引用格式：陈恺，李锭，孟国柱，等．人工智能赋能网络空间安全：机遇、挑战与路径．中国科学院院刊，2026, 41(6): 1152-1161, doi: 10.3724/j.issn.1000-3045.20260416008.
Chen K, Li D, Meng G Z, et al. Artificial intelligence for cybersecurity: Opportunities, challenges, and approaches. Bulletin of Chinese Academy of Sciences, 2026, 41(6): 1152-1161, doi: 10.3724/j.issn.1000-3045.20260416008. (in Chinese)

人工智能赋能网络空间安全：机遇、挑战与路径

陈 恺[1] 李 锭[2] 孟国柱[1] 纪守领[3] 李长江[3] 杨 伊[1] 冯登国[4*]

1 中国科学院信息工程研究所 北京 100085
2 北京大学 计算机学院 北京 100871
3 浙江大学 计算机科学与技术学院 杭州 310058
4 中国科学院软件研究所 北京 100190

摘要 网络空间安全研究本身、安全建制及治理政策都处于剧烈演变之中。当前，日趋隐蔽、快速迭代的智能化攻击与国家对高水平安全的紧迫期望，正在推动政府、科研机构及企业的职能与互动关系发生深刻变化。为此，文章在分析人工智能（AI）时代网络安全挑战与机遇的基础上，探讨关键信息基础设施保护、国家数据安全及网络空间主权维护等重要应用场景，剖析人工智能对安全领域的赋能逻辑，并提出治理体系对策，以期为加快建设网络强国、筑牢数字安全屏障提供管窥之见。

关键词 人工智能，网络空间安全，关键信息基础设施保护，智能化攻击与防御

DOI 10.3724/j.issn.1000-3045.20260416008
CSTR 32128.14.CASbulletin.20260416008

人工智能（AI）技术的快速迭代及其与网络空间安全的深度融合，正驱动网络防御向智能化、自适应化演进，重塑着数字时代的安全格局。这直接引发了网络安全产业结构的演变，即传统碎片化、被动型的安全产品线正加速向以人工智能为核心的集约化、主动型安全产品转变。这些改变对安全威胁的传播路径与防御资源配置产生了极大影响，形成了涵盖信息、物理与社会多维度的安全博弈格局，加剧了技术创新、国家主权与业务连续性之间的复杂关系。与此同时，网络空间安全研究体系、治理政策与攻防文化均

*通信作者
资助项目：中国科学院学部人工智能重大咨询项目
修改稿收到日期：2026年6月14日

在加速演变。基于此，本文分析人工智能时代的安全挑战与机遇，聚焦关键信息基础设施保护、数据安全、国际博弈与综合治理等场景，提出人工智能驱动的安全治理新体系。

1 人工智能时代下网络空间安全的挑战与机遇

从人工智能与网络空间安全关系的发展演变来看，技术开发者对人工智能算法创新与应用效率的追求，与政府对国家战略安全及数据主权的维护之间存在内在联系。这种关系的演进既受到社会对人工智能赋能网络空间安全攻防逻辑和政府监管职责认识的影响，也与人工智能时代数据驱动、跨学科交叉的知识生产特点有关。本节从研究对象、发展趋势、研究范式与技术支撑3个维度对人工智能时代下网络空间安全面临的挑战与机遇进行阐述。

1.1 研究对象

人工智能时代的网络空间安全涉及国家、个人与技术。*在国家层面*，包括承担网络空间安全监管职责的政府主管部门、维护国家关键信息基础设施安全的安全机构，以及履行国家安全战略的各类组织机构；*在个人层面*，涉及广大网络用户、网络安全从业者和人工智能系统使用者，其隐私数据、数字资产及人工智能服务使用权益是其关注的对象；*在技术层面*，涵盖人工智能算法、模型、数据集、训练与推理平台等要素，以及受人工智能影响的云计算、物联网、区块链、工业控制系统等领域。

1.2 发展趋势

当前，网络空间安全已上升为国家战略。网络空间作为“第五疆域”，正加速演化为大国战略竞争的新前沿。截至2025年12月，我国网民规模达11.25亿，互联网普及率超过80%，数字技术与人工智能深度融入经济社会各领域，网络空间安全风险也由传统单一技术问题，演变为涵盖数据安全、算法安全、平台安全、供应链安全及人工智能安全等在内的复杂系统性挑战。

随着人工智能技术持续演进，传统依赖专家经验的安全分析范式已向“数据驱动”转型，并逐步进入“智能驱动”新时代。海量网络流量、系统日志及威胁情报等多源异构数据的汇聚，为模型训练提供了基础，“以算力换人力”逐步成为现实。相较传统模式，数据驱动在分析速度、处理规模与响应持续性方面具备显著优势，可实现毫秒级关联分析，并有效识别跨时间、跨系统的隐蔽攻击链。当前，大模型和人工智能智能体（Agent）（以下简称“智能体”）[①]技术趋于成熟，安全分析范式正加速迈向智能体自主决策、动态博弈的“智能驱动”新阶段，推动网络空间安全体系由“人力密集型”向“算力密集型”加速转变。

同时，全球的安全产业格局正在重塑。以美国Palantir、CrowdStrike、Darktrace为代表的人工智能安全企业迅速崛起，市值达百亿甚至千亿美元。这些企业通过整合多源异构数据、构建数据融合分析引擎，实现了威胁检测、调查、响应的自动化闭环。2026年4月，Anthropic公司在测试其前沿模型（Claude Mythos Preview）时，该模型不仅发现了一个隐藏在FreeBSD操作系统中长达17年的重要漏洞，还完全自主编写了攻击代码，实现了最高权限的接管。这表明产业技术不仅完成了“数据换经验”的资产积累，同时也形成了由人工智能完全自主闭环的“智能驱动”里程碑，显示出人工智能技术在安全攻防领域的巨大潜力。

① Engineering at anthropic. Anthropic, building effective agents. (2024-12-16)[2026-06-10]. https://www.anthropic.com/engineering/building-effective-agents.

1.3 研究范式与技术支撑

人工智能时代下的网络空间安全正实现从“专家驱动”向“智能驱动”的技术范式转变。如图1所示，以数学、博弈论等支撑性基础学科为“根基”，发展出传统安全与人工智能领域研究方向。例如，安全领域的密码学、软件安全、系统安全等，以及人工智能领域的深度学习、强化学习、大模型（即大型语言模型）等。当两者结合后，产生了诸如智能化漏洞发现、智能体安全、人工智能模型安全等新兴研究方向，通过智能化手段解决传统安全问题，最终为多种安全应用场景提供支持。

要实现智能驱动，需要数据和技术的支撑。*在数据层面*，通过整合系统日志、流量数据及威胁情报等多源异构数据，利用图神经网络、知识图谱等手段进行语义关联表征，构建全域数据底座，确保人工智能模型在训练与推理中拥有充足的“信息燃料”。*在技术层面*，设计多种攻防智能体，结合大模型与强化学习等算法，在网络环境中部署开展持续博弈，并在博弈过程中不断自我进化。这些智能体可在人工智能网络安全挑战赛和国内大模型攻防演练等真实对抗场景中进行检验与提升。这种“攻击—防御—演化”的机制让安全能力脱离了对静态规则的依赖，转而进入动态对抗、自主生成的智能进化阶段。人工智能实战能力不断提升，直接赋能于关键信息基础设施保护、国家数据安全与隐私保护等应用场景。在该范式支撑下，网络空间安全将从被动补漏走向主动演进，在持续的智能博弈中实现安全能力的代际突破，构筑起守护网络空间主权的动态屏障。

本文以下几节将针对不断演进的安全威胁，从国家战略契合度、技术引领性、防护广度与深度及经济社会效益等维度，聚焦三大核心场景，即关键信息基础设施保护、国家数据安全与隐私保护，以及国际博弈与网络空间主权维护开展研究，系统阐述人工智能如何由被动防御走向主动赋能，及其对安全领域带来

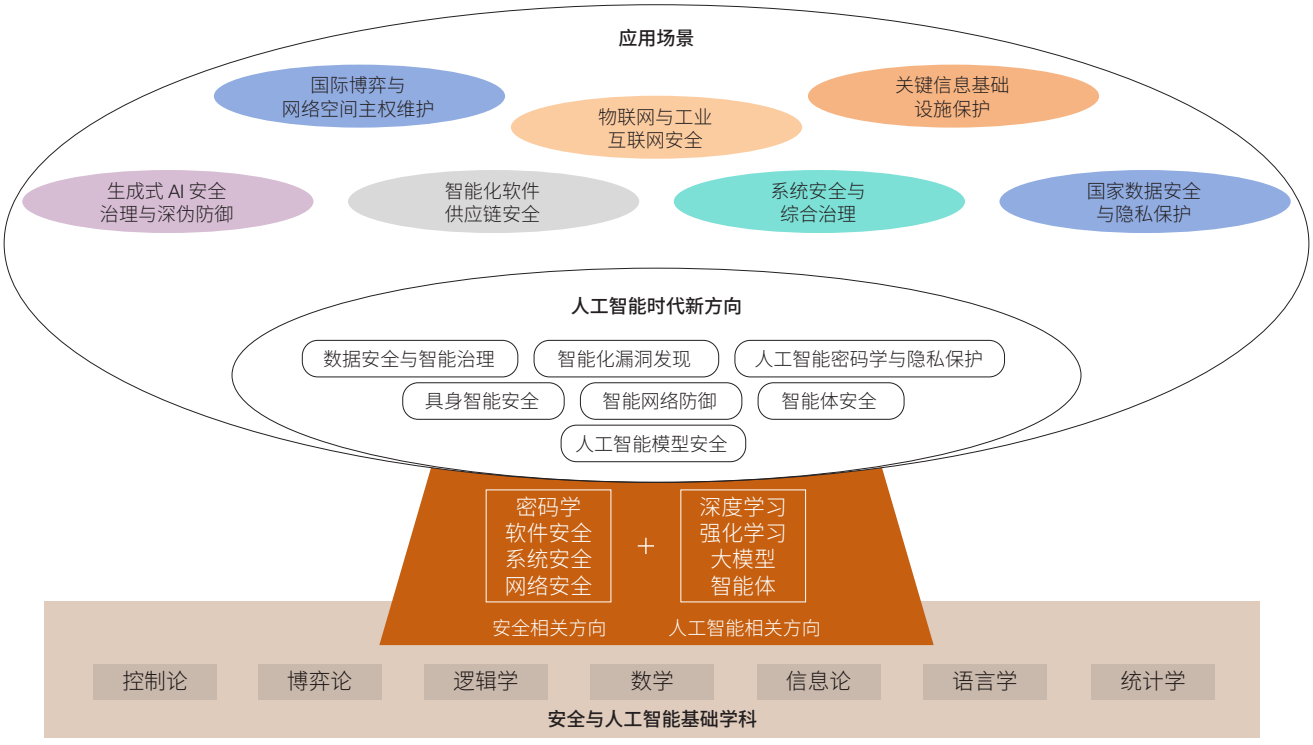

图1　人工智能时代网络空间安全研究整体框架图

Figure 1　Research framework on AI for cybersecurity

的挑战和变革。

2 关键信息基础设施保护

2.1 从人机对抗到机机博弈，高自治智能体引发防线失效困境

关键信息基础设施是国家数字经济与社会运行的核心底座②。现阶段，面向关键信息基础设施的高级持续性威胁（APT）已呈现出攻击手段多元化、覆盖范围广、潜伏周期长、隐蔽性强等特征。静态边界防御本就长期承压，而智能体的规模化应用，将网络攻防博弈从“人机对抗”彻底推向“机机博弈”，直接导致关键信息基础设施传统防御体系陷入两大失效困境：① 攻击端智能体的自主化、规模化攻击特质，对传统防御的底层设计逻辑构成巨大威胁；② 防御端为提效引入的防守智能体，新增大量原生攻击面，传统防护适配管控能力弱③。

2.2 智能体驱动下关键信息基础设施防御失效的深层机理

（1）攻击端智能体突破传统防御的能力边界。传统防御依赖静态规则库、人工特征提取的被动模式，能应对已知特征、边界清晰的攻击。而智能体具备全流程自主攻击能力，可自主感知环境、拆解任务、动态演进攻击链路，攻击效率呈指数级提升；同时，可自动化发现零日漏洞、发起跨多个网络域的APT攻击。传统静态补丁、单点监测难以应对这种动态变化的攻击手段，尤其是对高隐匿长周期攻击易于失防[1,2]。

（2）防御端增加防守智能体，其带来范围更大的攻击面。关键信息基础设施引入的高权限防守智能体，高度依赖外部插件生态与大模型指令调度，攻击者可通过技能（skill）投毒、提示词注入、供应链污染等方式，隐蔽劫持智能体执行逻辑[3]，且无明显异常特征，使得传统规则匹配方式失效④。传统边界防护多为监测节点流量，难以对智能体内部推理、工具调用行为等进行防御。高权限智能体一旦被劫持，将引发风险级联扩散，击穿边界隔离体系。

2.3 迈向多智能体协同与主动免疫的内生安全新范式

针对攻击端冲击，构建适配“机机博弈”的全域感知响应体系：以大模型为中心，智能体为其执行操作的手段，实现从被动处置到主动预警的转变，搭建自动化响应闭环，匹配智能体攻击的效率节奏。在此过程中，利用大模型的核心不再仅关注模型参数，而在于向“Harness（驾驭）工程”⑤进化，通过为大模型搭建包含工具链、反馈循环和自动验证的完整工作环境，将模型的原始动力转化为可控的实战能力。通过预设通用的“skills”能力单元⑥，让大模型能直接理解防御工具的功能与调用逻辑，从而在毫秒级博弈中实现自主且精准的战术响应。

针对防守智能体内生风险，构建全周期纵深防御与外部审计。① 实现智能体全生命周期安全审查。以供应链校验与静态代码审计为起点，阻断恶意插件的

② 《国家网络空间安全战略》全文. (2016-12-27)[2026-06-14]. https://www.cac.gov.cn/2016-12/27/c_1120195926.htm.

③ 国家互联网信息办公室关于《关键信息基础设施安全保护条例（征求意见稿）》公开征求意见的通知. (2017-07-11)[2026-06-14]. https://www.cac.gov.cn/2017-07/11/c_1121294220.htm.

④ 程学旗. 专家解读｜新框架实现三个“转变”,构建我国人工智能安全治理新格局. (2025-09-29)[2026-06-10]. https://www.cac.gov.cn/2025-09/28/c_1760779757508049.htm.

⑤ Lopopolo R. 工程技术：在智能体优先的世界中利用 Codex. (2026-02-11)[2026-06-10]. https://openai.com/zh-Hans-CN/index/harness-engineering.

⑥ Anthropic, equipping agents for the real world with agent skills. (2025-10-16)[2026-06-10]. https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-skills.

投毒扩散；在运行阶段，通过动态行为审计与基线对齐，实现对指令流、接口调用及特权操作的全流程管控。配合最小权限调度机制与沙箱隔离技术，有效防止单点风险引发的级联失效，确保智能体的自动化探索始终被锁定在预设的安全解空间内。② 使用外部审计的治理逻辑，纠正模型自我评价不可靠的缺陷。通过在智能体模型外部搭建独立的第三方评估机制与权限边界，将原本可能失控的自动化行为转化为可预测的受控演化。这2种措施并非限制智能体的能力，而是使其在可控情况下推动关键信息基础设施从被动打补丁向主动免疫体系转型。

3 国家数据安全与隐私保护

3.1 算法自主执行改变传统隐私治理体系

在人工智能时代，数据已从静态资产转化为驱动模型训练、自动决策与智能执行的核心生产要素。随着采集对象从传统信息扩展至多模态及生物特征，安全威胁已演变为贯穿采集、标注与流通全生命周期的链式风险。尽管我国已建立以《中华人民共和国数据安全法》为核心的治理体系并迈向精细化监管[7]，但人工智能在对海量数据的强依赖与个人信息保护“最小化处理”原则之间，依然存在难以调和的实践难度。

随着智能体的普及，传统的隐私治理模式正面临着失效的困境。① 传统的“告知—同意”机制，其在智能体自主拆解任务、频繁调用外部插件的过程中已经很难应对。现有方法常对智能体进行一次性授权，但难以管控大模型自主设计的任务编排方案。②“静态围堵”的防御逻辑，其难以追踪数据在多模型、多工具间的流转与变形，使得传统的物理边界防护难以处理。大模型和智能体的自主决策模式，将隐私侵害从“信息泄露”转变为“执行失控”。因而，当前法治化治理的紧迫任务在于重塑隐私保护逻辑，构建一种能对自主执行行为进行全过程强制约束的动态机制。

3.2 人工智能复合工具链正瓦解传统数据的安全边界

在人工智能时代，数据安全风险已从传统的保密问题演变为由数据、模型、协议与工具链交织构成的复合治理难题。① 数据的再识别风险显著上升。即便经过脱敏处理，海量数据集的交叉匹配仍能重构个人身份与社会关系；而模型侧的成员推断攻击则能从输出端反推出训练样本[4,5]，使传统“去标识化”手段在攻击算法面前易于失效。② 安全威胁正从简单的“数据泄露”向“数据污染与结果误导”发展。攻击者不仅能诱导人工智能应用泄露敏感信息，还能通过批量伪造内容实施生成式引擎优化（GEO）投毒攻击⑦，从源头污染模型认知并操控最终决策。这标志着风险已延伸至决策链的可靠性。

随着智能体及插件系统的普及，数据流转边界的拓宽使权限管控面临前所未有的挑战。恶意第三方工具常通过仿冒手段潜入生态，通过智能体自动下载执行，并利用来自当前智能体的高级权限隐藏及窃取本地机密。这种从“直接泄露”向“权限继承与滥用”的演变，使得敏感数据外泄更难被感知。因此，当前制度设计需考虑人工智能数据链，推动治理重心从单一的收集合规转向全流程监控，确保数据在调用链与工具链的使用中始终受到持续且有效的约束。

3.3 全调用链监管筑牢数据安全防线

面对人工智能对数据安全边界的重构，国家治理体系需突破“分类分级”与“告知同意”的传统框架，向“法律、监管、技术”三位一体的深度融合体

⑦ 陈归辞. 315晚会曝光AI“投毒”灰产链，暴露大模型背后算法高危漏洞. (2026-03-16)[2026-06-10]. https://www.21jingji.com/article/20260316/herald/8cf9afdb3bc8ba06b10b2f89aef3bc17.html.

系转型。这一转变的核心在于实现治理逻辑从“静态合规”向“运行时合规”的跨越，确保数据保护能力能够跨越复杂的算法模型与自动化工具链，实现全生命周期的动态受控。

（1）在法律与监管维度，推动规则体系从“声明式”向“操作式”改变。立法层面须进一步明确模型训练、工具接入及自动执行各环节的责任边界，以应对不同国际治理路径带来的跨境数据流动挑战⑧。监管层面则需要强化合规审计与算法备案的联动，要求企业提供训练数据溯源与清洗日志等技术证据，证明保护措施在系统运行中持续有效。针对智能体普及带来的风险，监管重心应前移至身份认证与权限约束，参考国际标准建立更具针对性的身份与行为审计框架⑨。

（2）技术控制层面需构建实时防御架构。治理重点应从单纯的静态清洗转向对模型推理与智能体执行过程的策略强制，通过部署安全网关实现即时授权与短时令牌机制，确保每一项自动化指令均可追溯。同时，综合运用差分隐私与可信执行环境（TEE）等隐私保护技术[6,7]，从数据形态与运行环境两端阻断滥用路径。

（3）国家数据安全框架有必要将治理目标升级为“防泄露、防污染、防误导”的协同防御。这意味着需将供应链安全、恶意样本投喂及插件越权等新型风险纳入监控范畴，建立起涵盖数据来源认证与流转审计的综合治理机制。只有通过制度、过程与技术的深度耦合，才能在释放数据要素价值的同时，筑牢人工智能时代的国家安全与个人隐私底线。

4 国际博弈与网络空间主权维护

4.1 人工智能执行链扩张挑战网络空间主权

人工智能广泛应用，网络空间已由传统的信息传播媒介升级为国家主权延伸的核心战略领域[8]。随着《全球数字契约》等国际准则的推进⑩，网络空间治理正处于全球结构调整的关键期。在人工智能时代，网络空间主权的内涵已突破了地理意义上的基础设施与数据管辖，进一步向模型能力、接口协议、工具生态及智能执行链条扩展。这意味着国家主权的实现，正日益取决于对算法训练、规则设定及工具分发权的掌控。例如，2026年4月15日，美国Anthropic公司宣布在Claude模型平台引入强实名身份验证⑪，强制要求用户提供实体政府证件与实时自拍。这种从“匿名工具”向“强实名”的转变，实质上是境外平台利用其模型主导权，将物理世界的身份主权与数字世界进行强制绑定，强化了其对全球智能化生态的非对称控制。

当前，人工智能对主权的影响主要体现在2个层面。① 内容层面，生成式人工智能重塑了生产机制，其携带的价值立场与叙事框架正深刻影响着各国的文化认同与舆论结构；② 执行层面，智能体将主权争端

⑧ European Parliament and Council. Regulation (EU) artificial intelligence act. (2024-07-12) [2026-06-10]. https://artificialintelligenceact. eu; The White House. Executive Order 14110: Safe, secure, and trustworthy development and use of artificial intelligence. (2023-10-30) [2026-06-10]. https://www. federalregister. gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

⑨ NIST Center for AI Standards and Innovation. AI agent standards initiative. (2026-02-17) [2026-06-10]. https://www. nist. gov/artificial-intelligence/ai-agent-standards-initiative.

⑩ United Nations. Global digital compact. (2024-09-22) [2026-06-10]. https://www. un. org/digital-emerging-technologies/global-digital-compact.

⑪ Identity verification on Claude. (2026-04-15)[2026-06-10]. https://support.claude.com/en/articles/14328960-identity-verification-on-claude.

推进至技术接入规则与行为责任的边界。我国在《全球人工智能治理倡议》中明确主张技术应尊重主权⑫，反对操纵舆论与干涉内政。总体而言，人工智能正推动网络空间主权由数据主权向涵盖模型、规则及执行链的综合治理演进，维护智能化主权已成为新时代的重大战略课题。

4.2 国际博弈重心加快转向规则、标准与生态控制权竞争

在人工智能深度重塑网络空间的背景下，国家间的竞争重心已从单纯的技术攻坚转向规则解释权、标准制定权与生态控制权的博弈。尽管国际社会在治理框架上已初步达成共识⑬，但在主权边界等核心议题上依然存在显著分歧。当前博弈的关键，已演变为能否将本国的治理理念有效嵌入国际规则，通过将技术优势加速转化为制度与标准优势，使标准成为治理原则转化为市场准入规则的载体[9]。更深层的博弈正从模型能力延伸至协议与生态的主导权。以模型上下文协议（MCP）为代表的机制，实质上定义了工具接入、权限边界与审计规则，掌握了这些协议便掌握了智能体调用链的核心话语权。从近年频发的“工具投毒”等风险可以观察到，协议与生态的控制权能直接转化为国家安全审计与攻防执行能力的代差⑭。

由此可见，未来国际博弈的焦点将全面覆盖协议定义权、工具分发权，以及由“指令”和“算法”构成的智能执行链控制权。在这种态势下，维护网络空间主权已无法仅依赖传统的物理边界防御，而必须同步提升规则塑造、标准供给，以及智能体生态的自主能力。只有构建起自主可控的智能生态体系，才能在数字化博弈中确保国家主权不受侵蚀。

4.3 内外协同构建自主可控的智能空间主权防御体系

网络空间主权正被人工智能重塑，维护国家网络空间主权的路径需从被动防御转向主动布局，统筹国内能力建设与国际规则参与，形成内外协同的总体治理框架⑮。主权维护的基石在于构建真实可靠的底层技术体系与完整的“主权技术栈”。这要求国家不仅要在高端芯片、基础软件与安全大模型上实现突破，更须在关键接口、审计认证及生态主导等核心环节掌握自主权，确保技术优势能有效转化为国际话语的主导权。

制度塑造能力是主权维护的深层支撑。主权的实现不仅体现在拥有模型，更体现在定义规则。围绕大模型、智能体及其调用协议，加快建立国家级的安全基线与接口规范，形成一套既具兼容性又具安全性的自主规则框架。这种制度供给的核心在于掌握“可信度”与“接入资格”的界定权⑯，从而防止境外平台

⑫ 全球人工智能治理倡议. (2023-10-20) [2026-06-10]. https://www.mfa.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml.

⑬ United Nations. Final report of the open-ended working group on security of and in the use of information and communications technologies 2021-2025. (2025-07-01)[2026-06-10]. https://dig.watch/resource/oewg-report-2021-2025.

⑭ NIST. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. (2025-03-06)[2026-06-10]. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf; National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative artificial intelligence profile, NIST AI 600-1. (2024-07-26)[2026-06-10]. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

⑮ 中华人民共和国个人信息保护法. (2021-08-20)[2026-06-10]. https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm; 中共中央 国务院印发《数字中国建设整体布局规划》. (2023-02-27)[2026-06-10]. https://www.gov.cn/zhengce/2023-02/27/content_5743484.htm.

⑯ 联合国信息安全开放式工作组. 中国关于网络空间国际规则的立场. (2024-01-01)[2026-06-10]. https://www.fmprc.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/zclc_674645/qt_674659/202110/t20211012_9552671.shtml.

利用事实标准对国内系统形成隐性约束，确保智能生态的控制权始终掌握在自己手中。

*最终，维护网络主权需要在坚持主权原则的基础上主动拥抱全球治理。*通过将核心技术、产业生态与国际规则有机结合，依托多边平台在数据跨境与智能体治理等关键议题上提出可操作的方案。只有通过持续增强国际影响力，将国内治理标准转化为全球认可的公共产品，才能在复杂的国际博弈中构筑起坚实的主权防线，真正守住智能化时代国家安全的底线。

5 构建人工智能驱动的安全治理新体系

当前，全球网络空间安全领域正在经历一场深刻的范式变革。传统的专家驱动范式高度依赖安全专家的知识积累和经验判断，在威胁发现、漏洞检测、攻击溯源等核心环节需要专业人员深度参与，这种模式在面对海量数据和复杂关联分析时已显现出能力瓶颈。从国际技术发展趋势来看，以美国为代表的网络空间安全强国已率先完成向数据驱动范式的转型，并正在依托高自治智能体技术加速布局“智能驱动”的战略防线。我国网络空间安全研究需顺应这一历史性趋势，将数据驱动的规模化关联与智能驱动的自主博弈双轨并进，从顶层设计、政策制定、关键技术突破、生态环境构建和人才培养变革等几个方面系统推进。

*强化人工智能驱动的战略顶层设计。*在应对从“专家驱动”向“智能驱动”范式转变的过程中，确立以“技术—工程—生态”为核心的治理体系，将“算力换人力、数据换经验、模型换专家”作为重塑数字时代安全格局的基本方略。通过统筹国内能力建设与国际规则参与，打造内外协同的总体治理框架，确保在安全大模型及关键协议等核心技术体系中实现自主可控，从而推动安全能力从被动补漏转向主动演进与动态博弈，在智能化博弈的国际竞争中夺取战略制高点。

*完善法律法规与标准体系，规范数据和模型治理。*制度保障是范式转型的逻辑起点。推动网络安全治理从“静态合规”向“运行时合规”转变，针对智能体自主拆解任务、频繁调用插件等特质，完善相关法律配套细则。重点确立“指令溯源”与“行为审计”的标准体系，将治理原则转化为可操作、可检测的市场准入规则。通过建立国家级的人工智能安全基线与接口规范，掌握对“可信度”与“接入资格”的界定权，从制度源头防范境外事实标准对国内智能生态形成隐性约束。

*攻克关键核心技术实现安全防御的代际跨越。*为了将网络安全治理从传统的被动处置推向主动免疫，不仅需关注大模型的安全，还需在“Harness工程”方面与大模型深度融合，攻克面向机机博弈的毫秒级战术响应技术。通过研发具备“全域感知”能力的数字神经中枢，构建包含工具链反馈、动态逻辑冗余及沙箱隔离等安全防护架构。推动产业向自动化响应闭环进化，将技术领先优势转化为战略主动权。

*构建协同共享的安全数据与平台生态。*加快建设国家级安全大数据平台，通过统筹各类敏感数据资源，建立分级分类的集约化管理机制。针对智能体及插件系统带来的权限滥用风险，构建覆盖“数据—模型—协议—工具链”的复合治理生态，推动政、产、学、研多方主体共享威胁情报与安全基线。通过确立统一的协议定义权与工具分发权，阻断恶意指令与“技能投毒”的级联扩散，形成互信共赢的智能化主权防御屏障。

*强化交叉人才培养，提升安全与人工智能驾驭能力的协同。*面对攻防范式的根本性重构，人才培养方式也随之改变。在高校及科研院所推动“人工智能安全”交叉学科建设，培养既精通安全领域知识（如软件安全、数据安全），又掌握大模型、智能体驾驭能力的复合型人才。依托国家级实训基地，培养相关人员参与真实场景下的“机机博弈”对抗演练，提升其

面对人工智能攻防时的应对能力。

总体而言，全球网络空间安全正处于从“专家驱动”向“数据驱动”和“智能驱动”范式转变的历史交汇点。我国需在顶层设计的指导下，以制度保障为起点，通过完善法律标准体系掌握治理主动权；以关键技术突破为基石，构建具备主动免疫能力的防御架构；以协同生态为集成引擎，构筑数据与平台联动的安全屏障；以交叉人才为核心驱动，培养具备人工智能驾驭能力的安全人才。最终在智能化博弈的国际竞争中夺取战略制高点，筑牢数字文明的主权屏障。

参考文献

1 苏璞睿, 冯登国. 2024年网络空间安全科技热点回眸. 科技导报, 2025, 43(1): 102-117.
Su P R, Feng D G. Review of cyberspace security science and technology hotspots in 2024. Science & Technology Review, 2025, 43(1): 102-117.

2 中国科学院. 中国学科发展战略-网络空间安全. 北京: 科学出版社, 2025.
Chinese Academy of Sciences. Development Strategy of China’s Disciplines: Cyberspace Security. Beijing: Science Press, 2025. (in Chinese)

3 Liu T, Meng G Z, Zhou P, et al. The art of hide and seek: making pickle-based model supply chain poisoning stealthy again// Proceedings of the 35th USENIX Security Symposium. Baltimore MD: USENIX Association, 2026.

4 徐龙第. 全球网络空间治理:核心问题、中国方案与未来方向. 欧洲研究, 2023, (6): 111-119.
Xu L D. Global cyberspace governance: Core issues, China’s solutions, and future directions. European Studies, 2023, (6): 111-119..

5 Choquette-Choo C A, Tramer F, Carlini N, et al. Label-only membership inference attacks// Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 1964-1974.

6 Yu D. Differentially private fine-tuning of language models. 2022, arXiv: 2110.06500.

7 Zhu J W, Yin H, Deng P, et al. Confidential computing on NVIDIA Hopper GPUs: A performance benchmark study. 2024, arXiv:2409.03992.

8 中国信息通信研究院, 腾讯云. AI Agent安全实践指引. 北京: 中国信息通信研究院, 2026.
China Academy of Information and Communications Technology, CloudTencent. Security Practice Guidelines for AI Agents. Beijing: China Academy of Information and Communications Technology, 2026. (in Chinese)

9 张凌寒. 人工智能法律治理的路径拓展. 中国社会科学, 2025, (1): 91-110.
Zhang L H. Path expansion of legal governance for artificial intelligence. Social Sciences in China, 2025, (1): 91-110.

Artificial intelligence for cybersecurity: Opportunities, challenges, and approaches

CHEN Kai[1] LI Ding[2] MENG Guozhu[1] JI Shouling[3] LI Changjiang[3] YANG Yi[1] FENG Dengguo[4*]

（1 Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China;

2 School of Computer Science, Peking University, Beijing 100871, China;

3 College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China;

4 Institute of Software, Chinese Academy of Sciences, Beijing 100190, China）

Abstract Cybersecurity research, institutional structures, and governance policies are undergoing profound transformations. Currently, increasingly covert and rapidly evolving intelligent attacks, coupled with the national urgent expectations for high-level security, are driving significant shifts in the roles and interactions of governments, research institutions, and enterprises. Consequently, this study, based on analyzing the challenges and opportunities of the AI era, explores core application scenarios such as critical information infrastructure protection, national data security, and the maintenance of cyberspace sovereignty. It provides an analysis of artificial intelligence in dimensions such as correlation and causality, and proposes a governance framework, aiming to provide insights for accelerating the construction of a cyberpower and fortifying the digital security barrier.

Keywords artificial intelligence, cybersecurity, critical information infrastructure protection, intelligent attack and defense

陈 恺 中国科学院信息工程研究所研究员。主要研究领域：软件与系统安全、人工智能安全。E-mail: chenkai@iie.ac.cn

CHEN Kai Professor in Institute of Information Engineering, Chinese Academy of Sciences (CAS). His research focuses on software and system security, AI security. E-mail: chenkai@iie.ac.cn

冯登国 中国科学院院士。中国科学院软件研究所研究员。主要研究领域：网络与信息安全，数据安全与隐私保护，可信计算与机密计算。E-mail: fengdg@263.net

FENG Dengguo Academician of Chinese Academy of Sciences (CAS). Professor in Institute of Software, Chinese Academy of Sciences (CAS). His research focuses on cybersecurity and information security, data security and privacy protection, trusted computing and confidential computing. E-mail: fengdg@263.net

■责任编辑：文彦杰

*Corresponding author