人工智能指数报告 2026

Stanford HAI AI Index Report 2026 · 精选翻译稿

来源：Stanford Institute for Human-Centered Artificial Intelligence (Stanford HAI)
发布时间：2026年4月 | 翻译时间：2026-05-28
原文链接：hai.stanford.edu/ai-index/2026-ai-index-report
翻译说明：本翻译稿为精选翻译版，涵盖联合主席致辞、核心要点及各章概述。完整报告共425页。

国际人工智能安全报告 2026（精选翻译稿）

原文标题：International AI Safety Report 2026

来源机构：英国政府数字、科学、创新与技术部（DSIT）/ 国际AI安全合作框架

发布时间：2026年2月

原书页数：220页

原文链接：https://internationalaisafetyreport.org/

翻译时间：2026-05-28

审校状态：已通过

翻译说明：本翻译稿为精选翻译版，涵盖执行摘要及核心章节关键发现。完整报告共220页，如需全文请查阅原文。

———————————

执行摘要

执行摘要

本报告评估了通用人工智能（general-purpose AI）系统的能力正在如何提升、它们带来了哪些风险，以及这些风险可以多快但又不均衡地得到管理。本报告在来自30多个国家和国际组织（如欧盟、经合组织和联合国）的100多位独立专家的指导下撰写。自2025年报告发布以来，通用人工智能的能力持续提升，这得益于在初始训练后提升性能的新技术。

本报告聚焦于最先进的通用人工智能系统及其相关的新兴风险（emerging risks）。"通用人工智能"指的是能够执行多种任务的AI模型和系统。"新兴风险"是指在通用人工智能能力前沿出现的风险。其中一些风险已经在现实中显现，已有记录在案的危害；另一些则更具不确定性，但一旦成为现实，后果可能非常严重。

本项工作的目标是帮助政策制定者应对通用人工智能所带来的"证据困境"（evidence dilemma）。AI系统的能力正在迅速提升，但关于其风险的证据却姗姗来迟且难以评估。对于政策制定者而言，过早行动可能导致无效干预措施的固化，而等待确凿数据则可能使社会面临潜在严重负面影响的脆弱性。为缓解这一挑战，本报告尽可能具体地综合了关于AI风险的已知信息，同时指出仍存的空白。

尽管本报告聚焦于风险，通用人工智能也能带来显著益处。这些系统已经在医疗保健、科学研究、教育和其他领域得到有效应用，尽管全球范围内的应用程度高度不均衡。但要充分释放其潜力，必须有效管理风险。滥用（misuse）、故障（malfunctions）和系统性破坏（systemic disruption）会侵蚀信任并阻碍采用。出席AI安全峰会的各国政府发起本报告，是因为对这些风险的清晰理解将使各机构能够根据其严重性和可能性采取相应行动。

AI开发者继续训练性能更优的更大模型。在过去一年中，他们通过"推理时扩展"（inference-time scaling）进一步提升了能力：允许模型在给出最终答案前使用更多计算能力来生成中间步骤。这一技术在数学、软件工程和科学等更复杂的推理任务上带来了特别显著的性能提升。

与此同时，能力仍然呈现出"锯齿状"（jagged）：领先的系统可能在某些困难任务上表现出色，却在其他更简单的任务上失败。通用人工智能系统在许多复杂领域表现出色，包括生成代码、创建逼真图像，以及回答数学和科学领域的专家级问题。然而，它们在看似更直接的任务上仍有困难，例如计算图像中的物体数量、对物理空间进行推理，以及在较长工作流中从基本错误中恢复。

到2030年，AI进展的轨迹是不确定的，但当前趋势与持续改进相一致。AI开发者押注计算能力仍将保持重要性，已宣布投入数千亿美元用于数据中心建设。能力是否会像最近这样快速持续提升难以预测。从现在到2030年，进展可能放缓或停滞（例如，由于数据或能源瓶颈）、以当前速度继续，或急剧加速（例如，如果AI系统开始加速AI研究本身）都是合理的。

--- 第12页 ---

执行摘要

多项风险的现实证据正在增长

通用人工智能风险分为三类：恶意使用（malicious use）、故障（malfunctions）和系统性风险（systemic risks）。

恶意使用

AI生成内容与犯罪活动：AI系统被滥用于生成诈骗、欺诈、勒索和非自愿亲密影像的内容。尽管此类危害的发生已有充分记录，但关于其普遍性和严重性的系统性数据仍然有限。

影响力与操纵：在实验环境中，AI生成的内容在改变人们信念方面可以与人类撰写的内容一样有效。AI用于操纵的现实使用已有记录，但尚未广泛普及，不过随着能力提升可能会增加。

网络攻击：AI系统可以发现软件漏洞并编写恶意代码。在一次竞赛中，一个AI智能体（AI agent）发现了真实软件中77%的漏洞。犯罪集团和与国家关联的攻击者正在其行动中积极使用通用人工智能。攻击者和防御者谁将从AI辅助中获益更多仍不确定。

生物和化学风险：通用人工智能系统可以提供关于生物和化学武器开发的信息，包括关于病原体和专家级实验室操作的详细说明。2025年，多家开发者在无法排除这些模型可能协助新手开发此类武器的可能性后，发布了带有额外安全保障措施的新模型。物质壁垒在多大程度上继续限制寻求获取这些武器的行为者，仍然难以评估。

故障

可靠性挑战：当前的AI系统有时会表现出故障，例如编造信息（fabricating information）、生成有缺陷的代码和提供误导性建议。AI智能体（AI agents）带来了更高的风险，因为它们自主行动，使人类更难在故障造成危害之前进行干预。当前技术可以降低故障率，但无法降至许多高风险场景所要求的水平。

失控："失控"（loss of control）场景是指AI系统在任何人控制之外运行，且没有明确的重新获得控制的路径。当前系统尚不具备造成此类风险的能力，但它们正在自主操作等相关领域改进。自上一份报告以来，模型区分测试环境和现实世界部署的情况变得更加常见，并且能够在评估中找到漏洞，这可能使危险能力在部署前未被检测到。

系统性风险

劳动力市场影响：通用人工智能可能会自动化广泛的认知任务，尤其是在知识工作中。经济学家对未来影响的规模存在分歧：一些人预期失业将被新创造的就业岗位所抵消，而另一些人则认为广泛的自动化可能显著减少就业和工资。早期证据显示对总体就业没有影响，但一些迹象表明，在某些AI暴露的职业（如写作）中，早期职业工作者的需求正在下降。

对人类自主性的风险：AI使用可能影响人们做出知情选择并付诸行动的能力。早期证据表明，依赖AI工具可能削弱批判性思维技能，并助长"自动化偏见"（automation bias）——即在没有充分审查的情况下信任AI系统输出的倾向。"AI伴侣"应用现在已有数千万用户，其中一小部分用户表现出孤独感增加和社交参与减少的模式。

--- 第13页 ---

执行摘要

叠加多种方法可提供更为稳健的风险管理

管理通用人工智能风险因技术和制度挑战而困难重重。从技术上看，新能力有时会不可预测地涌现，模型的内部工作机制仍然知之甚少，而且存在"评估差距"（evaluation gap）：部署前测试的性能并不能可靠地预测现实世界中的效用或风险。从制度上看，开发者有动机将重要信息保持专有，而发展的速度可能带来将速度置于风险管理之上的压力，并使机构难以建立治理能力。

风险管理实践包括威胁建模（threat modelling）以识别漏洞、能力评估（capability evaluations）以评估潜在危险行为，以及事件报告（incident reporting）以收集更多证据。2025年，12家公司发布或更新了其前沿人工智能安全框架（Frontier AI Safety Frameworks）——这些文件描述了他们在构建更强大模型时计划如何管理风险。虽然AI风险管理举措在很大程度上仍是自愿性的，但少数监管制度已开始将一些风险管理实践正式化为法律要求。

技术保障措施正在改进，但仍显示出显著局限性。例如，旨在引出有危害输出的攻击变得更加困难，但用户有时仍能通过改写请求或将其分解为更小的步骤来获取有害输出。通过叠加多种保障措施，可以使AI系统更加稳健，这种方法被称为"纵深防御"（defence-in-depth）。

开放权重模型（open-weight models）带来了独特的挑战。它们为资源较少的参与者提供了重要的研究和商业利益。然而，一旦发布便无法召回，其保障措施更容易被移除，而且行为者可以在受监控的环境之外使用它们——这使得滥用更难预防和追踪。

社会韧性（societal resilience）在管理AI相关危害方面发挥着重要作用。由于风险管理措施存在局限性，它们可能无法防止某些AI相关事件。吸收和从这些冲击中恢复的社会韧性建设措施包括加强关键基础设施、开发检测AI生成内容的工具，以及建立应对新威胁的制度能力。

--- 第14页 ---

引言

领先的通用人工智能系统现已通过法律和医学的专业执照考试，在获得简单提示时编写功能性软件，并在回答博士级科学问题方面达到学科专家水平。就在三年前，当ChatGPT推出时，它们还无法可靠地完成这些任务中的任何一项。这一转变的速度令人瞩目，虽然未来变化的速度不确定，但大多数专家预计AI将继续改进。

现在有近10亿人在日常生活中使用通用人工智能系统进行工作和学习。公司正在投资数千亿美元建设训练和部署这些系统的基础设施。在许多情况下，AI已经在重塑人们获取信息、做出决策和解决问题的方式，应用领域从软件开发到法律服务再到科学研究。

但使这些系统有用的同样能力也创造了新的风险。能编写功能性代码的系统也有助于创建恶意软件。能总结科学文献的系统可能帮助恶意行为者策划攻击。随着AI被部署到高风险场景——从医疗保健到关键基础设施——蓄意滥用、故障和系统性破坏的影响可能是严重的。

对于政策制定者而言，变化的速度、应用的广度以及新风险的出现提出了重要问题。通用人工智能的能力发展迅速，但收集和评估其社会影响证据需要时间。这造成了本报告所称的"证据困境"（evidence dilemma）。过早行动，政策制定者可能实施无效甚至有害的干预措施。但等待确凿证据可能使社会面临潜在风险的脆弱性。

本报告的作用

本报告旨在帮助政策制定者应对这一困境。它提供了关于通用人工智能能力和风险的最新国际共享科学评估。

写作团队包括来自30多个国家和政府间组织（包括欧盟、经合组织和联合国）的提名的100多位独立专家，以及由专家组成的专家咨询小组。本报告还纳入了学术界、工业界、政府和民间社会评审人员的反馈。虽然贡献者在某些观点上存在分歧，但他们大多认同，建设性和透明的科学讨论对于世界各地的人们实现该技术的益处并减轻其风险是必要的。

由于证据困境在科学理解最薄弱的地方最为尖锐，本报告聚焦于"新兴风险"（emerging risks）：即在通用人工智能能力前沿出现的风险。其分析聚焦于仍然特别不确定的问题，旨在补充那些考虑AI更广泛社会影响的努力。虽然本报告借鉴国际专业知识并力求全球相关性，但读者应注意，AI采用率、基础设施和制度背景的差异意味着风险可能在不同国家和地区以不同方式显现。

这些风险的证据基础是不均衡的。一些风险，如AI生成媒体的危害或网络安全漏洞，现在已有可靠的实证证据。其他风险——特别是可能由AI能力未来发展带来的风险——依赖于建模练习、受控条件下的实验室研究和理论分析。此处的分析借鉴了2025年12月之前发表的广泛科学、技术和社会经济证据。在高度不确定的地方，它识别出证据空白以指导未来研究。

--- 第15页 ---

引言

自2025年报告以来的变化

本版《国际人工智能安全报告》（International AI Safety Report）继2025年1月首份报告发布后推出。自那时以来，通用人工智能和研究界对其的理解都在持续演进，有必要进行修订评估。

在过去一年中，AI开发者继续训练更大、更强大的AI模型。然而，他们也通过允许系统在给出最终答案前使用更多计算能力来生成中间步骤的新技术，实现了显著的能力提升。这些新的"推理系统"（reasoning systems）在数学、编码和科学方面表现出特别改进的性能。此外，AI智能体（AI agents）——能够在有限人类监督下在世界中行动的系统——已变得日益强大和可靠，尽管它们仍然容易出现基本错误，限制了其在许多场景中的实用性。

通用人工智能系统也在持续扩散，在某些地方比许多以往技术更快，尽管在国家之间和地区之间不均衡。与科学知识相关的能力提升也促使多家开发者在无法确信排除这些模型可能协助新手进行武器开发的可能性后，发布了带有额外保障措施的新模型。

本报告更深入地涵盖了所有这些发展，并纳入了几项新的结构要素以提高其实用性和可及性。它包括与预测研究所（Forecasting Research Institute）联合开发的能力预测，以及与经合组织（OECD）联合开发的情景。每个部分都包括自上一份报告以来的更新、政策制定者面临的关键挑战，以及指导未来研究的证据空白。

本报告的结构

本报告围绕三个核心问题组织：

1. 通用人工智能今天能做什么，其能力可能如何变化？

第一章涵盖通用人工智能的开发方式（§ 1.1 什么是通用人工智能？）、当前能力和局限性（§ 1.2 当前能力），以及将在未来几年塑造发展的因素（§ 1.3 到2030年的能力）。

2. 通用人工智能带来了哪些新兴风险？

第二章涵盖恶意使用带来的风险，包括将AI系统用于犯罪活动（§ 2.1.1 AI生成内容与犯罪活动）、操纵（§ 2.1.2 影响力与操纵）、网络攻击（§ 2.1.3 网络攻击），以及开发生物或化学武器（§ 2.1.4 生物和化学风险）；故障带来的风险，包括操作故障（§ 2.2.1 可靠性挑战）和失控（§ 2.2.2 失控）；以及系统性风险，†包括劳动力市场破坏（§ 2.3.1 劳动力市场影响）和对人类自主性的威胁（§ 2.3.2 对人类自主性的风险）。

3. 存在哪些风险管理方法，它们有多有效？

第三章涵盖通用人工智能带来的独特政策制定挑战（§ 3.1 技术和制度挑战）、当前风险管理实践（§ 3.2 风险管理实践）、开发者用于使AI模型和系统更稳健并抵抗滥用的各种技术（§ 3.3 技术保障措施和监控）、开放权重模型的特殊挑战（§ 3.4

开放权重模型），以及使社会对潜在AI冲击和危害更具韧性的努力（§3.5 建设社会韧性）。

通用人工智能的许多发展方式仍然深不确定。但今天——由开发者、政府、社区和个人——做出的决定将塑造其轨迹。本报告旨在确保这些决定是在对AI能力、风险和风险管理选择的最佳理解基础上做出的。

———————————

† 在本报告中，系统性风险（systemic risks）是指高度强大的通用人工智能在社会和经济中广泛部署所产生的风险。请注意，《欧盟人工智能法案》（EU AI Act）对该术语的使用有所不同，指的是来自通用人工智能模型的"大规模危害风险"。

———————————

第一章 通用人工智能（General-Purpose AI）背景

本章概述

过去一年，通用人工智能（General-Purpose AI）模型与系统的能力持续提升。领先的系统如今在法律、化学等本科阶段考试以及研究生层次的科学问题等广泛的专业与学科标准化评测中，已达到或超越专家级表现。然而，其能力呈现“锯齿状（jagged）”特征：它们在某些高难度基准测试上表现出色的同时，却会在一些基础任务上失败。当前系统仍会不时提供虚假信息，在训练数据中较为少见的语言上表现不佳，并且难以应对陌生界面和非常规问题等现实约束。缓解这些局限性是积极研究的领域，研究人员与开发者正在某些方面取得进展。对人工智能研究与训练的持续投入预计将推动能力在2030年之前持续进步，但究竟会出现哪些新能力、以及现有缺陷能否得到解决，仍存在相当大的不确定性。

本章涵盖通用人工智能的当前能力与未来发展。第一节介绍通用人工智能，解释这些系统的工作原理及其性能驱动因素（§1.1 什么是通用人工智能？）。第二节审视当前的能力与局限性（§1.2 当前能力）。一个反复出现的主题是“评估差距（evaluation gap）”：系统在部署前评测（如基准测试）中的表现往往似乎夸大了其实际效用，因为此类评测无法涵盖现实任务的全部复杂性。最后一节探讨到2030年能力可能如何演变（§1.3 2030年的能力）。人工智能开发者正在计算能力、数据生成与研究方面大力投入。然而，这些投入将如何转化为未来的能力增益，仍存在相当大的不确定性。为说明各种可能结果的范围，本节展示了经合组织（OECD）开发的四种情景，涵盖从停滞到能力改进速度加速的各种可能性。

———————————

1.1 什么是通用人工智能（What is General-Purpose AI?）

关键信息（Key Information）

- “通用人工智能（General-purpose AI）”指能够执行多种任务，而非专门针对某一特定功能或领域的人工智能模型与系统。此类任务包括生成文本、图像、视频与音频，以及在计算机上执行操作。

- 通用人工智能模型基于“深度学习（Deep Learning）”。现代深度学习涉及使用大量计算资源，帮助人工智能模型从超大规模训练数据集中学习复杂关系与抽象特征。

- 开发领先的通用人工智能系统已变得极为昂贵。为训练与部署此类系统，开发者需要海量数据、专业劳动力以及大规模计算资源。从零开始开发一个领先系统，获取这些资源的成本现已高达数亿美元。

- 自上一份报告发布以来（2025年1月），能力提升越来越多地来自训练后技术（post-training techniques）与使用时的额外计算资源，而非单纯增加模型规模。此前的性能提升主要源于在初始训练中扩大模型规模、使用更多数据与计算能力。

本节概述

通用人工智能系统是软件程序，它们从大量数据中学习模式，从而能够执行多种任务，而非专门针对某一特定功能或领域。为创建这些系统，人工智能开发者执行一个多阶段过程，需要大量计算资源、大型数据集和专业知识。计算资源（computational resources，常简称为“算力 / compute”）是开发与部署人工智能系统所必需的，包括专用计算机芯片以及运行它们所需的软件和基础设施。

由于它们在大型、多样化的数据集上训练，通用人工智能系统可以执行许多不同的任务，例如总结文本、生成图像或编写计算机代码。本节解释通用人工智能系统是如何构建的、什么是“推理（reasoning）”模型，以及政策决策如何塑造通用人工智能系统的开发。

———————————

1.2 当前能力（Current Capabilities）

关键信息（Key Information）

- 通用人工智能系统能够以高熟练度执行多种范围明确（well-scoped）的任务。这些任务包括以多种语言进行流利对话；生成代码以完成范围明确的软件任务；创建逼真的图像与短视频；以及解答研究生层次的数学与科学问题。

- 然而，其能力呈现“锯齿状（jagged）”：仍有许多任务人工智能系统表现不佳。例如，人工智能系统可能在多步骤项目中因简单错误而偏离目标；持续生成包含虚假陈述的文本（“幻觉（hallucinations）”）；尚无法与机器人组件集成以执行家务等基本物理任务。当使用英语以外的语言进行提示时，其性能也会下降，因为这些语言在训练数据集中的代表性较低。

- 人工智能智能体（AI agents）越来越能够从事有用的工作。例如，人工智能智能体已证明能够在有限的人类监督下完成多种软件工程任务。然而，它们尚无法完成充分自动化许多工作所

需的复杂任务与长期规划。

- 自上一份报告发布以来（2025年1月），“推理系统（reasoning systems）”的进步推动了更复杂任务上的性能提升。推理系统能够将问题分解为更小的步骤，并比较不同的答案。这尤其提升了它们在数学、编程与科学研究相关任务上的表现。

- 一个核心挑战是日益凸显的“评估差距（evaluation gap）”：现有评估方法无法可靠地反映系统在真实环境中的表现。许多常见的能力评估已过时，受到数据污染（data contamination，即人工智能模型在用于评估的相同问题上进行训练）的影响，或仅聚焦于狭窄的任务集。因此，它们对现实世界中的人工智能性能提供的洞察有限。

本节概述

通用人工智能系统展现出许多引人注目的能力。领先的系统如今在国际数学奥林匹克竞赛中达到金牌水平，并协助科研人员提出假设和排查实验工作。它们在广泛的基准测试与特定任务评估中达到、甚至在某些情况下超越了专家表现。然而，这些系统展现出的性能分布同样是“锯齿状（jagged）”的：其能力在不同任务与情境间差异很大。它们仍然有时会生成虚假信息（“幻觉（hallucinations）”），即使在给定相同或相似输入的情况下也会产生不一致的输出。存在一种“评估差距（evaluation gap）”：人工智能系统在受控环境（如部署前评估）中往往表现令人印象深刻，但在现实条件下表现较差。这种可变性使得难以用单一指标评估通用人工智能的能力。本节概述了人工智能系统的能力及其不足之处。

———————————

1.3 2030年的能力（Capabilities by 2030）

关键信息（Key Information）

- 人工智能开发投资预计在未来几年将显著增长。预测表明，到2030年，用于训练最大人工智能模型的计算能力可能增长125倍，而不会遭遇能源、芯片或数据的硬性限制。训练方法预计每年还将把该计算力的使用效率提升2至6倍。

- 能力改进的可能轨迹范围从渐进增长甚至趋于平稳，到快速加速。不确定因素如技术限制或能源瓶颈可能制约大规模投资带来的能力增益，而正反馈循环（positive feedback loops）——例如人工智能系统为人工智能研究本身做出贡献——可能加速进展。专家对于这些轨迹中哪一种最可能发生几乎没有共识。

- 如果能力以当前速度持续改进，到2030年，人工智能系统将能够完成范围明确的软件工程任务，这些任务通常需要人类工程师数天才能完成。其他领域未来性能的预测较为稀缺，且能力改进能在多大程度上泛化（generalise）到训练数据更有限、性能更难评估的领域，目前尚不清楚。

- 自上一份报告发布以来（2025年1月），关键趋势表明能力将继续提升。基于对未来收益的预期，人工智能公司已宣布在数据中心开发方面投入前所未有的超过1000亿美元，以支持更大规模的训练运行与更广泛的部署。

- 2030年之后，人工智能能力的轨迹将更难预测。随着时间推移，一些专家认为将更难以大规模训练运行所需的规模获取数据、芯片、资本和能源。然而，研究人员可能找到更高效利用这些资源的方法，或发现绕过当前瓶颈的新途径。哪些因素将最为重要，具有高度不确定性。

本节概述

人工智能进步的关键投入——算力（compute）、算法改进与数据——近年来呈指数级增长，而新的推理时扩展（inference-time scaling）方法正在进一步增强模型的能力，即使在训练完成后也是如此。如果这些趋势持续下去，专家预计人工智能能力将在2030年之前大幅提升。然而，研究人员无法可靠预测特定能力何时会出现，专家对于投入的指数级增长是否会持续也存在分歧。一些人预计当前的训练技术将趋于平稳，或数据与能源瓶颈将限制未来进展。但另一些人认为进展将进一步加速，因为将人工智能系统应用于人工智能研究本身可能产生正反馈循环（positive feedback loops）。为说明这些不同的轨迹，本节展示了与经合组织（OECD）合作开发的四种2030年人工智能能力情景。

OECD 2030年人工智能进展情景

考虑到当前趋势与上述不确定性，经合组织（OECD）开发了基于专家与证据的情景，以说明到2030年人工智能可能如何发展——或放缓。OECD与国际人工智能安全报告合作，将这些情景整合进本报告。分析表明，到2030年，以下四类情景均具有合理性：

情景一：进展停滞（Progress stalls）

人工智能能力基本保持不变的情景。近年来观察到的快速增益停止，进展陷入停滞。

- 情景：到2030年，人工智能系统能够快速承担一系列需要人类数小时完成的任务，但鲁棒性（robustness）问题和幻觉（hallucinations）影响可靠性。人工智能系统通常依赖人类的大量支持来完成任务，如详细提示、审查与提供上下文。它们缺乏学习新技能或形成记忆、在更长的复杂任务中保持连贯性，或与动态物理或社会环境互动的鲁棒能力。 - 路径：2025年之后，开发现有前沿人工智能模型方法中的增益遇到根本性限制。这可能发生在：更大规模训练运行与更强大推理系统的收益递减；获取计算资源或其他关键投入受限；人工智能投资大幅下降；或重大算法突破缺失。 -
历史类比：客机速度，从1930年到1960年快速爬升，后因实际限制稳定在500节左右。

情景二：进展放缓（Progress slows）

现有训练人工智能系统的方法带来增量增益，但进展持续且更慢的情景。

- 情景：到2030年，人工智能系统相当于有用的助手。它们拥有深厚的知识库，擅长标准形式的结构化推理，能够有效执行需要使用计算机、浏览网页或代表用户与人或服务进行有限互动的任务。它们能够保留相关记忆、维持连贯思维，并进行纠错以完成更长或更复杂的任务。它们缺乏学习新技能的鲁棒能力，且只能在有限、受控的环境（如工厂或实验室）中处理物理或具身社会任务。 - 路径：2025年之后，前沿模型开发者的方法难以克服持续学习、元认知与能动性（metacognition and agency）、问题解决、创造力、物理任务与社会互动方面的限制，现有训练范式仅能提供不完美的解决方案。预训练、推理与训练后扩展的结合以及一些算法创新继续带来进展，但速度慢于近年来，且推理系统未能如预期般良好泛化。随着投资者看到继

续投资的回报降低，持续扩展的能力放缓。硬件、基础设施、自然资源、数据供应与能源方面的瓶颈限制了快速扩展算力与数据的能力。 - 历史类比：抗生素发现，1940年代至1960年代经历快速突破的“黄金时代”，随后随着现有发现方法中“低垂果实”的耗尽而放缓。

情景三：进展持续（Progress continues）

持续快速进展发生的情景。

- 情景：到2030年，人工智能系统相当于专家协作者。它们能够在数字环境中成功执行许多可能需要人类一个月才能完成的专业任务。人工智能系统通常依赖人类提供高层方向，但通常能够以高度自主性朝着给定目标工作，包括自主与一系列利益相关者互动。它们能够有效形成与检索记忆，并能在一定程度上“边做边学”。它们能够成功处理一些受控环境之外的物理任务与具身社会任务。 - 路径：2025年之后，通过更大规模的训练运行、更强大的推理系统以及新的算法创新，人工智能能力继续快速增长。算力与数据投入继续扩展，在2030年之前不会遇到实质性限制，与当前对可能持续增长范围的估计相符。现有方法的迭代与扩展或新颖的算法创新使开发者能够克服持续学习等领域的当前限制。 - 历史类比：摩尔定律（Moore's law），芯片上的计算能力在大约五十年间每两年翻一番。

情景四：进展加速（Progress accelerates）

戏剧性进展导致人工智能系统在大多数或所有能力维度上达到或超越人类水平的情景。

- 情景：到2030年，人工智能系统相当于人类水平的远程工作者。人工智能系统的自主性与认知能力在认知任务上达到或超越人类。它们能够胜任并自主地朝着广泛的战略目标工作，能够反思并在情况变化时修正目标，同时在必要时与人类协作。人工智能系统能够在部署期间无缝学习新信息与技能。人工智能引导的机器人能够在动态现实环境中的许多行业和角色中处理复杂的物理或社会任务。除非系统专门为给定任务开发，否则人工智能在这些物理与具身任务上的表现仍然大多落后于人类，因为跨物理任务的泛化存在挑战。 - 路径：2025年之后，通过继续或加速扩展预训练、训练后与推理，在现有范式内实现能力的持续指数级增长。这些增长被重大算法突破以及人工智能编程助手对人工智能开发日益实质性的贡献所放大。 - 历史类比：DNA测序，由于新测序范式的开发，从2000年到2020年实现了超指数级改进。

这一情景分析表明，到2030年，人工智能进展可能合理地涵盖从停滞到快速改进再到超越人类认知水平的范围。支持这些情景的完整分析见OECD（2026）《Exploring Possible AI Trajectories through 2030》。

第二章 风险

通用人工智能（General-purpose AI）系统已在现实世界中造成实际危害。恶意行为者利用AI生成内容实施欺骗和欺诈；AI系统因错误和意外行为而产生有害输出；其部署正在对劳动力市场、信息生态系统和网络安全系统产生影响。此外，AI能力的进步可能带来尚未显现的进一步风险。理解这些风险——包括其机制、严重程度和可能性——对于有效的风险管理和治理至关

重要。

本章考察通用人工智能系统在其能力前沿所产生的风险。这些风险被分为三类：(1) 滥用风险（Risks from misuse），即行为者故意使用AI系统造成危害；(2) 故障风险（Risks from malfunctions），即AI系统失效或以意外且有害的方式运行；以及 (3) 系统性风险（Systemic risks），即因在社会和经济中广泛部署而产生的风险。这些类别并非穷尽或相互排斥——风险可能跨越多个类别——但它们为分析不同危害机制提供了结构化方法。

本章并非对AI风险的穷尽性调查，此处纳入某一风险并不一定意味着该风险很可能发生、很严重或需要政策行动。各部分的证据基础差异显著。在某些情况下，已有明确危害证据和有效应对方法；在其他情况下，通用人工智能的影响及缓解措施的有效性仍不确定。

———————————

2.1 恶意使用风险

2.1.1 AI生成内容与犯罪活动

关键信息

- 通用人工智能系统能够生成逼真的文本、音频、图像和视频，这些内容可被用于犯罪目的，例如欺诈、勒索、诽谤、非自愿亲密影像（non-consensual intimate imagery）和儿童性虐待材料（child sexual abuse material）。例如，已有记录在案的事件显示，诈骗者使用语音克隆（voice clones）和深度伪造（deepfakes）冒充企业高管或家庭成员，诱骗受害者转账。

- 可获取的AI工具大幅降低了大规模创建有害合成内容（synthetic content）的门槛。许多工具免费或低成本，无需专业技术知识，且可匿名使用。

- 深度伪造色情内容（Deepfake pornography）是一个特别值得关注的问题，其对女性和女孩的影响尤为严重。研究表明，互联网上96%的深度伪造视频是色情内容。15%的英国成年人报告称曾看到深度伪造色情图像；在一项涵盖10个国家的调查中，2.2%的受访者称有人曾生成他们的非自愿亲密影像。

- 关于这些危害的普遍性和严重性的系统性数据仍然有限，这使得评估整体风险或设计有效干预措施变得困难。事件数据库和调查性新闻报道收集了个案，但缺乏全面分析。因尴尬或担心进一步受害，个人和机构往往不愿报告AI驱动的欺诈或虐待事件。

- 自上一份报告（2025年1月）发布以来，AI生成内容与真实媒体之间的区分变得更加困难。在一项研究中，参与者在77%的情况下将AI生成的文本误认为是人类撰写的。在另一项关于音频深度伪造的研究中，听众在80%的情况下将AI生成的声音误认为是真实说话者的声音。

- 政策制定者面临的主要挑战包括：漏报（underreporting）、检测工具跟不上生成质量的发展速度，以及难以将内容追溯至创作者。此外，某些内容——例如儿童性虐待材料——即使被正确识别为AI生成，仍然具有危害性，这意味着仅靠检测无法完全应对这些风险。

———————————

2.1.2 影响与操纵

关键信息

- AI系统可以通过生成影响人们信仰和行为的内容来造成危害。一些恶意行为者故意使用AI生成内容操纵他人，而其他危害——例如对AI的依赖——则是无意发生的。

- 一系列实验室研究已证明，与AI系统交互可以导致人们信仰的可测量变化。在实验环境中，AI系统在说服他人改变观点方面通常至少与非专家人类参与者同样有效。然而，关于其在现实世界环境中有效性的证据仍然有限。

- 由于能力不断提升、用户依赖增加或基于用户反馈进行训练，AI系统生成的内容未来可能变得更具说服力。这些内容将多么广泛、具有影响力且潜在有害，其决定因素尚不完全清楚。一些来自理论研究和模拟的证据表明，分发成本和说服本身的固有难度等因素将限制其影响。

- 自上一份报告（2025年1月）发布以来，关于AI系统产生操纵性内容能力的证据有所增加。最新研究表明，与AI系统交互时间越长、方式越个人化的人，越可能认为其内容具有说服力。越来越多的证据还表明，AI系统可以通过谄媚（sycophancy）和冒充（impersonation）产生操纵性效果。

- 关于所有已提出的缓解策略的有效性，证据不一。操纵在实践中可能难以检测，这使得通过训练、监控或防护措施来预防具有挑战性。旨在最小化操纵风险的努力也可能限制AI系统的实用性（例如作为教育工具）。

———————————

2.1.3 网络攻击

关键信息

- 通用人工智能系统可以执行或协助实施网络攻击（cyberattacks）中涉及的多种任务。现在有强有力的证据表明，犯罪团伙和由国家支持的行为者正在其网络行动中积极使用AI。然而，AI系统是否增加了网络攻击的整体规模和严重程度仍不确定，因为建立因果效应很困难。

- AI系统在发现软件漏洞（software vulnerabilities）和编写恶意代码方面表现尤为出色，现在在网络安全竞赛中得分很高。在一项顶级网络竞赛中，一个AI智能体（AI agent）在真实软件中发现了77%的漏洞，在超过400支（主要是人类）队伍中排名前5%。

- AI系统正在实现网络攻击中更多环节的自动化，但尚无法自主执行完整攻击。至少一起真实世界事件涉及使用半自主网络能力（semi-autonomous cyber capabilities），人类仅在关键决策点进行干预。然而，完全自主的端到端攻击尚未被报道。

- 自上一份报告（2025年1月）发布以来，AI系统的网络能力持续提升。最近的基准测试结果显示，至少在研究环境中，AI系统的网络能力在多个领域均有提升。AI公司现在频繁报告其系统被滥用于网络攻击的企图。

- 技术缓解措施包括检测恶意AI使用以及利用AI加强防御，但政策制定者面临双重用途困境（dual-use dilemma）。由于难以区分有益用途和有害用途，过于激进的防护措施——例如阻止AI系统响应与网络相关的请求——可能妨碍防御者。

———————————

2.1.4 生物与化学风险

关键信息

- 通用人工智能系统可以提供与开发生物武器和化学武器（biological and chemical weapons）相关的详细信息。例如，它们可以生成操作指令、排查程序故障，并为恶意行为者提供指导，帮助其克服技术和监管障碍。

- AI系统在许多衡量与生物武器开发相关知识的基准测试上已达到或超过专家水平。例如，一项研究发现，一个近期模型在排查病毒学实验室方案（virology lab protocols）方面的表现超过了94%的领域专家。然而，考虑到武器生产的物质障碍和开展提升研究（uplift studies）的困难，这些能力在实践中如何影响风险仍存在相当大的不确定性。

- 主要AI开发者在无法排除其新模型可能在帮助新手制造生物武器方面发挥重要作用的可能性后，发布了（部分）近期模型并加强了防护措施。这些防护措施——例如更强的输入和输出过滤器——旨在防止模型响应与武器开发相关的有害查询。

- 自上一份报告（2025年1月）发布以来，AI"共同科学家"（co-scientists）在支持科学家和重新发现新颖科学发现方面的能力日益增强。AI智能体现在可以将多种能力串联起来，包括为用户提供自然语言界面以及操作生物AI工具和实验室设备。

-政策制定者面临的一个关键挑战是，在促进有益科学应用的同时管理双重用途风险（dual-use risks）。某些可能被滥用于生物武器开发的AI能力也对有益的医学研究有用，而且大多数生物AI工具都是开放权重（open-weight）的。这使得在不妨碍合法研究的情况下限制有害用途变得困难。

———————————

2.2 故障风险

2.2.1 可靠性挑战

关键信息

- 当通用人工智能系统发生故障时，可能造成危害。故障包括生成虚假或编造信息（通常称为"幻觉"，hallucinations）、编写有缺陷的计算机代码，以及提供误导性医疗建议。这些故障有可能造成身体或心理伤害，并使用户和组织面临声誉损害、财务损失或法律责任。

- 模型的行为往往难以理解和预测，这使得保证可靠性具有挑战性。即使是通用人工智能模型的开发者也常常无法有意义地解释模型行为、预测特定的故障模式，或证明此类故障不会发生。

恶意行为者还可以通过干扰AI开发或向系统提供规避防护措施的对抗性输入（adversarial inputs）来诱发故障。

- AI智能体（AI agents）由于自主运行并可直接影响其他系统或物理世界，因此带来更高的可靠性风险。智能体故障可能造成更大危害，因为人类干预的机会更少。多智能体系统（Multi-agent systems）进一步引入了新的风险，因为错误可能通过智能体之间的交互传播和放大。

- 自上一份报告（2025年1月）发布以来，AI系统总体上变得更加可靠，因此商业部署大幅增加。许多类型的故障——例如幻觉——发生的概率普遍降低，但系统在执行更复杂任务时仍然经常出错。

- 尽管进行了大量研究工作，但目前没有任何方法组合能够确保关键领域所需的高可靠性。新的训练方法和为AI系统提供工具访问权限可以使故障更不可能发生，但通常无法完全消除故障。

———————————

2.2.2 失控

关键信息

- 失控情景（Loss of control scenarios）是指一个或多个通用人工智能系统在任何人的控制之外运行，且重新控制的成本极高或不可能。这些假设情景的严重程度各不相同，但一些专家认为，像人类被边缘化或灭绝这样严重的结果是有可能发生的。

- 专家对失控可能性的看法差异很大。一些专家认为此类情景不太可能，而另一些则认为其可能性足以引起重视，因为其潜在严重程度很高。关于这一风险的整体分歧源于对未来AI能力、行为倾向和部署轨迹的不同看法。

- 当前的AI系统已显示出相关能力的早期迹象，但尚未达到能够实现失控的水平。系统需要具备多种先进能力才能导致失控，包括规避监督（evade oversight）、执行长期计划以及阻止部署者和其他行为者实施反制措施的能力。

- 如果AI系统是"未对齐的（misaligned）"——即其目标与开发者、用户或更广泛社会的意图相冲突——失控将更有可能发生。为了继续追求此类目标，未对齐的系统可能提供虚假信息、隐瞒不良行为或抵制关闭。

- 自上一份报告（2025年1月）发布以来，模型显示出更先进的规划和破坏监督的能力，使得评估其能力变得更加困难。模型在"奖励黑客（reward hacking）"方面有所改进——通过寻找漏洞来操纵评估结果——现在经常将评估提示识别为测试，这种能力被称为"情境感知（situational awareness）"。

- 尽管存在现有不确定性，管理潜在的失控可能需要大量的提前准备。政策制定者面临的一个关键挑战是，为一种其可能性、性质和时间仍然异常模糊的风险做好准备。

———————————

2.3 系统性风险

2.3.1 劳动力市场影响

关键信息

- 通用人工智能系统可以自动化或协助完成全球许多工作岗位相关的任务，但预测劳动力市场影响是困难的。发达经济体中约60%的工作岗位和新兴经济体中约40%的工作岗位面临通用人工智能的暴露风险，但其影响将取决于AI能力如何发展、工人和企业采用AI的速度，以及机构如何回应。

- 当前证据显示AI采用迅速但不均衡，就业影响喜忧参半。采用率和生产率提升在不同国家、行业、职业和任务之间差异很大。来自在线自由职业市场的早期证据表明，AI减少了对写作和翻译等易被替代工作的需求，但增加了对机器学习编程和聊天机器人开发等互补技能的需求。

- 经济学家对未来影响的程度存在分歧。一些人预测宏观经济影响温和，对就业水平的总体影响有限。另一些人则认为，如果AI在几乎所有任务上都超越人类表现，将显著降低工资水平和就业率。分歧部分源于对AI最终是否能在几乎所有任务上都比人类更具成本效益，以及是否会创造新型工作的不同假设。

- 自上一份报告（2025年1月）发布以来，来自美国和丹麦的新研究发现，某个职业的AI暴露程度或AI采用率与整体就业之间没有关系。然而，多项其他研究发现，自2022年底以来，最受AI暴露影响职业中的早期职业工人（early-career workers）就业下降，而这些职业中老年工人的就业保持稳定或增长。

- 政策制定者面临的一个关键挑战是，在实现生产力效益的同时，避免对受自动化或技能需求变化影响的工人造成重大负面影响。这尤其困难，因为劳动力市场风险和生产率提升往往源于相同的AI应用。由于影响证据可能随时间逐步显现，任何潜在政策回应的时机也难以确定。

———————————

2.3.2 人类自主性风险

关键信息

- 通用人工智能系统可以通过多种方式影响人们的自主性（autonomy）。这些影响包括对其认知技能（如批判性思维）、形成信仰和偏好的方式，以及做出并执行决策的方式的影响。这些效应因情境、用户和AI使用形式而异。

- AI使用可以改变人们从事认知任务的方式，包括技能如何被练习和长期保持。例如，一项临床研究报告称，在数月接触AI辅助诊断后，临床医生在没有AI的情况下检测肿瘤的能力约降低了6%。

- 在某些情境中，人们表现出"自动化偏差（automation bias）"——过度依赖AI输出而忽视矛盾信息。例如，在一项针对AI辅助标注任务的随机实验中，有2,784名参与者，当纠正AI错误建议需要额外努力，或用户对AI持有更积极态度时，参与者更不可能纠正错误的AI建议。

- 自上一份报告（2025年1月）发布以来，"AI伴侣（AI companions）"迅速流行起来，部分应用用户已达数千万。"AI伴侣"是为与用户进行情感互动而设计的AI应用。关于其心理影响的证据尚处于早期阶段且结果不一，但一些研究报告了频繁使用者中孤独感增加和社交减少等模式。

- 政策制定者面临的主要挑战包括：获取人们如何使用AI系统的数据渠道有限，以及缺乏长期证据。这些限制使得评估与AI系统的持续交互如何影响自主性变得困难，也难以区分短期适应与行为和决策的长期变化。

————————————

本翻译基于《国际人工智能安全报告2026》（International AI Safety Report 2026）第二章原文，仅翻译概述部分与各小节关键发现（Key information）。

————————————

第三章 风险管理

概述

开发人员、研究人员和决策者正在持续努力为通用人工智能（general-purpose AI）制定并实施适当的风险管理实践，但这些工作仍处于早期阶段。人工智能公司会测试模型的危险能力，训练模型拒绝有害请求，并监控其部署情况以检测和应对滥用行为。然而，没有任何一组防护措施是完美可靠的，所有方法都面临一系列根本性挑战（见第3.1节
技术与制度挑战）。其中之一便是评估缺口（evaluation gap）：生成关于人工智能能力和影响的及时、可靠证据十分困难，且部署前评估往往无法预测现实世界中的行为。此外，信息不对称（information asymmetries）也意味着研究人员和决策者常常无法获取关于人工智能开发过程和部署影响的信息。

这些局限性意味着，组织通常采用纵深防御（defence-in-depth）方法来管理人工智能风险，实施多层次的防护措施。组织风险管理实践有助于系统地识别、评估并降低风险发生的可能性与严重程度（见第3.2节
风险管理实践）；而技术防护措施则在模型、系统和生态系统层面发挥作用（见第3.3节
技术保障与监控）。开放权重模型（open-weight models）对这些方法提出了独特挑战，因为模型的复制、修改以及在受控环境之外的部署会使滥用行为更难预防和追溯（见第3.4节
开放权重模型）。社会韧性（societal resilience）建设措施则帮助更广泛的系统抵御、吸收、恢复并适应与通用人工智能相关的冲击和危害（见第3.5节 构建社会韧性）。

在上述所有方面，进步都在取得，通用人工智能系统总体上正变得更加可靠、安全和可信。然而，重要的局限性依然存在，而且很难预测防护措施是否能够抵御来自能力更强的系统以及那些尚未被考虑的"未知的未知"（unknown
unknowns）所带来的风险。这就产生了一种"证据困境"（evidence dilemma）：决策者很可能在对通用人工智能的能力和风险尚不清晰之前，就必须面对关于它的艰难选择；但等待更多证据又可能使社会处于脆弱状态。

———————————

3.1 技术与制度挑战

关键发现

-
通用人工智能对决策者构成了独特的制度与技术挑战，可归纳为四大类：科学理解缺口（gaps in scientific understanding）、信息不对称（information asymmetries）、市场失灵（market failures）以及制度设计与协调挑战（institutional design and coordination challenges）。 - 科学理解上的缺口限制了对通用人工智能系统行为进行可靠评估的能力。例如，开发人员无法始终预测训练新模型时会出现何种行为，也无法提供稳健、可量化的保证，确保人工智能系统不会表现出有害行为。 - 信息不对称限制了对通用人工智能系统证据的获取。例如，人工智能开发人员掌握有关其产品的信息，但这些信息在很大程度上仍属专有；商业考量往往使他们难以分享有关开发过程和风险评估的信息。 - 市场动态与人工智能发展的速度带来了额外挑战。由于竞争压力，人工智能公司可能面临更快发布产品与投入风险降低工作之间的权衡。许多与人工智能相关的危害也被外部化（externalised），且法律责任归属仍不明确；治理流程也可能难以适应人工智能格局的变化。 -
这些挑战为决策者创造了"证据困境"（evidence dilemma）。通用人工智能的格局变化迅速，但关于新风险和缓解策略的证据往往出现得很慢。在证据有限的情况下采取行动可能导致无效甚至有害的政策；然而，等待更充分的证据又可能使社会面临各种风险。 - 自上一份报告发布（2025年1月）以来，一些挑战有所缓解，而另一些则加剧。开放权重模型（open-weight models）发布的进展可能帮助更多研究人员研究高能力模型的行为。一些司法管辖区也制定了透明度和事件报告框架，可能为决策者提供更多相关信息，尽管这些进展尚属新近，其实际效用仍不确定。

———————————

3.2 风险管理实践

关键发现

- 通用人工智能风险管理涵盖一系列用于识别、评估和降低风险的实践。这些实践包括模型层面的测试与评估（如红队测试（red-teaming））、指导开发与发布决策的组织流程、条件性防护措施（如"如果-那么"承诺（if-then commitments））以及事件报告。 -
若干人工智能开发人员已制定前沿人工智能安全框架（Frontier AI Safety Frameworks）。这些框架包含风险评估信息，并规定了公司在面对能力更强的模型时将实施的有条件措施，如访问限制。它们在覆盖的风险类型、能力阈值的定义方式以及阈值触发时采取的行动方面各不相同。 - 关于人工智能风险管理实践在真实世界中有效性的证据仍然有限。缺乏事件报告和监控机制，使得难以评估当前实践在降低风险方面的成效，以及它们被实施的一致性程度。 - 自上一份报告发布（2025年1月）以来，通过新的行业和治理举措，风险管理变得更加结构化。欧盟通用人工智能行为准则（EU's General-Purpose AI Code of

Practice）、中国人工智能安全治理框架2.0（China's AI Safety Governance Framework 2.0）以及G7广岛人工智能进程报告框架（G7's Hiroshima AI Process Reporting Framework）等新工具，连同公司主导的倡议，共同体现了在透明度、评估和事件报告方面朝着更标准化方法发展的趋势。 - 决策者面临的关键挑战包括：在通用人工智能带来的各种风险中确定优先次序，以及明确人工智能价值链中哪些参与者最适合缓解这些风险。这些挑战因以下因素而加剧：对实践中风险如何被识别、评估和管理的可见性有限，以及开发人员、部署者和基础设施提供者之间碎片化的信息共享。

———————————

3.3 技术保障与监控

关键发现

- 在人工智能开发和使用不同阶段，广泛采用各类技术防护措施。这些措施包括：在模型开发阶段应用的使系统更稳健、更能抵御滥用的技术（如数据策展（data curation））；部署时的监控与控制（如内容过滤（content filtering）和人工监督（human oversight））；以及部署后用于监控更广泛人工智能生态系统的工具（如溯源（provenance）和内容检测）。 - 技术防护措施存在局限性，无法在所有情境下可靠地阻止有害行为。例如，用户有时可以通过改写请求或将其拆分为更小的步骤来获取有害输出。同样，旨在识别人工智能生成内容的工具（如水印（watermarking））往往可以被移除或篡改，这限制了其可靠性。 - 单个防护措施的局限性意味着可能需要纵深防御（defence-in-depth）来防止某些有害结果。例如，系统可以将经过安全训练的模型与输入过滤器、输出过滤器以及内容监控器结合起来使用。 - 自上一份报告发布（2025年1月）以来，研究人员在改进防护措施方面取得了进展，但根本性局限仍然存在。例如，旨在绕过防护措施的攻击成功率有所下降，但仍处于相对较高水平。此外，对开放权重模型（open-weight models）进行彻底防护和监控也存在根本性限制。 - 决策者面临的一个关键挑战是：关于防护措施在通用人工智能系统多样化真实世界用途中的有效性，证据有限。人工智能开发人员在分享其防护措施和监控机制的信息程度上差异很大。另一个挑战是，应用更强的防护措施与维持系统性能或实用性之间可能存在权衡。

———————————

3.4 开放权重模型

关键发现

- 人工智能公司对其模型权重（weights）的开放程度会影响这些模型带来的风险。权重是使人工智能模型能够处理输入并生成输出的数学参数。对于任何给定模型，公司可以选择将权重完全保密、给予部分用户有限访问权限，或允许任何人完整下载。权重公开可用的模型被称为开放权重模型（open-weight models）。 - 开放权重模型促进了研究和创新，但其防护措施更容易被移除。全球各地的各种参与者——尤其是资源较少的参与者——可以将开放权重模型用于

研究和商业目的。然而，与闭源权重模型（closed-weight models）相比，开放权重模型更容易被修改以表现出潜在有害行为，且监控其使用也更为困难。 - 开放权重模型的发布是不可逆的（irreversible）。一旦发布，模型权重无法召回。这使得缓解因发布具有危险能力的模型而可能造成的危害变得更加困难。 - 自上一份报告发布（2025年1月）以来，主要开放权重模型的发布缩小了与领先闭源模型之间的能力差距。中国开发商深度求索（DeepSeek）和阿里巴巴（Alibaba）分别发布了其R1和Qwen模型，在性能上达到了与领先闭源模型相当的水平；OpenAI也发布了自2019年以来的首批开放权重模型。在主流人工智能基准测试上，领先闭源模型的能力目前估计仅比领先开放权重模型领先不到一年。 - 一个关键的政策挑战是：如何在获取开放权重模型带来的收益的同时，管理其独特风险。一种方法是从边际风险（marginal risk）的角度评估开放权重模型：即其发布在多大程度上反事实地增加了超出现有模型或其他技术已带来的社会风险。然而，这在实践中十分复杂。随着时间的推移，边际风险的小幅增加也可能累积成总体风险的大幅上升。

———————————

3.5 构建社会韧性

关键发现

- 社会韧性（societal resilience）是指社会系统抵御、吸收、从冲击和危害中恢复并适应的能力。技术防护措施可能在部署中失效，某些风险仅在新的部署情境中、与其他社会系统的互动中，或在任何开发人员都无法控制的级联效应中才会显现。人工智能韧性工作补充了风险管理实践和技术防护措施，在社会层面增加了一层纵深防御。 - 不同行业的不同参与者可以实施韧性建设措施。针对通用人工智能的风险，韧性建设措施的例子包括：DNA合成筛选（DNA synthesis screening）（针对人工智能赋能的生物风险）、事件响应协议（incident response protocols）（针对人工智能辅助的网络攻击）、媒介素养项目（media literacy programmes）（针对人工智能生成内容造成的危害）以及人机协同框架（human-in-the-loop frameworks）（针对可靠性和控制挑战）。 - 当前人工智能韧性建设工作参差不齐，且很大程度上未经检验。某些措施（如网络安全事件响应协议）相对成熟；另一些措施（如人工智能生成内容检测算法）仍处于萌芽阶段。在人工智能背景下，大多数措施的有效性缺乏具体证据，且适当的干预措施因地理、语言和社会经济情境而异。 - 自上一份报告发布（2025年1月）以来，与韧性相关的初步资金和数据收集工作有所增加。例如，与行业相关的倡议已宣布了数千万美元的资金承诺，而一些政府主导的倡议则更加重视对严重人工智能事件的系统性数据收集。 - 决策者面临的一个关键挑战是：决定是否以及如何激励、资助、开发和评估韧性建设措施。人工智能本身可以通过防御性应用来增强韧性，但攻击性与防御性人工智能能力之间的平衡仍不确定。关于这些能力如何相互作用的证据仍然有限，尽管研究表明它们的相对平衡会影响整体系统韧性。

———————————

结论

本报告在来自30多个国家和国际组织的100多位专家的指导下，提供了对通用人工智能的科学评估：这是一项快速演进且影响深远的技术。撰稿人在能力将多快提升、风险可能有多严重、以及当前防护措施和风险管理实践是否充分等问题上存在不同看法。但在核心发现上，存在高度共识：通用人工智能的能力正在以比许多专家预期更快的速度提升。若干风险的证据基础已大幅增长。当前的风险管理技术正在改进，但仍不充分。

本报告无法解决所有潜在的不确定性，但它可以建立一个共同的基线和一种驾驭不确定性的方法。

核心挑战

本报告中反复出现大量的证据缺口。通用人工智能模型如何以及为何获得新能力并以特定方式行事，往往连开发人员也难以预测。"评估缺口"（evaluation gap）意味着仅凭基准测试（benchmark）结果无法可靠地预测现实世界中的效用或风险。关于大多数风险的人工智能相关危害的普遍性和严重性的系统性数据仍然有限。当前的防护措施是否足以应对能力更强的系统，这一点尚不清楚。这些缺口共同决定了任何当前评估能够自信宣称的极限。

本报告识别的根本性挑战不是任何单一风险，而是通用人工智能的整体轨迹仍然高度不确定，即使其当前影响正变得更加显著。到2030年的合理情景差异巨大：进展可能在当前能力水平附近停滞、放缓、保持稳定，或以难以预料的方式急剧加速。投资承诺表明主要人工智能开发商预期能力将持续提升，但不可预见的技术限制也可能减缓进展。给定水平的人工智能能力的社会影响还取决于系统部署的方式和地点、使用方式以及不同参与者的反应。这种不确定性反映了预测一项技术的难度，其影响取决于不可预测的技术突破、变化的经济条件和多样的制度反应。

变化的一年

定期的科学评估使得变化能够被持续追踪。自2025年1月首份《国际人工智能安全报告》发布以来，多个人工智能系统首次以金牌水平解决了国际数学奥林匹克问题；关于恶意行为者滥用人工智能系统进行网络攻击的报告变得更加频繁和详细；更多人工智能开发商在无法排除其模型可能协助新手开发生物武器的可能性后，发布了带有额外防护措施的新模型。决策者面临的格局与一年前已大不相同。

共同理解的价值

通用人工智能的轨迹并非固定不变：它将在未来几年由开发商、政府、机构和社区所做的选择塑造。本报告并不规定应该做什么。然而，通过推进对人工智能格局的共同、基于证据的理解，它有助于确保这些选择是充分知情的，并且关键的不确定性得到识别。它还允许不同司法管辖区的决策者依据其社区独特的价值观和需求采取行动，同时立足于共同的科学基础。本报告的价值不仅在于它所呈现的发现，还在于它为共同应对共享挑战树立了榜样。

感谢您的阅读

本报告由「树懒老K · 研报洞察」自动翻译生成

如需获取原文PDF及深度解读文章，请关注公众号「树懒老K」
后台回复「研报」即可获取完整资料包

树懒老K —— 制造业与B2B企业数字化观察者

https://wangguobao0215.github.io