国外产业动态

长江产业经济研究院（南京大学） 2024 年 7 月

2024 年
第 2 期

目 录

主编：叶明

助理编辑：尹衍斐

长江产业经济研究院
Yangtze IDEI

南京大学长江产业经济研究院

地址：南京大学鼓楼校区北园丙丁楼

网址：idei.nju.edu.cn

邮箱：idei@nju.edu.cn

微信：长江产经智库

2024 年初人工智能发展现状[1]

B 如果说 2023 年是世界发现生成式人工智能（Gen AI）的一年，那么 2024 年就是企业真正开始使用这项新技术并从中获取商业价值的一年。在麦肯锡关于 Gen AI 的最新全球调查中，65% 的受访者表示，他们的组织正在定期使用 Gen AI，这一比例比十个月前的上一次调查高出近一倍。受访者对 Gen AI 影响的预期与去年一样高，四分之三的受访者预测 Gen AI 将在未来几年给他们所在的行业带来重大或颠覆性的变革。在此背景下，麦肯锡公司于 2024 年 5 月 30 日发布的报告，深入分析了 Gen AI 所带来的各种风险以及表现最出色的公司为减轻这些挑战和获取价值而采取的新做法。

一、人工智能采用率激增

人们对 Gen AI 的兴趣使得更广泛的人工智能也受到关注。在过去六年中，受访企业采用人工智能的比例一直徘徊在 50% 左右。今年，麦肯锡调查发现这一采用率已跃升至 72%。而这种现象确实是全球性的：2023 年调查发现，任何地区的人工智能采用率都没有达到 66%；然而，今年几乎每个地区都有超过三分之二的受访者表示他们的组织正在使用人工智能。此外，受访者的回答还表明，企业目前正在将人工智能应用于更多业务领域。半数受访者表示，他们的企业已在两个或更多业务领域中采用了人工智能，而在 2023 年，这一比例还不到三分之一。最新调查还显示大多数受访者（67%）预计他们的企业将在未来三年内加大对人工智能的投资。

Gen AI 正在渗透到受访者的个人生活中。与 2023 年相比，受访者更有可能在工作中使用 Gen AI，甚至更有可能在工作和个人生活中都使用 Gen AI。调查发现，所有地区的 Gen AI 使用率都有所上升，其中亚太地区和大中华区的增幅最大。同时，与中层管理者相比，最高级别的受访者在工作和工作之外使用 Gen AI 的比例大幅上升。从具体行业来看，能源和材料行业以及专业服务行业的受访者对 Gen AI 的使用增幅最大。

这些投资在哪里得到了回报？麦肯锡的最新调查首次按业务领域探讨了 Gen AI 的使用所创造的价值。受访者认为成本下降最多的领域是人力资源。受访者最常报告的是供应链和库存管理方面有意义的收入增长（超过 5%）。在分析新人工智能方面，受访者最常报告的是在服务运营以及在营销和销售中使用人工智能带来的有意义的收入增长。

---

1 资料来源：McKinsey 发布报告 The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. 2024.5.30.

二、Gen AI 可能存在的风险

随着企业开始认识到 Gen AI 的优势，他们也意识到了与该技术相关的各种风险。这些风险既包括数据管理风险，如数据隐私、偏见或知识产权（IP）侵权；也包括模型管理风险，其重点往往是输出不准确或缺乏可解释性；当然安全和使用不当也是该技术的风险之一。

与去年相比，最新调查的受访者表示，他们的组织认为不准确性和知识产权侵权与他们使用 Gen AI 有关，约有一半的受访者继续将网络安全视为一种风险。一些组织（44%）已经经历了使用 Gen AI 带来的负面后。受访者最常报告的影响其组织的风险是不准确性，其次是网络安全和可解释性。

最新调查还试图了解企业如何以及如何快速部署 Gen AI 的使用。报告发现，有三种方法可以用来解决 Gen AI 所产生的问题：接受者使用现成的、公开可用的解决方案；塑造者利用专有数据和系统定制这些工具；以及制造者从头开始开发自己的基础模型。在受访者的业务领域中，约有一半的 Gen AI 使用的是现成的、公开可用的模型或工具，很少或根本没有进行定制。能源和材料、技术以及媒体和电信领域的受访者更有可能对公开可用的模型进行大量定制或调整，或开发自己的专有模型来满足特定的业务需求。

三、如何应对挑战

Gen AI 是一项新技术，各组织仍处于探索其机遇并在各职能部门推广的早期阶段。因此，只有一小部分受访者（876 位受访者中的 46 位）表示，他们组织的税前利润中有相当大的一部分可以归功 Gen AI。尽管如此，这些人工智能领导者仍然值得仔细研究。毕竟，这些都是先行者，他们已经将其组织 10% 以上的税前利润归功于对 Gen AI 的使用。这些企业的人工智能相关实践可以为那些希望通过在自己的企业采用 Gen AI 创造价值的人提供指导。

首先，通过 Gen AI 获得高绩效的企业在更多的业务领域中使用 Gen AI 平均三个领域，而其他企业平均两个。它们与其他组织一样，最有可能在营销和销售以及产品或服务开发中使用 Gen AI，但在风险、法律和合规、战略和企业财务以及供应链和库存管理中使用 Gen AI 解决方案的可能性要比其他组织高得多。它们在会计文件处理、风险评估、研发测试、定价和促销等活动中使用 Gen AI 的可能性是其他公司的三倍多。总体而言，在所报告的业务领域中，约有一半的 Gen AI 使用的是公开可用的模型或工具。但这些企业不太可能使用这些现成的选择，而是实施这些工具的重大定

制版本或开发自己的专有基础模型。

这些绩优企业还有哪些不同的做法？首先，它们更加关注与 Gen AI 相关的风险。它们经历过 Gen AI 带来的每一种负面后果，从网络安全和个人隐私到可解释性和知识产权侵权。有鉴于此，它们比其他人采取了更多措施来降低风险。这些企业也更有可能遵循一套与风险相关的最佳实践。例如，在开发 Gen AI 解决方案的初期，它们让法律职能部门参与进来并嵌入风险审查的可能性几乎是其他组织的两倍。此外，它们还比其他公司更有可能采用其他各种最佳实践，从与战略相关的实践到与扩展相关的实践。

2024 年人工智能指数报告[2]

由斯坦福大学人工智能研究所发布的人工智能指数被全球公认为见解最可信、最权威的人工智能数据和。2024 年人工智能指数是迄今为止最全面的指数，它的发表正值人工智能对社会的影响前所未有的重要时刻。相比于往年，该报告扩大了研究范围，更广泛地涵盖了人工智能的技术进步、公众对该技术的看法以及围绕其发展的地缘政治动态等内容。该报告希望能够为决策者、研究人员、高管、记者和公众等提供对人工智能的复杂领域更彻底、更细致的了解。

一、人工智能的研究和发展

该报告指出，当前人工智能的发展有以下几个特点：工业继续主导前沿人工智能研究；出现了更多开放的基础模型；前沿人工智能模型变得更加昂贵；美国领先中国、欧盟和英国，成为顶级人工智能模型的主要来源；人工智能专利数量激增；中国主导人工智能专利；开源人工智能研究爆炸式增长；人工智能出版物的数量持续增加。

二、技术进步

人工智能在某些任务上例如图像分类、英语理解胜过人类，但并非全部例如数学和规划等。多模式人工智能如谷歌的 Gemini 和 OpenAI 的 GPT-4 应运而生。这些模型展示了灵活性，能够处理图像和文本，在某些情况下，甚至可以处理音频。同时，由于大语言模型的发展，机器人变得更加灵活。语言建模与机器人技术的融合产生了更灵活的机器人系统，这标志着机器人向更有效地与现实世界互动迈出了重要一步。

三、目前的不足

目前，我们仍然严重缺乏对大语言模型的有力和标准化评估；使用 AI 干涉选举的现象难以避免；大语言模型中依然存在巨大的漏洞；人工智能的风险包括隐私、数据安全和可靠性等正成为全球企业关注的问题，但在全球范围内，大多数公司迄今为止只减轻了这些风险的一小部分；大语言模型可以输出包含受版权保护的材料例如《纽约时报》的节选，这可能导致破坏版权；人工智能开发人员在透明度方面得分较低，这种开放性的缺乏阻碍了进一步理解人工智能系统的稳健性和安全性的努力；滥用人工智能的事件数量持续上升，例如使用人工智能生成泰勒·斯威夫特的色情作品；ChatGPT 有政治偏见。研究人员发现，ChatGPT 对美国民主党和英国工党存在重大偏见。这一发现

2 资料来源：Institute for Human-Centered AI, Stanford University 发布报告 Artificial Intelligence Index Report 2024. 2024.4.

引发了人们对该工具影响用户政治观点的可能性的担忧，尤其是在全球重大选举的年份。

四、人工智能与经济发展

尽管去年对于人工智能的总投资下降，但是对于生成式 AI 例如 Open AI 的投资增加；美国作为领先者，在人工智能私人投资方面进一步领先；美国和全球人工智能工作岗位减少，这可以归因于领先的人工智能公司发布的职位减少，以及这些公司中技术职位的比例降低；人工智能降低了成本，增加了收入，并显著提高业务效率；人工智能私人投资总额再次下降，而新资助的人工智能公司数量增加；人工智能在企业组织中的应用越来越多；中国在工业机器人领域占据主导地位；机器人装置的多样性更大，例如工业机器人、医疗机器人等；人工智能使工人更有效率，并带来更高质量的工作；《财富》500 强公司开始大量谈论人工智能，尤其是生成式人工智能。

五、人工智能与科技和医疗

得益于人工智能，科学进步进一步加速。2022 年，人工智能开始推动科学发现。然而，2023 年推出了更重要的科学相关的人工智能应用——从使算法排序更高效的 AlphaDev 到促进材料发现过程的 GNoME。同时，人工智能帮助医学取得重大进展。2023 年，推出了几个重要的医疗系统，包括增强疫情预测的 EVEscape 和帮助人工智能驱动的突变分类的 AlphaMissence。人工智能正越来越多地被用于推动医学进步。美国食品药品监督管理局批准了越来越多与人工智能相关的医疗设备，人工智能越来越多地被用于现实世界的医疗目的。

六、人工智能与教育

美国和加拿大计算机科学学士毕业生的数量继续上升，新硕士毕业生保持相对平稳，博士毕业生略有增长；人工智能博士向工业的迁移仍在加速。而学术人才从工业向学术的转变较少，这表明大学向行业的人才流失正在加剧；美国和加拿大的计算机科学教育变得不那么国际化，硕士类国际学生的减少尤为明显；越来越多的美国高中生选修计算机课程，但入学问题依然存在；人工智能相关学位项目在国际上呈上升趋势；英国和德国在计算机等领域的毕业生生产方面处于领先地位。

七、人工智能与政府和政策

美国的人工智能法规数量急剧增加；美国和欧盟推进了具有里程碑意义的人工智能政策行动，例如欧盟的《人工智能法案》以及拜登总统签署的一项关于人工智能的行政命令；人工智能逐渐吸

引美国决策者的注意力，联邦层面的人工智能立法显著增加；同时，全球的立法者对于人工智能的关注都显著增加，越来越多的监管机构将注意力转向人工智能。

八、人工智能与多样性

美国和加拿大的计算机科学领域的学士、硕士和博士学生继续在种族上变得更加多样化；欧洲计算机等领域的毕业生在所有教育水平上都存在巨大的性别差距。尽管过去十年中，大多数国家的性别差距有所缩小，但缩小的速度一直很慢。美国 K-12 的计算机教育日益多样化，反映了性别和种族代表性的变化。

九、人工智能与公众

世界各地的人们更加认识到人工智能的潜在影响，也更加紧张；西方国家的人工智能情绪持续低迷，但正在缓慢改善；公众对人工智能的经济影响持悲观态度；在人工智能乐观主义方面出现了人口统计学差异。对人工智能提高生存潜力的认知存在显著的人口差异，年轻一代普遍更乐观。此外，收入和教育水平较高的人比收入较低、教育程度较低的人更乐观地看待人工智能对娱乐、健康和经济的积极影响；ChatGPT 广为人知并应用广泛。

国际科学报告：先进人工智能的安全性[3]

该报告受英国政府委托，由图灵奖获得者、联合国科学顾问委员会成员 Yoshua Bengio 领头撰写。工作由一个国际专家顾问小组监督，该小组由包括英国在内的 30 个国家、应邀参加 2023 年在布莱切利公园举行的人工智能安全峰会的各国提名人以及欧盟和联合国的代表组成。该报告侧重于识别通用人工智能的风险，并评估和减轻风险的技术方法，包括有益地使用通用人工智能来减轻风险。

一、通用人工智能的发展

根据许多指标，通用人工智能能力进展迅速。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。如今，一些通用的人工智能模型可以就广泛的主题进行多回合对话，编写简短的计算机程序，或根据描述生成视频。同时，近年来人工智能开发人员通过不断增加用于训练新模型和改进现有算法的资源，快速提升了通用人工智能能力。通用人工智能能力的未来进展速度对管理新出现的风险具有重大意义。

二、当前面临的挑战

管理通用人工智能风险的方法通常基于这样一种假设，即人工智能开发人员和决策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，该假设面临一些关键挑战：

首先，开发人员对他们的通用人工智能模型是如何运作的仍然知之甚少。通用人工智能模型可以由数万亿个被称为参数的组件组成，它们的大部分内部工作是不可理解的，包括对模型开发人员来说。其次，通用人工智能主要通过在各种输入上测试模型或系统来进行评估。这些测试往往忽略了危险，高估或低估了能力；再次，原则上独立参与者可以审计公司开发的通用人工智能模型或系统。然而，公司通常不会为独立审计师提供必要水平的直接访问模型或严格评估所需的数据和方法的信息；最后，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不足以产生严格和全面的评估方法。

三、通用人工智能的不利影响

该报告将通用人工智能风险分为三类：恶意使用风险、故障风险和系统风险。

恶意使用：像所有强大的技术一样，通用的人工智能系统可以被恶意使用，造成伤害。可能的

3 资料来源：Department for Science, Innovation and Technology and AI Safety Institute of UK 发布报告 International Scientific Report on the Safety of Advanced AI. 2024.5.17.

恶意使用类型既包括证据相对充分的类型，如通过通用人工智能增强的"网络钓鱼"攻击、恶意使用通用人工智能进行虚假信息和操纵公众舆论等；也包括一些专家认为可能在未来几年发生的类型，例如一些专家对通用人工智能可能被用于支持生物武器等武器的开发和恶意使用表示担忧。尽管目前没有强有力的证据表明通用人工智能系统会带来这种风险，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，通用人工智能的故障也可能导致严重风险。这种故障可能有几种原因：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差；同时，人工智能系统中的偏差也是一个尚未解决的问题，例如通用人工智能输出可能在种族、性别、文化、年龄和残疾等受保护特征方面存在偏见。

系统性风险：通用人工智能技术的广泛开发和采用带来了几个系统性风险，从潜在的劳动力市场影响到隐私风险和环境影响。通用人工智可能会导致许多人失去目前的工作。通用人工智能的研发目前集中在少数西方国家和中国，这导致低收入国家和学术机构处于不利地位。通用人工智能的市场集中，例如在金融或医疗保健等关键部门的广泛使用，使社会更容易受到系统性风险的影响。通用人工智能开发和部署中日益增长的计算机使用迅速增加了与通用人工智能相关的能源使用以及二氧化碳排放。通用人工智能模型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知识产权法以及同意、补偿和数据控制系统构成了挑战。

四、如何应对挑战

虽然该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减轻方法。尽管取得了这一进展，但目前的方法并没有可靠地防止现实世界中甚至是明显有害的通用人工智能输出。

在训练通用人工智能模型以更安全地运行方面学者已经取得了一些进展。将通用人工智能系统的功能限制在特定的用例中，有助于降低意外故障或恶意使用的风险。通用人工智能系统部署后，有几种技术可用于识别风险、检查系统动作和评估性能。这些做法通常被称为"监控"。通用人工智能系统中的偏见缓解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，在通用人工智能系统中完全防止偏见是一项挑战，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，

什么是不应反映在输出中的不良偏见。隐私保护是一个活跃的研究和发展领域。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，当敏感数据被有意或无意地使用时，现有的降低隐私风险的技术工具很难扩展到大型通用人工智能模型，并且可能无法为用户提供有意义的控制。

五、结论

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发展的，由谁来开发，它被设计来解决哪些问题，社会是否能够充分利用通用人工智能的经济潜力，谁从中受益，我们面临的风险类型，以及我们在研究中投资多少来减轻风险——这些和许多其他问题取决于社会和政府今天和未来为塑造通用人工智能发展所做的选择。

人工智能的经济影响及其监管[4]

由国际货币基金组织发表的工作论文回顾了有关人工智能的经济影响以及监管人工智能的文献。其中，人工智能的经济影响包括增长、就业、生产力和收入不平等，而监管涵盖市场竞争、数据隐私、版权、国家安全、道德问题和金融稳定。

一、人工智能的经济影响

人工智能对劳动力市场的影响是深远而复杂的。理论上，AI 技术将影响几乎所有职业，并有可能改变增长模式。然而，实证研究对于 AI 对就业和生产力的具体影响尚未得出明确结论。现有文献普遍认为，高技能白领职业因任务暴露度高而面临更大的就业风险，但同时也有研究指出 AI 技术的增强潜力和其他"保护因素"，可能会减轻这种风险。例如，实验性论文表明，生成性 AI（Gen AI）可能为低技能工人带来生产力收益，这可能缓解工资降低的负面影响。政策制定者面临的挑战是如何平衡 AI 技术带来的潜在劳动力置换和生产率提升。一些建议性政策包括提高劳动力的适应性和技能培训，以及调整税收和财政政策，以促进技术选择，使人类劳动力与 AI 技术相辅相成。

人工智能对生产力和增长的影响同样重要，但相比劳动力市场的研究，这一领域的实证研究较少。理论上，AI 技术的采纳可能带来显著的生产率提升，但其实际效果仍不确定。公司层面的研究表明，采用 AI 的公司可能会实现每个工人的销售增长，但这种增长的具体幅度差异较大。此外，AI 技术的采纳被认为是实现生产率增长的关键途径，但目前关于 AI 采纳率的估计仍然有限。对于新兴市场和发展中经济体（EMDEs），AI 可能带来的增长和生产力溢出效应可能比就业置换效应更为重要。AI 作为发展工具，可以为发展中国家创造满足特定需求的新产品，并通过快速扩展的解决方案，有助于技术向较贫穷的人群传播。然而，要实现这些好处，需要在基础设施和教育上进行大量投资。同时，数据的可用性和对本地语言的模型训练是关键的制约因素。国际合作在促进 AI 技术的安全和最佳实践方面发挥着重要作用，不同国家和地区在 AI 监管方面采取了不同的方法和范围，这反映了它们在创新优势和潜在风险之间的权衡。

二、对于人工智能的监管

人工智能的快速发展也带来了一系列监管挑战，这些挑战要求政策制定者在促进创新和保护公共利益之间找到平衡。监管的主要关注点包括市场竞争、数据隐私、版权问题、国家安全、伦理问

4 资料来源：国际货币基金组织发表工作论文 The Economic Impacts and the Regulation of AI: A Review of the Academic Literature and Policy Actions. 2024.3.

题以及金融稳定。例如，AI 可能加剧市场垄断，因为大型技术公司通过访问大量数据和算法优势获得市场力量。此外，AI 系统处理个人数据的方式引发了隐私保护的担忧，而其在内容创造和分发中的使用也对版权法提出了新的问题。在国家安全方面，AI 的军事应用和网络攻击的潜力需要严格的监管框架来确保安全。伦理问题，特别是算法偏见和歧视，要求 AI 系统的设计和部署必须透明和公正。金融稳定也是监管者关注的重点，因为 AI 在金融领域的应用可能带来系统性风险。不同国家和地区采取了不同的监管方法，从欧盟的风险基础方法到美国的分散指导方针，再到中国的算法推荐和伦理审查，显示出全球在 AI 监管上的多样性和复杂性。

面对这些挑战，监管机构和国际组织正在采取行动制定相应的政策和框架。例如，欧盟提出了 AI 法案，旨在通过风险评估和透明度要求来规范高风险 AI 系统。美国则通过行政命令强调了 AI 的安全、安保和可信度，并鼓励发展伦理标准。中国则专注于通过临时措施来管理生成性 AI 服务，并强调了国家安全和社会公共利益的保护。此外，国际合作在 AI 监管中发挥着关键作用，如经合组织（OECD）的 AI 原则和联合国的 AI 咨询机构等。这些努力旨在确保 AI 技术的发展和应用能够符合全球共同的价值观和利益。

三、未来方向

研究和政策的未来方向需要关注几个关键领域。首先，需要更多的跨学科研究来更好地理解 AI 技术的经济和社会影响，以及如何设计有效的监管策略。其次，随着 AI 技术的不断演进，监管框架必须具有足够的灵活性和适应性，以应对新兴的挑战和风险。此外，国际合作在制定全球 AI 治理标准和原则方面将变得更加重要，以确保技术的健康发展不会受到不必要的限制，同时保护全球公民的权益。最后，政策制定者、学者和行业利益相关者之间的持续对话对于确保 AI 技术惠及社会各个方面至关重要。通过这些努力，我们可以期待一个更加安全、公平和可持续的 AI 技术未来。

利用人工智能应对全球挑战[5]

总统科学技术顾问委员会（PCAST）是一个由总统任命的联邦咨询委员会，旨在向总统提供相关的科学建议。2024 年 4 月，该委员会向美国总统拜登提交报告，探讨了应用人工智能应对重大社会和全球挑战的可能性。

一、人工智能带来的机遇与挑战

凭借精心设计、公平共享和负责任地使用的基础设施，人工智能将使科学家能够应对紧迫的挑战，包括在气候变化时期改善人类健康和加强天气预测。人工智能可以帮助探索激发和拓展人类创造力的长期科学奥秘，例如揭示宇宙的起源和进化。人工智能还将帮助研究人员解决持续的国家需求，从加速半导体芯片设计到发现新材料来满足我们的能源需求。此外，人工智能正在开始消除使科学研究缓慢而昂贵的障碍，例如，通过提供快速确定最佳候选药物进行测试的手段（从而减少昂贵的实验室试验的数量），帮助优化实验设计，以及比手工或使用传统数据科学方法更有效地揭示数据中的联系。

正如任何其他新工具或技术一样，实现人工智能的潜力需要解决其局限性。这些问题包括误导性或不正确的结果、偏见或不公平的持续存在以及模型训练数据中嵌入的模式的采样错误、对高质量训练数据的访问受限、保护知识产权和隐私的挑战、训练或部署模型或运行人工智能算法所需的大量精力，以及不良或邪恶行为者将现成的人工智能工具用于恶意目的的风险。

二、如何充分利用人工智能的潜力

扩大现有努力，广泛、公平地分享人工智能的基本资源。广泛支持可广泛访问的共享模型、数据集、基准和计算资源，对于确保学术研究人员、国家和联邦实验室、小型公司和非营利组织能够使用人工智能为国家创造利益至关重要。在美国，这方面最有希望的努力是国家人工智能研究资源（NAIRR），该项目目前是一个试点项目。PCAST 建议尽快将 NAIRR 试点扩大到设想的规模，并提供全额资金。全面的 NAIRR，加上联邦和州层面的行业合作伙伴关系和其他人工智能基础设施努力，可以成为国家或国际层面人工智能基础结构项目的垫脚石，以促进高影响力的研究。

扩大对联邦数据集的安全访问，以满足批准的关键研究需求，并提供适当的保护和保障。允许经批准的研究人员有限、安全地访问联邦数据集，以及允许向 NAIRR 等精心策划的资源中心发布此

5 资料来源：The President’s Council of Advisors on Science and Technology (PCAST) of US 发布报告 Supercharging Research: Harnessing Artificial Intelligence to Meet Global Challenges. 2024.4.

类数据集的匿名版本，其好处是巨大的。PCAST 强烈鼓励扩大现有的安全数据访问试点项目，并制定联邦数据库管理指南，其中包括尖端的隐私保护技术。使用现代人工智能技术实现此类数据集管理的自动化具有巨大潜力。PCAST 鼓励使用人工智能改善数据管理，将其作为 data.gov 等联邦数据共享举措的长期目标。PCAST 支持联邦机构授权负责任地共享其资助或开展的研究所产生的数据集的努力。同时 PCAST 鼓励进一步执行此类授权，包括共享根据联邦资助的研究数据训练的人工智能模型，以及支持所需行动的充足资源。

支持人工智能的基础和应用研究，包括学术界、工业界、国家和联邦实验室以及联邦机构之间的合作。联邦政府资助的学术研究和私营部门研究之间的界限是模糊的。许多研究人员在学术机构、非营利组织和 / 或私营公司的附属机构之间流动，在所有人工智能研发（R&D）中有很大一部分是私人企业支持的。为了充分利用人工智能对科学的潜在好处，必须支持涉及广泛有前景和富有成效的假设和方法的研究。这可能需要资助机构扩大其在如何与行业合作以及哪些研究人员可以得到支持方面的姿态，以促进不同部门之间的创新研究努力和合作。这种合作的例子可以包括从多个来源创建高质量的公共科学数据集，或者创建多模式的基础模型。

在科学研究过程的所有阶段，采用负责任、透明和值得信赖的人工智能使用原则。管理人工智能科学应用中不准确、有偏见、有害或不可复制的发现的风险，应从研究项目的初始阶段就进行规划，而不是事后考虑。PCAST 建议联邦资助机构考虑更新其负责任的研究行为准则，要求研究人员制定负责任的人工智能使用计划。这些计划应包括机构办公室和委员会提出的解决潜在人工智能相关风险的建议最佳实践，并描述任何自动化流程的使用监督程序。为了最大限度地减少研究人员的额外行政负担并建立责任文化，在列举主要风险后，机构应提供风险缓解的模型流程。与此同时，美国国家科学基金会（NSF）和美国国家标准与技术研究所（NIST）等机构应继续支持负责任和值得信赖的人工智能的科学基础研究。这项研究应包括开发标准基准，以衡量人工智能模型的特性，如准确性、再现性、公平性、弹性和可解释的人工智能，以及监测这些特性并在基准不在规定规范范围内时进行调整的人工智能算法。此类研究的另一个目标应该是开发工具来评估数据集中的偏差，并将合成数据与真实世界的数据区分开来。

鼓励采用创新方法将人工智能援助纳入科学工作流程。科学企业是一个极好的"沙盒"，可以在

其中实践、研究和评估人类与人工智能助手之间合作的新范式。目标不应该是最大限度地提高自动化程度，而是让人类研究人员负责任地利用人工智能辅助实现高质量的科学。资助机构应认识到这些新工作流程的出现，并设计灵活的程序、指标、资助模型和挑战问题，以鼓励用人工智能辅助的新方法来组织和执行科学项目。这些工作流程的实施也为来自各种学科的研究人员提供了机会，如人为因素、工业和组织心理学，以提高我们在人机团队领域的知识。更广泛地说，资助机构、学术界和学术出版业的激励结构可能需要更新，以支持更广泛的科学贡献，例如管理高质量和广泛可用的数据集，而这些数据集可能没有得到传统研究生产力指标的充分认可。