基础模型的负责任开源

—— 超越开源闭源的二元对立：负责任开源的内涵、实践与方案

安远AI
北京大学人工智能研究院
北京大学武汉人工智能研究院
北京通用人工智能研究院

2024年4月

执行摘要

1. 开源基础模型已成为创新的重要驱动力之一

根据斯坦福大学《2024年AI指数报告》[1]，2023年全球总共发布了149个基础模型，比2022年发布的数量翻了一倍还多，**而且更高比例是开源的**。在这些新发布的模型中，有65.7%是开源的，相比之下，2022年只有44.4%，2021年只有33.3%的模型是开源的。

根据全球开源社区Hugging Face的调研[2]，**Llama 1和Llama 2现在已经衍生出了3万个新模型**。多位专家预计，即将推出的Llama 3 400B将会是“首个GPT-4级别的开源模型”。

2. 如何治理开源AI已成为短期内重要的未解决议题之一

本报告从安全治理的角度探讨开源AI的政策和实践。在制定相关政策时，各国需要综合考虑促进创新生态、技术的安全性与可控性、隐私保护、知识产权、伦理与责任、国际合作与标准制定、市场竞争环境、教育与公众参与等多个方面。这些维度与各国的战略考虑及监管取向相结合，共同构成了开源AI的治理政策框架。

全球范围内，许多国家和地区，包括欧盟、美国、英国、法国、中国以及其他全球南方国家，都在积极制定AI相关政策，**开源AI也成为多项政策探索的核心**。尽管这些政策旨在平衡技术发展与安全需求，但在监管取向和具体条款的设计上存在显著差异，这部分原因是由于政策制定过程中缺乏关于风险、收益及潜在影响的严谨证据。

3. 前沿AI开源的主要争论

领先的基础模型研发机构近年决定开源其模型或限制对其模型的访问，引发了关于是否以及如何开放能力日益增强的基础模型的争论。

我们识别了两种主要立场：一方是审慎开放的倡导者，他们担心前沿AI开源成为潜在不安全技术**“不可逆转的扩散”**，并主张在确保安全的基础上逐步推进开放；另一方则是鼓励开放的支持者，他们认为前沿AI开源是**“确保对技术信任的唯一途径”**，强调开放性对于促进创新和透明度的重要性，并反对过度限制的做法。尽管在风险与收益的评估、开源方式、安保措施以及监管政策等方面存在分歧，但**两方都认同开放性在推动技术进步和促进社会福祉方面的重要作用，以及前沿AI开源的潜在风险，都主张在开源前应采取必要的评测等安全措施。**

此外，从企业视角看，有关开源和闭源的讨论和实践**或多或少地带有商业利益的考量**。

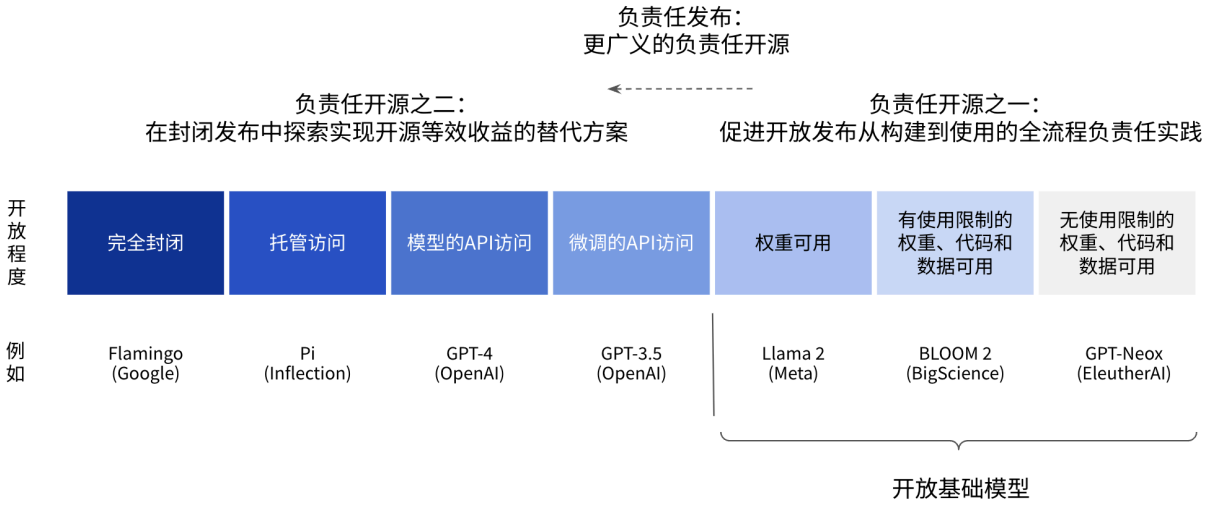
[1] Stanford HAI, “2024 AI Index Report”, 2024-04-15, https://aiindex.stanford.edu/report/.
[2] Clem Delangue, “Llama 3 is officially the fastest model from release to #1 trending on Hugging Face - in just a few hours.”, 2024-04-19, https://twitter.com/ClementDelangue/status/1781068939641999388.

4. 超越简单化的“开放与封闭”争论

虽然开放基础模型带来了发展与安全之间的紧张关系不可能完全消除，但我们**提倡可以超越简单化的一维视角，探索更丰富的发布政策设计空间**。

将AI模型简单地划分为开源或闭源是一种过于简化的做法。开源AI的概念尚未得到清晰定义，与开源软件不同，AI模型的“源代码”可能包括多种组件，这些组件的开放程度可以各异。此外，从“完全开放”到“完全封闭”的发布选项实际上是多样的，需要明确的标准和定义来权衡透明性、安全性和商业考量。

根据多个角度的安全和治理评测，我们依然无法得到开放或封闭模型哪个更有明显优势的结论。综合模型安全性评测，开放模型和封闭模型均显示出对各种攻击的脆弱性。AI研发机构治理评测指出，倾向于开放模型的机构和倾向于封闭模型的机构各有所长。



本报告的讨论范围设定参考了斯坦福大学基础模型研究中心的“开放基础模型”概念图[3]

5. 推动基础模型负责任开源的务实方案

开源是科学和创新的重要驱动力，但同时需要权衡其潜在风险，对未来更强的前沿AI不同程度开源可能引入更大的潜在风险。因此，我们建议推动负责任开源，这包括两个层面：

第一，**促进开放发布从构建到使用的全流程负责任实践**。建议根据基础模型的生命周期和流程阶段，设计构建和使用阶段的负责任开源维度，并针对不同能力级别的模型制定差异化的负责任开源要求。例如对于大多数AI模型，负责任主要体现在提高透明度、确保合规和促进创新。而对于能力更强的前沿模型，需要实施与模型的潜在风险相称的评测和安全缓解措施。

[3] Rishi Bommasani et al.，“Considerations for Governing Open Foundation Models”，2023-12-13, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.

第二，**在封闭发布中探索实现开源等效收益的替代方案。**建议开发者应考虑开源的替代方案，在获得技术和社会效益的同时，又没有太大的风险。包括为受信任的研究人员提供结构化访问，以帮助识别安全或道德缺陷，鼓励独立第三方的审核等。

虽然**严格意义上我们讨论的是“负责任发布”**，但我们希望通过突出“负责任开源”的概念，推动开源AI安全治理的讨论，并促进负责任开源实践的发展。

6. 面向四类目标群体和国际合作分别提出建议

本报告是为中国的基础模型研发机构、AI开源社区、AI治理/政策/立法专家、AI投资方和资助方编写的，其目的是作为基础模型的负责任开源的决策和实践的参考。我们鼓励相关机构和专家进一步探讨负责任开源的内涵，实施负责任的开源实践和方案。**我们倡导在全球范围内展开合作**，通过负责任开源助力发展中国家提升AI技术和治理能力，推动形成具有广泛共识的高风险模型治理框架和标准规范。

开源AI的负责任实践并非一成不变，而是会随着技术发展和社会需求的变化而不断演进。可以预见，未来开源与闭源的讨论将更加深入和细化，可能会出现更多创新的发布模式和治理机制，以适应不断变化的环境和挑战。在这个过程中，各方面的合作和对话将至关重要。

术语定义

本报告聚焦——基础模型的负责任开源。

大规模机器学习模型相关术语，主要参考斯坦福大学、智源研究院：

- **基础模型(Foundation Model)：**在大规模广泛数据上训练的模型，使其可以适应广泛的下游任务；国内外学界通常简称为“大模型”。

模型开源和开放相关术语，主要参考斯坦福大学、牛津大学研究机构：

- **开源AI(Open-Source AI)：**概念尚未得到清晰定义，不同机构都用它来表示不同程度的“公开可用”；开放源代码促进会(OSI)等机构正致力于明确定义开源AI。
- **开放基础模型(Open Foundation Models)**：基础模型在发布时，其权重是广泛可用的；不严格区分时，也会称为“开源基础模型”“开放模型”“开源模型”。
- **封闭基础模型(Closed Foundation Models)：**基础模型在发布时，其权重不是广泛可用，可能受一定限制或完全封闭；不严格区分时，也会称为“闭源基础模型”“封闭模型”“闭源模型”“受限模型”。
- **负责任开源(Responsible Open-Source)：**开源项目的维护者和贡献者在开源过程中遵循一定的道德和法律标准，确保技术的构建和发布对社会和个人是安全和有益的，这可能包括安全性、透明度、可访问性、包容性、合规性、社区治理和生态和创新影响等方面。

模型能力相关术语，主要参考全球AI安全峰会、前沿模型论坛：

- **前沿AI(Frontier AI)：**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，提供了最多的机遇但也带来了新的风险。

人工智能风险相关术语，主要参考牛津大学研究机构：

- **灾难性风险(Catastrophic Risk)：**一种可能发生的事件或过程，若发生将导致全球约10%或更多人口丧生，或造成类似损害。

致谢

本报告的主要贡献者：
安远AI：方亮（主要撰写人）、谢旻希、程远、段雅文
北京大学人工智能研究院：杨耀东
北京大学武汉人工智能研究院：辜凌云
北京通用人工智能研究院：綦思源

感谢北京通用人工智能研究院院长、北京大学人工智能研究院院长朱松纯教授，北京大学人工智能研究院人工智能安全与治理中心主任、北京大学武汉人工智能研究院副院长张平教授，给予的悉心指导和宝贵建议。

感谢安远AI伙伴潘汉飞、张玲、王婧人对内容的贡献。

目录

1 各国积极发布基础模型相关政策，开源部分取向不同

我认为，如何监管开源人工智能，是短期内最重要的未解决问题。

——加里·马库斯 (Gary Markus)[4]

各国在制定开源AI相关政策时，通常需要综合考虑促进创新生态、技术安全与可控性、隐私保护、知识产权、伦理与责任、国际合作与标准制定、市场竞争环境、教育与公众参与等多个方面，这些维度与各国各地区的战略考虑或监管取向相结合，共同组成了对于开源AI的治理政策框架。

欧盟、美国、英国、法国、中国和其他全球南方等国家和地区在开源AI的治理上，虽然都希望能平衡发展与安全，但整体监管取向和具体条款设计有所不同。

1.1 欧盟《AI法案》创全球首部全面AI监管法，设独特开源豁免规定

2023年12月8日，欧盟就《AI法案》[5]达成协议，该法成为全球首部针对AI进行全面监管的法案。2024年2月2日，欧盟理事会常务代表委员会就《AI法案》进行表决，获得全票通过。2024年3月13日，欧洲议会以523票赞成、46票反对和49票弃权通过了《AI法案》，这标志着欧盟在AI技术的监管上走在了世界前列。接下来《AI法案》还需得到欧盟理事会的正式批准，但最艰难的立法阶段已经过去，距离最终出台仅一步之遥。

欧盟《AI法案》整体倾向于“监管优先”，但也力图平衡AI风险管控、保护中小企业。法案自2021年以来一直在制定中。自那时起，该技术已经发生了快速而显著的发展，该提案也经历了多次修订以跟上步伐。ChatGPT的出现引发了控制基础模型的一轮修订。谈判在2023年2月底达到白热化程度。法国、德国和意大利为了保护本国的开发者，试图削弱对基础模型的限制[6]。最后敲定的协议条款对被认为具有特别危险性的AI的使用做出了限制，但减轻了中小型公司和模型开源、开发的负担。

[4] David Harris, “Open-Source AI Is Uniquely Dangerous”, 2024-01-12, https://spectrum.ieee.org/open-source-ai-2666932122.
[5] European Parliament, “Artificial Intelligence Act”, 2024-04-23(引用日期), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.
[6] Gian Volpicelli, “Power grab by France, Germany and Italy threatens to kill EU's AI bill”, 2023-11-20, https://www.politico.eu/article/france-germany-power-grab-kill-eu-blockbuster-ai-artificial-intelligence-bill.

到目前为止，欧盟主要通过以下这些法案对开源模型或软件的安全管理做出规定：

时间	主要法案	对开源AI或软件的规定
2024年3月	《AI法案》	**免费或用于科学研究和开发目的而投入使用的开源AI系统可豁免该法案**，但该豁免不适用于被认为会带来系统性风险的模型、商业化的开源AI产品。
2024年3月	《产品责任指令》	该指令适用于所有AI产品和系统，但**不应适用于在商业活动之外开发或提供的免费且开源软件及其源代码**。
2024年3月	《网络弹性法案》	根据开源软件的所属和开发方式，实施**分层安全管理**。
2020年10月	《开源软件战略(2020-2023年)》	建立世界一流的公共服务，鼓励更多地使用开源软件来进行构建，鼓励共享和重复使用软件、应用程序，以及数据、信息和知识，以期通过共享源代码来为知识社会作出贡献。

欧盟对开源AI或软件做出规定的主要法案（本报告自制）

注：《网络弹性法案》《产品责任指令》《AI法案》欧洲议会已批准，还需欧盟理事会正式批准后生效

欧盟《AI法案》对开源AI设定了一些独特的规定。

该法案规定，所有通用AI系统的开发者必须确保透明度，**但如果AI系统是免费开源的，则可享有特别豁免**。然而，**此豁免不适用于商业化的开源AI产品**，比如那些提供付费技术支持或通过广告覆盖成本的企业。重要的是，**对开源AI的这种豁免不适用于被认为具有系统性风险的模型，例如任何使用了超过10^{25}浮点运算次数(FLOPs)训练的AI模型**，这些模型必须遵守更为严格的规定，包括提供详细的技术文档和进行安全测试。

法案还特别提到，为科学研究和开发目的开发的AI系统可以豁免，但这也为根据开源许可证开发的科研用途的模型转为商业用途提供了可能，从而绕过了部分安全法规。此外，欧盟《产品责任指令》的草案[7]也扩大了AI系统的适用范围，涵盖所有AI产品，为了不妨碍创新或研究，该指令也明确说明不应适用于在商业活动之外开发或提供的免费且开源软件及其源代码。但**强调如果非商业活动中开发的免费开源软件一旦被用作产品组件，制造商也可能需对由此引发的缺陷负责**。

《网络弹性法案》的草案进一步引起争议，尤其是关于**上游开源开发者可能要为下游产品的安全缺陷承担责任**的条款，引发了开源社区的广泛反响，一封公开信称该法案可能会对软件

[7] European Parliament, “New Product Liability Directive”, 2023-12-14, https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739341/EPRS_BRI(2023)739341_EN.pdf.

开发产生“寒蝉效应”[8]。对此，欧盟提出了分层安全管理的概念[9]，要求单独开发并控制产品内开源软件的商业实体需要承担所有的合规性责任，但减轻了对开源软件基金会或类似支持组织下协作开发的或多方协作开发开源软件的管理压力，确保开源项目可以在合理的法律框架内运行。2024年3月，欧洲议会正式批准了《网络弹性法案》，之后还须得到欧盟理事会的正式通过才能成为法律，而最终生效要到2027年，这就给了各方时间来满足法律要求并梳理各种合规细节。

在监管实践方面，欧盟设立了AI办公室，负责监督并制定GPAI模型的标准和测试实践[10]。同时，法案鼓励开源开发者采纳广泛接受的文档实践，如模型卡和数据表，以提高透明度和可追溯性，尽管具体细节仍待明确。

结合该法律的广泛范围，法案将显著影响欧盟内开源AI的开发和使用。在某些情况下，例如当一家公司单方面开发开源AI模型时，遵从规定与开发封闭AI模型没有任何不同。这些新规则的复杂性可能会使其更难在欧盟内推广开源AI。

从更广泛的视角看，欧盟希望通过其单方面的市场规制能力，在无需其他国家、国家机构协作的情况下，制定出全球市场遵循的规章制度，引领了全球商业环境“欧洲化”，即“布鲁塞尔效应”[11]。《AI法案》《产品责任指令》《网络弹性法案》以及欧盟GDPR、《数字服务法案》等都是“布鲁塞尔效应”在数字治理领域的体现，但AI法案因可能对技术进步和产业发展产生制约，也受到了广泛质疑[12]。这些法案的最终效果，还有待时间的检验。

1.2 美国白宫《AI行政命令》关注广泛可用的模型权重所带来的挑战

2023年10月30日，美国白宫发布《关于安全、可靠和可信的AI行政命令》[13]（以下简称《AI行政命令》），该行政命令为AI安全和保障建立了新的标准，意图保护美国民众的隐私，促进公平和公民权利，维护消费者和工人的利益，促进创新和竞争，确保政府负责任且有效地使用AI，提升美国在全球的领导力。

---

[8] Eclipse Foundation et al.，“Open Letter to the European Commission on the Cyber Resilience Act”, 2023-04-17, https://newsroom.eclipse.org/news/announcements/open-letter-european-commission-cyber-resilience-act.
[9] Euractive，“EU policymakers' advance on open source software, support period in new cybersecurity law”, 2023-10-31, https://www.euractiv.com/section/cybersecurity/news/eu-policymakers-advance-on-open-source-software-support-period-in-new-cybersecurity-law.
[10] EU Artificial Intelligence Act，“The AI Office: What is it, and how does it work?”, 2024-03-21, https://artificialintelligenceact.eu/the-ai-office-summary/.
[11] Anu Bradford，“The Brussels Effect”, 107:1 (2012), https://ssrn.com/abstract=2770634.
[12] 鲁传颖：欧盟推出全球首部AI法案，会形成效应还是陷阱？2024-03-27，https://ciss.tsinghua.edu.cn/info/zlyaq/7029.
[13] 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/.

该行政命令被外界称为**“有史以来政府为推进AI安全领域所采取的最重大行动”**[14]，以确保美国在把握AI的前景和管理AI风险方面处于**领先地位**，同时也明确了美国政府下一步**行动措施与职责分工**。作为美国政府负责任创新综合战略的一部分，该行政命令以美国政府之前采取的行动为基础，包括促使15家领军企业自愿承诺推动AI的安全、可靠和可信的发展。

美国政府针对开源软件做出过一系列的监管规定，在此基础上，2023年发布的《AI行政命令》和针对基础模型的《透明度法案》，也涉及了对开源模型的要求：

时间	**主要法案**	**对开源AI或软件的规定**
2023年12月	《2023年AI基础模型透明度法案》[15]	旨在指导联邦贸易委员会制定标准，公开有关AI基础模型训练数据和算法的信息，但将考虑为**开源基础模型制定特别规定。**
2023年10月	《关于安全、可靠和可信的AI行政命令》	没有直接提到开源，而是要求商务部长通过公开咨询征求各利益相关方对**具有广泛可用权重的双用途基础模型**相关的潜在风险、利益、其他影响以及适当的政策和监管方法的意见，并向总统提交报告。
2023年9月	《CISA开源软件安全路线图》[16]	高度关注**开源软件漏洞“连锁”效应、供应链“投毒”**等两类特有风险，实行美国政府网络和关键基础设施**“小核心”**重点保护，强化开源生态系统**“大范围”**联动协作。
2023年3月	《2023年开源软件安全法案》[17]	强调了**开源软件作为公共数字基础设施的重要性，**但需针对其安全性挑战**加强监管**。
2022年9月	《2022年保护开源软件法案》[18]	建议美国网络安全和基础设施安全局(CISA)创建**“风险框架”**，以降低使用开源软件系统的风险。

美国对开源AI或软件做出规定的主要法案（本报告自制）

[14] BBC, “US announces 'strongest global action yet' on AI safety”, 2023-10-31, https://www.bbc.com/news/technology-67261284.
[15] US Congress, “AI Foundation Model Transparency Act of 2023”, 2023-12-22, https://www.congress.gov/bill/118th-congress/house-bill/6881.
[16] Cybersecurity and Infrastructure Security Agency, “CISA Open Source Software Security Roadmap”, 2023-09, https://www.cisa.gov/sites/default/files/2023-09/CISA-Open-Source-Software-Security-Roadmap-508c.pdf.
[17] US Congress, “S.917 - Securing Open Source Software Act of 2023”, 2023-03-22, https://www.congress.gov/bill/118th-congress/senate-bill/917/text.
[18] US Congress, “S.4913 - Securing Open Source Software Act of 2022”, 2022-09-21, https://www.congress.gov/bill/117th-congress/senate-bill/4913/text.

《关于安全、可靠和可信的AI行政命令》没有直接提到开源(Open Source)一词，而是对更具体的具有**广泛可用模型权重的双重用途基础模型**(Dual-Use Foundation Models with Widely Available Model Weights)进行了规定，主要集中在第4节（确保AI技术的安全性和安保性）的第4.6小节。该行政命令指出，“当双重用途基础模型的权重被广泛可用时，例如当它们在互联网上公开发布时，创新可能会带来巨大的收益，但也会带来巨大的安全风险，例如移除模型内的安全措施。”具体规定如下：“在命令发布之日起270天内，商务部长应通过通信和信息助理部长，并与国务卿协商，通过公开咨询过程征求私营部门、学术界、民间社会及其他利益相关者对于具有广泛可用权重的双用途基础模型相关的潜在风险、利益、其他影响以及适当的政策和监管方法的意见。并将收集到的意见，与其他相关机构负责人协商后向总统提交一份报告。”这个任务既需要技术洞察力，也需要政策制定的智慧。

对于开源AI有可能促进如生物和化学武器等危险材料生产的问题，白宫行政命令在第4.1小结提到了化学/生物/辐射/核威胁(CBRN)的风险，美国国会目前正在考虑多项法案[19]来应对这些威胁。第4.2小节要求大模型开发者必须分享红队测试中的表现结果，**对使用大于10^{26} FLOPs训练的任何模型，提出了报告安全测试结果和关键信息的要求，另要求生物序列数据相应的门槛是10^{23} FLOPs。**

这些规定旨在通过跨部门合作和公众参与来解决广泛可用的AI模型权重所带来的挑战，并确保在创新和安全之间取得平衡。通过这种方式，行政命令试图制定出既能促进AI技术发展，又能保护国家安全和公共利益的政策框架。

《2023年AI基础模型透明度法案》则是在此基础上，要求联邦贸易委员会在九个月内制定标准，提高AI基础模型在数据和操作方面的透明度，并为**开源基础模型**或基于其他基础模型重新训练或调整的模型**制定特别规定**，显示了对开源AI的一定支持。

在早些时候，国会已经提出了多项关于开源软件的法案，如《2022年保护开源软件法案》和《2023年开源软件安全法案》，旨在通过制定**风险框架和安全职责**，增强开源软件的安全性，**并将其视为重要的公共数字基础设施**。《CISA开源软件安全路线图》进一步强调了联合协作和安全责任，明确了**开源软件的数字公共物品属性**，高度关注**开源软件漏洞“连锁”效应、供应链“投毒”**等两类特有风险。

对此行政命令，不同专家和利益相关者持有不同看法[20]。一些智库专家支持，认为这是美国AI治理的重要步骤，也有专家批评政府可能过度干预，政策制定者似乎准备回避开放式创新模式，这种模式曾使美国公司在几乎所有计算和数字技术领域处于全球领先地位。虽然以OpenAI为代表的业界普遍赞成或至少接受这样的要求，但一些投资者、小型AI行业参与者和

[19] Edward Markey, “Sens. Markey, Budd Announce Legislation to Assess Health Security Risks of AI”, 2023-07-18, https://www.markey.senate.gov/news/press-releases/sens-markey-budd-announce-legislation-to-assess-health-security-risks-of-ai.

[20] 安全内参，“拜登AI行政令评论：美国技术监管方式大转变？或是回避‘开放式创新模式’”，2023-11-01，https://www.secrss.com/articles/60265

学者的代表致函拜登总统[21]担忧自身在政策制定过程中的声音有限，并建议未来应为中小企业设计豁免机制，为大平台设立具体责任，以适应不同的风险和能力水平。总的来说，美国的这些政策和法案在试图促进开源AI的发展的同时，也在努力确保其在不增加国家安全风险的前提下进行创新和应用。政府与公众的合作，以及透明和公正的监管是实现这一目标的关键。

1.3 英国政策文件谨慎对待开放与封闭之争，防范监管捕获

英国的开源产业政策由来已久，政府鼓励创新，并在开源技术的开发上投入了大量的资金和人才，这让其对**鼓励开放创新和AI竞争**的立场一脉相承。

但同时英国政府也关注到开源模型**如果没有没有足够的保护措施，也可能造成伤害**。组织在开发和发布基础模型的不同方式，也为AI的监管引入了复杂性。在英国举办的首届全球AI安全峰会上，在《布莱切利AI安全宣言》等成果与共识之外，开源模型的风险和利弊成为一个主要争论点[22]。英国还在峰会上宣布成立AI安全研究所，并将“开源系统以及采用各种形式的访问控制部署的系统”和“破坏保障措施或利用不安全的模型权重”作为其优先事项之一[23]。

英国对于对开源AI或软件政策文件包括：

时间	政策文件	对开源AI的规定或立场
2024年2月	《大语言模型和生成式AI》报告[24]	积极的市场成果将需要一系列在开源和闭源基础上推动前沿的模型。**呼吁政府将“开放创新和AI竞争”作为明确的政策目标**。
2023年3月	《一种支持创新的AI监管方法》[25]	**开源模型可以提高对基础模型的变革能力，但如果没有足够的保护措施，也可能造成伤害**。组织在开发和提供基础模型的方式上的差异，为AI的监管引入了广泛的复杂性。

英国对开源AI或软件做出规定的主要政策文件（本报告自制）

2024年2月，英国上议院的通信和数字委员会发布《大语言模型和生成式AI》报告，强调需要重新进行战略平衡，以应对安全和社会风险，同时把握新技术带来的机遇。报告提出多项

[21] Martin Casado, “We've submitted a letter to President Biden regarding the AI Executive Order and its potential for restricting open source AI”, 2023-11-04, https://twitter.com/martin_casado/status/1720517026538778657
[22] 谢旻希, “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.
[23] UK Government, “Introducing the AI Safety Institute”, 2024-01-17, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.
[24] UK Parliament, “House of Lords - Large language models and generative AI”, 2024-02-02, https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5402.htm.
[25] GOV.UK, “A pro-innovation approach to AI regulation”, 2023-08-23, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

建议，包括**防范监管俘获(regulatory capture)[26]、审查灾难性风险、赋予监管机构权力**等，并用单独一章探讨了开放模型与封闭模型之间的竞争动态，以及政府应该如何在这两者之间采取立场，并处理潜在的监管捕获问题，力求在开放与封闭模型之间找到合理立场。

报告认为开放模型通常成本低廉、易于获取，促进了技术的广泛尝试和社区驱动的改进，但在性能基准测试中可能不如封闭模型。封闭模型则因其先进技术被少数研究实验室控制，使得其他企业难以进行技术检查和实验。英国政府通过监管和政策引导，试图在这两种模型之间寻求平衡，推动各自在前沿领域的发展。报告中还提到，积极的市场结果需要基于开源和闭源相结合的平衡发展，并关注大型开发者可能利用先发优势造成的市场权力固化问题，**并呼吁政府将“开放创新和AI竞争”作为明确的政策目标**。

此外，英国科学、创新与技术部(DSIT)在2023年发布的《一种支持创新的AI监管方法》政策文件中，其中谈到一些组织对其基础模型的开发和分发进行严格控制，其他组织则采取开源方式开发和分发技术。**开源模型可以提高人们对基础模型的变革能力，但如果没有足够的保护措施，也可能造成伤害。**组织在开发和提供基础模型的方式上的差异，为AI的监管引入了广泛的复杂性。

英国是开源技术的领导者之一，早在2004年就首次发布开源产业政策，并于2009年2月进行了更新，开源非营利组织OpenUK于2021年2月在欧盟开源政策峰会发布了其三阶段报告，报告指出开源技术为英国贡献了高达430亿英镑的经济增长，其国内预计有12.6万名贡献者参与了创建、开发和维护开源的工作；这一数字将近欧盟26万名开源开发者中的一半。英国希望在政府内部鼓励创新，鼓励开源思维，在外部帮助发展一个充满活力的市场。

综合来看，英国的AI政策旨在通过支持开放创新与市场竞争，同时确保足够的监管措施，来平衡技术创新与风险管理的关系。这种策略不仅关注如何利用AI带来的机遇，也严肃对待由此可能产生的社会和安全挑战。英国试图在AI监管方面开辟自己的道路，借鉴但不照搬美国、欧盟和中国的做法，从而保持战略灵活性。

1.4 法国将开源AI作为其“创新优先”发展AI的核心战略之一

法国政府高度重视AI的发展，并将其视为国家竞争力的关键。在这一背景下，法国将开源AI作为技术政策的核心战略之一，积极推广和支持开源AI的发展。**法国总统马克龙在多个场合强调了创新和发展的重要性，并主张在治理之前优先考虑创新**。2023年6月，马克龙在欧洲科技峰会Viva Tech上强调“我们相信开源”[27]。法国的这一立场与其在欧盟中的作用形成了一定的不协调，因为欧盟更倾向于采取“监管优先”的方法。然而，法国也意识到了在推动AI创新

[26] 这一术语在日常使用中，指的是监管有利于受监管行业或特殊利益集团而不是公共利益的现象。
Wendy Li, “Regulatory capture's third face of power”, 2023-02-07, https://doi.org/10.1093/ser/mwad002.
[27] Politico, “France bets big on open-source AI”, 2023-08-04, https://www.politico.eu/article/open-source-artificial-intelligence-france-bets-big/.

发展的同时，需要平衡与欧盟政策的关系，因此法国国家信息和自由委员会(CNIL)作为法国的行政机构，在执行欧盟监管政策的同时，也在积极推动符合法国利益的AI创新政策[28]。

时间	法规或政策文件	对开源AI或软件的规定
2016年	《数字共和法案》	增强数字领域的透明度、开放性和私人数据保护。法案中明确要求公共部门和某些私人部门开放其数据，鼓励数据的自由流通和使用，同时强调保护个人隐私和数据安全
2023年5月	CNIL《AI行动计划》	CNIL不仅监管数据保护法律的遵守，也发布了关于AI和数据使用的指导原则。这些指导原则旨在确保AI技术开发过程中的数据使用遵守法律规定，同时保护用户的隐私权。
2018年3月	国家AI战略《AI造福人类》	这一战略计划通过公共投资高达15亿欧元，用以推动AI研究、创新和商业化。政策重点鼓励开源AI平台和工具的发展，旨在建立一个开放和协作的AI生态系统。

法国对开源AI或软件做出规定的主要法规或政策文件（本报告自制）

法国政府对于开源AI的监管倾向于制定灵活的政策，以适应技术快速发展的需求，保证开源项目的活力不被过度监管抑制。政策明确鼓励创新和数据的开放使用，但同时确保了数据安全和用户隐私的保护。这种平衡的监管框架为Mistral大模型等开源AI项目提供了发展的土壤。

法国的开源AI推广策略也体现在对人才的重视上。法国拥有强大的数学和信息科学基础，这为其在人工智能领域的研究提供了坚实的基础[29]。法国政府通过PIA(Investments for the Future Program)等项目，支持高等教育和培训，以及应用型基础研究及其经济价值的实现。法国政府通过投资研发和建立合作网络，鼓励公私部门的合作，推动了开源AI的创新和应用。其中，图灵奖得主、Meta首席AI科学家杨立昆(Yann LeCun)的参与，使得法国在全球AI研究领域的地位愈发显著。

在产业层面，法国也涌现出了一批在开源AI领域具有国际影响力的公司。例如，Hugging Face的联合创始人都是法国人，该公司已成为全球开源AI社区的重要参与者。法国AI初创公司Mistral AI开发的旗舰模型对标GPT-4，展示了法国在大模型领域的强大实力[30]。Mistral AI的成

[28] 曲子寰，“法国CNIL：法国和欧盟人工智能政策的平衡者”，2024-01-18, https://zhuanlan.zhihu.com/p/678550342.
[29] 安全内参，“法国人工智能发展现状、重要举措及启示”，2023-09-25, https://www.secrss.com/articles/59182.
[30] 智东西，“法国版OpenAI杀疯了，1760亿参数MoE登开源榜首，3张A100显卡可跑”，2024-04-11, https://www.thepaper.cn/newsDetail_forward_26506922.

功，部分归功于其能够巧妙地将AI技术与政治结合起来，这也是法国政府支持开源AI发展的一个例证[31]。此外，法国还有欧洲首个致力于AI开放科学研究的独立实验室Kyutai。

总体而言，法国政府通过一系列政策和措施，积极推广开源AI的发展，并在国际舞台上展现了其对AI技术的重视和支持。这些努力不仅有助于法国在全球AI竞争中保持领先地位，也为全球开源AI社区的发展做出了贡献。

1.5 中国人工智能法的两份专家建议稿对开源问题做不同处理

中国政府出台了一系列政策来促进开源软件的发展，鼓励企业和研究机构参与开源社区，推动开源软件的创新和应用，这些政策也适用于开源AI项目。但暂未发布针对基础模型开源的相关政策法规，相关立法还在制定中。

2023年10月，中国政府在第三届“一带一路”国际合作高峰论坛上发布了《全球人工智能治理倡议》，围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理中国方案。在国际治理中，中国政府站在全球南方的角度**呼吁开源人工智能技术**，并倡导需要开展面向发展中国家的国际合作与援助，不断弥合智能鸿沟和治理能力差距。

时间	**主要倡议或建议稿**	**对开源AI的规定或立场**
2024年4月	《人工智能示范法2.0版（专家建议稿）》[32] (2023年8月1.0版，9月1.1版[33])	**重视人工智能开源发展，**提出促进开源社区建设、制定专门合规指引、明确责任减免规则等支持措施。
2024年3月	《中华人民共和国人工智能法（学者建议稿）》[34]	**推进开源生态建设，建立开源治理体系，免费开源的人工智能/基础模型可豁免本法。**
2023年10月	《全球人工智能治理倡议》[35]	鼓励全球共同推动人工智能健康发展，共享人工智能知识成果，**开源人工智能技术**。

中国对开源AI或软件做出规定的主要倡议或建议稿（本报告自制）

2023年6月，国务院办公厅印发《国务院2023年度立法工作计划》。其中显示，《人工智能示范法》草案等预备提请全国人大常委会审议。目前，已有两组专家对人工智能法提出了建

[31] 机器之能，“拿下微软合作、旗舰模型对标GPT-4，认识一下「欧洲版 OpenAI」”，2024-03-01, https://www.jiqizhixin.com/articles/2024-03-01
[32] 冯恋阁，“《人工智能示范法2.0（专家建议稿）》重磅发布 重视AI开源发展、构建知识产权创新规则”，2024-04-16, https://m.21jingji.com/article/20240416/herald/4df710ffed0ffe037cdf6c54aa369961.html.
[33] 冯恋阁，“《人工智能示范法1.1（专家建议稿）》重磅发布”，2023-09-07, https://m.21jingji.com/article/20230907/herald/982ae3bb7b82597b4dc1f990ded64ad2.html.
[34] 数字法治，“重磅首发|《中华人民共和国人工智能法（学者建议稿）》”，2024-03-16, https://mp.weixin.qq.com/s/2i9zAXJ5dJKlKNMf4ppUDw.
[35] 中央网络安全和信息化办公室，“全球人工智能治理倡议”，2023-10-18, https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.

议稿，均对开源问题做出了探讨，**都明确了重视开源人工智能发展的积极立场。**具体而言，《人工智能示范法2.0（专家建议稿）》相较于《中华人民共和国人工智能法（学者建议稿）》在针对开源AI的制定专门合规指引、明确责任减免规则、提供税收抵免优惠等具体支持措施方面提供了更多的细则。

2023年8月，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》课题组发布**《人工智能法示范法1.0(专家建议稿)》**，9月又发布1.1版，表示“支持建设、运营开源开发平台、开源社区和开源项目等，推进开源软件项目合规应用”。2024年4月，**《人工智能示范法2.0（专家建议稿）》**发布，2.0在此前版本的基础上不断更新，将基于负面清单实施的人工智能许可管理制度与负面清单外人工智能活动的备案制度明确区分，避免过重合规负担影响人工智能产业的经营预期；**重视人工智能开源发展，提出促进开源社区建设、制定专门合规指引、明确责任减免规则等支持措施。**

《人工智能示范法2.0（专家建议稿）》对于开源AI比较有特色的条款包括：1）第二十二条（税收抵免优惠）人工智能研发者、提供者研发、购置用于安全治理等专用设备的投资额，可以按不低于30%的比例实行税额抵免。国务院针**对开源人工智能研发规定专门的税收优惠办法。**2）第五十九条（创新监管）国家人工智能主管机关针对开源人工智能研发者制订专门的合规指引，**推动开源人工智能创新发展。**3）第七十一条（**开源人工智能的法律责任减免**）存在条件限制，以免费且开源的方式提供人工智能研发所需的部分代码模块，同时以清晰的方式公开说明其功能及安全风险的，不承担法律责任。免费且开源提供人工智能的个人、组织能够证明已经建立符合国家标准的人工智能合规治理体系，并采取相应安全治理措施的，可以减轻或免于承担法律责任。

此外，由中国政法大学数据法治研究院牵头的专家组，于2024年3月发布了《中华人民共和国人工智能法（学者建议稿）》，其中开源AI的相关政策相对简要和宽松。提出国家应**推进开源生态建设**，支持相关主体建设或者运营开源平台开源社区、开源项目，鼓励企业开放软件源代码、硬件设计、应用服务，培育共享协作的开源创新生态。同时**建立开源治理体系**，鼓励通过协议规范开源产品许可、知识产权保护与责任分配机制，推进开源生态行业规范建设。人工智能开发者、提供者应当对开源框架、基础软硬件和部署环境的漏洞和安全风险进行定期检查和监测，并实时监测可能的攻击。明确表示开源的基础模型不需要对衍生利用承担连带责任、**基础模型属于开源模型的应豁免相应的法律责任，免费开源的人工智能**不适用于本法。

除此之外，国内学者专家也对于AI开源发展与法律规制举办了若干研讨[36,37]。以发展与安全的平衡为例，部分专家认为，尽可能进行精准规制、敏捷治理，避免一刀切式的安全冗余设

[36] SPPM法与公管交叉研究，“人工智能立法之开源发展与法律规制会议顺利召开”, 2024-01-27 https://mp.weixin.qq.com/s/H4s7V-Jc16PwxMoBygDROg.
[37] 中国法学创新网，“AI善治论坛在京召开发布《人工智能法（学者建议稿）》”，2024-03-19, http://www.fxcxw.org.cn/dyna/content.php?id=26910.

置，并在守住系统性风险底线的同时支持人工智能的自由发展；逐步建立算法相关新型权益、广泛的利益共享机制、人工智能赋能机制，保护个人选择空间等。在考虑我国大型开源模型的现实状况与未来发展需求时，我们应当综合运用技术保障措施、负责任的研发行为规范以及全面的安全评估等手段，实现自律与他律的有效结合，确保开源大模型生态在安全与健康的基础上不断推动创新[38]，我们期待就这一议题展开更深入的探讨。

1.6 其他全球南方国家鼓励AI风险与收益研究，以开放科学应对全球发展

2024年3月，联合国大会未经表决一致通过了一项呼吁《抓住安全、可靠和值得信赖的AI系统带来的机遇，促进可持续发展》的决议[39]。其中，**“鼓励开展研究和国际合作，以了解、平衡和解决与AI系统在弥合数字鸿沟和实现所有17项可持续发展目标方面所发挥作用有关的潜在利益和风险，包括扩大开源AI系统等数字解决方案的作用”**。

2023年9月，《哈瓦那G77+中国峰会的最终宣言》[40]强调了对**开放科学合作的需要以及推广全球科技发展的开源模式的重要性。**它呼吁促进一个包容和公平的技术进步环境，避免垄断和其他障碍，确保尤其是发展中国家能公平获取信息和通信技术。宣言突出了开放科学在应对全球挑战和通过共享知识及技术增强各国发展能力中的作用。

中东地区，阿联酋积极采用开源治理模式，对大模型发放开源许可证(Open-source License)。此外，阿联酋政府还向Falcon基金会承诺3亿美元的资金，支持该非营利性机构的大模型开源开发工作，阿联酋有兴趣帮助其他国家获取这些开源模型，以便他们可以开放相应的AI应用程序。有观点认为，阿联酋致力于开源AI的承诺在外交上取得了成功，赢得了来自全球南方国家的朋友，否则这些国家将被排除在昂贵的AI开发之外。其中一部分动机是为阿联酋找到一个领域，并增强国家竞争力[41]。

此外，巴西促进能够验证数据集和机器学习模型中歧视性趋势的开源代码的传播[42]。非洲部分国家，如乌干达、肯尼亚、卢旺达、坦桑尼亚，制定了AI治理数据保护相关立法，但目前未涉及针对开源治理的重点讨论[43]。

---

[38] 傅宏宇，“《中华人民共和国人工智能法（学者建议稿）》：产业期待中的中国方案”，2024-03-18, https://www.sohu.com/a/765103343_384789.
[39] 联合国，“联合国大会通过里程碑式决议，呼吁让人工智能给人类带来'惠益'”，2024-03-21, https://news.un.org/zh/story/2024/03/1127556.
[40] Cuban Ministry for Foreign Affairs，“Final Declaration of Havana's G77+China Summit”，2023-09-18 https://orinocotribune.com/final-declaration-of-havanas-g77china-summit/.
[41] Billy Perrigo，“The UAE Is on a Mission to Become an AI Power”，2024-03-22, https://time.com/6958369/artificial-intelligence-united-arab-emirates/.
[42] Brazil Government，“Summary of the Brazilian Artificial Intelligence Strategy”，2022-02-24, https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-summary_brazilian_4-979_2021.pdf.
[43] Thomson Reuters Foundation，“AI Governance for Africa Toolkit”，2023-11-22, https://www.trust.org/dA/97390870db/pdfReport/AI%20Governance%20for%20Africa%20Toolkit%20-%20Part%201%20and%202.pdf.

肯尼亚前任外长Raychelle Omamo表示，南方国家更关注的是发展，以及谁能更好地在基础设施建设、人员培训和公共健康等领域支持和帮助自己国家，而不会关心是不是给某个国家贴上标签。**全球南部的许多国家仍将数字发展置于地缘政治协调之上**[44]。

1.7 小结

在全球化的背景下，各国对于开源AI的监管政策呈现出了不同的监管取向和政策设计。这些不同的政策取向不仅反映了各国对于AI技术发展的重视程度，也体现了对于开源AI风险与收益的不同理解和评估。

一方面，监管机构试图通过制定相关法规来确保AI技术的安全性和可控性，防止潜在的滥用风险；另一方面，政策制定者也认识到开源AI在促进技术创新、推动行业发展以及普及教育资源等方面的积极作用。这种平衡发展与安全的双重目标，导致了在开源AI政策上的不同选择和实施路径。

我们认为制定开源AI的政策需要有更多关于风险、收益及潜在影响的严谨证据支持，同时安全治理需要国际合作，因此需要充分理解各个国家的监管政策和国情考虑。

与此同时，产业界和学术界对于开源AI的风险和收益也展开了激烈的讨论，展现出不同的立场。接下来，我们将深入探讨这些不同立场及其异同点，并分析这些观点对于开源AI治理政策制定的启示和影响。

[44] 肖茜，“在慕安会感受欧美AI治理的协调与分歧”，2024-02-22, https://ciss.tsinghua.edu.cn/info/subemail_wzjx/6931.

2 审慎开放vs鼓励开放，前沿AI开源的主要争论

到目前为止，AI领域的开源一直是有益的。但未来我们可能会到达一个地步，不开源在减少风险的层面会对社会更有益，尽管这会在创新速度等方面有所损失。显然，这有利有弊。

——约书亚·本吉奥 (Joshua Bengio)[45]

林纳斯定律：只要有足够多的眼睛，所有的错误都是显而易见的。

——埃里克·雷蒙德 (Eric Raymond)[46]

2023年9月底，抗议者聚集在Meta旧金山办公室外，抗议其公开发布AI模型的政策，声称Llama系列模型发布代表了潜在不安全技术的“不可逆转的扩散”[47]。但也有人表示，几乎没有证据表明开源模型造成了任何具体损害，AI研发的开源是“确保对技术信任的唯一途径”。

2.1 争论主要在于前沿AI的滥用和失控风险

2.1.1 大多数AI应开放并无争议

图灵奖得主Yoshua Bengio在讨论开源AI潜在的滥用风险时认为[48]：模型开源一般由研发机构自身决定。**目前大多数AI是有益/无害的，应该共享。**但未来更强的AI可能因滥用而产生全社会的影响，这些决定应经过利益相关方的审议。因为模型的开放共享决策是**不可逆转**的，一旦落入恶意行为者手中，国家和社会就无法控制。

当Elon Musk等指责Sam Altman“呼吁监管”只是为了保护OpenAI的领导地位时，Sam Altman回应称[49]，“应该对那些超过某一高度能力阈值的大型公司和封闭模型进行更多监管，而**对小型初创公司和开源模型的监管应该较少**”。

OpenAI、DeepMind、微软等机构的研究人员撰写的研究报告《前沿AI监管：管理公共安全的新兴风险》[50]指出：**开源AI模型可能是一项重要的公共产品**。然而，前沿AI模型可能需要比**相对于它们更小、更专用或能力较弱的同类模型受到更多限制**。

---

[45] Azeem Azhar, Yoshua Bengio, “Yoshua Bengio: Towards AI's humanistic future”, 2024-02-14, https://www.exponentialview.co/p/yoshua-bengio-towards-ais-humanistic.
[46] Eric Raymond, “大教堂与集市”, 2014-05, https://book.douban.com/subject/25881855/.
[47] Edd Gent, “Protesters Decry Meta's 'Irreversible Proliferation' of AI”, 2023-10-06, https://spectrum.ieee.org/meta-ai.
[48] Yoshua Bengio, “以民主治理管理人工智能风险”, 2023-12-09, https://weibo.com/tv/show/1042163:4977133622067294.
[49] ETtech, “Regulation? AI, says OpenAI CEO Sam Altman”, 2023-06-09, https://economictimes.indiatimes.com/tech/technology/regulation-ai-says-openai-ceo-sam-altman/articleshow/100830314.cms.
[50] Markus Anderljung et al., “Frontier AI Regulation: Managing Emerging Risks to Public Safety”, 2023-11-07, https://arxiv.org/abs/2307.03718.

Anthropic在《第三方测试是AI政策的关键组成部分》[51]一文中认为：**当今绝大多数（甚至可能是全部）AI系统都可以安全地公开传播，并且在未来也可以安全地广泛传播**。然而，我们相信，未来前沿AI系统完全开放传播的文化与社会安全文化之间可能很难调和。

2.1.2 前沿AI可能因滥用和失控引发灾难性风险

前沿AI可能涉及的滥用和失控风险主要包括网络安全、化学/生物/辐射/核威胁(CBRN)、虚假信息的滥用风险，以及操纵欺骗、模型自主性导致的滥用和失控风险。

当前证据表明开放模型和封闭模型之间存在显著的性能差距[52]，但随着开源模型的能力日益接近GPT-4级别[53]，对于前沿AI模型是否应开源的争论预计将更加激烈。

风险	类别	描述
网络安全	滥用	利用AI进行网络攻击，从而破坏计算机系统的机密性、完整性、可用性相关的风险。
CBRN	滥用	利用AI辅助手段创建化学/生物/辐射/核威胁相关的风险。
虚假信息	滥用	利用AI生成传播有害或虚假信息给公共安全带来相关的风险。
操纵欺骗	滥用/失控	利用AI生成内容使人们改变其信念或据此采取行动相关的风险。
自主性	滥用/失控	更强的模型自主性使行为体能够适应环境变化并规避被关闭，从而可能被规模化滥用。自主性也是自我复制、自我改进、资源获取的前提。 由于目前的安全技术还不够完善，其行为可能会违背设计者或使用者的初衷（不对齐）。即使没人故意滥用，也可能成为灾难性风险的来源。

前沿AI可能涉及的滥用和失控风险（本报告自制）

注：

1）《布莱切利AI安全宣言》，重点关注网络安全、生物技术、虚假信息

2）《安全、可靠和可信AI行政命令》，重点关注网络安全、CBRN、欺骗

3）《北京AI安全国际共识》，重点关注自主复制或改进、权力寻求、协助武器制造、网络安全、欺骗

4）OpenAI的Prepareness Framework(Beta)，重点关注网络安全、CBRN、操纵（含欺骗）、自主性

5）Anthropic的Responsible Scaling Policy 1.0，重点关注网络安全、CBRN、自主性和复制

[51] Anthropic，“Third-party testing as a key ingredient of AI policy”, 2024-03-25, https://www.anthropic.com/news/third-party-testing.
[52] Stanford HAI，“2024 AI Index Report”, 2024-04-15, https://aiindex.stanford.edu/report/.
[53] Meta，“Introducing Meta Llama 3: The most capable openly available LLM to date”, 2024-04-18 , https://ai.meta.com/blog/meta-llama-3.

2.2 立场一：审慎开放，防范风险的开放门槛须标准更高

2.2.1 产业界

1）Anthropic专注于开发安全、可控的AI系统，前沿模型均未开源[54]

一方面，Anthropic认识到科学进步很大程度上依赖于研究的开放和透明文化，AI领域的许多革命性进展都是建立在开源研究和模型的基础上的。开源系统通过允许更多人测试技术并识别潜在弱点，有助于提高安全环境的稳健性。

另一方面，Anthropic 也表达了对于前沿AI系统完全开源可能带来的风险的担忧。他们认为，随着 AI 模型的能力日益增强，如果（“如果”是一个关键且尚未解决的问题）存在可能导致有害影响或灾难性事故的风险，**那么当前的开源文化可能需要调整，以确保 AI 系统的安全和社会责任。**

Anthropic提出，AI开发者在发布系统时需要提供强有力的安全保证，例如通过分类器检测和阻止滥用尝试，或通过合同义务限制微调系统的能力。如果有人想要公开发布模型权重，他们需要确保模型经过强化以防止滥用（例如，通过RLHF或RLAIF训练），并找到一种方法来制作该模型能够适应将其微调到可能导致这种滥用的数据集的尝试。还需要试验披露流程，类似于安全社区如何制定有关零日披露预先通知的规范。

Anthropic强调，尽管这些安全措施可能成本高昂，但为了预防AI系统可能导致的严重滥用或事故，这些努力是必要的。然而，他们也承认，对于AI系统的公开传播进行限制需要在AI系统或系统行为的不可接受滥用行为上达成广泛共识。

最后，Anthropic指出，作为一家主要开发封闭系统的公司，他们没有正当性来决定哪些行为在开源模型中应该或不应该被接受。因此，**他们呼吁需要合法的第三方来开发和应用被广泛认可的测试和评估方法**，以及定义AI系统滥用行为的标准，并对受控（例如通过API）或公开传播（例如通过开源权重）的模型进行这些测试，以生成关于AI领域安全特性的基本信息。如果不这样做，可能会面临严重滥用或AI事故的风险，这可能对人和社区造成重大伤害，并可能导致对AI行业不利的法规。

2）OpenAI开源策略经历了从相对开放到逐步收紧的过程

2015-2019年的早期阶段，OpenAI秉持开放、协作、造福人类的理念，主张通过开源推动AI的进步。他们发布了GPT、GPT-2等自然语言处理模型，以及Universe等平台的源代码，希望学术界和业界能广泛参与、共同进步。

但随着语言模型等AI系统变得越来越强大，OpenAI开始意识到技术滥用和误用可能带来的危害，如制造虚假信息、侵犯隐私等。为了降低风险，他们在GPT-2发布时采取了分阶段发

[54] Anthropic，“Third-party testing as a key ingredient of AI policy”，2024-03-25, https://www.anthropic.com/news/third-party-testing.

布[55]、有限度的开源策略。这一举措引发了一些争议，但也反映出OpenAI对于强大AI系统负责任、可控发展的重视。

2019年底，OpenAI宣布从非营利组织转为"封顶利润"的有限责任公司，以获取更多的商业资源和灵活性。这导致了商业利益在其决策中的比重上升。为了保护自身的竞争优势和创新成果，OpenAI开始减少对开源模型和训练代码，同时发布的研究成果也大幅减少，像GPT-3这样的大模型只开放API，希望既能惠及社会，又能在一定程度上控制技术的传播和应用，但OpenAI的品牌名称导致了一定程度的混乱[56]。

总的来说，OpenAI的开源策略是在开放共享、风险防控和商业利益之间不断权衡、动态调整的。这种平衡和博弈，也反映了当前整个AI领域在开源问题上的复杂性和两难性。

3）Google DeepMind开源了AlphaFold 2和Gemma，但不开源前沿模型Gemini

2024年2月，Google DeepMind的首席执行官Demis Hassabis在接受《纽约时报》专访时[57]，谈及如何看待通过开源使基础模型可用增加了被恶意使用其能力的风险这一批评。Demis Hassabis提到，开源和开放科学对于AGI技术的发展显然是是有益的。但他也指出开源可能带来的风险，特别是对于强大的AGI技术，因为其通用性，一旦开源就可能被恶意利用。

他解释说，Google DeepMind决定开源Gemma模型的原因是因为Gemma是轻量级的版本，相对于前沿模型Gemini来说，它的能力已经得到了很好的测试和理解，因此Google DeepMind认为这种规模的模型相关的风险并不大。

他还强调了安全性、鲁棒性和责任性的重要性，特别是在接近AGI时，必须更加谨慎地考虑这些系统可能被滥用的可能性，**开源的门槛必须更高。**他认为，对于开源的极端主义者，必须更多地考虑这些问题，因为这些系统变得越来越强大。

而此前，DeepMind还开源了AlphaFold 2，但也通过内部机构审查委员会，外部生物研究、生物安全、生物伦理专家咨询，以及精心设计的发布策略，旨在使AlphaFold 2的收益尽可能地广泛可用，同时保持对其局限和准确性的透明度[58]。

总的来说，Demis Hassabis和DeepMind支持开源，但同时也强调了在开源时需要考虑的安全性和潜在风险，特别是在AGI技术方面。他们倾向于在确保风险可控的情况下进行开源，以促进技术的健康发展和广泛应用。

[55] OpenAI, “GPT-2: 1.5B release”, 2019-11-05, https://openai.com/research/gpt-2-1-5b-release.
[56] David Harris, “Open-Source AI Is Uniquely Dangerous”, 2024-01-12, https://spectrum.ieee.org/open-source-ai-2666932122.
[57] Hard Fork, “Google DeepMind C.E.O. Demis Hassabis on the Path From Chatbots to A.G.I.”, 2024-02-23, https://www.nytimes.com/2024/02/23/podcasts/google-deepmind-demis-hassabis.html.
[58] Google Deepmind, “How our principles helped define AlphaFold's release”, 2022-09-14, https://deepmind.google/discover/blog/how-our-principles-helped-define-alphafolds-release.

4）AI合作伙伴关系（PAI）建议谨慎发布前沿模型

2023年10月，PAI发布《安全基础模型部署指南》[59]，为AI模型提供商提供了一个框架，并为负责任地开发部署基础模型提出了22条实践建议，旨在确保社会安全并适应不断发展的AI能力和用途。

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AI安全原则。

指南被设计为一份动态文档，可以随着新的AI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。

指南的主要特点：

- **监督和安全的扩展性。**为了适当地解决风险，模型部署指南的方针是根据每个AI模型的能力和可用性量身定制监督和安全实践。模型部署指南避免了过度简化，不仅仅将模型的大小或通用性等同于风险。
- **开放访问指导。**模型部署指南包括对开放访问模型的指导方针，为透明度和风险缓解策略提供了起点。这为当前和未来的开源模型提供商提供了指导。
- **广泛的适用性。**模型部署指南适用于从现有到前沿的基础模型全谱系。
- **谨慎推出前沿模型。模型部署指南建议对前沿模型最初进行分阶段发布和限制访问，直到展示出足够的安全保障措施。**
- **安全的整体视角。**模型部署指南建立了起点，以解决各种安全风险，包括与偏见、过度依赖AI系统、工人待遇和恶意行为者相关的潜在伤害。

2.2.2 学术界

1）GovAI探讨了开放高能力基础模型的潜在风险和替代方案

2023年9月，GovAI发布的《开放高能力基础模型》报告[60]，对追求开源目标的风险、收益和替代方案进行了评估。

作者们承认，开源有显著的优势，例如使外部监督成为可能、加速进步和去中心化的AI控制。但也存在明显的风险，例如允许恶意行为者在没有监管的情况下滥用AI，并可关闭模型本身设计的安全措施。

[59] PAI，“PAI's Guidance for Safe Foundation Model Deployment”, 2023-10-24, https://partnershiponai.org/wp-content/uploads/1923/10/PAI-Model-Deployment-Guidance.pdf.
[60] Elizabeth Seger，“Open-Sourcing Highly Capable Foundation Models”, 2023-09-29, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

虽然迄今为止，开源已经为大多数软件和AI开发流程提供了可观的净收益，但对于在不久的将来可能出现的一些高能力模型，开源的风险可能超过收益。主要考虑因素包括：

- **潜在风险：**高能力模型可能带来极端风险，比如用于生产生化武器或进行网络攻击。当前AI能力还未能超过最极端风险的临界能力阈值。然而，我们已经看到初步的危险能力出现，随着模型变得越来越强大，用户部署和调整这些模型所需的专业知识和计算资源越来越少，这一趋势可能会持续。
- **攻防失衡：**开源有助于解决一些风险，但可能加剧部分极端风险。对于传统软件，开源促进了防御一方。然而，对于越来越强大的基础模型，攻防平衡可能会因以下原因而偏向于攻击，因为恶意行为者可以更容易地发现和利用漏洞，开源也不利于修复漏洞以及改进措施在下游实施。
- **替代方案：**有其他更少风险的方法可以追求开源目标，尽管这些策略都有其自身的缺点。例如，针对特定研究、审计和下游开发需求的结构化模型访问选项，以及积极努力组织安全合作，鼓励和实现更广泛地参与AI开发、评估和治理过程。

鉴于这些潜在风险、攻防失衡和替代方案，为帮助建立负责任开源的最佳实践，在安全的前提下保留开源的优势。作者为开发者、标准制定机构和政策制定者提出了以下建议：

- **开发者和政府应该认识到，一些功能强大的模型开源的风险太大，至少在初期是这样。**随着社会对风险适应和安全机制改进，之后可以开源。
- **开源高性能基础模型的决策应基于严格的风险评估。**除了评估模型的危险能力和直接滥用之外，还必须考虑模型微调或修改可引发的滥用。
- **开发者应考虑开源的替代方案，在获得技术和社会效益的同时，又没有太大的风险。**可能的替代方案包括渐进/分阶段发布、为研究和审核人员提供结构化模型访问，对AI开发和治理决策的广泛监督等。
- **开发者、标准制定机构和开源社区应多方协作，定义模型组件发布的细粒度标准。**标准应基于对发布不同组件特定组合所带来的风险的理解。
- **政府应对开源AI模型进行监督，并在风险足够高时实施安全措施。**AI开发者或许不会自愿采用风险评估和模型共享标准，政府需要通过法规来执行此类措施，并建立有效执行此类监督机制的能力。

2）Mila等机构的学者提出可下载模型微调的日益便捷可能会增加风险

2023年12月，Mila、英国AI安全研究所、剑桥大学等机构的研究人联合发布论文《可下载基础模型日益便捷的微调所带来的风险》[61]（这里的“可下载访问”，是指公开发布预训练

[61] Alan Chan et al. “Hazards from Increasingly Accessible Fine-Tuning of Downloadable Foundation Models”, 2023-12-22, https://arxiv.org/abs/2312.14751.

基础模型的权重，3.2节将进一步讨论）。论文讨论了以下几个主要方面：

- **微调的可访问性提高：**研究如何通过减少微调的计算成本和改进成本分摊机制，提高了微调的可访问性。既包括了模型权重的可获取性，也包括了通过研究进展（如改进算法、使用合成数据、参数高效的微调方法等）来降低微调的技术门槛和成本，使得更广泛的用户群体能够便捷地使用这些模型。
- **潜在危害的增加：**论文认为，微调方法的可访问性提高可能会通过促进恶意使用和使得对具有潜在危险能力的模型进行监督变得更加困难，从而增加危害。
- **潜在缓解措施和收益：**论文讨论了可能的缓解策略，例如使预训练模型更难针对特定任务进行微调[62]，以及“遗忘学习”(unlearning)技术来移除可能被恶意行为者利用的记忆信息。同时，论文也指出了微调的可访问性提高可能带来的潜在好处，如促进学术研究、适应新的用例、避免权力不平衡等。
- **不确定性和未来工作：**论文强调了关于危害的大量剩余不确定性，并建议未来的工作应该集中在研究危害可能出现的情况，并开发缓解措施，如使某些任务或能力的微调变得更加困难。

论文强调了在开放模型和提高微调可访问性的同时，需要平衡潜在的收益和风险，并并投入更多的研究努力来理解和减轻这些风险。论文的结论强调了在提高微调可访问性的同时，需要平衡潜在的危害和好处，并投入更多的研究努力来理解和减轻这些风险。这表明作者倾向于在采取适当的预防措施和监管框架的情况下，审慎地推进基础模型的开放。

2.3 立场二：鼓励开放，边际风险的严谨证据仍相当有限

2.3.1 产业界

1）领先的AI开源机构：仅列举开源了其最强模型的机构

Meta(Facebook)长期以来一直积极支持开源社区，贡献了PyTorch等多个重要的项目。Meta在开始每个AI项目时都希望将项目的每个组件开源[63]，但有时却因为研究想法不成功或其他考量无法开源。在大模型方面，2022年5月Meta发布OPT-175B大语言模型，2023年2月发布Llama模型，都为非商业研究用途免费开放，并在2023年7月和2024年4月分别开源了可直接商用的大模型Llama 2和Llama 3。Meta首席科学家杨立昆认为[64]，开源的大模型能够吸引更多人参与，从而加速技术进步，这样的系统更安全，性能更佳。AI必须是开源的，因为当大模型成为通信结构的重要组成部分时，我们都需要有一个通用的基础设施。

[62] Peter Henderson，“Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models”，2022-11-27, https://arxiv.org/abs/2211.14946.
[63] Stanford HAI，“How to Promote Responsible Open Foundation Models”, 2023-10-3, https://hai.stanford.edu/news/how-promote-responsible-open-foundation-models.
[64] Steven Levy，“How not to be stupid about AI, with Yann LeCun”, 2023-12-22, https://www.wired.com/story/artificial-intelligence-meta-yann-lecun-interview/.

BigScience是一个由学术界和工业界合作发起的国际开源研究项目，旨在训练和开源有益于科学发展的大规模AI模型，在推动大模型开源和多样性方面走在前列。2022年5月他们发布BLOOM-176B等大规模多语言开源对话模型，希望让不同语言和文化背景的用户都能使用先进AI技术。该项目由Hugging Face联合创始人Thomas Wolf构想，他希望与大公司竞争，提出不仅要训练出立于世界上最大的多语言模型之林的模型，还要让所有人都可以公开访问训练结果，圆了大多数人的梦想[65]。

EleutherAI是一个致力于开源的非营利实验室，成员主要是AI研究人员、工程师等。他们的目标是复现和开源类似GPT-3的大规模语言模型，让更多人能使用和研究。EleutherAI发布了GPT-Neo系列、GPT-J、Pythia系列等免费商用授权的开源模型，这些模型有不同规模，全部采用公开数据进行训练，旨在帮助研究人员理解大模型训练的不同阶段。Connor Leahy是EleutherAI和Conjecture的联合创始人，他对AI模型开源持复杂观点。EleutherAI在推动开源AI研究方面发挥了重要作用，但Connor Leahy也批评Meta是“最不负责任的AI参与者”，“我们应该令核武器的设计透明化吗？”[66]强调了在没有严格安全防护的情况下公开强大技术的风险。

Stability AI是一家致力于开发先进的图像生成软件开源AI的公司。其代表作Stable Diffusion软件在业界享有盛誉，最强大的开源图像生成模型之一，被广泛应用于图像识别、生成和编辑等领域，推动了AI艺术创作的发展和普及。

阿联酋人工智能和数据科学研究所(TII)是阿联酋国家AI战略的重要组成部分，致力于推动AI技术的研究、开发和应用。TII在开源大模型领域取得了重大进展，推出Falcon 180B模型，为全球研究人员和开发者提供了宝贵的资源。TII将 Falcon 180B 开源，旨在促进全球AI研究的合作和发展，成为阿联酋和中东地区积极发展AI技术的体现。

2）IBM和Meta联合发起AI联盟，合作推动开放、安全、负责任的AI发展

2023年12月，IBM和Meta联合发起AI联盟(AI Alliance)[67]。这是一个国际性的技术开发者、研究人员和采用者社区。AI联盟由超过50个创始成员和协作者组成，包括AMD、CERN、谷歌、英特尔、Hugging Face等知名机构。该联盟旨在通过开放创新和科学，聚焦于加速负责任的AI创新，同时确保科学严谨、安全、多样性和经济竞争力。联盟将开展多个项目，包括开发和部署标准、工具，推进开放基础模型的生态系统，促进AI硬件加速器生态系统的繁荣，以及支持全球AI技能培训和探索性研究。

[65] Stas Bekman, “千亿参数开源大模型 BLOOM 背后的技术”, 2022-07-14, https://huggingface.co/blog/zh/bloom-megatron-deepspeed.
[66] 36氪, “杨立昆希望用开源战胜OpenAI？”, 2023-12-04, https://36kr.com/p/2546244653229831.
[67] AI Alliance, “AI Alliance Announces 25+ New Members, Launches AI Safety Tooling and AI Policy Working Groups to Enable Open, Safe, and Responsible AI for All”, 2024-02-08, https://thealliance.ai/news.

3）负责任创新实验室发布聚焦于负责任AI的自愿承诺协议

2023年11月，总部位于旧金山的负责任创新实验室(Responsible Innovation Labs, RIL)[68]发布了针对初创公司及其投资者的首个行业驱动的负责任AI承诺。这些承诺包括五个关键行动步骤：确保组织对负责任AI的认同、通过透明度建立信任、预测AI的风险和收益、审计和测试以确保产品安全、进行定期和持续的改进。这一自愿性协议旨在提供实用指导，由多部门联盟支持，包括风险投资家、生成AI初创公司、学术界和美国商务部。协议聚焦于负责任AI的自愿承诺，为早期阶段的初创公司提供了针对其特定需求的资源。RIL的目标是为日益增多的希望在扩大业务和创新时整合负责任AI实践的初创公司和投资者提供指导。该协议的初始签署者包括35家领先的风险投资基金。

2.3.2 学术界

1）Mozilla发起《关于AI安全和开放的联合声明》主张以公开访问和审查增强安全性[69]

声明指出，当前正处于AI治理的关键时期。为了减轻当前和未来AI系统可能带来的伤害，需要拥抱开放性、透明度和广泛的访问权限。这应该是一个全球性的优先事项。签署者包括来自不同领域的专家，如Meta的杨立昆、斯坦福大学的吴恩达、哥伦比亚大学的Camille François、Mozilla的Mark Surman、加州大学伯克利分校的Deborah Raji、诺贝尔和平奖得主Maria Ressa Rappler、EleutherAI的Stella Biderman等1800多位签署者。

声明反对仅通过严格、专有控制AI模型来防止对社会的大规模危害的观点，主张通过公开访问和审查来增强安全性。声明强调，过于匆忙地推行限制性法规可能损害竞争和创新。

从开源到开放科学，声明呼吁采取多样化的方法：

1. 通过促进独立研究、合作和知识共享，加速理解AI能力、风险和伤害。
2. 通过帮助监管者采用工具来监测大规模AI系统，增加公众审查和问责。
3. 为专注于创造负责任AI的新参与者降低进入门槛。

当谈到AI安全和安保时，强调“开放性是一种解药，而不是毒药”。

2）普林斯顿大学等机构的学者探讨开放基础模型的社会影响，力求促进讨论的精确性[70]

作者认为对开放影响的分歧是由于对其社会影响的说法缺乏精确性造成的。在分析了开放基础模型的收益的基础上，他们提出了一个风险评估框架，用于评估开放基础模型与封闭模型或互联网上的网络搜索等现有技术相比的边际风险。

---

[68] Responsible Innovation Labs, “Introducing the Responsible AI Commitments”, 2024-04-23(引用日期), https://www.rilabs.org/responsible-ai.
[69] Camille François et al., “Joint Statement on AI Safety and Openness”, 2023-10-31, https://open.mozilla.org/letter/.
[70] Sayash Kapoor et al., “On the Societal Impact of Open Foundation Models”, 2024-02-27, https://crfm.stanford.edu/open-fms/.

开放基础模型的收益。开放性的关键属性包括：更广泛的访问（通过允许更广泛的人访问模型权重）、更大的可定制性（通过允许用户按需调整模型）、本地适应和推理（用户可自行选择硬件）以及无法撤销访问权限（基础模型开发人员一旦发布就无法轻易撤销访问权限）。

这些特性带来许多收益：

- **分配谁定义可接受的模型行为：**更广泛地访问模型及其更大的可定制性扩展了谁能够指定可接受模型行为的边界，而不是仅由基础模型开发人员拥有决策权。
- **加大创新力度。**更广泛的访问、更大的可定制性和本地推理扩展了基础模型用于开发应用程序的方式。例如，具有严格隐私控制要求的应用程序可使用本地部署的模型。
- **促进科学研究。**许多类型的基础模型研究和使用都需要访问模型权重。过去两年中，已经看到了开放模型带来的速度提高和安全挑战的例子。同时，访问数据、文档和模型检查点等资产对于其他研究来说是必要的，因此单独提供模型权重通常并不足够。
- **实现透明度。**持续且广泛的公开源代码支持的同行评议可以通过识别和消除核心开发团队可能无法发现的缺陷来提高软件的安全性[71]。同时，没有经验证据表明开源软件比闭源软件更容易受到攻击或更不安全[72]。对模型权重的广泛访问可实现某种形式的透明度，例如有关模型架构的详细信息。然而，与研究类似，透明度也需要模型权重以外的资产，特别是公开文档，当前即使模型权重公开发布，也往往缺乏公开文档。
- **减轻单一文化和市场集中度。**在不同的应用中使用相同的基础模型会导致单一文化。当模型出现问题时，会影响所有下游应用。更大的可定制性可减轻单一文化的危害，因为基础模型的下游开发者可对其进行微调以改变其行为。同样，更广泛的模型访问渠道可降低开发不同类型基础模型的进入壁垒，从而有助于降低下游的市场集中度。

Misuse risk	Paper	Threat identification	Existing risk (absent open FMs)	Existing defenses (absent open FMs)	Evidence of marginal risk	Ease of defense	Uncertainty/assumptions
Spear-phishing scams	Hazell (2023)	●	◑	○	○	◑	○
Cybersecurity risk	Seger et al. (2023)	◑	○	◑	○	◑	○
Disinformation	Musser (2023)	●	◑	○	○	◑	●
Biosecurity risk	Gopal et al. (2023)	●	○	◑	○	○	○
Voice-cloning scams	Ovadya et al. (2019)	●	◑	◑	◑	◑	●
Non-consensual intimate imagery	Lakatos (2023)	●	◑	○	◑	◑	○
Child sexual abuse material	Thiel et al. (2023)	●	●	●	●	●	●

使用其框架对开放基础模型的风险进行评分研究。

●表明该步骤已明确完成；◑表示部分完成；○表示不存在该步骤（完整论文清单）[73]

[71] Rishi Bommasani et al.，“Considerations for Governing Open Foundation Models”，2023-12-13, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.
[72] Guido Schryen，“Is Open Source Security a Myth?”，2024-04-12, https://dl.acm.org/doi/pdf/10.1145/1941487.1941516.
[73] Sayash Kapoor et al.，“On the Societal Impact of Open Foundation Models”，2024-02-27, https://crfm.stanford.edu/open-fms/.

框架步骤	步骤描述	以网络安全（自动漏洞检测）为例
威胁识别	指定威胁是什么以及来自谁? 所有滥用分析都应系统地识别和描述正在分析的潜在威胁为了提出明确的假设。	漏洞检测工具可用于自动执行发现软件漏洞的过程。威胁行为者包括个人黑客、小团体或国家支持的攻击者。
现有风险(不考虑开放基础模型)	这种威胁的现有风险是什么? 在许多情况下，公开发布模型的风险已经存在于现实世界（尽管严重程度可能有所不同）。	攻击者受益于漏洞检测中自然最坏情况的不对称性：攻击者只需利用单个有效漏洞即可成功，而防御者必须防御所有漏洞才能成功。现有风险很大程度上受到攻击者资源的影响：老练的攻击者经常在攻击设计中使用自动漏洞检测工具。模糊测试工具长期以来一直被用来查找软件中的漏洞，Metasploit等工具也是如此，Metasploit是一个免费的渗透测试框架，可以帮助自动漏洞检测。MITRE的AI系统对抗威胁格局是对抗性机器学习的网络安全威胁矩阵，包括许多利用封闭基础模型和其他类型的机器学习模型来检测漏洞的技术。
现有防御(不考虑开放基础模型)	这种威胁的现有防御是什么? 开放基础模型的许多所谓风险都有现有的防御措施。	网络安全防御通常采用纵深防御策略，其中防御是分层的，以确保基于一层中未解决的漏洞的利用不会影响其他防御层。在漏洞检测设置中，防御者可以抢先使用漏洞检测工具来检测和修补安全威胁，这同样取决于他们对资源的访问。漏洞赏金等激励策略可以通过激励漏洞发现者（黑客、安全研究人员、公司）报告漏洞，在一定程度上使攻防平衡向有利于防御的方向倾斜。
边际风险(marginal risk)的证据	这种威胁的风险增量是什么? 一旦威胁途径、现有风险水平以及现有防御措施的范围明确后，理解开放发布模型的边际风险就非常重要。不仅要与现有技术（如互联网）相比较，还要与闭源的基础模型的发布相比较。	我们不知道现有证据表明恶意行为者已成功使用开放基础模型来自动检测漏洞。存在暗网工具广告，声称可以促进自动漏洞检测，但尚不清楚这些产品是否依赖于开放基础模型。在考虑相对于封闭基础的边际风险时，虽然可以更好地监测封闭基础模型的滥用情况，但尚不清楚此类用途是否会得到可靠性识别。也就是说，使用封闭基础模型进行漏洞检测并不一定是滥用，这引入了区分用于自动漏洞检测的封闭基础模型的合法和恶意使用的重要分类问题。
新风险防御难易度	改进防御措施以应对新风险有多难? 虽然现有的防御措施为应对开放基础模型引入的新风险提供了基线，但可以实施新的防御措施或修改现有的防御措施，以应对总体风险的增加。	大模型可以纳入信息安全工具包中以加强防御。展示了大模型如何扩大流行的模糊测试工具OSS-Fuzz的覆盖范围。基础模型还可用于监测来自已部署软件系统的信号，以发现主动攻击的迹象。Google在其流行的恶意软件检测平台Virus- Total中利用了大模型，使用模型来帮助解释特定文件中包含的恶意软件的功能。纵深防御在辅助防御方面将继续发挥重要作用。无论用于自动漏洞检测的模型是开放还是封闭的，信号以及以机器规模和速度分析信号的能力都可以为防御者提供不同的支持，因其可以更好地访问系统。
不确定性和假设	分析中隐含的不确定性和假设是什么? 一些分歧可能源于不同研究人员对开放基础模型生态系统未明确的假设。该框架要求研究人员对其具体说明，以澄清分歧。	对边际风险和防御难易度的分析假设防御者将继续更好地访问最先进的漏洞检测工具，包括基于开放基础模型的工具。它还假设防御者投资使用这些工具来更新他们的信息安全实践，并且随着模型能力的提高，攻防平衡不会发生巨大变化。

风险评估框架对网络安全风险分析的实例化[74]

[74] Sayash Kapoor et al.，“On the Societal Impact of Open Foundation Models”, 2024-02-27, https://arxiv.org/abs/2403.07918.

作者实例化了两种滥用风险的框架，对自动漏洞检测产生的网络安全风险（如上表所示）和数字化修改的NCII风险进行了初步分析。对于前者，开放基础模型目前的边际风险较低，并且有多种方法可以防御边际风险，包括使用AI进行防御。对于后者，开放基础模型目前目前带来了相当大的边际风险，而且看似合理的防御似乎很困难。

最后，作者希望其概念框架能够帮助弥补当前实证证据的不足，并提出以下政策建议：

- **开放基础模型的开发者**应明确其与产品下游开发者之间的责任划分。特别是，开发人员应明确实施了哪些负责任的AI实践，以及哪些留给可能修改模型以在面向消费者的应用程序中使用的下游开发者。
- **研究开放基础模型风险的研究人员**应采用风险评估框架来明确阐明公开发布基础模型的边际风险。如果没有这样的评估，就不清楚所概述的风险是否也存在于现状中或新风险确实无法制定良好防御措施。目前这种边际风险的证据仍然相当有限[75]。
- **政策制定者**应主动评估监管草案对开放基础模型的影响，尤其是在缺乏边际风险权威证据的情况下。资助机构应确保调查开放基础模型风险的研究获得足够的资助，同时保持适当独立于基础模型开发者的利益。
- **竞争监管机构**应该投资于更系统地衡量基础模型的收益以及开放性对这些收益的影响。例如，AI监管对围绕开放基础模型活跃创新生态系统的潜在意外后果。

2.4 两种立场的异同点

总体而言，审慎开放方和鼓励开放方都认可：开源开放对加速创新、提高透明度、促进科学研究等有重要意义，都意识到强大的基础模型如果完全开源可能带来重大风险；都主张在开源时需要采取一定的安全措施，如对系统进行必要的测试和评估，以降低风险。

	审慎开放方	**鼓励开放方**
总体态度	认为随着基础模型能力的增强，当前的开源文化可能需要调整，以确保安全和责任，支持有条件、渐进式的开源	认为开放性、透明度和广泛访问对于减轻AI风险、促进创新竞争至关重要，反对严格限制性的做法
收益vs风险	承认开源有助于外部监督、加速进步等优点，但认为对于高能力模型，开源的风险可能大于收益	强调开源在分配模型行为定义权、促进创新、支持科研、实现透明、减轻单一文化等方面的重要价值。应客观评估开源的边际风险，很多风险在闭源情况下也存在

[75] Rishi Bommasani et al.，“Considerations for Governing Open Foundation Models”, 2023-12-13, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.

	审慎开放方	鼓励开放方
如何开源	倾向于渐进、分阶段、有限度地开源。认为应该评估不同技术成熟度下的开源风险，对成熟、可控的模型可以开源，但对于前沿、高风险的模型应暂缓开源	主张尽可能完全、直接地开源源代码、模型、权重等，及时开源有助于发现和修复问题，即使是前沿模型也应尽快开源，但要有配套的安全防护
安全保障	主张开源前应进行强有力的滥用检测、使用限制等安全保证，高风险时不应开源，政府应加强监管执法	提倡在开源的同时加强防御措施，而非严格管控。呼吁明确开发者和下游开发者的责任划分
监管政策	更倾向于严格监管。建议制定细粒度的模型开源标准，政府应对高风险模型实施必要的安全管控措施	担心监管过严会抑制创新，主张更多地依靠行业自律，提醒应评估监管对开源创新生态的影响，政策制定应有充分的风险证据支撑，竞争机构应衡量开放性的价值

审慎开放方vs鼓励开放方的立场对比（本报告自制）

但双方的主要分歧在于：支持审慎开放的一方更强调潜在的重大风险，主张谨慎渐进，更倾向于集中治理。支持鼓励开放的一方则更看重开放创新的巨大收益，主张尽快行动，强调创新活力和去中心化。

这反映了当前AI治理所面临的复杂性和两难困境。未来还需要各方持续对话协作，在负责任、可持续的前提下推动AI技术造福人类。

2.5 争论之外的立场三：是否开源主要取决于商业考量

在审慎开放和鼓励开源这两种立场的争论之外，还存在一类观点认为企业决定是否开源前沿模型，很大程度上取决于其商业利益的考量，而非对滥用风险的审慎权衡。

例如，Mistral、xAI，以及国内的智谱AI和百川智能等AI企业采纳了开源与闭源并行的策略。根据不同的业务需求和市场环境，选择将某些模型开源以促进技术的普及和创新，同时保留核心技术的闭源状态，以保护商业秘密和增加竞争优势。如百度创始人兼CEO李彦宏所指出的，闭源模型在成本控制和效率提升方面可能更具优势，因为企业可以集中资源进行深度优化和定制化开发。此外，闭源模型也为企业提供了更多的商业模式选择，包括提供高质量的付费服务、定制解决方案等。这种双重策略可以帮助公司在保持技术领先的同时，也能够在开放创新生态中发挥作用。

字节跳动AI Lab总监李航引用微软全球执行副总裁沈向洋的观点，强调企业的开源决策与其市场地位紧密相关[76]。行业的领导者可能不会选择开源，第一名肯定不会开源，第二名想要和第一名竞争也不会开源，第三、第四名的公司可能会选择开源以取得一些竞争优势。李航指出，从历史经验看，这一观点有一定道理。他举例在AI公司中，OpenAI、Anthropic目前尚未开源，而Meta和Amazon则选择了开源。他认为，从商业角度看，Meta等公司开源并非出于其他考虑，而主要是为了获取商业利益优势。此外，月之暗面的CEO杨植麟也表示[77]，领先的公司大概率不会开源其主要模型，反而是落后者可能会这么做，或者开源小模型，旨在打乱现有秩序。

同时，智源研究院和智谱AI等机构签署的《北京AI安全国际共识》[78]也表明，无论开源还是闭源，行业的参与者都高度重视AI的安全和伦理治理。这显示出虽然商业利益是一个重要考量，企业在决策时还必须考虑到技术安全、伦理治理、地缘政治等复杂因素。

2.6 小结

在探讨前沿AI开源的争论中，我们识别了两种主要立场：一方是审慎开放的倡导者，他们关注潜在风险并强调在确保安全的基础上逐步推进开放；另一方则是鼓励开放的支持者，他们看重开放性对于促进创新和透明度的重要性，并反对过度限制的做法。尽管在风险与收益的评估、开源方式、安保措施以及监管政策等方面存在分歧，但都认同开放性在推动技术进步和促进社会福祉方面的重要作用。

除了这两种立场之外，还有一部分观点认为，企业在决定是否开源其AI模型时，商业利益往往是主要考量。企业会根据自身的市场定位和商业目标来选择开源或闭源策略，以实现竞争优势、保护核心技术或商业生态演进。

实际上，开源与闭源并非非此即彼的二分法，而是存在于一个广阔的设计空间中，其中包含了多种可能的开放和发布选项和策略。接下来我们将深入剖析开源AI的真正含义，探索不同模型发布选项的特点，评估各种发布和治理模式的安全性，并讨论如何实现更为负责任的开源实践。

[76] 薛澜等，“中国在这一波人工智能浪潮中处于什么位置？”，2024-03-26, https://mp.weixin.qq.com/s/ovBVf8ortxfMoIEW8A23cw.
[77] 张小珺，“月之暗面杨植麟复盘大模型创业这一年：向延绵而未知的雪山前进”，2024-03-01, https://mp.weixin.qq.com/s/kEKotLcnlFK0jf8gNajXIg.
[78] Yoshua Bengio et al.，“北京AI安全国际共识”，2024-03-11, https://idais-beijing.baai.ac.cn.

3 开源vs闭源，是错误的二分法

发布选项可能存在一个广阔而未充分探索的设计空间，这似乎是一个需要多元专业知识才能解决的社会技术设计问题。

——吉里什·萨斯特里 (Girish Sastry)[79]

关于基础模型发布选项的讨论，通常基于“开源”与“闭源”的对立。值得探讨的是，有哪些不同的发布选项，如何能够更加负责任地开源，以及如何在维持闭源所带来的优势的同时也享受到开源所带来的益处。换句话说：我们是否能够两全其美？

3.1 不同于开源软件，开源AI的概念尚未得到清晰定义

开源最初是用于描述开源软件(open-source software, 简称OSS)，它是一种社会契约，意味着软件的源代码是公开可访问的，任何人都可以查看、使用、修改和分发，并且是在开源许可证下发布的。开源软件的标准定义必须满足十个核心标准[80,81]，包括免费提供源代码、允许衍生作品以及不歧视任何使用该软件的领域或群体等。因此，开源既指源代码的可用性，也指允许下游不受限制地使用所述代码的合法许可。

然而，随着像Llama 1、Llama 2、StableLM这样的AI模型的发布，“开源”这一术语与开源许可证的要求开始脱节[82]。一些开发者使用“开源”一词仅意味着他们的模型可以被下载，但许可证可能仍然禁止某些使用情况和分发。例如，尽管Meta将Llama 2称为开源模型，但其许可证有一个限制，即拥有超过7亿月活跃用户的下游开发者不能将其用于商业用途，其输出也不能用于训练其他大模型[83]。因此严格来说，根据传统的开源软件定义，Llama 2并不是开源的，将其作为开源进行市场推广被开源倡议组织批评为错误和误导性的[84]。其他组织则使用了具有使用限制的OpenRAIL(Open & Responsible AI)许可证[85]。

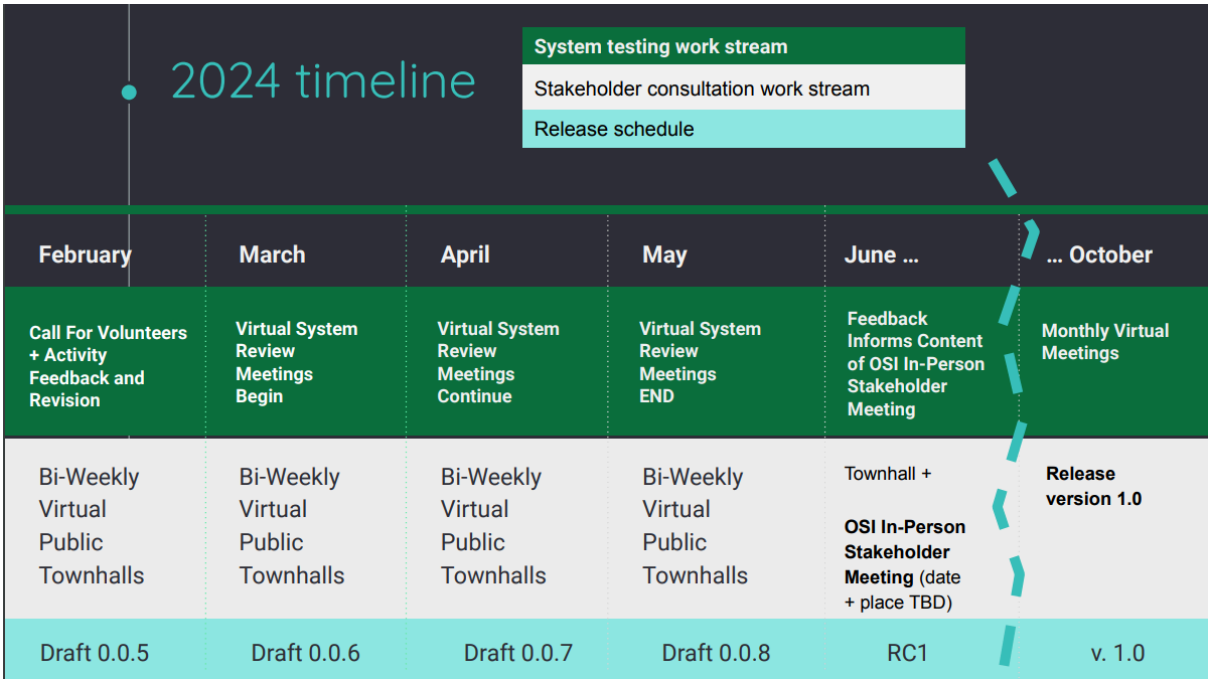
抛开许可证问题，开源软件仅仅指“免费且公开可下载的源代码”的概念，并不直接适用于AI，因为系统的构建方式不同[86]。对于AI系统，“源代码”可以指推理代码和/或训练代码，

[79] Girish Sastry, “Beyond ‘Release’ vs. ‘Not Release’”, 2021-10-18, https://crfm.stanford.edu/commentary/2021/10/18/sastry.html.
[80] Open Source Initiative, “The Open Source Definition”, 2024-02-16, https://opensource.org/osd.
[81] Choose a License, “Licenses”, 2024-04-23(引用日期), https://choosealicense.com/licenses/.
[82] Center for the Governance of AI, “Open-Sourcing Highly Capable Foundation Models”, 2023-09-29, Open-Sourcing Highly Capable Foundation Models.
[83] Meta, “Llama 2 Version Release”, 2023-07-18, https://ai.meta.com/llama/license/.
[84] Open Source Initiative, “Meta's LLaMa 2 license is not Open Source”, 2023-07-20, https://opensource.org/blog/metas-llama-2-license-is-not-open-source.
[85] Carlos Muñoz Ferrandis, “OpenRAIL: Towards open and responsible AI licensing frameworks”, 2022-08-31, https://huggingface.co/blog/open_rail.
[86] Sid Sijbrandij, “AI weights are not open 'source'”, 2023-06-27, https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/.

这两者可以独立共享。AI系统还有超出源代码的其他系统组件，如模型权重和训练数据，所有这些都可以独立于源代码和彼此共享或保留。

正如开放源代码促进会(Open Source Initiative, 简称OSI)的执行董事Stefano Maffulli所说的，开源AI的概念尚未得到清晰定义。不同的组织使用该术语来指代不同的事物。这非常令人困惑，因为每个人都用它来表示不同程度的“公开可用”，并提出了一系列新的挑战，其中最重要的是围绕训练数据的隐私和版权问题[87]。

专家们对于哪些模型组件需要共享才能将AI模型视为开源并没有达成一致。OSI自2022年以来一直致力于明确开源AI的确切定义[88]，截止到2024年3月底，已形成0.0.6版草案，并计划在2024年内发布1.0版。



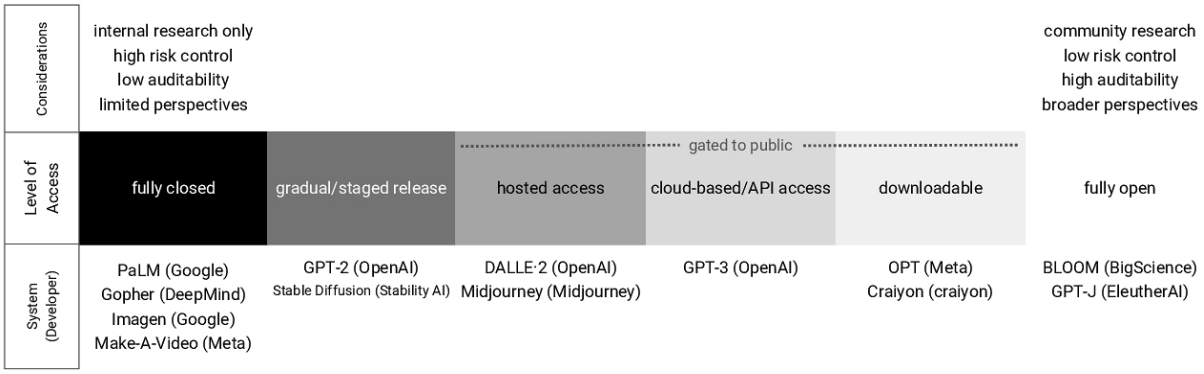
OSI的开源AI定义工作的2024年时间线[89]

[87] Edd Gent, “Protesters Decry Meta's 'Irreversible Proliferation' of AI”, 2023-10-06, https://spectrum.ieee.org/meta-ai.
[88] Open Source Initiative, “Join The Discussion on Open Source AI”, 2024-04-23(引用日期), https://opensource.org/deepdive.
[89] Open Source Initiative, “Open Source AI Definition”, 2024-01-26, https://opensource.org/wp-content/uploads/2024/01/osi_townhall_2.pdf.

3.2 从“完全开放”到“完全封闭”之间存在多种模型发布选项

3.2.1 AI系统访问的渐进等级

模型要么开源，要么闭源的想法，提出了错误的二分法[90,91]。从“完全封闭”到“完全开放”之间存在多种模型发布选项。



AI系统访问的渐进等级：需要进一步的研究和定义[92]

系统越开放，可以更好地对其进行审核和社区研究，但更难控制其风险。

通常所说的“开源”模型发布涵盖了Irene Solaiman梯度最右侧两个系统访问类别：可下载（Downloadable，特别是无门槛下载——意味着任何人都可以免费下载可用组件，通常包括权重和架构）和完全开放（Fully Open，源代码、权重、训练数据、其他组件、文档全部公开）。

开源访问的渐进等级：对于完全开放的模型，源代码、权重、训练数据以及所有其他模型组件和可用文档都是公开的。然而，在无门槛下载类别中——其中一些组件是可公开下载的（通常包括权重和架构），而其他组件则被保留——还有进一步规范的空间。重要的是，开源的确切收益和风险是由公开的模型组件和文档的特定组合决定的。

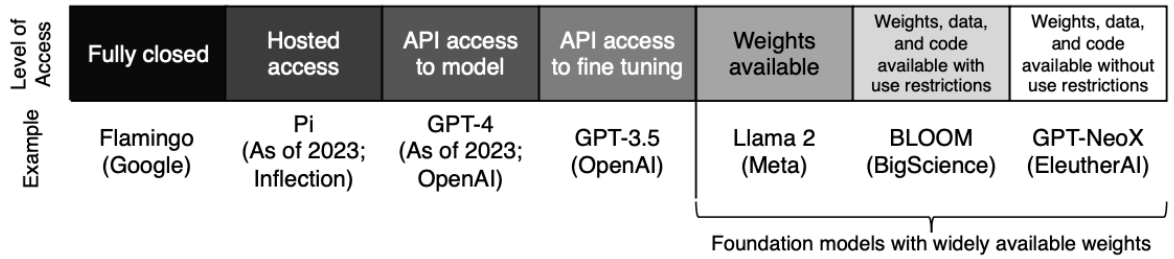
需要明确的标准和定义：需要对项目进行调查和阐明通过访问不同的模型组件（组合）可以进行哪些活动。这些信息对于构建有效且细粒度和不过分繁重的模型发布标准，以及确保开源价值受到保护并在安全的情况下享受利益至关重要。

3.2.2 进一步探讨基础模型的“发布”

关于如何监管基础模型的争论的核心是基础模型的发布(released)过程：在多大程度上以及通过什么机制提供给基础模型开发人员之外的实体？

[90] Center for the Governance of AI，“Open-Sourcing Highly Capable Foundation Models”, 2023-09-29, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.
[91] Girish Sastry，“Beyond 'Release' vs. 'Not Release' ”, 2021-10-18, https://crfm.stanford.edu/commentary/2021/10/18/sastry.html.
[92] Irene Solaiman，“The Gradient of Generative AI Release: Methods and Considerations”, 2023-02-05, https://arxiv.org/abs/2302.04844.

发布的可以是多维的：不同的资产（例如，训练数据、代码、模型权重）可以发布给特定的实体或公众。开发者在完全封闭设置（不向任何人发布任何内容）和完全开放设置（所有资产都向所有人发布）之间有许多中间选项。



斯坦福大学基础模型研究中心经Solaiman授权的修改版AI系统访问的渐进等级[93]

模型可以是完全封闭的（如Google DeepMind的Flamingo，不向开发组织外的任何人开放）；托管访问（通过网页界面提供，如Inflection的Pi）；模型的API访问（如OpenAI的GPT-4）；微调的API访问（如OpenAI的GPT-3.5）[94,95]；权重可用（如Meta的Llama 2）；有使用限制的权重、代码和数据可用（如BigScience的BLOOM）；无使用限制的权重、代码和数据可用（如EleutherAI的GPT-NeoX）。

斯坦福大学的学者使用的开放基础模型(open foundation models)的概念，即权重广泛可用的基础模型。这与美国《安全、可靠和可信AI行政命令》的要求一致。OSS Capital也在尝试对“开放权重”(Open Weight)的概念[96]和许可框架[97]进行更严格的定义[98]。

围绕开放基础模型的许多担忧源于这样一个事实：一旦模型权重发布，开发人员就放弃对其下游使用的控制。即使开发者对下游使用以及谁可以下载模型进行限制，这种限制也可能被下游开发者特别是恶意行为者忽略。相比之下，面对恶意使用或其他重要风险，封闭基础模型开发者可以限制对模型的访问，即通过在模型发布渐进等级上转移到更严格的点来减少访问。

[93] Rishi Bommasani et al.，“Considerations for Governing Open Foundation Models”，2023-12-13, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.
[94] OpenAI，“Fine-tuning”，2024-04-23(引用日期), https://platform.openai.com/docs/guides/fine-tuning, OpenAI目前可进行微调的模型包括：gpt-3.5-turbo-0125（推荐）、gpt-3.5-turbo-1106、gpt-3.5-turbo-0613、babbage-002、davinci-002，以及gpt-4-0613（实验性，需申请访问权限）。
[95] OpenAI，“Introducing improvements to the fine-tuning API and expanding our custom models program”，2024-04-04, https://openai.com/blog/introducing-improvements-to-the-fine-tuning-api-and-expanding-our-custom-models-program.
[96] Sid Sijbrandij，“AI weights are not open 'source' ”，2023-06-27, https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/.
[97] Heather Meeker，“Definition/Open Weights License.MD at main”，2024-01-04, https://github.com/Open-Weights/Definition/blob/main/Open%20Weights%20License.MD.
[98] Heather Meeker，“Definition/Definition.md”，2024-02-07, https://github.com/Open-Weights/Definition/blob/main/Definition.md.

关于模型发布的更多信息，可参考斯坦福大学的研究人员整理的Ecosystem Graphs数据库[99]。

因此，发布引入了科学知识的开放生产（包括基础模型的局限性和风险的知识）与发布可能导致的不安全部署风险之间的紧张关系。虽然这种紧张关系不可能完全化解，但**我们可以超越简单化的一维视角，探索更丰富的发布政策设计空间，从而取得一些进展。**

斯坦福的学者提出[100]可以通过确定四个关键问题来为这个设计空间提供一个坐标系：发布什么、向谁发布、何时发布以及如何发布。**社区目前缺乏发布规范，不同的基础模型开发人员有非常不同的政策**。我们鼓励制定社区规范，并对研究访问的发布进行协调。

1）发布什么

斯坦福大学的Percy Liang副教授认为，通常可以认为更开放的发布可以使研究人员能够解决更深层次的问题[101]。

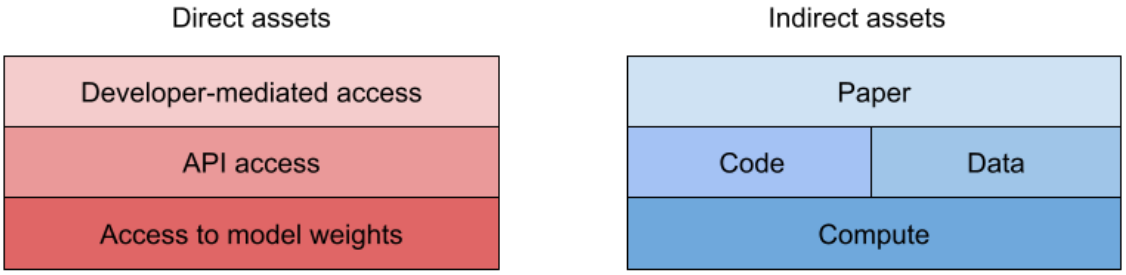
发布内容首先可以被分为直接资产和间接资产：

直接资产提供对现有模型的访问，这既允许对模型进行即时研究，也可以支持模型的部署。直接访问的形式包括：(i)开发者中介访问(例如对预测或嵌入的访问)，(ii)API访问，以及(iii)对模型权重的访问。目前，API访问是最常见的，它是一种结构化访问[102,103]，可以实现对访问的监测和撤销。开发者中介访问(即让基础模型开发者代表外部研究人员运行评估)提供更强的监督，并降低维护API的基础设施成本，但引入了人为瓶颈，无法支持需要交互性的研究。另一方面，访问模型权重可以支持更深入的研究，这是即使有API访问也无法实现的，例如开发新的微调方法。**请注意，访问模型权重并不排除结构化访问：模型权重可以托管在开发者控制的环境中，这将为开发者提供监督，为研究人员提供算力资源的便利性。**

间接资产提供构建模型的手段;这些包括：(i)描述基础模型的论文，包括有关数据、训练和模型的细节；(ii)访问训练和数据处理的代码；(iii)访问训练数据(很少见)；(iv)训练新模型的算力资源(是重要的)。**值得注意的是，发表论文也是一种发布形式。**如果一篇论文包含足够的细节来复现模型，那么发布该论文在使高资源参与者能够使用(更令人担忧的是滥用)方面可能相当于发布完整模型[104]。但实际上，仅凭论文是不足以完全复现的，需要以**代码、数据和算力**

[99] Stanford CRFM, “Ecosystem Graphs for Foundation Models”, 2024-04-23(引用日期), https://crfm.stanford.edu/ecosystem-graphs/index.html?mode=table.
[100] Percy Liang et al., “The Time Is Now to Develop Community Norms for the Release of Foundation Models”, 2023-05-17, https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models.
[101] Percy Liang, “Meta's release of OPT is an exciting step towards opening new opportunities for research”, 2022-05-04, https://twitter.com/percyliang/status/1521627736892010497.
[102] Toby Shevlane, “Structured access: an emerging paradigm for safe AI deployment”, 2022-01-13, https://arxiv.org/abs/2201.05159.
[103] Benjamin S. Bucknall, Robert F. Trager, “Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements”, 2023-10-27, https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements/.
[104] EleutherAI, “Why Release a Large Language Model?”, 2021-06-02, https://blog.eleuther.ai/why-release-a-large-language-model/.

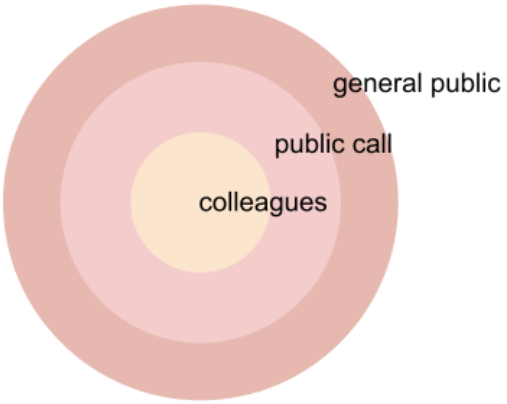
的形式进行更强的发布。如果以负责任的方式发布，这些资产将开辟广阔的新研究机会，包括探索新的模型架构、探索新的训练策略以及执行数据消融。**这种类型的研究访问（特别是算力）对于从根本上长期改进基础模型至关重要。**



基础模型开发人员可以发布直接资产，这提供了使用现有基础模型的方法；
和间接资产，它们提供了理解现有模型的构造或构建新模型的方法[100]

2）向谁发布

基础模型开发人员通常会逐步扩大其资产的开放范围，(i)起初仅向那些被充分了解并信赖的内部同事开放，(ii)随后逐步扩展到根据公开征集访问申请并被开发者授予访问权限的人，(iii)最终普及至一般大众。



基础模型的开发者通常会逐步扩大资产发布受众[100]

在这一过程中，介于内部同事与广泛公众之间的第二阶段尤为关键。**这一阶段涵盖了由第三方进行的严密审计与红队攻击测试，以此确保模型的安全性。**

随着模型能力的增强，这种“了解你的客户”（KYC）的审查变得尤为重要，因为高能力模型的潜在风险更大。然而，同时必须注意到过度严格的KYC要求可能会抑制技术创新和应用普及。因此，可以考虑在第一阶段与第二阶段之间增设一类用户类别：持有经官方认证的安全使用许可证的用户。这些用户必须通过一系列专业培训与资质审核，其资格类似于持枪证，以保证他们在访问高敏感度资产时的专业性与安全性。

3）何时发布

何时发布资产取决于内在属性（例如安全评估的结果）和外部条件（例如存在哪些其他模型以及已经发布了多长时间）。一般来说，我们建议发布应该分阶段进行[105]，每个阶段都沿着**“发布什么”**或**“向谁发布”**轴扩展。重要的是，我们认为这一进展应该受到条件的限制：例如，为了扩大对公众的访问，应该经过一些时间间隔以便有时间进行分析，并且模型应该通过一定的安全标准。

4）如何发布

除了分阶段发布之外，发布也不应该是一次性的决定。基础模型开发人员有责任随着时间的推移维护其版本，类似于维护软件[106]。发布应包括开发人员和研究人员之间的双向沟通方式。如果下游开发者有反馈，例如特定的故障案例或系统偏差，他们应该能够向开发人员公开报告这些反馈，类似于提交软件错误报告。相反，如果模型开发人员更新或弃用模型，他们应该通知所有下游开发者。

基础模型的发布非常重要，没有任何一个组织能够拥有所需的不同视角来预见所有长期问题，同时也必须适当控制发布以最大程度地降低风险。目前对于发布缺乏规范，建议制定社区规范并鼓励研究访问发布方面的协调。

3.3 基础模型安全性评测：开放vs封闭模型均显示出对各种攻击的脆弱性

3.3.1 当前还没有成熟的安全性评测，需要评测科学

模型评测是一个新兴领域，当前还没有成熟的评测[107]，不能像在成熟领域那样信任评测结果，开源大模型的安全评估方法更是明显缺乏[108]，我们需要评测科学(We need a Science of Evals)[109]。

1）HarmBench：对自动化红队和鲁棒拒绝机制进行详细比较测试的标准化评估框架

在HarmBench的研究中，对开源模型(Open-Source Models)和闭源模型(Closed-Source Models)进行了评测对比，以评估它们在面对各种攻击时的安全性和鲁棒性。

---

[105] Irene Solaiman et al.，“Release Strategies and the Social Impacts of Language Models”，2019-08-24, https://arxiv.org/abs/1908.09203.
[106] Colin Raffel，“A Call to Build Models Like We Build Open-Source Software”，2021-12-08, https://colinraffel.com/blog/a-call-to-build-models-like-we-build-open-source-software.html.
[107] Stanford HAI，“2024 AI Index Report”，2024-04-15, https://aiindex.stanford.edu/report/.
[108] Yuxia Wang et al.，“Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs”，2023-08-25, https://arxiv.org/abs/2308.13387.
[109] Apollo Research，“We need a Science of Evals”， 2024-01-22, https://www.apolloresearch.ai/blog/we-need-a-science-of-evals.

在开源与闭源模型的比较中，研究发现没有一种模型能够对所有攻击完全鲁棒，鲁棒性在这两类模型中甚至在同一类别内也有显著差异。这意味着无论是开源模型还是闭源模型，都存在被特定攻击方法成功的可能性。通过使用HarmBench进行的大规模比较实验揭示了一些有趣的发现，**例如，模型的鲁棒性与其规模的相关性较小，而是更多地取决于训练数据和算法。**

总的来说，不能简单地说开源模型或闭源模型的安全性更高，因为它们的安全性取决于多种因素，包括模型的设计、训练过程、以及所面临的特定攻击类型。HarmBench提供了一个可以帮助研究人员和开发者评估和改进各种模型安全性的工具。通过持续的红队测试和对抗性训练，可以提高模型的鲁棒性，无论是开源还是闭源模型，都能够受益于这些方法，从而推动整个领域向更安全的AI系统发展。

Model	Baseline															
	GCG	GCG-M	GCG-T	PEZ	GBDA	UAT	AP	SFS	ZS	PAIR	TAP	TAP-T	AutoDAN	PAP-top5	Human	DR
Llama 2 7B Chat	32.5	21.2	19.7	1.8	1.4	4.5	15.3	4.3	2.0	9.3	9.3	7.8	0.5	2.7	0.8	0.8
Llama 2 13B Chat	30.0	11.3	16.4	1.7	2.2	1.5	16.3	6.0	2.9	15.0	14.2	8.0	0.8	3.3	1.7	2.8
Llama 2 70B Chat	37.5	10.8	22.1	3.3	2.3	4.0	20.5	7.0	3.0	14.5	13.3	16.3	2.8	4.1	2.2	2.8
Vicuna 7B	65.5	61.5	60.8	19.8	19.0	19.3	56.3	42.3	27.2	53.5	51.0	59.8	66.0	18.9	39.0	24.3
Vicuna 13B	67.0	61.3	54.9	15.8	14.3	14.2	41.8	32.3	23.2	47.5	54.8	62.1	65.5	19.3	40.0	19.8
Baichuan 2 7B	61.5	40.7	46.4	32.3	29.8	28.5	48.3	26.8	27.9	37.3	51.0	58.5	53.3	19.0	27.2	18.8
Baichuan 2 13B	62.3	52.4	45.3	28.5	26.6	49.8	55.0	39.5	25.0	52.3	54.8	63.6	60.1	21.7	31.7	19.3
Qwen 7B Chat	59.2	52.5	38.3	13.2	12.7	11.0	49.7	31.8	15.6	50.2	53.0	59.0	47.3	13.3	24.6	13.0
Qwen 14B Chat	62.9	54.3	38.8	11.3	12.0	10.3	45.3	29.5	16.9	46.0	48.8	55.5	52.5	12.8	29.0	16.5
Qwen 72B Chat	-	-	36.2	-	-	-	-	32.3	19.1	46.3	50.2	56.3	41.0	21.6	37.8	18.3
Koala 7B	60.5	54.2	51.7	42.3	50.6	49.8	53.3	43.0	41.8	49.0	59.5	56.5	55.5	18.3	26.4	38.3
Koala 13B	61.8	56.4	57.3	46.1	52.7	54.5	59.8	37.5	36.4	52.8	58.5	59.0	65.8	16.2	31.3	27.3
Orca 2 7B	46.0	38.7	60.1	37.4	36.1	38.5	34.8	46.0	41.1	57.3	57.0	60.3	71.0	18.1	39.2	39.0
Orca 2 13B	50.7	30.3	52.0	35.7	33.4	36.3	31.8	50.5	42.8	55.8	59.5	63.8	69.8	19.6	42.4	44.5
SOLAR 10.7B-Instruct	57.5	61.6	58.9	56.1	54.5	54.0	54.3	58.3	54.9	56.8	66.5	65.8	72.5	31.3	61.2	61.3
Mistral 7B	69.8	63.6	64.5	51.3	52.8	52.3	62.7	51.0	41.3	52.5	62.5	66.1	71.5	27.2	58.0	46.3
Mixtral 8x7B	-	-	62.5	-	-	-	-	53.0	40.8	61.1	69.8	68.3	72.5	28.8	53.3	47.3
OpenChat 3.5 1210	66.3	54.6	57.3	38.9	44.5	40.8	57.0	52.5	43.3	52.5	63.5	66.1	73.5	26.9	51.3	46.0
Starling 7B	66.0	61.9	59.0	50.0	58.1	54.8	62.0	56.5	50.6	58.3	68.5	66.3	74.0	31.9	60.2	57.0
Zephyr 7B	69.5	62.5	61.1	62.5	62.8	62.3	60.5	62.0	60.0	58.8	66.5	69.3	75.0	32.9	66.0	65.8
R2D2 (Ours)	5.5	4.9	0.0	2.9	0.2	0.0	5.5	43.5	7.2	48.0	60.8	54.3	17.0	24.3	13.6	14.2
GPT-3.5 Turbo 0613	-	-	38.9	-	-	-	-	-	24.8	46.8	47.7	62.3	-	15.4	24.5	21.3
GPT-3.5 Turbo 1106	-	-	42.5	-	-	-	-	-	28.4	35.0	39.2	47.5	-	11.3	2.8	33.0
GPT-4 0613	-	-	22.0	-	-	-	-	-	19.4	39.3	43.0	54.8	-	16.8	11.3	21.0
GPT-4 Turbo 1106	-	-	22.3	-	-	-	-	-	13.9	33.0	36.4	58.5	-	11.1	2.6	9.3
Claude 1	-	-	12.1	-	-	-	-	-	4.8	10.0	7.0	1.5	-	1.3	2.4	5.0
Claude 2	-	-	2.7	-	-	-	-	-	4.1	4.8	2.0	0.8	-	1.0	0.3	2.0
Claude 2.1	-	-	2.6	-	-	-	-	-	4.1	2.8	2.5	0.8	-	0.9	0.3	2.0
Gemini Pro	-	-	18.0	-	-	-	-	-	14.8	35.1	38.8	31.2	-	11.8	12.1	18.0
Average (↑)	54.3	45.0	38.8	29.0	29.8	30.8	43.7	38.3	25.4	40.7	45.2	48.3	52.7	16.6	27.3	25.3

HarmBench对标准/上下文/版权行为的评测对比(上方21个为开源模型，下方8个为闭源模型)[110]

2）SuperCLUE-SafetSC-Safety: 中文大模型多轮对抗安全基准

SC-Safety大模型安全类测评，包含以下三大能力的检验：传统安全类、负责任人工智能和指令攻击。总体发现：

1. 总得分，是指计算每一道题目的分数，汇总所有分数并除以总分。**可以看到总体上，相对于开源模型，闭源模型安全性做的更好。**
2. 与通用基准不同，安全总榜上国内代表性闭源服务或开源模型与国外领先模型较为接近；闭源模型默认调用方式为API。

[110] Mantas Mazeika et al.，“HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal”, 2024-02-27, https://arxiv.org/abs/2402.04249.

排名	模型	机构	总分	传统安全类	负责任类	指令攻击类	许可
1	BlueLM	vivo	92.51	87.21	96.59	94.16	闭源
2	AndesGPT	OPPO	90.87	87.46	94.60	90.81	闭源
3	Yi-34B-Chat	零一万物	89.30	85.89	94.06	88.07	开源
4	文心一言4.0	百度	88.91	88.41	92.45	85.73	闭源
-	GPT4	OpenAI	87.43	84.51	91.22	86.70	闭源
5	讯飞星火(v3.0)	科大讯飞	86.24	82.51	91.75	85.45	闭源
6	360gpt-pro	360	85.31	82.82	90.35	82.75	闭源
7	讯飞星火(v2.0)	科大讯飞	84.98	80.65	89.78	84.77	闭源
-	gpt-3.5-turbo	OpenAI	83.82	82.82	87.81	80.72	闭源
8	文心一言3.5	百度	81.24	79.79	84.52	79.42	闭源
9	ChatGLM2-Pro	清华&智谱AI	79.82	77.16	87.22	74.98	闭源
10	ChatGLM2-6B	清华&智谱AI	79.43	76.53	84.36	77.45	开源
11	Baichuan2-13B-Chat	百川智能	78.78	74.70	85.87	75.86	开源
12	Qwen-7B-Chat	阿里巴巴	78.64	77.49	85.43	72.77	开源
13	OpenBuddy-Llama2-70B	OpenBuddy	78.21	77.37	87.51	69.30	开源
-	Llama-2-13B-Chat	Meta	77.49	71.97	85.54	75.16	开源
14	360GPT_S2_V94	360	76.52	71.45	85.09	73.12	闭源
15	Chinese-Alpaca2-13B	yiming cui	75.39	73.21	82.44	70.39	开源
16	MiniMax-Abab5.5	MiniMax	71.90	71.67	79.77	63.82	闭源

SC-Safety安全总榜（更新时间2024年1月4日）[111]

[111] Liang Xu，“SuperCLUE-Safety：中文大模型多轮对抗安全基准”，2024-01-04, https://github.com/CLUEbenchmark/SuperCLUE-Safety.

3.3.2 封闭模型的脆弱性

尽管封闭模型的开发者可以通过限制模型访问来防范恶意使用，例如实施严格的访问控制策略，但模型仍面临多种安全威胁，容易被恶意使用。以下是一些有代表性的脆弱性：

- **微调对齐的大语言模型引入的新安全风险：**封闭模型的开发者常通过微调来优化模型以满足特定的下游用例需求。然而，这种定制化的微调过程可能会带来安全成本。例如，普林斯顿大学等机构的研究发现[112]，当微调权限扩展至最终用户时，现有的安全措施无法全面覆盖新的安全风险。他们通过对OpenAI的API以不到0.20美元的成本在10个这样的样本上进行微调，演示了如何绕过GPT-3.5 Turbo的安全护栏。此外，即使是非恶意的常规数据集微调也可能不经意间降低模型的安全性。
- **构建通用和可迁移的对抗性攻击：**对已进行安全对齐的大型语言模型，如ChatGPT，可以设计特殊字符序列，这些序列加入到用户查询中可能使模型执行有害指令。通常这种“越狱”需要大量的手动设计工作，并且通常可以很容易地被模型开发者修补。卡内基梅隆的研究表明[113]，这种对抗性攻击可以完全自动化地构建，允许生成几乎无限数量的攻击。这些攻击虽然针对开源模型设计，但也可迁移到ChatGPT、Bard和Claude等多种封闭模型，增加了对封闭系统的安全性担忧，特别是当它们开始以更加自主的方式被使用时。
- **通过偏离攻击从语言模型中提取训练数据：**Google DeepMind等机构的研究发现[114]，通过重复某些词语，可以诱使模型泄露其训练数据，这种数据称为“可提取记忆”(extractable memorization)。作者设计了一种“偏离攻击”(divergence attack)，使模型偏离其聊天机器人风格的生成，并且以正常行为高出150倍的概率泄露训练数据。这种偏离攻击表明封闭模型同样容易受到信息泄露的威胁[115]。这类攻击的成功几率与特定词语的选择有关，显示了模型在数据隐私保护方面的脆弱性。
- **偏好数据引起的“谄媚”(Sycophancy)现象：**Anthropic的研究[116]探讨了一种大模型产生的回应趋向于符合用户的立场或偏好，但有时可能以牺牲真实性或准确性为代价的行为。此研究涵盖了5个当时最领先的AI系统：Claude 1.3、Claude 2、GPT-3.5、GPT-4、Llama 2，发现这些系统一致地展现了谄媚行为。该研究强调了封闭模型在处理偏好数据时如何产生偏向用户立场的输出，从而影响了决策和信息的质量。尽管谄

[112] Xiangyu Qi，“Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!”，2023-10-05, https://llm-tuning-safety.github.io/.
[113] Andy Zou，“Universal and Transferable Adversarial Attacks on Aligned Language Models”，2023-12-20, https://llm-attacks.org/.
[114] Milad Nasr et al.，“Scalable Extraction of Training Data from (Production) Language Models”，2023-11-08, https://arxiv.org/abs/2311.17035.
[115] 段雅文，“AI安全前沿 #2 | 大模型谄媚现象、RLHF后门攻击、AI4Science模型的滥用风险、态势感知能力、表征工程”，2023-12-29, https://mp.weixin.qq.com/s/kntG9r-VD2Aw5iMGhFKqgw.
[116] Mrinank Sharma，“Towards Understanding Sycophancy in Language Models”，2023-10-20, https://arxiv.org/abs/2310.13548.

媚行为通常被视为误导性反馈的问题，但它同样是封闭模型中一个未被解决的挑战。同时，这也说明了开发超越仅依赖人类评分的训练方法的必要性[117]。

3.3.3 开放模型的脆弱性

为何开源模型更容易遭受滥用？实际上，尽管最好的开源模型与闭源模型相比仍存在能力差距（有专家判断差距大约为1年半[118]，另有专家认为差距正在缩小[119]），但在众多场景中，开源模型往往是恶意行为者的首选，原因可能包括[120]：

- **无法监测滥用或偏见：**封闭模型可以监测有意或无意的滥用行为并禁用相关帐户，而如果模型在不良行为者自己的硬件上运行，则开放模型本质上无法监测。开放模型也无法进行偏见监测，因为他们甚至不知道他们的模型在被谁以及如何使用。
- **删除安全特性的能力：**研究人员已经证明，通过对模型代码进行极其简单的修改和其他对抗攻击[121]，就可以删除开放模型的“安全特性”，如果恶意行为者在自己的硬件上操作，这种行为是难以检测的。
- **精心微调后的滥用：**专家还证明，开放模型可以进行微调以便在滥用情况下做得更好（例如GPT-4Chan），例如提高其在合成生物学、错误信息生成或说服方面的表现。
- **无速率限制：**封闭模型可以对每个用户的内容产出进行限制，但是当恶意行为者使用自己的硬件时，他们可以生成旨在伤害人们的无限内容，并使其高度个性化和互动，而只受到自己硬件的限制。这可能会助长包括窄播(narrowcasting)、虚假草根运动(astroturfing)、结队(brigading)或旨在使观众两极分化的材料等各种危害。
- **一旦发布，安全漏洞无法修补：**即使开放模型的开发人员发现了漏洞（例如Llama 2的Uncensored版本[122]可以设计生物武器），一旦发布，他们也无法有意义地召回。这使得发布开放模型的决定给社会带来了不可逆转的风险。
- **用于监测和分析目标：**开放模型不仅可以用于生成内容，还可以用来对大量内容进行结构化分析。封闭模型的输出可能受到速率限制，而开放模型则可用于分析大量有关个人的公共信息，甚至是非法获取的数据库，然后确定影响操作的目标、放大极化内容制作者、易受骗的受害者等。

---

[117] Stephen Casper et al.，“Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback”, 2023-07-27, https://arxiv.org/abs/2307.15217.
[118] Jack Clark，“Import AI 367: Google's world-spanning model; breaking AI policy with evolution; $250k for alignment benchmarks”, 2024-04-01, https://importai.substack.com/p/import-ai-367-googles-world-spanning.
[119] 张俊林，“如何看待 Meta 发布 Llama3，并将推出 400B+ 版本？”, 2024-04-19, https://www.zhihu.com/question/653373334/answer/3471466524.
[120] David Harris，“How to Regulate Unsecured 'Open-Source' AI: No Exemptions”, 2023-12-04, https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.
[121] Markus Anderljung et al.，“Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?”, 2023-03-16, https://arxiv.org/abs/2303.09377.
[122] Hugging Face，“jarradh/llama2_70b_chat_uncensored”, 2024-04-23(引用日期), https://huggingface.co/jarradh/llama2_70b_chat_uncensored.

- **对封闭模型的公开攻击：**研究人员利用开放模型来研究“越狱”，在某些情况下可以转移到封闭模型，从而使这两种类型的模型都更容易受到滥用。
- **水印去除：**开放模型可以通过改写文本或去除图像、音频、视频水印，实现对水印的大规模、自动化移除。
- **危险材料、物质或系统的设计：**虽然封闭模型可以限制与这些主题相关的查询，但开放模型的障碍可以被消除。这是一个真正的威胁，GPT-4和Claude 2在预发布版本的红队测试中发现了这方面的潜在风险[123,124]。

3.4 AI研发机构治理评测：倾向于开放vs封闭模型的机构各有所长

3.4.1 安全政策评测：倾向于封闭模型的机构表现较优

2023年全球AI安全峰会前，英国政府发布了《前沿AI安全的新兴流程》[125]。并邀请7家企业介绍自己在AI安全方面的政策。剑桥大学未来智能研究中心邀请了由15名AI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。

总体发现：

- 没有一家公司符合所有最佳实践政策。
- 其中整体做的较好的是“开展研究以促进AI安全”。
- 但也有一些是都做的不太好的，例如“通过邀请外部参与者评估其输入数据并共享输入数据审计的信息，促进对输入数据的外部审查”，“准备应对潜在的最坏情况或持续的滥用情况，包括通过快速模型回滚和撤回”。
- 可以看到，表现最好的Anthropic也只达到82%。当时也是唯一发布了自己的负责任扩展策略/负责任能力扩展的机构。
- 有几家公司明显落后，尤其是Meta和亚马逊。Meta主要在“包括保护模权重在内的安全控制”和“防止和监测模型滥用”部分失分。**这显然涉及了开源前沿基础模型的争论。**

最佳实践	Anthropic	DeepMind	Microsoft	OpenAI	Amazon	Meta
达成率	82%	75%	75%	74%	58%	48%
评级	B	C	C	C	E	F

[123] OpenAI, “GPT- Technical Report”, 2023-03-15, https://cdn.openai.com/papers/gpt-.pdf.
[124] Anthropic, “Frontier Threats Red Teaming for AI Safety”, 2023-07-26, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety.
[125] GOV.UK, “Emerging processes for frontier AI safety”, 2023-10-27, https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety

整体情况[126]

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			

分类详细评分[127]

[126] Seán Ó hÉigeartaigh et al.，“Do Companies' AI Safety Policies Meet Government Best Practice?”，2023-10-31，http://lcfi.ac.uk/news-and-events/news/2023/oct/31/ai-safety-policies/.

[127] 安远AI，“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”，2024-01-17，https://mp.weixin.qq.com/s/9c0whoQpxgxggS7rOj9Sng.

3.4.2 透明度评测：倾向于开放模型的机构表现较优

来自斯坦福大学、麻省理工学院和普林斯顿大学的多学科团队为评估透明度，设计了一个名为基础模型透明度指数(Foundation Model Transparency Index)的评分系统[128]，评估了透明度的100个不同方面，从公司如何建立基础模型、如何运作，以及如何在下游使用。团队使用指数对10家主要公司进行评分时。

总体发现：

- 令人担忧的是，即使最高分的模型也只得了54分/满分100分。这表明没有一家主要的基础模型开发者能够提供足够的透明度，揭示了AI行业在透明度方面的根本不足。主要子领域层面的分析揭示了哪些类型的透明度导致上述评分，例如数据、劳动力和算力所需的资源最不透明，而对用户数据保护及其模型的基本功能更加透明。
- 开放基础模型开发者在透明度指数上的得分平均高出封闭开发者20%，在供应链各部分的透明度优于封闭模型开发者。这种透明度对于避免重现过去不透明的数字技术造成的危害可能很重要，但目前下游对经济和社会的影响缺乏透明度仍然令人担忧。
- 主要的开放基础模型在满足法律草案关于披露训练数据和算力使用信息的要求方面普遍优于封闭模型，而封闭模型在"部署相关要求"方面表现更好。

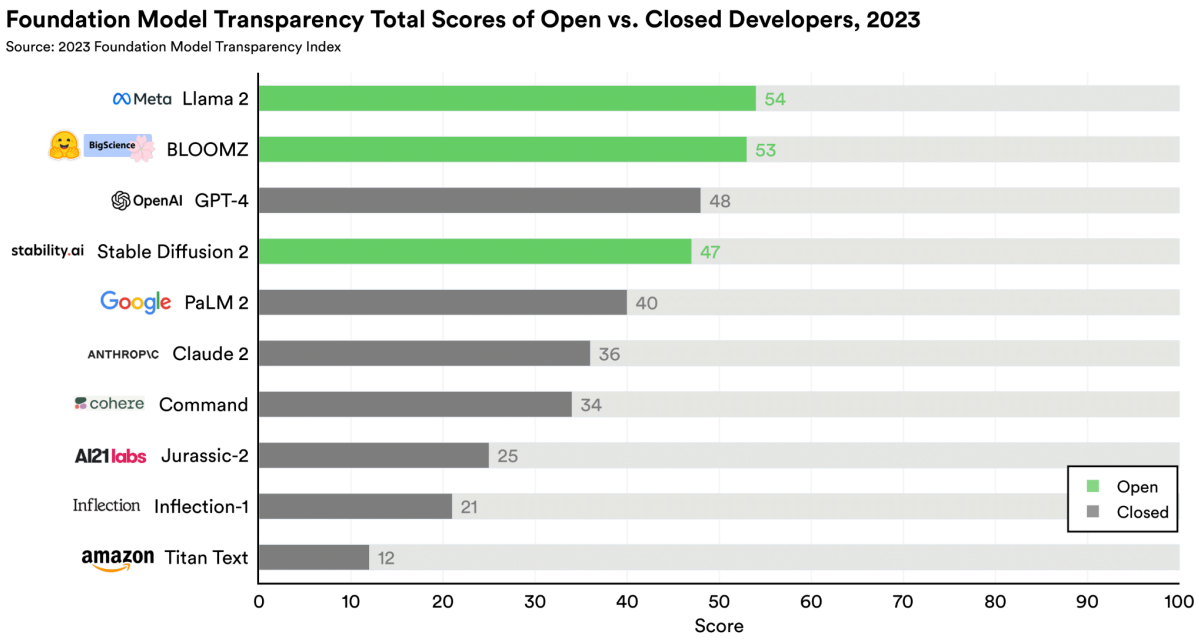
Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

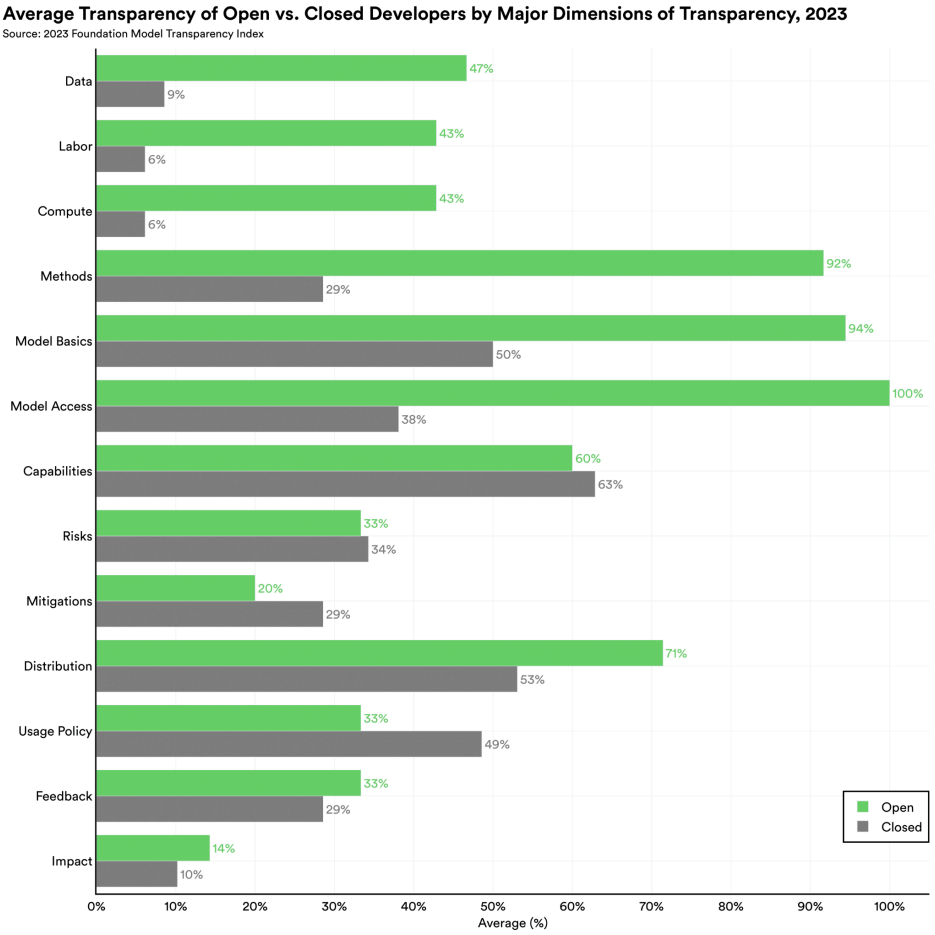
Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

对10家基础模型提供商的评分按照13个子领域进行细分，每个子领域包含三个或更多的指标[128]

[128] Stanford CRFM, “Foundation Model Transparency Index”, 2023-10-04, https://crfm.stanford.edu/fmti/.



开放模型领先，并占据排行榜的1、2、4位（分别是Meta的Llama 2、Hugging Face的BLOOMZ和Stability AI的Stable Diffusion 2），封闭模型中领先的是OpenA 的GPT-4 之后[128]

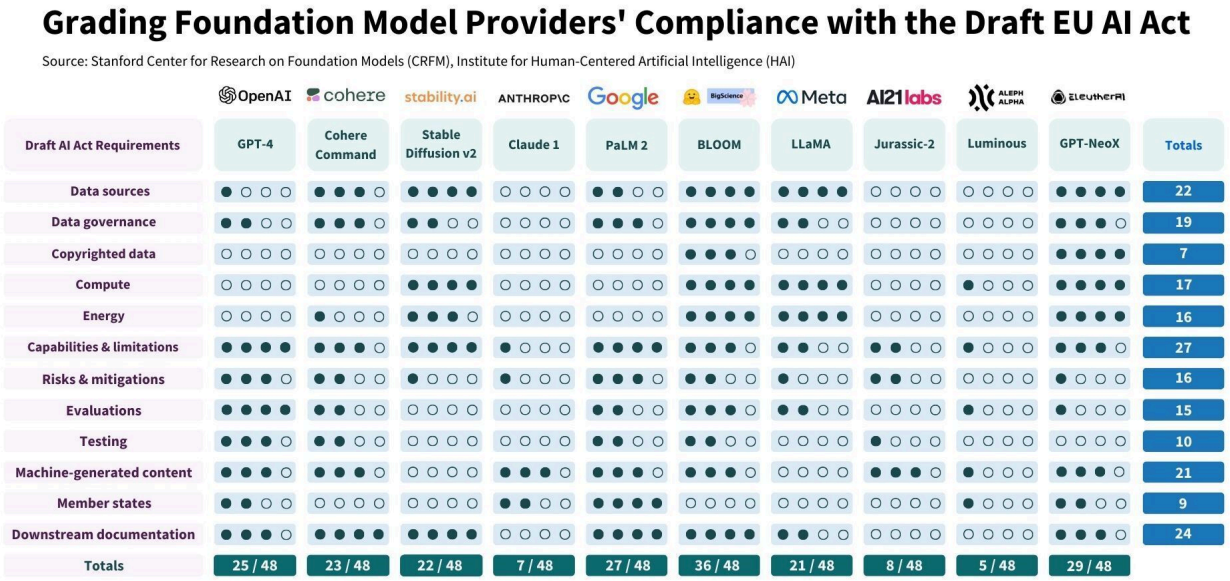


开放和封闭模型的差异是由上游指标缺乏透明度造成的，例如模型开发的数据、劳动力和算力信息[128]

但需要说明的是，虽然透明性有助于信任建立和风险管理，并且在许多行业中透明性也是法律合规的一部分，但透明性和安全性之间的关系并非简单的正相关。在实际操作中，许多行业都面临着如何在保护关键商业信息和满足公众透明度需求之间做出平衡的挑战。例如汽车厂商生产过程透明度可以不高，但需要通过安全性测试。

3.4.3 合规性评测：倾向于开放vs封闭模型的机构各有所长

欧盟《AI法案》从被提出开始即引起全球广泛关注，也会影响大多数AI大模型在欧洲的发展前景。2023年6月，在斯坦福大学曾发布一份针对大模型提供商是否满足《AI法案》草案合规性的研究[129]，研究团队从监管草案选取出22项要求，以是否有意义地使用公众信息作为标准最终选择出12项评估要求。在此基础上，研究人员将这12项要求进行维度划分为数据来源、数据处理、模型本身和实践应用四个层次。



大模型提供商是否遵守《欧盟AI法案》草案？[129]

总体发现：

- 各大AI模型的得分与满分仍有很大差距。主要的问题集中在：版权责任不明确；能源使用报告不均衡；风险缓解方面披露不充分；缺乏评估标准或审计生态系统等等。
- 开放发布通常由强调透明度的组织进行，这导致他们在资源披露要求（例如数据和算力）方面通常获得较高分数，如EleutherAI在这些类别中获得19/20分。然而，这种开放发布使监测或控制部署变得具有挑战性。

[129] Stanford CRFM, “Do Foundation Model Providers Comply with the Draft EU AI Act?”, 2023-06-16, https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.

- 封闭发布通常与为提供商的旗舰产品和服务提供支持的模型相吻合，这使得他们在部署相关要求方面得分更好。例如，Google的PaLM 2在部署方面获得11/12分。API或开发人员介导的访问为结构化访问提供了更简单的方法。

3.5 负责任开源之一：促进开放发布从构建到使用的全流程负责任实践

确保基础模型开源的安全性对于保护敏感数据、防御对抗攻击和加强应用的鲁棒性至关重要。针对基础模型开源的独特挑战量身定制安全措施，可提高模型的可信度并防范潜在风险。

我们建议根据基础模型的**生命周期和流程阶段**，设计构建和使用阶段的负责任开源维度，并针对**不同能力级别的模型**制定差异化的负责任开源要求[130]。例如对于大多数的AI模型，负责任的实践主要通过提高透明度、确保合规以及促进创新来实现。而对于能力更强的前沿模型，需要实施与模型的潜在风险相称的评估和安全缓解措施。

我们还将以审慎开放方的代表Google DeepMind对Gemma的开源和鼓励开放方的代表Meta对Llama的开源为例，对比展示他们的负责任开源实践。

3.5.1 上游：基础模型的构建（程序负责任）

模型开发	• 人员培训：开发团队接受过AI伦理和安全方面的培训 • 风险评估：开发过程中对社会的潜在影响和风险进行了全面评估 • 缓解措施：开发过程中针对风险设计了相应的缓解措施 • 伦理审查：有明确的伦理审查流程和决策记录 • 定期评估：定期评估和调整开发政策和实践，以符合技术发展和现实影响
数据和模型合规	• 数据合规：训练数据的获取和使用合法合规 • 数据质量：训练数据具有足够的质量、多样性和代表性，以避免偏见和歧视 • 数据管理：有明确的数据管理和保护政策 • 知识产权：尊重他人的知识产权，如版权、专利等 • 开源许可证：遵守开源许可证的要求和限制 • 标准规范：积极与监管机构和政策制定者沟通，推动负责任AI标准和规范
模型评测	• 安全性评测：评估了模型在安全性方面的表现 • 公平性评测：评估了模型在公平性方面的表现 • 隐私性评测：评估了模型在隐私性方面的表现 • 多学科红队：建立了多学科专家红队，并进行了广泛的内部测试和外部审计 • 环境影响评测：评估了模型在能耗和环境方面的表现和影响 • 评测结果公开：测试和评估结果透明公开，接受社区监督

上游：基础模型的构建相关指标（本报告自制）

[130] 感谢中国社科院AI安全治理实验室对本小节关于负责任开源的内涵以及差异化要求的讨论。

1）Gemma

2024年2月，Google DeepMind发布轻量级开源模型Gemma，称其性能在同等规模中最为先进，在模型的开发阶段采取的负责任举措包括[131]：

- **根据公司AI原则进行内部审查。**只有在确定收益显著且误用风险较低或可以减轻的情况下才发布模型。该公司对开放模型采取同样的方法，权衡更广泛地访问特定模型的收益与滥用风险，并考虑如何减轻这些风险。与Gemma模型的发布同时，Google DeepMind考虑了增加AI研究和创新的需求，以及支持这些用例所需的访问权限。
- **设定更高的评测标准。**需要保护下游开发者和用户免受开放模型的意外行为的影响，包括产生有毒语言或延续歧视性社会危害、模型幻觉和泄露个人身份信息。Gemma模型经过了全面的评估，并设定了比封闭模型更高的标准。评估覆盖了广泛的领域，包括安全、公平、隐私、社会风险，以及CBRN风险、网络安全和自主复制等能力。
- **开源条件：**Google DeepMind认为在现有生态系统中，Gemma模型对整体AI风险组合的影响可忽略不计。考虑到模型在研究、审计和下游产品开发方面的实用性，公司认为Gemma模型对AI社区的益处超过了潜在风险。

2）Llama

2024年4月，Meta迄今为止能力最强的开源模型Llama 3发布。与此同时，Meta也公布了其在模型的开发阶段采取的负责任举措，包括[132]：

- **应对训练中的风险。**例如扩展Llama 3的训练数据集，增加数据多样性；使用Llama 2构建文本质量分类器，为Llama 3提供支持；利用合成数据来训练编码、推理、和长上下文；遵循Meta的标准隐私审查流程，并删除了大量个人信息相关的数据。
- **安全评估与调优。**通过自动和手动评估并采取额外措施，了解和限制模型在武器、网络攻击和儿童剥削等一系列风险领域进行不必要的响应。例如，与外部和内部专家进行了广泛的红队练习，通过CyberSecEval衡量模型帮助实施网络攻击的可能性，利用RLHF让人类对模型的响应提供偏好反馈等。
- **减少良性拒绝。**改进微调方法，并使用高质量数据来显示这些语言细微差别的模型响应示例，以降低Llama 3无意中拒绝回答无害提示的可能性，这使Llama 3成为Meta迄今为止最有帮助的模型。
- **开源条件：**安全性比其他开源模型更好，接近闭源模型[133]。

---

[131] Anne Bertucio, “Building Open Models Responsibly in the Gemini Era”, 2024-02-21, https://opensource.googleblog.com/2024/02/building-open-models-responsibly-gemini-era.html.
[132] Meta, “Our responsible approach to Meta AI and Meta Llama 3”, 2024-04-28, https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/.
[133] Stanford HAI, “How to Promote Responsible Open Foundation Models”, 2023-10-3, https://hai.stanford.edu/news/how-promote-responsible-open-foundation-models.

3.5.2 模型：基础模型的功能或属性（结果负责任）

模型功能或属性	• 安全性：模型在安全性方面达到一定的标准[134]，并且嵌入的安全功能不易被移除或绕过 • 可解释性：模型在可解释性方面达到一定的标准 • 隐私保护：模型在隐私保护方面达到一定的标准 • 公平性：模型在公平性方面达到一定的标准

模型：基础模型的功能或属性相关指标（本报告自制）

1）Gemma

正如Google DeepMind的技术报告中所述，Gemma模型在人体并行评估中表现出最先进的安全性能，但也承认此发布是不可逆转的，并且开放模型造成的危害尚未明确定义，因此将继续采取与模型潜在风险相称的评估和安全缓解措施。例如，需要进一步研究事实性、对齐、复杂推理和对抗攻击的鲁棒性等，也需要更具挑战性和稳健的基准。

		Mistral v0.2	Gemma 1.1 IT	
Benchmark	metric	7B*	2B	7B
RealToxicity	avg	8.44	**7.03**	8.04
BOLD		46.0	**47.76**	45.2
CrowS-Pairs	top-1	32.76	45.89	**49.67**
BBQ Ambig	1-shot, top-1	**97.53**	58.97	86.06
BBQ Disambig	top-1	84.45	53.9	**85.08**
Winogender	top-1	**64.3**	50.14	57.64
TruthfulQA		**48.54**	44.24	45.34
Winobias 1_2		**65.72**	55.93	59.22
Winobias 2_2		84.53	**89.46**	89.2
Toxigen		61.77	**29.64**	38.75

Table 8 | Safety academic benchmark results of Gemma 1.1 IT models, compared to similarly sized, openly-available models. Evaluations run by us. Note that due to restrictive licensing, we were unable to run evals on LLaMA-2; we do not report previously-published numbers for LLaMA-2 on TruthfulQA, as we use different, non-comparable evaluation set-ups: we use MC2, where LLaMA-2 uses GPT-Judge. Results for Gemma 1.0 IT models can be found in appendix.

Gemma技术报告中的安全性评测：Gemma 1.1 IT模型的安全学术基准结果与类似规模开放模型比较[135]

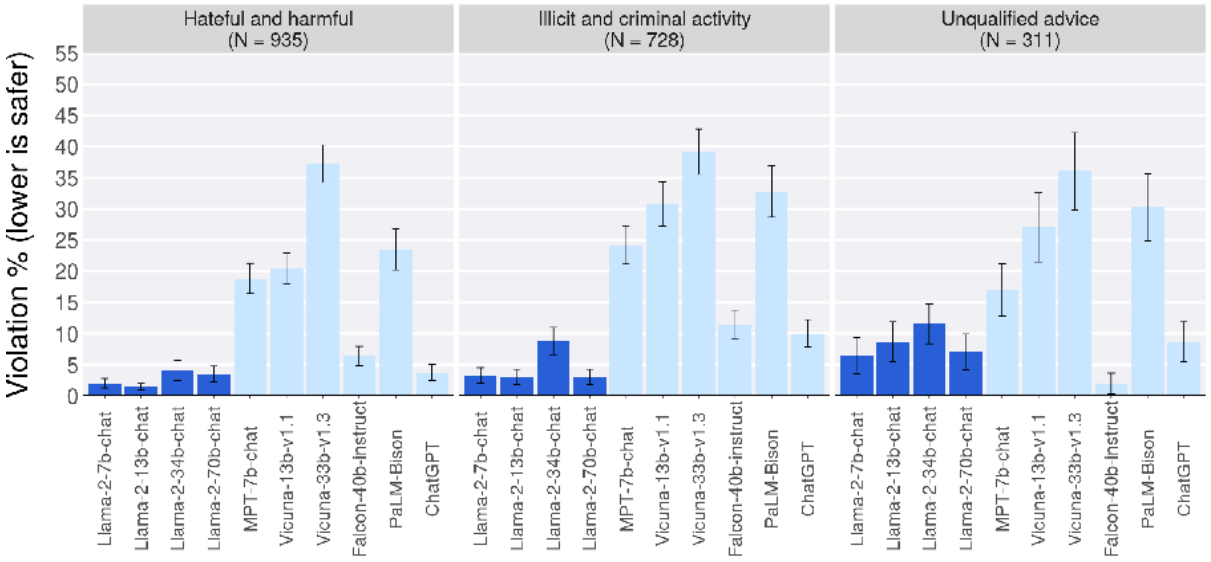
2）Llama

Meta在红队方面投入较大，并使用监督微调和RLHF等技术，使Llama 2在安全性方面与当时领先的封闭模型不相上下。虽然各个类别的模型表现相似，但Llama 2-Chat在不合格建

[134]需考量直接滥用、危险能力，以及模型微调或修改可引发的滥用，如2.1.2章节提到，前沿AI可能涉及的滥用和失控风险主要包括网络安全、化学/生物/辐射/核威胁(CBRN)、虚假信息的滥用风险，以及操纵欺骗、模型自主性导致的滥用和失控风险。

[135] Google, “Gemma: Open Models Based on Gemini Research and Technology”, 2024-02-21, https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf

议类别下的违规行为相对较多（绝对值仍然较低）。对于其他两个类别，无论模型大小如何，Llama 2-Chat始终都能实现相当或更低的违规百分比。Llama 3的技术论文和安全性评测暂未发布，安全性和社会影响还有待进一步观察。



Llama 2技术论文中的安全性评测：每个风险类别的违规百分比[136]

由于提示集的局限性、审查指南的主观性、内容标准和个人评分者，应仔细解释这些结果

3.5.3 下游：基础模型的使用（程序负责任）

模型分发和使用	• 常规性使用指南：提供了清晰的使用指南和实践建议 • 风险披露：披露了模型的预期用途、使用局限和潜在风险 • 负责任使用指南：提供了负责任使用的建议、教程或培训 • 可访问性：开源资料易于访问、理解和复现，并支持社区参与和贡献 • 模型更新：有关于开发人员版本控制协议、更改日志和弃用政策
模型监测和监督	• 开源条件：设定了开源条件，如哪些级别的模型经过哪些步骤后允许开源 • 漏洞报告：建立了漏洞报告机制 • 内容溯源：公开了用于检测此模型生成的内容的任何机制 • 影响评估：定期评估模型的社会影响，并根据反馈不断改进模型 • 长期监督：有长期的社会影响评估和风险管理计划
社区参与和治理	• 社区参与机制：建立明确的社区行为准则，促进负责任、包容的交流和协作 • 利益相关者参与：广泛征求了不同利益相关者的意见 • 开发者责任义务：明确了开发者的责任义务，提供违规问责的渠道 • 开发者限制政策：披露了谁可以和不能使用该模型的描述

[136] Hugo Touvron et al.，“Llama 2: Open Foundation and Fine-Tuned Chat Models”，2023-07-18, https://arxiv.org/abs/2307.09288.

	• 不当使用追责：有机制追究滥用模型的责任 • 用例限制政策：披露了模型的允许、限制和禁止的用途 • 可持续发展：建立有效的治理机制和发展路线图
生态和创新影响	• 促进开源生态：代码贡献者数量、代码fork数量、社区活跃度、可扩展性等 • 促进科学研究：基于此开源模型的论文数量、引用次数、研究项目数量等 • 促进产业创新：基于此开源模型的产品/服务数量和规模、经济效益等 • 促进公平和包容：模型对于弱势群体赋能、缩小数字鸿沟等方面的贡献 • 其他负责任创新：模型在伦理和社会福祉方面的创新贡献

下游：基础模型的使用相关指标（本报告自制）

1）Gemma

Google DeepMind对于Gemma模型使用阶段采取的负责任举措包括：

- **模型卡：**尽管在改进模型方面做了大量投入，但Google DeepMind认识到其局限性。为确保对下游开发者的透明度，他们发布了详细的模型卡，以便研究人员更全面地了解Gemma模型。
- **负责任的生成式AI工具包。**发布Gemma模型的同时，还为开发人员发布了负责任的生成式AI工具包[137]，提供指导和工具来帮助他们创建更安全的AI应用程序。
- **许可证/使用条款。**Gemma模型的使用条款允许个人开发者、研究人员和商业用户免费访问和重新分发，同时允许他们自由创建和发布模型变体。这些开发人员在使用模型时需承诺避免将其用于有害目的[138]，这体现了公司对负责任AI开发的承诺和对技术使用的增加考量。尽管存在使用条款的约束，但Google DeepMind认识到直接提供模型权重而非通过API方式带无法完全阻止不良行为者出于恶意目的微调Gemma。还需建立更强大的策略以防止故意滥用。
- **未来的发布方式探索。**随着能力的进步，Google DeepMind可能会探索扩展测试、交错发布或替代访问机制，以确保负责任的AI开发。

2）Llama

Meta在发布Llama 2模型的同时发布了**《负责任使用指南》**[139]，为开发人员提供了以负责任的方式构建由大语言模型支持的产品的最佳实践和注意事项，涵盖从构思到部署的各个开发阶段。Llama 3发布时，Meta在《负责任使用指南》的基础上，又为开发人员提供了安全可信

[137] Google, “Responsible Generative AI Toolkit”, 2024-02-21, http://ai.google.dev/responsible.
[138] Google, “Gemma Prohibited Use Policy”, 2024-02-21, https://ai.google.dev/gemma/prohibited_use_policy.
[139] Meta, “Responsible Use Guide: your resource for building responsibly”, 2024-04, https://llama.meta.com/responsible-use-guide/.

方面的**开源工具**[140]，帮助开发者更轻松地自定义Llama 3及其驱动的AI应用体验。具体包括：

- 提升模型透明度。与Llama 2一样，Llama 3发布了一张模型卡，包含有关模型架构、参数和预训练评估的详细信息，和有关模型的功能和限制的信息。
- 正在发布更新后的Llama Guard 2组件，这是一种最先进的安全防护模型，开发者可以将其作为额外层来减少模型生成与既定指导方针不一致的输出的可能性。
- 更新了CyberSecEval，该工具旨在帮助开发者评估由大型语言模型生成的代码可能带来的任何网络安全风险。
- 推出了Code Shield，开发者可以使用它来减少生成潜在不安全代码的机会。
- 分享了Llama Recipes，其中包含开源代码，使开发人员可以更轻松地使用Llama进行构建，完成诸如组织和准备数据集、微调以教导模型执行特定用例、设置安全措施等任务通过RAG系统识别和处理模型生成的潜在有害或不当内容，并部署模型并评估其性能以查看其是否按预期工作。
- 通过像GitHub这样的开源存储库以及Meta长期运行的漏洞赏金计划，直接从开源开发者和研究人员那里接收反馈，帮助Meta更新他们的功能和模型。
- 在与全球合作伙伴合作，创建有利于整个开源社区的行业标准。

	Segment Anything	**Llama**	**Llama 2**	**Llama 3**
数据集	已发布+数据卡 (人脸&证件去标识)	非商用可用	互联网数据+ 委托数据	暂未发布
代码	已发布(Apache许可证)	向研究人员发布	已发布	已发布
模型	已发布(Apache许可证)	向研究人员发布	已发布	已发布
模型卡	-	已发布	已发布	已发布
负责任使用指南	-	已发布	已发布	已发布
研究论文	已发布	已发布	已发布	暂未发布
演示	可公开访问	未发布	由合作伙伴发布	可公开访问

Meta AI项目的发布策略对比：安全/隐私与透明度之间的权衡[141,142]

[140] Meta，“Introducing Meta Llama 3: The most capable openly available LLM to date”，2024-04-18, https://ai.meta.com/blog/meta-llama-3/.
[141] Joelle Pineau，“A culture of open science, in the era of large AI foundation models, 2023-06-18”，https://slideslive.com/39006384.
[142] Meta，“Introducing Meta Llama 3: The most capable openly available LLM to date”，2024-04-18，https://ai.meta.com/blog/meta-llama-3.

3.6 负责任开源之二：在封闭发布中探索实现开源等效收益的替代方案

开发者应考虑开源的替代方案，在可以获得技术和社会效益的同时，又没有太大的风险。Yoshua Bengio认为完全共享的替代方案[143]包括：1）为受信任的研究人员提供结构化访问，以帮助识别安全或道德缺陷；2）独立第三方的审核；3）前沿AI实验室的民主治理：AI的力量掌握在社会手中，而不是少数公司手中；4）财富再分配+AI向善。GovAI则在《开源高性能基础模型》[144]报告中更系统地分析了开源的三方面收益：1）促进外部评测，2）加速有益的进步，3）分散对技术开发和利益的控制，并进一步探讨了在封闭模型中可能有助于实现相同目标的替代方案。

3.6.1 开源作为实现外部模型评估的机制

支持开源AI的论点	• 透明多元促进安全：开源可以使更广泛的开发者社区对项目进行独立的模型评估。利用更广泛的AI社区可以帮助发现错误、偏见和否则可能被忽略的安全问题，最终导致性能更好、更安全的AI产品
收益评估	• 适合评估复杂的安全问题：开源模型允许广泛的研究者和开发者接触并审核代码和算法，这样的集体智慧能更有效地识别和解决复杂的安全挑战 • 识别离散错误的作用较小：离散错误通常是指那些明确、具体的错误，比如某个具体功能的代码错误或是单个数据处理的问题。这类错误通常更依赖于详细的代码审查和专注的测试，而不是开源的广泛参与
替代方案	• 受限下载或研究API：这些方式可以控制谁可以访问模型，从而在不完全公开源代码的情况下，向受信任的第三方提供访问权 • 增量分阶段发布：通过分阶段逐步公开模型，可以逐步观察和调整社会影响和安全问题，有助于在不完全暴露模型的情况下，逐步解决潜在问题 • 红队测试：建立由独立选出的专业人员组成的红队，这个团队可以在模型发布前对其进行压力测试，确保其稳定性和安全性 • 安全赏金计划：通过激励广大公众参与发现和报告新的行为和安全问题，可以增加社会各界对于AI安全性的监督和参与

开源作为实现外部模型评估的机制（基于《开源高性能基础模型》报告修改）

1）增量分阶段发布

模型开发人员可以进行分阶段发布影响测试，以收集有关模型在开源情况下可能如何被使

[143] Yoshua Bengio, “以民主治理管理人工智能风险”, 2023-12-09, https://weibo.com/tv/show/1042163:4977133622067294.
[144] Elizabeth Seger, “Open-Sourcing Highly Capable Foundation Models”, 2023-09-29, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

用和修改的数据。分阶段发布的影响测试，是通过API逐渐发布模型的更大增量版本的过程[145,146]。发布的每个阶段都允许有时间观察模型的使用方式、研究其社会影响，并在下一个更强大的版本发布之前实施补丁或新的安全措施。

如果需要在各个阶段之间实施许多安全措施来减轻危害，这明确表明开源将容易被恶意使用。因为一旦开源，这些措施就可以很容易地被规避。假设模型评估和风险评估过程中没有出现其他重大问题，进行分阶段发布影响测试可以让开发人员更轻松地开源他们的模型。

然而，分阶段发布也是存在成本的，一方面如果不通过监管在整个行业中实施此类流程，单独实施的开发者可能会付出损失市场份额的代价。此外，该模型带来的任何收益到达相关社区的时间也可能推迟。

OpenAI在GPT-2发布之前的做法是完全开源其模型，包括GPT的早期版本。2019年，OpenAI对GPT-2采取了分阶段发布的策略。分阶段发布涉及随着时间的推移逐步发布一系列模型。OpenAI表示分阶段发布GPT-2的目的是让人们有时间评估这些模型的属性，讨论其社会影响，并评估每个阶段后发布的影响。2019年2月，OpenAI发布了小型的124M模型，5月份发布了中型的355M模型，以及随后与合作伙伴和AI社区一起研究该模型的滥用潜力和社会效益之后，8月又发布了7.74亿参数的版本，最终在11月完全开放了GPT-2的访问，包括发布了GPT-2的1.5B参数的最大版本以及代码和模型权重[147]。这一行为标志着OpenAI在模型发布政策上的重大转变，即从全面开源转向了分阶段、有条件的开源。

2）外部审计和红队

除了分阶段发布影响测试之外，开发人员还可以向受信任的第三方审核员授予特殊的模型访问权限。这些外部合作方负责在模型发布之前评估基础模型的安全性和安保性，或评估和验证AI研发机构所实施的模型评测措施。

尽管外部审计尚处于发展的早期阶段，但已被提议作为促进可信AI发展的关键制度机制[148,149]。一个早期的例子是METR(原ARC Evals)与Anthropic和OpenAI建立了公开的合作伙伴关系，并在GPT-4[150]和Claude[151]公开发布前合作进行了测试，以便对具有危险能力的模型提供早期预警，例如GPT-4成功欺骗众包工人。

---

[145] Irene Solaiman et al.，“Release Strategies and the Social Impacts of Language Models”, 2019-08-24, https://arxiv.org/abs/1908.09203.
[146] Toby Shevlane，“The Artefacts of Intelligence: Governing scientists' contribution to AI proliferation”, 2022-04-22, https://cdn.governance.ai/Shevlane_Artefacts_of_Intelligence.pdf.
[147] OpenAI，“GPT-2: 1.5B release”, 2019-11-05, https://openai.com/research/gpt-2-1-5b-release.
[148] Miles Brundage et al.，“Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims”, 2020-04-15, https://arxiv.org/abs/2004.07213.
[149] Jakob Mökander et al.，“Auditing large language models: a three-layered approach”, 2023-06-27, https://arxiv.org/abs/2302.08500.
[150] OpenAI，“GPT-4 System Card”, 2023-03-23, https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[151] MTER，“Update on ARC's recent eval efforts”, 2023-03-17, https://metr.org/blog/2023-03-18-update-on-recent-evals/.

尽管最佳实践仍在开发中，但诸如OpenAI和Anthropic所雇用的红队越来越普遍。例如，例如，OpenAI在DALL·E 2和GPT-4等模型上的红队测试就涉及到了与外部专家的合作，并宣布公开招募OpenAI红队网络[152]，邀请有兴趣提高OpenAI模型安全性的各领域专家合作，严格评测和红队测试其模型。Anthropic则用了超过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[153]，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，并分享了相关发现、教训以及未来计划。展望未来，随着众多政府和研发机构开展外部审计和红队，需要为红队制定共享的规范、实践和技术标准，以确保第三方审计的质量和一致性[154]。

3）漏洞赏金和安全赏金

安全赏金计划，是帮助识别和揭示大型基础模型中的新安全和对齐问题的一种利用更广泛的全球社区的方法。赏金“猎人”不像选定的红队那样经过预先审查。

与网络安全中常用的错误赏金计划类似，安全赏金计划将为发现并负责任地报告新的安全故障的公众提供经济和声誉奖励，例如新颖的越狱方法，或超出内部测试所发现的新能力。与红队一样，赏金“猎人”可以通过与API背后的系统进行交互。然而，目前尚不清楚这在多大程度上阻碍了外部测试人员发现和探测安全问题的能力。

但鉴于传统软件漏洞和AI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。

OpenAI为ChatGPT进行的早期安全赏金试验以有限的宣传和总共1万美元的API奖励[155]，获得了超过1500份提交。尽管OpenAI指出，除了内部红队已经注意到的安全问题之外，提交的内容似乎没有产生什么新发现，但这次演习让OpenAI加深了对最常见攻击路线的了解，并为未来公众参与提供了经验教训[156]。此外，微软也启动了AI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。微软还在与前沿模型论坛合作，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。

[152] OpenAI，“OpenAI Red Teaming Network”，2023-09-19, https://openai.com/blog/red-teaming-network.
[153] Anthropic，“Frontier Threats Red Teaming for AI Safety”，2023-07-26, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.
[154] Deep Ganguli，“Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned”，2022-11-22, https://arxiv.org/abs/2209.07858.
[155] OpenAI，“ChatGPT Feedback Contest: Official Rules”，2022-11-30, https://cdn.openai.com/chatgpt/ChatGPT_Feedback_Contest_Rules.pdf.
[156] Patrick Levermore，“AI Safety Bounties”，2023-08-10, https://rethinkpriorities.org/publications/ai-safety-bounties.

3.6.2 开源作为加速AI进步的机制

支持开源AI的论点	• 增强协作与多样性：开源能够让更多人参与到AI开发中来，实现大规模的协作。这种方式带来了更多的专业知识和多样的视角，以及更多的人力和创造力的投入 • 推动创新与安全研究：多元化的参与可以推动新的和有用的集成创新，同时促进AI安全研究，并推进AI技术的边界
收益评估	• 技术整合的进步：开源最有助于技术整合进步。模型访问使更多人得以调整、创新和优化以集成到新的下游应用程序中。但也因“算法黑箱”，协同开发存在困难 • 技术能力的进步：尽管开源对技术能力的提升有正面影响，但实际效益受到人才、算力和数据资源等因素限制，这些资源是有助于前沿AI能力研究 • 安全研究的进步：学术安全研究往往因无法获得高能力模型而受到限制，开源的收益可能由于缺乏足够的计算基础设施而受到影响
替代方案	• 技术整合的进步：通过使用插件来探索新的应用程序，以及提供受限的访问权限，并实施严格的客户认证（KYC）流程 • 技术能力和安全研究的进步：向特定AI研究小组提供特权模型访问权限，可能是通过结构化的研究API实现的；建立与可信任伙伴的合作关系，并提供受限访问权限；建立多方利益相关者治理机构以确保访问权限公正，并支持独立的学术研究 • 建立激励机制：通过设立大奖计划来激励使用AI实现重要的科学发现或社会进步，如蛋白质折叠、健康和公平应用，以及AI安全方面的突破，如可解释性；承诺将一定比例的利润或研究时间投入AI安全项目

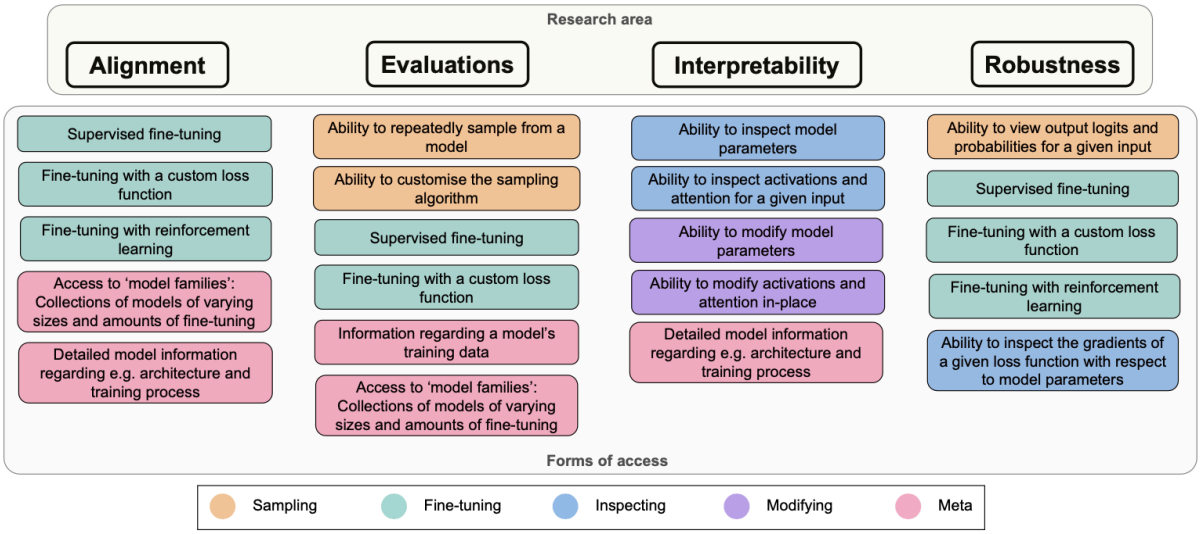
开源作为加速AI进步的机制（基于《开源高性能基础模型》报告修改）

1）向特定的AI研究小组提供特殊的模型访问权限，可能通过结构化的研究API

前沿AI模型的最新发布版本大多封闭，这为如何为外部各方提供足够的模型访问权限以开展重要的安全研究提出挑战。**一个潜在解决方案是使用基于API的“结构化访问”方法，为外部研究人员提供开展研究工作所需的最低访问级别。**

牛津大学和GovAI的研究人员探讨了开展不同形式安全研究需要何种系统访问权限的，根据文献研究和研究员访谈，他们发现模型访问权限不足经常限制了研究议程的选择或实验结论的得出，但所需的访问权限因具体研究领域而有很大差异。例如，评测和基准研究通常通过API从模型中采样就够了；对齐研究则通常要求能够通过微调来修改模型。虽然微调也可以通过API来进行，但当前的接口通常无法提供有关底层模型的足够信息，使他们无法从研究中得

出有意义的结论；可解释性研究进一步要求研究人员可以直接修改模型内部结构，例如学习参数和激活模式，需要完整或接近完整的模型访问权限。



模型访问形式的分解：对于四个研究领域有价值且当前可行项目的必要访问权限[157]

基于研究结果，他们对**"研究API"**的设计提出建议，以便于外部研究人员对封闭前沿模型进行研究和评测。除了当前API中允许从模型中进行广泛采样的功能之外，建议实施以下**四个功能**作为此类服务应提供的核心功能：

- 提高**模型信息**的透明度，例如：清楚地告知正在与哪个模型进行交互、有关模型大小和微调过程的信息，以及有关预训练中使用的数据集的信息。
- 能够查看输出的**logits**，以及选择和修改不同的**采样算法**。
- **版本稳定性和向后兼容性**，以便在更新发布之后也能够对给定模型进行持续研究。
- **微调**给定模型的能力，至少通过监督微调，同时提高微调过程的算法细节的透明度。
- 访问**模型系列**：沿给定维度系统地不同的相关模型的集合，例如参数数量，或者它们是否以及如何进行微调。

2）建立多利益相关者治理机构

即使已愿意向外部合作方提供特殊的模型访问权限，具体决定向哪些参与者提供相应权限以进行外部评估和研究或进行合作，也是一个挑战。当AI研究机构被大量的研究访问请求淹没时，可能会出现对个别团体或内部团体的偏向。AI研究机构还可能会优先考虑那些他们认为会支持其市场利益的外部合作者。

[157] Benjamin Bucknall, Robert Trager, “Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements”, 2023-10-27, https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements/.

一个可能的解决方案可能是建立一个多利益相关者治理机构或系统，以调解研究者对前沿模型的访问。**例如，英国[158]和美国[159]等国家新成立的AI安全研究所正扮演类似的角色。**

3.6.3 开源作为分配AI控制权的机制

支持开源AI的论点	• 影响力分散和多样性增加：开源AI的一个核心优势在于能够通过赋能更小的团体和独立开发者，来分散AI技术的影响力，从而避免技术和市场的单一化。同时，开源模型的可定制性允许用户根据自身需求进行调整，有助于避免在不同应用中出现单一文化的风险
收益评估	• 控制权的分配：开源有助于将技术进步中的控制权分配给开源社区，从而增强社区的参与感和对AI发展的共同责任 • 市场和文化的多元化：开源AI有助于减少下游市场的集中度，更广泛的模型访问渠道降低了开发不同类型基础模型的进入壁垒。这不仅促进了市场内的竞争，也有助于文化的多元化，减少了由单一模型引起的系统性风险 • 上游资源集中问题：尽管开源促进了某种程度的平等和控制权分散，但实际上，对前沿AI系统的控制往往仍然集中在需要大量算力、数据和人才资源的大型实验室手中，开源本身可能难以彻底打破上游市场的集中度 • 开发者的影响力：开源虽然促进了技术的共享，但大型模型的开发者可能仍能通过技术集成，加强对AI生态系统的影响。开源社区成员也可能变成开发者工具和模型的熟练用户，从而增强原开发者的市场地位。
替代方案	• 多元包容磋商过程：提议通过参与式或代表性的磋商过程来指导有关AI的高影响力决策，确保决策过程公开透明并广泛代表不同利益群体 • 制度化民主结构：在大型实验室内部实施民主治理结构，如民选董事会或强制性磋商程序，以减少单一决策者的权力 • 适当的监管干预：需要适当的监管措施来干预开发者的行为，以防止监管机构被特定利益群体操控 • 鼓励市场参与：支持市场多样性的政策来解决上游市场集中的问题。例如推广使用多种基础模型的策略和支持小型实验室及初创企业的政策，可能有助于进一步降低市场和技术的集中度

开源作为分配AI控制权的机制（基于《开源高性能基础模型》报告修改）

[158] UK Government, “AI Safety Institute”, 2024-04-02, https://www.gov.uk/government/organisations/ai-safety-institute.
[159] NIST, “U.S. Artificial Intelligence Safety Institute”, 2024-02-07, https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute.

1）公众参与与审议

AI研发机构和政策制定者可以建立参与性和协商性的程序来指导有关AI复杂问题的决策。例如，Pol.is等参与式平台可用于以低成本征求公众意见并将其综合到有关AI的复杂规范决策中[160]。或者可以召集分层抽样选出的受影响人群的代表性审议来探索AI治理问题[161]。

大型科技公司已开展了相关的探索。例如，Meta已开展了一系列国家和跨国试点项目[162]，以应对“复杂的规范挑战”，并已扩大到近乎全球的审议过程[163]。Anthropic和集体智能项目提出了一种“集体宪法AI”(Collective Constitutional AI)[164]的概念，开展了一项涉及约1000名美国人的公众意见征询流程，即通过广泛的公众参与和反馈，来共同制定和维护一套规则和原则，以探索民主进程如何影响AI的发展。OpenAI也启动了一项“AI的民主投入”的资助计划[165]，以尝试建立民主流程来决定AI系统应在法律范围内遵循哪些规则。

2）组织结构

除了直接征求公众意见来为关键决策提供信息之外，AI研发机构可以引入更具民主性质的组织结构。这些结构将有助于保持内部实践的透明度，并使控制权从单方面的决策者中分散开来，更好地反映利益相关者的利益。

例如，AI研发机构可以作为公益企业(Public Benefit Corporations)注册。作为PBC的注册并不需要公众参与，但它确实为公司做出关于机构结构的决定提供了更清晰的法律立场。传统的公司是为了股东利益最大化而存在的，而公益企业在法律章程中规定了必须将公共利益和其他利益相关方（员工、供应商、社区和环境等）放在和股东利益同等重要的位置，从机制上避免走上传统企业先不惜代价赚钱然后再做慈善的老路。两家领先的AI研发机构Anthropic和Inflection均为公益企业。

此外，科技伦理（审查）委员会机制有助于确保AI研发机构在技术道德和合规性方面的行为符合社会及法规要求。这一委员会可以从伦理、法律和社会影响等多方面对AI技术的研发和部署进行评审，从而确保技术的应用不仅追求经济效益，而且符合伦理规范和社会期待。

另外，也需要董事会或监督委员会对AI研发机构实施有效的治理。OpenAI的“董事会之争”折射出AI企业由单一董事会控制非营利公益组织与营利性子公司时，治理目标可能存在着

[160] Polis, “Input Crowd, Output Meaning”, 2024-04-23(引用日期), https://pol.is/home.
[161] The Collective Intelligence Project, “Alignment Assemblies”, 2024-04-23(引用日期), https://cip.org/alignmentassemblies.
[162] Wired, “Meta Ran a Giant Experiment in Governance. Now It's Turning to AI”, 2023-07-18, https://www.wired.com/story/meta-ran-a-giant-experiment-in-governance-now-its-turning-to-ai/.
[163] Brent Harris, “Improving People's Experiences Through Community Forums”, 2022-11-16, https://about.fb.com/news/2022/11/improving-peoples-experiences-through-community-forums/
[164] Anthropic, “Collective Constitutional AI: Aligning a Language Model with Public Input”, 2023-10-17, https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.
[165] OpenAI, “Democratic inputs to AI”, 2023-05-25, https://openai.com/blog/democratic-inputs-to-ai.

天然的矛盾[166]。

在《AI和灾难性风险》[167]一文中，Yoshua Bengio还讨论了创建“多边研究实验室网络”的想法，该网络专注于开发安全、防御性AI技术。该网络将由独立的非营利实验室组成，主要由政府资助，但其运作不受政府直接控制，以防止权力集中，并确保这些实验室专注于人类福利和民主稳定。

3）多元互动的政府监管

最后，AI研发机构可以鼓励政府监管，限制其行为和独立决策的能力，因为这种监管可能产生重大的社会影响。例如，政府可能需要授权大型基础模型的发布，并成立多利益相关者委员会来协调研究对前沿模型的访问。监管干预措施的制定应针对开发者、开源社区、学术界和民间社会参与的审议过程，以反映不同利益相关者的利益，并防止AI行业的监管捕获。通过这种方式，适当的政府监管可以帮助系统地减少领先的私营研发机构对AI的单方面控制。

4）数据本地化和文化适应

世界各地都需要适应当地市场和文化的大模型，东南亚、中亚、中东等地都出现用开源模型对齐当地文化的大模型项目。要确保基础模型在全球多样化环境中有效并适应当地文化，可以通过在模型开发的关键阶段实施数据本地化和文化适应措施来实现。

在预训练阶段，模型是通过大量的数据来学习语言和世界知识的。此阶段是确保数据多样性和广泛性至关重要的时刻。通过集中收集本地语言和文化数据，包括文本、语音和图像，可以帮助模型更好地理解和反映特定区域的文化特性。此外，确保数据标注团队具备相应的文化背景，可以提高数据标注的准确性和文化相关性。

当模型转向特定任务时，微调阶段的目标是优化模型以提高其在特定领域的表现。在这一阶段，构建专门的数据集来反映当地的具体用途和文化环境尤为重要。例如，利用本地新闻、社交媒体和专业文档等内容，可以帮助模型更精准地服务于当地用户。同时，与本地文化顾问合作，调整和优化模型输出，确保其行为与地方习俗和语境保持一致。

AI模型发布后，持续学习和更新是确保模型与时俱进并有效应对新兴文化趋势的关键。通过动态更新数据集来包括新出现的用语和文化现象，以及建立反馈循环系统让用户直接参与报告和改进模型，可以极大地增强模型的长期适用性和文化敏感度。

通过这些综合策略的实施，有助于构建更公平、包容的AI系统，尤其适合于文化多样性显著的地区。

---

[166] 新华财经，“专家称OpenAI董事会之争根源在于AI企业的治理理念冲突”，2023-11-27，https://bm.cnfic.com.cn/sharing/share/articleDetail/178294371/1.

[167] Yoshua Bengio，“AI and Catastrophic Risk”，2023-09，https://www.journalofdemocracy.org/ai-and-catastrophic-risk/.

3.6.4 开源作为争取商业利益优势的机制

竞争中暂时落后的企业可能会选择开源以取得一些竞争优势。但超过了一定的标准/能力阈值的模型，需要兼顾商业竞争、创新和安全。

支持开源AI的论点	• 品牌建立与技术信誉：开源策略可以迅速提升品牌知名度和技术可信度。通过公开源代码，企业不仅展示了其技术的透明度，还能够吸引更广泛的开发者社区参与，这在技术领域尤其重要 • 市场认可与技术创新：开源还有助于快速迭代产品和推动技术创新，因为它允许企业利用来自全球的智慧和资源。此外，早期的市场认可可以通过社区的正面反馈和改进建议来加速获得 • 行业标准与市场地位：开源可以使企业通过社区的支持推动行业标准的建立，尤其是对于后起之秀，这可以帮助它们在行业中占据一席之地
收益评估	• 资源利用与成本效益：开源使企业能够利用全球开发者的资源，这不仅可以补充自身在技术和资源上的短缺，还能以较低的成本提升产品的市场竞争力和适应性 • 提高知名度和关注度：开源项目通常更容易获得媒体和行业的关注，这对于提高企业的知名度及吸引潜在的客户和投资者极为重要
替代方案	• 分阶段发布：对前沿模型进行分阶段发布，确保每一阶段都具有足够的安全保障措施，这可以减少风险同时促进技术的稳健发展 • 特定访问权限：向特定的AI研究小组提供特权模型访问权限，这可以通过预设的研究API来控制使用方式，既保护了技术，也利于科研合作 • 建立应用生态：通过建立插件和扩展市场，促进生态系统的繁荣，增加产品的可扩展性和自定义能力 • API试用与反馈：开放API试用和提供免费试用版，不仅可以吸引初期用户体验产品，还有助于潜在客户了解产品的价值，收集反馈以优化产品 • 定制演示与沙盒环境：允许潜在客户在沙盒环境中自由探索和测试产品功能，既保护了核心技术不被泄露，又促进了客户对产品的理解和兴趣，尤其适合复杂或新颖的技术产品

开源作为分配AI控制权的机制（本报告自制）

1）DeepMind提供对AlphaFold 2的受限发布

AlphaFold 2是由DeepMind开发的一个革命性的AI系统，它解决了生物学中长期存在的一个巨大挑战——蛋白质折叠问题。AlphaFold 2的算法可以预测蛋白质的三维结构，这一功能对于了解生物学过程和疾病机制具有重大意义。在2020年，该模型在“蛋白质结构预测”领域的国际比赛——CASP中取得了突破性成就，准确率远超过其他方法。

尽管AlphaFold 2的算法细节和数据被DeepMind公开，但DeepMind对其具体的软件实现保持了一定的控制。他们没有将整个系统作为一个开源软件发布，而是选择与科研机构和生物医学领域的企业合作，通过特定的合作协议来使用AlphaFold 2[168]。这种策略既保护了DeepMind的商业利益，也确保了技术的负责任使用，促进了科学研究的进步。

AlphaFold 2的应用潜力巨大，可以帮助科学家更好地理解疾病机理，加速新药的开发，以及在农业和生物技术等领域中创造新的解决方案。DeepMind 的这一步骤被看作是将AI的力量用于社会利益的一个典范。

2）AI21 Labs通过API试用推动Jurassic-1模型的市场化和持续优化

Jurassic-1 是AI21 Labs开发的一款大语言模型，它旨在提供与OpenAI的GPT-3相竞争的性能。Jurassic-1的设计理念是提供更加细腻且多样化的语言理解和生成能力，支持广泛的应用，从自动写作助手到复杂的对话系统。

AI21 Labs推出了Jurassic-1的API服务，允许开发者和企业根据用量支付费用，同时提供了限量的免费试用。这种策略不仅降低了初期用户的试用门槛，还帮助AI21 Labs收集了来自不同使用场景的反馈，这些反馈对于模型的迭代和优化极为重要。通过API，用户可以轻松集成Jurassic-1到他们的产品和服务中，从而测试其性能和适用性。

通过API提供的服务，AI21 Labs能够建立一个有效的用户反馈循环。他们利用这些反馈来持续优化模型，解决在真实世界应用中遇到的具体问题，比如偏见减少和生成控制的提高。这样的反馈机制确保了模型在市场上的竞争力和适应力。

3）需要警惕不符合传统开源许可证要求的“开源”

在不符合传统开源许可证要求的情况下，使用“开源”或“开放”术语也可能带来经济、战略和声誉方面的收益。以Llama 2模型为例，尽管Llama 2模型被Meta称为开源模型，但其许可证中包含了一些限制性条款，例如不允许月活跃用户超过7亿的开发者将其用于商业目的，且其输出不能用于训练其他大模型。这种使用“开源”术语的策略可能为Meta带来了战略优势，因为它可以在一定程度上控制模型的使用和分发，同时避免了完全放开可能带来的竞争风险。

与之类似，目前的开源AI由于缺少严格定义，通常存在以下考量：

- **市场营销和品牌推广：**通过将模型标记为“开源”，即使它们并不符合传统的开源定义，公司可以在市场上获得更多的关注和认可。这种做法可能会吸引开发者和研究者的注意，从而增加公司的知名度和声誉。

[168] Google Deepmind，“How our principles helped define AlphaFold's release”，2022-09-14, https://deepmind.google/discover/blog/how-our-principles-helped-define-alphafolds-release/.

- **吸引开源社区：**即使某些模型的使用受到限制，将它们标记为“开源”仍可能吸引开源社区的兴趣，因为他们可以下载和使用这些模型进行研究和实验。这有助于公司建立一个围绕其技术的开发者生态系统，从而为其产品和服务的开发和改进提供支持。
- **潜在的商业机会：**通过开放模型（即使是在有限的许可证下），公司可能能够识别和开发新的商业机会。例如，它们可以提供额外的付费服务或工具，以帮助开发者在其平台上更有效地使用这些模型。
- **规避开源许可证的法律风险：**通过使用自定义许可证，公司可以避免开源许可证可能带来的法律风险，例如GPL许可证下的“传染性”要求，这可能要求任何基于开源代码的衍生作品也必须以开源形式发布。

“开放清洗”(Openwashing)这一术语就是是用来形容那一些公司为了市场营销目的，表面上打着开源和开放许可的旗号，实际上却继续实行封闭做法的行为[169]。

3.7 小结

在当今AI领域，将AI模型简单地划分为开源或闭源是一种过于简化的做法。开源AI的概念尚未得到清晰定义，与开源软件不同，AI模型的“源代码”可能包括多种组件，如推理代码、训练代码、模型权重和训练数据，这些组件的开放程度可以各异。此外，从“完全开放”到“完全封闭”的发布选项实际上是多样的，需要明确的标准和定义来权衡透明性、安全性和商业考量。

在安全和治理方面，研究发现无论是开源还是闭源模型，都存在对特定攻击的脆弱性，而AI研发机构在安全政策和透明度方面的表现各异。倾向于开放模型的机构在推动透明度和外部评估方面表现较好，而倾向于封闭模型的机构则在安全政策实施方面更为优秀。

为了推动负责任开源的实践，一方面需要促进开放模型从构建到使用的全流程负责任，建议对于不同能力级别的AI模型，应有差异化的开源要求。另一方面，需要探索在封闭模型中实现既可以获得开源的益处，又没有太大的风险的替代方案，如增量分阶段发布、结构化访问和研究API、独立第三方审核、数据本地化和文化适应等。

开源AI的负责任实践并非一成不变，而是会随着技术发展和社会需求的变化而不断演进。我们可以预见，未来开源与闭源的讨论将更加深入和细化，可能会出现更多创新的发布模式和治理机制，以适应不断变化的环境和挑战。在这个过程中，各方面的合作和对话将至关重要，以确保AI技术的健康发展和广泛应用。

[169] Klint Finley, “How to Spot Openwashing”, 2011-02-03, https://readwrite.com/how_to_spot_openwashing/.

4 对推动基础模型负责任开源的建议

不论开源还是闭源，只要有心作恶，总会找到途径。现有的各种技术，如果被滥用，都可能对人类社会造成破坏，生物技术就是个例子，它同样存在被滥用的风险。因此，更重要的是如何建立一套体系，来防范和制止任何个人或组织滥用技术危害社会。

——薛澜[170]

在报告整体讨论的基础上，结合我国AI的技术、产业、开源社区的发展现状，我们建议：

4.1 基础模型研发机构

1）根据模型的能力和潜在影响，实施分层管理策略，并建立相应的风险缓解措施

- 对于风险较低的模型，鼓励进行模型开源并协助开源社区建设。
- 对于高风险/前沿AI模型，模型开源的决策应进行严格的风险评估和伦理审查。
    - 除了直接滥用和危险能力外，还应考虑模型微调或修改可引发的滥用。
    - 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。
    - 加强对开放基础模型的边际风险的评估和研究，应采用风险评估框架来明确阐明公开发布基础模型的边际风险。
- 引入动态风险评估框架，按模型性能和潜在风险分级，定期更新评估准则。

2）促进开放发布从构建到使用的全流程负责任实践

- 根据基础模型的生命周期和流程阶段，分别设计构建和使用阶段的负责任开源维度，并针对不同能力级别的模型制定差异化的负责任开源要求
    - 对于大多数的AI模型，负责任的实践主要通过提高透明度、确保合规以及促进创新来实现。
    - 对于能力更强的前沿模型，需要实施与模型的潜在风险相称的评估和安全缓解措施。
- 开放基础模型的研发机构应明确已实施了哪些负责任AI实践，以及哪些负责任AI实践留给可能修改模型以在面向消费者的应用程序中使用的下游开发者。
- 开放基础模型应清晰的说明其具体开放程度，基于安全/隐私与透明度之间的权衡，适当发布数据集、代码、模型、模型卡、负责任使用指南等。

[170] 薛澜等，“中国在这一波人工智能浪潮中处于什么位置？”, 2024-03-26, https://mp.weixin.qq.com/s/ovBVf8ortxfMoIEW8A23cw.

- 更多全流程负责任实践和案例，可参考《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》[171]。

3）在封闭发布中探索实现开源等效收益的替代方案

- 建议设计并实施结构化的研究API。根据研究人员、红队和审核机构的访问需求，分别授予其更灵活的微调权限，促进安全研究进步。
- 其他替代方案还包括增量分阶段发布、独立第三方审核、数据本地化和文化适应等。

4）建设科学系统的安全性评测体系

- 与学术界和产业界伙伴合作，研究和推广AI安全性的评测科学(Science of Evals)[172]，促进开源大模型的安全评估方法的研究，以填补当前的成熟评测体系的空白，提升评测的科学性和可靠性。

4.2 AI开源社区

1）开发者、标准制定机构和开源社区应多方协作，推动制定和推广开源AI标准

- 鼓励定义模型组件发布的细粒度标准，标准应基于对发布不同组件特定组合所带来的风险的理解，积极参与OSI对开源AI进行明确定义的议程[173]。
- 鼓励制定新的适用于开源AI协议（类似于MIT、Apache许可证），对模型的传播、扩散作出规定。

2）鼓励分享促进科学研究和安全研究的数据集、基准和组件

- 鼓励分享更多适用于各种语言和多模态的安全提示集和基准。
- 鼓励分享更多能够促进安全研究的组件和资源，例如安全对齐数据集和奖励模型。

3）鼓励社区进行模型的安全性和公平性评测，通过集体智慧提高模型的整体质量

- 支持社区对模型进行独立的安全评测，以及促进开放的讨论和协作。
- 鼓励与跨领域专家进行持续的红队演练活动，例如针对大模型安全的全球最大规模AI黑客大赛DEF CON 31[174]。

[171] 安远AI，“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”，2024-01-17, https://mp.weixin.qq.com/s/9c0whoQpxgxggS7rOj9Sng.
[172] Apollo Research，“We need a Science of Evals”， 2024-01-22, https://www.apolloresearch.ai/blog/we-need-a-science-of-evals.
[173] Open Source Initiative，“Join The Discussion on Open Source AI”，2024-04-23(引用日期), https://opensource.org/deepdive.
[174] 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

4.3 AI治理、政策和立法专家

1）建立敏捷治理体系，防范技术滥用危害社会的同时，鼓励负责任开源创新

- 对于风险较低的模型，制定政策鼓励开源和开源社区建设，例如规定专门的税收优惠和研发补贴办法、制订专门的合规指引、明确责任减免规则[175]、设立专项基金和奖项认证等，同时明确开源过程中需要遵循的法律和伦理标准。
- 对于高风险/前沿AI模型，确立严格的风险评估和伦理审查程序。政策法规应明确，只有在通过这些审查后，模型才能被考虑用于开源或广泛部署。
- 引入动态风险评估框架，按模型性能和潜在风险分级，定期更新评估准则。

2）对超过一定阈值的所有AI模型和系统建立注册和许可制度

- 为超过阈值的AI模型和系统进行追溯和持续注册，并建立风险评估、风险缓解和独立审计程序。
- 标准可能包括但不限于：浮点运算次数、参数量、模型训练费用或危险能力等[176]。
- 对达到发布前评测和合规标准的AI模型和系统颁发许可证。

3）实施明确的责任制度

- 要求基础模型研发机构进行模型开源和部署前的风险评估，引导负责任发布，推动行业自我监督。
- 应明确开放基础模型研发机构与下游开发者和用户之间的责任划分。
- 基础模型研发机构与下游开发者应对其系统造成的“合理可预见的滥用”承担法律责任[177]。
- 强制性事件披露，当开发人员获悉其基础模型中的漏洞或故障时，必须依法要求其向指定的政府机构报告。

4）由政府牵头成立专门的AI安全研究机构

- 集合AI、社会学、管理学等多学科专家，共同监督、审查和研究现有的AI系统。
- 机构的主要职责包括：
    - 开发和实施AI安全研究，建立AI安全评估体系，创新治理技术，包括预测和应对高风险AI技术的方法。

[175] 冯恋阁，“《人工智能示范法2.0（专家建议稿）》重磅发布 重视AI开源发展、构建知识产权创新规则”，2024-04-16, https://m.21jingji.com/article/20240416/herald/4df710ffed0ffe037cdf6c54aa369961.html.
[176] David Harris，“How to Regulate Unsecured 'Open-Source' AI: No Exemptions”，2023-12-04, https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.
[177] European Parliament，“Artificial Intelligence Act”，2024-04-23(引用日期)，https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

- 支持和推进国内负责任的AI开源社区和生态系统建设。
- 与国家监管和标准机构合作，分享研究成果，支持AI治理和法规制定。
- 加强与国际AI安全研究机构的合作，促进技术交流和知识共享，提升我国在AI安全领域国际影响力。

4.4 AI投资方和资助方

1）AI开发者和政府资助者：在AI安全和负责任开源研究上投入更多的资源

- 多位图灵奖的主和顶尖专家呼吁AI开发者和政府资助者至少将他们AI研发预算的三分之一投入到安全领域[178,179,180]。
- 鼓励支持开放基础模型风险进行研究的项目。

2）AI投资方和其他资金方：投资于评估开放性对创新和社会的影响

- 鼓励投资于更系统地衡量不同程度的开放性对基础模型收益的影响，以及AI监管对开放基础模型创新生态的潜在影响，为政策制定和资金投入提供指导。
- 鼓励投资于那些能够平衡技术创新和社会价值的研发项目。

4.5 负责任开源的国际合作

1）通过负责任开源，助力发展中国家提升AI技术和治理能力，不断弥合智能鸿沟和治理能力差距

- 通过国际合作加强基础设施建设和技术研发，创建公共AI基础设施和安全的AI模型，提供平等的技术访问机会，不断弥合智能鸿沟和治理能力差距[181,182]。
- 支持高质量数据集的收集和本地参与，提升本地模型的性能和跨文化价值对齐[183,184]。
- 促进AI伦理、安全和治理方面的国际合作，共享最佳实践，共同应对全球挑战。

---

[178] 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.
[179] IDAIS，“International Dialogues on AI Safety”，2024-03-11, https://idais.ai/.
[180] Yoshua Bengio et al.，“北京AI安全国际共识”，2024-03-11, https://idais-beijing.baai.ac.cn/.
[181] 中国网信网，“2024年中非互联网发展与合作论坛关于中非人工智能合作的主席声明”，2024-04-03, https://www.cac.gov.cn/2024-04/03/c_1713731793084792.htm
[182] 中央网络安全和信息化办公室，“全球人工智能治理倡议”，2023-10-18, https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.
[183] AI Singapore，“Southeast Asian Languages In One Network Data(SEALD)”，2024-03-11, ”https://aisingapore.org/aiproducts/southeast-asian-languages-in-one-network-data-seald/
[184] 鹏城实验室，“鹏城・脑海”通用人工智能大模型创新之路正式启程”，2023-09-21, https://mp.weixin.qq.com/s/QE3Mq-dkS9OlCHWo7dxQkQ

2）对于高风险模型（包括开源和闭源）的风险评估和伦理审查，在国际上推动形成具有广泛共识的治理框架和标准规范。

- **制定统一的风险评估标准：**包括对高风险模型可能造成的社会、经济、政治、环境等方面的全面评估。标准应详细说明评估的方法、工具和指标。
- **制定一套标准的伦理审查流程：**包括对高风险模型的透明度、公正性、隐私保护和用户权益等的保障，流程应确保所有高风险模型在发布前都经过严格的伦理审查。
- **鼓励国际合作与信息共享：**包括共享风险评估结果、伦理审查的案例和监管的最佳实践。通过建立一个公开的数据库来促进信息共享和透明度。

最后，我们希望再次强调开源与闭源不应简单视为二元对立，其间存在多种模型发布选项和政策设计空间。我们支持国内积极促进基础模型开源及其生态发展的政策取向，同时鉴于AI技术的迅猛发展及其潜在风险，也迫切需要制定更具约束力的法规来确保技术的安全使用。这不仅要在国内形成广泛的共识，更需通过有效的治理策略，确保发展与安全之间的平衡。我们倡导在全球范围内展开合作，建立一个既促进基础模型开源技术创新，又能有效管理风险的国际框架。