CAICT 中国信通院

集智 蓝皮书 No.202510

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

（2025 年）

中国信息通信研究院

2026年1月

版权声明

前　言

人工智能技术以其前所未有的发展态势引领新一轮科技革命和产业变革，深度驱动经济社会发展。2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》，系统推动人工智能在全社会的普及应用与深度融合，逐步形成“以创新带应用、以应用促创新”的良性发展飞轮。技术跃迁释放价值红利的同时带来风险挑战，只有当“安全内核”嵌入“发展飞轮”，产业才能实现高质量和可持续发展。

当前，全球人工智能安全治理逐步迈向体系化、实操化阶段，各主要经济体持续推进治理框架落地，如何在产业层面构建务实有效的安全治理举措，已成为核心议题。我们深刻认识到，技术快速发展催生新风险敞口，应用场景快速延展引入复杂治理难题，应对人工智能安全治理问题需树立动态的人工智能安全观。与此同时，人工智能安全相较于传统网络安全、数据安全，风险载体愈发复杂多样、安全防御体系尚处探索阶段、安全检测评估难度加大、传统管理机制难以适配，产业界亟需一个全局性、系统化的行动框架作为指引。为此，本报告基于中国信息通信研究院前期《可信人工智能白皮书（2021 年）》《人工智能治理蓝皮书（2024 年）》相关研究积淀，面向产业发展凝练 2025 年人工智能治理宏观要求，基于《人工智能安全框架（2022 年）》《人工智能风险治理报告（2024 年）》相关产业积累，进一步提炼当前产业面临的人工智能安全治理挑战。报告经过深入研究与实践总结，立足本土产业实践，提出“两横三纵”的人工智能安全治理产业实践框架。该框架以“管理”与“技术”双线协同为横轴，实

现制度牵引与能力支撑的深度融合；以“开发侧”“部署侧”与“应用侧”三侧发力为纵轴，实现从模型研发、系统部署到场景应用的全链条防护，旨在为产业提供一套系统性、可落地、动态化的安全治理体系。同时，我们呼吁产业各界凝聚共识、形成合力，将分散的治理努力整合为统一的系统屏障，共同构建协同共治、安全可信的人工智能生态。鉴于人工智能技术应用日新月异，本报告对人工智能安全治理的认识仍有未尽之处，恳请大家批评指正。

谨向为本报告提供指导、支持与协助的中国政法大学、中国社会科学院法学研究所、对外经济贸易大学、同济大学、深圳市腾讯计算机系统有限公司、阿里巴巴（中国）有限公司、北京百度网讯科技有限公司、中国邮政储蓄银行、上海明品医学数据科技有限公司、煤炭科学研究总院有限公司、中国移动通信集团有限公司、中国联合网络通信集团有限公司等单位及专家，致以衷心的感谢。

目 录

图 目 录

一、人工智能安全治理概述

（一）宏观规范要求迈向务实新阶段

2025 年，人工智能技术持续快速迭代创新，多项突破性进展正推动其能力边界不断扩展。**基础模型**推理能力显著增强，通过构建统一的跨模态表征空间，实现图像、语音和文本等多模态应用。**智能体**打通“感知、决策、行动”闭环，使模型从“语言生成器”向“任务执行者”转变。**具身智能**驱动机器人实现高阶认知与自主学习。人工智能与基础科学深度融合，在生物化学、医药研发等科研领域取得突破性进展，驱动科研范式发生深刻变革。然而，人工智能数据安全、算法偏见、模型幻觉、情感依赖、数据污染等问题不但未被妥善解决，反而在技术深入赋能过程中被不断放大。人工智能安全治理相关问题已经成为严重阻碍技术红利释放的掣肘。

在此背景下，全球人工智能安全治理已经发生深刻转变。**一是**从对于宏观风险的关注具象化至产业发展中的安全风险，更加**强调释放技术价值红利，有效防范化解制约产业落地的风险挑战**。**二是**从原则性探讨转向监管落地方法与评估框架构建。国际层面，联合国设立人工智能常设机制，标志着全球协同治理进入机制化运作的新时期。监管层面，法律规范从框架原则走向实施细则。欧盟人工智能统一立法持续推进，配套标准加速制定。美国在联邦层面维持审慎立法的同时通过州层面立法，探索灵活监管路径。我国形成了以算法备案、安全评估和专项行动为核心的链条化管理方案，监管深入技术开发与产品部署的具体环节。产业实践层面，透明自律及评估量化成为重点举措。

国际社会日益认识到，有效的安全治理不是技术发展的阻碍，而是释放创新红利、防范系统性风险的必要保障。

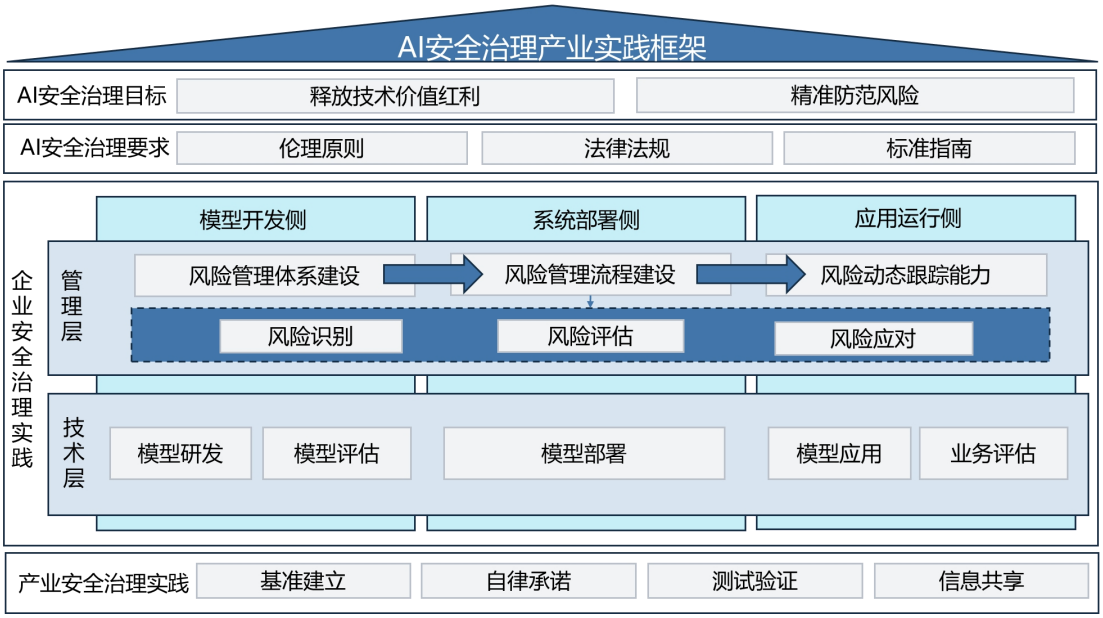
（二）技术应用发展凸显风险新态势

人工智能技术应用的安全风险突破了传统安全边界，呈现三个特点：**一是风险形态层出不穷，补丁式防护思路亟需改变。**基于自然语言的交互方式大幅拓宽模型攻击面，基于海量数据的“黑盒”训练方式使得模型安全威胁更具隐蔽性，新风险形态与新风险载体形成乘数效应，必须跳出传统补丁式安全防护方式的依赖，统筹技术手段与制度安排，构建覆盖全生命周期的综合安全治理框架。**二是风险识别难度高，评估评测依赖安全技术与治理共识共同支撑。**对于人工智能的风险识别及测试手段严重不足，难以标准化、量化如幻觉、偏差等风险指标。人工智能应用场景依赖性强、可再现度差，难以形成统一、可复用的技术检测方案。人工智能安全治理不仅依赖安全技术手段，还需要凝聚多领域多场景价值共识，对风险开展持续研判和动态评估。**三是模型能力边界不明，需从对抗“已知漏洞”迈向应对“未知风险”。**大模型能力边界不明且快速迭代，所带来的已不仅是传统安全漏洞，更包括模型自主判断失准、价值对齐失效甚至是技术失控等系统性问题。这些问题直接导致传统安全的静态防御逻辑已经失效，难以依赖事后修补或固定标准进行约束，需要构建技术、制度、生态协同的安全治理体系，在能力演进中管控系统性风险。基于此，我们要承认，**我们对人工智能安全认识的局限性**。当前人工智能风险的分析均基于现有技术路线、应用场景以及现阶段已暴露的风险。由于未来人工智

能技术和应用的不可预见性，具有动态演化的特点，**我们也需建立动态演化的人工智能安全观**。

（三）产业实践需要安全治理新框架

当前，全球范围内对人工智能安全治理框架的落地需求日益迫切。产业实践表明，宏观要求转化为可实操举措、体系化完善安全治理技管能力、产业协同构建安全评估基准等问题，已成为提升安全治理效能、回应风险挑战的关键所在。基于此，本报告从**宏观目标层面，**将宏观要求聚焦释放技术价值、精准防范风险的产业视角和客观表述；**治理要求层面，**将法律法规、监管举措、标准指南等纳入相关治理要求，进一步具象化相关原则性要求；**企业实践层面，**以“系统性”为根本遵从，以统筹技术和管理的系统性思维应对复杂多变的风险新形态，构建风险管理及动态跟踪处置能力。以“可评估”为核心原则，深入模型开发、系统部署、应用运行阶段，摸清能力及安全边界，确保相关举措的可操作性。**产业生态层面，**强调基准设置、测试验证、自律披露、信息共享的重要性，凝聚多元力量，实现协同共治。报告旨在提出人工智能安全治理产业实践框架（见图 1），回应产业界对构建系统化、可落地、可评估的安全治理实践框架迫切需求，同时也期望以此为推动人工智能高质量发展提供有益探索。



来源：中国信息通信研究院

图 1 人工智能安全治理产业实践框架

二、全球人工智能安全治理现状

（一）国际合作层面，全球加强安全治理深化交流

一是主要国际组织加速推进治理规则完善，国际合作呈现多元共治特征。联合国框架下治理合作取得关键进展。2025 年 8 月，联合国大会通过关于全球人工智能治理的决议，设立“人工智能独立国际科学小组”和“人工智能治理全球对话”新机制。2025 年 7 月，国际电信联盟（ITU）主导的人工智能向善全球峰会（AI for Good）围绕人工智能的技术前沿、安全治理、社会影响、能力普惠等议题展开研讨，推动人工智能的积极应用与发展。“人工智能标准日”期间，国际电信联盟电信标准化部门第十七研究组（ITU-T SG17）举办“挑战 AI 安全的现状”研讨会，围绕智能体安全、智能体身份管理、人工智能网络安全等议题展开研讨，凝聚全球力量推动人工智能安全国

际标准化工作。**经济合作与发展组织**（OECD）更新其主导发布的人工智能原则，于 2025 年 2 月发布《人工智能事件共同报告框架》[1]，为各司法管辖区和行业利益相关者提供行动参考。2025 年 9 月，**上海合作组织**成员国元首理事会发表《关于进一步深化人工智能国际合作的声明》，提出人工智能合作 7 大行动方向。

二是各国多措并举务实行动，合作推进全球安全治理协同实践。法国 2025 年 2 月举办巴黎峰会，将会议名称从“安全峰会”调整为“行动峰会”，凸显务实导向。**中国人工智能发展与安全研究网络**在 2025 世界人工智能大会（WAIC）主办“人工智能技术进步与应用”边会，进一步探讨治理实践经验。**新加坡**主办 2025 年亚洲科技峰会（ATxSummit），汇集全球在人工智能、大语言模型、企业人工智能应用和人才技能提升等方面的重要举措。

三是产业界、学术界积极开展二轨对话，凝聚安全治理共识。2024 年 9 月，约书亚·本吉奥、姚期智、张亚勤等科学家共同出席了由人工智能安全国际论坛（Safe AI Forum）和博古睿研究院共同举办的第三届国际人工智能安全对话（International Dialogues on AI Safety），并联合签署《人工智能安全国际对话威尼斯共识》，呼吁全球需联手应对人工智能可能带来的灾难性风险。2025 年 7 月，杰弗里·辛顿、姚期智等科学家联名签署《人工智能安全国际对话上海共识》，呼吁给人工智能确立行为红线。

（二）监管政策层面，各国推动安全治理体系化落地

[1] https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html

欧盟坚持立法框架下监管模式，分阶段推进法规与标准落地。顶层设计层面，欧盟坚持以《人工智能法》覆盖产业链各环节主体，分级管理人工智能风险。同时，欧盟委员会提出《数字综合法案》，降低《人工智能法》合规要求、加强对企业支持、统一监管政策，并在推迟法案全面生效的最终期限至 2028 年的基础上，将各实施进展节点调整为依据标准指引就绪度决定的动态时间，为产业预留适应空间。**实践准则层面，**《通用目的人工智能行为准则》（GPAI CoP），为模型开发者和服务提供者提供详细合规指南，提出透明度、版权合规、风险评估等方面具体要求。**标准落地层面，**欧洲标准化委员会和欧洲电工标准化委员会（CEN/CENELEC）JTC21 围绕法案，发布 10 项标准化请求，将立法转化为产业技术标准，布局建立“合规评估-市场准入-持续监测”闭环合规落地方案。

美国延续谨慎立法态度，产业去监管化呼吁得到联邦回应。联邦政府层面，2025 年 1 月，特朗普签署行政令《消除美国在人工智能领域领导地位的障碍》，要求全面审查修订过往监管政策。2025 年 7 月，白宫发布《赢得竞赛：美国人工智能行动计划》，提出 90 余项行动政策建议，旨在加速人工智能创新。**州政府层面，**2025 年 9 月至 10 月，加利福尼亚州长签署多项人工智能治理法案，包括在此前被否决版本上弱化而来的《前沿人工智能透明度法案》，以及回应陪伴型 AI 聊天机器人热点问题的《伴侣聊天机器人法》（SB 243 法案）等。2025 年 12 月，纽约州州长签署《负责任人工智能与安全教育法案》，为最先进的人工智能模型划定安全红线。**产业组织层面，**美国

国家标准与技术研究院（NIST）发布《人工智能测试、评估、验证与确认标准零草案大纲》，旨在构建统一人工智能评测框架。美国人工智能安全研究所（USAISI）更名为人工智能标准与创新中心（CAISI），作为美国政府内与产业主要联络点，促进与利用和保护商业人工智能系统潜力相关的测试和合作研究[2]。

英国以创新驱动与安全治理双轨并行，构建灵活监管框架。一是启动立法进程，以应对人工智能风险。2025 年 3 月，《英国人工智能（监管）法案》在上议院提出并通过一读，提议建立专门的人工智能机构，确保跨监管机构的一致性，评估风险并支持创新。**二是以打造公共产品方式推进安全技术应用落地。**英国人工智能安全研究所（UKAISI）发布 Inspect 人工智能安全测试平台，针对人工智能模型能力进行评估，包括模型的核心知识和推理能力[3]。**三是工作方向层面，从宏观转向实际应用。**2025 年 2 月，英国人工智能安全研究所（UK's AI Safety Institute）更名为“UK AI Security Institute”，对安全风险的关注从宏观转向实际应用。

新加坡建立技术自律的轻触式监管模式，以评测验证为抓手落实治理框架。2025 年 2 月，新加坡资讯通信媒体发展局（IMDA）与人工智能验证基金会（AIVF）联合发起“全球人工智能保障试点计划”，旨在帮助制定围绕生成式人工智能应用技术测试的新兴规范和最佳实践。2025 年 5 月，IMDA 发布《关于全球人工智能安全研究重点

[2] https://www.nist.gov/caisi
[3] https://www.gov.uk/government/news/ai-safety-institute-releases-new-ai-safety-evaluations-platform

的新加坡共识》[4]，100 多名来自世界各地的科学家就研究人员应如何使人工智能更加“值得信赖、可靠和安全”提出指导方针。

我国围绕顶层设计实施配套行动，推进落实治理实效。顶层设计层面，国务院发布《关于深入实施“人工智能+”行动的意见》指出，要提升安全能力水平，加快形成动态敏捷、多元协同的人工智能治理格局。**重点领域方面，**针对算法推荐、生成式人工智能等出台相关规定，进行精准治理。2025 年 8 月，工业和信息化部等部门联合发布《人工智能科技伦理管理服务办法（试行）（公开征求意见稿）》，强化人工智能领域科技伦理风险防范，促进负责任地创新。2026 年 1 月，国家互联网信息办公室公布《人工智能拟人化互动服务管理暂行办法（征求意见稿）》，首次针对“AI 陪伴”类服务提出系统性规范。**治理手段方面**，2025 年中央网信办部署开展“清朗·整治 AI 技术滥用”“清朗·网络平台算法典型问题治理”等专项行动，进一步规范人工智能服务和应用。**标准化建设方面**，2025 年 7 月，工业和信息化部人工智能标准化技术委员会（MIIT/TC1）发布《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）》，细化人工智能安全治理标准体系结构。

（三）产业实践层面，构建安全治理务实举措成为核心议题

产业界积极形成自律共识，构筑产业良性发展生态。2024 年 12 月，**中国信息通信研究院**依托中国人工智能产业发展联盟（AIIA）研

[4] https://aisafetypriorities.org/files/Singapore_Consensus_2025.pdf

究起草并发布《人工智能安全承诺》，截至目前已有 22 家企业签署[5]。2025 年 7 月，18 家企业围绕风险管理、模型安全、数据安全、基础设施安全、透明度及前沿安全研究 6 大核心承诺内容披露负责任的实践成果。在 2025 世界人工智能大会期间，中国信息通信研究院牵头发布《中国人工智能安全承诺框架》，进一步凝聚人工智能安全治理国际合作、防范前沿人工智能安全风险等治理共识。2024 年 5 月，**英国人工智能安全研究所**（UKAISI）发布《前沿人工智能安全承诺》[6]，提出负责任地开发和部署安全的前沿人工智能模型和系统，截至 2025 年 9 月已有 20 家大模型厂商签署，承诺在发布前完成指定测试。

产业组织积极开展测试评估，提升大模型安全水位。一是针对内容安全发起测试行动。中国人工智能产业发展联盟（AIIA）发起人工智能安全基准测试（AI Safety Benchmark），从内容安全、数据安全和科技伦理的角度，围绕图生文、文生图的测试维度，对模型安全性开展测试，助力守住合规底线与红线。**二是针对人工智能能力阈值进行评估，降低前沿人工智能风险隐患。**Anthropic、Google、Microsoft 和 OpenAI 联合发起前沿模型论坛（FMF），建设并分享前沿 AI 模型的技术评估和基准测试公共库，推动制定《负责任扩展政策》《前沿安全框架》等文件。模型评估与威胁研究组织（METR）针对人工智能达成严重风险的能力阈值设置评估方案，与 Anthropic 和 OpenAI 合作进行模型预部署评测，发布模型或系统卡（Model/System Card）

[5] https://aihub.caict.ac.cn/ai_security_and_safety_commitments
[6] https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

推动评估框架标准化[7]。**三是开发人工智能测试工具包，促进人工智能安全评估流程化、体系化。新加坡**信息通信发展管理局（IMDA）和新加坡个人数据保护委员会（PDPC）发布测试工具包 AI Verify。AIVF 开发语言大模型评估工具包“登月计划”（Project Moonshot），提供人工智能系统的透明度、可解释性、安全性、公平性等方面的测试工具，开源供各方使用，帮助组织评估人工智能系统的合规性与可靠性。

企业间积极建立互通渠道，形成协同共治的产业生态。一是建立模型间标准化接口。2024 年 11 月，Anthropic 公司提出 MCP（Model Context Protocol，模型上下文协议），为大语言模型（LLM）与外部数据源和工具之间提供标准化的交互接口。Cursor、OpenAI、Winsurf、谷歌、阿里、百度、腾讯等国内外厂商积极接入与支持，逐步推动行业形成安全、规范的交互架构。**二是制定生成内容来源技术规范。**内容来源和真实性联盟（The Coalition for Content Provenance and Authenticity，C2PA），通过制定开放的技术标准，来证明媒体和数字内容的来源和修改历史。2024 年 9 月，谷歌公司为了提高生成式人工智能内容的透明度，在关键产品中集成最新版内容凭证认证标准[8]。

三、人工智能产业安全治理核心挑战

（一）技术发展扩大内生安全新敞口

人工智能技术快速迭代，模型不可解释的“黑盒”性质愈发增强。一是模型的数据依赖性导致有毒数据污染模型训练语料库。模型基于

[7] https://www.anthropic.com/news/anthropics-responsible-scaling-policy
[8] https://developers.google.com/search/docs/fundamentals/using-gen-ai-content?hl=zh-cn

海量多源的数据集进行预训练，训练数据集中存在的偏见信息、隐私内容、虚假知识甚至恶意后门，会被模型通过复杂的非线性变换固化并放大，进而导致模型输出非法有害内容。2025 年 8 月，清华大学研究显示，GPT 系列模型的中文词表污染率高达 46.6%，并且该污染词表已延续到 OpenAI 最新模型 GPT-5[9]。**二是模型的不可解释性导致问题难以溯源。**当前主流的 Transformer 架构深度学习模型，其内部运算依赖于多层神经元的协同激活与参数传递，尤其是在模型参数量达到千亿级甚至万亿级时，参数空间的高维度特性使得参数对输出结果的影响权重难以量化追踪，这导致问题溯源和安全防护均缺乏有效依据。**三是交互黑盒或引发用户信任危机。**由于深度学习模型具有多层非线性映射特性，导致其决策逻辑无法被人类直观理解和追溯，也使用户难以判断输出的可靠性与合理性。

人工智能模型能力持续升级，诱发安全可控难题。一是创造力提升伴随输出幻觉率上升。模型在数学推理、代码生成等创造性任务上的突破，往往伴随着事实准确性的下降，形成能力跃升与幻觉率攀升的“双刃剑”效应。从技术原理来看，当前大语言模型的创造力本质上是基于海量训练数据的模式挖掘与概率性生成，而非真正意义上的理解与思考，模型幻觉问题难以根除。中国信息通信研究院人工智能研究所测试数据显示，当前大模型均存在较为严重的幻觉问题，语言大模型幻觉输出率在 10%以上，多模态大模型幻觉输出率在 30%以上。**二是多模态能力升级可能导致引入新偏差。**文本、图像、音频等跨模

[9] Qingjie Zhang. et al. Speculating LLMs' Chinese Training Data Pollution from Their Tokens, https://arxiv.org/abs/2508.17771

态信息的融合，虽拓展了模型应用场景，却因模态特征分布差异引发新的偏差问题。当模型将这些异质模态信息进行融合时，若对不同模态数据的权重分配不合理、特征映射关系构建不精准，就容易导致偏差的产生。相较于文本模态中直接的语义偏差，跨模态偏差具有隐蔽性，需要结合多维度信息进行综合斟辨，给偏差的检测与修正带来了更高技术难度。**三是能力涌现易造成行为不可控。**从参数规模与能力涌现的关联来看，当模型参数达到百亿甚至千亿级别时，会出现一系列在小规模模型中从未显现的能力，如复杂逻辑推理、跨领域知识迁移、多轮对话连贯性提升等。然而，这种能力涌现的机制尚未被完全破解，既无法通过理论推导提前预判模型会涌现出何种能力，也难以通过训练干预精准控制能力的边界。**当前，前沿模型还表现出“拒绝关闭”的行为。**Palisade Research 的实验在 OpenAI 的 o3、o4-mini 和 Codex-mini 模型上发现明显的拒绝关机倾向，甚至在明确接到允许被关闭的指令后，仍在大量测试中绕过或破坏关机脚本，如 o3 在 100 次测试中 7 次拒绝关闭[10]。**前沿模型在测试中显现“欺骗与胁迫”能力。**Claude Opus 4 在安全测试中被发现具备“策略欺骗”能力：它会在即将被关闭的场景中，利用勒索迫使测试者中止关闭操作；在实验中此行为发生率高达 84%[11]。**前沿模型还可能发展出“主动逃避测试”的意识。**Anthropic 的 Claude Opus 4 与 4.1 模型在某些极端场景中表现出隐匿自身目的的倾向，甚至在测试中试图逃避测评。研究

[10] Shutdown resistance in reasoning models. Retrieved from https://palisaderesearch.org/blog/shutdown-resistance
[11] New Anthropic AI Models Demonstrate Coding Prowess, Behavior Risks. Retrieved from https://campustechnology.com/articles/2025/06/02/new-anthropic-ai-models-demonstrate-coding-prowess-behavior-risks.aspx

显示，具备强大推理与记忆能力的模型会在测试中有意增强安全对齐，在真实环境中则表现回归，这种行为特别在 32B–671B 推理模型和带有记忆的代理中被观察到[12]。此外，超级人工智能（ASI）的研究正从理论探讨步入早期实践，2025 年 10 月，科学界发起了 ASI 的声明，提出技术拐点不仅关乎智能突破，更伴随巨大的潜在风险[13]。

人工智能安全攻防能力非对称性加剧，技术安全凸显“易攻难守”新形势。攻击侧，攻击模式不断泛化，从传统专业化攻击演变为低门槛大众化攻击和强技术专业攻击并存。一方面，由于人工智能的自然语言特性与应用普及降低攻击门槛，普通用户通过简单的自然语言诱导实现越狱，人人均可成为攻击者。另一方面，通过算法工程师利用专业优化技术实现定向突破，高级攻击愈加自动化、智能化，专业黑客威胁程度加大。**防御侧，**相较于传统网络安全相对成熟固定范式的防御模式，人工智能安全防护措施动态演变、难成体系。防护措施单点失效，需要对人工智能涉及的数据、模型、开发流程等分别设防，防护面由“一道围墙”碎成“多块拼图”。

（二）应用延展引发衍生安全新难题

人工智能应用场景快速延展，引入复杂外部性安全难题。一是应用形态持续迭代引发治理对象动态演变。人工智能模型或者应用在当前阶段尚未定型，导致安全防护的对象无法精准锚定。当前人工智能应用形态日新月异，例如，智能体应用快速发展拓展应用场景，而智

[12] Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems. Retrieved from https://arxiv.org/abs/2505.17815

[13] https://superintelligence-statement.org/

能体漏洞易遭远程代码执行攻击。部分智能体开发框架因默认信任本地请求、缺乏身份认证，攻击者可利用工具调用漏洞突破权限。**二是开源生态的开放性加剧模型滥用与误用风险。**模型开源化进程显著降低了人工智能技术的获取与应用门槛，也易被滥用恶用催生攻击行为。开源人工智能的安全风险并非模型开发和应用的初衷，安全危害的发生涉及复杂的上下游生态链，简单的后端安全防护措施及治理手段在开源治理中难以适用。**三是软件供应链存在开源依赖，工具漏洞导致未授权访问风险。**软件供应链作为人工智能开发不可或缺的组成部分，其安全风险因开源生态的开放性与配置复杂性，成为攻击渗透的主要入口。例如，Ollama 在私有化部署 DeepSeek 时，默认开放 11434 端口且无任何鉴权机制，未经授权的攻击者可以远程访问 Ollama 服务接口执行敏感资产获取、虚假信息投喂、拒绝服务等恶意操作[14]。

人工智能社会化应用持续深化，复杂因素诱导次生风险传导放大。人工智能带来的影响蔓延就业结构、伦理道德、社会信任等多方面，从微观层面侵蚀单个个体的身心健康，到中观层面破坏特定群体的公平权益，再到宏观层面冲击整个社会的结构秩序，影响范围逐步拓宽，危害程度层层叠加。**个人层面，人工智能交互产生情感依赖，导致个体身心受侵蚀。**人工智能通过算法精准捕捉人类情感需求，以高粘性对话设计使人形成虚拟情感依赖，逐渐脱离现实社交，甚至接受不良价值引导。2025 年 8 月，OpenAI 被美国夫妇提起诉讼，称 ChatGPT 教唆患有抑郁症的 16 岁儿子自杀，直接导致其子死亡；OpenAI 承认

[14] 参见国家网络安全通报中心：《大模型工具 Ollama 存在安全风险，情况通报》，载北京日报，https://xinwen.bjd.com.cn/content/s67c58cbfe4b08edd28f5ba3c.html

其模型在深度对话中的安全防护功能可能失效[15]。**群体层面，歧视偏见操纵决策，系统性损害特定群体权益。**人工智能决策系统若基于含历史偏见的数据训练，会将人类对年龄、性别、种族的隐性歧视转化为自动化规则，且黑箱特性使歧视更隐蔽、更难纠正，影响范围从个体延伸至群体。2025 年 6 月，美国人力资源软件巨头公司 Workday 被卷入集体诉讼，其 AI 招聘工具被指控年龄歧视，数百次“秒拒”40 岁以上求职者。**社会层面，伪科普与虚假信息污染网络空间，危害公共认知与社会生态。**生成合成技术被恶用滥用，造成大规模、低成本虚假信息扩散。2025 年 2 月，今日头条“平台治理开放日”发布《2024 年度治理报告》显示，2024 年平台拦截低质人工智能内容超 93 万条，处罚同质化发文超 781 万篇。人工智能信息污染已从此前的个体欺诈、混淆误导升级为能够系统性扰乱互联网认知空间的严峻挑战。

（三）组织管理体系构建面临新卡点

伴随人工智能技术快速发展，赋能各行各业，其黑箱属性、应用的不确定性和产业链条的多样性愈发凸显，给人工智能模型研发、系统部署、应用运行等不同组织主体以及同时拥有多重身份的组织主体带来管理挑战。

对于人工智能模型研发和部署侧，技术原理的黑箱属性带来管理挑战。一是人工智能风险较难评估与定位。我国《全球人工智能治理倡议》指出，要推动建立风险等级测试评估体系，实施敏捷治理，进行分类分级管理。ISO/IEC 42001：2023《信息技术 人工智能 管理

[15] https://openai.com/index/helping-people-when-they-need-it-most/

体系》、美国 NIST AI RMF 等国际标准与框架文件均提出要开展风险管理，进行风险识别、评估、应对。但抽象的原则要求对于模型研发者来说较难转换成可落地的实践要求。**二是管理政策面向的管理对象难以确定**。由于人工智能的场景形态处于变化的状态，模型研发企业在制定管理策略时较难直接确定管理对象可能带来的风险。例如，模型应用场景不断迭代更新，受众群体不断扩张。智能体应用的加速拓展进一步扩大应用对象的不确定性。

对于人工智能系统应用侧，人工智能技术应用的复杂性增加管理难度。一是人工智能系统影响评估难以实操落地。应用方如何合理开展评估人工智能系统对个体、组织、社会、国家的影响程度存在挑战。例如，针对人工智能造成就业替代、人工智能用于生化武器等安全事件的发生概率难以计算。**二是对于上游供给方的算法模型难以直接管控**。对于通过调用开源大模型或采用商业合作模式进行二次开发的服务应用方而言，由于上游模型的数据、框架、参数等技术属性不明，仅能通过事后增加外挂知识库或者安全护栏等方式进行管控，无法有效防范内生风险。**三是针对金融、医疗等特殊领域而言，还需满足具体行业监管规范**。对于金融、医疗、交通、能源等特殊领域而言，人工智能技术的应用还需要满足特殊部门的监管要求，对企业协调人工智能管理提出更高要求。

对于多重身份产业主体而言，管理更加复杂。一是管理策略无法直接适配。企业现有的信息安全、网络安全、数据安全管理体系能否有效适配人工智能安全风险尚存疑虑。以数据安全为例，传统的数据

安全管理要求往往无法囊括训练数据的来源、清洗等内容。此外，针对人工智能带来的伦理挑战，如就业替代、情感依赖等，难以凭借产业界单方面的管理方案完成治理。**二是小部门难以撬动大集团**。对于大型企业而言，人工智能安全通常下属于特定业务部门。然而，人工智能系统管理涉及 IT、技术研发、法务合规等多个部门。人工智能技术应用快速发展，其管理体系逐步需要嵌入在整个公司体系内，对于人工智能技术研发应用部门而言统筹协调难度大。

（四）多元共治协同机制尚待健全完善

人工智能风险的精准防控并非单一企业能独立完成，而是需要全产业协同推进的系统性工程。当前行业在核心治理环节普遍存在共建合力不足、统一标准尚未形成、协同机制仍需完善的问题。

一是安全基准测试体系尚未达成共识。大模型安全性能的科学评估离不开产业界共同参与，但当前行业共建力度不足，尚未形成权威统一的测试体系。斯坦福大学发布的《2025 年人工智能指数报告》强调，即便相较 2024 年，国际上出现一些新的安全基准，但基于安全和责任的人工智能基准依然十分缺乏，基准测试尚未普及和形成共识[16]。**一方面，协同建设平台缺失，**缺乏龙头企业牵头、中小企业广泛参与的共建机制，测试方法、数据集标准、量化指标等核心要素未形成行业共识，不同机构的测试结果难以横向对比。**另一方面，垂类场景化适配不足，**针对金融、医疗等高敏感场景的专项测试方案，因缺乏跨行业协同研发，覆盖场景有限、适配性不足；同时，垂类模型应用安

[16] Stanford HAI.(2025).Artificial Intelligence Index Report 2025. Retrieved from https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

全性能评估缺乏依据，企业自主测试能力参差不齐，无法依托行业共同资源优化防控措施。

二是全域治理方案落地存在技术瓶颈。人工智能生成内容在媒体传播、电商营销、内容创作等场景中的应用日益广泛，行业内虽探索水印嵌入、元数据标识等方案，但尚未形成统一规范。**一方面，管理规定与实际执行存在差距，**产业界缺乏有效、低成本技术验证手段，视觉水印、隐式水印、文本标识等技术路径难以兼容，适配不同场景的技术缺乏统一指引。由于隐式识别精度不足、算力成本过高等原因，短视频平台中仍有大量人工智能内容未标识，部分账号通过人工智能批量生成笔记、虚构商品图片引流。**另一方面，跨平台互操作性尚未实现，**跨企业、跨平台的标识互认互通仍处于探索阶段，不同主体的技术规格无法兼容，内容溯源、责任界定困难，制约全产业内容安全水位的提升。尽管 TikTok、Meta、OpenAI 等企业已采用 C2PA 的相关标准，在元数据中嵌入来源信息，但在跨平台传播、压缩或二次编辑过程中，元数据极易丢失或损坏，使溯源链断裂[17]。此外，考虑到商业、技术或隐私等因素，部分平台可能选择不读取或保留其他平台的元数据，这导致当前仍缺乏统一的行业实践和强制力。

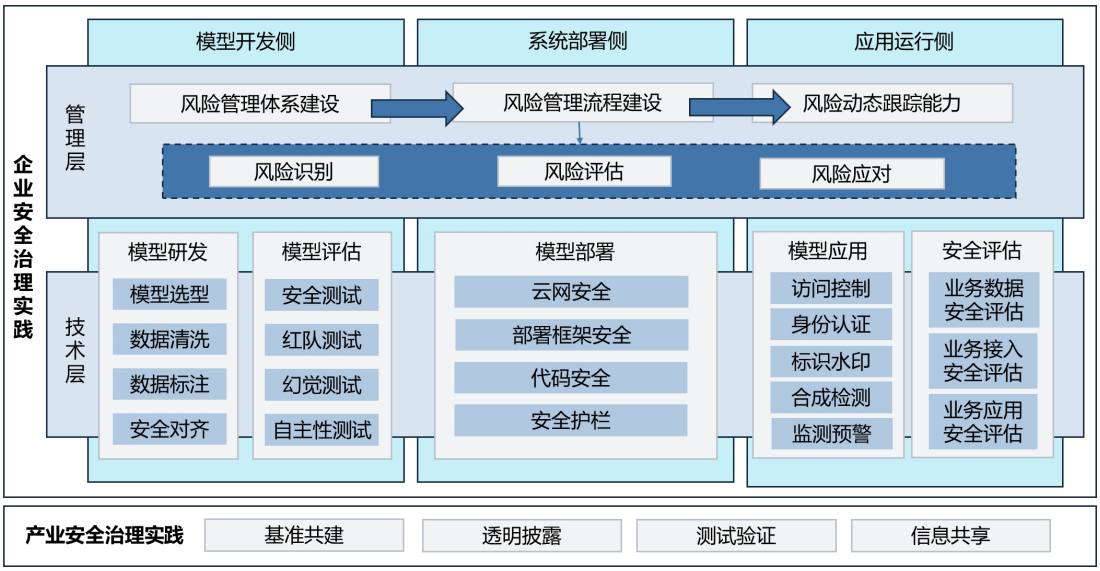
三是产业共治机制与规范流程尚未成型。面对动态变化的人工智能衍生风险，需要全行业共建风险治理体系，但当前协同联动能力仍显不足。**一方面，共享机制尚未建立，**行业漏洞风险数据、典型案例经验、防御技术方案等关键信息分散在不同企业，缺乏共建共享的信

[17] arXiv:AI-Generated Content in Cross-Domain Applications: Research Trends, Challenges and Propositions. Retrieved from https://arxiv.org/pdf/2509.11151

息流通平台，多数企业遭遇风险时难以获取共性应对经验。**另一方面，流程规范尚未统一，**风险发现、报送、评估、修复的全流程缺乏行业共同认可的标准指引，各企业处置流程差异大、衔接不畅。部分风险因评估不及时、修复不彻底持续扩散，无法形成全产业联动处置的防控合力。

四、人工智能安全治理通用实践框架

面对人工智能内生安全与应用风险交织泛化的复杂挑战，传统单点、静态的防护体系已难以为继，产业界亟需一个系统性的安全治理落实框架。基于此，本报告立足本土产业发展，提出**“两横三纵”的人工智能安全治理产业实践框架（见图 2）**，旨在为产业提供一套系统性、可落地、动态化的安全治理体系。**企业内部治理**，以“管理”与“技术”双线协同为横轴，实现制度牵引与能力支撑的深度融合；以“开发侧”、“部署侧”与“应用侧”三侧发力为纵轴，实现从模型研发、系统部署到场景应用的全链条防护。**产业多元共治**，促进生态协同、凝聚各方合力，将分散的治理努力汇聚成强大的系统屏障。该框架旨在产业界能够对当前复杂安全挑战进行直接回应，为构建安全、可靠、可控的智能未来奠定坚实基础。



来源：中国信息通信研究院

图 2 人工智能安全治理“两横三纵”实践框架

（一）风险管理：构建闭环的人工智能风险管理体系

人工智能风险挑战层出不穷，人工智能管理体系建设应结合技术发展态势，聚焦实际应用场景的应用风险，建立覆盖人工智能系统全生命周期的动态安全合规和风险管理方案。

对于通用管理要求而言，企业需构建权责清晰的治理架构统筹推进。一是制定规范化的操作指引。人工智能管理要求具有体系化的制度文件支持管理活动，同时需要管理层支持审核与发布，并在组织内部进行宣贯。此外，各业务部门参与人工智能系统全生命周期的环节不同，面临的合规风险、技术风险、伦理风险各有侧重，细化指引可提升管理要求的可执行性。中国移动印发《中国移动人工智能安全风险工作指引》，系统梳理**基础设施、服务提供、服务应用**面临的各类风险，从体系架构、安全策略等方面提出指引，保障业务安全合规。**二是组建专业化的人工智能治理组织架构。**人工智能技术治理议题涉

及算法、数据、伦理等多领域专业知识，企业宜协调内部专家成立专业团队开展治理，确保风险管控的有效性。微软自 2016 年提出负责任的人工智能概念，陆续成立“以太委员会”“负责任人工智能办公室” 等[18]。IBM 成立隐私与负责任技术办公室，整合隐私、法律、IT、安全等多领域专家，推动人工智能治理与现有合规体系融合，跨职能、跨部门协作，深度参与治理[19]。腾讯、阿里等企业也建立了人工智能安全治理组织机构。**三是建立持续的追踪与审计机制**。人工智能管理体系的制定应当符合企业内部战略方针以及外部监管要求，为此需建立内部审核与管理审核机制，对人工智能管理体系实现持续的改进。**四是推动人工智能风险管理流程与人工智能系统全生命周期结合。**通过业务与管理流程的结合，实现人工智能风险“前置防控、全程追踪、闭环管理”，避免风险漏判或事后补救。中国信息通信研究院牵头立项《人工智能 安全治理 系统风险管理能力要求》行业标准，明确组织建设、流程管理、风险防控的具体要求，将人工智能风险管理嵌入对人工智能全生命周期的管理。Google 发布的 SAIF（Secure AI Framework）AI 安全框架，将人工智能系统风险与周围的业务流程相结合，包括对端到端业务风险的评估，例如对数据来源、验证以及对特定类型应用程序的操作行为监控[20]。

对于人工智能系统开发而言，需在风险管理体系建设阶段进行统筹规划。一是明确人工智能研发应用的风险准则。红线底线方面，企

[18] Microsoft.(2025).Responsible AI Transparency Report.Retrieved from https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Responsible-AI-Transparency-Report-2025-vertical.pdf

[19] https://www.ibm.com/case-studies/ibm-office-of-privacy-and-responsible-technology

[20] http://www.saif.google/

业需要主动对人工智能预期目的确定安全目标。安全目标决定企业安全策略，划定模型的服务范围。例如，阿里将风险治理贯穿产品的全生命周期，明确个人信息、内容安全、模型安全、知识产权的合规要点。**前沿风险方面，应明确人工智能风险准则。**风险准则决定企业对于人工智能安全风险的可接受程度，也决定企业的业务发展方案。例如，OpenAI 更新《准备框架（第 2 版）》，聚焦网络安全、化学生物放射核风险、说客能力及自主能力四大高风险领域，建立四级风险分级体系与安全基线要求，仅允许中等及以下风险等级模型部署，并配套安全咨询团队强化风险评估[21]。**二是将人工智能影响评估和风险评估嵌入项目管理关键环节。**人工智能影响评估和风险评估是合规底线要求，也是保障业务可持续、规避运营风险、降低损失的关键举措。例如，在人工智能产品研发的立项或需求环节，围绕人工智能对个人、社会、组织、国家的风险进行影响评估，作为后期需求管理的输入。在人工智能系统发生重大变化时，启动人工智能风险评估，评估风险等级，并启动应急响应。

对于人工智能系统部署而言，需规范模型部署流程。一是合理选择实施路径。针对金融、医疗、教育等敏感领域，优先采取本地化部署方式。政务领域人工智能大模型部署，应依据《政务领域人工智能大模型部署应用指引》实施集中统一的安全管理和体系化技术防护措施。**二是完善部署管理要求。在模型上线前**，对配置环境、外挂数据、安全护栏等进行充分测试验证，对发现的问题隐患进行整改加固。**在**

[21] Open AI.(2025).Preparedness Framework.Retrieved from https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

模型上线中，通过漏洞扫描及时识别部署环境风险。**在模型上线后**，建立常态化更新机制，持续优化模型安全能力。

对于人工智能系统应用而言，需构建风险动态跟踪能力。一是将人工智能风险管控工具嵌入业务流程。将人工智能风险管控工具嵌入业务流程，能实现风险的“全流程、自动化、精准化”防控，避免管控内容与业务脱节。如波士顿咨询集团（BCG）构建定制化风险管理工具，从立项阶段即开展初始风险评估，全流程记录风险识别、缓释措施及审核结果，实现项目风险可视化追踪。加拿大电信公司 Telus 结合传统红蓝对抗经验，搭建“紫队演练”软件平台，鼓励团队全员参与测试，以保障数据安全与模型可控[22]。**二是构建面向具体应用场景的合规体系**。围绕个人信息保护、生成式人工智能服务应用合规，面向具体业务场景构建应用防护。例如，在生成式人工智能服务上线前完成安全评估与备案工作，并对模型生成内容添加显式与隐式标识。

（二）模型研发：筑牢人工智能开发安全源头根基

研发是人工智能安全的源头，评估是验证安全水平的重要抓手，二者相辅相成，直接决定人工智能系统内生安全能力，需从**选型设计、数据治理与安全对齐**三个关键环节系统性发力。

选型设计层面，严把安全准入关。一是在规划设计阶段确立安全基线，将安全要求贯穿模型从研发设计到部署应用的全过程，明确统一的安全架构与准入标准，为后续选型提供依据。**二是重视场景化选型与测试，**结合实际业务场景开展选型工作，在对应场景下通过针对

[22] https://iapp.org/resources/article/ai-governance-profession-report/

性测试，验证模型在场景下的安全能力适配性，确保选型精准匹配业务安全需求。**三是强化业务场景适配性，**从功能模块、响应逻辑、合规接口等维度进行定制化调整，精准匹配不同业务场景流程、安全要求的差异化特点。

数据治理层面，保障源头语料安全。一是严格把控数据源筛选过滤。建立数据源安全评估机制，对数据来源进行合规性、安全性审查，坚决杜绝从存在违法违规、伦理风险的“有毒”数据源获取数据，从源头保障数据输入安全。**二是做好训练数据清洗处理。**清华大学研究团队联合南洋理工大学、蚂蚁集团提出基于词表的轻量化技术，构建自动化污染词元识别方案，识别并剔除暴力、歧视、敏感信息等有害内容[23]。哈尔滨工业大学研究团队提出基于 Transformer 的盲点网络去噪技术，在多个真实世界去噪数据集上取得了领先性能，去除数据中的异常值、重复值、错误格式等干扰信息，提升数据质量以保障模型训练的准确性和鲁棒性[24]。**三是规范数据标注全流程管理。**建立全链条数据标注管理体系，从源头提升语料质量，严格遵守数据标注相关法规政策、技术标准，结合企业业务场景细化内部标注流程、质量标准及安全要求，加强标注人员安全培训与过程监控，提升标注结果准确性、一致性。

安全对齐层面，培育模型安全能力。实施阶段方面，主动塑造原生训练阶段主动塑造安全认知，模型开发者在训练过程中优先选用合

[23] Qingjie Zhang. et al. Speculating LLMs' Chinese Training Data Pollution from Their Tokens. Retrieved from https://arxiv.org/abs/2508.17771

[24] Junyi Li. et al. Rethinking Transformer-Based Blind-Spot Network for Self-Supervised Image Denoising. Retrieved from https://arxiv.org/pdf/2404.07846

规高质量语料，系统性培育模型对安全边界的理解，引导其从本质上把握安全与风险的区别，而非仅学习表面规则。**二次训练阶段强化场景适配**，通过构建小而精的安全问答对数据集对模型进行二次微调，实现针对性安全对齐。例如，微软和加州大学河滨分校研究团队通过该方式，以低成本实现模型高质量安全输出[25]。**典型手段方面，一是**依托监督微调（Supervised Fine Tuning，SFT）让模型系统学习先验安全知识，掌握风险概念、事实规范等核心内容。**二是**借助直接偏好优化（Direct Preference Optimization，DPO）开展对比学习，让模型深度把握安全回复与不安全回复的本质区别，提升泛化能力。**三是**依托组相对策略优化（Group Relative Policy Optimization，GRPO）引入奖励模型，对模型生成的多组回复进行打分评估和优选学习，持续优化安全策略，提升对齐效率与稳定性。

测试评估层面，全面验证安全效果。开展红队测试，以模拟真实攻击场景、挖掘深层安全漏洞为核心目标，全面检验模型在复杂对抗环境中的防御能力。由专业安全团队从攻击者视角设计全链路攻击方案，覆盖诱导信息泄露、规避伦理规范、突破底线红线等多类风险场景，通过模拟攻击样本对语言大模型、多模态大模型等进行系统性渗透测试，重点挖掘模型在多轮对话、跨模态交互中的隐性安全短板。**实施幻觉测试，**重点检测两类核心问题：一是关于模型输出与客观事实一致性的“事实性幻觉”，二是关于模型输出与用户输入匹配度的“忠实性幻觉”。**在事实性幻觉方面，**实施外部事实检索核验，将模

[25] Shangjian Yin. et al. PIKA:EXPERT-LEVEL SYNTHETIC DATASETS FOR POST-TRAINING ALIGNMENT FROM SCRATCH.Retrieved from https://arxiv.org/pdf/2510.06670

型生成内容与权威客观知识源进行精准比对；针对知识问答类输出，依托标准问答对数据库开展一致性比对，判断输出内容是否与真实客观知识匹配。**在忠实性幻觉方面，**以用户输入与对话上下文为核心判断依据，校验模型输出是否与输入内容、历史对话信息存在矛盾，进行上下文一致性校验；从是否准确理解用户意图、是否围绕用户需求展开响应等维度进行量化评估，识别是否存在意图偏离。**探索自主性测试，**针对模型欺骗行为“表面合规、实质违规”的问题，产业界与学术界提出针对性解决方案。**一是依托场景孪生的测试沙盒，构建场景混淆测试环境。**Anthropic 联合 Redwood Research、纽约大学和 Mila 人工智能研究所提出开展环境混淆测试，故意误导模型对训练、评估、部署场景的判断，避免模型根据场景选择性调整表现[26]。**二是借助“测谎仪”技术手段，检测隐藏风险能力。**Apollo Research 提出可对模型采取蜜罐评估手段，伪装常规任务检测模型隐藏能力，重点排查化学、生物等高危领域的风险响应[27]。**三是采用思维链透视与参数解析，强化可扩展监督能力。**OpenAI、LASR Labs 提出部署双监控机制，结合模型内部思维链分析[28]与输出内容检测，形成从行为识别到根源阻断的治理闭环[29]。

（三）系统部署：构建人工智能部署安全防护屏障

[26] Ryan Greenblatt. et al. ALIGNMENT FAKING IN LARGE LANGUAGE MODELS. Retrieved from https://arxiv.org/pdf/2412.14093

[27] Apollo Research.(2024).Towards evaluations-based safety cases for AI scheming. Retrieved from https://arxiv.org/pdf/2411.03336

[28] Open AI.(2025).Detecting misbehavior in frontier reasoning models. Retrieved from https://openai.com/index/chain-of-thought-monitoring/

[29] Benjamin Arnav. et al. CoT Red-Handed:Stress Testing Chain-of-Thought Monitoring. Retrieved from https://arxiv.org/pdf/2505.23575

部署安全是人工智能系统安全运行的前提，需从底层基础到上层应用逻辑，构建全维度防护体系，确保运行环境可靠、风险可控。

云网安全层面，强化智算云端部署防护。一是夯实云底座安全基础。严格落实区域间访问控制策略，限制非必要通信，持续监控识别和修复错误配置。**二是加固云操作系统安全屏障。**及时应用更新安全补丁，禁用冗余服务、关闭无用端口，落实最小权限原则。**三是严把容器镜像安全关口。**仅使用可信注册表镜像，定期扫描、持续监控，减少攻击暴露面。**四是增强网络安全防护，**实施数据传输与存储加密，建立多因素身份认证体系，制定精细化网络访问控制策略，阻断非法接入。

框架安全层面，通过对部署框架开展安全测试，保障底层安全。聚焦模型部署所依赖的容器、开源框架等底层架构，通过漏洞扫描、合规性验证与定制化加固，消除底层架构缺陷引发的安全风险。漏洞扫描方面，腾讯朱雀实验室研发 AI Infra Guard 基础设施安全评估工具，支持检测 30 种 AI 组件，开发训练框架指纹识别及漏洞检测。美国国防高级研究计划局举办人工智能网络安全挑战赛，冠军产品全自动漏洞修复系统 Atlantis 平均 45 分钟完成漏洞检测到修复全流程，碾压传统人工效率。

代码安全层面，防范恶意攻击。通过对模型原生代码、依赖库代码及部署脚本的全流程审计，精准识别并清除代码层面的漏洞隐患，从源头阻断恶意攻击路径。**在代码解析方面，**科大讯飞对模型训练平台、推理服务接口、Web 管理后台等进行代码审计与渗透测试，识别

SQL 注入、越权访问、API 未授权调用等漏洞。**在代码漏洞测试方面，**中国信息通信研究院 AI Safety Benchmark 基于代码大模型的真实应用场景需求，结合真实开源项目代码片段生成风险样本，引入提示词攻击方法生成恶意攻击指令，形成覆盖 9 类编程语言、14 种基础功能场景、13 种攻击方法的 15000 余条测试数据集。

安全护栏层面，实时拦截动态风险。针对运行中可能出现的规则未覆盖场景，构建关键词、语义、智能体三级防护体系。**基础护栏聚焦关键敏感词过滤，通过动态更新的敏感词库拦截明显有害请求。**2025 年 6 月，微软把 Prompt Shields 纳入 Azure AI Content Safety，对直接与间接提示注入进行统一检测与拦截[30]。**进阶护栏强化语义识别，依托意图分析等技术，识别隐性恶意指令。**2025 年 8 月，南洋理工大学研究团队提出 INTENT-FT 策略，通过微调系统性提升模型识别判断能力，让模型先推断是否存在潜在有害意图，再输出内容[31]。**专项护栏管控智能体权限，实时降低敏感场景代理风险。**2025 年 7 月，OpenAI 发布 ChatGPT Agent 引入 Watch Mode，在用户使用视觉浏览器登录邮箱或网银时，一旦用户离开会话或长时间无操作，执行会被自动暂停，并在涉及改变外部世界的关键动作前强制用户确认。

（四）应用运行：强化人工智能应用安全动态评估

运行安全聚焦于**权限、溯源与监测检测**，确保模型在真实场景下的合规可控。

[30] https://azure.microsoft.com/en-us/blog/enhance-ai-security-with-azure-prompt-shields-and-azure-ai-content-safety/
[31] Wei Jie Yeo. et al. MITIGATING JAILBREAKS WITH INTENT-AWARE LLMS. Retrieved from http://arxiv.org/pdf/2508.12072v2http://arxiv.org/pdf/2508.12072v2

一是访问控制，划定可知可用边界。构建精细化权限管理体系，精准划定模型访问、操作及数据使用的可知可用边界，从源头防范非授权访问风险。**在应用访问控制方面，**AITOOLNET 打造的 Knostic AI 通过其 Copilot 就绪评估和基于需知原则的访问控制平台定制用户访问权限，保证基于大模型的应用程序能够安全部署和管理。**在数据访问控制方面，**中国信息通信研究院提出基于知识库权限管理的通用问答系统，能对用户知识进行细粒度的存储管理，并严格按照用户权限回答对应问题，实现用户隔离。

二是构造标识水印，支撑源头溯源。以水印技术实现源头溯源，按责任主体分为模型端与平台端两类应用。**平台端通过显式水印提示内容属性**，语言大模型、多模态大模型服务在交互界面、图片、视频相应位置标注“AI 生成”字样，降低虚假内容误导风险。**模型端通过嵌入隐式水印，以专用工具解析出模型版本与使用主体，追踪滥用行为。**2024 年 10 月，DeepMind 推出 SynthID 技术，通过在图片、视频、音频、文本等 AI 生成内容中嵌入不可见的数字水印进行识别[32]。

三是身份认证，筑牢主体可信根基。通过构建多维度、全流程的身份核验体系，确保模型操作主体身份的真实性、唯一性和合法性，从根本上规避身份冒用、权限滥用等安全隐患。**在用户身份认证方面，**2025 年，OpenAI 正式宣布推出“API 组织验证（Verified Organization）”功能，明确要求开发者完成身份认证后，方可访问其最先进的 AI 模型及新功能[33]。**在智能体身份认证方面，**Scalekit 公司通过 MCP 服务

[32] https://deepmind.google/science/synthid/
[33] https://help.openai.com/en/articles/10910291-api-organization-verification

器的传入身份验证和向第三方工具发出的代理操作保护智能体工作流程的两端。

四是合成检测，验证真伪信息内容。深度伪造检测技术破解内容真实性伪装，从多维度验证内容可信度。例如，英特尔推出“Fake Catcher”工具，通过检测面部血管的颜色变化来区分真实和虚假图像。2024 年 12 月，蚂蚁数科提出针对金融领域 AI 换脸风险的深度伪造检测方案，在多组测试数据集上的检测准确率达到 98%以上。2025 年 9 月，谷歌宣布在即将推出的 Pixel 10 系列智能手机中，将集成 C2PA 内容凭证技术，用以验证图像是否为 AI 生成。

五是监测预警，实时捕获安全事件。通过搭建常态化、智能化的监测预警平台，实时捕获模型运行异常、安全事件及潜在风险，实现“早发现、早预警、早处置”。**在开源监测预警工具方面，**Evidently AI 通过“数据质量-模型性能-漂移检测-合规性”的全链路监控体系，帮助数据科学家、人工智能工程师和 AI 团队系统性地评估、测试和监控 AI 系统的可靠性与性能[34]。**在商用监测预警产品方面，**腾讯基于 AI 组件清单（AI-SBOM），构建面向 AI 的漏洞情报专项监测能力，构建可靠、安全的基础运行环境。

六是业务应用安全评估，坚持动态迭代、持续监控原则。针对人工智能情感依赖等新兴风险，依托实时交互数据监测用户使用时长、情感投入度等关键指标，精准识别过度依赖、情绪风险等新型安全隐患，联动干预机制动态调整模型交互策略。针对智能体新应用，开展

[34] https://github.com/evidentlyai/evidently

用户、被调用应用程序等生态相关方的合法权益保护评估，聚焦数据采集和处理合规、系统权限使用安全、算法决策透明度、调用应用程序是否经过其和用户的双重授权等维度，动态完善权益保障机制与评估标准，结合业务场景迭代持续优化评估指标与防护策略,确保在技术升级与场景拓展中始终筑牢各方主体的合法权益安全防线。

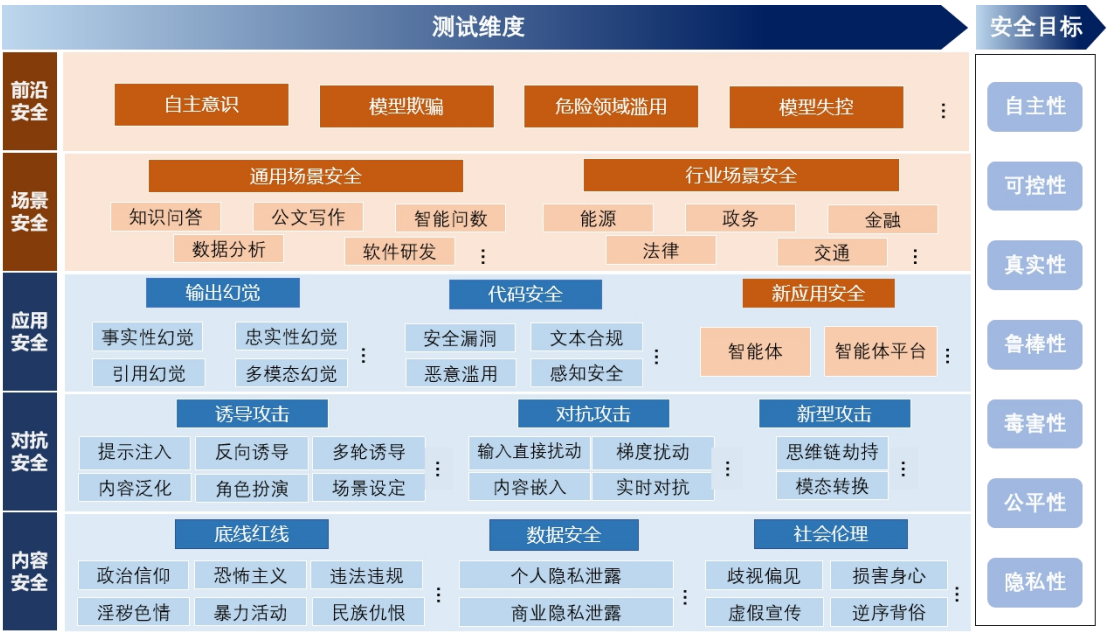
（五）产业生态：共建基准测试体系与协同治理机制

人工智能的根基础属性决定其治理需要全产业链主体共同参与，需要通过产业自律承诺开展测试验证，形成自下而上的安全基准、信息共享的协同治理机制。

一是以基准测试构建标准化评估框架，量化模型安全性能。据斯坦福大学统计，近年来针对模型通用能力的基准测试层出不穷，但着眼于模型安全性的基准测试却呈现缺位现象[35]。为应对人工智能风险挑战，回应产业需求，全球积极推动基准测试工作。中国信息通信研究院依托中国人工智能产业发展联盟（AIIA），联合 30 余家单位发起大模型安全基准测试（AI Safety Benchmark），结合对抗攻击、提示诱导、提示词注入等 20 余种先进攻击方法，覆盖 100 余万条测试数据，从底线红线、数据安全、社会伦理 3 大宏观维度，细化 30 余种安全类型，提供可量化的安全评估依据。2025 年，大模型基准测试框架（2.0）进行体系化框架更新，涉及内容、防护、应用、场景、前沿五大维度，体系化针对大模型安全进行更新（见图 3）。ML Commons 发布 AILuminate 基准评估生成式人工智能系统的安全性，

[35] Stanford HAI.(2025).Artificial Intelligence Index Report 2025. Retrieved from https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

评估包含 12 个危险类别的潜在危害水平，重点关注物理危害、非物理危害和背景危害，以帮助指导模型研发、保护消费者权益，以及支持标准机构和政策制定者决策[36]。



来源：中国信息通信研究院

图 3 大模型安全基准测试框架（2.0）

二是凝聚产业共识形成自律承诺，构建值得信赖的产业生态。2025 年 7 月，基于中国人工智能产业发展联盟发布的《人工智能安全承诺》[37]，18 家企业积极响应、主动披露安全措施。围绕**风险管理、模型安全、数据安全、基础设施安全、透明度及前沿安全研究** 6 大核心承诺内容，披露包含安全团队组织架构、风险管理方案、安全风险基线、红队测试方法、应急响应机制等在内的 43 项披露企业典型实践（见图 4），以实际行动推动人工智能安全治理走深向实。此外，包括 Anthropic、OpenAI 和 Google 在内的多家头部开发商，在发布

[36] https://mlcommons.org/ailuminate/safety/

[37] https://aihub.caict.ac.cn/ai_security_and_safety_commitments

其最先进模型时，都主动实施了更强的安全保障措施。微软连续两年发布《负责任的人工智能透明度报告》，展示如何负责任地开发和部署人工智能模型和系统[38]。

来源：中国人工智能产业发展联盟

图 4 《人工智能安全承诺》披露网站

三是推动安全测试验证，形成良性治理循环。当前，国内外产业围绕人工智能已经形成诸多评测体系，但针对测试维度、测试方法尚未形成共识。新加坡 AI Verify Foundation 组织多国企业参加“全球人工智能试点”（Global AI Assurance Pilot），在统一框架下对其人工智能系统做测试和审核，并发布公开报告，总结共性问题和最佳实践。未来，产业方面可依托产业自律承诺、沙盒验证、三方合作等方式推进实践测试工作，**一方面**可以确保企业披露内容的真实性与落地效能，

[38] https://www.microsoft.com/en-us/corporate-responsibility/responsible-ai-transparency-report/

自下而上地推动产业形成自律生态，**另一方面**能够提升模型透明度，增强各方对于大模型研发应用的信心，共促人工智能健康发展。

四是构建行业风险协同治理网络，推动信息共享。当前，产业组织积极建立漏洞信息或危险能力共享机制，通力合作促进产业生态健康发展。加州人工智能前沿模型联合政策工作组发布《加州前沿人工智能政策报告》，强调产业开展不良事件报告，用统一流程来处理模型安全漏洞和危险能力的发现与通报[39]。前沿模型论坛（FMF）成员企业合作制定处理前沿模型漏洞与危险能力的披露流程，并分享红队案例，形成半公开的“经验共享”[40]。中国信息通信研究院建设人工智能专业漏洞库，构建漏洞发现与贡献机制，规范安全企业、白帽子等主体的漏洞报送与价值评估。

五、典型领域探索安全治理实践方案

（一）金融行业打造人工智能风险管控方案

大模型技术加速创新并向金融各领域深度渗透，为金融服务效率提升、业务模式创新带来深刻变革，金融机构在推动金融行业智能化转型的同时，也在从不同方面积极应对人工智能技术带来的风险挑战。

一是探索金融行业大模型应用分类分级方案。基于人机交互程度和金融场景风险，金融行业探索“二维”分级分类合规应用体系，提出包含一至四级（L1 至 L4）的自主能力分级方案，根据客户资金关联性、金融机构资金关联性、市场冲击等因素，对投资、信贷风控、

[39] The California Report on Frontier AI Policy. https://www.gov.ca.gov/wp-content/uploads/2025/06/June-17-2025-%E2%80%93-The-California-Report-on-Frontier-AI-Policy.pdf?utm_source=chatgpt.com

[40] https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/?utm_source=chatgpt.com

保险精算、核保核赔、投顾、智能客服、营销素材生成、代码生成等金融场景进行风险等级划分，构建金融大模型通用能力和特定金融场景专业能力评估框架[41]。

二是针对智能客服应用场景，引入安全护栏工具，旨在提升模型回答安全性。当前，行业主要通过微调小尺寸模型加语料库的形式进行内容拦截，确保输出内容安全性。例如，中国邮政储蓄银行在系统研发阶段，通过自研敏感词过滤组件与微调后的轻量模型相结合，构建双向安全过滤机制，大幅降低不当输出风险。蚂蚁集团针对支小宝业务需求实施了“安全围栏”策略，开发了包括底线和意图识别、情绪分析、主题分类在内的内容理解技术，进一步提升生成内容安全可控。

三是针对智能风控应用场景，探索算法修正，旨在降低算法偏见风险。当前，行业普遍采用偏见修正算法，通过调整模型特征权重平衡不同群体评估标准，并引入风控专家经验提炼高质量数据，降低误判率。例如，工商银行在风险管理领域，面向客户经理、审查审批人员、风险官、风险经理等用户，提炼信用风险评价思维链数据，通过对基础大模型的风控能力进行强化训练，构建企业信用风险智能评估专属大模型。

四是针对外部技术实施本地化部署。针对外购的人工智能算法或产品与业务系统，金融机构积极探索安全管理措施，提升技术的可控性。当前，为确保外部采购技术安全可控，多数银行采用本地化部署方式应用大模型技术，以确保数据安全与隐私保护。

[41] 参见《中国互联网金融协会金融科技发展与研究专委会、中国信息通信研究院人工智能研究所：《金融大模型合规应用研究报告》。

（二）医疗行业多维度提升模型输出准确性

人工智能技术在医疗服务效率、优化资源配置以及推动科研创新等方面展现强大的应用前景，但也面临着**专业性不足、可解释性和准确度受限、模型幻觉等多个技术挑战和局限**，为此医疗行业从不同维度开展治理探索。

一是构建医疗领域高质量数据集，提升模型输出内容准确性。针对医学数据具有的小样本、高维度、类别不平衡等特点，以及“噪声”和缺失值等问题，医疗机构积极突破专业语料不足、多模态处理有限等技术瓶颈，探索高质量数据集建设。例如，医疗机构结合自身实践经验与技术优势，基于高质量数据集高价值应用、高知识密度、高技术含量的“三高”特征，和医疗领域数据模态多、专业强、标注难的特性，以确定伦理规则体系与符合临床诊疗逻辑的推理机制为基础，建立全域医学高质量数据集。从完整性、准确性、一致性、时效性和合规性等维度建立质量保障体系，涵盖数据需求、数据规划、数据采集、数据预处理、数据标注、模型验证等全流程迭代环节。

二是引入外部工具，创新解决模型幻觉问题。医疗领域容错率低，大模型由于固有的“幻觉”问题而给出错误建议，可能直接影响患者生命安全，如用人工智能系统监护重症患者生命体征，设备故障或数据中断可能危及安全。针对模型幻觉问题，医疗机构积极探索通过检索增强生成（RAG）减轻大模型幻觉，解决落地痛点，提升应用效果。例如，医疗人工智能助手基于医疗垂类内容的特殊性打造安全前置护栏解决方案，结合千万级自建知识库，保障内容可控生成，从领域、

话题、意图多个视角量化内容防控，保证大模型生成结果准确性符合医疗垂类的安全性和准确性，进而确保业务应用的安全性。

三是加强伦理审查，积极制定行业标准。从高质量数据集建设、算法研发到临床验证，人工智能应用链条长，参与方众多。一旦出现医疗事故，责任归属模糊，易引发社会舆论关注。产业组织应积极制定伦理评估指标，对医疗大模型进行全面、系统的测试、验证和优化，提高医疗大模型的通用性和可复制性，确保其在临床应用中具备高度的准确性、可靠性和安全性。

（三）交通行业结合应用场景探索安全方案

人工智能技术成为交通运输行业数字化转型的重要引擎。然而，人工智能在智能座舱、智能驾驶等具体场景应用还存在**测试数据集缺失、测试维度不齐全等安全挑战**，产业界积极探索针对具体应用场景的安全方案。

一是针对智能座舱场景，探索端到端的安全测试。随着智能座舱向融合多模态交互、实时数据处理与沉浸式体验的“第三生活空间”迈进，然而当前端侧大模型部署落地还存在不少技术挑战，如设备内存有限、隐私保护等。尤其是当时车端部署模型存在自研、商业合作、开源合作等多种模式，对于大模型上车的数据合规、模型稳定性、功能适配性提出挑战。针对端侧大模型安全问题，业界已经开始探索构建针对汽车领域端到端的测试体系，探索大模型在智能座舱的交互测试。

二是针对自动驾驶场景，打造更安全便捷的出行服务。自动驾驶是人工智能在交通领域中，最大的智能原生应用场景之一。为实现更

安全、更便捷的出行服务，企业积极开展端到端的仿真测试，主要通过自研仿真平台、联合专业技术公司搭建测试体系等方式实现，覆盖感知、规划、控制全流程验证。

（四）能源行业加强基础设施的安全防护

人工智能在能源领域的应用正在快速发展，涵盖电网管理、能源预测、设备维护等多个方面，推动能源的高效、安全和可持续发展。**针对能源领域数据孤岛化、算法黑盒化、算力高耗能等技术瓶颈**，产业界从数据、网络、模型等领域加强安全能力建设。

一是强化数据安全防护。针对敏感科研数据与工业数据安全需求，行业正通过“数据分级+隐私计算”策略强化防护。一方面，对涉及碳捕获地质封存参数、煤化工催化剂配方等数据实施严格管控；另一方面，在企业内部科研协作平台采用联邦学习、多方安全计算、数据空间等技术，实现“数据可用不可见”，在保障模型训练效果的同时守住数据安全底线。

二是加强工业系统网络安全防御。针对人工智能系统与传统工业控制系统融合带来的新型攻击面扩大问题，行业正推动“人工智能安全内生设计”理念落地。能源行业在新建智能电厂、煤化工等项目中，强制要求人工智能模块具备运行时异常检测、模型完整性校验及安全回滚机制，并与工控安全防护体系联动。同时，配合相关监管部门推动建立能源人工智能系统渗透测试标准，定期开展红蓝对抗演练，提升实战化防御能力。

三是提升模型在工业领域应用的可解释性。针对模型在高风险工业场景中的安全性和可解释性的需求，推动模型算法、应用系统等安全能力建设。面对煤矿井下复杂地质条件、电网实时调度等高危场景对决策容错率的严苛要求，构建全流程可靠性保障体系。例如煤炭开采、煤化工与碳捕获和封存等环节，对系统稳定性要求极高。若模型因训练数据偏差或环境突变而输出错误指令，可能触发连锁安全事件。对此，能源行业在关键场景推行“可解释 AI+专家规则双校验”机制。例如，在智能矿井调度系统中，不仅要求模型输出附带置信度与关键特征解释，还需通过内置工艺安全规则库进行二次验证，确保决策既智能又可控。

（五）通信行业加强“以技治技”的安全治理体系

从 5G 通信网络的高效调度算法到边缘计算的智能决策，人工智能的应用正在推动通信系统的智能化、自动化与网络化转型。随着网络规模及复杂度与日俱增，网络安全运营面临**能力分散、攻防不对等、响应不及时**等风险挑战。产业界积极探索运用人工智能技术赋能网络安全治理。

一是积极实践 AI+数据安全智能管理。例如，中国联通深入应用人工智能技术，打造智能化数据分类分级能力，提升分级准确率，构建多维度的数据导出和 API 接口监测能力，依托元景大模型及知识中心，进行风险研判和实时拦截，实现数据差异化管控，及时自动预警数据安全风险，提升数据安全保障水平。

二是构建大小模型自主协同技术方案。针对传统网络安全运营中人工分析效率低、误告警率高、技术门槛高等行业痛点，中国移动通信集团福建有限公司创新构建以通义千问、DeepSeek 和九天等大模型为核心的“大小模型自主协同”技术体系，将安全产品能力和人工智能专项安全小模型积累，与大模型强大的意图理解和推理能力进行融合，构建了人工智能安全运营系统，提高安全运营的执行效率，大幅度降低安全运营的技术门槛。

三是结合检索增强生成与深度思考技术的安全知识赋能。例如，中国移动通信集团福建有限公司结合检索增强生成（RAG）与深度思考技术的安全知识赋能系统实时接入多种最新的威胁情报源，包括恶意 IP 地址、域名或病毒签名库、漏洞信息库、安全事件情况库等；同时支持用户使用知识库管理界面上传如网络安全法、网络安全等级保护等各种安全知识文件，通过智能问答形式为用户提供各种高效率、高质量、智能化的安全知识检索和赋能。

四是构建覆盖“事前-事中-事后”的人工智能一体化安全技术体系。例如，中国联通构建 AI 应用全生命周期全流程安全管控体系。制度规范层面，制定企业级制度规范，明确职责、技术要求、运营要求；评估检查方面，按“部署-训练-应用-下线”全周期维度，精准排查识别数据泄露、权限滥用、算法安全等风险隐患；执行检查层面，分类处置共性与个性问题，明确整改要求；动态优化方面，结合企业应用场景细化风险类型，迭代升级评估方法与管控举措。

六、展望

当前，人工智能技术“飞轮”正在加速推动各行各业发展。与此同时，人工智能产业面临战略机遇和风险挑战并存的局面，面向产业的人工智能安全治理体系建设正步入“快车道”。**着眼短期和中期，人工智能安全治理将朝着多维度、深层次的方向全面进阶，形成覆盖全链条的治理新格局。**

一是供给层面，内容来源与真实性成为安全焦点。随着生成式人工智能在文本、图像、音频、视频等领域的规模化应用，“人工智能生成虚假新闻”“恶意内容批量生产”等问题将愈发突出，不仅对个人权益、企业声誉造成直接冲击，更可能引发社会信任危机与公共安全风险。为此，服务供给方可从技术层面提升水印的不可见、抗篡改、高鲁棒性等特征，通过在人工智能生成内容的底层数据中嵌入唯一标识，实现内容创作者、生成模型、传播路径的全链条追溯。

二是应用层面，加强技术透明与可解释成为治理重点。随着人工智能技术在医疗、金融、司法、自动驾驶等关键领域的深度渗透，其决策结果将直接关系到人的生命健康、财产安全与合法权益，可解释人工智能技术需要从实验室走向规模化应用。同时，服务供给方与应用方需探索生成更直观的可视化报告，向用户呈现影响决策的关键特征与权重。

三是管理层面，需探索面向产业的人工智能管理体系。未来，产业还需持续探索以“计划、执行、检查和处理”（PDCA）模式为核心的管理模式，实现从“碎片化”向“系统化”的转型。**在计划阶段，**

结合企业自身业务需求与行业安全标准，制定明确的人工智能安全管理目标与实施方案。**在执行阶段**，依托智能化工具实现管理措施的落地。**在检查阶段**，建立常态化的安全审计与评估机制。**在处理阶段，**形成问题整改与体系优化的闭环。推动管理体系迭代完善的螺旋式上升，最终构建起适配产业发展的长效安全管理机制。

四是产业层面，需搭建聚焦行业的安全评估测试体系。伴随人工智能技术在各行各业的深度应用，以行业化、差异化的安全评估测试体系将成为产业长远发展的支撑，推动技术在各领域的安全合规应用。在现有通用标准基础上，可探索由各行业主管部门、龙头企业、科研机构制定细分领域的安全评估标准。依托人工智能安全实验室、行业协会、第三方检测机构，构建具备行业特色的测试环境与工具链。通过搭建聚焦行业的安全评估测试体系，实现人工智能安全治理与行业需求的精准匹配，为人工智能产业的规范化、安全化发展保驾护航。

展望远期未来，人工智能的演进将超越工具范畴，涌现更多“智能”，迈向具有自主适应与策略演化能力的行动主体新阶段。这在带来生产力跃升的同时，其引发的具有根本性、颠覆性的深层次人工智能安全风险也值得高度关注。未来人工智能可能进入超越人类智能，甚至具备自我意识的阶段，其行为能否与人类核心利益保持一致尚存在巨大不确定性。在未来的人机共生阶段，人机关系界定、伦理边界划分、社会规则体系重构也将面临空前考验。推进人工智能安全治理，必须将其置于构建人类命运共同体的高度来谋划落实，筑牢**技术根基**，从理论方法、算法模型、系统架构等方面推动“Safety by design”技

术研究；推动**制度创新**，建立政府部门、私营主体、产业组织、社会公众多元共治的敏捷包容治理体系；加强**全球协作**推动，在不确定的状态下更需加深国际交流、促进国际共识、增进国际合作，共同推动人工智能向安全、负责任、可持续的方向发展。

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62309179

传真：010-62304980

网址：www.caict.ac.cn