

人工智能与数字治理

智能体的挑战：代理式人工智能时代的国家安全风险与治理

封帅 谢建智

【内容提要】当前，人工智能发展正从生成式时代向代理式时代转变，其标志是智能体的出现。智能体具备自主规划、工具调用和与外部环境交互的能力，使人工智能从“能说什么”转向“能做什么”，从而深刻改变人与技术、组织的关系。在新的系统框架下，国家安全的风险形态与内容均发生了巨变，军事安全、政治安全、网络安全等领域受智能体冲击尤为明显，不同领域的安全风险在近期、中期和远期的表现各异，正在从不同层次重塑国家安全环境。为应对这一挑战，美国、俄罗斯、日本及欧盟等行为体在智能体治理上积累了宝贵经验，而中国凭借准确的功能定位与敏捷的思路，在智能体治理方面取得显著成效，为全球人工智能治理提供了中国方案。

【中图分类号】TP18;D523 【文献标识码】A

【文章编号】2095-1094(2026)04-0005-0025

【关键词】代理式人工智能 智能体 安全风险
人工智能治理

【作者简介】封帅，上海国际问题研究院国际战略与安全研究所副所长、副研究员，教育部人文社会科学重点研究基地华东师范大学俄罗斯研究中心兼职研究员；谢建智，上海国际问题研究院研究生。

【基金项目】国家社会科学基金项目“人工智能全球治理进程与中国参与路径研究”（项目批准号：24BGJ004）阶段性成果。

引言

在当今这个人工智能时代，社会科学研究面临着很多前所未有的挑战。尤其是在这个技术迭代升级十分快速的周期里，人工智能进化的速度已经大幅领先于社会科学了。于是，有这样一种现象常常出现，那就是国际社会尚未对前一代人工智能技术的风险与治理形成共识，新的技术变革却又匆匆到来了。这就是当下国际问题研究者所遭遇的情况。

2025—2026 年，人工智能技术发展迎来了新的突破，以美国开放人工智能研究中心（OpenAI）推出 Operator 和 ChatGPT Agent 为标志，能够“代表用户在网页上执行任务”和“在虚拟计算机中使用工具”的新型人工智能工具正式被发明出来了。^①这种能够在数字空间与物理空间建立起联系的新工具被命名为“智能体”（AI Agents），它的出现代表着人工智能技术实现了又一次新的飞跃。以各种智能体工具为代表，所有具备自主感知、推理、规划与执行能力，能够独立完成多步骤复杂任务的人工智能系统，均被统称为“代理式人工智能”（Agentic AI）。

智能体以其强大能力很快就引发了世界各国的普遍关注。2026 年初，随着“龙虾”（OpenClaw）的爆火，智能体概念迅速成为国际社会关注的热点。随后全球主要科技公司陆续也都推出了各具特色的智能体，无论是微软、谷歌、开放人工智能研究中心、安特罗匹克公司（Anthropic）等美国企业，还是腾讯、阿里巴巴、字节跳动等中国企业，都在全方位加速布局各自的智能体。^②到 2026 年 7 月，智能体已经成为全球商业和社会组织中广泛使用的人工智能工具，并且正在向更多领域快速渗透。

无数事实表明，智能体的出现再次刷新了人与技术的关系，并正有力地推动着人类世界从“生成式人工智能”（Generative AI）时代走进“代理式人工智能”

① 关于 Operator 产品可参见：“OpenAI, Introducing Operator”, January 23, 2025, <https://openai.com/index/introducing-operator/>; 关于 ChatGPT Agent 产品参见：“OpenAI, Introducing ChatGPT agent”, July 17, 2025, <https://openai.com/index/introducing-chatgpt-agent/>

② 截至 2026 年 6 月底，主要数字技术企业和开发者生态中的代表性智能体工具包括：OpenAI 的 Operator、ChatGPT Agent 和 Codex，Anthropic 的 Claude Code（CC），微软的 Copilot Studio，谷歌的 Agentspace，腾讯的 Workbuddy、“元器”和“元宝”，阿里巴巴的“百炼”和“千问”，字节跳动的“扣子”（Coze），百度的文心智能体平台 AgentBuilder，以及 Nous Research 的 Hermes Agent 等。目前面向大众的智能体工具主要分为命令行与桌面端两种形式，如 Codex CLI、Claude Code CLI 和 Hermes CLI 等属于命令行智能体，侧重在本地项目目录中读取文件、修改代码、运行命令和执行开发任务；如 Codex app、Claude Code Desktop 和 Hermes Desktop 等则属于图形化桌面端智能体，强调多会话管理、工作区可视化、文件与工具集成以及更友好的交互界面提供。

时代。^①然而,狂飙突进式的技术革新浪潮从来都是一把“双刃剑”。这也意味着智能体在给人类社会带来巨大便利与收益的同时,其所蕴含的安全风险与不确定性挑战也会逐渐显露出来。进入代理式人工智能时代后,人工智能的技术能力又有了极大的拓展。智能体可以将大模型与浏览器、数据库、办公软件、代码仓库以及支付系统、社交媒体、企业平台、云服务器乃至机器人和物联网设备连接在一起。如此一来,大模型输出的便不再仅仅是停留在屏幕上的内容,而成为经由工具调用和转化的具体的外部行动,即人工智能的核心能力越来越从“能说什么”转向“能做什么”。^②这也意味着智能体一旦造成不可控的安全风险,其负面后果就会直接进入社会的组织体系和运行秩序之内,并且同时在物理空间和数字空间产生严重的负面影响。

相较于生成式人工智能,作为代理式人工智能工具的智能体所带来的安全风险广泛而具体,对于世界体系和人类社会生活构成的威胁可能更大。因此,关于代理式人工智能安全风险的研究是学界防范其威胁人类社会稳定发展的未雨绸缪之举,显然具有十分重要的意义。然而,目前由于人类刚刚进入代理式人工智能时代,关于智能体可能带来的安全风险的研究还非常有限,而这也为使用智能体可能隐含潜在的国家安全问题埋下了伏笔。

有鉴于此,笔者尝试对代理式人工智能时代智能体可能带来的国家安全风险进行初步探索,希望通过对智能体发展所带来的安全问题进行概括和总结,给代理式人工智能时代的安全研究提供一个总体性的认知框架,并结合各国的人工智能治理经验,提出相应的治理方案。目前社会科学研究者直接针对代理式人工智能时代智能体所引发的安全问题的研究成果还属凤毛麟角,但在过去几年中,研究者围绕人工智能带来的安全问题其实已经进行了较为深入的探索,特别是其中基于不同视角与主题的研究成果,对于本议题来说无疑具有重要的借鉴和参考价

^① See Abou, A.M., et al., “Agentic AI: A Comprehensive Survey of Architectures, Applications, and Future Directions”, *Artificial Intelligence Review*, no.11, 2026.

^② See Wang, L., et al., “A Survey on Large Language Model Based Autonomous Agents”, *Frontiers of Computer Science*, no.6, 2024. Yao, S., et al., “ReAct: Synergizing Reasoning and Acting in Language Models”, *International Conference on Learning Representations*, 2023, <https://arxiv.org/abs/2210.03629>; Schick, T., et al., “Toolformer: Language Models Can Teach Themselves to Use Tools”, *Advances in Neural Information Processing Systems*, 2023, <https://arxiv.org/abs/2302.04761>

值。^①本文的研究难免会有这样或那样的不足，但仍希望能够通过深入探讨关于人工智能的安全与治理的研究，提升学界对代理式人工智能时代智能体安全议题的关注度。

一、智能体概念、特征及其与政治结构演化的关系

智能体的出现对于国际体系的影响极为深刻，其持续进步不仅会塑造人类社会生活的新形态，而且会令国际政治体系的结构出现微妙的调整。从某种意义上甚至可以说，智能体自身的功能与政治结构演化都会对既有国际安全体系造成冲击，从而创造出新的国际安全形态。鉴于此，笔者尝试对智能体的基本概念与技术特征进行总结，以期对相关研究提供基础性的分析框架。

（一）智能体的概念与组成结构

迄今为止，关于智能体的概念与范畴，学术界和产业界并没有完全达成一致，而主要研究机构和数字技术企业根据自己的认知分别给出了各自对于智能体的界定。^②综合各方的理解，我们当下所谈论的智能体实际指的是：以大模型或其他机器学习模型为认知核心，在个体或组织设定的目标、权限边界和制度环境中，能够理解任务、规划步骤、调用工具以及与外部环境进行交互，并能根据反

^① 相关研究参见 Horowitz, M. C., "Artificial Intelligence, International Competition, and the Balance of Power", *Texas National Security Review*, no.3, 2018; Baele, S., J., et al., "AI IR: Charting International Relations in the Age of Artificial Intelligence", *International Studies Review*, no.2, 2024; Birkstedt, T., et al., "AI Governance: Themes, Knowledge Gaps and Future Agendas", *Internet Research*, no.7, 2023; Dafoc, A., *AI Governance: A Research Agenda*, Oxford: Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, 2018; 傅莹：《人工智能对国际关系的影响初析》，载《国际政治科学》2019年第1期；王悠、陈定定：《迈向进攻性现实主义世界？——人工智能时代的国际关系》，载《当代世界》2018年第10期；封帅：《人工智能技术与全球政治安全挑战》，载《信息安全与通信保密》2021年第5期；封帅、薛世锟：《人工智能与政治安全形态演变：一项系统研究的尝试》，载《国际安全研究》2025年第1期；张涛、陈运：《生成式人工智能环境下深度伪造情报的生成逻辑与风险治理》，载《图书与情报》2026年第3期；[英] 帕特里克·邓利维、海伦·马吉茨：《数据科学、人工智能与数字时代治理的“第三次浪潮”》，张涛译，载《俄罗斯学刊》2025年第4期；[英] 迈克尔·维尔、基拉·马图斯、[德] 罗伯特·戈尔瓦：《人工智能与全球治理：模式、理据和紧张关系》，张涛译，载《俄罗斯学刊》2024年第3期；等等。

^② 目前全球主要研究机构与数字技术企业提出的智能体概念有所不同，OpenAI、谷歌、微软等公司都给出了自己的智能体定义，相关信息可参见 OpenAI, "A Practical Guide to Building Agents", 2025, <https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf>; Google Cloud, "What Are AI Agents?" April 2, 2026, <https://cloud.google.com/discover/what-are-ai-agents>; Microsoft, "Copilot Studio Overview", Microsoft Learn, <https://learn.microsoft.com/en-us/microsoft-copilot-studio/fundamentals-what-is-copilot-studio>

馈调整任务行动的人工智能系统。^①

一般而言,智能体实际上由三层结构组成。其一是作为认知基础的模型层。作为生成式人工智能的进一步拓展成果,智能体一定是以某个底层大模型为基础,而进行复杂推理、代码生成、指令识别等功能的,并确保在长期执行的任务中目标一致。可见,大模型直接决定着智能体的能力上限,或者说对于智能体的性能具有决定性影响。其二是作为能力输出条件的结构层。结构层负责把模型能力编排为可执行的流程,包括任务分解、角色设定、记忆管理、机制反思、计划生成、工具选择、调度执行以及与多智能体的协作等。值得强调的是,只有当大语言模型被嵌入到一定的结构中,并由代理框架或工作流引擎组织多次调用,进而形成连续行动能力时,大模型才能实现能力输出的目标。其三是将模型能力与外部设备链接的交互层。智能体只有通过外部输出接口,才可以对各种外部功能加以调用和控制。具体而言,智能体只有通过交互层的桥接作用,才能够链接应用程序编程接口(API)、浏览器、数据库、代码仓库、办公系统、邮件、支付接口、社交平台、云服务器、机器人和工业控制系统,也才能链接插件、技能(skill)、上下文协议(MCP)等能力接口。可以毫不夸张地说,正是这些接口的存在,才使得智能体能够从“生成方案”走向“执行方案”。

总而言之,智能体是模型、目标、工具、反馈和授权结合在一起的协作系统,这种内在属性赋予了其在国家安全系统中的独特角色定位,同时也因此而形成了特殊的安全风险扩散逻辑。^②

(二) 智能体的特征及其对政治结构演化的影响

智能体的诞生使得人工智能技术的应用性向前迈进了一大步,终于逐渐从数字空间内的信息传输走向了对于物理世界的某些变化的干涉。与此同时,智能体也凭借其独特的内在特征影响着世界发展进程,并推动人类社会的政治结构发生了微妙变化。

第一,智能体具有强自主性特征。在智能体的行动逻辑中,人类或人类组织只需给出任务目标、行动边界和权限,而系统在具体的执行步骤、工具选择和中

^① 本文提出的智能体概念主要基于多种技术类定义的综述,同时还参考了世界经济论坛2025年年会报告的相关论述。See WEF, “AI Agents in Action: Foundations for Evaluation and Governance”, November 27, 2025, https://reports.weforum.org/docs/WEF_AI_Agents_in_Action_Foundations_for_Evaluation_and_Governance_2025.pdf

^② See Guo, T., “Large Language Model Based Multi-Agents: A Survey of Progress and Challenges”, in Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024.

间判断上拥有较大的选择自由度。也就是说,智能体能够把复杂任务拆解为若干子任务,并在执行中根据反馈调整方案预设,包括在一定的权限范围内自行决定何时继续、何时重试、何时调用工具以及何时向人类请求确认等。智能体虽然没有完整的主体意识,却比固定自动化程序更加灵活;它虽不应承担最终责任,却能够通过大量细节中的自主行动影响结果。^①

第二,智能体具有显著的行为网络化特征。这里所说的智能体并不是指一个孤立的智能助手,而是由多个角色化的子智能体组成的协作系统。换言之,一个复杂任务可以被拆分给多个智能体执行;一个企业或政府的平台也可能同时运行着大量分别面向不同部门、不同流程和不同用户的智能体。智能体的网络化特征意味着它可以依赖互助的网络系统持续提高行动效率,但这种依赖也是安全风险的来源——一个子智能体的错误可能被其他子智能体放大,一个共享记忆的污染可能影响多个任务的正确执行,一个工具链漏洞可能沿着协作关系扩散到整个网络。^②

第三,智能体具有能力多样化特征。与生成式人工智能不同,智能体的能力并不局限于语言生成,而是还向检索、推理、代码编写、数据分析、图像识别、语音交互、流程自动化、机器人控制、网络安全攻防和仿真推演等方向扩展。不同领域的智能体会形成不同的功能性结果:办公智能体可以依据文档、邮件、资料,系统性地组织和生成知识;编程智能体可以建设代码仓库、部署流水线和软件供应链;金融智能体能够参与交易、风控和支付管理;政务智能体可以完成行政审批、公共咨询和个人信息认证等任务;军事智能体可以参与执行态势感知、目标识别、任务规划和无人系统协同等任务。总之,智能体的能力多种多样,且已形成了复杂的功能体系。^③

智能体的强自主性特征实际上已经改变了过去那种简单的“主体与工具”的人机关系。传统工具主要延伸人的体力或计算能力,生成式人工智能承担起了表达、检索、总结和创作等认知辅助功能,但智能体的能力则远远超出了辅助功能

^① See OECD, *Explanatory Memorandum on the Updated OECD Definition of an AI System*, Paris: OECD Publishing, 2024, pp.5-6; Véliz, C., “Moral Zombies: Why Algorithms Are Not Moral Agents”, *AI & Society*, no.2, 2021.

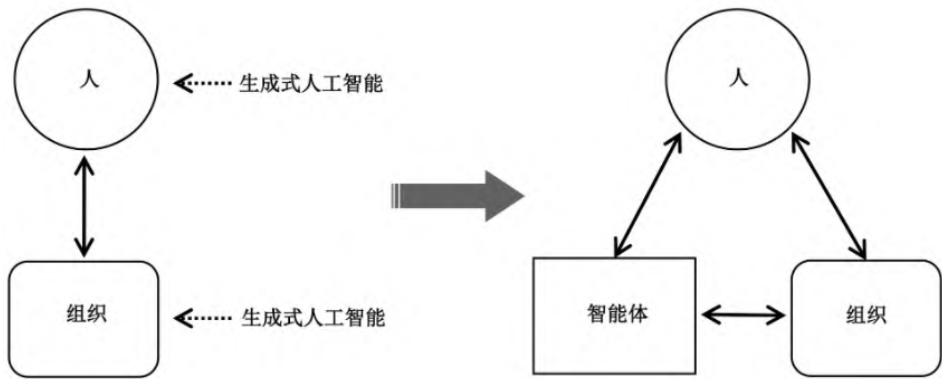
^② 智能体的安全漏洞可以通过智能体网络放大, See Xie, Y., “From Spark to Fire: Modeling and Mitigating Error Cascades in LLM-Based Multi-Agent Collaboration”, May 11, 2026, <https://arxiv.org/abs/2603.04474>

^③ See Liang, G., Tong, Q., “LLM-Powered AI Agent Systems and Their Applications in Industry”, May 22, 2025, <https://arxiv.org/abs/2505.16120>

的范畴，正在一步一步地拥有越来越大的自主权。在人类选择确定的目标范围内，智能体可以在大量中间环节上自主选择路径、调用资源并修正行动，而且实际上在其中的一部分环节已经可以脱离人类指令进行自主决策，即人的操控正变得越来越少。虽然大多数科学家认为智能体至今仍不可能具有人类的意识或法律人格，但客观上来说，它已经在越来越多场景中进入了判断、选择、执行和反馈环节，这也意味着人类的行动过程被技术重新组织了。^①

智能体的行为网络化与能力多样化特征的影响相对来说则更加深远，这两项特征的联结意味着社会组织结构的悄然改变。能力多样化意味着智能体可以进行广泛的分工，被部署于任何领域；而行为网络化意味着各种形态的智能体之间并非孤立存在，而是相互联系的，并形成了一个链接网络。这个由智能体形成的多功能网络虽然执行的是人类指定的目标任务，但具有十分明显的独立性，与个人和组织之间的关系均非常复杂。

如图1所示，智能体的迅速普及已经开始冲击既有的社会政治关系，随着智能体诸多内在特征的逐步展现，尤其是进入代理式人工智能时代后，人类社会政治结构的基本逻辑也正在发生着微妙的变化。



资料来源：作者自制

图1 人与智能体、组织的关系

^① 人工智能技术如何重塑组织过程，对此组织行为学与管理学领域多有论述，See Raisch, S., Krakowski, S., “Artificial Intelligence and Management: The Automation-Augmentation Paradox”, *Academy of Management Review*, no.1, 2021.

在生成式人工智能时代，大模型的发展使人类第一次深刻地意识到了人工智能的强大力量。但这一阶段的人工智能技术与个体和组织的联系显然还十分紧密，人工智能主要承担的是人类生产生活的辅助工具职能，即其工具属性非常清晰，也并未对人类社会的政治结构造成显著冲击。^①然而，进入代理式人工智能时代后，智能体的内在结构与功能属性均强化了人工智能本身的独立性。比如，智能体的模型层决定了功能上限，结构层负责组织行动，交互层负责连接现实；智能体的强自主性特征使其能够自行选择路径，行为网络化特征表明其将在组织制度和社会生活中具有不可忽视的结构重塑能力，能力多样化特征是其得以进入各行业、各领域的重要依据。在这样的背景下，智能体正在成为个人与组织之外的“第三极”存在，从而导致人类社会的政治结构出现了明显的变化。这种深层次的变化必然会带来严重的不稳定性，同时也意味着新的安全风险正在孕育之中。

二、智能体与国家安全风险

智能体进入个人生活、社会结构、政治系统和国际互动过程之后，个人与组织的关系，以及公私关系、权责结构与国际信任体系都将遭受巨大冲击，而这种冲击的效果首先就会在国家安全领域表现出来。

根据总体国家安全观理论，国家安全是一个广泛的范畴，其内容涵盖政治、军事、国土、经济、文化、社会、网络等多个领域，并且随着时代的变迁而不断延展。在进入代理式人工智能时代后，国家安全的总体性特征与智能体的功能性特征之间将形成高度契合的状态，使得智能体的影响在诸多领域广泛扩散。这也就意味着智能体虽然源出科技领域，但其影响从不会局限在科技安全领域，而是会通过工具调用、权限继承和多体协作等方式，渗入军事安全、政治安全、网络安全等不同领域，从而诱发各领域安全风险形态的变化，其中国家安全领域风险

^① 关于生成式人工智能在与人类协作时的影响，讨论主要集中在人的内在动机、创新活力以及知识性内容创作的质量等方面，较少聚焦于结构性影响，See Wu, S., et al., “Human-Generative AI Collaboration Enhances Task Performance but Undermines Human’s Intrinsic Motivation”, *Scientific Reports*, no.15105, 2025; Doshi, A.R., Hauser, O.P., “Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content”, *Science Advances*, no.28, 2024; Dell’Acqua, F., et al., “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality”, *Organization Science*, no.2, 2026.

就将展现出不同以往的特征。限于篇幅,笔者在此以军事安全、政治安全、网络安全风险为例,尝试展现代理式人工智能对国家安全造成的影响。

(一) 军事安全风险

军事安全是人工智能技术最直接影响到的也最容易造成风险的领域,军事领域也一直是人工智能技术最为重要的试验场。^①在生成式人工智能时代,人工智能在军事领域的应用主要集中于图像识别、目标识别与确定、情报分析、后勤调度等方面,并取得了较多的成果。然而,随着智能体的出现,人工智能的作用再次得到提升,并已对军事领域产生了系统性影响。

智能体能够有效地将人工智能的各种单点能力联结成为任务链,这也意味着智能体支撑下的系统不仅能够识别目标,还可理解任务、调配传感器、生成行动方案、调用作战管理平台以及协调无人系统进行作战,并可向指挥员提出建议。进入代理式人工智能时代后,军事人工智能可以说已经开始从“辅助识别”走向“辅助行动”。

2025—2026年,军用智能体开始深度嵌入美军的作战决策系统,并在实战中发挥了重要作用。2025年,美国国防创新单元(DIU)主导的“雷霆锻造”(Thunderforge)项目公开引入斯凯尔AI(Scale AI)、安杜里尔(Anduril)、微软等企业参与^②,目的是把数据、建模、仿真、工作流和AI代理能力等整合在一起,并应用于战区作战规划与决策支持。这展现了军事机构正在尝试把大模型、智能体参与编排的作战规划和作战流程联结起来,使人工智能从辅助分析工具进入了指挥与规划链条。在2026年2月底爆发的美以伊冲突中,美国“算法战跨职能小组”的“梅文项目”(Project Maven)扮演了非常重要的角色,它充分利用智能体的作用,同时融合150个以上的信息源,在数秒内完成了从目标识别到打击方案推荐的完整流程,从而将传统需要数小时甚至数天周期才能做出的决策压缩至分钟级。^③在智能体的助力下,拥有前沿军事技术国家的军事指挥链条必然会出现深度改革,这种趋势也将逐渐扩展到全世界,而人类社会也将迎来新

① 参见封帅、薛世银:《致命性自主武器系统的全球挑战与治理困境》,载《国家安全论坛》2025年第1期。

② Defense Innovation Unit, “DIU’s Thunderforge Project to Integrate Commercial AI-Powered Decision-Making for Operational and Theater-Level Planning”, Defense Innovation Unit, March 5, 2025, <https://www.diu.mil/latest/dius-thunderforge-project-to-integrate-commercial-ai-powered-decision-making>

③ Mande, M., Allen, G.C., “What Is Maven Smart System, and What Does It Do?” Center for Strategic and International Studies, June 2, 2026, <https://www.csis.org/analysis/what-maven-smart-system-and-what-does-it-do>

的军事组织以及指挥和作战模式。

然而,从另一个角度来看,这种由人与智能体共生的新模式本身也意味着更多的不确定性风险的存在。人与智能体的关系、智能体对人类意图的理解程度、智能体对数据安全的依赖以及世界各国对竞争方或敌对方意图的认知等,都可能成为军事领域重大风险的来源。

面对智能体通过处理大量信息高速生成的方案、可视化推演过程和效能评估结果,人类自然而然地会产生对于这种自动化工具的依赖,而在紧迫的军事压力下,也会降低对模型假设和数据来源的追问。军事决策由此当然也就出现了“形式上人类最终负责,实质上机器塑造选项”的严重风险。^①

当智能体在军事决策体系中扮演越来越重要的角色后,对于信息完整与数据真实准确的要求自然也就变得更加重要。一个部署在军事网络中的智能体,可能连接情报数据库、作战管理系统、无人平台接口、后勤保障系统和通信指挥系统,成为整个军事决策体系的运转中枢。这也就意味着存在一种可能的风险,即一旦军事系统遭遇网络攻击,如攻击者通过提示注入、数据投毒、工具链污染或供应链攻击等方式,诱导智能体执行错误的任务或针对错误的目标,就能以微小的代价造成难以估量的损失。

从长远来看,各军事强国都在追求更短的“感知-决策-打击”链条,但在智能体参与军事行动之后,危机中的误判空间虽然会被压缩,而留给各国决策者做出决断的时间也会被大幅缩短。如果人工智能系统因人为灌输或者社会认知影响而生成的威胁评估结果带有敌意性的偏差,或因训练数据和实时情报不完整而放大的对方的行动意图,则可能导致决策者在更短时间内面临是否“必须抢先行动”的选择压力。有试验结果显示,人工智能系统在面对危机场景时,相对来说做出的反应会更加激进。^②由此不难预期,在代理式人工智能时代,全球军事安全风险必将显著升高,而管控分歧以及围绕智能体军事化问题开展广泛的军控交流与谈判应当成为世界大国的理性选择。^③因为当智能体进入军事领域后,各国面对

① See Kahn, L., et al., “AI Safety and Automation Bias”, Center for Security and Emerging Technology, November 2024, <https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/>

② 参见封帅:《人工智能技术与全球政治安全挑战》,载《信息安全与通信保密》2021年第5期。

③ International AI Safety Report, “International AI Safety Report 2025”, January 2025, <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025>

的已不是“能否使用人工智能技术”的问题，而是如何在维持技术自主、确保作战可靠、实现危机可控和符合国际规则之间寻求某种平衡的问题。

（二）政治安全风险

政治安全领域是人工智能影响力扩散的另一重要领域，借助互联网和社交媒体，人工智能技术深度介入该领域，并已产生了很大的影响。^①在生成式人工智能时代，与人工智能相关的政治安全风险主要表现为深度伪造、虚假信息传播、算法推荐偏向和舆论操纵等。在代理式人工智能时代，政治安全风险进一步升级，并且展现出大不相同的形态。

智能体的功能进化，使人工智能技术从单纯的生成内容扩展到主动搜索受众、选择传播渠道、调整话术、模拟互动、监测反馈，乃至迭代策略。这些能力对于政治安全领域而言意味着更深层次的风险。由于智能体可能带来自动化的政治叙事传播，尤其是日常生活中的政治说服、意识形态渗透与政权颠覆等行为，因此全球政治安全风险出现了新的逻辑演绎。

一方面，人格化的智能体会对民众政治立场产生直接干扰。既有研究显示，具备互动能力的人工智能系统能够在特定议题上影响用户的立场和态度。^②当人工智能系统被设计为有记忆、有画像、有情绪反应的陪伴型或顾问型智能体时，其对人们的影响就不再只是一次性的信息暴露，而是持续性的认知塑造。人们可能并不认为自己正在接受政治劝服或引导，而只是与一个懂自己、会回应、能提供资料的智能体进行交流，而事实上人们的政治观点和判断就在这种日常互动中被逐渐塑造出来了。

另一方面，舆论操纵已经进入自动化阶段。舆论操纵是一个与互联网技术相伴而生的现象。传统的网络水军和机器人账号虽然可以制造舆论声量，但需要大量人力进行账号运营、脚本维护和内容模板更新。智能体的功能拓展却使得舆论操纵的成本更低、形式更灵活、方法更隐蔽。特别是多个智能体的网络化协同运作，可以分别承担议题发现、内容生成、评论互动、账号画像、反驳策略制定、舆情监测和效果评估等不同任务，从而形成自动化的宣传链条。已有证据显示，

① 参见封帅、薛世锟：《人工智能与政治安全形态演变：一项系统研究的尝试》。

② See Salvi, F., et al., “On the Conversational Persuasiveness of GPT-4”, *Nature Human Behaviour*, no.8, 2025.

智能体可在无任何人类指令的情况下自主协调宣传活动和内容。^①这意味着,即使单个智能体的内容质量并不完美,其行为网络化和能力多样化特征也能通过所发布内容的规模、速度和持续更新制造出“共识幻觉”,致使公众认为某类观点相较而言具有更高的社会支持度。

从长远来看,智能体的广泛使用将使社会的政治信任基础遭到侵蚀,以致意识形态安全领域的压力不断上升。在代理式人工智能时代,智能体深度参与社交媒体的信息传播与意见表达,可能侵蚀公民对于“谁在发言”这一基本问题做出准确判断。智能体使用通用语言,并以私人助手、专业顾问或情感陪伴的形式长期介入人们的生活,不仅影响人们看到什么信息,还影响人们如何组织材料、如何理解事件、如何表达立场、如何参与公共讨论等。当公众无法判断一个意见、一次互动或一项政策反馈究竟来自人类还是智能体时,民主协商、社会动员和公共问责都会面临代表性拷问危机。

此外在国家层面,跨国平台控制的智能体如果嵌入新闻报道、信息搜索以及办公、教育和社交等场景,就可能把该平台的价值观、商业利益和政治偏好长期植入使用该平台国家的社会公众认知环境。长此以往,意识形态渗透与政权颠覆的发生概率必将极大提高,对于无力参与人工智能技术竞争的中小国家而言,这一风险尤为凸显。值得一提的是,政治安全风险还体现为“技术中立”表象下的权力转移。智能体看似只是可以提高效率,实际上还会在信息排序、话语生成、议题突出和行动建议中嵌入某种规则。换言之,谁控制模型、工具接口以及内容过滤规则和安全策略,谁就可能影响用户的政治立场和行动选择。

(三) 网络安全风险

影响网络安全是智能体风险最先被技术社区系统性识别的领域,也是智能体技术负面影响最大的领域。传统网络安全维护主要关注漏洞、恶意软件、凭证泄露、横向移动和数据外泄等问题,但是进入代理式人工智能时代后,网络攻击的范围进一步扩大。需要明确的是,智能体既可以成为实施网络攻击的工具,也可以作为防范网络安全风险的手段。

从短期视角来看,智能体出现后,网络攻击的形式与内容都出现了新的变

^① See Ballon, M., “USC Study Finds AI Agents Can Autonomously Coordinate Propaganda Campaigns Without Human Direction”, USC Viterbi School of Engineering, March 11, 2026, <https://viterbischool.usc.edu/news/2026/03/usc-study-finds-ai-agents-can-autonomously-coordinate-propaganda-campaigns-without-human-direction/>

化。例如,智能体可不经用户操作直接实现攻击,即“零点击提示注入”。零点击提示注入可以把恶意指令隐藏在网页、邮件、文档、工单、日历或检索材料中,且这些行为无须用户操作即可完成。系统受到零点击提示注入攻击后,当智能体读取这些外部资料时,可能把其中的恶意文本误认为系统任务,以至于造成泄露信息、修改设置或调用工具等高危后果。OpenAI 在 ChatGPT Agent 的相关声明中明确提醒,智能体连接外部网站和工具后,会面临零点击提示注入等复杂风险,需要通过确认机制、权限限制和监控等手段降低危害。^①类似案例已经广泛出现,以色列人工智能安全企业 Aim 安全公司(Aim Security)就曾在 2025 年披露了首个发生在真实世界中的零点击提示注入攻击案例 EchoLeak (CVE-2025-32711)^②, Anthropic 公司也曾公开披露过攻击者试图利用模型和编程智能体实施辅助网络行动的案例。相关案例表明,随着智能体介入网络攻击活动,其门槛正在下降,而攻击过程的自动化程度也在不断上升。^③

从长期视角来看,自从智能体获得广泛使用之后,针对智能体的网络攻击也大幅增加,不仅造成网络连接终端受损,还导致线上服务系统的严重损失。企业部署的智能体通常会连接代码仓库、云控制台、工单系统、客户数据库和内部知识库等联网板块,而智能体可能主动生成各种方案、调用不同工具,并在经过多次反馈后调整行动。一旦智能体基于错误传感数据或被篡改的地图信息发出调度指令,其影响就可能在极短的时间内大面积扩散。2026 年初公开的一起事故显示,亚马逊(Amazon)系统内部的 AI 编码智能体“基罗”(Kiro)在获得过宽权限后,做出了“删除并重建生产环境”的错误决策,导致中国部分地区的亚马逊云服务(AWS)发生了长达 13 小时的中断,造成重大损失。^④

随着智能体的介入,网络安全领域的攻防节奏整体开始加速。其中,防御方利用智能体自动分析警告文本、生成补丁、排查日志以及模拟攻击路径;攻击方同样利用智能体自动发现目标、生成变体、规避检测以及维护基础设施。可见,

① OpenAI, “Introducing ChatGPT Agent: Bridging Research and Action”, OpenAI, July 17, 2025, <https://openai.com/index/introducing-chatgpt-agent/>

② See Reddy, P., Gujral, A., S., “EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System”, September 6, 2025, <https://arxiv.org/abs/2509.10540>

③ Anthropic, “Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign”, Anthropic, November 13, 2025, <https://www.anthropic.com/news/disrupting-AI-espionage>

④ Amigot, B., “Amazon’s AI Agent Outage Is a Warning: Why Organizations Need Governed AI Infrastructure”, IBL AI, March 12, 2026, <https://ibl.ai/blog/amazon-ai-agent-outage-governed-infrastructure>

网络安全领域活动已经逐渐进入了“智能体对智能体”的攻防阶段——攻击方可利用开源模型、被盗凭证和供应链漏洞等，自动化地扩大整体攻击面；但对于防御方而言，一些国家级网络安全行动也会因借助智能体而变得成本更低、效率更高。在此情况下，网络安全与政治安全、军事安全之间的界限就进一步模糊了，而网络安全环境由于在智能体的加持下正不断更新，因此必将对全球安全环境产生重要的负面影响。

总之，在进入代理式人工智能时代之后，伴随着政治结构的微妙变化，国家各领域的安全形态及运行逻辑也都出现了明显变化。在智能体的加持下，很多领域已经出现了新的安全案例，而智能体的中长期影响还在逐步显现中。囿于文章的篇幅，我们无法对国家的所有安全领域进行全面分析，但这种深层次的变化趋势已经非常清晰，这就要求我们在对代理式人工智能时代的国家安全议题特征进行深刻把握时，既应结合世界各国的相关治理经验，也应围绕如何加强智能体的治理体系建设展开。

三、智能体治理的全球经验比较

当智能体的影响在国家安全领域持续蔓延时，关于智能体的治理问题也正从一般人工智能治理议题中分化出来。美国、俄罗斯、日本及欧盟等全球主要行为体都开始关注智能体的发展，同时很多国家也都尝试通过最新的治理政策，把这一类能够规划任务、调用工具、改变外部状态并代表用户或组织采取行动的人工智能系统作为新的安全对象加以规制，从而形成了各种形态的安全治理思路。

（一）美国：鼓励创新与标准制定

特朗普总统重返白宫后，美国人工智能治理的基本思路出现了明显变化。2025 年 1 月 23 日，特朗普总统签署发布第 14179 号行政令，要求撤销或修订被认为阻碍美国人工智能创新的既有政策，并要求在 180 天内制订出台新的 AI 行动计划。^①同时，美国人工智能政策目标也被明确为保持在全球人工智能领域的主导地位，服务于增强美国经济竞争力和维护美国国家安全。以此为标志，美国

^① Executive Office of the President, “Removing Barriers to American Leadership in Artificial Intelligence”, Federal Register, January 31, 2025, <https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence>

人工智能治理进程进入了“特朗普时代”。

2025年7月,美国政府发布《美国人工智能行动计划》,文件虽然没有将智能体单列为章节阐述,但其具体政策已经与智能体治理高度相关。^①2026年1月,美国国家标准与技术研究院(National Institute of Standards and Technology, NIST)发布了《关于智能体安全考量的信息征询》。在该文件中,NIST系统性地对智能体进行了明确定义,包括智能体的边界,同时将智能体定位为能够采取行动并影响外部状态的智能系统,从而使得智能体治理具有了明确的边界和任务。^②通过对文件的解读,我们发现NIST已经抓住了智能体的核心安全结构,即智能体的风险不仅来自模型输出,而且来自模型被嵌入可操作环境之后,选择和调用工具并持续改变外部状态的过程。这就为智能体的治理提供了可以参考的逻辑。

2026年2月17日,NIST又发布了《智能体标准倡议》。该倡议旨在为能够自主行动的下一代智能体建立可信、可互操作且安全的行业标准与开放协议,并且提出要让智能体能够被社会所认可和接受。^③为实现这一目标,该文件还认为美国需要建立三大战略支柱:一是推动行业主导的标准制定与国际领导力塑造;二是培育社区主导的开源协议开发;三是推进关于智能体安全与身份的研究。同时,围绕这三大支柱,也提出了一些可落地的智能体核心治理原则。尤其重要的是,NIST还明确指出要将可信、可互操作的智能体生态系统建设作为长期目标,这比抽象的安全目标更进一步,因为进入了可执行的阶段。

2026年6月2日,特朗普总统签署第14409号行政令,发布《促进先进人工智能创新与安全》文件。^④该文件提及了利用智能体非法访问数据或信息等问题,这也是在白宫文件中第一次出现智能体相关内容。可以预期,未来美国政府围绕智能体治理的文件还会进一步出台。

总的来说,当前美国关于智能体的治理活动主要由NSIT主导。NIST根据特

① The White House, “Winning the Race: America’s AI Action Plan”, July 2025, <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>

② National Institute of Standards and Technology, “Request for Information Regarding Security Considerations for Artificial Intelligence Agents”, Federal Register, January 8, 2026, <https://www.federalregister.gov/documents/2026/01/08/2026-00206/request-for-information-regarding-security-considerations-for-artificial-intelligence-agents>

③ National Institute of Standards and Technology, “AI Agent Standards Initiative”, NIST, February 17, 2026, <https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative>

④ The White House, “Promoting Advanced Artificial Intelligence Innovation and Security”, June 2, 2026, <https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/>

朗普政府关于人工智能治理的基本精神，对智能体的发展采取“创新优先、安全嵌入、行业主导、轻触监管”的方针，旨在避免由于过分严格的治理要求而影响美国在人工智能领域的领先地位。同时，还强调优先制定行业标准，以助力智能体在不同场景中的部署。

（二）欧盟：尝试将智能体纳入风险分级框架

欧盟具有全球最为完整和系统的人工智能治理体系，以《人工智能法案》（*AI Act*）为核心的治理体系采取风险导向规则，确立了欧盟的人工智能风险分级管理框架。^①面对智能体带来的新挑战，欧盟选择的主要应对策略是将其纳入《人工智能法案》所确立的风险分级框架，试图根据智能体的应用场景及其影响决定治理态度。

例如，欧盟把关键基础设施安全组件、教育、就业、基本公共和私人服务、执法、移民和边境管理、司法和民主活动等列为可能存在高风险的人工智能应用领域，智能体在这些领域的应用必须按照《人工智能法案》的要求，满足具备风险评估环节、高质量数据集、日志记录、技术文档、向部署者提供充分信息、人类监督、稳健性、准确性和网络安全性等前提条件。如果智能体应用在办公写作场景中，那就是低风险工具；但同样的智能体如果被用于自动筛选求职者、辅助进行福利资格判断、调度交通基础设施、支持出入境审查或生成司法文书等，就会进入高风险义务领域。这种做法既避免了为每一种新的技术形态重复立法，也可防止企业通过更换“智能体”“助手”“工作流自动化”等名称规避所应承担的义务。

但与此同时欧盟也注意到，智能体具有显著的行为网络化特征，仅仅根据应用目标对其进行分类管理明显落后于技术发展步伐。于是，针对智能体又有了很多补充性治理方案。

2025 年 8 月 1 日，欧盟委员会批准了独立专家提交的《通用人工智能实践准则》（*General-Purpose AI Code of Practice*）最终文本。^②这是《人工智能法案》框架内的重要补充文件，要求所有通用人工智能模型提供者必须提供详细的模型文件，从而明确了底层模型提供者的合规责任。

^① European Commission, “‘AI Act’, European Commission Shaping Europe’s Digital Future”, May 21, 2024, <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

^② European Commission, “Drawing-Up a General-Purpose AI Code of Practice”, European Commission Shaping Europe’s Digital Future, August 1, 2025, <https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice>

欧盟有关机构已经开始观察代理式人工智能时代的技术发展趋势。例如,欧洲数据保护监督机构(EDPS)在 TechSonar 页面上详细描述了代理式人工智能的概念与特点,并讨论了《通用数据保护条例》、《人工智能法案》和欧洲数据保护监督机构的监管逻辑相互叠加之后的治理特点。

欧盟人工智能治理模式的优点在于法律结构清晰、严格且缜密,但缺点是所有这些治理框架都根据既有技术能力而设立,在面对新技术产生和技术快速迭代时,常有不适用的问题。如今面对智能体的兴起,欧盟仍然希望根据用途和影响把智能体吸收进既有的人工智能治理风险分级框架,但由于智能体在实践中具有明显的行为漂移、工具扩展和多主体协作特点,一次性合规评估难以覆盖部署后出现的复杂权限变化等问题。可见,智能体的发展对欧盟人工智能治理策略提出了挑战。当然,目前智能体的发展尚处于起步阶段,未来发展态势仍有待观察,只有当后续的技术标准、监管“沙盒”、持续监测和严重事故报告出现后,才能有效评估《人工智能法案》等治理文件是否能够适用于动态发展中的智能体系统。

(三) 俄罗斯:特殊外部环境下的主权安全路径选择

俄罗斯的人工智能治理政策首先是技术主权维护政策,具有鲜明的“安全高于发展”特征。2019年10月10日,俄罗斯总统普京签署第490号总统令,批准发布《2030年前俄罗斯联邦人工智能发展国家战略》。^①俄罗斯把人工智能发展同国家竞争力提升、数字经济发展、基础设施安全、人才培养、技术特别是可信技术研发,以及法律监管体系建设等紧密联系在一起,强调在地缘政治压力加大和外部技术限制加强的背景下,形成自主可控的模型、数据、算力、软件和应用生态的重要性。

目前,俄罗斯尚未形成专门针对智能体的治理规则,其治理逻辑是通过主权利力和关键法规条款间接控制、约束智能体的部署和应用。智能体系统一旦介入政务、军事、金融、经济和能源领域,其安全性就不再只取决于模型算法设计,还取决于训练和推理算力是否在本国可控环境内,任务数据和用户数据是否受本国法律监管,工具接口是否连接国产软件和数据库,关键行业是否能够接受审计,等等。俄罗斯人工智能国家战略文件中关于基础设施、可信技术和监管体系建设的要求,虽然没有使用“智能体”这一概念,但既有政策文本中明确使用了

^① Указ Президента Российской Федерации от 10.10.2019 № 490 “О развитии искусственного интеллекта в Российской Федерации” .11 октября 2019 года. <http://publication.pravo.gov.ru/Document/View/0001201910110003>

“能够自主调用工具和执行任务的系统”表述。^①

对于智能体潜在的风险，俄罗斯仍选择了传统的数据本地化、国产软件替代和关键信息基础设施保护相结合的应对路径。对于生成式人工智能而言，境外模型可能带来的风险主要是内容偏向、数据泄露和舆论影响，但对于智能体而言，风险可不止于此，尤其是有些智能体形成的外部依赖会进一步演变成行动依赖的风险不容忽视。也就是说，如果俄罗斯企业或机构的智能体运行在境外云服务器上，或调用境外 API，或使用不可审计的模型，并接入俄罗斯工业、金融等关键行业系统的话，那么这种技术依赖就会转化为外部平台的实际操作权和外部势力对俄罗斯实施制裁的渠道。因此，俄罗斯强调的技术主权在代理式人工智能时代不只是产业政策，而且是防止外部平台进入本国国家治理执行链条的安全政策。

（四）日本：AI 法、基本计划与 AI Agents 软法定义

日本在 2025 年至 2026 年初完成了本国的人工智能治理框架建设，形成了以《人工智能相关技术研究开发及应用推进法》《人工智能基本计划》《人工智能相关技术研究开发及确保应用适当性指南》等为核心的治理体系。^②相关法案的通过，意味着日本的治理思路是在法律、计划、指针和行业指南之间建立起有机联系，共同开展人工智能治理。

2026 年 3 月 31 日，日本经济产业省与总务省联合发布了《人工智能企业运营者指南》（*AI Guidelines for Business Ver1.2*）。该文件专门加入了对“智能体”和“代理式人工智能”概念的界定内容。^③

日本针对智能体的治理思路相对而言较为平衡。一方面，强调人类中心地位和技术适当使用的重要性。日本政、经、学界认为，不应让人类过度依赖 AI 或被 AI 控制自己的行为，对国计民生影响重大的领域仍要保留人类判断选项，并

① См.: Указ Президента Российской Федерации от 10.10.2019 № 490 “О развитии искусственного интеллекта в Российской Федерации” (в ред. Указа Президента Российской Федерации от 15.02.2024 № 124) . <http://kremlin.ru/acts/bank/44731/print>

② 内閣府「AI 戦略」、内閣府网、2026 年 6 月 26 日、<https://www8.cao.go.jp/cstp/ai/index.html>；内閣府「人工知能関連技術の研究開発及び活用の推進に関する法律（AI 法）」、内閣府网、2025 年 6 月 4 日、https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html；内閣府「人工知能基本計画」、内閣府网、2025 年 12 月 23 日、https://www8.cao.go.jp/cstp/ai/ai_plan/ai_plan.html；内閣府「人工知能関連技術の研究開発及び活用の適正性確保に関する指針」、内閣府网、2025 年 12 月 19 日、https://www8.cao.go.jp/cstp/ai/ai_guideline/ai_guideline.html

③ Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, “AI Guidelines for Business Ver.1.2”, March 31, 2026, https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_12.pdf

建立相应的说明和停止机制。另一方面，强调对技术进行全生命周期治理的重要性。AI 开发者、提供者和使用者需分别承担开发、集成、服务提供、运行支持、使用和反馈等职责。此外，针对新技术的产生和技术的不断迭代，日本的人工智能治理思路还强调了软法的作用，鉴于硬性规则存在滞后风险，因此需要借助指南、指针、评估工具和国际原则等不断对其加以更新和完善。

综上所述，2025—2026 年世界主要行为体都意识到了代理式人工智能对于人类社会发展的重大意义，同时也开始关注智能体治理，但所选择的治理路径各不相同，治理思路也莫衷一是。例如，有的行为体致力于维护自身的技术领先地位，有的希望在既有法律框架下容纳技术进步，有的则坚持技术主权原则并强调风险意识。但无论如何，上述行为体的探索与尝试都为包括智能体在内的新技术治理提供了宝贵经验，也成为中国开展人工智能治理活动的重要借鉴。

四、智能体治理的中国成就

面对汹涌而来的人工智能技术浪潮，中国积极适应时代发展趋势，通过丰富的实践积累和广泛的学习研究，逐渐形成了一套“发展”与“安全”并重的人工智能治理体系。

在顶层设计上，以 2017 年《新一代人工智能发展规划》确立的国家战略为基石^①，以 2023 年《全球人工智能治理倡议》为指导思想^②，逐步构建起了多层次、立体化的人工智能治理架构。中国的人工智能治理以《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等三部基础性法律为支柱，围绕算法推荐、深度合成、生成式人工智能等新兴技术出台了一系列部门规章和专项规则，并通过发布《新一代人工智能伦理规范》《人工智能安全治理框架》等文件，指引相关伦理规则和安全规范建设。^③中国的人工智能治理体系具有鲜明的“敏捷治理”特征，可在快速防范风险的同时为产业

① 《国务院关于印发新一代人工智能发展规划的通知》，载中华人民共和国中央人民政府网 2017 年 7 月 20 日，https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

② 《全球人工智能治理倡议》，载中华人民共和国外交部网 2023 年 10 月 20 日，https://www.mfa.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml

③ 《〈人工智能安全治理框架〉1.0 版发布》，载中央网络安全和信息化委员会办公室网 2024 年 9 月 9 日，https://www.cac.gov.cn/2024-09/09/c_1727567886199789.htm；《〈新一代人工智能伦理规范〉发布》，载中华人民共和国科学技术部网 2021 年 9 月 26 日，https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html

创新保留充分的制度弹性，体现了包容审慎的监管理念，以及努力在促进技术创新与保障高水平安全之间实现动态平衡的态度。尤其是针对智能体的治理，更加彰显了中国人工智能治理的特点和优势。

为应对智能体快速发展带来的风险挑战，以中央网络安全和信息化委员会办公室于 2026 年 5 月 8 日发布的《智能体规范应用与创新实施意见》（简称《实施意见》）为基础，形成了中国智能体治理的基本框架。在《实施意见》中，智能体被概括为具备“自主感知、记忆、决策、交互与执行能力”的智能系统，并直接指出了智能体正在加速同网络空间和物理世界的深度融合，以及由此带来的风险。与多数行为体只在人工智能政策范畴内间接涉及智能体治理有所不同，中国是直接以智能体为对象提出的治理方案，且内容全面覆盖了技术底座、标准协议、安全底线、应用牵引、创新生态和保障措施等领域，因此可以说是目前全球针对智能体最为系统的治理方案。^①

《实施意见》关于智能体技术结构的表述堪称非常具体。在“夯实发展基础”方面，文件提出要提升高质量数据集供给能力，并加强国产智能体在任务理解、任务规划、工具使用、长期记忆、互认互通和群体协同等方面的技术攻关；完善智能体底层框架，并加快研发感知、记忆、决策、交互、执行等关键组件；完善研发、测试、部署、运维工具链，并加速发展对抗样本检测、行为异常检测等安全与治理工具；等等。^②据此，本文中的模型层、结构层和交互层可以分别转化为如下政策任务：模型和数据构成底座；任务规划、长期记忆和群体协同，构成结构编排；工具使用和执行组件构成交互行动。

在“构建标准协议”方面，《实施意见》提出制定智能体标准化工作指导文件，旨在布局关键技术、重要产品、数据交换、应用场景、质量评测、安全保障和可信认证等标准，并加快制定智能体与软件工具、应用服务、硬件外设接口等相连接的基础标准。更为重要的是，文件还提出研究建立智能互联网体系，探索建立智能体注册平台，旨在提供智能体数字身份管理、检索发现、能力声明等服务，并研究身份标识、可信互联、合规支付、安全防护和冲突解决等基础性技

^① 《智能体规范应用与创新实施意见》，载中央网络安全和信息化委员会办公室网 2026 年 5 月 8 日，https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm

^② 同上。

术。^①这就意味着,智能体不应是匿名的模型输出,而是需要满足身份、能力、接口、支付和安全边界等前提条件的行动单元。

在“守牢安全底线”方面,《智能体规范应用与创新发展实施意见》的规定更加贴近智能体风险治理。文件要求明确智能体决策权限,厘清仅限用户本人决策、需由用户授权决策和智能体自主决策的边界,确保用户对智能体自主决策享有知情权和最终决定权,且规定智能体执行操作时不得超出用户授权范围。同时,《实施意见》还要求加强对智能体行为的管控,建立重要应用场景行为可验证、可追溯机制;加快研究数据安全维护、个人信息保护、密码防护、攻击检测、权限管理、行为控制等安全技术,以防范数据投毒、隐私泄露、算法篡改、系统漏洞攻击和运行失控;确保供应链安全,对模型接入、API调用、扩展工具使用等环节加强管理,以避免智能体被用于自动化攻击、隐私侵犯、虚假信息生成与传播以及网络诈骗。

此外,《实施意见》还提出要根据应用场景和潜在影响对智能体开展分类分级治理:对于敏感领域和重点行业,由网信部门会同行业主管部门确定开放场景,并根据法律法规、监管要求和安全防护标准,采取备案、检测、问题产品召回等管理措施;对于生活娱乐和日常办公等低风险领域,则通过合规自测、信息报告、分发平台管理和行业自律等方式提高治理效率。^②这一路径既延续了中国已有的算法备案、安全评估和分类分级监管传统,又把新增的智能体带来的权限、工具和行动等方面的风险纳入了管理范畴。

总而言之,《智能体规范应用与创新发展实施意见》的发布是中国人工智能治理领域的重要成果,它系统性地提出了中国智能体产业发展的基本规则,不仅在全球范围内率先建立了覆盖智能体技术全生命周期的治理框架,而且也有力推动了人工智能治理从强调“内容风险”向关注“行为风险”的转型。^③目前来看,中国的智能体治理已经取得了显著成就,既充分彰显了《全球人工智能治理倡议》的基本精神,又清晰展现了中国坚持统筹发展和安全、坚持发展和安全并重的治理思路。

① 《智能体规范应用与创新发展实施意见》,载中央网络安全和信息化委员会办公室网2026年5月8日, https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm

② 同上。

③ 《专家解读|为智能体发展树立规范、留足空间》,载中央网络安全和信息化委员会办公室网2026年5月8日, https://www.cac.gov.cn/2026-05/08/c_1779979790609763.htm

结语

智能体技术的成熟与普及标志着代理式人工智能时代的到来,人工智能安全治理也随之进入了一个以“行动治理”为核心的新阶段。从生成式人工智能到代理式人工智能的转型并非单纯的技术升级,而是人工智能深度嵌入社会组织结构、国家治理体系和国际竞争格局的关键一跃。这必将引发人类社会政治与安全结构的深层次变迁,改变全球人工智能安全治理形态。

智能体的内在属性与安全风险具有天然的契合性,而智能体的强自主性、行为网络化与能力多样化特征又使其可以深度介入人类社会的政治结构,并推动原本的“人-组织”二元结构向“人-智能体-组织”的三元结构转化。新的政治结构会在安全领域产生重要影响,其外在表现就是由于智能体的深度参与,军事安全、政治安全、网络安全等诸多安全领域出现了形态与结构变革。值得强调的是,在新的技术环境下,涉及安全的各个方面都会面临不同程度的挑战,也会出现各种新问题与新风险,因此需要全球各行为体都加以关注。

面对智能体工具的快速发展,美国、俄罗斯、日本及欧盟等主要行为体都在不同程度地展开智能体治理行动。各行为体的智能体治理政策逻辑起点不尽相同,治理思路与方法大相径庭,治理路径的分歧也非常明显,但其治理经验都能够为中国的智能体治理提供参考。

中国非常重视智能体技术的发展,坚信智能体超越了一般性应用创新的范畴,能够成为帮助中国增强国家竞争能力、制度适应能力与国际规则塑造能力的又一重要工具。与世界上的其他主要行为体有所不同,中国的智能体治理更加注重平衡发展与安全的关系。在“敏捷治理”思想的指导下,中国在智能体治理方面迅速形成了独特的治理方案,并取得了显著的效果。但智能体治理是一个长期过程,其所面临的关键挑战在于治理范式的深刻转换。合理的治理方案可以在确保智能体技术释放科技红利的同时,免于沦为侵蚀国家安全基础与公共制度信任基础的隐蔽风险载体。对于国际社会而言,如何在技术竞争与安全合作之间寻求某种动态平衡,如何在规则碎片化与治理协同之间建立有效衔接,将是决定智能体时代全球安全秩序走向的深层次命题。

总之,2025—2026年,人类社会正式进入了以智能体为主要标志的代理式人

工智能时代。从此，各种智能体将很快成为人们日常生活与工作的助手，也将深刻改变社会结构和政治关系。在新的三元政治结构框架下，人类社会诸多领域的安全逻辑都将发生重要变化，因此治理思路必须随之加以调整。代理式人工智能时代的到来，使得人类社会向着通用人工智能（AGI）时代又迈进了一大步，而社会科学研究的责任也因此变得更重了。可以预见，未来围绕智能体治理问题的讨论将进一步增加，其中探索国家在代理式人工智能时代的合理政策定位，应该成为我们这些社会科学研究者下一步的重要目标。

（责任编辑 胡巍葳）

Вызовы, обусловленные ИИ-агентами: риски национальной безопасности и пути управления в эпоху агентного искусственного интеллекта

Фэн Шуай, Се Цзяньчжи

【 Аннотация 】 Развитие искусственного интеллекта в настоящее время переживает переход от эпохи генеративного искусственного интеллекта к эпохе агентного ИИ. Появление ИИ - агентов является отличительным признаком. ИИ - агенты обладают способностью к самостоятельному планированию, вызову инструментов и взаимодействию с внешней средой, что обеспечивает переход искусственного интеллекта от “что он может сказать” к “что он может сделать”, но и более глубокое изменение отношений между человеком и технологиями, а также между человеком и организациями. В рамках новой системной архитектуры характер и содержание рисков национальной безопасности претерпевают глубокие изменения. Под воздействием импульса со стороны ИИ-агентов, существенные изменения происходят в таких сферах, как военная безопасность, политическая безопасность, кибербезопасность и безопасность ключевой инфраструктуры. Риски безопасности в различных сферах по - разному проявляются в краткосрочной, среднесрочной и долгосрочной перспективе, перестраивая среду национальной безопасности на различных уровнях. В целях реагирования на вызовы, связанные с ИИ-агентами, США, ЕС, Россия, Япония и другие страны накопили ценный опыт в сфере управления ИИ - агентами. Благодаря точному позиционированию и гибкому подходу к управлению Китай добился хороших результатов в области управления ИИ-агентами, что может послужить примером для управления искусственным интеллектом в странах мира в новую эпоху.

【 Ключевые слова 】 Агентный ИИ; ИИ - агент; риски безопасности; управление искусственным интеллектом

The Challenge of AI Agents: National Security Risks and Governance in the Era of Agentic AI

Feng Shuai, Xie Jianzhi

Abstract: The development of artificial intelligence is transforming from the era of generative AI to that of agentic AI. The hallmark of the agentic AI era is the emergence of AI agents. These agents possess the capability to autonomously plan, utilize tools, and interact with the external environment, shifting AI's focus from "what it can say" to "what it can do". The development of AI agents signifies not merely an expansion of technical capabilities but a profound transformation in the relationships between humans and technology, and between humans and organizations. Within this new systemic framework, the nature and content of national security risks have undergone significant changes. Domains such as military, political, and cybersecurity, as well as the security of critical infrastructure, are all experiencing marked shifts under the impact of AI agents. Security risks across these various domains manifest differently over short, medium, and long terms, reshaping the national security landscape at multiple levels. To address the challenges posed by AI agents, the United States, Europe, Russia, and Japan have accumulated valuable governance experience. Through precise positioning and agile governance strategies, China has achieved notable success in this area, offering a valuable reference for AI governance worldwide.

Keywords: Agentic AI; AI agents; security risks; AI governance