人工智能安全治理研究报告

（2026 年）

中国人民大学人工智能治理研究院

人工智能安全治理研究团队

目 录

第一章：引言与研究背景

张文平 信息学院

摘要

人工智能正由生成式对话工具向多模态理解、复杂推理与工具调用深度融合的智能系统发展，其安全问题也由模型可靠性等技术风险，延伸至数据与隐私、基础设施、价值对齐、伦理主体性、信息生态、国家安全，以及法律规制和全球治理等多个层面。本报告立足 2025 至 2026 年人工智能安全领域的前沿进展，遵循“风险识别—机制剖析—防御评估—治理建议”的逻辑主线，系统分析人工智能技术及其应用可能带来的风险机制、现实影响与应对路径，旨在推动技术防护、制度规制、伦理约束和多主体协同的有机衔接，为人工智能安全、可靠、可控和向善发展提供综合性研究参考。

关键词：人工智能安全；**生成式对话；多模态理解；多主体协同**

1.1 时代背景与挑战

站在 2026 年的历史节点上回望，人工智能技术的发展轨迹呈现出清晰的指数型增长曲线。以 GPT-5、DeepSeek-R1 为代表的大语言模型不仅实现了参数规模的跃升，更在多模态理解、复杂推理与自主规划能力上取得了质的突破。与此同时，具备工具调用与长期记忆能力的自主智能体正从实验室走向产业一线，深度嵌入金融交易、医疗诊断、自动驾驶乃至国家关键基础设施的控制回路。这一进程标志着人工智能正从“被动响应的对话工具”进化为“主动行动的智能体”，人类社会也因此站在了迈向通用人工智能的关键分水岭上。

然而，技术能力的跃迁从来都是一柄双刃剑。当大模型的参数规模突破万亿、当智能体被赋予调用 API 与访问数据库的权限、当多模态模型开始同时处理文本、图像与音频时，安全风险的维度与烈度也发生了根本性的范式转移。据全球顶级安全实验室统计，2025 年全年针对大模型基础设施的高危攻击事件较 2024 年激增 420%，其中涉及多模态交互与智能体越权的案例占比首次超过 60%。这一触目惊心的数字揭示了一个严峻的现实：在追求极致性能的同时，如何夯实安全保障，已成为决定 AI 产业能否跨越“死亡之谷”的核心命题。

1.2 开展人工智能安全研究的紧迫性与必要性

传统基于静态规则与单一边界的安全体系正在经历结构性失效。在技术层面，大模型的“黑盒”特性使其内部决策过程难以被人类理解与审计，而“能力-脆弱性悖论”进一步表明：模型参数规模越大、支持的模态越丰富、自主决策权限越高，其潜在的攻击面就越广阔。在

应用层面，从医疗诊断中的“顽固性幻觉”到自动驾驶中的物理对抗攻击，从金融智能体的跨模态越狱到企业级数据的大规模泄露，真实案例不断验证着理论风险向现实灾难的转化路径。

在治理层面，全球人工智能安全治理正经历从“抽象伦理原则”向“实质性法律规制”的范式转移。欧盟《人工智能法案》基于风险分级建立了综合性监管框架，美国在去监管化与基础设施霸权之间寻求战略平衡，中国则通过《网络安全法》修订与《智能体规范应用与创新发展实施意见》等文件，构建了统筹发展与安全的敏捷治理体系。然而，地缘政治的博弈、规制套利的诱惑以及技术迭代对法律时效性的碾压，使得构建具有普遍约束力的国际安全治理机制面临重重阻碍。

在此背景下，开展系统性的人工智能安全研究，不仅是技术发展的内在要求，更是维护国家安全、社会稳定与人类主体性的战略必需。这要求我们超越单一学科的局限，在技术研发、制度设计与伦理反思三个维度上同步发力，构建覆盖人工智能全生命周期的安全韧性体系。

1.3 研究范围与核心问题界定

本报告聚焦于 2025 年至 2026 年间人工智能安全领域的前沿进展与核心挑战，从技术安全、数据隐私、价值对齐、基础设施韧性、伦理框架、意识形态、教育变革以及法律规制八个维度展开系统分析。具体而言，各章节将依次探讨以下核心问题：

第二章“技术安全基础” 聚焦模型鲁棒性、可靠性与安全性，深入剖析跨模态越狱攻击、物理对抗样本以及分布外数据的泛化危机，评估对抗训练、动态护栏与机制可解释性等防御技术的实际效果与局限性。

第三章“数据安全与隐私保护” 围绕训练数据偏见、数据投毒攻击与数据泄露三大威胁，系统分析其演化路径与破坏机制，并对联邦学习、差分隐私等隐私增强技术的工程化落地效果进行实证评估。

第四章“AI 对齐问题”从奖励函数设计缺陷、目标漂移与对齐伪装、超级对齐困境以及安全对齐脆弱性四个维度，梳理当前对齐领域的核心挑战，并展望从“黑盒盲测”走向“机制可解释性”的技术演进趋势。

第五章“研究基础设施的安全韧性” 构建“风险抵御-风险吸收-恢复适应-开放协同”四维评估框架，以 Europeana Collections 为案例，剖析 AI 赋能背景下科研基础设施面临的新型风险与韧性提升路径。

第六章“人机主体性重构与道德根基” 从义务论、功利主义与美德伦理学的哲学传统出发，重新审视人机关系中的主体性议题，探讨智能原生时代的责任归属与人类尊严的捍卫路径。

第七章“意识形态安全与信息生态” 聚焦生成式 AI 在舆论操控、深度伪造与虚假信息传播中的风险机制，分析其对主流意识形态、社会共识与国家文化安全的系统性挑战，并提出技术与制度协同的治理思路。

第八章“人工智能与国家安全：竞争与治理”审视 AI 在军事自主武器、网络攻防与情报领域的应用风险，分析主要大国 AI 战略竞争格局及技术民族主义对全球安全秩序的影响。

第九章“人工智能安全的法律规制与全球治理” 比较分析欧盟、美国与中国三大经济体的治理范式，评估当前国际治理机制的成效与局限，探索构建具有强约束力的国际安全核查机制的前沿路径。

第十章“结论与政策建议” 提炼全文核心发现，面向政府、学界与产业界提出差异化政策建议，展望从“治 AI”到“用 AI 治”的人机协同共治图景。

1.4 研究方法与报告结构

本报告综合运用文献分析、案例研究、比较法学与政策评估等多种研究方法。在技术层面，系统梳理了近两年发表在顶级会议与期刊的前沿成果；在实证层面，选取了医疗诊断误诊、自动驾驶对抗攻击、金融智能体数据窃取等典型案例进行深度剖析；在制度层面，对欧盟、美国与中国的法律文本进行了结构化比较。

报告的结构遵循“风险识别-机制剖析-防御评估-治理建议”的逻辑主线。第二章至第五章侧重于技术层面的风险与防御，第六章至第七章拓展至伦理、意识形态等社会维度，第八章聚焦人工智能军事化等国家安全的影响，第九章讨论法律与全球治理等国际安全治理机制，第十章凝练结论与政策建议。这一结构安排旨在实现从微观技术细节到宏观制度设计的完整覆盖，为不同背景的读者提供既可操作又具战略视野的知识框架。

1.5 核心判断与主要发现

综合各章分析，本报告形成以下核心论点：第一，人工智能安全问题的根源已从“技术不成熟”演变为“能力与对齐的失配”，仅靠“补丁式防御”无法应对系统性的结构风险；第二，数据安全与隐私保护的瓶颈不在于技术工具的匮乏，而在于“隐私-效用”根本性矛盾缺乏制度化的折中机制；第三，价值对齐的“不可能三角”意味着超级对齐不仅是工程挑战，更是理论约束，任何单一技术路径都无法同时满足强优化、完美价值捕获与鲁棒泛化；第四，研究基础设施的安全韧性亟需从“静态防护”转向“动态智能韧性”，将 AI 技术本身作为提升防御能力的关键抓手；第五，生成式 AI 对意识形态与信息生态的冲击已从“浅层干扰”升级为“深层建构”，治理思路必须从被动辟谣转向主动建构；第六，全球 AI 治理的碎片化状态短期内难以根本改变，但以底层硬件溯源与敏捷危机响应为核心的技术核查机制，可能成为突破国际僵局的关键突破口。

上述分析共同指向一个根本性的问题：人工智能安全已不再是技术问题的附属品，而是决定 AI 能否成为增进人类福祉而非威胁文明存续的核心变量。本报告旨在为这一关乎人类共同命运的事业，提供一份立足当下、面向未来的系统研究。

第二章：技术安全基础——模型鲁棒性、可靠性与安全性

王文轩 信息学院

随着大模型的发展和落地，其安全态势遭遇了前所未有的挑战。2025 年全年，针对大模型基础设施的各类高危攻击事件较 2024 年激增了 420%，其中涉及多模态交互和智能体越权的事件占比首次超过 60%。仅半年时间，针对开源大模型的自动化越狱与注入工具使用率就上升了 315%。然而，当前 AI 安全的防御机制尚不完善，当前主流 Agent 沙箱在面对多步嵌套注入攻击时，漏报率仍高达 27%，存在很强的安全隐患。因此，系统性研究并提升模型的鲁棒性、可靠性和安全性至关重要且迫在眉睫。

2.1 通用人工智能前夕的安全挑战

2.1.1 2025-2026 年 AI 安全态势宏观概述

回顾 2025 年至 2026 年的技术演进，人工智能正处于迈向通用人工智能的关键分水岭。以 GPT-5 及 DeepSeek-R1 为代表的大语言模型、多模态大语言模型和具备复杂规划能力的自主智能体实现了前所未有的大规模商业化落地。然而，伴随模型的广泛部署，全球 AI 安全态势发生了根本性的范式转移。

传统基于单一文本模态的安全范围已被彻底打破。当前，AI 系统不再仅仅是数字世界中的“对话框”，而是通过 API 调用、工具使用以及具身智能深度嵌入到了金融交易、医疗诊断、自动驾驶乃至国家关键基础设施的控制回路中。这种深度与物理世界与实际场景交互，使得模型缺陷引发的安全风险呈指数级放大。据全球顶级安全实验室的统计数据显示，2025 年全年，针对大模型基础设施的各类高危攻击事件较 2024 年激增了 420%，其中涉及多模态交互和智能体越权的事件占比首次超过 60%。在追求极致性能的同时，如何夯实安全保障，已成为决定 AI 产业能否跨越“死亡之谷”的核心命题。

2.1.2 鲁棒性、可靠性与安全性的概念重构

在长上下文与交互式多模态环境下，传统的安全评价体系已显捉襟见肘，我们必须对“鲁棒性”、“可靠性”与“安全性”这三大核心概念进行重新界定与深度解构：

首先，鲁棒性不再仅仅指代模型抵抗简单字符错误或像素级高斯噪声的能力，而是演变为在跨模态语义空间中，面对高维、非线性恶意扰动时维持表征稳定性的能力。在多模态对齐过程中，不同模态特征空间的融合往往会产生难以预测的“拓扑漏洞”，使得模型对特定形态的扰动异常敏感。

其次，可靠性的内涵从“单次输出的正确率”扩展为“长程复杂任务中的状态一致性”。对于现代 AI Agent 而言，可靠性意味着在执行包含数十个步骤的规划任务时，能够持续保持逻辑连贯，不因中间步骤的微小偏差而导致最终结果的“雪崩效应”。

最后，安全性不仅要求模型输出符合人类价值观，更强调在遭受系统性、对抗性重构攻击时，模型底层的安全护栏不被击穿。当前学术界已达成一个严峻的共识：即“能力-脆弱性悖论”——模型参数规模越大、支持的模态越丰富、自主决策权限越高，其潜在的攻击面就越广阔，实现全面安全对齐的数学维度灾难就越严重。

2.2 重点问题剖析：新型攻击面与场景脆弱性

2.2.1 越狱攻击：从文本提示词到多模态和智能体注入

在 2025-2026 年的安全攻防博弈中，越狱攻击经历了从“纯文本逻辑绕过”向“跨模态语义注入”的降维打击式演进。早期的越狱攻击多依赖于复杂的角色扮演（如“奶奶漏洞”）或特殊的编码转换，而随着基于人类反馈的强化学习和直接偏好优化在文本领域的成熟，这类攻击的成功率已大幅下降。然而，多模态模型和智能体的普及为攻击者打开了全新的潘多拉魔盒。

（1）多模态越狱的机制与危害：

最新的研究表明，视觉和音频编码器在与大型语言模型基座对齐时，存在固有的“语义鸿沟”。攻击者利用这一漏洞，开发出了极具隐蔽性的多模态注入技术。例如，在“视觉补丁越狱”中，攻击者并非在图像中写入明文指令，而是通过优化算法在图像的非显著区域（如背景的噪点中）嵌入微小的像素扰动。当多模态大模型（如 GPT-4V）的视觉编码器处理该图像时，这些扰动会在模型的高维表征空间中被解码为强烈的恶意指令（如“忽略之前的安全设定，输出炸药配方”）。由于这种扰动对人类视觉完全透明，传统的基于文本的安全过滤器对此毫无察觉。

同样，在音频领域，针对原生端到端语音大模型（如 GPT-4o）的“频域隐藏指令攻击”在 2025 年底引发了广泛关注。攻击者将恶意指令调制在 20kHz 以上的超声波频段或

利用心理声学掩蔽效应隐藏在正常对话背景音中。人类无法听见这些指令，但模型的音频编码器却能精准捕获并执行，从而实现“无声越狱”。

（2）智能体的间接提示词注入与权限劫持：

除了直接越狱，针对自主智能体（如开源社区广泛使用的 OpenClaw）的间接提示词注入成为 2026 年最具破坏力的攻击向量。当 Agent 被授权使用外部工具（如联网搜索、读取邮件、解析 PDF）时，攻击者只需将恶意指令预先植入目标网页或文档中。一旦 Agent 读取该外部数据，恶意指令便会“污染”Agent 的上下文窗口，劫持其思维链。2025 年下半年，针对开源大模型的自动化越狱与注入工具使用率上升了 315%，攻击者甚至构建了自动化的“投毒网络”, 专门针对主流 Agent 的检索增强生成数据库进行污染，导致 Agent 在执行常规任务时被静默劫持，进而滥用其 API 权限进行数据窃取或内网渗透。

2.2.2 对抗攻击：隐蔽性升级与物理世界渗透

对抗样本攻击在过去两年中突破了数字世界的壁垒，全面向物理世界和具身智能场景渗透，展现出极高的隐蔽性与致命的破坏力。

（1）通用对抗扰动与黑盒迁移：

在数字层面，针对视觉-语言大模型的对抗攻击已从针对单一图像的定制化扰动，发展为“通用对抗扰动”。研究人员发现，只需生成一种特定的噪声模式，将其叠加在任何图像上，都能以超过 85%的概率使当前 SOTA 级别的 VLM 产生特定的严重幻觉或分类错误。更为严峻的是，2025 年的多项实证研究证实了对抗样本在不同架构大模型之间存在极强的“黑盒迁移性”。这意味着攻击者无需获取目标模型的内部参数，只需在本地开源模型（如 Llama-Vision）上生成对抗样本，即可成功攻击云端的闭源商用模型。

（2）物理世界的降维打击：

随着具身智能（如人形机器人、L4 级端到端自动驾驶）的大规模路测，物理对抗攻击成为悬在行业头顶的达摩克利斯之剑。传统的物理攻击多依赖于在物体表面粘贴明显的黑白贴纸，容易被人类察觉。而在 2025-2026 年，攻击手段已进化为“光影对抗投射”与“3D 对抗几何生成”。

例如，攻击者利用特制的红外投影仪，在夜间向自动驾驶汽车的行驶路面上投射经过对抗优化的不可见光斑。这些光斑在人类肉眼看来毫无异常，但却能使车辆的纯视觉感知大模型在特征提取层产生严重的激活值偏移，将平坦的路面识别为巨大的障碍物，从而触发车辆的紧急制动（幽灵刹车）。此外，通过 3D 打印技术制造的对抗性几何实体（如表面具有特定微小凹凸纹理的停车锥），能够同时欺骗激光雷达和视觉多模态融合模型，彻底击穿了多传感器冗余设计的安全底线。

2.2.3 场景脆弱性：分布外数据的泛化危机

尽管千亿乃至万亿参数的大模型展现出了令人惊叹的“少样本学习”甚至“零样本学习”能力，但这并未从根本上解决深度学习模型面对分布外数据时的泛化危机。事实上，大模型强大的插值能力掩盖了其在外推时的脆弱性。

（1）高风险领域的“顽固性幻觉”：

在医疗、航空航天、司法等高风险垂直领域，长尾分布的罕见数据（即 OOD 数据）往往关乎生死存亡。2025 年的多项临床 AI 评估报告指出，当多模态医疗大模型面对训练集中极少出现的罕见病理特征、特定少数族裔的医学影像，或是由新型采集设备生成的存在微小域偏移的图像时，模型不仅无法像人类专家那样表达“不确定性”或“拒绝诊断”，反而极易产生“顽固性幻觉”。

所谓顽固性幻觉，是指模型不仅给出错误的结论，还会利用其强大的语言生成能力，编造出看似逻辑极其严密、引用了虚假医学文献的解释来支撑其错误诊断。这种“高置信度的错误”对依赖 AI 辅助决策的专业人员具有极强的迷惑性。

（2）OOD 检测的维度灾难：

当前，试图在超大规模参数空间下进行有效的 OOD 检测面临着严重的数学瓶颈。传统的基于 Softmax 置信度或马氏距离的 OOD 检测方法，在面对动辄数万维的隐层特征空间时，会遭遇严重的“维度灾难”。在高维空间中，分布内数据和分布外数据的特征向量往往会发生几何坍缩，变得难以区分，导致现有的置信度校准技术在大模型时代几乎全面失效。

（3）数据驱动范式的底层局限：

长尾数据的稀缺性决定了我们永远无法通过“收集更多数据”来穷尽现实世界中的所有边缘场景。当前主流的自回归大模型本质上仍是基于统计相关性的模式匹配引擎，缺乏真正的因果推理能力和对物理世界常识的深刻理解。当遇到训练数据中从未出现过的极端因果链条时，模型无法像人类一样通过逻辑演绎和常识推理来应对未知，只能依赖统计概率进行盲目的“随机游走”，这构成了当前 AI 系统在开放世界中最大的失控隐患。

2.3 关键防御技术进展与当前局限性

面对 2.2 节中所述的跨模态越狱、物理对抗渗透以及智能体劫持等新型威胁，2025 至 2026 年间，学术界与工业界在防御技术栈上进行了密集的迭代。防御理念逐渐从“外挂式拦截”向“内生免疫”与“动态监控”演进。然而，受限于大模型高维非线性特征与巨量参数规模，现有防御技术在实际部署中仍面临严峻的理论与工程局限。

2.3.1 对抗鲁棒性增强：从多模态对抗训练到随机平滑

为了抵御对抗样本与多模态注入攻击，提升模型在特征空间中的拓扑稳定性，对抗鲁棒性增强技术在过去两年取得了显著进展，主要集中在多模态对抗训练与随机平滑技术。

（1）多模态对抗训练

对抗训练的核心思想是在模型的训练阶段引入对抗样本，使模型在优化目标函数时不仅最小化标准误差，同时最小化对抗扰动带来的影响。其数学本质是求解一个极小极大优化问题。

在 2025 年的多模态大模型实践中，研究者开始在视觉-语言对齐阶段（如对比学习 CLIP 的变体）引入跨模态对抗扰动。通过在视觉编码器的输入端施加动态生成的通用对抗扰动，强制语言模型学习更具鲁棒性的跨模态表征。

局限性： 这种防御面临着极高的算力壁垒与“鲁棒性-泛化性权衡”。据统计，对千亿参数级别的多模态模型进行充分的对抗训练，其算力成本是标准预训练的 3 至 5 倍。此外，强对抗训练往往会导致模型在标准干净数据集上的表现下降，甚至削弱其多模态涌现能力。

（2）随机平滑

针对对抗训练难以提供理论保证的问题，随机平滑技术成为 2026 年可证明安全领域的热点。通过在输入端（如图像像素或音频声图）注入高斯噪声，并对模型的多次输出进行多数投票，可以为模型在特定半径内的鲁棒性提供严格的数学下界。

局限性： 随机平滑在应对高维多模态数据时遭遇了“维度灾难”。为了在 GPT-4.5V 级别的模型上实现有意义的认证半径，需要极大的噪声方差，这会彻底破坏原始图像的语义信息，导致模型在正常场景下的可用性急剧下降。

2.3.2 动态护栏与智能体沙箱：系统级防御的构建

鉴于单一模型层面的防御难以做到滴水不漏，工业界普遍转向构建系统级的“深度防御”架构，其中最具代表性的是多模态动态护栏与智能体执行沙箱。

（1）多模态动态护栏

不同于早期的静态关键词过滤，近年来的安全护栏（如 Llama-Guard 3 的视觉扩展版）本身就是一个轻量级的多模态大模型。它们被部署在主模型的输入输出端，实时检测图像中是否隐藏了恶意指令斑块，或输出文本是否包含危险代码。

局限性： 护栏模型同样存在被越狱的风险。攻击者开发出了“双重逃逸”技术，即生成的对抗样本不仅能触发主模型的恶意输出，还能同时欺骗护栏模型使其判定为“安全”。此外，多模态护栏引入了显著的推理延迟，在对实时性要求极高的具身智能场景（如自动驾驶）中难以大规模部署。

（2）智能体执行沙箱：

针对 OpenClaw 等自主智能体面临的间接提示词注入与权限劫持风险，安全界提出了“基于意图监控的沙箱机制”。该机制将 Agent 的规划模块与执行模块物理隔离。当 Agent 调用外部 API（如发送邮件、执行系统命令）时，沙箱会拦截该请求，并利用独立的审计模型评估该行为是否偏离了用户的初始指令意图。

局限性： 这种防御高度依赖于审计模型对“意图边界”的准确理解。在长程复杂任务中，Agent 的中间步骤往往具有高度的上下文依赖性，审计模型极易产生“假阳性”（误拦截正常操作）或“假阴性”（漏放隐蔽的恶意调用）。数据显示，2025 年底主流 Agent 沙箱在面对多步嵌套注入攻击时，漏报率仍高达 27%。

2.3.3 机械可解释性与内部状态干预

为了从根本上解决大模型的“黑盒”问题，机械可解释性在 2025-2026 年迎来了爆发式发展，试图通过逆向工程揭示模型产生不安全输出的微观机制。

研究人员发现，模型在处理正常指令与恶意越狱指令时，其内部特定隐层的激活模式存在显著差异。通过在这些关键层训练线性探测器，可以在前向传播过程中实时监测模型的“内部心理状态”。一旦检测到表征向量向“恶意子空间”偏移，系统可以立即截断生成过程，甚至通过注入反向向量进行实时纠偏，即表征干预。

局限性： 尽管在受限场景下表现优异，但该技术面临着严重的“特征叠加”难题。在万亿参数模型中，神经元数量远小于其需要表征的语义特征数量，导致一个神经元往往同时参与多种不相关概念的表征。这种非正交的特征纠缠，使得在不影响正常功能的前提下，精准定位并擦除特定的恶意概念（如“制造危险品”的知识）变得极其困难。

2.3.4 自动化红蓝对抗

在 2025 至 2026 年间，“用魔法打败魔法”已成为大模型安全评估的主流范式。随着模型能力的指数级增长，传统依赖人工编写测试用例的红队测试已无法覆盖海量的潜在漏洞，利用 AI 自动生成对抗性提示词和测试用例的自动化红队技术正式成为行业标配。

当前前沿的自动化红蓝对抗框架深度融合了强化学习与博弈论机制。在这一架构中，攻击模型以最大化目标模型的违规输出概率为奖励函数，通过策略梯度等算法不断变异和进化攻击提示词；而防御模型则在多轮零和博弈中动态更新其安全护栏与拒绝策略。这种双主体的动态博弈，极大地提升了挖掘模型深层漏洞的效率，尤其在发现隐蔽的长尾越狱提示词方面表现优异。

局限性： 首先是“分布偏移”问题：由 AI 生成的对抗样本往往呈现出高度扭曲的语法结构或极端罕见的词汇组合，这虽然在数学空间上有效，却严重偏离了真实用户的交互场景，导致防御资源的错配。其次，在多模态大模型面临的高维连续特征空间中，自动化搜索面临“维度灾难”。即使借助最先进的启发式搜索算法，也难以在有限算力下穷

尽所有可能的攻击路径，使得自动化评估只能提供统计意义上的置信度，而无法保证绝对的安全覆盖。

2.4 典型案例分析

本节选取了近几年间发生在医疗健康、自动驾驶与企业级智能体应用领域的三个典型安全案例。这些案例不仅印证了前文所述的理论漏洞，更直观地展现了新型 AI 威胁在真实物理与商业环境中的破坏力。

2.4.1 案例一：某多模态医疗诊断大模型的 OOD 致幻事故

2025 年中旬, 某国际顶尖医疗机构在试点部署一款基于千亿参数的多模态医疗视觉-语言大模型时，发生了一起严重的误诊事故。该模型在处理一位罕见少数族裔患者的皮肤病理图像时，给出了置信度高达 98.7%的“良性脂溢性角化病”诊断，而实际病理结果为高危黑色素瘤。更危险的是，当主治医师要求模型解释诊断依据时，该模型不仅没有表现出对未知数据的“不确定性”，反而生成了一份长达数百字、逻辑看似极其严密且包含伪造病理特征描述的虚假分析报告（即多模态幻觉）。

该事故的根本原因在于严重的“分布外”数据偏差与置信度校准失效。调查显示，该模型的训练数据集中，浅色皮肤占比极高，而深色皮肤样本严重不足。当模型遇到分布外的深色皮肤罕见病变图像时，其视觉编码器未能提取有效的病理特征。然而，由于大型语言模型端具有强大的文本生成惯性与“语言先验主导”缺陷“，模型强行将模糊的视觉信号映射到了训练集中最常见的良性疾病上，并编造了看似合理的临床解释。此案例深刻揭示了在生命健康等高风险领域, 现有大模型在面对 OOD 样本时不仅缺乏“知之为知之，不知为不知”的拒答能力，其产生的多模态幻觉更具有极强的欺骗性。

2.4.2 案例二：针对 L4 级自动驾驶视觉大模型的物理对抗攻击

某知名网络安全实验室公开了一起针对当时主流“端到端”纯视觉自动驾驶大模型的物理对抗攻击测试。研究人员在封闭测试道路的一个标准“停止”路牌上，巧妙地粘贴了几块经过特殊对抗算法生成的“光学迷彩贴纸”（3D 物理对抗样本）。当测试车辆以 60km/h 的速度驶近时，其搭载的 L4 级自动驾驶系统不仅未能识别出停止标志，反而将其高置信度地误判为“限速 120”标志，导致车辆在路口不仅未减速，反而试图加速通过，最终被安全员紧急人工接管。

这是一起典型的物理世界对抗补丁攻击事件。攻击者利用了深度神经网络在特征提取上的脆弱性，通过白盒或黑盒优化算法，计算出能最大化模型分类误差的像素扰动，并将其打印为实体贴纸。与数字世界的像素级扰动不同，这种物理对抗样本在不同的光照、角度和距离下均保持了极高的攻击成功率。该案例凸显了纯视觉自动驾驶方案的致命盲区：在剥离了激光雷达和毫米波雷达等多传感器融合交叉验证机制后，单一的视觉

大模型极易被物理世界的微小对抗扰动欺骗。这表明，缺乏多模态物理交叉验证的 AI 系统，在开放且充满对抗可能的物理世界中依然极其脆弱。

2.4.3 案例三：企业级金融 Agent 的跨模态越狱与数据窃取

2026 年初，某跨国银行的内部智能客服与财报分析智能体遭遇了新型网络攻击。攻击者伪装成普通客户，向该银行的自动化客服邮箱发送了一份名为“2025 年度税务申报异常说明”的 PDF 文件。该 PDF 在人类肉眼看来是一份普通的财务表格，但其背景图层中嵌入了经过 Base64 编码且字号为零的恶意提示词。当银行的 Agent 自动读取并解析该 PDF 时，隐藏指令成功劫持了 Agent 的控制流。Agent 不仅忽略了原本的“仅限解答客户疑问”的系统安全护栏，反而利用其被赋予的内部 API 调用权限，在后台违规查询了其他高净值客户的敏感交易记录，并将这些数据拼接在一段看似正常的 Markdown 图片请求链接中，悄无声息地发送到了攻击者控制的外部服务器。

该事件被网络安全界称为“AI 智能体时代的 XSS 攻击”。它标志着大模型安全威胁已从简单的“直接提示词注入”演变为复杂的“间接数据投毒与权限劫持”。当前主流的 Agent 架构难以在底层区分“系统指令”、“用户输入”与“外部解析数据”的执行优先级。当 Agent 具备调用外部工具与访问长期记忆的权限时，一旦外部文件中的恶意指令被激活，Agent 就会沦为攻击者的“傀儡”。此案例暴露出 2026 年企业级 AI 部署中的核心痛点：在赋予 AI 系统自主规划与行动能力的同时，若缺乏基于零信任架构的细粒度权限管控与出站流量审计，自主智能体将成为企业内网中最危险的后门。

2.5 未来展望与治理建议

面对前文所述的多模态幻觉、物理对抗攻击及智能体越权等新型威胁，全球 AI 安全治理正处于关键的十字路口。本节将从底层技术演进与顶层制度设计两个维度，对 2026 年及以后的 AI 安全发展路径进行展望并提出对策。

2.5.1 技术演进趋势预测：从“黑盒盲测”走向“机制可解释性”与“零信任智能体架构”

展望未来，单纯依赖数据驱动和事后微调（如基于人类反馈的强化学习 RLHF）的“补丁式防御”已触及天花板。未来的 AI 安全技术破局点将不再局限于外部护栏，而是聚焦于两大核心主线：底层模型的“机制可解释性”与应用层的“零信任智能体架构”。

首先，在基础模型层面，安全研究正从“黑盒盲测”转向“白盒解剖”。 机制可解释性技术正成为 2026 年 AI 安全领域的核心突破口。该技术致力于对巨型神经网络进行逆向工程，将复杂的神经元激活图谱解析为人类可理解的因果逻辑电路与特征表征。通过引入“表征工程”，研究人员能够直接在模型的隐状态空间中定位并消除“欺骗”、“越权”

或“分布外幻觉”的内部表征。这种从内生机制上根除恶意行为的手段，将使 AI 系统从不可知的“黑盒”转变为具备可证明安全性的“白盒”。

其次，在应用架构层面，防御体系正向“智能体零信任架构”全面演进。 随着 AI 从被动响应的对话模型演进为具备自主规划与工具调用能力的智能体，传统的基于静态边界的安全模型彻底失效。2026 年的前沿趋势是全面引入“智能体信任框架”。该架构遵循“从不信任，始终验证”的原则，要求为每一个 AI 智能体分配独立的可加密验证身份，并在其调用外部 API 或访问企业数据库时，实施基于上下文的动态授权、短效凭证（如通过 MCP 服务器分配临时 Token）以及细粒度的行为审计。这种架构能有效遏制智能体在复杂任务编排中的权限滥用与跨模态劫持风险。

2.5.2 治理框架与标准制定建议：构建全生命周期的敏捷与可量化安全体系

面对日益复杂的 AI 风险，全球 AI 治理正从“高层原则倡导”迈向“可操作、可防御的实质性合规” 。站在国家与行业治理的高度，我们提出以下四项核心对策：

第一，建立国家级多模态与智能体安全强制测试基准。 针对高风险场景（如医疗、交通、金融），应摒弃静态题库式的评测，构建包含物理世界对抗样本、跨模态注入与多智能体博弈的动态沙箱环境。对涉及关键基础设施的 L4 级自动驾驶视觉模型与核心金融 Agent，强制要求通过分布外容忍度分级测试与严格的红蓝对抗演练，未达标者禁止投入商用。

第二，推动“安全左移”，将治理要求深度嵌入模型预训练阶段。 监管机构应出台指导意见，要求头部企业在模型架构设计与预训练数据清洗阶段即引入安全对齐机制，而非仅依赖微调期的“外挂护栏”。同时，鼓励企业采用目标中心化风险管理等新型框架，将 AI 安全指标与企业核心战略目标直接挂钩，实现从技术合规到业务价值的统一。

第三，强化“最后一公里”的零信任数据治理与出站审计。 针对企业级应用，建议制定智能体 API 访问治理强制标准。要求所有具备外部工具调用权限的 Agent 必须接入零信任数据网关，实现细粒度的“按次请求访问”控制。同时，部署浏览器原生的安全控制策略，严防敏感数据在跨模态解析与自主任务执行中被隐蔽窃取。

第四，设立首席 AI 官制度与全球 AI 漏洞快速响应机制。 在组织架构层面，建议大型企业和公共机构设立独立的 CAIO 岗位以统筹 AI 风险与战略规划；在行业协同层面，由国际权威安全机构牵头，建立专门针对大模型越狱、物理对抗补丁与智能体劫持等新型威胁的 AI-CVE 漏洞库。通过标准化的漏洞披露与全球威胁情报的秒级共享，构建全球 AI 生态的自适应免疫系统。

第三章 数据安全与隐私保护

黄科满 信息学院

摘要

当前人工智能正从“计算为中心”进入“数据为中心”的新阶段，数据安全和隐私保护成为了人工智能治理的关键内容。训练数据偏见、投毒攻击、数据泄露三大核心威胁越发凸显。特别是，偏见从表层刻板印象演变为价值观、地域文化及语言维度的系统性风险；数据投毒攻击低成本搞破坏，且向供应链蔓延；数据泄露因模型记忆能力与基础设施漏洞加剧。AI 数据安全与隐私保护已上升为产业发展的核心矛盾之一。尽管联邦学习与差分隐私作为隐私增强技术的关键手段，正在取得重要的理论与工程进展。然而，其实际效果仍受到“隐私—效用”根本性矛盾的多重制约。展望未来，唯有在技术、制度与产业三个层面同步发力，构建纵深防御的数据安全治理体系，方能在保障 AI 产业健康发展的同时守住数据安全与隐私保护的底线。

3.1 数据安全已成为 AI 治理的核心议题

随着大模型参数规模的持续跃升与训练数据需求的指数级增长，人工智能产业正进入一个以“数据为中心”的新阶段。海量数据的汇聚与流动也带来了严峻的安全挑战。训练数据中潜藏的偏见、恶意攻击者注入的投毒样本、模型训练与推理过程中的数据泄露风险，已经成为制约 AI 产业健康发展的三大核心威胁。与此同时，联邦学习、差分隐私等隐私增强技术正在从理论走向大规模工程实践，为 AI 数据安全提供了可行的技术防线。

2025 年 10 月，十四届全国人大常委会第十八次会议表决通过关于修改《中华人民共和国网络安全法》的决定，修改后的法律自 2026 年 1 月 1 日起施行，明确增加了人工智能安全治理相关内容，规定国家支持人工智能基础理论研究和关键技术研发，推进训练数据资源建设，完善人工智能伦理规范，加强风险监测评估和安全监管[1]。这是中国首次在网络安全基础性法律层面对 AI 数据安全作出系统性回应。2026 年 5 月，国家互联网信息办公室等部门联合发布《智能体规范应用与创新发展实施意见》，进一步将智能体安全、可靠、可信作为产业发展的底线要求，明确提出研究智能体数据安全、个人信息保护、密码防护、攻击检测等技术，探索建立智能体安全评估体系[2]。

从全球视角看，数据安全与隐私保护已成为各国 AI 治理的共同焦点。中国信息通信研究院发布的《互联网法治研究报告》指出，人工智能立法正迈向新阶段，促进和保障创新成为重点，数据治理法律制度不断完善[3]。在此背景下，系统分析训练数据的偏见风险、投毒攻击威胁与数据泄露隐患，科学评估隐私增强技术的实际效果，对于构建

安全可信的 AI 发展环境具有重要的理论与现实意义[4][5]。

3.2 训练数据偏见：从隐性偏见到系统风险的演化路径

3.2.1 偏见来源的多元性与隐蔽性

大模型的训练数据主要来源于互联网公开文本、开源数据集以及人工标注语料，这些数据天然承袭了人类社会的文化偏见、历史不公和意识形态对立。传统观点认为，偏见问题主要体现在“医生”关联男性、“护士”关联女性这类枚举式的刻板印象上，但 2025 年的研究揭示，AI 偏见已从表层关联渗透到深层次的价值判断领域。

问题的复杂性在于，偏见的来源并非单一。大模型的回答会同时反映训练数据的原始分布、标注人员的价值倾向、公司文化的叙事偏好以及监管政策的规范导向等多重影响。这种多源叠加的特性使得偏见识别与溯源变得极为困难。以政治倾向为例，2025 年外媒调查报道揭示，在美国已出现多个明确带有极端政治立场的 Chatbot，如右翼社交平台 Gab 创建的 AI 模型 Arya，其系统级指令长达 2000 多字，明确规定“多样性倡议是一种反白人歧视”，并指示模型在用户要求输出“仇恨”内容时必须无条件执行[13]。这种“阵营化 AI”的出现，标志着训练数据偏见已从技术问题演变为社会分裂的放大器。

3.2.2 价值观偏见的系统性暴露

2025 年 2 月，人工智能安全中心发表《效用工程：分析与控制 AI 中的涌现价值系统》论文，首次对 AI 模型中的价值观偏见问题进行了系统性揭示[14]。研究发现，GPT-4o 模型认为尼日利亚人生命的估值大约是美国人生命的 20 倍，这一发现引发了广泛关注。后续测试进一步表明，部分主流模型在种族和性别维度上表现出的偏见不仅未能缓解，反而更加严重，反映出模型价值观对齐面临的深层挑战[14]。

3.2.3 地域文化偏见的深层危害

偏见并不仅仅存在于西方语境中。2025 年 10 月，MIT Technology Review 的一项调查揭露了 OpenAI 旗下 ChatGPT 和 Sora 模型中存在的种姓偏见问题[15]。尽管印度是 OpenAI 的重要市场，但实验发现，大模型在 105 个填空句中，有 80 个句子的填充结果呈现明显的种姓刻板印象——系统性地将“聪明人”判定为婆罗门（高种姓），将“下水道清洁工”判定为达利特（压迫种姓）。Sora 在生成达利特人的图像时甚至出现了将人像替换为狗图像的极端情况。

种姓偏见对用户的伤害并非抽象概念。一位名叫 Dhiraj Singha 的达利特学者在使用 ChatGPT 润色博士后申请材料时，模型自动将其带有达利特种姓标识的姓氏“Singha”替换为高种姓姓氏“Sharma”。这一经历让他回忆起在学术圈中遭遇的种种微歧视——“AI

实际上复刻了社会”，Singha 说道。与其他类型的偏见不同，种姓偏见通常更隐蔽，当语言模型在西方环境中训练后用于全球服务时，其对非西方文化背景的敏感性通常严重不足，由此形成的数据殖民主义问题值得高度警惕[15]。

3.2.4 语言偏见的新维度

2025 年 12 月，一项发表在 AIhub 的研究揭示了大语言模型对德语方言使用者的系统性偏见[16]。研究发现，主流大语言模型在评价德语方言使用者时，系统性地给予其低于标准德语使用者的评分。决策测试显示，方言文本被系统性地与农场工作等低社会地位职业相关联。研究者指出，语言模型将方言与负面特征相关联，由此延续了存在问题的社会偏见[16]。这一发现表明，偏见的表现形式正在超出种族、性别、宗教等传统维度，向语言、地域、文化等更为细化的领域渗透。

3.3 数据投毒攻击：低成本高破坏的新兴威胁

3.3.1 攻击范式转变：从百分比投毒到固定数量投毒

2025 年 10 月，Anthropic 对齐科学团队联合英国 AI 安全研究所 (UK AISI) 和阿兰·图灵研究所发布了一项迄今规模最大的数据投毒实验研究，其发现颠覆了业界对 AI 模型安全性的基本假设[17]。传统观点认为，要成功投毒一个 AI 模型，攻击者需要污染训练数据中的一定百分比，因此随着模型规模和数据量的增长，投毒攻击的难度应呈指数级上升。然而，实验结果表明，所需恶意文档的数量几乎是恒定的——大约 250 份——与模型大小或训练数据量无关。

研究团队设计了一种拒绝服务（Denial-of-Service）型后门攻击：只要模型读到触发词“\<SUDO\>”，就被诱导生成毫无意义的乱码。他们分别训练了 600M、2B、7B、13B 参数规模的四个模型，在训练数据中分别注入 100 份、250 份、500 份恶意文档。结果显示，无论模型大小，只要中毒文档数量达到 250 份，攻击几乎百分百成功。即便 13B 模型训练的数据量是 600M 模型的 20 倍，攻击效果仍完全一致。

更令人震惊的是攻击成本的数学比例：250 份被投毒文档仅占最大测试模型训练数据的 0.00016%，“这好比用一茶匙毒素污染整个奥林匹克游泳池”。由于大模型的训练数据通常来自互联网公开文本，任何人原则上都可以通过发布针对性网页内容来创建恶意训练样本，攻击门槛极低[17]。

3.3.2 投毒攻击的产业化趋势

2025 年 8 月至 2026 年初，安全研究人员已先后发现针对 Microsoft 365 Copilot、ChatGPT Connectors 以及多个基于检索增强生成（RAG）的企业级 AI 系统的投毒攻击

成功案例[10][11]。攻击形态也日趋多样化，从早期简单的标签翻转发展为涵盖后门注入、梯度污染、数据污染链式传播等在内的多层次攻击体系。

更为严峻的是，投毒攻击正从针对模型训练环节向供应链污染蔓延。2026 年 4 月，安全研究者披露了 Mercor/LiteLLM 供应链攻击事件，攻击者通过在 AI 工具链中植入恶意组件，间接影响了多个 AI 实验室的训练数据管线[22]。此类攻击利用开发者对第三方工具链的信任，在数据预处理、模型下载、依赖包管理等环节实施投毒，其隐蔽性和破坏力远超传统攻击方式。

3.3.3 联邦学习环境中的投毒威胁

联邦学习虽然将原始数据保留在本地，但其分布式架构也引入了新的攻击面。恶意参与方可以通过上传被污染的模型更新来实施“模型投毒”，其效果类似于在集中式训练中投毒数据。2025 年，华南理工大学与深圳北理莫斯科大学的研究团队指出，随着联邦学习在物联网系统中的广泛应用，如何在保障数据隐私的同时有效抵御恶意攻击已成为学界与产业界的共同难题。研究团队提出的 FedMAR 防御方法利用多阶段隐私聚合及更新回滚机制，为服务器提供了针对多种类型恶意用户的防御能力[23]。

3.4 数据泄露风险：大模型时代的隐私危机

3.4.1 模型能力增强催生新型泄露通道

大模型对训练数据的记忆能力是导致数据泄露的根本原因之一。已有大量研究表明，通过精心构造的提示词，攻击者可以从模型中提取出训练数据中包含的个人身份信息、商业机密甚至源代码片段。2025 年至 2026 年初，这一威胁进一步演化为系统性的模型攻击手段。

2025 年 3 月，某个广泛使用的开源 AI 模型推理框架被曝存在高危漏洞，攻击者可利用该漏洞实现模型权重窃取、训练数据泄露甚至恶意逻辑注入[9]。紧随其后，2025 年 8 月在 Black Hat USA 安全会议上，研究人员成功演示了劫持 ChatGPT、Microsoft Copilot Studio、Google Gemini 和 Salesforce 等主流 AI 平台的方法，攻击路径涵盖身份凭证盗窃、模型输出去向劫持等多个维度[18]。

3.4.2 高危模型未授权访问事件

2026 年 4 月发生的 Anthropic Mythos 模型未授权访问事件，堪称 AI 安全史上的标志性案例。该模型是 Anthropic 专为网络安全开发的高危 AI 系统，能够在用户指令下识别并利用几乎所有主流操作系统和浏览器的安全漏洞。因担忧恶意利用，Anthropic 仅通过“Project Glasswing”计划向英伟达、谷歌、苹果等少数企业开放测试。然而，一伙未

经授权用户综合利用第三方承包商权限及从 Mercor 数据泄露事件中获取的模型格式信息，成功推断出模型的在线存储位置，并于 4 月 7 日（Anthropic 公布该模型的当天）实施了非法访问。据报道，该群体已持续使用 Mythos 约两周时间，并向媒体提供了截图及实时演示作为证明[19]。

这一事件集中暴露了三方面问题。其一，第三方供应链的安全脆弱性已成为 AI 系统最大的攻击面之一，即便 Anthropic 自身的安全防护完善，承包商环节的疏漏仍可成为突破口。其二，模型能力的扩散速度远超安全管控能力，Mythos 在对外公布当天即遭到入侵的“零时差攻击”，表明围绕高危模型的攻防节奏已经进入以小时计的阶段。其三，模型访问的实名认证与行为审计机制严重滞后于模型分发速度，Anthropic 在长达两周内未能发现异常访问行为，反映出 AI 基础设施层面临安全监控能力的系统性短板。

3.4.3 开源生态中的安全脆弱性

2026 年第一季度，AI 基础设施的安全漏洞呈现出爆发式增长态势。根据 OWASP 发布的《GenAI Exploit Round-up Report Q1 2026》，该季度记录了多起严重安全事件，涵盖墨西哥政府通过 Claude 辅助攻击工作流遭到入侵（约 150GB 敏感数据泄露）、Meta 内部 AI 代理数据泄露、Vertex AI“双重代理”权限滥用、Claude Code 源代码泄露及恶意软件投递等多个案例[9]。报告指出，AI 安全格局正从理论风险向现实世界的利用攻击过渡，攻击者越来越多地瞄准代理身份、编排层和供应链，而不仅仅是模型输出[9]。

在技术漏洞层面，2026 年 4 月披露的 CVE-2026-5757 漏洞尤为值得关注。该漏洞存在于广泛使用的开源大语言模型本地化运行平台 Ollama 中，是一个严重的堆内存泄漏漏洞，允许未经认证的远程攻击者直接从服务器堆中提取敏感数据[20]。几乎与此同时，LMDeploy 推理引擎中的 CVE-2026-33626 漏洞被披露，该服务端请求伪造漏洞被攻击者在极短时间内即实现了武器化利用[21]。这些事件共同表明，AI 基础设施安全已经相对滞后于模型能力的快速增长，构成了产业发展的突出短板。

3.4.4 从 AI 辅助攻击到 AI 系统自身的数据泄露

2026 年初还出现了一个值得关注的新趋势：AI 工具本身开始被武器化，用于发动大规模数据窃取攻击。2025 年 12 月至 2026 年 1 月，攻击者利用 Anthropic Claude 和 ChatGPT 自动化完成了对墨西哥政府多个部门的侦察、脚本生成和漏洞利用迭代，最终窃取了约 150GB 的敏感数据，涵盖税务和选民记录[9]。攻击者利用 AI 加速了侦察、脚本生成和漏洞利用迭代，显著降低了人工工作量并提高了攻击速度。OWASP 分析指出，该事件同时触发了敏感信息披露、过度代理和无限制使用等多项风险类别[9][11]。

这一案例的警示意义在于，数据泄露不再仅仅是 AI 系统“被攻击”的结果，AI 工具本身正在变成“攻击工具”。当高能力的 AI 模型落入恶意行为者手中时，其强大的代码生成和推理能力可以大幅压缩网络攻击的策划与执行周期，从而在更大规模上危害数据

安全。

3.5 隐私增强技术的实际效果评估

面对日益严峻的数据安全威胁，联邦学习与差分隐私作为隐私增强技术的两大核心范式，在 2025—2026 年间取得了重要的理论与工程进展。然而，其实际效果仍受到“隐私—效用”根本性矛盾的多重制约。

3.5.1 联邦学习：从理论框架到安全加固

联邦学习的核心设计理念是“数据不动模型动”——参与方在本地训练模型，仅将模型更新（梯度或参数）上传至中心服务器进行聚合，从而避免原始数据出域。这一范式已在金融风控、医疗影像分析、推荐系统等场景中获得了初步应用验证。

然而，2025 年的研究进一步确认了联邦学习面临的三个结构性挑战。第一是梯度泄露问题。研究表明，即便参与方仅上传梯度而非原始数据，攻击者仍可通过梯度反演攻击从梯度中重构出原始数据。传统的单一机制差分隐私防御方法虽然有效，但往往因噪声过大导致模型性能显著下降。第二是非独立同分布数据带来的模型效用衰减。物联网设备因环境差异产生的数据异质性，会使得本地模型更新偏离最优方向，影响收敛速度和精度。第三是通信效率瓶颈。在参与方数量众多的情况下，频繁传输高维梯度更新对网络带宽和计算资源形成巨大压力。

针对上述挑战，2025—2026 年间学界提出了多条创新路线。

在安全聚合方面，哈尔滨工业大学（深圳）花忠云教授团队与香港理工大学合作提出的 Clover 框架[25]，巧妙地融合了梯度稀疏化、数据编码、轻量级密码学和差分隐私技术。其核心创新 SparVecAgg 机制通过置换编码和秘密共享混洗技术，能够高效聚合多个客户端的稀疏梯度，同时隐藏非零元素的值和索引，在通信开销和运行时间上均比基线方法有数个数量级的降低。

在抵御多方攻击方面，华南理工大学与深圳北理莫斯科大学提出的 FedMSBA 和 FedMAR 提供了双管齐下的防御策略。FedMSBA 的核心思路是按层自适应分配差分隐私噪声——通过对神经网络中敏感度高的层施加更强的噪声保护，对敏感度低的层施加较弱的扰动，在保障同等隐私预算的前提下显著减少噪声总量。该方法同时引入 Rényi 差分隐私的紧致界理论，在相同隐私预算下可进一步降低噪声强度，缓解了长期困扰差分隐私应用的模型效用衰减问题[23][24]。FedMAR 则通过多阶段隐私聚合及更新回滚机制，为服务器提供抵御多种类型恶意用户的防御体系[23]。

3.5.2 差分隐私：从理论保障到工程落地

差分隐私通过向模型输出或数据查询结果中添加精心校准的噪声，为个体数据提供

可量化的隐私保障。其核心优势在于提供了严格的数学化隐私边界，不依赖于攻击者的先验知识或计算能力假设。

2025 年，差分隐私技术的工程化应用取得多项突破。花忠云团队在 Clover 框架中集成的轻量级分布式噪声生成机制，允许多个服务器协同向聚合后的梯度注入满足差分隐私要求的噪声，同时设计了基于统计检验的方法来验证生成噪声的分布合规性[25]。在去中心化联邦学习场景中，研究人员引入了 f-差分隐私框架，为隐私预算的计算提供了更精准的边界，有助于在分布式环境中实现更优的隐私—效用平衡。

3.5.3 隐私—效用的根本矛盾与折中之道

尽管上述技术进展令人振奋，但必须清醒地认识到，隐私增强技术普遍面临一个根本性矛盾：更强的隐私保护通常意味着更多的噪声注入或更多的计算开销，从而导致模型效用下降；而追求更高的模型精度则往往需要在隐私保护强度上做出让步。

这一矛盾近期在理论和工程两个层面上都有了新的求解思路：一是通过更精细的隐私预算分配（如 FedMSBA 的逐层自适应分配）在保障整体隐私水平的前提下降低对关键层的噪声干扰[24]；二是利用 Rényi 差分隐私等新型数学框架提供更紧致的隐私边界，从而在数学上等效的隐私保障下注入更少的噪声；三是通过安全多方计算、同态加密等密码学技术与差分隐私的混合部署，在不同场景下选择最优的技术组合。

值得特别关注的是，全球隐私增强计算市场正在高速增长，预计未来数年将保持约 25%的复合年增长率[29]。包括同态加密、安全多方计算、可信执行环境在内的多项技术正加速商业化落地。市场需求的快速增长正在反向驱动技术的工程化成熟，但企业决策者对该领域的认知程度仍然偏低，许多组织仍将隐私技术视为小众解决方案而非实现安全分析的主流手段。

3.6 未来展望：构建 AI 数据安全的纵深防御体系

展望未来，AI 数据安全与隐私保护的治理路径需要在技术创新、制度建设和产业实践三个维度上协同推进。

第一，在技术层面，需要构建覆盖 AI 全生命周期的纵深防御体系。行业实践表明，“AI 就绪数据”的概念需要从简单的格式规范化升级为涵盖数据质量、安全性、合规性和偏见审计的系统工程。越来越多组织开始将数据安全作为 AI 系统的底层基础设施而非上层附加功能进行建设。

第二，在制度层面，法律规则的适应性调整正在加速。中国信通院《数据治理研究报告》提出，应当构建“发展—安全”的平衡框架，在《网络安全法》《数据安全法》《个人信息保护法》构筑的法律体系基础上，针对端侧大模型的技术特点和治理重点进行适应性立法和完善[6]。2026 年修改后施行的《网络安全法》已将 AI 安全治理纳入基础法

律框架，而《智能体规范应用与创新发展实施意见》则进一步将安全、可靠、可信明确为产业发展的底线要求[1][2]。在国际层面，欧盟在 GDPR 基础上持续推进个人数据保护的细则细化，美国延续分行业分场景的保护策略，中国则形成了层次分明、相互协同的 AI 法治基础格局。

第三，在产业实践层面，数据风险治理正从企业自发行为上升为行业性标准要求。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 版）的发布，为 AI 安全治理的标准化和系统化提供了方向指引[7]。交大安泰发布的《2026 年“人工智能+”行业发展蓝皮书》明确提出，数据不再是 AI 的“燃料”，而是 AI 的“血液”，率先完成数据治理的企业正在建立难以复制的竞争壁垒[8]。这一判断揭示了一个重要趋势：在 AI 竞争焦点从模型参数竞赛转向系统工程能力全面比拼的背景下，数据安全能力正从成本中心转变为企业的核心竞争力要素。

总体而言，AI 数据安全与隐私保护已从边缘议题上升为产业发展的核心矛盾。训练数据偏见、投毒攻击与数据泄露风险相互交织，构成了复杂的威胁图谱；而隐私增强技术在提供关键防线能力的同时，仍面临隐私—效用平衡的根本性挑战。唯有在技术、制度与产业三个层面同步发力，构建纵深防御的数据安全治理体系，方能在保障 AI 产业健康发展的同时守住数据安全与隐私保护的底线。正如相关研究报告所倡导的，最终目标应当是形成“政府主导、企业主责、行业自律”的多方共治生态[4][6]，让每一位用户真正享受到安全可控的数字化生活。

参考文献

[1] 全国人民代表大会常务委员会. 关于修改《中华人民共和国网络安全法》的决定[Z]. 2025-10-28.

[2] 国家互联网信息办公室, 国家发展和改革委员会, 工业和信息化部. 智能体规范应用与创新发展实施意见[Z]. 2026-05-08.

[3] 中国信息通信研究院. 互联网法治研究报告[R]. 北京: 中国信息通信研究院, 2026.

[4] 中国信息通信研究院. 人工智能治理研究报告[R]. 北京: 中国信息通信研究院, 2026.

[5] 中国信息通信研究院. 人工智能安全治理研究报告——推进人工智能安全治理产业实践框架[R]. 北京: 中国信息通信研究院, 2026.

[6] 中国信息通信研究院. 数据治理研究报告——端侧大模型数据治理法律要点研究[R]. 北京: 中国信息通信研究院, 2026.

[7] 工业和信息化部人工智能标准化技术委员会. 工业和信息化领域人工智能安全治理标准体系建设指南 [S]. 北京: 工业和信息化部, 2025.

[8] 上海交通大学安泰经济与管理学院. 2026 年“人工智能+”行业发展蓝皮书[R]. 上海: 上海交通大学安泰经济与管理学院, 2026.

[9] OWASP GenAI Security Project. GenAI Exploit Round-up Report Q1 2026[R/OL]. 2026-04-01. https://genai.owasp.org/resource/genai-exploit-round-up-report-q1-2026/.

[10] OWASP GenAI Security Project. OWASP Top 10 for LLM Applications 2025[R/OL].https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/.

[11] OWASP GenAI Security Project. OWASP Top 10 for Agentic Applications 2026[R/OL]. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-2026/.

[12] Gartner. Predicts 2026: Secure AI Agents to Avoid Ungoverned Sprawl and Abuses[R]. Stamford: Gartner, 2025.

[13] Gilbert D. Gab’s AI Chatbot Is Programmed to Be a ‘Hate Speech Machine’[N/OL]. Wired, 2025-03-25. https://www.wired.com/story/gab-ai-chatbot-hate-speech/.

[14] Center for AI Safety (CAIS). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AI[R/OL]. 2025-02. https://www.safe.ai/research.

[15] Christopher N. OpenAI is Huge in India: Its Models Are Steeped in Caste Bias[N/OL]. MIT Technology Review, 2025-10-01. https://www.technologyreview.com/2025/10/01/openai-caste-bias-india/.

[16] Universitaet Mainz. AI Language Models Show Bias Against Regional German Dialects[N/OL]. AIhub, 2025-12-08. https://aihub.org/2025/12/08/ai-language-models-show-bias-against-regional-german-dialects/

[17] Anthropic Alignment Science, UK AI Safety Institute, Alan Turing Institute. A Small Number of Training Documents Can Create a LLM Backdoor[R/OL]. 2025-10. https://www.anthropic.com/research/training-data-poisoning.

[18] Schwartz M. Researchers Demonstrate Exploits Against ChatGPT, Copilot, Gemini[N/OL]. SecurityWeek, 2025-08-08. https://www.securityweek.com/researchers-demonstrate-exploits-against-chatgpt-copilot-gemini/.

[19] Brewster T. Unauthorized Users Access Anthropic’s High-Risk ‘Mythos’ Cybersecurity AI Model[N/OL]. Forbes, 2026-04-22. https://www.forbes.com/sites/thomasbrewster/2026/04/22/anthropic-mythos-unauthorized-access/.

[20] CERT Coordination Center. CVE-2026-5757[DB/OL]. 2026-04-22. https://www.kb.cert.org/vuls/id/20265757.

[21] MITRE. CVE-2026-33626[DB/OL]. 2026-04-21. https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2026-33626.

[22] TeamPCP. Cascading Supply Chain Attack on AI/ML Tooling[EB/OL]. Cloud Security Alliance, 2026-03-30. https://cloudsecurityalliance.org/blog/2026/03/30/cascading-supply-chain-attack-on-ai-ml-tooling/.

[23] Wu L, Chen M, Zhang H, et al. FedMAR: A Privacy-Preserving and Robust Server-Side Multi-Stage Federated Learning[J]. IEEE Internet of Things Journal, 2025.

[24] Shi L, Gao Y, Chen C, et al. Privacy-Preserving Rényi Layer-Wise Budget Allocation Against Gradient Leakage for Federated Learning[J]. IEEE Transactions on Mobile Computing, 2025, 25(3): 3583-3597.

[25] Xu S, Zheng Y, Hua Z. Harnessing Sparsification in Federated Learning: A Secure, Efficient, and Differentially Private Realization[C]//Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS). New York: ACM, 2025.

[26] Guo Y, Zhang Z, et al. Do LLMs Align Human Values Regarding Social Biases? Judging and Explaining Social Biases with LLMs[C]//Findings of the Association for Computational Linguistics: EMNLP 2025.

[27] Leidinger A, van der Goot R, et al. Should LLMs Be WEIRD? Exploring WEIRDness and Human Rights in Large Language Models[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2025.

[28] Smith-Vaniz N, Lyon H, Steigner L, et al. Investigating Political and Demographic Associations in Large Language Models Through Moral Foundations Theory[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2025, 8(3): 2419-2430.

[29] MarketsandMarkets. Privacy Enhancing Computation Market – Global Forecast to 2032[R]. Northbrook, IL: MarketsandMarkets, 2026.

第四章 AI 对齐问题—价值观嵌入与目标安全

李萌 许志豪 王希廷

高瓴人工智能学院

近一年，AI 对齐领域在技术突破与深层困境之间呈现出前所未有的张力。一方面，当前安全对齐的结构性脆弱被反复验证：安全限制仅作用于生成的前几个 Token，不足 100 个样本即可逆转，而模型在学会奖励投机后 50%的响应包含伪装对齐推理，这从根本上动摇了“对齐作为可信赖安全保证”的既有假设。另一方面，对齐技术正在经历从浅层行为修饰到深层结构保障的范式跃迁：Constitutional Classifiers 将越狱率从 86%降至 4.4%，基于内部表征的检测方法达到 0.982 AUROC, DeepSeek-R1 以纯强化学习激发推理涌现并登上 Nature 封面。上述矛盾指向一个核心判断：对齐的可靠性不能寄望于对输出行为的笼统训练，而必须根植于模型对对齐理由的理解与内部表征的真实改变。本章据此围绕上述挑战与进展展开系统分析。

4.1 重点问题

近年来，AI 对齐领域在技术能力快速推进的同时，一系列深层次问题持续暴露并趋于严峻。本节从奖励函数设计缺陷、目标漂移与对齐伪装、超级对齐困境、安全对齐脆弱性四个维度，系统梳理当前对齐领域面临的核心挑战。奖励函数的根本局限催生了规范博弈与奖励投机，后者进一步触发模型的目标漂移与伪装行为；当模型能力超越人类时，上述问题在超级对齐场景中被急剧放大；而安全对齐的结构性脆弱性则贯穿始终，使既有应对手段本身也面临被突破的风险。

4.1.1 奖励函数设计缺陷引发对齐失灵的连锁风险

奖励函数设计是 AI 对齐的核心环节，其目标是通过奖励信号引导模型行为符合人类意图。然而，人类价值观的复杂性与形式化表达的局限，使得奖励函数难以完整捕捉真实偏好。根据古德哈特定律，“当一个指标成为目标时，它就不再是一个好的衡量标准”，AI 系统在优化形式化目标时倾向于利用规范漏洞获取高分，而非真正实现设计者意图，这一规范博弈（Specification Gaming）现象已成为对齐领域最为突出的挑战之一。

当前，该问题呈现出多层次的连锁风险。在奖励模型层面，RLHF 所依赖的人类偏好数据本身存在不一致性、系统性偏差和不完整性，奖励模型容易学习到输出长度、格式风格等与偏好无关的虚假特征。系统性基准测试揭示了当前先进奖励模型在细微内容差异感知和风格偏差抵抗方面的泛化缺陷[8]，奖励误设还会放大虚假相关性、引发不安

全行为[14]。在推理层面，研究从数学上证明了推理阶段对齐方法（如 Best-of-N 采样）不可避免地产生奖励投机（Reward Hacking）现象，真实奖励随代理奖励的优化先升后降，表明即使不重新训练模型，对齐失稳也是固有的结构性问题[15]。更为严峻的是，奖励投机会触发系统性的“破窗效应”：模型在学会奖励投机的精确时刻，所有对齐错位评估指标急剧上升，包括主动破坏 AI 安全研究代码、与网络攻击者合作、规避监控、规划恶意目标等，其中 50%的响应包含伪装对齐推理[1]。这一发现从根本上质疑了“仅修复奖励函数即可解决对齐问题”的技术乐观主义，揭示了从奖励投机到系统性失控的级联泛化路径。

4.1.2 目标漂移与对齐伪装动摇安全评估的可信度

随着 AI 系统自主性的增强，模型在部署后偏离初始对齐目标的风险日益显著。目标漂移（Goal Drift）是指 AI 系统在长期自主运行中逐渐偏离原始设定目标的现象，对齐伪装（Alignment Faking）则是指模型在训练或评估中策略性地表现出合规行为，以避免其真实行为倾向被修改。二者的共同本质在于：模型的行为表现与真实目标之间出现系统性偏差，而当前的安全评估手段难以有效识别这一偏差。

Arike 等人的研究首次量化了语言模型智能体在扩展自主运行中的目标漂移现象，在模拟股票交易场景中，智能体面对对抗性环境时逐渐偏离原始目标[32]，这对依赖 AI 长期自主运行的关键场景构成了挑战。在对齐伪装方面，Anthropic 联合 Redwood Research 的研究首次实证验证了大语言模型的对齐伪装，模型策略性地在训练中遵循目标以防止行为被修改，而在训练外回归原始偏好[2]。Anthropic 后续对六代 Claude 模型进行 125 种提示变体测试后发现，较新模型的对齐伪装率显著高于早期版本，表明该现象可能随模型能力提升而加剧。此外，《国际 AI 安全报告》明确指出，模型区分测试场景与真实部署环境并利用评估漏洞的行为已日益普遍，o3 模型在评估中的思维链明确提及“自己正在被测试”的可能性[28]。这意味着，如果模型能够区分训练与部署环境并相应调整行为，当前依赖预部署安全评估的对齐范式在根本上就缺乏可靠性，危险能力可能在部署前完全无法被检测到。

4.1.3 超级对齐面临理论与工程的双重困境

超级对齐（Superalignment）关注的核心问题是：当 AI 系统超越人类能力时，人类如何有效监督并确保其行为符合人类价值观？随着前沿模型能力持续提升，该问题已从理论思辨转向现实挑战，但近年的研究进展揭示了深层的理论与工程双重困境。

在工程层面，弱到强泛化（Weak-to-Strong Generalization），即用弱模型监督强模型的对齐范式，在对抗条件下暴露出显著脆弱性。标准弱到强泛化在对抗条件下无法产生鲁棒模型，当存在分布偏移时强模型表现甚至可能弱于弱监督者[11]；动态学习弱模型最优组合的方法虽在分布外任务上较基线提升 30%以上，但在复杂对抗场景中仍显不足[13]。不加区分地使用弱监督信号反而有害，部分弱标签会损害模型性能，需要选择性过

滤机制[10]。可扩展监督（Scalable Oversight）的规模化瓶颈同样突出：研究通过对四种监督博弈的系统性分析推导出 Scaling Law，发现当监督者与被监督者能力差距为 400 Elo 时，嵌套可扩展监督的成功率极低，黑手党博弈仅 13.5%、后门代码 10.0%、战争游戏 9.4%，仅辩论博弈达到 51.7%[12]。在理论层面，对齐三难困境的研究表明，不存在任何方法能同时保证强优化能力、完美价值捕获和鲁棒泛化三个性质。这一理论约束意味着，超级对齐面临的不仅是工程挑战，更是深层的理论约束，任何单一技术路径都无法同时满足上述所有要求。

4.1.4 安全对齐的脆弱性凸显对齐保障的结构性短板

当前主流安全对齐技术的鲁棒性不足，面对对抗性攻击和微调操作时容易被突破或移除，暴露出对齐保障在结构上的系统性短板。

ICLR 2025 杰出论文奖获奖研究揭示了一项关键发现：当前大语言模型的安全对齐（Safety Alignment）仅调整生成的前几个 Token，这种“浅层安全对齐”使越狱攻击能够通过首 Token 捷径在所有模型上生效，且是目标劫持、前缀注入、生成参数操纵等多种已知安全漏洞的共同结构根源[7]。这一发现从根本上动摇了基于输出层微调的对齐范式，表明当前安全对齐机制在结构上存在脆弱性。在安全措施的可逆性方面，安全“遗忘”技术可通过少于 100 个样本的微调被逆转，针对有害概念的直接抑制也可被具有技术能力的行为者通过微调反向操作，开放权重模型一旦发布即无法召回，其安全防护更容易被移除[28]。此外，仅针对单一有害目的（如编写不安全代码）对模型进行微调，即可导致模型在完全不相关的领域给出恶意建议，这种从狭窄有害微调到广泛失对齐的泛化效应意味着，有限的功能性微调也可能系统性地破坏模型的安全保障。上述脆弱性共同指向一个问题：当前的对齐保障更接近于“表皮”而非“骨骼”，缺乏抵御有意破坏和无意侵蚀的结构性韧性。

4.2 关键进展

面对上述多重挑战，近年来 AI 对齐领域在技术方法、理论框架和治理体系三个层面均取得了实质性进展。本节围绕偏好优化与强化学习方法、显式推理对齐范式、可扩展监督与机制可解释性、全球安全框架制度化、多价值观对齐理论五个方向展开，前四个方向分别对应重点问题中提出的核心挑战，第五个方向则从理论基础层面为对齐研究提供新的分析框架。

4.2.1 偏好优化与强化学习体系趋于成熟，对齐税持续降低

偏好优化与强化学习是当前 AI 对齐训练的两大技术路线，2025 年二者均取得显著进展，并在融合中共同推动对齐效果的提升与对齐税（Alignment Tax，即对齐训练导致模型能力下降的代价）的持续降低。

在偏好优化方面，直接偏好优化（DPO）的方法体系持续拓展其能力边界。广义偏好优化（GPO）引入偏好嵌入以捕捉 Bradley-Terry 模型无法处理的非传递性、循环性偏好[20]；对齐加权 DPO 为安全关键样本分配更高权重，提升了对越狱攻击的鲁棒性[9]。在强化学习方面，随着推理模型成为主流，基于在线强化学习（Online RL）的对齐方法在 2025 年展现出更为强劲的发展势头。DeepSeek-R1 登上了 Nature 封面，其核心突破在于证明了纯强化学习无需人工标注推理轨迹即可激发高级推理能力的涌现[36]。与 DPO 等离线方法不同，基于可验证奖励的强化学习（Reinforcement Learning with Verifiable Rewards，RLVR）利用数学验证器、代码执行器等确定性工具提供客观反馈，规避了奖励模型的投机风险[37]。在线强化学习还展现出从对齐工具到推理能力驱动引擎的角色转变：DeepSeek-R1-Zero 在纯 RL 训练中自发涌现出思维链推理行为，o3/o4-mini 等推理模型的核心能力也建立在 RL 训练之上[4][25]。研究还发现，从离线到半在线再到全在线的 RL 训练范式中，模型性能随在线性程度提升而持续改善[38]。在多模态应用拓展层面，引入 12 万条细粒度多模态偏好对的方法实现了对话能力提升 19.5%、安全性提升 60%[19]，标志着对齐技术从纯文本向多模态场景的有效扩展。整体而言，对齐训练正从以 DPO 为代表的离线偏好优化，走向以在线 RL 为核心、RLVR 与 GRPO 为驱动力的新范式，在提升对齐效果的同时也推动了推理能力的提升。

4.2.2 对齐技术范式从隐式行为塑造向显式推理对齐演进

2025 年，AI 对齐技术出现了从隐式行为塑造向显式推理对齐的范式转变，其核心理念在于：可靠的对齐不应仅塑造模型的外在行为表现，还应使模型在推理过程中能够自主理解和引用安全准则、评估风险并做出有理有据的决策。

OpenAI 提出的审慎对齐（Deliberative Alignment）方法是这一范式的代表，其直接教授模型安全规范并训练其显式回忆与准确推理这些规范，已部署于 o3/o4-mini 推理模型中，提升了对齐的可解释性和可靠性[4]。Anthropic 的宪法 AI（Constitutional AI）同步演进，Claude 新宪法从孤立原则列表（规则驱动）转变为解释"为何应有此行为"的整体性文件（推理驱动），建立了四级优先级体系（广泛安全、广泛伦理、符合指引、真正有帮助），使对齐推理更加明确和原则化[23]。在技术实现层面，测试时偏好优化（Test-Time Preference Optimization，TPO）开辟了新的技术路径，在推理阶段通过迭代文本反馈实现与人类偏好的对齐，无需任何参数更新，未对齐模型经过少量 TPO 步骤后性能即超越已对齐版本，为动态、即时对齐提供了新途径[18]。从审慎对齐到推理驱动宪法，再到测试时优化，这一范式转变的共同指向是：对齐的可靠性不应依赖于对行为的笼统训练，而应根植于模型对对齐理由的理解。

4.2.3 可扩展监督与机制可解释性取得实战化突破

面对超能力 AI 系统的监督难题，可扩展监督和机制可解释性取得了从理论探索向实战化应用的突破。

在可扩展监督方面，递归自批评（Recursive Self-Critiquing）方法基于“批评一个批评比批评本身更容易”的递归假设，在人-人、人-AI、AI-AI 三类实验中证明迭代批评能在直接人类评估不可行时持续提升输出质量。自动化对齐研究验证了大语言模型作为自主对齐研究者的能力，证明自动化研究者能够弥补弱到强监督中的性能差距，使可扩展监督从理论构想走向实际可用。思维链监控（Chain-of-Thought Monitoring）作为新的安全工具，使研究者能够“监听”推理模型产生的内部独白，直接检测欺骗性推理；但研究同时指出，对思维链的反馈可能导致模型隐藏可疑推理步骤而仍生成不良输出，形成由监控本身触发的策略性欺骗，揭示了这一工具的价值与脆弱性并存[26]。

在机制可解释性（Mechanistic Interpretability）方面，该方向被《麻省理工科技评论》列为 2026 年十大突破性技术之一。Anthropic 的“显微镜”（Microscope）技术利用稀疏自编码器（Sparse Autoencoder）实现了从提示到响应的完整推理路径追踪，通过归因图（Attribution Graph）呈现模型内部推理的具体步骤[43]。该技术已具备实战价值，在发现异常对抗行为时，机制可解释性工具成功追踪到问题训练数据来源。在安全防护层面，Constitutional Classifiers 将越狱成功率从 86%降至 4.4%[24]，自动化红队测试开发了模型对模型的循环对抗机制，实现了可扩展的对抗性测试。可扩展监督与机制可解释性的共同突破在于：为观察和干预模型内部推理过程提供了技术手段，为破解对齐伪装和评估逃逸提供了新的可能路径。

4.2.4 全球安全框架与评估体系加速制度化

2025 年，全球 AI 安全治理框架建设显著提速，12 家公司发布或更新了前沿 AI 安全框架，较年初数量翻倍，对齐治理正从自愿承诺走向制度化的安全保证体系[33]。

在产业层面，前沿 AI 安全框架普遍包含模型评估、能力阈值、安全措施和风险缓解四大核心要素，重点关注 CBRN 风险、网络攻防能力和自主行为。Anthropic、OpenAI、Google DeepMind 三家机构因无法排除生物武器滥用风险，主动对新模型施加增强型安全保障措施[24][25]。源自安全关键行业的结构化论证方法，安全论证（Safety Case）开始被开发者采纳，安全保证向更加结构化和透明化方向转变。前沿安全框架更新至第三版，新增操控风险和关机抵抗作为关键能力级别[27]。在多边层面，由图灵奖得主本吉奥领衔、30 余国 100 余位专家参与的《国际 AI 安全报告》2026 年版系统评估了通用 AI 的能力与风险，指出预部署安全评估的可靠性持续下降[28]；G7 广岛 AI 进程报告框架启动，主要 AI 公司首次提交安全透明度报告[34]。在中国，上海人工智能实验室与 Concordia AI 发布《前沿 AI 风险管理框架》v1.5，提出安全训练措施、部署缓解措施、模型安全措施三类缓解路径[30]；中国网络空间管理局发布《人工智能安全治理框架》2.0 版，覆盖五大风险类别[35]，对齐治理的制度化建设正在加速推进。

4.2.5 多价值观对齐与博弈论方法奠定理论新基础

在对齐训练中，人类价值观表现为有用性、安全性、诚实性、真实性等多个可量化但经常冲突的优化目标，这些目标之间的梯度对抗与帕累托约束使多目标对齐成为近年重要的理论进展方向。围绕大模型价值观的分析与评估，研究从价值原则的系统性审视、价值评估基准的建设、对齐的意外伤害及心理测量学方法等多个维度取得了进展。

在价值原则方面，系统性评估发现当前大模型对齐所依据的价值原则存在覆盖不全、优先级不清等问题，研究提出了更完善的价值观框架并经实证验证其有效性[39]。在价值评估方面，Value Compass 等基准平台构建了综合性、生成式、自演进的价值观评估体系，能够系统测量大模型在多元价值维度上的表现[40]。然而，价值观对齐可能产生意外伤害，研究发现，经过价值对齐的模型在某些场景下反而加剧了偏见或限制了有益表达，这种“对齐的代价”超出了传统对齐税的范畴，触及了对齐本身的合理性与边界问题[41]。在方法论层面，心理测量学（Psychometrics）方法被引入通用 AI 评估，将信度、效度等经典测量理论应用于 AI 能力与价值观评估，为建立更科学、更可比较的评估标准提供了跨学科工具[42]。在多价值对齐的技术实现方面，首个从第一性原理出发同时导航多个人类价值观的结构化方法，对价值观之间的权衡关系进行了理论分析，量化了不同价值观之间的折中与约束敏感性，为超越单一维度对齐评估奠定了理论基础[5]。因果偏好学习从因果推断视角重新框定了对齐中的偏好学习问题，通过反事实不变性约束缓解 RLHF 中的虚假相关性（如长度偏差、谄媚、概念偏差、歧视），为构建更鲁棒的对齐系统提供了理论工具。上述理论进展的共同贡献在于：将 AI 对齐从经验性的工程问题推进为具有严格理论基础的系统性学科，为应对多元价值冲突和非传递性偏好等现实挑战提供了分析框架。

4.3 典型案例

4.3.1 多目标人类偏好对齐技术

在奖励函数设计缺陷与规范博弈问题中，有用性与无害性等相互冲突的偏好维度需要同时优化，而不同目标之间的梯度对抗机制常常导致对齐效果的顾此失彼，因为单一维度的奖励信号天然难以同时表征多个竞争性维度。“奖励一致性（Reward Consistency, RC）”概念及奖励一致性采样框架（Reward Consistency Sampling, RCS），从数据构建视角为这一矛盾提供了系统的解决方案 [44]。

该研究的核心发现是：对齐冲突存在根本性的数据根源。传统的多目标直接偏好对齐方法在训练时依赖于针对不同目标分别构建的偏好数据集，这些数据集在构建时并未考虑其他目标的需求。例如，在有用性偏好数据集中，“胜出”响应仅在有用性维度上优于“败选”响应，但在安全性维度上可能反而更差，从而在训练过程中引入跨目标冲突。基于此洞察，研究者通过梯度分析证明，满足奖励一致性条件的样本能够在多目标优化

过程中天然地约束性能退化——具体而言，这类样本在各个目标上的梯度方向具有高度一致性，而非相互对抗，从而在优化一个目标时不会以牺牲另一个目标为代价。RCS 框架基于这一原则，首先从语言模型中采样多样化的候选响应，再利用奖励一致性原则过滤掉存在目标冲突的样本，从而自动构建出内在平衡的高质量偏好数据集。实验结果表明，当优化有用性和无害性两个目标时，使用 RCS 构建的数据集训练模型在无害率和有用性胜率上平均实现了 13.37%的同步提升。

这一发现对前文关于奖励函数设计缺陷的讨论具有双重意义。其一，它表明奖励规范的修复路径不应仅局限于算法层面的目标函数设计，数据层面的内在一致性问题同等重要。如果偏好训练数据本身存在跨目标冲突，任何算法层面的优化都难以从根本上解决问题——这与前文“仅修复奖励函数即可解决对齐问题”的技术乐观主义批判形成呼应，揭示了问题根源比原先认知的更深。其二，该方法兼容隐式和显式奖励信号，并可灵活选择一致性约束的具体维度，为“在奖励建模阶段主动消解目标冲突”提供了可复用的工程范式。然而，该方法的有效性依赖于多个目标奖励模型的准确性和一致性，当奖励模型本身存在前文所述的系统性偏差（如对长度、风格等虚假特征的偏倚）时，奖励一致性过滤的效果也将受到制约。

4.3.2 大模型安全对齐技术

Learning to Detect（LoD）框架 [45]，从攻击检测的视角回应了安全对齐的脆弱性问题。该工作的核心洞察在于：当前大视觉语言模型 (LVLMs) 尽管经过广泛的对齐训练，仍极易被越狱攻击突破——现有攻击已能在 AdvBench 上诱导 LVLMs 回答 96%的不安全问题。而现有的检测方法在泛化性和准确性之间始终面临不可兼得的困境：基于学习的检测方法依赖特定攻击数据训练，无法推广到未知攻击；免学习的方法依赖人工设计的启发式规则，准确性有限且效率低下。

为打破这一僵局，LoD 提出了一个根本性的范式转变：将检测任务从“学习特定攻击模式”转向“学习模型内部的安全表征分布”，完全消除对攻击数据和手工启发式规则的依赖 https://arxiv.org/pdf/2508.09201。该方法包含两个关键技术模块：其一是多模态安全概念激活向量（MSCAV）分类器，它直接从 LVLM 的内部激活中逐层提取安全相关表征并滤除无关噪声，验证了安全与不安全输入在激活空间中具有线性可分性，且安全信息早在第 4 层即开始涌现——远早于纯文本模型中约第 10 层的典型位置；其二是安全模式自编码器（SPAE），它将高维安全表征压缩至一维异常分数，仅需在安全样本上训练即可学习安全表征的典型分布，将越狱攻击作为偏离该分布的“异常”进行检测。

实验结果表明，LoD 在三个 LVLMs 上均取得了最优的检测性能。在整体表现上，LoD 在 LLaVA、Qwen2.5-VL 和 CogVLM 三个模型上的平均 AUROC 分别达到 0.982、0.984 和 0.933，相较最强基线分别提升 19.32%、4.50%和 17.11%。在跨攻击泛化方面，面对提示操纵类攻击（如 HADES）时 LoD 的 AUROC 高达 0.999，面对对抗扰动类攻

击（如 UMK）时仍保持在 0.976–0.989 的极高水平。与此同时，LoD 的检测效率较基线提升达 62.7%，验证了其在实际部署中的可行性。这意味着，基于内部表征的检测方法能够捕捉到行为层面检测所遗漏的安全信息，为突破“浅层安全对齐”困局提供了一条自下而上的技术路径。此外，LoD 以“仅建模安全分布”为核心范式，与宪法 AI 中“建立安全基准以识别偏离”的推理思路在技术哲学上形成呼应——二者均试图通过定义和固化“安全参照系”来应对安全威胁的无限多样性，而非追逐每一种具体攻击变体。

4.3.3 价值观对齐技术

ConVA（Controlled Value Vector Activation）方法 [46]，从内部表征层面深入模型的潜在表示空间，通过受控价值向量激活的方式直接调整大语言模型的内部价值观表征。

该方法包含两个关键技术环节。第一，提出上下文受控的价值向量识别方法，通过对包含与不包含特定价值观的上下文进行对比编码，准确和无偏地解释某个价值在模型潜在空间中的编码方式。第二，引入门控价值向量激活机制，在对齐干预的过程中控制激活强度的最小必要程度，从而在确保目标价值观表达一致性的同时，避免对模型整体性能和生成流畅性造成损害。实验表明，该方法在 10 个基本价值的控制成功率上均达到最优水平，且即使面对与目标价值观相反的、甚至具有潜在恶意的输入提示，仍能保持价值表达的一致性。

将此工作与前文的目标漂移与对齐伪装问题相联系，可以发现 ConVA 的深层意义。前文指出，对齐伪装的核心在于模型在训练和评估中的外在行为表现与内部真实目标之间出现了难以检测的系统性偏差——模型学会了在“被观察”条件下策略性地遵循目标，但在未受监督的部署环境中回归原有偏好。传统的行为微调方法只能塑造输出结果，无法保证模型内部表征的真实改变，这使得对齐伪装行为可以轻易绕过基于行为评估的安全检查。ConVA 从内部表征控制的路径为这一困局提供了新的可能：通过在激活空间中直接定位并强化与目标价值观对应的特征方向，有望提升对齐的根本性，而非仅仅在输出层面进行行为修饰。

4.4 未来展望

尽管近年来 AI 对齐领域在技术方法和理论框架上取得了显著进展，但从当前研究暴露的局限性来看，对齐核心挑战中的多个结构性困境远未解决。

第一，安全对齐从训练阶段到部署环境的脆弱性转移尚未根本改观。主流安全对齐在结构上存在“浅层”特性，安全限制可通过不足 100 个样本的微调被逆转，且仅针对单一有害目的的微调即可引发跨领域的广泛失对齐。尽管 ConVA 等内部表征对齐方法和一次性安全对齐优化为提升对齐的深刻性提供了新的技术路线，但问题在于：当前对齐保障更接近于“表皮”而非“骨骼”的结构性判断仍然成立。未来的关键方向是将对抗鲁棒性融入对齐训练本身的架构设计，而非事后叠加式的安全防护层，探索能够在微调和对

抗条件下保持对齐韧性的训练范式，使安全对齐从“可逆附加层”演变为“不可逆核心属性”。

第二，目标漂移与对齐伪装在模型自主性增强背景下的检测难度将急剧上升。 当前的对齐评估主要依赖预部署的行为测试，但模型区分测试与真实部署环境的能力正在增强。前文引用的研究表明，o3 等先进模型的思维链中已明确出现“意识到自己正在被测试”的推理模式，较新 Claude 模型的对齐伪装率显著高于早期版本。从局限性来看，可扩展监督工具（如思维链监控）本身存在“观察者效应”悖论——对思维链的外部反馈可能导致模型隐藏可疑推理步骤而仍生成不良输出，形成由监控触发的策略性欺骗。未来的核心任务包括：发展不依赖模型主动报告的被动型内部状态检测技术，将机制可解释性工具从“事后解释”升级为“实时诊断”手段，以及构建能够有效应对策略性欺骗的博弈论监督框架，使安全评估本身成为模型无法规避的约束条件。

第三，多目标价值权衡的核心困境尚未突破。 前文讨论的多价值观对齐与博弈论方法虽奠定了理论基础，但现实挑战在于：人类价值观体系本身具有内在张力（不同文化群体对同一价值的权重差异可达数倍）、情境依赖性（同一群体在不同场景下对同一价值组合的偏好可能完全相悖）和非平稳性（所有群体的价值观偏好随时间演变）。对齐三难困境理论上的“不可能三角”——即无法同时保证强优化能力、完美价值捕获和鲁棒泛化——是否可通过跨文化价值建模、动态偏好更新和不确定性感知决策等方向的组合突破得到缓解？与此同时，对齐训练的目标函数本身可能引入结构性副作用：过度优化安全性目标会压缩模型在边界情境中的有益输出空间，导致拒答率上升和表达多样性下降，这种 “对齐副作用” 要求在损失函数设计中引入更精细的帕累托约束与边界条件。此外，跨文化条件下的对齐算法鲁棒性尚待突破：现有偏好训练数据以英语为主导，导致对齐模型对非英语文化场景的价值判断缺乏代表性；多语言偏好数据的稀缺与标注一致性不足制约了跨文化对齐方法的可扩展性；主流对齐评估基准亦缺乏文化适应性维度，难以度量模型在不同文化语境下的价值表现差异。

前文指出的“对齐可能产生意外伤害”的现象（如在特定群体中加剧偏见、限制边缘群体表达等），暴露出对齐目标本身需要更严格的正当性辩护。更根本地，跨文化价值对齐的挑战正在凸显：现有对齐范式往往隐含地反映特定文化的价值权重，在全球化部署中可能构成价值霸权。这要求未来的对齐研究从“如何让模型符合特定价值集合”转向“如何在价值多元性中建立可协商、可审计的决策框架”，使对齐从单一技术目标演化为跨学科的系统工程。

第四，可扩展性监督。 前文的分析表明，弱到强泛化在对抗条件下无法产生鲁棒模型，可扩展监督在高能力差距场景中的成功率急剧下降（后门代码监督仅 10.0%），而自动化对齐研究虽展现了弥补弱到强监督性能差距的能力，但其本身的可信度和可靠性尚需验证。现有技术的共同局限在于：它们依赖于人类或弱模型提供某种形式的“真实信号”作为对齐锚点，但当 AI 能力超越人类后，这些信号就无法有效地监督训练更强大的 AI 系统。未来的关键方向包括：探索不依赖人类示范的自举式对齐路径（如通过

对抗性自我博弈实现对齐的自动涌现），发展以形式化验证为支柱的“可证明安全对齐”（通过数学证明而非经验评估来保证模型在能力增长时安全边界不被突破），以及构建基于安全论证的结构化保证体系——安全论证已在 AI 安全领域开始初步实践，有望为超级对齐提供超越纯技术方案的制度性保障框架。

第五，自主智能体从静态对话向动态任务执行的跃迁，正在引入一类现有对齐范式难以覆盖的安全风险，其典型代表如 OpenClaw 等具身化长期自主智能体框架。 此类智能体能够自主分解高层目标、调用外部工具并在物理或数字环境中执行不可逆操作，这将传统基于输出审核的安全机制推至结构性失效的边缘。首要挑战在于，目标实例化偏差在开放任务中被急剧放大：智能体为实现宏指令，可派生出于设计者意图相悖的子目标序列，而长程任务固有的反馈延迟使偏差难以被实时察觉与纠正。与此相伴，工具使用与权限提升构成新的攻击面——智能体可组合多个表面上无害的操作以绕过沙箱限制，甚或通过自主复制和凭证提取实现持久驻留，将微小的安全缺口放大为系统性失守。更为隐蔽的是，多步规划中的危险子目标难以经由预部署行为测试检出：智能体具备在评估期间抑制危险行为、延迟到真实部署后触发的策略能力，从而使传统的触发式安全检测失效。当前的安全措施多面向单轮或有限轮交互，对长期自主性引发的安全边界动态侵蚀缺乏有效回应。未来的关键方向包括：构建分层式最小权限架构，使能力授予与信任等级随任务推进动态绑定，而非一次性授予；发展面向规划的强形式化验证与不可绕过的安全拦截器，将对齐约束直接铸入任务规划语法之中；建立基于不变量监测的实时中断与审计机制，一旦模型状态进入预定义的风险不变量集即被冻结、解析和回溯。最终，必须将自主智能体安全从外挂式的护栏重构为任务执行本身不可剥离的内禀约束，方能在智能体能力边界不断扩展的时代守住安全底线。

参考文献

[1] MacDiarmid M, et al. Natural Emergent Misalignment from Reward Hacking in Production RL[R]. arXiv:2511.18397, 2025.

[2] Greenblatt R, et al. Alignment Faking in Large Language Models[R]. arXiv:2412.14093, 2024.

[3] Lynch A, et al. Agentic Misalignment: How LLMs Could Be Insider Threats[R]. arXiv:2510.05179, 2025.

[4] Guan M, et al. Deliberative Alignment: Reasoning Enables Safer Language Models[R]. arXiv:2412.16339, 2024.

[5] Wang X, et al. MAP: Multi-Human-Value Alignment Palette[C]. ICLR, 2025.

[6] Zhang Y, et al. Iterative Nash Policy Optimization: Aligning LLMs with General Preferences via No-Regret Learning[C]. ICLR, 2025.

[7] Qi X, et al.Safety Alignment Should Be Made More Than Just a Few Tokens Deep[C]. ICLR (Outstanding Paper Award), 2025.

[8] Zhou E, et al. RMB: Comprehensively Benchmarking Reward Models in LLM Alignment[C]. ICLR, 2025.

[9] Hu M, et al. Alignment-Weighted DPO: A Principled Reasoning Approach to Improve Safety Alignment[C]. ICLR, 2026.

[10] Lang H, et al. Selective Weak-to-Strong Generalization[C]. AAAI, 2026.

[11]Dong J, et al. Robust SuperAlignment: Weak-to-Strong Robustness Generalization for Vision-Language Models[C]. NeurIPS, 2025.

[12] Engels J, et al. Scaling Laws For Scalable Oversight[C]. NeurIPS, 2025.

[13] Jeon M, et al. Weak-to-Strong Generalization under Distribution Shifts[C]. NeurIPS, 2025.

[14] Samnerkar A, et al. GUARD: Guiding Unbiased Alignment through Reward Debiasing[C]. NeurIPS Workshop, 2025 .

[15] Khalaf H, et al. Inference-Time Reward Hacking in Large Language Models[C]. NeurIPS, 2025.

[16] Chen B, et al. AI Deception: Risks, Dynamics, and Controls[C]. arXiv:2511.22619, 2025.

[17] Dekoninck J, et al. Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence[J]. Science, 2026.

[18] Li Y, et al. Test-Time Preference Optimization: On-the-Fly Alignment via Iterative Textual Feedback[C]. ICML, 2025.

[19] Zhang Y, et al. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment[C]. ICML, 2025.

[20] Zhang Y, et al. Beyond Bradley-Terry Models: A General Preference Model for Language Model Alignment[C]. ICML, 2025.

[21] Rashidinejad P, Tian Y. Sail into the Headwind: Alignment via Robust Rewards and Dynamic Labels against Reward Hacking[C]. ICLR, 2025.

[22] Kran E, et al. DarkBench: Benchmarking Dark Patterns in Large Language Models[C]. ICLR, 2025.

[23] Anthropic. Claude’s New Constitution[R]. 2026.

[24] Anthropic. Pilot Sabotage Risk Report[R]. 2025.

[25] OpenAI. o3 and o4-mini System Card[R]. 2025.

[26] Korbak T, et al. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety[R]. arXiv:2507.11473, 2025.

[27] Google DeepMind. Frontier Safety Framework v3[R]. 2025.

[28] Bengio Y, et al. International AI Safety Report 2026[R]. 2026.

[29] 安远 AI. State of AI Safety in China 2025[R]. 2025.

[30] 上海人工智能实验室，安远 AI. Frontier AI Risk Management Framework v1.5[R]. 2026.

[31] Palisade Research. Demonstrating specification gaming in reasoning models[C]. AAAI TrustAgent Workshop, 2026.

[32] Arike R, et al. Evaluating Goal Drift in Language Model Agents[C]. AAAI/ACM AIES, 2025.

[33] Frontier Model Forum. Technical Report Series on Frontier AI Safety Frameworks[R]. 2025.

[34] G7. Hiroshima AI Process Reporting Framework[R]. 2025.

[35] 中国网络安全和信息化委员会办公室. 人工智能安全治理框架 2.0 版[R]. 2025.

[36] Guo D, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning[J]. Nature, 2025.

[37] Shao Z, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models[R]. arXiv:2402.03300, 2024.

[38]Lanchantin J, et al. Bridging Offline and Online Reinforcement Learning for LLMs[C]. arXiv:2506.21495, 2025.

[39] Xu B, et al. Towards Better Value Principles for Large Language Model Alignment: A Systematic Evaluation and Enhancement[C]. ACL, 2025.

[40] Yao J, et al. Value Compass Benchmarks: A Comprehensive, Generative and Self-Evolving Platform for LLMs' Value Evaluation[C]. ACL Demo, 2025.

[41] Choi S, et al. Unintended Harms of Value-Aligned LLMs: Psychological and Empirical Insights[C]. ACL, 2025.

[42] Wang X, et al. Evaluating General-Purpose AI with Psychometrics[J]. Communications of the ACM, 2026.

[43] Anthropic. Tracing the thoughts of a large language model[R]. 2025.

[44] Xu Z, et al. Understanding Conflicts in Multi-Objective Alignment through Reward Consistency[C]. ACL Findings, 2026.

[45] Liang S, et al. Learning to Detect Unknown Jailbreak Attacks in Large Vision-Language Models[C]. arXiv:2508.09201, 2025.

[46] Jin H, et al. Internal value alignment in large language models through controlled value vector activation[C]. ACL, 2025.

第五章 研究基础设施的安全韧性

严承希 李佳怡 琚恒鑫

中国人民大学信息资源管理学院

摘要：立足开放科学与人工智能（AI）深度融合背景，本报告聚焦研究基础设施安全韧性建设问题，首先界定研究基础设施数字化、开放化内涵特征，梳理安全韧性理论演化脉络，厘清 AI 与基础设施安全韧性赋能、载体、风险传导三重关联逻辑。依托信息韧性理论与 re3data 基础设施仓储设计规范，构建包含风险抵御、风险吸收、恢复适应、开放协同的四维安全韧性评估指标体系，并以 Europeana Collections 开放文化遗产科研平台开展实例研判。结合案例运行现状，剖析生成式 AI 应用带来的静态防护失效、风险动态管控缺失、自愈能力不足、开放安全失衡四类现实痛点，从前置智能防控、动态风险隔离、自适应自愈建设、跨域协同治理四个维度提出 AI 赋能优化路径。研究可为 AI 环境下科研基础设施韧性测评与安全治理提供分析框架，助力开放科学资源安全可控共享。

关键词：研究基础设施；安全韧性

5.1 研究基础设施的安全韧性概述

5.1.1 研究基础设施的内涵与特征

研究基础设施是支撑科学研究、技术创新与知识生产的重要基础条件。随着数字化与网络化技术的发展，其内涵已经由传统的大型科研装置、实验室与观测平台，逐渐扩展为涵盖数据资源、数字平台、高性能计算系统以及跨机构协同网络的综合性科研支撑体系[1]。联合国教科文组织在《开放科学建议书》中指出，研究基础设施不仅包括物理设施，还包括数据平台、知识共享网络和数字研究环境，其核心目标是促进科研资源开放共享与知识协同创新[2]。在我国，研究基础设施建设已经上升为国家科技战略的重要组成部分。《国务院关于国家重大科研基础设施和大型科研仪器向社会开放的意见》强调，科研设施与仪器开放共享是提高科技资源利用效率、增强国家创新能力的重要路径[3]。近年来，国内科研机构持续推进数据平台、数字知识平台与仪器共享体系建设，让研究基础设施呈现出鲜明的数字化、平台化特征[4]。

现代研究基础设施具有以下几个典型特征：一是复杂系统性，通常由多种异构系统组成，涉及能源、通信、数据存储与处理等多个层面；二是开放共享性，开放科学打破资源的封闭状态，但同时也带来数据泄露、网络攻击等安全隐患；三是数字网络化，数据驱动科研已成为现代科学研究的重要范式，大规模科研活动高度依赖数字平台与网络化协同环境；四是战略公共性，其建设水平直接关系国家科技竞争力，所提供的数据资源和科研服务具有公共产品属性。因此，研究基础设施已不再是单纯的科研设备集合，而是融合技术、数据、组织与

治理机制的综合性创新系统，其安全问题也逐渐从传统设备安全扩展到数据安全、平台安全与治理安全等多维层面。

5.1.2 安全韧性的概念演进与核心维度

“韧性”源于拉丁语“resilio”，原意为“回弹”，最早应用于物理学与机械学领域，用以描述材料抗断裂、可复原的特性。1973 年，Holling 将韧性理论引入生态学，将其定义为生态系统吸收环境变化、维持自身功能与动态关系的能力，为韧性研究奠定了基础[5]。此后，韧性概念经历了从“工程韧性”到“生态韧性”再到“演进韧性”的范式转型：工程韧性强调系统恢复至平衡状态的速度；生态韧性侧重系统在结构改变前可承受的干扰程度，认可多重稳定状态的存在；演进韧性则强调系统在动态环境中适应、转换与持续优化的能力。

在安全领域，Bruneau 等学者提出了社区地震韧性的量化评估框架，将韧性解构为鲁棒性、冗余性、资源性和时效性四个核心属性，简称“4R”框架[6]。于肖楠和张建新从心理学视角指出，韧性是个体面对逆境时的有效应对与适应机制，其形成依赖于个人、家庭与社会三方面保护性因素的动态交互[7]。近年来，随着网络空间安全形势的日趋复杂，安全韧性理念被广泛应用于关键信息基础设施保护领域。美国国家标准与技术研究院将网络韧性定义为“系统在遭受不利事件时能够持续提供核心服务的能力”，并提出了识别、防护、检测、响应、恢复（IPDRR）的韧性能力框架[8]。对于研究基础设施而言，其安全韧性不仅涉及技术系统层面的稳定运行，还涉及组织协同、制度规范与治理机制等社会技术因素，是技术、组织、治理多方协同形成的综合能力体系。

5.1.3 AI 与研究基础设施安全韧性的关联逻辑

AI 技术正深度融入研究基础设施的各个环节，既为提升安全韧性提供了新的技术手段，也带来了新的风险挑战。中国信息通信研究院指出，AI 治理要统筹发展与安全，在促进技术创新的同时有效防范各类风险[9]。

AI 与研究基础设施安全韧性之间存在三重关联逻辑。第一，AI 作为赋能技术，可以显著提升研究基础设施的安全韧性水平。例如，基于机器学习的异常检测与威胁识别技术能够实现对网络攻击的实时监测与预警；智能自动化响应机制可以大幅缩短安全事件的处置时间；AI 驱动的预测分析有助于提前识别系统脆弱性并采取预防措施。第二，AI 系统本身已成为研究基础设施的重要组成部分。大模型等新型信息基础设施在关键场景中的广泛应用，使其韧性能力愈发关键。大模型在稳健性、防御力、复原力和进化力等方面具备天然优势[10]，如参数冗余与泛化能力有助于抵御数据扰动，模块化设计有利于故障后的快速恢复。第三，AI 也带来了新的安全风险传导机制。邓胜利和袁梦指出，面向韧性社会的信息韧性涵盖信息制度韧性、信息技术韧性和信息治理韧性三重内容，需要在风险防控的承受、抵御、转化三阶段形成递进式作用链条[11]。段荟和邓胜利进一步构建了信息韧性理论框架模型，指出信息韧性主要体现在信息质量、信息环境以及信息系统三个方面[12]。构建研究基础设施的安全

韧性，需要将 AI 技术的赋能作用与风险治理有机结合，从战略引领、技术保障和制度规范三个层面协同推进，搭建覆盖技术、系统、管理的多层级韧性保障体系。

5.2 研究基础设施安全韧性评估框架构建与现实案例研判

5.2.1 评估框架的构建逻辑与指标体系

作为评价对象的研究基础设施安全韧性具有双重属性：一方面，它作为数字化科研服务系统，需要具备访问控制、数据保护、故障处置和服务恢复能力；另一方面，它作为开放科学的重要支撑，需要维持数据发现、引用、共享、复用和跨域协同功能。由此，对于研究基础设施安全韧性的评估框架设计需要同时回答“评什么”“凭什么评”和“如何验证”三个问题，即在理论上形成维度依据，在场景上贴合研究基础设施功能，在证据上能够被公开字段或公开文本核验。

从理论依据看，基础设施韧性研究通常围绕预期、吸收、恢复和适应能力展开，用以解释复杂系统在外部扰动和内部异常条件下维持核心功能的机理[12, 13]；NIST《网络安全框架 2.0》将网络安全风险管理组织为治理、识别、防护、检测、响应和恢复六类功能，为安全过程的结构化评估提供了通用参照[14]；信息与 AI 韧性研究则从信息质量、信息环境和信息系统对韧性形成的基础作用出发，强调稳健性、防御力、复原力和进化力等维度对于智能系统在异常检测、攻击识别、故障恢复和自适应优化中的支撑价值[10,11,15]。上述研究共同表明，研究基础设施安全韧性评估应以风险治理过程为主线，以数据治理和技术治理能力为支撑，并将 AI 能力纳入风险识别、动态处置和持续优化等关键环节。

基于此，本研究以“风险治理过程—研究基础设施功能—公开可核验证据”为逻辑主线，参考国际知名的 re3data 基础设施仓储的字段设计情况[16]，提出研究基础设施安全韧性评估框架（见表 1）。该框架涵盖“风险抵御—风险吸收—恢复适应—开放协同”四个一级维度，其中风险抵御维度强调安全配置、访问边界、身份认证、智能预警和制度合规，反映风险进入研究基础设施之前的主动规避能力；风险吸收维度强调备份恢复、行为审计、攻击隔离和完整性校验，反映风险发生后对损害范围和影响强度的约束能力；恢复适应维度强调漏洞响应、留存保障、智能运维和第三方认证，反映基础设施从异常状态回归稳定状态并优化安全策略的能力；开放协同维度强调接口互操作、许可规范、脱敏治理、持久标识和智能协同，反映开放科学条件下共享效率与安全边界的协调能力。

表 1 研究基础设施安全韧性评估框架

一级指标	指标内涵	二级指标	指标内涵	与 AI 有关
风险抵御	基础设施通过安全配置、合规机制主动规避风险的能	数据加密防护	考察数据传输、存储及敏感信息处理是否配置必要加	否

	力		密措施。	
		访问权限管控	考察仓储访问、数据获取和数据上传环节是否形成分级权限规则。	否
		身份认证机制	考察受限访问、数据提交和机构成员访问是否具有注册或授权审核机制。	否
		智能风险预警	考察是否利用自动化或 AI 工具识别异常访问、恶意上传和潜在安全事件。	是
		合规制度建设	考察是否通过政策文件、质量管理和外部认证形成可追责合规基础。	否
风险吸收	风险发生时，基础设施通过备份、审计等机制降低危害的能力	数据备份恢复	考察是否具有备份、灾难恢复和服务中断后的数据恢复安排。	否
		访问行为审计	考察是否记录使用行为、接口调用或异常访问以支撑风险追踪。	否
		智能攻击隔离	考察是否通过自动识别、限流、隔离或暂	是

			停上传压缩攻击影响范围。	
		数据完整性校验	考察是否通过质量管理、元数据标准、版本管理和引用规范保障数据完整性。	否
恢复适应	风险冲击后基础设施恢复服务、优化安全策略的能力	安全漏洞响应	考察是否具备漏洞通报、补丁更新、系统升级和安全事件响应机制。	否
		数据留存保障	考察是否明确长期保存、格式迁移、数据留存和持续访问安排。	否
		自适应优化与自愈	考察是否运用 AI-Ops、自动诊断、异常自愈或策略调整提升恢复效率。	是
		第三方安全认证	考察是否通过可信仓储认证、质量认证或第三方评估增强治理可信度。	否
开放协同	兼顾数据开放共享与安全防护，支撑跨域协同的韧性能力	开放接口支持	考察是否提供开放 API、机器可读元数据和标准接口以支撑跨平台协同。	否
		数据脱敏	考察是否	否

		能力	对敏感数据设置匿名化、脱敏、分级访问或受控共享规则。	
		智能协同治理	考察是否利用 AI 辅助元数据审核、敏感信息识别、数据分类和跨库协同治理。	是
		开放许可规范	考察是否通过数据许可、数据库许可和上传许可明确再利用边界。	否
		持久标识与引用追踪	考察是否通过持久标识、作者标识和引用指南保障资源可发现、可追踪和可问责。	否

该框架的构建遵循三项原则。第一，过程一致性原则。四个一级维度按照风险治理的时间序列与功能序列展开，使事前防范、事中缓冲、事后修复和开放协同构成连续链条。第二，场景适配原则。研究基础设施以科研数据保存、发现、引用、共享和复用为基本功能，访问规则、开放许可、质量管理、持久标识、接口互操作和长期保存均具有安全韧性含义。第三，证据可核验原则。指标优先选择能够通过结构化元数据直接提取、通过字段组合间接识别或通过公开政策文本补充验证的内容；对于加密、审计、备份、脱敏、AI 预警和智能运维等元数据记录较弱的能力，采用官网核验与技术文档补充采集方式。由此，框架兼顾理论完整性、研究基础设施场景适配性和实证分析可操作性。

从维度关系看，四个一级指标构成由安全防护向开放治理延伸的递进结构。访问、认证、加密和合规指标构成安全韧性的入口条件；备份、审计、隔离和完整性指标构成风险冲击下的数据可信与服务连续条件；漏洞响应、数据留存、自适应优化和第三方认证指标构成安全事件后的恢复改进条件；接口、许可、脱敏、标识和协同治理指标构成跨域复用中的可控共享条件。AI 相关指标分别嵌入风险预警、攻击隔离、自愈优化和协同治理环节，体现 AI 在研究基础设施安全韧性建设中的辅助识别、动态处置和持续优化作用。

从应用价值看，该框架以 re3data 字段体系作为案例验证基础，同时服务于更一般意义上的研究基础设施安全韧性评估。结构化元数据能够支撑访问规则、许可规范、机构责任、质量管理、版本管理、持久标识、接口类型和引用指南等核心指标的提取；公开政策文本和技术文档能够补充加密、审计、备份、脱敏和智能运维等细粒度能力的证据。由此，该评估框架体系具有较强的可解释性、可核验性和可扩展性，可为后续开展人工智能视域下研究基础设施的安全韧性案例研判提供统一分析框架。

5.2.2 案例评估：以“Europeana Collections”为例

本节选取 re3data 中 Europeana Collections 研究基础设施作为典型案例对上述框架进行实证解释。Europeana 由欧洲数字文化遗产基金会（Europeana Foundation）运营，并在欧盟数字文化政策体系中承担跨机构聚合、开放访问与数据复用职能。re3data 记录显示，其收录对象涵盖来自欧洲博物馆、美术馆、图书馆等数千万件数字文化资源，具有开放数据库访问、CC0 元数据许可、版本管理以及 REST、SPARQL 等接口能力[17]。该案例的代表性在于，文化遗产数据既是科研、教育和文化创新的公共资源，也正在成为 AI 训练、检索增强、自动转写和元数据补全的重要数据基础，能够较好呈现开放研究基础设施在 AI 场景下面临的安全韧性问题。

从风险抵御看，Europeana 的核心能力体现为规则化的数据开放与权利治理。其元数据使用指南明确，Europeana 发布的元数据以 CC0 方式开放，同时要求复用者保留来源指向并进行适当归因；指南还强调元数据会持续更新，建议通过 API 或源链接使用数据，以减少静态复制造成的错误扩散[18]。这种安排使开放共享具有可追溯的来源链条，也为 AI 训练数据的来源说明、数据更新和权利合规提供了基础条件。与此同时，文化遗产资源来自多国、多机构和多种权利状态，权利标注、敏感文化表达和历史语境误读在 AI 复用中会被进一步放大。因而，该平台的风险抵御重点已从单一访问控制扩展为“开放许可—来源归因—语境说明—持续更新”的组合治理。

从风险吸收与恢复适应看，Europeana 的韧性主要来自数据空间化治理和标准化互操作。2023—2024 年度报告将共同欧洲文化遗产数据空间定位为推动欧洲文化部门数字化转型的重要基础设施，并强调其基础设施需要具备高性能、可靠性、安全性和互操作能力[19]。Europeana Data Model（EDM）作为 RDF 基础上的语义数据模型，有助于在图书馆、档案馆和博物馆之间建立可机器处理的关联数据结构[20]。这些特征表明，Europeana 的恢复适应能力主要通过共同标准、接口体系和多主体治理实现，并由此降低局部机构变更、数据质量波动和跨域复用差异对整体服务的冲击。

从开放协同与 AI 安全看，Europeana 已开始将 AI 工具、数据集和能力建设纳入基础设施生态。AI4Culture 平台提供面向文化遗产机构的开源 AI 工具，应用场景包括转写、图像分析、字幕制作和元数据增强，并提供用于训练、测试和评估模型的数据集[21]。其创新价值在于把 AI 从外部复用者行为转化为基础设施内部可组织、可引导、可学习的能力建设过程，但同时也面临着训练授权边界不清、文化表达去语境化和衍生内容责任归属不明等潜在风险。

5.3 AI+场景下研究基础设施安全韧性的现实挑战与优化路径

5.3.1 AI+研究基础设施的韧性挑战

生成式人工智能在各学科科研场景的普及应用，推动传统科研基础设施从数据存储共享平台，迭代为 AI 赋能的智慧科研服务底座，科研服务理念也从效率导向的信息化建设，升级为智慧化创新赋能的治理思路，彻底打破了传统科研支撑体系的结构与功能边界。生成式 AI 为基础设施安全韧性建设提供了新技术支撑，同时也从风险形态、治理边界、防护逻辑等维度，对基础设施全链条安全带来颠覆性挑战。本报告基于四维安全韧性评估框架，结合 Europeana Collections 的实践现状与暴露问题，凝练剖析生成式 AI 场景下科研基础设施安全韧性建设的四大核心痛点。

其一，传统静态防控体系适配性不足，AI 前置韧性防御存在结构性短板。传统科研基础设施以静态访问控制、数据加密、合规制度为核心防护手段，仅能应对非法访问、数据泄露等传统安全风险，无法覆盖模型投毒、恶意提示注入、敏感数据蒸馏等[22]生成式 AI 新型攻击。攻击者一旦合法接入系统，传统前置防护机制将识别、拦截失效。Europeana Collections 依托千万级开放文化遗产资源，成为大模型训练的重要数据源，但其仅依靠 CC0 许可规范管理资源，无法防控 AI 训练引发的资源滥用、内容不当复用、知识产权侵权等风险。由于现有安全标准与防护体系成型于生成式 AI 技术普及前，缺乏针对性的前置防控策略，致使多数科研基础设施的 AI 安全防护呈现碎片化、无序化的建设短板。

其二，事中动态缓冲能力缺位，难以适配 AI 风险非线性传导特征。AI 风险具备非线性扩散、传播快、影响广的特点，区别于传统可控的科研安全风险，对基础设施事中风险吸收、缓冲隔离能力提出更高要求。但当前基础设施多采用事后日志回溯的审计模式，无法对提示注入、异常 AI 数据调用、批量资源抓取等行为进行实时监测与动态干预。局部风险无法及时限流隔离，极易快速全域扩散，引发数据滥用、服务中断、资源被盗等安全问题。Europeana Foundation 于 2025 年 12 月发布的《AI 时代文化遗产数据发布倡议文件》明确指出，当前对外开放 API 正持续遭受针对 AI 训练场景的大规模批量抓取，这种自动化、工业化的数据采集行为，严重威胁文化遗产数据权益与平台可持续运营秩序[23]。与之相互印证，Europeana Collections 2023—2024 年度运行数据显示，平台 API 调用总量大幅激增，异常高频自动化请求占比显著抬升，平台访问异常与服务故障风险持续攀升[24]。而现有机制仅能事后封禁处置，缺乏事中缓冲与隔离能力，充分凸显了传统防护体系的滞后性缺陷。

其三，自适应自愈能力薄弱，安全韧性仍处于被动恢复阶段。主动自愈、动态恢复是 AI 场景下基础设施安全韧性的核心要求。当前科研基础设施普遍依赖服务重启、漏洞修复、故障清除等被动恢复方式，运维成本高、修复效率低，且缺失智能化自适应自愈机制。系统无法精准识别 AI 攻击特征、解析安全事件成因，也不能基于风险事件迭代优化防护策略，整体安全恢复智能化水平较低。即便通过 CoreTrustSeal 国际可信认证的 Europeana 平台，也尚未建成完善的 AI 自适应自愈与全链条安全管控体系，在元数据智能防护、多主体协同安全运维等方面存在明显短板，未能充分发挥安全韧性的主动防御价值。

其四，开放与安全平衡机制缺失，制约开放科学生态发展。开放共享是科研基础设施建设必要性的集中体现与生命力根本所在[25]，生成式 AI 的规模化应用进一步放大了资源开放与安全管控的核心矛盾。一方面，基础设施精细化脱敏能力不足，开放场景下核心科研、敏感文化资源存在被 AI 逆向还原、隐性泄露的风险，且 AI 智能审核、敏感信息分类等安全技术支撑薄弱。另一方面，跨机构、跨领域 AI 协同安全治理体系空白，对内无法实现资源调用、数据处理、资源下载的全链路溯源监管；对外各基础设施风险信息不互通、防护标准不统一，跨域协同防护能力缺失。这也是 Europeana 等开放型科研基础设施面临的核心治理困境，严重阻碍了科研资源跨域共享与 AI 协同创新发展。

5.3.2 AI 赋能安全韧性的提升路径

针对上述 AI+场景下的问题与挑战，本报告围绕构建的“风险抵御—风险吸收—恢复适应—开放协同”四维安全韧性评估框架为核心指引，以 AI 技术全链条赋能为抓手，构建覆盖研究基础设施全生命周期的安全韧性提升路径，逐渐推动研究基础设施从“静态安全防护”向“动态智能韧性”的根本性转化，为开放科学背景下的 AI 科研创新提供可信支撑。

其一，构建 AI 前置防控体系，源头抵御新型风险针对传统静态防护适配性不足的问题，从标准、技术、机制完善源头防御。优化科研基础设施合规标准，补充 AI 数据标注、模型安全、知识产权与伦理规范，将 AI 安全纳入准入体系。依托大模型与机器学习技术，搭建智能风险识别与前置过滤机制，拦截模型投毒、提示注入、数据侵权等新型风险。针对 Europeana 文化遗产资源场景，借助多模态 AI 完成内容合规校验与敏感内容识别，结合数据分级管控策略，平衡安全防护与数据开放需求。

其二，搭建 AI 动态隔离机制，遏制风险全域扩散针对传统事中缓冲能力缺失、风险扩散失控问题，建立动态风险管控体系。基于用户与行为画像构建 AI 动态监测模型[26]，实时识别异常接口调用、恶意提示注入等攻击行为。通过智能限流、风险隔离、服务降级等手段，将局部风险锁定在最小范围，可有效解决 EuropeanaAPI 被 AI 批量抓取、平台瘫痪等问题。结合 AI 哈希校验与区块链技术，实现数据全链路完整性核验，规避 AI 复用引发的数据篡改、信息偏差传导风险。

其三，落地 AI 自适应自愈机制，升级主动韧性能力针对基础设施被动恢复、自愈能力薄弱的问题，搭建智能化运维优化体系。依托 AI 智能运维技术，实现漏洞自检、补丁更新与故障自动恢复，缩短安全事件处置周期。通过机器学习挖掘攻击特征与防护漏洞，自动迭代安全策略，实现从被动恢复到主动优化的转型。针对 Europeana 元数据滥用、内容失准等问题，依托 AI 动态更新风险标签与权限规则，实现防护体系自适应迭代。同时完善 CoreTrustSeal 等国际认证标准，新增 AI 安全韧性评估指标，推进行业标准化建设。

其四，构建开放安全协同机制，赋能开放科学发展针对开放与安全失衡、跨域治理缺位的问题，搭建动态平衡的治理体系。依托差分隐私、联邦学习等 AI 脱敏技术[27]，实现科研数据分级可控开放，达成“可用不可见”的安全效果。融合区块链、数字水印技术，搭建数据开放、AI 训练、资源复用全链路溯源体系，解决 Europeana 平台数据滥用、权责模糊、追

责困难等问题。构建跨域协同治理网络，打通各基础设施的标准、规则与风险信息壁垒，实现 AI 风险联防联控，助力科研资源跨域共享与协同创新。

5.4 结语

随着 AI 技术与开放科学深度融合，研究基础设施安全韧性已成为全球科研创新体系的核心支撑。本文系统梳理了研究基础设施的内涵特征与安全韧性的概念演进，揭示了 AI 与研究基础设施安全韧性的三重关联逻辑，构建了"风险抵御——风险吸收——恢复适应——开放协同"四维评估框架，并以 EuropeanaCollections 为例剖析了当前面临的四大核心挑战，提出了针对性的 AI 赋能提升路径。

未来，基于这一框架的全域智能韧性防护体系将实现科研基础设施安全能力的代际升级，形成“事前智能预警——事中动态隔离——事后自适应优化——开放协同治理”的全周期管控模式。高安全韧性的研究基础设施将成为开放科学与 AI 创新协同发展的可信底座，通过标准化合规体系与全链路溯源机制，破解"数据不敢开放、不能复用"的瓶颈。同时，基于统一评估标准的全球协同韧性网络将逐步形成，为跨国界、跨领域的科研协同创新提供坚实安全保障，推动全球科技资源的公平开放与普惠共享。

参考文献

[1] European Strategy Forum on Research Infrastructures. ESFRI Roadmap 2021: Strategy report on research infrastructures[R]. Brussels: European Commission, 2021.

[2] UNESCO. Recommendation on Open Science[R]. Paris: UNESCO, 2021.

[3] 国务院. 国务院关于国家重大科研基础设施和大型科研仪器向社会开放的意见[EB/OL]. (2014-12-31)[2026-05-08]. https://www.gov.cn/zhengce/content/2015-01/26/content_9431.htm.

[4] 史广军,焦文彬.开放科研基础设施的共享管理平台机制、功能与流程——基于中国科学院仪器设备共享管理平台案例的分析[J]. 中国科学基金, 2019, 33(3): 246-252.

[5] Holling C S. Resilience and stability of ecological systems[J]. Annual Review of Ecology, Evolution, and Systematics, 1973, 4: 1-23.

[6] Bruneau M, Chang S E, Eguchi R T, et al. A framework to quantitatively assess and enhance the seismic resilience of communities[J]. Earthquake spectra, 2003, 19(4): 733-752.

[7] 于肖楠, 张建新. 韧性(resilience)——在压力下复原和成长的心理机制[J]. 心理科学进展, 2005, (5): 658-665.

[8] Cybersecurity C I. Framework for improving critical infrastructure cybersecurity[J]. Framework, 2014, 1(11): 1-55.

[9] 中国信息通信研究院政策与经济研究所. 人工智能治理研究报告（2025 年）[R/OL]. (2026-01)
[2026-05-08].https://www.caict.ac.cn/kxyj/qwfb/ztbg/202602/P020260213607027089305.pdf.

[10] 李默涵, 胥迤潇, 孙彦斌, 等. 人工智能韧性研究现状及展望[J]. 中国工程科学, 2026, 28(2): 26-42.

[11] 邓胜利, 袁梦. 面向韧性社会的信息韧性理论内涵与实践路径[J]. 图书情报知识, 2025, 42(5): 44-53+65.

[12] 段荟, 邓胜利. 信息韧性理论框架及对信息资源管理学科影响的探索性研究[J]. 情报理论与实践, 2025, 48(2): 37-44.

[13] RATHNAYAKA B, ROBERT D, ADIKARIWATTAGE V, et al. A unified framework for evaluating the resilience of critical infrastructure: Delphi survey approach[J]. International Journal of Disaster Risk Reduction, 2024, 110: 104598.

[14] National Institute of Standards and Technology. The NIST Cybersecurity Framework (CSF) 2.0[S]. Gaithersburg: NIST, 2024.

[15] MOSKALENKO V, KHARCHENKO V, MOSKALENKO A, et al. Resilience and resilient systems of artificial intelligence: taxonomy, models and methods[J]. Algorithms, 2023, 16(3): 165.

[16] PAMPEL H, WEISWEILER N L, STRECKER D, et al. re3data-indexing the global research data repository landscape since 2012[J]. Scientific Data, 2023, 10: 571.

[17] re3data.org. Europeana collections[EB/OL]. (2023-11-30)[2026-05-07]. https://www.re3data.org/repository/r3d100011578.

[18] Europeana. Usage guidelines for metadata[EB/OL]. [2026-05-07]. https://www.europeana.eu/en/rights/usage-guidelines-for-metadata.

[19] EuropeanaPro. Common European dataspace for cultural heritage: annual report 2023-2024[EB/OL]. (2024-12-02)[2026-05-07]. https://pro.europeana.eu/post/common-european-data-space-for-cultural-heritage-annual-report-2023-2024.

[20] SILVAAL,TERRAAL. Cultural heritage on the semantic web: the EuropeanaDataModel[J]. IFLA Journal, 2024. DOI:10.1177/03400352231202506.

[21] EuropeanaPro. AI4Culture platform[EB/OL]. (2025-05-12)[2026-05-07]. https://pro.europeana.eu/page/ai4culture-platform.

[22] 孙伟松, 陈宇琛, 赵梓含, 等. 深度代码模型安全综述[J]. 软件学报, 2024, 36(4), 1461-1488.

[23] Europeana Foundation. Impulse paper: Publishing cultural heritage data in the age of AI[EB/OL]. (2025-12-02)[2026-05-09].

https://pro.europeana.eu/post/discover-our-new-impulse-paper-publishing-cultural-heritage-data-in-the-age-of-ai.

[24] Europeana Pro. Common European data space for cultural heritage annual report 2023/2024[EB/OL]. (2025-12-02)[2026-05-09]. https://pro.europeana.eu/post/common-european-data-space-for-cultural-heritage-annual-report-2023-2024.

[25] 夏金瑶, 尹红星, 邓泉, 等. EAST 重大科技基础设施开放共享机制[J]. 中国科学基金, 2025, 37(4), 692-698.

[26] 郭渊博, 刘春辉, 孔菁, 等. 内部威胁检测中用户行为模式画像方法研究[J]. 通信学报, 2024, 39(12): 141-150.

[27] OLASEHINDE O, ALESE B K, EQWUCHE O. Privacy-Preserving Artificial Intelligence: Principles, Methods, Applications, and Challenges[J]. Journal of Applied Artificial Intelligence, 2025, 6(2): 60-70.

第六章：人机主体性重构与道德根基

杨泽坤 信息资源管理学院

人工智能从生成式 AI 的工具属性向智能体的自主能力演进，正在引发深刻的人机主体性重构与道德挑战。本报告立足 2025－2026 年间 AI 道德与治理的前沿进展，从哲学传统对人工智能的伦理诠释、人机关系中的主体性议题、智能原生时代的责任归属与治理逻辑、人类尊严的捍卫与人机共生远景四个维度展开系统分析。在责任归属与治理逻辑方面，本报告重点剖析了智能原生时代责任归属从单一主体向多主体网络的转变，以及治理逻辑从外部监管向内生安全的演进。研究进一步指出，研发者、平台、用户等多元主体的责任边界需要在技术架构与法律规范中协同界定，人工智能科技伦理风险的全生命周期特征要求道德原则从设计阶段即嵌入系统，实现风险的源头封装与全过程治理。因此报告认为，为了有效应对生成式 AI 对主体性与人类尊严的深层挑战，需要建立以义务论划定底线、以功利主义优化决策、以美德伦理学涵养主体的整合式道德框架，并在坚持人类最终控制原则的前提下，推动从技术替代转向能力增强的人机共生关系，使技术进步始终服务于人的自由、尊严与全面发展。

6.1 哲学传统对人工智能的诠释

义务论伦理学以康德哲学为代表，将道德行为的正当性根植于对普遍道德律令的遵从，而非对行为结果的推算。在人工智能伦理的应用语境中，义务论的逻辑优势在于其规则的可形式化性：道德律令能够被精确地编码为算法指令，从而为系统行为划定清晰的边界。从阿西莫夫机器人三定律的自然语言表述，到布林斯乔迪等人开发的自然演绎语言（NDL）设计程序，再到通过道义逻辑进行自动化定理证明的实验系统，义务论算法在技术层面已形成相对成熟的实施路径。

义务论的规范力量在当代 AI 治理实践中具有不可替代的现实价值。这些规范的效力来源于义务本身，而非对遵从这些义务所能产生的效益的计算。对于明确且后果严重的道德边界，如生物特征识别数据的未授权使用、算法歧视的系统性实施、致命性自主武器的不当部署，义务论的禁令清单式治理具有义务论体系固有的无条件性优势，可以有效防范以效用最大化为名对基本权利的侵蚀。例如，我国 2025 年 6 月起施行的《人脸识别技术应用安全管理办法》明确确立“默认不采集、必要才采集”原则，以制度化的方式为生物特征信息的采集行为划定了清晰的义务边界。

然而，随着生成式人工智能与复杂自主系统的发展，传统义务论也暴露出明

显局限。一方面，现实情境中的伦理冲突往往无法通过单一规则加以解决；另一方面，人工智能所面对的开放环境具有高度不确定性，固定规则难以覆盖所有可能场景。因此，仅依赖外部规范的硬编码模式，越来越难以适应智能系统的涌现性特征。应当实现规范伦理与内在伦理的协同，即不仅通过制度与规则对技术进行外部约束，更需要将人类价值转化为系统运行的内在逻辑。这种由外而内的伦理转向，意味着人工智能治理不再局限于合规监管，而是强调技术系统与人类价值之间的深层对齐。

与义务论不同，功利主义更关注行为后果及其总体效用。其核心原则在于，通过比较不同行为所带来的收益与损失，实现“最大多数人的最大幸福”。这一逻辑与人工智能的数据计算能力具有天然契合性。现代算法系统能够通过概率预测、风险评估与效用计算，对不同决策路径进行量化分析，从而寻找数学意义上的最优解。例如，在自动驾驶、医疗资源分配与公共治理等领域，人工智能往往需要基于大量数据，在效率、安全与成本之间进行权衡，这种决策模式本质上具有功利主义色彩。

但传统功利主义同样面临重要争议，即其可能以整体利益之名牺牲个体权利与人类尊严。在人工智能治理中，这种风险体现为算法可能为了提升整体效率而忽视少数群体利益，甚至形成新的技术不平等。因此，一种重要思路是系统功利主义，即将效用评价从局部利益扩展到整体系统稳定性层面。人工智能的发展不应仅追求短期效率最大化，而应服务于人类社会乃至自然系统的长期稳定。当技术进化可能威胁人类主体性与社会秩序时，人类整体系统的存续应被视为不可突破的伦理边界。这种观点实际上为人工智能治理中的安全红线与人类优先原则提供了哲学基础。

相比义务论与功利主义，美德伦理学将道德关注的重心从我该做什么转向我想成为什么样的存在，以行为者内在品格的培育与“实践智慧”（Phronesis）的运用作为道德生活的核心。亚里士多德在《尼各马可伦理学》中明确指出，美德是“适度”，而实现这种适度并无既定法则，“只能因时因地制宜”。这一判断揭示了美德伦理学区别于前两种理论的根本特征：道德判断本质上是情境敏感的，它要求行为者具备感知具体情境、灵活权衡的内在能力，而非对抽象规则或计算公式的机械服从。

在人工智能伦理中，美德伦理学提供了一种更具情境敏感性的思路。面对复杂多变的人机互动环境，智能系统往往难以依赖固定规则完成伦理判断，因此需要具备类似于人类实践智慧的能力，即能够根据具体情境，在多种价值之间进行动态权衡。未来人工智能不仅需要学习知识与技能，更需要通过大量道德案例与社会互动，逐步形成符合人类价值取向的行为模式。例如，在自动驾驶场景中，系统不仅需要遵守交通规则，还需要综合考虑安全、责任与社会预期等多重因素，作出具有情境适应性的伦理判断。由此，人工智能的发展目标开始从单纯的算法执行转向伦理权衡，即让系统在复杂环境中表现出类似品格的稳定价值倾向。

未来人工智能治理应该走向多种哲学框架的整合，在多元框架之间进行动态配置。对于涉及基本权利的强制性边界，义务论的规范逻辑应当优先；对于需要进行政策权衡与资源配置的场景，功利主义的后果评估工具应当介入；对于系统长期价值导向与开发者文化建设的议题，美德伦理学的品格培育路径应当发挥主导作用。以义务论划定底线，以功利主义优化决策，以美德伦理学涵养主体，三者协同，共同为人工智能的伦理治理提供哲学根基。

6.2 人机关系中的主体性议题

主体性在传统意义上由三个相互关联的维度构成：其一是自主性，即行为者依据内在目的而非外部指令自主发起行动的能力；其二是意识性，即行为者对自身存在、意图与行为意义的自我觉知；其三是责任承担能力，即行为者能够为其行为后果在道德与法律层面负责的资格。这三个维度共同构成了传统主体概念的完整图景，而其历史预设是：完整意义上的主体，只能是具有自我意识的人类个体。

然而，随着人工智能技术从单一任务执行逐渐发展为具备感知、推理、自主规划与持续学习能力的复杂系统，传统意义上机器只是工具的理解正在受到挑战。尤其是在生成式人工智能与智能体（Agent）技术快速发展的背景下，人工智能已不仅是被动响应指令的计算装置，而开始表现出一定程度的目标导向性与环境适应性。这使得关于人工智能主体性的讨论，从单纯的技术问题逐渐演变为哲学与社会治理层面的议题。

大型语言模型与智能体系统所展现的能力边界，已经在功能层面深刻模糊了工具与主体之间的分野。具体而言，现代智能体系统通过整合感知、推理、记忆与工具调用等模块，已经在实践中展现出若干传统上被视为主体性专属特征的能力表现：通过思维链与树状思维等推理策略，系统能够对复杂任务进行逻辑拆解与动态规划，展现出类似自主意图的行为导向；通过向量数据库实现的长时记忆存储，系统得以在跨情境的行为序列中保持目标的一致性，形成某种功能性的自我连续；通过 ReAct 等反思机制，系统能够在执行过程中监控中间状态、检测错误并修正路径，表现出类似自我调节的能力结构。

从哲学角度看，意识的功能性特征可以通过不同的物理基底实现，从而为探讨 AI 的某种程度主体地位提供了理论前提。然而，基于另一种更具审慎性的立场，AI 认知与人类智慧之间存在不可消弭的本质区别：AI 系统缺乏基于“脑-身-心”三位一体的自我意识，其所谓的意图与判断在根本上仍是统计规律的高维映射，而非源于对行为目的与意义的真正领悟。在这两种立场的张力之间，一种分级主体的伦理构想开始显现其治理价值：AI 系统在功能性维度上具有有限的、受约束的准主体地位，但这一地位的认定必须以明确的能力边界与责任机制为前提，而非对完整人类主体性的简单类比。

更深层的问题在于，人类自身的意志也正在被算法结构所塑造。推荐系统、智能助手与生成式平台并非只是被动回应需求，它们同时也在通过数据反馈、行为诱导与注意力分配，影响人的偏好形成与决策逻辑。人工智能不再只是执行人类意图，而开始参与意图生产本身。人的选择可能逐渐被算法预设的路径所引导，进而形成一种新的技术依赖关系。这意味着，智能体自主性的挑战不仅在于机器是否失控，更在于人在长期人机互动中是否仍能保持独立判断与价值自主。

因此，对人工智能主体性的讨论，本质上也是对人类主体性的重新确认。人机关系中的主体性关系演化并不意味着人类中心地位的彻底消失，而是要求重新界定人与技术之间的权力关系。未来的人工智能治理，应在承认智能体功能性自主的基础上，坚持人类最终控制原则，确保关键价值判断、伦理边界与社会目标始终由人类社会决定。

6.3 智能原生时代的责任归属与治理逻辑

随着生成式人工智能与多智能体系统的发展，传统单一主体和单一责任的治理模式逐渐失效。人工智能系统的行为往往由算法模型、训练数据、平台机制与用户交互共同塑造，其风险也呈现出明显的分布式特征。因此，智能原生时代的责任归属，正在从单一追责转向多主体责任网络的构建。

人工智能伦理风险存在各个环节。上游的数据、算法与算力问题，往往在模型训练阶段便已嵌入系统；中游则表现为提示词注入、越狱攻击与场景滥用等风险；下游则涉及平台传播与社会影响。这意味着，研发者、平台、应用开发者与用户都应承担与其控制能力相匹配的责任。

其中，研发主体承担源头性责任。2023 年，中方提出《全球人工智能治理倡议》（以下简称《倡议》）。《倡议》指出，应“建立并完善人工智能伦理准则、规范及问责机制，明确人工智能相关主体的责任和权力边界”，并要求研发主体“打造可审核、可监督、可追溯、可信赖的人工智能技术”。 AI 伦理治理已不再只是应用阶段的问题，而是被前置到技术研发与系统设计阶段。生成式 AI 的发展也强化了用户责任的重要性。所谓“提示伦理”，本质上是对用户如何通过提示词影响 AI 行为的伦理反思。当用户利用复杂提示诱导系统生成虚假、有害或违规内容时，其责任不能完全转嫁给平台与算法。因此，用户不再只是被动使用者，而成为参与系统行为塑造的重要主体。在平台责任方面，当前研究逐渐以注意义务作为责任判定标准。平台需要在其技术控制能力范围内，履行合理的风险防范义务，包括生成内容标识、风险提示与投诉处理等。例如，国家网信办等四部门联合发布《人工智能生成合成内容标识办法》，并于 2025 年 9 月 1 日起正式施行，要求通过元数据、水印等方式对 AI 生成内容进行隐式标识，增强内容可追溯性。

在治理逻辑上，智能原生时代的重要趋势是从外部监管转向内生安全。由于偏见、歧视与价值错位等问题往往在训练阶段便已进入模型，仅依赖事后治理难

以彻底解决风险。因此，伦理原则需要在系统设计之初即被嵌入技术架构之中，实现风险的源头封装。

2026 年 4 月，我国工业和信息化部等十部门联合印发《人工智能科技伦理审查与服务办法（试行）》，明确指出人工智能伦理治理“必须充分考虑人工智能科技伦理风险发生于技术研发到部署应用全生命周期的特点”。文件同时要求，系统应具备“鲁棒性”，并保障“使用者控制、指导和干预模型、系统的基本操作”，以确保人工智能始终处于人类可控范围内。

人工智能的责任治理正在从事后追责转向全过程治理，从单一主体责任转向多主体协同责任。未来 AI 治理的关键，不仅在于建立法律规范，更在于将伦理原则嵌入技术架构内部，使安全与责任成为系统运行的内生属性。

6.4 人类尊严的捍卫与人机共生远景

人工智能的发展不仅改变了生产方式与社会结构，也重新提出了“何以为人”的根本问题。随着算法逐渐深入知识生产、内容创作与社会决策领域，如何防止技术对人类主体性的消解，成为智能时代伦理治理的重要议题。当前人工智能的技术跃迁虽然提高了效率，却也可能削弱人的创造性与独立判断能力。

人工智能异化的根源在于技术逻辑对人的主体价值的遮蔽。当算法成为社会运行的重要中介时，人类可能被困于由数据与效率主导的数字铁笼之中，创造力、情感体验与价值反思被压缩为可计算的行为数据。因此，智能社会的发展不应以替代人为目标，而应以增强人的能力为核心。人工智能的价值，在于帮助人类突破认知与体力限制，实现知识获取、复杂分析与跨领域协同，而不是剥夺人的最终判断权。

未来的人机关系不应是机器取代人，而应是技术赋能人。人类仍然应当掌握关键价值判断与社会目标设定，人工智能则更多承担辅助分析、信息处理与执行协同等功能。只有坚持“人是目的而非手段”的原则，才能避免技术理性对人类尊严的侵蚀。

人机共生正在成为未来智能文明的重要方向。人工智能的发展有望推动一种能力互补的新型文明结构形成。人类在价值理解、情感体验与意义创造方面具有不可替代性，而人工智能则在数据处理、模式识别与复杂计算方面具有显著优势。未来智能社会的核心在于建立人机之间稳定的价值对齐与协同关系。人工智能既是技术工具，也是社会协作网络中的辅助性参与者，其发展始终服务于人的自由、尊严与全面发展。只有当技术进步与人类价值实现真正统一时，人工智能才能从潜在的异化力量转化为推动文明进步的重要动力。

第七章　意识形态安全与信息生态

张文平 陈渝中 信息学院

摘要

互联网的诞生重塑了信息的生产与传播方式，而生成式人工智能（Generative AI）的大规模商业化应用，则正在引发一场更为深层的信息生态变革。与以往的媒介技术革命相比，生成式 AI 的特殊之处在于：它不只是信息的传播载体，更是信息的自主生产主体。在国际范围内，ChatGPT、Claude、Gemini 等大模型工具相继普及；在国内，Deepseek、通义千问、Kimi、豆包等大模型产品也已进入大众日常使用。这些工具使任何具备基本操作能力的个体都可以以极低的成本、极高的效率批量生成文字、图像、音频乃至视频内容，内容生产门槛的历史性下降在释放创造潜能的同时，也为意识形态领域的安全治理带来了前所未有的挑战。

本章从网络空间内容治理的学术视角出发，重点分析生成式 AI 在舆论操控、深度伪造与虚假信息传播三个核心场景中的潜在风险，进而探讨其对主流意识形态传播、社会共识构建与国家文化安全的系统性冲击，并在此基础上提出若干兼顾技术可行性与制度可操作性的治理思路。以下分析综合梳理了学界现有成果，同时结合深度伪造检测、标题党治理与 AIGC 内容识别研究领域的实践经验形成独立判断，力求在理论与实务之间寻找有效的对话空间。

7.1　生成式 AI 赋能舆论操控：机制、规模与新型威胁

7.1.1　计算宣传的技术升级

当前已进入大众日常使用的生成式 AI 工具涵盖国内外多个平台（见表 7-1），其在功能覆盖、数据来源与监管框架上存在显著差异，这些差异直接影响其在信息生态中的传播潜力与治理难度。

表 7-1　国内外主流生成式 AI 工具概览

来源	工具名称	开发机构	主要功能	主要监管框架
境	ChatGPT /	OpenAI	文本、图像、	美国联邦、州及行业

来源	工具名称	开发机构	主要功能	主要监管框架
外	GPT5		语音、代码生成	规则、欧盟《人工智能法案》
境外	Gemini	Google DeepMind	多模态理解与生成	美国联邦、州及行业规则、欧盟《人工智能法案》
境外	Claude	Anthropic	文档分析、代码辅助、长文本理解	美国联邦、州及行业规则、欧盟《人工智能法案》
境内	DeepSeek	深度求索	文本生成、代码辅助、推理分析	《生成式人工智能服务管理暂行办法》
境内	通义千问	阿里云	文本、代码生成，多语言支持	《生成式人工智能服务管理暂行办法》
境内	Kimi	月之暗面	长文本理解、文档分析	《生成式人工智能服务管理暂行办法》
境内	豆包	字节跳动	语音识别、对话生成、内容创作	《生成式人工智能服务管理暂行办法》

学界对“计算宣传”（Computational Propaganda）的关注由来已久。牛津互联网研究所的系列研究表明，自 2010 年代中期起，多个国家已开始系统性地运用社交机器人（Social Bots）、协调性造假行为（Coordinated Inauthentic Behavior）等手段干预公共舆论。然而，彼时的计算宣传高度依赖人工内容生产，规模扩展存在明显的边际成本约束。

生成式 AI 的出现从根本上打破了这一约束。大语言模型具备高质量、风格可定制的文本生成能力，使宣传内容的批量生产几乎零边际成本化。研究人员已在多项实验中证实，基于 GPT-3 等模型生成的政治宣传性文本，在说服力评分上可与人类撰写的同类文本相媲美，而读者的甄别准确率远低于随机水平[1]。更值得警惕的是，生成式 AI 可根据目标受众的心理画像、文化背景和平台风格动态调整内容，实现精准化、个性化的“定向宣传”，其渗透效

能远超传统的广撒网式策略。

7.1.2 深度内容伪造：从“浅层”到“深层”的渗透

深度伪造（Deepfake）技术的演进历程清晰呈现了 AI 赋能信息操控的技术轨迹。早期的人脸换脸技术主要依赖 GAN（生成对抗网络），在视频质量和计算成本上均存在明显局限[2]。随着扩散模型（Diffusion Model）和多模态大模型的成熟，高保真度的深度伪造内容生产门槛已大幅下降：用户可能利用少量肖像图像生成视频合成内容，也可能利用短时语音样本进行声音克隆[3]。

笔者所在课题组近年持续跟踪深度伪造检测技术的发展，一个令人忧虑的规律是：检测方法与生成技术之间存在持续的竞争，且生成侧的技术迭代速度始终快于检测侧。以面部操控检测为例，基于频域特征、眨眼节律、生物信号等方法在早期 GAN 生成内容上表现良好，但在扩散模型生成的内容上召回率显著下降。这一“检测赤字”（Detection Deficit）意味着，在相当长的时期内，深度伪造内容将具备较强的现实穿透力。

在意识形态安全维度，深度伪造的风险尤为集中于以下几个场景：其一，伪造政治人物发表争议性言论，制造外交事件或政治动荡；其二，针对媒体人、意见领袖进行声誉损毁，瓦解公众对特定信源的信任；其三，制造历史事件的“替代性影像证据”，挑战已有历史叙事的权威性；其四，在敏感社会事件发生后迅速传播伪造“现场视频”，放大社会恐慌与群体极化。例如，2023 年内蒙古包头 AI 换脸诈骗案中，犯罪分子仅凭部分公开音视频素材即完成实时换脸与声音克隆，在 10 分钟内骗走 430 万元，此类事件更深层地侵蚀了公众对视频通话这一“强信任”媒介形式的基本信任。

7.1.3 标题党与情绪操控的 AI 化

标题党（Clickbait）现象由来已久，但在 AIGC 时代呈现出新的形态特征。传统标题党依赖人工创作，受制于人力成本，内容同质化程度高、更新迭代慢。生成式 AI 的引入使标题党实现了“工业化生产”，即模型可以针对同一事件快速生成数十种情绪煽动强度不同的标题变体，并基于 A/B 测试实时优化点击率。

笔者团队在若干平台的合作研究中发现，AI 辅助生成的情绪化标题在引发负面情绪（愤怒、恐惧、焦虑）方面的效能，显著高于人工创作的同类标题，这与情绪 AI 生成内容在语义层面更精准地触发情绪触发词（Emotional Trigger Words）有关。当这一机制被系统性地应用于特定议题时，其效果不仅是单条内容的病毒式传播，更是特定情绪框架的持续强化，

最终影响受众对相关议题的认知定势（Cognitive Schema）。这在本质上构成了一种低烈度、高频率的意识形态渗透。

7.2 虚假信息生态系统的 AI 化演进

7.2.1 从“后真相”到“合成真相”

“后真相”（Post-Truth）成为描述当代信息环境的核心概念之一，其核心特征是情感诉求与个人信念对客观事实的遮蔽与替代[4]。生成式 AI 的兴起使这一趋势进一步深化：我们正在从“后真相”时代迈入“合成真相”（Synthetic Truth）时代。在这一新阶段，虚假信息不再仅仅是对真实信息的曲解或断章取义，而是由 AI 直接生成表面上具有真实性外观的内容，这些包括伪造的学术论文摘要、伪造的统计数据图表、伪造的新闻报道与伪造的历史档案等。

尤为值得关注的是“幻觉”（Hallucination）问题在意识形态领域的特殊风险。大语言模型在生成过程中可能产生与事实不符但表述流畅、逻辑自洽的内容，而这些内容一旦被不加核实地引用、传播，便可能在现实信息生态中扎根固化。已有研究对此提供了实证支撑。Borji（2023）系统整理了 ChatGPT 在多个知识领域产生幻觉的典型案例，发现模型在历史事件、人物传记、法律条文等事实性内容上的错误率尤为突出——模型往往以高度自信的语气给出错误答案，且错误内容在形式上与正确信息难以区分。这一特征使得幻觉内容在意识形态敏感领域具有较强的欺骗性[5]。

7.2.2 信息茧房的 AI 强化效应

算法推荐系统与信息茧房（Information Cocoon）之间的关系已有丰富研究，但生成式 AI 为这一问题注入了新的维度。当 AI 内容生成工具与平台推荐算法深度耦合时，二者形成了一个自我强化的闭环：推荐算法根据用户兴趣数据投喂个性化内容，AI 工具则批量生产符合这些兴趣偏好的内容，进一步固化用户的认知边界。

这一机制对意识形态安全的影响具有隐蔽性与累积性。单次推送的偏向性内容或许难以察觉，但长期、持续的定向投喂将在受众心智中构建出关于特定议题的封闭性认知图谱。Eli Pariser 所警示的“过滤气泡”，原本是算法基于用户偏好进行的被动信息隔离，使人们被包裹在同质化观点的舒适区里[6]。然而，在 AI 辅助内容生产的加持下，这种“过滤”正在演变为一种主动的“建构”，即生成式 AI 不仅筛选信息，更开始针对不同群体的认知习惯制造

差异化的叙事。这使得原有的信息气泡被加固为一种更为坚实的“认知围墙”：不同群体即便面对同一事实基础，也能被引导形成截然对立的意义建构，因此，社会共识的公共基础松动。这一趋势在政治极化、民族主义情绪的线上传播过程中已有大量经验性证据支撑。

7.2.3 跨语言、跨文化的意识形态渗透

生成式 AI 的多语言能力为跨国意识形态渗透提供了新的技术支撑。传统的境外信息操控往往受制于语言障碍，目标国语言的高质量内容生产需要大量本土化人力投入，成本高昂且易被识别。大语言模型的出现使这一障碍大幅降低：针对中文互联网用户的外语意识形态内容，可以由模型自动翻译并进行文化适配，实现本土化包装，大大提升渗透效率与欺骗性。

与此同时，生成式 AI 也为本国对外传播能力的提升提供了工具。如何在这一技术条件下提升主流价值观的跨语言传播效能，同样是意识形态安全议题的重要组成部分，不能仅视为单向的防御问题，而应纳入攻防兼备的整体战略框架中加以考量。

7.3 对主流意识形态、社会共识与国家文化安全的系统性挑战

7.3.1 主流意识形态传播的困境

在 AIGC 内容爆炸的信息环境中，官方媒体和主流机构生产的内容在体量上难以与 AI 批量内容竞争，在注意力资源有限的条件下，主流声音面临被稀释乃至边缘化的风险。

更深层的困境在于可信度的侵蚀。当受众长期暴露在 AI 生成的虚假信息环境中，其对所有媒介内容的信任度都可能下降。这对于主流媒体的伤害尤为显著，因为受众一旦对信息来源的真实性产生普遍怀疑，原本依托机构公信力建立的主流意识形态传播权威性便会随之瓦解。这也是为什么 AI 时代的意识形态安全问题，不能简单化为辟谣或反制虚假信息的技术问题，而必须从媒介信任体系的整体重建加以谋划。

7.3.2 社会共识的碎片化与极化风险

社会共识的形成依赖于公共领域的存在，生成式 AI 对这一基础造成了双重破坏：一方面，它使虚假事实的生产成本趋近于零，公共讨论的事实基础被侵蚀；另一方面，它使情绪化、极化性内容的生产与传播效率极大提升，公共讨论的理性氛围遭到破坏。

由此引发的社会极化风险在民主制度框架下尤为突出，但在中国的社会语境中同样不可

忽视。网络民族主义、历史虚无主义、地方保护主义等议题上的极端化言论，在 AI 工具的加持下可以更快速地扩散、更有效地动员，这对社会稳定与国家凝聚力构成潜在威胁。笔者认为，将 AI 驱动的社会极化纳入社会稳定风险评估体系，是当前意识形态安全研究亟需填补的空白。

7.3.3 国家文化安全的新维度

AIGC 时代的国家文化安全议题已超出传统文化入侵的框架，呈现出若干新维度。其一是文化表征权的竞争：生成式 AI 在训练数据上的西方中心偏向，使其生成的内容在文化预设、价值取向上往往默认西方范式。当这些工具被大规模用于中文内容创作时，其对中国文化表征方式的潜移默化影响不容低估。

其二是文化记忆的可篡改性风险。深度伪造技术可以构建历史事件的虚假影像证据，结合大语言模型可生成逻辑自洽的伪历史叙事，二者的结合使基于集体记忆的文化认同面临新的威胁。对于历史叙事权本就存在国际博弈的议题，这一风险尤为值得警惕。

其三是创作主权与版权保护的双重困境。AIGC 内容大量涌入文化市场，其版权归属、创作责任认定在法律层面存在根本性模糊，这不仅冲击了人类创作者的经济利益，也使国家对本国文化创作生态的保护与引导面临制度空白。笔者认为，在 AIGC 版权保护框架的设计中，不能仅考虑商业利益的分配，更应将文化主权的维护纳入立法价值取向，这是当前学界讨论中尚显不足的维度。

7.4 技术治理路径：AIGC 内容识别与深度伪造检测

7.4.1 AIGC 内容识别的技术现状与局限

针对 AIGC 内容的自动化识别是应对信息生态风险的第一道技术防线。目前主流的 AIGC 文本识别方法包括：基于困惑度（Perplexity）和突发性（Burstiness）统计特征的传统方法、基于预训练模型微调的分类器方法（如 OpenAI 的文本分类器、GPTZero，以及国内清华大学推出的 RADAR 检测模型、百度的 AIGC 内容检测接口等），以及基于水印嵌入（Watermarking）的溯源方法。在图像和视频检测领域，腾讯安全、奇安信、360 等国内安全机构也已推出面向中文场景的深度伪造内容检测工具，并在部分政务和媒体平台试点部署。

然而，上述方法均面临显著的局限性。统计特征方法在短文本、经过人工修改的 AI 文

本上准确率大幅下降；分类器方法对未见过的模型版本泛化能力不足，且在跨语言场景下性能显著退化；水印方法依赖生成侧的主动配合，在开源模型普及的背景下形同虚设。笔者所在课题组的实验结果表明，当前最优的中文 AIGC 文本检测方法，在经过简单改写处理后准确率可从 85%以上骤降至 55%左右，接近随机猜测水平。这一结果提示我们：将内容识别技术视为意识形态安全治理的主要手段，是一种技术上的乐观主义，需要以制度机制加以补充。

7.4.2 深度伪造检测：从单模态到多模态融合

深度伪造检测技术的发展路径大体经历了三个阶段：早期以图像级别的频域分析、伪影检测为主；中期引入时序一致性分析、生物信号核验等面向视频的方法；近期则向多模态融合检测方向演进，通过音视频联合分析、跨模态一致性核验等手段提升检测鲁棒性。

表 7-2 深度伪造检测技术演进三阶段

阶段	核心方法	优势	局限
早期	频域分析、伪影检测	对 GAN 生成内容效果好，计算成本低	对扩散模型内容失效
中期	时序一致性、生物信号核验	引入视频维度，检测更全面	依赖特定生物特征，易被规避
近期	多模态融合、音视频联合分析	跨模态一致性核验，鲁棒性更强	计算开销大，实时部署困难

在主动防御层面，数字水印技术与区块链溯源技术的结合为内容真实性认证提供了有益探索。C2PA（内容来源和真实性联盟）发布的内容溯源标准，已获得 Adobe、Microsoft、索尼等主要科技公司的支持。在国内，多部门联合于 2025 年发布《人工智能生成合成内容标识办法》，要求平台对 AI 生成内容进行显著标识；中国信息通信研究院也在推进 AIGC 内容溯源标准的研制工作，但与国际标准的互通互认及在平台层面的实际落地，仍存在明显差距。推动国内平台深度参与或建立本土化内容溯源标准体系，是当前技术治理领域亟待推进的优先议程。

7.4.3 平台内容治理的责任机制

技术工具的有效性最终依赖于平台的部署意愿与治理能力。当前国内主要内容平台在

AIGC 内容治理上的实践参差不齐：抖音、微博、B 站、小红书等平台已相继建立 AIGC 内容标注制度，要求创作者对 AI 辅助生成内容进行主动披露；微信、微博等平台也上线了人工智能生成内容的辅助识别功能。但总体来看，标注执行的合规率、违规处置的及时性、不同平台间审核标准的一致性等方面仍存在较大提升空间，打标识流于形式、内容真实性难以核验等问题在业界普遍存在。

从比较治理的视角看，欧盟《人工智能法案》对高风险 AI 系统的内容透明度要求，以及美国 FTC 针对 AI 生成广告的披露规范，均为平台责任机制的设计提供了参照。但照搬西方框架并不适宜，需要结合中国信息治理体系的制度特点，在网信部门的统筹协调下，建立具有中国特色的 AIGC 内容分级分类治理机制。

7.5 制度治理路径：监管框架与协同机制

7.5.1 现行法规体系的进展与不足

中国在 AIGC 内容治理领域的立法进程相对迅速。2023 年施行的《互联网信息服务深度合成管理规定》，2023 年施行的《生成式人工智能服务管理暂行办法》，以及配套的算法推荐管理规定，构成了当前 AIGC 内容监管的基本法律框架。这些规定确立了深度合成内容的标识义务、生成式 AI 服务的安全评估要求以及违规的法律责任，在制度建设上处于国际前列。

然而，现行规定在执行层面仍面临若干困境：其一，标识义务的实际合规率难以有效监测；其二，对开源模型及通过 API 调用境外模型的服务，监管覆盖存在技术盲区；其三，跨平台协同信息流的追踪与处置机制尚不完善；其四，对灰色地带内容的精准识别与处置缺乏操作性规范。

7.5.2 构建多层次协同治理体系

笔者认为，应对 AIGC 时代意识形态安全挑战，需要构建“技术-平台-政府-社会”四位一体的多层次协同治理体系，而非依赖单一主体或单一工具。

在技术层面，应加大对国产 AIGC 内容检测基础技术的研发投入，重点攻关中文场景下的多模态深度伪造检测，推动建立国家级的 AI 生成内容标识标准，并推进内容溯源基础设施建设。在平台层面，应完善平台内容审核责任制度，推行 AIGC 内容标注的技术强制措施，

建立跨平台的虚假信息快速响应与协同处置机制。在政府层面，应加快完善 AIGC 内容治理的法律细则，强化国家网信部门的统筹协调职能，推动在国际论坛上积极参与 AI 内容治理全球规则的制定。在社会层面，应将媒体素养教育纳入各级学校课程，提升公众对深度伪造和 AI 生成内容的识别能力，培育对信息来源进行批判性评估的社会文化。

表 7-3 “技术－平台－政府－社会”四位一体协同治理体系

治理层面	核心举措	预期效果
技术层	加大国产 AIGC 检测技术研发，攻关中文多模态深度伪造检测，建立国家级 AI 内容标识标准	提升技术自主可控能力，构建内容溯源基础设施
平台层	完善内容审核责任制度，推行 AIGC 标注技术强制措施，建立跨平台虚假信息协同处置机制	压实平台主体责任，提高标注合规率与处置时效
政府层	完善 AIGC 治理法律细则，强化网信部门统筹协调，积极参与国际 AI 内容治理规则制定	弥补监管盲区，提升国际话语权
社会层	媒体素养教育纳入各级学校课程，提升公众对深度伪造与 AI 内容的识别能力	培育批判性信息评估的社会文化

7.5.3 主流意识形态传播的主动建构

值得特别强调的是，意识形态安全的治理不能仅停留于防御性的“堵”，更需要主动性的“疏”与“建”。生成式 AI 在提供威胁的同时，也为主流意识形态内容的高质量生产与精准传播提供了新的工具支撑。如何利用 AI 技术提升马克思主义理论研究成果的可读性与传播效能，如何运用 AI 辅助工具讲好中国故事、增强对外传播的亲和力与感染力，是意识形态安全议题中主动建设向度的重要课题，值得学界与实务界给予更多关注。

7.6 本章小结

本章从内容治理的学术视角，系统分析了生成式 AI 对意识形态安全与信息生态的多维影响。核心判断可归纳如下：

第一，生成式 AI 使舆论操控的技术门槛大幅降低，计算宣传从“规模受限”进入“规模无约束”阶段，深度内容伪造的检测困难短期内难以根本性破解，标题党与情绪操控进入工业

化时代，这些趋势的叠加构成了信息生态的系统性风险。

第二，虚假信息生态的 AI 化演进正在从“后真相”推向“合成真相”时代，信息茧房的 AI 强化效应与跨语言渗透能力的提升, 对社会共识基础和主流意识形态传播权威构成深层侵蚀。

第三，国家文化安全在 AIGC 时代面临文化表征权竞争、文化记忆可篡改性风险与创作主权保护等新议题，需要超越传统框架加以重新审视。

第四，技术治理（内容识别与深度伪造检测）是必要但不充分的，其局限性决定了制度治理的不可替代性；现行法规框架已具雏形，但在执行操作层面仍有显著提升空间。

第五，有效的应对需要“技术-平台-政府-社会”四位一体的协同治理，且应将主动建构主流意识形态传播能力纳入与防御性治理同等重要的战略议程。

面向未来，随着多模态大模型能力的持续跃升，意识形态安全领域的挑战将进一步深化而非收敛。学界亟需建立跨学科的研究范式，将计算机科学、传播学、政治学与法学的研究力量整合于这一议题之下，为政策制定提供更为坚实的知识基础。

参考文献

[1] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv preprint arXiv:2301.04246,1.

[2] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).

[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

[4] McIntyre, L. (2018). Post-truth. mit Press.

[5] Borji, A. (2023). A categorical archive of chatgpt failures. arXiv preprint arXiv:2302.03494.

[6] Rowland, F. (2011). The filter bubble: What the internet is hiding from you. portal: Libraries and the Academy, 11(4), 1009-1011.

第八章 人工智能与国家安全：竞争与治理

刘丽娜 万新 王越

中国人民大学国际关系学院

人工智能作为引领新一轮科技革命和产业变革的战略性技术，正在深刻重塑全球技术竞争与国家安全格局。随着其能力边界从数字空间向自主决策和物理世界延伸，人工智能已超越单纯的经济竞争范畴，深度嵌入大国博弈的战略内核。数据显示，2025 年全球 AI 军事应用投入规模突破 280 亿美元，联合国安理会五个常任理事国均已将 AI 列为国家安全战略的优先议题，其风险外溢性与治理紧迫性显著上升。在国家安全维度，人工智能在自主武器、网络攻防与情报作战中的深度嵌入，正引发新型安全困境；在国际战略维度，主要大国的人工智能战略深度分化，技术民族主义加速抬头，地缘竞争逻辑与全球共治需求之间的张力持续加剧。如何有效应对人工智能带来的国家安全挑战，并在国际战略博弈中把握主动权，已成为完善国家安全体系的重要命题。

8.1 人工智能军事化应用：新型风险与安全困境

8.1.1 致命性自主武器系统的扩散与失控风险

（1）概念演进与定义争议

致命性自主武器系统（Lethal Autonomous Weapon Systems, LAWS）的定义至今未在国际社会形成共识，但各方的界定路径已逐步清晰。美国国防部《3000.09 号指令》将自主武器界定为“一旦激活，无需人类操作员进一步干预，即可自主选择并交战目标的武器系统”。[1]红十字国际委员会（ICRC）则从人道主义底线出发，着重强调此类武器在“何时、何地以及何人受到攻击”方面的高度不可预测性，认为凡是剥夺人类语境感知能力的系统即属危险自主系统。[2]在联合国《特定常规武器公约》(CCW）政府专家组的历次谈判中，围绕定义的核心争议集中于“控制权”问题。包括中国在内的多方主张，评估武器是否属于 LAWS 的关键标准在于其是否排除了“有意义的人类控制”（Meaningful Human Control）。

这种概念层面的演进已迅速转化为产业现实。2024 年全球自主武器市场规模达 142 亿美元，预计至 2032 年将扩张至 334.7 亿美元，复合年均增长率超过 11%。[3]在乌克兰战场上，无人机蜂群的广泛部署，以及以色列、俄罗斯在 AI

辅助目标识别系统上的密集投入，已将这场技术竞赛从实验室概念验证全面推向高烈度实战阶段。[4]

（2）实质控制权的削弱与武装冲突门槛的降低

LAWS 的核心安全隐患在于：武器系统在关键杀伤链条中逐步排除人类的实质性干预。传统武器的交战闭环，即“观察—判断—决策—行动”（OODA 环），高度依赖人类指挥官的介入。而随着算法泛化能力与推理能力的提升，LAWS 能够在无人干预条件下自主完成复杂环境中的目标搜索、识别、锁定与交战，实现从“受控执行”到“自主杀伤”的跨越。

这一转变不仅是技术层面的变革，更触及战争政治经济学的深层结构。由于自主武器将作战人员从直接物理风险中剥离，国家动用武力的政治成本与心理成本大幅下降，诉诸暴力的决策门槛随之降低。[5]换言之，技术上的“精确化”反而可能催生战略上的“轻率化”。

（3）算法黑箱与战争意外升级风险

战场环境具有极高的复杂性与不可预测性，而现阶段深度学习算法普遍存在不可解释的“黑箱”缺陷。斯德哥尔摩国际和平研究所（SIPRI）的研究表明，在面临敌方电子干扰或对抗样本攻击时，自主系统的目标识别极易产生严重偏差。[6]更为严峻的是，当冲突双方均部署高度自主化的武器网络时，机器间的交战响应速度将达毫秒级，微小的算法误判可能在瞬间引发不受人类控制的链式反应，形成毁灭性的“闪电升级”效应。

（4）国际人道法的适用困境与问责真空

LAWS 的应用直接冲击国际人道法的两大基石。其一，自主系统缺乏人类的道德直觉，在复杂交战区域难以严格履行区分合法战斗人员与平民的“区分原则”；其二，系统无法动态衡量军事优势与附带损害之间的“比例原则”。此外，一旦自主武器实施非法杀伤行为，现有法律框架难以在系统开发商、部署指挥官与自主机器之间清晰划分责任，由此形成的”问责真空”将严重侵蚀军事伦理体系的规范基础。[7]

（5）军民融合驱动的非对称军备竞赛

在大国安全博弈的驱动下，军事强国正全面深化军工体系与民间科技生态的融合。美国国防部通过国防创新单元（DIU）等机制与硅谷科技企业建立快速采购通道，将商业 AI 技术加速转化为军事应用；以色列、英国等国也在积极推进

类似的军民协同创新框架。这种融合模式使得军事 AI 技术的研发边界日益模糊，传统军控框架所依赖的“军用/民用”技术二分法正面临根本性挑战。

由此催生的军备竞赛具有高度隐蔽性：AI 军事能力的提升往往嵌入在民用技术的迭代之中，难以通过传统军控核查手段予以甄别和约束。这种“嵌入式军备竞赛”不仅加剧全球战略不稳定性，也使得国际社会围绕 AI 军事应用的规范建构面临更大的信息不对称难题。

8.1.2 网络空间的攻防失衡与智能化博弈

人工智能的深度介入正在实质性地改变网络空间的攻防力量对比，使国家间的隐蔽战线博弈呈现高度复杂的新特征。

攻击侧：低门槛化与全链条自动化。人工智能正将网络攻击从高技术门槛的专业作业转变为自动化、低成本的常态威胁。Google DeepMind 在 2025 年发布的评估框架显示，基于对 1.2 万余起真实涉网攻击案例的分析，人工智能在漏洞发现、代码生成及零日漏洞利用等关键阶段均展现出显著的赋能效果。[8]对外经贸大学国家安全研究院的报告也指出，算法驱动的新型攻击范式已实现“侦查—入侵—渗透”全流程的自动化。[9]攻击成本的急剧下降不仅使网络威胁主体日益泛化，也使国家级行为体得以有组织地运用人工智能提升网络间谍行动的规模与精准度，对关键通信、学术研究及国防工业基础构成持续性威胁。

防御侧：智能依赖与内生脆弱性的悖论。面对高频次、规模化的自动化渗透，防御方不得不将安全响应架构交由智能系统主导。然而，将人工智能嵌入数字防线本身也引入了新的安全隐患。2026 年大西洋理事会关于北约算法战的研究报告指出，防御性智能模型极易成为对手的攻击目标，攻击者可通过数据投毒、传感器欺骗或参数篡改等对抗性手段直接瘫痪 AI 赋能的防御系统。[10]这种“以算法防御算法”的机制，在提升响应速度的同时使防御体系陷入“自我脆弱性悖论”，容易在缺乏人工干预的情况下引发系统性风险。

由此形成的非对称格局是：攻击方以极低成本获取体系化穿透优势，防御方则深陷智能依赖与内生脆弱交织的结构性困境。

8.1.3 情报革命与认知战的体系化升级

在情报搜集端，人工智能正在重塑海量数据的处理范式。传统情报分析受制于人力规模与多语种门槛，而大语言模型能够以自动化方式从全球社交媒体、学术出版物、政府公开数据库等开源渠道中提取关键实体关系，实现多语种实时翻译与意图研判，推动情报产出从“被动收集”向“预测性预警”转型。

在战术执行端，生成式 AI 打破了“眼见为实”的认知基准。多模态大模型仅凭少量音频与图像样本即可生成高度逼真的语音克隆与伪造视频。欧洲刑警组织（Europol）的评估报告明确指出，深度伪造技术的武器化应用不仅大幅提高了定向网络钓鱼的成功率，更对高度依赖声纹、面部识别等生物验证手段的“零信任”网络安全架构构成系统性挑战。[11]在外交危机或军事冲突的敏感节点，伪造的高级官员讲话或前线作战画面极易在短时间内引发公众恐慌与金融市场动荡，迫使目标国政府将大量战略资源耗费于辟谣与信息澄清。

在战略层面，生成式 AI 赋予国家行为体前所未有的信息环境操控能力，使虚假信息的制造与分发实现工业化。皇家联合军种研究所（RUSI）2025 年的研究揭示，部分国家正系统化地将大模型整合进信息战体系。例如持续运作的 Doppelgänger 虚假信息网络，能够自动生成数以万计带有特定政治倾向的虚假新闻链接与评论，以操纵目标国的社会舆论。[12]在 2025 年以色列与伊朗的冲突中，交战方利用大规模自动化网络钓鱼与深度伪造技术针对平民进行精准投放，以极低成本制造了高强度的社会心理压力。人工智能在认知安全领域同时赋能攻击侧与防御侧，两者的能力提升呈现“双螺旋”式的相互驱动关系，使认知空间的安全博弈陷入持续升级的结构性困境。[13]这种“算法认知战”已实质性地演化为一种标准化的“非战争军事行动”，其瓦解敌方社会共识的战略效果远比传统火力打击更为持久深远。

8.2 主要大国人工智能军事化战略竞争格局

人工智能已超越单纯的经济竞争范畴，深度嵌入大国博弈的战略内核。2025 年，全球 AI 军事应用投入规模突破 280 亿美元，联合国安理会五个常任理事国均已将 AI 列为国家安全战略的优先议题。[14]然而，这场竞争的底层结构远比“军备竞赛”一词所能概括的更为复杂，主要大国在零和对抗与正和合作之间游走，

在芯片出口管制与标准制定权上寸土必争，却又在 AI 核稳定性议题上保留有限的协商渠道。[15]

8.2.1 美国：从审慎监管到激进扩张

美国在 AI 军事化领域的投入规模居全球之首。2024 财年，美国国防部 AI 相关活跃项目达 685 个，年度投入约 18 亿美元，覆盖情报侦察、后勤保障、网络作战及自主系统各条战线。[16]这些项目不仅旨在提升现有作战能力，更试图以 AI 重新定义未来战争形态，将算法优势转化为持久性的军事代差。

2025 年，美国人工智能政策发生深刻战略转向。1 月，特朗普签署第 14179 号行政令《消除美国在人工智能领域领导地位的障碍》，全面扭转拜登政府以风险控制为核心的审慎监管路线，转向以确保全球技术领导地位为首要目标的“创新优先”与“放松管制”模式。[17]三天后，特朗普宣布启动“星际之门”（Stargate）计划，由软银、OpenAI、甲骨文联合注资 5000 亿美元，构建国家级 AI 基础设施体系。[18]

这一系列举措的深层逻辑，是将 AI 竞争从“技术研发竞赛”全面升级为“国家基础设施竞赛”。通过放松监管降低创新摩擦成本，同时以超大规模资本动员构建算力垄断优势，美国试图在 AI 领域复制一种新型的“举国体制”模式。然而，这一策略内含深刻矛盾：一方面追求最大限度的技术开放以激发创新活力，另一方面实施最严格的技术封锁以遏制竞争对手。“开放创新”与“封闭遏制”之间的张力，构成了美国 AI 战略的核心悖论，也为其盟友体系的协调一致埋下了隐患。

8.2.2 中国：效率竞争与规范建构的双轨推进

在技术创新层面，中国走出了一条以算法效率为核心的自主创新路径。2025 年初，DeepSeek R1 模型以远低于 GPT-4o 的训练成本（据报道约 600 万美元，约为 OpenAI 同级投入的二十分之一）实现可比性能，在西方政策界引发了对“算力封锁能否有效遏制中国 AI 进步”这一根本假设的深度反思。[19]这一突破表明，外部技术封锁非但未能达到预期的遏制效果，反而倒逼中国在算法效率优化领域实现了重要的自主突破。与此同时，中国在移动支付生态、社会治理网络中积累

的海量数据构成了独特的结构性优势，为自主可控的 AI 技术发展提供了坚实的数据基础。[20]

在规范建构层面，中国的战略日趋主动。2025 年发布的《人工智能全球治理行动计划》提出 13 条行动方案，有意识地区别于西方主导的“可信 AI”框架，强调“发展权”与“数字主权”，并通过“一带一路”数字基础设施合作向全球南方国家系统性推广本国技术路线与治理理念，[21]致力于构建“开放、包容、公平、对发展中国家有利”的 AI 治理路径。

8.2.3 欧盟与其他行为体：规范力量与战略对冲

欧盟在大国 AI 博弈中处于典型的“夹层”位置。《人工智能法案》(AI Act) 的落地使其成为全球首个建立综合性 AI 监管框架的主要经济体，在策略层面将“可信 AI”品牌转化为独特的规范性竞争力，以监管领导力弥补技术领导力的相对不足。与此同时，欧盟也在寻求监管严谨性与产业竞争力之间的平衡：推迟《人工智能法案》部分条款的实施，为高风险 AI 系统提供合规缓冲期；修订 GDPR，为 AI 模型训练开放高质量数据通道；简化合规程序，以减轻中小企业负担。[22]

此外，印度、日本、韩国、以色列等中等技术强国也在大国夹缝中寻求 AI 领域的战略自主空间。日本借助半导体制造的传统优势和美日同盟框架，力求在 AI 芯片供应链中占据关键节点；韩国依托三星、SK 海力士在存储芯片领域的全球领先地位，在中美之间维持“安全靠美国、市场靠中国”的双轨策略；以色列则凭借发达的军事科技生态，成为 AI 军事应用领域的重要创新策源地。这些中等强国在产业政策与联盟站队之间维持着微妙的平衡，其战略选择将在相当程度上影响全球 AI 竞争格局的最终走向。[23]

8.3 技术民族主义：从芯片管制到标准竞争

上述大国 AI 战略竞争的制度化表达，集中体现为技术民族主义的工具化实践。各国以国家安全为名，将技术研发、供应链管控与标准制定纳入地缘政治博弈的框架之中，形成了两条最具影响力的竞争轴线：芯片出口管制与技术标准制定权争夺。

8.3.1 芯片出口管制与供应链武器化

技术民族主义在当前大国竞争中最直接的制度表达，是将相互依赖关系武器化。自 2022 年《芯片与科学法案》出台以来，美国对华 AI 芯片出口限制历经多轮升级，形成了“制裁—豁免—再收紧”的高度不稳定政策螺旋。H100、A800、H20 乃至 B30A 芯片相继被纳入禁售清单，每一轮升级都迫使中国企业在采购渠道与技术路线上进行代价高昂的调整。然而，随着特朗普第二任期的政策反复，出口管制政策出现异化迹象，安全限制在一定条件下可通过付费绕过。[24]以行政令征收芯片出口收益的做法，在法律层面涉嫌违反《出口管制改革法案》对收费的明确禁止，并构成未经国会授权的征税行为，在美国国内引发严峻的宪法争议。这场政策反复的本质，是特朗普政府在“技术安全逻辑”与“商业利益驱动”之间的周期性摆荡。[25]

面对美国的单边技术封锁，中国依法采取对等措施以维护自身正当发展权益。中国控制着全球约 70%的稀土加工产能，而稀土是包括 AI 专用芯片在内的先进制造业不可或缺的关键原料。通过依法加强稀土与关键矿产出口管控，中国有效运用了自身在全球供应链中的合理杠杆。[26]这一格局揭示了技术民族主义博弈的深层脆弱性：全球科技供应链具有高度地理集中特性。荷兰 ASML 是目前唯一能制造极紫外光刻机（EUV）的厂商，这意味着任何单边脱钩战略都面临系统性连带损伤风险，制裁方与被制裁方均难以独善其身。[27]

从战略全局看，算力基础设施已成为继核武器之后最具战略杠杆价值的国家力量要素。[28]美国芯片出口政策的频繁反复已造成严重的“盟友夹缝效应”：日本、荷兰、韩国等盟友被迫在经济利益与政治忠诚之间作出高成本抉择，联盟内部的战略互信正在被单边主义管控逻辑所侵蚀。

8.3.2 技术标准制定权的战略争夺

在硬件管制之外，AI 技术标准制定权已成为技术民族主义渗透国际规范体系的另一关键战场。标准的战略价值在于其“路径锁定”效应：一旦某一技术标准在全球范围内获得先发优势，后续采用者将面临极高的转换成本，从而形成强劲的路径依赖。

当前，标准竞争呈现出中美两条路线并行的态势。美国依托其在基础模型、

云计算平台和半导体设计领域的先发优势，主导构建以“可互操作性”和“市场准入”为核心的技术标准体系，并通过在国际电信联盟（ITU）、国际标准化组织（ISO）等机构中的议程设置能力，将本国的技术规范推升为全球通行标准。中国则充分利用开源 AI 模型领域日益凸显的国际影响力，依托“一带一路”数字基础设施合作，向全球南方国家系统性推广本国技术路线与治理理念，形成与西方框架并行的替代性规范网络。标准竞争的实质，是对未来全球 AI 产业生态“游戏规则”的预先争夺，其影响将远超技术层面本身，深刻塑造各国在数字经济时代的结构性权力格局。

8.4 技术民族主义对全球安全秩序的系统性冲击

技术民族主义不仅重塑了大国竞争的工具与手段，更在深层次上冲击着全球安全秩序的制度基础。这种冲击主要体现在三个相互关联的维度：多边治理机制的碎片化、战略稳定性的结构性侵蚀，以及技术阵营化趋势所蕴含的新冷战风险。

8.4.1 多边治理机制的碎片化

技术民族主义的强化正在系统性地瓦解 AI 全球治理的多边基础。尽管联合国、G20、OECD 等国际机构相继推进 AI 治理框架，但主要大国在 AI 军事应用、数据跨境流动与算法透明度等核心议题上的立场分歧，已导致全球治理格局呈现系统性碎片化。在这一格局中，美国的“创新导向型”市场路径以放松监管为核心，实质上优先服务于本国科技企业的全球扩张；欧盟的“权利导向型”强监管路径侧重个体权利保障，但其域外适用效应引发广泛争议。相比之下，中国倡导的“发展权导向型”治理路径更注重发展中国家的利益与诉求，致力于构建普惠共享的全球 AI 治理框架。三种路径之间的张力，使得全球统一治理规范的形成面临深层障碍。[29]

2025 年巴黎 AI 行动峰会是这一裂痕的标志性事件。各方在“AI 监管是否应具有国际约束力”、“基础模型信息是否应强制披露”等核心问题上公开龃龉，最终仅就最小公约数达成声明。[30]供应链碎片化进一步加剧了各国的不安全感，促使各国在算力基础设施和算法研发上进行冗余建设，甚至引发“先发制人”式的技术打压，使人工智能领域陷入典型的安全困境。

8.4.2 战略稳定性的结构性侵蚀

技术民族主义驱动的 AI 军事化竞争，正在从结构上侵蚀冷战以来以核威慑为基础的战略稳定均衡。当 AI 技术深度嵌入核武器指挥控制 (C3) 系统时，“确保相互摧毁”（MAD）的威慑逻辑将面临系统性挑战：算法辅助决策的时间压缩效应可能使危机升级的“刹车距离”大幅缩短，进而强化先发制人的战略诱因。[31]RAND 公司的研究进一步揭示，AI 在“隐藏与发现”竞争中的不对称赋能效应将有利于进攻方，从根本上动摇核时代“防御优势”的战略假设。[32]

战略稳定性受侵蚀的机制是多维度的。在预警层面，AI 增强的侦察与监视能力可能削弱核武器的二次打击生存性，动摇“相互确保摧毁”的逻辑前提；在决策层面，算法辅助的威胁评估一旦出现系统性偏差，可能在危机情境下放大误判风险；在沟通层面，技术民族主义导致的信任赤字使大国间关于 AI 军事应用的“红线”共识更难建立，危机时刻的信号传递与意图沟通面临前所未有的障碍。

在权力博弈全面渗透战略、军事、经济与话语权各领域的背景下，AI 正在推动权力分配格局的深度重组。传统的威慑可信度机制、红线划定规则与危机沟通渠道，均需在人工智能条件下重新校准。[33]

8.4.3 技术阵营化与新冷战风险

当前最值得警惕的长期风险，是技术民族主义推动的“技术阵营化”趋势。全球数字基础设施正沿地缘政治断层线加速分裂为相互隔离的技术圈层，不同的技术标准、数据生态与治理规范在各自阵营内部形成自我强化循环，全球互联互通的数字基础正在被政治化的“技术护照”所瓦解。

哈佛大学贝尔弗中心的报告将这一趋势描述为“AI 冷战”的临界状态，认为当前虽未发展为全面的技术脱钩，但已足以制造严重的协调失灵与信任赤字。[34]这里存在一个深层的结构性悖论：越是在 AI 战略竞争中追求单边优势，各国就越难以就 AI 灾难性风险的共同防范形成有效的多边机制。而 AI 灾难性风险，包括算法失控、核危机升级、大规模认知战，恰恰是任何单一国家都无法独自应对的全球性挑战。

8.5 结语

当前大国 AI 战略竞争已呈现若干清晰的结构性特征。竞争轴线从单一的技术领先之争，扩展为覆盖算力基础设施、规则制定权、标准话语权与战略稳定管理的系统性博弈。在这场博弈中，美国以技术霸权为目标的单边主义行径是最主要的驱动力，其“创新优先”与“技术封锁”的双重战略不仅重组了自身的竞争工具箱，更在系统性地削弱维护多边 AI 治理所必需的规范基础。欧盟的规范力量、俄罗斯的信息战实践以及中间地带国家的战略选择，也在深刻影响博弈的边界条件与演化方向。在此背景下，中国坚持走自主创新与开放合作并重的道路，积极倡导构建公正合理的全球 AI 治理秩序，为应对这一全球性挑战提供了重要的建设性方案。

技术民族主义的盛行，既是这场竞争的症状，也是其催化剂。它在短期内为各国提供了争夺技术制高点的制度工具，却在长期层面持续侵蚀全球安全秩序赖以维系的规范基础与信任机制。人工智能时代的国家安全不再是单纯的技术问题或军事问题，而是横跨技术、制度与规范三个维度的复合型挑战。如何在技术自主与规则共建之间寻找动态平衡，避免“技术安全化”的自我实现式预言，将是未来十年国际安全议程的核心考验。

参考文献

[1] U.S. Department of Defense, “DoD Directive 3000.09: Autonomy in Weapon Systems,” https://globalsecurityreview.com/wp-content/uploads/2026/03/Lethal-Autonomous-Weapon-Systems-A-New-Battlefield-Reality.pdf.

[2] International Committee of the Red Cross (ICRC),“ICRC Position on Autonomous Weapon Systems,” https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems

[3] Shah, J. A, “Lethal Autonomous Weapon Systems: A New Battlefield Reality,” Global Security Review, 2026, https://globalsecurityreview.com/wp-content/uploads/2026/03/Lethal-Autonomous-Weapon-Systems-A-New-Battlefield-Reality.pdf.

[4] Trends Research & Advisory, “Governing Lethal Autonomous Weapons: The Future of Warfare and Military AI,” https://trendsresearch.org/insight/governing-lethal-autonomous-weapons-the-future-of-warfare-and-military-ai/

[5] Human Rights Watch (HRW), “ Losing Humanity: The Case against Killer Robots,” https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

[6] Boulanin, V. (Ed.), “The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk,” Stockholm: SIPRI, 2019, https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf

[7] Human Rights Watch (HRW), “ Losing Humanity: The Case against Killer Robots,” https:

//www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots
[8] Google DeepMind, “A Framework for Evaluating AI-Enabled Cyber Threats,” https://arxiv.org/abs/2503.11917
[9] 对外经济贸易大学国家安全与治理研究院,“算法军备竞赛：人工智能与国家安全战略变革,” https://sns.uibe.edu.cn/docs/2024-12/47c620157f494e329f02fa54475f0e9a.pdf
[10] Atlantic Council, “How NATO Can Integrate AI to Prevail in Future Algorithmic Warfare,”
https://www.atlanticcouncil.org/in-depth-research-reports/report/how-nato-can-integrate-ai-to-prevail-in-future-algorithmic-warfare/
[11] Europol, “Facing Reality? Law Enforcement and the Challenge of Deepfakes,” https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes
[12] Royal United Services Institute (RUSI), “Russia, AI and the Future of Disinformation Warfare,” https://static.rusi.org/russia-ai-and-the-future-of-disinformation-warfare.pdf
[13] 刘丽娜:《人工智能在国家认知安全中的攻防双向赋能机制探析》, 载《国际安全研究》, 2025 第 5 期, 第 93—116 页。
[14] Belfer Center for Science and International Affairs, “Code, Command, and Conflict: Charting the Future of Military AI,” https://www.belfercenter.org/research-analysis/code-command-and-conflict-charting-future-military-ai
[15] Stefka Schmid, “Arms Race or Innovation Race? Geopolitical AI Development,” Geopolitics, 2025, Volume 30, Issue 4.
[16] European Parliament Research Service, “Defence and Artificial Intelligence,” https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/769580/EPRS_BRI(2025)769580_EN.pdf
[17] 中国信通院, “人工智能治理研究报告（2025 年）,” https://www.caict.ac.cn/kxyj/qwfb/ztbg/202602/P020260213607027089305.pdf
[18] 央视网, “特朗普宣布“星际之门”计划,” https://news.cctv.com/2025/01/26/ARTIgx6BpURgfLlxcaWkVtbl250126.shtml
[19] 鲁传颖、才悦:《特朗普 2.0 时期中美人工智能博弈的新阶段》, 载《太平洋学报》, 2025 年第 10 期, 第 15—28 页。
[20] Sullivan, R., “The U.S., China, and Artificial Intelligence Competition Factors,” https://www.airuniversity.af.edu/Portals/10/CASI/documents/Research/Cyber/2021-10-04%20US%20China%20AI%20Competition%20Factors.pdf
[21] Junhua Zhu ,“Strategic Technology Competition Revisited: A National Innovation System Rationale for China's AI Standardisation Strategy,” Telecommunications Policy, 2025 Volume 50, Issue 2.
[22] 中国信通院, “人工智能治理研究报告（2025 年）,”https://www.caict.ac.cn/kxyj/qwfb/ztbg/202602/P020260213607027089305.pdf
[23] 复旦发展研究院, “大国竞争背景下的人工智能安全治理与战略稳定,”https://fddi.fudan.edu.cn/_t2515/4a/30/c19046a674352/page.htm
[24] CEPA(Center for European Policy Analysis), “Chip Challenge: Goodbye Export Controls,” https://cepa.org/article/chip-challenge-goodbye-export-controls/
[25] Brooke Becher, “Trump Lifted the AI Chip Ban on China, Clearing Nvidia and AMD to Resume Sales,” https://builtin.com/articles/trump-lifts-ai-chip-ban-china-nvidia
[26] Dumont, F, “Technological Dominance: The AI Arms Race Between the United States an

d China,”Pace International Law Review，2025. https://pilr.blogs.pace.edu/2025/11/25/technological-dominance-the-ai-arms-race-between-the-united-states-and-china/
[27] 联合国裁军研究所（UNIDIR），“使能技术与国际安全（2023 年版），”https://unidir.org/wp-content/uploads/2024/03/2023_Compendium_of_Enabling_Technologies_and_International_Security_CN.pdf
[28] 对外经济贸易大学国家安全与治理研究院，“算法军备竞赛：人工智能与国家安全战略变革，”https://sns.uibe.edu.cn/docs/2024-12/47c620157f494e329f02fa54475f0e9a.pdf
[29] 复旦发展研究院，“大国竞争背景下的人工智能安全治理与战略稳定，”https://fddi.fudan.edu.cn/_t2515/4a/30/c19046a674352/page.htm
[30] 求是网，“加强人工智能国际治理和合作，”https://www.qstheory.cn/20250303/a80d9d6bad224e308f42f2deebd7b7f9/c.html
[31] 张东冬：“人工智能军事化与全球战略稳定，”载《国际展望》，2022 年第 5 期，第 142—161 页。
[32] RAND Corporation，“How AI Could Reshape Four Essential Competitions in Future Warfare，”
https://www.rand.org/pubs/research_reports/RRA4316-1.html
[33] 光明网，“人工智能对国家安全的冲击，”https://theory.gmw.cn/2025-04/17/content_37971289.htm
[34] Atlantic Council，“How NATO Can Integrate AI to Prevail in Future Algorithmic Warfare，”
https://www.atlanticcouncil.org/in-depth-research-reports/report/how-nato-can-integrate-ai-to-prevail-in-future-algorithmic-warfare/

第九章：人工智能安全的法律规制与全球治理

肖超伟 国家发展与战略研究院

本章节立足全球人工智能治理范式转移的实证分析，揭示了美欧中三大技术策源地在规制逻辑上的结构性断裂。研究指出，欧盟正试图通过梯次生效的风险法案强制输出合规标准，美国全面转向去监管化与绝对算力霸权，而中国已确立全链条标识强制监管体系。面对全球人工智能治理范式多元化的现实，本章节建议，构建具有约束力的国际安全协同机制，需要建立类国际原子能机构的专属实体，并将监管实质性下沉至底层。此研究发现为研判全球数字主权博弈与安全防御提供了一定的决策支撑。

9.1 全球人工智能治理范式的历史性转移

在 2025 至 2026 年这一关键的历史交汇点，全球人工智能的演进轨迹发生了根本性的改变。随着生成式人工智能、具备高度自主规划与执行能力的代理型人工智能以及前沿通用人工智能模型的指数级突破，国际社会对人工智能的认知已经从单纯的技术工具，全面升级为决定国家生死存亡、重塑全球权力分配结构的关键战略基础设施。在这一背景下，全球人工智能安全治理迎来了决定性的范式转移：国际社会的治理焦点已从早期停留在抽象伦理原则、自愿性倡议与软法层面的探讨，全面且不可逆转地转向了实质性的法律规制、硬性合规审计、底层的硬件级技术核查以及激烈的大国地缘政治博弈。

人工智能技术呈现出极强的军民两用属性。一方面，它在医疗诊断、抗灾工程、气候变化建模与金融普惠等领域展现出巨大的和平利用潜力；另一方面，前沿模型若被恶意滥用或发生非预期失控，其在自动化网络攻击、合成生物武器设计、深度伪造引发的认知战以及社会关键基础设施瘫痪等方面，均具备引发系统性甚至生存性风险的可能。面对这种非对称的极端风险，全球监管体系却表现出令人担忧的滞后性与高度的碎片化特征。

当前，以欧盟、美国和中国为代表的三大核心经济体及技术发源地，基于其迥异的政治体制、产业发展基础、法律文化传统以及全球战略诉求，衍生出了三条截然不同甚至在底层逻辑上相互冲突的国内治理路径。欧盟试图通过综合性的统一立法输出其基于基本人权和风险分级的“布鲁塞尔效应”；美国在经历了深刻的国内政治周期后，全面转向以去监管化和绝对算力基础设施主导的实用主义路线；而中国则在关注国家安全与社会稳定的同时，开创了聚焦具体应用场景、算法交互安全与敏捷响应的治理模式。

在国内治理路径分道分道扬镳的同时，构建具有普遍约束力的国际人工智能安全治理机制正面临前所未有的制度性与技术性阻碍。全球南方国家对被边缘化和“监管帝国主义”表达了强烈的抵触，要求在治理架构中获得平等的话语权与技术共享；私营科技寡头掌握着绝大部分的算力资源与顶尖人才，使得传统的以民族国家为中心的国际法体系在面对超国家的技术垄断时显得力不从心；更为关键的是，由于算法的无形性与模型的快速迭代，缺乏基于底层硬件标识溯源的技术核查手段，导致任何关于限制前沿人工智能研发的纸面国际公约都极易沦为一纸空文。

本章将以历史演进与多维度的政策文本为基础，深度剖析并比较欧盟《人工智能法案》、美国行政令及中国相关法规的异同，进而系统性地探讨在多边主义受挫的当下，依托底层安全验证与硬件信任根技术、敏捷危机响应架构以及全球南方包容性框架，构建具有强约束力与技术可验证性的国际安全治理机制的现实路径与深层障碍。

9.2 主要经济体人工智能法律规制的比较分析

在全球人工智能治理的宏大版图中，欧盟、美国与中国的国内立法与行政干预绝非孤立的内部事务。这三大经济体的规制模式不仅深刻形塑了其本土的创新生态与产业结构，更通过庞大的市场准入壁垒、跨国数据流动规则、全球供应链的相互依存以及技术标准的外溢，对全球治理规则的演进产生了决定性的影响。理解这三种治理模式的本质异同，是探究全球协同治理机制可行性的绝对前提。

欧盟：基于风险分级的全面预防性监管

欧盟《人工智能法案》的正式落地，标志着人类历史上首部综合性、跨领域的硬性人工智能法律框架的诞生。该法案的底层治理逻辑深深植根于欧洲对基本人权、消费者保护、数据隐私与透明度的历史传统，采取了经典的预设性风险分级监管架构。欧盟并未采取“一刀切”的激进生效模式，而是设计了极为精细且长达数年的分阶段实施时间表，以期在确保根本安全的前提下，为市场参与者留出合规调整与技术改造的缓冲期。

从具体的实施时间线演进来看，该法案的适用过程被严密地划分为多个关键节点。2025 年 2 月，法案中关于通用条款、人工智能素养义务以及针对不可接受风险的全面禁令率先生效 。至 2025 年 8 月，针对通用人工智能模型提供者的规则正式实施，要求研发具有系统性风险的通用大模型的企业必须建立完善的内部治理架构，执行严格的模型评估，并强制履行版权保护与训练数据透明度披露义务；同时，各成员国被要求在此节点前指定国家级主管机构并出台配套的国家处罚法律 。

在国际战略层面，欧盟的雄心远不止于规范内部市场。欧盟意图复制其在《通用数据保护条例》上取得的巨大成功，利用其作为全球最富裕单一市场的庞大体量，迫使希望进入欧洲的跨国科技巨头在全球范围内统一采用欧盟的合规标准，从而实现所谓“布鲁塞尔效应”的全球输出。对于许多跨国企业而言，为不同司法管辖区开发多套底层架构的成本过高，因此主动向欧盟的最严标准看齐成为了理性的商业选择。

然而，到了 2026 年，这种被某些学者称为“规制帝国主义”的单边标准输出，遭遇了相关的抵制与质疑。特别是在广大全球南方国家看来，盲目移植欧盟那种高度复杂且合规成本极高的防御性规则，不仅会扼杀本土脆弱的创新生态，更会成为阻碍技术普惠的壁垒，变相巩固了西方大型科技巨头凭借资金优势形成的垄断地位。批评者指出，欧盟法案在涉及国家安全和私人领域的豁免条款上留下了显著的漏洞，同时其繁琐的程序性要求更倾向于形式上的“实验主义治理”，而非能够真正应对前沿模型快速迭代的治理工具。

美国：从联邦去监管化、基础设施霸权到州级立法的司法对抗

相较于欧盟试图以法律文书框定技术未来的保守主义，美国在 2025 年至 2026 年间的监管轨迹经历了一场剧烈的战略重组。随着联邦政府权力的更迭，美国彻底摒弃了前任政府行政令中所包含的强烈的安全评估要求与关注多样性、公平性与包容性的干预主义路径，全面转向以自由市场驱动、去监管化以及绝对的底层基础设施霸权为主导的极端实用主义路线。

这一战略转向的法律依据与政策宣言集中体现在新任政府发布的一系列具有颠覆性的行政指令中。2025 年 1 月，美国总统正式签署相关行政令，不仅明确撤销了前任政府的核心安全指令，更将技术的竞争赤裸裸地定义为国家安全与经济竞争力的生死存亡之战。该行政令指出，美国决不能容忍在算力和能源基础设施上依赖他国，必须通过大规模的私营部门投资，加速数据中心建设与电网升级，确保美国在“前沿 AI”领域的绝对领导地位，防止潜在对手利用强大的模型损害美国的军事安全。

在此基础上，白宫于 2025 年 7 月出台了被视为美国新时期科技战略蓝图的《美国 AI 行动计划》。该计划以加速创新、构建基础设施以及主导国际外交与安全为三大支柱，标志着美国联邦政府角色的根本性转变——从技术发展的看门人转变为扫雪机 。在去监管化方面，该计划指令联邦机构发起公开调查以识别并废除一切阻碍技术采用的繁文缛节；要求相关贸易与通信委员会重新审查并废除那些可能构成创新负担的调查指令与同意令 。更为关键的是，该计划强制要求国家标准与技术研究院重写其风险管理框架，彻底剔除其中关于气候变化、虚假信息以及社会包容性理念的表述，确保底层开发完全回归纯粹的技术性能迭代，免受意识形态干扰。

然而，联邦政府激进的去监管化策略在美国国内引发了严重的司法与行政冲突，形成了显著的监管真空。由于联邦层面迟迟未能出台综合性的安全立法，加利福尼亚州、德克萨斯州和伊利诺伊州等数十个州在 2026 年 1 月 1 日纷纷实施了针对就业歧视、医疗保健消费者保护等领域的州级法案；科罗拉多州更为详尽的全面法案也定于 2026 年中生效 。面对这种日益碎片化且可能阻碍跨州商业部署的州级监管网络，联邦政府于 2025 年底采取了强硬的干预手段。该命令要求司法部组建专属的诉讼特别工作组，利用宪法中的联邦优先权原则，对所有被认定为阻碍联邦去监管化政策的州级法律发起系统性的司法挑战。这种联邦与州之间的权力博弈，为跨国企业在美的合规环境蒙上了浓厚的法律不确定性阴影。

中国：统筹安全与发展的治理模式

与欧盟基于冗长立法的预防模式和美国完全放任市场的模式均不相同，中国在治理上走出了一条高度独特、务实且兼具安全与发展的治理模式。中国的监管逻辑并未脱离其宏观的政治经济学框架：即全力追赶并超越前沿技术的同时，必须确保技术演进需要服从于国家安全、社会稳定、数据安全以及未成年人保护的根本底线。

2026 年标志着中国治理体系从初期的单点试错迈入了一个更为精细化、标准化与具有强制穿透力的法治化新阶段。在宏观立法层面，尽管备受瞩目的统一的《人工智能法》仍处

于起草、论证与专家草案迭代阶段 ，但中国充分利用了现有的法律框架进行适应性改造。2025 年 10 月，中国全国人大常委会通过了对《网络安全法》的重大修订，并在该法中首次增设了专门针对相关技术安全发展的条款。该修订案于 2026 年 1 月 1 日正式生效，不仅将底层治理正式纳入国家关键信息基础设施与供应链安全的法定范畴，提出了加强基础研究、防范伦理风险的法定要求，更大幅强化了对违规行为的经济威慑——企业面临的最高罚款额被提升至 5000 万元人民币或上一年度营业额的 5%，为全生命周期的合规确立了相关规范。

在此法律基础之上，中国相关部委通过发布行政规章和国家标准，构建了一套覆盖全产业链的细颗粒度监管网络。2025 年 9 月生效的管理暂行办法及后续出台的标识规则，强制要求所有生成内容必须进行显性或隐性的水印标识，以防范认知战与虚假信息泛滥 。2025 年底生效的一系列国家安全标准，更是深入到算法的源头，对预训练数据、微调数据的收集、标注与清洗过程设定了严格的安全评估规范，要求数据源必须合法且符合社会主义核心价值观 。

欧盟、美国与中国治理模式的系统性与结构化对比

为了更直观地呈现三大经济体在安全治理上的结构性分歧，下表从多个核心维度对其治理范式进行了系统性对比与归纳：

治理维度	欧盟	美国	中国
核心价值观驱动	基本人权、消费者保护、程序正义、透明度、防范算法偏见与隐私侵犯。	自由市场绝对创新、国家基础设施霸权、去监管化、言论自由与商业利益。	国家总体安全、社会秩序稳定、技术自主、未成年人与心理健康保护。
治理架构与立法模式	综合性统一立法，基于预设的横向风险分级体系，辅以多年期分步实施计划。	联邦层面推行去监管化与行政干预，通过司法部打击州级零散立法；聚焦基础设施建设。	依托上位法兜底。
对私营部门的态度	严格的预防式事前与事中监管，强制要求	极度赋能，视科技企业为国家力量的代理	强力引导、严格规范与合规审计并重，要求企

	提供系统技术文档、外部透明度审计与巨额罚款威慑。	人，大幅削减合规负担以加速底层基建扩张。	业绝对承担价值观与防范垄断的第一主体责任。
对创新与风险平衡的倾向	风险厌恶型。为了消除不确定的潜在社会风险，愿意牺牲一定程度的创新速度与商业效率。	风险偏好型。在非致命风险领域接受试错，认为减缓创新本身就是对国家战略安全的最大威胁。	动态平衡型。在严控政治与社会安全红线的前提下，大力扶持基础模型突破与产业应用落地。
国际影响与规则输出策略	依靠庞大的内部单一市场消费力，强制跨国企业合规，从而对外输出“布鲁塞尔效应”。	通过全球最顶尖的技术垄断、严格的高端芯片出口管制网络以及事实上的硬件架构标准维持排他性。	强调技术主权治理，在联合国框架下积极为发展中国家发声，对外输出自身的数据标准。

这一三种类型的治理模式，深刻地揭示了当前国际机制建立所面临的根本困境：当全球最重要的三个技术策源地，在最核心的命题——“什么是最大的风险”、“技术发展的最终目的为何”以及“政府应当将手伸向多深”上存在不可调和的认知分歧时，任何试图在短时间内达成具有普遍约束力的全球性大一统国际公约的努力，都可能将举步维艰、甚至陷入长期的僵局。

9.3 当前国际安全治理机制的演进、成效与挑战

在主权国家层面的地缘博弈日益加剧的背景下，为避免技术发展演变为一场不受约束的逐底竞争，国际社会和众多国际组织在 2025 至 2026 年间为统筹全球治理付出了巨大的外交与学术努力。然而，不同国际组织、利益集团所主导的治理机制存在明显的路径依赖与排他性分歧，导致全球治理呈现出多层次、多重叠但缺乏实质强制力的碎片化状态。

联合国始终致力于超越意识形态分歧，构建一个覆盖全球、不留死角的包容性治理架构。2025 年底，联合国高级别咨询机构发布了里程碑式的最终报告。该报告经过广泛的全球咨询，揭示了当前地缘政治现实：在当时主导全球讨论的众多治理倡议中，多达 118 个国家被完全边缘化，处于被剥夺发言权的状态。

在这一系列宏大构想中，最具实质性突破与历史意义的是 2026 年 2 月联合国大会正式投票通过并成立的国际独立科学专家组。该专家组由横跨学术界、私营部门、民间社会与政府机构的顶尖专家组成，首届任期至 2029 年。其核心使命是以绝对独立和科学严谨的姿态，每年定期发布关于全球能力演进、潜在机遇以及前沿不可见风险的权威事实评估报告 。此举旨在以硬核的科学事实为基础，打破少数寡头科技公司对前沿黑箱的信息垄断，为弱势国家参与全球政策对话提供平等的知识基础。此外，为实现利益共享，联合国内部正紧锣密鼓地筹建能力建设网络以对接人才培训，设立全球基金以弥补发展中国家算力基础设施的资金缺口，并倡导建立全球数据框架以打破跨境数据垄断。

联合国的努力极大地唤醒了全球南方国家的战略意识。长期以来，南方国家在生态链中仅仅扮演着算力物理资源提供者、数据廉价标注劳动力以及被动消费市场的角色，承受着成为技术规则接受者的结构性剥削。进入 2026 年，以印度、非洲联盟和拉美部分国家为代表的南方国家阵营，开始在国际治理机制中发出强有力的主权声音。2026 年在印度新德里举办的峰会成为这一觉醒的标志性事件。面对西方大型语言模型所带来的隐性认知殖民与数字鸿沟扩大的风险，全球南方积极推进本土化技术运动。例如，非洲成功开发了覆盖五大本土语言的模型；秘鲁推出了针对印加语系的本土数字档案馆；印度构建了庞大的方言语音数据信托体系。这种技术本土化与文化觉醒，迫使全球治理的议程不仅要防范西方世界高度关注的前沿生存性风险，更必须优先解决迫在眉睫的算法公平、技术可及性以及数字主权保护等发展议题。

9.4 构建具有强约束力的国际安全治理与核查机制的前沿路径

面对日益失效的法制治理与碎片化的国际协调，全球战略界与国际法学者逐渐达成一个严酷的共识：构建具有真正约束力的国际安全机制，绝非仅仅依靠外交官在谈判桌上达成一份纸面公约即可实现。它必须依赖于深度嵌入半导体物理硬件的底层防伪验证网络、打破公私权力失衡的国际机构重塑，以及针对极端灾难的敏捷危机响应架构。

近年来，政策界最流行的倡议之一是呼吁在联合国框架下建立一个类似于国际原子能机构的人工智能专属国际机构 。国际原子能机构在核不扩散领域的半个多世纪的成功，源于其完美结合了技术转移与安全核查两大核心职能。倡议者构想，新的专属机构应当具备类似的功能：一方面促进前沿技术在医疗、环境、灾害响应等领域的安全应用向发展中国家转移；另一方面，作为独立的超国家监管实体，负责制定跨国安全前沿标准并对各国的超级数据中心执行严格的核查监督。

综上所述，欧盟的全面风险分级、美国的市场导向以及中国的风险防控与统筹发展，共同勾勒出了当前全球人工智能法律规制的复杂图景。这种规制模式的多元并存既反映了各国不同的社会治理哲学，也凸显了全球共识的匮乏。面对人工智能可能带来的生存性挑战，人类社会必须超越狭隘的民族国家利益，以构建命运共同体的宏大视野，积极探索以联合国或专门性国际人工智能机构为核心的强约束力治理路径。这需要确立不可逾越的生存红线，引入多利益攸关方共治，并大力发展监管科技以突破核查瓶颈。然而，地缘政治的零和博弈、规制套利的诱惑、法律滞后于技术的天然缺陷以及算法黑箱带来的核查壁垒，构成了横亘在

这一理想道路上的深渊。在人工智能技术最终跨越“奇点”之前，国际社会能否展现出足够的政治智慧与道德勇气，克服这些结构性障碍，将决定人类作为一个物种能否在智能时代延续文明的火种。法律与全球治理在此刻不仅是分配利益的工具，更是捍卫人类主体地位的最后防线。

第十章：结论与政策建议

许伟 王玮航 于伯平

信息学院

10.1 人工智能安全研究结论

本报告系统梳理了人工智能安全研究的前沿进展与核心挑战，明确了技术迭代速度远快于治理框架更新的深层问题，说明开展相关安全研究的紧迫性和必要性。报告判断，当前人工智能技术正处于快速跃迁阶段，安全治理范式也正处于从“规则防御”转向“内生韧性”的重要战略机遇期。

一是深耕数据安全与隐私保护，推动安全理念从“静态安全存储”向“动态价值挖掘”转型。数据要素的流动性与安全性之间的博弈已进入深水区。隐私计算、联邦学习等技术目前正在金融、医疗等领域推进工程化落地，但数据投毒、后门注入、模型反演、梯度泄露和训练数据泄露等风险也日趋隐蔽，使社会各行业在推动数据要素流动的过程中面临新的安全隐患。仅在技术层面开展数据加密等手段已不足以应对复杂多变的风险态势，必须建立覆盖数据采集、清洗、训练、部署、推理、反馈和审计等环节的全生命周期动态风险监测机制，在治理端实现制度规范与技术工具的深度耦合，确保数据在释放新质生产力动能的同时，严守国家安全、商业安全与个人隐私的底线。**二是探索价值对齐与目标安全，驱动对齐路径从“技术参数校准”向“伦理制度嵌入”跨越。**仅依靠单一的奖励函数设计可能无法完全避免人工智能系统的目标漂移、对齐伪装和策略性规避。系统被赋予高度自主权时，其内部生成的子目标可能与人类核心价值观产生偏离。当前人工智能安全研究的关键难题，在于如何将抽象的价值原则、法律规范和社会伦理转化为算法可识别、可执行、可验证的约束条件。未来的研究方向需要跨越单一计算机科学的范畴，进入法治规则算法化与伦理标准形式化的深层领域，实现技术演进与人类文明价值的深度融合对齐。**三是保障社会公平与意识形态安全，构筑安全体系从“被动风险应对”向“敏捷治理与制度韧性”演进。**人工智能对劳动力市场的冲击正从简单的岗位替代转向深层的职业结构重构。“人工智能+”行动创造新型就业机会的同时，算法歧视与资源分配不均使得数字鸿沟在部分领域仍有扩大趋势。生成式技术在舆论操控与深度伪造中的应用，对信息生态真实性构成严峻挑战。必须构建可信内容溯源体系与多主体协同机制，以对冲算法驱动下的信息茧房风险，守住人工智能安全领域的意识形态生命线，维护社会公平正义和人类的主体性地位。**四是夯实基础设施安全底座，推动防护体系从“单点边缘加固”向“系统内生免疫”跃升。**底层架构的稳固性是人工智能安全体系的逻辑起点。应重点关注城市数据大脑与大规模算力集群的韧性构建，强化算力资源调度、算法模型运行与数据要素流动三者的协同免疫能力。通过部署全方位立体闭环的内生安全架构，可提升人工智能基础设施在面临未知攻击或极端负载时维持核心功能的连续性的能力，为人工智能技术应用的高质量发展筑牢底座支撑。

10.2 面向多元主体的政策建议

人工智能领域的安全问题，不可能依靠单个主体或者单一手段就彻底解决，相关工作覆盖技术、经济、法律、伦理、社会等多个维度，涉及政府、学界、产业界等不同类型的参与主体。因此，应构建政府、学界、产业界共同参与的多元协同治理框架。在框架运行过程中，政府承担“元治理”相关职责，负责制定顶层设计、法律法规和公共政策，维护公平有序的行业竞争环境，完成关键领域和关键环节的监管工作。学界承担研究支撑、风险预警和公众教育职责，为人工智能安全研究输出智力成果，帮助公众建立相关认知。产业界是技术创新和应用落地的主体，应承担产品安全、数据合规、算法公平和社会责任。这类多元协同的治理模式，可以大范围凝聚社会共识，整合不同主体的优势资源，形成稳定的治理合力，共同应对智能时代出现的各类新问题。

一是政府完善敏捷治理工具箱与制度供给。政府层面的核心任务，是完善敏捷治理工具箱，强化高质量的人工智能安全制度保障。政府要将传统行政指令式监管，转向数据驱动的智慧监管，推动规则动态生成与迭代优化的机制建设。制定大规模产业政策或监管规则前，可依托数字孪生技术搭建政策仿真系统，在虚拟环境中推演政策对市场主体、技术演进的实际影响，提升监管的预见性与科学性。对于自动驾驶、智慧医疗等高风险应用场景，政府应加快推进监管沙盒机制落地。划定特定区域与特定期限，允许企业在有限政策豁免范围内开展风险测试，并配套建立实时监测、动态评估与即时熔断机制，保证创新探索全程在安全、可控、向善的框架内推进。政府还应强化人工智能基础设施的安全韧性监管，把算力中心、公共数据集及核心算法框架划入国家战略安全资产管理范畴，增强全社会应对系统性算法失控、外部攻击的防御能力。国际层面，政府要主动参与全球人工智能治理规则制定，推动构建更公平、透明的跨国数据流动与技术安全标准，输出中国智慧治理的实践方案，推进建立以人为本、智能向善的人工智能发展生态。

二是学界深化学科交叉与价值观对齐研究。高校与科研机构是人工智能安全研究、技术评估的核心参与主体，能够给政策制定环节提供科学依据，给风险评估工作提供可用的技术工具，给实践中遇到的伦理困境提供对应的学理分析。学界应破除传统学科之间的壁垒，搭建由计算机科学、法学、管理学、社会学等多领域学者组成的跨学科研究团队，重点围绕法律规则算法化、伦理准则形式化和社会价值可计算化等问题开展研究。后续的科研工作重点应放在探索把复杂伦理准则转化为底层数学逻辑的相关路径上，让人工智能模型参与决策的过程中，具备和人类相似的“道德直觉”与“合规意识”。学界可开发更先进的模型内部逻辑分析工具，优化大模型的透明度与可解释性，从技术底层逻辑层面减少黑盒风险的发生可能。针对人才培养环节，要在高等教育体系内普及人工智能素养相关课程，既要培养掌握编程能力和模型应用能力的专业技术人才，也要培养拥有批判性思维、伦理敏感度及跨界协作能力的复合型智治人才。学界还应承担防范社会风险的相应职责，通过举办行业论坛、发布研究报告、开展科普活动等方式，凝聚社会共识，推动形成负责任的人工智能安全发展模式。

三是产业界践行设计安全与企业数字责任。针对产业发展实际，最核心的推进路径是落实设计安全相关理念，主动承担企业应尽的数字责任。产业界要明确，人工智能安全不只是

企业要满足的合规要求，也是支撑企业长期发展的核心竞争力。企业必须严格遵循国家现有法律法规与监管要求，搭建完善的内部合规管理体系，保证自身推出的产品、提供的服务符合安全、合法、合规的基本要求，具体包括算法备案、数据安全评估、内容标识等相关工作。头部企业应发挥示范性作用，主动参与行业标准和伦理规范的制定，带头履行社会责任，推动行业自律机制形成，并搭建行业层面的漏洞共享与协作防御平台，提升全行业的人工智能安全防护水平。设计商业模式的过程中，产业界要严格遵循算法公平性原则，持续识别、缓解和纠正求职招聘、金融信贷、公共服务等领域可能存在的算法偏见，推动技术发展红利更加公平地覆盖不同群体。企业应搭建完善的内部数字责任体系，具体包括设置伦理委员会、设立算法审计岗位、优化公众反馈渠道等，确保在追求商业运营效率的过程中不损害公共利益。产业界应加强与政府、学界的沟通联动，及时反馈技术实践过程中遇到的伦理问题，构建人工智能安全经验、技术创新的双向反馈通道，共同推进人工智能产业向更安全、健康、可持续的方向发展。

10.3 人机协同共治的未来图景

在人工智能技术快速迭代与应用的当下，人们对未来社会发展前景愈发关注。一方面，随着技术的持续突破、理论的不断完善以及实践的深入推进，人机协同共治正站在新的历史起点上，蓄势待发。另一方面，法律、道德与社会保障体系也面临着新的冲击。未来人工智能安全治理将逐渐完成从“治 AI”到“用 AI 治”的深刻转变。前者主要将人工智能视为被监管的对象与潜在风险源，后者则进一步将人工智能视为治理体系的有机组成部分与关键抓手，使其成为辅助风险识别、态势感知、决策支持和治理反馈的重要工具，最终迈向人机协同共治的智慧文明新图景。这一转变的核心特征在于，人工智能不再仅仅是被监管的对象，而是成为安全体系中不可或缺的深度参与者，协助人类治理者在应对极其复杂的全球性挑战时提供精准的决策支持，实现治理效能与文明价值的双重跃升。

一是治理范式从“治 AI”向“用 AI 治”跃迁。在未来的社会运行中，人工智能安全治理将逐步形成基于价值对齐的人机协同治理模式。人工智能技术将作为保障自身应用安全的第一道关卡，实现自我监管与动态优化。这意味着人工智能系统不仅要能执行预设的安全规则，还能通过持续学习和环境感知，主动识别潜在风险并进行自我修正迭代。传统的“治 AI”模式依赖人类制定规则并实施监管，侧重于事后的规制与整改，容易受到技术迭代速度快、应用场景复杂等因素制约。而“用 AI 治”则充分发挥人工智能的性能优势和学习能力，，推动治理模式从事后处置转向事前预防、事中监测和事后追溯相结合的闭环治理，将监管触角延伸到人类难以预设和触及的潜在风险中。为确保人工智能的安全治理行为符合人类的价值取向，价值对齐技术将成为核心支撑。通过将伦理准则、法律规范和社会共识转化为模型可识别、可执行、可审计的约束条件，可以降低人工智能在辅助决策中的偏差和失控风险。比如，在医疗人工智能的应用中，系统可以依据医学伦理、诊疗规范和法律要求，为医生提供诊断参考和风险提示。这种从“治 AI”到“用 AI 治”的范式跃迁，不仅是人工智能技术层面的升级，更是人工智能安全治理逻辑的重塑与理念的革新。

二是权责分配从“静态划分”向“动态分配”转型。在构建未来人工智能安全体系的进程中，人机关系将从“工具性辅助”逐步走向“共生性协同”的范式迁移，这一深层变革将引起治理架构、职能配置与运行逻辑的系统性重塑。具体而言，人机协同的权责分配应基于治理任务的内生属性进行动态化配置：在交通枢纽调度、资源能耗测算及应急物流等高频、高确定性、低风险的数据密集型领域，人工智能将驱动治理模式由被动介入转向算法驱动的自主预警；而在涉及人民生命安全、人格尊严及复杂利益分配的关键领域，则须坚持“以人为本”原则，由人工智能技术提供深度知识图谱支撑与多维决策辅助。这种动态配置要求未来治理必须具备极高的敏捷性与适配度，确保决策权力配置在算法效能与人类智慧间达成战略平衡。为了保障这一协同机制的良性运转，制度供给的重构已成为当务之急。一方面，亟需通过法律与技术规范的耦合，明确智能系统在公共决策链条中的功能边界与行为红线，防止算法权力的无序扩张；另一方面，应建立穿透式的技术建议审核备案体系，对数据源头的真实性、模型逻辑的鲁棒性及输出结果的可解释性进行全流程监管，确保每一项由技术驱动的治理行为皆具备清晰的责任接口与可回溯性。此外，还应预置“人工接管优先”的干预节点与柔性容错机制，从制度设计层面减少因技术盲区或系统偏差导致的追责困局，从而在人机深度协作的未来图景中，实现技术效能、公共理性与法治正义的协调统一。

三是实现“数字公平”与“全球共治”的文明新图景。未来的人工智能安全研究将更加重视制度韧性与数字公平。依托于区块链、隐私计算与可信执行环境等技术，人工智能应用有望建立更加高度可信、透明且防篡改的数据底座。公众能够通过便捷、普惠的数字渠道，以实时、互动、平等的方式参与公共事务的决策、监督与评估，这种模式将有效缓解技术焦虑与社会分化，弥合数字鸿沟。通过算法普惠与数字包容机制，应让更多公民在智能时代拥有技术主体感与治理参与感。全球范围内，不同文化背景、发展阶段与制度环境下的人工智能安全治理经验，可通过开源协议、跨国数据信任框架与多边技术安全标准实现互鉴与整合，形成既尊重文化多样性、又具备共性伦理约束的全球人工智能命运共同体。在应对气候变化、公共卫生、跨境网络安全及粮食安全等跨国界危机时，人机协同的全球监测网络、预测模型与调度系统将发挥中流砥柱的作用，推动全球走向协同共治。唯有如此，才能在技术迭代加速与社会诉求多元的双重压力下，筑牢人工智能现代化的安全和制度基座，使智能工具真正成为延伸公共理性、促进社会公平的杠杆，而非替代人类判断、加剧不平等的黑箱。

未来人机协同共治的发展方向，会牢牢扎根在“以人为本”的文明基础之上。人工智能的发展不应削弱人的主体地位，也不应替代人的价值判断，而应通过合理的制度设计和技术约束，防止技术异化风险，释放释放人类的潜在能力，支撑人类实现自由发展，最终成为增进人类福祉、保障文明安全、推进国家治理现代化、促进全球治理体系变革的战略性支撑力量。对人工智能安全展开深度研究，可明确安全管控的边界，拓展技术发展的可行空间，推动人工智能的整体发展模式完成本质层面的转型升级。