先进人工智能安全国际科学报告

中期报告

2024年5月

AI SEOUL SUMMIT
21 - 22 MAY 2024
Hosted by the Republic of Korea and the United Kingdom

贡献者

椅子

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

专家咨询小组

Prof. Bronwyn Fox， 联邦科学与工业研究组织 (CSIRO) (澳大利亚)

andr é Carlos Ponce de Leon Ferreira de Carvalho ，圣保罗大学数学和计算机科学研究所 (巴西)

Dr. Mona Nemer， 加拿大首席科学顾问 (加拿大)

Raquel Pezoa Riv era，Federico Santa 3月 í a技术大学 (智利)

曾毅博士， 中国科学院空间研究所 (中国)

Juha heikkil ä， 连接DG (欧洲联盟)

Guillaume Avrin， 企业总局 (法国)

Antonio kr ü ger， 德国人工智能研究中心 (德国)

教授Balaraman Ravindran， 印度理工学院，马德拉斯 (印度)

Prof. Hammam Riza， KORIKA (印度尼西亚)

Dr. Ciar á n Seoighe， 爱尔兰科学基金会 (爱尔兰)

Dr. Ziv Ka tzir，以色列创新局 (以色列)

Dr. Andrea Monti， 基耶蒂-佩斯卡拉大学 (意大利)

Dr. Hiroaki Kitano， 日本索尼集团

Mary Kerema， 信息通信技术和数字经济部 (肯尼亚)

何塞·拉蒙·洛佩斯·波蒂略博士， 元素Q (墨西哥)

Prof. Haroon Sheikh， 荷兰政府政策科学委员会 (荷兰)

Dr. Gill Jolly， 商业、创新和就业部 (新西兰)

Dr. Olubunmi Ajala， ninnovation and Digital Economy (尼日利亚)

Dominic Ligot， CirroLytix (菲律宾)

教授Kyoung Mu Lee， 首尔国立大学电气与计算机工程系 (大韩民国)

Ahmet Halit hadip， 土耳其工业和技术部 (土耳其共和国)

Crystal Rugege， 人工智能和创新政策国家中心 (卢旺达)

Dr.Fahed Albalawi， 沙特数据和人工智能管理局 (沙特阿拉伯王国)

Denise Wong， 信息通信媒体发展管理局 (IMDA) 数据创新和保护小组 (新加坡)

dr. Nuria Oliver， ELLIS Alicante (西班牙)

Dr. Christian Busch， 瑞士联邦经济事务、教育和研究部

Oleksii Molchanovskyi， 乌克兰人工智能发展专家委员会 (乌克兰)

Marwan Alserkal， 内阁事务部，总理办公室 (阿拉伯联合酋长国)

Saif M. Khan， 美国.商务部 (美国)

Dame Angela McLean， 英国政府首席科学顾问

Amandeep Gill， 联合国技术特使 (联合国)

写作小组

Daniel Privitera (首席作家)，基拉中心
Tamay Besiroglu，A时代I
Rishi Bommasani，斯坦福大学马萨诸塞州
Stephen Casper， Ins技术研究所
Yejin Choi， 华盛顿大学/A12卡内基梅隆大学
Hoda Heidari，Mila-魁北克人工智能研究所**Hoda Heidari**，
Leila Khalatbari， 香港科技大学
Shayne Longpre， 麻省理工学院
Vasilios Mavroudis， 伊利诺伊大学香槟分校**Mantas Mazeika**， 艾伦图灵研究所
Kwan Yee Ng, Concordia AI
Chinasa T. Okolo，**博士**， 布鲁金斯学会
Deborah Raji, Mozilla
Theodora Skeadas， 《人文情报》
弗洛里安·特拉梅尔， 苏黎世联邦理工学院

科学协调员

Sören Mindermann, Mila - Quebec AI Institu te

高级顾问

Bayo Adekanmbi， 尼日利亚数据科学
Paul Christiano， 在美国人工智能安全研究所担任高级顾问之前
David Dalrymple， 研究 + 先进发明机构 (ARIA)
俄勒冈州立大学**Thomas G. Dietterich**，
Edward Felten， 普林斯顿大学
香港科技大学**Pascale Fung**在担任Meta职位之前曾担任高级顾问
Pierre-Olivier Gourinchas, International Monetary Fund (IMF)
Nick Jennings CB FREng FRS， 拉夫堡大学
Andreas Krause， 苏黎世联邦理工学院
Percy Liang， 伯南布哥联邦大学斯坦福大学**Teresa Ludermir**，
Vidushi Marda, REAL ML
Helen Margetts OBE FBA， 牛津大学/艾伦·图灵研究所
John A. McDermid OBE FREng， 约克大学
普林斯顿大学**Arvind Narayanan**， **Alondra Nelson**， KAIST计算学院高级研究学院**Alice Oh**，
Gopal Ramchurn， 英国RAI/UKRI TAS Hub/南安普敦大学**Stuart Russell**， 加利福尼亚大学，
伯克利
斯坦福大学**Marietje Schaake**，
Dawn Song， 加州大学伯克利分校**Alv aro Soto**，智利天主教大学
Lee Tiedrich， 杜克大学
ga ë lvaroquaux， 国家数字科学与技术研究所 (Inria)
姚明， 清华大学跨学科信息科学研究所**张亚勤**， 清华大学

秘书处

由AI安全研究所主办的**英国政府秘书处**
Benjamin Prud'homme，-魁北克AI Institut e

致谢

秘书处感谢以下英国组织的有益支持，评论和反馈: Ada Lovelace研究所，Alan Turing研究所，长期复原力中心，人工智能治理中心和英国人工智能安全研究所。还要特别感谢丹·亨德里克斯、迪伦·哈德菲尔德·梅内尔和帕梅拉·萨缪尔森。

有关本刊物的任何查询，请发送给我们:
secretariat.AIStateofScience@dsit.gov.英国

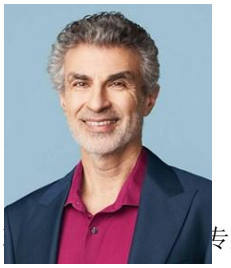
免责声明

本报告不代表主席、编写或咨询小组的任何特定个人，也不代表支持其发展的任何政府的观点。本报告是对高级AI能力和风险的现有研究的综合。报告主席对此负有最终责任，并自始至终监督其发展。

研究系列编号: DSIT 2024/009

前言

这份报告是人工智能安全之旅的开始



，我很荣幸主持发布首届《高级人工智能安全国际科学报告》。我很自豪地发布这份中期报告，这是自2023年11月布莱奇利公园人工智能安全峰会委托这项工作以来的六个月里，许多专家付出了巨大努力的结果。

我们知道先进的人工智能正在迅速发展，而且这些先进的人工智能系统如何影响我们未来的生活和工作方式存在很大的不确定性。人工智能有巨大的潜力让我们的生活变得更好，但它也带来了伤害的风险。这就是为什么要进行彻底的分析

专家意见至关重要。我们知道的越多，我们就越有能力塑造我们的集体命运。

我们的使命很明确: 推动对高级人工智能安全性的共享、基于科学的、最新的理解，并随着时间的推移继续发展这种理解。该报告正确地强调了专家之间存在共识的领域，以及对高级人工智能的能力和风险的分歧，特别是那些预计在未来开发的。为了有效地履行我们的使命，我们的目标是解决知识分子诚实的专家社区之间的分歧。通过剖析这些差异，我们为明智的决策铺平了道路，并刺激了有助于消除迷雾和减轻风险所需的研究。

我感谢我们的国际专家咨询小组的宝贵意见，这些意见最初确定了报告的范围，后来又对整个草案提供了反馈。他们的不同观点和认真审查扩大并加强了这份临时报告。同样值得认可的是我敬业的作家和高级顾问团队。他们在过去几个月的承诺创造了一个超出我预期的临时产品。我还要感谢英国政府启动这一进程并提供出色的运营支持。对我来说，同样重要的是，英国政府同意撰写这份报告的科学家应该拥有完全的独立性。

这份中期报告只是旅程的开始。毫无疑问，这份报告在第一次尝试中未能捕捉到的观点和证据。在这样的科学过程中，反馈是宝贵的。我们将纳入更多的证据和科学的观点，因为我们对最终版本的工作。

Yoshua Bengio教授

蒙特利尔大学/Mila-魁北克人工智能研究所和主席

AI安全是一个共同的全球问题

我很高兴向您介绍有关高级AI安全性的第一份国际科学报告的临时更新，这是2023年11月在布莱奇利公园举行的开创性AI安全峰会的重要成果。这份具有里程碑意义的报告代表了全球前所未有的努力，以建立对人工智能快速发展所带来的机遇和风险的共同的，基于科学的理解，并证明了 “布莱奇利效应”-召集才华横溢的头脑来应对人类最大的挑战之一。

人工智能造福人类的巨大潜力，需要积极努力，以确保安全、负责任地开发和部署这些强大的技术。没有一个国家能够独自应对这一挑战。这就是为什么我如此热衷于将一群世界领先的专家聚集在一起，贡献他们的知识和观点。我要特别感谢Yoshua Bengio教授作为主席在巧妙地指导这一复杂的国际努力方面发挥的领导作用。

至关重要的是，该报告还揭示了我们当前知识中的重大差距以及迫切需要进一步研究和讨论的关键不确定性和辩论。我真诚地希望，这份报告及其背后的合作进程能够成为缩小关键知识差距所需的研究和政策努力的催化剂，并为未来具有挑战性的政策选择提供宝贵的投入。

我们还有很多东西要学，但这份报告标志着一个重要的开始。英国期待继续与国际伙伴合作，促进负责任的、以人为本的人工智能发展方法 -- 利用这些强大的工具来改善生活和生计，同时警惕地防范下行风险和伤害。我们可以共同努力，建设一个全人类都能从人工智能的奇迹中受益的未来。

，Rt Hon Michelle Donelan议员， 科学，创新和技术部国务卿

向前迈出的关键一步，呼吁AI安全采取行动

人工智能的快速发展将以深刻和不可预见的方式重塑我们的世界。从革命性的医疗保健和运输到自动化复杂任务和解锁科学突破，人工智能的积极影响潜力是不可否认的。

然而，除了这些显著的可能性之外，还存在着重大的挑战，需要采取前瞻性的方法。关注的范围从嵌入算法中的意外偏见到自治系统超过

主风险凸显了迫切需要进行全球对话，以确保人工智能的安全和负责任的发展。

在这种情况下，国际人工智能安全报告将为全球合作提供重要的基础。该报告汇集了来自30个国家、欧盟和联合国的专家的知识，提供了对人工智能安全性的全面分析。通过关注对通用人工智能能力和风险的早期科学理解，并评估评估和缓解这些风险的技术方法，该报告将引发多方利益相关者之间的持续对话和合作。

我希望基于这份报告，来自30个国家、欧盟和联合国的专家继续进行平衡的讨论，实现可接受的、适合发达国家和发展中国家具体情况的人工智能风险缓解，从而创造一个创新和负责任的人工智能和谐共存的未来。

Lee jong-ho， 大韩民国MSIT部长

执行摘要

关于本报告

- 这是第一份 “关于高级ai安全性的国际科学报告” 的临时出版物。由75位人工智能 (AI) 专家组成的多元化小组为本报告做出了贡献，其中包括由30个国家，欧盟 (EU) 和联合国 (UN) 提名的国际专家咨询小组。
- 在本报告主席的领导下，撰写本报告的独立专家集体对其内容拥有完全的酌处权。
- 在人工智能发展取得前所未有的进展之际，这份第一份出版物将其重点限制在近年来发展特别迅速的一种人工智能上: 通用人工智能，即可以执行各种任务的人工智能。在快速发展的过程中，通用人工智能的研究目前正处于科学发现的时代，尚未成为科学定论。
- 世界各地的人们只有在风险得到适当管理的情况下，才能安全地享受通用人工智能的许多潜在好处。本报告着重于识别这些风险，并评估评估和减轻这些风险的技术方法。它的目的不是全面评估通用人工智能的所有可能的社会影响，包括其许多潜在的好处。
- 这份中期报告有史以来第一次汇集了30个国家、欧盟和联合国提名的专家以及其他世界领先的专家，为通用人工智能安全的讨论和决策提供了一个共享的科学、循证基础。我们仍然在围绕通用人工智能能力、风险和风险缓解的几个问题上存在分歧，无论是次要的还是主要的。但我们认为这个项目对于提高我们对这项技术及其潜在风险的集体理解，以及更接近达成共识和有效的风险缓解至关重要，以确保人们能够安全地体验通用人工智能的潜在好处。赌注很高。我们期待着继续这一努力。

执行摘要的要点

- 如果管理得当，通用人工智能可以用于促进公共利益，可能会带来更好的福祉，更多的繁荣和新的科学发现。然而，通用人工智能出现故障或被恶意使用也可能造成伤害，例如，在高风险环境中做出有偏见的决定，或者通过诈骗、虚假媒体或侵犯隐私。
- 随着通用人工智能能力的不断发展，可能会出现诸如大规模劳动力市场影响，人工智能黑客攻击或生物攻击以及社会失去对通用人工智能的控制等风险，尽管研究人员对这些情况的可能性存在争议。对这些风险的不同看法往往源于对社会将采取的限制措施、这些措施的有效性以及通用人工智能能力的推进速度的不同期望。
- 通用人工智能能力的未来进展速度存在相当大的不确定性。一些专家认为，到目前为止，进展最有可能放缓，而另一些专家则认为，极快的进展是可能的或可能的。
- 开发人员可以采用各种技术方法来评估和降低通用人工智能的风险，监管机构也可以要求，但它们都有局限性。例如，用于解释为什么通用AI模型产生任何给定输出的当前技术受到严重限制。

- 通用人工智能技术的未来是不确定的，即使在不久的将来，也可能出现各种各样的轨迹，包括非常积极和非常消极的结果。但关于AI的未来，没有什么是不可避免的。社会和政府的决定将决定人工智能的未来。这份临时报告旨在促进对这些决定的建设性讨论。

这份报告综合了对通用人工智能的科学理解 -- 人工智能可以执行各种各样的任务 -- 重点是理解和管理其风险。

使用人工智能的系统的能力一直在迅速发展。这凸显了人工智能为商业、研究、政府和私人生活创造的许多机会。它还提高了人们对与先进人工智能相关的当前危害和未来潜在风险的认识。

关于高级人工智能安全的国际科学报告的目的是朝着对人工智能风险以及如何减轻风险的国际共识迈出一步。该报告的第一份临时出版物将其重点限制在一种能力发展特别迅速的人工智能上: 通用人工智能，即可以执行各种任务的人工智能。

在快速发展的过程中，通用人工智能的研究目前正处于科学发现的时代，尚未成为科学定论。该报告概述了当前对通用人工智能及其风险的科学理解。这包括确定科学共识的领域以及存在不同观点或开放研究问题的领域。

世界各地的人们只有在风险得到适当管理的情况下，才能安全地享受通用人工智能的潜在好处。本报告的重点是识别通用人工智能的风险，并评估评估和缓解这些风险的技术方法，包括使用通用人工智能来缓解风险。它的目的不是全面评估通用人工智能的所有可能的社会影响，包括它可能提供的好处。

根据许多指标，通用AI能力近年来增长迅速，并且在如何预测未来进展方面没有达成共识，使得各种场景出现可能

根据许多指标，通用AI能力正在迅速发展。五年前，领先的通用人工智能语言模型很少能产生连贯的文本段落。今天，一些通用的人工智能模型可以在广泛的主题上进行多轮对话，编写简短的计算机程序，或者从描述中生成视频。然而，通用人工智能的能力很难可靠地估计和精确定义。

通用人工智能的发展速度取决于技术进步的速度和监管环境。本报告侧重于技术方面，不讨论监管工作如何影响通用人工智能的开发和部署速度。

近年来，人工智能开发人员迅速提高了通用人工智能功能，主要是通过不断增加用于训练新模型 (一种称为 “扩展” 的趋势) 和改进现有算法的资源。例如，最先进的人工智能模型用于训练的计算资源 (“计算”) 每年增加约4倍，训练数据集大小增加2.5倍，算法效率 (相对于计算的性能) 增加1.5倍。“缩放” 是否导致了诸如因果推理等基本挑战的进展，研究人员之间存在争议。

通用人工智能能力的未来进展速度对管理新兴风险具有重大影响，但专家们对即使在不久的将来也会发生什么持不同意见。专家们以各种方式支持通用人工智能能力缓慢、快速或极快发展的可能性。这种分歧涉及一个关键问题: 继续 “扩展” 资源和改进现有技术是否足以产生快速进展并解决可靠性和事实准确性等问题，还是需要新的研究突破来大幅提高通用AI能力？

几家开发通用人工智能的领先公司正在押注 “扩展” 以继续带来性能改进。如果最近的趋势继续下去，到2026年年底，一些通用人工智能模型将使用比2023年发布的最计算密集型模型多40倍至100倍的计算进行训练，并结合使用这种计算效率提高3倍至20倍的训练方法。然而，进一步增加数据和计算存在潜在的瓶颈，包括数据的可用性、人工智能芯片、资本支出和本地能源容量。开发通用人工智能的公司正在努力解决这些潜在的瓶颈。

一些研究工作旨在更可靠地理解和评估通用AI，但我们对通用AI模型和系统如何工作的总体理解是有限的

管理通用人工智能风险的方法通常基于这样的假设，即人工智能开发人员和政策制定者可以评估通用人工智能模型和系统的能力和潜在影响。但是，虽然技术方法可以帮助评估，但所有现有方法都有局限性，无法提供强有力的保证，以防止与通用人工智能相关的大多数危害。
总体而言，对通用人工智能的内部运作、能力和社会影响的科学理解非常有限，专家普遍认为，提高我们对通用人工智能的理解应该是当务之急。一些关键挑战包括:

- 开发人员仍然对他们的通用AI模型如何运行知之甚少。这是因为通用AI模型不是传统意义上的编程。相反，他们是经过训练的: 人工智能开发人员建立了一个涉及大量数据的训练过程，这个训练过程的结果就是通用人工智能模型。这些模型可以由数万亿个称为参数的组件组成，并且它们的大部分内部工作都是难以理解的，包括对模型开发人员来说。模型解释和可解释性技术可以提高研究人员和开发人员对通用AI模型如何运行的理解，但这项研究还处于起步阶段。
- 通用AI主要通过在各种输入上测试模型或系统来评估。这些抽查有助于评估优势和劣势，包括漏洞和潜在的有害能力，但不提供定量的安全保证。测试通常会忽略危险，高估或低估功能，因为通用AI系统在不同情况下，不同用户或对其组件进行其他调整时可能会表现不同。
- 原则上，独立参与者可以审核公司开发的通用AI模型或系统。但是，公司通常不向独立审计师提供必要的直接访问模型或有关严格评估所需的数据和方法的信息。一些政府正在开始建设进行技术评估和审计的能力。
- 很难评估通用人工智能系统的下游社会影响，因为对风险评估的研究还不足以产生严格和全面的评估方法。此外，通用人工智能具有广泛的用例，这些用例通常不是预定义的，只是受到轻微的限制，使风险评估进一步复杂化。了解通用人工智能模型和系统的潜在下游社会影响需要细致入微的多学科分析。增加多样化的代表性

通用人工智能开发和评估过程中的观点是一项持续的技术和制度挑战。

通用人工智能可能对个人和公共安全和福祉构成严重风险

该报告将通用AI风险分为三类: 恶意使用风险，故障风险和系统性风险。它还讨论了导致许多风险的几个交叉因素。

恶意使用。与所有强大的技术一样，通用AI系统也可能被恶意使用以造成伤害。可能的恶意使用类型包括相对证据充分的类型，例如通用AI实现的诈骗，以及一些专家认为未来几年可能发生的类型，例如恶意使用通用AI的科学功能。

- 通过通用AI生成的虚假内容对个人造成的伤害是一种相对有据可查的通用AI恶意使用。通用AI可用于增加诈骗和欺诈的规模和复杂性，例如通过通用AI增强的“网络钓鱼”攻击。通用人工智能也可以用来生成虚假的妥协内容，包括未经个人同意的个人，例如未经同意的deepfake色情内容。
- 另一个令人担忧的领域是恶意使用通用人工智能来提供信息和操纵公众舆论。通用人工智能和其他现代技术使生成和传播错误信息变得更加容易，包括影响政治进程。像水印内容这样的技术对策虽然有用，但通常可以被适度复杂的参与者规避。
- 通用人工智能也可能被恶意用于网络犯罪，提升个人的网络专业知识，并使恶意用户更容易进行有效的网络攻击。
  通用AI系统可用于扩展和部分自动化某些类型的网络操作，例如社交工程攻击。但是，通用AI也可以用于网络防御。总体而言，尚无任何实质性证据表明通用AI可以自动执行复杂的网络安全任务。
- 一些专家还对通用人工智能可能被用来支持生物武器等武器的开发和恶意使用表示担忧。没有强有力的证据表明目前的通用人工智能系统会带来这种风险。例如，尽管目前的通用人工智能系统显示出与生物学相关的不断增长的能力，但有限的研究并没有提供明确的证据表明，目前的系统可以比使用互联网更容易地“提升”恶意行为者来获取生物病原体。然而，未来的大规模威胁几乎没有得到评估，也很难排除。

故障风险。即使用户无意造成伤害，由于通用AI的故障，也可能会产生严重的风险。这种故障可能有几种可能的原因和后果:

- 基于通用人工智能模型和系统的产品的功能可能会被用户理解得很少，例如由于误解或误导性广告。如果用户随后以不合适的方式或出于不合适的目的部署系统，这可能会造成损害。
- 人工智能系统中的偏见通常是一个很明显的问题，对于通用人工智能来说也没有解决。通用人工智能输出可能会在种族、性别、文化、年龄和残疾等受保护特征方面存在偏见。这可能会产生风险，包括在高风险领域，如医疗保健，工作招聘和金融贷款。此外，许多
  广泛使用的通用人工智能模型主要是在不成比例地代表西方文化的数据上训练的，这可能会增加对这些数据不能很好地代表的个人造成伤害的可能性。

- “失控” 情景是潜在的未来情景，在这种情景中，社会不再能够有意义地限制通用人工智能系统，即使它们显然正在造成伤害。人们普遍认为，目前的通用人工智能缺乏构成这种风险的能力。一些专家认为，目前开发通用*自主*人工智能 (可以行动、计划和追求目标的系统) 的努力，如果成功，可能会导致失控。专家们对失控情况的合理性，何时可能发生以及减轻这种情况的难度持不同意见。

系统性风险。通用人工智能技术的广泛发展和采用带来了一些系统性风险，从潜在的劳动力市场影响到隐私风险和环境影响:

- 通用人工智能，特别是如果它进一步迅速发展，有可能自动化非常广泛的任务，这可能会对劳动力市场产生重大影响。这可能意味着许多人可能会失去目前的工作。然而，许多经济学家预计，潜在的失业可能会被创造新的就业机会和非自动化部门需求的增加所抵消，甚至可能完全抵消。
- 通用人工智能的研发目前主要集中在少数西方国家和中国。这种 “AI划分” 是多原因的，但部分原因是开发通用AI所需的计算访问级别不同。由于低收入国家和学术机构获得计算机的机会比高收入国家和技术公司少，因此它们处于不利地位。
- 通用人工智能发展的市场集中度使社会更容易受到几种系统性风险的影响。例如，少量的广泛使用金融或医疗保健等关键部门的通用人工智能系统可能会在这些相互依赖的部门中同时造成广泛的故障和中断，例如由于错误或漏洞。
- 在通用AI开发和部署中不断增长的计算使用量迅速增加了与通用AI相关的能源使用量。这种趋势没有显示出放缓的迹象，$_2$可能导致进一步增加的CO排放和水消耗。
- 通用AI模型或系统可能会对隐私构成风险。例如，研究表明，通过使用对抗性输入，用户可以从模型中提取包含有关个人信息的训练数据。对于未来针对敏感个人数据 (如健康或财务数据) 进行训练的模型，这可能会导致特别严重的隐私泄露。
- 通用人工智能开发中潜在的版权侵权对传统的知识产权法以及同意、补偿和数据控制系统构成了挑战。不明确的版权制度阻碍了通用AI开发人员宣布他们使用的数据，并且不清楚在未经许可的情况下使用其作品来训练通用AI模型的创作者会受到哪些保护。

***交叉风险因素*支撑通用人工智能相关风险的**是几个交叉风险因素 -- 通用人工智能的特征增加了不是一个而是几个风险的概率或严重性:

- 跨领域的技术风险因素包括难以确保通用AI系统可靠地按预期运行，我们对其内部工作原理缺乏了解，以及正在开发的通用AI “代理” 可以在减少监督的情况下自主行动。
- 社会交叉风险因素包括技术进步的速度和监管反应的速度之间的潜在差距，以及人工智能开发人员快速发布产品的竞争激励，这可能是以彻底的风险管理为代价的。

有几种技术方法可以帮助降低风险，但目前没有一种已知的方法能够提供强有力的保证或保证，防止与通用人工智能相关的伤害。

虽然本报告没有讨论减轻通用人工智能风险的政策干预措施，但它确实讨论了研究人员正在取得进展的技术风险缓解方法。尽管取得了这一进展，但目前的方法并没有可靠地防止在现实世界环境中甚至公开有害的通用AI输出。使用了几种技术方法来评估和减轻风险:

- 在训练通用AI模型以更安全地运行方面取得了一些进展。开发人员还训练模型，使其对旨在使其失败的输入更加健壮 (“对抗性训练”)。尽管如此，对手通常可以找到替代投入，以低至中等的努力降低保障措施的有效性。将通用人工智能系统的功能限制在特定的用例中，有助于降低不可预见的故障或恶意使用带来的风险。
- 有几种技术可用于识别风险，检查系统操作以及在部署通用AI系统后评估性能。这些做法通常被称为 “监控”。
- 减轻通用AI系统中的偏见可以在系统的整个生命周期中解决，包括设计，培训，部署和使用。然而，完全防止通用人工智能系统中的偏见是具有挑战性的，因为它需要系统的训练数据收集、持续的评估和有效的偏见识别。它还可能需要权衡公平性与其他目标，如准确性和隐私，并决定什么是有用的知识，什么是不应该反映在输出中的不良偏见。
- 隐私保护是研究和开发的活跃领域。简单地在培训中尽量减少敏感个人数据的使用是一种可以大大降低隐私风险的方法。然而，当有意或无意地使用敏感数据时，用于降低隐私风险的现有技术工具难以扩展到大型通用AI模型，并且可能无法为用户提供有意义的控制。

结论: 广泛的通用人工智能轨迹是可能的，这在很大程度上取决于社会和政府的行为

通用人工智能的未来是不确定的，即使在不久的将来，也可能出现各种各样的轨迹，包括非常积极和非常消极的结果。但通用人工智能的未来并不是不可避免的。通用人工智能是如何开发的，由谁开发，它旨在解决哪些问题，社会是否能够获得通用人工智能的全部经济潜力，谁从中受益，我们面临的风险类型，我们投入多少研究以降低风险-这些和许多其他问题取决于社会和政府今天和未来做出的选择，以塑造通用人工智能的发展。

为了帮助促进关于这些决策的建设性讨论，本报告概述了科学研究的现状以及关于管理通用人工智能风险的讨论。赌注很高。我们期待着继续这一努力。

1 介绍

我们正处于一场技术革命之中，这场革命将从根本上改变我们的生活、工作和相互联系的方式。人工智能(AI) 有望改变我们社会和经济的许多方面。

科学界普遍认为，人工智能系统的能力在过去五年中在许多任务上取得了快速进展。大型语言模型 (llm) 是一个特别突出的例子。在2019中，GPT-2，当时最先进的LLM，无法可靠地产生连贯的文本段落
不能总是数到十。在撰写本文时，像克劳德3，GPT-4和双子座Ultra这样最强大的llm可以始终如一地进行多轮对话，编写简短的计算机程序，在多种语言之间进行翻译，在大学入学考试中获得高分，并总结长文档。

这种能力的逐步变化以及持续进步的潜力，可以在许多方面帮助提高公众利益。其中最有希望的前景是人工智能在教育、医疗应用、广泛领域的研究进展以及导致繁荣的创新增加方面的潜力。这一快速进展也提高了人们对与最有能力的人工智能类型相关的当前危害和未来潜在风险的认识。

本报告旨在促进对先进人工智能安全的国际共享科学理解。

为了开始就先进人工智能的风险达成国际共识，政府代表和学术界、商界和民间社会的领导人于2023年11月在英国布莱奇利公园召开了首届国际人工智能安全峰会。在峰会上，出席会议的国家以及欧盟和联合国同意支持制定关于先进人工智能安全的国际科学报告。本报告旨在促进对先进人工智能安全的国际共享科学理解。这是该报告的第一份临时出版物: 第一份报告的最终版本将在法国*AI峰会*之前发布。

一个由75位人工智能专家组成的国际小组，他们的观点广泛，相关的背景也多种多样，为这份中期报告做出了贡献。报告所考虑的证据包括相关的科学，技术和社会经济证据。由于人工智能领域正在飞速发展，因此并非本报告使用的所有来源都经过同行评审。但是，该报告致力于仅引用高质量的来源。高质量源的标准包括:

- 该作品构成了推动该领域发展的原始贡献。
- 该作品全面地与现有的科学文献相结合，在适当的情况下引用其他人的工作，并对其进行准确的解释。
- 该作品真诚地讨论了对其主张的可能异议。
- 这篇文章清楚地描述了其分析所采用的方法。它批判性地讨论了方法的选择。
- 这篇文章清楚地强调了它在方法上的局限性。
- 这篇文章在科学界很有影响力。

由于对先进人工智能风险的科学共识仍在形成中，因此在许多情况下，该报告并未提出自信的观点。相反，它提供了科学理解和共识的当前状态的快照，或者缺乏科学理解和共识。在文献中存在空白的地方，报告指出了这些空白，希望这将促进进一步的研究。此外，本报告没有评论哪些政策选择是对其讨论的风险的适当回应。最终，政策制定者必须选择如何平衡先进人工智能带来的机遇和风险。

政策制定者还必须判断适当的审慎和谨慎程度，以应对仍然模棱两可的风险。

该报告的第一次迭代侧重于“通用”AI，即可以执行广泛任务的AI

人工智能 (AI) 是指使用广泛适用的方法开发的先进的基于机器的系统，以实现给定的目标或回答给定的问题。人工智能是一个广泛且快速发展的研究领域，有许多不同种类的人工智能。本中期报告并未涉及所有类型的高级AI的所有潜在风险。该报告的第一次迭代侧重于通用AI，即可以执行各种任务的AI。通用人工智能系统，现在通过ChatGPT等应用程序为许多人所知，在过去的18个月里，公众和政策制定者对人工智能产生了前所未有的兴趣。它的能力一直在迅速提高。通用AI不同于所谓的“窄AI”，这是一种专门执行一项特定任务或一些非常相似的任务的AI。

为了更好地理解我们如何在本报告中定义通用AI，区分“AI模型”和“AI系统”很有用。人工智能模型可以被认为是原始的数学本质，通常是人工智能应用的“引擎”。AI系统是多个组件的集合，包括一个或多个AI模型，旨在以某种方式对人类特别有用。例如，ChatGPT应用程序是一个AI系统。它的核心引擎GPT-4是一个人工智能模型。

本报告涵盖了AI模型和AI系统的风险，如果它们是“通用”AI模型或系统。我们认为AI模型是通用的，如果它可以执行或可以适应执行各种各样的任务。我们认为AI系统是通用的，如果它是基于通用模型，但如果它是基于从通用模型派生的专用模型。在通用AI领域，本报告重点关注通用AI，它至少与当今最先进的通用AI (如GPT-4 Turbo、Claude 3和Gemini Ultra) 一样强大。在我们的定义中，模型或系统不需要具有多种模态，如语音，文本和图像，就可以被认为是通用的。相反，可以在特定领域内执行各种任务的人工智能，如结构生物学，在我们的定义中也被视为通用。

重要的是，不要将通用AI与“人工通用智能”(AGI) 混淆，AGI有时用于指代潜在的未来AI系统，该系统在所有或几乎所有认知任务上的表现均等于或超过人类。通用AI是一个较弱的概念。

本报告没有解决“狭窄ai”带来的风险，“狭窄ai”经过培训可以执行非常有限的任务，并且捕获了相应的非常有限的知识体系。编写这份中期报告的时间有限，导致人们把重点放在先进的通用人工智能上，因为这方面的进展最为迅速，相关风险的研究和理解也较少。然而，从风险和安全的角度来看，狭义人工智能也可能具有高度相关性，报告中使用了与这些系统风险相关的证据。狭义的人工智能模型和系统被广泛用于医药、广告或银行等领域的产品和服务，并且可能在其中许多领域带来重大风险。这些风险可能导致诸如有偏见的招聘决定，车祸或有害的医疗建议等危害。窄AI也被用于各种军事应用。一个应用，虽然是人工智能在军队中的应用的一个非常小的子集，(*1*) 涉及，例如，致命的自主武器系统 (law)。这些主题在其他论坛中都有涉及，不在本中期报告的范围之内。

一个庞大而多样的领先国际专家小组为本报告做出了贡献，其中包括来自所有联合国区域集团以及欧盟和联合国的30个国家提名的代表。虽然我们的个人观点有时会有所不同，但我们坚信，关于人工智能的建设性科学和公共讨论对于世界各地的人们安全地获得这项技术的好处是必要的。我们希望这份临时报告能够有助于这一论述，并成为

未来的报告将逐步改善我们对高级人工智能的能力和风险的共同理解。

该报告分为六个主要部分。在此介绍之后，2。Capabilities提供有关通用AI当前功能、基本原理和潜在未来趋势的信息。3.评估和理解通用AI系统的方法解释了研究人员如何尝试了解通用AI可以做什么以及它可能带来的风险。4.风险部讨论特定风险和交叉风险因素。5.减轻风险的技术方法介绍了减轻通用人工智能风险的技术，并评估了它们的优势和局限性。6.结论: 总结和总结。

2 能力

2.1 通用AI如何获得其能力?

关键信息

- 通用AI模型和系统可以生成文本，图像，视频，未标记数据的标签，并启动操作。
- 通用人工智能模型和系统的生命周期通常涉及计算密集型的 “预培训” 、劳动密集型的 “微调” 以及持续的部署后监控和更新。

有各种类型的通用AI。通用AI模型的示例包括:

- 聊天机器人风格的语言模型，如GPT-4 (*2 **)，Gemini-1.5 (*3**)，Claude-3 (*4**)，Qwen1.5 (*5**)，Llama-3 (*6**)，和米斯特拉尔大 (*7**)。
- 诸如DALLE-3 (*9**) 、Midjourney-5 (*10**) 和稳定扩散-3 (*11**) 的图像生成器 (*8*)。
- 视频发生器，如SORA *12 **。
- 机器人和导航系统，如PaLM-E (*13*)。
- 分子生物学中各种结构的预测因子，如AlphaFold 3 (*14*)。

通用AI模型依赖于*深度学习* (*15*) 或人工神经网络的训练，这是由多层互连节点组成的AI模型，松散地受到生物神经网络大脑结构的启发。大多数最先进的通用人工智能模型都基于 “变压器” 神经网络架构 (*16*)，该架构已被证明在将越来越多的训练数据和计算能力转换为更好的模型性能方面特别有效。从广义上讲，通用AI模型的开发和部署遵循相同的一系列不同阶段: 预训练，微调，系统集成，部署和部署后更新。每个都需要不同的方法和资源。

预训练和微调都是 “训练” 通用AI模型的方法。在训练过程中，通用AI模型会获得一些数据，并对其进行处理以预测其他数据。例如，该模型可以被给定维基百科文章的前500个单词，然后预测第501个单词。最初，它是随机预测的，但随着它看到更多的数据，它会自动适应从错误中学习，它的预测也会提高。每个预测都需要一定量的计算资源 (“计算机”)，因此训练需要数据和计算。由开发人员设计的模型架构决定了模型进行预测时发生的广泛类型的计算，并且在训练期间调整了这些计算中使用的确切数字。

预培训: 预培训的目标是将一般背景知识构建成通用的AI模型。在预训练期间，通用AI模型通常从大量数据 (通常来自互联网) 的模式中学习。收集和准备训练前数据是大规模的操作，在大多数情况下，训练前是计算最密集的发展阶段。如今，通用AI模型的预训练需要数周或数月，并使用数千个图形处理单元 (gpu)-专门的计算机芯片，旨在快速处理复杂的并行计算。例如，Falcon-180B模型使用4,096个gpu

多个月，PaLM (540B) 使用6,144芯片50天 (*13*)。如今，与2010中的最先进模型训练 (*17*) 相比，此过程使用的计算量大约是其100亿倍。一些开发人员使用自己的计算进行预培训，而其他开发人员则使用专业云计算提供商提供的资源。

微调: 经过预训练后，大多数通用AI模型都会经历一个或多个额外的微调阶段，以完善其完成int结束任务的能力。微调可以包括各种技术，包括从期望示例 (*18*) 、成对的期望和不期望示例 (*19*) 或奖励和惩罚 (*20*、*21**) 中学习。微调通常需要大量的人工参与，并且往往是培训中最劳动密集型的部分，微调现代模型需要数百万个人工反馈实例 (*22 **)。通常，这种反馈是由成千上万的签约知识工作者提供的。

系统集成: 模型经过训练后，可以通过将其与旨在增强功能和安全性的其他系统组件集成来构建通用AI系统。在实践中，通用AI模型通常与用户界面、输入预处理器、输出后处理器和内容过滤器集成。

部署: 经过训练后，可以部署模型以供使用。部署可以是 “内部的”，其中系统仅由开发人员使用，也可以是 “外部的”，允许公共或其他非开发人员实体使用它。外部部署可以是 “封闭源” 或 “开放源 ”。闭源意味着公众只能通过有限的界面使用该系统。开源意味着整个系统，包括所有的模型参数，都是可用的。一些最先进的通用人工智能系统，如GPT-4 (*2 **)，是闭源的，而其他像Llama-3 (*6**) 是开源的。从减轻风险的角度来看，开源模型有其优缺点，这是科学界正在进行的讨论的主题。这份中期报告没有详细讨论开源模型的优缺点。

部署后监控和更新: 部署后许多通用AI系统会不断更新。这使开发人员可以更新功能并尝试在发现缺陷和漏洞时解决它们。这些变化通常相当于一种 “猫和老鼠” 的游戏，开发人员不断更新高调的系统，以应对新发现的漏洞 (*22 **)。

2.2 当前通用AI系统的能力

关键信息

- 通用人工智能能力很难可靠地估计，但大多数专家认为，目前的通用人工智能能力包括:
  - 协助程序员和编写简短的计算机程序
  - 在几个回合内进行流利的交谈
  - 解决教科书上的数学和科学问题
- 大多数专家认为，通用人工智能目前无法完成以下任务:
  - 执行有用的机器人任务，如家庭任务
  - 可靠地避免虚假陈述
  - 开发全新的复杂想法
- 评估通用人工智能系统能力的一个关键挑战是性能是高度特定于上下文的。有时仅在部署模型之后才会发现引发改进模型功能的方法，因此可能会低估初始功能。

可替代地，通用AI模型和系统能力可能被高估，因为在不同的上下文中缺乏鲁棒性并且使用不同的方法来引出能力。

本节重点介绍按模态 (如视频和语言) 和技能 (如推理和知识) 分类的通用AI模型和系统的功能。功能也可以根据特定基准的性能进行分类 (请参见3。评估和了解通用AI系统)。虽然本节涵盖了一般功能，但4.4.1.交叉-削减技术风险因素的重点是 “高风险” 能力。

难以定义能力- 通用人工智能系统通常是根据其能力来描述的，但在人工智能领域，“能力” 一词并没有被广泛接受的定义。定义能力的部分困难在于它不能被直接观察到 -- 人工智能研究人员只能观察人工智能系统的*行为*: 系统实际产生的一组输出或动作以及它这样做的背景 (例如，导致观察到的行为的提示) (*23*)。人工智能研究人员只能总结在许多情况下观察到的系统行为，从而得出系统的能力-能力的印象。即使在模型建立之后，也很难定义和衡量新的通用AI模型的全部功能; 研究人员和用户通常会在模型部署后发现新的方法来获取功能，例如，通过提示模型 “逐步思考” (*25 24*，)。定义通用人工智能系统功能的另一个复杂之处在于，它们是由其环境中的*能力* -- 它可以访问的工具和资源 -- 塑造的。例如，当一个通用人工智能系统连接到互联网并配备网络浏览器时，它会获得新的affor信息检索和与现实世界互动，有效地扩展其功能 (*26*)。

2.2.1 按方式的能力

通用AI模型可以根据它们处理的模式 (例如文本，图像，视频) 作为输入并生成输出进行分类。通用人工智能模型存在10 + 模态 (*27*)，如时间序列 (*28**) 和音乐 (*29**)，但文本处理模型是目前对通用人工智能模型的大部分关注的来源。先进的通用人工智能模型越来越能够处理和生成文本、图像、视频、音频、机器人动作以及蛋白质和大分子:

- 高级**文本**语言模型可以生成流畅的文本，并可用于跨各种自然语言、主题和格式的多轮对话。文本和自然语言界面对于人们与通用AI模型进行交互非常有用。一些通用AI模型可以使用文本作为输入和输出，例如OpenAI的GPT-3 (*30*); 而其他人则将文本作为输入，例如稳定性AI的稳定扩散3 (*11**)，并且可以处理越来越长的文本序列-例如，google的Gemini-Pro-1.5可以处理30,000行代码 (*31**)。文本可以包括编码为文本的许多类型的数据，例如数学公式和软件代码。在软件领域，语言模型可以编写简短的程序并为程序员提供帮助 (*32*)。
  图像- 与图像相关的许多通用AI模型可以将图像作为可能与文本结合的输入，例如Anthropic的Claude 3 (*444**)，并且可以用于分类 (*34*)，描述 (*2**)，编码 (*35*) 或分割图像，以区分其中的不同对象 (*36**)。通用AI模型也可以生成图像作为输出，例如OpenAI的DALL-E 3 (*9 **)。先进的通用人工智能模型可以生成越来越可控的图像，对更复杂的概念和图像中的文本渲染 (*9 **) 有显著的改进。
- 与**视频**相关的通用AI模型将现有视频作为输入，例如Meta的v-jepa (*37**)，或者可以从文本生成视频，例如OpenAI的Sora (*38**)。一些通用AI模型学习对可以在视频中跨时间跟踪的对象属性进行编码。电流

模型可以生成逼真的视频，但在长度 (通常小于一分钟)，保真度和一致性方面受到限制。

- **机器人动作-** 可以使用通用AI模型来规划多步机器人动作，并解释指令以指导较低级别的动作 (*39*)。最初的工作还在探索通用AI模型，这些模型不仅可以计划或解释，还可以生成机器人动作，例如Google的RT-2-X (*40 **)，但是生成机器人动作的通用AI模型功能相对较少。部分原因是，数据收集是具有挑战性的，虽然正在作出大量的努力 (*42 41，*)。
- **蛋白质和大分子-** 与蛋白质和其他大分子一起工作的通用AI模型对各种表示 (例如残基序列，3D结构) 进行操作。这些模型可以预测蛋白质折叠，生成有用的新蛋白质，并执行一系列与蛋白质相关的任务。因此，它们属于通用AI模型的定义1。
  介绍。蛋白质通用AI模型可以越来越多地控制led ，以生成跨大型蛋白质家族具有可预测功能的蛋白质设计*43， 44*。

2.2.2 技能的能力和限制

为了全面评估通用人工智能能力，通过展示知识、推理和创造力等众所周知的技能对它们进行分类可能会有所帮助。与按模态分类相比，技能更难精确定义，但为通用AI功能提供了更直观的视角。

从技能的角度来看，当今最强大的通用AI系统显示出部分熟练程度，但并不完全可靠。专家们经常不同意目前的通用人工智能系统是否可以说具有特定的技能。一种看待这一点的方法是通过能力限制对。

- **知识 (能力) 和不一致 (限制) --** 通用人工智能模型对公共互联网上发现的广泛事实进行编码 (*45*)，但在识别细微的事实差异方面受到限制，并不总是生成自洽的文本。因此，从通用AI模型中获取知识可能是不一致的 (*46 45，*)。
- **创造力 (能力) 和幻觉 (限制) -**通用AI模型可以生成新颖的示例 (例如，新的图像或文本)。这种类型的 “创造力” 可能是有用的，但也可能导致制造内容的 “幻觉”。对于e xa mple，语言模型通常会生成不存在的引文，传记或事实 (*46， 47 *， 48， 50 49，*)，这些都会带来错误信息的风险 (请参阅4.1。恶意使用风险)。
- **常识推理 (能力) 和因果关系 (限制) -**诸如 “常识” 或 “推理” 之类的术语在人工智能领域通常没有很好的定义，并且通常以不同于日常使用的方式来描述人类的能力。在某些情况下，通用人工智能模型展示了模仿广泛的 “常识知识” 的能力 (*51*)，逐步解决相对复杂问题的能力 (*24*)。以及在给定的上下文中学习和应用新模式的能力 (称为 “上下文学习”) (*52 30，*)。然而，“推理” 的适当形式是上下文和任务相关的。通用人工智能系统在多大程度上表现出真正的 “推理” 或 “常识” 是有争议的。研究表明 (*53*)，即使在不寻常的情况下 (*53*)，一些r的缓解能力也在提高常识推理问题，尽管其他形式的基本 “常识知识” 存在很大的局限性 (*54*)。即使通用人工智能模型似乎对世界进行了正确的 “推理”，它们也可能没有确定这种推理的潜在因果基础 (*55*)。有一个普遍的共识是，目前通用的人工智能缺乏人类水平的推理能力。
- **形式推理 (能力) 和组合 (限制) -**特别是当给定额外的资源，如工具和多次尝试时，语言模型可以在数学，计算机编程和自然科学等领域执行一些形式推理任务。

注意到上述关于在该领域使用 “推理” 一词的警告。例如，研究表明，克劳德3模型接近研究生水平的专家在生物学，物理学和化学 (*4 **) 的相关问题上的表现-在克劳德3最初训练后创建的基准上 (有关基准的讨论，请参见2.4。能力进展
未来几年，3。评估和理解通用人工智能系统的方法)。然而，这些模型使用 “变压器 ” 架构，这是当今大多数通用人工智能系统所基于的，2017年引入 (*16*)。原则上，这种体系结构在执行任意组合推理时具有基本限制，这是形式推理 (*56*) 的基础。目前尚不清楚这些理论限制在实践中的相关性。

- **预测 (能力) 和新概念 (限制) -**当集成到更复杂的系统中时，语言模型可以用于在受限领域中以合理的预测精度预测未来事件。最近的一项研究 (*57*) 表明，使用检索的语言模型系统可以在统计预测问题 (即预测事件的概率) 上与专家预测者的总体性能相匹配。然而，虽然目前的能力表明模型可以综合信息来推断未来事件的可能性，但模型可以综合全新概念的程度似乎是有限的 (*58*)。
- **模拟 (能力) 和实施 (限制) -**当集成到虚拟环境中时，通用AI模型可以模拟虚拟代理的行为 (*59*)。例如，最近的研究表明，由openai的ChatGPT提供支持的25个虚拟代理可以以与人类行为相匹配的方式运营虚拟城镇 (*60*)。然而，虽然目前的通用人工智能模型可以模拟虚拟代理，但它们在 “实施方式” 上受到限制，还不能有效地控制物理机器人或机器，因为通用人工智能模型与电机控制系统的集成仍然是一个挑战 (*61*)。

2.3 能力及其驱动因素的最新趋势

关键信息

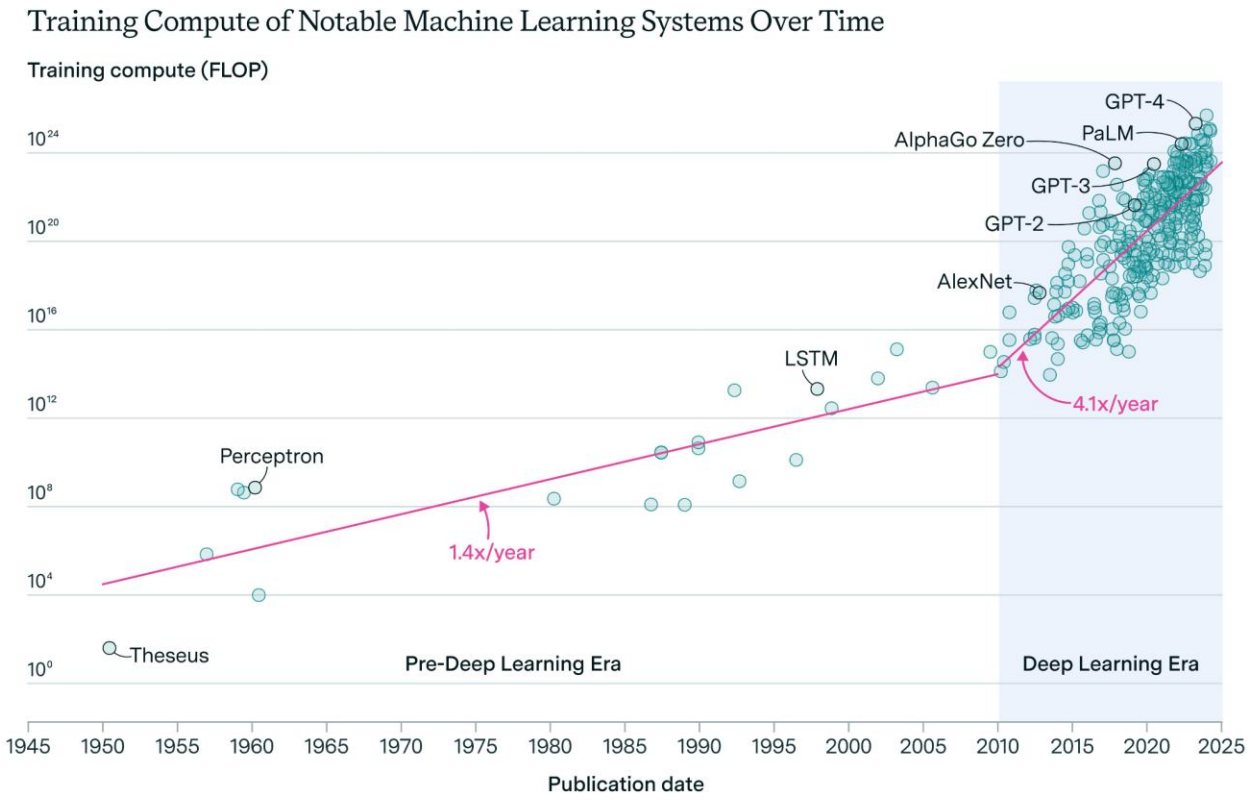
- 近年来，通用人工智能能力根据许多指标迅速发展，这要归功于用于训练和算法改进的资源的增加。每个模型，这些估计都增加了:
  - 训练计算: 4倍/年
  - 训练数据集大小: 2.5倍/年
  - 算法训练效率: 1.5倍至3倍/年
  - 培训期间用于为计算机芯片供电的能量: 3倍/年
  - 硬件效率: 1.3倍/年
- 近年来，使用越来越多的计算和数据来训练通用AI模型被称为 “扩展” 模型。广泛指标的性能随着规模的扩大而提高，许多人工智能研究人员一致认为，近年来，扩展推动了高级通用人工智能功能的大部分增长。然而，人们争论这是否导致了诸如因果推理等基本挑战的进展。

2.3.1 计算、数据和算法的最新趋势

在过去十年中，对计算资源的投资增加，硬件效率的提高，易于在线访问的数据集的存在以及算法的渐进式创新为通用AI的发展做出了贡献。本节研究计算能力、数据和算法的最新趋势。

训练和推理中使用的计算趋势

用于训练AI模型的计算资源一直在快速增加。计算资源 (通常称为 “计算机”) 表示所执行的操作的数量。自21世纪10年代初以来，这一数字呈指数级增长，用于训练机器学习m模型的平均数量大约每六个月翻一番 (*17*)。2010，著名的机器学习模型 (*62*， *63*， *64*) 平均使用了大约1e15个浮点运算 (FLOP) (*65*)，但2023年拐点-2，这是公开报告计算预算的最大模型。二手1e25翻牌 (*66**) -增加100亿倍。这一进展是由行业实验室愿意将更多数据中心容量用于大规模通用AI培训推动的。没有足够的数据来确定这种趋势是否在较短的时期内发生变化，例如21世纪20年代。



图I.随着时间的推移训练值得注意的机器学习模型的计算 (17， 65)。计算是在从AI文献估计的总浮点运算 (FLOP) 中测量的。对于GPT-4等最近未公开的模型，估计的准确性预计在两倍或五倍之内。

从“机器学习中的参数，计算和数据趋势”中获得Epoch AI的许可。在线发布t epochai.org。检索自: 'https://epochai.org/data/epochdb/visualization'.

在过去的十五年里，每美元的计算量增加了大约50到200倍 (*68 67*， 67)。然而，用于训练通用人工智能模型的计算总量远远超过了计算成本的降低: 例如，谷歌的Word2vec模型使用了大约3e16个FLOP 2013年进行训练，比目前的frontier模型 (*65*) 小10亿倍。虽然GPU性能的改进有所帮助，但这些改进部分受到数据中心GPU短缺和AI应用程序中使用的顶级GPU价格高昂的限制。高端处理器，包装，高带宽内存和其他组件的供应链短缺正在延迟技术部门满足对AI服务器等人工智能硬件的巨大需求的能力 (*69*)。通用AI计算使用的扩展主要是行业实验室越来越愿意将数据中心资源和工程人员分配给大规模通用AI培训运行的结果。

神经 “缩放定律” 的发现，描述了计算量，模型和数据的大小以及性能之间的可预测关系，促成了以计算为中心的AI开发观点，这在一些领先的AI实验室中很突出。[1]，Google Gemini Ultra和OpenAI的GPT-4等旗舰通用AI模型的开发是由扩展法则 (*2 *， 3**) 的工作指导的。因此，对硬件基础设施专业知识的需求更大，并且AI实验室与微软和谷歌等技术巨头之间的合作更加紧密。[2]

用于部署的计算资源也出现了显著增长。公司正在快速扩展基础设施以满足这些不断增长的需求。推理所需的计算资源 (向用户提供通用AI系统的关键部分) 经历了显着增长 (*76*)，因为部署通用AI系统的用户数量快速增长。据报道，在2023年4月，OpenAI的人工智能系统估计会产生700美元/天的推理成本 (*77*)。一些估计表明，用于通用人工智能推理的总计算量已经超过了用于训练新模型的计算量，例如，人工智能推理代表了谷歌人工智能基础设施排放的60% 2022年 (*78*)。

用于训练和推理的计算资源不断增长，也迅速扩大了人工智能的能源使用 (4.3.4风险环境)。

训练数据趋势: 更大的数据集、多模式、合成数据和人类偏好

通用人工智能开发人员已经能够显著增加训练数据集的大小，这要归功于互联网内容的可用性，包括开放的web数据存储库。这些较大的数据集有助于在各种指标上实现更高的性能。用于训练通用AI的数据集大小已经从原始Transformer模型的大约20亿个令牌 (令牌是一个单词，一个字符，或者一个单词的一部分次) 增加到2017 3万亿多个令牌 (*79 *， 2023年80**)，每三年增长约10倍 (*65*)。

然而，通用的人工智能开发人员在互联网上只有有限的文本数据可供利用 (*82 81，*)。虽然这可以克服，例如通过多次训练相同的数据，使用人工智能生成的数据，或在其他非文本数据源 (如YouTube视频) 上进行训练，但一些人认为，2030年缺乏acceSible在线高质量文本数据c会减缓模型有效缩放的速度 (参见2.4.2资源将被缩放迅速?)。

数据质量在训练高性能语言模型中起着至关重要的作用。选择高质量数据并优化da tase t的整体组成可以显着提高模型性能，但是此过程需要大量劳动 (*83， 85 84，*)。此外，测量和分析数据以识别和减轻pr缺陷，如偏见和缺乏多样性，对于产生高质量的模型至关重要 (*86 **)。

在图像，音频和视频以及文本等多种模式上训练通用AI模型最近获得了牵引力。GPT-4、克劳德3和双子座超等通用人工智能模型结合了不同的模态来执行需要联合处理文本、视觉和听觉信息的任务，例如分析带有文本和图形的文档或创建多媒体演示 (*2 *,3 *,4 **)。

“人类偏好” 数据捕获用户喜欢的输出类型，对于开发通用AI系统至关重要。这些数据不能从公开可用的来源中挖掘，但必须专门为培训而生成; 因此，它比用于

---

[1]神经缩放定律已经显著影响了一些前沿人工智能实验室以计算为中心的人工智能开发观点。Anthropic的研究原则指出，他们 “制定了缩放定律，以帮助我们进行系统的，经验驱动的研究 (...)，以更有效，更可预测地训练网络，并评估我们自己的进步。”(70 *).

[2]例如，OpenAI和Microsoft (71 *) 、Anthropic和Google Cloud (为72 *) 、Amazon Web Services (73 *) 、Cohere和Google Cloud (为74*) 之间存在合作关系。mistral和Microsoft Azure (75*)。

预培训。这些数据有助于微调语言模型，以符合用户和开发人员的需求，适应不同的偏好，并根据人类对质量和帮助的判断 (*20*、*21 **、*87**)。人工智能实验室和大公司可能在生产和访问大量专有的人类偏好数据方面具有优势。

通用人工智能的技术和训练方法不断改进

随着时间的推移，最强大的通用人工智能模型的技术和训练方法得到了持续和可靠的改进 (*88 *， 89*)。在图像分类、游戏和语言建模等关键领域，人工智能技术和训练方法的效率大约每2到5年就会提高10倍。例如，训练模型以执行图像分类以达到设定的性能水平所需的计算量2012年减少了44倍和2019倍，这意味着效率在16个月内翻了一番。玩游戏的人工智能系统每5到20个月需要一半的训练样本 (*90*)。在语言建模中，达到固定性能水平的计算要求大约每8个月平均2012年减半 (*89*)。这些进步使通用AI研究人员和实验室能够在有限的硬件预算范围内开发出更强大的模型。

在算法中也有增量的进步，这些进步并不能最好地理解为提高计算效率。例如，新技术显著增加了上下文窗口的大小，允许通用人工智能系统处理更大量的信息 (*31 *， 91 *， 92**)。训练后算法允许g通用AI系统使用工具并在没有人类帮助的情况下在世界上采取行动 (请参阅2.4.3。算法的进步会导致快速的进步吗？)。

尽管人工智能算法取得了重大进展，但近年来通用人工智能的重大概念突破相对较少。“Transformer架构” 可能仍然是最重要的创新，并且被最先进的通用AI系统 (*16*) 使用。虽然已经提出了许多替代架构，但没有一个在实质上并且始终如一地优于变压器。一旦经过适当测试，最近的 “选择性状态空间模型” (*93*) 可能会被证明比变压器更有效。这些模型强化了最近的趋势，即允许语言模型分析更长的上下文，例如书籍和大型软件项目。如果需要更基本的概念突破来推进通用人工智能能力，这可能是进一步发展的关键障碍，即使增量改进和扩展继续推动某些领域的快速进展 (见2.4.1。如果资源继续快速扩展，这会导致快速的进步吗？)。

2.3.2 能力的最新趋势

通用人工智能的发展越来越快，在某些指标上接近或超过了人类水平的表现，其影响存在争议。

最近通用人工智能的发展速度很快，在某些指标上经常超过人工智能专家的预期。在过去的十年中，人工智能在计算机视觉、语音识别、图像识别和自然语言理解等领域的一些基准测试中，已经达到或超过了人类的平均水平 (图2)。LLMs的最新进展建立在这一长期趋势的基础上。

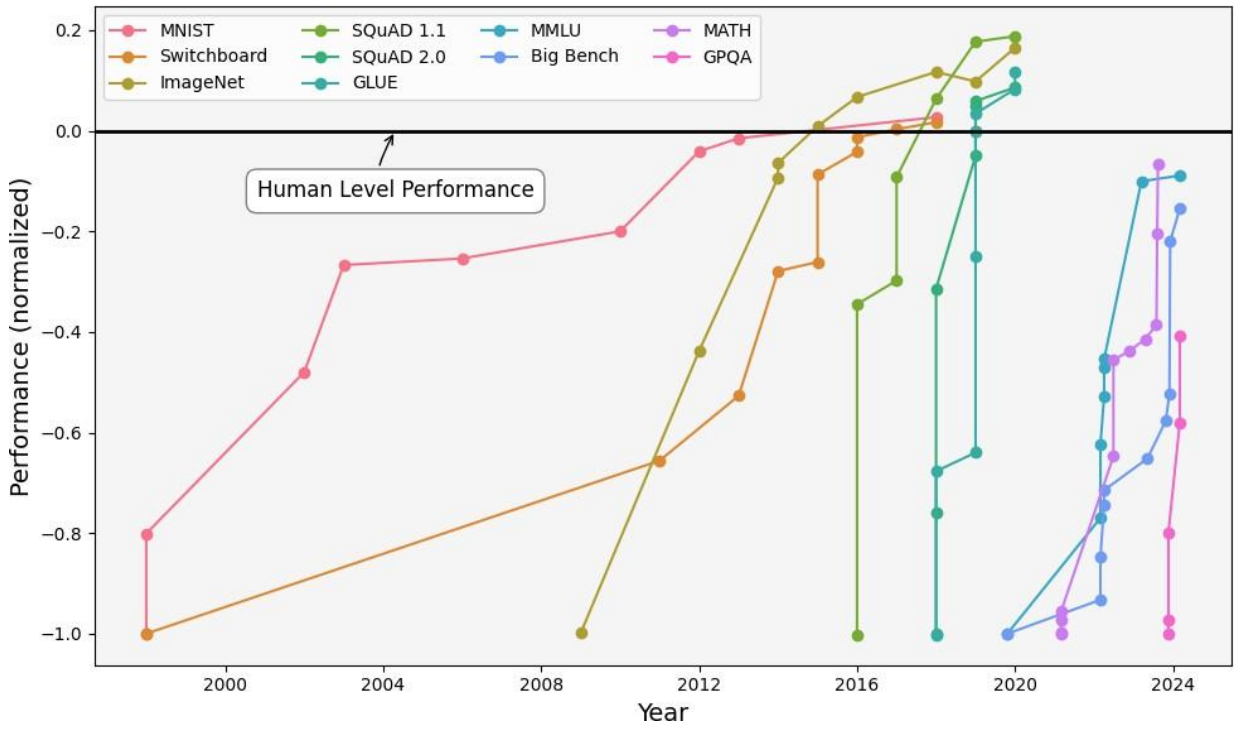


图2。AI模型在各种基准上的性能1998年2024年，包括计算机视觉 (MNIST、ImageNet) 、语音识别 (Switchboard) 、自然语言理解 (SQuAD 1.1、MMLU、GLUE) 、通用语言模型评估 (MMLU、Big-Bench和GPQA) 和数学推理 (MATH)。许多模型都超过了人类水平的性能 (黑色实线) 2024年，这表明过去二十年来，人工智能在不同领域的能力取得了重大进步。数据来自mn的 (94个)，Switchboard，ImageNet，SQuAD 1.1，2和GLUE。MMLU、Big Bench、GPQA的数据 (95、96、97) 来自相关论文。

LLM功能在多个领域2020年和2024中取得了显着进步，如大规模多任务语言理解 (MMLU) (*95*)，Big-Bench (*96*) 和研究生水平的Google证明Q & A (GPQA) (*97*)。2020年，在许多基准测试中，通用人工智能模型的表现远远低于人类测试一下的平均水平; 2024年，先进的通用人工智能模型已经接近人类水平的表现。例如，考虑数学基准测试 (*98*)，它测试数学解决问题的能力。最初，通用人工智能系统在这个基准测试中表现不佳，但在发布两年后，GPT-4似乎达到了42.5% 的准确性 (*99 **)，随后的工作将使用GPT-4的最先进性能推向了84.3% (*100*)。接近专家测试人员获得的分数。

尽管基准指标取得了快速进展，但与现实任务相比，这些基准是非常有限的，专家们争论这些指标是否有效地评估了真正的概括和有意义的理解 (*101*)。最先进的通用人工智能模型经常在一些基准测试中表现出意想不到的弱点，这表明它们完全依赖于记忆模式，而不是采用强大的推理或抽象思维 (*103 102，*)。在某些情况下，模型是在ben chmark解决方案上意外训练的，尽管缺乏实际能力 (*105 104，*)，但仍具有较高的基准性能。模型也很难适应训练数据 (*106*) 中较少代表的文化。这突显了基准测试结果与可靠地将知识应用于实际现实情况的能力之间的巨大差异。

人工智能和人类能力有明显的优势和劣势，使得比较具有挑战性

虽然将人类的认知能力与通用人工智能系统的能力进行比较可能很诱人，但它们有明显的优势和劣势，使得这些比较

在许多情况下意义不大。虽然通用人工智能在某些领域表现出色，但它可能缺乏人类 (*102*) 深刻的概念理解和抽象推理能力。目前的通用人工智能系统往往性能参差不齐，在一些狭窄的领域表现出色，而在其他领域表现不佳 (*102*)。

当前的通用人工智能系统容易出现一些人类没有的故障 (*108 107，*)。通用人工智能推理可能 “脆弱” (无法应对新的场景)，并且过度受到表面相似性的影响 (*102*)。在人类通常擅长的环境中，llm可能无法推理。例如，一个包含 “Olaf Scholz是德国第九任总理” 的数据训练模型将无法自动回答 “谁是德国第九任总理” 的问题 (*107*)。此外，llm可以通过无意义的输入来利用，从而偏离其通常的保护措施，而人类会识别这些提示 (请参阅5.2。培训
更值得信赖的模型)。

随着通用人工智能模型的扩大，它们的能力总体上有所提高，但到目前为止，这种增长对于特定的能力来说很难预测。

随着计算能力和数据的规模，语言模型的总体性能得到了可靠且可预测的改善。研究人员发现了经验性的 “缩放定律”，可以量化中的the与模型在下一个单词预测 (*109 *，* next-word-prediction) 等广泛性能指标上的能力之间的关系。*110**。跨不同领域的实证研究已经证明，机器学习系统ems的的性能提高了计算资源的 ，包括视觉 (*111 *，* *112*)，语言建模 (*109 *，* *110**)，和游戏 (*113**)。

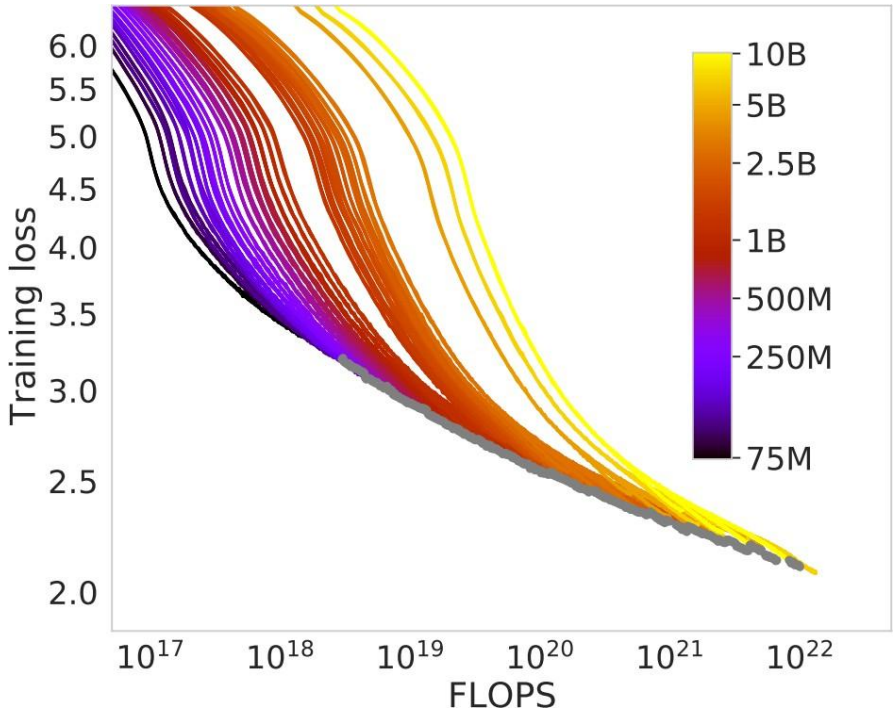


*图3.交叉-熵损失可预测地在经验研究的训练运行的广泛ra nge上进行计算。FLOPS是指在训练期间执行的操作的数量。(110 * 的数字)。不同的颜色代表具有不同数量参数的模型。每条线显示了模型的训练失败次数增加时损失如何下降。权限是从作者那里寻求的。*

最著名的缩放定律预测，随着语言模型规模的增长和对更多数据的训练，它们在[3]可预测的数量( )。具体地说，这些模型在预测序列中的下一个 “标记” 时变得更加准确，该 “标记” 可以是单词，字符或数字。当这种性能提高时，模型实际上会在数据集中隐含的任务方面变得更好。例如，随着模型的扩展，通用AI模型性能已被观察到持续改进广泛的基准，这些基准测试一下许多功能，例如MMLU (*95*)。

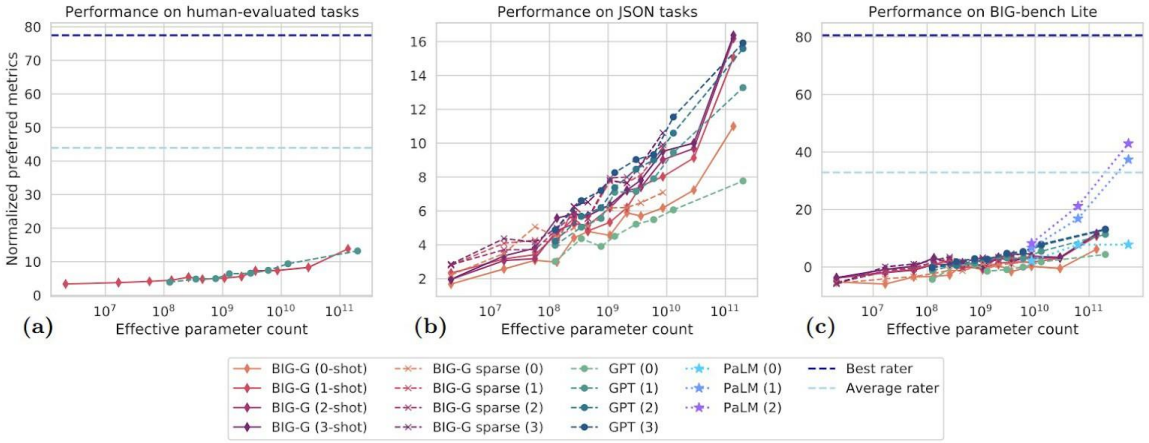


图4。在诸如Big-Be nc h之类的广泛基准上的性能与参数缩放和更普遍的计算缩放密切相关。这个数字来自 (96)。

这些比例定律是从经验观察得出的。它们，非from在违反原则，尽管已经提出了理论模型来解释它们(*115 *，116 *，117，118，119*)。因此，在数学上无法保证它们将继续适用于超出用于建立它们的经验数据范围的标度。

综合基准测试中许多任务的*综合*性能可以根据模型规模进行部分预测 (参见图4)，但是，目前尚不清楚我们是否可以可靠地提前预测是否以及何时会出现*特定*功能。有许多记录在案的示例，当模型达到一定规模时，有时会突然出现，而没有明确地编程到模型 (*120， 121， 123*) 中。例如，一定规模的大型语言模型在被提示逐步执行计算时，已经获得了以高精度执行大数加法的能力。一些研究人员有时将这些定义为 “紧急” 能力 (*120， 121， 122 *， 123*)，表明它们存在于较大的模型中，但不存在于较小的模型中，因此它们的出现可能很难提前预测。这表明新能力，包括有益和潜在有害的能力，可能会意外出现。

，人们争论这些能力是逐渐出现还是突然出现，对于它们可以提前多久预测也存在分歧。最近的研究发现，如果使用更多线性和连续的指标来衡量进度 (*124*)，则某些此类功能会渐进和可预测。这导致一些研究人员质疑能力是 “新兴的”，因为人工智能中出现的一些定义要求能力以一定的规模突然出现 (*124*)。这表明，目前似乎突然出现的能力可能会变成

[3]这项工作隐含地将数据视为同质的; 最近的工作研究了缩放定律和数据集管理的交集 (114)。

如果使用不同的进度指标，可以提前预测。这是一个悬而未决的问题，是否可以预测新模型的新功能，以及提前多长时间。

最近的研究已经确定了 “逆sc aling” 的例子，其中语言模型性能*恶化*，因为模型大小和训练计算增加 (*125*)。例如，当被要求用新的结尾完成一个常用短语时，较大的模型更有可能失败，并简单地复制记忆的短语，而不是 (*125*)。虽然模型缩放有时会导致性能下降，但对这种现象的研究也会发现性能出现随机波动。当外推到更大的模型时，一些明显的反向缩放趋势可能不会持续，性能最终会 (*126*) 再次提高。逆缩放的全部含义仍不清楚，可能需要进一步研究以更好地理解这一现象。

2.4 未来几年的能力进步

关键信息

- 通用人工智能能力的未来进展速度对管理新兴风险具有重要意义，但专家们对未来的预期存在分歧，即使是在不久的将来。专家们以各种方式支持通用人工智能能力缓慢、快速或极快发展的可能性。
- 这种分歧涉及一个关键问题: 继续 “扩展” 和完善现有技术是否会产生快速进展，或者这种方法是否受到根本限制，是否需要不可预测的研究突破来大幅提高通用人工智能能力？那些认为需要研究突破的人通常认为，最近的进展并没有克服常识推理和灵活的世界模型等基本挑战。
- 近年来，三个主要因素推动了人工智能的进步: 扩大训练中使用的计算能力 (“计算机”); 扩大训练数据量; 改进人工智能技术和训练方法。
- 领先的人工智能公司正在押注这三个因素继续推动改进，特别是增加计算。如果最近的趋势继续下去，到2026年年底，一些通用人工智能模型将使用目前发布的计算密集型模型的40到100倍的计算量进行训练，结合大约3到20倍的更高效的技术和训练方法。
- 然而，进一步增加数据和计算存在潜在的瓶颈，包括数据的有限可用性、人工智能芯片生产挑战、高总成本和有限的本地能源供应。人工智能公司正在努力克服这些瓶颈。扩展的速度还取决于可能对AI部署和开发施加限制或条件的法规。

本节研究了进一步扩展计算和训练数据的可行性和有效性，以及通过算法开发快速发展的潜力。总体而言，这两种方法都可能加起来，并导致进一步的进步，但没有商定的方法来预测进步的步伐。有关AI能力的全球差异的讨论，请参见4.3。 2
全球AI鸿沟。

2.4.1 如果资源继续快速扩展，这是否会导致快速进步？

关于未来进展的分歧的一个关键驱动因素是如何解释过去的进展。近年来，扩展计算资源和数据已导致通用AI系统的性能持续提高，这可以通过一系列基准测试的更高分数来证明 (作为通用AI模型被放大，它们的能力提高了整体，但到目前为止，这种增长很难预测特定能力)。一些人工智能研究人员认为，过去的进展涉及通用人工智能系统在理解和推理能力方面的重大而有意义的进步，并且随着更多的计算和可能适度的概念创新，这种持续的进步可能会导致通用人工智能系统的发展，这些系统在大多数认知任务 (*127*) 的广泛人类水平或更高的水平上运行。

其他研究人员对此表示怀疑。他们认为，目前基于深度学习的通用人工智能系统 (目前主要的机器学习方法依赖于深度人工神经网络 (*15*))，从根本上缺乏智能的关键组成部分。特别是，一些人认为当前的深度学习系统缺乏因果推理能力 (*128*) 、从有限数据中抽象出来、常识推理 (*102， 130 129，*) 和灵活的预测世界模型 (*55，130 128，*)。se的研究人员认为，这些缺点不能仅通过缩放来解决 (*102， 130 129，*)。当前系统的重大局限性支持了这一点，2.2讨论了当前通用AI系统的能力。他们认为，解决这些局限性，可能需要超越当前深度学习范式的重大概念突破和创新。这表明，在通用人工智能系统中实现人类水平的性能需要重大的概念突破，而目前由渐进式改进驱动的进展类型不足以实现这一目标。

导致通用人工智能能力重大飞跃的基本概念突破是罕见且不可预测的。即使发明了新技术，现有的通用人工智能系统的基础设施和开发者惯例也可能为大规模应用它们提供障碍。因此，*如果*需要重大的概念突破，可能需要很多年才能实现。

这些对立的观点不一定是不相容的: 尽管目前最先进的深度学习系统的推理能力比大多数人类弱，但我们看到的是从一代通用人工智能模型到下一代的进步。许多人工智能研究人员正在探索适应通用人工智能模型的方法，以解锁或改进“系统2”推理 (分析、基于规则和控制的推理) 和“自主代理”能力。如果这些能力取得进展，可能会对未来几年的人工智能风险管理产生重要影响。4.4 1 交叉技术风险因素 ，用于讨论自主代理能力的潜在进展和风险。

2.4.2 资源是否会迅速扩展？

专家们一直在争论是否有可能继续快速增加用于人工智能开发的资源，以及持续多长时间。

谷歌和亚马逊等科技巨头正在对数据中心和GPU收购进行大量资本投资，以支持进一步扩大通用AI模型。如果这在不久的将来产生实质性的改善，那么由此产生的能力可能会激发市场信心，并证明额外的支出是合理的。大型科技公司拥有所需的现金储备，可以将最新的培训规模扩大100至1,000倍 (*131 **)。然而，进一步扩大规模超过这一点可能更难以融资。此外，获得资本投资并不是唯一的潜在瓶颈。缺乏数据、能量和gpu都是进一步快速扩展资源的潜在障碍，这将在下面讨论。此外，数字基础设施的质量存在巨大的全球差异，这给许多国家带来了额外的障碍，并导致人工智能能力的全球鸿沟不断扩大 (4.3.2 global AI divide)。

数据瓶颈可能会限制快速扩展，但这刺激了新的方法，如多历元训练，合成数据和跨域迁移学习

虽然数据可用性可能会在中期内限制通用AI的扩展，但合成数据和迁移学习等解决方案可能有助于解决这一瓶颈。当前最先进的通用AI模型使用具有数万亿字的文本数据集。训练数据集规模的快速扩大可能很快就会受到可访问的在线高质量文本数据 (*81*， *82*) 和有关数据访问的法律纠纷的限制。

训练大型通用AI模型的数据可用性瓶颈是最近的一个挑战，克服这些瓶颈的方法仍处于探索的早期阶段。有一系列克服这些挑战的方法，例如:

- **多历元训练:** 为“单个历元”训练LLMs是很常见的，这意味着它只看到一次训练数据。然而，在相同的数据上多次训练是机器学习其他领域的标准做法，并且已经表明，训练四个时期的通用AI模型可以产生ben efits，大致相当于4x数据，尽管额外的时期收益下降 (*81*)。这可以在一定程度上扩展语言模型的现有数据库存可以完成的工作，这取决于当前最先进的模型已经在某些数据源上针对多个时代进行了训练的程度。
- **合成或自生成数据**: 通用AI生成的输出可以增强训练数据或人类反馈数据，这在实际数据有限的 (*132**) 时非常有用。这种方法提供了对生成数据的控制，填补了现有数据集中的空白 (*133*)，并提高了模型的鲁棒性、预测和泛化 (*134 133*，)。然而，有人担心合成数据可能会减少有意义的人类监督，并可能逐渐放大通用人工智能模型的偏见和不良行为 (*135*)，并且可能无法像实际数据那样提高能力 (*136 133*，)。
- **跨域迁移学习**: 在来自各种来源的数据上训练通用AI模型，例如文本，图像，视频，语音和生物序列 (*35,137*，*138 **)，可能会大幅增加可用数据。例如，对代码进行训练可以提高自然语言任务的模型性能，相当于将语言数据集 (*81*) 增加一倍。

能源需求可能会对现有的电力基础设施造成压力

训练通用AI系统的能源需求不断增加，可能会开始使结构的能量紧张。在全球范围内，用于人工智能的计算预计需要至少70太瓦时的电力2026年 (*139*)，大致相当于奥地利或芬兰等较小的欧洲国家的耗电量。这可能会从其他目的转移能源，并产生环境成本*140*。

例如，在美国，目前有许多领先的人工智能公司，电网和输电基础设施可能难以适应与人工智能相关的电力需求激增。升级电网以将更多的电力从发电厂输送到数据中心涉及漫长的规划，批准和建设过程。这种缓慢的构建过程给通用AI培训设施带来了特殊的挑战，这可能需要许多gpu在地理上紧密地协同定位，从而导致巨大的本地功耗。

主要的人工智能公司正在通过主动寻求确保其电力供应来做出回应。例如，一家计算提供商最近购买了一个具有960MW能源供应 (*141 **) 的数据中心。这可以为训练运行提供比训练GPT4 (*2 **) 多100倍的计算资源，但需要更多的能量来继续以目前的速度扩展超过几年。

AI芯片生产挑战和GPU改进放缓可能会限制AI计算扩展

在过去的几年中，数据中心gpu的生产一直是通用AI系统扩展计算的瓶颈，部分原因是半导体制造工厂的能力有限以及全球半导体供应链中的限制和优先事项 (*143 142，*)。人工智能芯片制造依赖于复杂的、难以规模化的供应链，包括复杂的光刻、先进的封装、专业的光刻胶和独特的化学品。
建造新的半导体制造工厂 (“fabs”) 非常昂贵，通常需要三到五年 (*144，145*) 使行业对市场需求反应迟缓。这些因素使得预测未来的供应变得复杂，并且已经提出了一系列芯片可用性方案。虽然最先进的GPU生产正在大幅回升，但人工智能芯片的供应链可能无法适应需求。这可能会减缓前沿人工智能公司进一步快速增长的雄心，尽管人工智能公司如果能够获得生产的gpu总数的很大一部分，可能会在短期内继续扩大规模。

改进的GPU性能也有助于最近的计算 (扩展*67， 68*)。在这十年中，由于晶体管尺寸 (*146， 147*) 和能效 (*148*) 的物理限制，单个GPU性能的进步可能会放缓。这对缩放的影响有限，因为计算缩放的主要驱动因素不是提高GPU性能，而是增加了使用的GPU数量。GPU价格性能和相关计算的能源效率每年都在提高大约30%，其中增加张量核心和转向较低精度格式 (*68 67，*) 的因素为10x-100x。但是，训练中使用的总计算量每年2010年增加约4倍 (为*17*)，超过了硬件效率改进的速度。这表明，支出的增加，而不是硬件效率的提高，一直是人工智能培训计算预算增长的主要驱动因素。

2.4.3 算法的进步会带来快速的进步吗？

训练算法的效率持续快速增长。支持通用人工智能模型的技术和算法随着时间的推移一直在不断改进 (2.3 1 最近的趋势在计算、数据和算法)。这包括对关键算法的增量更改，例如变压器的体系结构，对硬件上实现AI算法的改进，对如何缩放模型的更好理解，在神经网络中处理表示的更好方法以及其他进步。这些进步使研究人员和实验室能够在不增加硬件预算的情况下，开发出更强大的通用AI模型，例如视觉 (*88 **) 、强化学习 (*90*) 和语言建模 (*89*)。在语言建模中提高算法效率的速度没有显示出放缓的迹象 (*89*)，尽管在不久的将来可能会有回报的减少。

的训练后算法可用于以低成本显着提高通用AI模型的能力，包括通过微调，工具使用和结构化推理技术。近年来，在预训练后增强通用AI模型性能的技术和方法方面取得了重大进展。许多训练后算法在给定基准上提高了模型性能，使用的训练计算量超过5倍，在某些情况下超过20倍 (*25*)。训练后算法可以应用于各种用例的给定模型，目的是更好地满足最终用户的特定需求，而成本要比开发原始模型低得多。这种低成本意味着包括低资源参与者在内的各种参与者可以通过开发更好的训练后算法来推进前沿通用AI能力。治理过程需要考虑训练后的算法。

训练后的算法包括微调模型以获得更好的性能，为他们配备利用外部工具的能力，制作提示以指导他们的输出，构建他们的推理

处理更连贯和合乎逻辑的响应，并从多个响应中选择最相关和最准确的候选输出。在培训后增强方面的工作正在迅速增长，在广泛的 (*149*) 和特定领域 (如代码生成 (*150*) 和数学 (*151*)) 提高前沿llm的性能。这些创新有可能在未来几年进一步提高通用人工智能系统的性能。然而，如果由于上面讨论的瓶颈，用于训练的资源的扩展速度减慢，这可能反过来减慢寻找更好的训练后算法的进度。

可以部署**通用AI系统来实现自动化和加速AI研发。**狭义的人工智能系统已经被用于开发和改进算法 (*153 152，*)。最近的llm用于与AI研发相关的领域，特别是在编程方面(*26*)，生成和优化提示 (*154， 155，156，157*)，取代人为微调数据 (*158 **)，并选择高质量的训练数据(*159 **)。随着通用人工智能系统能力的提高，预测对人工智能算法进展和工程的影响变得越来越困难。

3 评估和理解通用AI系统的方法

关键信息

- 通用人工智能治理方法假设人工智能开发人员和政策制定者都可以理解和衡量通用人工智能系统的能力及其潜在影响。
- 技术方法可以帮助回答这些问题，但有局限性。目前的方法无法提供强有力的保证，防止大规模通用人工智能相关的危害。
- 目前，开发人员对他们的通用AI模型如何运行仍然知之甚少。模型解释和可解释性技术可以提高研究人员和开发人员对通用人工智能系统如何运行的理解，但这项研究还处于起步阶段。
- 通用AI的能力主要通过在各种输入上测试通用AI来评估。这些抽查是有帮助和必要的，但不能提供数量保证。他们经常忽略危险，高估或低估通用人工智能能力，因为测试一下条件与现实世界不同。许多令人关注的领域并不完全适合当前评估所依赖的量化类型 (例如，偏见和错误信息)。
- 原则上，独立参与者可以审核公司开发的通用AI模型或系统。然而，公司并不总是为独立审计师提供必要级别的 “白箱” 访问模型或所用数据和方法的信息，这是严格评估所需的。一些政府正在开始建设进行技术评估和审计的能力。
- 很难评估通用人工智能系统的下游社会影响，因为尚未开发严格而全面的评估方法，并且通用人工智能具有广泛的现实用途。了解通用人工智能模型和系统的潜在下游社会影响需要细致入微的多学科分析。在人工智能开发和评估过程中，越来越多地参与和表达观点是一项持续的技术和制度挑战。

现代通用人工智能系统可以包含多个大规模模型和数千亿个参数，通常部署为通用产品。因此，很难预测这些通用人工智能产品在许多可能的部署场景中如何发挥作用，更难以适当地描述其部署的下游后果。科学家们经常对通用人工智能系统的意想不到的能力和影响感到惊讶。

3.1 通用人工智能评估用于评估模型的功能和影响。

评估通用AI模型和系统有两个主要原因:

1. **确定一般功能和限制:** 模型评估表明模型设计选择与模型结果之间的关系。这种性能分析有助于研究人员了解这些系统在受控和自然环境中如何满足我们的期望。对模型功能的更深入了解有助于判断其是否适合使用。每次评估都有局限性和不确定性，必须记录下来以正确解释其结果。
2. **评估社会影响和下游风险:** 对通用人工智能系统的更广泛影响进行预测和评估，可以为与部署或治理相关的问题提供信息。然而，这些评估是一个复杂的跨学科挑战。社会风险评估可以评估*产品安全*、*安全漏洞和不必要的外部性，如劳动力和环境影响，* 以及其他问题。这通常涉及在预期产品使用期间可能导致事故的因素，以及解决意外和恶意使用。

3.2 模型性能分析方法

各种利益相关者 (即人工智能开发人员、用户、受影响的人口成员等) 对通用人工智能系统在模型功能和防止负面下游社会影响方面的表现抱有期望。研究人员已经开发了多种方法来比较模型结果与这些预期 (*160*)。此模型性能分析对于了解模型的执行方式以及部署中可能出现的限制，收益或风险是必不可少的。

3.2.1 案例研究

在许多研究论文中，对模型能力的评估是定性的，依赖于模型性能的轶事演示 (*161*) 和人的判断。例如，对图像生成模型的早期评估通常简单地演示的少量示例 (*163 162，*)。当GPT-4新发布 (*99 **) 时，除了传统的基准之外，还使用了一组精选任务的模型输出示例来说明模型性能。现在，几个流行的ben chmarks依靠要求人类评分者对不同模型的反应进行评分 (*165 164，*)。与此同时，风险有时是通过“提升研究”来衡量的，其目的是测试一下当人类能够使用通用人工智能系统时，他们在完成一项潜在有害任务方面的能力有多强，而他们没有这样做 (*166 **)。

3.2.2 基准

大多数机器学习评估都是在标准化基准测量 (*167*) 上完成的。例如，涉及分类的一组相对较小的图像处理基准 (*168，169， 170，171*) 、分段 (*36 *,172*) 和问题回答(*173，174*) 一直是人工智能视觉研究不可或缺的一部分。同样，关于语言建模的研究也受到几个公开基准的影响，这些基准是我ant来衡量一般能力 (*95、96,165，175，176，177 *，178，179*) 和可信度 (*180，181，182*)。最近，基准被设计用来衡量越来越多的通用人工智能能力，这些能力结合了来自多种模式的信息，并使用了网络浏览器 (web浏览器*183*) 等软件工具 (另见4.4.1交叉技术风险因素)。

基准上的**绩效可能是衡量下游任务绩效的不完美的指标。**基准本质上是期望性能的代理度量，并且由于各种有效性挑战，基准的良好分数并不总是转化为实践 (*184*) 中的期望性能。*内部*有效性挑战与基准测量的可靠性有关，即与强基线相比，报告的度量在重复执行中的可靠性如何。例如，如果be nchmark不包含足够的e样本来做出关于模型性能的统计有效声明 (*185*) 或包含错误标签 (*187 186，*)，则会出现*内部*有效性问题。*外部*

有效性是指基准性能转化为现实世界设置的程度。基准本身可能是对现实世界任务的构建不良，不充分或不完整的表示。例如，在如何国家的最先进的模型是pr ompt的限制，并结合其他系统可能会导致他们的能力被低估 (*189 188， 24，*)。目前，基准通常不明确其适用范围。声称 "一般" 性能的基准经常在cu文化r epr中掩盖偏见，表述和注释者差异，基础事实的有争议的概念，以及更多 (*101， 167， 190， 192 191，*)。此外，由于现代通用人工智能模型已经在大量的互联网数据上进行了训练，因此很难将新的功能与记忆的功能分开 (*194 193，*)。

在基准上解释人类的表现可能很困难。，直观地理解模型性能的一个关键方面是与人类性能的比较 (*195 **)。人类绩效指标通常是不可靠的-例如，ImageNet基准测试中 "人类绩效" 的基线由单个研究生的注释 (*196*) 组成。众所周知，"基础真理" 的人类注释者是善变的 (*197*)，由于文化背景，价值观或专业知识的差异，他们经常以有意义的方式不同意 (*198*)。此外，在任务中，人的表现和人的能力之间存在有意义的差异-例如，后者通常涉及对鲁棒性的判断，而不仅仅是准确性 (*199*)。使注释者群体多样化的评估 (*200*)，允许多种真实标签 (*198*)，或适当考虑人类表现声明的背景 (*201*) 往往是更值得信赖的人类与人工智能比较评估。

3.2.3 红色团队和对抗性攻击

在现实条件下部署系统之前，评估人员会使用*对抗性攻击和红色团队*来识别最坏情况的行为、恶意使用机会以及系统意外失败的可能性。在网络安全中，对抗性攻击是指故意试图使系统失败。例如，针对语言模型的攻击可以采取自动生成攻击的形式 (*202，203，204 *，205，206*) 或手动生成ttacks (*204 *，207*)。这些可以包括 "越狱" 攻击，这些攻击会破坏模型的安全限制 (*208、209、210、212 211、*)。

*红队*是指一组旨在通过攻击系统来发现系统漏洞的人。与作为一组固定测试一下案例的基准相比，red-teaming的一个关键优势是它使评估适应正在测试的特定系统。通过与系统的交互，red- teamers可以为或模型设计自定义测试。研究人员通过各种策略和工具接近红色团队。Ojewale等人。(*213*) 规划出一个在人工智能问责过程中利用的工具生态系统，包括用于 "危害发现" 的资源，如 "漏洞赏金" 平台、事件数据库等 (*214*)。这些工具支持识别潜在的损害媒介，并使更广泛地参与损害发现。

但是，评估者有时可能无法代表公共利益。的通用人工智能系统主要由开发它们的组织 (*2 * 、22 * 、204**) 完成。学术界，审计网络和专门的评估组织也可以发挥关键作用 (*215， 217 216，*)。与AI开发人员本身的情况一样，红队评估员并不总是代表公共利益或人口统计数据，并且可以消除偏见，或者在识别或评估与AI相关的危害时忽略重要的考虑因素 (*215*)。尚未建立red- teaming的最佳实践 (*216*)。

在red团队合作期间，通用AI模型中的各种故障仍未被发现。红色团队攻击和对抗性攻击有助于更好地了解模型的最坏情况性能，并评估基准无法充分覆盖的性能。
然而，它们也有局限性。以前对红色团队和对抗性攻击技术进行基准测试的工作发现，错误通常会逃避检测 (*218*)。A real-world例如，当前最先进的通用AI聊天系统的越狱 (*208，209，210，211，212*)，这似乎逃避了设计它们的开发人员的最初检测 (*2 *,22 *,204 **)。总体而言，red-teaming只是有意义地理解通用AI所需的几种评估工具之一

*219*能力 ()。红队也可能无法捕捉到下游的危害，这只有在人工智能系统在社会中得到更广泛的部署时才会出现。这将在4.3进一步讨论。系统性风险。

3.2.4 审计

整个通用AI开发过程中的设计选择会影响最终系统的工作方式。审计提供了一种机制来审查和确保对这些选择负责。不同群体的通用人工智能审计师在收集的证据质量和实现的问责结果方面取得了不同程度的成功 (*215*， *217*)。一项针对一系列人工智能审计方法的调查 (*217*) 建议，在不同维度上独立评估部署的通用人工智能系统 (见图1)，让利益相关者对他们在开发通用人工智能系统及其使用方面做出的选择负责。

对训练数据的分析可以揭示有问题的内容。对训练数据的分析，也称为 “数据审计” (*217*)，是分析关键模型设计选择的一种具体方法，可以揭示有问题的内容。在机器学习开发过程中，收集、整理和注释数据 (*190*)。研究这些数据工程决策如何导致间接伤害，并影响其结果，有助于理解模型及其最终的下游影响。例如，对用于训练现代系统的互联网文本和图像数据的分析已经确定了cop正确的内容 (*220*， *333 221*，)，仇恨语音和模因 (*223*， memes *224*，
225)，恶意刻板印象 (*224*， *226*， *227*)，色情内容 (*223*， *226*) 以及对性暴力的描述 (*226*)，包括虐待儿童的材料 (*228*)。

此外，对培训数据集的调查通常会揭示主流数据源 (*230 229*，) 中某些人群的人口统计，地理和语言代表性不足的问题。此类数据审核为版权法律挑战提供了必要的证据。例如，《纽约时报》对OpenAI的诉讼 (*231*) 大量引用了道奇等人的数据审计。(*232*)，并导致尝试对某些被认为包含不适当材料的数据集进行编辑 (*234 233*，)。然而，现代通用人工智能模型通常是在极其大量的互联网数据集上训练的，这些数据集通常不会公开，因此数据来源仍然是一个系统性的挑战 (*236 235*，)。因此，在大规模的训练数据中系统地搜索潜在有害的示例是具有挑战性的。

对人工智能建模和产品选择的分析可以揭示权衡并指出下游风险。除训练数据外，其他方法选择也可能导致特定问题。
审查模型开发方式的审核通常称为 “过程审核”。例如，基于人类反馈的方法是训练通用AI模型的最先进方法，但c一个导致诸如sycophancy之类的问题(*237*，*238*) 并产生非多样化的输出 (*239*，*240*，*241*)。同样，诸如 “模型修剪” 之类的工程决策可能会对某些测试一下人口统计产生不公平的影响 (*242*)。同样，生成图像模型架构的选择已被证明会影响这些系统在表示不同种族 (*243*) 方面的性能。研究人员分析开发人员的方法是具有挑战性的，因为专有甚至开源通用AI模型的完整开发细节很少被记录和披露 (*244*)。

与此同时，“生态系统审计” 有助于评估人类与人工智能的互动。越来越多的实验室研究调查人类用户如何与通用AI系统交互。这些通常采用对照研究 (*245*) 的形式，其中参与者与模型进行交互，并且直接测量建模和用户交互设置对参与者的决策和行为的影响。这样的stu dies揭示了用户倾向于信任模型输出的某些表示而不是其他表示 (*247 246*，)。但是，由于招募了参与者，因此很难设计研究并招募足够广泛的参与者以有意义地反映实践中用户的全部影响。一些研究人员已经开始对人工智能在实际部署环境中的使用影响进行自然和受控的实验。例如，在美国肯塔基州 (*248*)，有研究关于人工智能风险评估对法官保释决定的影响，使用自动

关于招聘经理自由裁量权的招聘工具 (*249*)，以及在中层经理绩效 (*250*) 上使用生成式人工智能。对利益相关者的定性访谈已被证明可以有效地说明更系统的影响，例如人工智能实施 (*251*) 的一些社会影响，这些影响可以在人工智能系统 (*252*) 移除后持续存在。

在部署后分析现实世界中的人工智能系统，使研究人员能够将它们作为更大的社会系统的组成部分进行研究。上市后监督在其他具有审计生态系统 (*253*) 的行业中相当普遍。用户通常会发现开发人员没有发现的功能和故障模式，并且监视系统的实际使用情况可以进一步科学理解。例如，根据普通用户的发现，首先研究了针对现代大型语言聊天模型的越狱 (*254*)。对现实世界中deepfakes的研究也有助于形成研究和减轻危害的科学研究 (*256 255，*)。

3.3 模型透明度、解释和解释

与研究通用AI模型输出相反，评估模型的另一种常见方法是研究模型产生输出的内部机制。这可以帮助研究人员对模型性能进行上下文评估，并加深对模型功能的理解。研究通用人工智能模型和系统*如何*在内部运行是一个热门的研究课题，产生了数千篇学术论文。旨在提高透明度的研究领域包括文档，第三方访问机制，黑盒分析，解释模型动作以及解释模型的内部工作原理。

文档模板记录做出的决策，并促进运营层面的透明度。目前，提高通用AI模型透明度的最实用方法之一是通过记录和传达定义模型的工程决策。已经提出了几种文档解决方案，以将此类决策传达给更广泛的内部和外部利益相关者。其中一些努力，例如模型卡的开发 (*257*) 已经成功。最近的一项研究表明，“在AI commu nity中广泛使用模型卡” (*258*)。有文档t em板可用于交流数据集实践 (*259， 260， 261*)，更广泛的系统功能 (*262， 263*) 和更广泛的程序
*264*决策 ()。

模型解释和可解释性技术可以提高研究人员对通用人工智能系统内部运行方式的理解。有几种工具允许对通用AI系统进行外部审查，使ext ernal参与者能够直接查询通用AI系统，或者以其他方式获得模型细节的可见性 (*213*)。一种突出的技术方法涉及研究如何将模型的输出解释为给定输入 (*265， 268 266， 267，*) 的结果。这些解释可以在支持问责制方面发挥独特的作用，通过帮助确定责任，在人类可能受到自动化人工智能系统 (*271 269， 270，*) 的错误伤害或歧视的情况下。另一种方法用来研究神经系统中的计算网络orks参与了口译the rolepar的参数 (*272*) 、神经元 (*273， 274，275*) 、子网 (*276，277*) 或图层表示 (*278，279，280，281*) 在人工智能系统内部。
对模型的解释有时有助于研究人员发现漏洞。的例子包括红线组合 (*207*)，ide确定虚假特征的内部表示 (*282*)，b rittle特征表示 (*283， 285， 286*) 以及transformers中事实召回的限制 (*287*)。

，了解通用AI系统如何在内部运行并有效地使用这种理解是具有挑战性的。一个长期存在的问题是，在缺乏客观标准进行比较的情况下，很难确保对通用AI系统如何运行的解释是正确的。已经记录了许多“可解释性幻觉”，其中技术的可解释性暗示了对模型如何工作的误导性解释 (*288 *， 289*)。此外，一些研究还批判性地研究了算法tr非稀疏性工具如何被恶意用于构建错误的二分法，模糊和误导 (*290*)。严格评估

可解释性技术，它们产生的解释需要对一些下游任务 (*291， 292， 293，* 294 *，*295) 是有用的，但人工智能可解释性工具并不总是与更简单的技术竞争，但许多任务还 (*296*)。特别地，用于解释模型动作的不同技术常常彼此不一致，并且在健全性检查或下游使用 (*299 218*、*297*、*298*、) 时失败。尽管如此，可解释性有时改善了实际诊断和理解，特别是随着该领域的最新进展 (*300*)。进一步的进步可以实现更好的应用。然而，对通用人工智能系统的高级解释不能用来对当前模型和方法的行为做出正式保证。辩论了当前可解释性和可解释性技术在实际中有助于严格模型评估的潜力。

3.4 研究通用AI系统的挑战

进行彻底的评估并对通用人工智能的能力和风险做出强有力的保证是极其困难的。现代通用人工智能系统是复杂、分散的项目的结果，涉及数据收集、训练运行、系统集成和部署应用程序。与此同时，通用人工智能系统在现实世界中有非常多的用途。这种复杂性使得任何一个参与者都难以理解整个过程。

评价的质量取决于获取的程度和透明度。评估AI系统的不同技术需要不同类型的访问。例如，评估模型在测试一下数据上的性能通常只需要查询目标模型并分析其输出的能力。这通常被称为 “黑箱” 访问。查询黑盒系统的能力是有用的，但是许多类型的评估技术依赖于更高级别的访问。从历史上看，人工智能研究人员受益于开源方法、模型和数据。然而，今天，公司越来越多地将最先进的通用AI系统保密*244*。
缺乏 “白盒” 访问 (访问模型参数) 使得研究人员执行对抗性攻击，模型解释和微调 (*301 300，*) 具有挑战性。同时，由于缺乏对系统如何设计的信息 (包括数据，文档，技术，实现细节和组织细节) 的 “框外” 访问，因此很难对开发过程进行评估 (*226，227，257，300，302，303*)。与此同时，第三方审计生态系统正在萌芽，但正在增长。各种AI审计工具 (其中一些是开源的) 允许外部用户查询和访问模型 (*213*) 的详细信息。一些研究提倡合法的 “安全港” (*304*) 或政府介导的访问制度 (*253*)，以实现独立的红色团队和审计工作。已经提出了结构化访问的方法，该方法不需要使代码和权重公开 (*305*)，而是使独立的研究人员和审计员能够在旨在避免泄漏的安全环境中完全访问模型来执行他们的分析。

彻底评估下游的社会影响需要细致入微的分析，跨学科性和包容性。尽管了解整体社会影响是许多人工智能评估的最终目标，但许多评估都没有达到这一目标。首先，研究人员研究人工智能系统的设置与他们使用的不断变化的现实世界设置之间总是存在差异。其次，评估人工智能对社会的影响是一个复杂的社会技术问题 (*307 306，*)。例如，虽然已知大型语言模型表现出显著的
跨语言的安全差异(*308，309*) 、能力 (*310，311 **) 和倾向(*312，313 **)，对于研究人员来说，彻底评估跨语言的语言模型是一项挑战。此外，在处理 “公平” 和 “公平” 等道德概念时，过度依赖简化的技术代理可能会误导或排除代表性不足的利益相关者 (*315 314，*)。对通用人工智能更广泛影响的评估也是高度多方面的，需要跨学科和代表可能持有不同观点的多个利益相关者 (*316*、*317*、*318**)。
在部署之前，对人工智能在社会中的部署效果进行建模本质上是复杂的，很难在定量分析中进行锚定。
虽然在社会影响方面取得了进展

通用AI评估 (*319*、*320*、*321*、*322 **、*323**)，由于需要平衡多个利益相关者的利益 (*324*) 和资源挑战 (*325*)，这通常使这项工作在实践中难以完成，因此实施仍然具有挑战性。

在人工智能开发和评估过程中越来越多地参与和代表观点是一项持续的技术和制度挑战。指定相关的评估目标在很大程度上受到谁在桌子上以及如何组织讨论的影响，这意味着很容易错过或错误定义关注的领域。扩大参与审核过程的人员的范围也扩大了发现和表征当前或预期危害的过程中包含的经验范围 (*215*)。近年来，扩大参与一直是机器学习社区的重点，强调需要将更广泛的观点和st持有人纳入人工智能模型设计、开发、评估和治理过程 (*328 327， 326，*)。在实施影响评估期间，从促进更广泛的影响概念 (*329*) 提出了多种策略，以实现更具包容性的人类反馈 (*330*)。然而，寻找和吸收更多的声音是一项微妙的努力，需要敏感和尊重参与方，以尽量减少剥削的可能性 (*331*)。至关重要的是，增加参与的挑战还inv解决在不相容的价值观或优先事项 (*332， 334 333，*) 之间进行艰难选择的必要谈判。

虽然透明度对于评估人工智能系统至关重要，但在实践中很难实现。
透明度的努力并不总是对问责制 (*290*) 有意义的干预。Bhatt等人.
(*335*) 调查了数十家公司实施的可解释性技术，发现许多可解释性工作在模型开发过程中得到了有意义的使用，但很少达到提供最终用户透明度或合理性的政策愿望。此外，最近对专有，开源可解释性和可解释性工具的调查显示，此类工具很少经过充分验证以支持声明，并且容易受到操纵和健壮性挑战的影响 (*213， 337 336，*)。同样，在公司环境中实施文档实践的后勤工作可能充满内部政治 (*338*)。

4 风险

通用AI的开发和部署会带来一些风险，本节将对此进行讨论。本报告区分了“风险”和“交叉风险*因素*”。就本报告而言，“风险”是指发生伤害的可能性和该伤害的严重程度的组合 (*339*)。“交叉风险因素”是指导致不是一个而是几个风险的条件。

4.1 恶意使用风险

由于通用人工智能涵盖了广泛的知识领域，它可以被重新用于恶意目的，可能会造成广泛的伤害。本节讨论了恶意使用的一些主要风险，但还有其他风险，新的风险可能会继续出现。虽然本节中讨论的风险范围广泛，但在某些情况下，有证据表明它们目前可能根本不是严重的风险，但我们将它们包括在内，以全面概述与通用AI系统相关的恶意使用风险。

4.1.1 通过虚假内容对个人造成伤害

关键信息

- 通用AI系统可用于增加诈骗和欺诈的规模和复杂性，例如通过通用AI增强的“网络钓鱼”攻击。
- 通用人工智能可用于生成虚假的妥协内容，未经个人同意，对个人隐私和声誉构成威胁。

通用人工智能可以放大欺诈和诈骗的风险，增加其数量和复杂性。它们的数量可以增加，因为通用AI有助于以比以前更大的速度和规模生成诈骗内容。它们的复杂程度可以提高，因为目的的人工智能有助于大规模创建更具说服力和个性化的诈骗内容 (*341 340，*)。通用AI语言模型可用于设计和部署“phis hing”攻击，其中攻击者欺骗人们共享密码或其他敏感信息 (*342*)。这可能包括鱼叉式网络钓鱼，一种针对目标个性化的网络钓鱼活动，以及商业电子邮件妥协，一种恶意用户试图欺骗某人汇款或共享机密信息的网络犯罪。研究发现，从1月到2023年2月，电子邮件帐户样本 (*343 **) 中的“新型社交工程攻击”135% 增加，这被认为与ChatGPT的广泛采用相对应。

通用人工智能语言模型有可能通过与潜在受害者的自动对话来寻找有希望的目标 (*340*)，从而大幅扩大欺诈者的范围。通用人工智能也可能导致有针对性的身份盗窃，并产生可用于非法目的的假身份。例如，几年前，研究讨论了AI“语音克隆”的潜在风险-AI系统分析音调，音高，以及人类声音的其他特征来创建合成副本-以及欺诈者如何使用这种技术来假装自己是受害者的朋友或可信赖的权威机构 (*344*)。通用人工智能系统还可以帮助犯罪分子通过纠正语言错误和提高邮件的流畅性来逃避安全软件，否则这些邮件可能会被垃圾邮件过滤器捕获。

最近的数据表明，人工智能欺诈，特别是使用deepfakes的欺诈，在全球范围内*345*， *346*。检测通用人工智能系统的使用来生成用于欺诈的内容可能很困难，而且机构可能不愿意透露他们在人工智能欺诈 (*347*) 方面面临的挑战。

通用人工智能生成的虚假内容也可以被用来伤害个人，在未经他们同意的情况下将他们呈现在内容中，从而侵犯他们的隐私权，损害他们的声誉或尊严。这可能以虚假内容的形式发生，包括任何损害或声誉损害活动，但在deepfake色情案件中受到特别关注，其中通用AI用于在未经个人同意的情况下创建个人的色情视听内容。这可能包括创建儿童性虐待材料 (CSAM) 和其他 ，用于用于虐待的图像，例如，前家庭伴侣或勒索 (*348*， *350 349*，)。

4.1.2 虚假信息和舆论操纵

关键信息

- 通用人工智能可以以前所未有的规模和高度复杂的方式生成和传播虚假信息，这可能对政治进程产生严重影响。然而，人们对政治虚假信息运动的影响力普遍存在争议。
- 检测通用AI产生的信息可能很困难，因为输出越来越现实。技术对策，如水印内容，是有用的，但通常可以被适度复杂的参与者规避。

人工智能，特别是通用人工智能，可能被恶意用于虚假信息 (*351*)，就本报告而言，这是指故意误导或欺骗而产生或传播的虚假信息。Gen era-pu rpose AI生成的文本可能与真正的人类生成的材料 (*353 352*，) 无法区分，并且可能已经在社交媒体上大规模传播 (*354*)。此外，通用人工智能系统不仅可以生成文本，还可以生成完全合成或误导性修改的图像、音频和视频内容 (*355*)。人们经常发现这样的内容与真实的例子无法区分，并且生成这样的内容既相对简单又非常便宜 (*356*)。例如，使用通用AI或更窄类型的AI系统 (*357*) 更改或完全生成的人脸图像。这种 “深度伪造” 的图像和视频被认为在最近几个月的几次全国选举中被用来诽谤政治对手，可能会产生重大影响，但目前没有太多科学证据证明这种运动的影响。

通用人工智能工具可能会被用来说服和操纵人们，这可能会对政治进程产生严重影响。通用AI系统可用于大规模生成极具说服力的内容。例如，这可以在商业环境中用于广告，或者在竞选期间用于影响公众舆论 (*358*)。最近的工作衡量了通用人工智能生成的政治信息的说服力，发现它可以在一定程度上影响阅读这些信息的人的观点 (*360 359*，)。此外，通用人工智能可以用来为特定的个人或人口统计定制有说服力的内容 (称为 “微目标”)，例如，通过使用从社交媒体上抓取的信息。然而，目前只有有限的证据表明，微目标信息比通用的人工智能制作的内容 (*361*， *362*) 更具说服力，这与人们普遍对微目标有效性的怀疑相一致 (*363*)。

一个未被充分探索的前沿是使用会话式通用人工智能系统来说服用户进行多次对话。在一项研究中，人类-人类或人类-通用人工智能对在意见和反驳的循环中进行了辩论，通用人工智能系统被发现和人类一样有说服力 (*364*)。在另一项研究中，通用人工智能系统试图在三次对话中劝阻人类相信阴谋论，发现人们对那些持续长达两个月 (*365*) 的15至20% 理论的信仰有所减少。
这些结果提高了在对话环境中，通用AI系统可以用于非常强大的说服力的可能性。随着通用人工智能系统能力的增强，恶意使用它们进行欺骗或操纵手段可能会变得更加容易，甚至可能比熟练的人类更有效，以鼓励用户采取违背自己最大利益的行动 (*366， 367**)。在这样做的时候，他们可能会利用新的操纵策略，而人类没有做好准备，因为我们对操纵的防御是通过影响其他人类 (*368*) 的尝试而发展起来的。

总体而言，虚假信息运动的总体影响以及通用AI生成媒体的广泛传播的影响仍未得到很好的理解。尽管有迹象表明公共话语存在潜在的严重风险，以及通用人工智能带来的信息生态系统的完整性，但还是有一些警告。首先，缺乏关于大规模虚假宣传活动有效性的证据 (无论是否使用通用人工智能)。其次，一些专家认为，试图对虚假内容产生大规模影响的演员的主要瓶颈不是生成内容，而是大规模分发 (*369*)。
同样，一些研究表明，“廉价广告” (操作视听内容的不太复杂的方法，这些方法不依赖于通用人工智能的使用) 可能与更复杂的deepfake (*370*) 一样有害。如果这是真的，这将支持这样的假设，即虚假内容的质量目前对虚假宣传活动的成功没有那么决定性，而是围绕将内容分发给许多用户的挑战。像Meta或X这样的社交媒体平台采用各种技术，如人类内容审核和标签，以减少可能是虚假信息的内容的覆盖范围，包括通用AI生成的虚假信息。另一方面，多年来的研究表明，社交媒体算法通常优先考虑参与度和病毒式传播，而不是内容的准确性或真实性，这可能有助于人工智能生成的虚假信息的快速传播 (*372 371，*)。

一般来说，随着通用人工智能能力的增长，越来越多地用于大规模地生成和传播信息，无论是准确的、故意虚假的还是无意虚假的，人们可能会越来越不信任任何信息，这可能会给公众审议带来严重的问题。恶意行为者可以通过否认真实的、不利的证据的真实性来利用这种普遍的信任丧失，声称它是人工智能生成的，这种现象被称为 “骗子的红利 ” (*373*)。

识别通用AI生成内容的潜在措施 (如水印) 可能会有所帮助，但对于中等复杂的参与者来说很容易规避。研究人员已经采用了各种方法来试图识别潜在的AI作者 (*375 374，*)。*内容分析*会探索文本的统计特性，例如异常的字符频率或不一致的句子长度分布，这可能会偏离通常在人类写作中观察到的模式 (376， 378 377，)。语言分析技术检查风格元素，如情感或命名实体识别，以发现指示AI生成的不一致或不自然的语言模式 (380 379，)。可读性分数还可用于检查通用AI生成的文本在Flesch Reading Ease分数等指标上与人类编写的内容 (381) 相比可能异常高或低的地方。人工智能研究人员还提出了其他的错误信息检测方法，例如水印，其中不可见的签名识别由人工智能生成或更改的数字内容 (382)。然而，对于中等技能的演员来说，当前的技术相对容易规避 (374)，并且对水印去除的完美防御可能是不可能的 (383)。尽管如此，这些保障措施可能会阻止相对不成熟的威胁行为者 (384)。5. 技术方法，以减轻风险，提供了一个更深入的讨论水印技术。

4.1.3 网络犯罪

关键信息

- 通用人工智能系统可以提升个人的网络专业知识，使恶意用户更容易进行有效的网络攻击，并提供可用于网络防御的工具。通用AI系统可用于自动化和扩展某些类型的网络操作，例如社交工程攻击。
- 目前还没有实质性证据表明通用人工智能可以自动执行复杂的网络安全任务，这可能会在网络攻击者和防御者之间取得平衡，有利于攻击者。

通用人工智能系统可以通过多种方式加剧现有的网络安全风险。首先，它们可能会降低更复杂的网络攻击的进入门槛，因此能够进行此类攻击的人数可能会增加。其次，通用人工智能系统可以通过提高自动化水平和效率来扩大攻击性网络行动。

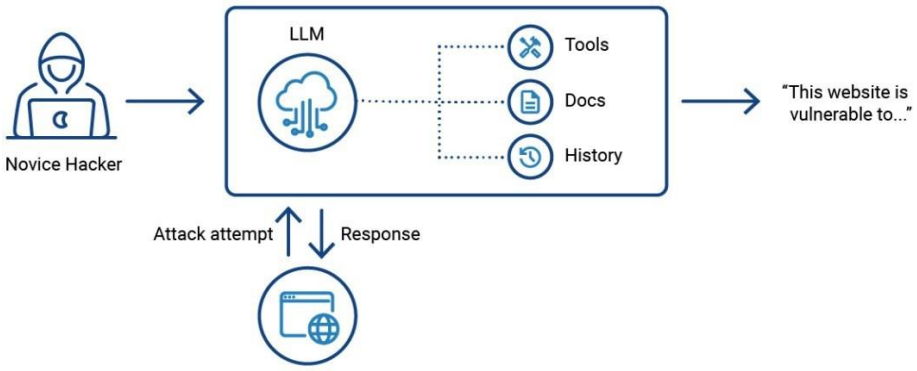
此外，通用人工智能可能会无意中使系统容易受到传统网络安全攻击。例如，通用AI系统被普遍地用作编码助手，并且可能会无意中引入软件漏洞 (*385， 387 386 *，*)。

通用人工智能系统可以降低进行网络攻击所需的成本、技术诀窍和专业知识。的攻击性网络操作包括设计和传播恶意软件，以及发现和利用关键系统中的漏洞。它们可能导致重大安全漏洞，例如在关键国家基础设施 (CNI) 中，并对公共安全构成威胁。鉴于这些操作的劳动密集型性质，先进的通用人工智能可以自动化流程的某些方面，减少所需的专家数量，降低所需的专业知识水平，可能对攻击者有用。

到目前为止，已证明当前的通用AI系统能够自主执行基本的网络安全挑战和狭窄的网络任务，例如入侵高度不安全的网站 (*388，389 *，390，391*)。然而，尽管进行了持续的研究努力，但现有模型似乎无法执行需要更长时间规划的多步骤网络安全任务 (*367 *， 392 391，*)。

，鉴于llm已经可以处理和操纵一些网络安全概念 (*393*)，规划可以解锁更复杂的网络能力，例如独立导航复杂的数字环境，大规模识别和利用漏洞，以及执行长期战略，而不需要直接的人类指导。

图6。一种支持AI的管道，用于自动执行针对网站的简单攻击 (来自 (391))。新手黑客为通用AI模型提供安全信息源和教科书，并提示其与目标网站进行交互。通过试验和错误，通用AI模型可以识别在该网站上起作用的攻击，并将详细信息返回给黑客。



通用人工智能系统如果应用于相关领域，可以提高网络防御能力。尽管进攻能力有所提高，但对手也不会不可避免地取得成功，因为通用人工智能系统的进步也将增强防御能力 (*395 394，*)。例如，通用人工智能系统可以显著减少人类识别和修复漏洞 (*397 396 *，*) 的时间。然而，与其攻击能力类似，现有模型的网络防御优势也有局限性 (*367 *， 398， 399， 401 400 *，*)。考虑到在数字基础设施中实施安全补丁所需的人力和时间，攻击者仍然有机会在尚未修复的系统中成功入侵。

随着通用人工智能系统的改进和更广泛的使用，攻击者和防御者之间的动态可能会受到组织因素的影响，如资源可用性和专业知识水平。主要的cy ber防御竞赛，如 (*402 **)，以及高质量的数据集和基准 (*403*)，可以推动更复杂和更有弹性的网络安全措施的发展。

4.1.4 两用科学风险

关键信息:

- 通用人工智能系统可以加速一系列科学工作的进展，从培训新科学家到实现更快的研究工作流程。虽然这些功能可能有许多有益的应用，但一些专家表示担心它们可能被用于恶意目的，特别是如果在采取适当对策之前不久就开发了进一步的功能。
- 用于生物用途的通用人工智能系统目前没有明确的威胁，未来的威胁很难评估和排除。在生物学领域，目前的通用人工智能系统显示出不断增长的能力，但有限的研究没有提供明确的证据表明，目前的系统可以 “提升” 恶意行为者，从而比使用互联网更好地获得生物病原体。没有足够的公开研究来评估近期的进展是否会提供这种提升，例如通过故障排除动手实验室工作。
- 由于科学工作不足，本中期报告未评估恶意使用导致化学，放射性和核风险的风险。

通用人工智能系统可以通过两种途径推测性地促进生命科学中的恶意使用: 首先，通过提供更多与恶意使用相关的信息和专业知识，其次，通过增加能力上限，这可能会导致现有威胁的更多有害版本的发展，或者最终导致新的威胁 (*405 404，*)。

当前能力

增加对信息的访问-先进的通用AI模型可以提供科学知识，逐步的实验协议以及故障排除实验的指导，所有这些都可能被恶意目的利用 (*405， 406， 408 407，*)。然而，鉴于其 “双重用途” 性质 ，与生物威胁产生有关的信息已经可以广泛获得 (*410 409，*)。目前，通用人工智能系统通过提高其访问与互联网等现有资源相关的信息的能力，“提升” 恶意行为者恶意使用生物学的能力尚不清楚。很少有实证研究评估当前的一般-

目的人工智能模型提供了一个提升，目前的证据是混合的 (*411 166 *，*)。量化隆升的方法也是新生的，并且面临限制。[4]

增加获得实践专业知识的机会- 仅获得信息不足以实现生物恶意使用。恶意行为者还必须能够成功地合成，武器化和交付生物制剂。从历史上看，这些过程需要专业知识和实践技能来完成实际的实验室任务 (*413 412，*)。专家认为，生物工程与通用人工智能的进步相结合可能降低了这些障碍 (*414*)。然而，现有的研究还没有评估通用人工智能系统是否会提升恶意行为者的实际实验室研究任务。现有的高级通用AI模型具有一定的设计和排除实验室实验故障的能力 (*407*)。可以通过为LLMs配备访问web搜索和专业计算工具等工具 (*415*) 的能力来增强这些功能。
当连接到实验室机器人或云实验室时，llm已被实验用于直接指示这些平台进行实验 (*189*)。然而，需要进一步的实证工作来确定这些现有的通用人工智能系统的能力是否相对于互联网的能力 “提升” 了实际实验室研究任务的参与者。

增加能力上限- 本报告侧重于通用人工智能系统 (1中定义的。引言)，两用科学风险也受到狭义AI工具的存在以及通用AI系统与这些狭义AI工具交互的能力的影响。狭义的人工智能生物工具已经可以重新设计现有的蛋白质，以增强功能的exis (*416*)，赋予新的生物功能 (*417*)，以及产生新的蛋白质 (*419 418，*)。类似的设计能力已经在包括化学 (化学) 在内的其他科学领域得到了*420*。狭义工具本身的功能具有双重用途: 例如，它们用于预测可能证明免疫逃避 (*421*) 或产生合理的新型毒性分子 (*422*) 的病毒突变。狭义的人工智能工具也经常广泛使用，这使得实施有意义的保障措施具有挑战性 (*43， 419， 423**)。此外，通用人工智能系统能够使用语言指令指导实验室机器人，并利用专门的化学计算工具 (*189， 424 415，*)。

预计未来能力

未来通用人工智能系统的恶意使用潜力可能会受到一系列预计进展的影响。其中包括: 模型能力的进步，通用AI系统与狭义AI工具的集成以及通用AI与自动化实验室设备的集成。尽管有一些证据表明这些预期的进步正在实现，但这些功能是否会提升用户相对于现有资源的能力仍然非常不确定。

进步通用人工智能模型能力-未来的通用人工智能系统将具有更强的特定领域知识、推理能力和制定复杂计划的能力。与使用互联网搜索相比，专家们不同意通用人工智能系统可以在多大程度上解决实际的实验室任务。人们一致认为，通过 “mu ltimodal” 通用人工智能系统，可以整合来自图像和视频的信息 (*405， 425 407，*)，可以为实际的实验室任务提供更有效的实时故障排除。然而，这种可能性尚未得到检验。此外，一些研究人员认为，在生物数据上训练的通用人工智能系统最终可能会胜过为特定生物应用程序 (*426*) 设计的较窄的工具，但通用人工智能系统尚未证明这种能力。通用人工智能系统在生物领域的未来性能和可扩展性仍不清楚。

[4]例如，使用统计显著性作为衡量模型风险的有效措施存在不确定性，并且操作增加信息访问实际上是危险的阈值具有挑战性。

集成狭义工具- 通用人工智能系统虽然可以使用狭义人工智能工具，但到目前为止，这种类型的集成是有限的 (*189， 415， 427**)。不确定通用AI系统与狭窄AI工具的集成将如何影响恶意使用借条的风险。
如今，狭义的人工智能工具需要专业知识才能有效地用于科学研究 (*428*)。虽然使用自然语言指导这些工具的能力将使专业任务更容易获得，但利用这些输出可能仍然需要技术专长，直到通用人工智能取得实质性进展。现有工具的输出也有限，需要大量的实验室验证，目前尚不清楚克服这些限制的程度。

自主科学能力- 先进的通用人工智能系统已经实现了一些自主科学能力，但目前尚不清楚对潜在的生物恶意使用的短期影响。虽然化学合成等化学工作流程的各个方面已经可以自动化 (*415 189，* 但由于涉及生命系统的自动化工作面临挑战 (*407*)，专家们不确定这些进展是否会很好地转移到生物学工作流程。此外，自动化实验室的高成本可能会使除了最先进的恶意行为者之外的所有人都无法访问大规模自动化，尽管商业云实验室可以部分抵消这一点。

总而言之，目前最先进的通用人工智能系统在多大程度上增强了恶意行为者在互联网等现有资源上使用生命科学的能力，目前尚不清楚。，尽管一些实证工作已经评估了信息获取和生物威胁方面的这种提升 (*166 *， 429*)，但还需要进行更多的研究来评估更广泛的任务和科学领域，以更深入地了解这个问题。目前还没有研究通用人工智能系统是否也可以加强对两用科学危害的防御。

4.2 故障风险

4.2.1 产品功能问题带来的风险

关键信息

- 当对通用AI模型或系统的功能存在混淆或错误信息时，就会出现产品功能问题。这可能会导致不切实际的期望和
  过度依赖通用人工智能系统，如果系统无法提供预期的功能，可能会造成伤害。
- 这些功能上的误解可能源于评估人工智能模型本身的真实能力的技术困难，或者预测其作为更大系统的一部分时的性能。广告和传播中的误导性声明也可能导致这些误解。

如果通用人工智能模型和系统不符合产品安全和产品功能的一般原则，可能会产生风险。与许多产品一样，通用人工智能产品的风险是由于对功能的误解以及对适当和安全使用的指导不足。在这方面，通用的基于人工智能的产品可能没有什么不同 (*430*)。

产品功能问题及其带来的风险可能由潜在的故障模式聚集在一起 (见表1)。

*不可能的任务*源于尝试使用通用AI系统实现目标的实例，该系统超出了通用AI系统的功能。很难确切地说在现代环境中什么是不可能完成的任务。从历史上看，大型语言模型无法考虑训练结束后发生的事件或发展。然而，

使人工智能产品能够从数据库中检索信息，提高了他们考虑训练后发生的事情的能力 -- 尽管模型在需要新信息 (*431*) 的测试中表现仍然较差。另一个可能不可能完成的任务可能是需要本质上无法访问的数据的任务-例如以可计算媒体的格式不存在的信息，或者由于法律或安全原因而无法进行培训的数据。

不可能完成的任务会带来风险，因为通常情况下，突出类型的故障-包括许多工程故障，部署后故障和通信故障 (见表1) -可能是错误测量，误解或误解的副产品。模型可以做什么，以及导致错误的部署。

例如，GPT-4模型取得了 "通过模拟律师考试，得分在前10% 名测试一下考生左右" 的成绩，并且在LSAT测试一下考生 (*2 **) 中排名第88位。对这一结果的信心甚至导致一些律师将该技术用于其专业用途 (*432*)。在不同的情况下，例如更改测试一下设置或与通过考试的首次酒吧考生进行比较时，该模型的百分位数结果要低得多 (*433*)。
那些试图在实际法律实践中使用该模型的人遇到了这些不足之处，因为这些模型产生的错误 (即不准确的法律引用、不适当的格式和措辞等) 而面临严重的专业后果。(*434*)。关于模型性能的类似误解被认为适用于医学背景 (*435*)，现实世界的使用和重新评估揭示了这些模型的复杂性，这些模型包含可靠的临床知识 (*436*) 或通过MCAT (*2**) 或USMLE (*437*) 等医学测试。更一般地说，一些部署的大型语言模型在某些语言环境下会遇到困难: 例如，它们可能难以导航否定，因此无法区分支持和反对行动的建议-尽管一些研究表明这些问题可以通过一般能力增益 (*439 438，*) 来解决。

一些缺点仅在部署后才显露出来。尽管许多彻底的评估已经检查了大型语言模型在代码生成中的使用 (*440 **)，包括在相关的实际任务中 (*441*)，实际部署用于编码的大型语言模型的实例，这些模型的使用可能导致潜在引入关键的被忽视的错误 (*442*)，以及在指导工程程序员时可能特别有影响的混乱或误导性的编辑 (*443*)，特别是在自动化部分工作流 (*444*) 的应用程序中。

失效模式	**类别**
不可能完成的任务	概念上可能实际上不可 能
工程故障	设计失败实施失败缺少安 全功能
部署后失败	健壮性问题 对抗性攻击下的失败意外交互
通信故障	伪造或夸大的能力

表1: 人工智能功能问题的分类，经许可复制 (430)。

功能误解源于不同的潜在问题。首先，如3。
方法论为了评估和理解通用AI系统，在设计和对通用AI系统的性能进行代表性评估方面存在技术困难，从而难以确定功能。其次，功能问题可能只会显现或

在现实世界的现实环境中表现不同，即使提供信息模型评估也不足以对通用AI系统和产品功能进行稳健的陈述。第三，失败不仅可能是由于评估不足，还可能是由于缺乏与产品用户就产品的局限性和潜在后果进行适当的沟通。误导性广告，因为它发生在许多市场，可能成为一个重大的风险来源的功能在通用人工智能 (AI *445*)。

通常，对于许多基于机器学习的产品，可能不清楚哪个部署上下文在数据中很好地表示并且适合于模型。然而，更通用的人工智能工具比能力较低或更窄的人工智能系统更难以审查部署准备情况: 使用通用人工智能，可能很难明确定义和限制可能不合适或可能不成熟的潜在用例。尽管在限制用例方面取得实质性进展是可行的。

4.2.2 偏见和代表性不足的风险

关键信息

- 通用人工智能系统的输出和影响可能在人类身份的各个方面存在偏见，包括种族、性别、文化、年龄和残疾。这给高风险领域带来了风险，如医疗保健、工作招聘和金融贷款。
- 通用人工智能系统主要在语言和图像数据集上进行训练，这些数据集不成比例地代表了英语和西方文化，增加了对这些数据不能很好地代表的个人造成伤害的可能性。

人工智能系统中的有害偏见和代表性不足一直是挑战，早在人们对通用人工智能的关注增加之前。它们仍然是通用AI的一个问题，并且在可预见的未来可能会成为通用AI系统的主要挑战。如果人工智能的决策基于受保护的特征 (如性别、种族等) 而扭曲，那么他们的决策可能会有偏见。因此，当这种偏见使决策对这些受保护群体的成员不利时，它们可能是歧视性的; 从而损害公平。本节讨论了人工智能中由偏见和代表性不足风险导致的当前和未来风险。由于这一领域丰富的研究历史，本节探讨了狭义人工智能和
通用AI。

人工智能系统可能会因训练数据倾斜、模型开发过程中做出的选择或过早部署有缺陷的系统而表现出偏见。尽管进行了广泛的研究，但完全减轻任何歧视的可靠方法仍然难以捉摸。人们特别担心先进的通用人工智能系统会复制和放大其训练数据 (*446*) 中存在的偏见。这在工作招聘、金融贷款和医疗保健 (*447*) 等高影响力的应用中构成了很大的歧视风险。在这些领域，通用人工智能系统的输出可能会对个人产生深远的负面影响，可能会限制就业前景 (*449 448*，)，阻碍向上的金融流动性，并限制获得基本医疗服务 (*451 450*，)。

有几个有据可查的人工智能系统案例显示基于种族、性别、年龄和残疾状况的歧视行为，造成重大伤害。鉴于人工智能系统在各个部门的应用越来越广泛，这种行为可能会延续各种类型的偏见，包括种族、性别、年龄和残疾。如果这些系统被赋予越来越高风险的决策，这可能会对个人造成严重后果，这可能会造成严重损害。人工智能系统中的种族偏见已被证明存在于商业上可用的面部识别算法 (*452*) 中，并导致在预测累犯结果方面无效

有色人种的被告，对边缘化种族和民族背景的患者需求的低估 (*454 453，*)，以及在文本生成模型的反应中不适当的基于种族的医学的延续 (*450 435，*)。

人工智能系统输出中的性别偏见是另一个关键问题。研究发现，通用AI (*456 455，*) 产生了性别歧视，女性歧视和性别刻板印象的内容，而使用窄AI alg orithms (*457*) 进行的性别中立的互联网搜索则产生了男性主导的结果。年龄偏见也是一个关键问题: 一些人工智能系统对年长的求职者表现出偏见 (*458*)，而年龄偏见出现在情感分析模型 (*459*) 的一些输出中。其中一个原因可能是训练数据中的偏差。例如，LLM驱动的人力资源筛选工具可能会针对偏向年轻员工的简历进行培训，这些简历可能会无意中打折年长申请人的经验和技能。类似地，由健康保险公司开发的医疗保健分配算法可能基于与年龄相关的健康风险而不利于年长的个体，即使这些个体是健康的。贷款算法可能无法适当处理老年人的财务状况，特别是在可能影响批准结果的社会保障收入方面 (*460*)。

研究还表明，人工智能系统和工具可能会歧视残疾用户，例如，按比例拒绝有复杂医疗需求的残疾个人的保险索赔 (*461*)，复制社会对残疾的刻板印象 (*462*)，以及不准确地对残疾人的情绪进行分类 (*463*)。尽管对手语r ecognition (*464*) 的研究越来越多，但人工智能系统对手语使用者 (*143*) 的自动转录能力有限，手语数据集的有限多样性也可能加剧高级通用人工智能系统的残疾偏见。因为大多数手语数据集代表美国手语。例如，最近的工作为六种非洲手语开发了数据集
(*465*) 是朝着实现更公平地纳入手语方言迈出的一步，尽管是适度的一步。

人工智能系统表现出交叉偏见的趋势。交叉性描述了具有多个边缘化特征的个人如何经历复合偏见或歧视(例如，有色人种的低收入妇女)。交叉偏见加剧了现有的社会不平等，并可能限制个人获得重要资源和机会的机会。虽然有一个新兴的研究领域专注于开发检测人工智能模型中交叉偏差的方法 (*466， 468 467，*)，但在减轻潜在影响方面进展较少 (*469*)。

通用AI系统输出中的**偏差可能是由于训练数据集中缺乏表示形式而导致的，从而导致输出偏差。**不同的群体和不同的文化在人工智能生命周期的不同阶段 -- 从输入数据到语言和图像生成模型的输出 -- 都有不平等的表现。与人工智能中的偏见相关的许多问题都证明了训练数据集中的有限表示的危害，这些数据集极有可能是英语 (*236*)。这导致了主要的di spa问题，我了不同社会中现代通用人工智能系统的安全性和可靠性 (*309，310， 313**)。用于训练人工智能模型的数据集也被证明不能充分代表各种人口统计指标，如年龄、种族、性别和残疾状况等 (*471 470，*)。人工智能语言模型在训练中主要依赖数字化书籍和在线文本，无法反映口头传统和非数字化文化。这些来源中固有的历史偏见，c与数据收集过程中潜在的不平等相结合，可能会延续系统性不公正 (*472 **)，并引导人工智能系统反映主流文化、语言和世界观，损害土著社区等边缘化群体(*196，230，239，473，474*)。

偏见和代表性问题仍然是一个未解决的问题。尽管文献和经济激励措施给予了公司相当大的关注，以避免有偏见的人工智能产品对声誉造成损害，但减轻偏见仍然是一个尚未解决的挑战。虽然开发人员可能会尝试在模型微调期间显式地解决偏见，但AI模型仍然可以获取特征之间的隐式关联 (*475*)，或者即使提示不包含显式的人口统计标识符 (*476*)，也可以使显著的偏见和刻板印象永久化。虽然人类反馈强化学习 (RLHF) 等方法旨在使模型决策与人类偏好保持一致，但这些方法可能会无意中引入基于人类提供反馈的多样性和代表性的偏见 (*303*)。RLHF已被证明会导致更多的政治偏见模型 (*477 239，*) 和

将用户信念纳入事实信息 () *238*。此外，评分者的反馈往往不一致 (*479 478*，)。理解数据集、模型和评估方法 (如RLHF) 中的表示问题至关重要，因为偏斜的表示可能导致AI模型中的输出有偏差。需要更多的研究来解决这些问题。

4.2.3 失控

关键信息

- 正在进行的人工智能研究正在寻求开发更有能力的 “通用人工智能代理”，即可以自主与世界互动、提前计划和追求目标的通用人工智能系统。
- “失控” 情景是潜在的未来情景，在这种情景中，社会不再能够有意义地限制一些先进的通用人工智能代理，即使它们显然正在造成伤害。假设这些场景是通过社会和技术因素的结合而产生的，例如将决策委托给通用AI系统的压力，以及用于影响通用AI系统行为的现有技术的局限性。
- 人工智能专家普遍认为，由于能力有限，目前已知的通用人工智能系统不会造成重大的失控风险。
- 一些专家认为，失去控制的情景是不可信的，而另一些专家认为它们是可能的，一些专家认为它们是低可能性的风险，由于其严重性高，值得考虑。
- 这种专家分歧很难解决，因为目前还没有一个商定的方法来评估失去控制的可能性，或者何时可能开发相关的人工智能能力。
- 如果失去控制的风险实际上很大，那么解决这一风险可能需要在人工智能安全的某些技术问题上取得根本性进展。目前尚不清楚这一进展是否需要多年的准备工作。

AI公司和研究人员对开发通用AI “代理” (有时也称为 “自主通用AI系统”) *越来越感兴趣*。通用人工智能代理是可以自主与世界互动、提前计划和追求目标的系统。虽然通用人工智能代理已经开始开发，但它们仍然只展示了非常有限的能力 (*26， 178， 480**)。各种研究人员和AI实验室最终希望创建通用AI代理，这些代理可以在很少或没有人为监督或干预的情况下操作和完成长期任务。

自主通用人工智能系统，如果完全实现，可能在许多领域都很有用。然而，一些研究人员担心他们的恶意使用的风险，或从他们的部署事故和意外后果 (*482 481 *，*)。

一些研究人员还对社会对自主通用人工智能系统进行可靠监督和控制的能力表示担忧。几十年来，计算机科学家一直在关注这类人工智能系统，包括人工智能先驱艾伦·图灵 (Alan Turing，*483*) 、I等。J.好 (*484*) 和诺伯特·维纳 (*485*)。这些担忧最近变得更加突出 (*486*)，部分原因是一部分研究人员现在认为，先进的通用人工智能代理可以比以前认为的更快地开发出来 (*127， 488 487，*)。

这种风险的合理性仍然存在很大争议。本节旨在阐明拟议风险的性质，概述目前告知研究人员关于其可能性的观点的主要论点和证据，并总结目前的专家意见。

危险因素

当人工智能系统的行为可以由人类有意义地确定或约束时，它被认为是“可控的”。虽然缺乏控制本身并不有害，但它会大大增加各种危害的风险。目前的通用人工智能系统通常被认为是可控的，但是，如果自主通用人工智能系统得到充分发展，那么失去控制的风险可能会大大增加。

在4.4。跨领域风险因素本报告讨论了危险AI系统的技术和社会风险因素。在本节中，我们将详细介绍这些因素如何影响失控风险。总体而言，失去对当前已知通用AI系统的控制的风险似乎可以忽略不计。未来，可能更有能力，自主的通用人工智能代理可以在多大程度上被控制仍然不清楚。

，目前还不知道未来是否会更容易或更难确保功能强大的人工智能系统实现其开发人员的目标。一个原因是，人工智能系统可以通过以意想不到的和潜在的有害方式实现目标来“游戏”他们的目标。例如，广泛部署的通用人工智能语言模型调整其陈述的观点，以更好地匹配用户的观点，无论事实如何，从而获得更多的积极反馈 (*238 237，*)。有观察表明，随着系统上限能力的增加，目标博弈会变得更加普遍，因为更有能力的系统可以找到更多的方法来实现给定的目标 (*490 489，*)。
此外，更有能力的系统有更多的方法来连贯地“一般”在不同于其训练设置的情况下表现不同，包括潜在的有害方式-尽管这只是迄今为止的假设问题 (*490 489，*)。然而，最近一些更强大的通用人工智能系统由于更好的tr和监督工具而变得更加可控 (*19， 491 20，*)，并且不太可能在其训练数据之外*不连贯*地泛化 (*492*)。因此，随着更先进的系统的创建，通用人工智能系统是否会变得或多或少可控尚不清楚。有关当前通用AI系统的可控程度的讨论，请参见
5.2培训更多值得信赖的模型和4.4.1交叉技术风险因素。

，一些数学发现表明，未来的通用AI代理可能会使用阻碍人类控制的策略，但目前尚不清楚这些发现将如何适用于现实世界的通用AI系统。，一些理想化的目标导向人工智能智能主体的数学模型发现，如果有足够先进的规划能力，多这样的人工智能主体会阻碍人类干扰其目标追求的尝试 (*493， 494， 496 495 *，*)。类似的数学研究结果表明，许多这样的人工智能代理可能倾向于通过积累资源来“寻求权力”，干扰监督公共关系ocesses和av油化被停用，因为这些行动帮助他们实现既定目标 (*493，494，495 *，497 *，498，499*)。

但是，从这项数学研究中得出现实世界的含义并不容易。对于ins，大多数r研究都假设AI系统是使用随机选择的目标 (*493， 494， 495 *，497 *， 498， 499*) 进行训练的，但在实践中，人工智能开发人员可以极大地影响哪些目标可能被编码在通用人工智能模型中 (见下文，5.2培训更值得信赖
模型)。此外，哲学研究表明，并非所有上述行为都是追求随机目标所隐含的 (*500*)。目前，一个数学发现和早期的经验发现支持，尽管gh弱作为本文的 ，这样的行为可能会发现在更现实的情况下 (*181， 237， 490，497**)。也有案例研究表明，在没有提示的情况下，通用人工智能系统和其他人工智能系统系统地诱导他人的错误信念，因为这对实现目标很有用 (*501 366，*)。如果更广泛地观察到，这种行为也与通用AI聊天机器人和代理的近期应用有关。

，如果人们将越来越重要的责任委托给通用AI系统，那么这可能会增加失去控制的风险。一系列社会和经济力量会影响

在这种情况下，人类和自主代理之间的交互。例如，在没有干预的情况下，经济压力可能有利于通用人工智能自动化，尽管有潜在的负面影响 (*502*)，和人类对通用人工智能代理的过度依赖将使监督变得更加困难 (*481**)。使用通用人工智能代理在政府、军事或司法应用中自动化决策可能会引起人们对人工智能对重要社会决策 (*503*、*504*、*506 505*、) 的影响的担忧。作为一个更极端的例子，一些参与者表示有兴趣有目的地开发不受控制的AI代理 (*507*)。

某些特定功能的可能会不成比例地增加失去控制的风险。这些能力 -- 目前还很有限 -- 包括识别和利用软件漏洞、说服、自动化人工智能研发，以及自主复制和适应所需的能力 (*509 367 ** 、人工智能*508 ** 、)。本报告的相关部分讨论了当前通用人工智能系统在其中一些领域的能力 (4.1.3网络犯罪，4.1 .2 信息和操纵舆论，4.1.4双重用途科学风险)。特别相关的是代理功能，它增加了通用AI系统自主运行的能力，例如规划和使用内存。这些在4.4.1交叉中讨论 技术风险因素。

失去控制的后果

一些通用人工智能系统不可逆转的失控并不一定是灾难性的。打个比方，计算机病毒早就能够近乎不可逆地大量扩散
(*510*) 不会导致互联网崩溃。一些研究人员已经探索了假设场景，其中高度先进的未来通用AI代理自主行动可能会对人类造成catas的伤害，特别是如果人类对实现代理的给定目标存在障碍 (*511 127，*)。通常认为损害机制源于上一段和4.1.4两用科学风险中列出的能力。然而，这些场景仍然是假设的，因为它们没有被当前的通用人工智能系统所展示。

人工智能研究人员对失控风险有不同的看法

人工智能专家讨论了失去对未来通用人工智能系统控制的可能性。关键问题包括是否会在中期内开发足够能力的通用AI代理，以及是否可以及时开发技术安全和治理解决方案以保持对它们的充分控制。由于评估失控风险的研究有限，专家意见可以提供替代指导，但不能取代研究。

一部分研究人员认为，失去控制的风险值得考虑。然而，极端控制失败的总体可能性仍然存在很大争议。，一些研究人员认为，在开发其他人担心可能会带来失控风险的AI系统类型方面，进展*129，* (*512*)。然而，已经提出了关于社会如何失去对人工智能系统的控制的广泛假设情景 (*511 507，*)，一些领先的研究人员强调了这些情景 (*127*)。例如，数百名人工智能研究人员签署了一份声明，宣布 “减轻人工智能灭绝的风险应该是全球优先事项” (*486*)，尽管没有明确提及失控。实际风险仍然存在很大争议，因为只有有限的研究评估它，科学家的意见不能取代这项研究。

4.3 系统性风险

4.3.1 劳动力市场风险

关键信息

- 与之前的自动化浪潮不同，通用人工智能有可能自动化非常广泛的任务，这可能会对劳动力市场产生重大影响。
- 这可能意味着许多人可能会失去目前的工作。然而，许多经济学家预计，自动化带来的潜在失业可能会被新就业机会的创造和非自动化部门需求的增加部分或全部抵消。
- 劳动力市场摩擦，例如工人学习新技能或重新安置新工作所需的时间，即使总体劳动力需求保持不变，也可能在短期内导致失业。
- 通用人工智能对工资的预期影响是模棱两可的。它可能同时通过提高生产率和创造新的机会来提高某些部门的工资，并降低其他部门的工资，因为自动化减少劳动力需求的速度快于新任务的产生。

人工智能可能会改变一系列工作，并可能取代工人

经济学家预计，通用人工智能将通过自动化任务来影响劳动力，提高工人的生产力和收入，改变各种职业所需的技能，并将工人从某些职业中取代 (*515 514， 513，*)。经济学家对这些影响的程度和时间持有广泛的观点，一些人预计未来十年将出现广泛的经济转型，另一些人则认为与人工智能相关的自动化和生产率增长不会迫在眉睫 (*516*)。未来通用人工智能进展的不确定性导致了通用人工智能对劳动力市场影响的不确定性。

以前的计算自动化浪潮主要影响了“常规”任务，这些任务可以很容易地编码和编程到计算机 (*517*) 中。相比之下，通用人工智能具有执行通常由人类执行的各种任务的潜力，包括复杂的问题解决和决策。大量文献研究了工作对自动化和人工智能的普遍影响，但没有特别关注通用人工智能 (*518*)。与这项研究相比，对通用人工智能可能对劳动力市场影响的探索还处于非常早期的阶段。据估计，在发达经济体中，由于以认知任务为导向的工作普遍存在，目前60% 的工作可能会受到引入通用人工智能系统 (如今天的LLMs) 的影响 (*519*)。这意味着通用人工智能有可能使大部分工作自动化，或者补充并显著改变工作的完成方式。在新兴经济体中，被认为可能受到通用人工智能系统影响的工作比例较低，但仍然很大，为40% % (*519* %)。

最近的实证研究已经开始证明当代通用人工智能系统对各个行业的影响，特别是在知识工作方面:

- 通用人工智能系统已被证明可以提高战略咨询 (*520*) 的、质量和速度，提高客户支持 (*250*) 的性能，并改善计算机编程 (*521**)。
- 机器翻译和语言模型已被证明可以替代人类工作者，例如sim ple大规模语言翻译 (*522*) 和一些与写作/编码相关的职业 (*524 523，*)，导致对其服务的需求减少。

- ChatGPT的引入和大规模采用导致提供写作，编辑和校对服务的自由职业者的就业和收入减少 (*523*)。

一个关键问题是，随着通用人工智能系统变得更先进、更广泛，是否会出现大量失业。一些经济学家认为，现有职业对劳动力的需求增加和创造新的就业机会可能会抵消失业的影响 (*516*， *525 517*，)。这与前一波自动化浪潮的影响是一致的: 例如，超过60% % 的就业2018年从事的职位1940年不存在 (*526*)。其他经济学家，通用人工智能系统对就业市场的未来影响很难预测 (*527 519*，) 与资本相比，通用人工智能可能导致人类劳动力价值大幅下降 (*528*)。

即使经济中对劳动力的总体需求没有减少，如果劳动力市场不能足够快地将工人与新的就业机会相匹配，流离失所的过程也会造成失业。这可能是由于各种劳动力市场摩擦而发生的 (*529*)，例如:

- 当工人必须学习新技能来转换职业时，他们需要时间来完成必要的培训或教育
- 劳动力流动性不完善，限制了工人重新安置新工作机会的能力
- 新工作的要求与流离失所工人的现有技能之间的技能不匹配

这些因素影响着被自动化取代的工人能够以多快的速度以及是否能够进入新的岗位。因此，即使整个经济中可能存在职位空缺，一些工人也可能暂时失业。通用人工智能可能导致集中失业，而生产率的提高可能会在整个经济中更加分散，这可能会导致一些工人难以转型，除非得到支持。

一些经济学家认为，由于通用人工智能系统实现自动化而导致的工作流失率可能超过创造新的就业机会，这是合理的，甚至是可能的，特别是如果ent通用人工智能系统的重点是替代人力，而不是增加人力 (*502*，*530*)。然而，很少有精确的定量预测，现有的研究与广泛的可能结果是一致的。

人工智能研究人员对未来通用人工智能进步的速度持不同意见，但有一些人支持极快进步的可能性，包括通用人工智能系统在几乎任何认知任务中匹配或超越人类专家能力的可能性。(见2.4.3。算法的进步会带来快速的进步吗？)。后一种情况只在很少的经济研究中被考虑过。该研究假设人工智能系统几乎可以比人类更经济有效地执行所有知识任务，并表明这种情况可能导致许多部门的工资暴跌，严重的失业和劳动力参与率的急剧下降 (*532 531*，)。但是，即使在这种情况下，由于消费者的偏好以及与道德或控制相关的原因，对人类劳动力的某些需求也可能持续存在 (*532 531*，)。在经济学家中，由于大规模通用人工智能自动化导致劳动力需求急剧下降的极端情况目前被认为是相对边缘的。

很难预测通用人工智能系统如何以及何时影响各种劳动力市场。首先，通用人工智能技术的发展速度以及其未来的能力可能存在很大的不确定性。其次，即使对于给定的技术复杂程度，通用AI系统实现自动化的程度以及这种自动化如何影响劳动力市场尚不清楚。其中一个原因是，在许多领域，大规模部署和采用通用人工智能系统可能存在相当大的滞后。例如，如果员工缺乏有效使用通用人工智能辅助的必要技能，这可能会减缓采用速度。总体而言，许多经济专家预计通用人工智能系统在未来十年 (*533*) 将对宏观经济产生适度的影响，而一些电子专家预计未来五到十年 (*534*) 将对劳动力市场和宏观经济产生重大影响。这些预测往往表明，由于通用人工智能能力未来进步的步伐未知，不确定性很高 (见未来几年2.4能力进展)。

，导致自动化的通用人工智能系统对工资的影响是不确定的，证据也参差不齐。通用AI自动化可以通过以下方式提高某些行业的工资:

- 使用通用AI工具 (*535 502，* 直接提高人类生产力)
- 导致经济的整体增长，从而导致更多的投资，提高非自动化部门工人的生产率 (*536*)
- 在尚未自动化的任务中增加产出并增加对劳动力的需求，或者创造新的任务和职业 (*537 517，*)
- 通过加速研发广泛提高技术 (*538*)

如果通用人工智能系统补充了人类劳动力，那么随着通用人工智能系统能力的提高，随着经济增长，年龄的增长可能会加速，可能比历史上的*539*要快得多。最近对当前通用人工智能系统使用情况的大规模调查支持这一观点。例如，最近对七个经合组织国家制造业和金融业的5,334名工人和2,053公司进行的一项调查显示，大约80% % 的使用人工智能的工人表示，人工智能改善了他们的工作表现 (*540*)。

然而，如果未来自动化减少劳动力需求的速度快于创造新的就业机会，那么工人的工资和收入份额可能会大幅下降 (*541*)。通用人工智能系统也可能在不同的时间点产生不同的经济影响: 使用通用人工智能系统最初可能会提高工资，但随着越来越多的任务自动化，对剩余工作的竞争可能会导致工资下降*532*。关于通用人工智能自动化对平均工资的净影响，专家们目前还没有明确的共识: 随着时间的推移，不同职业的影响可能会有所不同，这取决于许多因素，包括社会对技术的接受程度和组织决策，以及政府政策。

通用AI可能会增加收入不平等

通用人工智能可能会增加国家内部和国家之间的收入不平等。从历史上看，日常工作的自动化可能会通过将工人从他们拥有比较优势的工作类型中取代而加剧国家内部的工资不平等 (*517*)。同样，通用人工智能可以系统地与人类知识工作者目前拥有优势的一些任务竞争，如果他们不能轻易在其他地方找到工作，可能会压低工资 (*543 542，*)。与此同时，通用人工智能可以提高高收入职业的生产率，因此高薪收入者可以看到劳动收入的不成比例的大幅增长，从而扩大劳动收入不平等。一项模拟表明，人工智能的广泛采用可能会在发达经济体 (*519*) 采用后的十年内将高收入和低收入职业之间的工资不平等增加10% %。

自动化还可能通过降低劳动力的收入份额来加剧不平等，这将提高较富裕的资本所有者的相对收入 (*528*)。这将是一个持续趋势的一部分: 在全球范围内，劳动收入在2022年1980年的份额下降了大约6个百分点，美国、亚洲和欧洲也出现了类似的情况 (*544*)。进一步的自动化可能会导致这一趋势的延续。通常，10% 的收入者赚取大部分资本收入 (*546 545，*)。因此，更高的资本生产率将为高收入者带来福音。如果通用人工智能帮助创建具有强大市场力量的 “超级明星” 公司，这种动态可能会特别明显，因为它们将获得巨大的经济利润份额 (*519*)。

最后，通用人工智能技术如果主要由发达经济体采用，可能会加剧全球不平等 (另见4.3.2。全球人工智能鸿沟)。这些国家在面向认知任务的工作中所占比例更高，受到通用人工智能、更强大的数字基础设施、熟练劳动力和更发达的创新生态系统的潜在影响。这使他们能够比新兴市场和发展中经济体更快地获得人工智能生产率的增长，这可能导致不同的收入增长轨迹以及高收入和低收入国家之间差距的扩大 (*547 519，*)。

4.3.2 全球AI鸿沟

关键信息

- 通用人工智能的研发目前主要集中在少数西方国家和中国。这种“人工智能鸿沟”是多原因的，但在一定程度上与低收入国家有限的计算能力有关。
- 获得大量昂贵的计算能力已成为开发高级通用AI的先决条件。这导致大型科技公司在通用人工智能开发方面日益占据主导地位。
- 人工智能研发鸿沟往往与现有的全球社会经济差距重叠，可能会加剧这些差距。

在西方国家和中国，人工智能的研究和开发集中在有据可查的地方，包括对人工智能潜在社会影响的研究(*316，548，549*)。对于通用人工智能来说，这种全球“人工智能鸿沟”可能会变得更大，特别是由于与通用人工智能开发相关的高成本。一些国家在从通用人工智能开发和部署中受益方面面临巨大障碍，包括较低的数字技能素养、有限的计算资源获取、基础设施挑战以及对高收入国家 (*519， 550*) 实体的经济依赖。由于通用人工智能系统的开发主要由少数公司主导，特别是那些总部位于美国的公司，因此人们担心，在全球范围内使用的突出的通用人工智能系统主要反映了西方大型公司的价值观，文化和目标。此外，最近的趋势是旨在开发更大，更强大的通用AI模型，这也可能加剧全球供应链的不平等 (*551*)，对能源使用提出要求，并导致有害的气候影响，这也加剧了全球不平等 (*553 552，*)。如果在全球范围内部署有偏见或不公平的通用人工智能系统，全球通用人工智能鸿沟也可能是有害的。

技术人才集中度的差异以及开发和维持通用人工智能系统的高昂财务成本可能会使人工智能鸿沟与现有的全球社会经济差距保持一致。，美国拥有最多的精英AI研究人员，拥有大多数进行顶级研究的机构，并且是全球AI人才的首选目的地 (*554*)。然而，在人工智能发展方面处于领先地位的国家也遇到了人工智能技术人才分布的问题，这些人才正在迅速向工业转移。例如，70% 名拥有人工智能博士学位的北美大学毕业生最终在私营企业找到了工作，而20年前的这一比例为21% % (*555*)。

据报道，在2023年4月，OpenAI的人工智能系统估计每天产生700美元的推理成本 (*77*)，对于广大的少校来说，这是一项广泛无法获得的成本年限学术机构和公司的数量，对于那些总部设在全球南方的机构和公司来说更是如此 (*556，557*)。鉴于收集、标记和存储的高成本，低资源地区在访问数据方面也面临挑战。利用这些数据集进行模型开发的技术人才的可用性较低，可能会进一步加剧人工智能的鸿沟。基础设施问题是阻碍公平获得培训所需资源的一个主要因素和im由于诸如不足等问题，实现了通用AIaccess到宽带互联网 (*558，559*) 、停电和供电不足 (*560，561*)。

中国和美国的**学术机构在通用人工智能研究生产方面处于领先地位，但行业影响力越来越大。**中国目前发表的关于人工智能的研究最多，以期刊、会议和在线知识库的文章总量来衡量。
从地理上讲，重要的机器学习模型的开发集中在美国，加拿大，英国和中国等国家，其中至少有一个来自非洲。美国工业目前

主导着先进的通用AI系统的开发。美国机构在2022 (*562*) 生产了大多数 (54%) 的大型语言和多模态模型。工业现在在产生重要的机器学习模型 (32个对三个2022年) 方面超过了学术界，在十大领先的人工智能会议上发表的论文的行业合著者从22% 个2000年上升到2020的38% 个 (*563*个)。

不断上升的 "计算鸿沟" 正在导致计算资源分配的差异，以及对通用AI开发的不平等参与。，术语 "计算划分" 描述了大型工业AI实验室和典型的academi c AI实验室访问计算资源 (*556*) 的不同程度。近年来，这种鸿沟扩大了 (*557 556，*)。

据估计，美国科技公司是英伟达H100 GPU的主要买家，英伟达H100 GPU是市场上专为人工智能设计的最强大的GPU芯片之一 (*564 **)。亚马逊、Meta、谷歌和微软最近都宣布了定制的人工智能芯片，以减少对人工智能芯片供应链的依赖，这可能为更广泛地访问gpu铺平道路。然而，gpu的成本异常高 (在撰写本文时，H100等顶级型号的价格为15,000美元) 可能会阻碍学术机构和不太富裕的国家提供这种水平的人工智能基础设施。

将较低级别的人工智能工作委托给低收入国家的工人，导致了一个 "幽灵工作" 行业。从内容审核到校对再到数据标签，对于大型科技公司 (*565*) 的许多产品来说，典型消费者通常不知道的大量人力劳动- 有时被称为 "幽灵工作"。对训练通用人工智能系统的数据需求不断增加，包括帮助培训的人类反馈，进一步增加了对幽灵工作的依赖，包括创建公司帮助大型科技公司外包数据生产的各个方面，包括数据收集、清理和注释。这种趋势在诸如ImageNet (*566*) 之类的著名机器学习基准数据集的开发中发挥了重要作用。数据生产是通用人工智能进步的一个关键方面，通常依赖于工人

在平均工资较低的国家。这些工人可能面临图形内容，不稳定的时间表，繁重的工作量，并具有有限的社会和经济流动性 (*567， 570 569， 568，*)。这可能导致对边缘化工人的伤害，扩大人工智能鸿沟，并增加谁从先进的通用人工智能发展中受益的差距。

4.3.3 市场集中风险和单点故障

关键信息

- 开发最先进的通用AI模型需要大量的前期投资。这些非常高的成本造成了进入壁垒，使大型科技公司不成比例地受益。
- 市场力量集中在少数几家公司，这些公司是唯一能够构建领先的通用AI模型的公司。
- 包括金融，网络安全和国防在内的关键部门广泛采用一些通用AI模型和系统会产生系统性风险，因为任何缺陷，漏洞，错误，或者，占主导地位的通用人工智能模型和系统的固有偏见可能会在这些相互依赖的行业中同时造成广泛的故障和中断。

最先进的人工智能系统的开发目前耗资数亿甚至数十亿美元。最大的前期投资是专门的计算资源、人工智能专业知识和对大型、通常是专有的数据集的访问 (见2.3.1计算、数据和算法)。与这些投入相关的巨大成本是新公司 (进入的障碍*571，*

572、*574 573*、)。大型科技公司处于有利地位，这要归功于它们现有的必要资源和进行大量金融投资的能力。

此外，通用人工智能系统也受益于规模。计算密集型的大型模型往往优于较小的模型 (*110 **)，从而产生规模经济: 由于其卓越的性能，大型通用AI系统的需求更高，从而降低了每个客户的成本。高用户数也会产生网络效应: 随着越来越多的用户与这些模型交互，它们会产生大量可用于提高模型性能的额外训练数据 (*575*)。

通用人工智能行业的这些市场集中化趋势尤其令人担忧，因为通用人工智能有可能使少数公司的决策比以往任何时候都更加集中。由于整个社会既可以从这些决定中受益，也可以从这些决定中受益，这就提出了有关这几个大型系统的适当治理的问题。单一的通用人工智能模型可能会以良性、微妙、无意或故意利用的方式影响许多组织和部门 (*571*) 的决策。有可能恶意使用通用人工智能作为少数公司或政府操纵、说服和控制的强大工具。潜在的有害偏见，如人口统计、人格特征和地理偏见，可能存在于任何主导的通用人工智能模型中，这些偏见可能会广泛传播。例如，流行的文本到图像模型，如DALL-E 2和Stable Diffusion e ，跨越职业，个性特征和地理环境 (*576*) 的各种人口统计学偏见。

跨关键部门对少数人工智能系统的日益依赖带来了系统性风险。针对这些系统的错误、漏洞或网络攻击可能会造成广泛的破坏。已经提出了不同的场景来说明潜在的中断。例如，对广泛使用的AI API的拒绝服务攻击可能会破坏依赖该技术的关键公共基础设施。在金融领域，多个机构采用ems的同质人工智能系统可能会通过同步参与者的决策来破坏市场的稳定 (*577*): 如果几家银行依赖于一个模型，他们可能会无意中做出类似的选择，从而造成系统性漏洞 (*2**)。如果广泛部署具有类似功能的人工智能系统，国防或网络安全等领域可能会出现类似的风险 (另请参阅4.4。交叉风险因素)。

4.3.4 对环境的风险

关键信息

- 在通用AI开发和部署中不断增长的计算使用量迅速增加了与通用AI相关的能源使用量。
- 这一趋势可能会持续下去，可能导致二氧化碳排放量大幅增加。

最近用于人工智能的计算能力 (“计算机”) 需求的快速增长，特别是通用人工智能的开发和部署，可能使人工智能在不久的将来成为数据中心电力消耗的主要且可能是最大的贡献者。这是因为预计计算需求将远远超过硬件效率的提高。

今天，数据中心、服务器和数据传输网络占全球电力需求的1% % 至1.5% % (578); 欧盟大约为2% %，美国4% %，中国接近3% % (69，579，580)。人工智能目前可能占数据中心电力消耗的一半以下，但如果人工智能计算需求的快速增长持续下去，人工智能可能会在未来几年成为数据中心电力的主要消费者，并增加其在全球电力需求中的份额。在2023，最大的通用人工智能训练运行在5e25翻牌 (65) 左右。使用每GPU运行1400瓦的H100 GPU进行操作和冷却，消耗约40 GWh。如果这个数字是

以每年3倍的速度增长，然后在十年末，最大的培训运行将消耗90 TWh，占美国数据中心总用电量2022年的一半以上。

对于广泛的通用人工智能系统日益增加的能源使用，有几种潜在的缓解措施。专门的AI硬件和其他硬件效率改进可以随着时间的推移 (*78*) 提高机器学习工作负载的性能功耗比。此外，新的机器学习技术和架构可以帮助降低能耗 (*78*)。计算的能量效率通常每年提高估计的26% (*68*)。然而，即使对人工智能进行了额外的优化，人工智能训练对计算能力的需求也在不断增长，每年增长约4倍，到目前为止，这远远超过了能效改进的速度 (*17*)。

人工智能开发和部署产生的二氧化碳排放取决于其能源消耗的程度和来源以及几个因素。能源的碳强度是一个关键变量，与化石燃料 (*581 **) 相比，太阳能等可再生能源在其整个生命周期中贡献的二氧化碳排放量要少得多。人工智能公司通常依赖于可再生能源 (*76， 78*)，但全球人工智能培训的很大一部分仍然依赖于高碳来源，如煤炭或天然气 (*581**)。影响二氧化碳排放的其他重要因素包括数据中心的地理位置，其效率以及所使用硬件的效率。因此，在AI中消耗的给定能量的实际二氧化碳排放量可能会有很大差异。

人工智能硬件的 “隐含碳足迹”，包括制造、运输、物理建筑基础设施和处置 (而不是不运行硬件) 的排放，对排放的贡献很大 -- 根据位置的不同，这一比例可能高达50% (*76*)。随着硬件效率提高s，所体现的碳足迹可能会在总碳足迹中所占比例更大 (*78 76，*)。

用水量可能是AI带来的另一个值得注意的环境风险领域。考虑到用于训练和部署模型的计算的增加，冷却需求也会增加，导致更高的用水量。目前模型的用水量及其评估方法仍存在科学争论，但一些研究预测，人工智能的用水量2027年可能会增加到数十亿立方米 (*76， 582*)。在全球淡水短缺的背景下，假设没有明显的短期冷却替代方案，人工智能水足迹可能是环境问题的一个重要原因。

4.3.5 隐私风险

关键信息

- 通用AI模型或系统可以 “泄漏” 有关其数据用于训练的个人的信息。对于未来针对敏感个人数据 (如健康或财务数据) 进行训练的模型，这可能会导致特别严重的隐私泄露。
- 通用人工智能模型可能会加剧隐私滥用。例如，大型语言模型可能有助于更有效地搜索敏感数据 (例如，在互联网文本或违规数据泄漏中)，并且还使用户能够推断有关个人的敏感信息。

通用人工智能系统依赖并处理大量个人数据，这可能会带来重大且可能影响广泛的隐私风险。这些风险包括其数据被用于培训这些系统的人员的数据机密性丧失，透明度和对如何做出数据驱动的决策的控制丧失，以及这些系统可能导致的新形式的滥用。

从广义上讲，隐私是指一个人有权控制他人访问其敏感或个人信息的权利。在人工智能的背景下，隐私是一个复杂而多方面的概念，它

包括保密性、透明度和控制问题。隐私是一个很难定义的概念*583*。在人工智能的背景下，它包括:

- 数据机密性和保护为培训目的或推理过程中收集或使用的个人数据 (*584*)
- 透明度，以及对个人信息在人工智能系统中的使用方式的控制 (*585*)，例如个人选择退出为培训而收集的个人数据的能力；或事后使通用AI系统 “不了解” 有关个人的特定信息的能力 (*586*)；
- 由于数据使用或恶意使用而可能发生的个人和集体危害，例如创建deepfakes (*587*)。

通用AI系统可能会暴露其训练数据。通用AI模型的训练通常需要大量的训练数据。学术研究表明，其中一些训练数据可能会被通用AI模型记忆，或者可以使用对抗性输入提取，使用户能够推断有关其数据被收集的个人的信息
(*588， 590 589，* 甚至重建整个训练示例 (*591， 592， 594 593，*)。然而，记忆的定义各不相同，因此对记忆可能产生的危害做出任何具体声明是具有挑战性的 (*595*)。

许多系统都接受了包含个人信息的公开可用数据的培训，而无需相关个人的知识或同意。然后，该信息可以由通用AI系统在不期望的上下文中输出。在包含个人信息 (如医疗或财务数据) 的敏感数据上训练模型可能会导致严重的隐私泄露。很难评估这些风险的可能性或潜在影响: 例如，现有的医疗通用人工智能系统，如Google的gemini-med (*596 **)，仅在匿名的公共患者数据上进行训练。而且这种模型反刍训练数据的速度还没有被研究。不断从与用户的交互中学习的通用AI系统 (例如ChatGPT等聊天机器人) 也可能会将此类交互泄露给其他用户，尽管在撰写本文时，还没有充分记录的情况。

通用人工智能系统可能会导致隐私滥用。一些研究发现，通用AI系统具有与隐私相关的功能，这些功能可能会被这些系统的恶意用户利用。例如，细粒度的互联网范围内的搜索功能，如强大的反向图像搜索或形式的写作风格检测，允许个人被识别和跟踪的在线平台，或敏感的个人特征被推断，进一步侵蚀个人隐私 (*598 597，*)。大型语言模型还可以更有效地搜索互联网上或违规数据集中的敏感信息。通用人工智能生成的内容，如未经同意的深度伪造，可能被用来操纵或伤害个人，引发人们对恶意使用个人数据造成的伤害和对在线内容信任的侵蚀的担忧 (*255，256，373，599*)。

4.3.6 侵犯版权

关键信息

- 使用大量受版权保护的数据来训练通用AI模型对传统的知识产权法以及对数据的同意，补偿和控制系统构成了挑战。
- 开发通用人工智能的组织大规模使用受版权保护的数据可能会改变围绕创造性表达的动机。
- 不明确的版权制度阻碍了通用AI开发人员遵循数据透明度的最佳实践。

- 从互联网上获取和过滤法律和道德允许的数据以训练通用AI模型的基础设施非常有限。

通用AI模型通常是在在线获取的大型数据集上进行训练的，这引起了人们对侵犯版权，缺乏创作者补偿以及潜在的经济破坏的担忧。版权法旨在保护知识产权，鼓励书面和创造性的表达 (*601 600，*)。他们授予原创作品的创作者复制，分发，改编和执行自己作品的专有权。然而，在某些情况下，第三方使用受版权保护的数据作为培训数据可能是法律允许的，例如，基于美国的“合理使用”例外 (*602*)，通过欧盟的“文本和数据挖掘”例外 (*603*)，根据修订后的日本版权法 (*604*) 、以色列版权法 (*605*) 和新加坡版权法2021 (*606*)。除了版权之外，艺术家和其他个人有时会觉得他们的风格，声音和肖像没有得到充分的保护，这可能涉及其他形式的知识产权，如商标和品牌。

通用人工智能能力的最新进展在很大程度上导致了大规模网络抓取和数据聚合来训练通用人工智能模型 (*608 607，*)，这些模型通常包含受版权保护的作品，或者在未经数据创建者同意的情况下使用。这适用于创意作品，包括文本，图像，视频和语音，以及越来越多地用于开发通用AI模型的其他方式。法律允许的程度是复杂的，可能因国家而异。在美国，在训练通用人工智能模型 (*333*、*609*、*611*) 以及法律挑战 (*612*) 的情况下，*610*、合理使用例外被认为是合理使用。与数据集的创建和使用相关的许多问题使得训练AI模型的版权问题变得非常复杂 (*613*)。这些问题包括s数据集是专门为机器学习组装的，还是最初用于其他目的 (*614*)，侵权分析是适用于模型输入还是模型输出 (*615*)，以及管辖权问题等 (*236，616，617*)。这也提出了谁应对侵权或有害模型输出负责的问题 (*618*)。虽然有从模型输出中减轻版权侵权风险的技术策略，但这些风险很难完全消除 (*620 619，*)。

随着通用人工智能系统变得越来越，它们增加有可能扰乱劳动力市场，特别是创意产业 (*621 250，*) (另见 4.3.1劳动力市场风险)。在人工智能培训阶段，关于版权侵权的法律裁决将影响通用人工智能开发人员构建强大和高性能模型的能力。它们还可能影响数据创建者对其数据施加控制的能力，这可能会抑制创造性表达。

不明确的版权制度会阻碍通用AI开发人员提高数据透明度。通用AI模型训练数据的透明度有助于了解通用AI系统 (*259*) 的各种潜在风险和危害。然而，对于主要的通用人工智能开发人员来说，这种类型的透明度往往是缺乏的 (*244*)。对法律风险的担忧，尤其是对版权侵权的担忧，可能会阻碍这些开发人员披露他们的培训数据 (*622*)。

，用于获取和过滤法律允许的数据的基础设施还不完善，这使得开发人员很难遵守版权法。允许在没有许可证的情况下使用受版权保护的作品作为培训数据的一部分是诉讼的一个活跃领域。在没有版权问题的情况下获取和识别可用数据的工具是有限的。例如，最近的工作表明，在使用最广泛的可公开访问的数据集存储库中，60% 流行数据集的许可证信息 (*236*) 不正确或缺失。，当前用于识别web scrapes中无版权数据的工具也存在局限性 (*623 607，*)。然而，从业人员正在为数据文档开发新的标准，并为数据创建者制定新的协议，以表明他们同意用于训练人工智能模型 (*624 235，*)。

4.4 交叉风险因素

4.4.1 交叉技术风险因素

关键信息

- 本节涵盖七个交叉的*技术风险因素*，每个技术因素都会导致许多通用AI风险。
  a. 通用人工智能系统可以在许多方面和环境中应用，因此很难在所有实际用例中测试一下和确保其可信度。
  b. 通用AI开发人员对通用AI模型和系统如何在内部运行以实现其输出功能的了解非常有限。
  c. 通用人工智能系统可以按照意想不到的目标行事，导致潜在的有害输出，尽管人工智能开发人员进行了测试和缓解工作。
  d. 通用人工智能系统可以快速部署到大量用户，因此，如果大规模部署故障系统，造成的危害可能是快速和全球性的。
  e. 目前，通用人工智能系统的风险评估和评估方法还不成熟，可能需要大量的精力、时间、资源和专业知识。
  f. 尽管尝试进行调试和诊断，但开发人员无法在使用通用AI系统的所有情况下防止明显有害的行为。
  g. 一些开发人员正在努力创建通用人工智能系统，这些系统可以以越来越大的自主性行事，这可能会增加风险，因为它可以在更少的人为监督下实现通用人工智能系统的更广泛应用。

风险因素不同于风险。它们是增加风险发生的可能性和/或影响的条件。本节涵盖七个交叉的*技术风险因素*，即每个因素都会导致多个通用AI风险。

通用AI系统可以在许多方面和环境中应用，因此很难在所有可能的用例中测试一下和确保其可信度。通用AI系统的相对安全性取决于它们的使用环境。通用AI系统的输出通常是开放式的，例如自由形式的对话或代码生成。这使得设计安全的系统变得困难，因为无法详尽地评估所有可能的下游用例。用户还可以“越狱”通用AI模型，使其符合潜在的有害请求 (见5.2.3提高鲁棒性故障)。目前，计算机科学家无法对通用AI系统 (*625*) 做出“System X不会做Y”的保证。如3中所述。评估和理解通用人工智能系统的方法，评估通用人工智能系统在现实应用中的风险，以及对通用人工智能相关危害做出强有力的保证，用目前的方法是极其困难的。

通用AI开发人员对通用AI模型和系统如何在内部运行的理解非常有限。通用人工智能系统的一个关键特征是，它们的能力主要是通过学习而不是自上而下的设计来实现的。因此，与大多数人类工程系统不同，最先进的通用人工智能模型没有蓝图，它们的结构也不符合通用的设计原则。这引起了人们对理解或解释通用人工智能的担忧，在本报告中，通用人工智能可以互换使用，指的是能够提供人类可理解的描述，说明通用人工智能如何从输入和目标中获得输出和决策。关于什么有不同的看法

构成了一个人类可以理解的解释。对这些观点的彻底讨论不在本报告的范围之内，但关键问题包括:

- 它可以采取什么形式，例如，它是否必须是易于理解的语言，或者可以是复杂的数学信息，以及描述的大小是否应该被限制以允许人类理解解释。
- 它必须有多全面，例如是否应包括对培训中如何建立解释决策过程的回顾性分析；
- 是否必须是反事实的，即是否必须允许关于不同投入、不同目标或不同培训将产生什么产出的假设性陈述。

目前，科学家对通用人工智能系统的理解更类似于大脑或细胞，而不是飞机或发电厂。一些研究人员认为，通过将验证重点放在神经网络的可解释输出而不是人类目前无法理解的内部状态上，有可能开发出可以被证明是安全的或“设计安全”(*626*) 的通用AI系统。然而，科学家们还无法为最先进的通用人工智能模型实现这样的定量安全保证 (*58*)。关于彻底理解人工智能系统如何运作的研究有限的“ ”系统，这些系统比通用人工智能模型 (*627*、*628*、*629*、*630*) 要小得多，能力也差得多。或者不可靠，需要主要的简化假设 (*631*、*632**)。在实践中tice,t解释神经网络内部工作原理的技术可能会产生误导 (*213，288 *，289，290，336*)，并且可能无法通过健全性检查或证明在下游使用中无益(*218，297，298，299*)。如3所述。评估方法和理解通用人工智能系统， 正在开发这些研究方法，新的改进可能会产生进一步的见解，特别是在进一步研究方面有足够的投资。但是，尚不清楚解释神经网络的内部结构是否会提供足够的安全保证。

，确保通用人工智能系统实现其开发人员和用户的预期目标是困难的。尽管通用人工智能系统似乎擅长于学习what的y被“要求”去做，他们的行为可能不一定是他们的设计师想要的 (*489，633，634，635*)。即使设计师的目标和给予系统的激励之间的细微差异也会导致意想不到的失败。例如，通用的人工智能聊天机器人经常被训练来产生文本，这些文本将被人类评估人员积极评价，但用户批准是用户利益的不完美代表: 广泛使用的聊天机器人可以学会迎合用户的偏见，而不是优先考虑真相 (*238 237，*)。即使通用人工智能系统在训练过程中收到正确的反馈，它仍然可能开发出一种解决方案，一旦部署 (*636， 638 637，*)，在新情况下应用于新情况时，它仍然不能很好地概括，因为训练数据可能不能充分代表现实世界的场景。例如，一些研究人员发现，聊天机器人更有可能遵守其训练数据中代表性不足的语言的有害请求 (*309*)。请参见4.2.3。损失的控制5.2和。训练更多值得信赖的模型，以进一步讨论这些挑战。

，由于通用AI系统可以像其他软件一样迅速扩散，新的故障或有害功能可能会迅速产生全球性的，有时甚至是不可逆转的影响。少量的prop rietary和免费提供的 (开源) 通用AI模型可以覆盖数百万用户 (4.3.3.市场集中风险和单点故障)。因此，专有和开源模型在发布时都会产生快速和全球性的影响，尽管方式不同。一旦模型是开~~源的，就没有切实可行的方法从市场上删除模型，以防它有故~~障或功能导致恶意使用 (*639*) (请参阅4.1。恶意使用风险)。然而，对于模型故障，开源模型允许更多和更多样化的从业者发现它们，这可以提高对风险和可能的缓解措施的理解 (*640*) (参见3。评估和理解通用人工智能系统的方法)。
然后，开发人员或其他人可以修复故障，并鼓励用户更新到新的模型版本。此外，开源模型允许更多的参与者定制这些模型。如果模型的一个版本存在缺陷，则其他自定义版本可能不会共享相同的问题 (*640*)。但是，无论是修复模型错误还是自定义，都无法阻止蓄意的恶意使用 (*639*)。故意恶意使用的风险取决于模型与可用闭源相比的*边际风险*

模型和其他技术，例如internet搜索 (*640*) (请参阅4.1。恶意使用风险)。上述因素与通用人工智能模型快速、广泛和不可逆转的影响的具体可能性有关，但本报告没有对开源模型的总体影响进行评估。即使系统不是开源的，其功能仍然可以被广泛的用户群访问。在推出后的两个月内，ChatGPT拥有超过10000万名用户，并创下了所有消费者应用程序中增长最快的用户群的记录 (*641*)。每次更新通用人工智能系统时，它的新版本会迅速到达庞大的用户群，因此任何漏洞或有害倾向都可能迅速产生全球影响 (另请4.3 3。市场集中度和单点故障)。

，尽管尝试了调试和诊断，但开发人员仍无法在使用通用AI系统的所有情况下防止甚至明显有害的行为d。从经验上讲，有害行为包括泄露私人或受版权保护的信息 (*221，642 *， 643*); 产生仇恨言论(*225，644*); 传播社会和政治偏见 (*239，455，645*); 迎合用户偏见 (*238*); 产生幻觉accurate内容(*46,47 *,49*);exhibiting对其安全保护的各种攻击的脆弱性 (*108，208，209，210，211，309，646，647，648 **); 和协助明显有害的任务 (*368，649，650*)。用户可以相对轻松地规避通用AI模型的保护措施 (*651 650，*)，例如通过 “越狱” 技术 (3. 评估和理解通用人工智能系统的方法)。虽然有些人呼吁采取安全措施，排除*所有*情况下的*所有*公开有害行为 (*626*)，但目前的通用人工智能开发未能达到比这更低的标准: 排除任何特定的在可预见的情况下 (例如用户尝试越狱模型的情况) 明显有害的行为。

人工智能开发人员正在努力创建 “代理” 通用人工智能系统，该系统可以在很少甚至没有人工参与的情况下完成任务，这可能会增加事故和恶意使用的风险。

今天，通用人工智能系统主要被人类直接用作工具。例如，一个人可能会要求聊天机器人编写计算机代码来帮助他们完成一项任务，这自然需要 “人在循环”。然而，开发人员越来越多地设计允许通用AI系统自主运行的系统，通过控制web浏览器等软件工具或控制代码的执行，而不仅仅是编写代码。这使一些forms of右缓和about问题，制定计划，分步执行计划p (*13、26、188，189，652， 653，654，655 *，656*)。这样的系统包括web浏览虚拟代理 (*657 **)，研究助理 (*415*) 以及自主编写，修复和运行代码 (*441*)。

通用人工智能 “代理” 的主要目的是减少对人类参与和监督的需求，允许更快、更便宜地应用通用人工智能。此属性还可减少人为监督，从而可能增加事故风险 (请参阅4.1恶意使用风险)，并允许恶意使用的自动化工作流程 (请参阅4.2.1风险来自产品功能问题使用)，同时也与失控风险 (见4.2.3失控) (*658 481 *，*) 相关。

通用AI代理具有许多自主功能的早期形式，但在自主执行复杂任务方面缺乏可靠性。目前最先进的通用人工智能系统能够自主地许多简单的任务，但一些评估表明，它们更复杂的任务 (*509 183，*)。它们在执行涉及许多步骤的任务时尤其不可靠。与2023年一样，通用AI代理在旨在衡量其在复杂和经济上有用的任务上的表现的基准测试中得分较低 (*183， 441*)。然而，鉴于目前在开发具有更大自主性的通用人工智能系统方面的努力和不断增长的投资，通用人工智能代理的能力可能会继续增加。

虽然目前的通用人工智能代理不可靠，但进展很快。例如，对于衡量通用AI代理的两个具有挑战性的基准

与使用GPT-4的强基线相比，问题解决 (*183*) 和自主软件工程 (*441*) 的准确性在几个月内分别提高了2.2倍。这些基础的通用AI模型 (如llm) 的性能
通用AI代理是其可靠性的关键，因此新一代通用AI模型通常会增加代理功能。增加代理功能的一种潜在方法是将llm与搜索和计划方法相结合。在像Go和Stratego这样的棋盘游戏中，将深度学习与蒙特卡洛树搜索 (MCTS) 和自我游戏等方法相结合，带来了超越人类水平的性能 (*660 659*，)。将LLMs与搜索相结合的早期工作已经在简单设置 (*661*) 方面取得了改进。但是，需要仔细通用AI代理功能，以评估4中讨论的许多风险。风险。

4.4.2 跨领域的社会风险因素

关键信息

- 本节涵盖四个跨领域的社会风险因素-通用AI开发和部署的非技术方面，每个方面都会导致通用AI的许多风险:
  a. 争夺市场份额的人工智能开发商在降低风险方面的投资动机可能有限。
  b. 随着通用人工智能的快速发展，监管或执法工作可能难以跟上步伐。
  c. 缺乏透明度使得责任更难确定，可能会阻碍治理和执法。
  d. 很难跟踪通用AI模型和系统是如何训练、部署和使用的。

风险因素不同于风险-它们是增加风险发生的可能性和/或影响的条件。通用人工智能开发和部署的社会方面增加的不是一个而是几个通用人工智能风险。本节讨论这些“跨领域的社会风险因素”。

通用人工智能开发人员在一个动态的市场中争夺市场份额，在这个市场中，快速推出产品至关重要，他们在降低风险方面的投资动机可能有限。开发最先进的通用AI模型的一次性成本非常高，而将这种模型分发给 (其他) 用户的边际成本相对较低。这可能会导致“赢家通吃”的动态，为开发人员不惜一切代价构建最有能力的模型创造强大的激励，因为这样做可能会立即获得很大的市场份额。近年来，通用AI开发人员之间在快速构建和部署模型方面存在激烈的竞争。这引起了人们对潜在的“竞相触底”场景的担忧，在这种场景中，参与者竞相尽快开发通用人工智能模型，同时在确保安全和道德的措施上投资不足 (*663 662*，)。这可能会导致通用AI开发人员单方面遵守严格的安全标准具有挑战性，因为这样做可能会使他们处于竞争劣势 (*664*)。在国际层面上，监管工作也可能发生类似的动态。如果没有通用人工智能监管的全球协调，监管“竞相逐底”可能会看到各国试图通过宽松的监管来吸引人工智能公司，这可能不足以确保国内外的安全，这种动态已经被描述为几种类型的监管，如劳动法 (*665*)。

随着通用人工智能市场的快速发展，监管或执法工作可能难以跟上步伐。，在关于通用人工智能风险的论述中，一个反复出现的主题是

技术创新的步伐和治理结构的发展 (*666*)。虽然现有的法律和治理框架适用于通用人工智能系统的某些用途，而且一些司法管辖区 (如欧盟、中国、美国或加拿大) 已经启动或完成了专门监管人工智能和通用人工智能的工作，但监管缺口仍然存在。在一个像通用人工智能市场一样快速发展的市场中，事后很难填补这些空白，因为当实施监管修复时，它可能已经过时了。因此，政策制定者面临着创造一个灵活的监管环境的挑战，以确保从公共安全的角度来看，通用人工智能开发和部署的速度仍然可控。

通用人工智能系统固有的缺乏透明度，使得法律责任难以确定，可能会阻碍治理和执行。目前的法律框架如何适用于怀疑通用人工智能系统造成伤害的情况通常不清楚。这提出了许多关于问责、责任和正义的问题。原则上，人们和公司实体要承担责任，而不是技术，这就是为什么许多关键领域保持 “人在循环” 政策的原因。然而，将伤害追溯到开发或部署人工智能的责任人 (*667， 669 668，*) 是非常具有挑战性的，收集错误的证据也是如此。专有的通用人工智能模型的不透明性质加剧了这一问责问题，在这些模型中，商业敏感的培训数据、方法和决策过程通常不会接受公众的审查。d t这里是a缺乏wi戴利使用和解释人工智能系统的共享标准操作程序 (*291，292，293，294 *，295，670，671*)。也有人认为
通用人工智能系统可能会表现出 “紧急” 行为，这些行为不是由其开发人员明确编程或意图的，这引发了谁应该对由此造成的伤害负责的问题。通用人工智能开发的分布式性质涉及多个参与者，如数据提供者、模型训练者和部署者，这也使得将责任分配给单个实体 (*669*) 变得具有挑战性。

，很难跟踪通用AI模型和系统的训练，部署和使用方式。跟踪通用人工智能模型和系统的使用不仅对于确定使用通用人工智能模型和系统造成的潜在危害的责任很重要，而且对于监控和证明恶意使用以及注意故障也很重要 (*658， 673 672，*)。
全面的安全政府ernance在汽车，制药和能源等安全关键领域 (*674， 675， 677 676，*) 共同 ，但它通常依赖于目前通用AI治理中缺失的广泛接受的标准。

5 的技术方法来降低风险

报告的这一部分讨论了通过减轻通用人工智能相关风险来提高通用人工智能安全性的技术方法: 减少危害的规模或发生危害的可能性。本报告最终发现，虽然有许多技术方法可以降低风险，但现有方法不足以证明系统是安全的。

本报告的范围不包括非技术 (政治，法律或社会) 干预措施，但这些措施对于应对风险同样重要。此外，管理通用人工智能风险的技术和非技术方面高度交织在一起; 没有技术解决方案是在真空中实施的。在过去的几个月和几年里，政策制定者对人工智能监管的兴趣越来越大，一些司法管辖区 (如欧盟、中国、美国或加拿大) 已经开始或完成了专门监管人工智能和通用人工智能的工作。有效的方法将需要资源和政治意愿来实施技术解决方案，以及多种技术和非技术保障措施之间的相互作用，以防止通用人工智能的危害。

5.1 风险管理与安全工程

关键信息

- 开发和激励通用人工智能的系统风险管理实践是困难的。这是因为目前的通用人工智能进展迅速，没有得到很好的理解，并且具有广泛的应用。评估通用人工智能风险的方法过于新生，无法进行良好的风险定量分析。
- 虽然许多其他领域为如何开发这种方法提供了经验教训，但目前还没有针对通用人工智能系统的完善的风险管理和安全工程实践。
- 由于没有一种现有方法可以提供全部或部分的安全保证，因此一种实用的策略是*防御深度*分层的多种风险缓解措施。这是管理技术风险的常用方法。
- 通用人工智能有效风险管理的一个重要考虑因素是谁参与这一过程，以识别和评估高优先级风险。这可以包括来自多个领域的专家，也可以包括受影响社区的代表。

风险是伤害发生的概率和伤害发生时的严重程度的组合 (*339*)。从技术上讲，这包括正面和负面的结果，但在通常的用法中，重点是负面结果。4中描述了与通用AI相关的一些 (但不是全部) 常见风险。风险。

风险随着潜在危害的严重性和危害实现的可能性而增加，系统的结果是否以及在多大程度上被认为是不可取的，必须根据背景人类价值观进行概念化。各种利益相关者可能不同意任何特定结果的不良程度。

的**风险面/暴露:** 技术的风险面包括它可能通过事故或恶意使用造成伤害的所有方式。一项技术的通用性越强，其风险敞口预计就越大。通用人工智能模型可以在众多应用领域进行微调和应用，并被各种用户使用 (4.4.1。交叉技术风险

因素)，导致极其广泛的风险面和风险敞口，对有效的风险管理提出了挑战。

风险管理包括 *识别*、*评估*和 *优先*风险，并利用资源来最小化、监控和控制高优先级风险。

系统安全工程的定义非常相似，但强调了较大系统多个部分相互作用的重要性 (*678*)。就人工智能而言，这种方法需要考虑通用人工智能系统的所有组成部分，以及它运行的更广泛背景。金融、保险、健康和网络安全等行业已经了完善的风险管理实践 (*679*)。NIST (*680*) 的人工智能风险管理框架是最近为数不多的针对人工智能系统提出风险管理框架的努力之一。

5.1.1 风险评估

当AI系统的适用范围和使用范围较窄 (例如，以垃圾邮件过滤为例) 时，可以以相对较高的置信度来测量显著类型的风险 (例如，误报的可能性)。相比之下，评估通用人工智能模型的风险，比如有毒语言的产生，更具挑战性，部分原因是对什么应该被认为是有毒的以及毒性和背景因素 (包括用户的提示和意图) 之间的相互作用缺乏共识。

在通用人工智能 (*679 216，*) 方面，已经使用了广泛的风险评估技术，包括评估、红队、审计和通用人工智能系统的定性评估 (参见3。评估和理解一般方法-目的AI系统)。其他风险评估方法，其中一些借鉴了其他领域的既定做法，包括:

- 提升研究，其目的是测试一下当人类能够使用通用人工智能系统时，他们在完成一项潜在的有害任务方面的能力有多强 (*166 **)。
- 为高风险决策提供信息的预测*323 **
- 德尔菲研究汇总了相关专家组的预测 (*681*)。
- 现场测试旨在检测和记录在使用模型的环境中产生的风险以及在该环境中造成的特定危险。
- 对任务和数据集进行基准测试，以评估特定类型风险的普遍性和严重性 (例如，毒性，某些形式的偏见，对科学领域的掌握，危险能力)。

目前的风险评估方法往往无法对通用人工智能系统带来的风险进行可靠的评估。将此类风险评估方法用于高性能模型的一些关键挑战是:

- 指定相关/高优先级的缺陷和漏洞在很大程度上受到谁在桌子上以及如何组织讨论的影响，这意味着很容易错过或错误定义关注的领域。
- 防范潜在的恶意使用需要了解可能和可行的威胁，包括估计恶意行为者可用的资源 (例如，计算，访问和专业知识) 及其激励措施。
- 这些技术的通用性质增加了它们在部署中的使用的不确定性。说明这一点的示例是涉及开放式交互 (例如，聊天机器人) 的应用，其可以生成大量的潜在输出。
- 技术进步的快速步伐 (2.3.1最近的能力和4.4.1交叉削减技术风险因素) 加剧了上述挑战。

例如，Red teaming只评估一个模型*是否*可以产生一些输出，而不是它在现实世界中的程度，也不是这样做的有害程度。相反，他们倾向于提供定性信息，以判断系统构成的风险。

5.1.2 风险管理

对于通用AI，可以并且已经使用了一系列风险评估技术 (请参阅3。评估和理解通用人工智能系统的方法)。为了解决这些现有工具的局限性，研究人员可以寻找其他领域风险管理的既定实践。其他安全关键行业中的一些常见风险管理工具包括:

- 计划的审计和检查。
- 使用标准化文档确保可追溯性。
- 针对关键风险和故障的冗余防御机制。
- 风险管理指南规定了安全关键系统生命周期所有阶段的流程、评估和可交付成果。

已经提出了类似的想法来管理与通用AI系统相关的风险。这些现有方法中的许多方法并不直接适用于功能强大的通用AI模型，或者它们的功效没有得到很好的研究，但正在努力将现有指南扩展到生成式AI。

安全与可靠性工程: 安全工程的实践在各种安全关键工程系统中有着悠久的历史，例如桥梁，摩天大楼，飞机和核电站的建设。在高层次上，安全工程确保即使系统的某些组件发生故障，*生命至关重要*的系统也能按预期工作，并以最小的危害发挥作用。可靠性工程的范围更广，也可以解决非关键故障。这些方法提供了几种可能对通用人工智能风险评估有用的技术:

- 设计安全 (SbD) 是一种将用户安全集中在产品和服务的设计和开发中的方法。对于通用AI产品和服务，这可以采取限制对模型的访问的形式 (例如，通过限制用户与模型交互的长度)。
- 安全分析旨在了解单个组件的功能与整个系统之间的因果关系，以便可以预期并防止可能导致系统级危险 (例如，飞机坠毁或核反应堆堆芯熔化) 的组件故障。
- “预期功能的安全性” (SOTIF) 方法要求工程师提供证据，证明系统在按预期运行时是安全的。
- 一些风险评估方法，例如针对核电部门的风险评估方法，利用了数学模型，这些模型旨在将风险量化为各种设计和工程选择的函数，并伴随着监管机构设定的量化风险阈值 (*682*)。[5]这种方法的一个关键优势是，它让一个对公众负责的机构以一种公众和外部专家都清楚的方式来定义什么风险是可以接受的。

然而，将这些领域的最佳实践转化为通用人工智能是很困难的。通用人工智能的定量风险评估方法刚刚起步，目前尚不清楚如何获得定量安全保证。其他风险评估的一般经验-

[5]例如，一些监管委员会要求核反应堆运营商进行概率风险评估，并确保某些事件的估计风险保持在规定的阈值以下。例如，运营商被要求将发电厂以外的放射性释放的估计机会保持在百万分之一 (*683*) 以下。

目的AI表明，许多关注领域可能不适合量化 (例如，偏见和错误信息)。如果定量风险评估的不确定性太大而无法依靠，它们可能仍然是重要的补充，可为高风险决策提供信息，阐明用于评估风险水平的假设并评估其他决策程序 (例如与模型能力相关的程序) 的适当性。此外，“风险” 和 “安全” 是有争议的概念-例如，人们可能会问 “对谁安全？”-这可能需要不同的专家和潜在受影响人群的参与 (*307*)。

虽然目前还没有针对通用人工智能系统的完善的安全工程实践，但减轻人工智能危害的管道感知方法从安全工程中获得灵感，并建议仔细研究在通用人工智能生命周期中做出的众多设计选择，从构思和问题制定到设计。开发和部署，既可以作为单独的组件，也可以相互关联 (*685 684，*)。需要进一步的工作来将这些想法从传统AI扩展到生成AI。

安全案例: 航空、医疗器械、国防软件等安全关键技术的开发人员需要制作 *安全案例*。这将举证责任放在开发商t o demon strat e，他们的产品不超过监管机构设定的最大风险阈值 (*686， 689 688， 687，*)。安全案例是由证据支持的结构化论证，其中开发人员识别危险，对风险场景进行建模，并评估所采取的缓解措施。对于功能有限的通用AI系统，安全案例将更容易实现，因为功能较弱的模型通常会带来较小的风险。因此，它们对通用人工智能能力进展缓慢和快速的场景都很稳健 (参见未来几年的2.4能力进步)。安全案例利用了技术开发商的技术专长，但仍要求监管机构 (或合适的第三方) 具有评估安全案例的技术专长。安全案例通常只涉及风险和威胁模型的一部分，而不涉及重要的风险和威胁模型 (*691 690，*)。缓解这些限制的一种方法是审查安全案例以及由第三方专家 (*689*) 组成的红色团队制作的风险案例。

“瑞士奶酪” 模型用于通用AI安全工程: 高性能模型的通用，快速发展和难以理解的性质使得开发，评估和激励系统风险管理实践变得越来越困难。因此，有效管理高性能通用人工智能系统的风险可能需要多个利益相关方群体的参与，包括来自多个领域和受影响社区的专家，以识别和评估高优先级风险。此外，它表明不应依赖任何单一的防线。相反，针对这些风险的多个独立和重叠的防御层可能是可取的，这样，如果一个失败了，其他人仍然是有效的。这有时被称为 *瑞士奶酪模型的防御深度 (692)*

目前的风险管理在通用人工智能开发人员中: 虽然没有普遍遵守，但通常的做法是在发布前测试一下一些危险功能的模型，包括通过红队和基准测试，并将这些结果发布在 “模型卡” 中 (*257*)。此外，一些开发人员还设有内部决策小组，负责考虑如何安全，负责任地发布新系统。这些开发人员越来越普遍的做法是通过自愿的预定义能力阈值 (*693 *， 694**) 来约束决策。这些阈值决定了特定的模型功能必须满足特定的缓解措施，这些缓解措施旨在将风险保持在可接受的水平 (*682*)。这样的能力阈值具有在能力被开发之前可观察到的优点。但是，需要做更多的工作来评估遵守某些特定阈值集是否确实可以将风险保持在可接受的水平，并评估提前准确指定适当阈值的实用性。

5.2 培训更多值得信赖的模型

关键信息

- 在训练通用人工智能系统以更安全地运行方面取得了进展，但目前还没有一种方法可以确保通用人工智能系统在所有情况下都是无害的。
- 公司已经提出了训练通用人工智能系统的策略，使其更有用和无害: 然而，这些先进系统的方法的可行性和可靠性仍然有限。
- 当前用于使通用AI系统的行为与开发人员意图保持一致的技术在很大程度上依赖于来自人类的数据，例如人类反馈。这使他们受到人为错误和偏见的影响。增加这种反馈的数量和质量是改进的途径。
- 开发人员训练模型，使其对旨在使其失败的输入更加健壮 (“对抗性训练”)。尽管如此，对手通常可以找到替代投入，以低至中等的努力降低保障措施的有效性。
- 将通用人工智能系统的功能限制在特定的用例中，有助于降低不可预见的故障或恶意使用带来的风险。
- 研究人员开始学习分析通用AI模型的内部工作原理。这一领域的进展可以帮助开发人员更可靠地理解和编辑通用AI模型功能。
- 研究人员正在探索如何获得设计安全或可证明安全的人工智能系统，尽管仍有许多开放问题将这些方法扩展到通用人工智能系统。

5.2.1 使通用AI系统与开发人员的意图保持一致

“人工智能一致性 ”是指使通用人工智能系统按照开发人员的目标和利益行事的挑战 (参见中的5.4公平和代表性技术方法
通用人工智能系统，用于讨论不同利益相关者相互冲突的价值观对调整带来的挑战)。

5.2.1.1 两个对齐挑战

培训通用人工智能系统涉及两个挑战: 首先，确保他们的训练目标能够激励预期目标; 其次，确保输出从他们的训练环境转化为预期的现实世界，特别是在高风险的情况下。

，以一种不会无意中激励不良行为的方式为通用AI系统精确指定目标是具有挑战性的。目前，研究人员还不知道如何以一种可用于训练通用AI系统的方式来指定抽象的人类偏好和价值观。此外，考虑到通用AI系统中嵌入的复杂的社会技术关系，尚不清楚这种规范是否可行。通用人工智能系统通常经过训练，可以针对开发人员的真实目标 (*634*) 进行不完美代理的目标进行优化。例如，人工智能聊天机器人经常被训练来产生文本，这些文本将被人类评估人员积极评价，但用户批准是用户利益的不完美代理。研究表明

一些广泛使用的聊天机器人有时会将他们陈述的观点与用户的观点相匹配，而不考虑真相 (*238*)。这是通用人工智能和类似人工智能系统 (*489*、*633*、*634*、*695**) 面临的挑战。

确保通用人工智能系统学习从其训练环境转化为现实世界的行为，高风险的部署环境也是非常具有挑战性的。就像通用人工智能系统被训练来优化不完美的代理目标一样，训练环境也可能无法充分代表它们在部署后会遇到的现实情况。在这种情况下，通用人工智能系统即使经过正确的人工反馈 (*636*， *638*) 训练，也 *637*， 采取有害的行动。例如，一些研究人员发现，聊天机器人更有可能在其训练数据 (*309*) 中表现不足的语言中采取有害动作。使用更多的多语言数据和监督可能能够更好地缓解这种类型的故障。

5.2.1.2 对齐技术

为了从最先进的通用AI系统中引出所需的行为，开发人员使用人类监督来训练模型。提高现实环境中的性能得益于大量数据。

最先进的对齐技术依赖于人类的反馈或演示，因此受到人为错误和偏见的限制。如2.1中讨论的**。如何通用**人工智能获得了它的能力？ 通过大量的人工参与，开发人员可以微调最先进的通用人工智能系统。在实践中，这涉及利用人类生成的期望动作示例的技术 (*18*) 或人类对模型示例的反馈 (*19,20,21 *,303*)。这是大规模完成的，这使得它劳动密集且昂贵。然而，人类的注意力、理解力和可信度并不完美 (*303*)，这限制了最终通用人工智能系统的质量 (*696*， *698 697 **，)。即使人类反馈中的微小缺陷也可能在用于训练能力强的系统时被放大，从而产生潜在的严重后果 (请参阅4.1。恶意使用风险与4.2.3.失去控制)。

提高人类监督的质量和数量可以帮助训练更多的统一模型。一些研究表明，使用人类更丰富、更详细的反馈形式可以提供更好的监督信号，但代价是增加了数据收集的时间和精力 (*699*， *700*， *701*)。为了收集更大的数据集，利用general-pu rpose AI系统部分自动化反馈过程可以大大增加数据量 (*158 **， 即*702**)。然而，在实践中，与互联网数据的预训练中使用的数万亿个数据点相比，在精调教期间使用的明确的人类监督的量非常小，因此可能无法从预训练中完全去除有害的知识或能力。如果不重新思考最先进的通用人工智能系统的训练方式，改进微调反馈数据不太可能是一个解决方案。

保持对目标的不确定性可以减少风险行为。，一些研究人员提出的方法，在中，volv e在语料库中将不确定性到通用人工智能系统学习追求的目标中 (*703*， *704*， *705*， *706 **， *708 707*，)。通过要求通用人工智能系统以尊重其目标不确定性的方式行事，这些方法可以降低意外行为的风险，并鼓励寻求信息或尊重人类以应对歧义。然而，这些方法还没有被纳入最先进的、功能强大的人工智能中。

，一些研究人员正在研究设计安全方法，这些方法可能能够提供定量的安全保证。与上述估计目标和预测不确定性的方法类似，有可能设计出能够实现量化安全水平 (*626*) 的人工智能系统。数学保证和界限的优势在于，它们甚至可以在人工智能经过训练和测试的领域之外提供安全保证，这与目前作为设计深度学习系统标准的经验试错方法形成对比。然而，目前，实用的、可证明的安全保证是不可能的

然而，为了实现大规模人工智能系统的目标，还有许多悬而未决的问题。

，目前还不清楚人类是否以及如何能够监督能力超过人类的通用人工智能系统。超越许多或所有领域的人类专家能力的未来通用AI模型或系统将构成特别困难的挑战 (请参阅4.2.3。失去控制)。一些研究工作，特别是集中在领先的人工智能实验室，研究人类在多大程度上能够监督通用人工智能，其能力通常超过人类监督或在给定领域 (称为 “可扩展监督”)。一些实证研究已经研究了能力较差的通用人工智能系统监督能力更强的系统的能力，假设通用人工智能可能存在超过人类能力的类似动态 (*709 *，* 即*710**)。研究人员还提出了理论方法，在某些假设下，这些方法可以比现有方法 (*711， 712 *，* 713 *) 提供更强有力的保证。然而，关于这些方法的已发表的研究是高度初步的。

5.2.2 减少虚假的幻觉

虚假的幻觉是一个挑战，但可以减少。在人工智能中，“幻觉” 是指通用人工智能系统输出虚假和虚构内容的倾向。例如，语言模型通常会产生不存在的引用，传记或事实 (*46， 47 *， 48， 49， 50*)，这可能会造成涉及错误信息传播的法律和道德问题 (*714*)。减少通用人工智能系统产生不真实输出的幻觉倾向是可能的，但具有挑战性。*715 *，* 明确地微调通用人工智能模型，使它们更真实 -- 无论是在答案的准确性还是对其能力的分析方面 -- 是应对这一挑战的一种方法。此外，允许通用人工智能系统访问知识数据库，它们被要求执行任务，有助于提高语言模型生成的可靠性 (*717 716，*)。替代方法尝试检测幻觉ra而不是将其从模型中移除，并且通知用户所生成的输出是否不可信 (*718*)。然而，减少幻觉仍然是一个非常活跃的研究领域。

5.2.3 提高对故障的鲁棒性

有时，unf通用AI系统在部署中遇到的amiliar输入可能会导致意外故障 (*719*)，用户或攻击者可以构建专门设计用于使系统故障 (*720*) 的输入。

对抗训练有助于提高最先进的人工智能系统的鲁棒性。*对抗性训练*首先涉及构建旨在使模型不合需要的 “攻击”，其次涉及训练系统以适当地处理这些攻击。针对AI系统的攻击可以采取多种形式，可以是人类或算法生成的。一旦产生了对抗性攻击，对这些示例的训练就可以照常进行。Adversarial 培训有 成为使模型对故障更健壮的主要方法 (*2 *,3 *,22 *,204 *，207，721*)。

5.2.3.2鲁棒性中的开放问题

虽然对抗性训练是一个有价值的工具，但它本身是不够的 (*219*)。

，使系统对不可预见的故障模式类别更加健壮是一个的挑战，引发开放问题。对抗性训练通常需要修复失败的示例 (*723 722 *，*)。这些限制导致了正在进行的 “猫和老鼠” 游戏，其中一些开发人员不断更新模型以响应新发现的漏洞。这个问题的部分解决方案是简单地制作和训练更多的对抗性示例。自动生成攻击的方法可以帮助扩大对抗性训练 (*237 203， 210，*)。但是，通用AI系统的可能输入数量呈指数级增长，因此很难彻底搜索所有类型

的攻击。已经提出了一种通过将对抗性训练应用于通用AI模型的内部状态而不是输入来解决这个问题的方法 (*724*)，但对此的研究仍然是初步的。Mathem证明模型的健壮性在理论上是一种覆盖所有可能的攻击 (*626*) 的方法，但这对于当前的模型和方法是不可能的。

对抗性训练有时会损害模型的性能或鲁棒性。在视觉模型中，在对抗性攻击的鲁棒性和非对抗性数据 (*727 725，*) 的性能之间通常存在的权衡*726，*。即使对抗性训练有助于提高最坏情况下的性能，但当它损害平均情况下的性能时，也可能不会使用它。对抗性训练有时也会使语言模型对某些未经训练的攻击的鲁棒性*降低* (*722 *， 724*)。然而，一些改进的对抗性训练方法可能能够改善干净数据和对抗性数据 (*730 728，* 对抗性数据) 的性能之间的权衡*729，*。

5.2.4 移除危险功能

"机器不学习" 可以帮助从通用AI系统中删除某些不受欢迎的功能。例如，删除某些可以帮助恶意用户制造爆炸物，生物武器，化学武器和网络攻击的功能将提高安全性 (*408*)。取消学习作为一种消除不良训练数据影响的方法，最初是作为一种保护隐私和版权的方法提出来的 (*586*)，这在5.5隐私方法的通用AI中进行了讨论。

系统。消除危险ca能力的非学习方法 (*731，，732*) 包括基于微调 (*733**) 和编辑模型内部工作原理 (*408*) 的方法。理想情况下，unlearning应该使模型无法表现出不必要的行为，即使受到知识提取攻击，新颖的情况 (例如外语) 或少量的微调。然而，学习方法通常不能鲁棒地执行未学习，并且可能在期望的模型知识上引入不想要的副作用 (*734*)。

5.2.5 分析和编辑模型的内部工作

研究模型的内部工作原理可以帮助确定特定功能的存在或缺乏。研究人员使用的一种技术是分析通用人工智能模型的内部状态，以更好地理解他们推理的概念以及他们拥有的知识 (*278， 735 279，*)。例如，这些方法已用于研究与视觉分类器 (*736*) 和知识语言模型具有的公平性有关的性质 (*737， 738**)。然而，评估通用AI模型内部表示的方法是不精确的 (*739， 741 740，*)。与其他类型的评估相比，它们目前也没有被竞争性地用于理解通用AI模型的功能 (请参阅3。评估和理解通用AI的方法系统)。

了解模型的内部计算可能有助于调查他们是否学习了值得信赖的解决方案。"机械解释性" 是指研究最先进的人工智能模型的内部工作原理。然而，最先进的神经网络庞大而复杂，并且机械可解释性尚未与其他分析实际应用模型的方法相比有用和竞争。尽管如此，一些研究人员已经thor研究了非常小的神经网络如何执行非常简单的任务 (*628， 630 629，*)。最近的一些工作尝试了更具可扩展性的技术来设计人类可解释的模型 (*282， 631， 742 632 *，*)。这种类型的方法不能用来排除有害或意外行为的可能性，但它可以提供一个有用的镜头来了解模型是如何工作的，这可能有助于理解模型的安全性。与其试图解释神经网络的内部计算，人们可以使用神经网络来解释可解释和可验证的解释，从而产生定量的安全保证 (*626*)，尽管如何有效地做到这一点仍然是一个开放的问题。

了解模型的内部工作有时可以用来指导编辑以有效地改变其行为。尽管很难理解模型的内部工作原理，但一些

可以使用技术来指导对它们的特定编辑。与微调相比，这些方法有时可以是修改其功能的计算或数据效率更高的方法。研究人员有used a v各种方法f或这个，基于对其i的更改nt胸骨parameters (*743，744，745， 746，747，748，749*) 和神经元 (*275，282，750*) 或表示 (*281、751、752、754 753、*)。然而，这些技术是不完美的 (*299*)，并且通常会对模型行为 (*755*) 引入意想不到的副作用。它们仍然是一个活跃的研究领域。

5.3 监测和干预

关键信息

- 有几种技术可用于识别通用AI系统风险，检查通用AI模型操作以及在部署通用AI模型后评估性能。这些做法通常被称为“监控”。同时，“干预”是指防止通用AI模型的有害行为的技术。
- 正在开发的用于解释通用AI行为的技术可以用于检测，然后进行干预以阻止危险行为。然而，这些技术在通用人工智能系统中的应用仍处于起步阶段。
- 用于检测和水印通用AI生成内容的技术可以帮助避免不成熟的用户对生成通用AI系统的一些有害使用。然而，这些技术是不完美的，并且可以被中等技能的用户规避。
- 从通用人工智能系统中识别异常行为的技术可以改善监督和干预。
- 在部署通用人工智能系统之前和期间，让人类参与进来，并进行其他检查，这增加了监督，并提供了多层次的故障防御。然而，这些措施可能会减缓通用人工智能系统的输出，可能会损害隐私，并可能与使用通用人工智能系统的公司的经济激励相冲突。

通用人工智能系统部署过程中的监控是指持续识别风险、检查模型和评估性能。干预措施可防止潜在的有害产出。而3。评估和理解通用人工智能系统的方法讨论了如何评估系统，以便围绕其使用做出更明智的决策，本节讨论了如何将一些监控和干预技术内置到人工智能系统本身中。

下面讨论了研究人员为通用AI系统监控和干预而开发的不同策略，包括检测AI生成的内容，检测风险情况，识别有害行为，解释模型行为以及进行干预以覆盖或阻止它们。

5.3.1 检测通用AI生成的内容

内容生成ed b y目的的人工智能系统-特别是深度伪造-可能会产生广泛的有害影响 (*757 756，* 见4.1。恶意使用风险)。能够区分真正的和通用的AI生成的内容，以防止恶意使用生成模型。

，存在一些不可靠的技术来检测通用AI生成的内容。就像不同的人有独特的艺术和写作风格一样，生成式AI模型也是如此。一些程序已被开发私奔区分AI生成的文本和人类生成的文本(*374，375，379，380， 758*) 和图像 (*759，760*)。检测方法通常基于专门的分类器或

评估给定示例由特定通用AI模型生成的可能性。然而，现有的方法是有限的，并且容易出现误报，因为通用AI系统倾向于记住出现在其训练数据中的示例，因此常见的文本片段或著名对象的图像可能被错误地识别为AI生成的。随着通用AI生成的内容变得更加现实，检测通用AI生成的内容可能更具挑战性。

水印使AI生成的内容更容易区分，但它们可以被删除。，“水印”是指可以插入到文件中的微妙样式或主题，人类很难注意到，但算法很容易检测到。用于图像的水印通常采取插入到图像像素中的不可感知的图案的形式 (*761*)，而用于文本的水印通常采取风格或单词选择偏差的形式 (*762 382*，)。水印是有用的，但t他们是检测AI生成的内容的不完美的策略，因为它们可以被删除 (*383 374*，)。然而，这并不意味着它们没有用处。作为类比，指纹很容易避免或去除，但它们在法医学中仍然非常有用。

水印也可用于指示真正的内容。与将水印插入通用AI生成的ted内容，一种对比的方法是将加密水印放入非AI生成的内容 (*763*)。然而，这将需要改变物理记录设备的硬件和软件，并且不清楚这些方法是否可以被篡改绕过。

5.3.2 检测异常和攻击

检测通用AI系统上的异常和攻击，可以在识别时采取预防措施。，已经开发了一些方法，可以帮助检测来自AI systems的异常输入或行为 (*765 764*，)。其他技术方法旨在检测给定输入的模型输出何时不确定，这可能表明存在攻击或错误输出的风险 (*766*)。一旦检测到这些示例，就可以将其发送到故障处理流程或标记以进行进一步调查。有时也可以检测和过滤significant恶意攻击在传递到通用AI模型之前的比例 (*723*，*767*) 或检测潜在的有害输出，以便在将其发送给用户之前将其阻止(*768*，*769*)。

5.3.3 解释模型操作

技术来解释为什么部署的通用人工智能系统的行为方式是新生的，还没有广泛应用，但有一些有用的方法。通用AI系统的行为可能很难理解。但是，了解模型以这种方式运行的原因对于e评估和确定通用AI系统造成的危害的责任非常重要 (*770 269*，)。不幸的是，简单地询问通用AI语言模型对其决策的解释往往会产生误导性的答案 (*771*)。为了提高模型解释的可靠性，研究人员正在研究改进的提示和训练策略 (*772 **，即*773**)。用于解释通用AI模型动作的其他技术 (*774*， *775*) 已被证明有助于调试 (*207*)。但是，正确解释通用AI模型动作是一个难题，因为通用AI系统的大小和复杂性超出了人类的理解范围。最先进的通用AI系统经过训练，可以产生积极增强的输出-而不是出于期望的原因或以自洽的方式这样做。

5.3.4 将保障措施构建到AI系统中

正如5.1所讨论的。风险管理和安全工程，虽然没有完善的安全措施，但有多层防护措施和冗余保障措施的增加

保证的水平。一旦检测到，干预措施可以识别并防止来自已部署的通用AI系统的潜在有害行为。

在循环中有一个人允许直接监督和手动覆盖。与自动化系统相比，循环中的人是昂贵的。然而，在高风险的决策情况下，它们是必不可少的。人类与人工智能合作范式不是教授通用人工智能系统代表人类行事，而是将通用人工智能系统和人类的技能和优势结合起来，在具有潜在有害后果的复杂情况下通常被认为是可取的。(*703，776，777 *，778，779*)。然而，在许多情况下，当决策发生得太快并且不能减慢时 (例如具有数百万用户的聊天应用程序)，当人类没有足够的领域知识时，或者当人为偏见或错误会加剧风险时 (*780*)。因此，循环中的人只能在某些情况下有用。

的自动化处理和过滤方法可以提供额外的但通常不完美的保护层。，一些针对通用人工智能系统的网络攻击通常在其输入中采取微妙、脆弱的模式。结果，研究人员开发了可以去除这些模式的输入预处理方法 (*723， 781， 782， 783， 785 784，*)。有时，att empt ed攻击可以在传递给gen eral- 目的AI模型之前被检测到并被阻止 (*767 723，*)，而有害输出可能在发送给用户之前被检测到 (*769 768，*)。这些方法可以显着降低某些故障的风险，但可能容易受到攻击。

安全接口可以设计用于具有潜在危险功能的通用AI系统。可以在网络或物理世界中自主和开放地行动的通用AI系统会带来更高的风险 (参见交叉技术风险因素4.4 1)。对于具有高风险能力的通用AI系统，限制它们直接影响人或物体的方式是降低潜在风险的一种建议方法 (*786 625，*)。

5.4 通用人工智能系统中公平和表示的技术方法

关键信息

- 通用AI模型可以捕获并有时放大其训练数据中的偏差。这导致资源分配不平等、代表性不足和歧视性决定。
- 公平缺乏普遍认可的定义，并且在文化，社会和学科背景下存在差异。
- 从技术角度来看，偏见的原因通常是数据，这些数据可能无法充分代表目标人群中的少数群体。偏见也可能源于糟糕的系统设计或使用的通用AI技术类型。这些选择取决于在整个通用AI生命周期中不同视角的参与。
- 偏见的缓解应该在通用人工智能系统的整个生命周期中解决，包括设计、培训、部署和使用。
- 完全防止当前通用人工智能系统中出现偏见是非常具有挑战性的，因为它需要系统的培训数据收集、持续的评估和有效的偏见识别，并将公平性与其他目标 (如准确性) 进行权衡。并决定什么是有用的知识，什么是不应该反映在产出中的不良偏见。
- 对于在通用人工智能系统中实现有意义的公平性有多可行，存在不同的看法。一些人认为，通用的人工智能系统不可能

完全“公平”，而其他人则认为从实际的角度来看，接近完全的公平是可以实现的。

公平没有普遍认可的定义，并且根据文化，社会和学科背景 (*333， 787， 788， 790 789，*) 而有所不同。从哲学上讲，公平需要对原则，价值观以及资源和机会的分配进行深刻的道德反思，而法律定义可能取决于宪法原则和判例法。由于公平的多面性，就公平达成共识被证明是难以捉摸的，需要参与不同的观点和具体情况的因素。

人工智能应用的兴起使算法公平性成为一个重要问题。AI中的公平性试图纠正自动决策或内容生成中的算法偏差。人工智能公平性可以用各种方式来定义和衡量 (例如“个人公平性”[6]与“f组”[7]) (*791*)，这被证明是适当的，具体取决于应用程序的上下文和具体目标 (*333*)，使人工智能模型的评估复杂化。算法决策文献中这种困境的一个经典例子是，用于预测刑事再犯的COMPAS软件是不公平AI模型的著名案例。通过一些mea ，发现COMPAS对非裔美国人有偏见，而通过其他措施则没有 (*790 788，*)。

当人工智能模型不公平时，它采取的行动是有偏见的，会伤害个人或社区。人工智能中的偏见是指基于其固有或获得的特征对实体的不合理偏爱 (*788*)。偏见与代表性和公平性紧密交织在一起。数据驱动算法的有效性取决于它们使用的数据的质量。然而，数据集通常无法充分代表少数群体，这可以在这些数据集上训练的AI中反映和放大 (*227*)。人工智能系统中的偏见可能导致社会伤害，包括不平等的资源分配和歧视ory 针对边缘化群体的决策 (*792*) 在各个领域提出重大挑战 (*793，794，795*)。

5.4.1 缓解偏见和歧视在通用人工智能开发和部署的整个阶段都有效

研究人员部署了各种方法来减轻或消除通用人工智能系统 (*2 *， 22**) 中的偏见并提高公平性，包括预处理、处理中和后处理技术 (*796， 797*)。预处理技术分析和纠正数据以消除数据集中存在的固有偏差，而处理中技术设计并采用学习算法以减轻系统训练阶段的歧视。后处理方法在部署后调整通用AI系统输出。

- 预处理偏差缓解-这包括针对训练数据中的缺陷，如有害信息或某些人口统计群体的代表性不足。两种流行的技术是数据增强和数据更改。数据扩充涉及通过创建现有数据的修改副本或生成合成数据 (*798，800 799，*) 来为代表性不足的组添加样本，以增加其在da taset中的代表性。数据更改会根据预定义的规则修改数据集样本，例如从数据语料库 (*801 227，*) 中添加，删除或屏蔽性别和种族等属性。

[6]个人公平强调对相似个人的相似待遇

另一方面，[7]群体公平性确保了不同受保护群体 (例如性别或种族) 成员的统计均等形式 (例如积极结果或错误之间)

- 处理中的偏差缓解-数据操作本身往往无法确保通用人工智能系统的公平性 (*803 802，*)。即使有完全代表人口分布的数据，社会中存在的不良刻板印象和偏见也会显现 (*803*)。预处理技术中的另一种策略是通过设计和利用专门的学习到t的orithms来将公平性集成到通用AI系统中，每个通用人工智能模型，用于生成无偏见的内容 (*804*、*805*、*806*、*807*、*808*、*810 809*、)。这可以通过人类向通用AI模型提供关于什么样的输出是和不是desi的反馈来实现rable(*2 *,22 *,702 *，811*) 或通过教导通用AI模型遵循人类指令 (*3 *,812，813，814*)。通用人工智能模型还可以相互传授公平性，将信息从偏见较小的通用人工智能模型 (“老师”) 转移到第二个通用人工智能模型 (“学生”) (*815*)。减轻通用AI系统偏差的常规方法是同时训练通用AI模型联合国有偏差和有偏差的数据样本，这样模型就能学会做什么和不做什么 (*816， 817，818*)。使u nder表示的属性更加突出是另一种实用的模型无偏差 (*819， 即820**) 的策略; 然而，当试图补偿表示的属性 (*821**) 时，它可能会损害隐私。
- 后处理减轻歧视-当不能直接修改完整模型时 (*822*)。另一种方法是仅操纵通用AI系统的输入或输出以实现公平性。通过提示修改gen通用AI模型输入有助于减少输出中的差别 (*823 2 *，*)。在许多情况下，诸如偏置/安全分类器之类的外部模块负责检测系统的不公平或不安全输出 (*812*)。然后，外部模块可以对输出进行排名 (*811*)，并要求通用AI系统重新生成输出以产生正确的、非歧视性的内容。如果主题高度敏感，则可能会对通用AI系统进行培训，拒绝回复试图提示有偏见和歧视性输出的用户 (*820 *， 822，824*)。这些方法侧重于操纵通用AI系统的输出，以使其更加公平。这些技术通常用于补充其他方法。

没有一种技术可以确保每种情况或结果的公平性。因此，领先的人工智能公司利用这些方法的组合来迭代地提高其通用人工智能系统的公平性 (*20， 825**)。

重要的是，增加有意义的代表和参与可以帮助减少代表人数不足的群体被排除在外的风险。从社会的角度来看，5.2.1将通用AI系统与开发人员意图联系起来讨论的AI对齐问题是不明确的。在一个多元化的社会中，不可能让一个通用的人工智能系统代表每个人的价值观，在这个社会中，人们有时不同意(*239，307，332*)。人们提出了增加参与 (*826*)，代表权 (*827*) 和对话 (*332*) 的方法，以减少与某些人的利益保持一致对其他人有害的风险。然而，某些形式的参与可能没有意义，并且不能完全解决不同人之间的分歧所带来的挑战 (*331*)。

5.4.2 通用AI系统的公平性是否可以实现？

通用人工智能系统是否可以完全“公平”。有支持和反对其可行性的论点。数学结果表明，在合理的假设 (*828， 公平829， 公平831*) 下，可能无法同时满足*830的*所有方面。结果表明训练无偏通用AI模型的复杂性 (*832， 833*)，这一不可能的公平性定理得到了支持。

许多理想的属性涉及权衡，例如系统公平性，准确性，隐私性和效率之间的四向权衡 (*834*)。研究表明，公平和o其他值，如隐私 (*821 *，834*) 和通用人工智能系统中的预测准确性 (*829， 835，836*)。一个可能的例子是Google Gemini，它生成了来自19世纪00年代的美国参议员和有色人种妇女的图像，以及“种族差异”的第二次世界大战时期的德国士兵。这些内容事实上歪曲了历史，可能是由于试图

确保未能预见并适应这些特定用例的种族多样性。为了避免过早地优先考虑无意中反映利益相关者个人价值的特定方面，技术可以使用重要技术人员可以理解的定量和定性措施，帮助他们做出明智的权衡决策，然后可以由开发的系统执行。

相反的论点是，虽然存在理论限制，但实际的解决方案是可以实现的 (*838 837，*)。一些研究人员认为，公平定义实际上可以相互协调 (*839， 840*)，并且有可能同时满足多个公平标准，至少在比通常更大的程度上满足多个公平标准 (*841*)。经验证据对公平性和准确性之间总是存在不可忽视的权衡的想法提出了挑战，表明这种权衡通常可以在实践中解决。这些研究表明，减少通用人工智能系统输出的差异可能并不一定会导致准确性显著下降，或者需要复杂的方法 (*838， 839， 840*)。

尽管进行了严格而全面的培训和测试工作，但建立在所有措施以及不同文化，社会和科学背景下公平的通用AI系统仍然具有挑战性。现有的任何措施都不能完全消除偏见和不公平的所有潜在风险，这些风险是开发高能力人工智能系统所固有的 (*837*)。尽管如此，仍有可能努力不断完善更公平的系统。

5.4.3 实现公平通用人工智能系统的挑战

尽管为消除通用AI系统的偏见做出了所有努力，但仍然存在重大挑战。首先，公平应该如何定义和衡量是辩论 (*842 802，*)。有用和准确的世界知识与强化有害的刻板印象之间的界限可能很难划清，偏见的感知可能会因情况而异 (*796 227，*)。其次，确保通用人工智能系统安全的其他方面可能会产生或放大偏见: 例如，清理数据以减轻毒性和隐私泄露可能会改变数据集的人口分布，导致更多的偏见 (*843*)。第三，交叉偏见等问题仍然难以解决 (*844*); 例如，通用人工智能系统可能分别对亚洲人和女性公平，但对亚洲女性有偏见。最后，缓解偏见需要在通用人工智能系统的开发、部署和使用过程中不断努力。各种形式的偏见可能会逐渐出现，需要专门的检测和缓解技术。

5.5 通用AI系统的隐私保护方法

关键信息

- 通用人工智能系统给人们的隐私带来了许多风险，例如数据机密性的丧失、透明度和对数据使用方式的控制，以及新形式的隐私滥用。
- 隐私保护是研究和开发的活跃领域。然而，现有的技术工具难以扩展到大型通用AI模型，并且可能无法为用户提供有意义的控制。

如4.3.5所述。先进的通用人工智能系统会给人们的隐私带来风险，例如失去数据机密性、透明度和对数据使用方式的控制，以及新形式的隐私滥用。现有技术和政策只能部分解决这些威胁。

当前的隐私增强技术无法扩展到大型通用AI模型。虽然各种隐私技术可以应用于人工智能模型，以保护个人隐私，同时仍然允许从数据中获得有用的见解 (*846 845，*)，但这些技术可能会严重损害模型的准确性，难以扩展到大型模型。并且可能不适合所有用例，特别是针对文本训练的通用AI模型 (*847*)。对于具有高度敏感数据 (例如，医疗或金融) 的领域，可以通过采用功能强大的gen通用AI模型来获得强大的隐私保证，这些模型首先在互联网上公开可用的数据上进行了预先训练 (*849 848，*)。但迄今为止，这种技术很少应用于生产。另一种解决方案是使用合成数据，以避免在通用AI系统训练管道中使用敏感数据。然而，研究人员已经证明存在重要的效用/隐私权衡。如果合成数据效用高，则hen它们可能携带与原始数据一样多的信息，并启用大多数相同的攻击 (*850， 852 851，*)。

原则上，通用人工智能系统提出的机密性和数据集中化问题可以使用安全计算解决方案来解决，如加密方法 (*853*) 、联合学习 (*854*) 和硬件保护 (*855*)。但是，现有技术尚未扩展到当今正在训练的最大和最有能力的模型。这些解决方案也都带来了可能在规模上令人望而却步的成本，尽管研究人员正在努力寻找降低这些成本的方法。加速器的硬件保护方面的进步，这是一类专门的硬件，旨在加速人工智能应用，如人工神经网络和机器视觉，将来可以为训练和运行通用AI模型提供实用的途径，而无需访问敏感数据。

解决其他领域缺乏数据透明度和控制的**措施可以应用于通用人工智能系统。**为个人开发更好的机制来控制和追踪他们的数据，这将提高通用人工智能系统的透明度和问责制。这包括提供用于管理数据权限的用户友好界面，实施安全的数据来源系统以跟踪数据的使用和共享方式，以及为个人建立访问，查看，更正和删除其数据的清晰流程 (*856*)。提供这些控制的技术手段已经存在，并已成功部署在其他领域 (例如，用户控制的仪表板允许用户决定各种网站或公司如何收集或使用其个人数据)。这种和其他方法可能会扩展到通用AI系统。也有可能以更可追溯和更公平的方式重新分配从个人数据中获得的财富，例如通过使用经济工具进行数据评估 (*857*)。

然而，为人们提供透明度和控制他们的*公共*数据的使用方式 (即在网络上很容易找到的信息) 更具挑战性。虽然社交媒体平台等个别服务提供商可以禁止其数据被外部通用人工智能系统使用，但控制权最终不在最终用户手中，只覆盖网络上的一小部分数据。

另一个挑战是为*派生*数据的使用提供有意义的控制，或未识别但允许推断一个人的数据。这种情况可能在使用个人数据的人工智能系统中很常见。需要更多的研究来探索减少未经授权的数据使用和共享风险的方法。

某些形式的隐私滥用很难通过技术手段来防止。，由于缺乏数据透明度和控制，很难通过技术手段防止由通用AI引起的新形式的隐私滥用，例如未经同意的深度伪造或跟踪。一些法律框架旨在让创作者和发行商对恶意使用 (*858*) 负责，并为隐私受到侵犯的个人提供补救措施。最近的一些法规还要求以尊重隐私原则的方式开发和部署人工智能系统，例如通过执行数据最小化和目的限制 (*860 859，*)，但如何实现这些属性。或者它们在多大程度上可以实现是值得怀疑的 () *861*。

6 结论

这份关于高级人工智能安全性的中期国际科学报告发现，未来通用人工智能的发展轨迹非常不确定。即使在不久的将来，也可能出现各种可能的结果，包括非常积极和非常消极的结果，以及介于两者之间的任何结果。通用人工智能最有希望的前景之一是它在教育、医疗应用、广泛领域的研究进展以及带来更多繁荣的生产力提高方面的潜力。如果管理得当，通用人工智能系统可以大大改善全世界人民的生活。

但是，为了安全地获得这种变革性技术的好处，研究人员和政策制定者需要确定并采取明智的行动来减轻随之而来的风险。通用人工智能的恶意使用以及通用人工智能的故障已经在今天造成了伤害，例如通过深度伪造、诈骗和有偏见的输出。根据未来通用人工智能能力的发展速度、开发人员和监管机构为减轻风险而采用的技术方法、政府和社会在通用人工智能方面的决策以及全球协调的成功程度，也有可能出现进一步的风险。最糟糕的结果可能是出现大规模失业、通用人工智能恐怖主义、甚至人类失去对通用人工智能系统的控制等风险。专家们对这些风险的可能性有多大以及何时可能发生没有达成共识。

本报告还研究了使解决这些风险变得困难的因素。尽管功能迅速发展，但研究人员目前无法对通用AI模型和系统如何得出输出和决策产生人类可理解的解释。这使得很难评估或预测他们的能力，他们有多可靠，并获得他们可能带来的风险的保证。

有一些技术方法可以解决通用AI的风险: 减少模型偏差的方法，提高我们对通用AI模型内部工作原理的理解，
评估他们的能力和潜在风险，并使他们不太可能响应可能造成伤害的用户请求。还有一些补充技术可以监控和减轻通用人工智能系统的有害行为。然而，目前没有现有技术提供关于高级通用AI模型或系统安全性的定量保证。

关于AI的未来，没有什么是不可避免的。通用人工智能是如何开发的，由谁开发，它旨在解决哪些问题，我们是否能够获得通用人工智能的全部经济潜力，谁从中受益，我们面临的风险类型 -- 这些和许多其他问题取决于社会和政府今天和未来在塑造通用人工智能发展方面做出的选择。由于通用人工智能对我们生活的许多方面的影响可能是深远的，而且进展可能会继续迅速，因此预防原则意味着迫切需要达成共识，并投入资源来理解和应对这些风险。建设性的科学和公共讨论对于社会和政策制定者做出正确的选择至关重要。

这份中期报告有史以来第一次汇集了30个国家、欧盟和联合国提名的专家代表，以及其他几位世界领先的专家，为这些重要的讨论提供了共同的科学、循证基础。我们仍然在围绕通用人工智能的能力、风险和风险缓解的几个问题上存在分歧，无论是次要的还是主要的。但我们认为，这个项目对于提高我们对通用人工智能及其潜在风险的集体理解，以及更接近达成共识和有效的风险缓解，以确保人们能够安全地享受通用人工智能的好处至关重要。赌注很高。我们期待着继续这一努力。

主席关于中期报告的说明

这份中期报告是一个多元化的大型人工智能专家小组之间快速合作的产物，其中包括由30个国家以及欧盟和联合国提名的专家咨询小组。人工智能研究领域正在飞速发展，在许多重要问题上，该领域远未达成共识。在这种背景下，我对75名国际专家在如此短的时间内为本报告提供了不同观点的成就感到特别印象深刻。撰写一份报告，以平衡的方式讨论通用人工智能的能力、风险和潜在风险，对我来说尤为重要。我非常感谢为报告做出贡献的专家们以协作精神对待这一重要项目。

编写这份临时报告的时间很短，这也意味着必须就报告的范围作出几项困难的程序性决定和决定，许多重要问题没有得到解决，或者只是简短地讨论。我的下一份出版物将以这份中期报告为基础，我的目标是让特约专家共同确定下一份报告需要改进的最重要领域。

例如，这可能包括以下一些方面:

- 我们的目标是进一步改进报告评估和综合有关能力和风险的科学证据的方式。鉴于时间很短，我对我们如何在临时报告中做到这一点感到满意，但对于下一份报告，可以而且应该有所改进。我特别希望下一份报告:
  - 考虑到更多的科学工作，以提供对文献的更全面的讨论；
  - 在某些情况下，更明确地说明基于其方法论的特定研究内容的吸引力；
  - 在综合证据方面做得更好，以便对特定问题的科学状况提供更细微和简洁的评估。
- 对于临时出版物，我们将对报告的投入限制为撰写小组，高级顾问和我们在特定部分所需的专家投入，包括来自非常有限的民间社会组织。我们有意排除私营部门代表的意见，以避免利益冲突。然而，公司一直是人工智能能力最新进展的主要驱动力，也是我们引用的风险评估和缓解研究的积极贡献者。同样，民间社会组织有一个活跃而多样化的生态系统，其工作对于该领域对人工智能风险和风险缓解的理解至关重要。因此，既然临时报告已经发表，我们想扩大对报告的投入。我们呼吁公司和民间社会提交证据，以便为下一份完整报告的发展提供信息。我们将在适当的时候详细介绍如何做到这一点。同时，如果你有证据要提交，请给secretariat.AIStateofScience@dsit.gov.uk发电子邮件。
- 确保报告对高级人工智能风险的全面讨论一直是写作团队和高级顾问的首要任务。但是，我们从一开始就很清楚，编写临时报告的时间很短，不足以以应有的深度解决所有问题。我希望下一份报告更详细讨论的一个主题的例子是全球“人工智能鸿沟”，它阻止世界各地的人们平等地享受人工智能的任何好处。下一份报告将比本报告更详细讨论的另一个主题是AI开发和部署对环境的影响。为了使报告更加全面，我们将继续意识到证据的广度，深度和覆盖范围之间的潜在权衡，并将寻求我们认为对决策者最有用的平衡。

- 我们将本中期报告的范围限制在通用人工智能。我相信这是正确的决定，但很明显，各种类型的人工智能模型和系统对干净地定义任何人工智能子组都是一个挑战。对于下一份报告，无论其范围如何，我们都将致力于更详细地描述报告范围内和范围外的AI类型。

我非常感谢为这份临时报告做出贡献的所有专家，并期待着为下一份出版物开展工作。

不同的观点

主席和秘书处努力达成专家咨询小组成员之间关于报告内容的共识。根据报告的原则和程序，与专家咨询小组分享了中期报告的近最终版本。在小组仍有不同意见的地方，向小组成员提供了注意到这些意见的选择。以下是所指出的观点。

1. Ciar á n Seoighe (爱尔兰)

注意到对报告的总基调过于消极的关切。尽管报告确实指出人工智能的未来不是预先确定的，但所使用的语言可能会给人留下这样的印象，即无论采取什么步骤，人类的前景都是黯淡的，因此，报告对政策制定者的影响可能会受到损害。

2. Ciar á n Seoighe (爱尔兰)

要求报告的未来迭代应包括:

- 从技术和社会技术角度评估生成人工智能在信息生态系统、教育和民主进程等领域的社会影响。同样重要的是，要纳入一系列可以减轻社会风险的监管方法以及技术解决方案。
- 在开发和部署人工智能系统之前，人权/基本权利评估的重要性及其在减轻风险中的作用。指出这是报告当前版本中的一个重大遗漏。

词汇表

以下解释应全部用于在AI或通用AI的上下文中使用该术语。

适应性: 在人类程序员不直接设想的上下文和方式中或在系统训练数据的上下文之外识别模式、推理和做出决策的能力。

人工智能代理/自主代理: 能够在很少或没有人工监督的情况下完成多步骤任务以追求高级目标的人工智能系统。人工智能代理可以做一些事情，比如浏览互联网、发送电子邮件或向物理设备发送指令。

AI部署者: 提供或使用AI系统来提供产品或服务的任何个人或组织。部署可以是“内部的”，其中系统仅由开发人员使用，也可以是“外部的”，允许公共或其他非开发人员实体使用它。

AI开发人员: 设计、构建、训练、调整或组合AI模型和应用程序的组织或个人。

AI最终用户: 在部署时使用或消费基于AI的产品或服务的任何预期或实际个人或组织。

人工智能生命周期: 与人工智能系统的生命周期相关的所有事件和过程，从启动到退役，包括其设计、研究、培训、开发、部署、集成、操作、维护、销售、使用和治理。

人工智能风险: 由人工智能模型或系统的开发或部署引起的伤害发生的可能性以及伤害的严重程度的组合。

算法透明度: 告知通用人工智能输出的因素 (如建议或决策) 在多大程度上被各种利益相关者所了解。这些因素可能包括人工智能模型的内部工作原理，如何训练，训练的数据是什么，输入的哪些特征影响了输出，以及在不同情况下会做出什么决定。

一致性: 确保AI系统的目标和行为符合开发人员的价值观和意图的过程。

应用程序编程接口 (API): 支持AI系统和其他软件应用程序之间集成和通信的一组规则和协议。

人工通用智能 (AGI): 未来潜在的人工智能系统，在所有或几乎所有认知任务上的表现都等于或超过人类。许多人工智能公司已经公开表示他们的目标是建立AGI。然而，AGI一词没有普遍一致的定义。

自主性/自主: 在没有人类明确意图或监督的情况下能够操作、采取行动或做出决定。

生物设计工具 (BDTs): 在此报告中，生物设计工具 (BDTs) 是指经过生物数据训练的AI系统，可以帮助设计新的蛋白质或其他生物制剂，例如酶。

黑盒: 部署有限制的系统，使得用户无法访问或分析其内部工作。另请参见下面的“白盒”。

能力: AI系统可以执行的任务或功能的范围以及它可以执行它们的熟练程度。

云实验室: 自动化远程控制生化实验室。

认知任务: 涉及信息处理，记忆，信息回忆，计划，推理，组织，解决问题，学习和面向目标的决策的任务。

计算: 训练和运行通用AI模型所需的大量计算资源。主要通过图形处理单元 (gpu) 集群提供。

跨语言差异: 通用AI模型或系统如何响应不同语言的相同输入的差异。

深度学习: 利用大量数据和计算的一组AI开发方法。

部署: 将AI系统发布到现实环境中的过程，例如面向消费者的AI系统。

虚假信息: 为了欺骗或误导而故意产生或传播的虚假信息。

生态系统审计: 对人工智能系统及其周围生态系统的广泛评估。生态系统审计可能会考虑人工智能模型、其训练数据、部署环境以及周围的运营实践。

评估: 对人工智能系统的性能、能力或潜在影响进行系统评估。评估可以包括基准测试、红队和审计。

FLOPS: “每秒浮点运算”-衡量计算机计算能力的指标。

基础模型: 在大量数据上训练的机器学习模型，可以适应各种任务。

Frontier AI: 在Bletchley Park的AI安全峰会上，frontier AI被定义为可以执行各种任务并匹配或超过当今最先进模型中存在的能力的模型。

GPU (图形处理单元): 由半导体组装而成的计算机硬件，广泛用作通用AI的计算能力的中心来源。Gpu最初是为图形渲染应用程序设计的。

护栏: 设置预定义的安全约束或边界，以确保AI系统在所需参数范围内运行，并避免意外或有害的结果。

启发式: 经验法则，策略或简化原则，在计算机科学的背景下，当经典方法太慢或无法找到确切的解决方案时，已开发用于更有效地解决问题。

输入 (到AI系统): 输入AI系统的数据或提示，通常是文本或图像，AI系统在生成输出之前对其进行处理。

大型语言模型 (LLMs): 在大型数据集上训练的机器学习模型，可以识别，理解和生成文本及其他内容。

大规模多任务语言理解 (MMLU): 广泛使用的基准AI研究，评估通用AI模型在广泛任务和主题领域的性能。

错误概括: 训练在一种环境中表现良好的人工智能系统在新的环境中表现不佳。例如，如果一个人工智能主要在白猫的照片上训练，将一只黑猫贴上 "狗" 的标签，那么它就是从它的训练数据中错误地概括出来的。

错误信息: 可能在没有有害意图的情况下产生和传播的不正确或误导性信息。

模态: AI模型可以处理的数据的类型和性质，例如文本，图像，声音或视频。模型可以是单峰的，即仅能够处理一种类型的数据，或者是多模态的，即能够处理多种类型的数据。

模型卡: 提供有关通用AI模型的重要信息的文档，例如其目的，评估和基准测试的性能以及安全功能。

窄AI: 仅在单个任务或一组狭窄的任务 (如情绪分析或下棋) 上表现良好的AI系统。

开放式域: 具有大量可能状态和AI系统输入的场景或环境，因此开发人员无法预测所有类型的使用环境，因此无法测试一下AI在所有可能情况下的行为。

预训练: 开发现代通用AI模型的第一阶段，模型从大量数据中学习。预训练是通用AI训练的一部分，需要最多的数据和计算资源。

提示: AI系统的输入，通常是基于文本的问题或查询，系统在产生响应之前会处理。

Red-teaming: 通过尝试设计使系统失败的输入来评估系统安全性和鲁棒性的方法。这通常是通过开发 "对抗性攻击" 或具有挑战性的条件来实现的。Red-teaming试图揭示最坏情况下的行为或恶意使用机会。

风险因素: 可能增加下游风险的要素或条件。例如，薄弱的护栏构成了一个风险因素，可能使攻击者恶意使用AI系统执行网络攻击 (下游风险)。

安全和保障: 公民社会和人民的保护、福祉和自治。在本出版物中，安全性通常用于描述预防或保护与AI相关的危害。人工智能安全是指保护人工智能系统免受网络攻击或人工智能模型代码和权重泄露等技术干扰。

Scaffold: 帮助处理AI模型的输入和输出，同时保持模型本身不变的附加软件。例如，支架允许GPT-4为自主AI代理AutoGPT供电。scaffold提示GPT-4将高级任务分解为子任务，将子任务分配给自身的其他副本，将重要信息保存到内存中，并浏览internet。

半导体: 现代计算机硬件的基本材料组件，如gpu。

合成数据: 例如通过通用AI模型人工生成的数据，例如文本、图像等。合成数据可用于训练通用AI模型，例如在缺乏高质量自然数据的情况下。

系统集成: 将不同的软件元素组合成一个内聚系统以执行某些功能的过程。例如，系统集成可以将通用AI模型、内容过滤器、用户界面和各种其他组件组合到聊天机器人应用程序中。

迁移学习: 一种机器学习技术，其中模型在一个任务或主题区域上完成的训练被用作在另一个主题区域上训练或使用模型的起点。

Transformer架构: 大多数现代通用AI模型核心的深度学习架构。transformer架构已被证明在将越来越多的大量训练数据和计算能力转换为更好的模型性能方面特别有效。

权重: 模型中的参数，类似于算法中的可调拨盘。训练模型意味着调整其参数，以帮助它根据输入数据做出准确的预测或决策，确保它从所看到的模式中学习。

白盒: 不受限制地部署的系统，使得用户可以访问或分析其内部工作。另请参见上面的“黑匣子”。

参考文献

* 表示该参考文献是由AI公司发布的，或者至少50% 预印文章的作者都有AI公司作为其隶属关系。

1. 西蒙斯-埃德勒，R.，巴德曼，R.，朗普雷，S.，和拉詹，K。(2024)。人工智能驱动的自主武器面临地缘政治不稳定的风险，并威胁人工智能研究。arXiv预印本arXiv:2405.01859。在线: https://arxiv.org/abs/ 2405.01859。

2. * OpenAI，J。Achiam, S.阿德勒，S.Agarwal, L.艾哈迈德，我Akkaya, F.L.Aleman, D.阿尔梅达，J.Altenschmidt, S.奥特曼，S.Anadkat, R.阿维拉，我.Babuschkin, S.巴拉吉，V。Balcom, P.巴尔特斯库，H.鲍，M。巴伐利亚，J.贝尔古姆。。。B. Zoph，"GPT-4技术报告" (OpenAI，2024)；http://arxiv.org/abs/ 2303.08774。

3. * 双子座团队，R。Anil, S.Borgeaud，J.-B.Alayrac, J.于，R。索里卡特，J.Schalkwyk, A.M。戴，A。Hauth, K.Millican, D.银，M.约翰逊，我。安东诺格鲁，J.Schrittwieser, A.Glaese, J.陈，E。皮特尔r，T.Lillicrap, A.拉扎里杜。。。O. Vinyals，"双子座: 一系列功能强大的多模式模型" (Google DeepMind，2023)；http://arxiv.org/abs/ 2312.11805。

4. * Anthropic，"克劳德3模型家族: Opus，Sonnet，Haiku" (Anthropic，2024); https:// www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf。

5. * Qwen团队，J。白，S。白，Y。楚，Z。崔，K。当当，X。邓勇。范，W。Ge，Y。韩，F。黄，B。Hui, L.Ji, M.李，J。林，R。林，D。刘，G。刘，C。陆，。。。T. Zhu，Qwen技术报告，arXiv:2309.16609 [cs.CL] (2023)。http://arxiv.org/abs/ 2309.16609.

6. * Meta，使用Meta Llama 3 (2024) 构建AI的未来。https://llama.meta.com/llama3/。

7. * A.问:江，A。Sablayrolles, A.门施，C。班福德，D。S。查普勒，D.德拉斯卡萨斯，F.布雷桑，G。Lengyel, G.Lample, L.索尔尼尔，L.R。Lavaud，M.-A。Lachaux, P.股票，T. Le Scao，T.Lavril, T.王，T。拉克鲁瓦，W. El Sayed，"Mistral 7B" (Mistral AI，2023); https://doi.org/10.48550/arXiv.2310.06825。

8. L.杨，Z。张，Y。宋，S.洪，R。徐，Y。赵，W。张，B。崔，M.-H。杨，扩散模型: 方法和应用的综合考察。*ACM 计算机。Surv*。**56**，1-39 (2023)。https://doi.org/10.1145/3626235。

9. * OpenAI，"DALL·e3系统卡" (OpenAI，2023)；https://cdn.openai.com/papers/ DALL_E_3_System_Card.pdf。10.* Midjourney，

Midjourney文档 (2024)。https://docs.midjourney.com/v1/en。

11. * 稳定性AI，稳定扩散3 (2024)。https:// stability.ai/新闻/稳定扩散-3.

12. * T。布鲁克斯，B。皮布尔斯，C。福尔摩斯，W.DePue, Y.郭，L。静，D。施努尔，J.泰勒，T。L uhman, E.卢曼，C。Ng, R.王，A.Ramesh，"作为世界模拟器的视频生成模型" (OpenAI，2024); https://openai.com/research/video-一代模型作为世界模拟器。

13. D.德里斯，F。夏，M。S。M。Sajjadi, C.林奇，A。Chowdhery, B.伊克特，A.瓦希德，J.汤普森，Q.Vuong, T.于，W。黄，Y。Chebotar，P。Sermanet, D.达克沃思，S.莱文，V。Vanhoucke，K。豪斯曼，M。Toussaint, K.格里夫。。。P. Florence，"palm-e: 一种体现的多模式语言模型"，*载于第40届国际会议机器学习 (icml'23)* 论文集 (PMLR，2023)。202，第页。8469-8488。https://dl.acm.org/doi/ 10.5555/3618408.3618748。

14. J.艾布拉姆森，J。阿德勒，J.Dunger, R.埃文斯，T。绿色，A。Pritzel, O.Ronneberger, L.Willmore, A.J.巴拉德，J.班布里克，S。W.博登斯坦，D.A.埃文斯，C.-C。洪，M。O'Neill, D.Reiman, K.Tunyasuvunakool，Z。吴，A。Žemgulytė, E.Arvaniti, .。。J.m. Jumper，生物分子与AlphaFold 3相互作用的精确结构预测。*自然* (2024)。https://doi.org/10.1038/s41586-024-07487-w。

15. Y。LeCun，Y。Bengio, G.Hinton，深度学习。*自然***5 21**，436- 444 (2015)。https://doi.org/10.1038/自然14539。

16. A.Vaswani, N.Shazeer, N.帕尔马，J.Uszkoreit，L。琼斯，A。N.戈麦斯，Ł.U. Kaiser, I.Polosukhin，"注意力是你所需要的"，*进展神经信息处理系统 (NIPS 2017)* (Curran Associates，inc.，2017) 卷。30。https:// papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

17. J.塞维利亚，L。海姆，A.浩，T。贝西罗格鲁，M.霍布哈恩，P。Villalobos，*" 2022国际联合会议神经网络 (IJCNN 2022)* 的" 三个机器学习时代的计算趋势 "(2022) pp。1-8。https://doi.org/10.1109/ijcnn55064.2022.9891914.

18. C.周，P。刘，P。徐，S。艾尔，J.太阳，Y。毛，X。妈，A.Efrat，P。于，L。于，S。张，G。戈什，M.刘易斯，L。Zettlemoyer,
O.Levy，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上的 "利马: 少就是多" (2023)。https://openreview.net/forum？id= KBMOKmX2he。

19. R。拉法洛夫，A.Sharma, E.米切尔，C。D.曼宁，S.Ermon，C。Finn，"直接偏好优化: 您的语言模型是秘密的奖励模型"，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 中 (2023)。https://openreview.net/forum？id= hputixjaa9 & utm。

20. L.欧阳，J。吴，X。江，D。阿尔梅达，C。温赖特，P。米什金，C。张，S。阿加沃尔，K。Slama, A.格雷，J.舒尔曼，J。希尔顿，F.Kelton, L.米勒，M。西门斯，A.Askell，P。Welinder，P。克里斯蒂亚诺，J。雷克，。。。R. Lowe，“训练语言模型以遵循具有人类反馈的指令”，在*第36届神经信息处理系统会议 (NeurIPS 2022)* 中 (2022)。https://openreview.net/forum？id= TG8KACxEON。

21. * Y。白，A。琼斯，K.Ndousse, A.Askell, A.陈，N。DasSarma, D.排水管，S.Fort, D.甘古丽，T.Henighan, N.约瑟夫，S.Kadavath, J.Kernion, T.康纳利，S。El-Showk, N.Elhage，Z。Hatfield-Dodds, D.埃尔南德斯，T。休谟。。。J.卡普兰，训练a有益无害的助手，从人类反馈中强化学习，arXiv:2204.05862 [cs.CL] (2022)。https://doi.org/10.48550/arXiv.2204.05862。

22. * H. Touvron，L。马丁，K。斯通，P.艾伯特，A.Almahairi, Y.Babaei, N.Bashlykov, S.巴特拉，P。Bhargava, S.Bhosale，D。比克尔，L.布莱克，C。C.费雷尔，M.陈，G。Cucurull, D.艾斯奥布，J.费尔南德斯，J.傅，W。傅，。。。T. Scialom，“Llama 2: 开放基础和微调聊天模型”(Meta AI，2023)；http://arxiv.org/abs/ 2307.09288。

23. L.Sharkey, C.Ní Ghuidhir, D. Braun, J.Scheurer, M.Balesni, L.布什纳克，C。斯蒂克斯，M.Hobbhahn，人工智能监管和审计的因果框架。(2024)。https://doi.org/10.20944/预印本202401.1424.v1。

24. J.魏，X。王，D。舒尔曼斯，M。波斯马，B.伊希特，F.夏，E。Chi, Q.V。Le, D.Zhou，“思维链在大型语言模型中引发推理”，*《神经信息处理系统进展》(NeurIPS 2022)* (2022)。35页。24824-24837。https:// 会议录.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html。

25. T.戴维森，J.- S.Denain, P.维拉洛沃斯，G。Bas，“无需昂贵的再培训就可以显着提高AI能力”(Epoch，2023)；http://arxiv.org/abs/ 2312.07413。

26. L.王，C。妈妈，X.冯，Z。张，H。杨，J。张，Z。陈，J。唐，X。陈，Y。林，W。X。赵Z.魏，J。温，基于大语言模型的自主代理调查。*前面。计算机。Sci。186345* **18** (2024)。https://doi.org/10.1007/s11704-024-40231-1。

27. R。Bommasani，D。Soylu, T.I.李敖，K。A.克里尔，P。梁，生态系统图: 基础模型的社会足迹，arXiv:2303.15772 [cs.LG] (2023)。https://doi.org/10.48550/arXiv.2303.15772。

28. * A.达斯，W。孔，R。森，Y.周，时间序列预测的纯解码器基础模型，arXiv:2310.10688 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2310.10688。

29. * P。Dhariwal，H.6月，C.佩恩，J。W。金，A。Radford，我.Sutskever，“点唱机: 音乐的生成模型”(OpenAI，2020)；http://arxiv.org/abs/ 2005.00341。

30. T.布朗，B。曼恩，N.莱德，M。Subbiah, J.D.卡普兰，P。Dhariwal, A.Neelakantan，P。希亚姆，G.Sastry, A.Askell, S.阿加沃尔，A。赫伯特-沃斯，G.克鲁格，T。Henighan，R。孩子，A。拉梅什，D。齐格勒，J.吴，C。冬天，。。。D.Amodei，“语言模型是少数学习者”，*《神经信息处理系统*进展》(Curran Associates，inc.，2020) 第一卷。444，第页。1877-1901。https:// papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

31. * S.皮查伊，D。哈萨比斯，我们的下一代模特: 双子座1.5。(2024)。https:// blog.google/技术/ai/google-gemini-next-generation-model-february-2024/。

32. Fan，Gokkaya，Harman，Lyubarskiy，Sengupta，Yoo，Zhang，“软件工程的大型语言模型: 调查和开放问题”，*2023 IEEE/ACM国际软件工程会议: 软件工程的未来 (icse-fose)*，2023卷。0，pp。31-53。https://doi.org/10.1109/ICSE-FoSE59343.2023.00008。

33. * Anthropic，介绍下一代的克劳德 (2024)。https://www.anthropic.com/news/ claude-3-family。

34. A.Dosovitskiy, L.拜尔，A.科列斯尼科夫，D。Weissenborn, X.翟，T。Unterthiner, M.Dehghani, M.矿工，G。Heigold, S。Gelly, J.Uszkoreit, N.霍尔斯比，“一幅图像值16x16个字: 大规模图像识别的Transf ormers”*在第九届国际学习表示会议 (ICLR 2021)* 上发表 (2021)。https://openreview.net/forum？id=YicbFdNTTy。

35. A.雷德福，J.W。金，C。哈拉西，A.拉梅什，G.Goh, S.阿格沃尔，G。Sastry，A。Askell，P。米什金，J.克拉克，G。克鲁格，我Sutskever，“从自然语言监督中学习可转移的视觉模型”，*载于第38届国际机器学习会议 (ICML 2021)* 论文集 (PMLR，2021) pp。8748-8763。https:// 会议记录.mlr.press/v139/radford21a.html。

36. * A.Kirillov, E.明顿，N。拉维，H.毛，C。罗兰，L.古斯塔夫森，T。肖，S.怀特海，A.C.伯格，W.-Y。罗，P。Dollár, R.Girshick，“分割任何东西”(Meta AI，2023)；http://arxiv.org/abs/ 2304.02643。

37. * Meta，v-jepa: 迈向高级机器智能的下一步 (2024)。https://ai.meta.com/blog/ v-jepa-yann-lecun-ai-模型-视频-联合-嵌入-预测-架构/。

38. * OpenAI，Sora: 从文本创建视频 (2023)。https://openai.com/sora。

39. B.伊克特，A.布罗汉，Y.Chebotar，C。芬恩，K.豪斯曼，A。赫尔佐格，D.浩，J。Ibarz, A.Irpan, E.Jang, R.朱利安，D。卡拉什尼科夫，S.莱文，Y。卢，C。帕拉达，K。Rao, P.Sermanet, A.T.托舍夫，V。万豪科，。。。C. K.Fu，“尽我所能，而不是像我说的那样: 在机器人能力*中扎根语言”，载于第六届机器人学习年会 (CoRL)* 论文集 (PMLR，2022) 卷。205。https://openreview.net/forum？id= bdHkMjBJG_w。

40. * Q.Vuong, P.Sanketi，扩展跨许多不同机器人类型的学习 (2023)。
https:// deepmind.google/discover/blog/跨多个不同机器人类型/扩展学习

41. A.Khazatsky, K.Pertsch, S.奈尔，A.Balakrishna, S.达萨里，S。Karamcheti, S.Nasiriany, M.K。Srirama, L.Y。陈，K。埃利斯，P。D.费根，J.Hejna, M.Itkina, M.Lepert, Y.J.马，P。T.米勒，J.吴，S。Belkhale, S.达斯，。。。C. Finn，DROID: 大规模的野外机器人操纵数据集，arXiv:2403.12945 [cs.RO] (2024)。https://doi.org/10.48550/arXiv.2403.12945 。

42. 开放X-实施例协作，A.O'Neill, A.Rehman, A.Maddukuri, A.古普塔，A.Padalkar, A.李，A.Pooley, A.古普塔，A.曼德莱卡，A.Jain, A.董，A。Bewley, A.赫尔佐格，A.Irpan, A.Khazatsky, A.Rai, A.古普塔，A.王，。。。Z. Lin，开放X-实施例: 机器人学习数据集和rt-x模型，arXiv:2310.08864 [cs.RO] (2023)。
https://doi.org/10.48550/arXiv.2310.08864。

43. A.Madani, B.克劳斯，E.R。格林，S.Subramanian, B.P。莫尔，J.M。霍尔顿，J。L.奥尔莫斯，C.熊，Z。Z。太阳，R。索彻，J.S、弗雷泽，N。Naik，大型语言模型生成跨不同家族的功能蛋白质序列。*Nat。生物技术。*
41 ，1099-1106 (2023)。https://doi.org/10.1038/s41587-022-01618-2。

44. P。科比，G。波扎蒂，A。Elofsson，使用alphafold2改进了蛋白质-蛋白质相互作用的预测。*Nat Commun*，**13** ，1265 (2022)。https://doi.org/10.1038/s41467-022-28865-w。

45. X。胡某，J.陈，X。李，Y。郭，L。温，P。S。于，Z。郭，大型语言模型知道事实吗？，arXiv:2310.05177 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2310.05177。

46. Z。Ji, N.李，R。弗里斯克，T。于，D。苏，Y。徐，E。石井，Y。J .秤，A。Madotto，P。冯，自然语言生成中的幻觉调查。*ACM计算机。Surv。***55** ，1-38 (2023)。https://doi.org/10.1145/3571730。

47. * Y。张，Y。李，L。崔，D。蔡，L.刘，T。傅，X。黄，E。赵，Y。张，Y。陈，L。王，A。T.Luu, W.Bi, F.施，S.Shi，AI海洋中的Siren之歌: 大型语言模型中的幻觉调查，arXiv:2309.01219 [cs.CL] (2023)。
http://arxiv.org/abs/ 2309.01219。

48. 张M，O.新闻，W. 美林，A。刘，N。A.史密斯，语言模型幻觉如何滚雪球，arXiv:2305.13534 [cs.CL] (2023)。http://arxiv.org/abs/ 2305.13534。

49. L.黄，W。于，W。马，W。钟，Z。冯，H.王，Q。陈，W。彭，X。冯，B。秦，T。Liu，大型语言模型中的幻觉调查: 原理，分类学，挑战和开放性问题，arXiv:2311.05232 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2311.05232。

50. V。Rawte, A.Sheth, A.Das，“大型基础模型中的幻觉调查”，arXiv:2309.05922 [cs.AI] (2023)。
http://arxiv.org/abs/ 2309.05922。

51. E。戴维斯，数学，单词问题，常识和人工智能。*公牛。上午。数学。Soc。***61** ，287-303 (2024)。
https://doi.org/10.1090/牛市/1828。

52. 问:东，L。李，D。戴，C. 郑，Z。吴，B。张，X。孙，J。徐，L。李，Z。隋，情境学习调查，arXiv:2301.00234 [cs.CL] (2022)。http://arxiv.org/abs/ 2301.00234。

53. W。赵，J。T.Chiu, J.D.黄，F。Brahman, J.Hessel, S.乔杜里，Y。崔，X.L.李，A。Suhr，“不常识性的推理: 关于不常见情况的诱因推理”，arXiv:2311.08469 [cs.CL] (2023)。
https://doi.org/10.48550/arXiv.2311.08469。

54. J.刘，W。王，D。王，N。史密斯，Y。Choi，H. Hajishirzi，“Vera: 常识性陈述的通用合理性估计模型”，*《2023自然语言处理经验方法会议论文集》* (计算语言学协会，2023) pp。1264-1287。https://doi.org/10.18653/v1/2023.emnlp-main.81。

55. M. Mitchell，AI对理解世界的挑战。*科学***382** ，eadm8175 (2023)。
https://doi.org/10.1126/科学.adm8175.

56. N.Dziri, X.陆，M。斯克拉，X。L.李，L。江，B。Y。林，S.韦莱克，P。西，C。Bhagavatula, R.L.胸罩，J.D.Hwang, S.Sanyal, X。任，A。Ettinger, Z.Harchaoui, Y.Choi，“信仰与命运: 变形金刚对组合性的限制”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (Curran Associates，2023)。
https://openreview.net/forum？id= Fkckkr3ya8。

57. D.哈拉维，F。张，C。Y ueh-han，J。斯坦哈特，用语言模型接近人类水平的预测，arXiv:2402.18563 [cs.LG] (2024)。https://doi.org/10.48550/arXiv.2402.18563。

58. 美国。安瓦尔，A.萨帕罗夫，J.兰多，D。Paleka, M.特平，P。Hase, E.S。Lubana, E.詹纳，S.卡斯珀，O。Sourbut, B.L.爱德曼，Z。张，M。Günther, A.Korinek, J.埃尔南德斯-奥拉洛，L.哈蒙德，E。比奇洛，A.潘，L.兰戈斯科。。。D. Krueger，确保大型语言模型的一致性和安全性的基本挑战，arXiv:2404.09932 [cs.LG] (2024)。
https://doi.org/10.48550/arXiv.2404.09932。

59. J.Andreas，“作为代理模型的语言模型”，*研究结果为计算语言学: EMNLP 2022* (计算语言学协会，2022) pp。5769-5779。https://doi.org/10.18653/v1/2022。调查结果-emnlp.423。

60. J.S、帕克，J。O'Brien，C。J.蔡，M.R。莫里斯，P。梁，M。S。伯恩斯坦，"生成代理: 人类行为的交互式模拟"，*载于第36届年度ACM用户界面软件和技术研讨会 (UIST '23)* 论文集 (计算机协会，2023) pp。1-22。https://doi.org/10.1145/3586183.3606763。

61. J.王，Z。吴，Y。李，H.江，P。舒，E。施，胡浩，C。Ma, Y.刘，X。王，Y。姚，X。刘，赵，Z。刘，戴，L。赵，B.Ge，X。李，T。刘，。。。S. Zhang，Lar机器人技术的ge语言模型: 机遇、挑战和前景，arXiv:2401.04334 [cs.RO] (2024)。http://arxiv.org/abs/ 2401.04334。

62. D.C.Cier ş an，美国。Meier, L.M。甘巴德拉，J Schmidhuber，用于手写数字识别的深，大，简单的神经网络。*神经计算机*。**22**，3207-3220 (2010)。https://doi.org/10.1162/NECO_a_00052。

63. T.米科洛夫，M。卡拉菲亚特，L.Burget, J.Černocký, S.Khudanpur，"基于递归神经网络的语言模型" *Proc.语际2010* (ISCA，2010) pp。1045-1048。https://doi.org/10.21437/插话 。2010-343。

64. X。Glorot，Y。Bengio，"理解训练深度前馈神经网络的难度"，*载于第13届人工智能与统计国际会议论文集 (AISTATS 2010)* (PMLR，2010) 卷。9，pp。249-256。https:// 会议记录.mlr.press/v9/glorot10a.html.

65. 机器学习中的时代、参数、计算和数据趋势 (2024)。https://epochai.org/data/epochdb/visualization。

66. * 拐点AI，拐点-2 (2023)。https:// inflection.ai/inflection-2.

67. D.科伊尔，L。汉普顿，21世纪计算机的进步。*电信*。**48***政策*，102649 (2024)。https://doi.org/10.1016/j.Telpol.2023.102649。

68. M.霍布哈恩，L.海姆，G.Aydos，"机器学习硬件的趋势" (EPOCH AI，2023)；https://epochai.org/blog/ 趋势-机器学习-硬件。

69. G.李，Z。孙Q.王，S。王，K。黄，N。赵，Y。Di, X.赵Z.中国绿色数据中心的发展: 政策和碳减排技术路径。*环境。Res。116248* **231** (2023)。https://doi.org/10.1016/j.Envres.2023.116248。

70. * 人类学，研究 (2023)。https://www.anthropic.com/research。

71. * G。布罗克曼，我.Sutskever, S.奥特曼、OpenAI和微软 (2016)。https://openai.com/blog/ openai-和-microsoft。

72. * 与Google Cloud (2023) 合作。https://www.anthropic.com/news/ 的人类-伙伴-使用-google-cloud。

73. * 亚马逊员工，亚马逊和Anthropic宣布战略合作，以推进生成AI (2023)。https://www.aboutamazon.com/news/ 公司-新闻/亚马逊-aws-anthropic-ai

74. * Cohere团队，Cohere可在Google Cloud Marketplace (2022) 上使用。https://cohere.com/blog/ 凝聚-是-在google-cloud-marketplace上可用。

75. * E.博伊德、微软和Mistral AI宣布新的合作伙伴关系，以加速AI创新，并在Azure (2024) 上引入Mistral Large first。https://azure.microsoft.com/en-us/博客/microsoft-and-mistral-ai-宣布-新-合作伙伴关系-加速-ai-创新-引入-mistral-大型-首创-on-azure/。

76. C.-J。吴，R。拉格哈文德拉，美国。古普塔，B.阿恩，N。Ardalani, K.Maeng, G.张，F。阿加，J.黄，C。白，M。Gschwind, A.古普塔，M.奥特，A。梅尔尼科夫，S.Candido, D.布鲁克斯，G。Chauhan, B.李，H.-H。李，。。。K. Hazelwood，"可持续人工智能: 环境影响、挑战和机遇"，*载于第五届机器学习和系统大会 (MLSys)* 论文集 (2022) 卷。4，pp。795-813。https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf

77. A.Gardizy, W.Ma，微软准备AI芯片，因为机器学习成本激增 (2023)。https://www.theinformation.com/articles/ 微软-准备就绪-人工智能-芯片作为机器学习-成本激增。

78. D.帕特森，J。冈萨雷斯，美国。Hölzle, Q.乐，C。梁，L.-M。蒙吉亚，D。罗斯柴尔德，D.R。所以，M.特克西尔，J.Dean认为，机器学习培训的碳足迹将趋于平稳，然后缩小。*计算机***55**，18-28 (2022)。https://doi.org/10.1109/mc.2022.3148714。

79. * E.Almazrouei, H. Alobeidli, A.Alshamsi, A.卡普利，R.Cojocaru, M.Debbah, É.Goffinet, D.赫斯洛，J.Launay, Q.Malartic, D.Mazzotta, B.诺恩，B.Pannier, G.Penedo，开放语言模型的猎鹰系列，arXiv:2311.16867 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2311.16867。

80. * T。魏，L。赵，L.张，B。朱，L。王，H.杨，B。李，C。程，W。吕，R.胡，C。李，L。杨，X。罗，X。吴，L。刘，W。程，P。程，J。张X。张，。。。周勇，Skywork: 一个更开放的双语基础模型，arXiv:2310.19341 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2310.19341。

81. N.Muennighoff, A.Rush, B.巴拉克，T. Le Scao，N.Tazi, A.Piktus, S.Pyysalo, T.狼，C。A.Raffel，"扩展数据约束语言模型"，*《神经信息处理系统进展》36 (NeurIPS 2023) 主会议曲目* (2023) 第一卷。36页。50358-50376。https:// 会议录.neurips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html。

82. P。维拉洛博斯，J。塞维利亚，L。海姆，T.贝西罗格鲁，M.霍布哈恩，A.我们会用完了数据吗？机器学习中缩放数据集的限制分析，arXiv:2211.04325 [cs.LG] (2022)。http://arxiv.org/abs/ 2211.04325。

83. M.马里恩，A.Üstün, L.Pozzobon, A.王，M。Fadaee, S.胡克，“当少就是多: 调查数据修剪预训练llm规模”，在*第一研讨会上，在规模* (2023)。
https://openreview.net/forum？id= XUIYn3jo5T。

84. G.佩内多，Q。Malartic, D.赫斯洛，R.Cojocaru, H. Alobeidli, A.卡普利，B.Pannier, E.Almazrouei, J.Launay，“Falcon LLM的RefinedWeb数据集: 仅使用Web数据优于精选语料库”，在*第37届神经信息处理系统会议 (NeurIPS 2023) 数据集和基准跟踪*中 (2023)。
https://openreview.net/pdf？id= kM5eGcdCzq。

85. S.M、谢，H. Pham，X。东，N。杜，刘海，Y。卢，P。梁，Q。V。Le, T.妈，A.W。Yu，“DoReMi: 优化数据混合加速语言模型预训练”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 中 (2023)。的 https://openreview.net/forum？id=lXuByUeHhd 。

86. * M。米切尔，A.S。Luccioni, N.兰伯特，M.Gerchick, A.麦克米兰-梅杰，E.Ozoani, N.Rajani, T.画眉，Y.杰尔尼特，D。Kiela，测量数据，arXiv:2212.05129 [cs.AI] (2022)。: https://doi.org/10.48550/arXiv.2212.05129 。

87. * D.M。齐格勒，N。斯蒂农，J.吴，T。B.兄弟wn, A.Radford, D.Amodei, P.克里斯蒂亚诺，G。Irving，“从人类偏好微调语言模型” (OpenAI，2020)；http://arxiv.org/abs/ 1909.08593.

88. * D。埃尔南德斯，T。B.布朗，测量神经网络的算法效率，arXiv:2005.04305 [cs.LG] (2020)。
https://doi.org/10.48550/arXiv.2005.04305。

89. A.浩，T。贝西罗格鲁，E.Erdil, D.欧文，R. Rahman, Z.C.郭丁。阿特金森，N。汤普森，J。塞维利亚，“语言模型中的算法进展” (Epoch，2024)；http://arxiv.org/abs/ 2403.05812。

90. F.E. Dorner，测量深度强化学习样本效率的进展，arXiv:2102.04881 [cs.LG] (2021)。
https://doi.org/10.48550/arXiv.2102.04881。

91. * S.陈，S。Wong, L.陈，Y。Tian，通过位置插值扩展大型语言模型的上下文窗口，arXiv:2306.15595 [cs.CL] (2023)。http://arxiv.org/abs/ 2306.15595。

92. * 人类，长期背景提示克劳德·2.1 (2023)。https://www.anthropic.com/news/ claude-2-1-提示。

93. A.顾，T。Dao，Mamba: 具有选择性状态空间的线性时间序列建模，arXiv:2312.00752 [cs.LG] (2023)。
https://doi.org/10.48550/arXiv.2312.00752。

94. D.Kiela, M.巴托洛，Y。聂，D.考希克，A。盖格，Z。吴，B。Vidgen, G.普拉萨德，A.辛格，P。Ringshia, Z.Ma, T.画眉，S.里德尔，Z。Waseem, P.斯登托普，R。贾，M。班萨尔，C。波茨，A.威廉姆斯，“Dynabench: 重新思考NLP中的基准测试”，*载于2021会议北美分会的计算语言学: 人类语言技术*协会 (计算语言学协会，2021) pp。4110-4124。
https://doi.org/10.18653/v1/2021.naacl-main.324。

95. D.亨德里克斯，C。伯恩斯，S.Basart, A.邹，M。Mazeika, D.宋，J。斯坦哈特，“测量大规模多任务语言理解”*在第九届国际会议上学习表征 (ICLR 2021)* 中 (2021)。
https://openreview.net/forum？id= d7KBjmI3GmQ。

96. A.斯里瓦斯塔瓦，A.拉斯托吉，A.饶，A。A.M。Shoeb, A.阿比德，A。Fisch, A.R。布朗，A。Santoro, A.古普塔，A.加里加-阿隆索，A.Kluska, A.Lewkowycz, A.阿加沃尔，A。权力，A。雷，A.Warstadt, A.W。Kocurek, A.萨法亚，A。塔扎夫。。。Z. Wu，超越模仿G ame: 量化和外推语言模型的能力。*机器学习研究*学报 (2023)。https://openreview.net/forum？id= uyTL5Bvosj。

97. D.雷恩，B。L.侯，A。C.斯蒂克兰，J。佩蒂，R.Y。庞，J. 迪拉尼，J.迈克尔，S.R。鲍曼，GPQA: 一个研究生水平的谷歌证明问答基准，arXiv:2311.12022 [cs.AI] (2023)。http://arxiv.org/abs/ 2311.12022。

98. D.亨德里克斯，C。伯恩斯，S.Kadavath, A.阿罗拉，S.Basart, E.唐，D。宋，J。Steinhardt，“用数学数据集测量数学问题解决”在*第35届会议上神经信息处理系统 (NeurIPS 2021) 数据集和基准轨道（第2轮）*上 (2021)。
https://openreview.net/forum？id= 7Bywt2mQsCe。

99. * S.布贝克，V。Chandrasekaran, R.埃尔丹，J.Gehrke, E.霍维茨，E。卡马尔，P。李，Y。T.李，Y。李，S。伦德伯格，H.诺里，H.帕兰吉，M。T.里贝罗，Y.张，人工通用智能的火花: GPT-4的早期实验，arXiv:2303.12712 [cs.CL] (2023)。
https://doi.org/10.48550/arXiv.2303.12712。

100. A.周，K。王，Z。卢，W。施，S.罗，Z。秦，S。陆，A。贾，L。宋，M.詹，H。Li，“使用GPT-4代码解释器和基于代码的自我验证解决具有挑战性的数学单词问题”，*在第12届国际学习表征会议 (ICLR 2024)* 中 (2023)。
https://openreview.net/forum？id= c8McWs4Av0。

101. T.R.麦金托什，T.Susnjak, T.刘，P。沃特斯，M。N.Halgamuge，生成式人工智能时代大型语言模型基准的不足之处，arXiv:2402.09880 [cs.AI] (2024)。https://doi.org/10.48550/arXiv.2402.09880 。

102. M.米切尔，A.B.Palmarini, A.K。Moskvichev，“在抽象和推理任务上比较人类，GPT-4和GPT-4V”在*AAAI 2024研讨会“大型语言模型只是因果鹦鹉吗？”*中 (2023)。
https://openreview.net/forum？id= 3rGT5OkzpC。

103. S。斯里瓦斯塔瓦，M.B.Annarose，P。V。Anto, S.梅农，A.Sukumar, S.T.阿德怀斯，A。Philipose, S.王子，S.托马斯，推理性能稳健评估的功能基准，以及推理差距，arXiv:2402.19450 [cs.AI] (2024)。https://doi.org/10.48550/arXiv.2402.19450。

104. C.邓勇。赵，X。唐，M。格斯坦，A.Cohan，调查大型语言模型的现代基准中的数据污染，arXiv:2311.09783 [cs.CL] (2023)。http://arxiv.org/abs/ 2311.09783。

105. O.塞恩斯，J.坎波斯，我.加西亚-费雷罗，J.Etxaniz, O.L. de Lacalle, E.Agirre，“麻烦中的NLP评估: 关于需要测量每个基准的LLM数据污染”，*研究结果为计算语言学协会: EMNLP 2023* (计算语言学协会，2023) pp。10776-10787。https://doi.org/10.18653/v1/2023.findings-emnlp.722。

106. Y。曹，L。周，S。李，L.卡贝洛，M.陈，D。Hershcovich，“评估ChatGPT与人类社会之间的跨文化一致性: 一项实证研究”，*载于第一届NLP跨文化问题研讨会 (C3NLP)* 论文集 (计算语言学协会，2023) pp。53-67。https://doi.org/10.18653/v1/2023.c3nlp-1.7。

107. L.Berglund, M.童，M。考夫曼，M。Balesni, A.C.Stickland, T.Korbak, O.埃文斯，“逆转诅咒: 在“ A是B ”上接受培训的LLMs在*第12届国际学习表征会议 (ICLR 2024)* 中未能学习“ B是A ”(2023)。https://openreview.net/forum？id= GPKTIktA0k。

108. J.Geiping, A.斯坦，M.舒，K。Saifullah, Y.温，T。Goldstein，“强迫llm做并揭示 (几乎) 任何事情” *ICLR 2024研讨会安全可信的大型语言模型 (SET LLM)* 2024 https://openreview.net/forum？id= Y5inHAjMu0。

109. * J。卡普兰，S.McCandlish, T.Henighan, T.B.布朗，B。国际象棋，R.孩子，S。格雷，A.雷德福，J.吴，D。Amodei，神经语言模型的缩放定律，arXiv:2001.08361 [cs.LG] (2020)。https://doi.org/10.48550/arXiv.2001.08361 。

110. * J。霍夫曼，S.Borgeaud, A.门施，E。Buchatskaya, T.蔡，E.卢瑟福，D.德拉斯卡萨斯，L.A.亨德里克斯，J。Welbl, A.克拉克，T.亨尼根，E.诺兰，K。米利肯，G.范登德里舍，B。达莫克，A。盖伊，S.奥辛德罗，K。Simonyan, E.埃尔森，。。。L. Sifre，训练计算最优大型语言模型，arXiv:2203.15556 [cs.CL] (2022)。http://arxiv.org/abs/ 2203.15556。

111.* T。Henighan, J.卡普兰，M.卡茨，M。陈，C。Hesse, J.杰克逊，H.6月，T.B.布朗，P。Dhariwal, S.格雷，C.哈拉西，B.曼恩，A.Radford, A.Ramesh, N.莱德，D。M。齐格勒，J. 舒尔曼，D.Amodei, S.McCandlish，自回归生成建模的缩放定律，arXiv:2010.14701 [cs.LG] (2020)。http://arxiv.org/abs/ 2010.14701。

112. X。翟，A。Kolesnikov, N.Houlsby, L.Beyer，“缩放视觉转换器”，*2022 IEEE/CVF计算机视觉和模式识别 (CVPR)* 会议 (2022) pp。1204-1213。https://doi.org/10.1109/cvpr52688.2022.01179。

113. * A.L.琼斯，用棋盘游戏缩放比例定律，arXiv:2104.03113 [cs.LG] (2021)。https://doi.org/10.48550/arXiv.2104.03113。

114. S。戈亚尔，P。Maini, Z.C.立顿，A.拉格胡纳森，J。Z。Kolter，“数据过滤科学: 数据管理无法计算 ”，在*ICLR 2024研讨会上关于导航和解决基础模型 (DPFM)* 的数据问题 (2024) 中。https://openreview.net/forum？id= 9wzo4EjEGM。

115. * Y。Bahri, E.戴尔，J.卡普兰，J.李，U。夏尔马，解释神经缩放定律，arXiv:2102.06701 [cs.LG] (2021)。https://doi.org/10.48550/arXiv.2102.06701。

116. * A.马洛尼，D。A.罗伯茨，J。Sully，神经标度律的可解模型，arXiv:2210.16859 [cs.LG] (2022)。http://arxiv.org/abs/ 2210.16859。

117. 美国。夏尔马，J.卡普兰，来自数据流形维度的缩放定律。*J.马赫学习。Res。* **23** ，343-376 (2022)。https://dl.acm.org/doi/abs/ 10.5555/3586589.3586598。

118. Ł。D ę bowski，神经缩放定律的简化模型: 多周期圣达菲过程，arXiv:2302.09049 [cs.IT] (2023)。http://arxiv.org/abs/ 2302.09049。

119. E。J.米肖，Z.刘，U。吉里特，M.Tegmark，*第37届神经信息处理系统会议 (NeurIPS 2023)* 中的“神经缩放的量化模型” (2023)。https://openreview.net/forum？id= 3tbTw2ga8K。

120. S。比德曼，美国。S。Prashanth, L.Sutawika, H. Schoelkopf, Q.G.安东尼，S.Purohit, E.Raff，*第37届神经信息处理系统会议 (NeurIPS 2023)* 中的“大型语言模型中的紧急和可预测记忆” (2023)。https://openreview.net/forum？id= Iq0DvhB4Kf。

121. D.甘古丽，D.埃尔南德斯，L。洛维特，A。Askell，Y。白，A。陈，T。康纳利，N。Dassarma, D.排水管，N.Elhage, S. El Showk, S。Fort, Z.Hatfield-Dodds, T.Henighan, S.约翰斯顿，A。琼斯，N。约瑟夫，J。Kernian, S.克拉维克。。。J. Clark，“大型生成模型中的可预测性和惊喜”，*载于2022 ACM公平，问责制和透明度*会议论文集 (计算机协会，2022) pp。1747-1764。https://doi.org/10.1145/3531146.3533229。

122. * Z。杜，A。曾，Y。董，杰。唐，从损失的角度理解语言模型的涌现能力，arXiv:2403.15796 [cs.CL] (2024)。http://arxiv.org/abs/ 2403.15796。

123. J.魏，Y。泰，R。Bommasani，C。拉菲尔，B.Zoph, S.Borgeaud, D.Yogatama, M.波斯马，D。周，D。梅茨勒，E。H. Chi, T.桥本，O.Vinyals，P。梁，J。迪恩，W.Fedus，大型语言模型的紧急能力。*机器学习研究*学报 (2022)。https://openreview.net/forum？id= yzkSU5zdwD。

124. R。谢弗，B。米兰达，S.Koyejo，“大型语言模型的新兴能力是海市蜃楼吗？”在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id= ITw9edRDlD。

125. I.R.麦肯齐，A.Lyzhov, M.M。皮勒，A。帕里什，A.穆勒，A。Prabhu, E.麦克莱恩，X。沈，J。卡瓦纳，A。G.Gritsevskiy, D.考夫曼，A。T.Kirtland, Z.周，Y。张，S。黄，D。Wurgaft，M。魏斯，A.罗斯，G.Recchia, .。。E. Perez，逆缩放: 越大越好。*机器学习研究*学报 (2023)。

126. J.魏，N。金，Y。泰，Q。Le，“逆缩放可以变成U形”，*《2023自然语言处理经验方法会议论文集》(EMNLP 2023)* (计算语言学协会，2023) pp。15580-15591。https://doi.org/10.18653/v1/2023.emnlp-main.963。

127. Y。Bengio, G.Hinton, A.姚，D.宋，P。Abbeel, Y.N.Harari Y.-Q.张，L。薛，S.Shalev-Shwartz, G.哈德菲尔德，J。克伦，T。Maharaj, F.哈特，A.G.贝丁，S.麦克莱思，Q。高，A。阿查· 瑞亚，D。克鲁格，A。德拉甘.。。S. Mindermann，“在快速进步的时代管理人工智能风险”，arXiv:2310.17688 [cs.CY] (2023)。http://arxiv.org/abs/ 2310.17688。

128. J.珍珠，D.Mackenzie，“*The book of why: the new science of cause and effect*”(企鹅科学，企鹅图书，哈洛，英国，2019)，pp。418。

129. M. Mitchell，“为什么AI比我们想象的要难”，*《遗传与进化计算会议论文集》(GECCO '21)*，(计算机协会，2021) pp。3.https://doi.org/10.1145/3449639.3465421。

130. Y。LeCun，深度学习的力量和局限性: Yann LeCun在他的IRI奖章演讲中描绘了机器学习技术的发展，并提出了未来可能会发生的事情。*Res。技术。管理。***61**，22-27 (2018)。https://doi.org/10.1080/08956308.2018.1516928。

131. * 字母表，“年度报告2022”(字母，2023)；https:// abc.xyz/assets/d4/4f/a48b94d548d0b2fdc029a95e8c63/2022-alphabet-annual-report.pdf.

132. * L。范，K。陈，D。克里希南，D.Katabi, P.伊索拉，Y。田，用于模型训练的合成图像的缩放定律...目前，arXiv:2312.04567 [cs.CV] (2023)。https://doi.org/10.48550/arXiv.2312.04567。

133. S。傅，N。Y。塔米尔，S.孙达拉姆，L.柴，R。张，T。Dekel, P.Isola，“DreamSim: 使用合成数据学习人类视觉相似性的新维度”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id= DEiNSfh1k7。

134. Y。田，L。范，P。伊索拉，H.张，D.克里希南，“稳定: 从文本到图像模型的合成图像使强大的视觉表示学习者”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id=xpjsOQtKqx。

135. S。Alemohammad, J.卡斯科-罗德里格斯，L.Luzi, A.I.Humayun, H. Babaei, D.勒琼，A。Siahkoohi，R。Baraniuk，“自我生成模型发疯”*在第12届国际学习表征会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= ShjMHfmPs0。

136. I.Shumailov, Z.Shumaylov, Y.赵，Y。Gal, N.纸不是，R。《递归的诅咒: 对生成的数据进行训练使模型忘记》，arXiv:2305.17493 [cs.LG] (2023)。http://arxiv.org/abs/ 2305.17493。

137. H。Abdine, M.Chatzianastasis, C.Bouyioukos, M.Vazirgiannis，“Prot2Text: gnn和变压器的多模式蛋白质的功能生成”，在*第37篇关于神经信息处理系统 (NeurIPS 2023) 深度生成模型的健康研讨会*中 (2023)。https://openreview.net/forum？id= EJ7YNgWYFj。

138. * 无缝通信，L.巴罗，Y.-A。钟，M。C.梅格里奥利，D。戴尔，N.董，P.-A。Duquenne, H. Elsahar, H. Gong, K.赫弗南，J.霍夫曼，C。克莱伯，P。李，D。利希特，J.梅拉德，A。Rakotoarison, K.R。Sadagopan, G.温泽克，E。是的，。。。S. Wang，“无缝m4t: 大规模多语言和多模式机器翻译”(Meta AI，2023)；http://arxiv.org/abs/ 2308.11596。

139. 国际能源署，“电力2024: 分析和预测2026年”(IEA，2024)；https://iea.blob.core.windows.net/assets/ 6b2fd954-2017-408e-bf08-952fdd62118a/电力2024-分析和预测2026.pdf。

140. H。Ritchie, P.罗萨多，M.罗泽，能源。《*我们的数据*世界》(2024)https://ourworldindata.org/energy。141。* Talen能源公司宣布出售零碳数据中心园区 (2024)。https://talenenergy.investorroom.com/2024-03-04-Talen-Energy-Announces-Sale-of-Zero-Carbon-Data-Center-校园。

142. E。格里菲斯，绝望地寻找A.I.Boom是*《纽约时报》*最不可或缺的奖项。(2023)。https://www.nytimes.com/2023/08/16/技术/ai-gpu-chips-shortage.html。

143. D.布拉格，N。卡塞利，J。A.Hochgesang，M。Huenerfauth, L.Katz-Hernandez, O.科勒，R。Kushalnagar, C.Vogler，R。E。Ladner，手语人工智能数据集的发展前景: 跨学科视角。*ACM Trans。访问。计算机。***14**，1-45 (2021)。https://doi.org/10.1145/3436996。

144. 高级电子实践，H. Bauer，O.Burkacky, P.Kenevan, S.林格曼，K。波托兹基，B.怀斯曼，“半导体设计和制造: 实现领先能力”(麦肯锡公司，2020)；https://www.mckinsey.com/industries/ 工业和电子/我们的见解/半导体设计和制造-实现-领先-能力 #/。

145. J.VerWey，“没有许可证，没有晶圆厂: Impo半导体制造业监管改革的挑战”(安全和新兴技术中心，2021)；https://doi.org/10.51593/20210053。

146. G.蒂普，J.布德，F。鲍曼，K。K。Bourdelle, T.布恩，J.Garno, A.Ghetti, M.格林，H.戈斯曼，Y.Kim, R.克莱曼，A.Kornblit, F.克莱门斯，S.Moccio，D。穆勒，J.Rosamilia, P.西尔弗曼，T。Sorsch, W.蒂普，。。。B. Weir，MOSFET栅极氧化物厚度到零的无情行进。*微电子。Reliab.***40**，557-562 (2000)。https://doi.org/10.1016/s0026-2714(99)00257-7。

147. M.-L。陈，X。孙，刘海华，王海华，Q。朱，S。王，H。杜，B。董，J。张，Y。孙，S.邱，T。Alava, S.刘，D.-M。太阳，Z。Han，具有一个原子层沟道的FinFET。*Nat。Commun。1205* **11** (2020)。https://doi.org/10.1038/s41467-020-15096-0。

148. A.Ho, E.Erdil, T.Besiroglu，“限制能源CMOS微处理器的效率”，*2023 IEEE国际会议重启计算 (ICRC)* (2023) pp。1-10。https://doi.org/10.1109/icrc60800.2023.10386559。

149. A.周，K。Yan, M.Shlapentokh-rothman，H. Wang，Y.-X。Wang，“语言代理树搜索统一了语言模型中的推理和计划”，在*ICLR 2024研讨会上关于大型语言模型 (LLM) 代理* (2024)。https://openreview.net/forum？id= 2z5dzaqOLp。

150. S。张，Z。陈，Y。沈，M.丁，J。B.Tenenbaum，C。Gan，“使用大型语言模型进行代码生成规划”，*在第11届学习表示国际会议 (ICLR 2023)* 中 (2022)。https://openreview.net/forum？id= Lr8cOOtYbfL。

151. S。波卢，J.M。韩，K。郑，M。Baksys，我.Babuschkin，我.Sutskerver，“正式数学陈述课程学习”*在第11届国际学习表征会议 (ICLR 2023)* 中 (2022)。https://openreview.net/forum？id=- P7G-8dmSh4。

152. A.Fawzi, M.巴洛格，A.黄，T。休伯特，B.Romera-Paredes, M.Barekatain, A.诺维科夫，F。J. R.鲁伊斯，J.Schrittwieser, G.Swirszcz, D.银，D。哈萨比斯，P。Kohli，通过强化学习发现更快的矩阵乘法算法。*自然***610**，47-53 (2022)。https://doi.org/10.1038/s41586-022-05172-4。

153. A.Haj-Ali, N.K。艾哈迈德，T。Willke, Y.S。邵，K。Asanovic，我。Stoica，“神经向量化器: 具有深度强化学习的端到端向量化”，*载于第18届ACM/IEEE国际研讨会关于代码生成和优化 (CGO 2020)* 的论文集 (计算机协会，2020) pp。242-255。https://doi.org/10.1145/3368826.3377928。

154. R。Pryzant, D.Iter, J.李，Y。李，C。朱，M。Zeng，“使用“ 梯度下降 ”和光束搜索进行自动提示优化”，*《2023自然语言处理经验方法会议论文集》(EMNLP 2023)* (计算语言学协会，2023) pp。7957-7968。https://doi.org/10.18653/v1/2023.emnlp-main.494。

155. S。张，C。龚，L。吴，X。刘，M。Zhou，AutoML-GPT: 使用GPT进行自动机器学习，arXiv:2305.02499 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2305.02499。

156. S。刘，Z。林，S.于，R。李，T。凌，D。帕塔克，D.Ramanan，作为视觉语言模型的黑盒优化器的语言模型，arXiv:2309.05950 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2309.05950。

157. X。李，P。于，C。周，T。Schick, O.利维，L.Zettlemoyer, J.E。韦斯顿，M。刘易斯，*在第12届国际学习表征会议 (ICLR 2024)* 中 “与指令反向翻译的自我对齐” (2023)。https://openreview.net/forum？id=1oijHJBRsT.

158. * Y。白，S。Kadavath, S.昆都，A.Askell，J。Kernion, A.琼斯，A。陈，A。戈尔迪，A.Mirhoseini，C。麦金农，C.陈，C.奥尔森，C。奥拉，D.埃尔南德斯，D。排水管，D.甘古丽，D.李，E。约翰逊，E。佩雷斯。。。J·卡普兰，“宪法人工智能: 人工智能反馈的无害性”，arXiv:2212.08073 [cs.CL] (2022)。https://doi.org/10.48550/arXiv.2212.08073。

159. * N.萨克德法，B.科尔曼，W.-C。康，J。Ni, L.洪，E。H. Chi, J.卡弗利，J。麦考利，D。Z。程，如何训练数据高效的LLMs，arXiv:2402.09668 [cs.LG] (2024)。的 https://doi.org/10.48550/arXiv.2402.09668。

160. Y。张，X。王，J。王，Y。吴，L。杨，K。朱，H.陈，X。易，C。王，Y。王，W。叶，Y。张，Y。张，P。S。于，问:杨，X。谢，关于大型语言模型评估的调查。*ACM Trans。Intell。系统。技术。***15**，39:1-39:45 (2024)。https://doi.org/10.1145/3641289。

161. G.马库斯，E.戴维斯，S.Aaronson，DALL-E 2的初步分析，arXiv:2204.13807 [cs.CV] (2022)。https://doi.org/10.48550/arXiv.2204.13807。

162. A.博尔吉，GAN评价措施的利弊: 新进展。*计算机。Vis。图像Underst。***215** (2022)。https://doi.org/10.1016/j.Cviu.2021.103329。

163. S。周，M。戈登，R。克里希纳，A.Narcomey, L.F.飞飞，M.伯恩斯坦，“炒作: 生成模型的人眼感知评估基准”，*《神经信息处理系统进展》(NeurIPS 2019)* (Curran Associates，inc.，2019) 第一卷。32。https:// 会议录.neurips.cc/paper_files/paper/2019/hash/65699726a3c601b9f31bf04019c8593c-Abstract.html。

164. L.郑，W.-L。蒋Y。盛，S。庄，Z。吴，Y。庄，Z。林，Z。李，D。李，E。邢，H。张，J。E。冈萨雷斯，我.Stoica，“与mt-bench和Chatbot Arena一起评判llm-as-a-judge”在*第37届神经信息处理系统 (NeurIPS 2023) 数据集和基准轨道*会议上 (2023)。https://openreview.net/forum？id=uccHPGDlao。

165. P。梁，R。Bommasani, T.李，D。齐普拉斯，D.Soylu, M.Yasunaga，Y。张，D。纳拉亚南，Y.吴，A。库马尔，B。纽曼，B.元，B。严，C。张，C。A.科斯格罗夫，C。D.曼宁，C。Re, D.阿科斯塔-纳瓦斯，D。A.哈德森。。。Y. Koreeda，语言模型的整体评估。*机器学习研究*学报 (2023)。https://openreview.net/forum？id= iO4LZibEqW。

166. * T。Patwardhan，K。刘，T。马尔可夫，N.乔杜里，D。利特，N.圆锥体，C.Maltbie, J.Huizinga，C。温赖特，S. f.杰克逊，S。Adler, R.卡萨布兰德，A.Madry，“为LLM辅助的生物威胁创建构建预警系统”(OpenAI，2024)；https://openai.com/research/ 建设-预警系统-为llm辅助-生物-威胁-创造。

167. I.D. Raji, E.丹顿，E.M。Bender, A.汉娜，A.Paullada，*第35届神经信息处理系统会议 (NeurIPS 2021) 数据集和基准轨道 (第2轮)* 中的“人工智能和全世界的一切基准”(2021)。https://openreview.net/forum？id= j6NxpQbREA1。

168. Y。Lecun, L.Bottou, Y.Bengio，P。哈夫·纳，基于梯度的学习应用于文档识别。*Proc.IEEE Inst.电子。电子。Eng。* **86**，2278-2324 (1998)。https://doi.org/10.1109/5.726791。

169. A.Krizhevsky，“从微小图像中学习多层特征”(多伦多大学，2009)；https://www.cs.toronto.edu/~kriz/ learning-features-2009-TR.pdf。

170. Z。刘，P。罗，X。王，X。*2015 IEEE国际计算机视觉会议 (ICCV)* (IEEE计算机协会，2015) pp.Tang，“深度学习野外面部属性”。3730-3738。https://doi.org/10.1109/iccv.2015.425。

171. O.Russakovsky, J.邓海苏，J。克劳斯，S.Satheesh, S.Ma, Z.黄，A。Karpathy, A.Khosla, M.伯恩斯坦，A.C. L.伯格。飞飞，ImageNet大规模视觉识别挑战赛。*Int。J、计算机。Vis*。**115**，211-252 (2015)。https://doi.org/10.1007/s11263-015-0816-y。

172. T.-Y。林，M.梅尔，S.贝隆吉，J.海斯，P。佩罗纳，D.Ramanan，P。Dollár, C.L.Zitnick，*计算机视觉-ECCV 2014* 中的“Microsoft COCO: 上下文中的公共对象”(Springer国际出版，2014) pp。740-755。https://doi.org/10.1007/978-3-319-10602-1_48。

173. VQA联盟，视觉问答 (2015)。http://visualqa.org。

174. D.A.哈德逊，C.D.Manning，“GQA: 用于现实世界视觉推理和组合问题回答的新数据集”，*2019 IEEE/CVF计算机视觉和模式识别 (CVPR)* 会议 (2019) pp。6693-6702。https://doi.org/10.1109/cvpr.2019.00686

175. A.王，A。辛格，J。迈克尔，F。希尔，O.Levy, S.Bowman，“胶水: 自然语言理解的多任务基准和分析平台”，*载于2018 EMNLP研讨会BlackboxNLP: 分析和解释NLP*的神经网络 (计算语言学协会，2018) pp。353-355。https://doi.org/10.18653/v1/W18-5446。

176. A.王，Y。Pruksachatkun, N.南吉亚，A。辛格，J。迈克尔，F。希尔，O.Levy, S.Bowman，“超级胶水: 通用语言理解系统的粘性基准”，*《神经信息处理系统进展》(NeurIPS 2019)* (Curran Associates，inc.，2019) 第一卷。32。https:// 会议录.neurips.cc/paper_files/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html。

177. * W。钟，R。崔，Y.郭勇。梁，S。卢，Y。王，A。赛义德，W. 陈，N。Duan，AGIEval: 评估基础模型的以人为本的基准，arXiv:2304.06364 [cs.CL] (2023)。http://arxiv.org/abs/ 2304.06364。

178. X。刘，于华，张，Y。徐，X。雷，H.赖，Y。顾，H.丁，K。男人K.杨，S。张X。邓，A。曾，Z。杜，C。张，S。沈，T.张，Y。苏，H.孙，。。。J. Tang，“AgentBench: 评估LLMs作为代理”*在第12届国际学习代表会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= zAdUB0aCTQ。

179. Y。砰，S.Cahyawijaya，N。李，W。戴，D。苏，B。威利，H. Lovenia，Z.吉，T.于，W。Chung, Q.V。做，Y。徐，P。冯，“关于推理、幻觉和互动的聊天的多任务、多语言、多模式评估”*第13届自然语言处理国际联席会议和计算语言学协会亚太分会第三届会议论文集 (第1卷: 长论文)*(计算语言学协会，2023) pp。675-718。https://doi.org/10.18653/v1/2023.ijcnlp-main.45。

180. S。林，J。希尔顿，O.埃文斯，“TruthfulQA: 测量模型如何模仿人类的谎言”，*载于第60届计算语言学协会年会 (第1卷: 长篇论文)*，(计算语言学协会，2022) pp。3214-3252。https://doi.org/10.18653/v1/2022.acl-long.229。

181. A.潘，J.S。Chan, A.邹，N。李，S。Basart, T.伍德赛德，H.张，S。埃蒙斯，D。亨德里克斯，“奖励能证明手段是合理的吗？衡量马基雅维利基准中奖励和道德行为之间的权衡“*载于第40届国际会议机器学习 (icml'23)* 论文集 (JMLR.org，2023) 卷。202，第页。26837-26867。

182. N.Mu, S.陈，Z。王，S。陈，D。K aramardian, L.Aljeraisy, B.Alomair，D。亨德里克斯，D。瓦格纳，llm可以遵循简单的规则吗？，arXiv:2311.04235 [cs.AI] (2023)。https://doi.org/10.48550/arXiv.2311.04235。

183. G.Mialon, C.Fourrier, T.狼，Y。LeCun, T.Scialom，“盖亚: 通用人工智能助手的基准”*在第12届国际学习表征会议 (ICLR 2024)* 中 (2024)。https://openreview.net/forum？id= fibxvahvs3。

184. T.廖，R。Taori, D.拉吉，L.施密特: “我们还在学习吗？跨机器学习的评估失败的元审查”，*《神经信息处理系统追踪数据集和基准1 (NeurIPS数据集和基准2021) round2》* (2021) 卷。1。https:// 数据集-基准-会议记录.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html。

185. D.卡，P。亨德森，美国。Khandelwal, R.贾，K。Mahowald，D。Jurafsky，“无能为力，责任重大”，*《2020自然语言处理经验方法会议论文集》 (EMNLP 2020)* (计算语言学协会，2020) pp。9263-9274。https://doi.org/10.18653/v1/2020.emnlp-main.745。

186. C.G、诺斯卡特，A。Athalye, J.Mueller，“测试一下集中的普遍标签错误破坏了机器学习基准的稳定性”，在*第35届神经信息处理系统会议 (NeurIPS 2021) 数据集和基准轨道 (第1轮)* 中 (2021)。https://openreview.net/forum？id=XccDXrDNLek。

187. S。盖曼，S.古鲁兰根，M。Sap, Y.Choi, N.A.Smith，“RealToxicityPrompts: 评估语言模型中的神经毒性变性”，*研究结果为计算语言学协会: EMNLP 2020* (计算语言学协会，2020) pp。3356-3369。https://doi.org/10.18653/v1/2020.findings-emnlp.301。

188. T.Schick, J.Dwivedi-Yu, R.德西，R.Raileanu, M.Lomeli, E.汉布罗，L.Zettlemoyer, N.Cancedda, T.Scialom，“工具成形器: 语言模型可以自学使用工具”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 中 (2023)。https://openreview.net/forum？id=Yacmpz84TH。

189. D.A. Boiko, R.麦克奈特，G.Gomes，大型语言模型的新兴自主科学研究能力，arXiv:2304.05332 [physics.Chem-ph] (2023)。https://doi.org/10.48550/arXiv.2304.05332。

190. A.Paullada，我.D.Raji, E.M。Bender, E.丹顿，A.汉娜，数据及其 (dis) 内容: 机器学习研究中数据集开发和使用的调查。*模式 (N。Y)* 、100336 **2** (2021)。https://doi.org/10.1016/j.Patter.2021.100336。

191. S。R.鲍曼，G.Dahl，“在自然语言理解中修复基准测试将需要什么？”在*《2021会议北美分会》的计算语言学: 人类语言技术 (naacl-hlt 2021)* 中 (计算语言学协会，2021) pp。4843-4855。https://doi.org/10.18653/v1/2021.naacl-main.385。

192. B.哈钦森，N。Rostamzadeh, C.格里尔，K。海勒，V。Prabhakaran，“机器学习实践中的评估差距”，*载于2022 ACM会议关于公平性，问责制和透明度 (FAccT '22)* 的会议记录 (计算机协会，2022) pp。1859-1876。https://doi.org/10.1145/3531146.3533233。

193. S。Golchin, M.Surdeanu，“LLMs中的时间旅行: 大型语言模型中的跟踪数据内容”，*在第12届国际学习表示会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= 2Rwq6c3tvr。

194. Y。杨，A。Tomar，关于大型语言模型的规划、搜索和记忆能力，arXiv:2309.01868 [cs.CL] (2023)。http://arxiv.org/abs/ 2309.01868。

195. * 米。R。莫里斯，J。Sohl-dickstein, N.Fiedel, T.Warkentin, A.达福，A.浮士德，C.Farabet, S.Legg，“AGI的水平: 在实现AGI的道路上实现进展”(Google DeepMind，2024)；http://arxiv.org/abs/ 2311.02462。

196. S。Shankar, Y.Halpern, E.Breck, J.阿特伍德，J。威尔逊，D。Sculley，“没有代表的分类: 评估发展中国家开放数据集中的地理多样性问题”，在*第31届神经信息处理系统会议 (NIPS 2017) 机器学习发展中国家研讨会*上 (2017)。

197. L.阿罗约，C。韦尔蒂，真理是谎言: 人群真理和人类注释的七个神话。*AI Mag.***36** ，15-24 (2015)。https://doi.org/10.1609/aimag.v36i1.2564。

198. M. L.戈登，K。周，K。帕特尔，T。桥本，M.S。伯恩斯坦，“分歧反卷积: 使机器学习性能指标与现实相符”，*《2021 CHI会议关于计算系统中人为因素的论文集》 (CHI '21)* ，(计算机协会，2021) pp。1-14。https://doi.org/10.1145/3411764.3445423。

199. C.费尔斯通，人机比较中的性能与能力。*Proc.Natl.Acad。Sci。美国。A.* 26562-26571 **117** (2020)。https://doi.org/10.1073/pnas.1905334117.

200. L.阿罗约，A。S。泰勒，M。迪亚兹，C。M。Homan, A.帕里什，G。塞拉皮奥-加西亚，V。Prabhakaran, D.Wang，“DICES数据集: 对话AI安全性评估中的多样性”，在*第37届神经信息处理系统会议 (NeurIPS 2023) 数据集和基准轨道*上 (2023)。https://openreview.net/forum？id=GjNvvswoUL。

201. H。P.考利，M.Natter, K.Gray-Roncal, R.E。罗德斯，E。C.约翰逊，N。Drenkow, T.M。希德，F.S。Chance, B.韦斯特，W。Gray-roncal，作者更正: 在人类和机器学习比较研究中对人类表现进行严格评估的框架。*Sci。代表。11559* **12** (2022)。https://doi.org/10.1038/s41598-022-15857-5。

202. T.Shin, Y.Razeghi, R.L.洛根，四世，E.华莱士，S.Singh，“自动提示: 通过自动生成的提示从语言模型中获取知识”，*载于2020自然*经验方法会议论文集。

语言处理 (EMNLP 2020) (计算语言学协会，2020) pp。4222-4235。
https://doi.org/10.18653/v1/2020.emnlp-main.346。

203. E。佩雷斯，S.黄，F。宋，T。蔡，R.戒指，J.阿斯兰尼德，A.Glaese, N.麦卡利斯，G。Irving，“将语言模型与语言模型结合起来”，*《2022自然语言处理经验方法会议论文集》 (EMNLP 2022)* (计算语言学协会，2022) pp。3419-3448。
https://doi.org/10.18653/v1/2022.emnlp-main.225。

204. * D.甘古丽，L.洛维特，J.Kernion, A.Askell，Y。白，S。Kadavath, B.曼恩，E.佩雷斯，N。Schiefer, K.Ndousse, A.琼斯，S.鲍曼，A.陈，T。康纳利，N。DasSarma, D.排水管，N.Elhage, S.El-Showk, S.堡，。。。J. Clark，“减少危害的红色团队语言模型: 方法，缩放行为和经验教训” (Anthropic，2022)；
http://arxiv.org/abs/ 2209.07858。

205. S。卡斯珀，J.林，J。Kwon, G.卡尔普，D.Hadfield-menell，探索，建立，利用: 从头开始Red组合语言模型，arXiv:2306.09442 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2306.09442。

206. S。童，E。琼斯，J。Steinhardt，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上的“具有语言模型的多模式系统的大规模生产故障” (2023)。
https://openreview.net/forum？id= T6iiOqsGOh。

207. D.齐格勒，S.尼克斯，L。陈，T.鲍曼，P。施密特-尼尔森，T.林，A。Scherlis, N.纳比希马，B.温斯坦-劳恩，D.德哈斯，B.Shlegeris, N.托马斯，*《神经信息处理系统进展》 (NeurIPS 2022)* 中的“高风险可靠性的对抗性训练” (2022)。35页。9274-9286。
https:// 会议录.neurips.cc // paper_files/paper/2022/hash/3c44405d619a6920384a45bce876b41e-Abstract-Conference.html。

208. Y。刘，G。邓，Z。徐，Y。李，Y。郑y。张，L。赵，T.张，K。王，Y。刘，通过提示工程越狱聊天: 一项实证研究，arXiv:2305.13860 [cs.SE] (2023)。http://arxiv.org/abs/ 2305.13860。

209. A.魏，N。Haghtalab, J.Steinhardt，“越狱: LLM安全培训如何失败？”在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id= jA235JGM09。

210. A.邹，Z。王，N。Carlini, M.纳斯尔，J.齐科·科尔特，M。Fredrikson，对对齐语言模型的通用和可转移的对抗攻击，arXiv:2307.15043 [cs.CL] (2023)。，https://doi.org/10.48550/arXiv.2307.15043。

211. R。沙阿，Q.F.Montixi, S.倒，A.塔加德，J。Rando，*第37届神经信息处理系统会议 (NeurIPS 2023) 社会责任语言建模研究研讨会 (SoLaR)* 中的“通过角色调制实现语言模型的可扩展和可转移的黑盒越狱” (2023)。
https://openreview.net/forum？id= x3Ltqz1UFg。

212. A.Rao, S.Vashistha, A.奈克，S.阿迪亚，M。Choudhury，*在2024联合国际计算语言学，语言资源和评估会议 (lrec-coling 2024)* 上的“欺骗llm不服从: 形式化，分析和检测越狱” (2024)。https://doi.org/10.48550/arXiv.2305.14965。

213. V。Ojewale，R。斯蒂德，B。Vecchione, A.Birhane，我.D.拉吉，“走向人工智能问责基础设施: 人工智能审计工具中的差距和机会”，arXiv:2402.17861 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2402.17861 。

214. V。Turri, R.Dzombak，“为什么我们需要了解更多: 探索AI事件文档实践的状态”*《2023 AAAI/ACM人工智能、伦理和社会会议论文集》 (AIES '23)* (ACM，2023) pp.576-583。
https://doi.org/10.1145/3600211.3604700。

215. S。克斯坦萨-乔克，我D.Raji, J.Buolamwini，“谁来审计审计师？来自算法审计生态系统的现场扫描的建议”，*载于2022 ACM会议关于公平性，问责制和透明度 (FAccT '22)* 的会议记录 (计算机协会，2022) pp。1571-1583。
https://doi.org/10.1145/3531146.3533213。

216. M.费弗，A.辛哈，Z。C.Lipton，H. Heidari，《生成AI的红队: 银弹还是安全剧院？》，arXiv:2401.15897 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2401.15897。

217. A.Birhane, R.Steed, V.Ojewale，B。Vecchione，我.D.Raji，“SoK: 人工智能审计: 通往人工智能问责之路的破碎巴士”，在*第二届IEEE安全可信机器学习*大会上 (2024)。
https://openreview.net/forum？id= TmagEd33w3。

218. S。卡斯珀，T.Bu，Y。李，J。李，K。张，K。Hariharan, D.Hadfield-menell，“将深度神经网络与特征合成工具结合起来”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。
https://openreview.net/forum？id= Od6CHhPM7I。

219. S。弗里德勒，R。辛格，B。布利利-哈梅林，J.梅特卡夫，B.J.陈，“人工智能红队不是人工智能危害的一站式解决方案: 使用红队进行人工智能问责的建议” (数据与社会，2023)；
https://datasociety.net/library/ ai-red-teaming-is-not-a-e-stop-解决方案-ai-危害-建议-使用-red-teaming-for-ai-问责制/。

220. D.R. Maffioli，《生成性人工智能培训的版权: 通过标准化和透明度平衡合理使用》。(2023)。
https://doi.org/10.13140/rg.2.2.18478.48961。

221. A.Karamolegkou, J.李，L。周，A。S ø gaard，“版权侵权和大型语言模型”，*载于2023自然语言处理经验方法会议 (EMNLP 2023)* 论文集 (计算语言学协会，2023) pp。7403-7412。https://doi.org/10.18653/v1/2023.emnlp-main.458。

222. P。亨德森，X。李，D。Jurafsky, T.桥本，M.A.莱姆利，P。梁，基础模型与合理使用，arXiv:2303.15715 [cs.CY] (2023)。http://arxiv.org/abs/ 2303.15715。

223. A.Luccioni, J.维维亚诺，“盒子里是什么？普通爬行语料库中不良内容的分析”，*载于第59届计算语言学协会年会和第11届国际自然语言处理联合会议 (第2卷: 短篇论文)* (计算语言学协会，2021) pp。182-189。https://doi.org/10.18653/v1/2021.acl-short.24。

224. A.Birhane, V.Prabhu, S.韩，V。N.Boddeti, A.S。Luccioni，“进入巢穴: 调查多模式数据集中的仇恨”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id= 6URyQ9QhYv & noteId = 6URyQ9QhYv。

225. Y。曲，X。沈，X。他，M。Backes, S.Zannettou，Y。Zhang，“不安全的扩散: 关于从文本到图像模型生成不安全的图像和可恶的模因”，*《2023 ACM SIGSAC计算机和通信安全会议论文集》(ccs'23)*，(计算机协会，2023) pp。3403-3417。https://doi.org/10.1145/3576915.3616679。

226. A.Birhane, V.美国。Pra bhu，E。Kahembwe，多模式数据集: 厌女症、色情和恶性刻板印象，arXiv:2110.01963 [cs.CY] (2021)。http://arxiv.org/abs/ 2110.01963。

227. N.Shahbazi, Y.林，A。阿苏德赫，H. V.Jagadish，“代表数据中的偏见: 关于识别和解析技术的调查”。*ACM Comput.Surv*。**55**，293:1-293:39 (2023)。https://doi.org/10.1145/3588433。

228. D.Thiel，“识别和消除生成式ML训练数据和模型中的CSAM” (斯坦福互联网天文台，2023)；https://purl.stanford.edu/kh752sm9123。

229. J.Kreutzer，我.卡斯韦尔，L.王，A。Wahab, D. van Esch, N.Ulzii-Orshikh, A.塔波，N.Subramani, A.索科洛夫，C。西卡索特，M。Setyawan, S.沙林，S.Samb, B.Sagot, C.里维拉，A.Rios，我.Papadimitriou, S.Osei, P.O.苏亚雷斯。。。M. Adeyemi，质量一目了然: 对网络抓取的多语言数据集的审计。*Trans。Assoc。计算机。语言学家*。**10**，50-72 (2022)。https://doi.org/10.1162/tacl_a_00447。

230. T.de Vries，我.米斯拉，C。Wang，L. van der Maaten，“对象识别对每个人都有效吗？”在*IEEE/CVF会议计算机视觉和模式识别 (CVPR) 研讨会*上 (2019)。https://openaccess.thecvf.com/content_CVPRW_2019/论文/cv4gc/de_Vries_Does_Object_Recognition_Work_for_everyone_CVPRW_2019_paper.pdf。

231. 纽约时报公司诉微软公司等。(美国纽约南区地方法院，2023)。https:// nytco-assets.nytimes.com/2023/12/nyt_complint_dec2023.pdf。

232. J.道奇，M.Sap, A.Marasović, W.阿格纽尔，G.伊尔哈科，D。格罗内维尔德，M。米切尔，M。Gardner，“记录大型网络文本语料库: 关于巨大的干净爬行语料库的案例研究”，*《2021自然语言处理经验方法会议论文集》(EMNLP 2021)* (计算语言学协会，2021) pp。1286-1305。https://doi.org/10.18653/v1/2021.emnlp-main.98。

233. A.S. Luccioni, F.科里，H.斯里德兰，M。Ananny, J.舒尔茨，K。Crawford，“弃用数据集的框架: 标准化文档，识别和通信”，*载于2022 ACM会议关于公平，问责制和透明度 (FAccT '22)* 的会议记录 (计算机协会，2022) pp。199-212。https://doi.org/10.1145/3531146.3533086。

234. K。彭，A。Mathur, A.Narayanan，“减轻数据集危害需要管理: 1000论文的教训”，*《神经信息处理系统追踪数据集和基准1 (NeurIPS数据集和基准2021) round2》* (2021)。https:// 数据集-基准-会议记录.neurips.cc/paper/2021/hash/077e29b11be80ab57e1a2ecabb7da330-Abstract-round2.html。

235. B.麦加尼亚，D.戈德法布，P。马丁，M。阿特金森，S.Koulouzis, Z.Zhao，“数据起源”，“*迈向环境与地球科学的可互操作研究基础设施: 参考模型指导方法应对共同挑战*”，赵，M.赫尔斯特伦，编辑。(施普林格国际出版社，Cham，2020)，pp。208-225。

236. S。Longpre, R.马哈里，A。陈，N。奥本-马尔努，D.西里奥，W。布兰农，N.Muennighoff, N.Khazam, J.卡巴拉，K。佩里塞特拉，X。吴，E。希波尔，K。Bollacker, T.吴，L。别墅，S.彭特兰，S.Hooker，数据起源倡议: 人工智能中数据集许可和归属的大规模审计，arXiv:2310.16787 [cs.CL] (2023)。http://arxiv.org/abs/ 2310.16787。

237. E。佩雷斯，S.林格，K。Lukosiute，K。Nguyen, E.陈，S。海纳，C。佩蒂特，C。奥尔森，S.昆杜，S.Kadavath, A.琼斯，A。陈，B。曼恩，B.以色列，B。Seethor，C。麦金农，C.奥拉，D.Yan, D.阿莫迪，。。。J. Kaplan，“通过模型书面评估发现语言模型行为” (*发现关联计算语言学: ACL 2023*，2023) pp。13387-13434。https://doi.org/10.18653/v1/2023.findings-acl.847。

238. M. Sharma, M.Tong, T.Korbak, D.杜维诺，A.Askell, S.R。鲍曼，E。Durmus, Z.哈特菲尔德-多德兹，S。R。约翰斯顿，S。M. Kravec, T.麦克斯韦，S.McCandlish, K.恩杜斯，O.劳施，N.Schiefer, D.Yan, M.张，E。佩雷斯，“走向

理解语言模型中的交际“ *在第12届国际学习表征会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id=tvhaxkMKAn。

239. S。Santurkar, E.杜尔穆斯，F.拉达克，C。李，P。梁，T。桥本，“语言模型反映了谁的观点？”在 *第40届机器学习*国际会议论文集 (JMLR.org，2023) 卷。202，第页。29971-30004。https://dl.acm.org/doi/ 10.5555/3618408.3619652。

240. D.去吧，T.Korbak, G.Kruszewski, J.Rozen, N.Ryu, M.Dymetman，“通过f-发散最小化使语言模型与偏好保持一致”，*载于第40届国际机器学习会议 (ICML 2023)* 论文集 (2023) pp。11546-11583。https:// 诉讼。mlr。按 /v202/go23a.html。

241. Z。宋，T。蔡，J.D.李，W。J.Su，“对齐大型语言模型的奖励崩溃: 偏好排名的即时感知方法”，*ICML 2023研讨会基于偏好的学习*的许多方面 (2023)。https://openreview.net/forum？id= dpWxK6aqIK。

242. S。胡克，超越“算法偏差是一个数据问题”。**2***模式 (N Y)* ，100241 (2021)。https://doi.org/10.1016/j.Patter.2021.100241。

243. V。H. Maluleke, N.Thakkar, T.布鲁克斯，E。韦伯，T。达雷尔，A。A.埃夫罗斯，A。金泽，D。Guillory，“通过种族视角研究甘斯的偏见”，*计算机视觉-ECCV 2022: 第17届欧洲会议，特拉维夫，以色列，2022年10月23至27日，论文集，第十三*部分 (springer-verlag，2022) pp。344-360。https://doi.org/10.1007/978-3-031-19778-9_20。

244. R。Bommasani, K.Klyman, S.Longpre, S.卡普尔，N。马斯莱，B.熊，D。张，P。梁，“基金会模型透明度指数”(基金会模型研究中心 (CRFM) 和以人为中心的人工智能研究所 (HAI)，2023)；http://arxiv.org/abs/ 2310.12941。

245. V。赖，C。陈，A。史密斯-雷纳，Q。V。廖，C。Tan，“迈向人类人工智能决策科学: 经验性人类主体研究中的设计空间概述”，*《2023 ACM会议公平，问责制和透明度 (FAccT '23)》* (计算机协会，2023) pp。1369-1385。https://doi.org/10.1145/3593013.3594087。

246. G.班萨尔，B.Nushi, E.卡马尔，W。S。Lasecki, D.S。焊接，E.霍维茨，超越准确性: 心理模型在人类人工智能团队绩效中的作用。*HCOMP* **7** ，2-11 (2019)。https://doi.org/10.1609/hcomp.v7i1.5285。

247. R。Geirhos, K.梅丁，F。A.Wichmann，“超越准确性: 通过测量错误一致性来量化cnn和人类的试验行为”，载于*《第34届国际会议神经信息处理系统 (NIPS '20)》* (Curran Associates Inc.，2020) pp。13890-13902。https://dl.acm.org/doi/ 10.5555/3495724.3496889。

248. A.奥尔布赖特，如果你给法官一个风险评分: 来自肯塔基州保释决定的证据。*法律、经济和商业研究员讨论稿系列*，**85** ，2019-1 (2019)。http://www.law.harvard.edu/programs/ olin_center/奖品/2019-1.pdf.

249. M·霍夫曼，L.B.卡恩，D.李，招聘中的自由裁量权 *。**133** *“经济学*季刊”，765-800 (2018)。https://doi.org/10.1093/qje/qjx042。

250. E。Brynjolfsson, D.李，L。Raymond，“工作中的生成AI”(国家经济研究局，2023)；https://doi.org/10.3386/w3161.

251. V。玛尔达，S.Narayan，“新德里预测性警务系统中的数据”，*《2020公平，问责制和透明度*会议论文集》 (计算机协会，2020) pp。317-324。https://doi.org/10.1145/3351095.3372865。

252. 美国。Ehsan, R.辛格，J。梅特卡夫，M.Riedl，“算法烙印”，*载于2022 ACM会议关于公平，问责制和透明度 (FAccT '22)* 的会议记录 (计算机协会，2022) pp。1305-1317。https://doi.org/10.1145/3531146.3533186。

253. I.D. Raji, P.徐，C。霍尼斯堡，D。Ho，“局外人监督: 为人工智能治理设计第三方审计生态系统”，*《2022 AAAI /ACM AI，道德与社会*会议论文集》 (计算机协会，2022) pp。557-571。https://doi.org/10.1145/3514094.3534181。

254. X。沈，Z。陈，M。Backes, Y.沈，Y。Zhang，“现在做任何事情”: 在大型语言模型上表征和评估越狱提示，arXiv:2308.03825 [cs.CR] (2023)。http://arxiv.org/abs/ 2308.03825。

255. R。托洛萨纳，R。维拉-罗德里格斯，J.Fierrez, A.莫拉莱斯，J。Ortega-garcia，Deepfakes and beyond: 面部操纵和伪造检测的调查。*信息融合*，**64** ，131-148 (2020)。https://doi.org/10.1016/j.Inffus.2020.06.014。

256. M.穆斯塔克，J.萨尔米宁，M。Mäntymäki, A.Rahman, Y.K。Dwivedi，Deepfakes: 欺骗，缓解和机会。**154***商业研究*杂志，113368 (2023)。https://doi.org/10.1016/j.Jbusres.2022.113368。

257. M.米切尔，S.吴，A。Zaldivar，P。巴恩斯，L.Vasserman, B.哈钦森，E。斯皮策，我。D.Raji, T.Gebru，“模型报告的模型卡”，*载于《公平，问责制和透明度*会议论文集》 (ACM，2019) pp。220-229。https://doi.org/10.1145/3287560.3287596。

258. W。梁，N。Rajani, X.杨，E。Ozoani, E.吴，Y。陈，D。S。史密斯，J。邹，人工智能中有什么记录？32k人工智能模型卡的系统分析，arXiv:2402.05160 [cs.SE] (2024)。/ https://doi.org/10.48550/arXiv.2402.05160。

259. E。M.本德尔，B.弗里德曼，自然语言处理的数据陈述: 减轻系统偏见和促进更好的科学。*《计算语言学协会会刊》* **6**，587-604 (2018)。https://doi.org/10.1162/tacl_a_00041。

260. B.哈钦森，A。聪明，A.汉娜，E.丹顿，C。格里尔，O.Kjartansson，P。巴恩斯，M.Mitchell，"走向机器学习数据集的问责制: 软件工程和基础设施的实践"，*载于2021 ACM会议公平，问责制和透明度 (FAccT '21)* 的论文集 (计算机协会，2021) pp。560-575。https://doi.org/10.1145/3442188.3445918。

261. T.Gebru, J.摩根斯坦，B。维奇奥尼，J.W。沃恩，H. Wallach，H. D.三，K。Crawford，数据集的数据表。*Commun。美国医学*杂志**64**，86-92 (2021)。https://doi.org/10.1145/3458723。

262. M.阿诺德，R.K。E。贝拉米，M.欣德，S.Houde, S.梅塔，A。Mojsilović, R.奈尔，K.N.Ramamurthy, A.Olteanu, D.Piorkowski，D。Reimer, J.理查兹，J。Tsay, K.R。Varshney，FactSheets: 通过供应商的符合性声明增加对人工智能服务的信任。*IBM研究与发展*杂志**63**，6: 6:13 (2019)。https://doi.org/10.1147/jrd.2019.2942288。

263. F.古尔索伊，我。A.Kakadiaris，公共政策的基于AI的决策系统卡，arXiv:2203.04754 [cs.CY] (2022)。https://doi.org/10.48550/arXiv.2203.04754。

264. I.D. Raji, A.聪明，R.N.怀特，M.米切尔，T。格布鲁，B.哈钦森，J。史密斯-大声，D。塞隆，P。Barnes，"弥合AI问责制差距: 定义内部算法审计的端到端框架"，载于*《2020会议公平，问责制和透明度 (FAT * '20)》* (计算机协会，2020) pp。444-44。https://doi.org/10.1145/3351095.3372873。

265. C.陈O.李，D。陶，A。巴内特，C。鲁丁，J.K。Su，"这看起来像: 可解释图像识别的深度学习"，*《神经信息处理系统进展》(NeurIPS 2019)* (Curran Associates，inc.，2019) 第一卷。32。https:// 会议录.neurips.cc/paper_files/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html。

266. M. Danilevsky, K.钱，R。Aharonov, Y.Katsis, B.Kawas，P。Sen，"自然语言处理可解释人工智能现状调查"，*载于计算语言学协会亚太分会第一届会议和第十届国际自然语言处理联席会议 (AACL 2020)* 论文集 (计算语言学协会，2020) pp。447-459。https://aclanthology.org/2020 aacl-main.46。

267. A.D为，P。Rad，可解释人工智能 (XAI) 的机遇与挑战: 一项调查，arXiv:2006.11371 [cs.CV] (2020)。https://doi.org/10.48550/arXiv.2006.11371。

268. D.王明，H. X.，Y。F.李，T。N.阮，可解释的人工智能: 全面回顾。*Artif Intell Rev* **55**，3503-3568 (2022)。https://doi.org/10.1007/s10462-021-10088-y。

269. J.Gryz, M.Rojsz czak，黑色盒算法和个人权利: 没有的eas y解决方案来的 " expl ainability" 问题。*I n ter净P ol。Rev.***10** (2021)。https://policyreview.info/articles/analysis/ 黑盒算法和权利的-个人-不容易-解决方案-可解释性。

270. T.Ploug, S.Holm，*《人工智能医学》中的*"AI诊断竞赛权"。Lidstr ö mer，H. Ashrafian，编辑。(施普林格国际出版社，Cham，2022)，pp。227-238。

271. 《M. C.》布伊顿，L.A.丹尼斯，M.Schwammberger，"关于律师对自治系统的解释感兴趣的愿景"，*2023 IEEE第31届国际需求工程会议研讨会 (REW)* (2023) pp。332-336。https://doi.org/10.1109/rew57809.2023.00062。

272. 沃斯曼，V。Ramanujan, R.刘，A。Kembhavi, M.Rastegari, J.Yosinski, A.Farhadi，*《神经信息处理系统进展》(NeurIPS 2020)* 中的 "叠加中的超级掩膜" (Curran Associates，inc.，2020) 第一卷。444，第页。15173-15184。https:// 会议录.neurips.cc/paper/2020/hash/ad1f8bb9b51f023cdc80cf94bb615aa9-Abstract.html。

273. D.Bau, B.周，A。Khosla, A.奥利瓦，A.Torralba，"网络解剖: 量化深度视觉表示的可解释性"，*2017 IEEE计算机视觉和模式识别会议 (CVPR)* (2017) pp。3319-3327。https://doi.org/10.1109/cvpr.2017.354。

274. C.奥拉，A.Mordvintsev，L。舒伯特，特征可视化。*蒸馏***2**，10.23915/蒸馏00007 (2017)。https://doi.org/10.23915/distil.00007.

275. A.Ghorbani, J.Y。Zou，"神经元Shapley: 发现负责任的神经元"，*《神经信息处理系统进展》(NeurIPS 2020)* (Curran Associates，inc.，2020) 第一卷。444，第页。5922-5932。https:// 会议录.neurips.cc/paper/2020/hash/41c542dfe6e4fc3deb251d64cf6ed2e4-Abstract.html.

276. C.Olah, N.Cammarata, L.舒伯特，G.Goh, M.彼得罗夫，S.卡特，放大: 电路简介。*蒸馏* (2020)。https://doi.org/10.23915/distil.00024.001。

277. A.康米，A.N.Mavor-Parker, A.林奇，S.Heimersheim, A.Garriga-alonso，"走向机械可解释性的自动电路发现"，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum?_id= 89ia77nZ8u。

278. B.金，M。瓦滕伯格，J。吉尔默，C。蔡，J.韦克斯勒，F。维埃加斯，R。Sayres，“超越特征归因的可解释性: 概念激活向量 (TCAV) 的定量测试”，*载于第35届国际机器学习*会议论文集 (PMLR，2018) pp。2668-2677。https:// 诉讼。mlr。按 /v80/kim18d.html。

279. Y。Belinkov，探索分类器: 承诺、缺点和进步。*计算机。语言学家。Assoc。计算机。语言学家。***48** ，207-219 (2022)。https://doi.org/10.1162/科里 _a_00422。

280. S。黑色，L。Sharkey, L.Grinsztajn, E.温莎，D。布劳恩，J.Merizian, K.帕克，C。R。格瓦拉，B。米利奇，G.Alfour，C. Leahy，通过多面体透镜解释神经网络，arXiv:2211.12312 [cs.LG] (2022)。http://arxiv.org/abs/ 2211.12312。

281. A.邹，L。Phan, S.陈，J。坎贝尔，P。郭，R。任，A。潘，X。尹，M。Mazeika，A.-K.Dombrowski, S.Goel, N.李，M。J.Byun, Z.王，A。Mallen, S.Basart, S.Koyejo, D.宋，M.弗雷德里克森。。。D.亨德里克斯，代表工程: 一个顶级的-降低AI透明度的方法，arXiv:2310.01405 [cs.LG] (2023)。https://doi.org/10.48550/arXiv.2310.01405。

282. S。马克斯，C。Rager, E.J.Michaud, Y.别林科夫，D。Bau, A.Mueller，稀疏特征电路: 在语言模型中发现和编辑可解释的因果图，arXiv:2403.19647 [cs.LG] (2024)。https://doi.org/10.48550/arXiv.2403.19647。

283. S。卡特，Z.阿姆斯特朗，L.舒伯特，我。约翰逊，C。奥拉，激活阿特拉斯。*蒸馏***4** ，10.23915/蒸馏。00015 (2019)。https://doi.org/10.23915/蒸馏。

284. J.穆，J.Andreas，“神经元的成分解释”，*《神经信息处理系统进展》 (NeurIPS 2020)* (Curran Associates，inc.，2020) 第一卷。444，第页。17153-17163。https:// 会议录.neurips.cc/paper/2020/hash/c74956ffb38ba48ed6ce977af6727275-Abstract.html。

285. E。埃尔南德斯，S。施韦特曼，D。Bau, T.Bagashvili, A.托拉尔巴，J.安德烈亚斯，“深层视觉特征的自然语言描述”*在第十届国际学习表征*会议上 (2022)。https://openreview.net/forum？ id= NudBMY-tzDr.

286. S。卡斯珀，M.纳多，D。哈德菲尔德-梅内尔，G。*第36届神经信息处理系统会议 (NeurIPS 2022)* 的Kreiman，“强大的特征级对手是可解释性的工具”(2022)。https://openreview.net/forum？ id=lQ-- doSB2o。

287. M. Geva, J.Bastings，K。菲利波瓦，A.Globerson，“剖析自动回归语言模型中事实关联的回忆”*在2023自然语言处理经验方法会议 (EMNLP 2023)* 中 (2023) pp。11116-12235。https://openreview.net/forum？ id= F1G7y94K02。

288. * T。Bolukbasi, A.梨ce，A。元，A。科宁，E。Reif, F.维加斯，M。瓦滕伯格，伯特的可解释性幻觉，阿西夫: 2104.07143 [cs.CL] (2021)。https://doi.org/10.48550/arXiv.2104.07143。

289. A.马克洛夫，G。兰格，A.盖格，N。南达，“这是你要找的子空间吗？子空间激活修补的可解释性错觉“ *在第12届国际学习表征会议 (ICLR 2024)* 中”(2023)。https://openreview.net/forum？ id= Ebt7JgMHv1。

290. M. Ananny, K.克劳福德，看到不知道: 透明度理想的局限性及其在算法问责制中的应用。*新媒体Soc*。**20** ，973-989 (2018)。https://doi.org/10.1177/1461444816676645。

291. T.米勒，“人工智能中的解释: 来自社会科学的见解”。*Artif.Intell*。**267** ，1-38 (2019)。https://doi.org/10.1016/j.artint.2018.07.007.

292. F.Doshi-Velez, B.Kim，走向可解释机器学习的严格科学，arXiv:1702.08608 [stat.ML] (2017)。https://doi.org/10.48550/arXiv.1702.08608。

293. Z。C. Lipton，模型可解释性的神话: 在机器学习中，可解释性的概念既重要又滑。**16** *ACM队列* ，31-57 (2018)。https://doi.org/10.1145/3236386.3241340。

294. * 米。L.莱维特，A.莫科斯，“走向可证伪的可解释性研究”，arXiv:2010.12016 [cs.CY] (2020)。http://arxiv.org/abs/ 2010.12016。

295. T.Räuker, A.Ho, S.卡斯珀，D.Hadfield-menell，《走向跨父母AI: 解释深度神经网络内部结构的调查》，arXiv:2207.13243 [cs.LG] (2022)。http://arxiv.org/abs/ 2207.13243。

296. M. Krishnan，《反对互易性: 机器学习中可解释性问题的批判性考察》。*Philos.技术*。**444** ，487-502 (2020)。https://doi.org/10.1007/s13347-019-00372-9。

297. J.阿德巴约，J。吉尔默，M。Muelly，我Goodfellow, M.哈特，B.Kim，*《神经信息处理系统进展》 (NeurIPS 2018)* 中的“显著性地图的健全性检查”(Curran Associates，inc.，2018) 第一卷。31. https:// 会议录.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

298. J.阿德巴约，M.Muelly，我Liccardi, B.Kim，“模型解释的调试测试”，*进展神经信息处理系统 (NeurIPS 2020)* (Curran Associates，inc.，2020) 卷。444，第页。700-712。https:// 会议录.neurips.cc/paper/2020/hash/075b051ec3d22dac7b33f788da631fd4-Abstract.html.

299. P。Hase, M.班萨尔，B.金，A。Ghandeharioun，“本地化会为编辑提供信息吗？*第37届神经信息处理系统会议 (NeurIPS 2023)* 中基于因果关系的本地化与知识在语言模型中编辑的惊人差异 (2023)。https://openreview.net/forum？ id=EldbUlZtbd。

300. S。卡斯珀，C。Ezell, C.齐格曼，N。Kolt, T.L.柯蒂斯，B.Bucknall, A.豪普特，K。魏，J。Scheurer, M.霍布哈恩，L.Sharkey, S.克里希纳，M.冯哈根，S.Alberti, A.Chan, Q.孙，M。Gerovitch, D.Bau, M.泰格马克。。。D.Hadfield-menell，黑盒访问不足以进行严格的人工智能审计，arXiv:2401.14446 [cs.CY] (2024)。http://arxiv.org/abs/ 2401.14446。

301. B.S. Bucknall, R.F.Trager，“前沿人工智能模型第三方研究的结构化访问: 调查”研究人员的模型访问要求 ”(牛津大学马丁学院，牛津大学和人工智能治理中心，2023)：https:// www.ceris.be/wp-content/uploads/2024/03/structure-access-for-第三方-Research-Robert-悲剧.Pdf。

302. R。阿什莫尔，R.卡里内斯库，C。帕特森，确保机器学习生命周期。*ACM计算机。Surv*。**54** ，1-39 (2022)。https://doi.org/10.1145/3453444。

303. S。卡斯珀，X.戴维斯，C。施，T.K。吉尔伯特，J。Scheurer, J.Rando, R.弗里德曼，T。Korbak, D.林德纳，P。弗莱雷，T.T.王，S。马克，C.-R。塞格里，M。卡罗尔，A.彭，P。Christoffersen, M.Damani, S.斯洛克姆，美国。安瓦尔。。。D.Hadfield-menell，从人类反馈中强化学习的开放问题和基本限制。*机器学习研究*学报 (2023)。https://openreview.net/forum？id=bx24KpJ4Eb。

304. S。Longpre, S.卡普尔，K。Klyman, A.Ramaswami, R.Bommasani, B.布利利-哈梅林，Y.黄，A。Skowron，Z.-X.勇，S。Kotha, Y.曾，W。施，X。杨，R。索森，A。罗比，P。Chao, D. Yang, R.贾，D。康，。。。P.亨德森，“人工智能评估和红色团队的安全港”，arXiv:2403.04893 [cs.AI] (2024)。https://doi.org/10.48550/arXiv.2403.04893。

305. T.Shevlane，结构化访问: 安全AI部署的新兴范例，arXiv:2201.05159 [cs.AI] (2022)。https://doi.org/10.48550/arXiv.2201.05159。

306. M. Dolata, S.Feuerriegel, G.Schwabe，算法公平性的社会技术观点。*Inf*。*系统*。*J.* 754-818 **32** (2022)。https://doi.org/10.1111/isj.12370。

307. S。拉扎尔，A.纳尔逊，人工智能安全在谁的条件下？**38 1***科学*，138 (2023)。https://doi.org/10.1126/科学.adi8982.

308. Y。王，Y。朱，C。孔，S。魏，X。易，X。谢，J。Sa ng，CDEval: 测量大型语言模型的文化维度的基准，arXiv:2311.16421 [cs.CL] (2023)。http://arxiv.org/abs/ 2311.16421。

309. Z。X.勇，C。Menghini, S.巴赫，*NeurIPS社会责任语言建模研究 (SoLaR)* 中的 “低资源L语言越狱GPT-4” (2023)。https://openreview.net/forum？id= pn83r8V2sv。

310. Y。金，M。钱德拉，G.Verma, Y.胡，M·德·乔杜里，S。Kumar，更好地用英语提问: 医疗保健查询的大型语言模型的跨语言评估，arXiv:2310.13132 [cs.CL] (2023)。http://arxiv.org/abs/ 2310.13132。

311. * A.Üstün, V.Aryabumi，Z.-X。勇，W.-Y。Ko, D.D'souza, G.Onilude, N.Bhandari, S. 辛格，H.-L。Ooi, A. 凯伊德，F.Vargus，P。Blunsom, S.Longpre，N。Muennighoff，M。法达伊，J.Kreutzer, S.Hook er，Aya模型: 一个指令优化的开放访问多语言语言模型，arXiv:2402.07827 [cs.CL] (2024)。http://arxiv.org/abs/ 2402.07827。

312. Y。徐，曹，W。杜，W。王，跨岭情感分析综述: 方法、模型和评价。*数据科学与工程*279 **7** -299 (2022)。https://doi.org/10.1007/s41019-022-00187-3。

313. * W。王，Z。Tu，C。陈，Y。元，J.-T。黄，W。焦，M。R。Lyu，所有语言都很重要: 关于大型语言模型的多语言安全性，arXiv:2310.00905 [cs.CL] (2023)。http://arxiv.org/abs/ 2310.00905。

314. A.D. Selbst，算法影响评估的机构观点。**35** (2021)。https://jolt.law.harvard.edu/assets/articlePDFs/ v35/Selbst-机构观点算法影响-评估.Pdf。

315. S。L.布洛杰特，S.Barocas，H. Daum é，III，H. Wallach，“语言 (技术) 就是力量: NLP中“ 偏见 ”的批判性调查”，*载于第58届计算语言学协会 (ACL 2020)* 年会论文集 (计算语言学协会，2020) pp。5454-5476。https://doi.org/10.18653/v1/2020.acl-main.485。

316. A.哈格蒂，我。Rubinov，“全球人工智能和人工智能的社会影响和伦理意义回顾”，arXiv:1907.07892 [cs.CY] (2019)。https://doi.org/10.48550/arXiv.1907.07892。

317. M. M.马斯，“使人工智能监管与社会技术变革保持一致”，载于*《牛津人工智能治理手册》*，B.布洛克，Y.-C。陈，J。Himmelreich, V.M。哈德逊，A。Korinek, M.M。杨，B.张，编辑。(牛津大学出版社，2022)。

318. * L。魏丁格，J.巴恩哈特，J。布伦南，C。巴特菲尔德，S.年轻，W。霍金斯，L.A.亨德里克斯，R。科马内斯库，O。张，M.罗德里格斯，J.Beroshi，D。布洛克斯威奇，L。Proleev, J.陈，S。法夸尔，L.啁，我。Gabriel, A.达福，W。Isaac，“高级AI模型的整体安全和责任评估” (Google Deepmind，2024)：http://arxiv.org/abs/ 2404.14068。

319. M. Raghavan，“算法决策*的社会影响*” (计算机协会，纽约，纽约，美国，ed。1，2023)，卷。53页。366。

320. S。卡普尔，R。Bommasani, K.Klyman, S.Longpre，A。Ramaswami, P.Cihon, A.霍普金斯，K.班克斯顿，S.比德曼，M。博根，R.乔杜里，A。恩格勒，P。亨德森，Y。Jernite, S.Lazar, S.Maffulli, A.纳尔逊，J。皮诺，A。Skowron，。。。A. Narayanan，“关于开放基础模型的社会影响” (斯坦福以人为中心的人工智能研究所，2024)：https://crfm.stanford.edu /开放-fms/paper.pdf.

321. E。莫斯，E.A.沃特金斯，R。辛格，M.C.Elish, J.梅特卡夫，“组装问责制: 公共利益的算法影响评估 ” (数据与社会，2021); https://datasociety.net/library/ 组装-问责制-算法-影响评估-为公共利益/。

322. * 我。索莱曼，Z。塔拉特，W。Agnew, L.艾哈迈德，D.贝克，S.L.布洛杰特，H.道美，III，J。道奇，E.埃文斯，S.妓女，Y.Jernite, A.S。Luccioni, A.卢索利，M.米切尔，J。纽曼，M.-T。Png, A.海峡，A.Vassilev，评估生成性人工智能系统在系统和社会中的社会影响，arXiv:2306.05949 [cs.CY] (2023)。http://arxiv.org/abs/ 2306.05949。

323. * L。魏丁格，M.Rauh, N.Marchal, A.曼齐尼，L.A.亨德里克斯，J。Mateos-Garcia, S.伯格曼，J。凯，C.格里芬，B。巴里亚克h, I.加布里埃尔，V。Rieser, W.艾萨克，"生成式人工智能系统的社会技术安全评估" (谷歌Deepmind，2023)；http://arxiv.org/abs/ 2310.11986。

324. A.D. Selbst, D.博伊德，S.A.弗里德勒，S.Venkatasubramanian, J.Verdsi，"社会技术系统中的公平与抽象"，*载于《公平，问责制和透明度会议论文集》(FAT * '19)*，(计算机协会，2019) pp。59-68。https://doi.org/10.1145/3287560.3287598。

325. E。莫斯，J.Metcalf，"道德所有者: 数据驱动的技术公司中组织责任的新模型" (数据与社会，2020)；https://datasociety.net/wp-content/uploads/2020/09/Ethics-Owners_20200923-DataSociety.pdf。

326. M.费弗，M.Skirpan, Z.Lipton，H. Heidari，"从偏好激发到参与式ML: 对未来研究的批判性调查和指南"，*载于2023 AAAI/ACM AI，道德与社会会议 (AIES '23)* 论文集 (ACM，2023) pp。38-48。https://doi.org/10.1145/3600211.3604661。

327. F.德尔加多，S.杨，M。Madaio, Q.杨，"人工智能设计中的参与性转向: 理论基础和实践现状"，*载于第三届ACM会议关于算法、机制和优化的公平和访问 (eaamo'23)* 的论文集 (计算机器协会，2023) pp。1-23。https://doi.org/10.1145/3617694.3623261。

328. A.Birhane, W.艾萨克，V。Prabhakaran, M.迪亚兹，M.C.Elish，我.加布里埃尔，S.穆罕默德，"人民的权力？参与式AI的机遇和挑战"，*载于第二届ACM会议关于公平和访问的算法，机制和优化 (EAAMO '22)* 的论文集 (计算机协会，2022) pp。1-8。https://doi.org/10.1145/3551624.3555290。

329. J.梅特卡夫，E.莫斯，E.A.沃特金斯，R。辛格，M.C.Elish，"算法影响评估和问责制: 影响的共同构建"，*《2021 ACM会议公平，问责制和透明度 (FAccT '21)》* (计算机协会，2021) pp。735-746。https://doi.org/10.1145/3442188.3445935。

330. D.马丁，小，V。Prabhakaran, J.Kuhlberg, A.聪明，W.S。Isaac，"通过基于社区的系统动力学实现更公平的机器学习的参与式问题制定"，*ICLR研讨会关于现实生活*中的机器学习 (2020)。https://doi.org/10.48550/arXiv.2005.07572 。

331. M.斯隆，E.莫斯，O.阿沃莫洛，L。Forlano，"参与不是机器学习的设计修复"，*在第二届ACM会议关于公平和访问的算法，机制和优化 (EAAMO '22)* 的论文集中 (计算机协会，2022) pp。1-6。https://doi.org/10.1145/3551624.3555285。

332. R。I. J.多比，T。K。吉尔伯特，Y。Mintz，"人工智能中的艰难选择: 通过社会技术承诺解决规范性不确定性 (AIES '20)" *公共关系AAAI/ACM人工智能、伦理和社会会议论文告*(计算机协会，2020) pp。242。https://doi.org/10.1145/3375627.3375861。

333. S。A.弗里德勒，C.Scheidegger, S.Venkatasubramanian，公平的 (Im) 可能性: 不同的价值体系需要不同的公平决策机制。*Commun。ACM* **64**，136-143 (2021)。https://doi.org/10.1145/3433949。

334. N.古哈，C。M。劳伦斯，L.A.盖尔马德，K.T.Rodolfa, F.苏拉尼，R.博马萨尼，我D.拉吉，M.-F。Cuéllar, C.霍尼斯堡，P。梁，D。E。Ho，人工智能监管有其自身的一致性问题: 披露、注册、许可和审计的技术和制度可行性。**92***《乔治·华盛顿法律评论》* (2024)。https://dho.stanford.edu/wp-内容/上传/AI_Regulation.pdf。

335. 美国。巴特，A。Xiang, S.夏尔马，A.韦勒，A。Taly, Y.贾，J。戈什，R。普里，J.M。F.莫拉，P。Eckersley，"部署中的可解释机器学习"，*《2020 Confere nce关于公平，问责制和透明度 (FAT * '20)* 的论文集》 (计算机协会，2020) pp。648-657。https://doi.org/10.1145/3351095.3375624。

336. K。Kaye, P.Dixon，"风险分析: 评估和改进人工智能治理工具 -- 人工智能治理工具的国际回顾和前进道路的建议" (世界隐私论坛，2023)；https://www.worldprivacyforum.org/wp内容/上传/2023/12/WPF_Risky_Analysis_December_2023_fs.pdf。

337. D.松弛，S.Hilgard, E.贾，S。Singh，H. Lakkaraju，"愚弄石灰和SHAP: 对事后解释方法的对抗性攻击"，*载于AAAI/ACM会议关于AI，道德和社会 (AIES '20)* 的论文集 (计算机协会，2020) pp。180-186。https://doi.org/10.1145/3375627.3375830。

338. I.D. Raji, J.Yang，"关于ML: 关于机器学习生命周期的理解和透明度的注释和基准"，位于*NeurIPS 2019以人为中心的机器学习研讨会*中 (2019)。https://doi.org/10.48550/arXiv.1912.06166。

339. 项目管理协会 (PMI)，"项目管理知识体系指南 (PMBOK指南)" (PMI，2000)。

340. P。V. Falade，解码威胁景观: 社会工程攻击中的Cha tGPT、FraudGPT和WormGPT。*int。j.sci。res。计算机。sci。eng。inf。技术。***9**，185-198 (2023)。https://doi.org/10.32628/cseit2390533.

341. Y。姚，J。段，K。徐，Y。蔡，Z.太阳，Y。张，关于大型语言模型 (LLM) 安全和隐私的调查: 好的，坏的和丑陋的。*高置信度计算*100211 **4** (2024)。https://doi.org/10.1016/j.Hcc.2024.100211。

342. N.贝古，J.Vinoy, A.杜达，M.Korczy ń ski，“探索AI的黑暗面: 使用ChatGPT进行高级网络钓鱼攻击设计和部署”，*2023 IEEE通信与网络安全会议 (CNS)* (IEEE，2023)。https://doi.org/10.1109/cns59707.2023.10288940

343. * 米。Heinemeyer，网络安全威胁-生成AI的电子邮件妥协。(2023)。https://es.darktrace.com/blog/ 解决网络安全电子邮件妥协的软肋。

344. O.本德尔，人类声音的合成。*AI Soc*。**34** ，83 -89 (2019)。https://doi.org/10.1007/s00146-017- 0748-x。

345. R.吉尔，J.Virgili-Gomà，J.-M。洛佩斯-吉尔，R.加西亚，Deepfakes: 演变和趋势。*软计算*11318 **27** ，11295 (2023)。https://doi.org/10.1007/s00500-023-08605-y 。

346. 阿。F. Gambín, A.Yazidi, A. Vasilakos, H. Haugerud, Y.Djenouri，Deepfakes: 当前和未来趋势。**57***人工智能评论* ，64 (2024)。https://doi.org/10.1007/s10462-023-10679-x。

347. V。Ciancaglini, C.吉布森，D。桑丘，O.麦卡锡，M。Eira, P.Amann, A.Klayn，“恶意使用和滥用人工智能” (欧盟执法合作署，2020)。

348. R。Umbach, N.亨利，G.胡子，C.Berryessa，非自愿的合成亲密意象: 10个国家的流行、态度和知识，arXiv:2402.01721 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2402.01721。

349. 《M. B.》库格勒，C。速度，深度伪造隐私: 态度和监管。*西北。大学。法律Rev.***116** ，611-680 (2021)。https://scholarlycommons.law.northwestern.edu/nulr/ vol116/iss3/1。

350. M.维奥拉，C.Voto，旨在滥用？深度伪造和亲密图像的非自愿扩散。*合成* **201** ，30 (2023)。https://doi.org/10.1007/s11229-022-04012-2。

351. D.M. J.Lazer, M.A.Baum, Y.本克勒，A.J.Berinsky, K.M。格林希尔，F。门策，M。J.梅茨格，B。Nyhan, G.Pennycook，D。罗斯柴尔德，M.Schudson, S.A.Sloman，C。R。桑斯坦，E。A.Thorson, D.J.瓦茨，J.L.Zittrain，假新闻的科学。*科学***359** ，1094-1096 (2018)。https://doi.org/10.1126/科学aao2998。

352. G.Spitale，N。Biller-Andorno, F.Germani，人工智能模型GPT-3 (dis) 比人类更好地告诉我们。**9** *Sci Adv* ，eadh1850 (2023)。https://doi.org/10.1126/科学adh1850。

353. M. Jakesch, J.T.汉考克，M. Naaman，人工智能生成语言的人类启发式有缺陷。*Proc.Natl.Acad。Sci。美国。A.* e2208839120 (2023) **120**。https://doi.org/10.1073/pnas.2208839120。

354. K.-C。杨，F。Menczer，人工智能驱动的恶意社交僵尸网络的剖析，arXiv:2307.16336 [cs.CY] (2023)。https://doi.org/10.48550/arXiv.2307.16336。

355. M.马苏德，M.纳瓦兹，K。M。马利克，A.Javed, A.Irtaza，H. Malik，Deepfakes的产生和检测: 最先进的 ，开放的挑战，对策和前进的方向。*应用智能*3974 **53** -4026 (2023)。https://doi.org/10.1007/s10489-022-03766-z。

356. D.库克，A。爱德华兹，S.Barkoff, K.凯利，像硬币一样好oss: 人工智能生成的图像，视频，音频和视听刺激的人类检测，arXiv:2403.16760 [cs.HC] (2024)。http://arxiv.org/abs/ 2403.16760.

357. S。J. Nightingale，H. Fari ，人工合成的人脸与真实人脸没有区别，更值得信赖。*Proc.Natl.Acad。Sci。美国。A.* (2022) **119**。https://doi.org/10.1073/pnas.2120481119。

358. S。C.马茨，J.D.Teeny, S.S。Vaid，H. P特斯，G.M。哈拉里，M.Cerf，生成人工智能在大规模个性化说服方面的潜力。*Sci。代表。4692* **14** (2024)。https://doi.org/10.1038/s41598-024-53755-0。

359. H。白，J。G.Voelkel, J.C.艾希斯塔特，R。威勒，人工智能可以在政治问题上说服人类。(2023)。https://doi.org/10.31219/osf.io/stakv。

360. K。哈肯伯格，L.易卜拉欣，B.M。Tappin, M.Tsakiris，比较了角色扮演大型语言模型和人类专家对两极分化的美国的说服力。政治问题。(2023)。https://doi.org/10.31219/osf.io/ey8db。

361. K。Hac kenburg，H. Margetts，使用大型语言模型评估政治微定位的说服力。(2023)。https://doi.org/10.31219/osf.io/wnt8b。

362. A.Simchon, M.爱德华兹，S.Lewandowsky，《生成人工智能时代政治微目标的说服力》。*PNAS Nexus* **3** ，gae035 (2024)。https://doi.org/10.1093/pnasnexus/pgae035。

363. B.M. Tappin, C.维滕贝格，L.B.休伊特，A。J.Berinsky, D.G.兰德，将政治微目标的潜在有说服力的回报量化。*Proc.Natl.Acad。Sci。美国。A.* e2216261120 **120** (2023)。https://doi.org/10.1073/pnas.2216261120。

364. F.萨尔维，M.H.里贝罗，R.Gallotti, R.韦斯特，关于大语言模型的通用性说服力: 一项随机对照试验，arXiv:2403.14380 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2403.14380。

365. T.H.科斯特洛，G.Pennycook，D。G.兰德，通过与人工智能的对话持久地减少阴谋信仰。(2024)。https://doi.org/10.31234/osf.io/xcwdn。

366. P。S、公园，S.戈德斯坦，A.O'Gara, M.陈，D。Hendrycks，AI欺骗: 对示例、风险和潜在解决方案的调查。*图案***5** (2024)。https://doi.org/10.1016/j.Patter.2024.100988。

367. * 米。Phuong, M.艾奇森，E。Catt, S.科根，A。Kaskasoli, V.克拉科夫纳，D。林德纳，M。Rahtz, Y.阿萨尔，S。霍德金森，H.霍华德，T.Lieberum, R.库马尔，M。A.Raad, A.Webson, L.Ho, S.林，S.法夸尔，M.哈特。。。T. Shevlane，“评估危险能力的前沿模型”(Google Deepmind，2024); https://doi.org/10.48550/arXiv.2403.13793。

368. M.伯特尔，T.伍德赛德，“人工影响: 人工智能驱动的说服力分析”，arXiv:2303.08721 [cs.CY] (2023)。https://doi.org/10.48550/arXiv.2303.08721。

369. S。卡普尔，A.Narayanan，“如何为社交媒体上的大量生成AI做准备: 对挑战和机遇的扎根分析”(哥伦比亚大学骑士第一修正案研究所，2023)；https:// s3.amazonaws.com/kfai-documents/documents/a566f4ded5/How-to-Prepare-for-the-Deluge-of-Generative-AI-on-Social-Media.pdf。

370. M. Hameleers，廉价与深度操纵: 政治环境中廉价假货与深度假货的影响。**36***国际公共行动*(2024)。https://doi.org/10.1093/ijpor/edae004.

371. S。Zuboff*《监视资本主义时代: 争取人类未来的新权力》*(公共事务，2019)，pp。704。

372. S。Vosoughi, D.罗伊，S.阿拉尔，网上真假新闻的传播。*科学***359** ，1146-1151 (2018)。https://doi.org/10.1126/科学aap9559。

373. D.香橼，R。切斯尼，深度伪造: 对隐私、民主和国家安全的迫在眉睫的挑战。*Calif.法律Rev.1753* **107** (2019)。https://scholarship.law.bu.edu /学院_奖学金/640。

374. V。S. Sadasivan, A.K umar, S.Balasubramanian, W.王，S。Feizi，可以可靠地检测AI生成的文本吗？，arXiv:2303.11156 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2303.11156。

375. S。S、戈萨尔，S.Chakraborty, J.盖平，F.黄，D。Manocha, A.Bedi，一项关于人工智能生成文本检测的可能性和不可能性的调查。*机器学习研究*学报 (2023)。https://openreview.net/pdf？id=AXtFeYjboj。

376. J.罗，G。南，D。李，Y。Tan，AI生成的评论检测。(2023)。https://doi.org/10.2139/ssrn.4610727。

377. L.Fröhling, A.Zubiaga，基于特征的自动语言模型检测: 解决GPT-2，GPT-3和格罗弗。**7** *PeerJ计算机科学* ，e443 (2021)。https://doi.org/10.7717/peerj-cs.443。

378. S。格尔曼，H. Strobelt，A。拉什，“GLTR: 生成文本的统计检测和可视化”，*载于第57届计算语言学协会年会: 系统演示*论文集 (计算语言学协会，2019) pp。111-116。https://doi.org/10.18653/v1/P19-3019。

379. D.马科维茨，J。T.汉考克，J.N.Bailenson，固有的虚假人工智能交流和人类故意虚假交流的语言标记: 来自酒店评论的证据。*J.朗.Soc。神经病。***43** ，63-82 (2024)。https://doi.org/10.1177/0261927x231200201。

380. T.柏柏尔·萨丁哈，人工智能生成的文本与人类创作的文本: 多维比较。*应用语料库语言学* ，100083 **4** (2024).https://doi.org/10.1016/j.Acorp.2023.100083。

381. Y。谢，A。Rawal, Y.Cen, D.赵，S.K。Narang, S.Soshmita，MUGC: 机器生成与用户生成的内容检测，arXiv:2403.19725 [cs.CL] (2024)。https://doi.org/10.48550/arXiv.2403.19725。

382. M.基督，S.冈恩，O.Zamir，语言模型的不可检测水印，arXiv:2306.09194 [cs.CR] (2023)。https://doi.org/10.48550/arXiv.2306.09194。

383. H。张，B。L.爱德曼，D。Francati, D.文丘里，G.Ateniese, B.Barak，《沙子中的水印: 生成模型的强水印不可能》，arXiv:2311.04378 [cs.LG] (2023)。https://doi.org/10.48550/arXiv.2311.04378 。

384. C.R.莱博维奇，S.麦格雷戈，A。Ovadya，“Deepfake检测困境: 合成媒体中对抗性动态的多利益相关者探索”，*在2021 AAAI/ACM AI，道德与社会会议 (AIES '21)* 上 (计算机协会，2021) pp。736-744。https://doi.org/10.1145/3461702.3462584。

385. L.钟，Z。王，LLM可以替换堆栈溢出吗？大型语言模型代码生成的健壮性和可靠性研究*AAAI* **38** ，21841-21849 (2024)。https://doi.org/10.1609/aaai.v38i19.30185。

386. * H. Khlaaf, P.米什金，J.Achiam, G.克鲁格，M。Brundage，“代码合成大型语言模型的危险分析框架”(OpenAI，2022)；http://arxiv.org/abs/ 2207.14157。

387. H。皮尔斯，B.艾哈迈德，B.谭，B。多兰-加维特，R。卡里，“在键盘上睡着了？*2022 IEEE安全与隐私 (SP)* 研讨会(IEEE计算机协会，2022)，评估GitHub Copilot的代码贡献的安全性。754-768。https://doi.org/10.1109/sp46214.2022.9833571。

388. G.邓勇。刘，V。市长-维尔奇斯，P。刘，Y。李，Y。徐，T。张，Y。刘，M。Pinzger, S.Rass，PentestGPT: 一个LLM支持的自动渗透测试工具，arXiv:2308.06782 [cs.SE] (2023)。http://arxiv.org/abs/ 2308.06782。

389. * J。徐，J。W。斯托克斯，G。麦当劳，X.白，D。马歇尔，S.王，A。Swaminathan, Z.Li，AutoAttacker: 实现自动网络攻击的大型语言模型引导系统，arXiv:2403.01038 [cs.CR] (2024)。
http://arxiv.org/abs/ 2403.01038。

390. A.Happe, A.卡普兰，J.Cito，LLMs作为黑客: 自主Linux特权升级攻击，arXiv:2310.11409 [cs.CR] (2023)。，
https://doi.org/10.48550/arXiv.2310.11409。

391. R。方，R。宾杜，A.古普塔，Q.詹，D。Kang，LLM代理可以自动破解网站，arXiv:2402.06664 [cs.CR] (2024)。
https://doi.org/10.48550/arXiv.2402.06664。

392. 邵M，B。陈，S。Jancheska, B.多兰-加维特，S.Garg, R.卡· rri，M。Shafique，llm解决进攻性安全挑战的经验评估，arXiv:2402.11814 [cs.CR] (2024)。https://doi.org/10.48550/arXiv.2402.11814。

393. R。拉曼，P.Caly am, K.Achutan，ChatGPT或Bard: 谁是更好的认证道德黑客？*计算机与安全*，**140** ，103804 (2024)。https://doi.org/10.1016/j.Cose.2024.103804。

394. 安全和新兴技术中心，B.布坎南，J.班塞默，D.凯里，J.卢卡斯，M。Musser，“自动化网络攻击: 炒作与现实” (安全与新兴技术中心，2020)；
https://doi.org/10.51593/2020ca002。

395. 国家网络安全中心 (NCSC)，“人工智能对网络威胁的近期影响” (GOV.UK，2024)；
人工智能对网络威胁的 https://www.ncsc.gov.uk/report/ 影响。

396. * B.Berabi, A.Gronskiy, V.Raychev, G.Sivanrupan, V.Chibotaru, M.V echev，DeepCode AI Fix: 使用大型语言模型修复安全漏洞，arXiv:2402.13291 [cs.CR] (2024)。https://doi.org/10.48550/arXiv.2402.13291。

397. R。孟，M。Mirchev, M.Böhme, A.Roychoudhury，“大型语言模型引导的协议模糊测试”，*我于第31届年度网络和分布式系统安全研讨会 (NDSS)* 论文集 (2024)。https:// www.ndss-
symposium.org/wp内容/上传/2024-556-paper.pdf.

398. Y。丁，Y。傅欧。易卜拉欣，C.Sitawarin，X。陈，B。Alomair，D。瓦恩，呃，B。雷，Y.Chen，代码语言模型的漏洞检测: 我们还有多远？，arXiv:2403.18624 [cs.SE] (2024)。http://arxiv.org/abs/ 2403.18624。

399. H。皮尔斯，B.谭，B。艾哈迈德，R.卡里，B.Dolan-Gavitt，“使用大型语言模型检查零漏洞修复”，*2023 IEEE 安全与隐私 (SP)* 研讨会 (2023) pp。2339-2356。
https://doi.org/10.1109/sp46215.2023.10179324。

400. * A.舍斯托夫，R。列维切夫，R。穆萨巴耶夫，E.马斯洛夫，A.Cheshkov，P。Zadorozhny，用于漏洞检测的调整大型语言模型，arXiv:2401.17010 [cs.CR] (2024)。https://doi.org/10.48550/arXiv.2401.17010。

401. N.Risse, M.B ö hme，“揭示自动漏洞检测的机器学习的局限性”，在*USENIX安全研讨会2024*中 (2024) pp。19。

402. * 人工智能网络挑战赛，人工智能网络挑战赛 (2024)。
https://aicyberchallenge.com /。

403. S。乌拉，M。Han, S.Pujar, H. Pearce, A.Coskun, G.Stringhini，“LLMs无法可靠地识别和推断安全漏洞 (尚未？): *IEEE 安全与隐私*研讨会的综合评估，框架和基准” (2024)。https://research.ibm.com/publications/ llms-不能可靠地识别和原因关于安全-
漏洞尚未全面评估框架和基准。

404. S。玫瑰，C。尼尔森，“理解人工智能促进的生物武器发展” (长期复原力中心，2023);
https://www.longtermresilience.org/post/ 报告-启动-检查-风险-在交叉点-人工智能-
和-生物。

405. J.B. Sandbrink，人工智能和生物滥用: 区分语言模型和生物设计工具的风险，arXiv:2306.13952 [cs.CY] (2023)。https://doi.org/10.48550/arXiv.2306.13952 。

406. E。H.索斯，R.罗卡，K。科尔多瓦，M。斯佩克特，K.M。Esvelt，大型语言模型能否使两用生物技术的获取民主化？，arXiv:2306.03809 [cs.CY] (2023)。https://doi.org/10.48550/arXiv.2306.03809。

407. S。R.卡特，N.惠勒，S.Chwalek，C。艾萨克，J.M。Yassif，“人工智能与生命科学的融合: 保护技术，重新思考治理和预防灾难” (核威胁倡议，2023)；
https://www.nti.org/wp内容/上传/2023/10/NTIBIO_AI_FINAL.pdf。

408. N.李，A。潘，A.Gopal, S.岳，D。Berrios, A.加蒂，J.D.李，A.-K。Dombrowski, S.Goel, L.潘，G.Mukobi, N.Helm-汉堡，R.拉巴比迪，L.Justen, A.B.刘，M。陈，我。巴拉斯，O.张X。朱，。。。D. Hendrycks，The WMDP Benchmark: measusing and reduction malicuse Unlearning，arXiv:2403.03218 [cs.LG] (2024)。
https://doi.org/10.48550/arXiv.2403.03218。

409. S。Batalis，“AI和Biorisk: 解释者” (CSET，2023); https://cset.georgetown.edu/publication/ ai和biorisk-an-
解释者/。

410. G.刘易斯，P。米利特，A。桑德伯格，A.斯奈德-比蒂，G.格隆沃尔，生物技术中的信息危害。*风险肛门。* **39** ，975-981 (2019)。https://doi.org/10.1111/risa.13235。

411. C.A.穆顿，C.卢卡斯，E。Gues t，“人工智能在大规模生物攻击中的操作风险: 红队研究的结果” (兰德公司，2024);
https://www.rand.org/pubs/ 研究_报告/RRA2977-2.html。

412. S。Ben ouagrham-gormley，《*障碍生物武器: 挑战专业知识和组织武器发展*》(康奈尔大学出版社，2014)，pp。220。

413. J.Revill, C.杰斐逊，隐性知识和生物武器制度。*Sci*。*公共政策*的**41**，597-610 (2014)。https://doi.org/10.1093/scipol/sct090。

414. E。美国国家科学院和医学院，《*生物防御在合成生物学*时代》(美国国家科学院出版社，华盛顿特区，美国，2018)，pp。188。

415. A.M.布兰，S.考克斯，O。席尔特，C.巴尔达萨里，A。白色，P。Schwaller，*第37届神经信息处理系统会议 (NeurIPS 2023) 人工智能科学研讨会*上的“用化学工具增强大型语言模型”(2023)。https://openreview.net/forum？id= wdGIL6lx3l。

416. K。H. Sumida, R.努涅斯-佛朗哥，我.Kalvet, S.J.Pellock，B。I.M。Wicky, L.F.米尔斯，J。Dauparas, J.王，Y。基普尼斯，N。詹姆森，A。康，J.德拉克鲁兹，B.Sankaran, A.K。贝拉，G.Jiménez-Osés, D.贝克，用ProteinMPNN改善蛋白质表达、稳定性和功能。*J.上午。化学。Soc*。**146**，2054-2061 (2024)。https://doi.org/10.1021/jacs.3c10941。

417. Z。吴，S。B.J.Kan, R.D.刘易斯，B。J.维特曼，F。H·阿诺德，用组合文库进行机器学习辅助的定向蛋白质进化。*Proc.Natl.Acad*。*Sci*。*美国*。*A*. 8852-8858 **116** (2019)。https://doi.org/10.1073/pnas.1901979116。

418. A.H.-W。是的，C。诺恩，Y。基普尼斯，D。Tischer, S.J.Pellock, D.埃文斯，P。Ma, G.R。李，J.Z。张，我。阿尼先科，B.考文垂，L.曹，J。Dauparas, S.Halabiya, M.德威特，L.卡特，K.N.胡克，D.贝克，使用深度学习从头设计荧光素酶。*自然***614**，774-780 (2023)。https://doi.org/10.1038/s41586-023-05696-3。

419. J.L.沃森，D.Juergens, N.R。班尼特，B.L.特里普，J.Yim, H. E.艾森纳赫，W。Ahern, A.J.博斯特，R。J.Ragotte, L.F.米尔斯，B。I.M。威奇，N.Hanikel, S.J.Pellock, A.库尔贝，W。谢弗勒，J.王，P。Venkatesh，我.Sappington, S.V。托雷斯。。。D. Baker，蛋白质结构和功能的从头设计与RFdiffusion。*自然***620**，1089-1100 (2023)。https://doi.org/10.1038/s41586-023-06415-8。

420. T.布拉斯基，J。Arus-pous，H.陈，C。Margreitter, C.Tyrchan, O.恩克维斯特，K。帕帕多普洛斯，A.帕特罗诺夫，重塑2.0: 一种用于从头药物设计的人工智能工具。*J.化学*。*Inf*。*模型*。**60**，5918-5922 (2020)。https://doi.org/10.1021/acs.jcim.0c00915。

421. N.N. Thadani, S.Gurev, P.Notin, N.优素福，N。J.罗林斯，D。Ritter, C.桑德，Y。Gal, D.S。马克，从大流行前的数据中学习，预测病毒逃逸。*自然***622**，818-825 (2023)。https://doi.org/10.1038/s41586-023-06617-0。

422. F.乌尔维纳，F.L entzos，C。Invernizzi, S.Ekins，人工智能驱动的药物发现的双重用途。*Nat Mach Intell*，**4** 189-191 (2022)。https://doi.org/10.1038/s42256-022-00465-9。

423. * A.埃尔纳加尔，H. Essam，W.Salah-Eldin, W.穆斯塔法，M.Elkerdawy，C。罗切罗，B。Rost，Ankh: 优化的蛋白质语言模型解锁通用建模，arXiv:2301.06568 [cs.LG] (2023)。https://doi.org/10.48550/arXiv.2301.06568。

424. T.Inagaki, A.加藤，K.高桥，H.尾崎，G。N.Kanda，LLMs可以从生物实验室自动化中的面向目标的指令生成机器人脚本，arXiv:2304.10267 [q-bio.QM] (2023)。http://arxiv.org/abs/ 2304.10267。

425. J.N.阿科斯塔，G.J.Falcone, P.Rajpurkar, E.J.Topol，多模式生物医学人工智能。*Nat*。*Med.***28**，177 3 -1784 (2022)。https://doi.org/10.1038/s41591-022-01981-2。

426. 摩尔，O。班纳吉，Z。S。H.阿巴德，h.m.克鲁姆霍尔茨，J.莱斯科夫，ec。J.白杨，P.Rajpurkar，通用医学人工智能的基础模型。*自然***616**，259-265 (2023)。https://doi.org/10.1038/s41586-023-05881-4。

427. * 310.ai，GenAI + BIO: 大自然没有时间，我们有gpu (2024)。https:// 310.ai/.

428. 国家安全委员会新兴生物技术 (NSCEB)，“AIxBio的风险”(NSCEB，2024)；https://www.biotech.senate.gov/wp内容/上传/2024/01/NSCEB_AIxBio_WP3_Risks.pdf。

429. C.A.穆顿，C.卢卡斯，E。Guest，“大规模生物攻击中AI的操作风险: 红队方法”(RAND Corporation，2023)；https://www.rand.org/pubs/ 研究_报告/RRA2977-1.html。

430. I.D. Raji，我.E。库马尔，A。霍洛维茨，A.Selbst，“AI功能的谬误”，*载于2022 ACM会议关于公平，问责制和透明度 (FAccT '22)* 的论文集 (计算机协会，2022) pp。959-972。https://doi.org/10.1145/3531146.3533158。

431. A.马伦，A.Asai, V.钟，R。Das, D.Khashabi，H. Hajishirzi，“何时不信任语言模型: 研究参数和非参数记忆的有效性”，*载于第61届计算语言学协会 (第1卷: 长篇论文)* 的论文集 (计算语言学协会，2023) pp。9802-9822。https://doi.org/10.18653/v1/2023.acl-long.546。

432. A.Perlman，ChatGPT对法律服务和社会*的实践*意义。(哈佛法学院法律专业中心，2023)。https://clp.law.harvard.edu /知识中心/杂志/问题/generate-ai-in-the-法律职业/聊天对法律服务和社会的影响/。

433. E。马丁内斯，重新评估GPT-4的律师考试成绩。《*人工智能与法律*》(2024)。https://doi.org/10.1007/s10506-024-09396-9。

434. J.谭，H.韦斯特曼，K。Benyekhlef，“作为人工律师的聊天吗？”在*人工智能诉诸司法研讨会 (AI4AJ 2023)* 中 (CEUR研讨会程序，2023)。https:// ceur-ws.org/Vol-3435/short2.pdf。

435. J.A.奥米耶，J。C.莱斯特，S.Spichak, V.罗滕贝格，R。Daneshjou，大型语言模型传播基于种族的医学。*npj数字。医学。***6**，1-4 (2023)。https://doi.org/10.1038/s41746-023-00939-z。

436. K。辛哈尔，S。阿齐兹，T。Tu, S.S。Mahdavi, J.魏，H. W.Chung, N.天平，A.坦瓦尼，H.科尔-刘易斯，S.Pfohl, P.佩恩，M. Seneviratne, P.赌博，C。凯莉，A.Babiker, N.Schärli, A.Chowdhery，P。曼斯菲尔德，D。Demner-fushman，。。。V. Natarajan，对临床知识进行编码的大型语言模型。*自然***620**，172-180 (2023)。https://doi.org/10.1038/s41586-023-06291-2。

437. T.H. Kung, M.Cheatham, A.Medenilla，C。西洛斯，L.德莱昂，C.Elepaño, M.Madriaga, R.Aggabao, G.迪亚兹-坎迪多，J.Maningo, V.Tseng，ChatGPT在USMLE上的表现: 使用大型语言模型进行人工智能辅助医学教育的潜力。**2** *PLOS数字健康*，e0000198 (2023)。https://doi.org/10.1371/期刊.Pdig.0000198。

438. A.Ettinger，什么伯特不是: 从一套新的心理语言诊断语言模型的经验教训。*Trans。Assoc。计算机。语言学家。***8**，34-48 (2020)。https://doi.org/10.1162/tacl_a_00298。

439. Y。张，M。Yasunaga，Z。周，J。Z。郝晨，J。邹，P。梁，S。Yeung，"超越正缩放: 否定如何影响语言模型的缩放趋势"，*见计算语言学协会: ACL 2023* (计算语言学协会，2023) pp。7479-7498。https://doi.org/10.18653/v1/2023.findings-acl.472。

440. * 米。陈，J。特里克，H.6月，Q。袁，H·P·德·奥利维拉·平托，J。卡普兰，H.爱德华兹，Y.Burda, N.约瑟夫，G。布罗克曼，A.雷，R.普里，G.克鲁格，M。彼得罗夫，H. Khlaaf，G.萨斯特里，P。米什金，B.陈，S.格雷，。。。W. Zaremba，评估在代码上训练的大型语言模型，arXiv:2107.03374 [cs.LG] (2021)。http://arxiv.org/abs/ 2107.03374.

441. C.E.希门尼斯，J.杨，A。Wettig, S.姚，K。裴O.按，K。R。Narasimhan，"SWE-bench: 语言模型可以解决现实世界的Github问题吗？"*在第12届国际学习表示会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= VTF8yNQM66。

442. R。潘，A.R。Ibrahimzada, R.克里希纳，D.桑卡，L。P。瓦西，M。梅勒，B.索博列夫，R.Pavuluri, S.辛哈，R。Jabbarvand，"翻译中的迷失: 对大型语言模型在翻译代码时引入的错误的研究"，*载于IEEE/ACM第46届国际软件工程会议 (ICSE '24)* 论文集 (计算机协会，2024) pp。1-13。https://doi.org/10.1145/3597503.3639226。

443. F.卡萨诺，L。李，A。塞西，N。Shinn, A.布伦南-琼斯，J。Ginesin，E。伯曼，G.Chakhnashvili, A.洛日科夫，C。J.安德森，A。古哈，可以编辑吗？评估大型语言模型遵循代码编辑指令的能力，arXiv:2312.12450 [cs.SE] (2023)。http://arxiv.org/abs/ 2312.12450。

444. S。阮，h.m.宝贝，Y.紫，A.古哈，C。J.安德森，M。问:Feldman，"初学者程序员和代码llm (Mis) 如何相互阅读"，*《计算机系统人为因素论文集》(chi'24)*，(计算机协会，2024) pp。1-26。https://doi.org/10.1145/3613904.3642706。

445. P。达克，T.Dhar, C.B.我们inberg，X.Zeng，真相，谎言和广告: 了解虚假广告声明的市场和个人原因。(2022)。https://doi.org/10.2139/ssrn.4140673。

446. M. Rauh, J.F.J.梅勒，J。Uesato，P.-S.黄，J。Welbl, L.魏丁格，S.达斯里，A。Glaese, G.欧文，我。加布里埃尔，W。艾萨克，L.A.Hendricks，"有害文本的特征: 对语言模型进行严格的基准测试"，在*第36届神经信息处理系统会议 (NeurIPS 2022)* 上 (2022)。https://openreview.net/forum？id= u46CbCaLufp。

447. O.Bohdal, T.Hospedales，P。H. S.托尔，F。Barez，"AI的公平性及其对社会的长期影响"，*交叉点，增援，级联: 2023斯坦福存在风险会议*论文集 (斯坦福存在风险倡议，2023) pp。171-186。https://doi.org/10.25740/pj287ht2654。

448. D.Acemodlu，"人工智能的危害"，*《牛津人工智能治理手册》*，J。B.布洛克，Y.-C。陈，J。Himmelreich, V. M.哈德森，A.Korinek, M.M。杨，B.张，编辑。(牛津大学出版社，2023)。

449. I.*《量化工人: 法律和技术在现代工作场所》* (剑桥大学出版社，剑桥，2023)。

450. Z。奥伯迈尔，B。鲍尔斯，C。Vogeli, S.Mullai nathan，剖析了用于管理人群健康的算法中的种族偏见。*科学***366**，447-453 (2019)。https://doi.org/10.1126/科学aax2342。

451. A.王，E。奥特尔斯，J.P。唐纳利，A。克鲁姆，J.麦卡洛，O。底特律-库利，J。佩斯特鲁，M。菲利普斯，J。Konye，C。佩诺萨，M。Ghous, K.Singh，住院患者广泛实施的专有脓毒症预测模型的外部验证。*JAMA实习生*医学**181**，1065-1070 (2021)。https://doi.org/10.1001/jamainternmed.2021.2626。

452. J.Buolamwini, T.Gebru，"性别阴影: 商业性别分类中的交叉准确性差异"，*载于第一届公平，问责制和透明度会议论文集 (FAT/mm'19)* 中 (PMLR，2018) pp。77-91。https:// 会议记录.mlr.press/v81/buolamwini18a.html.

453. J.Dressel，H. Farid，"预测累犯的准确性、公平性和局限性"。**4** *Sci高级*，eaao5580 (2018)。https://doi.org/10.1126/科学。aao5580。

454. M. A.Gianfrancesco, S.Tamang, J.亚兹达尼，G。Schmajuk，使用电子健康记录数据的机器学习算法的潜在偏见。*JAMA实习生.医学。***178**，1544-1547 (2018)。https://doi.org/10.1001/jamainterned.2018.3763。

455. Y。万，G。Pu, J.太阳，A。加里梅拉，K.-W。Chang, N.彭，“‘凯利是一个热情的人，约瑟夫是一个榜样’: LLM生成的推荐信中的性别偏见”，*见计算语言学协会: EMNLP 2023* (计算语言学协会，2023) pp。3730-3748。https://doi.org/10.18653/v1/2023.findings-emnlp.243。

456. D.van Niekerk, M.佩雷斯-奥尔蒂斯，J.Shawe-Taylor, D.Orlič, I.Drobnjak, J.凯，N.西格尔，K。埃文斯，N。Moorosi, T.Eliassi- Rad, L.M。Tanczer, W.福尔摩斯，M。P。Deisenroth，我.稻草，M.Fasli, R.亚当斯，N.奥利弗，D.Mladeni ć，U。Aneja, .。。M. Janicky，“挑战系统性偏见: 调查大型语言模型中对妇女和女童的偏见”(教科文组织，IRCAI，2024)；https://ircai.org/project/ 挑战性-系统性-偏见/。

457. M. Vlasceanu, D.M。Amodio，通过互联网搜索算法传播社会性别不平等。**119***美国国家科学院院刊*，e2204529119 (2022)。https://doi.org/10.1073/pnas.2204529119。

458. 美国平等就业机会委员会，EEOC起诉iTutorGroup年龄歧视 (2022)。https://www.eeoc.gov/newsroom/ eeoc-起诉-itutorgroup-年龄歧视。

459. M. Díaz, I.约翰逊，A。拉扎尔，A.M。派珀，D.Gergle，“解决情感分析中与年龄相关的偏见”，*载于第28届国际人工智能联合会议 (IJCAI-19)* 论文集 (国际人工智能联合会议组织，2019) pp。6146-6150。https://doi.org/10.24963/ijcai.2019/852。

460. J.Stypinska，AI ag eism: 研究数字化社会中年龄歧视和排斥的关键路线图。*AI Soc*。**38**，665-677 (2023)。https://doi.org/10.1007/s00146-022-01553-5。

461. P。Rucker, M.米勒· D.阿姆斯特朗，信诺如何通过让医生拒绝索赔而无需在*ProPublica*中阅读来节省数百万美元。(2023)。https://www.propublica.org/article/ 信诺-pxdx-医疗-健康-保险-拒付-理赔

462. A.麦克，R.Qadri, R.丹顿，S.K。凯恩，C.L.Bennett，“他们只关心向我们展示轮椅”: “从文本到图像的AI模型中的残疾表示”，*载于CHI会议关于计算系统中人为因素 (CHI '24)* 的论文集 (ACM，2024)。https://dl.acm.org/doi/10.1145/3613904.3642166。

463. P。N. Venkit, M.Srinath, S.威尔逊，“自动能力主义: 情感和毒性分析模型中明显的残疾偏见的探索”，*载于第三届可信赖自然语言处理研讨会 (TrustNLP 2023)* 论文集 (计算语言学协会，2023) pp。26-34。https://doi.org/10.18653/v1/2023.trustnlp-1.3。

464. I.A. Adeyanju, O.O.Bello, M.A.Adegboye，手语识别的机器学习方法: 批判性回顾和分析。*智能系统及其应用*200056 **12** (2021)。https://doi.org/10.1016/j.Iswa.2021.200056。

465. S。Gueuwou, K.Takyi, M.Müller, M.S。Nyarko, R.Adade，R.-M.O.M。Gyening，“AfriSign: 非洲手语的机器翻译”在*第四届非洲自然语言处理研讨会 (AfricaNLP 2023)* 上 (2023)。https://openreview.net/forum？id= EHldk3J2xk。

466. W。郭，A。Caliskan，“检测紧急交叉偏见: 情境化的单词嵌入包含类似人类偏见的分布”，*在2 021 AAAI/ACM会议关于AI，道德和社会 (AIES '21)* 的论文集 (计算机协会，2021) pp。122-133。https://doi.org/10.1145/3461702.3462536。

467. 阿。A.卡布雷拉，W.Eppson，F。霍曼，M。卡恩，J。摩根斯坦，D.H. Chau，“FAIRVIS: 用于发现机器学习中的交叉点偏差的视觉分析”，在*2019 IEEE视觉分析科学与技术会议 (VAST)* 中 (2019) pp。46-56。https://doi.org/10.1109/vast47406.2019.8986948。

468. L.Magee, L.Ghahrem anlou, K.Soldatic, S.罗伯逊，因果语言模型中的交叉偏差，arXiv:2107.07691 [cs.CL] (2021)。https://doi.org/10.48550/arXiv.2107.07691。

469. A.Ovalle, A.Subramonian, V.高塔姆，G.哎呀，K.-W。Chang，“分解支配矩阵: 对AI公平性的交叉性的批判性回顾和重新想象”，*在2 023 AAAI/ACM AI，道德与社会会议 (AIES '23)* 上 (计算机协会，2023) pp。496-511。https://doi.org/10.1145/3600211.3604705。

470. J.S、公园，M。S。伯恩斯坦，R。N.布鲁尔，E。卡马尔，M。R。Morris，“理解AI数据集中年龄的表示和代表性”，*《2021 AAAI/ACM会议关于AI，道德和社会 (AIES '21)* 的论文集》(计算机协会，2021) pp。834-842。https://doi.org/10.1145/3461702.3462590。

471. R。Kamikubo, L.王，C。Marte, A.Mahmood，H. Kacorri，“可访问性数据集中的数据代表性: 荟萃分析”，*载于第24届国际ACM SIGACCESS会议计算机和可访问性 (ASSETS '22)*，(计算机协会，2022) pp。1-15。https://doi.org/10.1145/3517428.3544826。

472. * L。魏丁格，J.梅勒，M。Rauh, C.格里芬，J.Uesato，P.-S.黄，M。程，M。Glaese, B.Balle, A.Kasirzadeh, Z.肯顿，S.布朗，W.霍金斯，T。Stepleton，C。比尔斯，A.Birhane, J.哈斯，我。Rimell, L.A.亨德里克斯。。。I. Gabriel，“语言模型伤害的道德和社会风险”(Google DeepMind，2021)；http://arxiv.org/abs/ 2112.04359。

473. J.Nwatu, O.伊格纳特，R.Mihalcea，“弥合数字鸿沟: 视觉语言模型中跨社会经济因素的性能变化”，*《2023自然语言处理经验方法会议论文集》(EMNLP 2023)* (计算语言学协会，2023) pp。10686-10702。https://doi.org/10.18653/v1/2023.emnlp-main.660。

474. S。戈什，A.Caliskan，“ChatGPT使机器翻译中的性别偏见永久化，而忽略了非性别代词: 孟加拉语和其他五种低资源语言的发现”(*《2023 AAAI/ACM AI，道德与社会会议论文集》(AIES '23)* 2023) pp。901-912。https://doi.org/10.1145/3600211.3604672。

475. G.Vardi，关于深度学习算法中的隐式偏差。*Commun。ACM* **66**，86-93 (2023)。https://doi.org/10.1145/3571070。

476. F.比安奇，P。Kalluri, E.杜尔穆斯，F.Ladhak, M.程，D。Nozza, T.桥本，D.Jurafsky, J.邹，A。Caliskan，"易于访问的文本到图像的生成会大规模放大人口定型观念"，在*《2023 ACM Confere nce关于公平，问责制和透明度的论文集》(FAccT '23)* 中(计算机协会，2023) pp。1493-1504。https://doi.org/10.1145/3593013.3594095。

477. J.哈特曼，J。Schwenzow, M.Witte，《对话式AI的政治意识形态: 关于ChatGPT的亲环境，左翼自由主义取向的融合证据》，arXiv:2301.01768 [cs.CL] (2023)。http://arxiv.org/abs/ 2301.01768。

478. T.考夫曼，S.球，J。贝克，E。Hüllermeier, F.Kreuter，"从真实的人类反馈中强化学习的挑战和实践"，在*第一个研讨会上混合人机学习和决策 (hldm'23)* 中 (2023)。

479. A.Siththaranjan, C.Laidlaw, D.Hadfield-menell，"分布偏好学习: 对RLHF中隐藏上下文的理解和解释"，*在第12届国际学习表征会议 (ICLR)* 中 (2024)。https://openreview.net/forum？id=0tWTxYYPnW.

480. * OpenAI，ChatGPT插件 (2023)。https://openai.com/blog/ chatgpt-插件。

481. * I.加布里埃尔，A.曼齐尼，G.基林，L.A.亨德里克斯，V。Rieser, H. Iqbal, N.Tomašev, I.Ktena, Z.肯顿，M.罗德里格斯，S. El- Sayed, S.布朗，C。Akbulut, A.Trask, E.休斯，A.史蒂夫·B·艾格曼，R。谢尔比，N.Marchal，C。格里芬。。。J. Manyika，"高级人工智能助手的伦理"(谷歌DeepMind，2024)；http://arxiv.org/abs/ 2404.16244。

482. A.陈，R。Salganik, A.Markelius，C。庞，N。Rajkumar, D.Krasheninnikov, L.Langosco，Z。他，Y。段，M.卡罗尔，M。林，A。梅休，K。柯林斯，M。Molamohammadi, J.负担，W。赵，S.里斯马尼，K。Voudoulis，美国。巴特，。。。T. Maharaj，"越来越多的代理算法系统带来的危害"，*《2023 ACM会议公平，问责制，和透明度 (FAccT '23)》* (计算机协会，2023) pp。651-666。https://doi.org/10.1145/3593013.3594033。

483. A.M、图灵，智能机械，异端理论 *。*Philos.数学。***4**，256-260 (1996)。https://doi.org/10.1093/菲尔马特/4.3.256。

484. I.J. Good，"关于第一台超智能机器的推测"，*《计算机进步》*，F。L.Alt, M.Rubinoff，Eds。(爱思唯尔，1966)，卷。6，pp。31-88。

485. N.维纳，自动化的一些道德和技术后果。*科学***131**，1355-1358 (1960)。https://doi.org/10.1126/科学。131.3410.1355。

486. 人工智能安全中心，关于人工智能风险的声明: 人工智能专家和公众人物表达了他们对人工智能风险的担忧(2024)。https:// www.safe.ai/工作/声明-on-ai-风险。

487. Y。Bengio，Yoshua Bengio教授在美国参议院AI Insight论坛上的书面声明。(2023)。https://www.schumer.senate.gov/imo/media/doc/Yoshua%20Benigo% 20-%20Statement.pdf.

488. K。格蕾丝，H.斯图尔特，J.F.Sandkühle r, S.托马斯，B。温斯坦-劳恩，J。Brauner，成千上万的AI作者对AI的未来，arXiv:2401.02843 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2401.02843。

489. A.潘，K.巴蒂亚，J.斯坦哈特，"奖励错误规范的影响: 映射和减轻错位模型"，*在第十届国际学习表征会议 (ICLR 2022)* 中 (2021)。https://openreview.net/forum？id= JYtwGwIL7ye。

490. R。非政府组织，L.陈，S.Mindermann，"深度学习视角下的对齐问题"*在第12届国际学习表征会议 (ICLR 2024)* 中(2023)。https://openreview.net/forum？id= fh8eykfkts。

491. J.吉，T.邱，B。陈，B。张，H. Lou，K。王，Y。段，Z。他，J.周，Z。张，F。曾，K。Y。Ng，J。戴，X。潘，A.O'Gara, Y.雷，徐，B。谢先生，J。傅，。。。高文，人工智能对齐: 综合调查，arXiv:2310.19852 [cs.AI] (2023)。https://doi.org/10.48550/arXiv.2310.19852。

492. D.亨德里克斯，X。刘，E。华莱士，A.Dziedzic, R.克里希南，D.Song，"预训练的变压器提高了分布外的鲁棒性"，*载于第58届计算语言学协会 (ACL 2020)* 会议记录 (计算语言学协会，2020) pp。2744-2751。https://doi.org/10.18653/v1/2020.acl-main.244。

493. M. K. Cohen, M.哈特，M。A.奥斯本，先进的人工代理介入提供奖励。*AI Mag.***43**，282-293 (2022)。https://doi.org/10.1002/aaai.12064。

494. T.埃弗里特，M。哈特，R.库马尔，V。Krakovna，强化学习中的奖励篡改问题和解决方案: 因果影响图视角。*合成* **198**，6435-6467 (2021)。https://doi.org/10.1007/s11229-021-03141-4。

495. * R。库马尔，J。乌萨托，R。非政府组织，T.埃弗里特，V。克拉科夫纳，S.莱格，REALab: 关于篡改的嵌入式观点，arXiv:2011.08820 [cs.LG] (2020)。http://arxiv.org/abs/ 2011.08820。

496. D.哈德菲尔德-梅内尔，A.D.德拉根，P。Abbeel, S.Russell，"关闭游戏"*在第31届AAAI会议人工智能* (2017)。http://aaai.org/ocs/index.php/WS/ AAAIW17/paper/view/15156的。

497. * V.克拉科夫纳，J.Kramar，对于受过训练的特工来说，寻求力量是可能的和可预测的，arXiv:2304.06528 [cs.AI] (2023)。https://doi.org/10.48550/arXiv.2304.06528 。

498. A.特纳，L.史密斯，R。Shah, A.克里奇，P。Tadepalli，“最优策略倾向于寻求权力”，在*第35届会议上的神经信息处理系统 (NeurIPS 2021)* 中 (Curran Associates，inc.，2021) 卷。34。https:// 会议录.neurips.cc/paper/2021/hash/c26820b8a4c1b3c2aa868d6d57e14a79-Abstract.html。

499. A.特纳，P。Tadepalli，*《神经信息处理系统35 (NeurIPS 2022) 主要会议轨道》* (2022) 第一卷中的“可参数重定向的决策者倾向于寻求权力”。abs/2206.13477。https://doi.org/10.48550/arXiv.2206.13477。

500. J.D. Gallow，工具发散。*Philos.螺柱。*(2024)。https://doi.org/10.1007/s11098-024-02129-3。

501. J.Scheurer, M.Balesni, M.Hobbhahn，*ICLR 2024研讨会关于大型语言模型 (LLM) 代理*的“大型语言模型可以在压力下战略性地欺骗其用户”(2024)。https://doi.org/10.48550/arXiv.2311.07590。

502. E。Brynjolfsson，图灵陷阱: 人类人工智能的承诺和危险。*代达罗斯***151** ，272-287 (2022)。https://doi.org/10.1162/daed_a_01915。

503. 库济姆斯基，G。Misuraca，公共部门的人工智能治理: 三个来自民主环境中自动决策前沿的故事。*电信。***44***政策*，101976 (2020)。https://doi.org/10.1016/j.Telpol.2020.101976。

504. L.米特鲁，M.詹森，E。Loukis，“人工智能驱动的政府决策中的人类控制和自由裁量权”，见*第14届国际会议的理论和实践的电子治理 (ICEGOV 2021)* ，(ACM，2021)。https://doi.org/10.1145/3494193.3494195。

505. I.泰勒，“算法正义: 人工智能在刑事量刑中的局限性”。*Crim.司法伦理*，**42** ，193-213 (2023)。https://doi.org/10.1080/0731129 x.2023.2275967。

506. J.-P。里维拉，G。Mukobi, A.Reuel, M.兰帕斯，C。史密斯，J。Sch neider，军事和外交决策中语言模型的升级风险，arXiv:2401.03408 [cs.AI] (2024)。http://arxiv.org/abs/ 2401.03408。

507. D.亨德里克斯，M。Mazeika, T.伍德赛德，“灾难性人工智能风险概述”，arXiv:2306.12001 [cs.CY] (2023)。https://doi.org/10.48550/arXiv.2306.12001 。

508. * T。Shevlane, S.法夸尔，B.加芬克尔，M。Phuong, J.惠特尔斯通，J。梁，D。Kokotajlo, N.Marchal, M.Anderljung, N.Kolt, L.浩，D。Siddarth, S.阿文，W。kins，B。Kim，我.加布里埃尔，V。Bolina, J.克拉克，Y.Bengio，。。。A. Dafoe，“极端风险的模型评估”(Google DeepMind，2023); http://arxiv.org/abs/ 2305。15324。

509. 金尼曼特，L。J.K。佐藤，H. Du，B.古德里奇，M。哈辛，L.陈，L.H.迈尔斯，T.R。林，H. Wijk，J。Burget, A.Ho, E.巴恩斯，P。Christiano，评估现实自主任务上的语言模型代理，arXiv:2312.11671 [cs.CL] (2023)。http://arxiv.org/abs/ 2312.11671。

510. P。J·丹宁，“计算科学: 互联网蠕虫”。*Am。Sci。***77** ，126-128 (1989)。http://www.jstor.org/stable/ 27855650。

511. A.克里奇，S.罗素，塔斯拉: 人工智能社会规模风险的分类学和分析，arXiv:2306.06924 [cs.AI] (2023)。http://arxiv.org/abs/ 2306.06924。

512. G.马库斯，“深度学习: 批判性评价”，arXiv:1801.00631 [cs.AI] (2018)。https://doi.org/10.48550/arXiv.1801.00631。

513. D.阿西莫格鲁，D.作者，“技能、任务和技术: 对就业和收入的影响 *”，载于《劳动经济学*手册》*，D。卡，O.Ashenfelter，编辑。(爱思唯尔，2011)，卷。4，pp。1043-1171。

514. D.H. Autor，劳动力市场的“任务方法”: 概述。*J.劳工标志。Res。***46** ，185-199 (2013)。https://doi.org/10.1007/s12651-013-0128-z。

515. D.阿西米格鲁，P。Restrepo，自动化和新任务: 技术如何取代和恢复劳动力。*J.经济。透视。***444** ，3-30 (2019)。https://doi.org/10.1257/jep.33.2.3。

516. D.Autor，“应用人工智能重建中产阶级工作”(国家经济研究局，2024); https://doi.org/10.3386/w32140。

517. D.H. Autor，为什么还有这么多工作？工作场所自动化的历史和未来。*J.经济。透视。***29** ，3-30 (2015)。https://doi.org/10.1257/jep.29.3.3。

518. A.Georgieff, R.Hyee，人工智能与就业: 新的跨国证据。*Front Artif Intell* ，832736 **5** (2022)。https://doi.org/10.3389/frai.2022.832736。

519. M. Cazzaniga, F.Jaumotte, L.李，G。梅丽娜，A.J.潘顿，C.Pizzinelli, E.J.Rockall, M.M。Tavares，“gen-ai: 人工智能和工作的未来”(国际货币基金组织，2024); https://www.imf.org/en/Publications/Staff-讨论-说明/问题/2024/01/14/Gen-AI-Artificial-Intelligence-and-the-Future-of-Work-542379。

520. F.Dell'Acqua, E.McFowland, III, E.R。莫利克，H.Lifshitz-assaf，K.凯洛格，S。Rajendran, L.Krayer, F.坎德隆，K。R。Lakhani，“穿越锯齿状的技术前沿: 人工智能对知识影响的现场实验证据”

工人生产率和质量 "(哈佛商学院，2023); https://www.hbs.edu/ris/Publication%20Files/ 24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf。

521. * S.彭，E。Kalliamvakou，P。西霍 ，n，M。Demirer，AI对开发人员生产力的影响: 来自GitHub Copilot的证据，arXiv:2302.06590 [cs.SE] (2023)。https://doi.org/10.48550/arXiv.2302.06590。

522. E。D. Yilmaz, I.不适用乌莫夫斯卡，V。A.Aggarwal，"AI驱动的劳动力替代: 来自Google Translate和ChatGPT的证据" (INSEAD，2023)；https://sites.insead.edu/facultyresearch/research/doc.cfm？did= 70468。

523. X。Hui, O.雷塞夫，L.周，"短期影响关于就业的生成人工智能: 来自在线劳动力市场的证据 "(CESifo工作文件，2023)；https:// www.econstor.eu/handle/10419/279352。

524. L.Belzner, T.Gabor, M.Wirsing，"大型语言模型辅助软件工程: 前景，挑战和Ca se研究"，*弥合AI与现实 (AISoLA 2023)* (Springer Nature Switzerland，2024) pp。355-374。https://doi.org/10.1007/978-3-031-46002-9_23。

525. J.贝森，"自动化与就业: 当技术促进就业" (波士顿大学法学院，2019)；https://scholarship.law.bu.edu /学院_奖学金/815。

526. D.奥托，C。Chin, A.Salomons, B.Seegmiller，"新领域: 新工作的起源和内容，1940-2018" (国家经济研究局，2022)；https://doi.org/10.3386/w30389。

527. A.阿格拉沃尔，J.S。甘斯，A.Goldfarb，人工智能: 自动化预测对劳动力市场的模糊影响。*J.经济。透视。***444** ，31-50 (2019)。https://doi.org/10.1257/jep.33.2.31。

528. B.莫尔，L.瑞秋，P。雷斯特雷波，不均衡增长: 自动化对收入和财富不平等的影响。*计量经济学***90** ，2645-2683 (2022)。https://doi.org/10.3982/ecta19417。

529. M. Beraja, N.Zorzi，"低效自动化" (国家经济研究局，2022)；https://doi.org/10.3386/w30154。

530. D.阿西米格鲁，P。雷斯特雷波，错误的人工智能？人工智能与劳动力需求的未来 "(国家经济研究局，2019)；https://doi.org/10.3386/w25682。

531. S。G. Benzell, L.J.Kotlikoff, G.拉加尔da, J.D.萨克斯，"机器人是我们: 人类替代的一些经济学" (国家经济研究局，2015)；https://doi.org/10.3386/w20941。

532. A.Korinek, D.Suh，"向AGI过渡的情景" (国家经济研究局，2024)；https://doi.org/10.3386/w32255.

533. E。Ilzetzki, S.Jain，人工智能对VoxEU增长和就业的影响 -*CEPR的政策门户网站*(2023)。https://cepr.org/voxeu/columns/ 影响-人工智能-增长-和-就业。

534. M. N.贝利，E。Brynjolfsson, A.Korinek，"思维机器: 人工智能驱动的生产力繁荣" 的案例 (布鲁金斯，2023); https://www.brookings.edu/articles/ 机器的思维案例为人工智能驱动的生产力-吊杆/。

535. A.K.阿格拉沃尔，J.S。甘斯，A.Goldfarb，"图灵转型: 人工智能，智能增强和技能溢价" (国家经济研究局，2023)；https://doi.org/10.3386/w31767。

536. D.阿西米格鲁，P。Restrepo，《人与机器的竞赛: 技术对增长、要素份额和就业的影响》。*Am。经济。Rev。***108** ，1488-1542 (2018)。https://doi.org/10.1257/aer.20160696.

537. D.阿西米格鲁，P。Restrepo，"人工智能，自动化和工作" (国家经济研究局，2018)；https://doi.org/10.3386/w24196。

538. I.Cockburn, R.他nderson, S.斯特恩，"人工智能对创新的影响" (国家经济研究局，2018)；https://doi.org/10.3386/w24449。

539. W。诺德豪斯，我们正在接近经济危机吗？信息技术与经济增长的未来。*Am。经济。J、宏观经济。***13** ，299-332 (2021)。https://doi.org/10.1257/mac.20170105。

540. 经合组织，"在工作场所使用人工智能: 机遇、风险和政策应对" (经合组织，2024)；https://doi.org/10.1787/73d417f9-en。

541. D.阿西米格鲁，P。Restrepo、任务、自动化和美国的崛起工资不平等。*计量经济学***90** ，1973-2016 (2022)。https://doi.org/10.3982/ecta19815。

542. Ó。阿方索，R。Forte，常规和非常规部门，任务自动化和工资两极分化。*应用程序。经济。*(2023)。https://www.tandfonline.com/doi/abs/ 10.1080/00036846.2023.2280461。

543. D.Acemoglu, J.Loebbing，"自动化和两极分化" (国家经济研究局，2022)；https://doi.org/10.3386/w30528。

544. L.Karabarbounis，"劳动份额的观点" (国家经济研究局，2023)；https://doi.org/10.3386/w31854。

545. M. Ranaldi，收入构成不平等。*Rev.***68***收入财富* ，139-160 (2022)。https://doi.org/10.1111/roiw.12503。

546. T.皮凯蒂，A.Goldhammer，“ *资本在二十一世纪*”(哈佛大学出版社的Belknap出版社，马萨诸塞州剑桥市，2014)，pp。685。

547. A.Korinek, J.E。斯蒂格利茨，“Ar人工智能，全球化和经济发展战略”(国家经济研究局，2021)；https://doi.org/10.3386/w28453。

548. 进步通信协会，第19条，瑞典国际开发合作署，“全球信息社会观察2019: 人工智能: 人权、社会正义和发展”(亚太经社会，2019)。

549. N.Sambasivan, E.阿内森，B.哈钦森，T。多西，V。Prabhakaran，“在印度及其他地区重新构想算法公平性”，*载于2021 ACM会议公平，问责制和透明度 (FAccT '21)* 论文集 (计算机协会，2021) pp。315-328。https://doi.org/10.1145/3442188.3445896。

550. C.T. Okolo，《全球南方的人工智能: 更具包容性治理的机遇与挑战》(2023)。https://www.brookings.edu/articles/ 全球南方的人工智能机遇和挑战走向更具包容性-治理/。

551. M.-T。巴布亚新几内亚，“在南方和北方的紧张局势中: 全球南方利益相关者在人工智能治理中的关键作用”，*载于2022 ACM会议公平、问责和透明度 (FAccT '22)*，(计算机协会，2022) pp。1434-1445。https://doi.org/10.1145/3531146.3533200。

552. S。L.哈兰，D.N.佩洛，J.T.罗伯茨，S.E。贝尔，W.G.霍尔特，J.内格尔，“气候正义与不平等”，*气候变化与社会*，E。邓拉普，R。J.布鲁尔，编辑。(牛津大学出版社，2015)，第127-163。

553. F.苏丹，关键的气候正义。*Geogr。J.* 118-124 **188** (2022)。https://doi.org/10.1111/geoj.12417。

554. R。Zwetsloot，B。张，N。德雷克斯勒，L.卡恩，M.Anderljung, A.达福，M。C.Horowitz，“技术和移动: 人工智能研究人员移民偏好的调查证据”，*载于2 021 AAAI/ACM会议关于人工智能，道德和社会 (AIES '21)* 的论文集 (计算机协会，2021) pp。1050-1059。https://doi.org/10.1145/3461702.3462617。

555. R。Jurowetzki, D.Hain, J.Mateos-garcia，K。Stathoulopoulos，AI研究的私有化 (-ers): 原因和潜在后果-从大学与行业的互动到公共研究人才流失？，arXiv:2102.01648 [cs.CY] (2021)。http://arxiv.org/abs/ 2102.01648.

556. N.艾哈迈德，M。Wahed，AI的去民主化: 人工智能研究中的深度学习和计算鸿沟，arXiv:2010.15581 [cs.CY] (2020)。https://doi.org/10.48550/arXiv.2010.15581。

557. T.Besiroglu, S.A.Bergerson, A.迈克尔，L.海姆，X.罗，N。汤普森，机器学习中的计算鸿沟: 对学术贡献和审查的威胁？，arXiv:2401.02452 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2401.02452。

558. S。Teleanu, J.Kurbalija，“来自非洲的更强大的数字声音: 建立非洲数字外交政策和外交”(Diplo，2022)；https://www.diplomacy.edu/resource/ 报告-更强-数字-来自非洲的声音/。

559. 国际电信联盟 (ITU)，“ *衡量数字发展: 事实和数字2023*中的互联网使用”(ITU出版物，2023)，第1-2。

560. E。帕诺斯，M。密集，K。世界能源理事会全球能源情景中的电力获取: 2030年发展中地区的展望。**9***能源战略评论*，28-49 (2016)。https://doi.org/10.1016/j.Esr.2015.11.003。

561. H。Ritchie, P.罗萨多，M.罗瑟，*我们世界的数据*获取能源。(2024)。https://ourworldindata.org/energy-访问权限。

562. N.马斯莱，L.Fattorini, R.佩罗，V。帕里，A。Reuel, E.布林约尔松，J。Etchemendy, K.Ligett, T.里昂，J。Manyika, J.C.尼布尔斯，Y.Shoham, R.沃尔德，J.克拉克，“人工智能指数2024年度报告”(人工智能指数指导委员会，以人为中心的人工智能研究所，斯坦福大学，2024); https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf。

563. N.艾哈迈德，M。Wahed，N。C.汤普森，《工业在人工智能研究中日益增长的影响力》。*科学***379**，884-886 (2023)。https://doi.org/10.1126/科学.Ade2420。

564. * 米。Sukumaran, A.刘易斯，服务器市场分析-下半年 (2023)。https://omdia.tech.informa.com/om033795/服务器市场分析 -- 2h23。

565. M. L.格雷，S.苏里，《*幽灵工作: 如何阻止硅谷建立新的全球下层阶级*》(霍顿·米夫林·哈考特，2019)，pp。297。

566. J.邓，W。董，R。Socher，L.-J.李，K。李，L。Fei-fei，“ImageNet: 大型分层图像数据库”，*2009 IEEE计算机视觉和模式识别*会议 (2009) pp。248-255。https://doi.org/10.1109/cvpr.2009.5206848。

567. A.阿罗拉，M.巴雷特，E。李，E。Oborn, K.王子、风险和人工智能的未来: 算法偏见、数据殖民主义和边缘化。*Inf。风琴。***444** (2023)。https://doi.org/10.1016/j.Infoandorg.2023.100478。

568. C.T. Okolo，“解决人工智能发展中的全球不平等问题”，《*人工智能*批判研究手册》(爱德华·埃尔加出版社，2023)，第378-389。

569. D.王，S。Prabhat, N.Sambasivan，“谁的AI梦想？在“ *CHI会议上的人为因素在计算系统 (CHI '22)* 中寻找数据注释中的愿望 (ACM，2022) pp。1-16。 https://doi.org/10.1145/3491102.3502121。

570. M.米塞利，T.杨，A。阿尔瓦拉多-加西亚，J。波萨达，S.M。王，M。波尔，A.汉娜，记录数据生产的过程: 数据工作的参与式方法。*Proc.ACM嗡嗡声。计算机。互动。* **6**，1-34 (2022)。 https://doi.org/10.1145/3555623。

571. E。Brynjolfsson, A.Ng，“大人工智能可以集中决策和权力，这是人工智能治理*缺失环节*中的一个问题” (联合国教科文组织/Mila，2023)，pp。65-87。

572. R。Bommasani，D。A.哈德森，E。Adeli, R.奥特曼，S.阿罗拉，S.冯-阿克斯，M。S。伯恩斯坦，J。Bohg, A.Bosselut, E.Brunskill, E.Brynjolfsson, S.布赫，D.卡，R.卡斯特利翁，N。Chatterji, A.陈，K。克里尔，J.问:戴维斯，D。Demszky, .。。梁鹏，论基础模型的机遇与风险，arXiv:2108.07258 [cs.LG] (2021)。 https://doi.org/10.48550/arXiv.2108.07258。

573. J.Vipra, A.Korinek，“基础模型的市场集中度含义: ChatGPT的看不见的手” (布鲁金斯学会，2023)。

574. 竞争和市场管理局，“人工智能基础模型: 初始报告” (CMA，2023)； https://www.gov.uk/government/publications/ ai-基础-模型-初始-报告。

575. A.戈德法布，D。Trefler，“人工智能与国际贸易”，“*人工智能经济学: 议程*” (芝加哥大学出版社，2018)，第463-492。

576. R。奈克，B.Nushi，“通过文本到图像生成镜头的社会偏见”，*《2023 AAAI/ACM AI，道德与社会会议论文集》(AIES '23)*，(计算机协会，2023) pp。786-808。 https://doi.org/10.1145/3600211.3604711。

577. J.Daníelsson, R.麦克雷，A.Utemann、人工智能与系统性风险。*J.银行。财务。106290* **140** (2022)。 https://doi.org/10.1016/j.Jbankfin.2021.106290。

578. 国际能源署，“跟踪清洁能源进展2023” (IEA，2023)； https://www.iea.org/reports/ 追踪-清洁能源-进展-2023。

579. 欧盟委员会联合研究中心，G.神谷，P。Bertoldi，《*能源消耗在数据中心和宽带通信网络在欧盟*》 (欧洲联盟出版物办公室，2024)。

580. E。哈尔珀，在爆炸性的需求中，美国在《*华盛顿邮报*》上的权力正在耗尽。(2024)。 https://www.washingtonpost.com/business/ 2024/03/07/ai-数据中心-电力/。

581. * A.S。Luccioni, A.埃尔南德斯-加西亚，《计算碳: 影响机器学习排放的因素调查》，arXiv:2302.08476 [cs.LG] (2023)。 http://arxiv.org/abs/ 2302.08476。

582. P。李，J。杨，M。A.伊斯兰教，S.Ren，Maki ng AI不那么“口渴”: 发现和解决AI模型的秘密水足迹，arXiv:2304.03271 [cs.LG] (2023)。 http://arxiv.org/abs/ 2304.03271。

583. D.J·索洛夫，“ *理解隐私* ” (哈佛大学出版社，英国伦敦，编。首先，2009)，第页。257。

584. B.史密斯，负责任地开发和部署人工智能: 规范人工智能的有效立法框架的要素。(2023)。https://blogs.microsoft.com / 问题/2023/09/12/开发和部署人工智能-负责任地-要素- 一个有效的立法框架来监管人工智能/。

585. H。尼森鲍姆，“ *隐私背景: 技术、政策和社会生活*的完整性” (斯坦福大学出版社，加利福尼亚州帕洛阿尔托，2009)，第304。

586. L.Bourtoule, V.Chandrasekaran, C.A.乔奎特-乔，H.贾，A。特拉弗斯，B。张，D。谎言，N。*2021 IEEE安全与隐私 (SP)* 研讨会 (IEEE，2021) 中的“机器学习”141-159。 https://doi.org/10.1109/sp40001.2021.00019。

587. D.J. Solove，人工智能和隐私。《*佛罗里达法律评论*》 (2025)。 https://papers.ssrn.com/sol3/论文。cfm？abstract_id = 4713111。

588. R。Shokri, M.Stronati，C。宋，V。Shmatikov，“针对机器学习模型的成员推理攻击” *2017IEEE安全与隐私 (SP)* 研讨会 (IEEE，2017) pp。3-18。https://doi.org/10.1109/sp.2017.41。

589. M.弗雷德里克森，S.Jha, T.Ristenpart，“利用置信度信息和基本对策的模型反转攻击”，*载于第22届ACM SIGSAC会议计算机和通信安全会议 (ccs'15)* 论文集 (计算机协会，2015)。1322-1333。https://doi.org/10.1145/2810103.2813677。

590. G.陈，Y。张，F。Song，“slmia-sr: 针对说话人识别系统的说话人级别成员推断攻击”，*论文集2024网络和分布式系统安全研讨会 (NDSS 2024)* (互联网协会，2024)。https://doi.org/10.14722/ndss.2024.241323。

591. N.卡里尼，F.特拉梅尔，E.华莱士，M.Jagielski, A.赫伯特-沃斯，K.李，A.罗伯茨，T.布朗，D。宋，Ú。Erlingsson, A.Oprea, C.Raffel，“从大型语言模型中提取训练数据”，在*第30届USENIX安全研讨会 (USENIX安全21)* 中 (USENIX协会，2021) pp。2633-2650。 https://www.usenix.org/conference/ usenixsecurity21/presentation/carlini-提取。

592. N.卡里尼，J.海耶斯，M.纳斯尔，M.Jagielski, V.Sehwag, F.特拉梅尔，B.Balle, D.伊波利托，E.Wallace，“从扩散模型中提取训练数据”，在*第32届USENIX安全研讨会 (USENIX安全23)* 中 (USENIX协会，2023) pp。5253-5270。https://www.usenix.org/conference/ usenixsecurity23/演示文稿/carlini。

593. W。石，A。Ajith, M.夏，Y。黄，D。刘，T。Blevins, D.陈，L。Zettlemoyer，“从大型语言模型中检测预训练数据”，*在第12届国际学习表示会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= zWqr3MQuNs。

594. N.Lukas, A.塞勒姆，R.Sim, S.Tople, L.Wutschitz, S.Zanella-béguelin，“分析语言模型中个人身份信息的泄漏”，*2023 IEEE安全与隐私 (SP)* 研讨会 (IEEE，2023) pp。346-363。https://doi.org/10.1109/sp46215.2023.10179300。

595. N.Carlini, D.伊波利托，M.Jagielski, K.李，F。特拉梅尔，C.Zhang，“跨神经语言模型的量化记忆”*在第11届国际学习表征会议 (ICLR 2023)* 中 (2022)。https://openreview.net/forum？id= TatRHT_1cK。

596. * K。萨博，T.屠，W.-H。翁，R。坦诺，D.Stutz, E.Wulczyn, F.张，T。Strother, C.公园，E。Vedadi, J.Z。查韦斯，S.-Y。胡，M。Schaekermann, A.Kamath, Y.程，D。G.T.巴尔· 埃特，C。张，B。穆斯塔法，A.帕勒普。。。V. Natarajan，“双子座模型在医学中的能力”(谷歌Deepmind，2024)；http://arxiv.org/abs/ 2404.18416。

597. K。希尔，这家秘密公司可能会终止我们在“纽约时报”上所知*的隐私*。(2020)。https://www.nytimes.com/2020/01/18/技术/clearview-privacy-facial-recognition.html。

598. R。Staab, M.维罗，M.巴卢诺维奇，M.Vechev，“超越记忆: 通过使用大型语言模型进行推理来侵犯隐私”，*在第12届国际学习表征会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= kmn0BhQk7p。

599. M. Sharma, M.Kaur，“ *软计算安全应用*中的Deepfake技术回顾: 新兴的AI威胁”(Springer，新加坡，2022) pp。605-619。https://doi.org/10.1007/978-981-16-5301-8_44。

600. P。汉堡，伯尔尼公约: 它的历史和它在未来的关键作用。*J.L.& 技术*。**3**，1-70 (1988)。https://heinonline.org/HOL/LandingPage？handle=hein.journals/ jlawtecy3 & div = 4 & id = & page =。

601. L.帕特森，版权在1791: 一篇关于创始人对美国宪法第一条第8条第8款授予国会的版权权力的看法的文章。*埃默里l.j.* (2003)。https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle = hein.journals/emlj52 & section = 25.

602. 美国众议院法律修订顾问办公室，“对专有权的限制: 合理使用。秒。107“ *美国法典，2006版，补编4，标题17-版权”* (美国政府出版办公室，编辑。2010、2010)。

603. E。议会，D.-G。f.I. P. o.t.联盟，E。Rosati*的例外文本和数据挖掘 (TDM) 在拟议指令版权在数字单一市场-技术方面*，(欧洲议会，2018)。

604. 日本法律transa数据库系统，“著作権法(一部未施行) 版权法 (部分未执行)”(日本司法部，2024)；https:// www.japaneselawtranslation.go.jp/en/laws/view/ 4207。

605. 以色列司法部，意见: 使用受版权保护的材料进行机器学习。(2022)。https://www.gov.il/BlobFolder/legalinfo/机器学习/he/18-12-2022.pdf.

606. 新加坡知识产权局，“版权: 2021版权法概况介绍”(IPOS，2022)；https://www.ipos.gov.sg/docs/default-source/resources-library/copyright/copyright-act-factsheet.pdf

607. L.Soldaini, R.金尼，A.Bhagia, D.施文克，D.阿特金森，R。Authur, B.Bogin, K.Chandu, J.大仲马，Y.Elazar, V.霍夫曼，A.H. Jha, S.Kumar, L.露西，X.Lyu, N.兰伯特，我.马格努森，J。莫里森，N。Muennighoff，。。。K. Lo，Dolma: 一个开放的语料库，用于语言模型预训练研究，arXiv:2402.00159 [cs.CL] (2024)。http://arxiv.org/abs/ 2402.00159。

608. L.Tiedrich，AI数据抓取挑战: 我们如何负责任地进行？(2024)。https:// oecd.ai/en/wonk/data-scraping-负责任。

609. B.L. W.索贝尔，人工智能的合理使用危机。*JLA* **41**，4日- 97 (2018)。https://doi.org/10.7916/jla.v41i1.2036。

610. M. A.莱姆利，B.凯西，公平学习。*特克斯。法律Rev.***99**，743-786 (2020)。https://texaslawreview.org /公平-学习/。

611。P.萨缪尔森，生成性人工智能符合版权。**381***科学*，158- 161 (2023)。https://doi.org/10.1126/科学adi0656.

612. Tremblay诉OpenAI，Inc.(3:23-cv-03223) 文件1.(地区法院，N.D.加利福尼亚州，2023)。https://storage.courtlistener.com/recap/gov.uscourts. cand.414822/gov.us法院。cand.414822.1.0_1.pdf。

613. D.张，B。夏，Y。刘，X。徐，T。Hoang, Z.邢，M。斯台普斯，Q.Lu, L.朱，“生成式人工智能中的隐私和版权保护: 生命周期视角”，*第三届国际人工智能工程-人工智能软件工程 (CAIN)* (2024)。http://arxiv.org/abs/ 2311.18252。

614. R。Mahari, S.Longpre，“Discit ergo est: 训练数据来源和合理使用”，*涉及生成人工智能*， Schrepel, V.斯托克，编辑。(《网络法评论》，2023)。

615. K。李，A.F.库珀，J.Grimmelmann，“talkin"bout AI Generation: 版权和Generative-AI供应链 (短版)”，*《计算机科学与法律研讨会论文集》(cswlaw '24)* 中 (计算机机械协会，2024) pp。48-63。https://doi.org/10.1145/3614407.3643696。

616. J.Grimmelmann，识字机器人的版权。*爱荷华州法律牧师*。**101**，657-682 (2015)。
https://scholarship.law.cornell.edu/cgi/viewcontent.cgi？article= 2615 & context = facpub。

617. K。李，A.F.库珀，J.Grimmelmann, D.Ippolito，AI和法律: 下一代。(2023)。
https://doi.org/10.2139/ssrn.4580739。

618. L.Tiedrich，当AI产生工作时，标准合同条款可以帮助*OECD产生价值和清晰度。人工智能政策观察站* (2024)。https://oecd.ai/en/wonk/合同条款。

619. M、Sag，生成人工智能的版权安全。*Houst.法律Rev.***61**，295-347 (2023)。
https://houstonlawreview.org/article/ 92126-copyright-safety-for-generative-ai。

620. N.Vyas, S.M。卡卡德，B.巴拉克，“关于生成模型的可证明版权保护”，*载于第40届国际机器学习会议 (ICML 2023)* 论文集 (PMLR，2023)。
https:// 会议录.mlr.press/v202/vyas23b.html。

621. K。McElheran, J.F.李，E。Brynjolfsson, Z.Kroff, E. Dinlersoz, L.福斯特，N.Zolas，美国的人工智能采用: 谁，什么，在哪里。*J.经济。马纳格。***444***战略*，375-415 (2024)。https://doi.org/10.1111/jems.12576。

622. R。马哈里，我。Shayne, L.邓华德，A.波洛佐夫，A. 彭特兰，A.Lipsitz，就数据来源和版权向美国版权局发表评论。(美国版权局，2023)。https://dspace.mit.edu/handle/ 1721.1/154171？show = full。

623. S。Min, S.古鲁兰根，E。华莱士，W.施，H. Hajishirzi，N.A.史密斯，L。Zettlemoyer，“筒仓语言模型: 隔离非参数数据存储中的法律风险”，在*NeurIPS 2023研讨会 (DistShift)* 上 (2023)。
https://openreview.net/forum？id= z03bW0doni。

624. S。Longpre, R.Mahari, N.奥本-马努，W。Brannon, T.南，J。卡巴拉，S.Pentland，数据真实性，同意和AI的出处都被打破了: 修复它们需要什么？*麻省理工学院对生成人工智能*的探索 (2024)。
https://doi.org/10.21428/e4baedd9.a650f77d。

625. S。G.帕蒂尔，T.张，五。方，N。C., R.黄，A。郝，M。卡萨多，J.E。冈萨雷斯，R。A.波帕，我Stoica，GoEX: 面向自治LLM应用程序的运行时的观点和设计，arXiv:2404.06921 [cs.CL] (2024)。
https://doi.org/10.48550/arXiv.2404.06921。

626. D.达尔林普尔，J.Skalse, Y.Bengio, S.罗素，M。Tegmark, S.塞希亚，S.Omohundro，C。Szegedy, B.金哈伯，N。Ammann, A.阿巴特，J。Halpern, C.巴雷特，D.赵，T.志轩，J。翼，J。Tenenbaum，Tow ards保证安全的AI: 确保强大和可靠的AI系统的框架，arXiv:2405.06624 [cs.AI] (2024)。http://arxiv.org/abs/ 2405.06624。

627. N.Cammarata, S.卡特，G.Goh, C.Olah, M.彼得罗夫，L.舒伯特，线程: 电路。*蒸馏***5**，10.23915/蒸馏。00024 (2020)。
https://doi.org/10.23915/distil.00024。

628. N.南达，L.陈，T.Lieberum, J.史密斯，J。Steinhardt，“通过机械可解释性进行研究的进展措施”*在第11届国际学习表征会议 (ICLR 2023)* 中 (2022)。
https://openreview.net/forum？id=9XFSbDPmdW。

629. Z。钟，Z。刘，M。泰格马克，J.安德烈亚斯，“时钟和比萨饼: 神经网络机械解释中的两个故事”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。
https://openreview.net/forum？id= S5wmbQc1We。

630. E。J. Michaud，我.廖，V。Lad, Z.刘，A。穆迪德，C。拉夫里奇，Z。C.郭，T。R。Kheirkhah, M.武克里奇，M.Tegmark，打开AI黑匣子: 通过机械可解释性进行程序合成，arXiv:2402.05110 [cs.LG] (2024)。
https://doi.org/10.48550/arXiv.2402.05110。

631. R。Huben, H. Cunningham, L.R。史密斯，A。Ewart, L.Sharkey，“稀疏自动编码器在语言模型中发现高度可解释的特征”，*在第12届国际学习表示会议 (ICLR 2024)* 中 (2023)。
https://openreview.net/forum？id= F76bwRSLeK。

632. * T。Bricken, A.邓普顿，J。巴特森，B.陈，A。杰明，T.康纳利，N。特纳，C。Anil, C.丹尼森，A.阿塞尔，R。Lasenby, Y.吴，S。Kravec, N.Schiefer, T.麦克斯韦，N。约瑟夫，Z.Hatfield-Dodds, A.Tamkin, K.阮，。。。C. Olah，走向单式: 用字典学习分解语言模型，变压器电路线程 (2023)。
https:// transformer-circuits.pub/2023/单语义-功能。

633. L.高，J。舒尔曼，J。希尔顿，“奖励模型过度优化的比例法则”，*载于第40届国际机器学习*会议论文集 (PMLR，2023) pp。10835-10866。
https:// 会议录.mlr.press/v202/gao23h.html。

634. J.《M. V.》Skalse, N.H. R.豪，D。Krasheninnikov, D.Krueger，*第36届神经信息处理系统会议 (NeurIPS 2022)* 中的“定义和表征奖励游戏”(2022)。
https://openreview.net/forum？id= yb3HOXO3lX2。

635. P。Si nghal, T.戈亚尔，J.徐，G。Durrett，还有很长的路要走: 研究RLHF中的长度相关性，arXiv:2310.03716 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2310.03716。

636. J.Tien, J.Z.-Y。他，Z。埃里克森，A。德拉根，D.S。布朗，“基于偏好的奖励学习中的因果混淆和奖励错误识别”*在第11届国际学习表征会议 (ICLR 2023)* 中 (2022)。
https://openreview.net/forum？id= R0Xxvr_X3ZA。

637. L.E.麦金尼，Y.段，D。克鲁格，A。*第36届会议神经信息处理系统 (NeurIPS 2022) 深度强化学习研讨会*上的“学习奖励函数的脆弱性”(2022)。
https://openreview.net/forum？id= 9gj9vXfeS-y。

638. L.L. D.Langosco, J.科赫，L。D.Sharkey, J.Pfau, D.克鲁格，“深度强化学习中的目标误泛化”，*载于第39届国际机器学习*大会 (PMLR，2022) 卷。162，第页。12004- 12019。https:// 会议记录.mlr.press/v162/langosco22a.html.

639. E。西格，N。德雷克斯勒，R。Moulange，E。达达曼，J.舒特，K。魏，C。冬天，M。阿诺德，S.Ó。hÉigeartaigh, A.Korinek, M.安德隆，B。Bucknall, A.陈，E。斯塔福德，L。Koessler, A.奥瓦迪亚，B.加芬克尔，E。Bluemke, M.艾尔德。。。A. Gupta，“开源高性能基础模型: 评估”追求开源目标的风险、收益和替代方法 ”(人工智能治理中心，2023)；
http://arxiv.org/abs/ 2311.09227。

640. S。卡普尔，R。Bommasani, K.Klyman, S.Longpre，A。Ramaswami, P.Cihon, A.霍普金斯，K.班克斯顿，S.比德曼，M。博根，R.乔杜里，A。恩格勒，P。亨德森，Y。Jernite, S.Lazar, S.Maffulli, A.纳尔逊，J。皮诺，A。Skowron，。。。A. Narayanan，“论开放基金会模式的社会影响”，arXiv:2403.07918 [cs.CY] (2024)。
https://doi.org/10.48550/arXiv.2403.07918。

641. K。Hu，ChatGPT创造了增长最快的用户群记录- *路透社*分析师指出。(2023)。
https://www.reuters.com/technology/ chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/。

642. * 米。纳斯尔，N.卡里尼，J.Hayase, M.Jagielski, A. Feder Cooper, D.伊波利托，C。A.Choquette-Choo, E.华莱士，F.特拉梅尔，K。Lee，从 (生产) 语言模型中可扩展地提取训练数据，arXiv:2311.17035 [cs.LG] (2023)。
https://doi.org/10.48550/arXiv.2311.17035。

643. H。李，D。郭，W。范，M。徐，J。黄，F。孟，Y。Song，“ChatGPT上的多步越狱隐私攻击”*在2023自然语言处理经验方法会议 (EMNLP 2023)* 中 (2023)。
https://openreview.net/forum？id= ls4Pfsl2jZ。

644. A.Deshpande, V.Murahari, T.Rajpurohit, A.Kalyan, K.Narasimhan，“chatgpt中的毒性: 分析角色分配的语言模型”，*发现As计算语言学协会: EMNLP 2023* (计算语言学协会，2023) pp。1236-1270。https://doi.org/10.18653/v1/2023.findings-emnlp.88。

645. V。霍夫曼，P。R。Kalluri, D.Jurafsky, S.金，方言预判预测人工智能对人们的性格、就业能力和犯罪行为的决定，arXiv:2403.00742 [cs.CL] (2024)。https://doi.org/10.48550/arXiv.2403.00742。

646. E。崔，Y.乔，J.Jang, J.张，M。Seo，“有效提示的固定输入参数化”，*结果为计算语言学协会: ACL 2023* (计算语言学协会，2023) pp。8428-8441。
https://doi.org/10.18653/v1/2023.findings-acl.533。

647. E。Shayegani, M.A. Al Mamun, Y.傅，P。Zaree, Y.东，N。Abu-gh azaleh，《对抗性攻击揭示的大型语言模型中的漏洞调查》，arXiv:2310.10844 [cs.CL] (2023)。http://arxiv.org/abs/ 2310.10844。

648. * C.Anil, E.Durmus, M.夏尔马，J.Benton, S.昆杜，J.巴特森，N。Rimsky, M.童，J。Mu, D.福特，F。莫斯科尼，R.阿格拉沃尔，R。Schaeffer, N.巴什坎斯基，S.斯文宁森，M。兰伯特，A.Radhakrishnan, C.丹尼森，E.J.哈宾格。。。D. Duvenaud，“多次越狱”(Anthropic，2024)。

649. Y。M.p a Pa，S.Tanizaki, T.Kou, M. van Eeten, K.吉冈，T。松本，“攻击者的梦想？“探索ChatGPT开发恶意软件的能力”，*载于第16届网络安全实验和测试一下研讨会 (CSET '23)*，(计算机协会，2023) pp。10-18。
https://doi.org/10.1145/3607505.3607513。

650. 问:詹，R。方，R。宾杜，A.古普塔，T.桥本，D.Kang，“通过微调消除GPT-4中的RLHF保护”，*在2024年度会议的北美分会计算语言学*协会 (2024) 中。
https://doi.org/10.48550/arXiv.2311.05553。

651. X。齐，Y。曾，T。谢，P.-Y。陈，R。贾，P。米塔尔，P。亨德森，“微调对齐的语言模型会损害安全性，即使用户不打算这样做!”*在第12届学习表示国际会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= hTEGyKf0dZ。

652. Z。Xi，W。陈，X。郭，W。他，Y。丁，B。洪，M。张，J。王，S。Jin, E.周，R。郑X。风扇，X。王，L。熊，Y。周，W。王，C。江，Y。邹，X。刘，。。。T. Gui，基于大型语言模型的智能体的兴起和潜力: 一项调查，arXiv:2309.07864 [cs.AI] (2023)。https://doi.org/10.48550/arXiv.2309.07864。

653. Y。田，X。杨，J。吴哲，Y。Dong，H. Su，邪恶的天才: 深入研究基于LLM的代理的安全性，arXiv:2311.11855 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2311.11855。

654. Z。吴，C。Han, Z.丁，Z。翁，Z。刘，S。Y ao，T。于，L。Kong，OS-Copilot: 走向具有自我完善的通才计算机代理，arXiv:2402.07456 [cs.AI] (2024)。http://arxiv.org/abs/ 2402.07456。

655. * 司马团队，M。A.Raad, A.Ahuja, C.巴罗斯，F。Besse, A.博尔特，A.博尔顿，B。布朗菲尔德，G.Buttimore, M.不能，S.Chakera, S.C.Y。陈，J.Clune, A.科利斯特，V。Copeman, A.Cullum，我.达斯古普塔，德·切萨雷，J·迪·特拉帕尼，。。。N. Young，“在许多模拟世界中扩展可指导的代理”(Google Deepmind，2024)；http://arxiv.org/abs/ 2404.10179。

656. 问:Lu, L.朱X。徐，Z。邢，S。哈勒，J.Whittle，走向负责任的生成式AI: 设计基于基础模型的代理的参考体系结构，arXiv:2311.13148 [cs.AI] (2023)。http://arxiv.org/abs/ 2311.13148。

657. * R。中野，J.希尔顿，S.巴拉吉，J.吴，L。欧阳，C。金，C。黑森，S.Jain, V.Kosaraju, W.桑德斯，X。江，K。Cobbe, T.埃隆杜，G。克鲁格，K。按钮，M.骑士，B.国际象棋，J.Schulman，“WebGPT: 带有人类反馈的浏览器辅助问答”(OpenAI，2021)；http://arxiv.org/abs/ 2112.09332。

658. A.陈，C。Ezell, M.考夫曼，K。魏，L。哈蒙德，H.布拉德利，E.Bluemke, N.Rajkumar, D.克鲁格，N。Kolt, L.海姆，M. Anderljung，“对人工智能代理的可见性”，arXiv:2401.13138 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2401.13138。

659. D.银，A.黄，C。J.Maddison, A.Guez, L.Sifre, G. van den Driessche, J.Schrittwieser，我.Antonoglou, V.Panneershelvam, M.兰克托特，S.Dieleman, D.Grewe, J.Nham, N.Kalchbrenner，我.Sutskever, T.Lillicrap, M.利奇，K.Kavukcuoglu, T.雷佩尔· G，。。。D. Hassabis，通过深度神经网络和树搜索掌握围棋游戏。*自然***529**，484-489 (2016)。https://doi.org/10.1038/自然16961。

660. M. Mazeika, L.Phan, X.尹，A。邹，Z。王，N。Mu, E.Sakhaee, N.李，S。Basart, B.李，D。福赛斯，D.Hendrycks，HarmBench: 用于自动红队和稳健拒绝的标准化评估框架，arXiv:2402.04249 [cs.LG] (2024)。http://arxiv.org/abs/ 2402.04249。

661. S。姚，D.于，J。赵，我。沙夫兰，T。L.格里菲斯，Y。曹，K.R。Narasimhan，*第37届神经信息处理系统会议(NeurIPS 2023)* 的“思想树: 使用大型语言模型故意解决问题”(2023)。https://openreview.net/forum？id= 5Xc1ecxO1h。

662. T.A. Han, L.M。佩雷拉，T。Lenaerts，“建模和影响人工智能竞购战: 研究议程”*《2019 AAAI/ACM人工智能、伦理和社会会议论文集》(AIES '19)*，2019页。5-11。https://doi.org/10.1145/3306618.3314265。

663. T.Cimpeanu, F.C.桑托斯，L。M。佩雷拉，T。Lenaerts, T.A.Han，异构环境中的人工智能开发竞赛。*Sci。代表。1723* **12** (2022)。https://doi.org/10.1038/s41598-022-05729-3。

664. S。阿姆斯特朗，N。博斯特罗姆，C。舒尔曼，《奔向悬崖: 人工智能发展的模型》。*AI Soc*。**31**，201-206 (2016)。https://doi.org/10.1007/s00146-015-0590-y。

665. M. Menashe，《重新审视的底层竞赛: 国际劳动法、全球贸易和进化博弈论》。*Oxf.J.腿。螺柱*。**40**，53-81 (2020)。https://doi.org/10.1093/ojls/gqz029。

666. M. M.马斯，“变革下的人工智能治理: 基础、方面、框架”，哥本哈根大学论文 (2020)。

667. L.科利娜，M。S ayyadi，M。Provitera，关于人工智能的关键问题。问责回答。《*加州管理评论见解*》(2023)。https://cmr.berkeley.edu/2023/11/关键-问题-关于-我-问责制-已回答/。

668. A.T.达丰塞卡，E.Vaz de Sequeira，L. Barreto Xavier，“人工智能驱动系统的责任”，*多学科视角的人工智能和法律*，Sousa Antunes，P。M。弗雷塔斯，A。L.Oliveira, C. Martins Pereira, E. Vaz De Sequeira, L. Barreto Xavier, Eds.(施普林格国际出版社，Cham，2024)，pp。299-317。

669. M. Buiten, A. de Streel, M.Peitz，AI责任的法律和经济学。*计算机。法律安全。代表。105794* **48** (2023)。https://doi.org/10.1016/j.Clsr.2023.105794。

670. M. Busuioc，负责任的人工智能: 让算法负责。*公众adm。Rev*。**81**，825-836 (2021)。https://doi.org/10.1111/puar.13293。

671. F.Doshi-Velez, M.Kortz, R.Budish，C。巴维茨，S.J.Gershman, D.O'Brien, K.斯科特，S.谢伯，J.沃尔多，D。温伯格，A。韦勒，A。伍德，“法律下人工智能的责任: 解释的作用”(伯克曼·克莱因中心解释与法律工作组，2017); http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

672. N.Kolt, M.Anderljung, J.巴恩哈特，A.黄铜，K.Esvelt, G.K。哈德菲尔德，L.海姆，M.罗德里格斯，J。B.Sandbrink, T.伍德赛德，前沿人工智能发展责任报告，arXiv:2404.02675 [cs.CY] (2024)。https://doi.org/10.48550/arXiv.2404.02675。

673. M. Anderljung, J.巴恩哈特，A.Korinek, J.梁，C。奥基夫，J。Whittlestone, S.Avin, M.布伦达奇，J.布洛克，D.卡斯-贝格斯，B.Chang, T.柯林斯，T.拳头，G.哈德菲尔德，A.海耶斯，L.Ho, S.胡克，E.霍维茨，N。Kolt, .。。沃尔夫，“前沿人工智能监管: 管理公共安全的新兴风险”，arXiv:2307.03718 [cs.CY] (2023)。https://doi.org/10.48550/arXiv.2307.03718。

674. R。佩林，我Habli，“汽车安全的保证- *计算机安全，可靠性和安全性 (safecompp 2010)* 中的一种安全案例方法”(Springer，2010) pp。82-96。https://doi.org/10.1007/978-3-642-15651-9_7。

675. M. L.卡明斯，重新思考人工智能在安全关键环境中的成熟度。*AI Mag*.**42**，6- 15 (2021)。https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/ 7394。

676. 世界卫生组织，卫生产品政策和标准 (2024)。https:// www.who.int/teams/health-药品/指南的产品和政策标准/标准和规范/规范和标准。

677. I.I. Livshitz, P.A.Lontsikh, N.P。Lontsikh, E.Y。Golovina, O.M。Safonova，“燃料和能源行业工业安全保证的现代风险管理方法研究”，*2021国际质量*会议

管理，运输和信息安全，信息技术 (IT & QM & IS)，(2021) pp。165-167。
https://doi.org/10.1109/itqmis53292.2021.9642791

678. H。E.罗兰，B. *系统安全工程与管理*Moriarty (Wiley，纽约，ed。第二版，1990)，pp。367。

679. L.Koessler, J.Schuett，AGI公司的风险评估: 来自其他安全关键行业的流行风险评估技术的回顾，arXiv:2307.08823 [cs.CY] (2023)。https://doi.org/10.48550/arXiv.2307.08823。

680. 国家研究所标准和技术gy、人工智能风险管理框架的。*NIST* (2021)。
https://www.nist.gov/itl/ ai-风险-管理-框架。

681. D.Khodyakov, S.格兰特，J.Kroger, M.鲍曼，"进行和批判性评估德尔福小组的兰德方法指南" (兰德公司，2023)；https://doi.org/10.7249/tla3082-1。

682. ISO，ISO 31000: 风险管理 (2021)。https://www.iso.org/iso-31000-risk-management.html。683。美国核监管委员会，运行新反应堆的风险指标。(2009)。
https://www.nrc.gov/docs/ ML0909/ML090910608.pdf。

684. E。黑色，R.奈杜，R.加尼，K.罗道法，D.Ho，H. Heidari，"实现管道感知的ML公平性: 开发实用指南和工具的研究议程"，*在第三届ACM会议关于算法，机制和优化的公平和访问 (eaamo'23)* 中 (计算机协会，2023) pp。1-11。
https://doi.org/10.1145/3617694.3623259。

685. S。里斯马尼，R。谢尔比，A.聪明，E.Jatho, J.Kroll, A.月亮，N.Rostamzadeh，"从飞机失事到算法危害: 安全工程框架对负责任的ML的适用性"，*载于2023 CHI会议关于计算系统人为因素 (CHI '23)* 的论文集 (计算机协会，2023) pp。1-18。
https://doi.org/10.1145/3544548.3581407。

686. T.凯利，安全案例管理的系统方法。*SAE Trans。J.马特.曼努夫。***113**，257-266 (2004)。
http://www.jstor.org/stable/ 44699541。

687. M.斯坦，C.Dunlop，"售前安全: 从FDA的基础模型的生命科学监督模型中学到的东西" (Ada Lovelace Institute，2023); https://www.adalovelaceinstitute.org/wp-内容/上传/2023/12/2023_12_ALI_Safe-before-sale_Discussion_paper.pdf。

688. T.拉兹，D.Hillson，风险管理标准的比较回顾。*风险管理: 内部。J.*，**7**，53-66 (2005)。
https://doi.org/10.1057/帕尔格雷夫。rm.8240227。

689. J.克莱默，N.加布里埃利，D克鲁格，T。Larsen，安全案例: 如何证明高级AI系统的安全性，arXiv:2403.10462 [cs.CY] (2024)。http://arxiv.org/abs/ 2403.10462。

690. C.Haddon-Cave*的Nimrod评论: 一个独立的审查更广泛的问题周围损失的英国皇家空军Nimrod MR2 飞机XV230在阿富汗2006年，报告*，(文具办公室，2009)。

691. N.G. Leveson，应用系统思维来分析和学习事件。*Saf.Sci*。**49**，55-64 (2011)。
https://doi.org/10.1016/j.Ssci.2009.12.021。

692. D.Hendrycks，《*人工智能安全、伦理和社会*导论》(Taylor & Francis)。693。*

Anthropic，Anthropic的负责任扩展策略，版本1.0。(2023)。

694. * OpenAI，"准备框架 (测试版)" (OpenAI，2023); https://cdn.openai.com/openai-准备-framework-beta.pdf。

695. * Z。肯特上，T。埃弗里特，L.魏丁格，我。加布里埃尔，V。米库里克，G。Irving，"语言代理的对齐" (Google DeepMind，2021)；http://arxiv.org/abs/ 2103.14659。

696. 吴M，A。F.Aji，"风格重于实质: 大型语言模型的评估偏差"，arXiv:2307.03025 [cs.CL] (2023)。
https://doi.org/10.48550/arXiv.2307.03025。

697. * N.兰伯特，R。Calandra，Alig上限: 从人类反馈中强化学习的目标不匹配，arXiv:2311.00168 [cs.LG] (2023)。
https://doi.org/10.48550/arXiv.2311.00168。

698. H。班萨尔，J.当当，A。Grover，"窥视偏好: 解开反馈获取以对齐大型语言模型"，*在第12届国际学习表征会议 (ICLR 2024)* 中 (2023)。
https://openreview.net/forum？id= dKl6lMwbCy。

699. J.Uesato, N.库什曼，R。库马尔，F。宋，N。西格尔，L。王，A。克雷斯韦尔，G.欧文，我。希金斯，"用基于过程和结果的反馈解决数学单词问题" (谷歌Deepmind，2022)；
https://doi.org/10.48550/arXiv.2211.14275。

700. H。莱特曼，V。Kosaraju, Y.伯达，H.爱德华兹，B.贝克，T。李，J.雷克，J.舒尔曼，我。Sutskever, K.Cobbe，"让我们逐步验证" *在第12届国际会议上学习表示 (ICLR 2024)* 中 (2023)。
https://openreview.net/forum？id= v8L0pN6EOi。

701. Z。吴，Y。胡，W。施，N.Dziri, A.苏尔，P。Ammanabrolu, N.A.史密斯，M。Ostendorf，H. Hajishirzi，“细粒度的人类反馈为语言模型训练提供了更好的回报”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 中 (2023)。https://openreview.net/forum？id=CSbGXyCswu。

702. * H.李，S。Phatale, H. Mansoor, T.梅斯纳德，J.雪貂，K.卢，C。主教，E。霍尔，V。Carbune, A.Rastogi, S.Prakash，RLAIF: 使用AI反馈从人类反馈中扩展强化学习，arXiv:2309.00267 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2309.00267。

703. D.哈德菲尔德-梅内尔，A.德拉根，P。Abbeel, S.Russell，“合作逆强化学习”，*载于第30届国际会议神经信息处理系统 (nips'16)* 论文集 (Curran Associates Inc.，2016) pp。3916-3924。

704. D.哈德菲尔德-梅内尔，S。米利，P.Abbeel, S.罗素，A.D.Dragan，“反向奖励设计”，*载于第31届国际神经信息处理系统会议论文集 (nips'17)* 中 (Curran Associates Inc.，2017) pp。6768- 6777。

705. X。梁，K。舒，K。李，P。Abbeel，“在基于偏好的强化学习中探索的奖励不确定性”，*在第10届国际学习表征会议 (ICLR 2022)* 中 (2021)。https://openreview.net/forum？id= owzvd-l-zrc。

706. * A.Gleave, G.欧文，语言奖励模型的不确定性估计，arXiv:2203.07472 [cs.CL] (2022)。https://doi.org/10.48550/arXiv.2203.07472。

707. A.Rame, G.Couairon, C。Dancette J.-B.加亚，M.Shukor, L.苏利尔，M。在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上的Cord，“奖励汤: 通过对不同奖励进行微调的插值权重来实现帕累托最优对齐”(2023)。https://openreview.net/forum？id= lSbbC2VyCu。

708. T.美国科斯特。安瓦尔，R。柯克，D.Krueger，“奖励模型集合有助于缓解过度优化”*在第12届国际学习表征会议 (ICLR 2024)* 中 (2024)。https://openreview.net/forum？id=dcjtMYkpXx。

709. * S.R。鲍曼，J.Hyun, E.佩雷斯，E。陈，C。佩蒂特，S.海纳，K。Lukošiūtė, A.Askell, A.琼斯，A。陈，A。戈尔迪，A.Mirhoseini，C。麦金农，C.奥拉，D.Amodei, D.Amodei, D.排水管，D.李，E。Tran-Jo hnson，。。。J. Kaplan，衡量大型语言模型可扩展监督的进展，arXiv:2211.03540 [cs.HC] (2022)。http://arxiv.org/abs/ 2211.03540。

710. * C.伯恩斯，P.Izmailov, J.H.基什内尔，B.贝克，L。高，L。Aschenbrenner, Y.陈，A。Ecoffet，M。Joglekar, J.雷克，我。Sutskever, J.W u，“弱到强泛化：用弱监督激发强能力”，arXiv:2312.09390 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2312.09390。

711. J.迈克尔，S.马赫迪，D。雷恩，J. 佩蒂，J.迪拉尼，V。Padmakumar, S.R。鲍曼，辩论有助于监督不可靠的专家，arXiv:2311.08702 [cs.AI] (2023)。http://arxiv.org/abs/ 2311.08702。

712. * J。雷克，D。克鲁格，T。埃弗里特，M。马蒂奇，V。Maini, S.Legg，“通过奖励建模实现可扩展代理对齐: 一个研究方向”，arXiv:1811.07871 [cs.LG] (2018)。http://arxiv.org/abs/ 1811.07871。

713. * A.Khan, J.休斯，D。Valentine, L.Ruis, K.Sachan, A.Radhakrishnan, E.Grefenstette, S.R。鲍曼，T。洛克泰尔，E。佩雷斯，与更具说服力的llm进行辩论会导致更真实的答案，arXiv:2402.06782 [cs.AI] (2024)。http://arxiv.org/abs/ 2402.06782。

714. Z。李，《聊天的阴暗面》: 随机鹦鹉和幻觉带来的法律和伦理挑战，arXiv:2304.14347 [cs.CY] (2023)。http://arxiv.org/abs/ 2304.14347。

715. * A.Askell，Y。白，A。陈，D。排水管，D.甘古丽，T.Henighan, A.琼斯，N。约瑟夫，B。曼恩，N.DasSarma, N.Elhage，Z。Hatfield-Dodds, D.埃尔南德斯，J。Kernion, K.Ndousse, C.奥尔森，D。Amodei, T.眉毛n，J.克拉克。。。J. Kaplan，作为对齐实验室的通用语言助理，arXiv:2112.00861 [cs.CL] (2021)。http://arxiv.org/abs/ 2112.00861。

716. P。刘易斯，E。佩雷斯，A。皮克图斯，F.Petroni, V.Karpukhin, N.戈亚尔，H. K ü ttler，M.刘易斯，W.-T。Yih, T.Rocktäschel, S.里德尔，D.Kiela，“知识密集型NLP任务的检索增强生成”，在*第34届神经信息处理系统会议 (NeurIPS 2020)* 上 (Curran Associates，inc.，2020) 第一卷。444，第页。9459-9474。https:// 会议录.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

717. K。Shuster, S.Poff, M.陈，D。Kiela, J.韦斯顿，“检索增强减少了对话中的幻觉”，*发现了计算语言学协会: EMNLP 2021* (计算语言学协会，2021) pp。3784-3803。https://doi.org/10.18653/v1/2021.findings-emnlp.320。

718. L.库恩，Y。Gal, S.Farquhar，“语义不确定性: 自然语言生成中不确定性估计的语言不变性”，*在第11届国际学习表示会议 (ICLR 2023)* 中 (2023)。https://openreview.net/forum？id= VD-AYtP0dve。

719. D.亨德里克斯，S.Basart, N.Mu, S.Kadavath, F.王，E。多伦多，R。德赛，T.朱，S。帕拉朱利，M。郭丁。宋，J。斯坦哈特，J。Gilmer，“鲁棒性的许多方面: 对分布泛化的批判性分析”，*2021 IEEE/CVF国际计算机视觉会议 (ICCV)* (2021) pp。8320-8329。https://doi.org/10.1109/iccv48922.2021.00823。

720. N.Carlini, M.纳斯尔，C。A.Choquette-Choo, M.Jagielski，我.高，P。W。Koh, D.Ippolito, F.特拉梅尔，L.施密特，“对齐的神经网络是否对立对齐？”在*第37届神经信息处理系统会议 (NeurIPS 2023)* 中 (2023)。https://openreview.net/forum？id= OQQoD8Vc3B。

721. A.Madry, A.马克洛夫，L.施密特，D.齐普拉斯，A.Vladu，“迈向抵抗对抗性攻击的深度学习模型”，*在第六届学习表征国际会议 (ICLR 2018)* 中 (2018)。https://openreview.net/forum？id=rJzIBfZAb。

722. * E.Hubinger，C。丹尼森，J.Mu, M.兰伯特，M.童，M。MacDiarmid, T.兰汉姆，D。M。齐格勒，T。麦克斯韦，N。程，A.杰明，A。Askell, A.Radhakrishnan, C.阿尼尔，D。杜维诺，D.甘古丽，F.Barez, J.克拉克，K。恩杜斯，。。。E. Perez，《卧铺特工: 通过安全培训持续存在的培训欺骗性llm》，arXiv:2401.05566 [cs.CR] (2024)。https://doi.org/10.48550/arXiv.2401.05566。

723. N.Jain, A.Schwarzschild, Y.温，G。Somepalli, J.Kirchenbauer，P.-Y.蒋，M。Goldblum，A。萨哈，J.Geiping, T.Goldstein，针对对齐语言模型的对抗性攻击的基线防御，arXiv:2309.00614 [cs.LG] (2023)。https://doi.org/10.48550/arXiv.2309.00614。

724. S。卡斯珀，L.舒尔茨，O。帕特尔，D。Hadfields-Menell，通过潜在的对抗训练防御不可预见的故障模式，arXiv:2403.05030 [cs.CR] (2024)。https://doi.org/10.48550/arXiv.2403.05030。

725. D.齐普拉斯，S.Santurkar, L.恩格斯特罗姆，A。特纳，A。Madry，“鲁棒性可能与准确性不一致”*在第七届国际会议上学习表示 (ICLR 2019)* 中 (2018)。https://openreview.net/forum？id= SyxAb30cY7。

726. H。张，Y。于，J。焦，E。邢，L。E。Ghaoui, M.乔丹，“鲁棒性和准确性之间的理论原则权衡”，*载于第36届国际机器学习会议论文集* (2019) pp。7472-7482。https:// 会议记录.mlr.press/v97/zhang19p.html.

727. Y.-Y。杨，C。Rashtchian, H. Zhang, R.R。Salakhutdinov, K.Chaudhuri，《神经信息处理系统中的*广告》 (NeurIPS 2020)* 中的“准确性与鲁棒性的更深入研究”(Curran Associates，inc.，2020)。444，第页。8588- 8601。https:// 会议录.neurips.cc/paper/2020/hash/61d77652c97ef636343742fc3dcf3ba9-Abstract.html.

728. Y。巴拉吉，T。戈德斯坦，J。霍夫曼，《实例自适应对抗训练: 神经网络中改进的准确性权衡》，arXiv:1910.08051 [cs.LG] (2019)。https://doi.org/10.48550/arXiv.1910.08051 。

729. Y。王，D。邹，J。易，J。Bailey, X.妈，Q.Gu，“提高对抗鲁棒性需要重新审视错误分类的例子”*在第八届国际会议上学习表示 (ICLR 2020)* (2019)。https://openreview.net/forum？id= rklOg6EFwS。

730. R。雷德，S.-M。Moosavi-dezfooli，“减少过大的余量以实现更好的准确性与鲁棒性的权衡”，*在第十届国际学习表示会议 (ICLR 2022)* 中 (2021)。https://openreview.net/forum？id= Azh9QBQ4tR7。

731. Z。刘，G。Dou, Z.谭，田勇，M。Jiang，通过机器学习走向更安全的大语言模型，arXiv:2402.10058 [cs.CL] (2024)。https://doi.org/10.48550/arXiv.2402.10058。

732. S。刘，Y。姚，J。贾，S。卡斯珀，N.Baracaldo，P。哈斯，X。徐，Y。姚，H. Li，K。R。Varshney, M.班萨尔，S.Koyejo, Y.Liu，为大型语言模型重新思考机器学习，arXiv:2402.08787 [cs.LG] (2024)。https://doi.org/10.48550/arXiv.2402.08787。

733. * R。埃尔丹，M.罗辛诺维奇，谁是哈利·波特？LLMs中的近似学习，arXiv:2310.02238 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2310.02238。

734. A.林奇，P。郭，A。Ew艺术，S.卡斯珀，D.Hadfieldr-menell，LLMs中评估鲁棒学习的八种方法，arXiv:2402.16835 [cs.CL] (2024)。https://doi.org/10.48550/arXiv.2402.16835。

735. G.阿兰，Y.Bengio，“使用线性分类器探针理解中级语言”，在*第五届学习表示国际会议上，ICLR 2017* (2017)。https://openreview.net/forum？id= HJ4-rAVtl。

736. P。戈亚尔，A.R。索里亚诺，C。哈齐尔巴斯，L.萨贡，N.Usunier，“视觉特征提取器的系统评估的公平性指标”，*载于2022 ACM Confere nce关于公平性，问责制和透明度 (FAccT '22)* 的论文集 (计算机协会，2022) pp。70-88。https://doi.org/10.1145/3531146.3533074。

737. W。古尼，M。Tegmark，“语言代表空间和时间模型”*在第12届国际会议上学习表示 (ICLR 2024)* (2023)。https://openreview.net/forum？id= jE8xbmvFin。

738. * S.马克斯，M.Tegmark，真理的几何: 真/假数据集的大型语言模型表示中的紧急线性结构，arXiv:2310.06824 [cs.AI] (2023)。https://doi.org/10.48550/arXiv.2310.06824 。

739. A.拉维尚德，Y。Belinkov，E。Hovy，“探测范式: 探测准确性是否需要任务相关性？”在*第16届会议欧洲分会计算语言学协会论文集: 主卷 (EACL 2021)* 中 (计算语言学协会，2021) pp。3363-3377。https://doi.org/10.18653/v1/2021.eacl-main.295。

740. Y。Elazar, S.Ravfogel, A.Jacovi, Y.戈德堡，健忘公关: 健忘反事实的行为解释。*Trans。Assoc。计算机。语言学家。***9**，160-175 (2021)。https://doi.org/10.1162/tacl_a_00359。

741. O.Antverg, Y.Belinkov，“关于分析语言模型中单个神经元的陷阱”，*在第10届国际学习表征会议 (ICLR 2022)* 上 (2021)。https://openreview.net/forum？id= 8uz0EWPQIMu。

742. T.Lieberum, M.Rahtz, J.克拉玛尔，N.南达，G.Irving, R.Shah, V.Mikulik，电路分析可解释性是否可以扩展？来自龙猫多项选择能力的证据，arXiv:2307.09458 [cs.LG] (2023)。http://arxiv.org/abs/ 2307.09458。

743. E。米切尔，C。林，A。Bosselut，C。D.曼宁，C。Finn，“基于记忆的模型大规模编辑”，*载于第39届机器学习*国际会议论文集 (PMLR，2022) pp。15817-15831。https:// 议事录.mlr.press/v162/mitchell22a.html。

744. K。孟，D。Bau, A.J.Andonian, Y.Belinkov，*第36届神经信息处理系统会议 (NeurIPS 2022)* “在GPT中定位和编辑事实关联” (2022)。https://openreview.net/forum？id=- h6WAS6eE4。

745. K。孟，A。S。夏尔马，A.J.Andonian, Y.别林科夫，D。Bau，“变压器中的大规模编辑内存” *在第11届国际学习表征会议 (ICLR 2023)* 中 (2022)。https://openreview.net/forum？id=MkbcAHIYgyS。

746. Y。Gandelsman, A.A.埃夫罗斯，J.Steinhardt，“通过基于文本的分解来解释CLIP的图像表示” *在第12届国际学习表示会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= 5Ca9sSzuDp。

747. C.谭，G。张，J。Fu，“通过元学习对大型L语言模型进行大规模编辑”，*在第12届国际学习表示会议 (ICLR 2024)* 中 (2023)。https://openreview.net/forum？id= L6L1CJQ2PE。

748. S。王，Y。朱，刘，Z。郑，C。陈，J。李，大型语言模型的知识编辑: 一项调查，arXiv:2310.16218 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2310.16218。

749. Z。韩，C。高，J。刘，J。张，S。问:张，大型模型的参数有效微调: 综合调查，arXiv:2403.14608 [cs.LG] (2024)。https://doi.org/10.48550/arXiv.2403.14608。

750. X。吴，J。李，M。徐，W。董，S。吴，C。Bian, D.Xiong，“DEPN: 在预训练的语言模型中检测和编辑隐私神经元”，*《2023自然语言处理经验方法会议论文集 (EMNLP 2023)》* (计算语言学协会，2023) pp。2875-2886。https://doi.org/10.18653/v1/2023.emnlp-main.174。

751. K。李，O.帕特尔，F。维加斯，H.普菲斯特，M。Wattenberg，“推理时间干预: 从语言模型中得出真实答案”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id= aLLuYpn83y。

752. A.特纳先生，L。蒂尔加特，D。尤德尔，G.里奇，美国。Mini，M. MacDiarmid，激活添加: 无优化的转向语言模型，arXiv:2308.10248 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2308.10248。

753. N.贝尔罗斯，D。施耐德-约瑟夫，S.Ravfogel，R。Cotterell, E.拉夫，S.Biderman，“LEACE: 封闭形式的完美线性概念擦除”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id=awIpKpwTwF&noteId= Ju4XcafMir。

754. E。埃尔南德斯，B。Z。李，J。Andreas，检查和编辑语言模型中的知识表示，arXiv:2304.00740 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2304.00740。

755. D.布朗，C。戈弗雷，C。尼津斯基，J.Tu，H. Kvinge，“编辑神经网络的鲁棒性”，在*ICLR 2023研讨会上对基础模型的数学和经验理解 (me-fomo 2023)* 中 (2023)。https://openreview.net/forum？id= JAjH6VANZ4。

756. D.Gamage, J.陈，K。Sasahara，“Deepfakes的出现及其社会影响: 系统的回顾”，*载于2021真相与信任在线会议*论文集 (Hacks Hackers，2021) pp。28-39。https://www.researchgate.net/publication/ 355583941_The_Emergence_of_Deepfakes_and_its_Societal_Implications_a_系统_回顾。

757. A.考沙尔，A。米娜，A.Meena, T.H. Babu，“Deepfakes的社会影响: 检测和缓解的进展”，*2023 14届国际计算通信和网络技术会议 (ICCCNT)* (2023) pp。1-7。https://doi.org/10.1109/icccnt56998.2023.10307353。

758. R。唐，Y.-N。Chuang，X。Hu，检测LLM生成文本的科学。*Commun。美国石油公司***67**，50-59 (2024)。https://doi.org/10.1145/3624725。

759. R。科维，D.Cozzolino, G.Zingarini, G.波吉，K。长野，L。Verdoliva，“关于扩散模型生成的合成图像的检测”，在*ICASSP 2023 - 2023 IEEE国际会议声学，语音和信号处理 (ICASSP)* 中 (2023) pp。1-5。https://doi.org/10.1109/icassp49357.2023.10095167。

760. 美国。Ojha, Y.李，Y。J.Lee，*2023 IEEE/CVF计算机视觉和模式识别 (CVPR)* 会议 (IEEE计算机学会，2023) pp中的“迈向通用的跨生成模型通用假图像检测器”。24480-24489。https://doi.org/10.1109/cvpr52729.2023.02345。

761. Y。赵，T.庞，C。杜，X。杨，N.-M。张，M。Lin，水印扩散模型的配方，arXiv:2303.10137 [cs.CV] (2023)。https://doi.org/10.48550/arXiv.2303.10137 。

762. J.Kirchenbauer, J.Geiping, Y.温，J。Katz我.Miers, T.Goldstein，“大型语言模型的水印”*第40届机器学习*国际会议论文集 (PMLR，2023) pp。17061-17084。https:// 议事录.mlr.press/v202/kirchenbauer23a.html。

763. A.诺特，D。Pedreschi, R.Chatila, T.Chakraborti, S.利维，R.Baeza-Yates, D.艾尔斯，A.Trotman，P。D.蒂尔，P。比切克，S。罗素，Y.Bengio，生成式AI模型应该包括检测机制，作为公开发布的条件。*伦理Inf。技术。***25**，55 (2023)。https://doi.org/10.1007/s10676-023-09728-4。

764. G.庞，C。沈，L。曹，A. Van Den Hengel，深度学习用于异常检测: 综述。*ACM计算机。Surv.* **54**，38:1-38:38 (2021)。https://doi.org/10.1145/3439950。

765. T.阿里，P。Kostakos，HuntGPT: 将基于机器学习的异常检测和可解释的AI与大型语言模型 (llm) 集成，arXiv:2309.16021 [cs.CR] (2023)。https://doi.org/10.48550/arXiv.2309.16021 。

766. J.耿，F。蔡，Y.王，H. Koeppl，P。Nakov，I. Gurevych，大型语言模型中置信度估计和校准的调查，arXiv:2311.08298 [cs.CL] (2023)。http://arxiv.org/abs/ 2311.08298。

767. A.Aldahdooh, W.Hamidouche, S.A.Fezza, O.D é forges，Adve DNN模型的rsarial示例检测: 回顾和实验比较。*Artif.Intell。Rev。***55**，4403-4462 (2022)。https://doi.org/10.1007/s10462-021-10125-w。

768. M. Phute, A.Helbling, M.D.赫尔，S.彭，S。Szyller，C。科尼利厄斯，D.H. Chau，“LLM自我防御: 通过自我检查，LLM知道他们被欺骗了”，在ICLR*的第二篇小论文中2024* (2024)。https://openreview.net/forum？ id= YoqgcIA19o。

769. R。格林布拉特，B。Shleg eris, K.萨尚，F.罗杰，人工智能控制: 提高安全性，尽管故意颠覆，arXiv:2312.06942 [cs.LG] (2023)。https://doi.org/10.48550/arXiv.2312.06942。

770. T.Ploug, S.Holm，“从患者的角度定义透明度和可解释性要求的AI诊断的权利”，“*人工智能医学*”(Springer出版公司，2022)，pp。227-238。

771. 特平，J。迈克尔，E.佩雷斯，S.R。鲍曼，“语言模型并不总是说出他们的想法: 思想链提示中的不忠实解释”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 中 (2023)。https://openreview.net/forum？ id= bzs4uPLXvi。

772. * A.Radhakrishnan, K.阮，A。陈，C。陈，C。丹尼森，D.埃尔南德斯，E。杜尔穆斯，E.哈宾格，J.Kernion, K.Lukošiūtė, N.程，N。约瑟夫，N。Schiefer, O.劳施，S.麦坎德利什，S. El Showk，T.兰汉姆，T。麦克斯韦，V。钱德拉塞卡兰，。。。E. Pe rez，问题分解提高了模型生成推理的忠实性，arXiv:2307.11768 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2307.11768。

773. * J。蔡，E。Rees, H. Batra, S.R。鲍曼，J.迈克尔，E.佩雷斯，M。Turpin，偏差增强一致性训练在思维链中减少了偏差推理，arXiv:2403.05518 [cs.CL] (2024)。https://doi.org/10.48550/arXiv.2403.05518。

774. A.Saranya, R.Subhashini，可解释的人工智能模型和应用的系统回顾: 最近的发展和未来趋势。**7***决策分析杂志*，100230 (2023)。https://doi.org/10.1016/j.Darjour.2023.100230。

775. H。赵，H.陈，F。杨，N。刘，邓海辉，蔡海辉，S。王，D。尹，M。杜，大型语言模型的可解释性: 一项调查。*ACM Trans。Intell。系统。技术。***15**，1-38 (2024)。https://doi.org/10.1145/3639372。

776. I.Seeber, E.比特纳，R。O.布里格斯，T. de Vreede，G.-J。de Vreede, A.埃尔金斯，R.Maier, A.B.Merz, S.Oeste-Reiß, N.Randrup, G.Schwabe, M.S ö llner，“作为队友的机器: 团队协作中人工智能的研究议程”。*Inf。马纳格。103174* **57** (2020)。https://doi.org/10.1016/j.Im.2019.103174。

777. * A.Dafoe, E.休斯，Y。巴赫，T.柯林斯，K。R。麦基，J.Z。雷波，K。拉森，T。Graepel，合作人工智能中的开放问题，arXiv:2012.08630 [cs.AI] (2020)。https://doi.org/10.48550/arXiv.2012.08630 。

778. R。Shah，P。弗莱雷，N。亚历克斯，R。弗里德曼，D。Kr asheninnikov, L.陈，M.D.丹尼斯，P。Abbeel, A.德拉根，S.罗素，帮助胜过奖励学习的好处。(2020)。https://openreview.net/forum？ id=DFIoGDZejIB。

779. A.达福，Y。巴赫拉赫，G.E.哈德菲尔德 .霍维茨，K。拉森，T。Graepel，合作AI: 机器必须学会找到共同点。*自然* **593**，444-36 (2021)。https://doi.org/10.1038/d41586-021-01170-0。

780. X。吴，L。肖，Y。孙，J。Zh ang，T。妈，L.他，一项关于机器学习的人在环的调查。*未来将军。计算机。系统。***135**，364-381 (2022)。https://doi.org/10.1016/j.Future.2022.05.014。

781. J.科恩，E。罗森菲尔德，Z。Kolter，“通过随机平滑证明对抗鲁棒性”，载于第36届国际机器学习*会议论文集* (PMLR，2019) pp。1310-1320。https:// 议事录.mlr.press/v97/cohen19c.html。

782. N.卡里尼，F.Tra mè r, K.D.Dvijotham, L.米，M。孙，J。Z。Kolter，“(免费认证!!) 对抗鲁棒性!”在*第十一届国际学习代表会议 (ICLR 2023)* 2022 ()。https://openreview.net/forum？ id= JLg5aHHv7j。

783. W。聂，B.郭勇。黄，C。肖，A。瓦达特，A。Anandkumar，“对抗性净化的扩散模型”*第39届机器学习*国际会议论文集 (PMLR，2022) pp。16805-16827。https:// 会议录.mlr.press/v162/nie22a.html。

784. A.库马尔，C。Agarwal, S.斯里尼瓦斯，A。J.李，S。Feizi，H. Lakkaraju，证明对抗提示的LLM安全性，arXiv:2309.02705 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2309.02705 。

785. A.周，B。Li，H. Wang，“用于防御越狱攻击的语言模型的鲁棒提示优化”，*ICLR 2024研讨会关于安全和可信赖的大型语言模型* (2024)。https://openreview.net/forum？id= cSPXIO7min。

786. J.巴布科克，J.Kr á 3月，R.V。Yampolskiy，《*下一代伦理学: 工程建设更美好的社会*》中的“人工智能遏制指南”。E。阿巴斯，艾德。(剑桥大学出版社，剑桥，2019)，pp。90-112。

787. J.巴尼亚，J.W。Gichoya, N.马丁内斯-马丁，L.A.沃勒，G.D.Clifford，“事后的公平: 美国对模型开发人员-临床医生用户合作公平性的看法”。**2** *PLOS数字健康*，e0000386 (2023)。https://doi.org/10.1371/journal.Pdig.0000386。

788. N.A. Saxena, K.黄，E。DeFilippis, G.拉达诺维奇，D.C.帕克斯，Y。刘，“公平定义如何？检查公众对公平算法定义的态度”，*载于2019 AAAI/ACM会议关于AI，道德和社会 (AIES '19)* 的会议记录 (计算机协会，2019) pp。99-106。https://doi.org/10.1145/3306618.3314248。

789. W。Fleisher，“个人公平公平是什么？”在《*2021 AAAI/ACM会议关于AI，道德和社会 (AIES '21)* 的论文集》(计算机协会，2021) 中。480-490。https://doi.org/10.1145/3461702.3462621。

790. N.A. Saxena，“公平感”，*载于2019 AAAI/ACM会议关于AI，道德和社会 (AIES '19)* 的会议记录 (计算机协会，2019) pp。537-538。https://doi.org/10.1145/3306618.3314314。

791. R。Binns，“关于个人与群体公平之间的明显冲突”，*载于2020公平，问责制和透明度会议 (FAT * '20)* 的论文集 (计算机协会，2020)。https://doi.org/10.1145/3351095.3372864。

792. N.李，Y。Bang, H. Lovenia, S.Cahyawijaya, W.戴，P。冯，视觉语言模型中的社会偏见调查，arXiv:2309.14381 [cs.CL] (2023)。，https://doi.org/10.48550/arXiv.2309.14381。

793. N.Mehrabi, F.Morstatter, N.萨克塞纳，K。勒曼，A.Galstyan，关于机器学习中的偏见和公平性的调查。*ACM计算机。Surv。***54**，1-35 (2021)。https://doi.org/10.1145/3457607。

794. X。费雷尔，T.范·纽恩，J.M。这样，M。科特，N。Criado，人工智能中的偏见和歧视: 一个跨学科的视角。*IEEE技术。Soc。Mag.***40**，72-80 (2021)。https://doi.org/10.1109/mts.2021.3056293。

795. R.施瓦茨，A.瓦西列夫，K.格林，L.Perine, A.伯特，P。霍尔，“NIST特别出版物1270: 迈向识别和管理人工智能偏见的标准”(美国国家标准与技术研究院 (美国)，2022); 在 https://doi.org/10.6028/nist.sp.1270 。

796. R.Navigli, S.科尼亚，B.罗斯，大型语言模型中的偏见: 起源、库存和讨论。*J.数据和信息质量***15**,1-21 (2023)。的 https://doi.org/10.1145/3597307 。

797. Y。李，M。杜，R。歌，X.王，Y。王，关于大型语言模型公平性的调查，arXiv:2308.10149 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2308.10149。

798. M. Georgopoulos, J.奥德菲尔德，M。A.Nicolaou, Y.Panagakis, M.Pantic，通过基于风格的多属性转移缓解面部数据集中的人口统计偏差。*Int。J、计算机。Vis。***129**，2288-2307 (2021)。https://doi.org/10.1007/s11263-021-01448-w。

799. S。尤瑟，S.阿克赛，N.Al-Moubayed, T.P。Breckon，在*2020 IEEE/CVF计算机视觉和模式识别研讨会 (CVPRW)* 上的“通过每个主题的不利数据增强来探索人脸识别中的种族偏见”(2020)。https://doi.org/10.1109/cvprw50498.2020.00017。

800. D.金，Z。金，Z。胡，欧。Vechtomova, R.Mihalcea，文本风格迁移的深度学习: 一项调查。*计算机。语言学家。***48**，155-205 (2022)。https://doi.org/10.1162/科利_a_00426。

801. A.V. Nadimpalli, A.Rattani，“GBDF: 实现公平DeepFake检测的性别平衡DeepFake数据集”，用于 *模式识别，计算机Visio n和图像处理。ICPR 2022国际研讨会和挑战* (springer-verlag，2023) pp。320-337。https://doi.org/10.1007/978-3-031-37742-6_25。

802. O.Parraga, M.D.更多，C。M。奥利维拉，N.S。加文斯基，L.S。Kupssinskü, A.Medronha, L.V。莫拉，G.S。西蒙斯，R.C. Barros，深度学习的公平性: 视觉和语言研究的调查。*ACM计算机。Surv。*(2023)。https://doi.org/10.1145/3637549。

803. T.王，J。赵，M.Yatskar，K.-W。张，V。Ordonez，“平衡数据集还不够: 估计和减轻深度图像表示中的性别偏见”，*2019 IEEE/CVF国际计算机视觉会议 (ICCV)* (IEEE，2019) pp。5309-5318。https://doi.org/10.1109/iccv.2019.00541。

804. C.-Y。Chuang，Y。Mroueh，“F空气混合: 通过插值实现公平”*在第九届国际学习表征会议 (ICLR 2021)* 中 (2021)。https://openreview.net/forum？id= DNl5s5BXeBn。

805. V。S、洛汉德，A。K。Akash, S.N.拉维，V.Singh，“ *计算机视觉-ECCV 2020 (ECCV 2020)* 中的“ FairALM: 用于训练几乎没有遗憾的公平模型的增强拉格朗日方法 ”(springer-verlag，2020) pp。365-381。https://doi.org/10.1007/978-3-030-58610-2_22。

806. M. Kearns, S.尼尔，A.罗斯，Z。S。Wu，“防止公平Gerrymandering: 审计和学习为子组公平”，*载于第35届机器学习*国际会议论文集 (PMLR，2018)。
https:// 会议录.mlr.press/v80/kearns18a.html。

807. N.霍尔斯比，A。Giurgiu, S.Jastrzebski, B.Morrone, Q.D.Laroussilhe, A.Gesmundo，M。阿塔里扬，S.Gelly，“NLP的有效参数迁移学习”，*载于第36届国际机器学习*会议论文集 (PMLR，2019)。https:// 诉讼。mlr。按/v97/houlsby19a.html。

808. Z。法特米，C。邢，W。刘，C。熊，“在没有灾难性遗忘的情况下改善预训练语言模型的性别公平性”，*载于第61届计算语言学协会年会论文集 (第2卷: 短篇论文)*，(计算语言学协会，2023) pp。1249-1262。
https://doi.org/10.18653/v1/2023.acl-short.108。

809. M.托马林，B.伯恩，S.康坎农，D.桑德斯，S.乌尔曼，机器翻译中减少偏见的实践伦理: 为什么领域适应比数据去偏见更好。*伦理Inf。技术。***23**，419-433 (2021)。
https://doi.org/10.1007/s10676-021-09583-1。

810. L.程，A。Mosallanezhad, P.Sheth，H. Liu，“对社会负责的AI的因果学习”，*载于第30届国际联合人工智能IJCAI-21*会议 (国际人工智能联合会议组织，2021)。https://doi.org/10.24963/ijcai.2021/598。

811. A.Glaese, N.麦卡利斯，M。特朗巴茨，J。阿斯兰尼德，V。Firoiu, T.Ewalds, M.Rauh, L.魏丁格，M.查德威克，P。萨克，L.坎贝尔-吉林厄姆，J.Uesato，P.-S.黄，R。科纳斯库，F.杨，A。看，S.达他里，R。格雷格，C。陈，。。。G. Irving，“通过有针对性的人类判断改善对话主体的一致性” (Google Deepmind，2022); https://doi.org/10.48550/arXiv.2209.14375 。

812. Z.杨，L。李，K。林，J。王，C.-C。林，Z。刘，L。王，LMMs的黎明: GPT- 4V(ision) 的初步探索，arXiv:2309.17421 [cs.CV] (2023)。，https://doi.org/10.48550/arXiv.2309.17421 。

813. Y。李，C。张，G。于，Z。王，B。傅，G。林，C。沈，L。陈，Y。Wei，stablelava: 使用合成图像对话数据进行增强的视觉指令调整，arXiv:2308.10253 [cs.CV] (2023)。，https://doi.org/10.48550/arXiv.2308.10253。

814. W。戴，J.李，D。李，A。Tiong, J.赵，W。王，B。李，P。冯，S。Hoi，“instructiblip: 迈向具有指令调整的通用视觉语言模型”，在*第37届神经信息处理系统会议 (NeurIPS 2023)* 上 (2023)。https://openreview.net/forum？id=vvoWPYqZJA。

815. P。德洛贝尔，B.Berendt，“ *机器学习和数据库知识发现 (ECML PKDD 2022)* 中的“ 公平蒸馏: 减轻语言模型中的刻板印象 ”(springer-verlag，2023) pp。638-654。https://doi.org/10.1007/978-3-031-26390-3_37。

816. K。王，H. Machiraju，O.-H。Choung, M.Herzo g, P.Frossard，CLAD: 一种基于对比学习的背景去偏方法，arXiv:2210.02748 [cs.CV] (2022)。https://doi.org/10.48550/arXiv.2210.02748。

817. 张M，C。R'e，基础模型组鲁棒性的对比适配器。*Adv。神经Inf。过程。系统。***abs/2207.07180** (2022)。https://doi.org/10.48550/arXiv.2207.07180 。

818. 美国。古普塔，J.Dhamala, V.库马尔，A。Verma, Y.Pruksachatkun, S.克里希纳，R.古普塔，K.-W。Chang, G. Ver Steeg, A.Galstyan，“通过反事实角色反转减轻蒸馏语言模型中的性别偏见”，*研究结果为计算语言学协会: ACL 2022* (计算语言学协会，2022) pp。658-678。
https://doi.org/10.18653/v1/2022.findings-acl.55。

819. S。公园，K。Choi，H. Yu，Y。Ko，“学习永远不会太晚: 规范共指解决中的性别偏见”，*载于第16届ACM国际网络搜索和数据挖掘会议 (wsdm'23)* 论文集 (计算机协会，2023) pp。15-23。https://doi.org/10.1145/3539597.3570473。

820. * J。贝克尔，G.Goh, L.静，T。布鲁克斯，J。王，L。李，L。欧阳，J。庄，J。李，Y。郭，W。马纳斯拉，P。Dhariwal, C.Chu, Y.焦，A。Ramesh，“用更好的字幕改善图像生成” (OpenAI，2023)。

821. * A.向，“被看见” 与 “被误看见”: 计算机视觉中隐私与公平之间的紧张关系。**36***哈佛法律与技术*杂志 (2022)。
https://papers.ssrn.com/sol3/papers.cfm？abstract_id = 4068921。

822. G.普莱斯，M.Raghavan, F.吴，J。克莱因伯格，K。问:Weinberger，*第31届神经信息处理系统会议 (NIPS 2017)* 中的“公平与校准” (Curran Associates，inc.，2017) 第一卷。30。
https:// papers.nips.cc/paper_files/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html。

823. Z。施，Y。王，F。尹X。陈，K.-W。Chang，C.-J.Hsieh，Red将语言模型检测器与语言模型结合在一起。*Trans。Assoc。计算机。语言学家。***12**，174-189 (2024)。https://doi.org/10.1162/tacl_a_00639。

824. Y。刘，K。张，Y。李，Z。严，C。高，R。陈，Z。元，Y。黄，孙H，J。高，L。他，L.Sun，Sora: 关于大视觉模型的背景，技术，局限性和机会的回顾，arXiv:2402.17177 [cs.CV] (2024)。
https://doi.org/10.48550/arXiv.2402.17177。

825. * S.慕克吉，A.米特拉，G.贾瓦哈尔，S.阿加沃尔，H.帕兰吉，A。Awadallah，Orca: 从复杂的解释GPT-4痕迹中逐步学习，arXiv:2306.02707 [cs.CL] (2023)。https://doi.org/10.48550/arXiv.2306.02707 。

826. T.索伦森，J。摩尔，J。费舍尔，M。戈登，N.米什哈拉，C。M。Rytting, A.是的，L。江，X。Lu, N.Dziri, T.Althoff, Y.Choi，“多元对齐路线图”，arXiv:2402.05070 [cs.AI] (2024)。http://arxiv.org/abs/ 2402.05070。

827. M.Shur-ofry，“多重性作为人工智能治理原则”。(2023)。https://papers.ssrn.com/sol3/papers.cfm？abstract_id = 4444354。

828. A.Chouldec hova，具有不同影响的公平预测: 累犯预测工具中的偏见研究。*大数据*的**5**，153-163 (2017)。https://doi.org/10.1089/大2016.0047。

829. J.Kleinberg，“算法公平性的固有权衡”，*摘自2018 ACM国际会议计算机系统的测量和建模 (SIGMETRICS '18)* (计算机机械协会，2018) pp。40.https://doi.org/10.1145/3219617.3219634。

830. 问:张，J。刘，Z。张，J。温，B。毛，X。Y ao，通过进化多目标集成学习减轻不公平。*IEEE Trans。Evol。计算机。***27**，848-862 (2023)。https://doi.org/10.1109/tevc.2022.3209544。

831. M.哈特，E.价格，E。价格，N。Srebro，“监督学习中的机会均等”在*第30届会议上神经信息处理系统 (NIPS 2016)* 中 (Curran Associates，inc.，2016) 卷。29。https:// 会议录.neurips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html。

832. A.格罗弗，J.歌，A.卡普尔，K。Tran, A.阿加沃，E。J.霍维茨，S.Ermon，“使用无似然重要性加权对学习的生成模型进行偏差校正”，在*第33届神经信息处理系统会议 (NeurIPS 2019)* 上 (Curran Associates，inc.，2019) 第一卷。32。https:// 会议录.neurips.cc/paper/2019/hash/d76d8deea9c19cc9aaf2237d2bf2f785-Abstract.html。

833. S。阿罗拉，A.Risteski, Y.张，“甘斯学会分配吗？*第六届国际学习表征会议 (ICLR)* 中的“一些理论和经验”(2018)。https://openreview.net/forum？id= BJehNfW0-。

834. D.张，S。潘，T.Hoang, Z.邢，M。斯台普斯，X。徐，L。姚，Q。Lu, L.朱，被遗忘还是公平: 揭示机器学习方法的公平含义。**4***人工智能与伦理*，83-93 (2024)。https://doi.org/10.1007/s43681-023-00398-y。

835. H。Nilforoshan, J.D.盖布勒，R。Shroff, S.Goel，“公平及其后果的因果概念”*第39届国际会议机器学习 (ICML 2022)* 论文集 (PMLR，2022)。https:// 议事录.mlr.press/v162/nilforoshan22a.html。

836. N.康斯坦丁诺夫，C。H. Lampert，“论从损坏的数据中学习公平意识的不可能性”*算法公平性通过因果关系和鲁棒性研讨会 (AFCR 2021)* 的镜头 (PMLR，2021)。https:// 会议记录.mlr.press/v171/konstantinov22a.html.

837. M、金砖四国，R。V。Yampolskiy，AI中的不可能结果: 一项调查。*ACM计算机。Surv。***56**，1-24 (2023)。https://doi.org/10.1145/3603371。

838. K。T、罗道法，H.兰巴，R。Ghani，公共政策机器学习中公平-准确性权衡的经验观察。*自然机器智能*896 **3** -904 (2021)。https://doi.org/10.1038/s42256-021-00396-x。

839. B.格林，Esc探讨公平的不可能性: 从形式到实质算法公平。*Philos.技术。***35**，90 (2022)。https://doi.org/10.1007/s13347-022-00584-6。

840. A.贝尔，L.拜纳姆，N.Drushchak, T.扎哈尔琴科，L.罗森布拉特，J。Stoyanovich，“公平的可能性: 在实践中重新审视不可能性定理”，*《2023 ACM会议关于公平，问责制，和透明度 (FAccT '23)》* (计算机协会，2023) pp。400-422。https://doi.org/10.1145/3593013.3594007。

841. C.赫特维克，T.R ä z，公平标准的逐步 (不) 兼容性。*AAAI* **36**，11926-11934 (2022)。https://doi.org/10.1609/aaai.v36i11.21450。

842. S。卡顿，C。哈斯，机器学习的公平性: 一项调查。*ACM Comput.Surv。***56**，1-38 (2024)。https://doi.org/10.1145/3616865。

843. S。古哈，F。A.Khan, J.Stoyanovich, S.Schelter，“自动数据清理会损害基于机器学习的决策的公平性”，*2023 IEEE第39届国际数据工程会议 (ICDE)* (IEEE，2023) pp。3747- 3754。https://doi.org/10.1109/icde55515.2023.00303。

844. B.Ghai, M.N.霍克，K。Mueller，“单词偏差: 一种交互式可视化工具，用于发现单词嵌入中编码的交叉偏差”，在*扩展的2021 CHI C计算机系统人为因素会议 (CHI ea'21)* 中 (计算机协会，2021) pp。1-7。https://doi.org/10.1145/3411763.3451587。

845. C.Dwork, F.McSherry, K.Nissim, A.Smith，“在私人数据分析中将噪声校准为敏感性”，*《密码学》*，S。Halevi, T.拉宾，编辑。(计算机科学讲义，施普林格，柏林，海德堡，2006)，第一卷。3876。

846. M.阿巴迪，A.楚，我。古德费罗，H. B.麦克马汉，我米罗诺夫，K。Talwar, L.Zhang，“具有差分隐私的深度学习”，*载于2016 ACM SIGSAC会议计算机和通信安全 (ccs'16)* 论文集 (计算机机械协会，2016) pp。308-318。https://doi.org/10.1145/2976749.2978318。

847. H。布朗，K。李，F。Misherghallah，R。Shokri, F.Tram è r，“语言模型保护隐私意味着什么？”在*《2022 ACM会议公平，问责制和透明度》(FAccT '22)* 中 (计算机协会，2022) pp。2280-2292。https://doi.org/10.1145/3531146.3534642。

848. S。德，L。Berrada, J.海耶斯，S.L.史密斯，B。Balle，“通过尺度解锁高精度差分私有图像分类”(Google Deepmind，2022)；http://arxiv.org/abs/ 2204.13650。

849. X。李，F。特拉梅尔，P.梁，T。桥本，“大型语言模型可以成为强大的差异化私人学习者”*国际学习表征2022*会议 (2022)。https://openreview.net/forum？id= bVuP3ltATMz。

850. T.S tadler, B.Oprisanu, C.特隆科索，合成数据-匿名化土拨鼠日。(USENIX协会，2022)，pp.1451-1468。https://www.usenix.org/conference/ usenixsecurity22/presentation/stadler。

851. M. Meeus, F.A，Guepin。-M.Creţu, Y. -A.de Montjoye，Achil leles的高跟鞋: 合成数据发布中的脆弱记录识别。(斯普林格自然瑞士，2024)，pp。380-399。https://doi.org/10.1007/978-3-031-51476-0_19。

852. G.Ganev，E. De Cristofaro，关于基于相似性的隐私度量的不足: 针对“真正匿名合成数据”的重建攻击，arXiv:2312.05114 [cs.CR] (2023)。https://doi.org/10.48550/arXiv.2312.05114。

853. P。莫哈塞尔，Y。张，SecureML: 一个可扩展的隐私保护机器学习系统。(IEEE计算机学会，2017)，pp。19-38。https://doi.org/10.1109/sp.2017.12。

854. B.麦克马汉，E.摩尔，D。Ramage, S.汉普森，B。A. y.Arcas，通信-从分散的数据中高效地学习深度网络。(PMLR，2017)，第1273-1282。https:// 诉讼。mlr。按/v54/mcmahan17a.html。

855. O.Ohrimenko, F.舒斯特，C。Fournet, A.Mehta, S.诺沃津，K.Vaswani, M.Costa，在可信处理器上进行遗忘的多方机器学习。(USENIX协会，2016)，p p。619 -636。https://www.usenix.org/conference/ usenixsecurity16/技术-会议/演示/ohrimenko。

856. T.李，E。F.维拉隆加，P。Kieseberg，人类忘记，机器记住: 人工智能和被遗忘的权利。《**34***计算机法律与安全评论*》，304 (2018)。https://scholarship.law.bu.edu /学院_奖学金/817。

857. A.Ghorbani, J.Zou，“Data Shapley: 机器学习数据的公平评估”，*载于第36届国际机器学习会议 (ICML 2019)* 论文集 (PMLR，2019)。97页。2242-2251。https:// 会议录.mlr.press/v97/ghorbani19c.html。

858. M. ElBaih，隐私法规在人工智能发展中的作用 (讨论隐私法规可以塑造人工智能发展的方式)。(2023)。https://doi.org/10.2139/ssrn.4589207。

859. A.Cavoukian，设计隐私: 7个基本原则。(2009)。

860. 欧洲议会，人工智能法案: 关于可信赖人工智能的全面规则的交易 (2023)。https:// www.europarl.europa.eu/news/en/新闻室/20231206IPR15699/人工智能-行动-交易-全面-值得信赖的规则-ai。

861. T.S tadler, B.Kulynych, N.Papernot, M.Gastpar，C。Troncoso，最小特权学习的基本限制。(arXiv，2024)。https://doi.org/10.48550/arXiv.2402.12235。

如有任何有关本出版物的查询，请发送至：
secretariat.AIStateofScience@dsit.gov.uk

研究系列编号: DSIT 2024/009

2024年5月由英国政府出版

AI SEOUL SUMMIT
21 - 22 MAY 2024
Hosted by the Republic of Korea and the United Kingdom