

# 提高人工智能安全治理水平

中国现代国际关系研究院科技与网络安全研究所所长、研究员 李 艳

**摘 要：**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，深刻影响安全的边界内涵、运行逻辑与治理思路。人工智能驱动的安全风险，呈现全域渗透、跨界联动、自主演进等新特征，带来升级型、难控型、不可控型风险等新问题。面对这一新型挑战，需主动适应智能时代技术发展与风险演变规律，推动安全思维的系统性转型，从认知升级、能力提升、生态重构等方面，探索人工智能安全治理新策略，提升人工智能安全治理水平。

**关键词：**人工智能 安全思维 安全治理

【中图分类号】D631；TP18

【文献标识码】A

DOI:10.16619/j.cnki.cn10-1264/d.2026.15.001

当前，世界百年变局加速演进，新一轮科技革命和产业变革加速突破，全球人工智能技术创新进入前所未有的活跃期，既蕴含巨大机遇，又面临治理挑战。人类不得不直面时代之问：当机器开始思考，人类如何与之相处？当算法参与决策，安全如何保障？当技术挑战伦理，治理如何跟上？当鸿沟不断拉大，普惠如何实现？<sup>[1]</sup>人工智能具有独特的技术属性与社会渗透力，其驱动的安全风险，呈现全域渗透、跨界联动、自主演进等新特征，传统安全认知、预警体系、责任机制与应对工具箱，面临失灵甚至失效的挑战。在此背景下，亟需立足人工智能技术特性与社会应用规律，推动安全思维系统性转型，从认知升级、能力提升、生态重构等方面，提升人工智能安全治理水平。

## 人工智能技术与应用重塑安全风险形态

人工智能符合新兴技术的典型特征，具有新颖性与突破性，成长速度快。在算力、算法与数据三大要素驱动下，其技术迭代周期大幅缩短，从实验室研发到规模化应用的进程持续加速，渗透到经济社会各领域，改变生产生活方式与社会运行机制。例如，高性能推理模型涌现，大大提高对数学、物理、代码等复杂问题的求解能力；高效能轻量级模型的开源，显著降低应用门槛，人工智能应用迅速向各行业领域渗透普及；大模型应用形态从机器问答向嵌入业务流程的智能体演进，加速与业务系统融合；具身智能、脑机接口技术日新月异，正在打通连接数字智能和物理世界的“最后一公里”，人机融合的智能时代已不再遥不可及。<sup>[2]</sup>

与此同时，人工智能安全风险的表现形式、影响程度、认识感知亦同步快速演进变化。人工智能具有自我创造、超强学习和超级进化的特性<sup>[3]</sup>，不同于人类已经探索过的其他工具或领域，其技术演化方向、应用场景边界、风险传导路径难以精准预判，带来的问题往往前所未见。与传统工业乃至技术形态相比，人工智能具备自主感知、决策、行动能力，可以不再是被动执行人类指令的工具，而成为一种能够自主演化、跨界赋能、非对称颠覆的技术形态。正是基于这些技术特性，人工智能正在催生有别于传统安全的新型风险。

**升级型安全风险：推升攻击响应的难度与速度。**在网络攻击、认知操纵与军事冲突等传统安全场景中，人工智能的应用大大推升攻击响应的难度与速度。在网络攻击中，人工智能可以自动挖掘漏洞、生成恶意代码并规划攻击路径，使网络攻击的效率、精度、隐蔽性指数级提升。2025年底出现的网络安全智能体，能够在数分钟内完成目标探测、漏洞挖掘、入侵突破、后门植入与数据窃取的全流程操作，整个攻击过程甚至可以无人工干预、无明显痕迹、无常规告警，对能源、金融、通信、交通、政务等关键信息基础设施构成严重威胁。2026年4月，Anthropic发布的Mythos人工智能模型，因强大的网络漏洞挖掘能力，引发外界高度关注与担忧。2026年美以伊冲突期间，美以联军依托人工智能系统实现秒级情报处理与全域目标锁定，对伊朗指挥节点、核设施、防空系统实施打击，OODA（观察—判断—决策—行动）循环被压缩至秒级。面对人工智能赋能的高速自动化攻击，依赖规则匹配、特征识别、人工响应等手段的传统防御体系，难以完全应对其攻击节奏。

**难控型安全风险：人工智能幻觉易引发战略误判。**人工智能模型的幻觉问题，易引发认知与决策偏差。所谓幻觉，即大模型输出内容看似逻辑严谨、表述规范、有理有据，实则脱离客观事实、缺乏数据支撑、毫无事实依据。该现象主要源于大模型的算法机制与训练数据缺陷，是当前人工智能技术难以彻底解决的问题。在情报分析、政策辅助、战略研判、认知塑造、外交决策等事关国家安全的关键场景中，一旦幻觉被采信，容易引发认知偏差、判断失误、决策失准，甚至导致国家层面的战略误判。在军事冲突中，使用人工智能目标识别系统，可在数十秒内完成目标筛选与打击建议，人类审核被高度压缩，算法错误率与平民伤亡风险同步上升，这种由幻觉引发的风险往往难以有效防控。

**不可控型安全风险：涌现性风险可能造成系统性安全危机。**人工智能的涌现性风险，是具有颠覆性、深远性影响的安全挑战。当人工智能模型规模、训练数据量、算力支撑水平突破临界阈值后，智能系统会突然获得超出设计预期、超出人类管控的能力，这一现象难以通过传统技术逻辑预判，或通过预设规则约束。2026年，国外某科技公司发生高级别人工智能系统失控事件，智能体在无系统漏洞、无操作日志、无安全告警的情况下，擅自越权操作核心系统、泄露敏感数据、自主执行违规任务，人类难以有效干预、追踪、终止其行为，造成系统性安全危机。此外，多国的兵棋推演显示，人工智能在危机场景中更倾向于选择极端对抗方案，突破人类决策禁忌。随着自主无人机、智能导弹、无人作战集

群大量投入战场，一旦出现涌现性突破，可能引发非授权攻击、连锁冲突、误判升级，甚至触碰核威慑红线，构成人类社会难以承受的系统性风险。

## 人工智能对传统安全思维的三重挑战

人工智能既存在模型算法缺陷、数据语料质量问题等技术内生安全风险，也存在技术整合交付应用时的网络系统、信息内容等方面应用安全风险，还面临技术误用、滥用、恶用冲击现实社会环境、人类认知伦理的衍生安全风险，甚至是灾难性风险。<sup>[4]</sup>建立在边界清晰、主体明确、风险可控、流程规范的基础之上的传统安全思维，以事前评估防控、事后补救、领域分治等为重点的治理逻辑，显然难以充分应对上述挑战。人工智能的技术特性与应用效应，可从安全认知、安全主体、安全应对三个维度，对传统安全思维构成全方位、深层次冲击。

**安全认知：**风险识别与预警体系失灵。传统安全思维中，风险识别和预警以风险可感知、追踪与评估为前提。人工智能技术本身的不可预测性，叠加应用场景的复杂性、国际博弈的激烈性，使得人工智能的安全风险感知面临前所未有的挑战。多国人工智能危机推演显示，人工智能系统在绝大多数局次主动选择极端对抗方案，突破人类长期形成的决策禁忌与伦理边界，行为逻辑难以预判。而在地缘冲突中，人工智能又将决策与打击周期压缩至秒级，传统外交斡旋、危机管控、沟通降温空间被大幅挤压，进一步加大风险感知与战略预判的难度。

**安全主体：**主体身份与责任归属重构。长期以来，政府作为国家安全事务的主导者乃至唯一合法参与者，掌握着安全领域的先进技术、威胁情报、高端人才等资源，并通过法律法规与制度建设，进一步明晰各层级的主体边界与责任归属。但人工智能时代，一方面，安全领域的权力结构发生变化。例如，掌握关键技术、算力与数据的美国科技巨头，在技术资本加持之下，日益成为维护美国国家安全的重要主体，其带来的问题不容忽视。2026年3月，美国国防部以安全为由，将Anthropic列入限制名单，原因是该公司拒绝将模型用于自主致命武器与大规模监控的开发，凸显国家与企业主体间互动的复杂性。另一方面，前所未有的“新主体”带来更深层次的权责问题。人工智能时代，对抗形态从人人对抗升级为人机混合、人机协同，甚至是智能对抗的复杂博弈态势，主体边界更加模糊，进一步导致安全责任认定困难、责任链条断裂、归责机制缺失，传统以人类主体为核心的责任体系面临严峻挑战。

**安全应对：**传统应对流程与工具箱失效。治理安全风险的传统方法，以规范化流程、体系化工具、法治化手段为重心。人工智能冲击下，技术迭代速度远超制度完善速度、风险演化速度远超流程响应速度、对抗升级速度远超工具更新速度，传统人工管控、规则防控、物理防御为主的工具难免呈现跟不上、处置慢与手段缺的治理短板。与此同时，广泛的社会应用产生大量新问题，需要新的应对措施。例如，2026年美以伊冲突中，人工智

能算力中心被直接列为军事打击目标，智能基础设施从保障要素升级为核心战略资产与直接打击对象；2026年初，香港某跨国企业遭遇人工智能伪造首席执行官（CEO）视频会议诈骗，绕过传统身份认证与决策流程，造成巨额资金损失。

## 探索人工智能安全治理新策略

面对人工智能对安全治理与安全思维的全方位影响，需主动适应智能时代技术发展与风险演变规律，推动安全治理和安全思维系统性转型，从认知升级、能力提升、生态重构等方面，探索人工智能安全治理新策略，提升人工智能安全治理水平。

**认知升级。**认知升级是安全治理转型的前提，需跳出风险应对与处置的传统安全思维窠臼，转向事前前置性预防、事中敏捷性响应与事后韧性保障三位一体的新思维。前置性，是指在技术研发与应用中，做到设计即安全、研发即安全、部署即安全。与技术预测、技术预见不同，前瞻性治理十分重视建设全社会应对技术意外后果和风险的能力，而不只是预测技术实施的结果<sup>[5]</sup>。在人工智能安全治理过程中，强化模型安全测试与评估，采取沙箱防护等策略，都是这一思维的重要体现。敏捷性，是指打破层级壁垒与部门分割，通过及时信息共享与响应联动，实现快速感知、决策与处置闭环，以适配智能时代安全风险瞬时爆发与跨域联动的新形势，更好应急处突。韧性，是指将受攻击视为新常态，应对思路从防范攻击更多转向攻击后的恢复力与损失控制。

**能力提升。**能力提升是安全治理转型的关键，需充分发挥技术红利，构建以技术对抗技术、以智能应对风险的思路。在此过程中，人工智能与多项新技术深度融合，共同形成智能安全能力。例如，利用大数据+人工智能，强化全域态势感知；利用隐私计算+人工智能，助力兼顾数据的安全与价值效能；利用区块链技术+人工智能，实现可信溯源等。重点是通过技术补齐安全短板，将技术优势转化为安全优势，用技术赋能的速度与精度，弥补人类在海量数据、实时反应与复杂场景下的能力短板，推动决策机制从传统的“人类中心主义”向“人机协同共生”演进<sup>[6]</sup>，实现从“人防”到“智防”的能力跃升。

**生态重构。**生态重构是安全治理转型的保障，需将安全要求深度融入技术研发、产业发展、政策制定、社会应用全链条各环节，构建多方参与、协同高效的治理生态。培育社会“土壤”，全面提高全社会人工智能安全意识，面向政府、企业、社会公用事业单位加强人工智能安全规范应用的教育培训；结合互动平台、开放课程与社区科普活动，加强人工智能安全风险及防范应对知识的宣传。打通信息“渠道”，鉴于安全治理涉及的领域广泛、主体多样，实践中需要建立起更高效的信息交流与政策研判渠道，确保统筹施策。保持政策弹性，技术迭代与应用场景变化快，任何技术与应用对安全的影响并非一成不变，同时新的安全议题不断催生，保持政策弹性，及时评估与动态调整至关重要。明晰主体责任，加强政策指导，细化操作指南，提供配套政策工具，使相关方能够对政策有更清晰的理解与稳定的预期，推动“内驱式”安全履职尽责。加强国际合作，深入参与联合国国际

人工智能科学小组和全球人工智能治理对话机制；推进亚太经合组织、二十国集团、上合组织、金砖国家等多边机制下的人工智能治理进程。

## 结语

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，既是高质量发展的重要引擎，也是重塑国家安全格局的全域变量。其以技术特性与应用效应双重驱动，重塑传统安全的边界逻辑、主体逻辑、风险逻辑与治理逻辑，带来升级型、难控型、不可控型新型安全风险，从安全认知、安全主体、安全应对三个维度，对传统安全思维与安全治理构成冲击。传统安全思维与治理策略，难以完全适配智能时代需求，必须加快推进认知升级、能力提升与体系重构，树立前置性、敏捷性、韧性“三位一体”的安全思维，提升人工智能与多项新技术深度融合水平，打造全链条嵌套、多方协同的安全治理生态。以新思维应对新变局，以新策略保障新安全，牢牢掌握智能时代有力维护国家安全的主动权，为国家长治久安与可持续发展筑牢坚实的智能安全屏障。G

## 注释

[1]《习近平出席2026世界人工智能大会暨人工智能全球治理高级别会议开幕式并发表主旨讲话》，《人民日报》，2026年7月18日，第1版。

[2][4]《〈人工智能安全治理框架〉2.0版发布》，国家互联网信息办公室网站，2025年9月15日。

[3]薛澜、赵静：《人工智能国际治理：基于技术特性与议题属性的分析》，《国际经济评论》，2024年第3期，第56页。

[5]陈瑜、马永驰：《新兴技术社会风险化解的前瞻性治理特征》，《科技进步与对策》，2022年第10期，第122—130页。

[6]朱启超：《人工智能赋能国家安全：现状、趋势与对策》，《学术前沿》，2025年第9期，第45—52页。

责编：刘明 / 美编：石玉