险识别—预案调整—演练验证—能力提升”的闭环迭代。

强化极端事件情景构建与应急预案准备，提高巨灾应对的制度刚性与底线责任。前沿AI引发的部分风险情景，虽发生概率较低，但可能造成跨区域、跨领域的系统性影响，具有类似巨灾事件的风险特征。针对这类极端情景，建议落实我国应急管理及重点行业相关制度要求，[160]强化风险情景构建和预案准备，明确相关突发事件提级响应的触发条件、处置流程和指挥权限，确保在常规治理机制失效或风险快速扩散时，国家应急体系能够及时启动响应，维持关键功能运行。

三、强化应急能力建设

建议3：健全以应急预案与演练为驱动的前沿AI风险应急能力建设与评估机制

研究构建前沿AI风险应急能力体系，逐步明确能力建设标准。面向前沿AI风险快速演化和高度不确定等特点，建议研究制定关键领域应急保障的能力指标，构建适应前沿AI风险的应急

[159] 相似建议见习近平经济思想研究中心 赵一帆，《人工智能安全治理的国际经验和启示》，《科技中国》杂志2025年第7期，2025年12 月3日，https://www.ndrc.gov.cn/wsdwhfz/202512/t20251203_1402185.html.

[160] 参见《现代化应急体系建设“十五五”规划》：“强化重点行业领域和超大特大城市巨灾情景构建研究，因地制宜完善巨灾应急措施。完善极端天气避险转移预案。综合运用应急演练、突发事件复盘分析等方式检验预案效果，及时修订完善预案”；另见《突发事件应急预案管理办法》第十四条：“国家有关部门和超大特大城市人民政府可以结合行业（地区）风险评估实际，制定巨灾应急预案，统筹本部门（行业、领域）、本地区巨灾应对工作。”

能力框架，探索明确应急能力的核心要素、建设重点和评价维度，推动相关标准规范在关键信息基础设施领域的落地。

建立技术工具储备与能力动态评估机制，提升能力体系适应性。围绕应对前沿AI复杂故障与极端风险的特殊需求，建议统筹推进应急技术工具库的储备与更新，包括自动化威胁分析平台、紧急阻断控制工具等。同时，建立专业技术能力评估机制，定期开展技术工具效能测评与资源保障能力检查，及时识别应急保障体系的短板，确保在突发场景下技术支撑手段的可用性与先进性。

夯实关键系统韧性，筑牢极端冲击下的底线保障。针对可能造成系统性影响的极端风险，建议推动关键领域由依赖单一防护措施向多层次保障体系转变。在技术层面，应加强前沿AI系统及关键信息基础设施的冗余设计，完善离线运行、物理隔离、人工接管和关键功能备份等机制；在基础支撑层面，应统筹能源储备、关键物资调度、应急物流和公共服务保障能力建设，提升社会系统在重大冲击下的承载能力。通过构建多层次、可替代的保障体系，确保即使发生前沿AI系统重大故障或关键技术失效，重要基础设施和公共服务仍能够维持基本运行，并支持后续恢复重建。

建议4：探索发展面向前沿AI风险的多层次保险与风险分担机制

明晰前沿AI风险可保边界，拓展保险协同处置机制。首先，建议厘清不同风险类型的可保性边界，对于具有较明确损失范围和风险规律的，可探索开发相关保险产品，将前沿AI风险纳入保险保障体系。在此基础上，明确事故报告、证据保全、损失评估、理赔衔接和责任追偿等程序，使保险机制服务于损失补偿和风险减量。

针对极端风险，搭建多方分担的综合保障框架。对于高度不确定、损失规模可能突破单一保险产品承保能力的极端风险，探索财政支持、商业保险、再保险、行业共保体、巨灾基金以及巨灾风险证券化等相结合的多层次风险分担机制。同时，应推动建立重大事故损失数据库，完善风险评估模型和赔付触发机制，明确保险赔付、政府救助、责任主体赔偿与恢复重建资金之间的衔接关系，形成覆盖事前风险减量、事中快速响应与事后恢复补偿的综合保障体系。

四、优化应急运行机制

建议5：事前阶段，提升前沿AI相关突发事件的早期预警与及时响应能力

探索建立适配前沿AI技术特征的早期风险指标体系。鉴于前沿AI驱动的突发事件具有演化速度快、传播路径复杂和早期信号隐蔽等特点，建议探索建立适配其技术特征的早期风险指标体系，着力识别可能预示事件升级的异常信号，实现对风险的早期研判与预警。

研究识别风险演化过程中的“关键应急状态切换节点”，并明确相应处置措施。在风险触及关键阈值时，应及时“切换”至人工接管、限制高风险功能或实施系统降级运行等针对性措施，有效延缓风险扩散和事态升级，为后续响应处置和恢复重建争取宝贵时间。

建议6：事中阶段，针对已知和未知风险完善差异化的响应机制

针对相对已知的风险情景，完善标准化响应流程。对于已纳入风险情景库、具有较明确演化路径的突发事件，建议细化处置流程、岗位职责和协同机制，推动相关单位依照预案快速响应，减少临时决策成本和处置偏差。针对前沿AI风险演化迅速等特征，在重点领域探索建立适度授权机制，根据风险等级和事件紧迫程度，将部分处置权限前置至一线运营和管理单位，并配套完善的过程记录方式、事后复盘要求和责任认定机制。同时，优化一线安全运维人员的职责配置，在特定紧急场景下赋予其必要的快速处置权限，提升基层快速响应和先期控制能力。

针对超出既有预案的未知风险，探索科学有效的决策程序。对于无法依据既有经验和预案快速判断、可能具有高度复杂性和不确定性的突发事件，应避免基层单位基于有限信息进行临场决策和处置，防止局部判断失误导致风险进一步扩散。基层单位应重点承担现场控制、信息采集、日志留存和快速报告等职责，及时将事件态势上报至应急指挥机构。针对事件性质、影响范围和发展趋势，由应急指挥机构组织相关领域专家开展联合研判，统筹制定处置方案并协调资源调度，确保在未知风险情景下加强应对韧性、减小影响损失。

建议7：事后阶段，推动防御能力持续演进

把握防御方累积优势，以持续复盘推动防御体系的进化迭代。事件处置结束后，不应仅停留在问题修复，而应将事后复盘作为提升防御体系能力的重要环节。建议通过对事件处置过程、风险演化路径和系统失效表现进行复盘，将暴露出的薄弱环节、跨系统联动风险和应对经验沉淀为应急能力建设的重要积累。不同于攻击方需要针对具体目标持续调整攻击方式，防御方能够通过每次事件处置积累系统运行经验和风险认知，并将其转化为预案优化、技术改进和响应机制完善的基础，在持续迭代中提升防御效能。

五、构建协同治理体系

建议8：强化专业技术协同，提升前沿AI风险应急支撑能力

建立跨领域风险术语与分类评估框架，提升跨领域协同应急的基础支撑能力。针对前沿AI风险典型的跨领域特征，风险术语和分类框架的不统一或成为实现联合研判前置阻碍，[161]建议

[161] 同注154 (Rand, 2026)。

开展跨部门风险研判，结合不同领域风险特点，推动建立适用于前沿AI风险识别、分类和评估的共通框架，并明确相关风险分类分级的基本判断标准。通过统一风险术语和分析框架，减少部门间认知差异，为跨领域联合应对机制提供基础支撑。

组建针对前沿AI风险防控与应急处置的专业技术队伍，构建制度化的专业技术协调体系。我国网络安全领域已形成以国家计算机网络应急技术处理协调中心（CNCERT）等专业机构为技术支撑核心的应急协同机制。前沿AI风险具有跨领域传导、技术机理复杂、演化速度快等新特征，其影响覆盖网络安全、公共卫生、金融稳定等多个领域，已超出传统网络安全事件的处置范畴。因此，建议各领域借鉴网络安全领域CNCERT的经验，逐步构建面向前沿AI的跨领域专业技术协调体系，培育专业技术支撑队伍。在制度定位上，建议明确其专业技术协调属性，不替代行业主管部门的法定监管职责，也不与现有网络安全应急机构形成职能重叠，重点承担前沿AI风险识别、技术研判、事件评估、监测态势汇聚、应急资源对接等，推动职能规范化。

健全专业技术支撑能力建设与评估机制。围绕核心能力建设搭建机制和评估体系，可逐步健全与专业技术协调体系相配套的支撑机制。一是建立标准化能力评估体系，制定前沿AI应急技术机构的能力标准与准入规则，统筹培育各级专业技术支撑力量；二是组建复合型专业人才队伍，覆盖大模型技术机理、安全对齐、风险评估及行业应用等专业领域，夯实跨领域技术协同的能力基础。

建议9：深化跨机构协同，实现治理优势互补

健全政府部门间的风险信息共享，探索联合监测与协同响应机制。针对前沿AI风险传播速度快、影响链条长的特点，建议推动重点行业主管部门、应急管理部门和技术支撑部门建立常态化监测信息共享机制，加强风险态势、事件线索和处置经验互通。在重大风险情景下，可探索建立跨部门联合研判和协同指挥机制，明确信息报送、资源调度和职责协同流程，提升复合型突发事件整体应对能力。

促进政府和企业间专业人员的双向交流，推动技术研发与应急实践深度融合。前沿AI企业技术人员和开发者掌握模型能力边界、系统运行机制等专业知识，有关政府部门则积累了突发事件组织协调和风险处置的信息与经验。[162]可探索建立专业人员交流、联合培训和能力互认机制，促进技术研发与应急实践的深度结合。一方面，推动企业在模型开发、部署和安全评估过程中更充分考虑实际风险场景；另一方面，在突发事件发生后，借助技术力量快速分析风险来源、识别干预路径，提升事件处置效率。

[162] 同注154 (Rand, 2026)。

建议10：加强跨区域与国际协同，建立防范前沿AI风险的底线性共识

推动形成面向前沿AI风险的基础性治理共识。前沿AI风险具有跨区域、跨境传播的特点，短期内难以形成全球统一的治理标准。应优先围绕防止重大风险扩散这一共同目标，推动建立基础性的风险分类分级标准和应对原则，重点加强对大规模网络攻击、金融市场异常波动等跨境风险场景的协调，减少风险跨区域传播和叠加放大。

完善危机沟通与跨境协同机制，提升极端风险应对能力。针对可能影响多个国家和地区的重大AI风险事件，应推动建立关键领域和重点事件下的信息沟通渠道，确保相关方能够及时同步风险态势和处置进展。积极参与国际多边交流与政策讨论，将降低AI驱动极端风险扩散作为国际安全合作的重要议题，推动形成更加有效的全球风险应对网络。