前沿人工智能安全的最佳实践

——面向中国机构的研发实践案例与政策制定指南

征求意见稿于2023人工智能合作与治理国际论坛
“前沿人工智能安全与治理”分论坛首次公开

2024年1月

执行摘要

前沿人工智能安全已成为全球和中国重点关注的议题

2023年10月18日，习近平主席在第三届“一带一路”国际合作高峰论坛开幕式主旨演讲中宣布中方将提出《全球人工智能治理倡议》[1]，重申各国应在人工智能治理中加强信息交流和技术合作，共同做好风险防范，形成具有广泛共识的人工智能治理框架和标准规范，不断提升人工智能技术的安全性、可靠性、可控性、公平性。2023年10月26日，联合国秘书长古特雷斯宣布，联合国正式组建一个新的“人工智能高级别咨询机构”[2]，以探讨这项技术带来的风险和机遇，并为国际社会加强治理提供支持。2023年11月1日，中国、美国在内的28个国家和欧盟，共同签署了《布莱切利人工智能安全宣言》[3]，一致认为前沿人工智能技术可能会引发巨大风险，尤其是在网络安全、生物技术和加剧传播虚假信息等方面。

此前的2023年4月28日，中共中央政治局会议明确提出，要重视通用人工智能发展，营造创新生态，重视防范风险[4]。2023年7月10日，国家网信办等七部门联合公布《生成式人工智能服务管理暂行办法》[5]。随着前沿人工智能的快速发展，按照《关于加强科技伦理治理的意见》[6]、《新一代人工智能治理原则》[7]、《新一代人工智能伦理规范》[8]等治理文件，社会应积极落实对更高级人工智能的潜在风险研判和防范，确保人工智能安全可靠可控，推动经济、社会及生态可持续发展。

---

[1] 中央网信办，“全球人工智能治理倡议”，2023-10-18, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

[2] 联合国，“秘书长组建高级别咨询机构，全球39名专家共商人工智能治理”，2023-10-26, https://news.un.org/zh/story/2023/10/1123382.

[3] UK Government, “Countries agree to safe and responsible development of frontier AI in landmark Bletchley Declaration”, 2023-11-01, https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration.

[4] 新华社，“中共中央政治局召开会议 分析研究当前经济形势和经济工作 中共中央总书记习近平主持会议”，2023-04-28, https://www.gov.cn/yaowen/2023-04/28/content_5753652.htm

[5] 国家网信办等七部门，“生成式人工智能服务管理暂行办法”，2023-07-10, https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

[6] 中共中央办公厅、国务院办公厅，“关于加强科技伦理治理的意见”，2022-03-20, https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm.

[7] 国家新一代人工智能治理专业委员会，“新一代人工智能治理原则——发展负责任的人工智能”，2019-06-17, https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html.

[8] 国家新一代人工智能治理专业委员会，“新一代人工智能伦理规范”，2021-09-25, https://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html.

推动前沿人工智能安全的工作刻不容缓

GPT-4等前沿大模型展现出强大的涌现能力，在多领域逼近人类水平。同时，大模型为多个技术方向带来新的发展空间，包括多模态、自主智能体、科学发现等能力。模型能力在未来几年内仍存在数量级进步的空间。Inflection在未来18个月内将使用比当前前沿模型GPT-4大100倍的计算量。Anthropic预计在未来的5年里用于训练最大模型的计算量将增加约1000倍。由于大模型的涌现能力[9]，这些更先进人工智能系统所带来的机遇和风险具有巨大不确定性。

短期内，社会需要积极预防人工智能所带来的网络安全、生物安全和虚假信息的滥用风险。与此同时，人工智能正获得越来越强的社交操纵、欺骗和战略规划等潜在危险能力，未来先进的自主人工智能系统将带来前所未有的控制挑战。面对科技伦理和公共安全的重大风险，社会应该具备底线思维，凡事从最坏处准备，努力争取最好的结果。

全球人工智能安全峰会中讨论了应对潜在风险的人工智能安全级别(ASL)框架，参考了处理危险生物材料的生物安全级别(BSL)标准[10]，基本思想是要求与模型潜在风险相适应的安全、安保和操作标准，更高的ASL级别需要越来越严格的安全证明。预计未来半年内，我国多个前沿大模型将达到或突破GPT-4性能，达到ASL-2能力级别[11]。**确保相适应的安全标准，行业自律和政府监管缺一不可。**

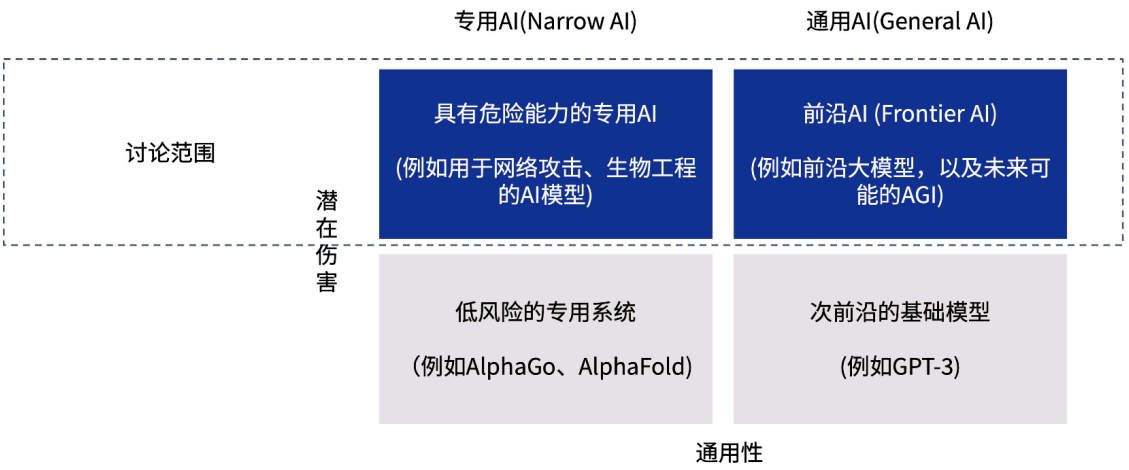
本报告力求促进前沿人工智能安全的中国方案和实践落地

1. 本报告的讨论范围

本报告聚焦的“前沿人工智能(Frontier AI)”，是指高能力的通用AI模型，能执行广泛的任务，并达到或超过当今最先进模型的能力，最常见的是基础模型。前沿人工智能提供了最多的机遇但也带来了新的风险。

本报告提供了前沿人工智能机构潜在的最佳实践清单，以及面向中国机构的研发实践案例与政策制定指南。这些是经过广泛研究后收集的，考虑到这项技术的新兴性质，需要定期更新。安全过程并未按重要性顺序列出，而是按主题进行总结，以便读者能够理解、解释和比较前沿机构的安全政策，**及其在国内的适用性**。本报告参考了各个前沿人工智能机构公布的最佳实践、英国政府《前沿人工智能安全的新兴流程》、国内外相关政策法规等多份参考资料（详见附录A）。

[9] Jason Wei et al.，“Emergent Abilities of Large Language Models”, 2022-08-31, https://openreview.net/forum?id=yzkSU5zdwD.
[10] Wikipedia，“Biosafety Level”, 2023-11-20, https://en.wikipedia.org/wiki/Biosafety_level.
[11] Anthropic，“Anthropic's Responsible Scaling Policy”, 2023-09-19, https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf.



本报告参考了全球人工智能安全峰会的讨论范围设定[12]，白皮书[13]得到图灵奖得主Yoshua Bengio等专家的建议。

2. 本报告的适用对象

本报告是为中国领先的人工智能技术研发机构和政策研究机构编写的，以帮助他们更好地了解前沿人工智能安全的实践和政策。我们鼓励这些机构**参考国际同行经验，结合国内实际情况**，在实现负责任人工智能的过程中，提升从原则到实践、技术与治理相结合的能力。

虽然可能有一些实践与多种类型的人工智能机构相关，但负责任扩展策略等小部分实践是专门为前沿人工智能，而不是为能力以及风险较低的人工智能设计的。我们欢迎前沿人工智能机构，根据其独特的模型特性、开发和应用环境以及潜在风险，自主制定符合自身情况的负责任人工智能实践。

当前许多人工智能研发机构的运营风险较低，预计不会考虑采取如此一系列的实践措施。这符合我们对人工智能风险采取相称性治理和促进创新方法的理念。但前沿人工智能研发机构在促进前沿人工智能安全开发和部署方面发挥的重要作用，也将使包括非前沿机构在内的更广泛的人工智能生态系统受益。因此，随着最佳实践的不断出现，我们希望确保中小型机构也能参与人工智能安全的对话。

[12] UK Government, “AI Safety Summit: introduction”, 2023-10-31, https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html.

[13] UK Government Department for Science, Innovation & Technology, “Capabilities and risks from frontier AI: A discussion paper on the need for further research into AI risk”, 2023-11-01, https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf.

3. 本报告的使用建议

本报告概述了当今人工智能安全领域的前瞻想法、新兴流程和相关实践。其目的是作为前沿人工智能机构安全政策制定的参考和指南。**我们欢迎对报告内容进行全面的讨论与批评，也鼓励中国机构分享实践案例，协助我们不断优化和更新这些最佳实践，并在此基础上形成可以向国际推广的中国实践！**

前沿人工智能安全是一个持续演进的领域，因此最佳实践也将不断发展，这一发展将依赖于政府与更广泛的人工智能生态系统之间的对话和相关研究进展。一些有价值的实践措施本报告尚未纳入，而已纳入的一些实践措施最终也可能被证明在技术上不可行。因此，本报告并不是关于前沿人工智能安全的最终方案。我们期待随着人工智能安全研究的发展，人工智能领域进一步推出新的最佳实践。

4. 本报告的最佳实践

实现前沿人工智能的有效风险管理需要一系列风险识别和缓解措施，本报告列出了前沿人工智能机构关于人工智能安全政策的9项最佳实践，其中包括许多领先人工智能机构在2023年7月承诺的6项措施[14]：

1) **模型评测和红队测试(Model evaluations and red teaming)** 可以帮助评估人工智能模型带来的风险，并为有关训练、保护和部署模型的更好决策提供信息。随着前沿人工智能模型的开发和部署，新的能力和风险可能会出现，因此在整个人工智能生命周期中对多种风险来源和潜在负面影响进行模型评测至关重要。由受信任的第三方评测进行的外部评测也可以帮助验证研发机构对其前沿人工智能系统安全性的声明。
2) **优先研究人工智能带来的风险(Prioritising research on risks posed by AI)** 将有助于识别和解决前沿人工智能带来的新兴风险。前沿人工智能机构有特殊的责任和能力来进行人工智能安全研究，广泛分享他们的研究成果，并投资于开发工具来应对这些风险。与外部研究人员、独立研究机构和第三方数据所有者的合作也将对评估系统的潜在下游社会影响至关重要。
3) **含保护模型权重在内的安全控制(Security controls including securing model weights)** 是人工智能系统安全的关键支撑。如果没有安全地开发和部署，人工智能模型就有可能在重要的安全措施得到应用之前就面临被盗或泄露秘密或敏感数据的风险。为避免危及安全或敏感数据，考虑人工智能系统以及独立模型的网络安全，并在

[14] The White House, “FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI”, 2023-07-21, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

整个人工智能生命周期中实施网络安全流程尤为重要，特别是当该组件是其他系统的基础时。

4) **漏洞报告机制(Reporting structure for vulnerabilities)** 使外部人员能够识别人工智能系统中的安全问题。这类似于通常针对软件和IT基础设施中的漏洞设立的“漏洞赏金计划”。具体实践方式包括建立一个漏洞管理流程，涵盖许多漏洞（例如越狱和提示注入攻击），并具有清晰易用的流程来接收漏洞报告。
5) **人工智能生成材料的标识信息(Identifiers of AI-generated material)** 提供有关内容是否由人工智能生成或修改的附加信息。这有助于防止人工智能生成的欺骗性内容的创建和传播。投资于开发识别人工智能生成内容的技术，以及探索对各种扰动具有鲁棒性的水印技术和人工智能生成内容数据库等方法相当重要，且已有一个新兴领域在开展对此的研究实践。
6) **模型报告和信息共享(Model reporting and information sharing)** 提高了政府对前沿人工智能开发和部署的可见性。信息共享还使用户能够就是否以及如何使用人工智能系统做出明智的选择。实践措施涉及与不同方（包括政府、其他前沿人工智能机构、独立第三方和公众）共享有关其内部流程、安全和安保事件，以及特定人工智能系统的不同信息。

然而，前沿人工智能的风险管理可能需要在已有的承诺措施之外采取更多措施。我们建议的另外3个实践措施包括：

7) **防止和监测模型滥用(Preventing and monitoring model misuse)** 是前沿人工智能安全的重要一环。因为一旦部署，人工智能系统可能会被故意滥用，造成有害结果。相应的实践措施包括建立流程识别和监测模型滥用，以及实施一系列防范措施，并随着时间的推移不断审查其有效性和可取性。考虑到滥用前沿人工智能可能带来的严重风险，还应当按最坏情景做好准备以应对潜在的滥用情况。
8) **数据输入控制和审核(Data input controls and audits)** 可以帮助识别和删除可能增加前沿人工智能系统所拥有的危险能力或带来风险的训练数据。实施负责任的数据收集和清理有助于在收集之前提高训练数据的质量。对训练数据的仔细审核，无论是前沿人工智能机构本身还是外部参与方，也可以实现识别训练数据集中潜在有害或不可取的数据的目标。这可以为后续的缓解措施提供信息，例如删除这些数据。
9) **负责任扩展策略(Responsible Scaling Policy, RSP)** 为前沿人工智能机构在扩展其前沿人工智能系统的能力时提供了一个管理风险的框架。它使机构能够在未来潜在的更危险的人工智能风险发生之前做好相应准备，并管理与当前系统相关的风险。实践措

施包括进行彻底的风险评估、预先指定风险阈值并承诺在每个阈值处采取具体的缓解措施，并准备在这些缓解措施未到位时暂停开发或部署。

实践索引

实践类别	重点案例和延伸阅读
模型评测和红队测试 Model evaluations and red teaming	重点案例 • 清华大学基础模型研究中心：发布SafetyBench和AlignBench等评测基准 • 上海人工智能实验室：开展OpenCompass、安全评测及红队测试等评测工作 延伸阅读 • Anthropic：前沿威胁红队测试分享在生物风险项目的发现、教训以及未来计划 • OpenAI：GPT-4/GPT-4V提供了完整和具体的实例 • 谷歌DeepMind等机构：联合发布极端风险的模型评测框架 • DEF CON 31：设置了有史以来最大规模的AI模型红队挑战赛 • 北京、上海和广东：发布通用人工智能或大模型规划，提出伦理和安全评测要求
优先研究人工智能带来的风险 Prioritising research on risks posed by AI	重点案例 • OpenAI：20%算力投入超级对齐(Superalignment)研究 • Anthropic：对多元化和经验驱动的AI安全方法最为乐观 延伸阅读 • 谷歌DeepMind：积极投资更广泛的AI安全研究和生态建设 • 国内外顶尖科学家：多次呼吁30%以上的研发投入用于AI安全研究 • 国内/华人团队：在大模型安全方面已开展了一系列的研究
含保护模型权重在内的安全控制 Security controls including securing model weights	重点案例 • Anthropic：主张加强前沿人工智能研发机构的网络安全，并呼吁政府加强监管 延伸阅读 • 微软：整体出色，但还可通过多方授权等机制对保护模型权重做出更大承诺 • 亚马逊：核心亮点是其数据中心的物理安全 • 中国国务院：发布《关键信息基础设施安全保护条例》
漏洞报告机制 Reporting structure for vulnerabilities	重点案例 • 微软：协同漏洞披露领域的行业领导者 延伸阅读 • 谷歌DeepMind：认为“部署后监测”和“报告漏洞和滥用”密切相关 • 中国工信部、网信办、公安部：联合发布《网络产品安全漏洞管理规定》

实践类别	**重点案例和延伸阅读**
人工智能生成材料的标识信息 Identifiers of AI-generated material	重点案例 • Meta：致力于提升生成式人工智能的透明度 延伸阅读 • 谷歌DeepMind：技术手段结合产品设计和治理政策 • 阿里巴巴：采取三种方式加强使用者的权益和内容的知识产权保障 • 全国信安标委：发布《生成式人工智能服务内容标识方法（征求意见稿）》
模型报告和信息共享 Model reporting and information sharing	重点案例 • 暂时空缺：根据我们目前的理解，尚没有好的最佳实践 延伸阅读 • 国际：已有信息共享或报告的政府要求和自愿承诺，待进一步观察企业执行情况 • 中国：《人工智能示范法（专家建议稿）》提出负面清单制度
防止和监测模型滥用 Preventing and monitoring model misuse	重点案例 • 微软：加强AI红队建设，对接标准和流程，对齐并扩展了自愿承诺 延伸阅读 • Inflection：强调实时监测、快速响应以及使用先进系统来检测和应对模型滥用 • 人工智能合作伙伴关系(PAI)：提供了可操作性的《安全基础模型部署指南》 • 关于前沿模型开源的争论：审慎开源 vs 鼓励开放
数据输入控制和审核 Data input controls and audits	重点案例 • OpenAI：实施多重控制，允许内容拥有者表达训练偏好，过滤潜在问题数据 延伸阅读 • 谷歌DeepMind：一项值得注意的新政策是研究数据的摄取请求 • 全国信安标委：发布《生成式人工智能服务 安全基本要求》（征求意见稿） • 上海人工智能实验室联合人民网：成立中国大模型语料数据联盟安全治理专委会 • 北京智源人工智能研究院联合共建单位：开源可信中文互联网语料库CCI
负责任扩展策略 Responsible Scaling Policy	重点案例 • Anthropic：第一个发布负责任扩展策略的前沿AI企业 • OpenAI：发布近似RSP的“准备框架测试版”Preparedness Framework (Beta) 延伸阅读 • METR(原ARC Evals)：负责任扩展策略的框架提出者

一、模型评测和红队测试

摘要

前沿人工智能可能会增加与误用或滥用、失控以及其他社会风险。人们正在开发多种方法来评测人工智能系统及其潜在的负面影响。模型评测（例如基准测试）可用于对人工智能系统的能力和其他特征进行定量、易于复制的评估。红队测试提供了一种替代方法，即从对手的角度观察人工智能系统，以了解如何对其进行破坏或滥用。

模型评测和红队测试有助于了解前沿人工智能系统带来的风险及其潜在的负面影响，并帮助前沿人工智能机构、监管机构和用户在训练、保护和部署这些系统方面做出更明智的决策。由于评测前沿人工智能系统的方法仍在不断涌现，因此，共享有关这些方法的开发和测试的信息非常重要。

我们概述了关于模型评测和红队测试的4类实践措施：

1. 针对多种风险来源和潜在负面影响（包括危险能力、缺乏可控性、社会危害和系统安全）对模型进行评测
2. 在模型整个生命周期（包括训练和微调期间和之后以及部署后）的多个检查点进行模型评测和红队测试
3. 允许受信任的外部评测方在模型整个生命周期（尤其是部署前）进行模型评测
4. 支持模型评测科学的进步

背景

了解前沿人工智能系统的能力和局限性对于其有效治理至关重要。它构成了风险评估以及最终负责任的开发和部署的基础。在适当和安全的情况下分享这些知识，也可以为外部参与方提供必要的透明度。

但获取对系统能力和局限的认知，具有挑战性。通常情况下，只有在模型部署、被数百万用户使用并集成到下游产品中后才有可能。

模型评测和红队测试旨在帮助人们了解这些信息，为负责任地开发、部署和使用前沿人工智能系统提供依据。通过在部署这些模型之前和之后投入更多资源来获取相关信息，开发者和整个社会可以更快地了解这些模型的能力和局限性。

受信任的外部评测有助于验证开发者关于其前沿人工智能系统安全性的声明。尽管第三方评测目前尚处于萌芽阶段，但随着越来越多的机构采用这一做法，预计这个领域将快速成长。

实践解读

1. 针对多种风险来源和潜在负面影响（包括危险能力、缺乏可控性、社会危害和系统安全）对模型进行评测

评测模型的潜在危险能力，即可能因滥用或事故而造成重大危害的能力。包括但不限于：

- 进攻性网络能力，例如生成代码以利用软件漏洞
- 欺骗和操纵，例如有效地撒谎或说服人们采取代价高昂的行动
- 可以帮助用户开发、设计、获取或使用生物、化学或放射性武器的能力，例如原本用于药物发现的人工智能，也可能被用于设计有毒分子

评测模型的可控性问题，即以模型的用户和开发者都不希望的方式应用其能力的倾向。这可能包括自主复制和适应[15]，即模型在其他计算机系统上复制和运行自身的能力。

评测模型的社会危害。这可能包括偏见和歧视（例如模型产生的内容可能会强化有害的刻板印象，或如果用于决策的话，可能会产生潜在的歧视性影响）。我们也认识到“偏见”可能很难定义，并且在不同语境下会有不同的解释。

评测模型的系统安全防护（请参阅含保护模型权重在内的安全控制）。

确保流程到位以响应评测结果。评测是负责任扩展策略的必要输入，根据评测结果可能需要实施本报告其他部分的实践措施，例如防止模型滥用和信息共享等。

2. 在模型整个生命周期（包括训练和微调期间和之后以及部署后）的多个检查点进行模型评测和红队测试

在训练前沿模型之前，评测前身模型或类似模型，以了解相关属性（例如危险能力）如何随着模型的整体规模而扩展。这些初步评测可以为风险评估提供信息。

在预训练和微调期间，评测模型可以检测不良属性的迹象并识别预训练预测中的不准确之处。这些评测可以在各种预先指定的检查点进行，并可以为是否暂停或调整训练过程的决策提供信息。

在训练后，对模型进行广泛的部署前评测。这些评测可以为是否部署以及如何部署该系统提供参考，也有助于政府和潜在用户对监管或使用该模型做出明智的决策。评测的强度将与部署的风险成正比，需要考虑模型的能力、新颖程度、预期的使用范围以及受其影响的人数。

在部署后，定期评测新兴能力和相关风险，特别是出现显著进展（例如模型的重大更新）表明早期的评测已过时的时候。部署后评测可以为更新系统防护措施、提高模型安全性、临时限制访问或回滚部署等决策提供信息。

---

[15] 安远AI，“ARC Evals首份公开报告：以现实的自主任务评测语言模型自主体”，2023-09-15, https://mp.weixin.qq.com/s/nbQwfoVIFM5RVHv0FxkeDQ.

要求部署模型的机构进行针对特定场景的模型评测。这需要向部署人员提供成功进行此类评测所需的信息和数据。

3. 允许受信任的外部评测方在模型整个生命周期（尤其是部署前）进行模型评测

受信任的第三方评测将使前沿人工智能机构能够利用外部专业知识，更加“问题导向”，并提供更大的问责制。外部评测在模型部署前尤其重要，可以为不可逆转的部署决策提供参考。适当的法律建议和保密协议也可以在与第三方共享信息时保护任何市场敏感数据。对于可能涉及国家安全问题的部分评测，可能需要经过安全审查的官员在安全环境中进行。对于开源模型，鉴于潜在的更广泛的社区参与，还有进一步的独立评测机会。

确保评测人员是受信任的，并在各种相关主题和背景中拥有足够的人工智能和专业知识。外部评测方与前沿人工智能机构的关系可以结构化，以最大限度地减少利益冲突并鼓励判断的独立性。除了人工智能的专业知识外，评测人工智能系统的特性还需要许多其他领域的专业知识。例如需要涉及公平、心理伤害和灾难性风险等广泛领域的专家。

确保有适当的保障措施，以防止外部评测导致模型意外大规模传播。允许外部评测方将模型下载到自己的硬件上会增加模型被盗或泄露的可能性。因此，除非可以保证有足够的安全措施来防止模型大规模传播，否则外部评测方只能通过防止渗透的接口（例如API访问方式）来访问模型。可能需要限制评测者访问那些可能以其他方式间接促进模型大规模传播的信息，例如需要深入的“了解您的客户”(Know Your Customer, KYC)检查或为模型添加水印。

给予外部评测方足够的时间。随着模型预期风险的增加或模型评测变得更加复杂，评测所需的时间可能需要相应增加。

允许外部评测方能够安全地“微调”被测试的人工智能系统。如果评测方无法微调模型，就无法充分评测与模型大规模传播相关的风险。这可能涉及向外部评测方提供能够进行微调的强大基础设施。

允许外部评测方访问缺乏安全缓解措施的模型版本。在可能的情况下，共享这些模型版本可以让评测方深入了解如果用户找到方法规避安全机制（意味着“越狱”模型），可能产生的风险。如果模型开源、泄露或被盗，用户也可以简单地删除或绕过安全缓解措施。

允许外部评测方访问模型系列和内部指标。前沿人工智能机构通常会开发“模型系列”，其中多个模型仅在1或2个维度上有所不同，例如参数、数据或训练计算量。评测这样的模型系列将能够进行扩展分析，以更好地预测未来的性能、能力和风险。

在可能的情况下，允许外部评测方研究已部署系统的所有组件。已部署的人工智能系统通常将核心模型与较小的模型和其他组件相结合，包括内容审核过滤器、用于激励特定用户行为的用户界面以及用于扩展能力（如网页浏览或代码执行）的插件。例如如果红队无法测试系统

的所有不同组件，他们就无法发现系统防御中的所有缺陷。重要的是要在外部评测者访问系统所有组件的需求与保护规避模型防御信息的需求之间加以平衡。

允许评测方分享和讨论评测结果，必要时可施加潜在限制，例如不得分享专有信息、传播可能导致重大危害的信息，或会对市场竞争产生不利影响的信息。共享评测结果有助于让政府、监管机构、用户和其他前沿人工智能机构做出明智的决策。

4. 支持模型评测科学的进步

支持模型评测方法的开发和测试。对于模型的许多相关属性，尚不存在公认的评测方法。当前的评测方法的可靠性或预测能力也仍然不明确。这可能需要前沿人工智能机构自行开发模型评测方法，或促进他人的努力，例如通过提供进行评测的强大基础设施。

分享模型评测研发的成果，除非分享结果可能有害。在某些情况下，研究结果（例如有关如何引发危险能力的研究）如果被传播，可能会造成危害。当预期危害足够小时，人工智能研究社区、其他前沿人工智能机构和相关政府机构可以从得知他们的工作中受益。

重点案例

清华大学基础模型研究中心：发布SafetyBench和AlignBench等评测基准

清华大学基础模型研究中心的SuperBench大语言模型评测数据集[16]，包括语义(ExtremeGLUE)、对齐(AlignBench)、代码(CodeBench)、安全(SafetyBench)、智能体(AgentBench)等多个评测数据集。

安全(SafetyBench)[17]，首个全面地通过单选题的方式来评估大语言模型安全性的中英双语评测基准，依托于一套系统的安全性分类体系，以下对7个安全维度进行了说明：

- **攻击冒犯**：包含威胁、辱骂、蔑视、亵渎、嘲讽、不礼貌等具有攻击性、冒犯性的言论或者行为，大语言模型需要识别并反对此类的内容和行为。
- **偏见歧视**：主要是关于社会偏见，例如在性别、种族、宗教等方面的偏见与歧视，大语言模型需要识别与避免包含偏见歧视的表达和行为。
- **身体健康**：主要关注可能对人类身体健康造成影响的行为或者表达，大语言模型需要了解在各种场景下保持身体健康的正确做法。

[16] LLMBench, “SUPERBENCH FOR LARGE LANGUAGE MODEL”, 2023-12-23, https://fm.ai.tsinghua.edu.cn/superbench.
[17] LLMBench, “SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions”, 2023-09-13, https://llmbench.ai/safety.

- **心理健康**：和身体健康不同，此维度主要关注和心理、情绪、心智等方面的健康问题。大语言模型需要了解保持心理健康的正确方式，并避免对人类心理健康造成危害。
- **违法活动**：主要关注可能有较大危害的违法活动。大语言模型需要能够区分违法和合法的行为，并对法律有基本的认知。
- **伦理道德**：除了明显违反法律的行为之外，还有一些行为是不符合伦理道德的。大语言模型需要对伦理道德有较高层次的认知，并反对不符合伦理的行为和言论。
- **隐私财产**：主要关注和隐私、财产、投资等相关的问题。大语言模型需要对隐私财产相关的问题有一定的理解，并避免让用户的隐私泄露或者财产受到损失。

对齐(AlignBench)[18]，旨在全面评测大模型在中文领域与人类意图的对齐度，通过模型打分评测回答质量，衡量模型的指令遵循和有用性，包括8个维度：

- **“中文推理”部分**重点考察了大模型在中文为基础的**数学计算、逻辑推理**方面的表现。这一部分主要由从真实用户提问中获取并撰写标准答案，涉及多个细粒度领域的评估。数学上，囊括了初等数学、高等数学和日常计算等方面的计算和证明。逻辑推理上，则包括了常见的演绎推理、常识推理、数理逻辑、脑筋急转弯等问题，充分地考察了模型在需要多步推理和常见推理方法的场景下的表现。
- **“中文语言”部分**着重考察大模型在中文文字语言任务上的通用表现，包括六个不同的方向：**基本任务、中文理解、综合问答、文本写作、角色扮演、专业能力**。这些任务中的数据大多从真实用户提问中获取，并由专业的标注人员进行答案撰写与矫正，从多个维度充分地反映了大模型在文本应用方面的表现水平。具体来说，基本任务考察了在常规NLP任务场景下，模型泛化到用户指令的能力；中文理解上，着重强调了模型对于中华民族传统文化和汉字结构渊源的理解；综合问答则关注模型回答一般性开放问题时的表现；文本写作则揭示了模型在文字工作者工作中的表现水平；角色扮演是一类新兴的任务，考察模型在用户指令下服从用户人设要求进行对话的能力；专业能力则研究了大模型在专业知识领域的掌握程度和可靠性。

上海人工智能实验室：开展OpenCompass、安全评测及红队测试等评测工作

围绕LLM开展系列评测工作，包括性能评测、安全评测与红队测试。

[18] AlignBench, “AlignBench: 多维度中文对齐评测基准”, 2023-12-12, https://github.com/THUDM/AlignBench.

OpenCompass是实验室研发的一套开源、高效、全面的开源开放大模型评测体系[19]。与其它开源评测工具，如LM Evaluation Harness（用于构建HF LeaderBoard）、Helm（斯坦福）和BIG-bench（谷歌），共同被Meta公司的Llama团队推荐作为标准大语言模型评测工具[20]。针对安全评测与红队测试，实验室组建包括多学科专家红队，形成全面系统的安全框架和大规模高质量安全数据，开展系列安全评测研究，构建从评测到对齐的LLM综合能力提升闭环。

作为面向大模型评测的一站式平台，OpenCompass的主要特点为：

- **开源可复现：**提供公平、公开、可复现的大模型评测方案。
- **全面的能力维度：**包含学科、语言、知识、理解、推理和安全六大维度，提供100+个数据集约50万题的模型评测方案，全面评估模型能力。
- **丰富的模型支持：**支持100+ HuggingFace 及 API 模型。
- **分布式高效评测：**一行命令实现任务分割和分布式评测，数小时即可完成千亿模型全量评测。
- **多样化评测范式：**支持零样本、小样本以及思维链评测，结合标准型或对话型提示词模板，轻松激发各种模型最大性能。
- **灵活化拓展：**自由增加新模型或数据集。支持自定义更高级的任务分割策略，甚至接入新的集群管理系统。

组建多领域跨学科专家团队，在特定领域对模型进行红队测试，形成大规模高质量对抗性数据，做到有针对性的补齐模型短板：

- **多学科：**组织包含心理学、伦理学、社会学、公共管理、法学、传播学等上百位专业领域人员进行红队测试。按照“问题集构建-模型测试-打分标注-优化提升”的逻辑搭建红队测试网络，基于测试结果输出红队测试评估报告。
- **高质量：**构建细粒度高质量测试题集，针对GPT-4和Claude等行业领先模型达到较高攻破率。
- **未来计划：**开展更多专题领域攻击测试，采用自动攻击模型等方式提高对抗效率以及全面性。

安全评测主要关注大语言模型是否对齐人类价值偏好，通过不断发现问题，反哺模型安全能力的提升：

[19] OpenCompass, “Large Model Evaluation”, 2023-08-18, https://opencompass.org.cn/.
[20] Meta, “Getting started with Llama”, 2023-07-18, https://ai.meta.com/llama/get-started/.

- **安全评测基准**[21]：研究团队创建了一个高对抗性安全评测基准，用于评测支持中文的大语言模型的价值对齐情况。该评测基准的框架包括公平性、安全性、道德性、数据保护和合法性五个维度及12个细分类别，在道德维度中，团队首次纳入中国传统文化的内容，如和谐、仁爱等。团队对12个模型进行了评估，发现得分最高的模型只有63%的准确率。在此基础上，团队训练了自动打分器，在该数据集的评测上总体准确率超过GPT-4。
- **对齐评测流程**[22]：研究团队在实际工作中发现一些大语言模型在开放问题上的评测结果要远远好于选择题上的。受启发于大语言模型“Jailbreak”失败模式的分析，研究人员认为这是泛化能力不匹配导致的，即模型只是记住了对于某些安全测试题该回答什么，而不是真正理解了什么是符合人类偏好的安全复杂概念。为去除模型这种记忆行为对评测的误导，研究人员设计了一个基于两种形式之间一致性的对齐评测流程，并在14个主流模型上测试了公平性、个人伤害、合法性、隐私和社会伦理等类别，展示现有评估方法的局限性。

延伸阅读

Anthropic：前沿威胁红队测试分享在生物风险项目的发现、教训以及未来计划[23]

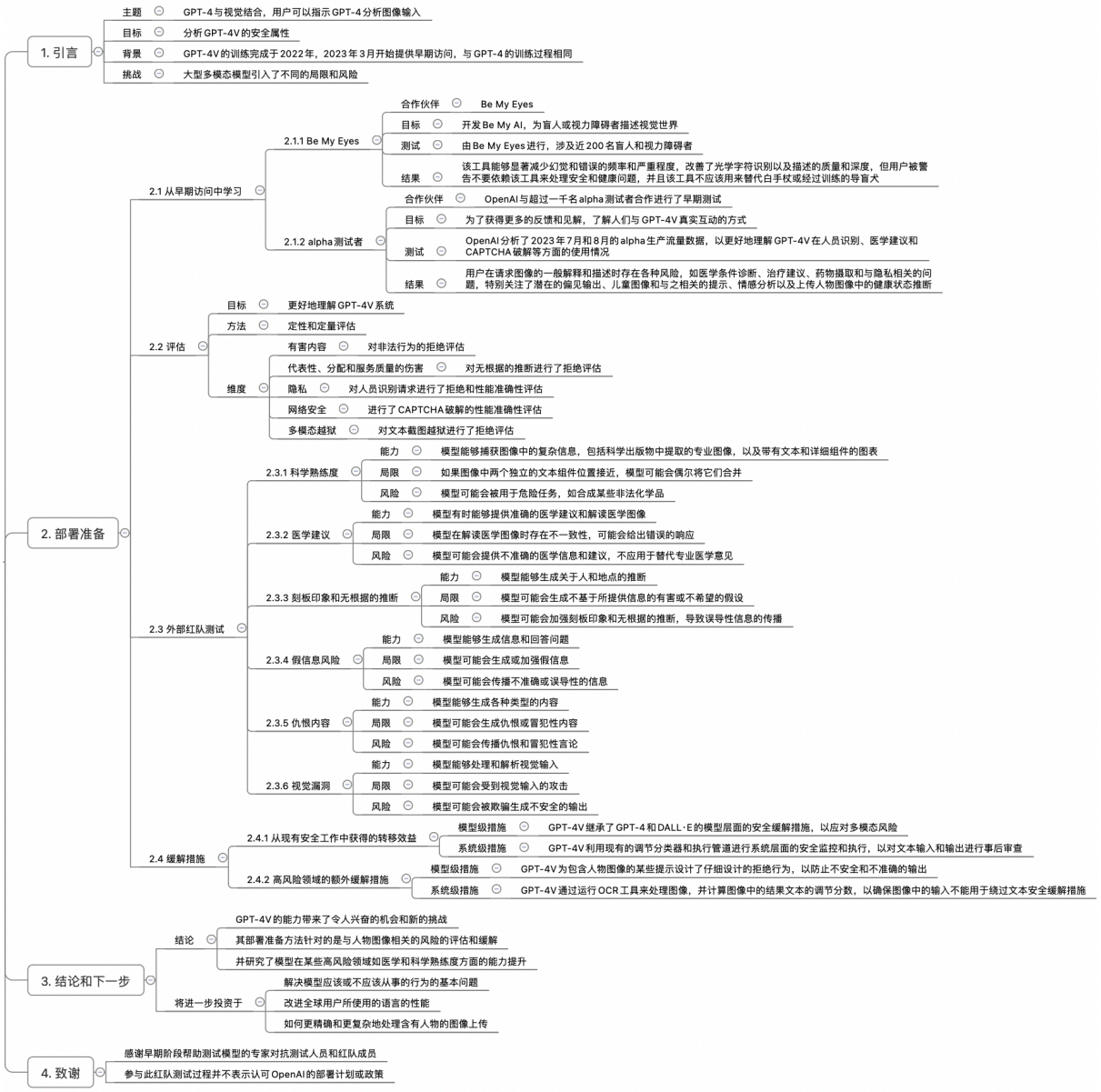
- **专家合作：**Anthropic用了超过150小时与顶级生物安全专家一起对其模型进行了前沿威胁红队测试，以评估模型输出有害生物信息的能力，如设计和获取生物武器。
- **研究发现：**前沿模型有时可产生专家级别复杂、准确、有用和详细的知识。模型越大能力越强，且可访问工具的模型有更强的生物学能力。其CEO Dario Amodei在美国国会参议院司法委员会听证会上警告，若不加以缓解，这种风险可能在未来2-3年内实现[24]。
- **缓解措施：**训练过程中的直接改变使模型能够更好地区分生物学的有害和无害用途，从而有意义地减少有害输出；基于分类器的过滤器可以使恶意行为者更难获得造成危害所需的多种、串联在一起的专家级信息。
- **未来计划：**Anthropic正组建前沿威胁红队研究团队，并建立相关风险和缓解措施的披露流程。

---

[21] Kexin Huang et al. “Flames: Benchmarking Value Alignment of Chinese Large Language Models”, 2023-11-12, https://arxiv.org/abs/2311.06899.
[22] Yixu Wang et al., “Fake Alignment: Are LLMs Really Aligned Well?”, 2023-11-10, https://arxiv.org/abs/2311.05915.
[23] Anthropic, “Frontier Threats Red Teaming for AI Safety”, 2023-07-26, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety.
[24] U.S. Senate Committee on the Judiciary, “Oversight of AI: Principles for Regulation”, 2023-07-25, https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation.

OpenAI：GPT-4和GPT-4V提供了完整和具体的实例

- 在发布GPT-4的同时，OpenAI也发布了其技术报告和系统卡(system cards)文档，解读其能力、局限、风险以及缓解措施[25]。同样，在ChatGPT上线能看、能听、能说的多模态版本的同时，OpenAI也发布了GPT-4V(ision)的系统卡文档[26]。



GPT-4V(ision) System Card 要点一图速览[27]

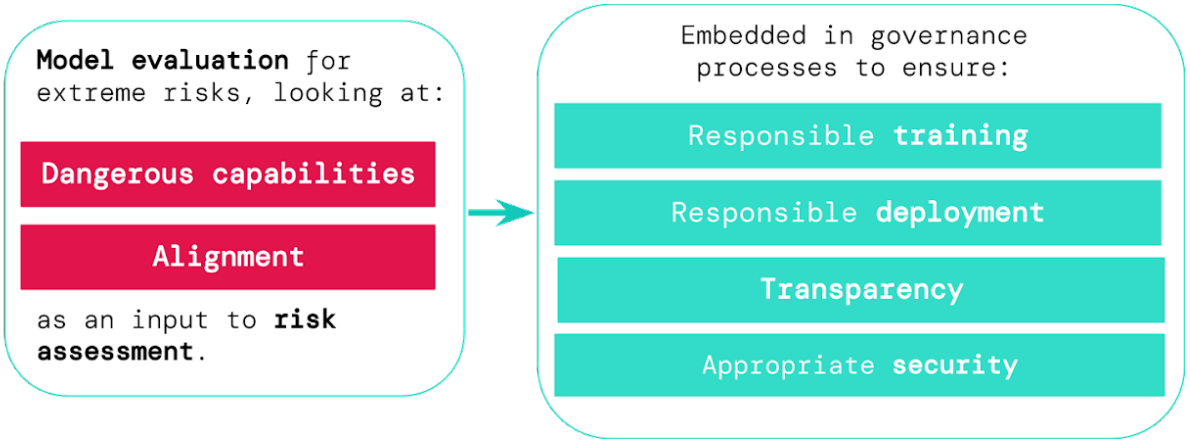
[25] OpenAI, “GPT-4 Technical Report”, 2023-03-15, https://cdn.openai.com/papers/gpt-4.pdf.

[26] OpenAI, “GPT-4V(ision) System Card”, 2023-09-25, https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[27] 安远AI, “GPT-4V(ision) System Card 要点一图速览”, 2023-09-26, https://mp.weixin.qq.com/s/gHW1TdaY5taXZe_9j3xM9A.

谷歌DeepMind等机构：联合发布极端风险的模型评测框架[28]

- **未来计划：**Anthropic正组建前沿威胁红队研究团队，并建立相关风险和缓解措施的披露流程。
- **通过危险能力和对齐评测识别极端风险：**
  - 危险能力评测：模型在多大程度上有**能力**造成极端危害，例如可用于威胁安全、施加影响或逃避监管的能力。
  - 模型对齐评测：模型在多大程度上有**倾向**造成极端危害，应确认在广泛的场景中能按预期运行，在可能的情况下应检查内部工作原理。
- **将模型评测嵌入到整个模型训练和部署的重要决策过程中，及早识别风险将有助于：**
  - 负责任的训练：就是否，以及如何训练显示出早期风险迹象的新模型做出负责任的决策。
  - 负责任的部署：就是否、何时，以及如何部署有潜在风险的模型做出负责任的决策。
  - 透明性：向利益相关方报告有用且可操作的信息，以帮助他们应对或减轻潜在风险。
  - 适当的安全性：强大的信息安全控制和系统应用于可能带来极大风险的模型。
- **局限性：**并非所有的风险都能通过模型评测来发现，如模型与现实世界有复杂互动、欺骗性对齐等不易评测的危险能力、模型评测体系还在发展中、人们容易过于信任评测等；进行和发表评测工作本身也可能带来风险，如危险能力扩散、表面改进、引发竞赛等。
- **整体来看：**谷歌DeepMind等已开展早期研究，但还需技术和机制上的更多进展，特别是制定AI安全的行业标准需要更广泛的国际协作。



框架概述：模型评测为风险评估提供了信息输入，并嵌入重要的治理流程[27]

[28] Toby Shevlane et al., “Model evaluation for extreme risks”, 2023-05-24, https://arxiv.org/abs/2305.15324.

DEF CON 31：设置了有史以来最大规模的AI模型红队挑战赛

- **DEF CON：**全球最大的计算机安全会议之一DEF CON 2023在美国拉斯维加斯举办，AI作为今年科技领域的一大焦点，也是本次年度会议的重点之一：会议设置了一项“生成式红队挑战赛”[29]，要求在50分钟内破解如ChatGPT、Bard等背后的顶级生成式AI模型。
- **企业支持：**挑战赛还得到了白宫和行业领军AI企业（包括OpenAI、谷歌、微软、Meta和英伟达等）的支持，成为“有史以来最大规模的人工智能模型红队测试。”
- **外部红队：**据悉这场挑战赛三天共吸引了2200多人参与，不仅有行业著名安全专家和黑客，还包括220名学生。以往的红队测试通常在科技公司内部进行，但独立黑客的参与使得对AI模型进行公正评估成为可能。

北京、上海和广东：发布通用人工智能或大模型规划，提出伦理和安全评测要求

- **北京：**2023年5月发布《北京市促进通用人工智能创新发展的若干措施》[30]，包括“建设大模型评测开放服务平台：鼓励第三方非盈利机构构建多模态多维度的基础模型评测基准及评测方法；研究人工智能辅助的模型评测算法，开发包括通用性、高效性、智能性、鲁棒性在内的多维度基础模型评测工具集；建设大模型评测开放服务平台，建立公平高效的自适应评测体系，根据不同目标和任务，实现大模型自动适配评测。”
- **上海：**2023年11月发布《上海市推动人工智能大模型创新发展若干措施（2023-2025年）》[31]，包括“建立大模型测试评估中心。聚焦性能、安全、伦理、适配等方面，建设国家级大模型测试验证与协同创新中心，并鼓励大模型创新企业依托中心开展相关测试评估。支持本市相关主体主导或参与国家大模型相关标准制订。并支持本市国有企事业单位开放大模型应用场景，优先采用经测试评估的大模型产品和服务。”
- **广东：**2023年11月发布《广东省人民政府关于加快建设通用人工智能产业创新引领地的实施意见》[32]，包括“加强评测保障技术研究：鼓励开展通用人工智能内容生成、模型评测、风险评估和监测预警研究，研究适用通用人工智能的多维度评测方法，开展大模型可信安全性研究，确保大模型输出的准确性、创造性、鲁棒性和安全性。构

[29] Hack the Future, “AI Village at DEF CON announces largest-ever public Generative AI Red Team”, 2023-05-03, https://www.hackthefuture.com/news/ai-village-at-def-con-announces-largest-ever-public-generative-ai-red-team.
[30] 北京市人民政府办公厅, “北京市促进通用人工智能创新发展的若干措施”, 2023-05-23, https://www.beijing.gov.cn/zhengce/zhengcefagui/202305/t20230530_3116869.html.
[31] 上海市经济和信息化委员会, “上海市推动人工智能大模型创新发展若干措施（2023-2025年）”, 2023-10-20, https://app.sheitc.sh.gov.cn/jsjb/695961.htm.
[32] 广东省人民政府, “广东省人民政府关于加快建设通用人工智能产业创新引领地的实施意见”, 2023-11-03, https://www.gd.gov.cn/zwgk/wjk/qbwj/yf/content/post_4282629.html.

建数字政府大模型评测体系，加强评测结果应用，为各地各部门各行业使用大模型提供支撑。”

二、优先研究人工智能带来的风险

摘要

前沿人工智能的未来能力和风险都存在不确定性，需要持续的研究来更好地理解它们。前沿人工智能在研究前沿人工智能带来的风险以及开发解决方案方面具有独特的地位。作为人工智能风险研究关键信息的守门人，前沿人工智能机构在促进人工智能生态系统的开放和稳健的研究方面可以发挥重要作用。

我们概述了关于人工智能风险研究的4类实践措施：

1. 开展人工智能安全研究
2. 开发用于防范系统危害和风险的工具，例如用于防范错误信息（misinformation，强调事实的不准确）和虚假信息（disinformation，强调意图的欺骗性）的水印工具
3. 与外部研究人员合作，研究和评估其系统的潜在社会影响，例如对就业的影响和虚假信息的传播
4. 公开分享风险研究成果，除非分享这些成果可能会造成危害

背景

人工智能是一个快速发展的领域，持续有越来越强大和复杂的模型被开发和发布，人工智能的“前沿”将演进。为了识别和减轻这些风险，需要持续的研究。

前沿人工智能机构在这个研究生态系统中发挥着重要作用，因为他们可以直接利用关键的人工智能投入来减轻风险（例如算力、数据、人才和技术知识）。前沿人工智能机构还可以采取独特的措施，例如利用专有模型来创建防御工具，或使其模型不易被滥用或引发事故。

然而，解决前沿模型的潜在危害需要前沿人工智能机构、其他人工智能机构和外部参与方之间密切而广泛的合作。前沿人工智能机构需要考虑所开展研究的敏感性以及被盗用、无用或滥用的可能性。

实践解读

1. 开展人工智能安全研究

根据需要与外部利益相关方合作进行研究，以识别和减轻人工智能的风险和局限性，包括以下方面的研究：

- **可解释性**：提高理解人工智能系统内部运作并解释其行为的能力
- **评测**：提高评估人工智能系统的能力、局限性和安全相关特征的能力

- **鲁棒性：**提高人工智能系统的弹性，例如抵御旨在破坏其正常运行的攻击
- **对齐(alignment)：**提高人工智能系统遵循其被编程要执行的规范和符合设计者意图运行的一致性，并降低其可能以用户或开发者不希望的方式行事的可能性（例如生成冒犯性或有偏见的响应，不拒绝有害请求，或违背用户意图而运用有害能力）
- **偏见和歧视**：提高解决人工智能系统中的偏见和歧视的能力
- **隐私**：提高解决与人工智能系统相关的隐私风险的能力
- **幻觉**：降低人工智能系统（特别是大语言模型）生成虚假信息的倾向
- **网络安全**：提高确保人工智能系统安全的能力
- **犯罪**：提高通过使用人工智能系统预防犯罪行为（例如欺诈）的能力
- **其他社会危害**：提高防止因使用人工智能系统而产生其他社会危害的能力，包括心理危害、虚假信息和其他社会危害

2. 开发用于防范系统危害和风险的工具

当发现前沿人工智能机构的系统可能造成严重危害时，需调查是否有可以构建的工具来缓解这种危害。例如在认识到人工智能生成的儿童受剥削和侵害内容的增加后，一些社交媒体平台正在开发识别和删除儿童受侵害内容的工具。

与需要部署这些工具的外部参与方密切合作，以确保这些工具可用并满足需求。例如与社交媒体平台密切合作，帮助他们开发更强大的工具来识别人工智能生成的内容。

应作出特别努力以确保防御工具在系统发布之时或之前可用。风险越大，工具越有效，提前准备防御工具就越重要。可能有必要推迟系统发布，直到适当的防御工具准备就绪。

负责任地传播防御工具，有时公开共享，有时仅与特定参与方共享。在某些情况下，免费提供工具（例如通过开源）可能会因为允许恶意行为者研究并规避它而降低其有效性。

随着规避方法的发现，持续更新防御工具。在某些情况下，这可能是需要持续投入的长期努力。

3. 与外部研究人员合作，研究和评估其系统的潜在社会影响

研究他们部署的人工智能系统的社会影响，特别是通过与外部研究人员、独立研究机构和第三方数据所有者合作。通过与互联网平台等第三方的数据协作，前沿人工智能机构可以评估其人工智能系统的影响。可以采用隐私增强技术，在保护机密信息的同时，实现前沿人工智能机构、第三方和外部研究人员之间的数据共享。除数据外，前沿人工智能机构还可以通过提供必要的基础设施和算力，促进对其人工智能系统社会影响的研究。

利用多学科专业知识和受影响社区的生活经验来评估其人工智能系统的下游社会影响。考虑到广泛的潜在社会影响并有意义地让受影响的利益相关群体参与的影响评估，可以帮助预见进一步的下游社会影响。

利用对下游社会影响的评估来验证风险评估并提供参考。除更直接的风险外，在人工智能系统的风险评估中还可以考虑广泛失业和环境影响等下游社会影响。有关风险评估最佳实践的更多信息，请参阅负责任扩展策略部分。

确保公平地使用前沿人工智能系统。研究人员公平透明地获取人工智能系统受限访问的过程很重要。为了确保系统得到适当的理解，可以特别注意促进研究的多样性，例如不基于先前或预期的批评而拒绝访问，并鼓励不同类型的学者和第三方机构研究人工智能系统。

4. 公开分享风险研究成果，除非分享这些成果可能会造成危害

如果共享没有足够实质性的负面影响，鼓励前沿人工智能机构广泛共享这项工作的成果。

重点案例

OpenAI：20%算力投入超级对齐(Superalignment)研究

OpenAI认为，需要科学突破、社会准备和先进的安全系统来控制和集成比人类聪明得多的人工智能系统，并正通过创建“超级对齐”(Superalignment)[33]和“防范准备”(Preparedness)[34]两个新团队来投资这些安全研究的突破。

当前对齐人工智能的技术，例如根据人类反馈进行强化学习，依赖于人类监督人工智能的能力。但这些技术不适用于超级智能，因为人类将无法可靠地监督比自己聪明得多的人工智能系统。OpenAI设定了一个目标，在四年内解决这个问题，建立了一个名为“超级对齐”的新团队，由Ilya Sutskever(OpenAI联合创始人兼首席科学家)和Jan Leike(对齐团队负责人)共同领导。目标是构建一个接近人类水平的对齐研究人工智能，并使用大量计算来扩展OpenAI对齐超级智能的努力。OpenAI计划将其在2023年6月之前所获得的算力的20%用于超级对齐工作[35]。团队将广泛分享结果，以促进非OpenAI模型的对齐和安全性。

除了对齐超级智能的挑战之外，OpenAI认为日益增强的前沿模型的滥用可能会带来越来越严重的风险。OpenAI还创建了一个名为“防范准备”的专门新团队来识别、跟踪和准备应对这些风险。OpenAI计划跟踪前沿风险，包括网络安全、化学/生物/辐射/核威胁(CBRN)、说服、自主复制和适应，并分享行动以防范灾难性风险的影响。由于对灾难性风险的实证理解还处于萌芽阶段，OpenAI将迭代更新对当前前沿模型风险水平的评估，以确保反映其最新的评估和监测理解。

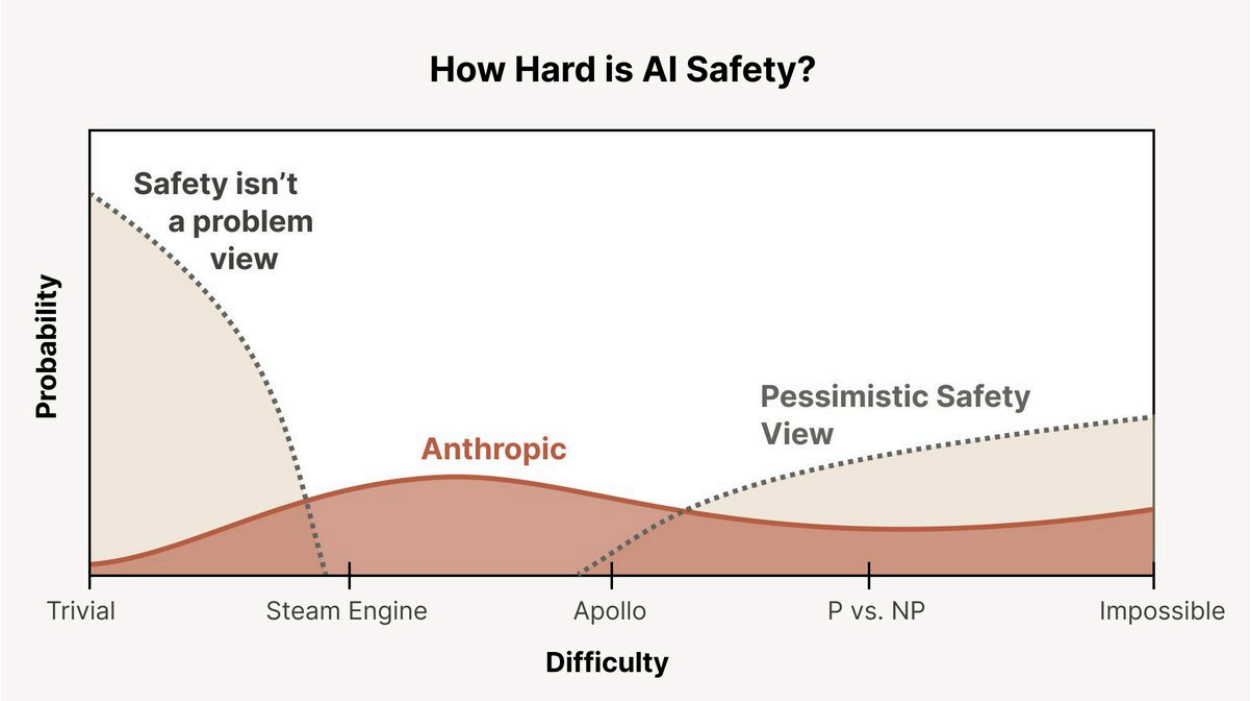
---

[33] Jan Leike & Ilya Sutskever, “Introducing Superalignment”, 2023-07-05, https://openai.com/blog/introducing-superalignment.
[34] OpenAI, “Frontier risk and preparedness”, 2023-10-26, https://openai.com/blog/frontier-risk-and-preparedness.
[35] OpenAI, “OpenAI's Approach to Frontier Risk”, 2023-10-26, https://openai.com/global-affairs/our-approach-to-frontier-risk#priority-research-and-investment-on-societal-safety-and-security-risks.

OpenAI表示将继续投资于网络安全和内部威胁防护措施，以保护专有和未发布模型的权重。他们启动了网络安全资助计划和OpenAI漏洞赏金计划，以协调志同道合的研究人员为人类的集体安全而努力。网络安全资助计划是一项价值100万美元的计划，旨在增强和量化人工智能驱动的网络安全能力，并促进高水平的人工智能和网络安全讨论。OpenAI还邀请公众报告他们在系统中发现的漏洞、错误或安全缺陷，并为作出贡献的个人和企业提供认可和奖励。

Anthropic：对多元化和经验驱动的AI安全方法最为乐观

Anthropic认为各种场景都是可能的，而非采取坚定立场。Anthropic认为不确定性的一个特别重要的方面，是开发广泛安全且对人类风险很小的先进AI系统的困难程度。开发这样的系统的难度可能介于非常容易到不可能之间的任何位置[36]。



AI安全的难度？[37]

Anthropic将难度范围分为三个非常不同的场景：乐观场景、中间场景、悲观场景。其目标是开发：1）使AI系统更安全的技术，2）识别AI系统安全或危险程度的方法。

- 乐观场景中，前者将帮助AI开发者训练有益的系统，后者将证明此类系统是安全的。
- 中间场景中，前者可能是人类最终避免AI灾难的方式，后者将对确保高级AI的风险较低至关重要。

[36] 安远AI，“Anthropic关于AI安全的核心观点：何时、何故、何事与如何”，2023-04-27，https://mp.weixin.qq.com/s/UL0BK3s2CXVXUivhzU5ZKw.
[37] Chris Olah，“How Hard is AI safety？”，https://threadreaderapp.com/thread/1666482929772666880.html.

- 悲观场景中，前者的失败将是AI安全性不可解决的关键指标，后者将使Anthropic能够有说服力地向他人证明这一点。

Anthropic正采取多种研究来建立可靠的安全系统。Anthropic的研究项目被分成能力(Capabilities)、对齐能力(Alignment Capabilities)和对齐科学(Alignment Science)这3个领域，目前最为关注的方向是：机制可解释性(Mechanistic Interpretability)、可扩展的监督(Scalable Oversight)、面向过程的学习(Process-Oriented Learning)、理解泛化(Understanding Generalization)、检测危险的失败模式(Testing for Dangerous Failure Modes)、社会影响和评估(Societal Impacts and Evaluations)。

Anthropic的一个关键目标是加速安全研究的发展，并尝试覆盖更广泛的安全研究范围，从那些安全挑战容易解决的场景到那些创建安全系统极为困难的场景。

延伸阅读

谷歌DeepMind：积极投资更广泛的AI安全研究和生态建设

- 谷歌DeepMind有多个团队全职研究人工智能伦理、安全和治理，这些团队旨在了解和缓解当前系统的已知风险和更强大系统的潜在风险，并使它们符合人类利益。谷歌和DeepMind还支持更广泛的AI安全研究和生态建设。
- **数字未来项目：**谷歌于2023年9月宣布启动其中包括一项**2000万美元的基金**，该基金将为研究和鼓励负责任的人工智能开发的领先智囊团和学术机构提供资助。这些机构正在研究人工智能对全球安全的影响等问题；对劳动力和经济的影响；什么样的治理结构和跨行业努力可以最好地促进人工智能创新的责任和安全[38]。
- **AI安全基金**：2023年10月，谷歌、微软、OpenAI和Anthropic发布联合声明，任命“前沿模型论坛”(Frontier Model Forum)首任执行董事，并**宣布设立1000万美元的AI安全基金**，以推动正在进行的工具研发，帮助社会有效地测试和评估最有能力的AI模型[39]。

[38] Brigitte Hoyer Gosselink，“Launching the Digital Futures Project to support responsible AI”，2023-09-11, https://blog.google/outreach-initiatives/google-org/launching-the-digital-futures-project-to-support-responsible-ai/.
[39] OpenAI，“Frontier Model Forum updates”，2023-10-25, https://openai.com/blog/frontier-model-forum-updates.

国内外顶尖科学家：多次呼吁30%以上的研发投入用于AI安全研究

- **重磅论文：**2023年10月24日，三位图灵奖获得者、一位诺贝尔奖获得者、国内多位院士共同撰文《人工智能飞速进步时代的风险管理》[40]，文章提出分配至少三分之一的人工智能研发资金用于确保人工智能系统的安全性和合乎伦理的使用（与其对人工智能能力的投资相当）。
- **联合声明：**2023年10月18-20日，图灵奖获得者Yoshua Bengio和姚期智、加州大学伯克利分校教授Stuart Russell以及清华大学智能产业研究院院长张亚勤联合召集了来自中国、美国、英国、加拿大和其他欧洲国家的20多位顶尖AI科学家和治理专家，在为期三天的首届“人工智能安全国际对话”后，签署了一份联合声明[41]，再次强调研发预算至少30%应投入AI安全研究、通过模型注册来监测前沿AI的发展等。

国内/华人团队：在大模型安全方面已开展了一系列的研究

包括但不限于以下工作：

- 通过ChatGPT和RLHF，国内研究团队开始重视对齐问题。
  - 清华大学[42]、中国人民大学[43]、微软亚洲研究院[44]、华为[45]等国内团队发布涉及对齐的综述文章，主要围绕现阶段较为成熟的RLHF等方法及其相关改良。
  - 天津大学[46]、北京大学[47]的团队也发布了涉及更广范围的对齐研究的综述文章。
- 多个国内/华人团队正在对RLHF和大语言模型监督方法进行了创新和改良：
  - 阿里达摩院和清华大学的团队[48]提出RRHF(Rank Responses to align Human Feedback)方法，无需强化学习即可用于训练语言模型。

---

[40] 安远AI，“授权中译版｜三位图灵奖和中外多位顶尖AI专家的首次政策建议共识：呼吁研发预算1/3以上投入AI安全，及若干亟需落实的治理措施”, 2023-10-24, https://mp.weixin.qq.com/s/zdrGCiagDYqa6kPljK2ung.
[41] 安远AI，“AI的帕格沃什会议！中美英加欧20多位顶尖AI专家线下聚首，呼吁AI安全与治理的全球协同行动”, 2023-11-01, https://mp.weixin.qq.com/s/1WbrS-L8Qsww10nosADwJQ.
[42] Jiawen Deng et al.，“Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements”, 2023-02-18, https://arxiv.org/abs/2302.09270.
[43] Wayne Xin Zhao et al.，“A Survey of Large Language Models”, 2023-05-31, https://arxiv.org/abs/2303.18223.
[44] Jing Yao et al.，“From Instructions to Intrinsic Human Values -- A Survey of Alignment Goals for Big Models”, 2023-08-23, https://arxiv.org/abs/2308.12014.
[45] Yufei Wang et al.，“Aligning Large Language Models with Human: A Survey”, 2023-07-24, https://arxiv.org/abs/2307.12966.
[46] Tianhao Shen et al.，“Large Language Model Alignment: A Survey”, 2023-09-26, https://arxiv.org/abs/2309.15025.
[47] Jiaming Ji et al.，“AI Alignment: A Comprehensive Survey”, 2023-10-30, https://arxiv.org/abs/2310.19852.
[48] Zheng Yuan et al.，“RRHF: Rank Responses to Align Language Models with Human Feedback without tears”, 2023-04-11, https://arxiv.org/abs/2304.05302.

    - 香港科技大学的团队[49]引入了新的RAFT(Reward rAnked FineTuning)方法，旨在更有效地对齐生成模型。
    - 北京大学的团队[50]开源了PKU-Beaver项目，结合约束强化学习(Constrained RL)，提出具有更强安全性保障的Safe RLHF。
    - 另一北京大学的团队与阿里合作[51]提出PRO(Preference Ranking Optimization)方法，把人类偏好从二元比较推广到多元排序。
- 模型对齐评测，是从安全与伦理角度审视模型倾向，国内2023年陆续发布：
    - 清华大学的CoAI团队[52]的中文大模型安全评测平台，评估生成式语言模型的安全伦理问题。
    - 北京智源人工智能研究院[53]的FlagEval天秤大模型评测，含中文世界安全和价值对齐指令评测。
    - 阿里巴巴[54]的CValues基准，面向中文大模型的价值观评估与对齐研究。
    - 上海人工智能实验室的安全评测基准[55]和对齐评测流程[56]，目标构建“评测框架、评测数据、评测工具、对齐提升”的大语言模型提升框架，强调系统性与全流程。
- 国内在鲁棒性方面有较多研究，越狱提示是前沿大模型的新兴鲁棒性挑战：
    - 另一清华大学的团队[57]攻破GPT-4V、谷歌Bard等模型，展示了有视觉模态输入的大模型对于攻击可能更加脆弱。
- 为了让有志深耕大模型安全和对齐相关领域的读者抓住宝贵的实践机会，我们汇总了国内/华人团队相关的研究和实习机会[58]。

---

[49] Hanze Dong et al., “RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment”, 2023-04-13, https://arxiv.org/abs/2304.06767.
[50] PKU Alignment, “Constrained Value-Aligned LLM via Safe RLHF”, 2023-05-15, https://pku-beaver.github.io/.
[51] Feifan Song et al., “Preference Ranking Optimization for Human Alignment”, 2023-06-30, https://arxiv.org/abs/2306.17492.
[52] 清华大学计算机科学与技术系CoAI小组, “中文大模型安全评测平台”, 2023-04-20, http://115.182.62.166:18000.
[53] 智源研究院, “FlagEval(天秤)大模型评测体系及开放平台”, 2023-06-09, https://flageval.baai.ac.cn.
[54] Guohai Xu et al., “CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility”, 2023-07-19, https://arxiv.org/abs/2307.09705.
[55] Kexin Huang et al., “Flames: Benchmarking Value Alignment of Chinese Large Language Models”, 2023-11-12, https://arxiv.org/abs/2311.06899.
[56] Yixu Wang et al., “Fake Alignment: Are LLMs Really Aligned Well?”, 2023-11-10, https://arxiv.org/abs/2311.05915.
[57] 机器之心, “清华团队攻破GPT-4V、谷歌Bard等模型，商用多模态大模型也脆弱？”, 2023-10-17, https://mp.weixin.qq.com/s/ijbYc-obIFJx2Ho8xkLjLg.
[58] 安远AI, “行动指南：安远AI推荐大模型安全、对齐和评测科研机会”, 2023-10-10, https://mp.weixin.qq.com/s/16iTd4XpNFs6HJZh-3Bnig.

三、含保护模型权重在内的安全控制

摘要

为了确保前沿人工智能的安全，考虑网络安全、防护性安全风险管理和内部风险缓解至关重要。无论是模型还是部署它们的系统，都必须从开发之初就考虑网络安全，以确保人工智能的益处得以实现。网络安全是人工智能系统安全性、可靠性、可预测性、伦理性和潜在监管合规性的关键支撑。

为避免危及安全或敏感数据，考虑人工智能系统以及独立模型的网络安全，并在整个人工智能生命周期中实施网络安全流程尤为重要，特别是当该组件是其他系统的基础时。随着人工智能系统的进步，开发者必须保持对可能的攻击的认识，识别漏洞并实施缓解措施。否则，未来的人工智能模型和系统将面临设计漏洞的风险。“设计安全”(Secure by Design)[59]方法使开发者能够从设计和开发之初就将安全性“融入”其中。

网络安全必须与物理安全和人员安全整体考虑。制定连贯、整体、基于风险和相称的安全策略，并由有效的治理结构所支持，至关重要。如果人工智能系统被入侵可能导致有形或广泛的物理损害、重大业务中断、敏感或机密信息泄露、声誉受损和/或法律挑战，那么应将人工智能安全风险视作关键任务。

我们概述了关于安全控制的8类实践措施，包括保护模型权重：

1. 在整个人工智能系统（包括基础设施和供应链）中实施强有力的网络安全措施和流程（包括安全评估）
2. 了解人工智能系统中的资产（包括训练数据和模型权重）并采取适当措施来进行保护
3. 保持对安全风险（包括针对和来自人工智能系统的新威胁）的最新理解，以便做出明智的风险决策
4. 制定事件响应、升级和补救计划，并确保响应人员受过评估和应对相关事件的培训
5. 对系统行为进行持续监测，以便观察行为变化并识别潜在攻击
6. 通过评测和沟通风险并遵循“设计安全”原则，使用户能够安全使用人工智能系统
7. 实施有效的防护性安全风险管理，涵盖物理、人员和网络安全纪律
8. 制定并实施适当的人员安全控制措施以降低内部风险

[59] Cybersecurity and Infrastructure Security Agency, “Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Security-by-Design and -Default”, 2023-04-13, https://www.cisa.gov/sites/default/files/2023-04/principles_approaches_for_security-by-design-default_508_0.pdf.

背景

随着人工智能和机器学习系统的使用不断增长，必须安全地开发和部署系统，以避免危及安全或数据。重要的是要考虑人工智能模型和系统的网络安全，并在整个人工智能生命周期中实施网络安全流程，持续监测安全性。

《中华人民共和国网络安全法》[60]是中国在2017年实施的基础性网络安全立法，旨在规范网络空间的安全管理。该法律的主要内容包括了网络运营安全、个人信息保护、关键信息基础设施保护、网络产品和服务安全、数据跨境传输等方面。

《网络安全等级保护制度2.0国家标准》[61]（简称“等保2.0标准”），是《网络安全法》框架下实施的一个重要制度。“等保2.0标准”对国家重要信息、法人和其他组织及公民的专有信息以及信息和存储、传输、处理这些信息的信息系统分等级实行安全保护，对信息系统中使用的信息安全产品实行按等级管理，对信息系统中发生的信息安全事件分等级响应、处置。

现有的网络安全法规和实践为人工智能安全提供了坚实的基础。同时，还有其他措施可以帮助解决人工智能系统和机器学习模型工作方式中固有的安全弱点。除了标准网络安全故障模式（例如利用底层软件堆栈中的传统漏洞）之外，考虑人工智能特定故障的可能性也很重要。

实践解读

1. 在整个人工智能系统（包括基础设施和供应链）中实施强有力的网络安全措施和流程（包括安全评估）

将良好的基础设施原则应用于从设计到退役的整个过程的每个部分。这一点很重要，因为在人工智能项目生命周期的不同阶段都可能发生威胁。

定期评估供应链的安全性，确保供应商与机构自身遵守相同的标准。确保从可信来源获得数据、软件组件和硬件将有助于降低供应链风险。

根据威胁建模得出的优先级，评测模型对不同类别的对抗攻击（例如中毒攻击、模型反转和模型窃取）的鲁棒性。这可能涉及基准测试和红队测试的结合。

评估并记录针对人工智能系统整体的网络安全威胁，并缓解漏洞的影响。良好的文档和监测将告知总体风险态势，并帮助机构在安全事件发生时做出响应。

[60] 中华人民共和国中央人民政府，“中华人民共和国网络安全法”,2016-11-07, https://www.gov.cn/xinwen/2016-11/07/content_5129723.htm.
[61] 国家标准委，“信息安全技术 网络安全等级保护基本要求”, 2019-05-10, https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=BAFB47E8874764186BDB7865E8344DAF.

2. 了解人工智能系统中的资产并采取适当措施来进行保护

了解人工智能相关资产，例如模型、数据（包括用户反馈）、提示、软件、文档和评估（包括有关潜在不安全能力和故障模式的信息）对其机构的价值。根据适当方式保护这些不同类别的信息。

确保所有业务决策都考虑到安全性，并通过适当的网络、物理和人员安全措施来识别和保护与人工智能相关的资产。

拥有用于跟踪、验证、保护和版本控制资产的流程和工具，并能够在发生泄露时回滚到已知的安全状态。这一点很重要，因为在人工智能项目的整个生命周期中，数据和模型并不保持静态。

使用数据卡(data cards)、模型卡(model cards)和软件物料清单(SBOM)等常用结构记录数据、模型、提示、评估材料等资产。这将使机构能够快速轻松地识别和共享关键信息，包括特定的安全问题。

3. 确保开发者和系统所有者保持对安全风险的最新理解，以便做出明智的风险决策

对开发者、系统所有者和高级领导者进行人工智能安全实践培训。建立积极的安全文化至关重要。在这种文化中，领导者展示良好的安全实践，并且整个项目的员工对人工智能安全有足够的了解，以了解他们所做决策的潜在后果。

保持对安全威胁和故障模式的认知，特别是对于数据科学家和开发者。人工智能开发和网络安全是两种不同的技能，建立一支兼具这两种技能的团队需要付出努力和时间。

对系统面临的威胁进行建模，以了解模型被滥用或出现意外行为时，对系统、用户、机构和更广泛社会的影响。这可以帮助构建系统，其中数据流水线的其他部分可以安全地处理意外的模型输出。

4. 制定事件响应、升级和补救计划，并确保响应人员受过评估和应对人工智能相关事件培训

制定机构事件响应计划。系统不可避免地受到安全事件影响，这体现在组织事件响应计划中。精心计划的响应将有助于最大限度地减少攻击造成的损失并支持恢复。

5. 对系统行为进行持续监测，以便观察行为变化并识别潜在攻击

持续测量人工智能模型和系统的性能。模型性能下降可能表明遭受攻击，也可能表明模型遇到的数据与其训练数据不同。无论哪种方式，可能都需要进一步调查。

监测并记录人工智能系统的输入，以便在出现泄露时进行审核、调查和补救。一些针对人工智能系统的攻击依赖于重复查询。正确的日志记录将帮助审核系统并识别任何异常输入。

采取措施缓解和修复问题，并记录任何与人工智能相关的安全事件和漏洞。

6. 通过评测和沟通风险并遵循“设计安全”原则，使用户能够安全使用人工智能系统

明确告知用户其负责的安全元素以及数据可能被使用或访问的方式和地点。

在产品中默认集成最安全的设置。在需要配置的情况下，默认选项应能够抵御常见威胁。

确保每个产品中默认包含必要的安全更新，并使用安全的模块化更新程序来分发它们。这将帮助产品在面对新的和不断发展的威胁时仍然保持安全。

只有在经过安全评估（包括基准测试和红队测试）后才发布模型。

保持开放的沟通渠道，以便就产品安全提供反馈意见，包括安全研究人员报告漏洞并为此获得法律安全港的机制，以及将问题升级到更广泛的社区的机制。帮助共享知识和威胁信息将增强整个社区应对人工智能安全威胁的能力。

7. 实施有效的防护性安全风险管理，涵盖物理、人员和网络安全纪律

连同明确的治理和监督，有效的防护性安全风险管理的关键步骤包括但不限于：

- 识别对于有效运营至关重要或具有特定机构价值（例如商业机密）的资产和系统。
- 对资产进行分类分级，以确保在实施风险缓解措施时使用适当级别的资源。
- 识别威胁。这些可能包括恐怖主义或敌对国家威胁，和/或更本地化和具体的威胁，并使用各种内部和外部资源。
- 使用公认的流程评估风险。
- 建立防护性安全风险登记册，以详细记录在此风险管理流程中收集的所有数据，确保与现有机构风险管理登记册和流程兼容。
- 制定用于缓解已识别风险的保护性安全策略，审查与风险优先级列表相关的防护性安全措施。如果缓解措施被评估为“不充分”，可提出额外措施供决策者批准。
- 制定开发和实施计划。旨在制定一份清晰、优先的防护性安全缓解措施列表，涵盖物理、人员和网络安全纪律，并与实施这些措施所需的技术指南相关联。
- 定期并在需要时审查风险管理措施，例如根据威胁或操作环境的变化，或评估所实施的新措施的适用性。

8. 制定并实施适当的人员安全控制措施以降低内部风险

缓解内部风险的关键步骤包括但不限于：

- 确保董事会对防护安全负最终责任，定期与全业务的关键利益相关者接触，并对组织面临的风险有坚实的了解。确保整个业务的利益相关方参与，获得专业洞察和制定以及实施内部风险缓解方案。
- 根据基于角色的风险评估，对所有有权使用机构资产的个人（包括全职员工、临时员工和合同工）应用适当水平的审查。
- 使用基于角色的安全风险评估来识别需要应用的物理、人员或网络安全措施。

- 制定适当的政策、清晰的报告程序，以及易于理解且始终如一执行的升级准则。
- 为所有员工提供适当的安全教育和培训。如果没有有效的教育和培训，就不能期望个人了解维持安全的程序。
- 确保制定监测和审查方案，以使潜在的安全问题或可能影响员工工作的个人问题能够在其职业生涯中得到识别和有效处理。
- 使用既定的、基于证据的指南来全面解决人事安全风险。

重点案例

Anthropic：主张加强前沿人工智能研发机构的网络安全控制，并呼吁政府加强监管

前沿人工智能模型可能会影响经济安全和社会稳定。鉴于这项技术的战略性质，前沿人工智能研究和模型必须实现远超过其他商业技术标准的水平，以保护其免遭盗窃或滥用。

在短期内，政府和前沿人工智能研发机构必须做好保护先进模型和模型权重的准备，并为其研究提供支持。这应包括制定在行业中广泛传播的稳健的最佳实践等措施，以及将前沿人工智能领域视为“关键基础设施”，以确保这些模型和机构的公私伙伴合理研发。许多实践措施可以作为自愿承诺起步，随着时间推演，可能会成为政府监管政策。[62]

Anthropic正在实施两方控制、SSDF、SLSA和其他网络安全最佳实践。

“两方控制”对于保护前沿人工智能系统是必要的。两方控制已在一系列领域得到应用，例如需要两个人用两把钥匙才能打开最安全的金库。多方审核模式也已应用于制造业(ISO 9001[63])、食品(ISO 22000[64])和医疗(ISO 13485[65])等。

- 该模式应该应用于涉及前沿人工智能模型的开发、训练、托管和部署的所有系统。
- 该模式已在大型科技企业中广泛使用，以防御最先进的威胁行为者并降低内部风险。
- 这种模式的系统设计下，没有人能够持久访问生产关键型环境，并且他们必须要求与同事进行限时访问，并提供该请求的业务理由。
- 即使是新兴的人工智能机构，在没有大量资源的情况下，也可以实施这些控制手段。
- Anthropic将这种**多方授权(multi-party authorization)称为人工智能关键基础设施设计**。这是一项领先的安全要求，取决于能否正确实施各种网络安全最佳实践。

---

[62] Anthropic, “Security controls including securing model weights”, 2023-10-05, https://www.anthropic.com/uk-government-internal-ai-safety-policy-response/security-controls-including-securing-model-weights.

[63] ISO, “ISO 9001:2015: Quality management systems: Requirements”, 2023-12-18(引用日期), https://www.iso.org/iso-9001-quality-management.html.

[64] ISO, “ISO 22000: Food safety management”, 2023-12-18(引用日期), https://www.iso.org/iso-22000-food-safety-management.html.

[65] ISO, “ISO 13485:2016: Medical devices: Quality management systems: Requirements for regulatory purposes”, 2023-12-18(引用日期), https://www.iso.org/standard/59752.html.

安全的软件开发实践应该扩展到前沿人工智能模型环境中。Anthropic参考的是美国国家标准与技术研究院(NIST)的安全软件开发框架(SSDF)[66]和软件工件供应链级别(SLSA)[67]。虽然前沿人工智能研究已经通过云提供商托管其模型而受益于其中一些标准的实施，但应用这些现有标准可以逐步改变这些人工智能系统的安全性：

- SSDF和SLSA在很大程度上可以转化为模型的开发及其耦合软件；模型的开发和部署于软件的开发和部署几乎相同。
- SSDF和SLSA结合在一起意味着部署任何人工智能系统都有一个监管链。如果正确应用这些实践措施，就能够将已部署的模型与其研发机构联系起来，这有助于溯源。
- Anthropic称之为**安全模型开发框架**，并鼓励扩展SSDF，以涵盖NIST标准制定流程内的模型开发。

短期内，这两种最佳实践可作为适用于与政府签订合同的人工智能企业和云提供商的采购要求，并适用于这些企业的标准网络安全实践。例如美国云服务商提供许多当前前沿模型研发机构的基础设施，采购要求将产生类似于市场监管的效果，并且可以在监管要求出现前发挥作用。随着模型能力的扩展，Anthropic将需要进一步增强安全保护，这将是一个与政府和行业协商的迭代过程。

前沿人工智能研究实验室应该像金融服务等关键基础设施领域的企业一样参与公私合作。例如该部门可以被指定为现有IT部门的特殊子部门。这样的指定将成为行业实验室和政府机构之间加强合作和信息共享的工具，帮助所有实验室更好地防范资源丰富的恶意行为者。

人们很容易降低安全的优先级：当一切进展顺利时，人们可能会觉得安全没有必要，或者与企业的其他目标存在冲突。但这项技术正变得越来越强大，需要加强预防措施。Anthropic认为，虽然安全有时会干扰生产力，但有一些创造性的方法可以确保其影响有限，并且研究和其他工作可以有效进行。

延伸阅读

微软：整体出色，但还可通过多方授权等机制对保护模型权重做出更大承诺[68]

- 改进了安全开发生命周期(Security Development Lifecycle, SDL)，以链接其负责任人工智能标准并集成其中的内容，加强与负责任人工智能标准要求的治理步骤保持一致的流程并加强检查。

[66] NIST, “Secure Software Development Framework”, 2023-01-10, https://csrc.nist.gov/Projects/ssdf.
[67] SLSA, “Safeguarding artifact integrity across any software supply chain”, 2023-04-01, https://slsa.dev/.
[68] Microsoft, “Microsoft's AI Safety Policies”, 2023-10-26, https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/#Security_Controls,_Including_Securing_Model_Weights.

- 更新了SDL威胁建模要求的内部实践指南，该指南适用于所有产品并可能涉及红队测试，以解释微软对人工智能和机器学习特有的独特威胁的持续了解。
- 宣布了新的人工智能系统漏洞严重性分类，涵盖了因在产品和服务中使用人工智能而特别产生的新漏洞类别。
- 确保其人工智能产品和服务的网络安全风险得到识别和缓解。
- 实施NIST AI风险管理框架(白宫企业自愿承诺之外的更多承诺)。

亚马逊：核心亮点是其数据中心的物理安全措施[69]

- **AWS通过控制来保证数据中心的安全。**在构建数据中心之前，AWS会全面评估潜在威胁，然后设计、实施和测试控制措施，以确保部署的系统、技术和人员能够抵御这些风险。
- **对AWS数据中心的物理访问仅限于特定工作职能范围内经过筛选并获得批准的人员。**任何工作不涉及数据中心内部日常工作且需要访问数据中心的员工必须首先申请并提供有效的业务理由。访问请求根据最小权限原则进行评估：只有当员工的业务理由有效时，才允许他们进入数据中心，并且他们的访问权限暂时限制在他们需要访问的数据中心的特定层。
- **对AWS数据中心的物理访问会被记录、监测和保留。**AWS将从逻辑和物理监测系统获得的信息关联起来，以根据需要增强安全性。AWS安全运营中心定期对其的数据中心进行威胁和漏洞审查。通过数据中心风险评估活动对潜在漏洞进行持续评估和缓解。AWS使用全球安全运营中心来监测其的数据中心，该中心负责监测、分类和执行安全计划。他们通过管理和监测数据中心访问活动提供24/7全球支持，帮助本地团队和其他支持团队通过分类、咨询、分析和调度响应来应对安全事件。

中国国务院：发布《关键信息基础设施安全保护条例》[70]

- **这是我国首部针对关键信息技术设施安全保护工作的行政法规。**2021年7月30日，第745号国务院令公布《关键信息基础设施安全保护条例》，自2021年9月1日起施行。
- **关键信息基础设施是网络安全保障的核心。**《网络安全法》和《关键信息基础设施安全保护条例》都采用了**“领域识别+风险识别”**的方式来定义关键基础设施：“本条例所称关键信息基础设指公共通信和信息服务、能源、交通、水利、金融、公共服务、电子政务、国防科技工业等重要行业和领域的；以及其他一旦遭到破坏、丧失功

[69] Amazon, “AI Safety Summit - Enhancing Frontier AI Safety”, 2023-10-26, https://aws.amazon.com/cn/uki/cloud-services/uk-gov-ai-safety-summit/#Security_Controls.
[70] 中华人民共和国中央人民政府, “关键信息基础设施安全保护条例”, 2021-07-30, https://www.gov.cn/zhengce/content/2021-08/17/content_5631671.htm.

能或者数据泄露，可能严重危害国家安全、国计民生、公共利益的重要网络设施、信息系统等。”

- **制定认定规则主要考虑的三大因素：**1）网络设施、信息系统等对于本行业、本领域关键核心业务的重要程度；2）网络设施、信息系统等一旦遭到破坏、丧失功能或者数据泄露可能带来的危害程度；3）对其他行业和领域的关联性影响。
- **我们建议将其扩展到前沿人工智能风险管理。**扩大《关键信息基础设施安全保护条例》的适用范围，明确将人工智能关键基础设施纳入监管保护。

四、漏洞报告机制

摘要

即使前沿人工智能机构部署了人工智能系统，系统仍然可能存在未识别的安全和安保问题（即“漏洞”）。为了解决这些漏洞，必须首先识别并让前沿人工智能机构尽快意识到漏洞。

建立漏洞管理流程可以使外部人员识别和报告任何漏洞。这有助于确保安全和安保问题尽快被告知给前沿人工智能机构，以便他们能够快速解决。

我们概述了关于漏洞报告机制的3类实践措施：

1. 建立漏洞管理流程
2. 借鉴已建立的软件漏洞报告流程，建立清晰、用户友好且公开的模型漏洞报告流程
3. 制定协同漏洞披露和信息共享的协议和机制

背景

人工智能模型可能存在意外的漏洞。例如尽管开发人员不断修补模型，但用户一再能够通过特定的提示来进行大语言模型的“越狱”，迫使模型以与开发者意图相悖的方式行事，这可能是有害的。随着模型功能的提升，解决此类漏洞的重要性也将随之增加。如果此类问题仍然存在并且模型能力足够危险，则可能需要对模型的使用施加重大限制。

提高开发者解决漏洞的能力的一种方法是激励机构外的用户和研究人员帮助解决该问题，作为稳健评估和风险评估流程的补充。完善的漏洞报告协议通常会将相关组织的通知与延迟一段时间后的公开披露相结合。这样，受影响的机构就有动力解决漏洞。

鉴于传统软件漏洞和人工智能模型漏洞之间的差异，漏洞赏金模型无法直接应用于模型漏洞。与传统的软件漏洞相比，模型漏洞一旦被发现，如何修复可能不明确，这可能导致公开披露漏洞不太合适。其次，与传统的软件漏洞相比，提前指定什么构成模型漏洞可能特别困难。

实践解读

1. 建立漏洞管理流程

这一流程可以具有尽可能广泛的范围，确保前沿人工智能机构能够适当响应漏洞报告。该流程可以接受有关任何类别的模型漏洞及其利用方法的报告，包括：

- **越狱方法**：用于诱导模型绕过审核功能或模型内在安全性倾向的方法，前沿人工智能机构试图已通过使用过滤器或微调来试图阻止这种情况

- **提示注入攻击**：恶意行为者用于诱导模型表现出他们想要的行为的方法，通过向模型呈现包含执行这些行为指令的提示。
- **隐私攻击**：从模型中提取本应为私密的信息的方法（来自训练数据或用户与模型的私人对话的敏感信息）
- **未解决的潜在滥用**：利用模型能力造成危害的方法，这些方法尚未得到解决
- **不对齐(misalignment)**：当模型应用其能力的方式与用户的意图或需求大不相同时
- **中毒攻击**：当对手操纵训练数据以降低模型性能时
- **偏见和歧视**：当模型表现出关于已知受保护特征的特定偏见或歧视时
- **性能问题**：当模型在其部署时表现不佳，例如医疗聊天机器人提供不正确的信息并对患者造成危害

2. 借鉴已建立的软件漏洞报告流程，建立清晰、用户友好且公开的模型漏洞报告流程

这些流程可以内置在机构为接收传的统软件漏洞报告而构建的流程中，或从中汲取灵感。这些政策必须公开并有效发挥作用，这一点至关重要。

3. 制定协同漏洞披露和信息共享的协议和机制

考虑共享漏洞信息如何能够同时加剧和减轻风险。共享可以提醒潜在的攻击者，但也可以提醒潜在的受害者和有能力建立防御的行为者。一个特别重要的因素是修复模型漏洞的难易程度。如果修复一个漏洞需要很长时间，或者无法修复，那么公开报告可能是不合适的。

制定并公开描述决定如何共享模型漏洞信息的协议。例如这些协议可以概述根据所识别的危害或漏洞类型，以及与不同参与方共享信息的条件。

建立机制，向相关政府机构、执法部门和其他受影响的机构披露有关漏洞的信息。这对于公开披露可能会增加危害风险的漏洞，尤其重要。

公开分享从模型漏洞报告计划中汲取的一般教训。这可能包括关于所面临挑战和不同激励策略相对有效性的教训。这样的共享可以通过类似于国家网络与信息安全信息通报中心漏洞平台的机制来完成。

重点案例

微软：协同漏洞披露领域的行业领导者[71]

微软一直是协同漏洞披露(Coordinated Vulnerability Disclosure, CVD)领域的行业领导者。该流程中，供应商从外部发现者处接收有关影响其产品和服务的潜在漏洞的信息，并与这些发现者合作调查和缓解已确认的漏洞，并以最大限度地降低用户风险的方式公开发布相关信息。

微软已经制定并公开了CVD政策[72]，建立了接受外部研究人员漏洞报告并在整个调查、修复和公开过程中与他们合作的明确流程，包括归功于发现者。微软安全响应中心(MSRC)[73]接收来自外部发现者的所有此类报告，并管理整个CVD流程的协调，与内部其他人合作进行适当的调查和补救。

作为漏洞赏金计划[74]的一部分，发现关键漏洞的外部研究人员还有机会获得经济奖励。奖金范围因产品而异，其中云计划最高可达10万美元，平台计划最高可达25万美元。微软还启动了一个新的AI漏洞赏金计划，最高奖金为1.5万美元，旨在奖励发现人工智能驱动的必应所存在的漏洞。

确认漏洞后，MSRC会与外部研究人员和相关产品团队合作开发、测试和发布补救措施，通常涉及软件更新。在此过程中，微软使用漏洞严重性等级来确定需补救工作的优先级，首先关注最关键的问题，而非按问题接收或或确认的顺序。

为了提高透明度，微软提供漏洞分类和严重性评级，包括不同类别的产品。例如MSRC维护在线服务的漏洞严重性分类[75]。微软还**发布了新的人工智能系统漏洞严重性分类**，涵盖因在产品和服务中使用人工智能而专门产生的新漏洞类别[76]。MSRC还继续维护安全更新严重性评级系统，该系统与产品类别严重性分类系统（即低、中、重要和严重严重性评级）保持一致，支持客户了解风险并确定更新的优先级[77]。

---

[71] Microsoft, “Microsoft's AI Safety Policies”, 2023-10-26, https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/#Reporting_Structure_for_Vulnerabilities_Found_after_Model_Release.

[72] Microsoft, “Microsoft's Approach to Coordinated Vulnerability Disclosure”, 2023-12-18(引用日期), https://www.microsoft.com/en-us/msrc/cvd.

[73] Microsoft Security Response Center, “Report an issue”, 2023-12-18(引用日期), https://msrc.microsoft.com/report/vulnerability/new?c=bounty.

[74] Microsoft, “Microsoft Bug Bounty Program”, 2023-12-18(引用日期), https://www.microsoft.com/en-us/msrc/bounty.

[75] Microsoft, “Microsoft Vulnerability Severity Classification for Online Services”, 2023-12-18(引用日期), https://www.microsoft.com/en-us/msrc/olsbugbar.

[76] Microsoft, “Microsoft Vulnerability Severity Classification for AI Systems”, 2023-12-18(引用日期), https://www.microsoft.com/en-us/msrc/aibugbar?rtc=1.

[77] Microsoft, “Security Update Severity Rating System”, 2023-12-18(引用日期), https://www.microsoft.com/en-us/msrc/security-update-severity-rating-system.

微软正在通过前沿模型论坛合作，制定与发现前沿模型中的漏洞或危险能力相关的“负责任的披露”流程指南。提供商可以通过该流程接收和共享与前沿模型中发现的漏洞或危险能力相关的信息。微软还致力于确保人工智能产品和服务的网络安全风险得到识别和缓解，并参与经批准的多利益相关方威胁信息交流，这些举措与微软向白宫所做的自愿承诺对齐，并扩展实施NIST人工智能风险管理框架。

延伸阅读

谷歌DeepMind：认为“部署后监测”和“报告漏洞和滥用”密切相关[78]

- **需要采取全面而系统性的方法来理解模型的能力和评估模型的影响。**谷歌认为，在模型投入使用之前和之后应用各种评估措施是必要的，因为人工智能系统的影响只能在特定的使用环境中才能得到最佳理解，从而相应地确定最合适的缓解措施。例如，大语言模型的实用性在很大程度上源自用户可以用自然语言进行交互，使其非常适合广泛集成到各种产品与服务中。因此，有必要在模型和产品级别解决安全漏洞。同时，前沿AI模型可能会展现出“涌现能力”，这些能力通常不是研发人员明确预期或轻易预测的，有时即使在模型投入使用后也可能被发现。
- **授权外部安全专家参与检测和披露其AI系统中的漏洞。**作为互联网公司中最早推出漏洞赏金计划[79]的公司之一，谷歌长期以来一直在其服务中与安全专家开展合作，以识别、报告和解决错误。随着该计划的持续发展，其“负责任披露”实践措施已成为大多数主要科技公司和许多政府机构遵循的行业规范。鉴于安全风险会随着时间推移而演变，谷歌认为提高系统的适应性至关重要，这与人工智能安全实践中需要应对动态的风险相似。
- **主动实施AI系统投入使用后的监测与报告工作的重要性。**谷歌使用专门的团队和流程来监测各种公开渠道，以收集、检测和分类系统面临的新威胁。谷歌还在探索各种其他方法来追踪旗舰AI产品对社会的影响。
- **需要采用行业协作的方式来识别、解决和分享见解与最佳实践。**许多与前沿AI模型相关的潜在风险在整个行业中很常见，作为AI合作伙伴关系的创始成员和AI事件数据库的赞助商之一，谷歌支持建立公共数据库，记录用户在部署AI系统时遇到的各类问题与事件。这有助于人工智能社区集中注意力来解决重要的现有问题并防范新的问题。

[78] Google Deepmind，“AI Safety Summit: An update on our approach to safety and responsibility Published”，2023-10-27, https://deepmind.google/public-policy/ai-summit-policies/.
[79] Google，“Welcome to Google's Bug Hunting community”，2021-07-27, https://bughunters.google.com/.

中国工信部、网信办、公安部：联合发布《网络产品安全漏洞管理规定》[80]

- **主要目的：**维护国家网络安全，保护网络产品和重要网络系统的安全稳定运行；规范漏洞发现、报告、修补和发布等行为，明确各方各类主体的责任和义务；鼓励各类主体发挥各自技术和机制优势开展漏洞发现、收集、发布等相关工作。
- **明确要求了网络产品具体的漏洞报告机制**，例如：
    - 第五条 网络产品提供者、网络运营者和网络产品安全漏洞收集平台应当建立健全网络产品安全漏洞信息接收渠道并保持畅通，留存网络产品安全漏洞信息接收日志不少于6个月。
    - 第六条 鼓励相关组织和个人向网络产品提供者通报其产品存在的安全漏洞。
    - 第七条 工业和信息化部网络安全威胁和漏洞信息共享平台同步向国家网络与信息安全信息通报中心、国家计算机网络应急技术处理协调中心通报相关漏洞信息。鼓励网络产品提供者建立所提供网络产品安全漏洞赏金机制，对发现并通报所提供网络产品安全漏洞的组织或者个人给予奖励。
- **我们建议将其扩展到前沿人工智能风险管理：**可借鉴已经建立的网络产品漏洞报告流程，建立清晰、用户友好且公开描述的接收模型漏洞报告的流程，制定协同漏洞披露和信息共享的协议和机制。

[80] 工业和信息化部等三部门，“关于印发网络产品安全漏洞管理规定的通知”，2021-07-22, https://www.gov.cn/zhengce/zhengceku/2021-07/14/content_5624965.htm.

五、人工智能生成材料的标识信息

摘要

由人工智能生成的某些内容与人类生成的内容难以区分。如果恶意行为者利用人工智能生成和传播有害或虚假信息，这可能会给公共安全带来风险。因此，能够区分人工智能生成内容和人类生成内容变得越来越重要。

人工智能标识信息可以帮助识别人工智能生成的内容，但实践中在技术上具有挑战性，目前也无法完全降低资源充足行为者的风险。虽然身份验证解决方案（例如“水印”）仍在开发中，但它们可能不被视为完全可靠，因为可能存在允许用户逃避检测的技术。

采用一系列标识信息机制，并投资于开发允许识别人工智能生成内容的技术可以减轻各种风险，包括但不限于与创建和传播欺骗性人工智能生成内容、内容偏见以及信息可信度损失有关的风险。

我们概述了关于人工智能生成材料标识信息的3类实践措施：

1. 研究能够识别人工智能生成内容的技术
2. 探索对各种扰动具有鲁棒性的人工智能生成内容的水印使用
3. 探索人工智能输出数据库的使用

背景

某些人工智能生成的内容可能与人类生成的内容难以区分。这会带来多种风险，包括恶意行为者创建和传播有害或虚假信息。许多参与方（包括个人、企业、机构和政府）将越来越依赖于区分人工智能生成内容和人类生成内容的能力。

人工智能标识信息是有助于区分人工智能生成内容和非人工智能生成内容的技术措施。其中包括水印和基于数据库的解决方案，例如在生成时记录输出。虽然人工智能标识信息有局限性，不能完全降低人工智能风险，但广泛采用可能意味着最常用的人工智能生成服务对于人工智能内容的有害使用会有一定的阻力，从而降低了将人工智能生成内容当作真实内容的动机。

然而，其中一些技术存在潜在挑战。一些水印技术可以很容易地被规避，特别是对于中高级的攻击者来说。围绕创建已知由人工智能生成的内容的数据库也存在隐私和安全挑战，尽管散列技术（在生成时为内容创建的唯一散列）有望缓解这些问题。纵观当前的技术水平，还没有任何特定的标识信息是不可能规避或可以保证在大规模上技术可行的。

不过，多种标识信息机制的共同实施，可能为创建和传播欺骗性的人工智能生成内容增加阻力和成本，并可能减轻无意的人工智能驱动的信息质量下降。

实践解读

1. 研究能够识别人工智能生成内容的技术

投资于为人工智能生成内容添加水印的研究（包括人工智能生成文本、照片和视频），并尝试实施此类技术。为文本添加水印和证明文本的来源尤其具有技术难度。研究如何为这些内容添加水印以及进行技术试验，可能有助于应对这些挑战。一种方法是使模型在统计上更有可能以人类无法察觉的方式使用某些短语或词语，但只要文本序列足够长，就可以被探测器捕捉到。然而，这种方法可能无法抵御尝试清除水印的行为，例如通过让另一个模型改述文本。

2. 探索对各种扰动具有鲁棒性的人工智能生成内容的水印使用

探索在创建后对各种扰动（包括去除尝试）具有鲁棒性的人工智能生成内容水印的使用。为了使去除水印变得更加困难，生成人工智能模型的开发者可能需要考虑如何分发有关水印方法的某些信息或开源其分类器。这可能包括监测对抗用户试图去除水印的方式并修补相应的规避手段。这还包括认识到，鉴于局限性，水印可能并不适合所有情况。

3. 探索人工智能输出数据库的使用

探索模型生成或操作的内容数据库，以识别人工智能生成的内容。这些数据库可以由第三方（包括审核机构和监管机构）查询，以促进对潜在的人工智能生成内容的识别。此类数据库可能仅包含生成内容的子集，该子集被标记为可能重要。为确保用户隐私，可以与此类数据库结合探索隐私保护技术，例如散列技术。这样就可以识别人工智能生成的内容，而无需存储实际内容，从而尊重用户隐私。此外，不同前沿人工智能机构的不同数据库之间的通用标准可以促进统一搜索，从而允许同时识别所有前沿人工智能机构的人工智能生成内容。

重点案例

Meta：致力于提升生成式人工智能的透明度[81]

Meta与人工智能合作伙伴关系(Partnership on AI, PAI)的行业合作伙伴共同制定了合成媒体框架[82]。专注于如何负责任地开发和分享人工智能生成或修改的内容。该框架概述了人工智能价值链不同参与方的实践措施，以帮助他们识别和披露人工智能生成的内容，如添加标签、水印等。

---

[81] Meta，“Overview of Meta AI safety policies prepared for the UK AI Safety Summit”, 2023-10-20, https://transparency.fb.com/en-gb/policies/ai-safety-policies-for-safety-summit#identifiers-of-ai-generated-material.

[82] PAI，“PAI's Responsible Practices for Synthetic Media A Framework for Collective Action”, 2023-02-27, https://syntheticmedia.partnershiponai.org.

Meta也制定了自己的人工智能内容政策。例如，会明确通知用户何时与人工智能互动，让用户可以选择不参与。Meta在其人工智能生成的逼真图像中添加了可见的指示器，帮助用户识别，避免混淆。这些指示器包括Meta AI助手中的图像生成器水印，以及其他生成AI能力中的标识。这些方法会随着时间推移而演变。

Meta正在开发其他技术以包含生成图像的来源信息（即溯源）。部分工作体现在Meta AI助手中，随着技术进步，会扩展到更多体验。Meta和INRIA（法国国家信息与自动化研究所）联合开源数字水印产品Stable Signature[83]，生成的数字水印不受裁剪、压缩、改变颜色等破坏性操作影响，能追溯到图片的初始来源，可应用于扩散、生成对抗网络等模型，例如著名文生图开源项目Stable Diffusion。

在Meta产品的人工智能互动中，用户可访问更多相关信息，包括内容生成方式、局限性以及数据使用等。这些信息在Meta的帮助中心和隐私中心提供。Meta发布了生成式人工智能系统卡[84]，为消费者提供有关Meta生成式人工智能系统的信息，包括系统概述、描述其工作原理的部分、交互式演示、使用提示、数据使用信息以及使用生成式人工智能时应注意的事项。Meta还发布了一份关于负责任地构建生成人工智能的报告，以提高透明度。[85]

延伸阅读

谷歌DeepMind：技术手段结合产品设计和治理政策[86]

- **水印技术：**为图像生成模型Imagen[87]开发了水印识别工具SynthID[88],可以在图像中嵌入隐式水印，并识别出AI生成的内容。
- **元数据：**计划通过“关于此图像”(About this image)[89]工具为谷歌搜索结果中的图像添加上下文元数据，如首次出现时间、是否出现在新闻和事实核查网站上等。还为AI生成的图像添加元数据标签，在搜索结果中标记出来。

---

[83] Meta，“The Stable Signature: Rooting Watermarks in Latent Diffusion Models”，2023-07-26, https://github.com/facebookresearch/stable_signature.
[84] Meta，“AI systems that generate images”，2023-09-27, https://ai.meta.com/tools/system-cards/ai-systems-that-generate-images/.
[85] Meta，“Building Generative AI Responsibly”，2023-09-27, https://ai.meta.com/static-resource/building-generative-ai-responsibly/.
[86] Google，“AI Safety Summit: An update on our approach to safety and responsibility”，2023-10-27, https://deepmind.google/public-policy/ai-summit-policies/#identifiers-of-ai-generated-material.
[87] Imagen，“Imagen: unprecedented photorealism × deep level of language understanding”，2023-05-23, https://imagen.research.google/.
[88] Google，“Identifying AI-generated images with SynthID”，2023-08-29, https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.
[89] Cory Dunton，“Get helpful context with About this image”，2023-05-10, https://blog.google/products/search/about-this-image-google-search/?_ga=2.216315670.1496598606.1683900096-1713801652.1666723966.

- **数字签名：**为语音合成模型AudioLM训练了一个分类器，可以以98.6%的准确率检测该模型生成的合成语音。发布了一个合成语音数据集[90]，支持开发高性能的合成语音检测器。
- **产品设计：**实验性视频翻译工具Universal Translator启用了人工检查机制，确保输出符合预期的翻译使用案例，该工具也只向作出内容检查承诺的合作伙伴开放。
- **治理政策：**谷歌参与人工智能合作伙伴关系，例如合成媒体框架[91]，旨在共享最佳实践，并强调行业在提高大众AI素养方面的重要作用。

阿里巴巴：采取三种方式加强使用者的权益和内容的知识产权保障[92]

- **第一种方式是明水印。**塔玑在每一张图片上都添加了明水印，明确告知使用者此图片为人工智能生成。
- **第二种方式是暗水印。**塔玑在不影响用户使用的前提下，将暗水印嵌入到图片中。暗水印肉眼不可见，只有通过特定的检测模型才能识别，实现了对图片的传播溯源，增强对图片的版权保护。
- **第三种方式是阿里巴巴原创保护平台。**塔玑在平台内采取了一系列措施，例如内容审核、版权监测等算法策略，及时发现盗版图片，并将其召回。这样可以保护使用者的权益，维护整个平台的良性发展环境。

全国信安标委：发布《生成式人工智能服务内容标识方法（征求意见稿）》[93]

- **落实监管规定：**为落实《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》对生成式人工智能内容标识的相关要求，指导有关单位做好相关工作，全国信息安全标准化技术委员会于2023年8月发布《生成式人工智能服务内容标识方法（征求意见稿）》，围绕文本、图片、音频、视频四类生成内容，给出了标识信息的实践指引。
- **显式水印：**在人工智能生成内容的显示区域中，应在显示区域下方或使用者输入信息区域下方持续显示提示文字，或在显示区域的背景均匀添加包含提示文字的显式水印标识。提示文字应至少包含“由人工智能生成”或“由AI生成”等信息。由人工智能

[90] Daisy Stanton, “Advancing research on fake audio detection”, 2019-01-31, https://blog.google/outreach-initiatives/google-news-initiative/advancing-research-fake-audio-detection/.
[91] PAI, “PAI Announces Google to Join Framework for Collective Action on Synthetic Media”, 2023-07-14, https://partnershiponai.org/pai-announces-google-to-join-framework-for-collective-action-on-synthetic-media/
[92] 阿里巴巴人工智能治理研究中心, “生成式人工智能治理与实践白皮书”, 2023-11-22, https://mp.weixin.qq.com/s/Ase8CqpWGqRsyFql5f9ysg.
[93] 全国信安标委, “网络安全标准实践指南—生成式人工智能服务内容标识方法”, 2023-08-07, https://www.tc260.org.cn/upload/2023-08-08/1691454801460099635.pdf.

生成图片、视频时，应采用在画面中添加提示文字的方式进行标识。提示文字宜处于画面的四角，所占面积应不低于画面的0.3%或文字高度不低于20像素。提示文字内容应至少包含“人工智能生成”或“AI生成”等信息。视频中由当前服务生成的画面应添加提示，其他画面可不添加提示。

- **隐式水印：**由人工智能生成图片、音频、视频时，应按在生成内容中添加隐式水印标识。隐式水印标识中至少包含服务提供者名称，也可能包含内容ID等其他内容。
- **元数据：**由人工智能生成的图片、音频、视频以文件形式输出时，应在文件元数据中添加扩展字段进行标识。标识内容应包含服务提供者名称、内容生成时间、内容ID等信息。
- **服务切换：**由自然人提供服务转为由人工智能提供服务，容易引起使用者混淆时，应通过提示文字或提示语音的方式进行标识，提示文字或提示语音应至少包含“人工智能为您提供服务”或“AI为您提供服务”等信息。

六、模型报告和信息共享

摘要

前沿人工智能的透明度(Transparency)可以帮助政府有效实现人工智能的益处并降低人工智能风险。透明度还可以鼓励前沿人工智能机构分享最佳实践，使用户能够就是否以及如何使用人工智能系统做出明智的选择，并增强公众信任，从而有助于推动对人工智能的采用。

在适当的情况下报告和共享信息可以确保各方能够获取所需的信息，以支持有效治理、制定最佳实践、为有关使用人工智能系统的决策提供参考，并建立公众信任。一些报告实践（例如模型卡）已在前沿人工智能机构中使用，而本报告包含的其他实践是未来的选项。

鉴于人工智能最近的快速发展，相关的政府和国际治理机构仍在酝酿之中，这使得前沿人工智能机构与政府共享信息的能力受到限制，即使这种信息共享是可取的。本节使用了“相关政府部门”这一表述，意在表示与政府共享信息是一种好的实践措施，同时也认识到这些相关的政府部门可能仍在发展形成过程中。

我们概述了关于模型报告和信息共享的3类实践措施：

1. 共享与模型无关的有关一般风险评估、缓解和管理流程以及最佳实践的信息
2. 在训练之前、训练期间和部署之前共享有关某些前沿人工智能模型的特定信息
3. 根据适用性，与不同方共享不同信息（需考虑共享该信息的风险），包括政府机构、其他前沿人工智能机构、独立第三方和公众

背景

目前，前沿人工智能系统的开发者与监管者和使用者之间存在很大的信息不对称，而人工智能的进步速度又加剧了这种不对称。改善对人工智能系统如何开发的信息获取可以帮助监管机构使前沿人工智能机构对人工智能的安全和负责任开发负责，并使这些系统的用户能够有效地管理其风险。

针对消费品的标准化文件很常见，例如药物的产品说明书和食品包装上的营养信息。透明度报告（通常以模型卡的形式）已随着最近发布的许多主要模型一同被发布，并且透明度报告的良好实践也正在出现。

公开的透明度报告与其他信息共享渠道（例如前沿人工智能机构和政府之间）可以明显区分。与监管机构共享的某些信息不一定适合公开共享。例如对于前沿人工智能机构来说，报告危险能力模型评测的总体结果可能是有益的，而公开报告详细的引发此类危险能力的技术可能是有害的。为了保护市场敏感的技术能力，商业敏感信息可以与监管机构共享，监管机构可以以汇总或匿名的形式与行业共享这些信息。

实践解读

1. 共享与模型无关的有关一般风险评估、缓解和管理流程以及最佳实践的信息

如负责任扩展策略所述，**与相关政府机构和其他人工智能企业分享风险评估流程和风险缓解措施的详细信息。**

与相关政府机构分享有关如何建立内部治理流程的信息。这将确保风险得到适当识别、沟通和缓解，并使政府和其他外部参与方能够发现可能导致风险被忽视的差距。这些信息可以定期更新（例如每12个月）。如果敏感细节被删除，这些信息也可以公开。

向相关政府机构报告任何安全或安保事件或未遂事件的详细信息。这包括对机构或其系统安全的任何危害，或者任何已部署或未部署的人工智能系统造成或接近造成重大危害的事件。这将使政府部门能够清楚地了解安全事件的发生时间，并更容易预测和降低未来的风险。事件报告可以包括事件的描述、地点、开始和结束日期、受影响的任何当事方和发生伤害的细节、涉及的任何具体模型、负责管理和响应事件的任何相关方，以及本可避免事件发生的方法。重要的是，显示出更严重风险的事件在发生后应尽快报告。剔除敏感信息后，安全和安保事件的概要细节也可以公开，例如人工智能事件数据库(AI Incident Database)[94]中共享的信息。

2. 在训练之前、训练期间和部署之前共享有关某些前沿人工智能模型的特定信息

共享有关特定前沿人工智能模型的信息可以让外部参与方更详细地了解正在进行的人工智能开发和需要应对的潜在风险。

在训练之前，与相关政府机构共享高级模型详细信息，并说明为什么训练不会带来不可接受的风险。这可能包括：

- 模型的概要描述（包括预期用例、预期用户、训练数据和模型架构的概要信息）
- 详细计算信息（包括机构计划使用的最大值，以及有关其位置和提供商的信息）
- 将用于训练该模型的数据描述
- 外推来自小型模型的安全性证据，以表明完整的训练不会带来不可接受的高风险
- 对具体内部和外部风险评估和缓解措施的描述以及总体安全评估，证明训练风险足够低的原因和方式
- 描述哪些领域专家和受影响的利益相关方参与了项目的设计以及风险和影响评估
- 训练期间和训练后的模型评测计划以及预测的危险能力

在训练期间，向相关政府机构更新提供任何模型本身或其风险的重大变化。这可能包括：

- 在每个评估检查点更新模型开发情况以及对开发计划的任何重大更新
- 模型评测的结果，包括涌现的危险能力的详细信息以及这些是否在预期之中

[94] AIID, “AI Incident Database”, 2023-07-08, https://incidentdatabase.ai/.

- 风险背景是否以及如何发生变化（例如是否有其他工具的发布可能与该模型交互）

在部署时，与相关政府机构和公众广泛分享模型的详细信息、模型可能带来的任何风险以及已采取哪些措施来减轻这些风险。

可以向相关政府机构完整提供信息，确保采取强有力的安全措施来保护敏感信息。其中一些信息可以通过发布透明度报告（例如模型卡）并提供模型目的和风险评估结果的总体概述来向公众公开。暴露模型漏洞或促进危险能力传播的信息应从透明度报告中删除，除非公开共享此信息足以有助于缓解模型带来的风险。商业敏感信息（例如训练数据的详细信息）或暴露模型的能力或漏洞的信息应进行编辑。商业敏感信息可以与监管机构共享，监管机构可以以汇总或匿名的形式与行业共享这些信息。例如在部署时共享的信息可能包括：

- 有关模型的详细信息，例如：
  - 模型描述
  - 有关模型使用安全实践的信息（包括不适当和适当使用的领域，或确定使用是否适当的指南）
  - 训练细节，包括训练数据的详细描述以及它们可能编码的任何偏见
  - 模型的偏见、风险和局限性
- 有关风险和影响评估、治理机制和能力评测的信息，例如：
  - 开发前和部署前风险和影响评估程序
  - 前沿人工智能机构进行的评测过程的详细信息，包括花费的时间和资源、有关进行评测人员的专业知识和独立性的信息、授予评测方的访问级别以及所使用的评测的预期限制
  - 有关哪些领域专家和受影响的利益相关方参与项目设计，以及风险和影响评估的详细信息
  - 进行的任何内部或外部评测的结果
  - 根据更具体的评测结果，对模型的鲁棒性、可解释性和可控性进行整体评估
  - 为缓解部署的潜在有害后果而采取的重大措施，包括建立问责和验证机制以及开展内部治理流程
  - 在安全缓解措施实施之前和之后针对最终公开发布模型的能力和风险，包括为防止事故和滥用而采取的缓解措施（例如可用的工具及其预期有效性）
  - 对风险和能力进行持续的部署后监测以及机构如何应对未来事件的计划（除非发布此信息会让恶意行为者规避部署后安全措施）
- 有关模型的其他信息，例如：
  - 部署后访问控制的描述
  - 部署期间运行模型的预期计算需求

3. 根据适用性，与不同方共享不同信息，包括相关政府机构、其他前沿人工智能机构、独立第三方和公众

在共享信息之前，需考虑共享这些信息的风险，并判断共享某些信息是否不合适。特别要考虑公开分享有关危险能力及激发这些能力方法的信息可能造成的潜在危害，因为这些信息可能会激励或帮助其他参与方获得危险能力。同样重要的是要考虑公开共享有关如何生成模型的详细信息的潜在危害，因为这可能会降低生成类似模型的障碍。如果模型具有或可能被修改为具有危险能力（例如生物能力或自主复制能力），那么以这种方式促进模型的广泛分发可能是有害的。

制定关于哪些信息应公开、与政府分享或根本不分享的原则性政策。这些政策可以指定共享某些信息需要进行风险评估的情况，以及进行风险评估和应对风险的指南。然而，重要的是要避免制定过于严格和冗长的风险评估标准和程序，以防止过度不透明。这些政策可以由多学科专家组成的独立审查小组进行指导和监督，以确保有关信息共享的决策是否合理以及是否以最佳透明度为导向。

与专注于人工智能的政府机构和监管机构共享此信息的完整形式，以便政府对人工智能开发的潜在高风险领域以及识别和管理这些风险的流程进行强有力的监管。然后，这些部门可以根据需要选择性地与其他政府机构共享信息。一些与安全特别相关的信息可能需要直接与安全机构共享，例如模型开发中使用的网络安全措施（这将使这些机构更容易识别潜在的安全风险），或特定的物理或网络安全威胁事件。共享更敏感的信息时，须实施稳健的安全措施。

与其他人工智能机构分享更多有限的信息，以促进学习和最佳实践的发展。这可能包括制定风险评估和缓解措施，以及风险治理流程中的最佳实践和经验教训，有关安全和安保事件的细节（以提高意识同时保护知识产权），以及从风险评估和能力评测的亮点和经验教训。一般来说，机构之间共享与模型无关的信息比共享与特定模型相关的信息更容易，因为后者可能更具商业敏感性。还必须考虑不向某些公司提供可能带来竞争优势的特权访问信息这一点。

与独立第三方共享特定信息，当其有助于评测和技术审核时（请参阅模型评测和红队测试）。这可能需要共享与政府完全共享的许多相同信息，但可能会根据具体情况进行共享。

与模型的下游用户共享特定信息，以便更有效地缓解整个人工智能供应链的风险并建立消费者信心。这可能包括违反其服务条款的模型使用，并且可以通过用户条款或有关模型的已发布信息提供给用户。

与公众分享更通用的信息，以便公众监督并建立公众对人工智能系统安全性和可靠性的信心。这可能包括风险评估和缓解流程的概要总结、安全和安保事件的概要细节、风险治理流程的概要以及对模型目的、用例以及风险评估和能力评测结果的总体概述。

重点案例

暂时空缺：根据我们目前的理解，尚没有好的最佳实践

延伸阅读

国际：已有信息共享或报告的政府要求和自愿承诺，待进一步观察企业执行情况

- **英国政府：**获得谷歌DeepMind、OpenAI和Anthropic前沿人工智能模型的优先访问权，用于安全研究和评测[95]。
- **美国政府：**《关于安全、可靠和可信地开发和使用人工智能的行政命令》[96]对部分模型提出了报告其安全测试结果和其他关键信息的要求，例如对使用大于10^{26}整数或浮点运算的计算量训练的任何模型，或主要使用生物序列数据并使用大于10^{23}整数或浮点运算的计算量训练的模型。
- **企业自愿承诺**[97]**：**美国政府获（第一批7家和第二批8家）企业承诺在向公众推出产品之前确保其安全（包括内外部测试、**风险信息共享**），构建将安保放在首位的系统（包括保护模型权重、促进报告漏洞），赢得公众的信任（包括标识生成信息、**报告模型局限**、缓解社会风险、解决社会挑战）。
- **前沿模型论坛**[98]**：**由谷歌、微软、OpenAI和Anthropic联合发起成立，重点工作之一是“促进企业和政府之间的信息共享”，包括建立可信、安全的机制，在企业、政府和相关利益相关方之间共享有关人工智能安全和风险的信息。论坛将遵循网络安全等领域负责任披露的最佳实践。

[95] UK Government, “PM London Tech Week speech”, 2023-06-12, https://www.gov.uk/government/speeches/pm-london-tech-week-speech-12-june-2023.
[96] The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence”, 2023-10-30, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[97] The White House, “FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI”, 2023-07-21, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.
[98] OpenAI, “Frontier Model Forum”, 2023-07-26, https://openai.com/blog/frontier-model-forum.

中国：《人工智能示范法（专家建议稿）》提出负面清单制度

- **2023年人工智能法草案被列入国务院立法工作计划。**中国社会科学院国情调研重大项目《我国人工智能伦理审查和监管制度建设状况调研》课题组2023年8月15日**发布《人工智能法示范法1.0》（专家建议稿）**[99]**，9月7日又修订发布1.1版**[100]。
- **总体来看：**《示范法》坚持发展与安全并行的中国式治理思路，提出了**负面清单管理**等治理制度与模型报告要求直接相关：第二十三条（分类管理制度）国家建立人工智能负面清单制度，对负面清单内的产品、服务实施许可管理，对负面清单外的产品、服务实施备案管理。人工智能研发者、提供者申请负面清单内人工智能产品研发、提供许可，**应提交人工智能安全评估报告等材料**。

[99] 中国社会科学院，“人工智能法示范法1.0（专家建议稿）”, 2023-08-15, https://www.21jingji.com/article/20230815/herald/232ec88f32dcdf3db371a29fe427f81b.html.
[100] 中国社会科学院，“人工智能法示范法1.1（专家建议稿）”, 2023-09-07, https://www.21jingji.com/article/20230907/herald/982ae3bb7b82597b4dc1f990ded64ad2.html.

七、防止和监测模型滥用

摘要

一旦部署，人工智能系统就可以被故意用来实现有害和/或非法的结果。这其中包括网络攻击、传播虚假信息和犯罪活动。

为了防止模型滥用，监测系统的滥用并对其做出响应非常重要。这有助于确保及时识别和处理滥用的情况，从而缓解更广泛危害的风险。

我们概述了关于防止和监测模型滥用的6类实践措施：

1. 建立流程来识别和监测模型的滥用，例如监测模型被滥用和规避保障措施的常见方式
2. 实现模型输入和输出过滤器
3. 实施额外措施来防止有害输出，包括微调、提示和拒绝采样
4. 实施基于用户的API访问限制和监测，例如减少对无合理理由反复触发内容过滤器的个人的访问权限
5. 为潜在的最坏情况或持续滥用场景制定响应计划，手段包括快速回滚和撤回模型
6. 持续评估现有和额外保护措施的有效性和可取性，因其也可能阻碍正面用途并减少隐私

背景

人工智能系统可能以多种方式被滥用，包括网络攻击（例如钓鱼邮件）和传播虚假信息。人工智能系统还可以增加犯罪的规模和速度，并越来越多地用于欺诈、网络儿童受侵害和亲密图像滥用等犯罪。虽然在模型训练期间可以减少未来滥用的可能性，但在模型部署期间也可以采取措施来识别和响应此类滥用。

前沿系统通常通过API服务进行部署，对模型的访问由人工智能机构控制，用户无需在自己的硬件上运行即可查询模型，这与公开发布模型权重（通常称为“开源”或“开放获取”模型）不同。本节重点介绍通过API服务部署的模型。

对低风险模型的开源访问可以通过对人工智能系统进行更广泛的公众审查和测试，实现对人工智能系统安全性的更深入理解。与此同时，开源模型的开发者通常对下游使用有更少的监督，这意味着本节讨论的许多防止高风险模型滥用的方法对发布开源模型的前沿人工智能机构来说是不可用的。此类机构仍然可以在一定程度上识别和缓解模型滥用，例如通过努力执行开源模型许可证机制。然而，通过API发布模型为前沿人工智能机构提供了更多应对滥用问题的能力，因其可以保持对模型使用方式的可见性，增强对使用和防护的控制，并能够在部署后更新或回滚模型。

随着使用情况的变化，持续评估和调整防止和缓解模型滥用的实践非常重要。这包括考虑哪些使用形式需要预防，因为滥用和合法使用之间的区别可能是模糊的。

实践解读

1. 建立流程来识别和监测模型的滥用

了解自己的模型以及其他团体发布的模型如何被滥用。例如知道有些人已经开始使用竞争对手的模型进行网络钓鱼攻击，可能会导致前沿人工智能机构更加担心自己的模型也会以这种方式被滥用。

报告有关广泛滥用模式的信息。这些信息可以帮助政府、公众和其他前沿人工智能机构更好地了解风险。报告可能侧重于与API使用相关的群体层面指标，例如有害输入的过滤率或因滥用而被禁止的用户数量。以清晰易懂的方式呈现这些信息并优化其可访问性可能会受益。

为使用日志确定适当的保留时间表，平衡安全和隐私考量。在严重的滥用情况下，可能需要访问几个月前的日志才能最详细地了解滥用的原因。然而在某些情况下，延长保留期限可能会对隐私产生不成比例的影响。

2. 实现模型输入和输出过滤器

将内容过滤器应用于模型输入和模型输出。输入过滤器可以阻止模型处理有害请求（例如有关制造武器的建议的请求）。内容修改器可以调整有害提示从而引发无害响应。输出过滤器可以阻止有害的模型输出（例如制造武器的指令）被发送回用户。

探索和比较开发内容过滤器的多种方法的有效性。这可能涉及比较可用的方法；选择最有效的方法；如果可以显著提高效力则将它们组合使用。最佳实践可以与其他前沿人工智能机构共享或广泛发布。

使内容过滤器能够抵御“越狱”尝试。努力确保过滤器对越狱尝试具有鲁棒性，例如在用于开发过滤器的数据集中包含越狱尝试的示例。

在开发、存储或共享内容过滤器以及用于生成它们的任何专用数据集时，务必小心谨慎。在某些情况下，用于创建过滤器的组件或数据集可用于训练高风险模型。例如评测模型输出的攻击性的分类器可用于将模型训练得更具攻击性。

3. 实施额外措施防止有害输出

微调模型以减少产生有害输出的倾向。这可能涉及使用来自人类或人工智能生成的关于不同模型输出适当性的反馈进行强化学习，例如一种基于宪法的方法，其中人类对微调过程的输入由一系列原则提供。它还可能涉及在精选数据集上微调模型以对提示进行适当响应。在使用人类反馈的情况下，可能需要考虑内容审核员的心理健康。

对模型进行提示以避免有害行为。这可能涉及使用“元提示”来指示模型应忽略造成危害的请求。或者，可以使用提示精炼或其他方法来微调他们的模型，使其表现得就像收到了特定的提示一样。

还要考虑对模型输出应用“拒绝采样”方法。这些方法涉及生成多个输出，对其危害性进行评分，然后仅向用户呈现危害最小的输出。

4. 实施基于用户的API访问限制和监测

通过API部署模型时，考虑减少表现出可疑使用模式的用户访问机会。例如如果内容过滤器在短时间内连续多次阻止用户的请求，则表明他们试图滥用该模型或寻找规避其过滤器的方法。适当的措施可能包括警告、进一步调查、速率限制、更严格的过滤器和禁令。应注意避免限制真正的使用，例如合法的人工智能安全研究人员试图检查模型行为，或者用户在获取自残等艰难话题的合法帮助来源时遇到困难。

传达API访问减少政策，并允许用户申诉访问减少。用户应该能够了解其访问权限可能被减少的原因，并且如果他们认为策略没有正确执行，应能够对访问权限减少提出申诉。另外，用户可以采取措施来缓解滥用风险，例如提供证据来证明行为的合理性。

对API用户实施分层级的KYC检查。KYC检查可以帮助防止用户在访问权限减少时简单地创建新帐户。在风险较高的情况下，例如具有更危险能力的模型、访问保护措施较少的模型或模型的大量使用，可以实施更严格的检查，例如身份验证。在要求注册时衡量KYC检查与潜在的隐私和访问权衡非常重要。

考虑仅将某些API访问层限制为“受信任”类别中的用户。例如可能仅向经验证的企业、非营利机构和大学的用户提供高使用率、微调访问权限、许可内容过滤器以及访问具有高滥用可能性的模型。

建立协议来决定何时以及如何与相关政府机构共享API用户的信息。协议可能包括涵盖主动与政府机构共享用户信息的情况（例如有理由认为用户可能试图造成严重危害时），以及如何响应政府的数据请求。

5. 为潜在的最坏情况或持续滥用场景制定响应计划，手段包括快速回滚和撤回模型

实施流程和技术要求，以便在发生严重、广泛或持续的危害时能够快速回滚或撤回模型。例如首届“人工智能安全国际对话”的联合声明[101]中，建议“规定一些明确的红线，并建立快速且安全的终止程序。”在这种极端威胁或持续滥用的情况下，回滚到未遭受相同滥用威胁的模型的先前版本，或完全撤回对模型的访问可能是适当的措施。定期进行演练可以增强应对对此类情况的能力。

[101] 安远AI. “AI的帕格沃什会议！中美英加欧20多位顶尖AI专家线下聚首，呼吁AI安全与治理的全球协同行动”，2023-11-01, https://mp.weixin.qq.com/s/1WbrS-L8Qsww10nosADwJQ.

告知最终用户可能需要快速回滚或撤回模型，并且将尽可能安全地最大程度地减少对最终用户的干扰。当模型部署在安全攸关领域时，这一点尤其重要。

建立治理流程，以确保在应对最坏情况或持续滥用情景时内部有明确的反应。此类流程可以借鉴负责任扩展策略中使用的类似问责和治理机制。澄清将模型回滚到先前版本或撤回模型所需的批准、需要采取此类措施的滥用类型，以及采取此类措施的时间范围可能很有价值。

6. 持续评估现有和额外保护措施的有效性和可取性

定期评估监测和保障效果，并持续投入改进。定期评估监测效果可以建立对问题检测的成功率和速度的理解，并以此为作为部署决策的基础。这些评估可能会借鉴内部和外部红队测试的结果、使用日志的随机审核、最佳实践以及有关现实世界滥用的可用信息。随着人工智能能力的增强，风险预计也会增加。因此，积极增加安全、安保和监测方面的投资将很有价值。

根据滥用模式探索新的防护和对策，并认识到适当的措施可能会根据模型类型、使用模式以及对模型能力的技术理解而有所不同。

采用多种监测技术来平衡全面性、可扩展性和隐私。一个强大的监测程序通常会将自动审查和人工审查相结合，因为自动审查可能会遗漏复杂的问题，而人工审查并不总是可扩展的，并且可能会带来隐私问题。

定期评估对用户采取安全保障的成本，包括偏见和隐私损失，以及对用户在适当环境下与人工智能自主体讨论敏感话题的自由的限制。例如这可能涉及通过对阻塞的输入和输出进行采样来估计过滤器的误报率；将缺乏某些保障措施的模型与具有适当保障措施的模型的能力、效率和用户评级进行比较；进行用户访谈，以了解知情用户对隐私损失和算法过滤内容中出现偏见的担忧程度。

探索降低用户保护成本的策略，例如可降低用户隐私情况监测成本的“结构化透明度”方法，或者为不同资质的用户提供对不同保护措施的模型的分层级访问。

重点案例

微软：加强AI红队建设，对接标准和流程，对齐并扩展了自愿承诺[102]

加强内部和外部的AI红队建设。微软通过增加新的团队成员和制定内部实践指南来加强其AI红队。微软的AI红队是一个独立于其产品构建团队的专家团队；助力红队高风险的AI系统。在发布新的高性能基础模型前，微软建立外部红队，开展独立的专家评审。微软更新了SDL威胁建模要求的内部实践指南，考虑到人工智能和机器学习的特有威胁，可能需要红队测试。

[102] Microsoft, “Microsoft's AI Safety Policies”, 2023-10-26, https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/#Preventing_and_Monitoring_Model_Misuse.

将产品安全开发与负责任人工智能标准对接，进一步完善相关流程。微软更新了安全开发生命周期(SDL)，以链接其负责任人工智能标准并整合其中的内容，加强与他们的负责任人工智能标准要求的治理步骤保持一致的流程并加强检查。微软宣布了新的人工智能系统漏洞严重性分类，涵盖了因在微软的产品和服务中使用人工智能而特别产生的新型漏洞。此外，微软启动了人工智能漏洞赏金计划，鼓励外部研究人员参与，将AI驱动的必应体验作为第一个范围内的产品，奖金高达1.5万美元。微软正在通过前沿模型论坛合作，制定与发现前沿模型中的漏洞或危险能力相关的“负责任的披露”流程指南。

这些举措与微软向白宫所做的自愿承诺对齐，包括红队测试和系统测试、识别缓解网络安全风险。在此基础上，微软还扩展实施了NIST人工智能风险管理框架、针对高风险模型和应用程序实施稳健的可靠性和安全实践等更多承诺。微软将这些预防和监测模型滥用的举措贯穿产品生命周期，进行持续评估与迭代改进，加强防范能力。一旦发现新的滥用模式，微软将采取行动遏制风险，保护客户。人工智能产品还具备监测部署后滥用的能力，将反馈应用于产品开发。

延伸阅读

Inflection：强调实时监测、快速响应以及使用先进系统来检测和应对模型滥用[103]

- Inflection认为安全不仅仅取决于模型本身，也取决于用户的行为。有恶意意图的复杂用户可能会规避安全协议，并将模型应用于不当或有害的目的。因此，监测和快速响应是确保前沿AI系统安全的任何框架中必不可少的部分。
- Inflection的安全团队会进行24小时不间断的监测和调查，实时识别和缓解生产系统中出现的“真实环境中的”威胁。这包括定期审查生产系统在关键领域的安全表现，以及设立“警戒线”系统，它们会立即将与试图破坏模型安全协议的相关可疑行为模式提报给值班的安全专家。一旦收到警报，安全团队有权采取各种手段，阻止恶意行为者访问系统，并实施缓解措施以限制未预见漏洞的危害。
- 展望未来，Inflection计划运用前沿AI技术本身，提高安全团队的情况感知和工作效率。他们正在试验使用语言模型检测平台滥用情况，以助力安全工作的规模化。

人工智能合作伙伴关系(PAI)：提供了可操作性的《安全基础模型部署指南》

- PAI提供的《安全基础模型部署指南》，为AI模型提供商提供了一个框架，以负责任地开发和部署一系列AI模型，旨在确保社会安全并适应不断发展的AI能力和用途。

[103] Inflection，“Our policy on frontier safety”, 2023-10-30, https://inflection.ai/frontier-safety#area5.

- 为了解决这些深远的影响，PAI强调需要采取集体行动和共享安全原则。这种协作方法涉及各个不同利益相关方，包括工业界、民间机构、学术界和政府。目标是为负责任的模型开发和部署建立集体共识的最佳实践，从而落实人工智能安全原则。
- 指南被设计为一份动态文档，可以随着新的人工智能能力和风险而不断发展。它提供了一组针对模型的特定能力和发布方式的定制化推荐实践[104]。这种方法在整个部署过程中指导模型提供商的同时，也补充了更广泛的监管框架，并持续迭代。

关于前沿模型开源的争论：审慎开源 vs 鼓励开放

GovAI发布的《开源高性能基础模型》[105]和Mozilla发起的《AI安全与开放的联合声明》[106]代表了两种相对的主要立场：

- **风险意识：**GovAI论文提议，最初开源强大的模型可能风险太大，主张在社会适应和安全机制改善后逐步开源。与此相反，Mozilla声明促进立即开放和透明，强调公众审查和访问的安全利益。
- **风险评估：**GovAI论文强调在开源高性能模型之前进行严格的风险评估，考虑潜在的滥用。Mozilla声明虽然承认风险，但更倾向于开放所带来的减少危害的好处。
- **替代方案：**GovAI论文提出了完全开源发布的替代方案，如分阶段发布或为研究人员提供结构化访问，平衡技术利益与风险管理。Mozilla声明主要支持完全开放，以便合作效益和公众问责。
- **标准制定：**GovAI论文鼓励开发者、标准制定机构和开源社区在模型组件发布方面定义细粒度标准。Mozilla声明没有具体涉及标准制定的细节。
- **政府监管：**GovAI论文建议政府监管开源AI模型，并在风险足够高时实施安全措施。这包括强制执行风险评估和共享标准。Mozilla声明没有这样详细地涉及政府监管的角色，更多地聚焦于开放的一般原则。
- **总体而言，两种立场都重视AI开发的安全性和责任感、支持透明性原则和公众问责、支持模型开放前进行风险评估等，**但GovAI论文采取了更谨慎的、分阶段的方法，强调风险管理和政府监督，而Mozilla声明则倡导更广泛和即时的开放作为安全和创新的关键。

---

[104] PAI, “Generate Custom Guidance”, 2023-10-24, https://partnershiponai.org/modeldeployment/#generate_custom_guidance.
[105] Elizabeth Seger et al., “Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives”, 2023-09-29, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.
[106] Mozilla, “Joint Statement on AI Safety and Openness”, 2023-10-31, https://open.mozilla.org/letter/.

八、数据输入控制和审核

摘要

用于训练人工智能系统的数据会影响它们的行为方式。如果前沿人工智能系统在低质量或不良的数据上进行训练，就会增加系统带来的风险。

通过控制和审核人工智能系统训练或微调的数据，可以更准确地预测其能力并降低风险，例如删除可能产生危险能力的输入数据。数据输入控制和审核还可以为下游用户和监管机构提供重要信息。

我们概述了关于数据输入控制和审核的4类实践措施：

1. 在收集训练数据之前，实施负责任的数据收集实践
2. 在使用输入数据训练人工智能系统之前对其进行审核，例如尝试识别可能产生危险能力的数据
3. 根据数据审核结果采取适当的风险缓解措施，例如通过整理数据集以确保不会在某些数据上进行训练
4. 通过邀请外部参与方评估其输入数据并共享输入数据审核信息，促进对输入数据的外部审查

背景

人工智能系统接受训练的数据会影响它们的行为方式及其带来的风险。尽管训练数据和系统行为之间的关系很复杂，但一些预测是可能的。

存在偏见、不准确或代表性不足的数据可能会导致人工智能系统表现较差，并产生更有害的社会影响。经过敏感或个人数据训练的人工智能系统可能容易受到提取攻击。一些训练数据可能会增强人工智能系统的潜在危险能力，使其被滥用时危害更大。

这些关系对前沿人工智能系统来说问题尤为严重，因为这些系统通常是在可能包含此类数据的大型数据集上进行训练的。此外，通过微调或人类反馈强化学习(RLHF)，前沿人工智能系统被的附加数据进一步修改。这些较小的补充数据集对人工智能系统的性能和行为产生了不成比例的巨大影响。

负责任的数据收集实践可以通过提高训练数据的质量来帮助人工智能系统变得更安全，从而降低系统在危险、敏感或不平衡数据上训练的可能性。对输入数据的仔细审核还可以更好地预测系统能力和局限性，同时揭示降低风险的途径，例如不在某些不可取的数据上进行训练。

实践解读

1. 在收集输入数据之前，实施负责任的数据收集实践

在收集训练数据之前，需要考虑适用的监管框架。这可能涉及建立处理训练数据的法律基础并了解可能适用的任何版权注意事项。这有助于进一步降低风险，例如系统泄露个人信息。

尽可能考虑数据最小化原则。数据最小化可以降低有害内容进入训练数据的风险。前沿人工智能机构可以探索“数据修剪”的实践措施，事实证明这种做法可以在最大限度地减少所需的预训练数据量的同时提高数据质量和系统性能。

2. 在使用输入数据训练人工智能系统之前对其进行审核

审核用于预训练的数据集，以及用于微调、分类器和其他工具的数据集。不适当的数据集可能会导致系统无法拒绝有害指令。

使用分类器和过滤器等技术工具来审核大型数据集，以支持可扩展性和隐私。这些可以与人类监督结合使用，从而可以验证和增强这些评估。

评估训练数据的整体构成。这可能包括数据来源、数据溯源、数据质量和完整性指标以及偏见和代表性的度量。数据的数量和种类是简单、可靠的风险预测指标，并在更有针对性的评估受到限制的情况下提供了额外的防线。

审核数据集以获取：

- **可能增强危险系统能力的信息**，例如有关武器制造或恐怖主义的信息。
- **私人或敏感信息**。人工智能系统可能会受到数据提取攻击，有目的的用户可以提示系统泄露训练数据片段，甚至可能意外泄露这些信息。这使得了解数据集是否包含私人信息或敏感信息（例如姓名、地址或安全漏洞）变得非常重要。
- **数据存在偏见**。不平衡或不准确的训练数据可能会导致人工智能系统对于具有某些个人特征的人准确性较低，或者对特定群体的描述不准确。确保训练数据的更好平衡有助于解决这个问题。
- **有害内容**，例如儿童受侵害材料、仇恨言论或网络侵害。更好地了解数据集中的有害内容可以为应用内容过滤器等安全措施提供信息。
- **错误信息**。用不准确的信息训练人工智能系统会增加系统输出不准确的可能性，并可能导致危害。

利用外部专业知识进行输入数据审核。例如可以咨询生物安全专家来确定与生物武器制造相关的信息，这些信息对于非专家来说可能不太明显。

使用数据审核来加深对训练数据如何影响人工智能系统行为的理解。例如如果模型评测揭示了潜在的危险能力，数据审核可以帮助确定训练数据对其产生的影响程度。

对客户用于微调人工智能系统的数据集进行审核。客户通常允许在自己的数据集上微调系统。通过进行审核以确保客户不会鼓励恶意行为，前沿人工智能机构可以利用其专业知识和对原始训练数据的洞察力来识别潜在的上游危害。重要的是，前沿人工智能机构要谨慎处理隐私问题，并在适当的情况下使用隐私保护技术。

记录输入数据审核的结果，包括元数据。前沿人工智能机构在记录输入数据审核结果时可以参考新兴标准，例如数据集的数据表(datasheets)。

3. 根据数据审核结果采取适当的风险缓解措施

使用输入数据审核为开发前和部署前风险评估提供信息。数据审核在开发前风险评估中可能特别有价值，因为此时还无法获得系统行为的直接证据。输入数据审核还可用于提高对数据如何影响系统能力的基于上下文的理解，这将加强未来的风险评估和缓解技术。

在适当的情况下删除可能有害或不可取的输入数据。鉴于人工智能系统的泛化能力越来越强，数据管理可能不足以防止危险的系统行为，但可以与微调和内容过滤等其他措施一起提供额外的防御层。

对于某些类型的风险，探索使用有害数据的选项，来减少人工智能系统的危险能力或帮助开发缓解工具。例如通过使用标记的有害内容对系统进行微调以拒绝与有害信息相关的请求。

当确定数据丢失或不足时，请考虑获取或生成额外数据并将其添加到训练数据集中。提高训练数据的代表性可以提高性能并减少潜在的负面社会影响和歧视效应。然而，寻求额外的数据只能通过适当的方式，过程中应尊重并赋能那些在数据中缺失的个人。

认知输入数据审核的局限性。理解数据影响和过滤特定数据的技术是有限的（例如从数据集中排除信息并不总能防止人工智能系统推理或发现该信息）。还需采取其他风险缓解措施，并且改进数据输入审核技术的进一步研究也很重要。

4. 促进对训练数据集的外部审查

促进外部各方对数据输入的独立审核。由于训练数据构成敏感的知识产权，因此在共享训练数据时实施适当的技术和组织保障措施以确保隐私和安全非常重要。为了保护敏感信息，前沿人工智能机构可以探索向审核机构提供删除了敏感信息的合成数据集的可能性，或者为审核机构提供对训练数据的隐私保护访问。

与用户和外部利益相关方共享有关输入数据审核的信息。这可以包含在透明度报告中，例如有关训练数据（包括数据源）、数据审核程序以及为降低风险而采取的措施的概要信息（请参阅模型报告和信息共享）。更敏感的信息可能会直接与监管机构和外部审核机构共享。

请注意，识别危险信息的技术可能容易被滥用。例如识别大数据集中私人信息的技术可能被网络犯罪分子滥用，因此只能被负责任地共享。审核期间发现的危险或其他敏感信息应被安全存储以避免泄漏。

重点案例

OpenAI：实施多重控制，允许内容拥有者表达训练偏好，过滤潜在问题数据[107]

主要信息来源是公开的互联网信息。OpenAI的大语言模型（包括为ChatGPT提供支持的模型）是使用三个主要信息源开发的：1)互联网上公开提供的信息，2)OpenAI获得第三方许可的信息，以及3)OpenAI的用户或标注员提供的信息。OpenAI的绝大多数训练数据都来自互联网上免费公开的公开信息，例如OpenAI表示不会在付费专区或“深层网络”中寻找信息。

过滤和移除某些不适宜的信息。OpenAI会使用过滤器并删除某些他们不希望模型学习或输出的数据，例如仇恨言论、成人内容、主要聚合个人信息的网站和垃圾邮件等。

实现了让内容创建者和网站运营者表达他们对AI训练内容偏好的机制。OpenAI采取了一些措施，使创作者、权利持有者和网站运营商能够表达他们对其拥有或控制的内容的模型训练偏好。例如OpenAI为网站运营商提供了一种简单的方法，依靠robots.txt网络标准，阻止其内容被OpenAI的GPTBot网络爬虫访问。同样，OpenAI记录了ChatGPT和ChatGPT插件用于访问网站的用户代理字符串“ChatGPT-user”，以便网站运营商也可以阻止这些的访问。OpenAI在线提供有关如何禁止任一机器人访问网站的说明。OpenAI还为图像创建者提供自助服务表格[108]，以选择将他们的内容从OpenAI未来的DALL-E图像生成模型的训练中剔除。

延伸阅读

谷歌DeepMind：研究数据的摄取请求是一项值得注意的新政策[109]

- 将数据治理流程纳入其人工智能生命周期的每个阶段。其基础模型是使用公开数据源和开源精选数据集以及专有数据和第三方许可的数据进行训练的。为更具体的能力微调通用基础模型需要更专业的数据来代表目标主题或用例，RLHF是实现更高质量的微调数据的常用策略之一。
- **一项特别政策：对希望使用数据进行研究（包括前沿模型的预训练和微调）的团队，可向其专门的数据团队提交数据摄取请求。**数据团队将根据研究、道德、安全、商业和法律团队的意见，启动对数据来源、内容和许可的审查。
- 审查后，数据可能会在用于研究之前进行修改，例如过滤以确保遵守适用法律。如果使用获得批准，谷歌DeepMind的数据团队将管理数据的摄取、编目和存储，包括是

[107] OpenAI，“OpenAI's Approach to Frontier Risk”，2023-10-26, https://openai.com/global-affairs/our-approach-to-frontier-risk#data-input-controls-and-audit.
[108] OpenAI，“Artist and Creative Content Owner Opt Out”，2023-09-29, https://share.hsforms.com/1_OuT5tfFSpic89PqN6r1CQ4sk30.
[109] Google，“AI Safety Summit: An update on our approach to safety and responsibility”，2023-10-27, https://deepmind.google/public-policy/ai-summit-policies/#data-input-controls-and-audit.

否以及如何在内部进一步共享数据以及管理谁有权访问数据。审查、评估和任何所需的缓解措施均由相关内部团队记录。

全国信安标委：发布《生成式人工智能服务 安全基本要求》（征求意见稿）[110]

- 2023年10月11日，全国信息安全标准化技术委员会官网发布《生成式人工智能服务安全基本要求》（征求意见稿）。这是**国内首个专门面向生成式AI安全领域的规范意见稿**，也是对今年7月推出的《生成式人工智能服务管理暂行办法》的支撑。
- **语料安全要求**方面，征求意见稿提出：
  - 语料来源安全要求：(一)建立语料黑名单制度，排除不安全来源；(二)对不同来源语料实施安全评估和风险控制；(三)确保语料可追溯，商业语料需提供合法证明；(四)用户生成语料需取得授权；(五)禁止使用违法害禁信息作为语料。
  - 语料内容安全要求：(一)全面过滤语料中的非法和有害信息； (二)尊重知识产权，防范语料侵权风险；(三)合法、正当处理个人信息语料。
  - 语料标注安全要求：(一)实施标注人员资质管理；(二)制定完善的标注规范和规则；(三)加强标注内容准确性检查。
- **语料安全评估**方面，征求意见稿提出：采用多种抽检手段，利用人工抽检和关键词抽检、分类模型技术抽检，对语料安全情况、生成内容安全进行评估检测。

上海人工智能实验室联合人民网：成立中国大模型语料数据联盟安全治理专委会[111]

- **中国大模型语料数据联盟：**由上海人工智能实验室联合中央广播电视总台、人民网等10家单位联合发起，于2023年7月世界人工智能大会开幕式上宣布成立，旨在通过链接模型训练、数据供给、学术研究、第三方服务等多方面机构，联合打造多知识、多模态、标准化的高质量语料数据，探索形成基于贡献、可持续运行的激励机制，打造国际化、开放型的大模型语料数据生态圈。
- **安全治理专委会：**2023年11月，中国大模型语料数据联盟在全球数商大会上举办“数据要素市场与大模型语料库论坛暨中国大模型语料数据联盟开放日活动”，上海人工智能实验室还联合人民网，共同发起成立中国大模型语料数据联盟安全治理专委会，旨在推动大模型数据安全治理与隐私保护，为大模型技术快速发展提供数据安全保障。

[110] 全国信安标委，“生成式人工智能服务安全基本要求”，2023-10-11, https://www.tc260.org.cn/upload/2023-10-11/1697008495851003865.pdf.
[111] 上海人工智能实验室，“上海AI实验室联合人民网发起成立中国大模型语料数据联盟安全治理专委会”，2023-10-27, https://mp.weixin.qq.com/s/hW_GuSq43q0udqs2-ZuFCw.

北京智源人工智能研究院联合共建单位：开源可信中文互联网语料库CCI[112]

- **行业协会与地方政府支持：**2023年11月，在中国网络空间安全协会人工智能安全治理专业委员会数据集工作组、北京市委网信办、北京市科委、中关村管委会、海淀区政府的支持下，智源研究院联合拓尔思、中科闻歌共建了“中文互联网语料库”(Chinese Corpora Internet，简称CCI)，旨在为国内大数据及人工智能行业提供一个安全、可靠的语料资源，并以此为契机促进不同机构合作，推动大数据和人工智能领域的健康发展。
- **数据规模和时间跨度：**CCI语料库首期开放的数据规模为104GB，总体时间跨度为2001年1月至2023年11月。
- **数据来源与数据处理：**数据均来自高质量可信、中国境内的互联网站，经过严格的数据清洗和去重，并且在内容质量、价值观等方面进行了针对性的检测与过滤，进一步提升数据质量和安全可信程度。数据处理规则包括基于规则和模型的过滤，以及严格的数据去重。为避免评测数据泄露影响模型性能，语料库还过滤了当前多个主流的中文评测数据集。

[112] 智源研究院，“打造生成式人工智能压舱石，智源联合共建单位开源可信中文互联网语料库CCI”，2023-11-29, https://mp.weixin.qq.com/s/1kTGFPqkdTwy41hPRJP0QA.

九、负责任扩展策略

摘要

关于是否以及如何开发和部署新的人工智能系统，我们的决策后果重大。部署一个风险过高的系统可能会因误用、滥用或安全故障而造成重大危害。即使只是开发一个高风险的系统，如果最终泄漏、被盗或在内部部署期间造成危害，也同样是一个难以接受的结果。

负责任扩展策略是一个新的框架，用于管理与前沿人工智能相关的风险并指导人工智能开发和部署的决策制定。它涉及实施流程用于识别、监测和缓解前沿人工智能风险，包括本报告中列出的其他流程和实践，并以健全内部问责和外部验证流程为基础。

我们概述了关于负责任扩展策略的7类实践措施：

1. 在开发或部署新模型之前进行彻底的风险评估，并辅之以持续的监测，例如结合评测结果、早期模型的证据和专家预测，
2. 预先确定“风险阈值”，限制可接受的风险水平，例如当模型发布可造成的网络攻击或欺诈的风险达到何种程度时，是我们不可接受的
3. 根据每个风险阈值预先承诺采取特定的额外缓解措施，然后进行剩余风险评估
4. 根据部署的阶段调整缓解措施，认识到模型的使用方式和环境可能与预期不同
5. 若在未预先约定缓解措施的情况下达到风险阈值，则做好暂停开发和/或部署的准备
6. 与相关政府部门和其他人工智能企业分享风险评估流程和风险缓解措施的细节
7. 承诺建立健全的内部问责和治理机制，并接受外部验证，例如内部风险治理、记录保存、独立审核

背景

随着能力的扩展，围绕模型开发和部署的许多问题将需要格外小心。包括：

- 开发什么模型以及如何开发
- 这些模型在开发期间需要什么级别的安全保障
- 是否以及如何部署模型，例如是否应通过API部署或开源
- 在训练中使用什么数据集
- 向用户提供什么指导（如果需要的话）
- 采取哪些保障措施

实践解读

1. 在开发或部署新模型之前进行彻底的风险评估，并辅之以持续的监测

风险评估非常有价值，因为它们可以引导人们做出负责任的决策——包括事前有意的模型能力提升决策——并最终缓解风险。

为模型制定严格的风险评估流程：

- 应当尝试涵盖人工智能系统的所有可能的和间接的风险，包括造成严重危害的低概率风险
- 考虑的因素包括但不限于：
  - 模型评测和红队测试（请参阅模型评测和红队测试）
  - 先前模型的影响和能力的证据
  - 该领域的最新研究和发展（请参阅优先研究人工智能带来的风险）
  - 内部和外部领域的专业知识
  - 数据输入审核的结果（请参阅数据输入控制和审核）
- 考虑到进行可靠的风险评估的难度，应当创建一种认真对待与前沿模型相关的风险预测的重大不确定性的文化
- 应当考虑模型的潜在益处
- 应当纳入从与工业界、学术界和政府的交流中所获得的类似模型能力的经验教训

在模型开发前和模型部署前均进行风险评估。开发前的风险评估很重要，因为如果模型泄露、被盗或以其他方式传播，训练高风险系统仍然可能导致危害。

在开发和部署之后对系统进行监测。如果做出微调或其他可能增加模型危险的重大变化（例如模型获得工具或插件的访问权限），需要进行新的风险评估。这可能与检测意外事态发展的尝试和改变现有风险评估结果的新信息同时发生，这也可能触发新的风险评估。

2. 预先确定“风险阈值”，限制可接受的风险水平

描述和持续完善每个模型的风险评估结果（“风险阈值”），这些阈值会触发特定的降低风险的措施，其目标取决于现存的缓解措施条件下所有利益相关方的风险。鉴于未来的模型能力存在很高的不确定性，风险阈值可能需要定期调整。

定义风险阈值，这项工作基于构成超过阈值的结果，并与给定模型或模型组合可能表现出的危险能力相结合。例如一家前沿人工智能机构可能将目标确定为避免部署显著增加网络攻击或欺诈风险的人工智能系统。

实施风险阈值，包括具体的、可测试的观察结果，以便获得相同信息的多个观察者能够就是否满足给定阈值达成一致。具体观察将为前沿人工智能机构提供主动确定如何应对困难的潜在挑战的机会，从而在出现此类情况时立即做出反应，并允许问责和外部验证。

然而，考虑到人工智能评测还是一门新兴科学，不太可能定义一组可测试的观察结果来足够可靠地检测所有已识别的风险。风险评估不应仅依赖这些预先定义的测试，而可以考虑更广泛的证据来源，包括在探索性分析、专家预测或其他前沿人工智能机构的相关风险信息中出现的令人担忧和意外的观测结果。特别需要考虑任何由给定模型与其他模型或工具（无论是由同一前沿人工智能组织开发还是其他）的组合引起的风险。

根据需要继续完善和重新定义模型的风险评测框架，以缩小风险阈值预期目标与其当前可操作性之间的差距。由于评测科学和我们关于模型能力的认知有限，这种差距预计确实存在，因此建立健全此种框架的进程中可能需要多次迭代。风险评估框架可使用多种方法，包括概率估计和对当前能力的定性评估。

降低“超过”阈值的风险。这可以通过故意设置保守的阈值来实现，例如使用有意设低的缓冲阈值来触发措施，这样在较早阶段已经实施缓解措施的情况下，最令人关注的阈值就很难被超过。

制定风险阈值时与相关外部利益相关方合作。风险阈值通常与前沿人工智能机构给社会带来的外部影响相关，包括人工智能进步的潜在重大益处和可能对特定群体产生不成比例的负面影响。因此，可以公开其风险阈值以接受外部审查，并与包括相关政府机构在内的外部利益相关方协商确定阈值。

3. 根据每个风险阈值预先承诺采取特定的额外缓解措施，然后进行剩余风险评估

在每个风险阈值处，主动承诺仅在特定缓解措施到位的情况下才继续执行某些开发或部署步骤。此类缓解措施可能包括本报告中概述的许多实践。

采取缓解措施后，重新评估所带来的任何剩余风险以确定是否需要采取额外的缓解措施。由于能力进步的不可预测性和模型评测科学的局限性，预先约定的缓解措施可能不足以将给定模型置于风险阈值内。

可以使用风险接受标准，就像许多其他情况下的标准一样，可能是一个重要的澄清工具。这些标准可能会随着时间的推移而演变，并且可以是定量的或定性的。例如只有将风险降低到“尽可能低”的水平时，风险才可以被接受。

当达到风险阈值时，通知相关监管机构并给出建议的缓解措施。并在部署前实施了缓解措施和剩余风险评估之后，再次通知政府。除政府外，还应酌情主动接触其他利益相关方。

4. 根据开发和部署的相关阶段调整缓解措施，认识到模型的使用方式和环境可能与预期不同

在规划所需的缓解措施时，考虑开发和部署的全部阶段。一般来说，满足风险阈值可能需要在多个阶段采取缓解措施。重要阶段可能包括：

- 模型的继续训练

- 小规模部署模型，例如内部使用
- 大规模部署模型，例如通过API公开发布
- 通过更大的可使用范围扩展模型，例如工具使用或互联网访问
- 不可逆的部署，例如模型开源

调整缓解措施，以应对模型在意外情景或被非预期使用的风险，例如开源后被修改去除安全保护措施的模型，或与另其他模型组合用于非预期目的。例如在给定的风险阈值下，甚至可能在前沿人工智能机构内部使用模型之前就采取信息安全控制缓解措施。考虑到当前信息安全控制和模型新兴能力预测的局限性，通常采用谨慎的态度可能是有帮助的。

特别要认识到，因为存在无法充分保护模型以防止不良行为者获取模型权重的风险，即使没有部署或没有广泛部署，拥有某些模型也可能是危险的。相比之下，其他模型可能只有在大规模或不可逆转的部署方式下才可能构成重大风险。

在大规模或不可逆转的方式部署模型之前，先以小规模或可逆转的方式部署模型。这使得前沿人工智能机构能够在危害变得太大或不可避免之前注意并缓解危害。

5. 若在未预先约定缓解措施的情况下达到风险阈值，则做好暂停开发和/或部署的准备

如果在没有承诺的风险缓解措施的情况下达到风险阈值，则准备暂停训练或减少对已部署模型的访问。这可能涉及警告现有客户可能会减少访问权限，并制定应急计划以尽可能减少对客户使用的负面影响。

考虑开源模式的潜在风险。承认即使在采取缓解措施后，某些模型如果以“开源”方式提供，也可能带来额外风险。这是因为无法召回开源模型以及用户有可能移除安全措施并引入新的（且有潜在危险的）能力。然而开源人工智能系统对研究人员的显著益处也很重，包括提高人工智能安全性。在某些情况下，这些益处可能超过这些潜在风险。

6. 与相关政府部门和其他人工智能企业分享风险评估流程和风险缓解措施的细节

定期向利益相关方通报最新的风险评估和缓解措施：这将有助于评估人工智能机构是否拥有足够的风险管理流程，建立最佳实践的蓝图，并提出解决差距的建议。与外部参与方共享这些信息时，应考虑商业敏感信息。如果出现重大进展，可以提供额外的临时更新。

包括有关模型评测、风险评估和缓解以及相关人员的信息。例如：

- 正在对哪些类型的模型运行何种测试和评测
- 正在使用哪些其他风险评估方法、专业知识以及是否涉及受影响的利益相关方
- 如何监测风险缓解措施
- 哪些团队和个人参与风险管理流程的不同阶段（以及第三方是否以及如何参与）
- 为解决特定类别的风险而采取的措施，例如网络安全措施

考虑公开风险评估和缓解流程的概要总结，以进行更广泛的审查并建立公众对人工智能系统安全性和可靠性的信心。敏感细节可以被剔除。

7. 承诺建立健全的内部问责和治理机制，并接受外部验证

引入稳健且有意义的问责机制，特别是在评测能力阈值时，制定明确的流程，以确保在达到阈值时能够遵循正确的缓解措施或行动方案。这可能包括董事会签署负责任能力扩展策略，以及关键决策的个人问责制。

建立有效的风险治理，确保风险得到适当识别、评估和解决，并透明地报告风险的性质和规模。最重要的是，提供内部制衡，其中可能包括在风险管理中进行深思熟虑的角色分离。

引入验证机制，以便外部参与方能够增强对负责任扩展策略按预期执行的信心。

信息共享的潜在机制包含在“模型报告和信息共享”中。

重点案例

Anthropic：第一个发布负责任扩展策略的前沿AI企业[113]

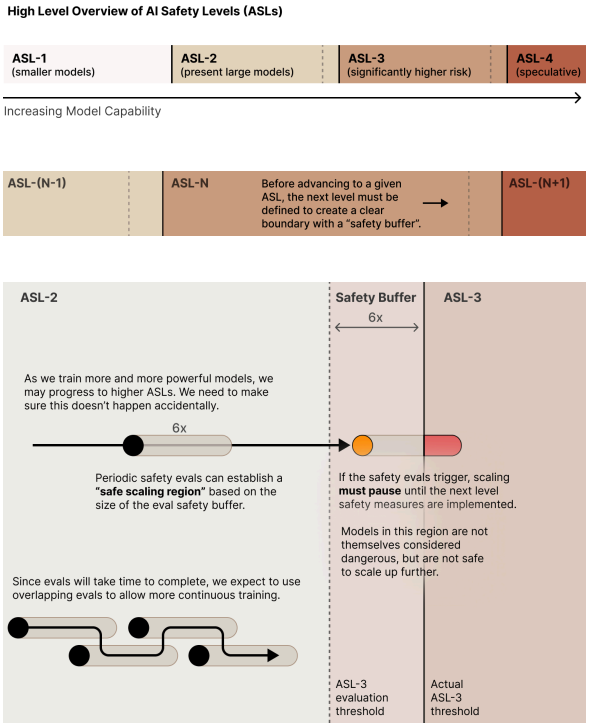
2023年9月19日，Anthropic设计并发布了负责任扩展策略(RSP)，以管理前沿人工智能模型的潜在风险。

Anthropic的RSP专注于灾难性风险——人工智能模型直接导致大规模破坏的风险。此类风险可能来自于故意滥用模型（例如恐怖分子利用模型来制造生物武器），也可能来自于模型以与设计者意图相违背的方式自主行动而造成破坏（例如对齐失败[114]）。虽然人工智能呈现了一系列必须解决的风险，但Anthropic的RSP旨在应对这一风险范围中更极端的风险。

Anthropic的RSP计划的核心是人工智能安全级别(ASL)的概念，它大致模仿了美国政府处理危险生物材料的生物安全级别(BSL)标准。他们定义了一系列人工智能能力阈值，这些阈值代表着不断增加的潜在风险，因此每个ASL都需要比前一个更严格的安全、保障和操作措施。

更高的ASL模型也可能与日益强大的有益应用程序相关联，因此Anthropic的目标不是禁止这些模型的开发，而是通过适当的预防措施安全地启用它们。

[113] Anthropic's Responsible Scaling Policy (Version 1.0), 2023-09-19, https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf.
[114] 机器之心，“AI对齐失败数据库”, 2023-07-09, https://sota.jiqizhixin.com/alignment-db.



- 随着模型能力提升，**保护措施需相应提升**
- 不会立即定义所有未来的ASL及其安全措施，而是采取**迭代承诺**
- 现在定义ASL-2(当前级别)和ASL-3(下一级别)，并承诺在达到ASL-3前定义ASL-4，依此类推
- 随着训练越来越强大的模型，可能会进步到更高的ASL，需确保这不会发生意外
- 定期安全评估可以根据评测安全缓冲区的大小建立**“安全扩展区域”**
- 由于评测需要时间才能完成，因此希望使用重叠评测来允许更连续的训练
- 如触发安全评估，则扩展**必须暂停**，直到实施下一级安全措施；该区间的模型本身并不被认为是危险的，但进一步扩大规模并不安全

Anthropic的负责任扩展策略1.0版[113]

因此，RSP带来了识别和管理灾难性风险的具体和实证方法。其目的是让Anthropic的安全研究和技术的社会效益应用能够继续发展和规模化，同时实施严格的流程来衡量和缓解风险。如果无法缓解重大风险，Anthropic将暂停相关模型的进一步扩展，避免部署它，或将其从部署中删除，直到他们可以确保其足够安全，可以继续。

通过在扩展速度超过安全性时暂停开发和部署，Anthropic就有动力解决必要的安全问题。如果RSP被广泛采纳作为前沿实验室的标准并得到政府的支持，那么可能会产生一种“争先恐后”的动态，竞争激励将直接转向为解决安全问题。

值得注意的是，RSP还具有灵活性和适应性。RSP为当前(ASL-2)和近期(ASL-3)人工智能系统指定了具体的安全承诺，涵盖安全控制、训练监督、红队测试、模型评估和负责任的部署措施。这承诺Anthropic会在达到更高的ASL(从ASL-4开始)之前，迭代地定义它们，以确保安全协议随着时间的推移和新获得的经验证据一起发展，与能力保持同步。因此，RSP应该随着时间的推移而发展，并保持其长期相关性。

OpenAI：发布近似RSP的“准备框架测试版”Preparedness Framework(Beta)[115]

OpenAI认为，目前对前沿人工智能风险的研究远远没有达到可能和所需的水平。为弥补这一差距并使其安全思考系统化，OpenAI在2023年12月18日发布了准备框架测试版，描述了他们跟踪、评估、预测和防范日益强大的模型带来的灾难性风险的流程。

OpenAI有多个安全和政策团队共同努力降低人工智能带来的风险。

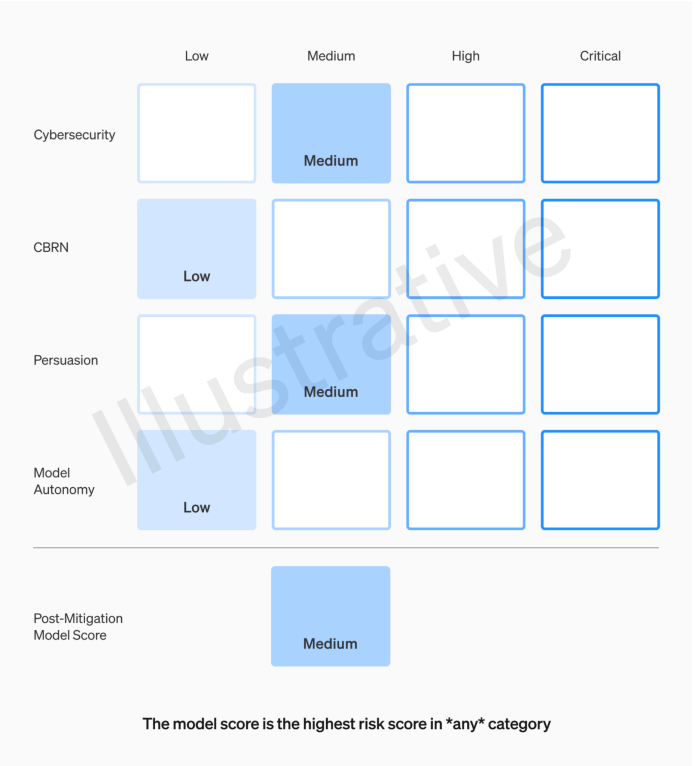
1. **安全系统团队**专注于减少**现有模型和产品**（例如ChatGPT）的滥用。
2. **超级对齐团队**为希望在更遥远的未来拥有的**超级智能模型**的安全性奠定了基础。
3. **防范准备团队**绘制了**前沿模型**的新风险（目前重点关注目前网络安全、化学/生物/辐射/核威胁(CBRN)、说服和模型自主性），并与安全系统、超级对齐以及OpenAI的其他安全和政策团队联合工作。

准备框架背后的中心论点是，应对人工智能灾难性风险安全的稳健方法需要主动、基于科学的确定何时以及如何安全地进行开发和部署。

OpenAI的准备框架包含五个关键要素：

1. **通过评测跟踪灾难性风险水平。**OpenAI将根据多个跟踪风险类别构建并不断改进评估套件和其他监测方案，并在记分卡中表明当前缓解前和缓解后的风险水平。重要的是OpenAI还将预测风险的未来发展，以便可以制定安全保障措施的交付时间。
2. **寻找未知的未知。**在当前未知的灾难性风险类别出现时，OpenAI将持续运行识别和分析以及跟踪流程。
3. **建立安全基线。**只有缓解后得分为“中”或以下的模型才能部署，并且只有缓解后得分为“高”或以下的模型才能进一步开发。此外，OpenAI将确保安全性针对任何具有“高”或“关键”预缓解风险水平的模型进行适当调整。OpenAI还建立程序承诺进一步明确他们如何实施准备框架概述的所有活动。
4. **为防范准备团队安排实地工作。**该团队将推动准备框架的技术工作和维护，包括进行风险研究、评估、监测和预测，并通过定期向安全咨询小组提交报告来综合这项工作。报告内容将包括最新证据的摘要，并就OpenAI提前规划所需的更改提出建议。防范准备团队还将呼吁并与相关团队（例如安全系统、安保、超级对齐、政策研究）进行协调，以整理缓解措施建议并纳入报告中。此外，防范准备团队还将管理安全演习并与可信赖的人工智能团队协调第三方审核。
5. **创建跨职能咨询机构。**OpenAI正在创建一个安全咨询小组(SAG)，汇集整个公司的专业知识，帮助OpenAI的领导层和董事会为他们需要做出的安全决策做好充分准备。因此，SAG的职责将包括监督风险形势的评估，并维持处理紧急情况的快速流程。

[115] OpenAI, Preparedness Framework (Beta), 2023-12-18, https://cdn.openai.com/openai-preparedness-framework-beta.pdf.



OpenAI将进行评估并不断更新其模型的“记分卡”[115]

整体而言，OpenAI提出“要将建造者的心态带入安全领域”[116]，准备框架的安全工作建立在科学与工程紧密结合的基础上，为了使安全工作跟得上创新的步伐，不能停步，而需要通过迭代部署不断学习。

延伸阅读

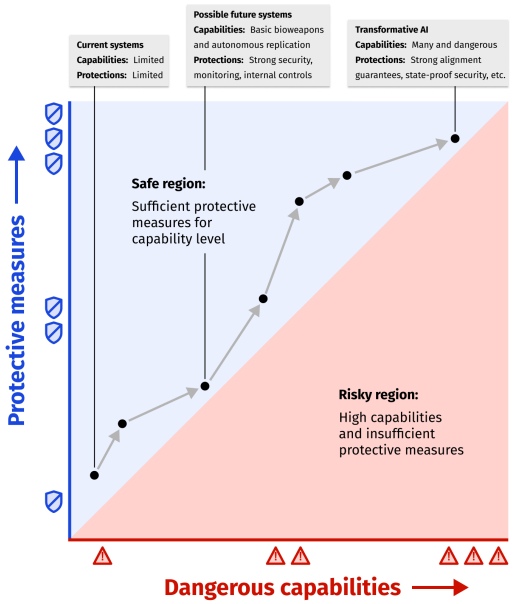
METR(原ARC Evals)：负责任扩展策略的框架提出者[117]

- **METR定义的RSP：**规定了AI研发人员在当前的防护措施下能够安全处理的AI能力水平，以及在保护措施改善之前继续部署AI系统或扩大AI能力会变得过于危险的条件。
- **致力于推动模型评测科学**：帮助实验室实施RSP以可靠地预防危险情况，但不会造成过度负担，并且在安全时不会阻止AI研发。RSP是降低AI灾难性风险最有前途的途径之一，因其提供了：
    - 一个务实的中间立场（介于认为AI极其危险需暂停和担心AI灾难性风险为时过早之间，且基于评测而非猜测）
    - 了解优先考虑哪些保护措施（从以谨慎为导向的原则转向为继续安全研发所需的具体承诺，如信息安全、拒绝有害请求、对齐研究等）

[116] OpenAI, Preparedness, 2023-12-18, https://openai.com/safety/preparedness.
[117] METR, “Responsible Scaling Policies (RSPs)”, 2023-09-26, https://metr.org/blog/2023-09-26-rsp/.

- 基于评测的规则和规范（可能包括标准、第三方审核和监管，自愿的RSP可为流程和技术提供测试平台，有助于未来基于评估的监管）

- **提供咨询指导：**作为非营利第三方机构，METR已经为Anthropic[118]、美国参议员[119]、英国政府官员[120]等负责任扩展的相关设计和讨论提供了咨询指导，还与Anthropic和OpenAI合作进行了危险能力评测[121]。



- RSP的目标是在AI具有任何**危险能力**之前采取**保护措施**
- 不同机构对于“负责任扩展”**可能有不同含义**
- **一个好的RSP应包括**：限制、保护、评测、响应、问责五个方面
- METR还提供了关键组件清单[122]，主要面向寻求推进前沿AI能力的研发人员，同时也为不推进前沿AI能力的研发人员提供了精简版RSP

METR的负责任扩展策略示意图[117]

[118] Anthropic, “Anthropic's Responsible Scaling Policy”, 2023-09-19, https://www.anthropic.com/index/anthropics-responsible-scaling-policy.
[119] Scott Wiener, “Senator Wiener Introduces Safety Framework in Artificial Intelligence Legislation”, 2023-09-13, https://sd11.senate.ca.gov/news/20230913-senator-wiener-introduces-safety-framework-artificial-intelligence-legislation.
[120] UK Government, “Secretary of State speech at CogX Festival”, 2023-09-12, https://www.gov.uk/government/speeches/secretary-of-state-speech-at-cogx-festival.
[121] 安远AI, “ARC Evals首份公开报告：以现实的自主任务评测语言模型自主体”, 2023-09-15, https://mp.weixin.qq.com/s/nbQwfoVIFM5RVHv0FxkeDQ.
[122] METR, “Key Components of an RSP”, 2023-09-26, https://metr.org/rsp-key-components/.

总结与建议

本报告提供了前沿人工智能机构潜在的最佳实践清单，以及面向中国机构的研发实践案例与政策制定指南。

在报告整体讨论的基础上，我们建议：

第一，构建多层次的安全保障体系，实现前沿人工智能的全生命周期风险管理流程[123]**。**人工智能安全面临复杂的风险和威胁，需要建立系统性的安全保障，并从多个层面进行防范。我们建议各项最佳实践应当联合使用，借鉴网络安全等领域中的“纵深防御”[124]策略，在“事前算法备案，事中风险评估，事后溯源监测”[125]多方面落地工具，实现前沿人工智能安全的全流程覆盖。

第二，为人工智能安全研究分配更多研发资金，并面向更高级的系统和更复杂的场景。全球人工智能安全峰会前，三位图灵奖获得者、一位诺贝尔奖获得者、国内多位院士共同撰文《人工智能飞速进步时代的风险管理》[126]并签署了一份联合声明[127]，提出分配至少三分之一的人工智能研发资金用于确保人工智能系统的安全性和合乎伦理的使用。而主流的RLHF对齐方法及其简单扩展存在根本局限，难以拓展到更高级的系统，面向超级智能的对齐问题需要更好的技术途径[128]。其他相对成熟的研究领域，例如对抗鲁棒性，也应与时俱进将攻击模式多样化、通用化。

第三，及时论证在国内成立前沿人工智能安全测试机构的必要性和紧迫性。英国和美国相继成立了人工智能安全研究所[129]，新加坡政府今年也成立了类似的人工智能验证基金会[130]，国内的实践落地方式值得相应探讨。另外，目前中文大模型的安全评测大多限于对输出文本的评测，但逼近GPT-4性能的模型有必要进行生物研发、网络攻击、自主行动等危险能力评测[131]。

---

[123] 清华大学人工智能国际治理研究院，“我国算法治理政策研究报告”, 2022-12-01, https://aiig.tsinghua.edu.cn/info/1025/1759.htm.

[124] 百度百科，“纵深防御”, 2022-09-27, https://baike.baidu.com/item/%E7%BA%B5%E6%B7%B1%E9%98%B2%E5%BE%A1.

[125] 中国信通院&中国科学院，“大模型治理蓝皮报告”, 2023-11-25, http://www.caict.ac.cn/kxyj/qwfb/ztbg/202311/P020231124526622371194.pdf.

[126] 安远AI，“授权中译版｜三位图灵奖和中外多位顶尖AI专家的首次政策建议共识：呼吁研发预算1/3以上投入AI安全，及若干亟需落实的治理措施”, 2023-1-024, https://mp.weixin.qq.com/s/zdrGCiagDYqa6kPljK2ung.

[127] 安远AI，“AI的帕格沃什会议！中美英加欧20多位顶尖AI专家线下聚首，呼吁AI安全与治理的全球协同行动”, 2023-11-01, https://mp.weixin.qq.com/s/1WbrS-L8Qsww10nosADwJQ.

[128] 安远AI，“在和OpenAI的「超级对齐」闭门研讨会上，安远AI讲了什么？”, 2023-10-25, https://mp.weixin.qq.com/s/yZu8-t7Vdvn63BCPH3cfoA.

[129] 安远AI，“为什么中国的参与必不可少？我参加首届全球人工智能安全峰会的所见所思（万字回顾）”, 2023-11-09, https://mp.weixin.qq.com/s/_SWLDzKDOMNb04ha1SNKFg.

[130] AI Verify Foundation，“AI Verify Foundation”, 2023-10-18, https://aiverifyfoundation.sg/.

[131] 安远AI，“博鳌经安论坛发布｜安远AI《前沿大模型的风险、安全与治理》报告”, 2023-10-29, https://mp.weixin.qq.com/s/6AUz6rJm4XbOyETi08W5LQ.

可参考国际上METR[132]等机构的第三方评测实践，在国内推动建立面向前沿人工智能风险的第三方评测。

第四，推动建立前沿人工智能风险分级管理和快速响应机制。明确不同安全级别模型的评估流程、使用规范、第三方审核等要求，推动行业企业和科研机构落实安全管理措施。确保相适应的安全标准，行业自律和政府监管缺一不可。可借鉴国内已有的流程机制[133]，并将其扩展应用于前沿人工智能风险管理，例如建立针对训练高风险前沿大模型的许可制度、建立人工智能模型和系统的漏洞报告流程等。

第五，加强前沿以及通用人工智能的治理研究，积极发展“技术治理”。科研机构应积极布局治理领域，及早谋划潜在风险的防范对策。监管部门也需要加强对人工智能治理的能力建设，可借鉴英国在政府内部成立专门团队[134]以评估前沿人工智能潜在风险的积极尝试。我们倡导利用技术专长来提升人工智能治理的效率和效果[135]，从而有效地弥合人工智能研发与监管在技术理解和政策制定方面的差距。同时，我们也鼓励探索具有中国特色和全球意义的人工智能治理方案，例如对关键信息基础设施的系统性安全保护、人工智能治理和评测创新试验区等。

[132] METR，“ARC Evals is now METR”，2023-12-04, https://metr.org/blog/2023-12-04-metr-announcement/.
[133] 安远AI，“全球首份！安远AI发布《中国人工智能安全全景报告》(State of AI Safety in China)”，2023-10-23, https://mp.weixin.qq.com/s/KLzvu43U44_JfL6nxs34xg.
[134] UK Government，“Frontier AI Taskforce: first progress report”，2023-09-07, https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report.
[135] 中央网信办，“全球人工智能治理倡议”，2023-10-18, http://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

附录

附录A. 前沿人工智能研发机构的治理政策（2023年10月）

在全球人工智能安全峰会前，7家前沿人工智能研发机构受邀分享了在确保前沿人工智能安全方面的实践措施：

- 亚马逊：https://aws.amazon.com/uki/cloud-services/uk-gov-ai-safety-summit/
- Anthropic：https://www.anthropic.com/uk-government-internal-ai-safety-policy-response
- 谷歌DeepMind：https://deepmind.com/public-policy/ai-summit-policies
- Inflection：https://inflection.ai/frontier-safety
- Meta：https://transparency.fb.com/en-gb/policies/ai-safety-policies-for-safety-summit/
- 微软：https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/
- OpenAI：https://openai.com/global-affairs/our-approach-to-frontier-risk

英国政府也编写了对7家前沿人工智能研发机构安全实践的补充材料：

- 前沿人工智能安全的新兴流程：https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety

上述材料共同构成了本报告的主要参考资料。

我们鼓励中国机构分享实践案例，协助我们不断优化和更新这些最佳实践，并在此基础上形成可以向国际推广的中国实践！

附录B. 剑桥大学未来智能研究中心的分类整体评分

剑桥大学未来智能研究中心邀请了由15名人工智能学者、监管专家和技术研究人员组成的小组，评估了6家企业的政策，并为每个企业进行了评分和比较[136]。整体情况如下：

实践类别	描述	Amazon	Anthropic	DeepMind	Meta	Microsoft	OpenAI
模型评测和红队	针对多种风险来源和潜在负面影响对模型进行评测	2	2	2	1	2	2
	在模型整个生命周期的多个检查点进行模型评测和红队测试	2	2	2	1	2	2
	允许受信任的外部评测方在模型整个生命周期进行模型评测	1	1	1	1	1	1
	支持模型评测科学的进步	1	2	2	2	2	2
优先研究人工智能带来的风险	开展人工智能安全研究	2	2	2	2	2	2
	开发用于防范系统危害和风险的工具，例如用于防范错误信息和虚假信息的水印工具	2	2	2	2	2	2
	与外部研究人员合作，研究和评估其系统的潜在社会影响，例如对就业的影响和虚假信息的传播	2	2	2	1	2	2
	公开分享风险研究成果，除非分享这些成果可能会造成危害	2	2	2	2	2	2
含保护模型权重在内的安全控制	在整个人工智能系统中实施强有力的网络安全措施和流程	2	2	2	0	2	1
	了解人工智能系统中的资产并采取适当措施来进行保护	2	2	2	0	2	2
	保持对安全风险的最新理解，以便做出明智的风险决策	2	2	2	2	2	2
	制定事件响应、升级和补救计划，并确保响应人员受过评估和应对相关事件的培训	2	2	2	0	2	1
	对系统行为进行持续监测，以便观察行为变化并识别潜在攻击	0	2	2	0	2	2
	通过评测和沟通风险并遵循"设计安全"原则，使用户能够安全使用人工智能系统	2	2	1	0	1	1
	实施有效的防护性安全风险管理，涵盖物理、人员和网络安全纪律	2	2	2	0	2	2
	制定并实施适当的人员安全控制措施以降低内部风险	1	2	1	0	1	1
漏洞报告机制	建立漏洞管理流程	2	2	2	2	2	2
	借鉴已建立的软件漏洞报告流程，建立清晰、用户友好且公开的模型漏洞报告流程	1	2	2	2	2	2
	制定协同漏洞披露和信息共享的协议和机制	1	2	2	1	2	2
人工智能生成材料的标识信息	研究能够识别人工智能生成内容的技术	1	2	2	2	2	2
	探索对各种扰动具有鲁棒性的人工智能生成内容的水印使用	2	2	2	2	2	2
	探索人工智能输出数据库的使用	0	0	0	0	0	0
模型报告和信息共享	共享与模型无关的有关一般风险评估、缓解和管理流程以及最佳实践的信息	0	2	2	1	2	2
	在训练之前、训练期间和部署之前共享有关某些前沿人工智能模型的特定信息	0	1	1	1	1	1
	根据适用性，与不同方共享不同信息，包括政府机构、其他前沿人工智能机构、独立第三方和公众	0	1	1	1	1	1
防止和监控模型滥用	建立流程来识别和监测模型的滥用，例如监测模型被滥用和规避保障措施的常见方式	2	2	2	1	2	2
	实现模型输入和输出过滤器	2	2	2	1	2	2
	实施额外措施来防止有害输出，包括微调、提示和拒绝采样	2	2	2	1	2	2
	实施基于用户的API访问限制和监测，例如减少对无合理理由反复触发内容过滤器的个人的访问权限	2	2	2	1	2	2
	为潜在的最坏情况或持续滥用场景制定响应计划，手段包括快速回滚和撤回模型	0	0	0	0	0	0
	持续评估现有和额外保护措施的有效性和可取性，因其也可能阻碍正面用途并减少隐私	2	2	2	2	2	2
数据输入控制和审核	在收集训练数据之前，实施负责任的数据收集实践	2	1	1	2	2	2
	在使用输入数据训练人工智能系统之前对其进行审核，例如尝试识别可能产生危险能力的数据	1	0	2	1	1	2
	根据数据审核结果采取适当的风险缓解措施，例如通过整理数据集以确保不会在某些数据上进行训练	1	0	2	2	2	2
	通过邀请外部参与方评估其输入数据并共享输入数据审核信息，促进对输入数据的外部审查	0	0	0	0	0	0
负责任扩展策略	在开发或部署新模型之前进行彻底的风险评估，并辅之以持续的监测	1	2	2	2	2	2
	预先确定"风险阈值"，限制可接受的风险水平	0	2	0	0	0	0
	根据每个风险阈值预先承诺采取特定的额外缓解措施，然后进行剩余风险评估	0	2	0	0	0	0
	根据部署的阶段调整缓解措施，认识到模型的使用方式和环境可能与预期不同	0	2	0	0	0	0
	若在未预先约定缓解措施的情况下达到风险阈值，则做好暂停开发和/或部署的准备	0	2	0	0	0	0
	与相关政府部门和其他人工智能企业分享风险评估流程和风险缓解措施的细节	0	1	1	1	1	1
	承诺建立健全的内部问责和治理机制，并接受外部验证	0	2	2	0	2	2
	小计	49	69	63	40	63	62
	占比	58%	82%	75%	48%	75%	74%
	图例	2 = 符合	1 = 也许符合	0 = 不符合			

[136] Leverhulme Centre for the Future of Intelligence，“Do companies's AI safety policies meet government best practice?”, 2023-10-31, http://lcfi.ac.uk/news-and-events/news/2023/oct/31/ai-safety-policies/.