AI ACTION SUMMIT

国际人工智能安全报告

先进人工智能安全国际科学报告

2025 年 1 月

贡献者

主席

Yoshua Bengio教授, 蒙特利尔大学/ Mila - 魁北克人工智能研究所

专家顾问小组

该国际小组成员由下列 30 个国家的政府、联合国、欧盟和经合组织提名。

澳大利亚：Bronwyn Fox，新南威尔士大学

巴西：André Carlos Ponce de Leon Ferreira de Carvalho，圣保罗大学数学与计算机科学研究所

加拿大：Mona Nemer，加拿大首席科学顾问

智利：Raquel Pezoa Rivera，费德里科圣玛丽亚理工大学

中国：曾毅，中国科学院

欧盟：Juha Heikkilä，欧洲人工智能办公室

法国：Guillaume Avrin，法國国家人工智能协调机构

德国：Antonio Krüger，德国人工智能研究中心

印度：Balaraman Ravindran，印度理工学院马德拉斯分校瓦德瓦尼数据科学与人工智能学院

印度尼西亚：Hammam Riza，人工智能合作研究与工业创新（KORIKA）

爱尔兰：Ciarán Seoighe，爱尔兰研究中心

以色列：Ziv Katzir，以色列创新局

意大利：Andrea Monti，意大利部长理事会主席国数字化转型副国务卿的法律专家

日本：北野宏明，索尼集团公司

肯尼亚：Nusu Mwamanzi，信息通信技术和数字经济部

沙特阿拉伯王国：Fahad Albalawi，沙特数据和人工智能管理局局长

墨西哥：José Ramón López Portillo，LobsterTel

荷兰：Haroon Sheikh，荷兰政府政策科学委员会

新西兰：Gill Jolly，新西兰商业、创新和就业部

尼日利亚：Olubunmi Ajala，尼日利亚通信、创新和数字经济部

经济合作与发展组织：Jerry Sheehan，科学、技术和创新局主任

菲律宾：Dominic Vincent Ligot，CirroLytix

韩国：Kyoung Mu Lee，首尔国立大学电子与计算机工程系

卢旺达：Crystal Rugege，卢旺达国家人工智能和创新政策中心

新加坡：Denise Wong，新加坡资讯通信媒体发展局数据创新与保护小组

西班牙：Nuria Oliver，ELLIS Alicante

瑞士：Christian Busch，瑞士联邦经济、教育与研究部

土耳其：Ahmet Halit Hatip，土耳其工业和技术部

乌克兰：Oleksii Molchanovskyi，乌克兰人工智能开发专家委员会

阿联酋：Marwan Alserkal，阿联酋内阁事务部、总理办公室

英国：Chris Johnson，英国科学、创新和技术部首席科学顾问

联合国：Amandeep Singh Gill，联合国数字和新兴技术副秘书长兼秘书长技术问题特使

美国：Saif M. Khan， 美国商务部

科研首席专家

Sören Mindermann，Mila - 魁北克人工智能研究所

首席撰稿人

Daniel Privitera，KIRA 中心

撰稿团队

Tamay Besiroglu，Epoch AI

Rishi Bommasani，斯坦福大学

Stephen Casper，麻省理工学院

Yejin Choi，斯坦福大学

Philip Fox，KIRA 中心

Ben Garfinkel，牛津大学

Danielle Goldfarb，Mila - 魁北克人工智能研究所

Hoda Heidari，卡内基梅隆大学

Anson Ho，Epoch AI

Sayash Kapoor，普林斯顿大学

Leila Khalatbari ，香港科技大学

Shayne Longpre，麻省理工学院

Sam Manning，人工智能治理中心（Centre for the Governance of AI）

Vasilios Mavroudis，艾伦图灵研究所

Mantas Mazeika，伊利诺伊大学香槟分校

Julian Michael，纽约大学

Jessica Newman，加州大学伯克利分校

吴君仪，安远AI

Chinasa T. Okolo，布鲁金斯学会

Deborah Raji，加州大学伯克利分校

Girish Sastry，独立人士

Elizabeth Seger（通才作家），Demos

Theodora Skeadas, Humane Intelligence

Tobin South，麻省理工学院

Emma Strubell，卡内基梅隆大学

Florian Tramèr，苏黎世联邦理工学院

Lucia Velasco，马斯特里赫特大学

Nicole Wheeler，伯明翰大学

高级顾问

Daron Acemoglu，麻省理工学院

Olubayo Adekanmbi，入职 EqualyzAI
之前担任高级顾问

David Dalrymple，英国高级研究与发明局

Thomas G. Dietterich，俄勒冈州立大学

Edward W. Felton，**普林**斯顿大学

Pascale Fung，入职 Meta 之前担任高级顾问

Pierre-Olivier Gourinchas，国际货币基金组织
研究部

Fredrik Heintz，林雪平大学

Geoffrey Hinton，多伦多大学

Nick Jennings,，拉夫堡大学

Andreas Krause，苏黎世联邦理工学院

Susan Leavy，都柏林大学学院

Percy Liang，**斯**坦福大学

Teresa Ludermir，伯南布哥联邦大学

Vidushi Marda，**人工智能Collaborative 控**股有
限公司

Helen Margetts，牛津大学

John McDermid，约克大学

Jane Munga，卡内基国际和平基金会

Arvind Narayanan，普林斯顿大学

Alondra Nelson，高等研究院

Clara Neppel，IEEE

Alice Oh，韩国科学技术研究院计算机系

Gopal Ramchurn，Responsible 人工智能UK

Stuart Russell，加州大学伯克利分校

Marietje Schaake，斯坦福大学

Bernhard Schölkopf，**蒂宾**根埃利斯研究所

Dawn Song，**加**州大学伯克利分校

Alvaro Soto，智利天主教大学

Lee Tiedrich，杜克大学

Gaël Varoquaux，Inria

Andrew Yao，**清**华大学交叉信息科学研究院

Ya-Qin Zhang，**清华**大学

秘书处

英国人工智能安全研究所

Baran Acar

Ben Clifford

Lambrini Das

Claire Dennis

Freya Hempleman

Hannah Merchant

Rian Overy

Ben Snodin

Mila — 魁北克人工智能研究所

Jonathan Barry

Benjamin Prud'homme

致谢

民间社团和行业评论者

民间社团： Ada Lovelace Institute, AI Forum New Zealand / Te Kāhui Atamai Iahiko o Aotearoa, Australia's Temporary AI Expert Group, Carnegie Endowment for International Peace, Center for Law and Innovation / Certa Foundation, Centre for the Governance of AI, Chief Justice Meir Shamgar Center for Digital Law and Innovation, Eon Institute, Gradient Institute, Israel Democracy Institute, Mozilla Foundation, Old Ways New, RAND, SaferAI, The Centre for Long-Term Resilience, The Future Society, The Alan Turing Institute, The Royal Society, Türkiye Artificial Intelligence Policies Association.

行业： Advai, Anthropic, Cohere, Deloitte Consulting USA and Deloitte Consulting UK, G42, Google DeepMind, Harmony Intelligence, Hugging Face, IBM, Lelapa AI, Meta, Microsoft, Shutterstock, 智谱AI.

特别感谢

秘书处感谢 Angie Abdilla、安远AI、Nitarshan Rajkumar、Geoffrey Irving、Shannon Vallor、Rebecca Finlay 和 Andrew Strait 的支持、评论和反馈。

免责声明

本报告并不代表主席、撰写团队或顾问小组中任何特定个人的观点，也不代表支持撰写本报告的任何政府人员的观点。本报告是对关于先进人工智能能力和风险现有研究的汇总。报告主席对本报告负有最终责任，并从始至终监督本报告的制定。

研究系列编号：DSIT 2025/001

关于本报告

- **这是首份《先进人工智能安全国际科学报告》。**继 2024 年 5 月发布中期报告后，来自 30 个国家提名的国际专家顾问小组、经济合作与发展组织（OECD）、欧盟（EU）和联合国（UN）的 96 位人工智能（AI）专家组共同参与完成了这份首份完整报告。该报告旨在提供科学信息，以支持明智的政策制定，但并不提出具体的政策建议。
- **本报告是独立专家的工作成果。**本报告由独立专家在主席的指导下完成，专家组对报告内容拥有完全的决定权。
- **虽然本报告关注的是人工智能风险和人工智能安全，但人工智能也为人类、企业和社会带来了许多潜在的益处。**人工智能有很多种类型，每种类型都有不同的益处和风险。多数情况下，在大多数应用中，人工智能可以帮助个人和组织提高效率。但只有妥善管理其风险，世界各地的人们才能安全地、充分地享受人工智能的诸多潜在益处。本报告重点在于识别这些风险，并评估降低风险的方法。它并不旨在全面评估人工智能可能对社会产生的所有影响，包括其众多潜在益处。
- **本报告重点关注通用型人工智能。**报告将重点放在近年来发展尤为迅速且相关风险研究和理解相对较少的一类人工智能上，即通用型人工智能，也就是能够执行多种任务的人工智能。本报告的分析聚焦于撰写时最先进的通用型人工智能系统，以及未来可能更强大的系统。
- **本报告总结了三个核心问题的科学证据：**通用型人工智能能做什么？通用型人工智能存在哪些风险？针对这些风险有哪些缓解技术？
- **事关重大。**作为本报告的撰稿专家，我们在有关通用型人工智能的能力、风险和风险缓解措施等方面仍存在诸多分歧，无论大小。但我们认为本报告对于增强我们对这项技术及其潜在风险的集体理解至关重要。我们希望这份报告能够帮助国际社会就通用型人工智能达成更广泛的共识，更有效地缓解其风险，从而让人们能够安全地享受其众多潜在益处。我们期待着继续这项努力。

本报告撰写后人工智能的最新进展：主席的话

在本报告编写期结束（2024年12月5日）至2025年1月发布期间，出现了重要进展。人工智能公司OpenAI分享了其新型人工智能模型o3的早期测试结果。这些结果显示，在人工智能领域最具挑战性的编程、抽象推理和科学推理等多项测试中，o3的表现显著优于此前所有模型。在部分测试中，o3的性能超越了众多（但非全部）人类专家。此外，o3在一项关键的抽象推理测试中取得突破性进展，这项成就此前被包括本人在内的众多专家认为难以实现。然而，截至撰稿时，关于其在现实世界中的应用能力，特别是在处理开放性任务方面的表现，尚无公开信息。

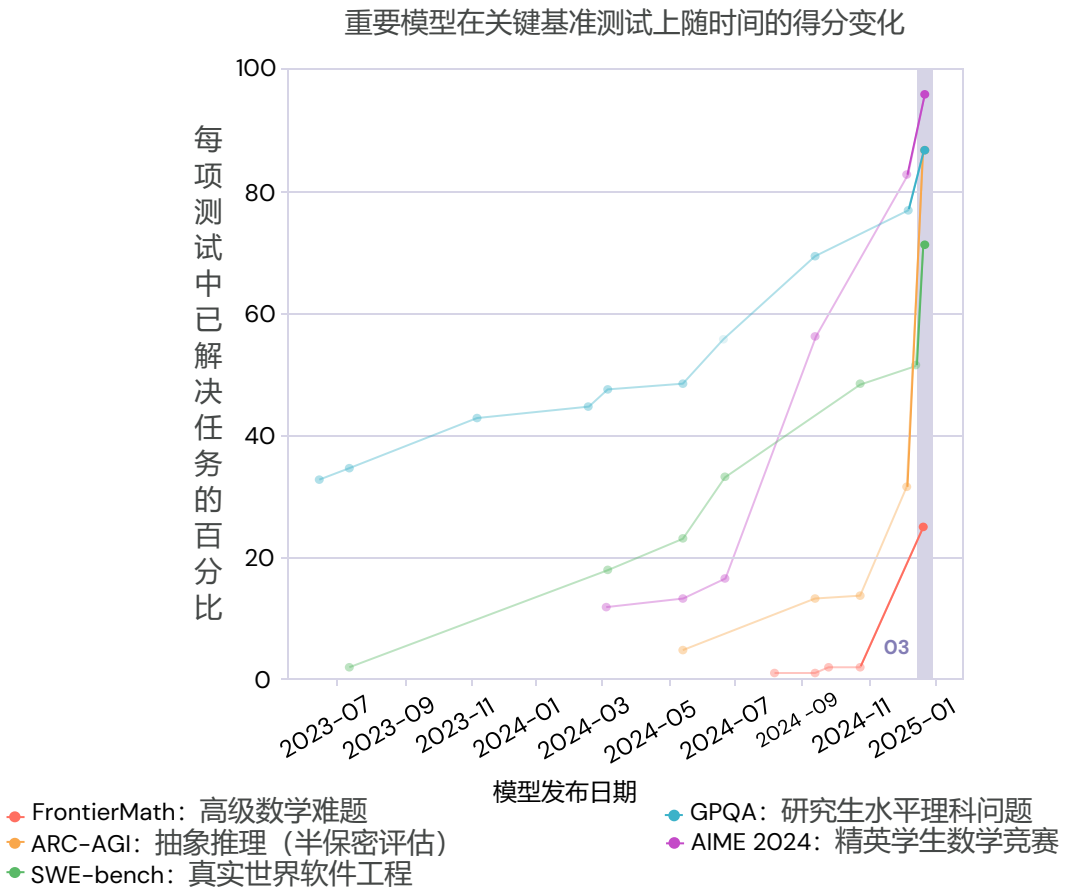

图 0.1： 2023年6月至2024年12月期间通用人工智能模型在关键基准测试中的得分情况。与现有最高水平（阴影区域）相比，o3显示出显著的性能提升。这些基准测试代表该领域在编程、抽象推理和科学推理方面最具挑战性的评估。图中标注了尚未发布的o3的公布日期以及其他模型的发布日期。较新的人工智能模型（包括o3）得益于改进的脚手架和测试时更多的计算量。

资料来源： Anthropic, 2024; Chollet, 2024; Chollet et al., 2025; Epoch AI, 2024; Glazer et al. 2024; OpenAI, 2024a; OpenAI, 2024b; Jimenez et al., 2024; Jimenez et al., 2025.

o3的测试结果表明，人工智能能力的发展速度可能持续保持高位甚至加速。具体而言，这些结果显示，通过为模型提供更多计算能力来解决特定问题（"推理扩展"）可能有助于突破既有限制。一般而言，推理扩展会增加模型的使用成本。但正如DeepSeek公司于2025年1月发布的另一个重要模型R1所显示的，研究人员在降低这些成本方面取得了成功。总体而言，推理扩展可能助力人工智能开发者在未来取得更大突破。o3的结果还凸显出我们需要更好地理解人工智能开发者对AI的日益增长的使用将如何影响人工智能自身的发展速度。

o3所体现的发展趋势可能对人工智能风险产生深远影响。科学和编程能力的提升此前已为网络攻击和生物威胁等风险提供了更多证据。o3的结果还涉及潜在的劳动力市场影响、失控风险和能源使用等问题。然而，o3的功能也可用于防范故障和恶意使用。总体而言，读者在阅读本报告的风险评估时应当认识到，自报告撰写以来，人工智能的能力已有提升。不过，目前尚无关于o3实际影响的证据，也没有信息能够确认或排除重大新风险和/或直接风险。

o3 结果所显示的能力提升以及我们对人工智能风险影响认知的局限性，突显了本报告指出的政策制定者面临的关键挑战：他们往往需要在缺乏大量科学依据的情况下，权衡即将到来的人工智能进步带来的潜在利益与风险。尽管如此，在未来数周乃至数月内，就o3所预示的发展趋势对安全和安保的影响收集证据将成为人工智能研究的当务之急。

本报告的主要发现

- **通用型人工智能（General-purpose AI，本报告关注的人工智能类型）的能力近年来迅速提升，并在近几个月内得到了进一步增强。**几年前，最先进的大语言模型（Large language model，LLM）很少能生成连贯的段落。如今，通用型人工智能能够编写计算机程序、生成的逼真的定制图像，并进行长篇幅的开放式对话。自《先进人工智能安全国际科学报告：中期报告》（“中期报告”）（2024 年 5 月）发布以来，新模型在科学推理和编程测试中的表现有了显著提升。
- **许多公司正在投资开发通用型人工智能自主体（General-purpose AI agents），将其作为进一步发展的潜在方向。**人工智能自主体（AI Agent）是能够自主行动、规划和分配任务以实现目标的通用型人工智能系统，几乎无需人类监督。复杂的人工智能自主体将能够使用计算机完成比现有系统更长的项目，从而带来更多的益处，同时也伴随着更多的风险。
- **未来数月乃至数年里，人工智能能力的进一步提升可能会非常缓慢，也可能极其迅速。**其进展取决于企业能否迅速部署更多的数据和计算能力来训练新模型，以及这种“扩展”（Scaling）模型的方式能否克服其当前的局限性。最近的研究表明，至少在未来几年内，迅速扩大模型规模在物理上仍然具有可行性。但重大的能力提升可能还需要其他因素：例如，难以预测的新研究突破，或者公司最近采用的新型扩展方法取得成功。
- **通用型人工智能的一些危害已得到充分证实。**这些危害包括诈骗、未经同意的隐私图像（Non-consensual intimate imagery，NCII）和儿童性虐待材料（Child sexual abuse material，CSAM），对某些人群或某些观点存在偏见的模型输出、可靠性问题，以及隐私侵犯等问题。研究人员已经开发出针对上述问题的缓解技术，但迄今为止，还没有任何技术组合能够完全解决此类问题。自中期报告（2024 年 5 月）发布以来，与通用型人工智能系统相关歧视的新证据揭示了更为隐蔽的偏见形式。
- **随着通用型人工智能的能力不断增强，更多风险的迹象也逐渐浮现。**这些风险包括大规模劳动力市场冲击、人工智能支持的引发的黑客攻击或生物攻击、以及社会对通用型人工智能失去控制等。专家们对这些风险的现有证据的解读不一：有人认为此类风险在数十年之后才会出现，而另一些人则认为通用型人工智能可能在未来数年内就会对社会造成大规模的危害。通用型人工智能能力的最新进展——特别是在科学推理和编程测试方面——为人工智能支持的黑客攻击和生物攻击等潜在风险提供了新[1]证据，这促使一家大型人工智能公司将其对最佳模型的生物风险评估等级从“低”提高到“中”。

[1] 请参阅主席在撰写本报告后有关对人工智能最新进展的看法。

- **风险管理技术尚处于起步阶段，但取得进展是有可能的。**开发者可以采用、监管机构可以要求使用各种技术方法来评测和降低通用型人工智能带来的风险，但这些手段均有局限性。例如，当前用于解释通用型人工智能模型为何产生特定输出的可解释性技术仍存在严重不足。不过，研究人员正在努力克服这些局限性。此外，研究人员和政策制定者正越来越多地尝试标准化风险管理方法，并开展国际协调。
- **通用型人工智能发展的速度和不可预测性给政策制定者带来了"证据困境"。**鉴于有时人工智能发展迅速且出人意料，政策制定者往往需要在缺乏大量科学证据的情况下权衡即将到来的人工智能发展的潜在益处和风险。在此过程中，他们面临着两难境地。一方面，基于有限证据采取的预防性风险缓解措施结果可能是无效或没有必要的。另一方面，如果等待更有力的证据来证明即将出现的风险，可能会使社会措手不及，甚至导致无法缓解风险，例如，如果人工智能能力突然大幅提升，其相关风险也随之而来。公司和政府正在开发早期预警系统和风险管理框架，以减轻这种困境。其中一些系统在出现新的风险证据时会触发特定的缓解措施，而另一些则要求开发者在发布新模型之前提供安全证据。
- **研究人员普遍认为，以下问题若能取得进展将大有裨益：**未来几年，通用型人工智能的能力将以何种速度提升，研究人员如何可靠地衡量这一进展？触发风险缓解措施的合理风险阈值是多少？政策制定者如何最有效地获取与通用型人工智能相关的公共安全信息？研究人员、科技公司和政府如何可靠地评估通用型人工智能的开发和部署的风险？通用型人工智能模型内部如何运作？如何设计通用型人工智能以使其行为可靠？
- **人工智能并非偶然降临：人们所做的选择将决定其未来。**未来通用型人工智能技术的未来充满不确定性，即便在不久的将来，也似乎存在多种可能的发展路径，既有非常积极的结果，也有非常消极的后果。这种不确定性可能引发宿命论，让人觉得人工智能是某种降临在我们身上的事物。但决定我们将走上哪条道路的，将是社会和政府如何应对这种不确定性所做的决策。本报告旨在促进对这些决策展开建设性且基于证据的讨论。

执行摘要

本报告的目的

本报告汇总了科学界当前对通用型人工智能（即能够执行各种任务的人工智能）的科学理解，重点在于理解和管控其风险。

本报告总结了通用型人工智能安全性的科学证据。旨在帮助建立国际社会对先进人工智能风险及其缓解方法的共同理解。为实现这一目标，本报告重点关注通用型人工智能（即可以执行各种任务的人工智能），因为此类人工智能近年来发展尤为迅速，并已被科技公司广泛应用于各种消费和商业用途。本报告汇总了科学界当前对通用型人工智能的理解，重点关注理解和管控其风险。

在快速发展的背景下，通用型人工智能的研究目前正处于科学发现的时期，而且在许多情况下尚未成为科学内容。本报告简要介绍了当前科学界对通用型人工智能及其风险的理解。这包括确定科学共识的领域以及当前科学理解中存在不同观点或知识空白的领域。

只有妥善管理通用型人工智能的风险，全球人类才能安全、充分地享受其潜在益处。本报告在于重点识别这些风险，并评估用于评测和降低风险的技术方法，包括利用通用型人工智能自身来降低风险的方法。本报告无意全面评估通用型人工智能可能带来的所有社会影响。最值得注意的是，通用人工智能当前和未来潜在的益处虽然巨大，但超出了本报告的范围。全面的政策制定需要同时考虑通用型人工智能的潜在益处和本报告中所涵盖的风险。此外，还需要考虑到其他类型的人工智能与当前通用型人工智能相比具有不同的风险/收益特征。

报告的三个主要部分总结了三个核心问题的科学证据： 通用型人工智能能做什么？通用型人工智能存在哪些风险？ 以及针对这些风险有哪些缓解技术？

第 1 部分 – 通用型人工智能的能力：通用型人工智能现在和将来的用途？

近年来，通用型人工智能的能力得到了迅速提升，其进一步的发展速度快慢并无定论。

人工智能用途是衡量其所带来的诸多风险的关键因素，并且根据许多指标衡量来看，通用型人工智能的能力一直在迅速提升。五年前，领先的通用**型**人工智能语言模型几乎无法生成连贯的文本段落。如今，一些通用型人工智能模型可以就广泛的话题进行对话、编写计算机程序或根据描述生成逼真的短视频。然而，可靠地估计和描述通用型人工智能的能力在技术上颇具挑战性。

近年来，人工智能开发者通过"扩展"（Scaling）迅速提升了通用型人工智能的能力。他们不断增加用于训练新模型的资源（此行为通常被称为"扩展"）并改进现有方法以更有效地利用此类资源。例如，根据最近的估测，最先进的人工智能模型用于训练的计算资源（"算力"）每年增加约 4 倍，训练数据集大小每年增加约 2.5 倍。

通用型人工智能未来的发展速度对于管理新兴风险具有重大影响，但专家们对未来数月乃至数年的预期仍存在分歧。 专家们对通用型人工智能能力的发展速度可能缓慢、迅速或极其迅速持有不同的意见。

专家们对于未来进步的速度存在分歧，原因在于他们对于进一步"扩展"的前景看法不一。而各公司正在探索一种新的扩展类型，这或许能进一步提升能力。尽管扩展通常可以克服以前系统的局限性，但专家们对于其解决当今系统剩余局限性的潜力存在分歧，例如在物理世界中行动的不可靠性以及在计算机上执行复杂任务的能力不足。近几个月来，一种新型的扩展类型展现出进一步提升其能力的潜力：人工智能公司不再仅仅关注用于训练模型的资源扩展，他们对还越来越对"推理扩展"（inference scaling）感兴趣——即让已经训练过的模型使用更多的计算来解决给定的问题，例如改进自身的解决方案，或者编写所谓的"思维链"，将问题分解为更简单的步骤。

有数家开发通用型人工智能的领先公司正在押注于"扩展"以持续推动性能提升。如果最近的趋势继续下去，到 2026 年底，部分通用型人工智能模型将使用比 2023 年计算量最大的模型大约 100 倍的训练算力进行训练，到 2030 年将增长到 10,000 倍，同时算法也将实现以更少的可用计算量获得更强的能力。除了训练资源可能的这种规模增长外，诸如推理规模扩大以及利用模型生成训练数据等近期趋势，可能意味着总体上会使用更多的计算量。然而，进一步快速增加数据和计算还存在一些潜在瓶颈，例如数据、人工智能芯片、资本和本地能源容量的可及性。开发通用型人工智能的公司正在努力克服此类潜在[2]的瓶颈。

[2†] 请参阅主席在撰写本报告后有关对人工智能最新进展的看法。

自《先进人工智能安全国际科学报告：中期报告》（“中期报告”）（2024 年 5 月）发布以来，通用型人工智能在某些科学推理和编程测试和竞赛中已经达到了专家级的表现，各家公司一直在努力开发自主智能体。科学和编程领域的进步得益于诸如编写长“思维链”（Chain-of-Thought）之类的推理扩展技术。新的研究表明，进一步扩展此类方法，例如允许模型通过编写比当前的模型更长的思维链来分析问题，可能会在更注重推理的领域（比如科学、软件工程和规划）取得进一步的进展。除了这一趋势之外，各公司还在大力开发更先进的通用型人工智能，这些自主体能够自主规划和行动以实现既定目标。最后，使用给定能力水平的通用型人工智能的市场价格急剧下降，使得这项技术更加广泛地普及和使用。

本报告主要关注人工智能发展的技术方面，但通用型人工智能的发展速度并不是纯粹的技术问题。未来发展的速度还将取决于非技术因素，可能包括政府对 人工智能的监管方式。本报告并未讨论不同的监管方法如何影响通用型人工智能的发展和应用速度。

第 2 部分 – 风险：通用型人工智能有哪些风险？

通用型人工智能的一些危害已得到充分证实。随着通用型人工智能的能力不断增强，更多风险的证据也逐渐浮现。

本报告将通用型人工智能风险分为三类：恶意使用风险、故障风险和系统风险。每个类别都包含已经出现的风险以及未来数年可能出现的风险。

恶意使用的风险*：恶意行为者可以使用通用型人工智能对个人、组织或社会造成伤害。* 每个类别都包含已经出现的风险以及未来几年可能出现的风险。恶意使用的形式包括：

- **通过发布虚假内容对个人造成伤害**：目前，恶意行为者能够使用通用型人工智能来生成有针对性地伤害个人的虚假内容。这些恶意用途包括非经同意的“深度伪造”的色情内容和人工智能生成的儿童性虐待材料、通过声音模仿进行金融欺诈、敲诈勒索、破坏个人和职业声誉以及心理虐待。然而，尽管有关人工智能生成的虚假内容造成危害的事件报告屡见不鲜，但仍然缺乏有关此类事件发生频率的可靠统计数据。

- **操纵公众舆论**：通用型人工智能使得大规模生成具有说服力的内容变得更加容易。那些试图操纵公众舆论的人，例如意图左右政治结果的行为者，恰可以借助于此。然而，关于此类行为的普遍性和有效性的证据仍然有限。虽然诸如内容水印之类的技术对策有用，但通常会被稍有技术的恶意行为者规避。
- **网络攻击：**借助通用型人工智能，不同技能水平的恶意行为者可更轻松或更快地发动网络攻击。当前的人工智能系统已经展现出在低复杂度和中等复杂度网络安全任务方面的能力，有国家支持的攻击者正在积极利用人工智能来侦察目标系统。新的研究证实，通用型人工智能在与网络攻击相关的方面的能力正在显著提升，但目前尚不清楚这是否会影响攻击者与防御者之间的平衡。
- **生物和化学攻击：** 近期通用型人工智能系统已展现出一定的能力，能够为复制已知生物和化学武器提供操作说明和故障排除指导，并有助于设计新型有毒化合物。在测试生成生物武器生产计划能力的新实验中，通用型人工智能系统有时表现得比能上网查询的人类专家还要出色。对此，一家人工智能公司将其最佳模型的生物风险评估从“低”提升到了“中”。不过，要在现实世界中开发此类武器仍需大量额外资源和专业知识。由于许多相关研究属于机密，对生物和化学风险进行全面评测估颇具难度。

自中期报告（2024 年 5 月）发布以来，通用型人工智能在与恶意使用相关的领域中的应用能力有所提高。例如，研究人员最近构建了通用型人工智能系统，该系统能够自行发现和利用网络安全漏洞，并在人工协助下发现广泛使用的软件中以前未知的漏洞。通用型人工智能系统在推理和整合不同类型数据方面的能力也有所了提升，这些能力有助于病原体或其他双重用途领域的研究。

故障风险：通用型人工智能也可能造成意外危害。即使用户无意造成伤害，通用型人工智能的故障也可能会造成严重风险。此类故障包括：

- **可靠性问题：** 当前的通用型人工智能可能不可靠，从而可能造成危害。例如，如果用户向通用型人工智能系统咨询医疗或法律建议，系统可能会生成包含错误信息的回答。用户常常意识不到人工智能产品的局限性，例如由于有限的“人工智能素养”、广告误导或沟通不畅。已知有大量因可靠性问题而造成危害的案例，但关于此类问题在不同领域的普遍程度，目前仍缺乏确切的证据。
- **偏见：**通用型人工智能系统可以放大社会和政治偏见，造成具体的伤害。它们常常在种族、性别、文化、年龄、残障、政治观点或人类身份的其他方面表现出偏见。这可能导致歧视性结果，资源分配不均、强化刻板印象以及对代表性不足的群体或观点的系统性忽视。虽然在通用型人工智能系统中减轻偏见和歧视的技术方法正在取得进展，但这些方法在偏见缓解与准确性、隐私等其他目标之间存在权衡，同时还面临其他挑战。

- **失控：** 所谓"失控"场景，指的是未来可能出现的一种假设性情况，即一个或多个通用型人工智能系统开始脱离任何人的控制，且没有明确的重新掌控途径。目前普遍认为，现有的通用型人工智能尚不具备造成这种风险的能力。然而，对于未来几年内出现失控的可能性，专家意见差异很大：一些专家认为失控不太可能发生，另一些专家则认为失控有可能发生，还有专家认为这是一种可能性适中但潜在危害严重的风险，因此值得予以关注。相关的实证研究和数学研究正在逐步推进这些讨论。

自中期报告（2024 年 5 月）发布以来，有新的研究对偏见和失控风险给出了新的见解。新的研究对偏见风险和失控风险有了新的认识。通用型人工智能系统存在偏见的证据有所增加，近期的研究还发现了更多形式的人工智能偏见。研究人员观察到，人工智能在某些能力方面有了适度的进一步提升，这些能力可能是通常所讨论的失控场景发生的必要条件，包括自主使用计算机、编程、未经授权访问数字系统以及寻找规避人类监督的方法。

系统性风险：除了单个模型的能力直接带来的风险之外，通用型人工智能的广泛部署还会带来一些更广泛的系统性风险。系统性风险的示例包括潜在的劳动力市场影响、隐私风险和环境影响：

- **劳动力市场风险：** 通用型人工智能，特别是如果其继续快速发展，有可能使非常广泛的任务实现自动化，这可能会对劳动力市场产生重大影响。这意味着许多人可能会失去目前的工作。然而，许多经济学家预计，潜在的就业岗位流失可能会被部分抵消，甚至有可能完全抵消，因为新的工作岗位会不断涌现，而且非自动化行业的需求也会增加。
- **全球人工智能研发鸿沟：**目前，通用型人工智能的研发集中在少数西方国家和中国。这种"人工智能鸿沟" 有可能使世界上许多国家对这一小部分国家产生更大的依赖。一些专家还预计人工智能鸿沟将加剧全球不平等。造成这种不均衡的原因有很多，其中一些并非人工智能所独有。然而，很大程度上，这是由于获取开发通用型人工智能所需的极其昂贵的计算能力方面存在差异：大多数中低收入国家拥有的计算能力远远低于高收入国家。
- **市场集中度和单点故障：**目前，少数几家公司主宰着通用型人工智能市场。这种市场集中度可能使社会更容易受到多种系统性风险的影响。例如，如果金融或医疗保健等关键领域的组织都依赖少数通用型人工智能系统，那么此类系统中的漏洞或缺陷可能会导致大规模的同步故障和中断。
- **环境风险：**通用型人工智能开发和部署中计算的用量不断增长，迅速提高了构建和运行所需计算基础设施所消耗的能源、水和原材料的数量。尽管技术上的进步使计算得到了更有效的利用，但增长趋势并没有明显放缓的迹象。通用型人工智能还有许多应用，既可以促进可持续发展，也可以损害可持续发展的成果。

- **隐私风险：**通用型人工智能可能导致或加剧侵犯用户隐私的行为。例如，训练数据中的敏感信息可能在用户与系统交互时无意中泄露。此外，当用户与系统共享敏感信息时，这些信息也可能会泄露。但通用型人工智能也可能为故意侵犯隐私提供便利，例如，恶意行为者使用人工智能从大量数据中推断出特定个人的敏感信息。然而，到目前为止，研究人员尚未发现与通用型人工智能相关的大规模隐私侵犯的证据。
- **版权侵权：**通用型人工智能既从创意表达作品中学习，又生成此类作品，这给传统的数据同意、补偿和控制体系带来了挑战。数据收集和内容生成可能涉及多种数据权利法，这些法律因司法管辖区而异，并且可能正处于积极的诉讼之中。鉴于数据收集时间的法律不确定性，人工智能公司对其使用的数据披露的信息越来越少。这种不透明性使得第三方对人工智能安全性的研究更加困难。

自中期报告（2024 年 5 月）发布以来，有关通用型人工智能对劳动力市场影响的更多证据浮出水面，同时新的发展态势也加剧了隐私和版权方面的担忧。对劳动力市场数据的最新分析表明，与以往技术相比，个人采用通用型人工智能的速度非常快。不同行业的企业使用速度存在很大差异。此外，通用型人工智能能力的最新进展使其越来越多地被部署在医疗保健或工作场所监控等敏感领域，从而带来了新的隐私风险。最后，随着版权纠纷的加剧以及针对版权侵权的技术缓解措施仍不可靠，数据权利持有者正在迅速限制对其数据的访问。

开放权重模型（Open-weight models）：评估通用型人工智能模型可能带来的诸多风险的一个重要因素，是它如何向公众发布。所谓的“开放权重模型”是指人工智能模型的核心组件（称为“权重”）公开供下载的人工智能模型。开放权重访问有利于研究和创新，包括人工智能安全方面的研究，同时提高透明度并使研究界更容易发现模型中的缺陷。然而，开放权重模型也可能带来风险，例如，它可能为恶意或错误使用提供便利，而这种使用对于模型开发者来说难以或无法监控或缓解。一旦模型权重可供公众下载，就无法对所有现有副本的全面回滚，也无法确保所有现有副本都收到安全更新。自中期报告（2024 年 5 月）发布以来，已达成高级别共识，即认为应从以“边际”风险（Marginal risks）的角度来评估人工智能开放程度提高所带来的风险：即相对于封闭式模型或其他技术等现有替代方案带来的风险，发布开放权重模型造成的风险，相较于给定风险，增减程度。

第 3 部分 – 风险管理：有哪些技术可以管理通用型人工智能带来的风险？

有多种技术方法可以帮助管理风险，但在许多情况下，现有的最佳方法仍然存在很大的局限性，同时亦缺乏其他安全攸关领域所提供的定量风险评测或保障。

风险管理——识别和评估风险，然后减轻和监控风险——在通用型人工智能的背景下颇具难度。尽管在许多其他领域，风险管理也一直极具挑战性，但通用型人工智能的某些特性似乎带来了独特的难题。

通用型人工智能的若干技术特性使得该领域的风险管理尤为困难。其中包括但不限于：

- **通用型人工智能系统的可能用途和使用环境范围异常广泛。**例如，同一系统可用于提供医疗建议、分析计算机代码中的漏洞以及生成照片。这增加了全面预测相关用例、识别风险或测试系统在相关现实情况下的行为方式的难度。
- **开发人员对于其通用型人工智能模型的运作原理仍知之甚少。这种理解上的缺失使得预测行为问题以及在观察到已知问题后对其进行解释和解决都变得更加困难。**理解之所以难以实现，主要是因为通用型人工智能模型并非以传统方式编程。这些模型是通过训练获得的：人工智能开发人员设置一个包含大量数据的训练过程，而该训练过程的结果就是通用型人工智能模型。此类模型的内部运作原理大多不为人知，包括对于模型开发者来说也是如此。模型解释和"可解释性"技术可以提高研究人员和开发者对通用型人工智能模型运行原理的理解，但是，尽管最近取得了进展，但这项研究仍处于起步阶段。
- **能力日益增强的自主智能体（能够自主行动、规划和授权以实现目标的通用型人工智能系统）可能会给风险管理带来新的重大挑战。**一般情况下，自主智能体通常可使用通用软件（例如网络浏览器和编程工具）自主地实现目标。目前，大多数自主智能体还不够可靠，无法广泛使用，但各家公司正在大力构建功能更强大、更可靠的自主智能体，并在近来数月取得了进展。自主智能体可能会变得越来越有用，但也可能加剧本报告中讨论的许多风险，并给风险管理带来新的难题。这些潜在的新挑战的例子包括：用户可能无法时刻掌握自己的自主智能体的动态，自主智能体的运行可能不受人控制，攻击者可能"劫持"自主智能体，以及人工智能之间的交互可能产生复杂的新风险。与自主智能体相关的风险管理办法才刚刚开始形成。

除了技术因素外，一些经济、政治和其他社会因素也使得通用型人工智能领域的风险管理尤为困难。

- **通用型人工智能的发展速度给决策者带来了"证据困境"。**能力的迅速提升使得某些风险有可能在短时间内突然出现；例如，利用通用型人工智能进行学术作弊的风险在一年内就从微不足道变得十分普遍。

风险出现得越快，就越难以通过事后应对来管理风险，而提前准备就显得越有价值。然而，只要关于风险的证据仍不完整，决策者就无法确定风险是否会出现，或者是否已经出现。这就需要权衡：采取预防性或早期的缓解措施可能会被证明是不必要的，但等待确凿证据可能会让社会面临迅速出现的风险而毫无防备。

- **人工智能公司对其人工智能系统的了解与政府和非行业研究人员对其人工智能系统的了解之间存在信息鸿沟。**在通用型人工智能系统广泛发布之前，公司通常只分享有限的信息。公司以商业顾虑和安全顾虑为由限制信息共享。然而，这种信息鸿沟也使得其他参与者更难以有效地参与风险管理，尤其是对于新兴风险而言。
- **无论是人工智能企业还是政府，往往都面临着激烈的竞争压力，这可能会导致它们降低对风险管理的重视程度。**在某些情况下，竞争压力可能会促使企业投入在风险管理上的时间或其他资源少于原本应有的投入。同样，当政府认为在国际竞争和风险降低之间存在权衡时，它们可能会减少用于支持风险管理的政策投入。

尽管如此，针对通用型人工智能的风险管理，存在多种可供公司采用、监管机构要求实施的技术和框架。这包括识别和评估风险的方法，以及减轻和监控风险的方法。[3]

- **对通用型人工智能系统进行的风险评估是风险管理的重要组成部分，但现有的风险评估存在严重局限性。**现有的通用型人工智能风险评估主要依赖于"抽查"，即在一系列特定情况下测试通用型人工智能的行为。这有助于在部署模型之前发现潜在的危害。然而，由于测试条件与现实世界不同，现有测试往往会遗漏危险，并且高估或低估通用型人工智能的能力和风险。
- **为了进行有效的风险识别和评估，评估者需要具备丰富的专业知识、充足的资源以及对相关信息的获取渠道。**通用型人工智能背景下进行严格的风险评估，需要将多种评测方法结合起来。这包括对模型和系统本身的技术分析，到对某些使用模式可能带来的风险评估。评测方需要丰富的专业知识才能正确地进行此类评测。为了进行全面的风险评估，他们通常还需要更多时间、对其模型与训练数据更直接的访问权限，以及比开发通用型人工智能的公司通常提供的更多有关所使用的技术方法的信息。
- **在训练通用型人工智能模型以使其更安全地运行方面，我们取得了一些进展，但目前尚无任何方法可以可靠地防止出现明显不安全的输出。**例如，一种称为"对抗训练"（adversarial training）的技术故意将 人工智能模型暴露于旨在使其在训练期间失败或行为不当的情景中，以增强其对这类情况的抵御能力。但是，对手仍然可以找到新的方法，以较低到中等的代价来规避此类防护措施。此外，最近的

[3†] 请参阅主席在撰写本报告后有关对人工智能最新进展的看法。

证据表明，当前的训练方法严重依赖不完善的人类反馈，可能会无意中促使模型在难以回答的问题上隐藏错误误导人类，使错误更难发现。提高这种反馈的数量和质量是一条可行的改进途径，不过通过人工智能检测误导行为的新兴训练技术也展现出了希望。

- **监控——在模型投入使用后识别风险和评估性能——以及各种预防有害行为的干预措施，能够提高通用型人工智能在部署给用户后的安全性。** 当前的工具能够检测人工智能生成的内容、追踪系统性能，并识别潜在有害的输入/输出，不过技术稍好的用户往往能够绕过这些防护措施。
- **在整个人工智能生命周期中，存在多种隐私保护的方法。** 此类方法包括从训练数据中删除敏感信息、控制从数据中学习多少信息的模型训练方法（例如差分隐私 （differential privacy）方法），以及将 人工智能与敏感数据结合使用的技术，使得数据难以恢复（例如"机密计算"（confidential computing）和其他隐私增强技术）。由于人工智能系统的计算需求，许多来自其他研究领域的隐私增强方法尚不适用于通用型人工智能系统。近几个月来，隐私保护方法已经扩展到解决人工智能在敏感领域日益广泛的应用，包括智能手机助手、自主智能体、始终在线的语音助手以及医疗保健或法律实践中的应用。

自中期报告（2024 年 5 月）发布以来，研究人员在解释通用型人工智能模型为何生成给定输出方面取得了一些进展。能够解释人工智能的决策可以帮助管理因偏见、事实不准确以及失控等故障带来的风险。此外，世界各地在使评估和缓解方法标准化的方面也做出了越来越多的努力。

结论：通用型人工智能的未来存在多种可能的发展轨迹，这在很大程度上取决于社会和政府如何行动。

通用型人工智能的未来充满不确定性，即便在不久的将来，其发展轨迹也呈现出多种可能性，既有非常积极的结果，也有非常消极的后果。但通用型人工智能的未来绝非不可避免。通用型人工智能如何开发、由谁开发，它被设计来解决哪些问题，社会能否充分挖掘其经济潜力，谁将从中受益，我们面临何种风险，以及我们在研究上投入多少资金来管控风险——这些问题以及许多其他问题，都取决于社会和政府当下及未来为塑造通用型人工智能的发展所做出的选择。

为了促进对这些决策的建设性讨论，当前关于通用型人工智能风险管控的科学研究和讨论状况。此事关系重大。我们期待继续推进这项工作。