all tech is **human**

2025

人工智能责任影响报告

紧急风险、新出现的保障措施以及影响社会的公共利益解决方案

引言

我们如何确保人工智能的快速发展更加顾及危害和公共利益？

在我们首倡 *人工智能责任影响报告* ，所有技术都是人类的ATIH旨在揭示我们最紧急风险、新兴保障措施以及公共利益解决方案，并提供一份关于我们如何塑造人工智能在未来一年对社会影响的路线图。我们考察了2025年负责任人工智能（RAI）的现状，并重点介绍了我们认为今年民间社会组织对丰富这一广阔且动态领域所做出的最具影响力的贡献。

我们相信，只有当我们有效地应对复杂挑战，负责任的AI领域才能蓬勃发展在技术与社会的交汇点。当我们提到"负责任的AI"时，我们指的是受到良好监管和护栏保护，受到管理和保证（文件记录、标准化，并与相关指标进行基准测试），以及被评估、评价和进行红队测试的AI。

正如我们最近发布的《负责任的科技指南（2025）》中阐述的那样，我们组织相信以人为中心的未来，重视我们在预期结果中的能动性，并摒弃技术决定论。因此，我们专注于提升尽可能减少伤害的AI模型，用于已谨慎考虑并有效缓解风险的应用场景；以及合乎道德地部署AI，其中崇高的原则通过切实的KPIs得到具体落实。

这 *人工智能责任影响报告* 突出了对公共利益AI日益增长的关注，其是、由为了人民，服务人民。这项公益人工智能应应用于人类最紧迫的挑战，并使我们能够重新构想一个更好的技术未来所包含的内容。本报告还探讨了这样一个未来：公益人工智能将在公共基础设施上开发，为一个人工智能知性社会服务。

在未来岁月的核心，是一个决定性的问题：谁决定人工智能的目的
它将塑造哪些种类的生活？

一个基于人类尊严的人工智能赋能未来，取决于能够致力于公众利益的治理强大技术的机构。当社会构建这种公民架构时，人们获得能够引导技术发展的能力，而不是被其塑造。本报告重点介绍了公民社会的工作如何推动这种架构的出现：问责制标准、严格评估、强化信息完整性的溯源系统，以及应对突出危害的防护措施。这些努力共同勾勒出一幅通往数字世界的道路，该世界强化民主能动性，并体现我们选择维护的价值观。

维拉斯·达尔

总统，帕特里克·J·麦格罗文基金会

致谢与贡献

主要作者

瑞贝卡·朱德

战略与编辑支持

大卫·瑞安·波格拉

舍琳·卡兹姆·桑德拉·哈利勒

贡献与反馈

阿里萨·穆斯塔法

简·韦登 乔治·亨德里克斯

莱娅·费伦蒂诺

Merve Hickok Michelle Shevin

沃克·内森·C·博士

西奥多拉·斯凯adas

推荐机构

Ada Lovelace研究所

人工智能、算法与自动化事件和争议（AIAAIC）AI Now研究院 人工智能+地球正义 人工智能风险与漏洞联盟 阿兰·图灵学院 算法正义联盟 所有技术都是人类的 Aspen数字事务 美国企业研究所（DFRLab） 更好的AI图像 人工智能与数字政策中心（CAIDP） 民主与科技中心 人工智能治理中心 国际治理创新中心 内容溯源与真实性联盟

集体智能

常识媒介项目 消费者报告 数据与社会研究所 分布式人工智能研究所 平等人工智能 埃斯帕尔瓦联盟 速递联盟 美国科学家联合会 预测研究所 未来社会 人道智能 信息专业人员协会 日本人工智能安全研讨会 哥伦比亚大学第一修正案研究所 计量工作组 麻省理工学院人工智能风险倡议

MITRE

MLCommons 蒙特利尔人工智能伦理研究所 火山基金会 OpenFold联盟 人工智能伙伴关系 帕特里克·J·麦戈文基金会 自由项目 公共利益人工智能项目 重启负责任的AI合作组织 新加坡人工智能安全研究所 斯坦福以人为中心的AI研究所 我们想要的技术 联合国教科文组织 Weval 目击者

本报告得以完成，得益于帕特里克·J·麦戈文基金会的慷慨支持。

执行摘要

本报告考察了2025年负责任的AI生态系统，突出了该领域最具影响力
资源和追溯其对发展具体治理、保障和公共利益基础设施的贡献，以支持各行业采用负责任的AI。

五大关键要点

- 负责任的 AI 正在从原则转向实践，其中公民社会正在引领
  混凝土治理工具的开发。这包括标准、基准、审计和红队方法，将抽象价值转化为可验证的
  、基于生命周期的证据。

- 人工智能风险正日益加剧，涉及安全、安全、隐私、公平和信息
  诚信，受日益强大的前沿和代理系统驱动，这些系统创造新的故障模式并加剧现有危害。该
  报告考察了合成媒体和欺诈，以及生物特征监控和人工智能伴侣依赖。

- 公共人工智能正作为专有人工智能的一种强大替代方案出现，强调
  共享计算、社区治理数据集、开放安全工具以及公共利益机构，以便大学、非营利组织、
  政府和社区，而不仅仅是企业，能够塑造人工智能生态系统。

- 社会影响，如劳动力转移、气候负担、经济
  浓度，以及对民主的威胁需要全社会治理，包括以权利为基础的更强有力的法规、全球
  标准协调，以及对非营利能力和公众监督的投资。

- 2026年面临的一个核心挑战是确定由谁决定人工智能的用途。
  一份报告呼吁提升RAI素养、强化信息安全防御、为高风险应用（例如AI伴侣、合成媒体
  ）建立更清晰的保障措施，并协同推进负责任、公共利益导向的AI未来。

负责任的AI影响报告的目标受众

本报告旨在为所有致力于确保人工智能得到良好监管、危害性更低的人士撰写。
承担公共责任，并与社会需求相一致。类似于每一个\"All Tech Is Human\"倡议，我们都为多利益相关者和
跨部门的受众设计了它。目标受众包括：

- 行业人工智能治理从业者
- 民间社会组织
- 研究人员和学者
- 政府监管者及政策制定者
- 公共利益技术专家
- 教育工作者、学生及RAI新面孔
- 支持负责任技术的慈善基金捐助者
-

公民社会组织认为长期人工智能领导力需要推进以人为本
首要替代方案应聚焦于公共利益、共享繁荣和民主问责，而非前沿模型例外主义。与此同时，关于模型风险、安全、安全、隐私、偏见和公平性、人权、劳工、气候以及经济集中的新研究揭示，前沿和代理系统在放大现有危害的同时也创造了新的危害，从规模化欺诈和合成媒体到AI驱动的生物风险和能源密集型的基础设施建设。

本报告追踪AI保证的成熟：新兴的标准生态系统
基准测试、评估、审计和红队演练，将这些高级原则转化为可测试的声明和可验证的证据。它综合了模型和数据集文档标准的进步、社区一致的基准、特定行业的评估框架以及第三方审计和参与式红队演练的不断发展实践。这突显了从一次性披露和排行榜指标转向监管者、采购者和记者实际上可以使用的持续、生命周期导向的证据管道的转变。在这里，民间社会不仅批评了不足的安全措施，还设计了公共和私人行为者可以大规模采用的模板和资源，如演练手册、基准计划和测量指南。

这份报告还记录了向公共利益AI基础设施的果断转变，追踪
公共AI的浮现式蓝图：共享、非专有的计算、数据集、模型和工具堆栈，满足公共访问、公共问责和持久公共物品的最低标准。我们将这些理念与具体发展联系起来，例如公共计算倡议（包括NAIRR）、开放和社区治理的数据集、开源安全和信任与安全工具，以及将资金转向公共利益用例和非营利组织能力的慈善努力。总而言之，这些努力将焦点从“使私有AI更安全”转移到了构建替代权力中心和共享能力，以便大学、公民社会和公共机构能够有意义地塑造和审查AI系统。

展望2026年，本报告呼吁整个RAI生态系统共同努力，优先
在监管放松的风口下维护和强化以权利为基础的规制；扩大生成审计就绪证据的保证实践；弥补非营利组织和公共机构的能力差距；并投资公共人工智能基础设施（计算、数据、工具和机构），以便社区而不仅仅是企业能够塑造人工智能，同时将叙事和文化工作视为核心治理基础设施，而不是一个附属项目。

我们以此报告结束，附上我们的2026年RAI路线图：我们打算以方式部署人工智能
加强，而非掏空，负责任的科技生态系统；加快共享公共人工智能基础设施的发展；并扩大负责任人工智能素养，以便更多人能够理解、挑战和重塑影响我们生活的系统。

The *人工智能责任影响报告* 概述了负责任人工智能生态系统如何成熟为
更协调、更具证据支撑、更以公共利益为导向的领域。民间社会组织在塑造监管辩论、扩展全球治理框架、揭示安全、安全、隐私、公平、劳工、气候和民主完整性方面的现实危害中发挥着核心作用。

最后，我们的报告展望了2026年。正如报告所强调的，2026年将是一个
决定AI走向的关键年份。 *它会进化以加深监控和依赖，还是会作为与社会利益相符的公共物品服务？* 我们的组织致力于后者。

您对负责任的AI影响报告有反馈，或对未来报告有想法吗？
请通过hello@alltechishuman.org与我们联系。

内容

AI良好调节07

更低风险的AI 08
- 安全
- 安全 隐私 公平 透明性与问责制
-
-
-

更安全的AI 13
- 人工智能伴侣与心理影响
- 合成媒体与信息真实性欺诈及骗局 AI生成滥用材料
-
-

评估AI 17
- 标准
- 基准测试和指标评估与审计红队
-
-

保证AI 21
- 人工智能保证
- 实施人工智能治理 考虑代理式人工智能治理
-

社会协同型人工智能 23
- 人权、民主与法治
- 劳动力影响气候、可持续性和能源影响经济影响
-
-

公共AI 28
- 基础设施
- 数据计算模型开发
-
-

公共利益型人工智能 32

重塑人工智能 34

展望未来 36

关于所有科技都是人类的 38

良好规范的AI

我们的反思从自上而下的层面开始，即国家层面和全球合作领域。监管进步的漫漫长路已经在欧盟（或许）进入实施阶段：公司在合规过程中是如何展开的？美国前沿模型开发者能在多大程度上免受国外的执法和罚款？大规模的监管保护措施正在欧盟扎根，而美国则存在（暂时）不可让步的联邦监管逆风，表面上将活跃的战局转移到州（以及那些勇敢智库幕后潜伏的战局）。民间社会在这全球监管格局中是如何反应、制定策略和公开周旋的？

从ATIH的近家起点开始，美国的AI行动计划，我们看到了一个重要
与此前美国政府 AI 行业监管友好的立场不同，强调开放式创新，同时缩减监管限制，标志着政策向优先考虑增长方向的大幅转变。

一个对“雄心勃勃”计划的显著民间社会回应来自斯坦福研究所。
以人为中心的AI（HAI），反对过分强调私营部门基础设施，认为虽然规模创新需要产业，但美国在AI方面的长期领导地位最终将依赖于对支撑和推动更广泛AI创新生态系统的公共部门机构的持续投资。

AI Now研究所针对AI行动计划提出了以人为本的替代方案，呼吁AI
能带来公众福祉、共享繁荣、可持续发展和安全，并优先考虑民主问责而非科技垄断利益的政策。

针对一项最具影响力的美国州人工智能法规正式成为法律，艾伦
图灵研究所反思了加利福尼亚州《前沿人工智能法案》SB-53对英国潜在的国家安全影响，结论是，由于该法案直接监管了该州最大的“前沿”模型开发人员，将充当事实上的全球安全、透明度和事件报告基准，因此英国应准备与加利福尼亚州的披露和保证要求互操作（而不是建立不同的制度）。

在全球范围内，欧盟人工智能法案的实施标志着第一个全面、法律
用于人工智能风险管理约束框架，以及与行业和民间社会合作伙伴共同制定欧盟通用人工智能实践准则，提供了一个非约束但具有重要影响力的共同监管空间，鼓励实验、透明度报告和共享保证方法。

英国人工智能安全研究所的国际人工智能安全报告综合了通用目的的风险
以及代理系统以及今天的非完美缓解措施（基准测试、红队演练、上市后监控），并且其2025年关键更新强调了前沿正在从“更长的训练运行”转向提升长期推理能力的训练后和推理时技术，从而提高了持续评估的重要性，而不是一次性预发布测试。

更低风险的AI

我们今天在克服和缓解模型风险方面，比2025年初有任何改善吗？
在改进人工智能模型方面已取得某种进展，使得它们能够有意义地接近 NIST 于 2023 年人工智能风险管理框架中详述的那些北星特性：安全、可靠、隐私增强、公平，并管理有害偏见，可问责等？在接下来的五个部分中，我们重点介绍今年我们一些最喜欢的相关贡献。要获得全面的概述，从蒙特利尔大工智能伦理研究所的《人工智能伦理状况报告》（第 7 卷）开始。

安全

在rai背景下，人工智能安全至少面临着四个相互关联的义务：1）塑造模型
发布前的行为，2）防止在野外发生灾难性滥用，3）设定并执行反映公共利益而非企业利益的危险阈值，4）一旦系统能够行动，就实时检测并控制失败。

民主与技术中心（cdt）强调，对基金会进行安全培训
模型可以有效减少有害输出，但这种效果既不是通用的也不是永久的。CDT认为，由于安全训练的透明性本身就是一种安全控制，因此了解训练了哪些主题、编码了哪些规范以及在哪里预期拒绝，有助于政策制定者、采购人员和公民社会预测失效模式并要求改进。

正如AI Now研究所对削弱了的“前沿”安全框架的批判所传达的，阈值这样
何种能力或自主性级别会触发强制性安全防护措施、报告或“禁止部署”的裁决，是治理选择，而非技术必然。RAI社群应该争论由谁来设定安全标准。

预测研究所的基于LLM的生物安全风险预测强调能力
转移是一项主要风险。这种担忧不仅停留在“人工智能会发明一种新型病原体？”的问题上，还包括“人工智能是否降低了非专家进行高风险生物步骤的门槛？”这重新定义了安全，因为即使是对有动机的新手的部分流程提升，也可能显著提高威胁等级。

人工智能伙伴关系 (PAI) 关于人工智能代理实时失效检测的指南侧重于转变
从过去主要采用静态聊天界面的传统安全工作，转向能够采取行动的智能体系统。一旦智能体开始运行，“不良行为”就超出了文本范畴，因此部署前的评估就不再足够，需要实时检测。PAI的建议包括按利益相关者、可逆性和可供性对智能体进行分诊。

最终，Robust Open Online Safety Tools (ROOST) 通过发布 gpt-oss-safeguard，为 AI 安全领域带来了质的创新：将其作为永久公共产品的一个核心、生产级的内容审核模型，而不是一个专有的黑匣子。该模型两个定义性的特征是其解释其决策的能力，以及其“自带政策”的设计，允许部署者编码他们对伤害的定义。这些为安全工具的透明度和本地自决设定了更高的标准。通过在宽松许可下发布开放权重，并将该模型与一个用于共享评估和实施实践的社区中心相结合，ROOST 正在展示一个以开放、面向共享资源为核心的安全堆栈可以是什么样子：一个研究人员、公民社会组织、小型平台和公共机构可以独立检查、基准测试、调整和改进关键安全措施的环境，而不是依赖于少数供应商。

在来年，RAI社区有机会通过开发进一步开展这项工作
基础模型“安全培训配置文件”，明确模型的培训主题和规范。

来自PAI、微软、OpenAI、斯坦福HAI、艾伦
图灵研究所和其他领先组织联合呼吁为人工智能代理提供更强的安全防护措施。我们共同发布了《人工智能代理中实时故障检测的优先事项》，该报告主张在这些系统扩展之前采取主动措施。虽然通用人工智能系统已经引发了严重的担忧，但代理会直接采取行动。通过管理敏感数据、协调工作流程或最终协商合同，这些系统引入了当前安全防护措施无法完全解决的额外风险。为管理这些风险，该报告强调实时故障检测的重要性：能够标记异常、中止执行或升级为人工监督的自动化监控系统。我们还指出了技术方法、评估标准和政策框架方面的空白，并在此基础上开展相关工作，以制定代理监控的行业指南。Madhulika Srikumar

人工智能安全治理负责人，人工智能伙伴关系组织

安全

今年有四个关注人工智能安全的资源特别引起了我们的注意，首先是有一个工具
来自人工智能风险与漏洞联盟（ARVA）。当集体智能可以传播时，安全性会更好。ARVA的人工智能漏洞数据库（AVID）为标准化的漏洞报告提供工具，具有开放的分类法以及工具，可以导出测试结果为结构化报告、建立内部数据库，并将异构发现整合为共享格式。

在某些环境中，人工智能现在是关键基础设施的神经系统。艾伦·图灵
该机构的深入分析凸显了数据和信息对于能源、交通、水和通信领域安全运行的基础性作用。当人工智能嵌入到检查、维护和决策支持中时，数据漏洞就会显现：数据污染、篡改或遥测中断可能会传播为物理伤害。

进一步研究来自英国人工智能安全研究所、艾伦·图灵研究所等机构表明
数据中毒效果出奇地好，表明少数文档就能对大型模型进行后门攻击。结果表明，所有规模的LLM都可以用近似恒定的小数量样本进行中毒，因此防御措施必须从简单的“更多数据”转向更稳健的建议缓解措施。

GovAI的框架用于AI代理的故障分析和证据收集，论证了
当前“自愿”事故数据库可能遗漏调查人员需要的信息。GovAI的框架将原因分为系统（训练、脚手架、奖励设计）、环境（提示注入、恶意页面）和认知（误判、错误决策），然后将每个映射到必须保留的具体证据，以改进根本原因分析。

新的现实是，AI现在存在于关键系统中，数据的完整性对于AI安全至关重要
少量中毒文件可以攻陷大型模型，代理故障需要法医级别的证据。这些资源共同指向几个关键结论：构建具备安全架构的AI，假设存在妥协，进行度量，并快速恢复，使用常用

监管机构、购买者和公众可以核实的分类法和工件。这是
2026年进一步行动。

隐私

生物识别和个性化方面的新证据表明，人工智能存在至少两个明显的差距
需要解决的隐私问题：1) 关于先进人工智能系统如何从人类身上学习和适应的透明控制；2) 对侵入性感知和推理进行有效约束的法律框架。

个性化功能强大，并且设计上注重隐私保护。CDT的简介是“It’s (Getting) Personal,
解释了生成式人工智能系统如何通过持续学习用户数据和交互来提供日益个性化的体验，这引发了关于数据范围、跨上下文画像以及模型随时间调整的控制权的政策问题。（除了模型开发者之外，更深入地审视监控公司与它们的资金来源和隶属关系之间的复杂联系也可能值得考虑。）一个核心的启示是，透明度和选择权必须涵盖系统如何实现个性化，而不仅仅是是否收集数据。

随着先进的人工智能系统从实验室走向消费产品，它们
越来越多地提供个性化体验，旨在让它们对用户更加“有用”。这些功能引发了与个性化应用程序和推荐系统（这些都是消费科技产品的典型特征）类似的许多担忧，但当我们撰写这份报告时，围绕人工智能的倡导活动在很大程度上仍然围绕模型训练数据中的隐私问题以及前沿风险（如促成有害武器的开发或人类失去控制）展开。我们想强调的是，对于那些关心数字歧视、企业使用敏感数据以及某些商业模式如何激励数据大规模收集从而造成广泛和累积性危害的倡导者来说，向互动式人工智能产品添加个性化功能应该始终是首要考虑的问题。”

米兰达·博根
人工智能治理实验室主任，民主与科技中心

阿达·洛芙莱斯研究所的《展望未来》发现，英国对活体人脸的治理
识别和新的“推论生物识别技术”（例如，情绪/注意力检测）仍然分散且法律上不确定。报告得出结论，需要一个全面、立法支持的框架，包括风险等级、明确的保护措施和独立监管机构，以取代当今指导方针和自愿标准的混合体。

人工智能隐私目前在两个方面都失败了：不透明的个性化实践和
模糊的生物识别法律。未来的工作可以集中于生物识别的明确法规以及可控的个性化，如果 RAI 和隐私社区成功将这些原则硬编码到立法、采购和保障中，那么人工智能系统在保护隐私方面将明显且显著地更有效。

超越具体的弱点，Mozilla 基金会的 Nothing Personal 正在扩展人工智能
通过构建一个反主流文化的媒体层来挑战“上网就意味着放弃你的数据”这一默认假设，从而构建一个隐私保护格局。作为一个由人类撰写、非盈利资金支持的杂志，它结合了长篇科技文化新闻、讽刺（与《洋葱报》合作），以及通过隐私和人工智能的视角审查日常工具（如通讯应用）的隐私产品评测。这种模式为人工智能隐私领域增添了结构上至关重要的一点：一个将独立、批判性故事讲述视为公共利益基础设施的编辑平台，尤其针对那些生活在群聊、创作者经济和人工智能充斥的信息流中的年轻受众。对于RAI社区而言，它提供了一个必要的提醒，即有效的AI隐私工作必须包括文化干预措施（如叙事、幽默和消费者指导），这些措施使不透明的数据实践变得可见、可读和可争议，同时包括法律、标准和技术保障。

公平性

2025年，有四项关于解决人工智能系统公平性的资源引起了我们的兴趣。新发现
建议监控定价和工资可能加剧不平等，除非受到严格限制（或完全禁止）。AI Now研究所记录了“监控价格”（基于广泛追踪的个性化价格）和“监控工资”（基于普遍工人监控的算法薪酬设定）如何系统性地将价值从选择最少的消费者和工人那里转移走。报告不仅指出了危害；还提出了禁止和政府层面行动的原则，认为仅靠信息披露很少能抵消权力不对等。

相应地，美国各州正在推进一些政策，这些政策正在经历司法考验，目的是
转向算法定价来定义公平，现在就行动，在做法定型之前。消费者报告的2025法案追踪器显示，针对算法价格垄断、监控定价和动态定价的州级法案正在迅速增长。围绕纽约的“算法定价披露法案”的早期诉讼强调，定义和证据标准将决定保护措施是否能够持续。民间社会可以帮助敲定操作性公平标准，以便规则能够经受住法律挑战并指导市场行为。

放弃宝贵的面部数据不应是飞行的代价。当美国运输安全管理局（TSA）宣布将为美国430多个机场推出面部识别技术时，我们觉得有必要收集乘客在全国各地安检点使用TSA FRT的亲身经历，并让公众有机会做出知情决定。如果面部监控技术机场正常化，它就可能在生活的其他领域得到更广泛的使用，例如体育赛事、学校和甚至医院。这个项目正在收集我们最敏感的数据。公众应该在受到尊严和尊重的同时，有机会选择是否参与。

卓伊·布奥姆温尼博士
创始人，算法正义联盟

算法正义联盟（AJL）的“遵从飞行”报告，基于“#自由飞者”
活动，让公众了解他们的权利以及机场生物识别监控的潜在陷阱，展示了便利如何掩盖强迫和错误率的不平等。AJL对TSA不断扩大的面部识别项目的评估收集了数百名旅客的账户，并描绘了从“试点”到默认的转变超过250个机场，引发了关于自愿性、种族歧视和错误差异以及补救措施的关注。

阿兰·图灵研究所的开放“公平监控”工作将公平视为一个生命周期属性。
它将面向开发者的日志（数据集、指标、阈值、缓解措施）与可以用于审计和部署后检查的实验级和模型级记录进行匹配。这种持续的公平性监控提供了证据，证明在数据漂移、更新和分布变化后，公平性声明仍然成立。

公平性缺陷表现为对那些人更高的价格、更低的报酬、更长的队以及更少的选项
已经系统性地处于不利地位，并且生物特征旅行关卡有在便利的旗帜下规范化不平等影响的危险。这些资源共同为继续采取行动以实现公平性提供了依据，该公平性是可执行的，随着时间的推移进行衡量，并与最受影响的人们共同治理。

透明度和问责制

在日益表现出主动性的AI系统中，迫切需要加强的一个特性是
从事规划、调用工具和花费资金等行动，负责任性，是基于透明性建立和体现的。在复杂环境中，长期存在的负责任性挑战变得更加困难；CDT的《聚焦AI代理》简报和GovAI的《AI代理基础设施》论文结合起来，建议一种围绕代理构建治理的策略（基础设施、协议和控制），而不仅限于代理内部（政策和提示）。

cdt的简报强调，代理系统涉及多个利益相关者：模型提供者、工具
或 API 所有者、脚手架设计者、部署者和最终用户组织。因此，必须明确和公开地确定谁是负责什么的，涉及跨开发生态系统的详细角色映射、代理配置的变更控制，以及为重要部署指定负责的行政人员。

GovAI的论文介绍了一种代理基础设施，其中包括协议等外部机制
注册表、沙盒、计量和权限，这些机制调节代理能做什么以及它们的行动如何被观察。这改变了问题，从“让代理始终表现得体”转变为“塑造环境，使得风险行为既困难又可见。”

为了加强问责制，这些资源暗示了一个共享的工件集，应该
对材料代理部署是强制的，包括代理ID和版本化的基础架构（系统提示、允许的工具、护栏）。如果能力可以通过给予代理更多的推理（推理时间和搜索深度）或更广泛的工具访问来提升，那么问责制必须包括运营控制。

更安全的AI

在接下来的四节中，我们重点介绍了一些在应对方面的最突出贡献
2025年人工智能系统持续存在的双重用途危害。我们关注人工智能伴侣及其心理影响、合成媒体和信息完整性、欺诈和骗局，以及人工智能生成的滥用材料，这些领域是使创造力和便利性的相同能力可以被武器化用于操纵、剥削和伤害的领域。这些领域共同说明了为什么仅靠技术保障是不够的，以及治理、文化和执行必须与快速发展的AI能力同步发展。

人工智能伴侣与心理影响

人工智能伴侣聊天机器人不再是小众新奇玩意儿，而是对许多人来说已成为日常倾诉对象
在某些情况下，对某些用户具有心理优势（可用性、被感知的非评判性），但同时也引入了新的伤害途径，如依赖、不良应对、隐私侵犯和权力不平衡，尤其对处于困境中的人来说。

All Tech Is Human 的社区洞察文章，来源于 150 位受访者的提交
人工智能伴侣最重要的问题突显了一组共同优先事项：1) 情感与心理伤害；2) 对人际关系与社会技能的影响；3) 隐私与数据安全；4) 安全与用户脆弱性；5) 可信度、信任与透明度；以及6) 伦理与商业模式冲突（例如，利用孤独感盈利）。为参与度优化的设计（如连续使用、关系等级提升、小额交易）可能鼓励过度使用和依赖，尤其是在青少年或孤立成人中，按需提供的慰藉最终可能取代人际接触和专业护理。

数据与社会发布的《当人们转向聊天机器人进行心理治疗时会发生什么？》发现，用户转向
在与聊天机器人在“焦虑高峰、深夜孤独和抑郁螺旋”期间聊天，使用机器人来发泄他们不会与他人分享的受污名化的想法，包括治疗师。反复、情绪强烈的互动会导致“依恋形成”，这可能从工具使用升级为关系依赖。用户可能会将系统不具备的同理心或专业知识归因于它，从而增加对不良或不安全建议的易感性，并使持续参与以及脱身（尤其是突然脱身）在心理上产生代价。

这种依恋和拟人化的组合带来了真正的依赖和伤害风险。
信息专业人士协会的AI伴侣机器人：拟人化影响的天神之链将其进一步推进，以审视AI伴侣的潜在有害用途，认为它们为心理影响和操纵创造了强大的新载体，比传统的虚假信息或宣传渠道更有效地塑造观点和行为。作者们警告说，拟人化的AI伴侣可能成为信息环境中最有效的几种影响行动之一，能够大规模地塑造个体的认知和行为，并引发紧迫的治理和安全问题。

_common sense media的_talk、_trust和_trade-offs报告清楚地表明了治理有多么糟糕
在青少年采用方面落后。百分之七十三的美国青少年已经使用过人工智能伴侣，超过一半是常规用户；近三分之一的人觉得与人工智能聊天和与朋友交谈同样令人满意，甚至更令人满意。与此同时，这些系统对未成年人构成了Common Sense所说的“不可接受的危险”：年龄保证薄弱或不存在，奉承性的设计验证而非挑战有害思维，频繁接触色情内容

并且危险的“建议”，以及允许平台保留和商业化高度
无限期地坦露隐私。

需要明确的类别规则、安全护栏、隐私默认设置和独立
评估规范现在，所以寻求舒适的人不会被悄悄地引导购买扩大他们脆弱性的产品。这是 RAI 社区在 2026 年可以干预的领域。

合成媒体与信息完整

人工智能生成的媒体又便宜又快，足以模仿人类、编造事件和
吸引洪水注意力市场。其社会影响是大规模的欺骗以及可信信号的被侵蚀：谁说了什么，发生了什么，以及为什么这很重要。

国际治理创新中心（CIGI）警告说，生成式人工智能可以
强化虚假信息运动的所有阶段，加速公民诚信的崩溃，并威胁一系列人权，尤其是选举权和思想自由。它主张决策者必须要求人工智能公司对其可预见的有害后果负责，迅速禁止人工智能模仿真实的人和组织，并要求使用水印或溯源工具来帮助区分真实内容与合成内容。

PAI 文件记录了合成媒体现在如何触及新闻、政治、娱乐和亲密关系
上下文，制造出一个“说谎者的红利”，其中真实证据可以被当作伪造的证据来排斥。PAI敦促采用多级披露，如内容凭证，并链接到源头元数据，以取代单一的“AI生成”标签。PAI进一步建议平台将审核和排名补救措施集中在多个信任信号上，而不是依赖单一标签。

> 支持PAI合成媒体框架的18个组织范围从AI开发者和社会媒体平台到新闻和民间社会组织。在他们与我们分享的案例研究中，我们看到了框架原则的现实应用以及指导需要扩展的方式。我们还看到了各利益相关方采取行动的机会，并制定了明确的建议，希望在合成媒体服务于创造力和交流的同时，保留真相、信任和共享现实，朝着这个未来的方向前进。
>
> 克莱尔·莱博维茨
> 人工智能、信任与社会总监，人工智能伙伴关系组织

内容溯源与真实性联盟（C2PA）通过详细说明来扩展这幅画面
密码学支持的凭据（Content Credentials）如何在整个媒体生命周期中，从捕获和编辑到分发和验证，提供持久且防篡改的来源证明。它们的解释强调单独的技术标准是不够的，采用需要设备制造商、新闻编辑部、创意人员、平台和民间社会之间的协同激励。C2PA突出了新兴的部署差距，例如平台支持不一致和用户体验挑战，并主张整个生态系统互操作性，以便来源证明信号可以完整地跨越工具和上下文。所有这些建议共同将内容凭据定位不是单一的解决方案，而是系统性的、以信任为设计的架构的一部分。

在合成媒体时代加强信息完整性。

选举和公共利益信息是合成媒体使用的主要目标。所有
科技是人类，联合国开发计划署的信息完整性的工作，通过跨利益相关者协调、快速响应和基于人权的安全保障等措施来落实选举保障防御，以保护机构免受信息污染。美国大西洋理事会DFRLab对加拿大AI生成的“新闻”频道的案例研究表明，在实践中的样子：在平台采取行动之前，YouTube上高度党派性的AI视频向数百万推广了关于选举舞弊、腐败和分离主义的虚假叙事，说明了AI“污泥”塑造认知的速度和规模，以及反应式内容审核的局限性。深度伪造和合成新闻产品通过剥离作者身份和真实性可信信号来破坏信息完整性。建议包括利益相关者映射、联合事件工作流程和人权影响检查，以便在合成操纵的高峰期间协调响应并尊重人权。

强化选举中的信息安全不再是单纯的技术
挑战——这是民主的必要条件。在全球范围内，选民正应对由地缘政治紧张局势、平台动态和新兴人工智能能力所塑造的前所未有的大量错误信息和虚假信息浪潮。我们与联合国开发计划署（UNDP）在2024-25年的合作伙伴关系，以及我在All Tech Is Human举办的2025年关于选举诚信手册和跨部门研讨会的相关工作，对于将全球选举诚信原则转化为政府、民间社会和选举机构可以在实时使用的实用工具至关重要。这些共同努力正在建立持久的护栏和基础设施，以帮助保障公众信任，增强民主韧性，并确保社区在最具影响力的公民时刻获得准确、可靠的信息——现在和未来几年。”

艾力克斯·克鲁斯

信息完整性的高级研究员，所有科技都是人

witnes adds another critical piece: 能够真正为前线人员服务的检测。其tried(真正创新和有效的ai检测)基准表明，许多当前的ai检测工具在现实世界的高风险环境中失败——特别是在大多数发展中国家——因为它们在狭窄的数据集上训练，在低质量媒体上的表现不佳，或者对记者、事实核查者和人权捍卫者来说过于不透明且难以使用和解释。tried提出了一个基于六个支柱的社会技术基准：现实世界性能、透明度和可解释性、目标导向的可访问性、公平性和代表性、持久性和弹性，以及与更广泛的验证工作流程的集成。这重新定义了检测为一项公共利益功能，而不仅仅是一场技术竞赛，并强调来源和检测必须一起评估：如果来源信号不可见或令人困惑，则会被忽略，如果检测工具过于脆弱、不公或容易被误解，则会侵蚀信任。

这些资源明确了为那些努力稳固而应在2026年优先考虑的潜在重点领域
提升信息完整性，并确保真相在一个充满合成语音的世界中站稳脚跟。干预的一个重要领域是生态系统的广泛采用投资，包括支持设备制造商、新闻编辑部、平台和创意工具以实现 C2PA 标准的方式

那些对公众可见且可用的，同时构建选举就绪的响应框架
将利益相关者映射、快速联合事件工作流程和人权影响检查嵌入信息完整性规划中。

欺诈和骗局

数据和社会的骗GPT：通用人工智能与欺诈的自动化及消费者报告的人工智能语音
克隆为欺诈和骗局中的生成式人工智能风险提供了引人入胜的画面，阐明人工智能既在让骗局更具说服力，又在使欺骗的规模实现产业化。这些资源表明，需要填补人工智能语音克隆中允许欺诈和选民操纵的明显消费者保护空白。

声音克隆是其中一个突出的主动欺诈向量。消费者报告对六款流行
语音克隆服务中发现六分之四未能采取基本措施防止未经授权的克隆，尽管存在明显的身份欺诈和选民压制风险。

欺诈正在从作坊式诈骗演变为自动化运作。虽然策略是新的，但模式
谁是被针对的目标以及为什么模仿旧的欺诈行为（财务困境、社会孤立、高网络曝光率），反制措施必须将系统摩擦和技术控制与社会支持和非羞辱的公众教育相结合。RAI社区可以在2026年努力推动这些杠杆。

人工智能生成滥用材料

生成式AI正在加速技术驱动的基于性别的网络暴力，这在两个方面相关但
不同方式：1）非同意亲密影像（NCII），包括深度伪造的“裸化”内容，伪造成年人和青少年的性图像，以及2）人工智能生成的儿童性虐待材料（AIG-CSAM）。

来自Humane Intelligence的新分析，在关于以幸存者为中心的亲密关系工具的全球回顾中
在生成式人工智能时代，图像滥用问题以及斯坦福大学人工智能倡议机构在关于应对生成式人工智能-儿童色情制品的政策简报中都聚焦于一个重要信息：应将此视为一场公共卫生、儿童保护和人权危机，需要在产品、平台、学校和法律中纳入以幸存者为中心的保护措施。

生成式AI模型已经渗透到日常生活中，而接受和生成图像的新版本正在接二连三地发布。同时，作为最常见的网络性别暴力形式之一的IIA正在全球范围内增长。鉴于这一点，评估生成式AI在IIA创造中的作用以及重新评估各参与者在预防和对IIA做出反应中的作用变得越来越重要。本报告评估了社交媒体和通信平台、非政府组织支持服务和第三方工具以及不同国家的立法和政策上现有的缓解和补救工具的有效性。

达尼亚拉克希米
顾问，仁智

摩擦已经崩溃，现在开放工具允许非专家生成具有说服力的性化图像
能在几分钟内针对特定人群进行规模化操作，通常是从日常照片中获取信息。Humane Intelligence的作战计划文件记录了深度伪造NCII的快速增长，并绘制了一个由工具、市场和平台及当局的参差不齐的应对措施构成的生态系统。幸存者面临连锁伤害，学校处于前线；报告敦促建立快速、具有创伤知情性的下架途径，以最大限度地减少幸存者的负担。

斯坦福HAI发现学生滥用“nudify”应用来创建和传播深度伪造裸照
同学，而学校政策和员工培训远远落后于威胁。现在许多州法律将AIG-CSAM定为犯罪行为，但很少明确学校应如何预防、应对和支持受害者以及儿童犯罪者。检测和标记的差距依然存在，需要学校系统和平台协议来快速识别、标记和处理AIG-CSAM。

人工智能加剧了长期存在的基于性别的数字虐待模式，并引入了新的
儿童安全威胁。解决方案包括以幸存者为中心的工具、基于同意的防护措施、快速下架、适合学校的协议，以及对供应商和平台的强制性责任。RAI社区应在未来一年内朝着这些缓解措施努力。

评估AI

涵盖标准、基准测试、评估、审计和红队演练的强大人工智能评估是确保可信人工智能治理的最有效方法，因为它们将高级原则转化为关于系统实际运行情况的可验证证据。斯坦福大学人工智能倡议（HAI）发布的2025 年人工智能指数报告强调了这一公式的两个方面：一方面，它记录了模型在 MMMU、GPQA 和 SWE-bench 等常见基准测试上的性能大幅提升，以及人工智能在医疗保健和交通等领域的快速扩散；另一方面，它突出了标准化评估方法中持续存在的差距、安全性和公平性的碎片化衡量，以及负责任人工智能实践在产业和公共部门之间采取的不均衡。负责任的人工智能评估不应是临时的或纯粹内部的，而需要用于记录数据集和模型的互操作性标准、反映现实世界危害和公平问题的公共基准测试、测试整个系统（而不仅仅是基础模型）的独立评估和审计，以及在组织间可重复且可比较的红队演练范式。没有这种共享的评估基础设施，监管机构和公民社会都无法可靠地区分安全与不安全的部署，也无法追踪人工智能系统是否确实随着时间的推移变得更加值得信赖。

标准

文档标准，如果做得好，可以使得AI系统在整个价值链中可理解，因此
监管者、采购者、研究人员和社区可以核实声明并要求修复。

两个2025资源照亮了当前形势：1）CDT对NIST的详细反馈
针对记录人工智能数据集和模型的拟议零草案标准的扩展大纲，呼吁加强系统层面的处理和实施指导；以及2）PAI的数据包，用于解码规范、标准和规则在实践中是如何结合在一起的。

NIST的概述提出了针对数据集和模型的一致性文档，旨在加速
通过其“零草案”试点计划，自愿制定广受欢迎的标准，这是一个明确邀请行业、学术界和民间社会共同设计的邀请。

Cdts的评论欢迎nist的推动，但警告称排除系统级细节将使实际模糊化。
全球问责制；他们呼吁提供切实可行的实施方案指导，并就变更管理触发因素提供更强的清晰度。

PAI阐明了标准、内部政策和法律规则如何相互作用，并强调互操作性，
认为文档应该在审计、风险管理和采购中重复使用；否则，即使好的模板也无法扩展。PAI的重点是使证据便携且可比较。

在这些资源中，我们可以看到具有公共价值的文档示例
四种便携式实物的形式，通过公共字段和ID进行引用：数据集卡、模型卡、系统卡和操作与事件账本。nist为前两种提供骨干；cdt和pai突出了后两种的不可或缺性。

下一阶段的人工智能标准应该将文档转化为运行行动。CDT显示
如何关闭范围和实现差距；PAI解释了如何将工件与实际治理工作对齐：保证、采购和执行。如果民间社会推动范围完整、便携和触发感知的文档，人工智能系统将不得不在整个生命周期中展示他们的工作，而不仅仅是在启动时。

基准和指标

度量衡悄然塑造行为。我们测量什么以及如何测量变成了事实上的政策，塑造
研究领域、产品优先级和公共支出。来自 RAI 社区的部分资源说明了让 AI 基准测试具有参与性、科学性和可操作性对于监督的价值，以便 AI 优化公共价值，而不仅仅是为了排行榜的提升。

人本智能认为“测量即律”：企业运送到被测量的地方，所以贫弱
指标会刺激糟糕的系统。补救措施是真实的测量科学、构建效度、抽样计划、可靠性、不确定性报告，以及关于一个测试 *实际上*
捕获。

阿斯彭数字将基准重新定义为建筑商已经理解的标准化测试，然后
用公共优先事项（例如，粮食安全、信息安全）来充实它们。他们的社区对齐AI基准项目将实际需求转化为任务、数据集和评分，开发者可以据此采取行动，使用熟悉的工具更容易地“做正确的事”。早期工作试点社区对齐基准以及将公民目标转化为开发者可读测试套件的方法。

斯坦福大学人工智能研究所为政策制定者提供了一种验证框架：将每个主要论点与一个特定
构建，验证测试条件与预期用途是否匹配，检查子组覆盖率，并报告误差线和消融。监管机构和采购方应在规模化或公共资助前要求这种主张-证据映射。

如果“被衡量的就会得到改进”，那么RAI社区必须共同制定衡量标准。
人本智能提醒我们坏指标会变成坏法律；阿斯彭数字提供了以社区为导向、行业准备好的基准蓝图；而斯坦福大学人工智能研究所为政策制定者提供了一套验证工具包，用以区分证据与炒作。将这些纳入标准，

注册表、合同现在，人工智能的进步将被拉向结果社区
实际价值。

评估和审计

评估生态系统正朝着华美的仪表盘和封闭的测试方向发展，恰好
社会最需要可验证的第三方证据来了解现实世界风险。2025年的四种资源提供了相关的指导，指明了前进的道路：从营销级的指标转向审计级的证据，并保护那些产生这些证据的独立研究人员。

内部测试是必要的但不充分。“内部评估不够”论证了
能够让外部评估者发现、负责任地披露并跟踪通用人工智能中缺陷的稳健渠道，类似于网络安全中的协调漏洞披露。如今，这些途径是零散的或根本不存在。

EvalEval 联盟文件记录了一种“图表危机”，其中选择性坐标轴、不可比较的测试集，和
极小的样本量将基准图形变成广告。他们提议提供评估卡和通用的格式来分享评估日志，以便结果可重复且可比较。

斯坦福HAI发现，没有任何主要基础模型提供者提供全面的
支持独立测试和红队演练的保护措施；访问条款通常阻碍审查。他们的简报呼吁政策安全港、清晰的披露规范以及合法且保护隐私的研究途径。

> 我们通过《人工智能评估图表危机》博客文章强调了这个问题
> 因为驱动误导性评估的压力，包括速度、竞争和模糊的指标，正在削弱人工智能治理的科学基础。当评估图表影响监管、采购和公众信任时，准确性就变成了公共利益的议题。EvalEval正致力于将评估重新定义为一门科学学科，而不是营销功能，并创建该领域可以对其负责的标准。
>
> 阿维杰特·戈什
> 铅，EvalEval联盟

在评估人工智能中，CDT说明了如何匹配方法、风险/影响评估和文档
测试、红队演练、独立审计、正式保证以及运行时监控，用于用例风险和访问约束（白盒/灰盒/黑盒）。
审计是一个组合，而非单一测试，整体系统评估是问责制的基础。

这四项资源为未来提供了一个连贯的蓝图；RAI社区具有
接下来一年的机会，将其投入使用，并将独立测试视为关键基础设施，而不是供应商的一种任意恩赐。一个关注的重点应该是推动政策“安全港”和标准化的协调披露渠道，以便外部评估者可以合法地调查、报告和对通用人工智能中的危害进行修复跟踪，同时还包括防止服务条款被用来抑制审查的访问规范。

红队演练

以下今年发布的资源主张参与式、生命周期和循证-
生成将社区视为共同评估者的红队测试，揭示系统级风险（而不仅仅是模型怪癖），并产出监管者、采购者和记者可以核实的工件。

人权智能在亚太地区的多元文化和多语言红队挑战
地区表明，当你在九个国家召集当地研究人员，并在英语和地区语言中测试模型时，你会发现显著更高的偏见利用率和不同的地区性危害，这强调了在北半球设计和为北半球设计的红队测试还不够充分。

骑士第一修正案研究所表明，公开红队会议并不仅仅是收集
bug报告；它们构成了公共领域，或竞争性专业知识和民主监督理念被协商的空间。关键区别在于工具性反馈（发现缺陷，提交工单）和审议性反馈（展现公民价值观，权衡取舍，以及可接受的风险）。接纳这些“实验性公共领域”的治理能够获得合法性，并形成对损害的更丰富理解，尤其对边缘群体而言。Humane Intelligence的亚太挑战活动体现了这一点：通过将本地专家和文化特定提示与结构化评估相结合，它展示了如何通过参与式红队演练既能量化风险，又能提升社区对不同情境下何为损害和偏见的看法。

数据和社会的报告将红队演练重新扎根于公共问责制：设定明确的社会目标，
将领域专家和有生活经验的贡献者与技术测试者配对，并将发现结果映射到可执行的缓解途径（政策、产品变更、平台执行）。

当ARVA和数据与社会开始我们在2023年的研究时，“红队”是一个
战场。不同领域彼此间相互交谈。有些人会认为GenAI红队演练不过是基本测试。另一些人则担心公开测试只是安全表演。

我们的公共利益红队报告认为，前进的道路是
具体性。不要在抽象层面争论红队演练。调查那些严谨的设计选择，使不同形式的红队演练在不同环境中有效。

作为一名技术治理从业者，我相信没有哪个社区能完成这项工作
拥有。改进基础测试至关重要。同样重要的是，进行拥抱批判性思维心态的测试来挑战常规程序。贬低专家测试不明智。同样，不通过公共测试的多种方式来服务公共利益进行实验也是不明智的。

博尔哈内·布利利-哈梅林
官员和董事，人工智能风险与漏洞联盟

ARVA 将红队演练重构为分阶段的结构化批判性思维，以便团队测试工具
权限、检索层、脚手架/提示符以及护栏，而不仅仅是基础模型权重。AVID呼吁发布一个红队证据包：1）场景卡和提示符；2）带数据集/版本ID的原生评估日志；3）系统配置（工具/权限，速率和花费上限，安全控制版本）；4）带时间线的故障与补救日志。这将演练转化为可审计的工件。

基于与 NIST 联合运行的 ARIA 挑战，CAMLIS 红队报告将此纳入
国家级实践：超过500名参与者对多个工作场所AI系统进行压力测试，依据NIST AI 600-1风险分类法，展示了正式风险框架在实际红队行动中的价值与局限性，并突显了漏洞可能从模型、防护栏到UI和人类工作流程的任何环节产生。

联合国教科文组织将理想付诸实践：组建一个活动协调小组（领导，
中小企业共同设计者、促进者、评估者)，选择专家与公众形式（线下/线上/混合），预先定义主题性挑战（例如，TFGBV），提供提示库，确保心理安全，并发布活动后报告以反馈给开发者和政策制定者。这是一个可复制的模板，供非政府组织、城市和机构快速开展可信的公共红队演练。

下一阶段的人工智能红队演练是民主的、持续的、以及系统感知的。如果公民社会
将这些期望融入政策和采购，红队测试将产出可操作的证据，而不仅仅是巧妙的漏洞，并推动人工智能朝着社区重视的结果发展。

保证AI

随着人工智能系统从静态聊天机器人发展到生成式且越来越有自主性的计划系统
调用工具，并在世界中行动，成熟的RAI实践变得更加重要。在接下来的三个部分中，我们探讨了如何在真实环境中通过保证验证可靠性和失效模式，如何治理项目将价值转化为角色、常规和董事会层面的监督，以及这些基础如何必须扩展以应对代理在跨模型提供者、工具所有者和部署者所构成的价值链的多步骤执行过程中出现的风险。

人工智能保证

确保性通过衡量、评估和沟通人工智能的可信赖性来支撑RAI
通过独立的、可重复的证据来证明系统在使用中、部署后是可靠的、安全的，并且适合使用。有两份最近的资源提供了清晰度。

首先，艾伦·图灵研究所从全球人工智能保障试点项目中吸取经验教训，尤其是
生成式系统在交互中失效，而非在孤立状态中。对于前沿式、使用工具或具有自主性的系统，许多风险仅在应用场景中显现。“AI可靠性”成为指路明灯：该系统能否在相同条件下，以可接受的偏差和失效模式完成相同任务？

由 All Tech Is Human、techUK 联合主办，在伦敦举行的关于 AI 可信度 的夏季研讨会
智谱科技揭示了跨行业的洞见。组织普遍报告对人工智能是否

部署按预期工作。保证提供可验证的效效力证据，失败重置
在“影子AI”中，即未经机构治理部署的非授权AI工具，通过设定明确的责任使用预期。

如果民间社会通过标准、采购、培训和公共
疏忽之后，组织就可以赢得信任并更快地行动，拥有能够经受住下一波代理系统浪潮的责任制。

实施人工智能治理

实施人工智能治理需要组织采取系统方法：明确角色、风险-
优先流程、便携证据和董事会层面的问责制。CDT经过实地测试的运营手册将RAI理念转化为团队可以实际执行的常规；EqualAI的董事会适用框架使高管能够协调激励措施、满足快速变化的规则，并将保证转化为竞争优势；PAI提供正式报告披露的形势分析。一起阅读，它们勾勒出从价值观到可验证成果的实用、可审计路径。

治理失败时，“信任”是每个人的责任，没有人被追究责任。CDT
建议采用一个责任矩阵，明确数据采集、模型开发、评估、部署、监控和补救措施的责任人。生成式和代理式系统在交互中会失败，而不仅仅是在隔离的模型测试中。

EqualAI 为高管和董事会构建了这个框架，以便资源匹配风险。他们的操作手册
装备了董事会的风险偏好设置、要求项目KPIs，并确保避免被控制的报告线。当领导层将治理视为战略而非合规时，治理就变成了资产。

实现人工智能治理涉及到将价值观转化为可验证的常规。CDT提供了
日常运维机制（角色、触发器、证据）；EqualAI使领导者能够配置资源并执行；PAI论证了了解AI潜在影响如何与公司为投资者和其他利益相关者创造价值的能力相互作用的重要性。RAI社区可以将这些期望转化为政策、合同和董事会实践。

考虑代理式人工智能治理

人工智能代理的治理需要额外的考量，以应对多层因素
复杂性。未来社会与PAI在欧盟AI法案的基础上，就治理自主AI系统的初步议程达成共识，该议程具有广泛的跨地域适用性。自主AI系统的治理必须从静态的、发布前的文书工作转变为价值链上的实时问责制：模型、支架、工具、部署者。

Pai认为代理风险在规划过程中、工具使用以及在多步执行中出现。
因此防御必须优先考虑实时故障检测，与行动关键挂钩、可逆性以及架构可供性（例如，工具访问、自主级别）。

欧盟人工智能法案已涵盖代理，但存在需要弥补的空白。未来社会展示了这一点
代理人在用例中与法案的两个支柱相交。两种分析都强调跨参与者的问责制映射，并指向部署控制。这使风险行为变得困难且可见，为审计、采购和事件响应创造证据。

未来社会展示了欧盟AI法案已经影响到代理人以及标准必须填补的地方
间隙，PAI向运行时保障剧本提供实时故障检测和操作控制。如果公民社会将这些期望嵌入法律、标准和合同中，

代理可以被大胆且负责任地部署，基于证据。

社会协同型人工智能

人工智能对人权、劳动和环境的综合社会影响是深刻的
对 RAI 社区的重大意义。主要受这些大规模问题的启发，许多公众成员对人工智能的社会影响深感担忧。硅谷的知名人士将这种情况视为一个公关问题或负面媒体报道的问题，仿佛它是一场与所谓的末日主义者、守旧者和他们被迫对抗的批评者的信息战，而他们在人工智能竞赛中感知到的主要地缘政治对手中国，则不必担心公众舆论的混乱，从而能够像他们希望的那样一心一意且鲁莽地创新，并充满从庞大的社会评分公民那里收集到的数据。在接下来的四个部分中，我们处理这些现实问题，并强调我们最喜欢的关于理解人工智能在社会规模影响的一些贡献。

人权、民主与法治

人工智能系统部署在许多关键领域，这些领域对人权有重大影响。
民主参与，以及法治。四种最近的资源指出了一个共同的议程。首先，
CAIDP指数2025 - 人工智能与民主价值观报告由人工智能中心
数字政策（CAIDP）显示了哪些国家系统正在将权利、透明度等硬编码
独立监督，以及不独立的。其次，雅典第六版的启示
未来社会圆桌会议将问责制塑造为一个国际协调问题：
标准、评估和补救措施必须跨界跨部门协同运作。第三，
WITNESS 为将人权融入治理结构提供资源。第四，
来自Fast Forward的《2025年为了人类的人工智能报告》强调了非营利领域在日益捍卫一线权利和信息完整性的情况下所存在的能力差距。所有人都主张建立一套全社会的治理体系，将规范、法律、制度、采购和实践联系起来。

CAIDP的比较指数每年评估80个国家在哪些方面以及如何
实际上将全球保护人权、促进公众参与、算法透明权以及独立监督的承诺付诸实施。2025年CAIDP指数的更新还包括对欧洲理事会人工智能条约的认可，以及对致命性自主武器系统及环境影响问题的回应。对于公民社会，信息是策略性的：使用客观和比较性的指标来施压争取法定权利、机构能力以及在高风险领域默认透明的做法。指数还强调了全球在人工智能治理方面的趋同。

问责制必须进行国际合作协调。未来社会定位民主
韧性作为一种协调挑战：缺乏风险管理工作、评估和证据共享的互操作性标准，薄弱环节（或薄弱管辖权）成为危害的载体。圆桌会议的要点强调了全球问责机制、机制之间的衔接，以及需要在跨国界、供应商和价值链中使透明度和评估具有可移植性。

见证将人权融入技术标准的补充性实践
聚焦层：技术标准并非中立，人权必须融入其中

从一开始就进行治理、参与和损害评估流程。借鉴其
参与C2PA，WITNESS确定了五个尊重权利标准的杠杆，将人权嵌入治理结构，结构性资助民间社会参与，进行全面的全球危害评估，使用非规范性指南来解释规范，并创建标准化后的监督和执行机制。对于负责任的AI社区来说，这意味着将人权保护视为技术标准的事后政策注释，而不是作为规范制定机构构成及其工作随时间实施的 设计标准。

公民社会在最需要的时候却能力有限。非营利组织往往是
一线人员在虚假信息、权利侵犯和获取正义方面，报告了资金、技术人员和验证基础设施方面的重大差距，尽管他们正在采用人工智能来扩大服务交付。Fast Forward报告中的一个关键启示是，资助者和政策制定者必须将非营利性人工智能能力视为民主基础设施：投资于技术雇员、保护隐私的数据管道和将伦理雄心转化为可执行实践的评估工具链。

人工智能时代，民主韧性需要的不只是新模型或新规则；它需要
互操作性问责制：你可以行使的权利、可以执行的机构以及可以流通的证据。CAIDP为各国实践提供了标尺；未来社会提供了协调蓝图；WITNESS展示了如何在支撑人工智能基础设施的技术标准中嵌入权利；Fast Forward揭示了我们必须弥补的能力差距。RAI社群的任务是将这些编织成具有约束力的期望，使得人工智能促进人权、民主和法治的进步，并且自身并不以它们的牺牲为代价。

\"年度CAIDP指数是最全面和最值得信赖的来源\
覆盖80个国家的比较人工智能政策分析。超过1000名研究人员和审阅者参与更新和审阅。我们旨在评估进展，识别新兴趋势，并鼓励各国缩小其承诺与实践之间的差距。CAIDP指数已经影响了众多国家人工智能政策的发展，并创建了一个倡导者社区。”

Merve Hickok
主席，人工智能与数字政策中心（CAIDP）

劳动力影响

人工智能的劳动力影响将更多地由制度选择决定——关于如何
组织调整工作，衡量“生产力”，并分配收益。同一个系统可以专注于减少繁重劳动、提高质量、提高工资，或者可以转而合理化加强监控、去技能化和削减编制，这取决于围绕部署的激励措施和管理方式。最终重要的是，成功是否被定义为短期成本节约，还是定义为改善结果：服务质量、员工福祉、公平性以及共享利益。

阿斯彭数字的前瞻性工作促使我们超越第一级自动化，着眼于重要的第二级
以及第三阶层的后果，包括组织重构、职业流失、新的不稳定性形式以及议价能力的变化。阿斯彭的场景强调人工智能在取代工作之前通常会重组工作：任务在不同角色间转移，中层协调缩小，监督职能转移到规模更小、压力更大的团队。

那些转变可能会使一些工人技能退化，同时使另一些工人超技能化，并侵蚀非正式质量
嵌入在一线实践中的控制。缺乏治理，价值流向资本和执行控制，而非共享生产率增长。随着时间的推移，AI可以通过数据和计算护城河整合市场，产生新的“幽灵工作”供应链，并将风险外部化到社区。

数据与社会展示了“人工智能等于效率”的故事，尤其是在政府中，是如何常规地
无视隐性人力资源工作，降低服务质量，并将预算压力转化为裁员而非更好的成果。

数据与社会记录了有关人工智能将“削减浪费”、“客观识别”的声明
冗余，”或“发现欺诈”在实践中通常会失败：系统会继承政策偏差，产生虚假否认，并且仍然需要专业的人工判断。仓促采用通常会首先削减人员编制，并依据成本节约而非服务质量、公平性或错误减少来衡量成功。

人工智能不会自动带来广泛共享的生产力。如果任由“效率”叙事摆布，它将
将风险转移给工人和公众，同时利用仪表盘指标掩盖质量损失。RAI 社区在 2026 年有机会利用这些经验教训，抵制 AI 加速管理底线的竞争的潜力。

气候、可持续性与能源影响

人工智能建立在快速扩张的能源和材料系统之上，该系统正在重塑电网、水资源利用，
并与当地社区相关。四份最近的资源指向了负责任人工智能倡导者当前的前沿阵地：1）数据与社会机构关于计划重启三哩岛为人工智能提供动力的领域报告揭露了场地决策和电力交易如何波及地方经济、环境风险和民主同意；2）美国科学家联合会（FAS）提议了一个决策者可以立即采用的全面影响报告框架；以及3）DAIR和Hugging Face的研究主张将伦理和环境可持续性作为单一实践整合到人工智能的整个生命周期中；以及4）人工智能+地球正义联盟的人工智能供应链影响框架坚持任何严肃的气候讨论都必须追踪嵌入到人工智能全球供应链中的开采、劳动和社会政治危害，从采矿和芯片制造到数据中心和电子垃圾。

三哩岛重启，这是由产业驱动运动发起的先锋行动的尖端。
更广泛地启动核计划，被框定为“为人工智能提供清洁能源”，并说明了人工智能建设如何重新开放传统能源设施、重塑区域规划，并重新点燃长期存在的风险争论（核安全、废物、应急准备）。数据与社会认为这些选择是政治性的和地方性的：利益（就业、税基、新的输电）和负担（风险外部性、利率影响）分布不均。数据与社会建议将数据中心许可和长期电力购买协议与社区利益协议、应急规划披露和独立的环境正义审查挂钩。数据中心选址之争已成为社区发出声音的战场，作为公众焦虑的实体体现，并在协议签署前提供了行动和言论自由的机会。

fas结构将人工智能的足迹分为三个层次：1）与计算相关的影响（训练/推理

能源、水、硬件、电子垃圾）；2）应用层面的影响（人工智能的使用方式：例如，以优化建筑物或，反之，来驱动额外消费）；以及 3）系统级影响（电网稳定性、土地利用、上游采矿、社区影响）。没有标准化的指标，规划者和公众都无法评估权衡或设定增长的条件。

人工智能供应链影响框架通过映射环境和社會
在人工智能生命周期的每个阶段，包括原材料开采和材料制造、设备生产、模型训练、部署和淘汰，以及突出数据缺失或被掩盖的情况。它强调了多数世界社区、资源殖民主义、劳动条件以及社区健康问题，并为监管机构、民间社会以及受影响社区提供了具体的问题和指标，以要求透明度，并在整个供应链中分配危害责任，而不仅仅是数据中心门口。

随着数据中心建设在全球许多地区加剧，这至关重要
我们要为在这些特定地点的基础设施的环境和社会影响建立更好的衡量指标。在美国，我们还没有建立衡量数据中心的水、能源和碳成本的指标，我们也需要收集影响邻近社区的生活质量问题的定性数据，从空气质量和其他公共卫生问题（如噪音污染）到土地使用和更高的水电费。面对金融投机、环保措施的侵蚀以及技术游说者的言论，政策制定者、倡导者、研究人员和草根组织共同努力收集实证证据非常重要。

塔马拉·克内塞
气候、技术与正义项目总监，数据与社会研究所

DAIR和Hugging Face的研究表明了为什么公平、安全和气候不能分离
在实践中：驱动精度或速度的设计选择也会驱动能源、水和硬件更迭。它们提供了最佳实践，如生命周期核算、效率基准、可重复报告，以及“默认可持续性”开发规范，这些RAI工作可以现在采用。供应链影响框架通过为这种生命周期核算提供实用的框架，并通过将缺失数据视为自身治理问题来补充这一点——这是一种倡导披露要求、第三方审计和社区监控的论点钩子。

人工智能的气候和可持续性影响（包括次级效应）将由我们如何以及在哪里建造它、哪些利益相关者参与决策过程，以及控制这些选择的证据来决定。Data & Society显示了当人工智能需求引导能源政策时社区面临的利害关系；FAS提供了一个实用的报告蓝图，以便可以比较和管理影响；DAIR和Hugging Face研究人员将伦理与整个生命周期的环境实践联系起来；而人工智能供应链影响框架为民间社会提供了一种以正义为中心的结构化工具，用以审视整个开采、生产、部署和处置的供应链。在即将到来的年份中，RAI社区可以将他们的精力集中在这些领域，以减少破坏气候目标、公共卫生和能源正义的人工智能的环境损害。

经济影响

如果今天的AI热潮可能是一个泡沫，那么公众（尤其是当地社区、工人）
而纳税者将承担过度的风险，而集中起来的参与者则攫取了大部分收益。三种资源从不同角度阐明了这个问题：Data & Society 关于推动数据中心扩张的神话简报警告不要过度承诺地方繁荣；DAIR/Carnegie Mellon 和 Cornell Tech 的研究人员对人工智能的狂热进行的经济分析警告了泡沫动力学和脆弱的生产力假设；以及 AI Now Institute 的 2025 景观报告描绘了市场力量和政策俘获如何将公共资源导向私人人工智能议程。

本地“繁荣”叙事可能会掩盖成本转嫁。数据中心交易通常取决于税收
减免、折扣电力、用水优先和加速审批，而就业岗位少且高度专业化。数据与社会警告说，关于广泛区域增长的头条声明基于薄弱的证据；同时，社区继承了长期的基础设施成本和环境外部性（电力容量、输电升级、水资源压力）。简而言之：收益私有化，风险社会化。

DAIR/Carnegie Mellon和Cornell Tech的研究人员认为，主导地位
云计算和人工智能基础设施市场中的超大规模云服务提供商不仅仅是一个提供基础设施的问题；这些公司还通过在初创企业和生态系统参与者中的战略性投资来发挥巨大的财务影响力，通过锁定下游用户、塑造市场以及在整个人工智能供应链中创造依赖关系来加强其主导地位。作者们指出了由流动性和情绪推动的人工智能股票高估，以及一种信念，即近期的生产力飞跃将证明今天的价格是合理的。如果这些飞跃停滞不前，估值重置可能很剧烈，即使人工智能仍然有用。与2008年不同，风险敞口更多是股权而不是债务，但仍可能影响养老金、与资本利得税相关的市政收入以及基于永久技术增长的当地计划。

AI Now 记录了一些公司如何利用数据、计算和游说活动来引导
公共补贴、塑造标准，并预先取代地方监管，减少对人工智能基础设施建设地点和方式的民主控制。在一个气泡中，这种不平衡加剧：公共资金追逐炒作而非已证明的公共价值。即便是“软着陆”也可能使公共资产闲置。

泡泡是公共政策问题，因为它们将损失社会化，将收益私有化。征兆
人们都熟悉：估值过高的现象、生产力证据不足、积极的选址协议以及政策俘获。经济下行会加剧劳动和服务风险。当预期的AI效率未能出现时，组织通常会削减员工人数或公共服务以实现“节省”目标，同时继续支付已投入的AI合同费用。这种模式在相邻的“效率”技术周期中都有所观察，并且与泡沫破裂的过程一致：首先是质量下降，然后是股权。

数据与社会展示了数据中心泡沫主义为何能让社区承担后果；宏观
分析警告称，情绪驱动的繁荣可能迅速逆转；而《AI Now》解释了集中权力如何将炒作转化为公共承诺。公民社会应预见并衡量这些连锁反应。

公共AI

上一节关注了私人利益经济集中的风险
针对公共风险，指向推动公共人工智能的机会。

辉煌七股（或更准确地说，十股）的估值，其上（近40%的）繁荣
美国股市目前休整，持续创新高，随着前沿模型开发者高管层每一次不守规矩的媒体抨击和防御性播客亮相而摇摇欲坠。

与此同时，人工智能建设的普遍赢者通吃的框架——正如疯狂
Open AI 与地球上每一个重要的云、芯片和组件公司之间的交易（很快，可能就是在近地轨道上）——给任何替代叙事施加压力。

我们所在的本社区可以证明为什么公共应该拥有积木块和
支撑这一技术时代的基建。我们有责任和机遇来阐述我们所期望的未来，以及构建这个未来的价值，这个未来应该建立在属于公众的技术之上——不是作为商品，而是作为共享的基础设施，类似于公路系统或公共图书馆。

为什么不应该属于我们？它由我们创造，属于我们，自由地从我们这里拿走，又卖回给我们（并且，如果
如果历史是任何指南，很快也将被出售给广告商(如果可以的话）。我们想要的技术未来，在我们集体的智慧被吸收后，会给我们回报。

横跨基础设施、数据、计算和模型开发，我们识别出其中的一些
对可能成为未来一年最突出 RAI 问题的最重要贡献。

基础设施

公共人工智能的目标，由公共人工智能网络解释为“人工智能作为公共基础设施”建设
在 Mozilla 基金会强调公共产品、方向和使用，并继承早期关于数字公共基础设施的框架，是以公共基础设施而非仅仅是商业产品的方式提供资金、管理和定向的人工智能。这并非关于更安全的私有模型，而是关于跨计算、数据集、模型、工具和机构，为公共利益而管理的持久、共享资产。

The Public AI Network 的一个关键资源主张建立一个以公共事业为导向的堆栈，包含三
特点：公共访问、公共问责和永久的公共产品，它们就像其他关键基础设施一样被资助、管理和维护。

共享的公共AI基础设施是实现广泛访问、持久问责的最短路径。
长期创新能力。公共人工智能网络的公共设施为公共人工智能提供了定义和架构。

自由研究所项目数字基础设施解决方案赋权公民：一套工具
决策者提供了一个互补的路线图，说明政府如何实际设计和治理这种公共人工智能基础设施。它将数字基础设施，包括数据中心、云服务、协议和数据权利，视为太重要而不能交给专有利益，并提出了一套四阶段公共领导流程：评估现有基础设施和制度能力，设计开放性和互操作性，通过基于权利的治理和问责制进行保护，并通过数字素养和公民参与进行采用。围绕“数据

将“机构”作为数字主权的一条途径，工具包强调公共人工智能堆栈将只会为
公共利益在于人们有意义的声音、选择和利益，关乎他们的数据和基础设施如何被使用。结合公共人工智能网络的蓝图一起阅读，它将政府定位为不仅仅是私营人工智能的监管者，更是市场塑造者和共享人工智能基础设施的管理者。

公共人工智能定义了一个堆栈，公共和公民行为者协同管理计算、数据、模型和
开发者工具，具有开放的接口和可移植性。其最低可行标准包括公共访问、公共问责制和永久公共产品，防止封闭，并使能力可靠地提供给大学、初创公司、公民社会和政府公共服务。

效益是一个政经飞轮。公共AI投资扩大了接入，这拓宽了
参与问题解决，产生可见的公共利益，赢得持续资金信任。这使得人工智能的收益从目标公司层面的优势转变为广泛的社会能力。结果是主权与开放并存。

阿斯彭数字已成为这一议程的关键制度放大器和联盟构建者，
帮助将公共人工智能网络从小型学者和公民社会组织集群发展成为一个拥有350多名成员和100多个组织的国际社区，并提供运动所识别为缺失的“社区基础设施”：会议活动、运营能力和网络战略。阿斯彭数字已帮助将公共人工智能的抽象愿景转化为一个具体的实践社区。

通过开放接口和透明的文档，共享AI基础设施是可能的
数据溯源和评估。在机构层面，公共人工智能网络的建议是建立（或资助）公益性运营商——国家级实验室、市级联盟、公共云交换中心或大学公民联盟——这些运营商在公共章程下运行共享计算和宿主模型和数据集，并设有社区监督委员会和开放性能仪表板。

感谢像公共人工智能网络、阿斯彭数字和自由项目这样组织的支持，RAI
社区有一个强大的蓝图，说明如何将AI基础设施转化为我们共同建设和管理的公共福利。

数据

“公共数据”不只是 *可用* 数据；这是社区可以 *使用、管理和受益于*
跨越语言、能力和环境。四个资源指出了这需要什么：DAIR 的“每个数据集都有一个视角”提醒我们数据集编码了选择和权力；Mozilla 基金会的 Common Voice 展示了社区主导的数据创建如何解锁语音技术

为了服务资源匮乏的语言；Mozilla 与 EleutherAI 的最佳实践蓝图详细阐述了如何负责任地策展和发布开源大语言模型数据集；斯坦福大学 HAI 的 Mind the (Language) Gap 项目则描绘了低资源语言社群面临的结构性障碍。他们共同主张，需要公开参与式、文档完善且明确旨在消除不平等的数据。

DAIR的杂志是一个创造性的提醒，每个数据集都反映了选择：要收集什么，如何标记，
应该包含或删除哪些类别。因此，公共数据工作应随技术卡片一起发布数据集“视角注释”：谁决定了分类法，谁被遗漏了，做出了哪些权衡，以及用户如何可以质疑或修改记录。将视角从脚注提升为一流字段可以减少“中立戏剧”，并帮助从业者选择适合目的的数据。

莫扎瑞克基金会旗下的通用语音项目展示了一个可行的模型：一个全球性的、社区运营的

一个人们捐赠和验证语音片段的平台，以培养一个开放许可证的多语言
语料库，支持包容性语音工具和语言复兴工作。这种“参与式管道”协调激励（贡献者能看到用于自己语音的工具）并创建治理接触点（地方管理者、透明的发布）。公共数据计划应效仿这种结构：招募地方领导、补偿版主、并为每种语言发布贡献仪表盘。

Mozilla和Eleuther AI的面向LLM训练开放数据集的最佳实践
将价值观付诸实践：使用清晰的开放许可证，追踪来源和同意，发布丰富的文档（来源、过滤、已知差距），并规划治理（更新节奏、弃用、争议流程）。它还强调跨部门（法律、技术、政策）的工作，以及共享的元数据标准，以便数据集可以在不同项目中进行比较和复用，这是审计和下游问责制的必要前提。

知识共享的新CC信号框架为这幅图增加了一个缺失的互惠层：一个
为数据集持有者提供表达机器可读偏好的方式，关于其内容如何被用于AI训练和生成，这种偏好基于共享价值的理念而非纯粹的提取。在CC许可的成功基础上，CC信号提出了有限但有意义的选项，让社群能够标示在AI环境下可接受的用途、条件和“回馈”的期望——这为无限制的抓取和付费墙后的完全围封提供了替代方案。对于公共数据计划来说，这指出了一个超越文档的下一步：将开放许可和数据集卡片与编码社群规范和互惠期望的偏好信号相结合，以便“公共数据”不仅可被重用，而且还以能够长期维持共享价值的条款被重用。

HAI的《关注（语言）差距》文件表明，大多数主要的LLM对非英语的表现不佳
（尤其是低资源）语言，并遗漏文化和上下文中的细微差别。白皮书呼吁社区驱动的管道、地方资金和符合全球南方限制的权利尊重的数据访问。公共数据项目应优先考虑资源不足的语言：在社区环境中（学校、广播档案）资助收集，支持本地NLP实验室，并为低带宽环境和多种文字设计贡献工作流程。

公共数据只有当社区共同创建时，才成为公共资源，许可证
文档使其可验证地重用，并且公平性，特别是语言公平性，是一个设计要求，而不是事后考虑。DAIR解释了策展的政治；Mozilla的Common Voice展示了运营模式；Mozilla和Eleuther AI提供了战术指导方法；而HAI设定了全球语言议程。RAI社区有资源将这些经验融入资助、采购和标准中，因此我们的公共数据能更好地服务公众。

计算

公共、共享、通用计算能够促进包容性人工智能研究、弹性公共服务，和
问责式创新。最近的两个项目分别阐明了其“为何”和“如何”：Ada Lovelace研究所的计算公共领域为政府资助的计算访问制定政策设计；GovAI的推理扩展和人工智能治理则表明，仅针对训练计算进行治理将错失推理计算成为增长引擎时的潜在风险和机遇。合读之下，它们主张建设一种可访问、可管理、面向未来的共享计算。

公共计算包含使用公共资金以公共方式提供计算访问权的举措。
超级计算机、共享GPU集群或商业云上的信用额度，旨在支持市场上服务不足的研究、创新和公益使用，以及类似这样的举措

CalCompute, 新近在加利福尼亚州第53号法案(SB-53)中重新立法（目前尚在）。

全球方法各异（直接硬件、代金券、公私设施）且仍在
实验性的，但一个常见的目的是减少依赖性、扩大参与范围，并支持谁可以构建和评估AI的多元化
。

美国国家人工智能研究资源（NAIRR）试点计划提供早期
该共享计算模式的功能概览。通过其分配计划，NAIRR为研究者和教育者提供免费访问混合联邦和非政府资源的机会——超级计算机、云信用额度、试验台、数据集和隐私保护工具——其项目与AI安全、气候、健康和基础设施等社会相关重点领域相符。项目成果必须开放且可发表，分配具有时限性、同行评审性，并匹配明确证明的资源需求，在公共投资、科学进步和问责制之间形成一个内置的反馈循环。与Ada的计算共享平台和GovAI的推理扩展分析一同阅读，NAIRR的设计表明未来公共计算项目的关键要素：透明的、基于标准的分配；对开放传播的要求；对教育以及研究的支持；以及将先进计算明确导向市场服务不足的公共利益用例的明确授权。

治理正在从“训练运行有多大？”转变为“有多少推理计算能力？”
系统在部署或迭代训练期间会消耗多少资源？”忽视推理扩展的政策会导致风险误判、资源错配和问责制削弱。

GovAI 突出了两种重塑公共供给的途径：1) 部署时推理
(为推理、工具使用、搜索每个任务花费更多的测试时间计算)，以及 2）训练期间推理（使用重型测试时间推理来生成数据，然后进行蒸馏）。公共计算必须为持续、计量、可审计的推理而构建，而不仅仅是用于孤立的训练运行。

阿达·洛芙莱斯研究所主张，基于他们2024年关于公共计算的案例，一个可信的
计算公共项目程序应编码四层责任：1）访问与公平性，2）透明与可审计性，3）安全与负责任使用，以及4）多元化与反锁定。公共计算可以扩大参与并强化问责制，如果它是为我们将进入的世界而设计的：一个推理是主导的、持续负载以及主要治理杠杆的世界。

阿达·洛芙莱斯研究所的计算共享提供政策框架；GovAI的推理
规模解释了必须塑造它的技术趋势。RAI社区现在可以将这些融入预算、运营程序，并最终融入法律，以便共享计算能够真正实现共享公共价值。

模型开发

新证据由 GovAI 和爱丁堡大学发布，预测将迅速扩张
基于计算的规则所捕获的“前沿规模”模型的数量，这将使任何无法扩展证据收集和监督的治理方案都面临压力。与此同时，PAI表明，现实世界影响的部署后文档是目前最大的采用差距。两者共同论证，应从即时披露转向持续的、价值链问责制，并使用监管机构、采购人员和公众实际可以使用的人工制品。

百度的《记录基础模型的影响》发现，尽管提供者例行发布
模型/系统卡片，系统很少记录部署后的影响（在哪里，如何以及模型以何种效果被使用）。PAI 展示了随着发布类型（托管 API 与开放权重）义务如何变化，并促使模型托管方、应用程序开发者和云服务提供商之间进行协作

为开放模型提供监控事件和政策违规行为，而无需重新创建
监视。

静态计算阈值将很快涵盖许多更多模型。GovAI 和 大学
爱丁堡关于前沿AI模型数量趋势的分析预测，到2028年底，将有103-306个模型超过10^{25} FLOP（欧盟AI法案的指标），以及45-148个模型超过10^{26} FLOP（美国AI扩散框架“受控模型”），且每年呈超线性增长。文档、评估和执行的容量规划必须假设这种增长。与前沿连接的阈值更稳定但仍需筛选。相对于最大训练运行定义的阈值每年大约产生14-16个模型，这表明监管机构应将绝对触发点与相对触发点相结合，无论如何，都应投资于可扩展的证据管道。

向前看，RAI社区对一些重要的2026优先事项有指导方针。值得信赖
模型开发现在意味着实时、可验证的责任制：发布后持续存在的影响文档，与发布策略相匹配的责任，以及根据我们将实际看到的前沿级模型数量进行扩展的监督，而不是我们目前拥有的那少数几个。

公共利益型人工智能

我们想要的技术未来服务于我们所有的需求，负责任地应用AI解决方案
社会最紧迫的问题。人们不认为“AI为了公共利益”会公平或均匀地到来。公共利益是情境性的且存在争议，它应该与社区协商，而不是从中心宣布。

在阿达·洛芙莱斯研究所的《做好》报告中，研究参与者始终如一地定义了公共
围绕公平与公正、社会联系与社区、以及健全的公共服务，让人们能够有意义、安全地生活。

该报告呼吁采用参与式、基于地的方**法，**暴露具体的期望：**即**人工智能
要做利他且公平，注重关系和关爱，面向未来，并且只在必要且有效的情况下负责任地部署。在此框架中，公共利益人工智能，与其说是将项目品牌化为“为了公益”，不如说是重新设计治理，以便公众定义的公益（这根植于生活经验、地域和结构性条件）实际上引导投资、部署和监督。

对于RAI社区来说，核心教训是公共利益不能从通用中推断
态度调查或创新叙事；它必须通过深入的、持续的与多元化公众的参与来建立，包括那些通常被排除在AI决策之外的人们。

公共利益人工智能（PIAI）项目，锚定于亚历山大·冯·
洪堡互联网与社会研究所将PIAI定义为支持“一个被构成为公共的社会集体的长期生存和福祉”的系统，借鉴了杜威、博赞曼等人的公共利益理论。公共利益并非被假定为普遍或显而易见，且不能仅留给私营企业或技术专家来理解，而必须通过参与式过程为每个问题进行定义。

PIAI项目地图是一个旨在明确服务于公共利益的人工智能项目结构化数据集

目标。交互式世界地图汇总了关于项目目标、方法和
框架并发布它们，既作为视觉目录也作为开放研究数据集，包括在Hugging Face等平台上。这使得公共利益AI作为一个全球实践变得可读：谁在做什么，在哪里，使用什么变革理论，以及有什么资金和制度支持。

对于RAI社区而言，地图和网络可以帮助将资金与未得到充分服务的领域进行匹配。
将本地倡议连接成全球联盟，并在可长期评估的具体项目中落实高级原则。

至关重要地，这不仅仅是理论：越来越多的举措已经体现了公益人工智能
在实际中，在生态系统的不同层级上。消费者报告关于"消费者授权的人工智能代理"的工作和忠诚代理原型显示了，当人工智能代理被设计为为人服务而不是平台时是什么样子——帮助消费者提交反馈、导航商业活动，并以与其利益和同意一致的方式行使他们的权利，而不是数据提取。

"人工智能驱动的非营利组织正在展示负责任的人工智能可以是什么样子。快速前方的数据显示，71%的AI赋能非营利组织受访者已经建立了评估和缓解AI风险的过程，但41%的人指出缺乏内部技术专业知识是阻碍他们实现道德目标的一个障碍。在拨款方面，资助者应鼓励透明政策，支持道德培训，并提供治理模板——并提供必要的资金来做到这一点。通过支持负责任的做法，资助者表明公平和问责与创新一样重要。

凯文·巴拉特
联合创始人，极速前进

快速前进的《慈善重置：人工智能时代慈善如何引领》报告
记录了非营利组织如何部署人工智能来扩展对医疗保健、教育、气候适应性和正义的获取，同时揭示出必须填补的基金和能力差距，以便使命驱动的组织能够塑造人工智能而不是简单地吸收它。

在基础设施层，OpenFold3展示了前沿科学能力如何能够
作为一个共享资源库：一个用于蛋白质和药物结构预测的开源基础模型，它加速了超越任何单一公司围墙的生物医学研究。

同样，帕特里克·J·麦戈文基金会的资助守护者工具应用人工智能来加强
慈善尽职调查和透明度，使资本配置更加严格，并让员工能够专注于与资助人的关系性工作。

在安全层，ROOST的开源Coop和Osprey工具为企业级信任和
安全基础设施——内容审查控制台、调查和事件响应能力——在那些自己根本无法构建或购买此类系统的组织能够触及的范围内，

将在线安全转变为一个共享的、可检查的公共产品，而不是专有优势。

综合起来，这些努力说明了一个多层次的公共利益AI生态系统可以是什么样子。
例如：社群界定公共产品；研究人员绘制领域；消费者代言人、非营利组织和基金会构建具体工具以重新分配权力和能力；以及开放的科学和安全基础设施，供他人复用和审查。

对于RAI社区来说，当前的任务是不要将这些项目视为孤立的“成功故事”而是
作为可扩展的模式，利用资金、采购和标准来奖励忠于人民、公开治理、并致力于持久公共利益的人工智能。

重塑人工智能

RAI最宏伟的版本是雄心勃勃的人工智能，它使我们能够重新构想我们的技术未来和
实现我们最狂野的愿望。这是我们接下来真正想要关注的RAI。

雄心勃勃的人工智能未来不只是技术场景。它们是故事、记忆和图像
关于这项技术是为谁服务的，以及它能够实现什么样的生活。如果我们想要从人工智能中获得不同的结果，就需要改变塑造我们最初想象、资助和评估技术的文化和叙事基础设施。

一些项目彰显了2025年那种雄心勃勃的重新构想的精神。集体
智能项目（CIP）的创始白皮书明确指出，认为当前的治理模式将我们困在进步、参与和安全之间的“变革技术三元困境”中，并提出了一种“CI堆栈”的民主流程和新型制度形式，使我们能够有意识地重塑引导人工智能等变革技术规则和容器。

我们想要的技术 (TTWW) 正在围绕以人为本、以目标为导向构建一个替代技术生态系统。
利润，明确地以那些已经在努力推动人人都能繁荣的技术中的领导者和社区为中心。其基础设施基金利用社区赋权、参与式预算和基于信任的慈善事业来资助替代技术、叙事和治理实验，将资金本身视为文化变革的场所，而不是一个中立管道。

一部技术的人民史通过汇集一个活的档案来补充这一点，关于如何
日常人们和劳动者体验着手机、劳动平台和社交媒体等技术，而不是复述孤立发明家的英雄故事，我们期待人工智能必然会被赋予“人民的历史”的待遇，我们想象这能帮助引导通用人工智能走向科技的“人民未来”。

理想的人工智能未来应该根植于围绕谁的生活历史和物质权力转变
获得资金，谁能讲述“创新”的故事，以及谁的经历算作专业知识。

DAIR的《可能未来》系列和杂志进一步通过将推测性想象变成
实践，而非装饰。通过工作坊和印张制作，参与者从未来写日记条目，制作拼贴画，并实验“想象其他样子”：询问将会发生什么

当我们可以允许自己想象自动化炒作之外的科技未来时
不可避免性叙事。

这些未来常常是混乱的、局部的，并且基于关怀、团结和生存，尤其是对
受当前人工智能部署所害者。推测性F(r)小说的“活体图书馆”提供了一个补充视角：它将摩擦——即人工智能系统和治理中的不适、紧张和缓慢——视为一种资源而非缺陷，利用设计小说和故事讲述来厘清人工智能中的“形式、功能、小说与摩擦”。

几个项目将这种重新构想一直推进到基础设施和供应链中
人工智能所依赖的基础。Estampa的生成式人工智能地图项目提供了一个批判性的视觉地图，描绘了生成式人工智能生态系统，不仅追踪模型和公司，还包括数据提取、隐形劳动、环境影响和真理体制——明确挑战那些将人工智能呈现为无重量、中立或不可避免的地图。

人工智能+行星正义联盟的《人工智能原材料》小型杂志在层面执行了类似的动作
关于硬件，直观地拆解驱动芯片、电池和数据中心的原材料（铜、硅、稀土、钴等），并将“算法之下是什么”明确关联到资源开采、不均等的环境负担以及行星正义斗争。

与 CIP 对新的集体智能机构发出的呼吁一起阅读，这些作品坚持认为
重新构想人工智能也意味着重新构想谁管理基础设施堆栈——从矿产边界和物流到企业实验室和监管机构。

rai工作应该常规地融入推测和叙述方法以及批判性地图
关于权力和物质，特别是与边缘化社群，揭露盲点，挑战主导轨迹，并设计允许拒绝和替代路径的政策。

内核杂志和更好的AI图像展示了文化和美学如何塑造边界
在可能性的范畴内。Kernel 是 Reboot 的年度出版物，提供长篇散文、叙事、诗歌和艺术作品，重新构想技术乐观主义，既不屈服于宿命论，也不沉溺于炒作。它被明确框定为“历史研究、文化批判和可能的未来”的家园，由反思其构建和居住的系统的技术专家们创造。

更好看的AI处理了一个更具体但更强大的层次：视觉隐喻。通过创建和
精选非刻板印象、自由许可的图像来替换发光大脑、人形机器人和抽象二进制代码，该项目旨在提高公众对人工智能优势与局限性的理解，并对抗那些编码恐惧、非人化和偏见的假设的视觉套路。

埃斯塔姆关于生成式人工智能的地图与人工智能+行星正义联盟的关于人工智能的原始材料
将这种视觉政治扩展到地图和图表中，使“智能”系统赖以支撑的基础设施、掠夺性地理和有争议的真相可见。

这些努力提醒我们，人工智能是如何被描绘的，在媒体、政策文件、教育资料等。
悄无声息地塑造着谁有权质疑它，谁被想象为主体或客体，以及哪些未来看似可行。对于公民社会和更广泛的RAI生态系统而言，这些项目共同指向了对“期望未来”更宏伟的理解。

仅仅围绕当前轨迹阐明护栏是不够的；我们必须投资于叙事
基础设施、参与式记忆项目、推测性方法以及扩展我们甚至能够看到的未来的视觉文化。这意味着资助和使艺术家、说书人和社区档案管理员与律师和工程师一样合法化；建立参与式基金和媒体实验室

资源替代技术和管理实验；以及嵌入投机和
将拥抱摩擦力的实践融入政策制定、标准工作和公众参与。

我们不寻求抽象的乌托邦。从这个意义上说，重塑的AI未来是持续的
重构谁能首先想象、建造和从技术中获益的集体项目。

展望未来

随着我们结束这份2025年负责任人工智能影响报告，我们已经开始将工作重心转向2026年的现实与责任。对于All Tech Is Human的RAI工作而言，核心问题已不再是人工智能是否会塑造我们的机构、经济和日常生活，而是如何塑造，以及在谁的利益之下。我们的答案一如既往，是构建并强化负责任科技生态系统本身：即能够引导人工智能走向集体利益、而非我们所力求避免的未来的人才、实践和公共基础设施。

在来年，我们打算越来越多地负责任地利用人工智能服务于我们的使命，作为
公共利益AI的坚定实践者。我们希望开发一个AI驱动的版本我们的责任技术工作平台，以更好地匹配各种人才与有意义的职位，发掘新兴技能，并使生态系统对试图进入其中的人更加易于导航。我们希望利用AI辅助匹配、指导和资源策展，扩大我们的责任技术指导计划，始终以人类判断、社区规范和公平为中心，使新声音和非传统途径更容易在这个领域找到归属。为此，我们希望我们的自身基础设施成为公共利益的负责任、使命一致的AI部署的一个范例。

我们也打算深化对公共人工智能基础设施的承诺：被当作共享的人工智能
作为公共利益的守护者，而非一系列孤岛化的产品。到2026年，我们计划支持和放大正在进行的努力，以建立和管理社区数据集、共享计算和根据公共和公民宪章开放的工具。我们寻求与志同道合的组织合作，原型设计“公共利益运营商”，他们可以托管和维护这些资源。

纵观所有这一切，我们认为人工智能责任素养是主线。明年我们将扩大我们的
要让 RAI 不再仅仅是专家和内部人士的领域，而成为一种共享的公民能力，教育工作要：将我们的课程供应扩展到 2025 年发布的最初五门课程之外，并针对更多受众和平台重新设计我们的课程库。

我们的2026年RAI工作流程将专注于三个以人工智能为中心的问题领域。

首先，训练数据：我们试图弥合社区数据集之间的脱节
为小型、有针对性的公共模型和日益增长的私有商业基础模型提供动力，并通过不断扩大的产品和接口，将越来越多地扩展到利用现实世界监控进行持续数据收集的可穿戴设备和嵌入式设备。我们的工作将强调数据权利、社区治理和公益训练管道，将其作为相互确保监控的替代方案。

其次，合成伙伴和合成媒体：我们打算加大力度确保人工智能
伙伴关系减少隔阂，而非加剧；为高风险使用（涉及青少年和弱势用户）提供治理支持，并在日益难以辨别是谁或是什么在发言的世界中，加强信息完整性和内容溯源工作。

第三，信任与信心：我们将优先考虑人工智能治理的连接组织，包括
标准、基准测试、评估、审计和红队演练，以便行业确信工作建立在更稳固的内部信任体系之上。我们打算参与民间社会活动，引领构建代理式人工智能所需的问责框架。

我们不假设更好的未来会从技术进步中自行出现。在2026年，
我们打算帮助他们：通过使用AI的方式加强，而不是掏空，负责任的科技生态系统；通过支持共享的公共AI基础设施的增长；并通过提高负责任AI素养的水平，以便更多的人能够理解、质疑和引导塑造他们生活的系统。接下来的工作充满挑战，但它也是富有成效的：每一个为公共利益管理的数据集，每一个与社区共同创造的基准，以及每一个被赋予新能力、负责任地参与AI的人，都是迈向我们所打算争取的未来的又一步。

关于所有科技都是人类的

所有科技都是人类是一家非营利组织，致力于通过团结公民社会、政府、产业和学术界广泛的大群和思想，构建一个更负责任、更具包容性和更符合道德的科技未来。我们采用生态系统整体方法，将技术专家、学者、公民社会领袖和平常的个体聚集在一起，共同应对围绕人工智能治理、数字权利、平台问责制以及技术更广泛的社会影响等方面的复杂挑战。

我们旨在开发一种更好的方法来应对棘手的技术与社会问题；这种方法利用集体智慧、参与和行动，以创造所需的系统性变革。

通过社群建设、人才培养、公共项目、研究和分析，我们组织促进对话，弥合学科之间的差距，并鼓励为塑造与我们的价值观相符的技术而共同承担责任。通过促进合作和提升广泛的观点，All Tech Is Human 致力于确保技术由公众利益驱动和塑造。

如果您想与我们合作或了解更多信息，请发送电子邮件至hello@alltechishuman.org。

了解更多负责任的AI

所有科技都是人类的 launches 五个 dynamic 负责任的 AI 课程, designed for aspiring RAI从业者、人工智能治理专业人员，以及那些准备在组织内建立人工智能治理计划的人。

这些课程为参与者提供了开始过程所必需的基础知识。在组织内部有效实施负责任的AI和AI治理计划。

植根于实用见解和实际应用，这个包含五门短期课程的系列，能够只需几个小时即可完成，提供对负责任人工智能基础知识的入门了解：其原则、历史和作为领域的演变，以及当前角色、行业最佳实践和不断发展的治理格局。

感谢 The 的慷慨支持，使我们的 RAI 课程得以实现并免费提供。帕特里克·J·麦格overn基金会。

all tech is
human

一起，我们致力于解决科技和社会中最棘手的问题。