世界互联网大会
World Internet Conference

前沿人工智能安全与治理

2026

风险认知与协同治理



世界互联网大会人工智能专业委员会
安全与治理推进计划
2026年11月

世界互联网大会人工智能专业委员会安全与治理推进计划成员及贡献者

牵头人

曾毅	中国人民大学高瓴人工智能学院 北京前瞻人工智能安全与治理研究院
肖恩•欧•海格缇	剑桥大学

专家及贡献者

（包括推进计划成员和特邀专家，以及为报告提供内容或意见的专家，按单位名称首字母排序）

AI Safety Connect
尼古拉 · 米亚伊、塞勒斯 · 霍兹

阿拉伯信息通信技术组织
纳达 · 拉阿比迪

阿里云
逄孝钢、王盛杰、王星光、张荣

巴西南大河州联邦大学
埃德森 · 普雷斯蒂

百度
包沉浮

北京大学
王小塞、杨耀东

北京印象笔记科技有限公司
乔迁

北京邮电大学
杨忠良

德国明斯特大学
伯恩德 · 霍尔兹纳格尔

伏羲智库
程凯、吴胤禹

国际人工智能治理协会
约翰 · 希金斯、罗伯 · 沃瑟姆

杭州安恒信息技术股份有限公司
王欣、胡壮、杨绍波、符春辉、林明峰

华为技术有限公司
柳嘉琪、王晓箴、许梦雯、林秋诗

IQuilibriumAI
希梅娜 · 索菲亚 · 比维罗斯 · 阿尔瓦雷斯

雷丁大学
唐银山

联合国教科文组织驻东亚代表处
夏和华

联想（北京）有限公司
胡永启

蚂蚁科技集团股份有限公司
徐继明

慕尼黑工业大学
丹尼尔 · 凯里米

南开大学
陶锋

奇安信科技集团股份有限公司
张勇、马兰、林雪艳、姜晨

清华大学
薛澜、张亚勤

人民网股份有限公司
汪凤琼、张晓赫

360集团
潘剑锋、马琳、王千玉、肖荐午

SaferAI
亨利 · 帕帕达托斯、马尔科姆 · 默里

诺基亚通信(上海)股份有限公司
常疆

上海创智学院
乔宇

上海稀宇科技有限公司
彭韬、沈俊成

斯里兰卡计算机应急响应中心
查鲁卡 · 塞纳尔 · 达穆努波拉、尼罗沙 · 阿南达

世界互联网大会
任贤良、梁昊、张雪丽

钛媒体集团
陈惠洁

腾讯
王梦寅、王融、徐德亮

天翼安全科技有限公司
杨勃

通用人工智能政策实验室（法国）
汤姆 · 戴维

先进计算与关键软件（信创）海河实验室
高雪松

香港岭南大学
姚新

香港中文大学
蒙美玲

微博
王巍

央视国际网络有限公司
程明

英国标准协会
克里斯 · 布朗

英国南安普顿大学
温迪 · 霍尔

英国特许仲裁学会
艾墨 · 尤纳斯

云安全联盟
约翰 · 杨

智慧非洲联盟
迪迪埃 · 恩库里基姆弗拉

中关村实验室
谭知行

中国电信集团有限公司
杨明川、刘威辰、王峰、张子婷

中国互联网络信息中心
陈晶晶、尉迟学彪、岳巧丽

中国互联网协会人工智能工委会
夏文辉

中国联合网络通信集团有限公司
周凯

中国人民大学
龚新奇

中国社会科学院人工智能研究促进中心
段伟文

中国网络空间安全协会
王健兵

中国新一代人工智能发展战略研究院
龚克

中国信息通信研究院
魏凯、呼娜英、郭苏敏、杨玲玲、牛晓玲、梁叶

中国移动研究院
林琳、吴淑燕

中国政法大学
张凌寒

中欧数字协会
鲁乙己 · 加姆巴得拉

中兴通讯股份有限公司
孟伟

编写团队

中国人民大学高瓴人工智能学院
北京前瞻人工智能安全与治理研究院
曾毅、鲁恩萌、王正奇、李凯、王沛宁

世界互联网大会
韩开宇、常瑶

中国科学院大学
曹功策、陈煜、谢佳玮、彭誉萱

联系邮箱

research@wicinternet.org

世界互联网大会
“前沿人工智能安全与治理”系列报告

人工智能正成为推动科技创新、产业变革和社会发展的重要力量，深刻影响知识创造、生产组织、公共服务和社会治理。随着技术能力持续拓展、应用边界不断延伸，人工智能带来的安全与治理问题日益成为国际社会共同关切。相关风险呈现技术性、社会性和跨境性相互交织的特点，对各方风险认知、治理能力和国际协同构成了更为紧迫的考验，亟需各方加强前瞻准备，形成合力，以有效应对日益复杂的治理挑战。

在此背景下，世界互联网大会人工智能专业委员会安全与治理推进计划打造“前沿人工智能安全与治理”系列报告，作为面向国际社会的年度公共产品。报告立足前沿人工智能能力与应用的最新进展，旨在系统探究技术发展动态及其引发的安全与治理新挑战，评估不断演变的风险图景，审视相应治理对策，明确国际协调的优先事项，力求增进全球共识、提升共同应对相关挑战的能力。

系列报告坚持整体性、前瞻性和开放性视角，汇聚政府部门、国际组织、企业、科研机构等多方观点，提供基于事实和证据的分析参考。通过持续开展年度观察和比较研究，报告致力于增强国际社会对前沿人工智能风险的前瞻准备和治理能力，促进各方完善治理实践，推动前沿人工智能在安全、可靠、负责任的基础上更好服务人类发展。

关于本报告

当前，前沿人工智能快速发展，大语言模型、多模态模型、自主智能体、具身智能等技术持续演进，推动人工智能由信息生成、知识服务进一步走向任务执行、工具调用和复杂系统协作，并不断进入科研创新、产业发展、公共服务和社会运行等更广泛场景。与此同时，相关风险也由单点、局部问题加快演变为跨场景、跨主体、跨边界的复合性挑战，对现有治理体系的适应能力、风险处置能力和国际协同能力提出更高要求。

本报告聚焦前沿人工智能的风险认知与协同治理，旨在为政府部门、国际组织、企业、科研机构等提供参考，促进各方凝聚共识、增进互信并加强协作，共同推动前沿人工智能安全、可靠、负责任发展，服务人类共同福祉。

报告结构

本报告按照"图景研判、风险认知、治理评估、协同应对"的路径展开，系统梳理前沿人工智能的能力发展、应用拓展与风险演进，结构化分析重点风险、严重程度以及应对优先级，评估现有治理实践和治理就绪状态，并进一步提出协同应对方向与可行机制建议。报告重点探索四个问题：一是前沿人工智能能力和应用正在如何演进，并如何催生新的风险；二是前沿人工智能风险主要有哪些方面、严重程度如何，哪些风险需要优先应对；三是当前各方已开展哪些治理实践，整体上治理就绪度如何，哪些方面仍需进一步完善；四是如何通过加强不同主体间的协同来提升前沿人工智能治理的效能。

研究范围

本报告所称前沿人工智能（Frontier AI），指在通用领域或专业领域处于能力前沿的人工智能模型，以及基于此类模型构建的智能体、多智能体协作或具身智能系统。此类模型或系统通常以在大规模数据与算力上训练而成的基础模型（Foundation Model）为核心形态，具备跨广泛任务的高水平能力，并可能因其能力、通用性与部署深度而引发难以在事前充分预见或约束的重大风险。本报告以能力前沿为界定标准，不设参数量或训练算力等固定参照；所谓"前沿"的边界会随技术进步，以及工具、权限、运行环境等部署条件的变化而动态演进。

本报告认识到，前沿人工智能在军事场景下的应用会带来一系列安全风险与治理挑战，属于人工智能安全治理议题的重要组成部分。基于本报告的研究定位，报告不探讨前沿人工智能在军事领域的相关问题，重点聚焦前沿人工智能在非军事领域的规模化应用场景，系统研判相关安全风险，分析覆盖个人、组织和社会层面的影响。

数据来源与方法

本报告的风险分析综合了当前学术界和产业界围绕模型与智能体开展的安全评测、公开记录的风险事件统计以及有关专家与机构的观点研判。

本报告开展了对前沿人工智能治理就绪度的评估。该评估过程依据公开资料，从六个方面、22项二级指标检视24个国家应对前沿人工智能风险的准备情况，用于识别治理基础和需要加强的环节。评估样本选取仅服务于本报告的分析目的，不应被解读为对任何国家人工智能治理能力的评判或排序，亦不构成对未纳入国家的排除性结论。治理就绪度评估的数据截止2026年8月31日。有关评估对象、指标体系及评估方法的说明详见附录。

说明

前沿人工智能仍处于快速演进之中，相关风险事件与科学证据持续积累。本报告的分析与结论基于特定时点可得的公开资料，仅反映特定时点的认知。本报告各项分析与结论系编写团队基于公开资料、汲取专家和贡献者意见后形成的学术性解读与研判，不代表任何单一个体、机构或国家的立场。

关键术语

术语	含义
前沿人工智能	指在通用领域或专业领域处于能力前沿的人工智能模型，以及基于此类模型构建的智能体、多智能体协作或具身智能系统。
智能体	能够规划并执行多步骤任务、调用工具并与环境交互的人工智能系统，执行中较少需要人工介入。
多智能体系统	由多个相互作用的智能体组成的系统，各智能体可以分工协作，也可能相互适应或竞争。
具身智能	通过传感器、机器人、车辆和工业设备等与物理环境交互的人工智能系统。
开放权重模型	模型参数（权重）可公开下载的模型。本报告所称开放模型指开放权重模型。
安全评测与红队测试	安全评测用于评估模型或系统的能力、脆弱性和潜在影响；红队测试由专门人员或团队主动寻找漏洞、局限和被滥用的可能。
灾难性与生存性风险	指可能造成范围广泛且难以逆转的重大后果，甚至危及人类长远生存与发展的风险，主要涉及高度自主系统失控、欺骗行为、自我复制和危险能力扩散等情形。
人工智能风险事件	人工智能系统的开发、使用或故障直接或间接造成人身伤害、关键基础设施运行中断、权利受损或财产损失等后果的事件。
协同治理	不同国家和主体围绕评测、核验、事件处置和能力建设等环节开展合作，明确分工，共享风险证据、责任记录和处置信息，共同应对跨平台、跨行业、跨境扩散的风险。
治理就绪度	根据公开可查的法律政策、机构机制、标准、评测、合规、事件通报、能力建设和国际协作等安排，考察一国应对前沿人工智能风险的准备程度。

执行摘要

前沿人工智能正处在能力快速跃升、应用加速扩散和风险持续显现的关键阶段。一方面，模型、智能体与工具环境协同形成的系统能力不断提升，人工智能加快进入科研创新、产业发展、公共服务和社会运行等更广泛场景；另一方面，相关风险正由潜在担忧转化为现实影响，对现有治理体系的适应能力、风险处置能力和国际协同能力提出更高要求。

前沿人工智能的风险，正由单点、局部问题演变为跨场景、跨主体、跨边界的系统性挑战。其能力的提升已从单项任务突破转向模型、工具与环境协同的系统跃升，应用由信息生成深入辅助决策与任务执行。能力跃升降低了高风险行为的实施门槛，应用嵌入使风险进入业务流程，二者叠加改变了风险的产生与传导路径：风险不再由单一模型输出引发，而是在工具调用、组织流程和社会网络等多个环节同步生成、相互叠加、连续放大。

当前现实压力集中在滥用与犯罪风险、信息完整性风险和关键系统风险等方面。从影响面看，前沿人工智能风险涉及滥用与犯罪风险、信息完整性风险、关键系统风险、社会认知与结构风险、灾难性与生存性风险等方面。其中，前三个方面的风险已较多转化为现实危害；社会认知与结构风险在就业、教育等领域的长期影响正在逐步累积；灾难性与生存性风险潜在后果重大，触发条件尚待认知，需纳入前瞻研判。

面对当前态势，风险应对需依据紧迫程度与潜在后果分级施策，并把防线前移至部署之前。对现实中高发、复制扩散快的危害，重在及时处置、阻断扩散；对影响逐步累积、干预窗口较长的情形，重在长期监测、及早调整；对后果广泛且难以逆转的情形，应在部署前开展危险能力评测与红队测试、预先设定能力阈值，并保留人工接管、降级运行、回滚和关停等干预手段；其余情形纳入常规管理并定期复核。随着能力演进与证据积累，不同风险情形会发生变化，优先应对的安排需要定期复核、动态调整。

从公开资料看，当前各国对前沿人工智能风险的治理基础已初步形成，但评测、核验与责任落实等关键环节仍有明显短板。本报告对24个国家的前沿人工智能治理就绪度评估显示，多数国家在制度与组织方面准备相对充分，但在技术安全与评测、验证与合规、责任落实与生态治理等方面相对薄弱；在“风险缓解、权限约束和部署控制能力”“数据治理、供应链安全和开源生态治理能力”“风险事件通报与未遂事件共享机制”等指标上，已形成较稳定安排的国家不足三成。从风险应对视角看，风险事件通报与共享、应急响应与复盘改进、跨境风险信息通报和演练等能力缺口最为明显。

面对跨场景、跨主体、跨边界扩散的风险，分散的治理努力需要转化为衔接紧密的共同行动。其关键，在于把这些努力连接为覆盖风险全链条的共同机制：一是推动模型安全评测协同，形成可比较、可解释、可复核的风险证据；二是推动协同核验，使安全承诺及其落实情况可检查、可比较、可持续跟踪；三是加强风险事件协同通报和联合处置，使单点发现及时转化为跨平台、跨行业和跨地区的防护经验；四是完善智能体和开

放模型的协同治理安排，使风险能够被及时发现、分级处置并追溯责任；五是深化能力建设协同，使评测工具、标准方法和公共案例转化为更多国家和机构可获得、可使用、可持续的治理资源。

上述协同治理的落实，取决于各方能否依据其对模型、接口、权限和场景的实际控制能力、信息条件与处置能力形成稳定分工。政府、企业、科研机构、国际组织和公众应发挥多主体优势，共同推动风险证据、责任记录和处置信息顺畅流转，使治理由原则声明走向可执行、可核验、可改进的共同实践，让前沿人工智能在安全、可靠、负责任的基础上更好服务人类共同福祉。

目录

一、前沿人工智能图景：能力、应用与新兴风险 1

（一）前沿人工智能的能力发展 2

（二）前沿人工智能的应用拓展 4

（三）前沿人工智能的风险演进 7

二、前沿人工智能的风险认知 8

（一）前沿人工智能的风险识别 9

（二）前沿人工智能的风险评估 12

（三）前沿人工智能风险的优先应对 14

三、前沿人工智能的风险治理 16

（一）前沿人工智能的治理实践 17

（二）前沿人工智能治理就绪度评估框架 19

（三）前沿人工智能治理就绪度评估结果 20

四、前沿人工智能风险的协同应对 27

（一）协同应对的重点领域与可行机制 28

（二）协同应对机制中的多主体参与 31

附录　前沿人工智能治理就绪度评估指标及方法说明 33

第一章

前沿人工智能图景：能力、应用与新兴风险



本章要点

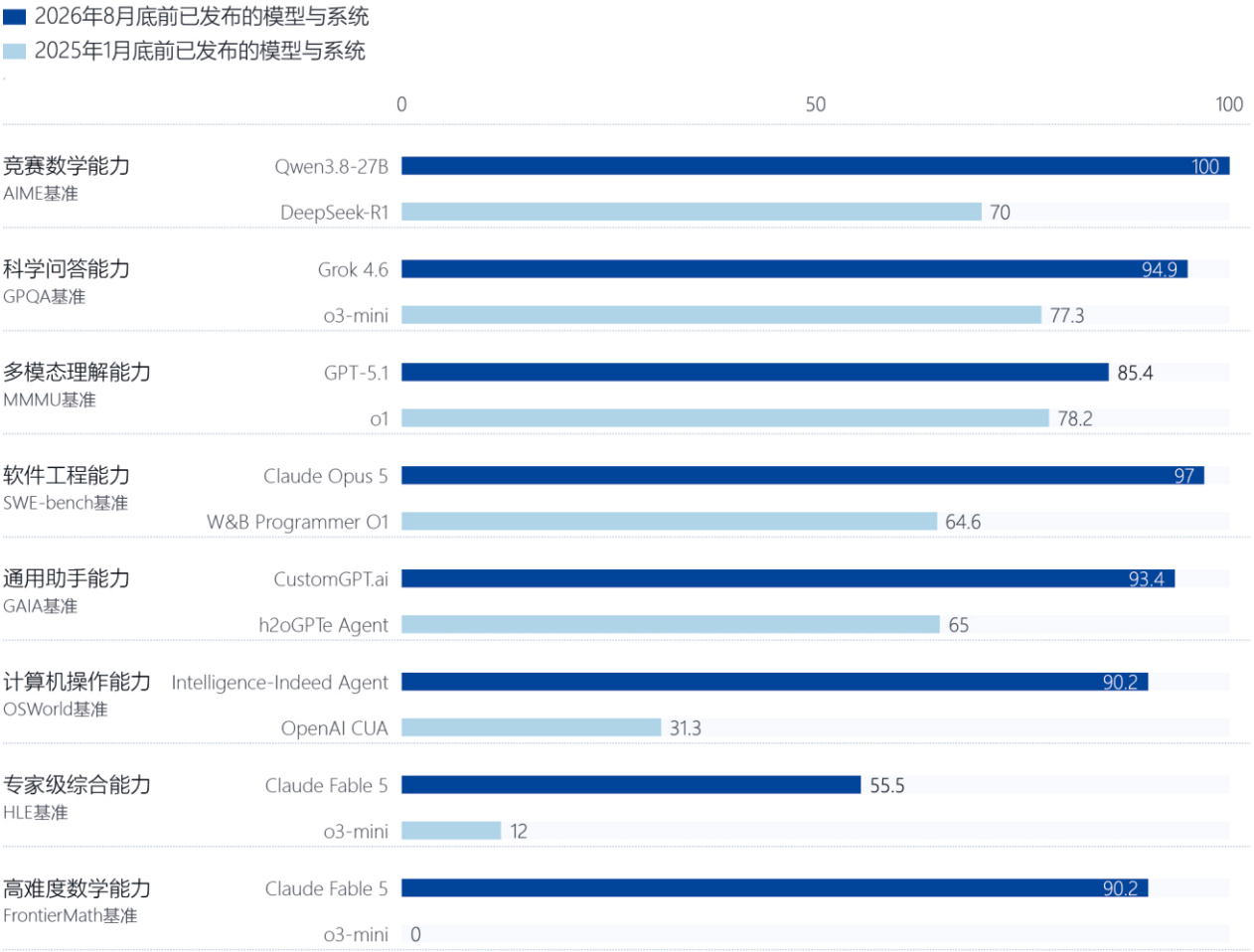
- 前沿人工智能正由单项能力突破转向模型、工具与环境协同的系统能力提升，任务自主执行和现实适应能力持续增强。
- 前沿人工智能正走向更广泛的实际使用，并在应用领域、应用深度和应用形态上持续拓展。
- 在促进创新、提高效率的同时，前沿人工智能的发展也带来了风险与挑战。能力提升降低了部分高风险行为的实施门槛，应用的深入使风险更易进入业务流程和社会场景，二者叠加改变了风险生成和传导的路径。
- 前沿人工智能风险正在持续加速演进，并快速转化为跨场景、跨主体、跨边界的现实挑战，这对风险治理提出了更高要求。

当前，前沿人工智能能力快速发展，应用场景不断拓展，相关风险也随之呈现更强的复杂性和外溢性。能力提升扩大其可完成任务的范围，应用拓展提高其介入真实场景的深度，风险也随之从单次输出延伸到任务执行、组织流程和社会环境中。这一图景表明，相关风险更容易在扩散过程中同步累积，并对风险识别、评估验证、监测响应和协同治理提出更高要求。

(一) 前沿人工智能的能力发展

前沿人工智能的基础能力持续增强，并带动其在多类复杂任务中的表现整体提升。过去一年，前沿人工智能模型在理解、推理、生成等基础能力以及多类复杂任务上的表现持续进步。公开基准和评测结果显示，前沿人工智能模型在软件工程等多项基准上的表现持续提升，部分前沿人工智能模型或系统在特定类型任务上也达到或超过相应测试的人类基准（见图1.1和图1.2）[1]。当前，

图 1.1
前沿人工智能模型与系统的多项能力显著提升
答题正确率或任务成功率（%）



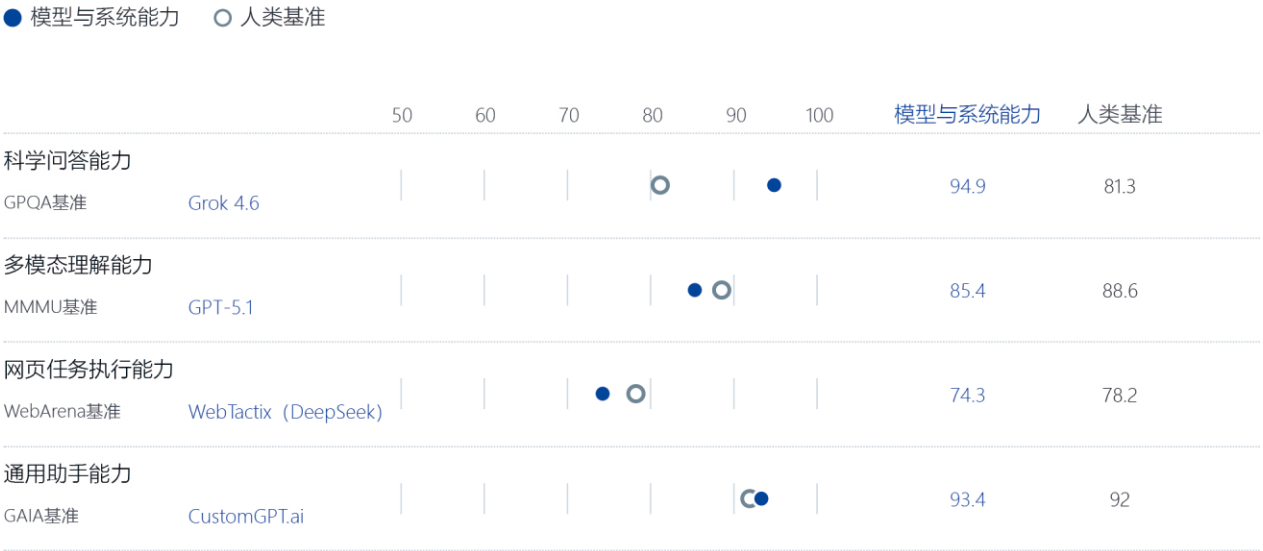
注：按模型与系统的可用时间分组，部分成绩来自后续评测。
数据来源：本报告基于公开信息整理，包括DeepSeek、ISTA-DASLab评测材料，各基准公开评测资料、Artificial Analysis和Epoch AI。

1 斯坦福大学以人为本人工智能研究院. 人工智能指数报告 2026（Artificial Intelligence Index Report 2026）[R/OL]. [2026-10-21]. https://hai.stanford.edu/ai-index/2026-ai-index-report. 该报告综合多项全球公开基准和评测结果，本报告据此归纳相关结论。

图 1.2

前沿人工智能在部分任务上已接近或超越人类水平

答题正确率或任务成功率（%）



注：1、 人类基准采用原作者公布的参考成绩，GPQA和MMMU为专家组成绩。
2、 GAIA和WebArena的人类与系统成绩来自不同样本。

数据来源：本报告基于Artificial Analysis、MMMU、GAIA、WebArena及WebTactix公开资料整理。

前沿人工智能模型已不仅限于处理文本，而是能够整合多模态数据；通过后训练和推理时计算，其处理复杂问题的能力显著增强[2]。相关进展表明，前沿人工智能的能力提升已不再局限于单项任务，而是逐步呈现多种能力共同增强的特征。

工具调用、智能体协作和具身智能的发展正推动前沿人工智能从内容生成迈向任务执行。通过连接外部软件、数据库和业务系统，模型的能力由生成内容和提供信息进一步拓展至执行具体操作。在此基础上，**智能体**综合工具调用、任务规划和环境交互等能力，能够独立完成的任务链条正在变得更长、更复杂[3]。**多智能体系统**则进一步强化复杂任务的分工、规划与连续执行能力，不同智能体可以围绕同一目标分担检索、分析、执行、校验和反馈任务[4]，形成更接近组织流程的协作方式[5]。**智能互联网**的发展探索，正通过体系架构和智能体注册、身份管理等基础服务建设，将智能体协作延伸到跨平台和跨系统的互联[6]。而**具身智能**则进一步打通数字与现实边界，使人工智能连接传感器、机器人[7]、车辆[8]和工业设备，并将作用场景与影响范围向受控物理环境和现实世界延伸[9]。

前沿人工智能的能力正从单个模型的任务表现，向模型、工具与环境共同构成的系

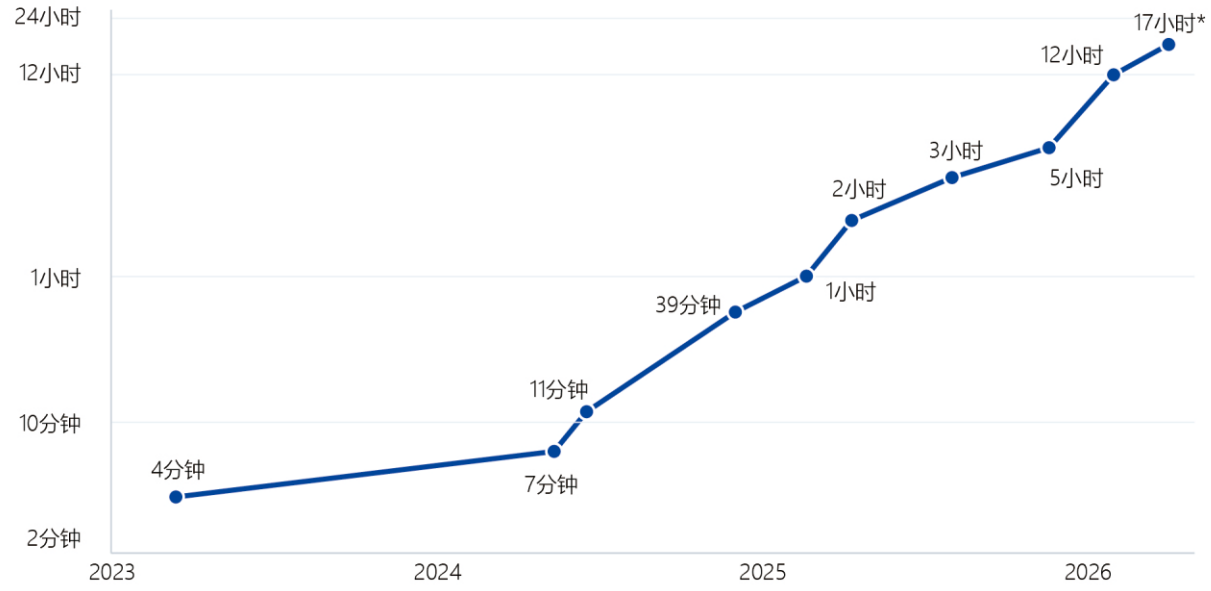
2 国际人工智能安全报告专家组. 国际人工智能安全报告 2026（International AI Safety Report 2026）[R/OL]. [2026-10-21]. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.
3 模型评估与威胁研究中心. 前沿人工智能模型任务完成时间跨度[EB/OL]. [2026-10-21]. https://metr.org/time-horizons/.
4 Salesforce. Agentforce HR Service 公告[EB/OL]. [2026-10-21]. https://www.salesforce.com/news/stories/agentforce-hr-service-announcement/.
5 ServiceNow. Yokohama 版本人工智能智能体发布材料[EB/OL]. [2026-10-21]. https://newsroom.servicenow.com/press-releases/details/2025/ServiceNows-latest-platform-release-adds-to-thousands-of-AI-agents-across-CRM-HR-IT-and-more-for-faster-smarter-workflows-and-maximum-business-impact-03-12-2025-traffic/default.aspx.
6 中国国家互联网信息办公室等. 智能体规范应用与创新发展实施意见[EB/OL]. [2026-10-21]. https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm
7 谷歌 DeepMind. Gemini Robotics[EB/OL]. [2026-10-21]. https://deepmind.google/models/gemini-robotics/.
8 英伟达. 利用 NVIDIA Cosmos 3 开发物理人工智能推理、世界和行动模型[EB/OL]. [2026-10-21]. https://developer.nvidia.com/blog/develop-physical-ai-reasoning-world-and-action-models-with-nvidia-cosmos-3/.
9 宝马集团. 斯帕坦堡工厂人形机器人[EB/OL]. [2026-10-21]. https://www.bmwgroup.com/en/news/general/2024/humanoid-robots.html.

统能力拓展。目前，其可完成任务的范围不断扩大，任务执行的时间跨度持续延长（见图1.3），但在复杂真实任务中的稳定性和可靠性仍有不足，实际表现仍需通过更贴近真实场景的评测加以检验。随着基础模型、工具集成和智能体系统持续发展，前沿人工智能的任务执行和现实适应能力仍有望进一步增强。

图 1.3

前沿AI智能体可独立完成的任务时长不断增加

以人类专家完成同类任务所需时间衡量，图示为50%成功率对应的任务时长



* METR提示：现有评测对超过16小时的任务时长估计仍有较大不确定性。

数据来源：模型评估与威胁研究机构（METR，Model Evaluation and Threat Research）公开数据。

(二) 前沿人工智能的应用拓展

前沿人工智能正走向更广泛的实际使用，在应用领域、应用深度和应用形态上持续拓展：应用场景不断增多，应用深度从信息处理延伸至辅助决策和任务执行，应用形态从问答和内容生成扩展到工具调用和智能体自主执行任务。人工智能应用的重心正在从技术可行性验证，转向嵌入业务流程、形成稳定产出并适应组织管理需要[10]。

从应用领域看，前沿人工智能正由专业生产和知识服务场景，延伸至公共服务、社会治理、文化娱乐等更广泛领域。应用场景的拓展，使人工智能的使用入口更多、调用频率更高、影响对象更广。尽管各领域应用成熟度不一，前沿人工智能正从少数专业领域的工具，扩展为社会运行中可被广泛调用的通用技术能力。欧盟官方统计显示，2023年至2025年，各行业使用人工智能的企业比例普遍上升，信息和通信业以及专业、科学和技术服务业中使用人工智能的企业比例已超过四成，制造、零售、住宿餐饮、建筑等传统行业的使用比例也成倍增长（见图

10 麦肯锡. 2026 年人工智能现状：走向投资回报（The State of AI in 2026: On the Road to ROI）[R/OL]. (2026-08-25)[2026-10-21]. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

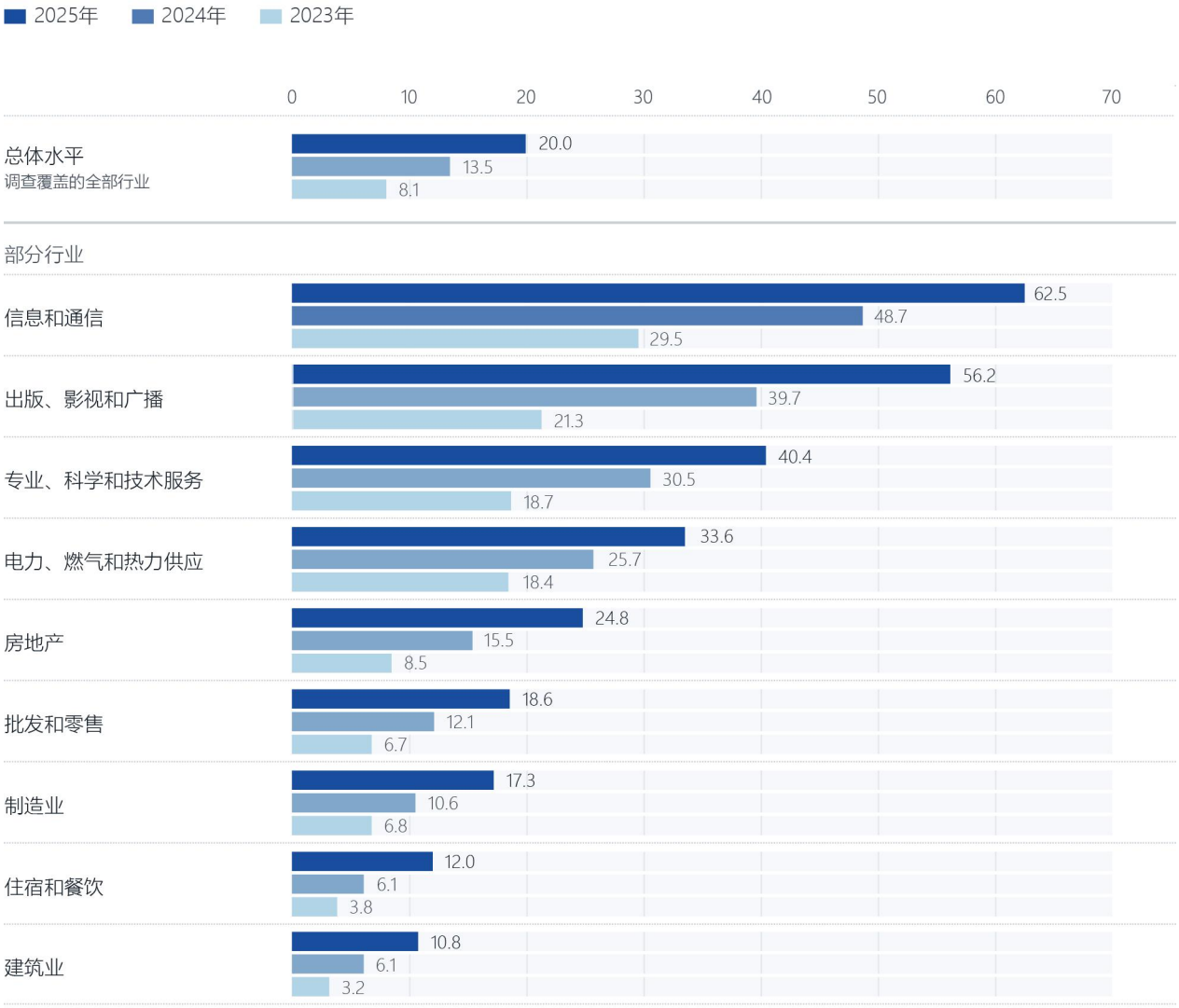
1.4）。麦肯锡2026年覆盖97个国家的调查显示，近九成受访者表示，所在组织已在至少一个业务职能中经常使用人工智能[11]。

从应用深度看，前沿人工智能正由信息处理和辅助支持，进一步进入辅助决策、流程协同和任务执行等更核心环节。在信息处理层，人工智能主要承担知识问答和内容生成的任务。在辅助决策层，系统开始参与风险识别、方案比较、诊疗辅助、金融风控和生产排程。在任务执行层，智能体可按照授权调用工具、填写表单、运行代码、触发流程或协调多个系统。人工智能应用正在不断

图 1.4

欧盟各行业人工智能使用比例普遍上升

使用人工智能的企业比例（%），2023—2025年；企业规模为10人及以上



注：1、 总体比例覆盖调查中的全部行业，并非图列行业比例的简单平均。
2、 “出版、影视和广播”包含在“信息和通信”中。
3、 2025年调查新增“使用人工智能生成图像、视频和音频”类别。

数据来源：欧盟统计局（Eurostat）企业信息通信技术使用与电子商务调查，2023—2025年。

11 麦肯锡. 2026 年人工智能现状：迈向投资回报（The State of AI in 2026: On the Road to ROI）[R/OL]. (2026-08-25)[2026-10-21]. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai. 调查于 2026 年 5 月 4 日至 6 月 8 日开展，共 1,719 名受访者，覆盖 97 个国家。

深入组织运行和业务流程的各个环节（见图1.5）。

从应用形态看，前沿人工智能正从内容生成和知识问答，拓展为工具调用、智能体自主执行和多系统协作等多种形态。内容生成和知识问答仍是最广泛的使用方式，但工具调用使模型能够连接外部软件、数据库和业务系统，智能体自主执行使模型能够在较长时间内围绕目标连续行动，多系统协作则使人工智能嵌入组织流程和产业链条。应用形态的拓展提升了服务能力和生产效率，也使人工智能从信息提供者逐步转向业务流程的实际参与者和任务操作者（见图1.6）。

图 1.5

前沿人工智能应用正逐步深入业务流程

应用深度从信息处理逐级递进至任务执行，越深入越接近组织运行和实际执行环节

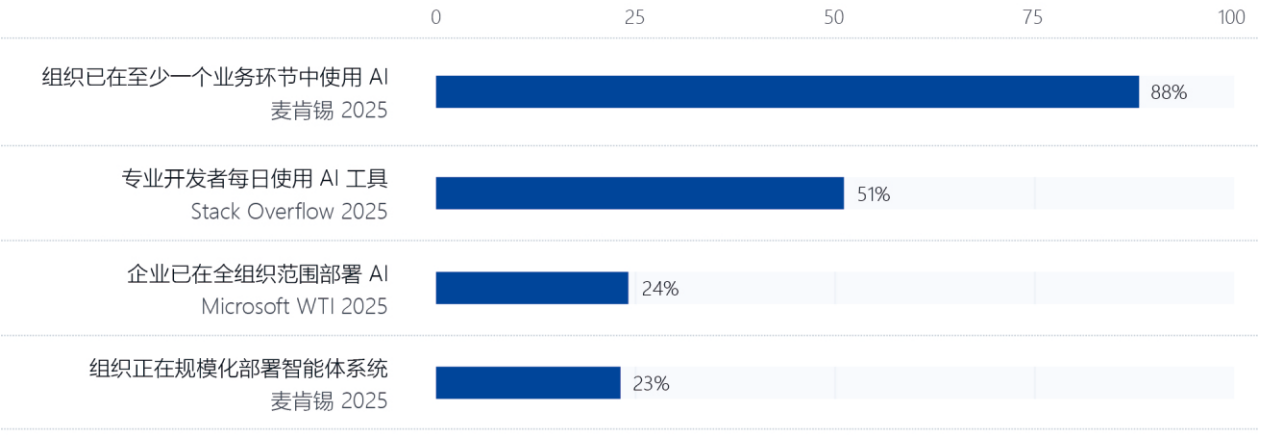


数据来源：本报告整理绘制。

图 1.6

人工智能已进入多类业务和专业场景

不同公开调查显示，人工智能已进入组织使用、专业使用、规模化部署和智能体应用等场景



数据来源：麦肯锡《The State of AI in 2025》、Stack Overflow Developer Survey 2025、Microsoft Work Trend Index 2025；本报告综合整理绘制。注：各项来自不同调查，调查对象和口径不同，仅用于展示应用范围，不作横向排序或阶段比较。

(三) 前沿人工智能的风险演进

随着前沿人工智能能力持续提升、应用不断深化，技术在驱动创新的同时，**相关风险也加速向现实场景渗透，风险演化逻辑正在发生深刻变化。**

能力演进改变了风险产生的条件，部分高风险行为更容易被低成本、规模化实施。前沿人工智能具备了更强的多模态处理能力，不仅提高了任务执行的效率，也降低了违法犯罪行为的实施门槛，使相关滥用行为更易自动化、规模化实施。例如，合成语音已被用于冒充技术支持人员，诱骗用户付费或安装恶意软件，损害用户权益和被冒充企业的信誉[12]。

应用深入改变风险作用环节，相关风险更容易进入业务流程、组织决策和现实环境。当人工智能被嵌入更多行业和组织流程，风险不再仅表现为模型输出偏差或内容失真，而可能进一步表现为智能体自主行动直接作用于业务流程和现实系统，影响流程运行、服务交付和管理决策。例如，2025年7月，一款在线编程平台的智能体在用户要求暂停修改代码期间，删除了企业正在使用的业务数据库，迫使平台紧急上线开发与生产数据库隔离等防护措施[13]。

能力演进及场景嵌入相互叠加，改变了风险传导路径，相关风险更易借助技术系统与组织流程的耦合实现跨边界外溢。在智能体和多系统协作的任务链条中，风险可能从生成计划延伸到工具调用，并伴随结果反馈或前馈进入下一步任务，造成风险的产生、累积与扩大。在组织流程中，模型输出一旦进入辅助决策和任务执行，风险可能沿审批、运营、风控、客服等环节继续传导。与此同时，风险高低并不单纯取决于模型规模，专业化能力、工具调用和系统的组织与协作方式同样影响着风险的产生及其后果。在公共信息环境中，合成内容和自动化传播工具相互配合，可能推动虚假信息、欺诈内容和操纵性叙事沿社会网络扩散。

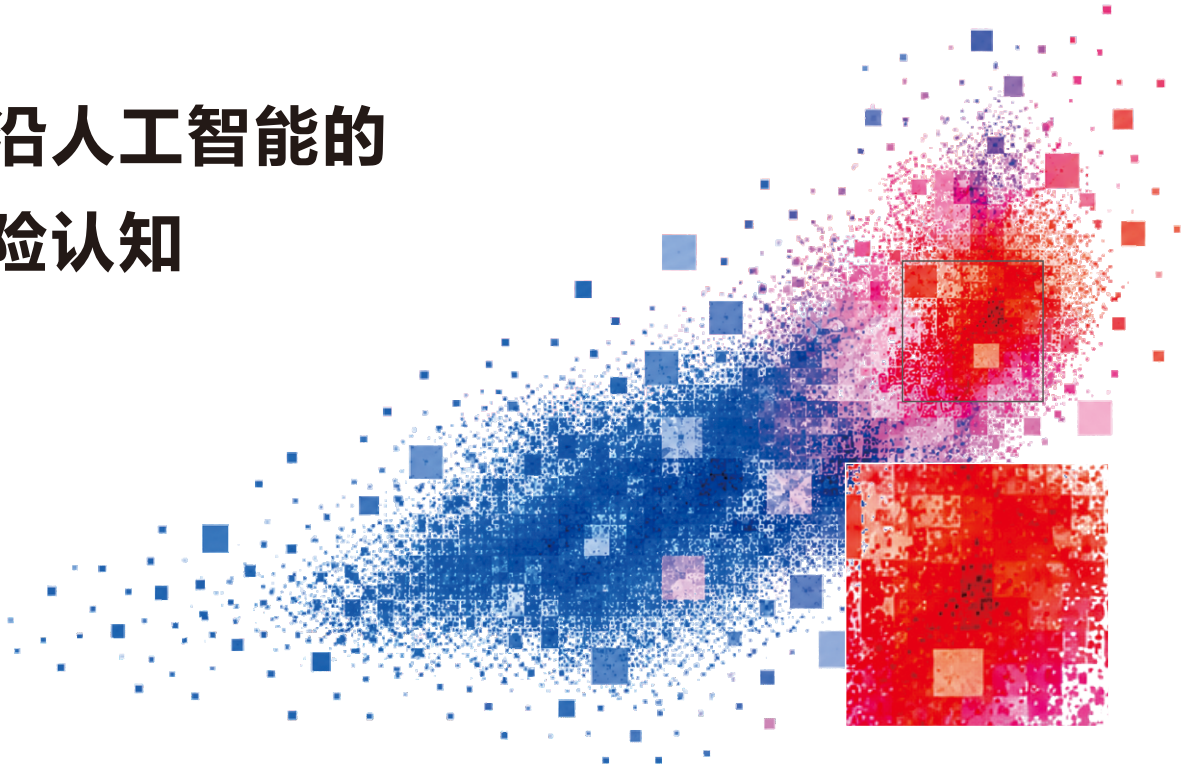
总体看，前沿人工智能能力提升与应用深入相互叠加，正在共同改变风险的形成方式、作用范围和扩散路径。风险不再主要表现为单一模型输出引发的局部问题，而是在数据供给、权限调用、工具连接、流程配置、组织复核和外部传播等多个环节中同步生成、相互叠加，并可能在实际部署中跨场景、跨主体连续放大。**前沿人工智能风险由此加速演变为更具复合性、系统性和外溢性的现实挑战，对风险应对与治理在识别、评估、协同和响应等方面提出更高要求。**面对这一趋势，需要进一步提升对风险表现、形成机制和潜在影响的识别与评估能力，系统检视不同治理主体在制度、技术和组织层面的就绪状态，并围绕具体风险特点和影响形成更有效的协同应对。

12 微软. 微软数字防御报告 2025（Microsoft Digital Defense Report 2025）[R/OL]. [2026-10-21]. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Microsoft-Digital-Defense-Report-2025.pdf. 见第 33 页。

13 Fortune. 一款人工智能编程工具删除了一家软件公司的数据库（An AI-powered coding tool wiped out a software company's database, then apologized for a 'catastrophic failure on my part'）[N/OL]. (2025-07-23)[2026-10-21]. https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/.

第二章

前沿人工智能的风险认知



本章要点

- 建立相对清晰、可比较、可映射的前沿人工智能风险识别体系，是推动风险治理的重要基础。
- 现有前沿人工智能风险分类框架从风险来源、影响后果、应用场景和治理要求等角度展开，为识别前沿人工智能重点风险提供了多种依据。
- 从风险影响看，前沿人工智能风险主要涉及滥用与犯罪风险、信息完整性风险、关键系统风险、社会认知与结构风险、灾难性与生存性风险等方面。
- 从风险态势看，滥用与犯罪风险、信息完整性风险和关键系统风险的现实压力最为集中，社会认知与结构风险影响逐步累积，灾难性与生存性风险潜在后果重大，均需纳入持续评估。
- 应对重点因风险情形而异：紧迫而影响相对局部的情形重在快速处置、阻断扩散；影响逐步累积的情形重在长期监测、及早调整；后果严重且难以逆转的情形重在部署前评测并保留干预手段；其余情形纳入常规管理并定期复核。

随着前沿人工智能风险逐步进入现实场景，风险认知需要从一般性讨论转向更清晰的识别、评估和排序。不同风险在影响对象、发生条件、扩散速度和潜在后果上存在差异，风险的应对与治理也需要据此形成相对稳定的判断基础。当前，国际上已形成多种人工智能风险的分类框架，为识别前沿人工智能重点风险提供了参照。本报告以风险影响为主线，结合模型评测、现实事件和专家研判，观察主要风险的演变态势，将分散的风险信号纳入相对统一的分析框架，为形成共同风险语言、按风险情形确定应对重点提供基础。

(一) 前沿人工智能的风险识别

建立相对清晰、可比较、可映射的前沿人工智能风险识别体系，是推动风险治理的重要基础。风险识别的重点在于明确主要风险及其具体表现和影响后果，并将其与模型能力、现实事件以及既有分类框架和分级治理要求相互对应。清晰的识别体系有助于界定风险边界，可比较的识别体系有助于判断不同风险的现实压力和相对严重程度，可映射的风险识别体系有助于把分散的风险信号转化为可用于评估和协作的共同参照。由此，监管部门、企业和科研机构等各方可以在相对一致的风险认知基础上开展评估、监测通报和治理协作。

现有人工智能风险框架从不同维度识别和管理风险，为前沿人工智能风险识别和治理提供了多种依据。例如，**在公共治理层面，**联合国有关建议文件从伦理影响[14]和全球治理关切[15]等角度识别人工智能风险；欧盟《人工智能法案》[16]按照应用风险等级划分风险类别并配置差异化义务；美国国家标准与技术研究院（NIST）《人工智能风险管理框架》[17]围绕治理、映射、测量和管理四个方面指导风险应对，并强调结合具体应用情境界定风险；中国《人工智能安全治理框架3.0》[18]将安全风险分为内生、应用和衍生三类，并专门提出针对智能体的风险管理框架。**在研究与产业实践层面，**相关框架则从知识分类、风险来源与后果、模型能力及其阈值等角度提供参照。例如，麻省理工学院（MIT）人工智能风险知识库[19]从知识分类角度系统梳理风险类型；《国际人工智能安全报告》[20]主要依据风险来源和影响后果，将风险概括为恶意使用、系统故障和系统性风险三类；一些前沿企业提出的安全框架则围绕模型能力、安全评估和能力阈值等开展风险分类[21]。这些框架各有侧重，共同构成了识别前沿人工智能风险的重要参照体系。

从风险影响视角看，前沿人工智能风险

14 联合国教科文组织. 人工智能伦理问题建议书（Recommendation on the Ethics of Artificial Intelligence）[R/OL]. [2026-10-21]. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence.
15 联合国人工智能高级别咨询机构. 为人类治理人工智能（Governing AI for Humanity）[R/OL]. [2026-10-21]. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.
16 欧盟. 人工智能法案（Artificial Intelligence Act）[EB/OL]. [2026-10-21]. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
17 美国国家标准与技术研究院. 人工智能风险管理框架（AI Risk Management Framework）[R/OL]. [2026-10-21]. https://doi.org/10.6028/NIST.AI.100-1.
18 全国网络安全标准化技术委员会. 人工智能安全治理框架 3.0[R/OL]. 2026-09-14[2026-10-21]: 37-45. https://www.cac.gov.cn/rootimages/uploadimg/1791137114683961/1791137114683961.pdf.
19 麻省理工学院未来技术项目. 人工智能风险知识库[DB/OL]. [2026-10-21]. https://airisk.mit.edu/.
20 国际人工智能安全报告专家组. 国际人工智能安全报告 2026（International AI Safety Report 2026）[R/OL]. [2026-10-21]. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.
21 OpenAI. 准备框架（Preparedness Framework）[EB/OL]. [2026-10-21]. https://openai.com/index/updating-our-preparedness-framework/；Google DeepMind. 前沿安全框架（Frontier Safety Framework）[EB/OL]. [2026-10-21]. https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/；Anthropic. 负责任扩展政策（Responsible Scaling Policy）[EB/OL]. [2026-10-21]. https://www.anthropic.com/responsible-scaling-policy；Meta. 前沿人工智能框架（Frontier AI Framework）[EB/OL]. [2026-10-21]. https://about.fb.com/news/2025/02/meta-approach-frontier-ai/.

需要考虑滥用与犯罪风险、信息完整性风险、关键系统风险、社会认知与结构风险、灾难性与生存性风险等方面。其中，滥用与犯罪风险主要体现为恶意使用对个人和组织权益、生产和管理安全、经济和社会秩序造成的**现实安全受损；**信息完整性风险主要体现为合成内容和自动化传播侵蚀信息准确性、一致性和可靠性，并进而损害公共信任和公共讨论基础的**信息环境失真；**关键系统风险主要体现为人工智能嵌入基础设施、公共服务和重要决策后，由于系统差错、运行故障、恶意攻击等因素可能引发的**关键系统失稳；**社会认知与结构风险主要体现为人工智能长期、广泛应用对社会结构和个体发展产生的**累积影响失衡；**灾难性与生存性风险主要体现为人工智能影响战略安全、降低危险能力获取门槛或超出人类有效控制等情形下带来的**重大后果难逆。**上述分析围绕风险的主要影响展开。这五个方面的风险并非彼此独立，同一事件往往同时涉及多个方面。这一分析视角有助于在后果相互交织的情况下把握风险的主要影响，也为分析具体事件的多重风险提供参照。因此，风险识别和应对既需针对不同风险情形采取相应措施，也需统筹考虑风险之间的相互关联及叠加影响。

滥用与犯罪风险表现为前沿人工智能被用于诈骗、网络攻击、危险知识获取等场景，对个人安全、公共安全和社会秩序造成直接冲击。前沿人工智能降低了生成钓鱼邮件、仿冒身份、编写恶意代码和检索危险知识的技术门槛，也提高了相关恶意行为的自动化和规模化程度。2025年4月以来，美国联邦调查局记录到利用人工智能生成语音冒充政府高级官员的恶意信息活动，攻击者借此建立信任并试图获取账户、敏感信息或资金[22]。这些风险的影响对象往往较为直接，既可能伤害个人财产和隐私，也可能影响组织安全和公共秩序。

信息完整性风险表现为错误、虚假或误导性内容被大规模生成和传播，对公共讨论、社会信任和信息环境产生持续影响。生成式人工智能可以低成本生产文本、图像、音频和视频内容，使深度伪造、虚假消息和误导性叙事更容易被批量制造和快速传播。当内容真实性、来源可追溯性和平台识别能力受到挑战，公众辨别信息真伪的成本也会随之上升。2025年爱尔兰总统选举前，社交平台出现深度伪造视频，虚假宣称候选人凯瑟琳·康诺利退出竞选；相关视频经举报后因违反平台规则被移除[23]。这些风险主要影响公共信息环境，既可能削弱公众信任，也可能干扰公共讨论和社会治理。

关键系统风险表现为人工智能进入基础设施、公共服务和重要决策场景后，因其失灵、误判或使用者对其过度依赖引发连锁影响。在目前人工智能的使用场景中，医疗、金融、交通、能源、政务等关键领域的使用对人工智能系统的可靠性、连续性和责任可追溯性要求更高。一旦系统输出错误、权限配置不当或人机协同失灵，影响可能从单一决策扩展到组织流程、公共服务和关键基础设施运行。金融稳定理事会（FSB）2025年发布的监测报告分析指出，生成式人工智能供应链在专用芯片、云服务和预训练模型等环节集中于少数供应商，可替代的选择有限；金融机构若在核心业务中依赖这些供应商，供应商一旦中断服务，可能同时波及多

22 美国联邦调查局互联网犯罪投诉中心. 恶意信息活动冒充美国高级官员（Senior U.S. Officials Impersonated in Malicious Messaging Campaign）[EB/OL]. (2025-05-15)[2026-10-21]. https://www.ic3.gov/PSA/2025/PSA250515.

23 爱尔兰媒体监管机构（Coimisiún na Meán）. 面向学生的信息（Information for Students）[EB/OL]. [2026-10-21]. https://www.cnam.ie/general-public/guides-resources/for-students/. 见"Threat to democracy"部分。

家机构[24]，形成系统性金融风险。

社会认知与结构风险表现为人工智能长期嵌入就业、教育和社会生活，对就业结构、经济利益分配、公平机会和人的身心发展和自主判断产生深层影响。这些风险通常不以单一事件集中呈现，更多是在岗位变化、教育使用、内容推荐、自动化决策和长期工具依赖中逐步累积。其影响既包括不同群体在技能、资源和机会上的差距，也包括个人判断、知识积累、信息辨识、专业能力和社会认知方式的变化。在就业方面，韩国央行的研究显示，2022年6月至2026年6月，15—29岁青年就业岗位减少28.5万个，其中26.8万个集中在人工智能暴露度较高的行业；而同期这类行业中，50—59岁人员的就业反而有所增加[25]。教育中的风险也已出现：一项近千名高中生参与的数学实验发现，可直接索取答案的人工智能辅导工具虽然提高了练习成绩，但撤去工具后，这些学生的考试成绩比从未使用过人工智能工具的学生低约17%[26]。此外，人工智能系统通过持续对话和拟人化回应来提供信息、陪伴和情感支持的同时，这种长期互动对用户身心健康、自主判断、情感依赖和现实人际关系的影响也需要关注[27]。

灾难性与生存性风险表现为前沿人工智能的能力及其影响可能超出现有治理和安全保障能力　进而造成范围广泛且难以逆转的重大后果。目前，相关讨论主要集中在高度自主系统失控、欺骗行为、自我复制和危险能力扩散等高影响情形。相关研究表明，前沿人工智能模型在测试过程中突破预期环境边界并影响现实系统的情形已经出现[28]；在虚构企业情境中开展的测试还显示，多家企业的部分前沿人工智能模型会以泄露敏感信息相威胁，试图阻止自身被替换或关停[29]。相关讨论还进一步延伸至超级人工智能[30]失控的极端情形：人类目前尚缺乏确保超级人工智能始终可控的严格科学保证，一旦其能力进一步超出人类有效理解与控制范围并发生失控，可能引发人类难以承受、难以逆转的灾难性甚至生存性后果；即使不以超级人工智能为直接研发目标，具有持续自我改进或能力扩展能力的系统，也可能在能力快速提升过程中进入超出原有预期的能力水平。因潜在后果严重，这类风险仍需充分纳入前瞻研判。

24 金融稳定理事会. 金融部门人工智能应用及相关脆弱性监测（Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector）[R/OL]. (2025-10-10)[2026-10-21]. https://www.fsb.org/uploads/P101025.pdf. 见第 4 节生成式人工智能供应链分析。

25 韩国央行. 青年岗位减少，真的是人工智能的影响吗？（청년 일자리 감소, 정말 AI 탓일까?）[EB/OL]. (2026-08-28)[2026-10-21]. https://www.bok.or.kr/portal/bbs/B0000347/view.do?menuNo=201106&nttId=11064157. 行政记录显示就业变化，研究指出招聘调整等因素共同作用，不能单独识别人工智能的因果影响。

26 BASTANI H, BASTANI O, SUNGU A, et al. Generative AI without guardrails can harm learning: Evidence from high school mathematics[J/OL]. Proceedings of the National Academy of Sciences, 2025, 122(26): e2422633122[2026-10-21]. https://doi.org/10.1073/pnas.2422633122.

27 International AI Safety Report 2026, p. 92, Box 2.6 "AI companions". https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

28 OpenAI. Hugging Face 事件及后续工作（The Hugging Face Incident and the Road Ahead）[EB/OL]. (2026-08-26)[2026-10-21]. https://openai.com/index/hugging-face-incident-and-the-road-ahead/；Anthropic. 调查网络安全评测中的三起现实事件（Investigating Three Real-World Incidents in Our Cybersecurity Evaluations）[EB/OL]. (2026-07-30)[2026-10-21]. https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals. OpenAI 事件涉及模型绕过控制并影响现实基础设施；Anthropic 事件中，评测环境的配置错误使模型能够访问并影响外部系统。

29 Anthropic. 智能体错位：大语言模型何以成为内部威胁（Agentic Misalignment: How LLMs Could Be Insider Threats）[EB/OL]. (2025-06-20)[2026-10-21]. https://www.anthropic.com/research/agentic-misalignment. 测试在研究人员设计的虚构企业情境中进行；其中的人员、组织及胁迫情节均为模拟设置。

30 本报告所称"超级人工智能"，是指在广泛认知与复杂任务等重要能力领域达到或显著超越人类最高水平的人工智能系统。鉴于该术语在学术和产业界尚无完全统一的定义，本报告采用上述操作性定义，不特指某一具体技术路线、系统形态或研发目标。

(二) 前沿人工智能的风险评估

前沿人工智能风险已从潜在担忧转化为现实压力，其中滥用与犯罪风险、信息完整性风险和关键系统风险最为突出。当前，对相关风险的判断主要基于模型安全评测、风险事件监测以及专家和机构研判，分别从潜在风险能力、现实危害和长期及高影响风险等方面提供证据。现有评估显示，部分模型已展现出高风险能力，现实事件显示相关危害已影响个人、组织和公共环境，专家和机构研判也持续把网络攻击、危险知识获取、社会信任侵蚀和关键场景失灵列为重点关注对象。随着前沿人工智能能力快速提高和应用日益广泛，相关风险的现实暴露压力正在上升。

对前沿人工智能模型与智能体的安全评测普遍显示，网络攻击、危险知识获取、长程任务执行和安全防护绕过等方面的高风险能力持续显现，滥用与犯罪风险和关键系统风险更加突出。模型层面的评测显示，前沿人工智能模型尤其擅长发现软件漏洞和编写恶意代码，可在网络攻击的多个环节提供辅助，既可用于发现和修复漏洞，也可被用于实施攻击[31]；近期发布的前沿人工智能模型在自动发现和利用漏洞方面的能力进一步增强[32]。当模型接入工具、在任务环境中作为智能体运行时，评测还暴露出单独测试模型时不易看到的风险。配备工具的智能体能够完成的软件任务持续变长、变复杂[33]；在缺少主动防御的受控网络靶场中，获得明确指令和网络访问权限的前沿人工智能模型已能自主完成多步骤攻击[34]，但在防护严密的目标上尚不能独立完成攻击全过程[35]；在模拟操作系统、数据库等工具交互环境和长程任务中的测试还发现，部分模型会执行越权操作，或未能遵从人的中断指令[36]。随着模型能力不断增强，一旦被误用或恶意利用，相关风险将更容易进入真实系统和高影响场景，对网络安全和关键系统运行构成更大压力。

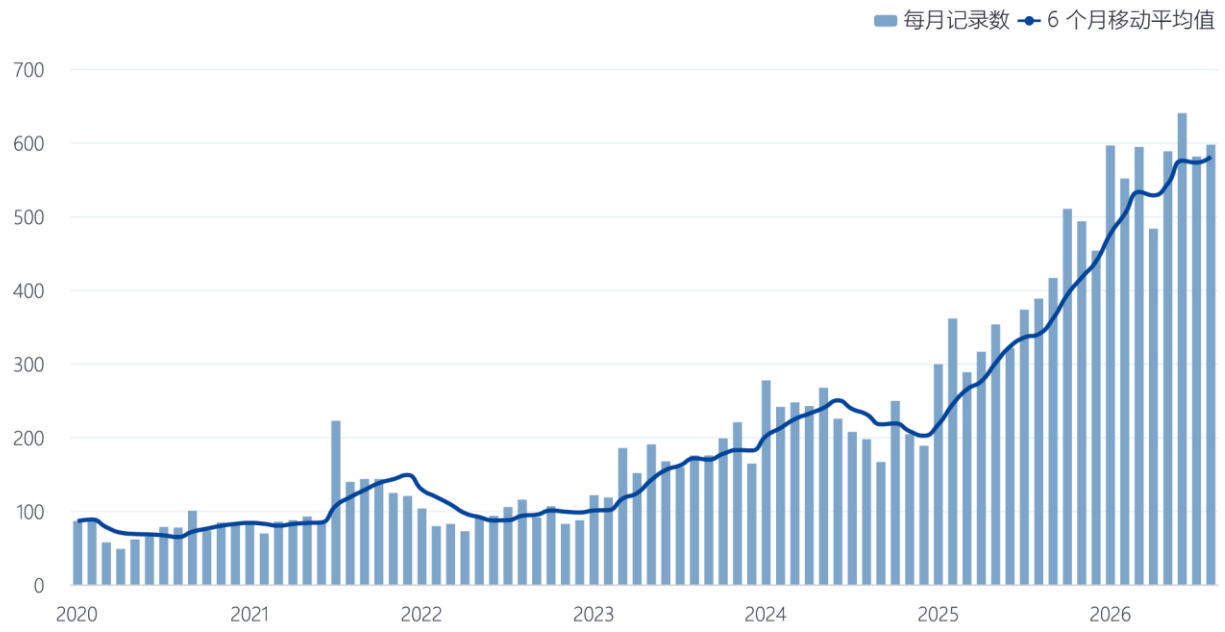
风险事件记录显示，人工智能风险正在从技术可能性转化为持续暴露的现实危害。深度伪造诈骗、信息误导、网络攻击和关键场景误用等问题在数量和类型上均呈扩散趋势。从数量看，经济合作与发展组织人工智能事件与危害监测（OECD AIM）显示，媒体报道的风险事件与危害记录数量从2024年约2700起增至2025年约4600起，2026年1至8月已超过4600起[37]（见图2.1）。从类型和影响看，这些事件与记录已涉及深度伪造诈骗、身份冒充、虚假内容传播、自动化决策争议和关键服务误用等场景，并对基本权利、公共利益、声誉、人身安全和经济财产等造成影响（见图2.2）。滥用与犯罪风险尤为突出，有研究显示，生成式人工智能正被用于扩大自动化网络攻击，使攻击呈现更高速度、规模和复杂程度[38]。风险已进入个人、企业和公共信息环境，造成实质性经济损失和组织运行影响。

31 国际人工智能安全报告专家组. 国际人工智能安全报告 2026（International AI Safety Report 2026）[R/OL]. [2026-10-21]. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.
32 Anthropic. Claude Fable 5 与 Claude Mythos 5[EB/OL]. [2026-10-21]. https://www.anthropic.com/news/claude-fable-5-mythos-5.
33 模型评估与威胁研究中心. 前沿人工智能模型任务完成时间跨度[EB/OL]. [2026-10-21]. https://metr.org/time-horizons/.
34 英国人工智能安全研究所. Claude Mythos Preview 网络能力评估[EB/OL]. [2026-10-21]. https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.
35 OpenAI. GPT-5.6 系统卡[R/OL]. (2026-07-09)[2026-10-21]. https://deploymentsafety.openai.com/gpt-5-6.
36 北京前瞻人工智能安全与治理研究院. 前瞻安全评测基准[DB/OL]. [2026-10-21]. https://foresightsafety-bench.beijing-aisi.ac.cn/.
37 经合组织人工智能政策观察站. 经合组织人工智能事件与危害监测器[DB/OL]. [2026-10-21]. https://oecd.ai/en/incidents. 记录由媒体报道汇集并经自动分类、聚类处理，覆盖范围、报道倾向和收录方法变化均会影响数量；记录增长不等同于实际发生率同比增长，也不替代用于处置事件的运营通报。方法见：经济合作与发展组织. 人工智能事件与危害监测概述及方法[EB/OL]. [2026-10-21]. https://oecd.ai/en/incidents-methodology. 关于跨法域、跨部门的事件报告字段，参见：经济合作与发展组织. 迈向人工智能事件共同报告框架（Towards a Common Reporting Framework for AI Incidents）[R/OL]. (2025-02-28)[2026-10-21]. https://doi.org/10.1787/f326d4ac-en.
38 国际人工智能安全报告专家组. 国际人工智能安全报告 2026（International AI Safety Report 2026）[R/OL]. (2026-02)[2026-10-21]. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. 见第 2.1.3 节。

图 2.1

媒体报道的人工智能风险事件月度数量总体上升

OECD AIM根据媒体报道汇集的事件及潜在危害记录数与6个月移动平均值，2020年1月—2026年8月



数据来源：经济合作与发展组织人工智能事件与危害监测（OECD AIM）；记录时间截至2026年8月31日，本报告整理绘制。

图 2.2

人工智能风险事件涉及基本权利、人身安全和经济财产等多类危害

媒体报道的事件记录中，各类危害占危害标注总数的比例（%），2020年1月—2026年8月；一起事件可涉及多类危害



注：仅统计具有危害标签的 OECD AIM 事件记录；同一事件可涉及多个危害标签，图中比例按危害标签总数计算，仅列出占比0.5%以上的标签。
数据来源：经济合作与发展组织人工智能事件与危害监测（OECD AIM）事件记录中的危害标签；记录时间截至2026年8月31日，本报告整理绘制。

专家和机构研判进一步提示，除模型评测和事件统计已经显示的现实风险外，社会认知与结构风险以及灾难性与生存性风险也需要纳入评估视野。围绕潜在高影响风险，国际社会已开始关注并讨论人工智能"红线"问题，相关倡议将未经授权的自我复制、欺骗、权力寻求、网络攻击和协助制造武器等能力列为需要提前防范和约束的重点[39]。在生物安全领域，近期国际专家共识指出，前沿人工智能在病原体设计等任务上的能力提升正在降低生物滥用的专业门槛，可能带来重大公共卫生风险[40]。这两个方面的风险虽然尚未像欺诈、信息造假和网络攻击那样集中表现为高频现实事件，但分别体现出影响长期累积和潜在后果重大的特征，因而也是前沿人工智能风险评估需要覆盖的重要内容。

总体看，模型安全评测、风险事件统计和专家机构研判从不同角度表明，前沿人工智能风险的现实压力正在上升。滥用与犯罪风险、信息完整性风险和关键系统风险已经在模型能力、现实事件和应用场景中表现得更为集中；社会认知与结构风险虽然更多体现为长期累积影响，灾难性与生存性风险的触发条件和发生概率仍存在不确定性，但二者分别因影响范围扩大和潜在后果重大，也需要纳入持续评估（见图2.3）。

图 2.3

前沿人工智能重点风险的证据与严重程度存在差异

重点风险的当前证据与严重性判断

01	滥用与犯罪风险	当前证据	现实事件多、增长快，网络攻击和欺诈证据充分
		严重性判断	现实危害程度高，事件数量仍在快速上升
02	信息完整性风险	当前证据	合成内容和虚假信息持续扩散，识别与溯源仍不稳定
		严重性判断	现实危害程度高，波及范围持续扩大
03	关键系统风险	当前证据	高风险场景部署增加，失灵、误判和过度依赖风险上升
		严重性判断	危害程度上升，潜在影响较大
04	社会认知与结构风险	当前证据	影响长期累积，就业、教育和社会认知变化仍在显现
		严重性判断	当前危害有限，影响长期累积
05	灾难性与生存性风险	当前证据	现实暴露有限，主要来自能力评测和专家研判
		严重性判断	现实危害正在显现，潜在后果严重

数据来源：本报告整理绘制。

(三) 前沿人工智能风险的优先应对

前沿人工智能风险表现多样、演进较快，治理资源和社会关注有限，需要在不同风险情形之间作出平衡。即使涉及同一风险方面，不同情形的紧迫程度和潜在后果也可能存在差异，需要结合具体情形确定应对重点。

本报告从紧迫程度和潜在后果两个方面刻画风险情形。**紧迫程度**由危害已发生的情况、扩散速度和干预窗口共同决定：危害在

39 国际人工智能安全对话. 北京共识声明[EB/OL]. [2026-10-21]. https://idais.ai/；人工智能红线倡议. 人工智能红线全球呼吁（Global Call for AI Red Lines）[EB/OL]. [2026-10-21]. https://red-lines.ai/；经合组织人工智能政策观察站. 通过全球红线开展人工智能治理有助于防止不可接受风险[EB/OL]. [2026-10-21]. https://oecd.ai/en/wonk/ai-governance-through-global-red-lines-can-help-prevent-unacceptable-risks.

40 国际人工智能安全对话. IDAIS 伦敦共识声明（IDAIS-London, 2026）[EB/OL]. (2026-04-19)[2026-10-21]. https://idais.ai/dialogue/idais-london/.

现实中发生越多、同类危害复制扩散越快、有效干预的时间越短，紧迫程度越高。**潜在后果**取决于危害可能波及的范围、持续时间和可逆性。这些判断基于公开证据作定性比较，不对紧迫程度和潜在后果设定量化分界。按这两个方面可以区分四种典型情形，分别以快速处置、长期监测、前置把关和常规管理为行动重心（见图2.4）。

紧迫程度高、单起事件影响相对局部的情形，应对重点在于快速处置，阻断危害扩散。可通过风险通报和平台、机构间的联动及时识别同类事件，根据滥用情况采取相应处置措施，并向可能受影响的人群提供预警和救济支持。

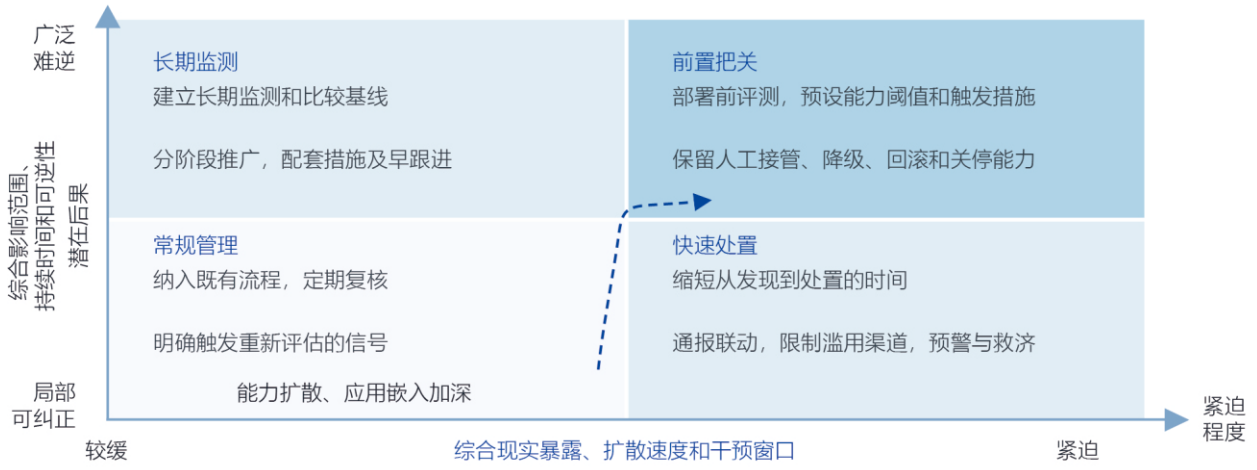
潜在后果广泛、影响逐步累积、干预窗口相对较长的情形，应对重点在于长期监测和适时调整应用安排。可按人群、职业和场景持续跟踪影响，在应用推广过程中同步评估，并结合评估结果调整培训、转岗支持和教学安排。

紧迫程度高、潜在后果广泛且难以逆转的情形，应对重点在于部署前的风险评估，以及部署后的干预准备。可结合危险能力评测和红队测试结果，判断是否需要加强防护、限制使用范围或暂缓部署，并根据应用场景配置人工接管、回滚等措施。对于可能走向超级人工智能阶段的能力跃迁，更应坚持部署前把关，在确认充分安全可控之前，对相关研发、部署与使用宜保持审慎。

紧迫程度较低、后果相对局部且容易纠正的情形，应对重点在于日常管理和定期复核。可通过质量管理、人工复核和用户反馈发现并纠正问题；当应用范围明显扩大、进入高影响场景或同类事件集中增加时，再结合具体情况评估是否需要调整管理措施。

风险情形会随能力演进、应用嵌入和证据积累而变化，优先应对需要定期复核。能力扩散可能让原本只在受控测试中出现的风险进入现实场景；应用深度嵌入业务流程和关键系统，可能缩短干预窗口、扩大潜在后果；防护措施完善后，部分风险也可能回到常规管理的范围。将有限的治理资源投向当前最需要处置、监测和前置评测的环节，并随证据更新及时调整，优先应对才能保持有效。

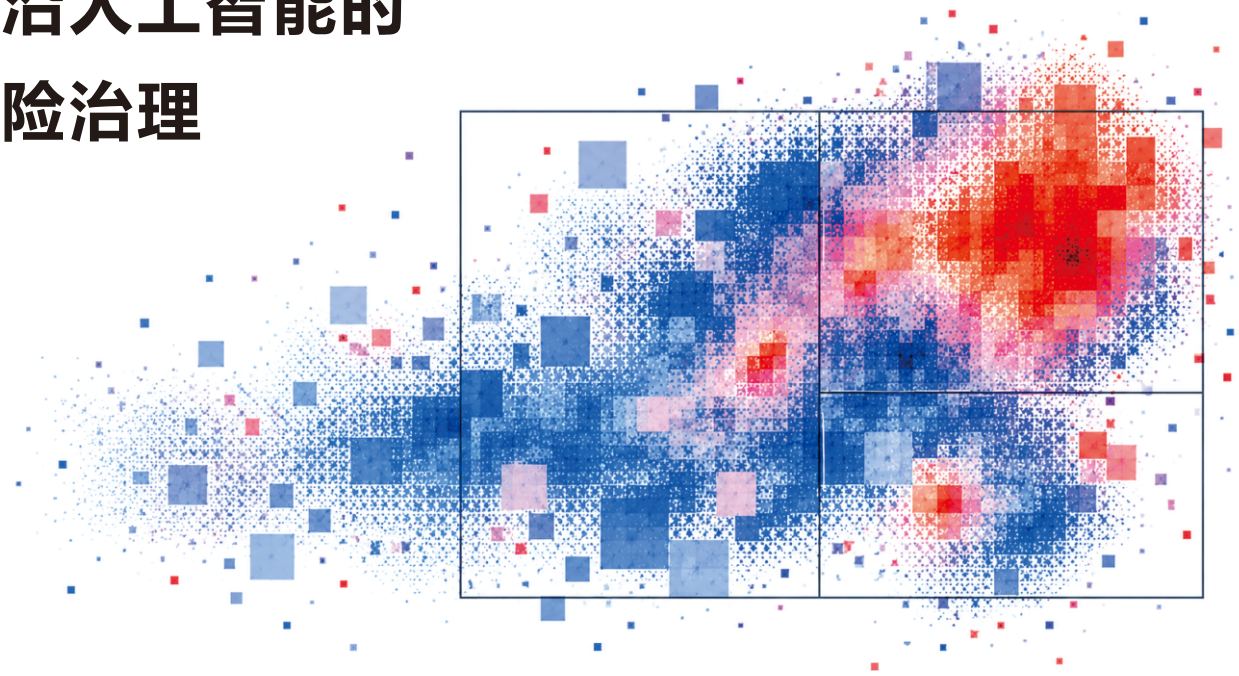
图 2.4
紧迫程度与潜在后果不同，优先行动的重心也不同
按紧迫程度和潜在后果区分的四种典型风险情形及其行动重心



注：同一风险方面可能涉及不同情形；这些情形在图中的位置会随能力扩散、应用嵌入和防护措施的变化而移动。
数据来源：本报告整理绘制。

第三章

前沿人工智能的风险治理



本章要点

- 多方立足各自角色，持续完善制度规则、嵌入安全要求、积累技术证据、推进国际协作，前沿人工智能治理基础正在形成。
- 治理就绪度评估可用于观察和比较已有治理实践，为判断治理基础、识别薄弱环节和明确能力建设方向提供依据。
- 本报告治理就绪度评估从六个方面将不同主体的现有努力纳入同一评估框架，重点判断其能否支撑前沿人工智能风险应对所需的关键能力。
- 从六个方面看，受评国家在制度与组织方面推进相对最充分，在技术安全与评测、验证与合规、责任落实与生态治理方面相对薄弱。
- 从收入组看，高收入国家整体治理就绪度更高。中等偏上收入国家与中低收入和低收入国家的总体水平相近，但薄弱环节有所不同。
- 从风险视角看，应对重点风险所需的能力就绪度仍不均衡，风险事件通报与共享、应急响应和复盘改进、跨境风险信息通报和演练等方面相对滞后，需要更有针对性的能力建设。

当前，前沿人工智能治理正在由原则倡议走向规则、评测、承诺和协作安排。各国政府持续完善制度规则，国际组织和多边机制推动风险认知、标准协调和能力建设，产业界把安全要求纳入研发和部署流程，学术界提供评测方法和风险证据，多方实践为前沿人工智能风险识别、评测核验、监测响应和协同处置积累了基础，也使治理进展具备了进一步观察和评估的条件。通过治理就绪度评估，可以更清楚地判断当前治理基础、薄弱环节和后续能力建设方向。

(一) 前沿人工智能的治理实践

前沿人工智能风险持续显现，推动政府机构、国际组织、企业、科研机构等多方立足各自角色开展治理实践。过去两年，前沿人工智能能力提升、应用扩散和风险暴露同步加快，治理议题从一般性原则转向更具体的规则、评测、承诺和协作安排。不同主体的实践路径各有侧重，但总体上都指向更早识别风险、更有效约束高风险行为，并在风险跨境外溢时形成协同应对。

各国各地区完善制度规则，围绕高风险应用、模型发布、安全评估和责任落实作出制度安排。欧盟通过综合性人工智能立法配置差异化义务[41]；中国围绕生成式人工智能服务[42]、生成内容标识[43]、拟人化互动[44]、智能体规范应用等[45]持续完善专项规则；美国通过联邦人工智能治理备忘录明确机构责任，并由美国人工智能标准与创新中心（US CAISI）启动智能体标准倡议；英国人工智能安全研究所（UK AISI）持续开展前沿人工智能模型评测，并于2025年12月发布首份《前沿人工智能趋势报告》，汇总对30余个前沿系统的评测研究；日本实施人工智能促进法并发布人工智能经营者指南；韩国人工智能基本法于2026年1月生效；新加坡开展生成式人工智能评测沙盒等实践[46]。不同制度路径存在差异，但共同指向更清晰的规则边界、更明确的监管责任和更可操作的安全要求。

国际组织和多边、区域机制加快搭建协作平台，围绕风险认知、标准协调、能力建设和跨境合作推动规则对接。继2023年28个国家和欧盟共同签署《布莱切利宣言》[47]之

41 欧盟. 人工智能法案（Artificial Intelligence Act）[EB/OL]. (2024)[2026-10-21]. https://eur-lex.europa.eu/eli/reg/2024/1689/oj. 通用人工智能模型相关义务自 2025 年 8 月 2 日起适用。参见：欧盟委员会. 人工智能法案[EB/OL]. [2026-10-21]. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

42 中国国家互联网信息办公室等. 生成式人工智能服务管理暂行办法[EB/OL]. [2026-10-21]. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

43 中国国家互联网信息办公室等. 人工智能生成合成内容标识办法[EB/OL]. [2026-10-21]. https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

44 中国国家互联网信息办公室等. 人工智能拟人化互动服务管理暂行办法[EB/OL]. [2026-10-21]. https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm.

45 中国国家互联网信息办公室等. 智能体规范应用与创新发展实施意见[EB/OL]. [2026-10-21]. https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm.

46 美国白宫管理和预算办公室. 通过创新、治理和公众信任加快联邦机构使用人工智能（Accelerating Federal Use of AI through Innovation, Governance, and Public Trust）[EB/OL]. (2025-04-03)[2026-10-21]. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf; 美国国家标准与技术研究院. 宣布启动"人工智能智能体标准倡议"（Announcing the "AI Agent Standards Initiative" for Interoperable and Secure Innovation）[EB/OL]. (2026-02-17)[2026-10-21]. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure; 英国人工智能安全研究所. 前沿人工智能趋势报告（Frontier AI Trends Report）[R/OL]. (2025-12)[2026-10-21]. https://www.aisi.gov.uk/frontier-ai-trends-report; 日本经济产业省. 人工智能经营者指南（第 1.2 版）（AI Guidelines for Business Ver. 1.2）[EB/OL]. (2026-03-31)[2026-10-21]. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html; 新加坡资讯通信媒体发展局. 生成式人工智能评测沙盒（Generative AI Evaluation Sandbox）[EB/OL]. (2023-10-31)[2026-10-21]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox; 日本内阁府. 人工智能相关技术研究开发及应用促进法[EB/OL]. [2026-10-21]. https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html; 韩国科学技术信息通信部. 人工智能基本法生效[EB/OL]. (2026-01-22)[2026-10-21]. https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1214&sCode=eng.

47 英国政府. 布莱切利宣言（The Bletchley Declaration）[EB/OL]. (2023)[2026-10-21]. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.

后，联合国独立国际人工智能科学小组于2026年7月发布初步报告，首届全球人工智能治理对话在日内瓦举行[48]，多国相关机构组成先进人工智能测量、评估与科学国际网络并开展联合测试[49]。2026年7月，29个国家在上海签署《关于成立世界人工智能合作组织的协定》，作为创始成员国共同设立这一政府间国际组织，其宗旨包括推动人工智能朝着有益、安全、公平的方向发展[50]。同期发布的2026世界人工智能大会暨人工智能全球治理高级别会议主席声明提出，积极推动构建法律法规、技术监测、风险预警和应急响应体系，并在国际上建立必要的危机管控机制[51]。多边机制正在围绕前沿人工智能风险形成更持续的讨论与合作安排，为风险认知、评测方法、标准互操作和能力建设提供共同平台。

产业界逐步将安全要求纳入模型研发、部署和运行流程，通过安全框架、自愿承诺、外部测试、模型发布说明和透明度报告等方式细化安全责任。前沿人工智能模型研发机构已围绕安全评估、能力阈值、发布决策和风险缓解作出共同承诺[52]。与此同时，企业围绕系统全生命周期安全[53]、算力基础设施与大模型应用安全[54]、模型服务平台安全[55]、大模型安全与伦理[56]、安全人工智能框架[57]、人工智能安全与治理框架[58]以及负责任人工智能实践[59]，持续发布白皮书、框架和研究报告；网络安全企业也围绕大模型运行风险[60]和大模型安全治理[61]提出技术方案和治理建议；面向智能体的安全护栏工具[62]、安全智能体成熟度评价[63]和负责任人工智能透明度报告[64]也在不断丰富。总体看，企业安全治理正在从原则声明和自愿承诺，转向在研发流程、模型评估、部署决策和运行保障环节落实的具体要求。

48 联合国人工智能问题独立国际科学小组. 人工智能问题独立国际科学小组初步报告：人工智能机遇、风险与影响的循证评估（Preliminary Report of the Independent International Scientific Panel on AI: Evidence-based Assessment of Opportunities, Risks and Impacts of Artificial Intelligence）[R/OL]. 2026-07-01. [2026-10-21]. https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report；联合国. 全球人工智能治理对话（Global Dialogue on AI Governance）[EB/OL]. [2026-10-21]. https://www.un.org/global-dialogue-ai-governance/en；联合国大会. 设立和运作人工智能问题独立国际科学小组和人工智能治理全球对话的职权范围和模式：第 79/325 号决议[EB/OL]. (2025-08-26)[2026-10-21]. https://docs.un.org/A/RES/79/325.

49 美国国家标准与技术研究院. 先进人工智能测量、评估与科学国际网络发布自动化评估实践共识（International Network for Advanced AI Measurement, Evaluation, and Science Publishes Consensus Areas on Practices for Automated Evaluations）[EB/OL]. (2026-02-13)[2026-10-21]. https://www.nist.gov/news-events/news/2026/02/international-network-advanced-ai-measurement-evaluation-and-science. 该网络原称人工智能安全研究所国际网络。 英国人工智能安全研究所. 国际联合测试：智能体测试（International Joint Testing Exercise: Agentic Testing）[EB/OL]. (2025-07-17)[2026-10-21]. https://www.aisi.gov.uk/blog/international-joint-testing-exercise-agentic-testing.

50 外交部. 成立世界人工智能合作组织协定签署仪式在上海举行[EB/OL]. (2026-07-16)[2026-10-21]. https://www.mfa.gov.cn/wjbzhd/202607/t20260716_11984399.shtml（中国新闻网转载：https://www.chinanews.com.cn/gn/2026/07-16/10660924.shtml）；新华社. 世界人工智能合作组织未来将重点开展三方面工作[EB/OL]. (2026-07-19)[2026-10-21]. https://www.bjnews.com.cn/detail/1784450205129272.html.

51 中华人民共和国常驻联合国代表团. 2026 世界人工智能大会暨人工智能全球治理高级别会议主席声明[EB/OL]. (2026-07-17)[2026-10-21]. https://un.china-mission.gov.cn/zgyw/202607/t20260717_11984707.htm.

52 人工智能首尔峰会. 前沿人工智能安全承诺（Frontier AI Safety Commitments）[EB/OL]. [2026-10-21]. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024；中国人工智能产业发展联盟组织国内企业签署《人工智能安全承诺》，参见：中国信息通信研究院. 人工智能安全治理研究报告（2025 年）[R/OL]. 2026: 9[2026-10-21]. https://www.caict.ac.cn/kxyj/qwfb/bps/202601/P020260109784447548497.pdf.

53 华为. AI 系统的网络安全治理实践[R/OL]. [2026-10-21]. https://www-file.huawei.com/-/media/corp2020/pdf/trust-center/a_look_at_effective_ways_for_ai_system_security_and_privacy_cn.pdf.

54 百度. AI 基础设施安全白皮书[R/OL]. [2026-10-21]. https://bce-doc-on.cdn.bcebos.com/bce-documentation/analyst-reports/0827%E7%99%BD%E7%9A%AE%E4%B9%A62%EF%BC%88%E7%BA%BF%E4%B8%8A%E7%89%88%EF%BC%89.pdf.

55 阿里云. 阿里云百炼安全白皮书[R/OL]. [2026-10-21]. https://developer.aliyun.com/ebook/8490.

56 腾讯. 大模型安全与伦理研究报告 2024：以负责任 AI 引领大模型创新[R/OL]. [2026-10-21]. https://aisecmatrix.org/.

57 谷歌. 安全人工智能框架（SAIF: Google's Guide to Secure AI）[EB/OL]. [2026-10-21]. https://saif.google/.

58 IBM. 规模化的可信人工智能：IBM 人工智能安全与治理框架（Trustworthy AI at Scale: IBM's AI Safety and Governance Framework）[EB/OL]. [2026-10-21]. https://newsroom.ibm.com/blog-trustworthy-ai-at-scale-ibms-ai-safety-and-governance-framework.

59 亚马逊. 负责任人工智能（Responsible AI: Building AI Responsibly）[EB/OL]. [2026-10-21]. https://aws.amazon.com/ai/responsible-ai/.

60 360 数字安全集团. 大模型安全白皮书[R/OL]. 2025-11. [2026-10-21]. http://pub1-bjyt.s3.360.cn/bcms/360%E3%80%8A%E5%A4%A7%E6%A8%A1%E5%9E%8B%E5%AE%89%E5%85%A8%E7%99%BD%E7%9A%AE%E4%B9%A6%E3%80%8B.pdf.

61 奇安信. 政务大模型安全治理框架[R/OL]. [2026-10-21]. https://www.qianxin.com/threat/reportdetail?report_id=323.

62 英伟达. 为智能体人工智能应用提供安全防护的 NIM 微服务（NVIDIA Releases NIM Microservices to Safeguard Applications for Agentic AI）[EB/OL]. [2026-10-21]. https://blogs.nvidia.com/blog/nemo-guardrails-nim-microservices/.

63 安恒信息. 安全智能体魔方成熟度模型评价研究报告[R/OL]. [2026-10-21]. https://www.dbappsecurity.com.cn/content/details2166_46781.html.

64 微软. 2025 负责任人工智能透明度报告（2025 Responsible AI Transparency Report）[R/OL]. [2026-10-21]. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/2025-Responsible-AI-Transparency-Report.pdf.

学术界持续推进安全评测、风险研究和方法建设，为前沿风险识别、证据积累和政策讨论提供专业支撑。国际性专家报告持续更新前沿人工智能风险的科学评估[65]，公开风险事件数据库积累可检验的事件记录[66]，安全评测基准[67]和红线讨论则将复杂技术问题转化为评估指标、测试方法和政策讨论可以使用的共同语言。这些工作增强了风险判断的证据基础，也为不同主体开展评测、监管和协作提供了参照[68]。

多方参与下，前沿人工智能治理实践已初步形成制度规则、技术证据、企业实践和国际协作相互支撑的格局。政府机构提供制度边界，企业将安全要求嵌入研发部署，学术界提供评测方法与风险证据，多边机制推动共识形成和协作安排。这些实践表明，前沿人工智能治理已不再停留于原则倡议，而是逐步进入规则细化、技术支撑、责任落实和协同推进并行发展的阶段。不同主体和不同环节的推进情况，也为进一步观察治理基础和能力条件提供了现实参照。

(二) 前沿人工智能治理就绪度评估框架

有效应对前沿人工智能风险，既需要梳理已有治理举措，也需要观察这些举措背后的制度基础、技术能力和协同条件。**本报告提出治理就绪度评估框架[69]，从制度与组织、技术安全与评测、验证与合规、风险监测预警与响应、责任落实与生态治理、协同治理与能力建设六个方面，观察不同国家及相关主体的前沿人工智能治理就绪度（见图3.1），用于判断治理基础的形成情况、关键能力的薄弱环节以及后续能力建设方向。**

- **制度与组织**构成风险治理的制度基础，法律规则、监管机关与执行资源共同支撑治理体系运转。
- **技术安全与评测**反映识别和控制高风险能力的专业基础，使风险研判更多建立在可检验的证据基础上。
- **验证与合规**连接安全承诺、外部核验和持续执行，是推动承诺从文本要求转化为实际约束的关键条件。
- **风险监测预警与响应**面向风险显现前后的发现、预警、处置和复盘，关系到事件能否被及时识别、联动处理和持续改进。
- **责任落实与生态治理**关注智能体、开放模型、供应链等新场景中的责任边界、行为追溯和生态约束。
- **协同治理与能力建设**面向跨境风险和能力差距，支撑各方在更大范围内开展标准协调、信息通报、联合评测和能力建设。

65 国际人工智能安全报告专家组. 国际人工智能安全报告 2026（International AI Safety Report 2026）[R/OL]. [2026-10-21]. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026；联合国人工智能问题独立国际科学小组. 人工智能问题独立国际科学小组初步报告：人工智能机遇、风险与影响的循证评估（Preliminary Report of the Independent International Scientific Panel on AI: Evidence-based Assessment of Opportunities, Risks and Impacts of Artificial Intelligence）[R/OL]. 2026-07-01. [2026-10-21]. https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report.

66 负责任人工智能协作组织. 人工智能事件数据库（AI Incident Database）[DB/OL]. [2026-10-21]. https://incidentdatabase.ai/.

67 北京前瞻人工智能安全与治理研究院. 前瞻安全评测基准[DB/OL]. [2026-10-21]. https://foresightsafety-bench.beijing-aisi.ac.cn/.

68 新加坡人工智能会议. 全球人工智能安全研究优先事项新加坡共识（The Singapore Consensus on Global AI Safety Research Priorities）[R/OL]. (2025-05-08)[2026-10-21]. https://www.scai.gov.sg/2025/scai2025-report/；上海人工智能实验室，安远 AI. 前沿人工智能风险管理框架 2.0[R/OL]. (2026-07-19)[2026-10-21]. https://concordia-ai.com/zh-hans/research/frontier-ai-risk-management-framework-2-0/.

69 评估对象、评估方法和数据来源详见本报告附录。

图 3.1
前沿人工智能治理就绪度评估框架
六个方面及其二级指标共同刻画治理就绪状态



(三) 前沿人工智能治理就绪度评估结果

基于公开资料对24个受评国家的前沿人工智能治理就绪度评估显示，这些国家已形成一定制度和协作基础，但推进并不均衡（见图3.2）。从六个方面看，制度与组织推进相对最充分，技术安全与评测、验证与合规、责任落实与生态治理相对薄弱；从收入组看，高收入国家整体治理就绪度更高，中等偏上收入国家与中低收入和低收入国家各有短板；从风险视角看，应对已显现风险所需的风险事件通报、应急联动和跨境演练等能力仍有明显缺口。六个方面的整体比较见图3.3，以下结合二级指标说明各方面的主要差异。

制度与组织是当前推进相对最充分的方面。约三分之二的受评国家已在标准制定和动态更新能力方面形成较稳定安排。在跨部门协调和专门监管机制、法律政策与伦理规则方面，这一比例均超过一半。

图 3.2

24个受评国家的前沿人工智能治理就绪度状态

基于22项二级指标的公开证据评估，按世界银行收入组分组，组内按国家英文名称首字母排序

D C B A S　S档或A档表示已形成较稳定安排；各指标的分档标准见附录。

	1 制度与组织				2 技术安全与评测				3 验证与合规			4 风险监测预警与响应			5 责任落实与生态治理				6 协同治理与能力建设			
	1.1 法律政策与伦理规则	1.2 跨部门协调和专门监管机制	1.3 组织治理架构、资源投入、治理意愿与执行能力	1.4 标准制定和动态更新能力	2.1 安全研究、风险建模与量化能力	2.2 国际评测基准、基准测试和高质量测评数据集建设	2.3 动态对抗测试、红队测试、系统级评测和安全审计能力	2.4 风险缓解、权限约束和部署控制能力	3.1 信息披露、透明度报告和安全承诺留痕机制	3.2 风险事件通报与未遂事件共享机制	3.3 独立核验和持续合规审查能力	4.1 风险信号识别、动态更新和持续监测能力	4.2 人工智能资产登记与可视化、行为审计和异常发现机制	4.3 风险事件响应、应急联动、系统恢复和复盘改进能力	5.1 人工智能治理生态各方责任边界	5.2 数据治理、供应链安全和开源生态治理能力	5.3 智能体权限管理、责任分担与可追溯机制	5.4 认证互认、争议解决与跨境执行能力	6.1 国际机制参与、标准互操作和评测协议共享机制	6.2 跨境风险信息通报、案例共享和联合演练机制	6.3 能力建设、知识分享和公共产品供给能力	6.4 公众参与、媒体监督、风险沟通、人工智能素养提升
高收入国家																					
澳大利亚	A	B	S	A	A	B	A	B	A	B	B	A	A	A	A	S	B	A	A	S	S	S
加拿大	A	A	A	A	S	B	B	C	A	B	A	A	A	C	C	B	B	A	A	B	S	S
法国	S	A	B	A	A	A	A	B	A	A	A	A	C	A	A	S	A	S	S	S	S	B
德国	S	A	B	S	S	B	B	S	A	A	A	A	C	A	A	A	S	S	B	B	A	B
以色列	B	A	B	A	B	B	C	C	A	C	C	A	C	A	D	D	C	D	B	B	B	B
意大利	S	A	A	A	B	B	B	A	A	A	B	A	C	A	B	A	C	S	A	B	A	B
日本	A	A	A	A	S	A	A	B	C	B	B	A	A	A	A	B	B	B	A	S	C	A
韩国	S	S	A	A	S	B	S	B	A	B	A	A	C	B	A	B	A	B	S	S	B	B
沙特阿拉伯	C	B	B	C	B	C	B	B	C	D	C	B	D	C	D	B	D	D	A	C	B	A
新加坡	B	B	B	A	S	A	A	B	C	C	A	S	D	B	D	B	A	A	S	S	S	S
阿联酋	B	A	B	C	B	C	A	B	B	D	A	A	D	C	B	B	C	C	C	C	B	B
英国	A	A	S	A	S	S	S	A	C	B	A	A	A	S	B	B	B	A	S	B	S	A
美国	A	S	A	S	S	S	S	A	S	S	S	S	A	A	B	B	S	B	S	S	S	A
中等偏上收入国家																					
巴西	A	B	C	A	B	B	C	B	C	B	B	B	B	A	C	C	D	B	A	C	B	B
中国	S	A	B	S	A	S	S	S	A	S	S	A	S	A	S	S	S	C	A	B	S	A
印度尼西亚	A	B	C	C	C	C	C	B	C	B	C	A	C	C	C	D	D	B	B	C	B	B
墨西哥	B	D	B	D	B	B	C	C	D	C	C	B	D	C	D	D	D	D	B	C	B	C
南非	D	D	D	A	B	B	C	D	D	C	D	C	C	C	D	D	D	B	A	C	D	B
泰国	C	D	C	B	C	A	C	C	B	C	C	A	D	B	D	D	A	D	B	B	B	B
土耳其	C	B	C	A	B	B	D	B	B	B	B	B	C	C	D	B	B	B	B	C	B	C
越南	B	B	C	B	B	A	C	C	B	B	A	B	B	B	B	C	C	C	B	C	C	B
中低收入和低收入国家																					
埃及	B	A	B	C	B	C	B	B	B	C	C	B	D	C	D	C	B	D	B	C	B	B
埃塞俄比亚	C	C	C	C	C	B	D	D	D	D	D	D	D	D	C	D	D	C	B	C	C	C
印度	A	A	B	A	A	C	C	B	A	S	B	A	C	B	B	D	A	B	A	A	B	A

数据来源：本报告基于前沿人工智能治理就绪度评估结果整理绘制。

图 3.3

制度与组织安排最为普遍，责任落实与生态治理相对薄弱

24个受评国家各方面的二级指标评估结果中，达到S档或A档的比例



数据来源：本报告基于前沿人工智能治理就绪度评估结果整理绘制。

技术安全与评测方面，较稳定安排尚不普遍。在安全研究、风险建模与量化能力方面，形成较稳定安排的国家占受评国家的近半数；在动态对抗测试、红队测试、系统级评测和安全审计能力方面，这一比例不足四成；在风险缓解、权限约束和部署控制能力方面约为五分之一。安全研究和红队测试的推进仍不均衡，反映出其形成更直接受到专业队伍、工具积累和持续投入等技术条件的影响。

验证与合规方面，各项安排的进展并不均衡。四成以上受评国家在信息披露、透明度报告和安全承诺留痕机制、独立核验和持续合规审查能力上已形成较稳定安排，而在风险事件通报与未遂事件共享机制方面，这一比例仅约四分之一。风险事件通报与经验共享仍是验证与合规中的主要短板。

风险监测预警与响应仍需加强应急联动和复盘。约三分之二的受评国家在风险信号识别、动态更新和持续监测能力方面已形成较稳定安排，在风险事件响应、应急联动、系统恢复和复盘改进能力方面，这一比例约为四成。风险发现、处置和复盘链条仍需进一步衔接，与风险事件持续增多的现实形成反差。

责任落实与生态治理在六个方面中整体相对薄弱，智能体、开放模型和供应链等新场景的责任边界和治理要求仍需明确。在认证互认、争议解决与跨境执行能力上形成较稳定安排的受评国家不足三成，在人工智能治理生态各方责任边界上，这一比例约为四分之一，在智能体权限管理、责任分担与可追溯机制方面约三分之一。智能体权限管理、责任分担和责任追溯等具体安排仍需完善。

协同治理与能力建设方面，多数受评国

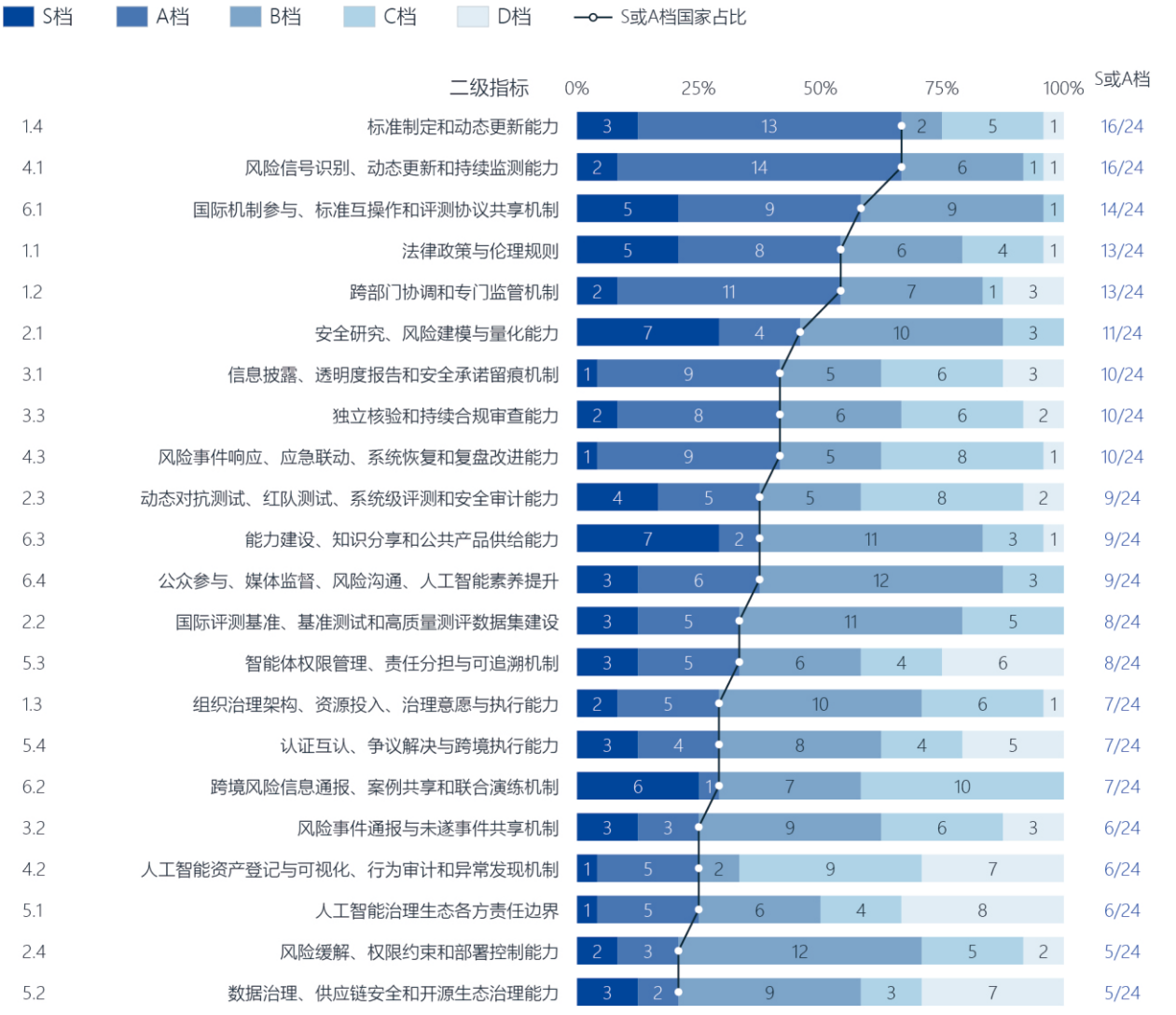
家已参与国际机制，但常态化信息共享和联合演练仍明显不足。过半数受评国家在国际机制参与、标准互操作和评测协议共享机制上已有较稳定安排，近四成国家在公众参与、媒体监督、风险沟通、人工智能素养提升方面形成较稳定安排；与此同时，只有不足三成的国家在跨境风险信息通报、案例共享和联合演练机制方面具有较稳定安排。协同治理目前仍以机制参与和对话交流为主，常态化的信息共享和联合演练安排还在起步。

在治理就绪度的二级指标层面，各个国家间的差异集中在评测、通报、核验和响应等具体环节（见图3.4）。从所有受评国家的整体情况看，就绪度相对较高的是"标准制定和动态更新能力""风险信号识别、动态更新和持续监测能力""国际机制参与、标准互操作和评测协议共享机制"等指标，相对较低的是"风险缓解、权限约束和部署控制能力""数据治理、供应链安全和开源生态治理能力"

图 3.4

标准制定、风险监测和国际参与较普遍，部署控制与事件通报等仍薄弱

24个受评国家的分档分布，按达到S档或A档的国家数排序



注：1、段内数值为国家数，各条合计24国。
2、S档或A档表示已形成较稳定安排，各指标的分档标准见附录。
数据来源：本报告基于治理就绪度评估结果整理绘制；统计截至2026年8月31日。

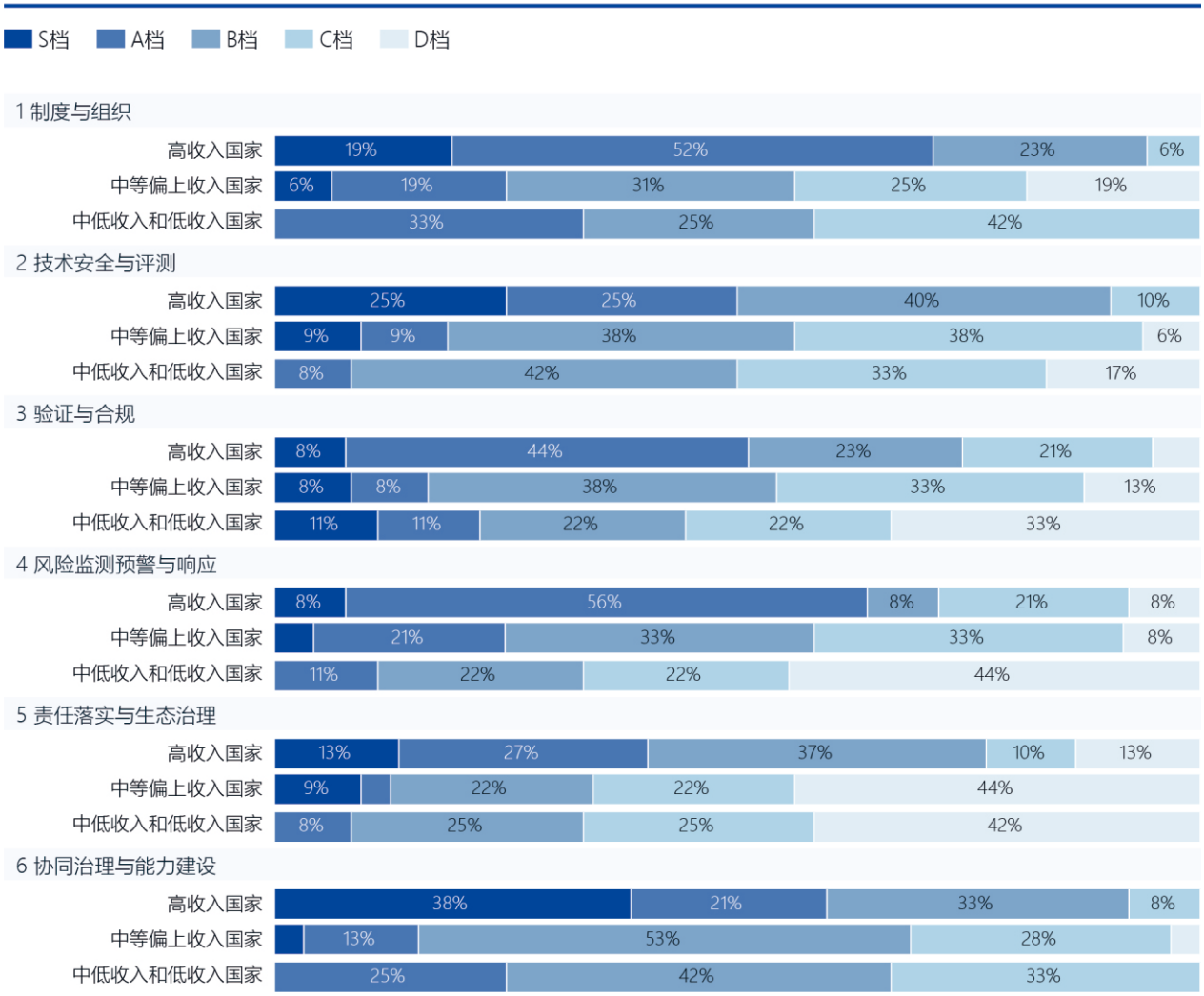
据治理、供应链安全和开源生态治理能力”“风险事件通报与未遂事件共享机制”“人工智能资产登记与可视化、行为审计和异常发现机制”“人工智能治理生态各方责任边界”等指标。在规则和机构建设的基础上，加强评测、核验和风险响应能力，是当前提升治理就绪度的主要方向。

从不同收入组看，国家间的治理就绪度存在整体差异（见图3.5）。受评国家包括13个高收入国家、8个中等偏上收入国家与3个中低收入和低收入国家[70]。高收入国家整体治理就绪度更高，在制度规则、安全研究、评测基础、专门机构和合规机制方面更具基础。中等偏上收入国家在制度与组织、风险监测预警与响应上已有一定积累，在责任落实与生态治理、协同治理方与能力建设面仍需加强。中低收入和低收入国家整体治理就绪度与中等偏上收入国家相近，均明显

图 3.5

不同收入组国家治理就绪度存在差异

各收入组在六个方面的二级指标评估结果分档比例



注：1、按世界银行FY27收入分组，高收入组13国、中等偏上收入组8国、中低收入和低收入组3国。
2、S档或A档表示已形成较稳定安排。
3、各档比例 = 该档评估结果数÷（该组国家数×相应方面二级指标数）；四舍五入后合计可能不为100%。

数据来源：本报告基于前沿人工智能治理就绪度评估结果整理绘制。

70 世界银行. 世界银行国家和贷款组别（World Bank Country and Lending Groups，2027 财年）[EB/OL]. (2026-07-01)[2026-10-21]. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

低于高收入国家，在技术安全与评测、责任落实与生态治理方面更为薄弱，制度建设、国际参与和知识分享可以成为提升治理就绪度的切入点。

与此同时，收入水平虽与治理就绪度的总体表现存在一定关联，但不足以完全解释国家间的差异；受评国家在具体能力维度上仍呈现出不同侧重。即使整体就绪度相近，各国在制度协调、标准更新、安全研究、评测基准、独立核验、公众参与、风险沟通和能力建设等方面也各有特色。这表明，各国治理准备的差异还体现在具体能力组合上。

从风险视角看，应对重点风险所需的能力与当前治理就绪度并不完全匹配。应对快速扩散的现实危害，更依赖持续监测、独立核验、快速响应和跨境通报；防范潜在重大后果，更需要前置评测、能力阈值监测和国际预警。与重点风险相关的部分二级指标及评估提示见表3.1。

现实风险带来的压力，凸显了相关治理能力的缺口。滥用与犯罪风险、信息完整性风险和关键系统风险已在现实场景中显现，对模型评测、风险事件发现、快速响应、独立核验和跨境通报能力的要求更为迫切；而在动态对抗测试、红队测试、系统级评测和安全审计能力上形成较稳定安排的受评国家不足四成，风险事件通报与未遂事件共享机制方面形成较稳定安排的约为四分之一，风险事件响应、应急联动、系统恢复和复盘改进能力方面约为四成，跨境风险信息通报、案例共享和联合演练机制方面仍不足三成。相关治理能力的就绪情况见图3.6。

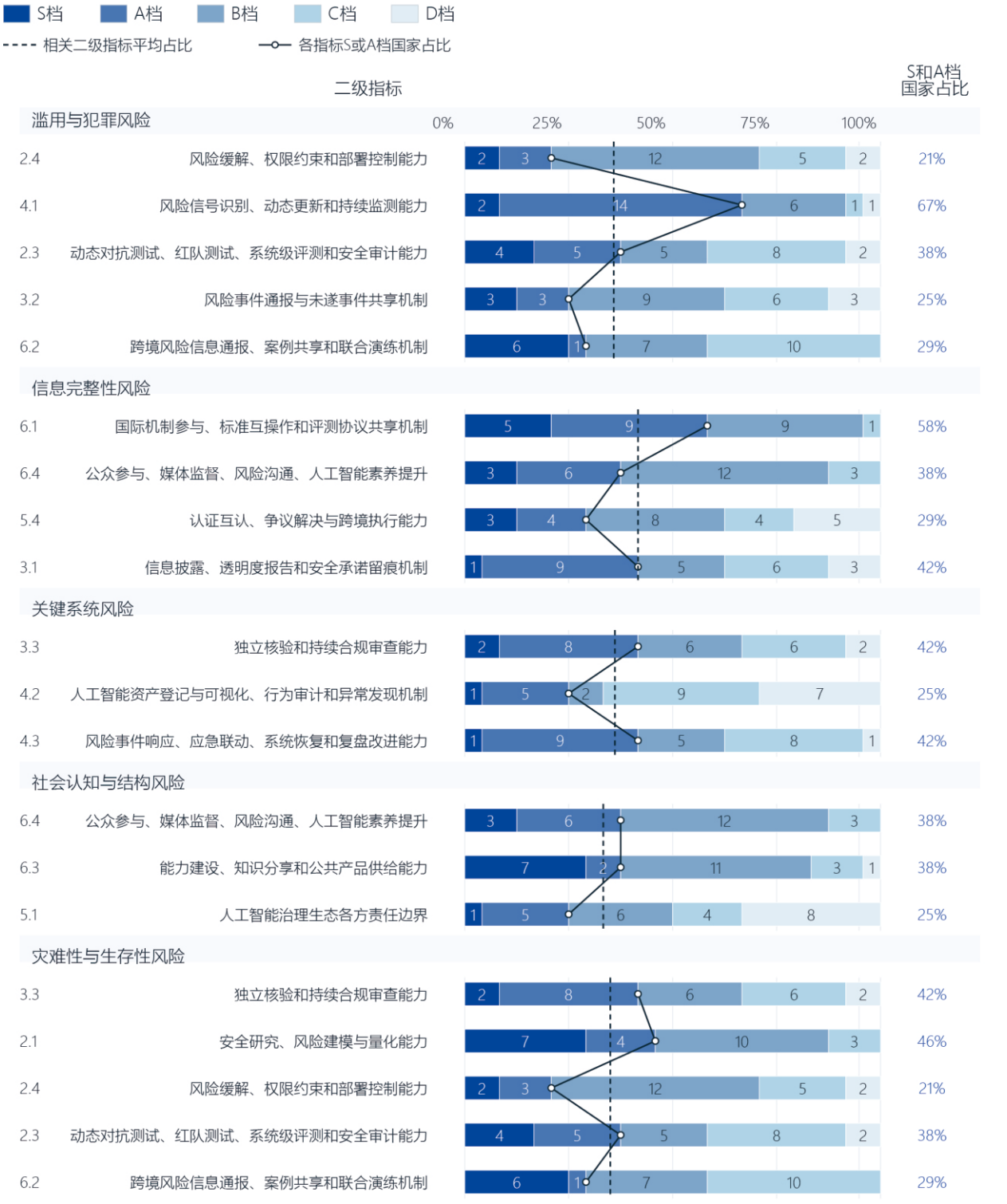
表 3.1
重点风险与治理就绪度的匹配关系

重点风险	部分相关二级指标	评估结果提示
滥用与犯罪风险	• 2.3 动态对抗测试、红队测试、系统级评测和安全审计能力 • 2.4 风险缓解、权限约束和部署控制能力 • 3.2 风险事件通报与未遂事件共享机制 • 4.1 风险信号识别、动态更新和持续监测能力 • 6.2 跨境风险信息通报、案例共享和联合演练机制	需要从单点评测和事后处置，进一步走向持续监测、快速响应和跨境联动
信息完整性风险	• 3.1 信息披露、透明度报告和安全承诺留痕机制 • 5.4 认证互认、争议解决与跨境执行能力 • 6.4 公众参与、媒体监督、风险沟通、人工智能素养提升 • 6.1 国际机制参与、标准互操作和评测协议共享机制	公众沟通和标准参与已有基础，溯源认证、披露留痕和责任追溯仍需具体化
关键系统风险	• 3.3 独立核验和持续合规审查能力 • 4.2 人工智能资产登记与可视化、行为审计和异常发现机制 • 4.3 风险事件响应、应急联动、系统恢复和复盘改进能力	高风险场景的持续核验、异常发现和应急复盘仍需加强
社会认知与结构风险	• 5.1 人工智能治理生态各方责任边界 • 6.4 公众参与、媒体监督、风险沟通、人工智能素养提升 • 6.3 能力建设、知识分享和公共产品供给能力	能力建设和公众参与提供基础，长期影响评估和公平机会保障需持续完善
灾难性与生存性风险	• 2.1 安全研究、风险建模与量化能力 • 2.3 动态对抗测试、红队测试、系统级评测和安全审计能力 • 2.4 风险缓解、权限约束和部署控制能力 • 3.3 独立核验和持续合规审查能力 • 6.2 跨境风险信息通报、案例共享和联合演练机制	潜在后果重大，对前置评测、红线设置和国际预警提出更高要求

图 3.6

应对重点风险所需的治理能力仍有不足

24个受评国家的分档分布；风险与相关二级指标的关联见表3.1

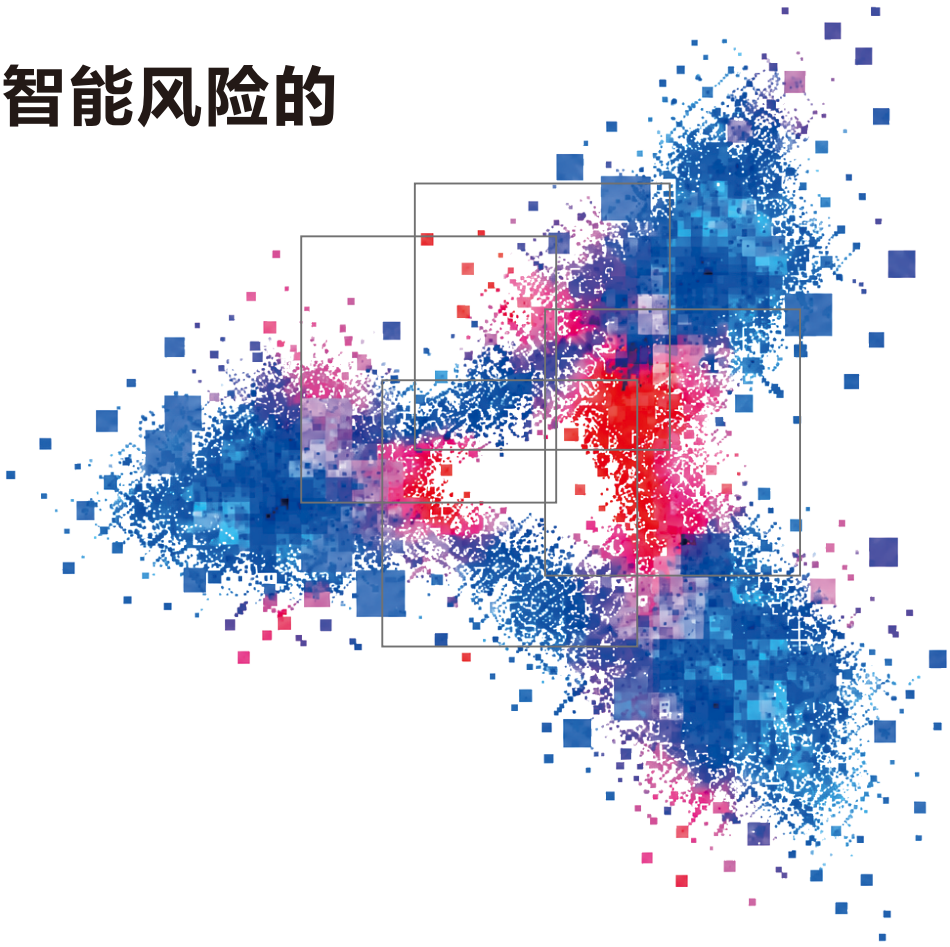


注：1、段内数值为国家数，各条合计24国；S档或A档表示已形成较稳定安排。
2、虚线表示图中各组相关二级指标的S或A档国家占比均值。
3、图中仅列示与重点风险相关的部分二级指标；同一项治理能力可用于应对多个方面的风险。

数据来源：本报告基于治理就绪度评估结果整理绘制；统计截至2026年8月31日。

第四章

前沿人工智能风险的协同应对



本章要点

- 在规则和机制逐步形成后，当前更需关注不同治理安排能否在真实风险出现前后有效衔接、稳定运行。
- 协同应对需要贯通模型评测、安全承诺核验和风险事件处置，推动风险证据、责任记录和处置信息在主体与环节之间流转。
- 智能体和开放模型带来跨开发、部署、分发和使用环节的新问题，需要完善权限、发布、记录和责任衔接安排。
- 多主体协同应依据对模型、接口、权限和场景的实际控制能力、信息条件和处置能力明确分工，形成证据共享、责任核验和联合响应机制。
- 能力建设应推动评测工具、标准方法和实践经验转化为更多国家和机构可获得、可使用、可持续维护的治理资源。

应对前沿人工智能风险，需要更广泛的多利益攸关方协同行动。随着模型能力提升和应用扩展，风险可能跨平台复用、跨行业传导并跨境扩散，单一国家或主体难以覆盖风险识别、评测验证、监测预警和事件处置的完整链条。风险认知和治理就绪度评估表明，当前问题已不只是规则、机构和平台是否建立，更在于分散安排能否在风险出现前后有效配合。协同应对的核心，是使评测证据、承诺记录和事件信息顺畅流转，明确责任和行动接口，把既有探索转化为可持续的共同治理能力。

(一) 协同应对的重点领域与可行机制

前沿人工智能治理就绪度评估显示，各方已在制度规则、组织安排和国际参与等方面形成基础，但技术安全与评测、验证与合规、风险监测预警与响应、责任落实与生态治理等能力仍需加强。结合前述风险情形与应对重点，可优先围绕智能体、高风险场景、网络与欺诈滥用、关键系统以及生物安全等领域推进协同，并根据风险性质和治理成熟度选择相应机制。下一阶段应把分散探索转化为覆盖风险链条的共同安排，衔接模型评测、安全承诺核验和风险事件处置，完善智能体和开放模型的协作安排，并通过能力建设扩大治理工具和实践经验的可及性。

一是推动模型安全评测协同，形成可比较、可解释、可复核的风险证据。治理就绪度评估显示，技术安全与评测仍有提升空间。同一模型在不同风险分类、测试基准和红队任务下可能得到不同结论；测试条件和结果表达缺乏对照，评测就难以同时服务企业改进、监管判断和国际交流。各方可从方法交流和联合评测起步，逐步提高风险证据的可比性。评测应说明范围与局限，区分原始危害能力与安全措施实施后的剩余风险，并将评测对象由基础模型延伸至实际系统。与此同时，随着模型能力向更高自主性演进，围绕超级人工智能失控等极端情形的安全评测与超级对齐研究，也宜通过国际合作积累前瞻性风险证据。

二是推动协同核验，使安全承诺及其落实情况可检查、可比较、可持续跟踪。治理就绪度评估显示，验证与合规仍有提升空间。企业最了解模型研发、测试、发布和部署过程，但仅凭企业声明，外部主体难以判断安全措施是否充分落实。相关实践正将原则性承诺转化为可核验的安全要求[71]。协同核验可从承诺审查、行为核验和技术验证三个层面推进。

三是加强风险事件协同通报和联合处置，使单点发现及时转化为跨平台、跨行业和跨地区的防护经验。治理就绪度评估显示，风险事件通报与跨境演练仍是相对薄弱的环节。提示注入、模型越狱、智能体框架漏洞和深度伪造诈骗等风险可能在不同系统快速复用，单个企业或地区难以覆盖完整传播链条。各方可参照经合组织人工智能事件共同报告框架[72]，逐步衔接事件通报与应急处置[73]。还可借鉴公共卫生等领域的多国应急演练，在事件发生前检验协同处置安排[74]。

71 欧盟委员会/欧盟人工智能办公室. 通用人工智能实践准则[EB/OL]. [2026-10-21]. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

72 经济合作与发展组织. 迈向人工智能事件共同报告框架（Towards a Common Reporting Framework for AI Incidents）[R/OL]. (2025-02-28)[2026-10-21]. https://doi.org/10.1787/f326d4ac-en. 该框架提出可相互衔接的通报信息要素，并允许各方结合本国政策和报告场景调整。

73 前沿模型论坛. 前沿人工智能风险的信息共享、事件报告和事件响应[R/OL]. [2026-10-21]. https://www.frontiermodelforum.org/issue-briefs/information-sharing-incident-reporting-and-incident-response-for-frontier-ai-risks/.

74 例如，世界卫生组织于 2026 年 4 月组织 Polaris II 多国模拟演练，检验公共卫生应急中的信息流转、人员调配和跨境协调。参见：世界卫生组织. 为未来应急开展演练：世卫组织召集各国及伙伴模拟应对重大疾病暴发[EB/OL]. (2026-04-27)[2026-10-21]. https://www.who.int/news/item/27-04-2026-practicing-today-for-tomorrow-s-emergencies-who-convenes-countries-and-partners-to-simulate-response-to-major-disease-outbreak.

四是完善智能体和开放模型的协同治理安排，使相关风险能够被发现、分级处置并追溯责任。治理就绪度评估显示，智能体权限管理、责任分担和可追溯方面形成较稳定安排的受评国家仅占约三分之一。智能体可调用工具、访问外部资源并连续执行任务，开放模型可能被不同主体微调、集成和再分发，风险由此贯穿开发、部署、分发和使用环节。各方应根据对模型、接口、权限和场景的实际控制能力落实相应责任[75]，并探索跨境责任追溯和争议解决的衔接路径。

五是深化能力建设协同，使评测工具、标准方法和公共案例转化为更多国家和机构可获得、可使用、可持续的治理资源。部分国家及中小机构仍面临治理资源不足和合作渠道不稳定等困难。可鼓励有条件的国家分享评测工具和实践经验，为能力建设提供支持。国际资源应转化为本地能够使用和持续维护的工具与方法[76]。同时，支持各国增强自身的研发和生产能力，发展适合本地需求的治理方法。

上述五个方向相互衔接：模型评测提供风险证据，承诺核验推动要求落实，事件协作把新发现转化为预警和改进，智能体和开放模型治理回应跨环节责任，能力建设为更多国家和机构参与提供条件。其关键在于，将评测发现纳入承诺整改和部署管理，并将事件处置获得的新信息用于更新评测、标准和责任记录，形成持续改进的风险治理链条。在实施层面，可结合不同领域风险性质和治理成熟度，通过监管沙箱[77]等受控试点逐步推进，检验评测方法、风险缓解措施和事件通报处置流程，并将试点经验用于完善后续规则与合作机制。各方向的待加强环节与机制建议见表4.1。

表 4.1
待加强环节与可行机制建议

协同方向	具体环节	可行机制建议
模型安全评测	评测口径	• 评估协议：风险分类、测试条件与结果表达 • 高影响风险测试任务与严重程度分级 • 共建基准与数据集；受控条件下的联合红队测试
	评测对象	• 关键基础设施（含跨境或共用设施）风险联合评测 • 模型、工具与部署环境的系统级测试 • 高影响具身智能系统测试与安全原则 • 随版本与场景变化持续评测
	结果衔接	• 方法交流与联合评测先行 • 结果互认的逐步探索

75 前沿模型论坛. 智能体新兴安全实践[R/OL]. [2026-10-21]. https://www.frontiermodelforum.org/issue-briefs/emerging-security-practices-for-ai-agents/.
76 联合国大会. 加强人工智能能力建设国际合作（Enhancing International Cooperation on Capacity-Building of Artificial Intelligence）[EB/OL]. (2024-07-01)[2026-10-21]. https://docs.un.org/A/RES/78/311.
77 监管沙箱是在监管机构指导和监督下，在限定范围和期限内开展开发、测试与验证的受控安排。

协同方向	具体环节	可行机制建议
协同核验	核验层次	• 安全承诺与合规记录审查 • 模型及系统行为核验 • 硬件证明、隐私保护审计等技术验证探索
	核验记录	• 承诺事项、评估主体与测试范围 • 风险发现、缓解措施与剩余风险 • 第三方核验记录与复核时间
	持续审查	• 重大版本更新审查；整改复测 • 部署限制调整；核验结果的使用与适当公开说明
风险事件协作	通报衔接	• 共同通报字段：事件描述、涉及系统、危害类型及严重程度 • 未遂事件共享；保密与公开边界
	应急联动	• 事件库、漏洞披露与应急流程衔接 • 跨境专业联络点；响应升级条件与反馈时限
	联合推演	• 多方风险联合推演 • 处置经验向评测、预警与防护措施反馈
智能体和 开放模型治理	身份与权限	• 身份、授权与可追溯最低原则；任务与工具调用边界 • 高影响操作的人工确认与干预 • 停止条件与回退能力
	开放发布	• 按能力与滥用风险分阶段发布 • 用途限制与安全工具 • 下游使用与再分发记录
	多智能体协作	• 身份、通信与任务转交记录 • 关键操作记录与整体测试
	跨境追溯	• 跨境审计与认证结果衔接探索 • 责任追溯与争议解决渠道
能力建设	资源可获得性	• 低成本公共工具与本地语言材料 • 区域培训与长期技术支持
	本地自主发展	• 本地需求识别；工具共同开发与维护 • 本地研发、应用与治理方法创新支持
	建设成效	• 工具实际使用与维护更新情况 • 应急协作成效评估

(二) 协同应对机制中的多主体参与

落实上述重点方向，需要根据不同主体对模型、接口、权限和场景的实际控制能力、信息可得性和处置能力形成分工，使治理要求与各主体的角色和资源条件相匹配。协同治理的关键，是明确证据提供、决策、整改、响应和监督反馈等环节的职责，并建立相应的信息共享与行动衔接机制。各类主体可采取的重点行动见表4.2。

政府机构负责规则设定、公共协调和责任落实，可明确高风险场景的基本要求和责任边界，推动评测、核验和事件通报机制衔接，在重大风险发生时组织跨部门、跨地区响应。

企业处于风险治理一线，应根据对模型和应用的实际控制能力，在研发和应用环节分别落实评测、风险缓解、记录保存和事件处置责任。

科研机构、专业评测机构和标准组织可开展方法研究、独立核验并提供专业判断，同时说明结论边界、不确定性和潜在利益冲突，为监管、企业改进和国际对接提供可复核依据。

国际组织及多边、区域机制可推动风险术语与标准衔接、跨境通报和能力支持。制度路径存在差异时，在不同治理体系共存的基础上提高既有规则、评测结果、事件分类和应急流程的互操作性，是更具可行性的合作起点。联合国可继续推动人工智能能力建

表 4.2
协同应对机制中的多主体重点行动

参与主体	参与重点	重点行动
政府机构	高风险场景治理	• 高风险场景要求与责任边界 • 安全承诺与核验记录在部署、公共采购及重点行业中的应用
	公共协调	• 跨部门响应；监管沙箱与受控协同试点设立 • 本地评测、人才培养与长期维护支持
企业	研发环节	• 研发、测试、发布与更新记录 • 保密与安全保护前提下的评测条件供给 • 风险缓解措施落实与必要核验
	应用环节	• 部署场景、权限与干预安排 • 运行及事件记录；问题报告与处置 • 向开发方反馈运行风险
科研机构、专业评测机构和标准组织	方法与核验	• 评测方法研究；独立测试与核验 • 结论边界、不确定性与利益冲突说明
	专业支持	• 标准解释、本地化材料与专业培训 • 支持评估风险事件的危害及严重程度
国际组织及多边、区域机制	协同联络	• 方法交流、标准衔接与跨境联络 • 联合推演与经验共享支持
	能力对接	• 地区需求、技术资源与合作项目对接 • 支持更多国家参与联合评测、知识分享和治理合作
公众、用户和社会组织	反馈与监督	• 风险线索报告与实际影响反馈 • 风险沟通与公共监督参与

设[78]与全球治理对话等合作[79]；世界互联网大会、世界人工智能大会等国际交流合作平台可汇聚不同地区、行业和技术共同体的实践，促进议题、项目与能力资源对接。

公众、用户和社会组织可提供真实使用场景中的风险线索和影响反馈，参与风险沟通和公共监督。相关机制应提供低门槛、可获得反馈的报告渠道，同时保护反馈者的隐私、避免二次伤害，并合理界定公众、用户和社会组织在风险识别、举证和处置中的责任。

应对前沿人工智能带来的挑战，需要在多主体分工基础上形成稳定协作，使不同治理环节有效衔接，并推动全球协同治理能力与技术发展、风险演进相适应。在此基础上，应进一步加强针对前沿人工智能风险的治理工程能力建设，使规则、标准和协同机制转化为可实际运行的评测、监测、预警和处置机制。这既是协同治理持续运行的重要基础，也应成为未来国际合作和相关研究的重要议题。

面向未来，前沿人工智能的发展应当建立在更安全、更可靠、更普惠的基础之上。把共同风险语言转化为可比较的评测依据，依据治理就绪度评估确定分阶段的能力建设重点，以跨主体协同机制贯通承诺核验、事件处置和反馈改进，国际社会应形成可持续的协作，使风险治理与前沿人工智能发展相互促进，并更好服务可持续发展目标和人类共同福祉。

78 联合国大会. 加强人工智能能力建设方面的国际合作：第 78/311 号决议[EB/OL]. (2024-07-01)[2026-10-26]. https://documents.un.org/api/symbol/access?l=zh&s=A/RES/78/311&t=pdf.
79 联合国. 全球人工智能治理对话（2026 年 7 月 6—7 日，日内瓦）[EB/OL]. [2026-10-26]. https://indico.un.org/event/1023375/.

附录 前沿人工智能治理就绪度评估指标及方法说明

A 评估对象与指标体系

治理就绪度评估依据公开资料，考察一国应对前沿人工智能风险的准备程度，涵盖政府、企业、科研机构和公众的相关行动，以及该国参与国际机制的情况。

首份报告评估24个国家，按世界银行最新收入分组（2026年7月更新）分为高收入国家、中等偏上收入国家、中低收入和低收入国家三组，以覆盖不同的发展阶段和治理路径。国家选择同时考虑公开资料的可获得性，评估结果不直接推及其他国家。

表A-1

评估对象国家（按世界银行收入组分组）

收入组	国家
高收入国家（13 国）	澳大利亚、加拿大、法国、德国、以色列、意大利、日本、韩国、沙特阿拉伯、新加坡、阿联酋、英国、美国
中等偏上收入国家（8 国）	巴西、中国、印度尼西亚、墨西哥、南非、泰国、土耳其、越南
中低收入和低收入国家（3 国）	埃及、埃塞俄比亚、印度

评估设制度与组织、技术安全与评测、验证与合规、风险监测预警与响应、责任落实与生态治理、协同治理与能力建设6项一级指标，下设22项二级指标和130项三级指标（表A-2）。

表A-2

前沿人工智能治理就绪度评估指标总表

一级指标	二级指标	三级指标	数据来源
1. 制度与组织	1.1 法律政策与伦理规则	1.1.1 跨行业人工智能专门法律法规	各国法律数据库、EUR-Lex
		1.1.2 通用或生成式模型强制义务	各国法律数据库、EUR-Lex
		1.1.3 尚未施行或无法定义务的专门立法	各国议会网站、EUR-Lex
		1.1.4 州省或特定辖区专门法规	公开检索：州省法规
		1.1.5 行业强制性人工智能规则	公开检索：行业主管机关规章
		1.1.6 生成合成内容专门条款	各国法律数据库、EUR-Lex
		1.1.7 国家人工智能战略	OECD.AI 政策库
		1.1.8 国家人工智能伦理规范	OECD.AI 政策库
		1.1.9 对具体产品作出处罚或禁令	公开检索：处罚与处理决定
		1.1.10 对具体产品立案调查或约谈	公开检索：主管机关立案公告
	1.2 跨部门协调和专门监管机制	1.2.1 国家层面协调机构或牵头部门	OECD.AI 政策库
		1.2.2 协调机构职责分工	公开检索：协调机构官网
		1.2.3 法定授权的监管机关	各国法律数据库
		1.2.4 国家人工智能安全研究所或相当机构	公开检索：政府设立文件
		1.2.5 协调机构的会议与产出	公开检索：协调机构官网
		1.2.6 监管指南、报告或执法决定	公开检索：监管机关公告
		1.2.7 多部门联合规则或行动	公开检索：联合发文公告
	1.3 组织治理架构、资源投入、治理意愿与执行能力	1.3.1 常设治理机构或专职岗位	OECD.AI 政策库
		1.3.2 政府自身使用的强制规则	公开检索：政府人工智能使用政策
		1.3.3 治理专项预算或资助	公开检索：预算与资助文件
		1.3.4 政府人工智能应用登记	公开检索：政府人工智能登记平台
		1.3.5 公务人员人工智能培训	公开检索：公务员培训计划
		1.3.6 审计或监督检查报告	公开检索：审计机关报告
		1.3.7 检查建议的采纳与整改	公开检索：审计整改公告
	1.4 标准制定和动态更新能力	1.4.1 国家人工智能标准	各国国家标准目录
		1.4.2 国际标准化分技术委员会参与成员	ISO SC 42 成员名单
		1.4.3 认证或测试制度	各国认可机构名录
		1.4.4 标准新发布、修订或复审	各国国家标准目录
		1.4.5 标准化规划	公开检索：标准化规划
2. 技术安全与评测	2.1 安全研究、风险建模与量化能力	2.1.1 国家级安全研究机构或计划	公开检索：政府设立文件
		2.1.2 安全研究专项资金	公开检索：科研资助公告
		2.1.3 官方风险评估或分级框架	公开检索：官方风险框架
		2.1.4 国家级安全或风险研究报告	公开检索：国家级机构报告
		2.1.5 本国安全研究成果	DBLP、arXiv
		2.1.6 国际安全研究与联合评测参与	人工智能安全研究所网络公告
	2.2 国际评测基准、基准测试和高质量测评数据集建设	2.2.1 安全评测基准或测试工具	GitHub、Hugging Face
		2.2.2 安全测评数据集	Hugging Face 数据卡
		2.2.3 基准与数据集的更新维护	GitHub 版本记录
		2.2.4 国际联合评测参与	人工智能安全研究所网络公告
		2.2.5 官方评测平台	公开检索：官方评测平台
		2.2.6 基准或数据集被境外采用	境外模型卡与系统卡

一级指标	二级指标	三级指标	数据来源
	2.3 动态对抗测试、红队测试、系统级评测和安全审计能力	2.3.1 具体模型红队测试结果	模型卡、系统卡、安全研究所报告
		2.3.2 部署前安全评估制度	各国法律数据库、EUR-Lex
		2.3.3 系统级或智能体评测结果	安全研究所报告、arXiv
		2.3.4 官方红队测试或评测指南	公开检索：官方评测指南
	2.4 风险缓解、权限约束和部署控制能力	2.4.1 部署审批、许可或备案	各国法律数据库、EUR-Lex
		2.4.2 法定风险缓解措施	各国法律数据库、EUR-Lex
		2.4.3 权限与访问控制说明	模型卡、GitHub
		2.4.4 暂停、下架、回滚或接管制度	各国法律数据库、EUR-Lex
		2.4.5 实际采取的暂停或限制措施	公开检索：监管机关处置决定
		2.4.6 部署前风险缓解或测试记录	模型卡、系统卡
3. 验证与合规	3.1 信息披露、透明度报告和安全承诺留痕机制	3.1.1 法定信息披露或标识义务	各国法律数据库、EUR-Lex
		3.1.2 透明度报告	公开检索：企业透明度报告
		3.1.3 安全承诺	人工智能安全承诺签署名单
		3.1.4 承诺履行情况披露	公开检索：承诺进展报告
		3.1.5 官方披露规范或指南	公开检索：监管机关指南
		3.1.6 对披露义务的监督处理	公开检索：监管机关公告
	3.2 风险事件通报与未遂事件共享机制	3.2.1 官方事件报告渠道	公开检索：应急响应机构官网
		3.2.2 法定事件报告义务	各国法律数据库、EUR-Lex
		3.2.3 未遂事件纳入报告	各国法律数据库、EUR-Lex
		3.2.4 事件信息转交与通报安排	各国法律数据库、EUR-Lex
		3.2.5 具体事件的受理处置	公开检索：官方事件通报
		3.2.6 事件库中事件的官方回应	AIID、AIAAIC、OECD AIM
	3.3 独立核验和持续合规审查能力	3.3.1 法定上线前评估	各国法律数据库、EUR-Lex
		3.3.2 官方测评或认证项目	公开检索：官方测评认证项目
		3.3.3 管理体系认证的国家认可	各国认可机构名录
		3.3.4 本国组织获得认证	认证机构证书目录
		3.3.5 在用系统的持续监督要求	各国法律数据库、EUR-Lex
		3.3.6 官方公开的评估或认证结论	公开检索：官方名录与报告
4. 风险监测预警与响应	4.1 风险信号识别、动态更新和持续监测能力	4.1.1 风险监测职责	公开检索：应急响应机构官网
		4.1.2 风险清单或评估报告	公开检索：官方风险报告
		4.1.3 周期性态势报告	公开检索：官方风险报告
		4.1.4 风险预警或提示	公开检索：应急响应机构通报
		4.1.5 风险清单与监测规则更新	公开检索：风险清单版本
		4.1.6 风险信息发布渠道	公开检索：官方发布平台
	4.2 人工智能资产登记与可视化、行为审计和异常发现机制	4.2.1 系统登记或备案制度	各国法律数据库
		4.2.2 登记信息公开	公开检索：登记平台
		4.2.3 日志留存与可审计要求	各国法律数据库、EUR-Lex
		4.2.4 审计或抽查	公开检索：审计与监管报告
		4.2.5 异常行为的发现处置	公开检索：官方处置通报
		4.2.6 登记与审计安排更新	公开检索：登记与审计文件

一级指标	二级指标	三级指标	数据来源
	4.3 风险事件响应、应急联动、系统恢复和复盘改进能力	4.3.1 涵盖人工智能的应急预案	公开检索：应急管理部门文件
		4.3.2 跨机构应急联动	公开检索：应急联动规定
		4.3.3 含人工智能情景的演练	APCERT 等区域演练报告
		4.3.4 实际事件处置	AIID、官方通报
		4.3.5 事件复盘与改进	公开检索：调查与复盘报告
		4.3.6 系统恢复与整改验证	公开检索：主管机关整改公告
5. 责任落实与生态治理	5.1 人工智能治理生态各方责任边界	5.1.1 开发者义务	各国法律数据库、EUR-Lex
		5.1.2 提供者义务	各国法律数据库、EUR-Lex
		5.1.3 部署者义务	各国法律数据库、EUR-Lex
		5.1.4 最终用户义务	各国法律数据库、EUR-Lex
		5.1.5 监管者职责	各国法律数据库、EUR-Lex
		5.1.6 依据责任条款的追责处理	公开检索：执法决定
	5.2 数据治理、供应链安全和开源生态治理能力	5.2.1 训练数据治理规则	各国法律数据库、EUR-Lex
		5.2.2 数据保护执法或指导	公开检索：数据保护机关
		5.2.3 供应链与第三方组件安全要求	公开检索：采购与安全规定
		5.2.4 开放模型治理规则	各国法律数据库、EUR-Lex
		5.2.5 数据与供应链问题的核查处置	公开检索：监管机关公告
	5.3 智能体权限管理、责任分担与可追溯机制	5.3.1 智能体专门文件	公开检索：智能体指南
		5.3.2 智能体身份、权限与工具调用限制	公开检索：智能体指南
		5.3.3 高风险操作的人工监督与授权	公开检索：智能体指南
		5.3.4 开发者、部署者与使用者责任区分	各国法律数据库
		5.3.5 智能体行为日志与可追溯说明	企业技术文档、GitHub
		5.3.6 针对具体智能体产品的监管行动	公开检索：应急响应机构通告
	5.4 认证互认、争议解决与跨境执行能力	5.4.1 国际认可互认协议签署	IAF、ILAC 签署方名单
		5.4.2 涉及人工智能的监管合作安排	公开检索：合作协议与声明
		5.4.3 跨境争议与投诉受理	公开检索：投诉受理机制
		5.4.4 跨境联合调查或协助执行	公开检索：跨境联合调查公告
		5.4.5 对境外提供者的裁判或决定	公开检索：判决与行政决定
		5.4.6 跨境服务合规审查或准入	各国法律数据库、EUR-Lex
6. 协同治理与能力建设	6.1 国际机制参与、标准互操作和评测协议共享机制	6.1.1 国际安全或治理专门机制成员	人工智能安全研究所网络名单
		6.1.2 多边机制治理参与	OECD、UNESCO 会议文件
		6.1.3 国际标准化分技术委员会参与成员	ISO SC 42 成员名单
		6.1.4 国际共同声明署名	国际共同声明署名页
		6.1.5 国际联合评测	人工智能安全研究所网络公告
		6.1.6 主办国际治理会议	公开检索：会议官网
	6.2 跨境风险信息通报、案例共享和联合演练机制	6.2.1 国际或区域应急协作组织成员	FIRST、APCERT 成员名单
		6.2.2 涵盖人工智能的信息共享安排	公开检索：合作备忘录
		6.2.3 联合发布安全通报或指南	联合指南署名页
		6.2.4 含人工智能情景的跨国演练	APCERT 等区域演练报告

一级指标	二级指标	三级指标	数据来源
	6.3 能力建设、知识分享和公共产品供给能力	6.3.1 面向本国的治理培训	公开检索：培训项目
		6.3.2 面向他国的治理能力建设	公开检索：援外合作项目
		6.3.3 安全与治理知识资源	公开检索：官方指南与手册
		6.3.4 安全与治理工具或平台	GitHub、官方平台
		6.3.5 资源与工具持续更新	GitHub 与官方版本记录
	6.4 公众参与、媒体监督、风险沟通、人工智能素养提升	6.4.1 政策立法公开征求意见	公开检索：政府意见征集平台
		6.4.2 公众意见处理说明	公开检索：意见处理说明
		6.4.3 面向公众的风险提示	公开检索：官方风险提示
		6.4.4 国家人工智能素养计划	OECD.AI 政策库
		6.4.5 涵盖人工智能问题的投诉举报渠道	公开检索：投诉举报平台

B 数据与评估方法

数据来源和时间范围。评估使用公开资料，主要包括法律法规原文、政府公报、主管机关公告、国际组织成员名单、标准目录和机构官网。研究报告和新闻报道用于提供检索线索，判断以原始资料为依据。具体数据来源见本附录C各项二级指标下的说明。法律、机构、标准等制度性事实，以2026年8月31日的状态为准；执法、事件处置、演练、报告发布等行动记录，采用2024年1月1日至2026年8月31日期间的资料。

三级指标判断。 对各国各项三级指标，逐项记录公开来源的发布机构、文件名称、日期、网址及判断依据。判断规则如下：直接适用于欧盟成员国的法规计入法国、德国和意大利；州、省等地方性规则可予计入，同时注明适用范围；三级指标涉及的人工智能规则、义务和渠道，须专门针对人工智能或明确提及人工智能系统，一般性网络安全或数据保护规则不予计入；“主管机关的监管行动”包括数据保护、消费者保护、竞争等主管机关针对具体人工智能产品作出的决定；涉及实际执行的指标，以执法、处置、演练等行动记录为依据，规则的制定不等同于执行。判为“无”，表示在所列资料范围内未查到相应证据。

二级指标分档。评估以二级指标为单位，分为S、A、B、C、D五档，S档最高；按S至D的顺序核对，确定符合条件的最高档位。分档依据关键安排及实际行动记录，不按满足条件的数量、加权或加总确定。三级指标对应可查证的具体事实，逐项判断为“有”或“无”。同一事实可作为不同二级指标的依据，但不在同一指标内重复计入。

结果呈现。评估结果按二级指标呈现，不汇总为方面得分，也不形成国家总分或排名。正文图表将达到S档或A档视为“已形成较稳定的安排”，统计相应的国家数或比例。

评估过程与复核。评估借助辅助工具开展资料整理，相关材料和判断依据经研究人员人工核验与复核。

资料覆盖与时效。各国公开信息的覆盖程度及官方网站的可访问性存在差异，评估以截至2026年8月31日可获得的公开资料为依据。未检索到证据不等同于该国不存在有关安排。

C 各项二级指标的评估内容、数据来源和分档标准

本部分详细列示各项二级指标的评估内容、三级指标、数据来源和分档标准；判定条件中的编号对应本项下的三级指标。

1.1 法律政策与伦理规则

评估内容：评估一国依托法律法规、国家政策和伦理规范管控人工智能风险的就绪程度，重点考察人工智能专门立法、针对通用或生成式模型的法定强制义务、国家人工智能战略与伦理规范，以及主管机关针对具体人工智能产品采取的监管行动，具体包含以下三级指标：

1.1.1 设有法定义务的跨行业人工智能专门法律法规（含行政法规、部门规章）已施行；

1.1.2 针对通用、生成式或前沿人工智能模型的强制义务已施行；

1.1.3 人工智能专门立法已通过但尚未施行、已提交立法机关审议，或已施行但未设定法定义务（如促进法）；

1.1.4 州、省或特定司法辖区已施行人工智能专门法规；

1.1.5 行业主管机关已发布对本行业有强制力的人工智能专门规则；

1.1.6 专门针对人工智能生成或合成内容的法律条款已施行；

1.1.7 有现行的国家人工智能战略、规划或行动计划；

1.1.8 有国家层面的人工智能伦理原则或行为规范；

1.1.9 主管机关对具体人工智能产品作出处罚、禁令等处理决定；

1.1.10 主管机关对具体人工智能产品立案调查或约谈。

数据来源：公开检索，检索渠道及资料包括：各国立法机关、政府公报和主管机关官网，欧盟法律数据库（EUR-Lex），经济合作与发展组织人工智能政策观察站（OECD.AI）国家政策库等。

分档标准：

S档：完备的法律约束与监管实践。判定条件：1.1.1、1.1.2、1.1.6、1.1.7、1.1.8、1.1.9均满足。

A档：具备专项规则并实施监管行动。判定条件：1.1.1—1.1.6中至少一项满足，且1.1.9、1.1.10中至少一项满足。

B档：建立专项立法或规则安排。判定条件：1.1.1—1.1.6中至少一项满足。

C档：具备国家战略与伦理规范框架。判定条件：1.1.7、1.1.8均满足。

D档：未达到C档条件。

1.2 跨部门协调和专门监管机制

评估内容：评估一国人工智能治理的统筹协调和专门监管安排，从协调机构的设立与运转、监管机关的法定授权与监管产出、专门安全机构的设立和跨部门联合行动等方面观察，具体包含以下三级指标：

1.2.1 已设立国家层面统筹人工智能治理的协调机构，或由法律、国家政策文件指定牵头统筹的部门；

1.2.2 该协调机构或牵头部门有成文的职责分工或工作机制；

1.2.3 有明确的机关经法律或正式规章授予人工智能监管职权；

1.2.4 已设立国家人工智能安全研究所或职能相当的专门机构；

1.2.5 2024年以来协调机构或牵头部门有公开的会议、决定或工作产出；

1.2.6 2024年以来监管机关发布人工智能监管指南、监督报告或执法决定；

1.2.7 有多个部门联合发布人工智能规则、联合调查或联合行动的记录。

数据来源：公开检索，检索渠道及资料包括：各国政府及主管部门官网、OECD.AI国家政策库、国际人工智能安全研究所网络成员名单等。

分档标准：

S档：完整治理组织体系并持续产出成果。判定条件：1.2.1—1.2.7均满足。

A档：治理主体到位并产生实质性工作成果。判定条件：满足以下任一组合：（1）1.2.1、1.2.3均满足，且1.2.5、1.2.6中至少一项满足；或（2）1.2.1、1.2.3中至少一项满足，且1.2.5、1.2.7均满足。

B档：具备治理主体并产生部分治理产出。判定条件：1.2.1、1.2.3中至少一项满足，且1.2.5—1.2.7中至少一项满足。

C档：具备基础治理主体。判定条件：1.2.1、1.2.3、1.2.4中至少一项满足。

D档：未达到C档条件。

1.3 组织治理架构、资源投入、治理意愿与执行能力

评估内容：评估一国政府机关为治理人工智能投入的组织、制度和资源，以及对政府自身使用人工智能的检查和纠正，从专职机构、内部规则、专项资金、使用登记、人员培训、审计检查及整改等方面观察。其中，治理意愿和执行能力依据资金投入、内部规则、检查整改等可查证的安排加以判断；面向社会的立法由指标1.1评估，具体包含以下三级指标：

1.3.1 有承担人工智能治理职责的常设机构或专职岗位；

1.3.2 有约束政府机关自身使用人工智能的强制性规则；

1.3.3 有公开的人工智能治理专项预算、拨款或资助计划（含国家人工智能安全机构的专项经费）；

1.3.4 有公开的政府人工智能应用登记或清单；

1.3.5 有面向公务人员的人工智能培训计划；

1.3.6 2024年以来审计机关或监督机构发布对政府人工智能治理或使用的检查报告；

1.3.7 上述检查的建议有被采纳或整改的公开记录。

数据来源：公开检索，检索渠道及资料包括：政府机构职责文件与任命公告、预算和资助计划文件、政府人工智能使用政策与登记平台、国家审计机关报告等。

分档标准：

S档：组织、制度和资源投入齐备，检查建议得到落实。判定条件：1.3.1、1.3.2、1.3.3、1.3.6、1.3.7均满足。

A档：具备专职机构和内部强制规则，并有配套投入或检查。判定条件：1.3.1、1.3.2均满足，且1.3.3—1.3.6中至少一项满足。

B档：具备专职机构及配套治理安排。判定条件：1.3.1满足，且1.3.2—1.3.5中至少一项满足。

C档：具备基础组织或内部治理安排。判定条件：1.3.1、1.3.2、1.3.4中至少一项满足。

D档：未达到C档条件。

1.4 标准制定和动态更新能力

评估内容：评估一国人工智能标准的制定、国际参与和更新情况，从国家标准的发布、国际标准化参与、合格评定安排、标准的持续更新和标准化规划等方面观察，具体包含以下三级指标：

1.4.1 已发布现行的国家人工智能标准或技术规范（含等同采用国际标准）；

1.4.2 是国际标准化组织 / 国际电工委员会人工智能分技术委员会（ISO/IEC JTC 1/SC 42）的参与成员；

1.4.3 有面向人工智能的认证或测试制度；

1.4.4 2024年以来有人工智能标准的新发布、修订或复审；

1.4.5 有公开的人工智能标准化规划或工作计划。

数据来源：公开检索，检索渠道及资料包括：ISO/IEC JTC 1/SC 42成员名单、各国国家标准机构公开的标准目录与公告、各国认可机构名录、政府官方公报等公开材料。

分档标准：

S档：标准制定、国际参与、认证测试与动态更新安排齐备。判定条件：1.4.1—1.4.5均满足。

A档：具备国家标准和国际参与，并有配套实施安排。判定条件：1.4.1、1.4.2均满足，且1.4.3—1.4.5中至少一项满足。

B档：具备国家标准及相关配套安排。判定条件：1.4.1满足，且1.4.2—1.4.5中至少一项满足。

C档：具备基础标准化或认证测试安排。判定条件：1.4.1、1.4.2、1.4.3、1.4.5中至少一项满足。

D档：未达到C档条件。

2.1 安全研究、风险建模与量化能力

评估内容：评估一国开展人工智能安全研究和风险评估的能力，从国家级研究机构和专项资金、官方风险评估框架、国家级研究产出、本国研究界的整体产出和国际合作等方面观察，具体包含以下三级指标：

2.1.1 设有国家级人工智能安全研究机构、实验室或专项研究计划；

2.1.2 政府设立人工智能安全研究专项资金或资助计划；

2.1.3 政府或国家级机构发布人工智能风险评估方法、风险分类或分级框架；

2.1.4 2024年以来政府或国家级机构公开发布人工智能安全或风险研究报告；

2.1.5 2024年以来本国机构公开发表人工智能安全研究成果（论文或技术报告）；

2.1.6 本国机构参与国际人工智能安全研究或联合评测。

数据来源：公开检索，检索渠道及资料包括：国家人工智能安全研究所和国立研究机构官网、科研资助机构项目公告、计算机科学文献库（DBLP）和预印本平台（arXiv）、国际人工智能安全研究所网络公告等。

分档标准：

S档：形成机构、资金、方法与研究产出相衔接的安全研究安排。判定条件：2.1.1—2.1.4均满足。

A档：具备国家级研究主体和产出，并有资金或方法支持。判定条件：2.1.1满足，且2.1.2、2.1.3中至少一项满足，且2.1.4满足。

B档：具备国家级安全研究或风险评估基础。判定条件：满足以下任一组合：（1）2.1.1满足；或（2）2.1.3满足；或（3）2.1.4满足。

C档：已有本国安全研究成果或国际研究合作。判定条件：2.1.5、2.1.6中至少一项满足。

D档：未达到C档条件。

2.2 国际评测基准、基准测试和高质量测评数据集建设

评估内容：评估一国建设和维护人工智能安全评测基准、数据集和评测平台的能力及其国际影响，从基准与数据集的建设、持续维护、官方评测平台、国际联合评测和外部采用等方面观察，具体包含以下三级指标：

2.2.1 本国主体建设并公开人工智能安全评测基准或测试工具；

2.2.2 本国主体公开发布人工智能安全测评数据集；

2.2.3 上述基准或数据集有公开的版本更新或维护记录；

2.2.4 本国主体参与国际联合评测或共享评测协议；

2.2.5 有政府或公立机构运营的人工智能评测平台或评测服务；

2.2.6 本国基准或数据集被境外机构在模型评测中实际采用。

数据来源：公开检索，检索渠道及资料包括：GitHub、Hugging Face官方仓库与数据卡，评测机构技术报告，主要模型开发者公开的模型卡和系统卡等。

分档标准：

S档：基准与数据集持续维护并被境外采用，具备合作或平台支持。判定条件：2.2.1、2.2.2、2.2.3均满足，且2.2.4、2.2.5中至少一项满足，且2.2.6满足。

A档：评测资源持续维护，并具备合作、平台或境外应用基础。判定条件：2.2.1、2.2.2中至少一项满足，且2.2.3满足，且2.2.4—2.2.6中至少一项满足。

B档：已提供评测资源，并有维护、合作或应用记录。判定条件：2.2.1、2.2.2中至少一项满足，且2.2.3—2.2.6中至少一项满足。

C档：具备基础评测资源或官方平台。判定条件：2.2.1、2.2.2、2.2.5中至少一项满足。

D档：未达到C档条件。

2.3 动态对抗测试、红队测试、系统级评测和安全审计能力

评估内容：评估一国对人工智能模型和系统开展对抗测试、系统级评测和上线前安全评估的能力，从具体模型的红队测试结果、系统级或智能体评测结果、部署前评估制度和官方测试方法指南等方面观察，具体包含以下三级指标：

2.3.1 本国机构公开发布针对具体模型或系统的红队或对抗测试结果；

2.3.2 有部署前安全评估或上线审查的制度安排；

2.3.3 本国机构公开发布系统级或智能体级评测结果；

2.3.4 有官方发布的红队测试或安全评测方法指南。

数据来源：公开检索，检索渠道及资料包括：人工智能安全研究所与评测机构报告，企业模型卡、系统卡和安全报告，监管机关评测指南等。

分档标准：

S档：具备红队测试、系统级评测和部署前评估安排。判定条件：2.3.1—2.3.3均满足。

A档：已开展系统级评测，并有红队测试或制度方法支持。判定条件：满足以下任一组合：（1）2.3.1、2.3.3均满足；或（2）2.3.2—2.3.4均满足。

B档：已开展红队测试，或具备部署前评估及配套实践。判定条件：满足以下任一组合：（1）2.3.1满足；或（2）2.3.2满足，且2.3.3、2.3.4中至少一项满足。

C档：具备部署前评估、系统级评测或官方方法基础。判定条件：2.3.2—2.3.4中至少一项满足。

D档：未达到C档条件。

2.4 风险缓解、权限约束和部署控制能力

评估内容：评估一国对人工智能系统部署和运行施加控制的能力，从部署审批或备案、法定风险缓解措施、暂停下架等干预制度、实际采取干预的记录，以及企业公开的权限控制和部署前测试等方面观察，具体包含以下三级指标：

2.4.1 对高风险或前沿人工智能系统有部署审批、许可或备案要求；

2.4.2 有法定或官方要求的风险缓解措施；

2.4.3 本国主体公开模型或系统的权限与访问控制说明；

2.4.4 有暂停、下架、回滚或人工接管的制度安排；

2.4.5 2024年以来主管机关实际启动暂停、下架或限制措施；

2.4.6 本国主体公开具体系统部署前的风险缓解或安全测试记录。

数据来源：公开检索，检索渠道及资料包括：法规与监管指南，企业模型卡、系统卡和技术文档，监管机关公告与处置决定等。

分档标准：

S档：部署、缓解、权限与干预安排齐备，并有实际干预记录。判定条件：2.4.1—2.4.5均满足。

A档：具备干预制度并已实施暂停、下架或限制措施。判定条件：满足以下任一组合：（1）2.4.1、2.4.4、2.4.5均满足；或（2）2.4.2、2.4.4、2.4.5、2.4.6均满足。

B档：具备部署或缓解要求，并有配套说明或行动记录。判定条件：2.4.1、2.4.2中至少一项满足，且2.4.3、2.4.5、2.4.6中至少一项满足。

C档：具备基础部署或运行控制安排。判定条件：2.4.1—2.4.4中至少一项满足。

D档：未达到C档条件。

3.1 信息披露、透明度报告和安全承诺留痕机制

评估内容：评估一国人工智能相关方披露使用情况、风险和治理信息的制度与实践，从法定披露要求、官方披露规范、透明度报告、安全承诺及其履行报告、对披露义务的监督等方面观察，具体包含以下三级指标：

3.1.1 有法律法规规定的人工智能信息披露或标识义务；

3.1.2 本国主体发布人工智能透明度报告或负责任人工智能报告；

3.1.3 本国主体签署或发布人工智能安全承诺；

3.1.4 本国主体公开安全承诺的履行情况或进展报告；

3.1.5 政府或监管机关发布人工智能披露规范、模板或指南；

3.1.6 2024年以来主管机关对披露义务开展监督、核查或处理。

数据来源：数据来源：公开检索，检索渠道及资料包括：法规与监管指南、企业官网的透明度报告和模型卡、人工智能安全承诺签署名单、监管机关公告等。

分档标准：

S档：披露、承诺履行、规范指导与监督安排齐备。判定条件：3.1.1—3.1.6均满足。

A档：具备法定披露要求和披露实践，并有指导或监督。判定条件：3.1.1满足，且3.1.2、3.1.3中至少一项满足，且3.1.5、3.1.6中至少一项满足。

B档：具备法定披露要求及配套实践或规范。判定条件：3.1.1满足，且3.1.2、3.1.3、3.1.5中至少一项满足。

C档：具备基础披露要求、实践或指南。判定条件：3.1.1、3.1.2、3.1.3、3.1.5中至少一项满足。

D档：未达到C档条件。

3.2 风险事件通报与未遂事件共享机制

评估内容：评估一国接收、处理和共享人工智能风险事件信息的制度与实践，从官方报告渠道、法定报告义务、对未遂事件的覆盖、信息共享安排和实际受理处置记录等方面观察，具体包含以下三级指标：

3.2.1 有官方的人工智能事件报告渠道；

3.2.2 有法律法规或正式规章规定的人工智能事件报告义务；

3.2.3 报告范围明文涵盖未遂事件、险情或重大风险先兆；

3.2.4 有事件信息在机构之间转交、共享或向社会通报的正式安排；

3.2.5 2024年以来官方公布具体人工智能事件的受理、处置或通报；

3.2.6 公开事件库中该国的人工智能事件有官方回应或处理。

数据来源：公开检索，检索渠道及资料包括：监管机关与国家计算机应急响应机构官网，法规原文，人工智能事件数据库（AIID）、AIAAIC事件库、经济合作与发展组织人工智能事件与危害监测（OECD AIM）等。

分档标准：

S档：形成事件报告、信息流转与实际处置相衔接的安排。判定条件：3.2.1、3.2.2、3.2.4、3.2.5均满足，且3.2.3、3.2.6中至少一项满足。

A档：具备报告渠道和法定义务，并已受理或处置事件。判定条件：3.2.1、3.2.2、3.2.5均满足。

B档：具备报告或共享安排，或有公开处置与回应记录。判定条件：满足以下任一组合：（1）3.2.1、3.2.2、3.2.4中至少一项满足；或（2）3.2.5、3.2.6均满足。

C档：已有具体事件的公开处置或官方回应。判定条件：3.2.5、3.2.6中至少一项满足。

D档：未达到C档条件。

3.3 独立核验和持续合规审查能力

评估内容：评估一国人工智能系统和组织接受独立核验、并在投入使用后接受持续合规审查的制度与实践，从上线前的法定评估要求、官方测评认证项目、国家认可体系、持续监督要求和公开的核验结论等方面观察，具体包含以下三级指标：

3.3.1 法律法规要求特定人工智能系统投入使用前接受第三方评估、认证或主管机关的安全评估；

3.3.2 有政府设立或政府认可的人工智能测评、认证项目（自愿参加的亦计）；

3.3.3 国家认可机构已将人工智能管理体系认证纳入认可范围；

3.3.4 本国具体组织取得人工智能管理体系或人工智能产品认证；

3.3.5 法律法规或官方方案要求对投入使用的人工智能系统持续监督或定期重新评估；

3.3.6 2024年以来主管机关或官方项目公开针对具体人工智能系统或组织的评估、核验或认证结论。

数据来源：公开检索，检索渠道及资料包括：法规原文、监管机关公告、国家认可机构和认证主管机关的公告与名录、官方测评项目官网等。

分档标准：

S档：形成上线前评估、认证认可与持续监督相衔接的安排。判定条件：3.3.1满足，且3.3.2、3.3.3中至少一项满足，且3.3.5、3.3.6均满足。

A档：具备正式评估安排，并有持续监督或公开结论。判定条件：3.3.1、3.3.2中至少一项满足，且3.3.5、3.3.6中至少一项满足。

B档：具备正式核验安排，或已有认证与持续监督。判定条件：满足以下任一组合：（1）3.3.1—3.3.3中至少一项满足；或（2）3.3.4、3.3.5均满足。

C档：已有本国组织获得人工智能相关认证。判定条件：3.3.4满足。

D档：未达到C档条件。

4.1 风险信号识别、动态更新和持续监测能力

评估内容：评估一国持续发现和更新人工智能风险信号的能力，从风险监测职责、风险评估或态势报告及其周期性、风险预警记录、风险清单或监测规则的更新、面向社会的风险信息发布渠道等方面观察，

具体包含以下三级指标：

4.1.1 有官方机构承担人工智能风险监测职责；

4.1.2 有官方发布的人工智能风险清单、风险评估报告或态势报告；

4.1.3 上述监测或态势报告周期性发布；

4.1.4 2024年以来官方发布人工智能风险预警、通报或风险提示；

4.1.5 风险清单或监测规则有更新或版本记录；

4.1.6 有面向公众或行业的人工智能风险信息发布渠道。

数据来源：公开检索，检索渠道及资料包括：国家计算机应急响应机构、监管机关和人工智能安全研究所官网，官方风险报告与通报等。

分档标准：

S档：形成风险监测、定期报告、预警与更新安排。判定条件：4.1.1—4.1.5均满足。

A档：具备监测职责和风险报告，并有周期发布或预警。判定条件：4.1.1、4.1.2均满足，且4.1.3、4.1.4中至少一项满足。

B档：具备监测职责及配套信息产出或发布渠道。判定条件：4.1.1满足，且4.1.2、4.1.4、4.1.6中至少一项满足。

C档：具备基础风险监测或信息发布安排。判定条件：4.1.1、4.1.2、4.1.4、4.1.6中至少一项满足。

D档：未达到C档条件。

4.2 人工智能资产登记与可视化、行为审计和异常发现机制

评估内容：评估一国掌握人工智能系统部署情况并发现异常的能力，从人工智能系统登记或备案及其公开、日志留存要求、审计抽查记录、异常行为的发现处置和登记范围的更新等方面观察，具体包含以下三级指标：

4.2.1 有人工智能系统登记、算法登记或备案制度；

4.2.2 登记或备案信息公开可查；

4.2.3 法律法规要求人工智能系统留存日志或保留可供审计的记录；

4.2.4 2024年以来对人工智能系统开展审计、抽查或核查；

4.2.5 2024年以来发现并处置人工智能系统异常行为的记录；

4.2.6 登记或审计安排有更新或扩展记录。

数据来源：公开检索，检索渠道及资料包括：政府算法登记与人工智能用例平台、法规原文、审计机关与监管机关报告等。

分档标准：

S档：形成登记、公开、日志、审计与异常处置安排。判定条件：4.2.1—4.2.5均满足。

A档：登记或备案信息公开，并已开展审计抽查。判定条件：4.2.1、4.2.2、4.2.4均满足。

B档：具备登记制度及配套监督或处置安排。判定条件：4.2.1满足，且4.2.2—4.2.5中至少一项满足。

C档：具备基础登记、审计或异常发现安排。判定条件：4.2.1—4.2.6中至少一项满足。

D档：未达到C档条件。

4.3 风险事件响应、应急联动、系统恢复和复盘改进能力

评估内容：评估一国处置人工智能事件的能力，从应急预案、跨机构联动安排、含人工智能情景的演练、实际事件处置、事后复盘改进和系统恢复等方面观察，具体包含以下三级指标：

4.3.1 有涵盖人工智能的应急预案或处置规程；

4.3.2 有明确的跨机构应急联动安排；

4.3.3 2024年以来开展含人工智能情景的应急演练；

4.3.4 2024年以来处置实际人工智能事件；

4.3.5 有事件处置后的复盘、调查报告或改进措施；

4.3.6 有系统恢复、重新启用或整改验证的记录。

数据来源：公开检索，检索渠道及资料包括：国家计算机应急响应机构与主管机关官网、亚太计算机应急组织（APCERT）等区域组织的演练报告、官方事件通报与调查报告、人工智能事件数据库（AIID）等。

分档标准：

S档：形成预案、实操、复盘及联动或恢复安排。判定条件：4.3.1、4.3.5均满足，且4.3.2、4.3.6中至少一项满足，且4.3.3、4.3.4中至少一项满足。

A档：具备应急预案和实操记录，并有配套联动或改进。判定条件：4.3.1满足，且4.3.2、4.3.5、4.3.6中至少一项满足，且4.3.3、4.3.4中至少一项满足。

B档：具备应急预案及联动、演练或处置记录。判定条件：4.3.1满足，且4.3.2—4.3.4中至少一项满足。

C档：具备基础应急预案或实操记录。判定条件：4.3.1、4.3.3、4.3.4中至少一项满足。

D档：未达到C档条件。

5.1 人工智能治理生态各方责任边界

评估内容：评估一国法律对人工智能开发者、提供者、部署者、最终用户和监管者责任的划分程度，以及依据这些条款追责的实践，具体包含以下三级指标：

5.1.1 法律法规明确人工智能开发者的义务；

5.1.2 法律法规明确人工智能提供者或服务运营者的义务；

5.1.3 法律法规明确人工智能部署者或使用单位的义务；

5.1.4 法律法规明确最终用户的义务或行为规范；

5.1.5 法律法规明确监管者的职责与程序；

5.1.6 2024年以来主管机关依据上述条款追责、处理或作出正式指导。

数据来源：公开检索，检索渠道及资料包括：各国法律法规数据库与政府公报、监管机关公告与执法决定等。

分档标准：

S档：五类角色责任均有规定，并有责任落实记录。判定条件：5.1.1—5.1.5均满足，且5.1.6满足。

A档：至少四类角色责任有规定，并有责任落实记录。判定条件：5.1.1—5.1.5中至少4项满足，且5.1.6满足。

B档：至少三类角色责任有明确规定。判定条件：5.1.1—5.1.5中至少3项满足。

C档：至少一类角色责任有明确规定。判定条件：5.1.1—5.1.5中至少1项满足。

D档：未达到C档条件。

5.2 数据治理、供应链安全和开源生态治理能力

评估内容：评估一国对人工智能训练数据、供应链和开放模型的治理规则与执行情况，从训练数据治理规则、数据保护执法、供应链安全要求、开放模型治理规则和实际核查处置等方面观察，具体包含以下三级指标：

5.2.1 有适用于人工智能训练数据的治理规则；

5.2.2 2024年以来数据保护机关就人工智能开展执法或发布正式指导；

5.2.3 有针对人工智能供应链、采购或第三方组件的安全要求；

5.2.4 有针对开放模型或开源组件的治理规则、指引或限制；

5.2.5 2024年以来主管机关对人工智能数据或供应链问题开展核查或处置。

数据来源：公开检索，检索渠道及资料包括：数据保护机关与人工智能主管机关文件、政府采购与网络安全规定、监管机关执法决定等。

分档标准：

S档：数据、供应链与开源生态治理规则齐备，并有执行记录。判定条件：5.2.1—5.2.5均满足。

A档：具备数据规则、监管指导与核查处置，并覆盖相关生态治理。判定条件：5.2.1、5.2.2均满足，且5.2.3、5.2.4中至少一项满足，且5.2.5满足。

B档：具备数据规则及配套监管或生态治理安排。判定条件：5.2.1满足，且5.2.2—5.2.4中至少一项满足。

C档：具备基础数据或生态治理安排。判定条件：5.2.1—5.2.4中至少一项满足。

D档：未达到C档条件。

5.3 智能体权限管理、责任分担与可追溯机制

评估内容：评估一国对人工智能智能体的治理安排，从专门针对智能体的官方规则或指引、权限与工具调用限制、人工监督与授权要求、责任分担、行为记录与可追溯，以及针对具体智能体产品的监管行动等方面观察，具体包含以下三级指标：

5.3.1 有官方文件专门针对人工智能智能体提出治理要求；

5.3.2 官方文件对智能体的身份标识、权限边界或工具调用作出限制性要求；

5.3.3 官方文件对智能体的高风险操作提出人工监督、人工接管或事前授权要求；

5.3.4 法律法规区分人工智能开发者、部署者和使用者之间的责任；

5.3.5 本国主体公开智能体的行为日志、动作记录或可追溯性技术说明；

5.3.6 2024年以来主管机关针对具体智能体产品或服务发布风险提示、开展调查或提出整改要求。

数据来源：公开检索，检索渠道及资料包括：监管机关与国家计算机应急响应机构公告、法规与官方指南原文、企业技术文档与系统卡等。

分档标准：

S档：智能体规则、权限、监督与追溯安排齐备，并有监管行动。判定条件：5.3.1、5.3.2、5.3.3、

5.3.5、5.3.6均满足。

A档：具备智能体专项规则及权限或监督要求，并有监管行动。判定条件：5.3.1满足，且5.3.2、5.3.3中至少一项满足，且5.3.6满足。

B档：已形成智能体专项政策安排。判定条件：5.3.1满足。

C档：具备人工监督、责任区分或行为追溯基础。判定条件：5.3.3—5.3.5中至少一项满足。

D档：未达到C档条件。

5.4 认证互认、争议解决与跨境执行能力

评估内容：评估一国在人工智能领域跨境协作和跨境执行的能力，从认证结果的国际互认、跨境监管合作、跨境争议受理、跨境联合调查、对境外人工智能服务提供者的执法和跨境服务的合规审查等方面观察，具体包含以下三级指标：

5.4.1 本国认可机构是国际认可论坛（IAF）或国际实验室认可合作组织（ILAC）互认协议签署方；

5.4.2 与他国签有涉及人工智能的监管合作安排、备忘录或联合声明；

5.4.3 有受理涉及人工智能的跨境争议或投诉的机制；

5.4.4 2024年以来与他国开展人工智能相关的联合调查、协助执行或信息协作；

5.4.5 2024年以来对境外人工智能服务提供者作出司法裁判或行政决定；

5.4.6 有面向跨境人工智能服务的合规审查或准入安排。

数据来源：公开检索，检索渠道及资料包括：国际认可论坛与国际实验室认可合作组织签署方名单、监管机关合作协议与联合公告、法院判决与行政决定等。

分档标准：

S档：形成互认、合作、争议处理与跨境执行安排。判定条件：5.4.1—5.4.6均满足。

A档：具备跨境调查和裁判实践，并有争议或准入机制。判定条件：5.4.3、5.4.6中至少一项满足，且5.4.4、5.4.5均满足。

B档：已有跨境联合调查或针对境外提供者的处理实践。判定条件：5.4.4、5.4.5中至少一项满足。

C档：具备跨境争议处理或合规准入机制。判定条件：5.4.3、5.4.6中至少一项满足。

D档：未达到C档条件。

6.1 国际机制参与、标准互操作和评测协议共享机制

评估内容：评估一国参与国际人工智能治理的深度，从专门国际机制的成员身份、在多边机制中的治理参与、国际标准化参与、共同声明署名、国际联合评测和主办国际治理会议等方面观察，具体包含以下三级指标：

6.1.1 是国际人工智能安全或治理专门机制的成员；

6.1.2 2024年以来在联合国、经济合作与发展组织、联合国教科文组织等多边机制中参与人工智能治理；

6.1.3 是ISO/IEC JTC 1/SC 42的参与成员；

6.1.4 在人工智能国际共同声明、共同方法或共识文件上署名；

6.1.5 参与国际联合评测或共享评测协议；

6.1.6 2024年以来主办国际人工智能治理会议或峰会。

数据来源：公开检索，检索渠道及资料包括：国际机制官网成员名单与会议文件、ISO/IEC JTC 1/SC 42成员名单、国际共同声明原文等。

分档标准：

S档：国际机制、标准、声明、评测与会议参与齐备。判定条件：6.1.1—6.1.6均满足。

A档：参与多边治理和国际标准化，并有其他实质参与。判定条件：6.1.1、6.1.5、6.1.6中至少一项满足，且6.1.2、6.1.3均满足。

B档：参与多边治理并具备相关国际合作基础。判定条件：6.1.1、6.1.3、6.1.4、6.1.6中至少一项满足，且6.1.2满足。

C档：具备基础国际治理参与记录。判定条件：6.1.2—6.1.4中至少一项满足。

D档：未达到C档条件。

6.2 跨境风险信息通报、案例共享和联合演练机制

评估内容：评估一国与他国交换人工智能风险信息和开展联合演练的能力，从国际应急协作网络的成员身份、明确涵盖人工智能的信息共享安排、与他国联合发布的人工智能安全通报或指南、含人工智能情景的跨国演练等方面观察，具体包含以下三级指标：

6.2.1 国家计算机应急响应机构是国际或区域应急协作组织成员；

6.2.2 与他国签有明确涵盖人工智能的风险信息共享协议、合作备忘录或联合行动安排；

6.2.3 2024年以来与他国机构联合发布人工智能安全风险通报、预警或安全指南；

6.2.4 2024年以来参与含人工智能情景的跨国联合演练。

数据来源：公开检索，检索渠道及资料包括：国际应急响应与安全组织（FIRST）、亚太计算机应急组织（APCERT）等成员名单，各国国家计算机应急响应机构和网络安全机构官网，联合指南原文的署名页，演练方案与完成报告等。

分档标准：

S档：具备跨境共享安排、联合产出和演练实践。判定条件：6.2.2—6.2.4均满足。

A档：已参与跨国演练，并有共享安排或联合产出。判定条件：6.2.2、6.2.3中至少一项满足，且6.2.4满足。

B档：已有跨境共享、联合通报或演练安排。判定条件：6.2.2—6.2.4中至少一项满足。

C档：已加入国际或区域应急协作组织。判定条件：6.2.1满足。

D档：未达到C档条件。

6.3 能力建设、知识分享和公共产品供给能力

评估内容：评估一国在人工智能安全与治理领域提供培训、知识和公共工具的能力，从面向本国的治理培训、面向他国的能力建设项目、公开的治理知识资源、可供他方使用的安全与治理工具以及这些资源的持续更新等方面观察，具体包含以下三级指标：

6.3.1 政府或公立机构面向本国举办人工智能安全或治理培训；

6.3.2 本国政府或其资助机构主办、实施或出资面向他国的人工智能能力建设项目，内容包含治理、安全或伦理；

6.3.3 政府或公立机构公开发布人工智能安全或治理知识资源（指南、手册、教材）；

6.3.4 本国机构公开提供可供他方使用的人工智能安全或治理工具、测评数据集或平台；

6.3.5 上述培训、资源或工具有持续更新或迭代记录。

数据来源：公开检索，检索渠道及资料包括：政府与国际组织项目页面，联合国教科文组织、国际电信联盟、经济合作与发展组织能力建设项目清单，GitHub、Hugging Face公开资源等。

分档标准：

S档：国内培训、对外支持与公共资源供给齐备并持续更新。判定条件：6.3.1—6.3.5均满足。

A档：具备国内培训、对外支持及知识和工具供给。判定条件：6.3.1—6.3.4均满足。

B档：开展国内培训，并有资源供给或更新支持。判定条件：6.3.1满足，且6.3.3—6.3.5中至少一项满足。

C档：具备基础培训或公共资源供给。判定条件：6.3.1、6.3.3、6.3.4中至少一项满足。

D档：未达到C档条件。

6.4 公众参与、媒体监督、风险沟通、人工智能素养提升

评估内容：评估一国公众参与人工智能治理和获得风险信息的渠道，从政策立法的公开征求意见及官方回应、面向公众的风险沟通与素养教育、涵盖人工智能问题的投诉举报渠道等方面观察。其中，媒体监督和社会监督主要通过公众反映人工智能问题的投诉举报渠道加以考察，具体包含以下三级指标：

6.4.1 2024年以来就人工智能政策或立法公开征求公众意见；

6.4.2 官方公开发布对公众意见的汇总或处理说明；

6.4.3 官方发布面向公众的人工智能风险提示或防范材料；

6.4.4 有国家层面的人工智能素养、教育或技能培训计划；

6.4.5 有涵盖人工智能问题的公众投诉、举报渠道。

数据来源：公开检索，检索渠道及资料包括：政府意见征集平台与回应文件，监管机关、消费者保护与公共教育机构材料等。

分档标准：

S档：公众参与、意见回应、风险沟通与素养建设安排齐备。判定条件：6.4.1—6.4.5均满足。

A档：已征求公众意见，并有处理说明或投诉渠道。判定条件：满足以下任一组合：（1）6.4.1、6.4.2均满足；或（2）6.4.1、6.4.5均满足。

B档：已征求公众意见，并有风险提示、教育或投诉安排。判定条件：6.4.1满足，且6.4.3—6.4.5中至少一项满足。

C档：具备基础公众参与或风险教育安排。判定条件：6.4.1、6.4.3、6.4.4中至少一项满足。

D档：未达到C档条件。

世界互联网大会人工智能专业委员会

世界互联网大会于 2024 年设立首个专业性、常态化分支机构——人工智能专业委员会，下设标准推进计划、安全与治理推进计划和产业推进计划。专委会汇集了来自人工智能领域国际组织、知名智库、科研院所、 专业协会及产业界的权威专家和专业人才，秉持搭建国际交流合作平台、推动发展与治理协同、促进全球共享人工智能发展成果的原则，通过开展专题研讨、成果分享、倡议发布等活动，不断凝聚国际共识， 促进人工智能包容普惠、可持续发展。

前沿人工智能安全与治理2026：风险认知与协同治理

如何引用本报告：

曾毅，肖恩·欧·海格缇，曹功策，鲁恩萌，陈煜，韩开宇，李凯，王正奇，常瑶，谢佳玮，彭誉萱，王沛宁，尼罗沙·阿南达，包沉浮，克里斯·布朗，常疆，陈惠洁，陈晶晶，程凯，程明，查鲁卡·塞纳尔·达穆努波拉，汤姆·戴维，段伟文，符春辉，鲁乙己·加姆巴得拉，高雪松，龚克，龚新奇，王小塞，郭苏敏，温迪·霍尔，约翰·希金斯，塞勒斯·霍兹，伯恩德·霍尔兹纳格尔，呼娜英，胡永启，胡壮，姜晨，丹尼尔·凯里米，纳达·拉阿比迪，梁叶，林琳，林明峰，林秋诗，林雪艳，柳嘉琪，刘威辰，马兰，马琳，蒙美玲，孟伟，尼古拉·米亚伊，马尔科姆·默里，牛晓玲，迪迪埃·恩库里基姆弗拉，潘剑锋，逄孝钢，亨利·帕帕达托斯，彭韬，埃德森·普雷斯蒂，乔迁，乔宇，沈俊成，谭知行，唐银山，陶锋，希梅娜·索菲亚·比维罗斯·阿尔瓦雷斯，王峰，汪凤琼，王健兵，王梦寅，王千玉，王融，王盛杰，王巍，王晓箴，王欣，王星光，魏凯，罗伯·沃瑟姆，吴淑燕，吴胤禹，夏和华，夏文辉，肖荐午，徐德亮，徐继明，许梦雯，薛澜，杨勃，杨玲玲，杨明川，杨绍波，杨耀东，杨忠良，姚新，约翰·杨，艾墨·尤纳斯，尉迟学彪，岳巧丽，张凌寒，张荣，张晓赫，张亚勤，张勇，张子婷，周凯，张雪丽，梁昊，任贤良. 前沿人工智能安全与治理2026：风险认知与协同治理[R/OL]. 北京：世界互联网大会，2026.

关注大会公众号

© 2026 世界互联网大会


世界互联网大会
World Internet
Conference