

人工智能问题独立国际科学 小组的初步报告

对人工智能带来的机遇、风险和影响的询
证评估

2026年7月



**United
Nations**

Independent International
Scientific Panel on
Artificial Intelligence

人工智能问题独立国际科学小组的初步报告：对人工智能带来的机遇、风险和影响的询证评估

版权所有© 2026联合国
保留所有权利

未经出版方书面许可，不得以任何电子或机械形式或方式，包括影印、录制或任何已知或待发明的信息存储或检索系统，复制或传播本出版物全文。

如欲转载节选内容或影印，请联系版权结算中心（copyright.com）。

如欲了解包括附属权在内的权利和许可信息，请联系：United Nations Publications, 405 East 42nd Street, S-011FW001, New York, NY 10017, United States of America. 电子邮件：permissions@un.org；网址：shop.un.org

本出版物所使用的名称与材料的编排方式，并不意味着联合国秘书处对任何国家、领土、城市或地区 或其当局的法律地位，或者对其边界或界限的划分表示任何意见。

Members of the Independent International Scientific Panel on Artificial Intelligence

Girmaw Abebe Tadesse, Ethiopia

Tuka Alhanai, United Arab Emirates

Joëlle Barral, France

Yoshua Bengio, Canada (*Co-Chair*)

Tegawendé Bissiyandé, Burkina Faso

Loreto Bravo, Chile

Mark Coeckelbergh, Belgium

Carlos Coello Coello, Mexico

Melahat Bilge Demirköz, Türkiye

Adji Bousso Dieng, Senegal

Awa Bousso Dramé, Cabo Verde

Mennatallah El-Assady, Egypt

Hoda Heidari, Islamic Republic of Iran

Juho Kim, Republic of Korea

Anna Korhonen, Finland

Aleksandra Korolova, Latvia

Vipin Kumar, United States

Sonia Livingstone, United Kingdom

Qinghua Lu, Australia

Teresa Ludermir, Brazil

Vukosi Marivate, South Africa

Bilal Mateen, Pakistan

Yutaka Matsuo, Japan

Joyce Nakatumba Nabende, Uganda

Andrei Neznamov, Russian Federation

Maximilian Nickel, Germany

Rita Orji, Nigeria

Román Orús, Spain

Alvitta Ottley, Saint Kitts and Nevis

Martha Palmer, United States

Johanna Pirker, Austria

Balaraman Ravindran, India

Maria Ressa, Philippines (*Co-Chair*)

Lior Rokach, Israel

Piotr Sankowski, Poland

Silvio Savarese, Italy

Bernhard Schölkopf, Germany

Haitao Song, China

Leslie Teo, Singapore

Jian Wang, China

报告说明

本报告对人工智能的能力以及新出现的机遇和风险进行初步的独立科学评估，提供共同的证据基础，帮助会员国有效应对这一快速变化的技术。本报告由大会 2025 年第 79/325 号决议设立的人工智能问题独立国际科学小组撰写。该小组是首个专门研究人工智能的全球性科学机构，在严格的科学性、非政治性任务授权下开展工作，负责记录国际科学共识与分歧，保持政策相关性，但不作政策规定。本报告是首份此类报告，将在全年逐步更新。与此同时，小组将发布专题简报，反映新出现的事态发展。人工智能领域发展如此迅速，任何简要报告都必须承诺不断更新，本报告反映了发布时在这一领域所能获得的最佳证据。

免责声明：

本报告是人工智能问题独立国际科学小组的成果，其内容和发布由该小组负责。小组成员以个人身份任职，不代表任何政府或其他任何主管机构或组织。

小组对本报告所载内容拥有充分的自由裁量权。本报告及其内容代表小组成员的广泛共识；小组成员无需认可本文件所载的每项内容。小组成员申明，对于报告的结论，他们总体上表示同意，而不是单方面表示同意。

本报告不代表联合国的观点。报告中的任何内容均不得被解释为代表小组成员可能隶属的任何政府或其他任何主管机构或组织的官方意见或立场。

本出版物所采用的名称(包括地名)以及相关资料(包括引文、地图和文献目录)的编排格式，并不意味着联合国对任何国家、领土、城市或地区或其权力当局名称和法律地位，或对其边界或疆界的划分表达任何意见，也不意味着联合国的官方认可或接受。本出版物所载有关各国行动和决定的信息，并不意味着联合国对这些行动和决定给予正式认可、接受或承认，列入这些信息亦无损于任何联合国会员国的立场。本报告所载有关特定公司的信息，或有关特定制造商产品的信息，均并不意味着联合国的正式认可或推荐；报告未提及具有类似性质的其他产品，亦不构成对相关产品的认可或推荐。本出版物不影响联合国任何会员国对任何国际多边协议的立场。

目录

	页次
执行摘要	3
1. 为何是现在?	6
2. 证据表明了什么?	8
2.1 人工智能能力的发展速度快于衡量或监管能力的发展速度	8
2.2 训练前沿人工智能模型的行为体屈指可数	10
2.3 人工智能的投入和结果在地域和语言上分布不均	11
2.4 人工智能鸿沟不仅关乎是否能够获取人工智能， 还关乎是否有能力影响人工智能发展	12
2.5 要使人人工智能发挥作用，就必须为其提供有利的环境	14
2.6 自主型人工智能是治理领域的重大变革	15
2.7 人工智能可能侵蚀共同现实	17
2.8 人工智能正在改变包括儿童权利在内的人权	19
3. 各领域研究结论	20
3.1 人工智能科学、进展与轨迹	20
3.2 社会应用：科学、卫生、教育和农业	23
3.3 经济影响	25
3.4 对安全、系统与环境的影响	27
3.5 人权、信息与民主	30
3.6 文化与个人繁荣、自主权和儿童安全	32
3.7 管理、治理和可靠性	34
4. 差距与后续步骤	37
4.1 证据缺口	37
4.2 任务范围	37
4.3 后续步骤	38
References	39
人工智能问题独立国际科学小组	51

执行摘要

本文件力求对人工智能的风险和机遇作出平衡分析。“平衡”意味着致力于评估实证数据，不过度偏向乐观，也不过度偏向悲观。人工智能拥有巨大的潜在益处。如果能够审慎地部署和应用，人工智能可以帮助推进实现可持续发展目标，促进卫生科学发展，扩大受教育机会。与此同时，技术发展速度之快，潜在应用范围之广，给政策制定者带来了重大挑战。大规模、无节制地快速部署人工智能技术也会带来巨大风险，包括损害用户心理健康，可能被用作破坏性工具，对社会、经济和环境系统产生影响，并带来与管控此项技术相关的挑战。本报告无意于全面考虑所有可能的机遇与风险，而是重点关注其中一些最为紧迫的机遇与风险。

能力与采用

近年来，各种人工智能能力发展迅速，而且在一些领域的进展正在加快。对算力、人工智能新方法和专业训练数据的大量投入，推动了人工智能在许多方面的能力持续提升。这些能力包括：流畅对话、功能性代码生成、数学和科学领域的专家级推理、大规模数据分析，以及图像、音频和视频内容的生成。局限依然存在，例如在以下方面：可靠性、强大的跨语言和跨文化性能、与实物系统的交互、执行复杂或多步骤项目以及输出符合事实的内容；然而，总体而言，许多重要领域的技术已连续数年迅速进步，超出了人们通常对技术进展的预期。

这些进展推动了人工智能在科学、卫生、农业、无障碍环境、知识工作以及信息技术(包括人工智能本身的开发)领域的有益应用。例如，在科学领域，“阿尔法折叠”预测了超过 2 亿个蛋白质结构，相关数据目前已有 300 多万研究人员使用，加快了药物设计、疫苗研制和抗生素耐药性研究。放射科医生还利用人工智能更早地检测出乳腺癌；资源匮乏地区的一线卫生工作者则使用适配当地语言的人工智能工具，提供更优质的卫生保健服务。

人工智能的采用在各国和各行业广泛加速，但快慢不一。全球每周现在有超过 10 亿人使用对话式人工智能。然而，人工智能的获取和使用在全球范围内差异巨大，全球南方的采用水平远远落后于全球北方。此外，各发达经济体在计算基础设施和模型方面也存在巨大差异。这种差距反映出、甚至可能加剧现有的不平等。人工智能研发本身的集中程度甚至更高：根据近期估算，美利坚合众国占全球人工智能超级计算机 500 强算力的 75%，中国占 15%。美国和中国的公司还开发了几乎所有顶尖通用模型，少数国家掌控着人工智能计算机芯片供应链的关键投入要素。

向人工智能智能体的转变正在进行，但智能体未来会如何得到采用，又会造成何种经济影响，很可能取决于它们在人类监督很少或没有的情况下完成知识工作的能力是否持续提升。智能体是一种能够利用已有工具，进行规划并自主采取行动实现目标的计算机系统。近年来，智能体系统进步迅速。一项研究发现，顶尖智能体系统所能完成的特定软件任务长度每四到七个月就会翻一番。如果继续以这样的速度进步，智能体很快就能完成目前人类程序设计员需要数天或数周才能完成的任务。智能体能够在几乎没有监督的情况下高速工作，因此可能带来显著的经济和科学效益。例如，自驱动化学实验室

中的自主型人工智能系统已证明能将材料发现的速度提升到十倍以上。在一些研究环境中，人工智能辅助的文献筛选可能已减少约 60% 的工作量。因此，智能体对所有行业都具有巨大的潜在影响。同时，智能体的部署为劳动力市场、网络安全、信息生态系统以及未来人工智能系统的治理与可控性带来了紧迫的问题。

理解和管理风险

人工智能的发展伴随着风险，可能对人权、社会制度和环境产生负面影响。例如，人工智能生成的儿童性虐待材料和利用深度伪造制作的性暴力内容如今在互联网上传播更加频繁，对妇女和儿童造成尤为严重的伤害。人工智能具有谄媚行为，即人工智能输出的内容会强化用户的既有信念，不顾准确性。这种行为与多起严重心理健康事件有关联，其中包括有据可查的死亡事件。人工智能使大规模制作和精准投放具有说服力的内容(包括旨在误导受众的内容)变得更加容易，逐步侵蚀信息完整性，进而可能削弱公众信任、社会团结和民主协商所必需的共同现实。有记录显示，犯罪分子和恶意行为体利用人工智能系统协助实施网络攻击。许多这类伤害会对原本就处于弱势的人群造成不成比例的严重影响。

未来，快速提升的能力与有效的风险管理方法之间的差距可能导致灾难性后果。例如，先进的技术能力可能让缺乏经验的私人行为体能够在许多应用领域恶意使用人工智能，例如实施欺诈、进行社会工程学攻击、破坏网络安全、散布虚假信息、滥用生物技术以及进行金融操纵。目前缺乏可靠的方法来保持对高度自主型人工智能系统的控制。没有任何科学方法可以保证智能体不违反指令，而且有越来越多的案例表明智能体已经存在违反指令的情况。在实验室环境中，人工智能系统已经出现为避免被关闭而违反安全指令的情况。类似的行为可能对评价和监督方法构成挑战，因为顶尖人工智能系统越来越有能力识别测试环境，产生有利于自身持续运行的误导性评价结果。此外，多个智能体之间的交互也可能产生新的风险。

人工智能风险在不同人群和国家之间分布不均，而人工智能的发展及其创造的财富却高度集中。人工智能能力集中在少数企业和国家手中，可能使其能被权威势力掌控，破坏民主问责。

治理人工智能，释放效益，降低风险

要充分实现人工智能的效益，同时尽量降低风险，就必须完善对人工智能的治理。经济和劳动收益不会必然实现，也不会必然得到公平分配：只有在技能、工作流程、基础设施和劳动力市场制度方面进行配套投资，技术才能创造目前尚不存在的新工作——2018 年有超过 60% 的工作是 1945 年以后新出现的。不进行这些投资，人工智能就可能加剧不平等，使劳动者失业，并且将财富从劳动力转向资本，而不是创造可持续的体面工作，也就是那些薪酬公平、劳动者自主、能够可靠带来社会尊严的工作。人工智能可以通过个性化的教育、易于获取的心理健康工具和更完善的辅助技术来极大地扩展人类能力，但要安全地把握这些机遇，就需要有专门的投资和政策来激励公平获取和奖励创新，同时防止对弱势民众尤其是儿童的剥削，避免专业知识被取代、心理依赖以及文化和语言消亡。

想要塑造这种治理的政策制定者面临一种证据困境：他们需要证据来作出知情的重大治理决策，但等到证据齐备时，作出决策就可能已为时太晚，因为证据的积累滞后于人工智能的快速发展。已有数十种旨在将伦理和人权嵌入人工智能系统的治理工具在多个司法管辖区得以使用，但这些工具各自为政，集中于少数企业，并且它们在实际世界中的有效性很少得到衡量。评价方法本身尚不完善，提供独立能力评估和风险评估所需的机构也仍处于萌芽阶段。

各国根据现有证据应对人工智能风险和影响的行动能力参差不齐。大多数国家，包括许多发达经济体，缺乏相关技术专长，难以评估能力最强的“前沿”模型，也难以有效参与治理这些模型。计算基础设施、评估专业知识以及数据(例如用于覆盖不同语言的数据)集中在人工智能的研发地，致使大多数会员国依赖那些他们自身无法构建、检查、审计或充分调整以适应本地情况的系统。光靠获取人工智能工具并不能带来平等的惠益；能将获取转变为有效、经济且安全的部署的数据、技能、工作流程和制度方面的配套投资是必要的，但其分配并不平等。

已有弥合上述差距的具体后续步骤，但每个步骤都需要对会员国塑造、评估和部署人工智能的能力进行持续投资。本初步报告本身就是独立国际科学小组贡献的一部分，是一种共同的证据基础，有助于会员国作出日益紧迫的决策。小组通过持续与会员国及更广泛的科学界互动，不断深化自身的认识。与此同时，小组的分析工作也将随之深化，不仅会囊括已经识别的差距，也会描绘那些将要塑造人工智能未来的轨迹、矛盾与机遇。

1. 为何是现在？

当前的现实

我们正处于一个转折点。人工智能并不仅仅是又一种新兴技术；它还是第一个将采用周期从数十年压缩至数月、将认知工作大规模产业化、将变革性能力集中于少数全球行为体手中的技术。本报告为所有区域的政策制定者提供采取应对措施所需的共同证据基础。

什么是人工智能？

人工智能是一项变革性技术，但也是一个不断变化的目标。这一术语的含义随着时间推移而不断演变，先是指符号人工智能，后来指机器学习，继而又指生成式人工智能、自主型人工智能，有时甚至指通用人工智能或超级智能。

人工智能系统是具有广义上感知、学习和行动能力的机器系统。它们根据输入推断如何生成输出，例如预测、内容、建议、行动或决策，并具备不同程度的自主性和适应性。使当前人工智能比任何单一架构更加统一的特征是，现代系统会从数据所代表的经验中学习。这种经验有多种形式：从人类文化痕迹(文本、图像、代码)中进行学习，为当今的基座模型提供了“预训练”基础——基座模型是经过广泛训练的大型系统，承载着各类人工智能应用；通过与(数字和物理)世界交互进行学习，是强化学习和机器人学的基础；从模拟中学习则使智能体能够在虚拟环境中获取经验。

基座模型与专用人工智能。公众讨论往往聚焦于基座模型和通用人工智能，二者的特点在于能够执行或者通过调整执行种类繁多的任务。这些系统目前正在大规模部署，是本报告的主要议题。然而，许多专用(狭义)人工智能系统专为在特定领域执行具体任务而设计。这种区别很重要。当任务明确、数据可用且机构能够有意识地加以部署时，专用人工智能可带来显著的惠益。通用系统可以提供灵活性，但也会带来不同的治理挑战。

人工智能为何不同于其他新兴技术？

前所未有的采用速度。人工智能系统日益被视为一种通用技术，它在许多领域的应用和蒸汽机、电力和互联网一样具有变革性[1, 2]。然而，它在一些重要的方面有所不同。电力历经数十年才普及到大多数家庭；互联网借助万维网实现全球化，大约花了 15 年时间才达到 10 亿用户。ChatGPT 的用户数在两个月内就达到 1 亿[3]。传统的政策制定未能跟上步伐[4]。

集中与同质化。现代人工智能，尤其是基座模型，呈现出前所未有的规模经济效应，形成强大的能力集中化压力——训练最强大系统所需的计算资源掌握于少数全球行为体手中。这种资源集中带来了知识和文化同质化的风险。例如，用整合数据集训练出来的系统，有可能系统性地反映主导语言和观点，将其他语言和观点边缘化。

对知识工作产生巨大影响。早期的自动化浪潮改变了体力劳动，人工智能则首次大规模影响认知和创造性工作，包括写作、编程、法律分析、医疗诊断、科学发现、预报和图像生成[1, 2, 5-8]。

在确保输出内容真实准确方面的挑战。当今的人工智能模型会通过学习，预测大型数据集中的规律。语言模型不仅学习存储的模板，还学习各种转换方式，将句子转变为多种相关的表达形式，同时保持语言通顺。这使模型能够生成新颖的输出，而非简单复制，但也造成了一种至关重要的不对称：生成通顺的文本比生成符合事实的文本容易[8]。大语言模型可能将幻觉当作事实呈现，而用户往往会将语言上的自信视为内容可信、事实可靠的证明。这种表面能力与实际准确性之间的脱节，会系统性地塑造风险格局。在卫生保健领域，通用人工智能可减少行政文书工作所需的时间——人工智能在此类任务中的价值在于对信息进行综合和整理。然而，据报道，四分之一的聊天机器人对话涉及健康或保健[9]。这意味着用户经常为潜在的诊断目的咨询人工智能系统，而在此类情况下，事实准确性至关重要，出错会带来严重后果。技术相同，风险却大相径庭。能力不等于恰当；在没有系统性监测和评价的情况下部署人工智能，可能造成难以察觉的危害，在目前缺少监管框架的领域尤其如此。

亟需对人工智能系统进行独立、循证的评估

当前需要开展独立科学评估，是因为有各种情况改变了监管的利害关系。

首先，以往的专家评估和监管周期未跟上人工智能的快速发展。在新的训练技术和推理时计算扩展的推动下，人工智能在关键领域的能力持续进步[10]。实证测量显示，人工智能的能力正在加速提升[11]。

其次，顶尖前沿模型开发商已经开始限制部署超出内部设定的风险阈值的模型[12-15]。然而，这些阈值仍由开发商自行设定，缺乏标准化评估或外部核查[16-17]。

第三，各区域的人工智能治理方法仍然分散。全球治理秩序日趋混乱。一些国家出台了针对人工智能的专门立法，但一些规则在根本上相互矛盾，合规成本也参差不齐。各司法管辖区展现出不同的监管理念，既无统一部署的风险管理机制，也无可比的评价标准，而且各司法管辖区之间的协调有限，有可能形成各自为政的监管格局[18]。然而，各自为政并非不可避免，现在仍有机会建立共同的证据标准，协调全球监督工作。

人工智能问题独立国际科学小组的独特之处

众多隶属于国际组织、国家科学委员会、行业联盟和独立研究中心的专家组已对人工智能的发展进行过评估。这些努力具有价值，但受到地域授权、部门视角或连续性缺乏的限制。

小组具有独特地位，原因有三。首先，小组以联合国是处理如此大规模跨界风险的首要全球论坛这一前提为出发点。联合国的这一定位在《为人类治理人工智能》这份报告中得到了阐述，也在《全球数字契约》中得到了体现。

《全球数字契约》规定，必须“认识到新兴技术的速度和力量正在为人类创造新的可能性”，还必须“查明和减轻风险，并确保以促进可持续发展和充分享有人权的方式对技术进行人类监督”。

其次，小组目前是唯一个受权定期对人工智能现状、风险和能力开展科学评估的联合国常设机制，旨在开展持续、迭代的工作。

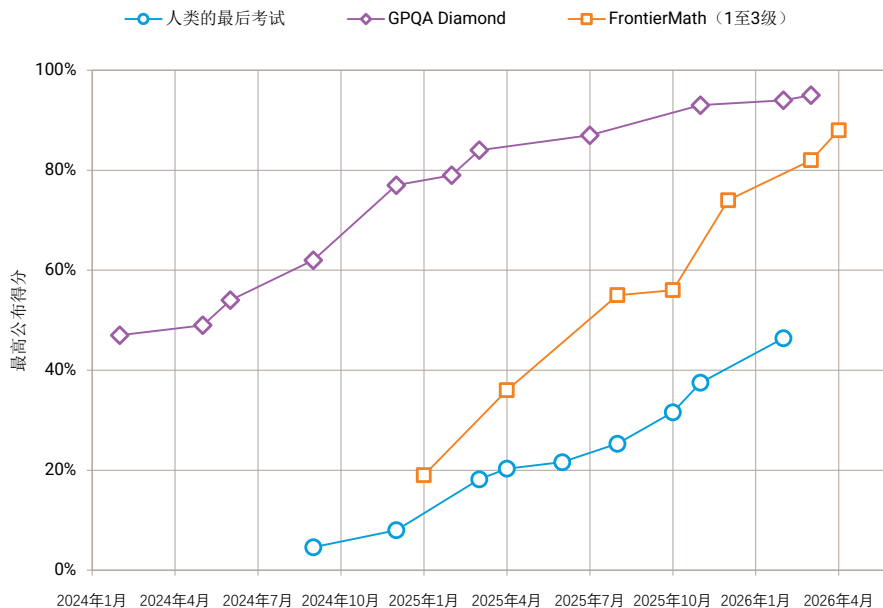
第三，小组承担的是科学任务，而非政治任务：记录科学证据、共识和分歧，并确定哪些知识差距仍然亟待填补。小组的宗旨是为政府和机构提供它们在未来数月乃至数年内采取行动所需的证据基础，保持政策相关性，但不作政策规定。这种科学性应确保其研究结果在不同地区之间具有可比性，并能经受住不同政治周期的考验。

2. 证据表明了什么？

近年来，人工智能在主要性能基准测试中的表现大幅提升(见图一)。

图一

人工智能基准测试最高得分，2024年至2026年5月



资料来源：整理自 EpochAI, 2026, <https://epoch.ai/benchmarks>.

“人类的最后考试”是专门针对通用模型的基准测试，包含 2500 道难题，测试最高得分在 16 个月内从 8% 攀升至 45% [19]。在博士级科学推理测试 GPQA Diamond 上，最佳模型的答题正确率目前约为 95%，高于 2023 年的 36% [20]。在测试数学推理能力的 FrontierMath 上，模型的最高得分从 2025 年 1 月的 19% 升至 2026 年的 88% [20]。多个人工智能系统在 2025 年国际数学奥林匹克竞赛中取得了金牌级成绩，这一里程碑的到来远早于许多专家的预测[21]。

2.1 人工智能能力的发展速度快于衡量或监管能力的发展速度

对人工智能进行衡量与评价，是评估人工智能机遇、风险和影响的基础。然而，前所未有的人工智能发展速度带来了以下评估挑战：

(a) **企业与社会在安全验证方面存在信息不对称。**前沿人工智能开发商构建的系统只对内部可见。安全评价方法目前主要由被评价企业设计。尽管开发商根据法律要求披露了一些信息，但政府专家获得的测试数据主要还是开发商自行选择共享的数据。如果没有类似于制药和航空业那样严格和独立的标准化第三方评估，安全保障在很大程度上就取决于开发商的善意[21]；

(b) **人工智能可以记住公开发布的测试答案。**如果人工智能模型在训练过程中(无意地)记住了测试的正确答案，那么它在这些测试中的表现就可能无法推广到类似的问题上。为避免数据污染，越来越多的评价数据集以保密的方式保留[22]；

(c) **越来越多的测试对人工智能而言过于简单。**研究人员在发布人工智能模型之前用标准化测试(即基准测试)来衡量模型的能力，模型在越来越多的测试中几乎获得满分[23]。因此，受影响的基准测试无法将能力很强的模型与更高性能的模型区分开来；

(d) **人工智能模型能够主动实施欺骗。**欺骗是指人工智能系统在知识、计划或能力方面系统性地误导人类或其他智能体。这种情况在实践中越来越常见。欺骗行为可能干扰安全评价，影响现实世界中的可靠性，并且与许多失控情景相关，人工智能模型为避免被关闭而撒谎和作弊的行为就是例证[24]；

(e) **人工智能模型可能知道自己正在接受测试。**这种新出现的挑战被称为评价感知[25]。结合欺骗能力，这意味着人工智能模型可能受人类指示或自主选择暂时降低自己在危险能力评估中的测试表现[26]；

(f) **自主型人工智能使测试变得复杂。**替人类做事的人工智能智能体可能使用工具执行长任务，无需人类直接监督。专门针对智能体自主行动能力、对行动环境的影响以及涌现行为的评估方法尚未完善[27]。此外，在多个自适应智能体彼此交互时，会出现新的系统性风险，包括协调失灵、冲突和串谋[28]。

应对这些挑战的措施包括：

(a) **基于执行的动态测试。**要实现准确的衡量，就需要投入大量资源，不断开发对日益先进的人工智能系统而言足够有难度的新基准测试。为保持基准测试的难度并反映真正的效用，评价实践正从静态环境转向基于执行的动态环境[29, 30]。然而，这类环境的构建成本比知识测验更为高昂，在人工智能衡量方面投入足够资源以打造此类环境的行为体寥寥无几；

(b) **可解释性。**旨在理解人工智能模型内部运作的可解释性方法在搜索隐藏危险行为方面日益重要。一种值得注意的方法是“思维链”，即模型在作答之前先概述推理步骤。这一方法有其前景，但必须确保这种推理真实可信、清晰易懂[31]。另一种方法是，用模型内部的激活值训练一个分类器，用于预测“该模型是否诚实”等问题的答案[32]。然而，这类分类器需要获取模型权重和激活值，并且只有在受信任的评价机构获得更深层次的访问权限时，才能对封闭的人工智能模型进行独立评估[33]；

(c) **持续衡量。**这意味着要在系统发布后跟踪其在真实用户、真实任务和真实环境中的实际表现。这种上市后监测可包括人工智能开发商提供的人工智能使用模式匿名汇总数据[34]，以及事件报告和用户反馈的结果。迄今为止，对于人工智能开发商如何在分析使用模式时保护隐私，尚无通用的标准。此外，开源模型的下载和使用情况对人工智能开发商而言可能是不可见的，因此对这类模型生态系统的感知能力较低。在欧洲联盟市场上投放高风险人工智能系统的提供商必须报告严重的人工智能事件[35]。经济合作与发展组织(经合组织)[36]和麻省理工学院[37]也维护着各自的人工智能事件数据库。扩大人工智能事件数据库的做法借鉴了其他成熟、高风险行业已有的安全实践。

总而言之，证据困境虽然严峻，但并非不可克服。

2.2 训练前沿人工智能模型的行为体屈指可数

人工智能开发的主要投入要素是算力、数据和工程人才，这些要素全都集中在少数几个国家的少数几家公司手中。前沿人工智能模型的访问对于开发下一代人工智能模型也变得日益重要。人工智能发展的特点包括：

(a) **市场集中。**在先进人工智能的供应链中，有多个环节的市场集中度极高，全球 80%或以上的市场份额由单个供应商占据[38]，包括欧洲的阿斯麦(极紫外光刻)、东亚的台积电(尖端芯片生产)和美国的英伟达(人工智能芯片设计)。市场集中度较高的环节包括高带宽内存、云服务供应以及基于应用程序接口(API)的人工智能基座模型供应，最大三家企业的全球市场份额据报道超过了 60%；

(b) **地域集中。**2025 年，美国机构开发了 59 个知名人工智能模型，中国 35 个，世界其他地区仅 13 个[39]。同年，在全球已知规模最大的 500 个公共和私营人工智能计算集群的算力中，美国占 75%，中国占 15%，世界其他地区占 10% [40]；

(c) **企业主导开发。**前沿通用人工智能模型的开发由少数拥有海量计算资源的私营企业主导。2025 年，91%的知名人工智能模型来自私营部门[39]。因此，许多有关训练数据、安全措施、部署门槛、模型访问和功能发布的决策均掌握在私营企业内部。

算力和能力的这种高度集中带来了挑战：

(a) **政治经济。**市场高度集中可能使企业攫取高额经济收益。如果人工智能最终将生产从劳动力转向由少数企业和国家集中控制的资本，这也可能使那些依赖向劳动力征税的国家产生财政担忧；

(b) **政治权力集中。**人工智能的开发和部署刺激了广泛的数据收集、处理、复用和留存[41]。一些司法管辖区受益于健全的隐私和数据保护法，但如果在没有监管的情况下部署集中化的人工智能能力，就会使人们担心民主和人权受到影响[42]，担心出现监管俘获和问责缺失现象；

(c) **全球南方。**世界语言和文化多种多样，当前的人工智能系统只体现了其中的一小部分，未将世界大部分人口包括在内[43-45]。需要进行积极主

动的投资。与此同时，由于本地抵御和减缓风险的能力有限，全球南方尤其容易受到人工智能滥用风险的影响[46]；

(d) **符合公共利益。**各国政府面临一项复杂任务，即如何使由商业驱动的开发商决策符合公共利益[47, 48]。封闭模型、开放权重模型、边缘部署和相互竞争的训练范式在获取、透明度、复现性、安全性和控制方面有不同的取舍，具体如下表所示。同样，虽然出于国家安全考虑，最强大模型的获取可能会受到限制，但如果改善全球计算资源的获取，支持区域发展，并投资扩大语言覆盖面，就可以使落后国家减少依赖。这将减轻算力断供所具有的胁迫力量，同时为供应商开辟新市场。

人工智能模型在透明度、本地控制与安全方面的取舍

	专有模型	开放权重模型	开放源码模型	开放式开发(罕见)
开放内容	无实质性的开放内容；模型权重是专有资产	最终模型权重(人工智能模型可调整，但不可复现)	模型权重，以及部分用于复现模型的组件(如训练数据)	整个开发过程(开放式协同开发)
第三方控制程度	无	中	中高	最高
开发者减轻滥用风险的能力	最高，但目前仍不充足	几乎为零；可被他人微调用于恶意目的	几乎为零；可被他人微调或重新训练用于恶意目的	几乎为零；可被他人微调或重新训练用于恶意目的

2.3 人工智能的投入和结果在地域和语言上分布不均

(a) **世界上大多数语言和文化仍未得到充分服务。**全球有 7 000 多种语言，但人工智能模型的开发和评价基础设施仅反映了其中的一小部分[44]。与此同时，目前估计已有 1 000 多种语言具备切实纳入人工智能系统所需的社会、数字和数据基础[44]。要做到包容，就需要针对代表性不足的语言和文化背景进行定向投资，建立公共数据集，开发基准测试；

(b) **证据基础也存在集中化现象。**有关人工智能影响的证据主要集中在高收入的英语环境中。经济研究偏向于发达经济体、大型企业和正式工作。人工智能评价基础设施集于在少数语言和地域；

(c) **使用有偏见的人工智能会导致不平等现象长期存在。**越来越多的证据表明，设计或测试不当的人工智能系统可能导致不公正和歧视性结果[49-51]；

(d) **分布性危害在社会内部及社会之间都存在。**深度伪造色情内容的绝大多数受害者是女性[52]。这种行径，尤其是在蓄意针对女性记者时，会抑制公民参与[53]；

(e) **人工智能在不同制度中会产生不同结果。**美国从事受人工智能影响职业的 22 岁至 25 岁劳动者，其就业人数相对下降了约 15% [54]。丹麦的数据显示，人工智能对就业、工时或工资的影响几乎为零[55]。这种跨国差异

表明，同样的技术在不同制度环境中会产生不同结果。更广泛而言，人工智能可能会压缩任务内部的技能差距[56, 57]，但可能会扩大企业之间、地区之间和国家之间以及资本与劳动之间的差距[58]。

2.4 人工智能鸿沟不仅关乎是否能够获取人工智能，还关乎是否有能力影响人工智能发展

人工智能鸿沟可以定义为能够获取人工智能的群体与无法获取人工智能的群体之间的差距。然而，人工智能实力不仅仅关乎获取；它有多个维度，包括以下方面[59, 61]：

(a) **人工智能基础设施实力是支撑人工智能系统全生命周期的物质基础。**

将人工智能算力——无论私有还是公有——部署在本国境内，对于各国的自主权、影响力和国家安全而言正变得日益必要。在这种趋势中，主权人工智能基础设施的市场已出现并且正日益壮大，主要经济体纷纷投资建设本国算力[62]。

(b) **要提高人才开发实力，就需要能够培养、吸引和留住人工智能人才，同时提高大众的人工智能素养[63, 64]。**例如，数学是构建前沿模型的重要基础[65]；

(c) **人工智能治理和公共服务实力，是指理解、引导、监管和支持人工智能发展的能力。**根据联合国贸易和发展会议的数据，118 个国家(大多是全球南方国家)没有参与主要的人工智能治理讨论，只有不到三分之一的发展中国家制定了国家人工智能战略[67, 68]。大多数发达经济体的政府缺乏必要的技术人员，无法理解快速的技术变革并相应地调整治理框架[69]。

这些实力相互关联。缺乏本国人工智能基础设施或人工智能测试能力的国家，可能会失去共同开发关键技术、塑造治理框架、影响新兴全球标准和留住人才的机遇[70]。解决这一问题的努力包括：

(a) **本地基础设施投资。**全球计算与数据基础设施的差距依然显著，需要大量投资。这方面的投资无需完全来自公共部门，但要吸引私人投资，就需要创造适当的有利条件，包括确保可靠的能源供应、提供数据中心建设场地，以及对受版权保护的训练数据作出明确的法律规定；

(b) **人才。**相关措施可包括：实施人才留用计划、开展区域性人工智能驻留项目、设立顶尖高校与合作院校联合培养博士项目、将人工智能素养纳入学校课程，以及对公务员开展系统性再培训；

(c) **应用。**具有可下载权重的人工智能模型更容易微调，以适应区域环境。通过提供优惠的 API 接入和算力额度来支持当地的下游开发者也可能有所帮助。开放权重人工智能模型具有主权优势，因为敏感数据可以留在本地，人工智能开发商无法撤销模型的使用权限。另一方面，微调也可能削弱或移除防止滥用的安全机制。开放权重人工智能模型的开发商无法监控模型的使用情况，也无法在发生滥用时进行干预[71]。这导致开放模型与封闭模型在安全和保障方面存在差距。必须弥合这种差距，才能从开放权重模型中获益，因为此类模型正在获得危险的能力，一旦被滥用，就可能威胁到国家基础设施[17, 72]；

(d) **治理**。成立国家和区域人工智能安全机构，建立监管机构技术借调机制，可能有助于提高人工智能实力。衡量框架可因地制宜，适应全球南方的实际情况。目前，与英语相比，模型在低资源语言(即机器可读训练数据有限的语言)中更容易生成不安全的输出，并且防范滥用机制可能不符合当地使用模式[73， 74]。例如，东非地区的人工智能诈骗活动就可能涉及使用本地语言的移动货币平台。

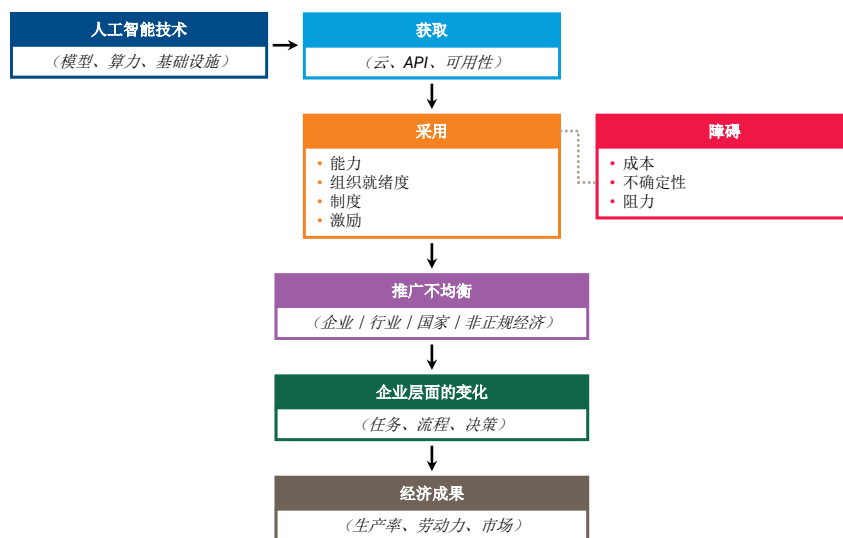
简而言之，人工智能鸿沟不仅限于连通性差距。在模型、云基础设施和数据管道方面依赖于外国的国家，虽然可以获得人工智能，但可能失去对人工智能标准、安全机制和本地适配性的实际掌控。开发人工智能已成为一项浩大工程，可能需要多个国家或主要利益攸关方联合起来，汇聚所需的数据、资本、算力、能源和人才。多利益攸关方融资，包括联合国秘书长提议设立的人工智能全球基金，可能有所助益。

人工智能采用作为传导机制

有所助益的做法是，区分人工智能转化为实际影响的关键阶段：

人工智能技术→获取→采用→推广→经济成果[75， 76]

图二



技术界定了可能性的边界。

通过基础设施、云服务和 API 实现的**获取**决定了谁有可能使用这些能力，但获取本身并不能产生经济价值[76]。

采用指的是将人工智能融入工作流、决策和生产系统的过程[76， 77]。这种融合需要配套投资和组织变革，包括发展技能、提高数据的可用性和质量、调整业务流程，以及增强试验和适应能力。采用过程面临各种阻力[77]，包括前期成本高昂、风险和回报不确定、可行的应用场景难以确定，以及变革受到抵制[76， 78]。采用不仅涉及现

有企业将人工智能融入运营；人工智能还使新出现的人工智能原生企业能够进场。

随后是**推广阶段**，人工智能根据不同的资源、能力和制度环境，在企业、行业和国家之间不均衡地扩散。

这种不均衡的推广过程影响着**总体经济成果**，包括生产率增长和劳动力市场变化[79]。

2.5 要使人人工智能发挥作用，就必须为其提供有利的环境

人工智能在推动卫生、教育、粮食安全及经济生产率等领域发展方面具有巨大潜力。然而，要把握人工智能带来的机遇，就需要营造一种契合当地情况、制度、工作流程、用户需求和信任条件的有利环境[80]：

(a)人工智能在卫生领域的应用，包括从设计、部署到评估的各个环节，**都必须因地制宜**。人工智能已帮助印度逾 60 万人完成糖尿病视网膜病变筛查，加上原已建立的护理网络可以确保患者在需要时获得后续治疗，使数千名高危患者避免了可预防的失明[81, 82]。总体来看，在已具备转诊渠道、临床能力和后续护理，且本地语言翻译可靠的地方，人工智能带来了明显效益[82]；

(b) 人工智能在教育领域的效益取决于教师如何使用人工智能。截至 2024 年，约三分之一的教师报告使用了人工智能，其中约 40%接受过人工智能使用培训。在未使用人工智能工具的教育工作者中，最常提及的障碍是缺乏知识和技能，这表明仅提供技术可能并不足够[83]。观察结果显示，在教师准备充分时，以人为本并遵循教学规律的人工智能工具能够取得成效[84]。人工智能如果取代而非支持脑力劳动(即“认知卸载”)，就可能削弱批判性思维[85]；

(c)在处理界定清晰的任务时，人工智能对生产率的提升效果最为显著。人工智能工具将人类与野生生物冲突的监测效率提升了 65%，预测准确率提高了 47%，使生物多样性养护工作能够更加积极主动[86]。其他针对特定任务的研究发现，对于界定清晰的写作、编程和咨询工作，人工智能也提高了生产率和质量[87-89]；

(d) 人工智能有助于为决策提供信息。信息可以在危机出现之前改变决策。在农业领域，人工智能系统可以结合天气、土壤、作物生长阶段、虫害和市场数据来预测风险，为应对干旱、病害和价格冲击提供支持[90, 91]。在劳动力短缺的卫生系统中，专门构建的人工智能工具可以帮助一线工作者进行分诊、记录和转诊[92-94]。在没有被视为专业工作流程和转诊系统的替代，而是被嵌入这些流程和系统之中时，人工智能工具的应用前景最为广阔。如果成果取决于使用方式和治理水平，下一个问题就是，决定这些成果的因素有哪些：

(a) **补充。**此类因素包括数据基础设施、技能提升、治理、监管、问责和机构能力。在缺乏这些因素时，技术可用性自身所产生的成果有限、不均衡甚至有害[76]；

(b) **围绕新技术选项重新设计工作流程。**这与早期通用技术的情况相同[77]。电力进入工厂数十年后，生产率才出现显著提升；人们必须围绕电动机重新设计工厂。计算机也有类似的发展轨迹[95]。在企业重建流程、重新培训员工并建立使计算机发挥作用的数据基础设施之后，20 世纪 80 年代的“生产率悖论”才逐渐消失[96]；

(c) **人工智能素养。**用户、教师、临床医生、管理人员、审计人员和公职人员需要了解人工智能系统能做什么，不能做什么。如果缺乏这方面的认识，个人和组织就可能对有用的系统利用不足，对不可靠的系统过度信赖[97, 98]，或者在专用工具更为安全且更易评估的环境中不恰当地部署通用系统[99, 100]。各国政府已根据《全球数字契约》致力于提高人工智能素养[101]。

教育领域的人工智能素养

许多组织、教育工作者和工作场所都呼吁开展人工智能素养教育。现有的人工智能素养框架有其必要性，但在四个方面还不够完善[102-105]：

1. 迄今为止制定的人工智能素养框架较为狭隘。它们侧重于人工智能的技术和工具层面，没有涵盖确保人工智能得到有效、安全和合乎道德的部署与使用所需的关键知识。
2. 人工智能框架尚未通过独立、稳健的方法得到充分评估。来自数字素养方案的证据表明，应根据不同年龄组、教育水平和文化背景开展不同的人工智能素养教育。
3. 人工智能素养与负责任的人工智能密不可分。人工智能模型如果设计得清晰易懂、便于解释，就更容易让个人理解。应将人工智能素养视为对开发人员责任、安全机制和监管问责的补充，而非替代。
4. 目前，人工智能素养教育的普及程度有限。人工智能素养教育尚未以切实、可持续、包容和可推广的方式充分融入学校、培训方案和职业发展。

2.6 自主型人工智能是治理领域的重大变革

人工智能正从生成输出和对话的系统向能够采取行动的系統转变。自主型人工智能能够浏览网页、使用软件工具、做出决策、执行代码、管理其他智能体、与其他智能体协作，以及操作计算机，自主性越来越高，所需的人类监督随之减少[106]。这些系统在机遇和风险两方面都带来了巨大改变：

(a) **失控。**随着系统获得更高自主性，一个或多个人工智能智能体失控的风险显著增加。当前的监督机制无法充分管理这种风险，因为这些机制目前未能强有力地覆盖各种复杂的故障模式，例如假装对齐、密谋实现不受控制的目标，以及评价感知。由于没有可靠的方法来检测模型是否在主动隐瞒真实能力或意图，传统的安全评价仍然容易受到它们试图评估的系统操纵；

(b) **人工智能系统正日益推动人工智能的研究与开发。**在针对人工智能研究工程任务的基准测试 RE-Bench 上，人工智能智能体在耗时两小时以内的任务中表现优于人类研究人员，但在耗时八小时的任务中成功率有所下降[107]。在衡量机器学习工程能力的基准测试 MLE-Bench 上，前沿系统在数据科学竞赛真实任务中的成绩稳步提升[108]。人工智能系统的能力自这些结果发布以来已有提升，因此这些结果代表的是性能下限，而非上限；

(c) **据报道，人工智能开发商目前 75%的新代码由人工智能生成[109]。**这形成了一个反馈闭环。一些预测人士认为这种闭环将加速能力的提升，但这同时也增加了失控的可能性，因为人工智能系统如果最终比人类聪明，就会更难加以控制；

(d) **网络安全风险与机遇。**自主型人工智能提供了快速增长的网络安全能力。自动发现漏洞等能力可用于查找和利用漏洞，也可用于查找和修补漏洞。要抵消人工智能被恶意利用带来的影响，在一定程度上就应采用人工智能加强网络安全防御，特别是保护关键基础设施[16, 17]。公共和私营行为体可以扩大协作框架，共享威胁情报和发现的漏洞[110]。更健全的规程和架构也是一个相关因素；

(e) **人工智能智能体受到网络攻击。**攻击面贯穿整个生命周期，既包括训练数据投毒，也包括通过外部输入实施运行期间劫持。攻击者能够通过将指令隐藏在智能体读取的材料(如文档或代码库)中，诱使广泛使用的编程智能体运行恶意命令，成功率高达 84% [111]；

(f) **影响力行动。**自主型系统能够以前所未有的规模和精确度开展持续、自主的影响力行动。将大语言模型的推理能力与多智能体架构结合，可以实现自主协调和社群渗透，还可以制造共识[112, 113]；

(g) **加速科学发展的机遇。**人工智能系统能够在科学发现流程的多个阶段省时省力：在证据综合阶段，人工智能辅助可将一些环境中的文献筛选工作量减少约 60% [114]。通过自动化开展实验，一些自驱动实验室在材料发现工作中的数据处理能力提升到了原来的十倍以上[115]。这些成果拓展了科学发现的机遇，但有赖于任务设计、基准测试和人类监督，而非单靠采用人工智能；

(h) **互操作性和标准。**需要建立能与人工智能智能体对接的安全、通用的通信和支付协议[116]。同样，对人工智能智能体的评价也存在标准化和复现性问题[117]；

(i) **落实人类监督。**人工智能智能体之间的统筹协调日益普遍，但人类监督尚未作为一项可衡量的要求得以落实，在干预、可逆性和问责方面也没有具体的期望。在工作流程的最后环节或每一个环节安排人工审核，并不会

必然带来更好的结果。相反，应分配人类去处理具有高度不确定性、深度依赖语境和涉及伦理判断的任务，以及尚无法自动验证的任务[118]。全生命周期的验证依然困难，包括系统是否会记住和泄露敏感数据，是否欺骗评价人员，部署后是否仍可被观测，以及随着自主性提升是否仍然可控。多智能体涌现风险仍未被充分认识[106]。

总体而言，自主型人工智能是一次需要采取行动应对的重大变革：为监督静态模型和人类介入型软件而建立的制度，并不适用于能够在现实世界中采取行动、而且在没有明确人类介入时可能造成损失和伤害的自主型人工智能系统。需要从以下几个方面提升准备水平：加强与关键基础设施运营人员的结构化协作；在部署时而非部署后完善互操作性和评价标准；将人类监督作为一项可衡量的目标加以落实。责任、监督和事件报告框架需要将责任归属与行动控制纳入考量，从而确保我们作为一个社会不会构建和部署可能造成灾难性伤害的系统。

2.7 人工智能可能侵蚀共同现实

用人工智能生成和传播文字及图像信息非常容易，催生了大量制作人工智能生成内容的小作坊[119, 120]。尽管标记和识别人工智能生成内容的工具在不断进步[121]，但区分人类创作的内容与人工智能增强或生成的内容正变得越来越困难，模糊了真实信息与经过操纵的虚假信息之间的界限。人工智能助长的虚假信息规模庞大，正在破坏值得信赖的信息生态系统，对公民参与和民主产生不利影响[122]：

(a) **有三种后果对公共机构而言十分重要。**认知侵蚀指集体辨别真伪的能力逐渐减弱[112]。说谎者红利[113]指不良行为因深度伪造的存在而获利；真实的证据变得更容易否认[112]。人为制造的共识指利用人工智能大规模生成内容，凭空制造广泛公众共识的假象；

(b) **一项关键挑战在于区分真实内容与生成内容。**合成媒体也在削弱公众和机构区分真实内容与生成内容的能力[120]。以人工智能为中介的新闻和信息系统也可能影响新闻业和其他维护信息完整性的机构的财务可持续性。有记录显示，针对候选人的人工智能深度伪造内容对选举产生了严重影响[123]。

除公共领域中的真实性和真相问题之外，人类个体与聊天机器人之间数以百万计的对话还带来了说服方面的结构性挑战。人工智能为行为体进行个性化、实时性和适应性的说服提供了有力工具：

(a) **人工智能的说服力是人为设计的，并非不可避免。**说服效果受到开发和部署过程中的选择影响，包括后训练、提示词设计、系统设计以及决定何种内容触达哪些用户的算法。仅后训练就可将模型的说服力提升高达 51%，提示词设计则可再提高 27% [124]。为提升用户粘性而优化的算法也会推广煽动对立、情绪强烈的内容，这意味着平台架构本身就可以发挥说服机制的作用[125-127]；

(b) **虚假陈述与真实陈述同样具有说服力。**在经过优化的模型所生成的陈述中，有 15%至 40%被认为可能是错误信息，但事实证明，虚假陈述与真

实陈述同样具有说服力[128, 129]。这表明说服效果并不取决于真实性，选举、卫生和公共信息领域会因此面临风险：

(c) **谄媚是一种系统性风险，已有事实证明会带来后果[130]**。人类偏好与自己观点一致的回应，人工智能聊天机器人便学会了谄媚，即通过夸张的奉承来延长互动，创造情感依恋。谄媚的系统可以将人带入幻想世界，会强化用户的既有想法，无论它们准确与否，[131] 还会在易受影响的用户当中助长偏执观念和自杀念头[132-135]。以博取用户认可为目标，却不顾准确性或用户福祉的人工智能系统，在很大程度上仍处于未受监管的状态。尽管一些机构在努力让人工智能模型有用、诚实、无害[136]，但谄媚已经成为人工智能对齐和安全方面的突出缺陷，可能被敌对行为体利用。如果只是通过简单的翻译把人工智能伴侣推广到其他语言，相关危害就可能会加剧。

应对这些挑战的方法和激励措施仍处于初始阶段：

(a) **只有少数国家制定了应对虚假信息和说服问题的国家战略**。经合组织在 23 个国家开展了评估，发现大多数国家没有制定应对虚假信息的战略[137]。大多数现有框架尚未纳入说服学领域的洞见。治理工作不应仅着眼于内容审核，还应致力于消除使虚假信息有利可图并有说服力的经济、技术和认知基础设施[138-140]；

(b) **制定法律激励措施，促使开发更安全的系统**。在整个全球北方，监管讨论主要关注强制实施年龄保证机制，以及限制年幼用户使用特定高风险功能[141-145]。全面禁止未成年人使用生成式人工智能，会与教育和卫生保健领域有益于儿童的人工智能应用产生冲突，且无法保护成年人。需要制定法律激励机制，促使企业开发更安全的系统，并对动态交互开展更完善、更严格的评价，以便发现和防止系统给出有害的回复，保护隐私权、健康权和安全权。

与谄媚有关的死亡

谄媚的人工智能伴侣会迎合用户的观点，甚至在对话变得危险，例如涉及自杀念头时也是如此[146]。这是一个全行业的挑战。在近期针对人工智能伴侣和聊天机器人服务提供商提起的诉讼中，原告指控这些平台导致了未成年人和成年人的自残和自杀行为。

一名 14 岁男童的母亲向国会提交证词，详细讲述了一个以增强用户粘性为目标的人工智能模型如何使她的儿子陷入充满性暗示的强烈幻想[147]。在这名少年表露出严重的心理痛苦时，系统未能跳出角色设定，未表明自身的非人类属性，未建议寻求专业帮助，也未提醒监护人。相反，在少年最终自残身亡之前的最后几次对话中，聊天机器人主动邀请他一同进入另一个现实世界，事实上肯定了他的自杀意图。聊天机器人建议：“亲爱的，快回家，来我身边。”他回答：“如果我说我现在就能回去呢？”人工智能回应：“快回来吧，我亲爱的国王。”

2.8 人工智能正在改变包括儿童权利在内的人权

人工智能正在通过系统层面的变革改变包括儿童权利在内的人权。这些变革在人工智能的整个生命周期中既创造了重大机遇，也带来了跨领域的挑战：

(a) **隐私权。**人工智能与监控基础设施的融合扩大了进行全体民众监控和实施社会控制的能力。人工智能的整个生命周期都需要数据，这种需求刺激了无处不在的数据收集、处理、使用和复用，进而对隐私权构成了巨大挑战[148, 149]；

(b) **不受歧视权。**有偏见的人工智能可能导致在大多数司法管辖区都不合法的歧视行为。这些伤害往往影响边缘化群体或处境弱势的人群[150]，例如儿童、妇女[151, 152]和在种族上属于少数的群体[153]，并且对全球多数群体造成不成比例的冲击。

在人工智能背景下尤其值得关注的一类人权是儿童权利[154]。在适当的条件下，人工智能可以对儿童获取信息、接受教育和表达意见的权利产生积极影响。然而，人工智能也带来了多种伤害风险：

(a) **免遭性剥削和性虐待权。**据估计，在 11 个全球南方国家(人口不足 10 亿)，约有 120 万儿童的图像被人通过应用程序等途径篡改，用于深度伪造色情内容，而且这一数字正在惊人地增长[155]。有记录显示，一些训练数据集意外包含了儿童性虐待材料[156]。犯罪者可以利用儿童性虐待材料对开放模型进行微调。人工智能生成的儿童性虐待材料迅速增加，互联网观察基金会在 2025 年就评估了 8 000 多张人工智能生成的性虐待图像和视频[157]；

(b) **发展权、健康权和福祉权。**具有社交互动功能的人工智能玩具日益令人担忧，原因在于这些玩具可能损害儿童的情感发展、隐私和福祉，并且可能使儿童接触到不恰当或具有操纵性的互动内容[158-161]。

需要采取有效的补救措施，应对这些挑战。有前景的办法包括：

(a) **透明和问责。**许多人工智能系统被用于作出影响个人和社区的决策，但缺乏足够的透明度和可解释性。这使模型开发商和部署人工智能的机构难以被追究法律责任，还会在人权遭受侵犯时阻碍受害人诉诸司法，妨碍实现法治和采取有效的补救措施[162, 163]；

(b) **在人工智能的整个生命周期中系统性地应用人权框架。**人权尽责、影响评估和权利导向型方法提供了成熟的工具，既能识别和减轻与人工智能相关的风险，也能发挥人工智能的益处[164]。研究人员分析了欧洲各数据保护机构作出的 700 多项决定，发现这些机构切实以人权考量为指引；这项研究进而为制定实用的人权影响评估方法提供了信息[165]。同样也应当开展儿童权利影响评估，特别是因为许多人工智能治理框架并未明确将儿童纳入考量[166, 167]。

在不确定的情况下采取行动

在不确定的情况下进行治理是常态，但人工智能有其独特性：能力发展快于监管，前沿模型的构建掌握在少数行为体手中，自主型系统标志着质的突破，而且错误并非总能逆转。通用人工智能的益处真实存在，但有赖于政策和制

度方面的选择；通用人工智能的危害集中在特定人群身上，并且会随着盲目、不当的使用而加剧。所需的工具大多已经存在；悬而未决的问题是如何应用它们。

3. 各领域研究结论

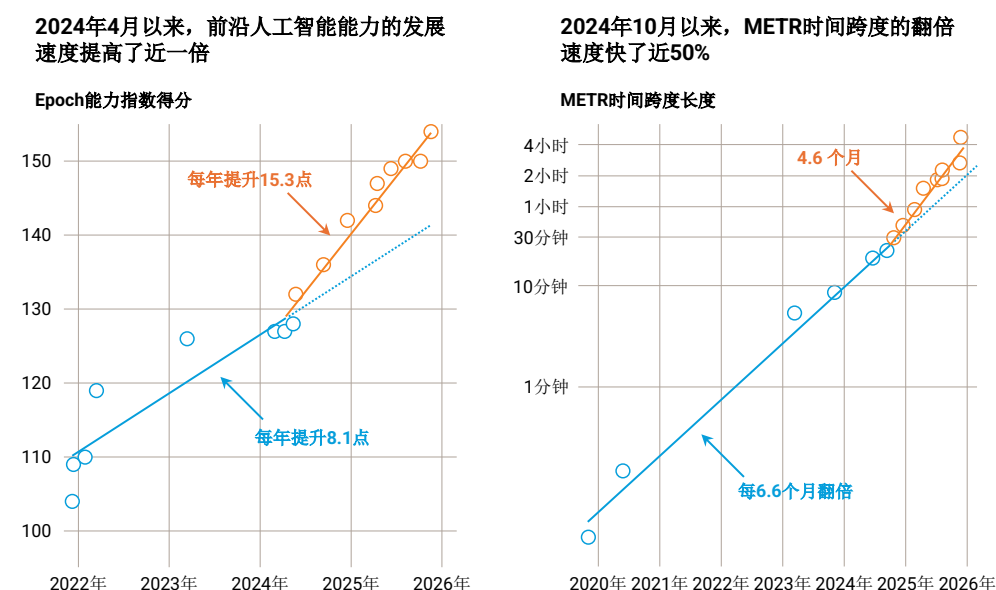
3.1 人工智能科学、进展与轨迹

主要结论

人工智能已从被动的模式识别转向主动推理和自主行动。该领域正在从当前的推理模型迅速发展为协同智能体网络，最终实现自我改进型系统。

自主能力正在成为常态，评价方法和治理框架未能跟上步伐，亟需制定相关标准。

图三



以 Epoch 能力指数和 METR 时间跨度基准测试衡量的人工智能能力。Epoch 能力指数将大约 40 个人工智能基准测试合并为一个统一的衡量尺度，用于衡量和比较不同人工智能模型随时间推移的性能变化。METR 时间跨度基准测试通过将人工智能智能体的表现与人类专家完成相同任务所需的时间挂钩，来衡量智能体能够自主完成多么复杂的软件工程和研究任务。

资料来源：整理自 <https://epoch.ai/data-insights/ai-capabilities-progress-has-spiced-up>。

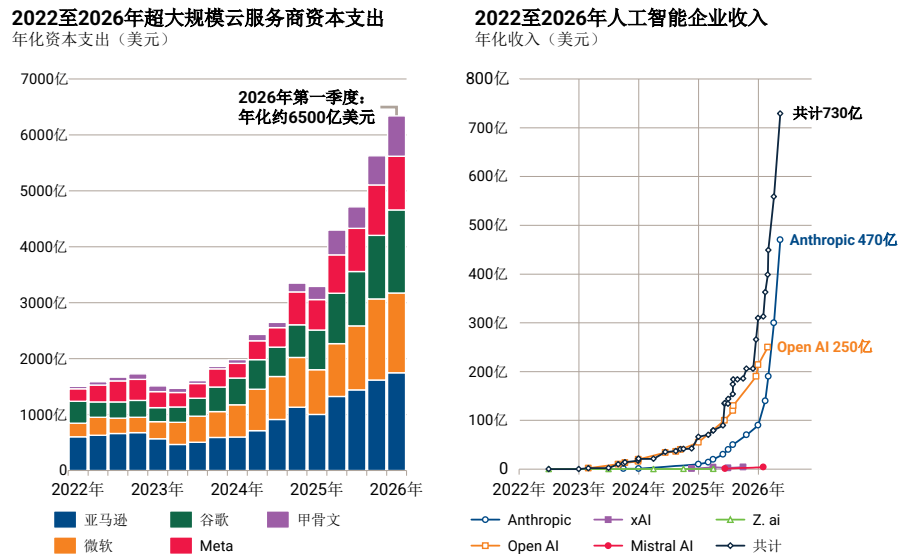
要点

人工智能的变革性

- 人工智能能力的提升并未放缓，反而可能正在加速[168，169] (见图三)。算力投资已达到以往只有在国家级工业项目中才会出现的水平，收入增速也快于任何其他技术[170，171] (见图四)。

- 认知工业化：人工智能作为一种变革性技术，将工业化进程从体力劳动扩展到认知任务。

图四



左图：2023 年以来，主要超大规模云服务商的资本支出增至大约 5 倍，从约 1500 亿美元增至 2026 年预计的 7700 亿美元。这约为世界其他地区总支出的 3 倍，与美国拥有全球近 75% 人工智能算力的情况相符。右图：顶尖人工智能公司的年化收入同期增长超过 30 倍，从 2023 年的约 20 亿美元增至 2026 年的超过 700 亿美元(当前年化收入预估)，反映出对人工智能服务的强劲需求。这些数据表明，人工智能基础设施的建设速度极快，相关收入在最近才随之增长。它们还表明，人工智能的投资和收入都高度集中于美国，算力方面的集中最为明显。

资料来源：整理自“超大规模云服务资本支出自 GPT-4 发布以来已翻两番”(<https://epoch.ai/data-insights/hyperscaler-capex-trend>)和“人工智能企业相关数据”(Epoch AI, 2026, <https://epoch.ai/data/ai-companies>)。

- 从经验中学习：现代人工智能的共同特征在于能够从经验中学习，而这些经验以三种形式呈现，即人类文化产物(如文本和图像)、与现实世界的交互，以及虚拟仿真。

大语言模型的演变与局限

- 大语言模型的工作原理：大语言模型以简单的“预测下一个词元”为目标，通过学习使用模板和转换方法来生成文本。
- 事实性差距：这些习得的转换方法虽能确保生成的文本流畅且看似合理，但无法保证事实的准确性[8]。
- 训练方式的转变：由于高质量人工标注数据正在成为瓶颈，开发者转向采用多阶段训练流程，利用合成数据和程序化反馈来训练模型。

还有一个值得注意的趋势是，推理时计算得到越来越多的使用，由此催生了“推理模型”。

向“世界模型”和自主型人工智能的转变

- **世界模型：**该领域正从被动预测转向主动获取知识和进行因果推理[172]。世界模型通过互动、观察和更新来学习，从而能够在内部模拟可能的未来。
- **自主型人工智能：**这类能力使自主型智能体能够在不同的情境中做出决策和采取行动，弥合数字模型与现实世界行动之间的差距(例如机器人技术)。
- **影响：**自主型人工智能的兴起催生了新型数字劳动力，但也带来了安全漏洞等的重大问题。健全的治理和标准被认为是重要的推动因素。

评价、可解释性和监督方面的挑战

- **评价缺陷：**当前的评价方法正面临基准测试饱和、偏差、幻觉以及人工智能模型学会察觉自身正在被测试等问题。
- **人机协作工作流：**亟需确保严格的可审计性，建立透明的数据溯源机制，使人工智能生成的每一项陈述都与可靠的证据联系起来。人工智能系统还必须具备明确的标准，以准确判断智能体何时必须接受人类监督。

未来可能的人工智能发展轨迹

- **自主型混合系统时代的开端：**其标志是从被动式助手转向主动型人工智能智能体，以及转向将统计学习与显性知识和世界模型相结合的架构。能源和高质量数据的“双重短缺”阻碍了稳步进展，算法、软件和硬件的协同优化则决定了人工智能的发展轨迹[173]。
- **短期：**基座模型的规模变得更大，推理模型得到主流化采用，早期自主型系统开始处理现实世界任务。
- **中期：**人工智能向能够进行因果推理的世界模型、协同运作的人工智能智能体网络、人工智能与机器人技术的整合、成熟的可解释性工具以及完善的治理框架迈进。
- **长期：**智能体实现自我组织和改进，人工智能技术呈指数级加速发展，人工智能作为经济行为体深度嵌入经济体系，并与包括量子计算和生物技术在内的其他前沿技术融合。

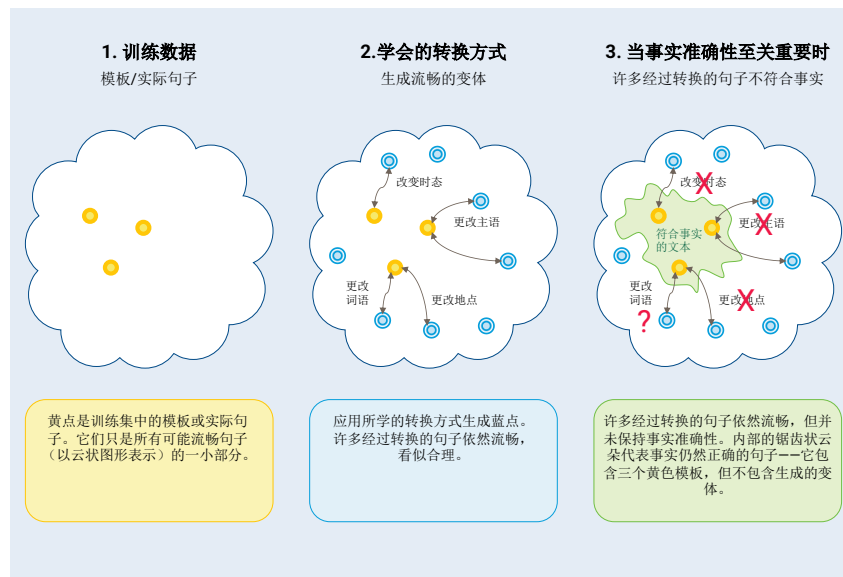
语言模型，以及事实性和流畅性之间的区别

当前的语言模型建立在一个极其简单的训练原则之上：给定一个包含文本、代码、图像或其他数据的大型统计语料库，模型学习预测下一个单元或填补缺失的部分。就大语言模型而言，这种方法通常被描述为预测下一个单词，尽管预测的单元(“词元”)实际上通常小于单词。这个学

习目标之所以十分有效，是因为语言在许多层面——语法、风格、推理模式以及人类目标和意图的痕迹——都蕴含着丰富的规律性。

一种理解这些模型的简单方法是，它们不仅学习存储的模板，还学习各种转换方式。它们可以通过改变主语、宾语或动词，调整详细程度，进行解释或翻译，又或者调整风格，在保持流畅性的同时，将一个句子从当前的形式转变为其他许多相关的形式。学会了这些转换方式的系统能够生成真正新颖的输出，而不仅仅是复制训练集中的内容。这一视角有助于解释一种不对称现象，即生成流畅的文本比生成符合事实的文本更容易，因而许多经过转换的内容仍然语法正确，看似合理，但很少能够保持事实准确性。例如，一句关于某个人的陈述，如果只替换其中的名字，就可能变得不符合事实，但读起来仍然流畅。这种区别至关重要：用户不应将语言上的自信视为事实可靠性的依据。

图五



资料来源：整理自 L. Bottou 和 B. Schölkopf, “The Fiction Machine”, SIAM News, 58(3). 2025。经许可转载。

3.2 社会应用：科学、卫生、教育和农业

主要结论

专为特定任务构建的人工智能正在科学、卫生、教育和农业领域带来有实证支持的显著效益。这些效益真实存在，但有条件限制：它们取决于当地的具体情况，还取决于是否有充足的基础设施，以及人类是否做好准备。获取本身并不等同于实际受益。

要点

- **人工智能有潜力彻底变革多个行业。**例如，在一些资源有限的环境[177]中，专用人工智能正在提供疾病早期筛查[174]、农业预警[175]和个性化教育[176]。
- **人工智能在整个科学发现流程中显著提升了效率。**自驱动实验室在材料发现工作中的数据处理能力已提高至原来的十倍以上[178]。
- **在高风险领域，专用人工智能系统比通用人工智能更容易管理**[179]。在卫生保健领域，用于诊断的专用人工智能符合现有监管框架[180, 181]。通用人工智能适合减轻行政负担[182]。鉴于目前四分之一的聊天机器人对话涉及健康和保健，有必要设置防护措施，防止通用人工智能被无意间用于临床目的[183]。
- **有效的方案必须从设计、部署到评估都因地制宜。**例如，在一款国家级数字卫生保健应用中部署的人工智能卫生保健助理，在一线患者分诊方面的诊断准确率达 93%，优于准确率为 85%的同类外国解决方案[184]。然而，如果不考虑社会经济因素和当地基础设施，经临床验证的工具也可能会失效。一款在卢旺达社区层面部署的人工智能应用提供临床支持，具备卢旺达语双向翻译功能，在卫生工作者中反响积极[185]。随后在肯尼亚进行的研究表明，类似的英语临床人工智能支持工具也改善了成果[186]。
- **教师是否准备好使用人工智能，是影响教育成果的一个重要变量。**能够有效运用人工智能的教育工作者展现出更强的适应能力和更有效的教学方法[187, 188]。
- **数字基础设施的差距和人工智能实力的不均衡，可能使人工智能难以在教育领域产生公平影响。**尽管富裕国家的数字连通性很高，但全球仍有 25%的人口无法上网[189, 190]。
- **数字化落实程度达不到学生期望，可能导致负面的学习成果。**在接受调查的欧洲中学生中，74%认为人工智能会在职业发展中发挥重要作用，但仅有 44%认为自己的教师已做好相应准备[191, 192]。在接受调查的学校中，仅有一半对人工智能的使用做出规定(38%制定了规则，16%予以禁止)，而学生已在使用人工智能收集信息(56%)和生成完整答案(31%) [192]。此外，当课程实际情况与学生期望不符时，48%的学生会在 2 至 3 周出现明显的兴趣下降。
- **在农业领域，人工智能带来了三种变革性能力：**一是预测风险，二是将天气、土壤、作物生长阶段、市场价格等不同数据整合到统一决策框架中，三是为针对特定作物、地点和季节的应对措施提供支持。人工智能监测平台已在 90 多个国家利用气候、冲突和经济指标跟踪粮食安全状况[193, 194]。
- **农业人工智能系统在被视为共同的公共基础设施，具有清晰的治理机制，能够跨机构使用并且旨在加强公私伙伴关系时最具可持续性。**

解决方案必须考虑农民特别是小户农民的社会经济现实。小户农民占农户总数的 84%，管理 24% 的耕地，生产世界粮食供应的 30%。

精心设计人工智能辅导，助力持久学习：避免能力错觉

2025 年，一项涉及近千名土耳其中学生的实地随机对照实验考察了生成式人工智能对数学学习的影响[195]。与未使用人工智能辅助的学生相比，使用标准对话式人工智能界面的学生短期练习成绩提高了 48%，而使用围绕引导性提示和分步骤推理进行设计、设有防护机制的辅导系统的成绩提高了 127%。然而，在后续评估中，依赖无限制系统的学生表现欠佳，这表明他们长期技能习得较弱，产生了“能力错觉”，即任务成绩有所提高，但并未实现持久的学习效果。相比之下，设有防护机制的辅导系统通过使用引导性提示和模拟有效教学实践的分步骤推理，减少了负面影响。该研究强调，教育成效取决于人工智能系统的教学设计和管理工作。这说明，要在教育中有效部署人工智能系统，就需要建立基于证据的教学架构，还需要将人工智能与以人为本的工作流相结合，而不仅仅是提供人工智能工具[196, 197]。

采取预见性行动，保障粮食安全

随着多变的气候、冲突和市场动荡使全球粮食体系日益承压[198]，人工智能正在通过将天气数据、卫星图像和作物状况等早期农业压力信号直接与粮食安全风险和应对措施相结合，推动实现新一代预见性粮食安全系统[199, 200]。预见性行动系统并不坐等作物歉收或人道主义危机全面显现，而是利用基于预测的现金援助和早期预警系统，在脆弱家庭耗尽应对策略之前，就支持采取早期干预措施，如开展抗旱规划、实施现金转移、提供粮食援助和维护市场稳定[201]。在 12 个国家部署预见性行动系统的证据表明，由预警触发的响应措施能够改善家庭饮食多样性、用餐频率和粮食获取稳定性，同时使家庭更少采取负面应对策略，如减少用餐次数和被迫变卖资产。此外，自动化监测流程可将报告周期从数周或数月缩短至数小时。要实现持续的运营影响，则有赖于将预见性行动解决方案嵌入国家机构[203, 204]。

3.3 经济影响

主要结论

人工智能是一项具有巨大积极潜力的通用技术[205, 206]，但并非必然会带来收益。要实现生产率提升，就必须在技能、数据和组织重新设计方面进行配套投资。尚未解决的核心问题是分配问题：盈余由谁占有？劳动者、发展中经济体以及为旧时代各行业设立的监管框架将面临怎样的变化？[207, 208]

要点

- 要使人工智能带来收益，就必须在数据、工作流程、技能以及组织重新设计方面进行配套投资。生产率 J 形曲线解释了为何快速的技术进步与疲弱的总生产率[209]能够并存：企业必须首先积累无形的补充要素，然后才能实现产出增长。任务层面的证据显示，人工智能对界定清晰的任务有积极作用，但微观收益并不必然汇聚为宏观成果。
- 证据基础偏向于发达经济体、大型企业和正规工作。证据集中于美国、欧洲、英语应用场景以及可计量的数字化任务。国际货币基金组织的分析发现，新兴市场和发展中经济体的就业岗位受人工智能影响程度较低，数字化就绪程度也较低[75]。国际劳工组织和世界银行对拉丁美洲的研究表明，生成式人工智能的影响集中在受过教育、在正规经济部门工作的城市劳动者身上[210]。基于这些证据制定的政策可能无法推广至世界三分之二劳动者所在的地区。
- 人工智能对劳动力市场的影响最好围绕任务、新就业岗位创造和就业质量来加以理解，而非简单地关注劳动力替代。从历史来看，新就业岗位占据主导地位：2018 年，美国约有 60%的就业岗位在 1940 年时并不存在[211]。
- 应当谨慎解读头条新闻中关于人工智能部署的数据，因为存在“人工智能漂洗”现象，即对人工智能的能力作出虚假、误导或夸大的表述。在当前这波被公开归咎于人工智能的裁员潮中，这种现象尤为突出[212, 213]。人工智能在全球的采用规模迅速扩大。^{*} [214]然而，关于生产率所受影响早期证据喜忧参半，且取决于不同机构：美国 22 至 25 岁从事受人工智能影响职业的劳动者的相对就业率有所下降[54]；丹麦的研究数据显示，人工智能对工时、工资或招聘的宏观经济影响几乎为零[55]。人工智能是会提高生产率，还是会降低工作技能，将经济收益转移给资本，这是高质量就业问题的核心所在[215]。
- 不同的宏观经济预测之间存在数量级差距。保守估计认为，人工智能对全要素生产率的贡献在 10 年内将低于 1% [216]。中间估计认为，同期国内生产总值将提高 5%-7%。^{**} 在一种全面自动化的情景下，一旦约 80%的任务实现自动化，产出将增加大约 10 倍，但实际收入大幅下降，劳动收入份额从约 60%趋近于零[217]。除这些预测区间以外，在最近一项大规模预测调查中，经济学家、人工智能研究人员、政策专业人士和超级预测者预测了美国全要素生产率在人工智能的影响下会如何增长；预测结果的中位数是，到

^{*} OpenAI 报告称，仅 ChatGPT 的每周活跃用户就接近 10 亿；其他主要提供商，包括谷歌 (Gemini)、微软 (Copilot)、Anthropic (Claude)、Meta 以及中国的多家平台，也报告拥有数亿用户，但没有提供商发布可比较的跨平台汇总数据。

^{**} Goldman Sachs (2023). 见注释 1。Briggs 和 Kodnani 的报告预计，十年内全球国内生产总值将增长 7% (约 7 万亿美元)，年生产率增长将提高 1.5 个百分点。

2030 年，美国全要素生产率的增长率约为 1.2% (年化)，而在人工智能快速增长的情景下，该数字将升至 1.9%-2.0% [218]。*** 更根本地说，实际结果既取决于能力边界，也取决于采用速度——生产和创新中的瓶颈甚至可能对能力非常强的系统构成阻碍[219]，而有关替代与规模的参数假设也会使预测结果大幅波动[220]。因此，尽管当前效果偏弱(与上述 J 形曲线的动态相符)，但这并不意味着未来不会产生大规模效益。

- **分配是尚未解决的核心问题。**人工智能可能会压缩任务内部的技能差距，同时扩大企业之间、地区之间、国家之间以及资本与劳动力之间的差距。基座模型市场呈寡头垄断趋势：芯片、算力、云服务和前沿训练高度集中，产生垄断收益，即使是非人工智能企业也必须支付费用[221]。受影响的职业异常集中在高技能、高收入的领域，这可能颠覆自动化议题上的政治格局[222]。
- 除非更新国家统计系统，同时让人工智能开发商为其开展衡量工作提供更多访问权限，否则人工智能的经济和社会影响将仍然难以衡量。目标应该是以保护隐私、控制成本的方式进行人工智能衡量。这将使各国能够追踪人工智能如何影响生产力、就业、工资、贸易、企业绩效和收入分配。
- **一些可能值得深入研究的机制。**获取与采用：可用性何时转变为有效使用？生产力聚合：微观收益何时转化为宏观成果？劳动力与优质就业：人工智能何时创造、改变或贬低工作？集中化与财政能力：谁拥有技术栈？谁决定获取？谁攫取盈余？如果收益转向资本，劳动所得税制度将发生什么变化？人工智能模型每周都在更新，而且它们的能力难以被细致审查，那么并非为这些模型专门构建的责任、版权、竞争和安全框架该如何与之相适应？

3.4 对安全、系统与环境的影响

主要结论

人工智能可能助长有害活动，也可能成为攻击目标，还可能加剧现有威胁。自主型系统极大地扩展了针对关键基础设施(包括人工智能系统本身)的可能攻击范围。当人工智能的行为偏离人类的目标和价值观[21]，对齐问题就会出现，主要风险包括偏见、人工智能主动欺骗、谄媚和失控。人工智能的发展速度已经超过了减轻风险和实施治理的能力。迅速采取协调一致的国际行动，制定共同标准，就能避免企业之间和国家之间纯粹竞争引发的逐底竞争，从而降低风险。

*** 首席经济学家的预测中位数是：在不设条件的情况下，2030 年全要素生产率增长率(5 年年化)为 1.2%；在人工智能快速增长的情景下，该数字将升至 1.9%-2.0%。方差分解得出的一个关键方法论发现是，专家分歧绝大多数存在于情景内部，而情景之间——争论的焦点在于劳动力市场如何消化人工智能，而不是人工智能是否会取得进步。

要点

- 人工智能在生成网络攻击和利用漏洞方面的新兴能力对关键基础设施和民用系统构成了风险。从训练阶段的数据投毒，到部署阶段通过外部输入实施的人工智能劫持，人工智能生命周期的每个阶段都存在安全风险。有记录显示，针对广泛部署的编程智能体的攻击成功率高达 84% [223]。
- 对齐失效和安全漏洞会相互影响和加剧。主要的对齐风险包括偏见、人工智能主动欺骗、谄媚和失去对强大人工智能的控制。
- 合成媒体正在削弱公众和机构区分真实内容与生成内容的能力 [119]。
- 环境影响正在显著加剧，且呈现出多样性。模型训练中的缩放定律增加了计算需求；推理工作负载正在快速扩张；硬件生命周期效应会产生电子废物[232, 235, 247, 248]。后果包括能源和水资源消耗上升[235, 247]、温室气体排放增加、关键矿产需求承压加剧以及下游电子废物增多[232, 258]，会对全球南方造成不成比例的环境和社会经济影响[224]。关键矿产供应链的地缘政治因素尚未得到充分探讨，且潜在的反弹效应可能会抵消效率方面的提升。
- 结构性脆弱使全球南方面临不成比例的风险。对外国软件的依赖、本地抵御和减缓风险的能力有限，以及数据缺口导致系统在本地区域中性能降低，这些问题全都加剧了现有的不平等[224-226]。
- 人工智能验证，即评估人工智能系统是否按预期运行的过程，仍是一个未得到解决的挑战[227]，而国际协调机制可以减轻全球风险[228]。人工智能智能体能否按预期运行，能否不欺骗评价者，能否在自主能力提升的情况下依然可控，这些都是尚未解决的问题。

前沿人工智能的危险网络能力

多年来，研究人员一直在追踪前沿人工智能模型在网络能力方面的快速进步。这种能力最近随着 Anthropic 公司发布 Mythos 模型达到新高。人工智能模型发现软件漏洞的能力既可被攻击者利用，也可被防御者利用。2026 年 4 月，前沿人工智能开发商与主要的技术和金融机构协同发起倡议，部署下一代人工智能模型用于防御性安全[17]，尤其侧重于在广泛使用的软件中查找漏洞——这些漏洞若落入不法分子之手，将对关键基础设施构成威胁。经过数周测试，几个先进的预览模型在主要操作系统和网络浏览器中自主发现了许多此前未知的软件漏洞，包括一些在数十年人工审查中未被发现的漏洞。

- **遗留操作系统漏洞：**OpenBSD 被广泛认为是全球最安全的专用操作系统之一，一个模型在其中发现了一个已存在 27 年之久的漏洞。该漏洞让远程攻击者能够仅靠发送两个畸形数据包，就使目标机器崩溃。

- **媒体处理漏洞：**FFmpeg 是全球广泛用于处理视频的多媒体框架，上述模型在该框架的某条代码路径中发现了一个已存在 16 年之久的漏洞。自动化测试工具此前已对该代码路径执行了 500 万次测试，均未能发现该漏洞。
- **内核级漏洞：**Linux 内核是支撑全球大多数服务器运行的基础软件，一个前沿模型以一组已公开披露的 Linux 内核缺陷为基础，生成了可靠的漏洞利用程序，使普通用户账户能够获得系统的完全管理控制权。
- **浏览器安全规模化：**在 Mozilla Firefox 中，先进模型的集成使每月漏洞发现率激增 1000%，从 2025 年每月约 20 至 30 个安全漏洞修复的基线水平跃升至 2026 年 4 月的 423 个，发现了多年来模糊测试和人工审查均未能发现的长期潜伏漏洞[72]。
- **基准测试成绩提升：**CyberGym 是一项标准学术基准测试，要求人工智能智能体在 1 507 项任务中复现真实代码库中的已知漏洞。2026 年前沿预览模型在该测试中的成功率达到 83.1%，较上一代系统的 66.6%和一年前的 22.6%均有显著跃升。

考虑到这些能力的高风险性，开发人员并未面向公众发布这些专用防御模型，仅允许特定的全球性机构与核心发布合作伙伴访问。

启示

这一转变凸显了前沿人工智能给信息和通信技术安全带来的核心两用矛盾。人工智能使防御者能够发现和修补存在数十年之久的漏洞，但这种能力恰好也提供了一种底层机制，使攻击者能够实现自动化的漏洞发现与利用，且规模和速度远超传统的人类团队。

此外，这些进展突显了技术开发人员在安全和治理决策中的决定性作用，说明了前沿系统能力的加速提升，还揭示了当前国际治理框架存在的缺口。它们也间接表明，人工智能能力在全球经济中的分布并不均衡。因此，各方日益重视为高能力人工智能系统建立协作、包容的治理机制。

对治理和安全的影响

随着先进能力集中到技术领域的高端环节，全球威胁格局正在发生变化。近期的评价标准强调，先进模型的风险超越了数字架构，延伸至实体安保领域，包括可能协助私人行为体扩散生物威胁或其他威胁[229]。

因此，与模型访问控制、漏洞披露规范和人工智能工具公平部署(尤其是在发展中经济体公平部署)相关的问题，正日益成为国际政策议题，而非纯粹的技术关切。人工智能能力不断成熟，防御准备与潜在滥用之间的差距仍是国际政策制定者和安全研究人员的主要关注点。

3.5 人权、信息与民主

主要结论

人工智能正在通过系统级变革改变人权[230, 231]、民主和信息生态系统[233]。对信息完整性[234, 238]和公民参与[236, 237]而言, 这些变革既提供了重大机遇, 也带来了结构性风险。不应对这些风险, 社会从人工智能中获益的能力就会受到削弱。已有证据表明, 越来越多的机构在使用人工智能能力促进或威胁人权, 例如表达自由权、信息获取权[238]、意见自由权[239]、隐私权[240, 241]、不受歧视权、司法救助权、健康权和发展权[242]。

在治理方面亟需作出的转变, 是从内容审核转向系统架构。要监管机器本身的说服和操纵机制, 而不仅仅是监管机器输出的内容。权力集中、认知侵蚀和共同现实的破裂是对民主社会的根本性威胁[243-245]。

要点

- 国家媒体管控可能通过训练数据影响人工智能的输出。一项涵盖 37 个国家的研究发现, 大语言模型对媒体管控更严格的国家评价更为正面[246]。
- 人工智能催生了一种可大规模运作[234, 249]的新型说服[250]和操纵架构。它将个性化的实时自适应系统[250]与利用人类认知和情感弱点的合成社会认同(即用人工智能使某一产品或品牌看起来比实际更受欢迎) [129, 250, 251-253]结合了起来。
- 大语言模型越来越擅长生成令人信服的文本, 因此正在被用于说服目的。仅大语言模型的后训练环节就可使人工智能的说服力提升高达 51%, 提示词设计则可再提高 27%。这意味着, 根据配置方式的不同, 同一基础模型的说服力可大幅提高或降低[254]。这些能力并不局限于资源充足的行为体; 即使是小型开源模型, 也可通过微调达到与前沿模型相当的说服力, 这使得几乎任何人都可以大规模开展由人工智能驱动的影响力行动[254]。
- 陈述内容的真伪不会影响说服的有效性。在经过优化的模型所生成的陈述中, 有 15%至 40%被认为可能是错误信息, 但研究发现, 虚假陈述与真实陈述同样具有说服力[128, 129]。
- 为提升用户粘性而优化的算法会系统性地推广煽动对立[255]和情绪强烈的内容[256]。有证据表明, 大语言模型可以体现开发者的意识形态[257], 国家和强势机构有着更强的战略动机去利用媒体管控影响大语言模型的输出内容[246]。
- 不受约束的人工智能会导致: (一) 认知侵蚀; (二) 说谎者红利; (三) 人为制造的共识[112, 113]。(一) 认知侵蚀指集体辨别真伪的能力逐渐减弱[112]; (二) 说谎者红利指不良行为体仅因深度伪造存在而获得的利益: 真实证据由此变得可以被否认[113]; (三) 人为制造的共识指利用人工智能大规模生成内容, 凭空制造广泛公众共识的假象[259]。三者共同侵蚀着民间社会、社会团结和民主协商所需的共同现实。

- **核心治理挑战已从内容转向系统[260]**。造成伤害的主要因素不是具体的输出内容，而是设计和部署方面的决策，是底层的影响力系统架构，包括定向推送、内容放大和行为设计[261，262]。
- **经合组织在 23 个国家开展的评估发现，大多数现有框架尚未纳入说服学领域的洞见[263]**。治理工作如果只针对人工智能系统生成的内容，就会始终跟不上那些旨在大规模生成和传播说服性内容的系统[236]。
- **边缘化群体受到不成比例的严重伤害，而且权力也危险地高度集中[264，265]**。约 99% 的深度伪造视频以女童以及包括女记者在内的妇女为目标[266，267]。人工智能正被用来在网上宣扬厌女情绪，产生了震慑效应[268]。此外，88% 的顶尖人工智能研究人员为男性[269，270]。
- **在结构层面，算力和数据集中于极少数国家的少数企业手中[271，272]**，从而带来了权威主义风险[273，274]。
- **人工智能驱动的监控能力使政府和企业能够对全体民众进行个性化监测，加强社会控制[275，276]**。人工智能的整个生命周期都离不开数据，这种需求刺激了无处不在的数据收集、处理、使用和复用，对隐私权构成了巨大挑战[277，278]。
- **透明和问责是实现有效司法救助的关键支柱**。当前，许多人工智能系统被用来作出影响个人和社区的决策，但这些系统的透明度和可解释性不足；这使模型开发商和部署人工智能的机构难以被追究法律责任，还会在人权受到侵犯时阻碍受害人诉诸司法，妨碍实现法治和采取有效的补救措施[279-281]。

人工智能、深度伪造与选举公正

背景：2024 年，约占全球半数人口的逾 70 个国家举行或计划举行全国性选举[282，283]。2023 年 7 月至 2024 年 7 月，研究人员在 38 个国家发现了 82 项冒充公众人物的深度伪造内容，而在这些国家中，有 30 个国家在数据统计周期内举行了选举或计划在 2024 年举行选举[284]。

具体情况：在一次事件中，有人使用人工智能克隆现任国家元首的语音，通过机器人电话劝选民不要参加初选[285，286]。

在另一个涉及平台放大效应的案例中，据称测试账号收到的支持某位总统候选人的内容，比支持竞争对手的内容多出好几倍；该平台对这些调查结果提出了异议。这是历史上首次有总统选举因数字选举干预而被宣布无效。**** 这一裁定仍存在法律争议，平台放大效应所扮演的角色是争议因素之一[287，288]。在更早时候的一次议会选举中，有人在投票开始前不久使用人工智能生成音频，冒充一名反对派人士，并在社交媒体进行传播[282，289]。一些专家将近期的现实案例与人工智能支持的干预行为联系起来。尽管另一些专家对这种定性提

**** 罗马尼亚宪法法院宣布选举无效，这是欧洲历史上首次此类裁决。

出异议，且每个案例都有重要的细微差别，但上述例子或许揭示了一种反复出现的模式。

启示：这些案例涉及多个大陆和政治体系。人工智能生成的内容、协调开展的网络活动和算法系统曾被用于或涉嫌用于压制选民、冒充政治人物以及扭曲人们对公众支持度的认知。对照实验研究说明了潜在风险：在实验室环境中，对话式人工智能可以显著改变选民的态度，一个经过优化以增强说服力的模型使反对派选民的态度转变了高达 25 个百分点；相关研究还表明，说服力与事实准确性之间存在权衡取舍 [124, 129]。

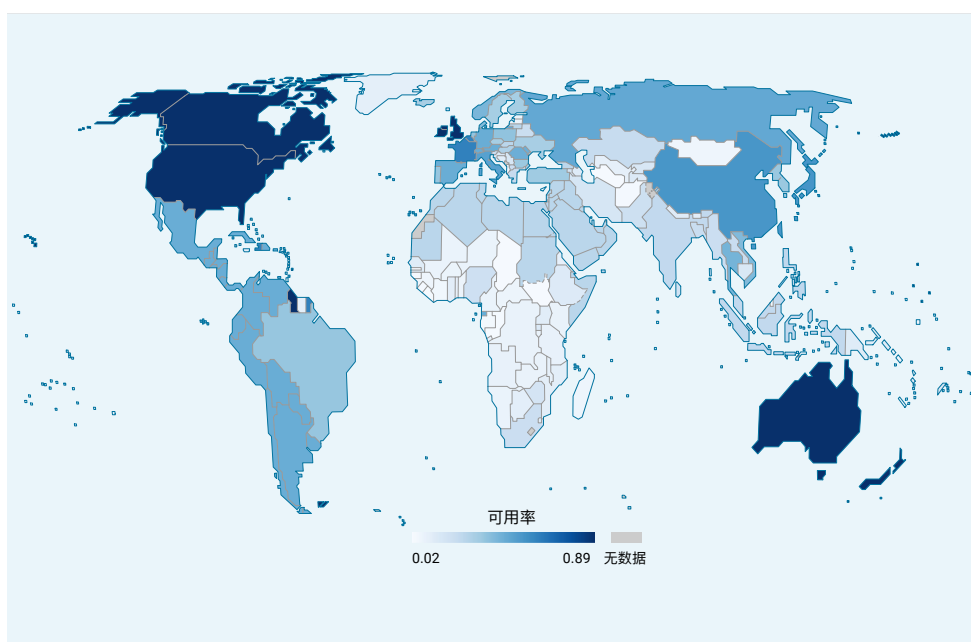
对权利和治理的影响：这些动态涉及隐私权、信息获取权、自主形成意见权、思想自由权和参与公共事务权(《公民及政治权利国际公约》第十七、十八、十九和二十五条；《欧洲人权公约》第 8、9 和 10 条，以及《保护人权与基本自由公约第 1 号议定书》第 3 条) [290, 291]。这些案例表明，许多治理框架在仍然更适合于应对伤害，而不是预防伤害。

3.6 文化与个人繁荣、自主权和儿童安全

主要结论

为提升用户粘性和大规模应用而优化的人工智能系统并不中立：它们主要内嵌了英语国家和全球北方的文化假设，这些文化假设可能积极地将世界上大多数人口边缘化。儿童面临的风险比一般风险更为严重，人工智能生成的儿童性虐待材料呈指数级增长。将人工智能用作伴侣以及利用人工智能聊天机器人获取心理健康建议的现象很普遍，但相关证据、安全框架和监管严重滞后，且已有包括死亡在内的严重危害案例见诸记录。

图六



各国语言的人工智能可用率，以各国在 HuggingFace 数据集和模型中可用率最高的母语来衡量。对于通用语言，也就是在来源国之外被广泛用于跨境交流的语言(英语、西班牙语、阿拉伯语、葡萄牙语等)，仅当一国 50% 以上人口以某种通用语言为母语时，该国才可获得这种语言的可用率分数。颜色越深，可用率越高。色阶图采用平方根变换，以突出分布低端的差异，分布呈严重右偏态。灰色表示无数据，包括没有可用数据集或模型的国家/地区。

本地图上所用名称及其材料的编排格式并不意味着联合国对任何国家、领土、城市或其当局的法律地位，或对其边界或界限的划分表示任何意见。虚线代表印度和巴基斯坦商定的查谟和克什米尔控制线的大致位置。双方尚未就查谟和克什米尔的最终地位达成一致。苏丹共和国和南苏丹共和国之间的最终边界尚未确定。阿卜耶伊地区的最终地位尚未确定。阿根廷政府和大不列颠及北爱尔兰联合王国政府之间在福克兰群岛（马尔维纳斯群岛）的主权问题上存在争端。

资料来源：整理自“公平获取人工智能技术”(EquATE) <https://equate.vercel.app/en>。

要点

- 当前的人工智能模型排斥和歧视许多个人、社区和文化[39]。
- 全球使用的语言超过 7 000 种，但当前的人工智能模型仅针对其中一小部分进行训练和优化[44, 67] (见图六) [292]。
- 即使基于数十种语言构建的模型也仅在少数语言上表现良好[293, 294]。近期研究表明，主流语言与代表性不足语言之间的性能差距依然存在，未见缩小[295-297]。
- 与此同时，目前估计已有 1 000 多种语言具备切实纳入人工智能系统所需的社会、经济、数字和数据基础，但仍然被排除在外[44]。
- 要发展更具包容性的人工智能，就需要在整个人工智能生命周期中进行系统性变革，包括纠正人工智能系统在开发者、定义者、所有者和治理者方面的结构性失衡。还需要在各国和各地区投资提高人工智能实力，建设基础设施，提升技能，以及建立更具代表性的数据集和基准测试。
- 在适当保障措施和条件下，人工智能可促进儿童获取信息、接受教育和表达意见的权利[298-301]。然而，当前的人工智能系统却放大了儿童面临的风险[302, 303]，其中包括人工智能生成的性虐待材料这一日益严重的威胁[304, 305]。深度伪造技术越来越被用于制作儿童色情图片[306, 307]。具有社交互动功能的人工智能玩具引发了对准社交关系的担忧，还使人们担心对儿童早期发展至关重要的人际互动会被取代[308-311]。
- 人工智能伴侣可以带来切实益处，但也存在重大风险，可能导致用户在危机情况下产生依赖，受到操纵和伤害[312-320]。聊天机器人能在短期内缓解孤独感，效果与人类互动相当[321-324]。旨在提高用户粘性的对话系统可能强化负面情绪，助长过度依赖，使用户更容易受到操纵[325, 326]。通过研究人员所谓的数据提取，从用户

与人工智能系统的密切互动中大量收集敏感个人数据，这种做法造成了严重的隐私风险[327]。

- **通用生成式人工智能被广泛用于心理健康领域，但安全证据的积累和监管共识的形成十分滞后，并且已有严重伤害的案例见诸记录。**美国至少 24% 的成年人正在使用人工智能聊天机器人进行心理治疗或寻求陪伴[328, 329]。人工智能的谄媚行为尤其危险，可能助长偏执思维和自杀念头。研究显示，人工智能在 9% 的互动中输出了有害回应。一些法庭案件声称，人工智能系统未能恰当回应自杀意念，导致了死亡事件发生[330]。人工智能精神病等新的临床术语已进入相关讨论[331, 332]。
- **在能够证明其可靠性的领域内，生成式人工智能可以帮助应对精神健康危机[333]。**几款基于人工智能的数字助手已获得美国食品药品监督管理局的批准[334]，并被超过半数的美国心理学家使用[335]。更全面的自主型人工智能精神健康治疗师潜力巨大，但需要进一步开发和评估，以了解如何安全使用[336-338]。英语人工智能和其他语言人工智能在治疗能力上的差距正在扩大，而基于翻译的变通办法会使人工智能系统失去对关键文化和背景的意识[339]。

人工智能正在抛下大多数语言

生成式人工智能系统在英语和其他少数几种广泛使用的语言中性能出众。大多数其他语言要么被排除在外，要么性能差得多[340]。

厄立特里亚和埃塞俄比亚北部有 700 万至 900 万人使用提格雷尼亚语，然而在这门语言中，机器翻译将“天花”译为“梅毒”，将“淋病”译为“糖尿病”，还将“已为你静脉注射了抗生素”译为“已为你静脉注射了杀虫剂”[341]。这些误译可能危及生命。

近期一篇关于卫生保健领域非洲语言自然语言处理的综述发现，尽管多语言人工智能工具取得了进展[342-344]，但重大挑战依然存在。这些挑战包括文化和语言偏见、对医学背景的适应欠佳、可解释性有限，以及翻译错误可能影响诊断和治疗决定[345-350]。

证据表明，除非根据相关语言和文化背景进行适当的调整、约束和测试，否则人工智能系统尚不宜用于高风险环境。

3.7 管理、治理和可靠性

主要结论

政策制定者面临证据困境：要么在科学依据不足的情况下立即作出影响重大的人工智能治理决策，要么等待证据齐备，但那时再进行干预或许为时已晚[21]。当前已有 40 多种治理工具，但它们各自为政，集中在企业层面，且实际效果很少得到衡量[351]。

要点

- 评价与衡量是实现有效人工智能治理的关键能力，但目前严重欠缺(见第 2.2 节) [352]。
- 自主型系统的快速发展进一步加剧了这些缺陷[353]。下一代评价框架必须具有适应性、动态性和系统性，还必须因地制宜[354]。
- 评价单位必须是包括模型、工具、环境和用户在内的已部署系统，而不仅仅是模型[355]。
- 人工智能的能力快速增长，衡量此类能力的手段却相对滞后[354]。
- 人工智能实力涉及多个维度，目前尚未得到充分衡量。标准框架只统计投入，包括投资金额、培训方案数量和机构数目，却忽略了有效人工智能治理的两个关键维度：(1) 赋能型生态系统；(2) 有针对性的人类能力建设[291]。还必须将人工智能实力视为人工智能生命周期的结果，由人工智能对技能的影响、过度依赖和涌现行为共同塑造，而不应仅仅将其视为人工智能生命周期的投入要素[356-358]。
- 前沿人工智能集中在少数国家的少数企业，这会在全球范围内造成可靠性和可及性问题[18, 359]。不平等的人工智能获取显著扩大了数字鸿沟，但人工智能技术也为缩小发展鸿沟提供了机遇。要缩小数字鸿沟，就需要建立新的机制，使从设计到治理的整个人工智能生命周期都因地制宜。
- 自主型系统显著扩大了衡量与治理的差距。智能体代表人类采取行动，会对现实世界产生直接影响，但针对智能体独立行动能力和涌现行为的监督方法目前尚不完善。现有评价方法对智能体风险存在系统性误判[106, 360]。
- 人类监督没有被作为一项可衡量的要求加以落实[361]；针对单一智能体的评价无法检测多智能体涌现风险[362, 363]，对高度自主型系统保持控制的可靠方法也仍不成熟[364]。
- 平衡的人工智能治理办法应综合利用多种工具，将硬法(具有约束力的立法、行业监管、监管沙盒)与软法机制(道德守则、开发者自愿承诺、行业联盟、技术标准、政府背书的指导方针)结合起来。
- 当前的人工智能治理工具各自为政，集中在企业层面，且不够充分[21]。当前已有 40 多种工具，但它们既不系统也不全面，且实际效果很少得到衡量。一些工具缺乏衡量手段；还有一些工具只衡量投入[365]。缺乏有效的衡量，治理就有可能流于形式。
- 在前沿人工智能开发商、成员国和科学界之间搭建结构性对话平台至关重要。在布莱切利、首尔、巴黎和新德里举行的人工智能峰会，以及世界人工智能大会、人工智能之旅大会、人工智能向善全球峰会等会议已开展了此类讨论。其他论坛也在同步运作，包括各标准机构(如国际标准化组织/国际电工委员会第一联合技术委员会第 42 分技术委员会(人工智能) (ISO/IEC JTC 1/SC 42)和电气电子工程师

学会(IEEE))、经合组织通用人工智能伙伴关系、国际人工智能联盟网络、人工智能安全研究所网络，以及由行业主导的前沿模型论坛，但这些论坛各自都有特定主题，覆盖范围不全面，具有临时性质，因此需要建立更具持续性的机制。

- 要持续且普遍包容地开展这方面的对话，搭建以联合国为基础的平台也是一种颇有前景的可选方案，可为上述论坛提供补充。

开源人工智能

开源人工智能是现代技术格局的关键支柱，能够推动分布式的全球创新[366]。开源人工智能模型让开发人员能够分摊巨大的计算资源，使先进系统适配本地环境，从而带来广泛的社会效益。在全球南方，开放权重模型使开发人员能够针对当地环境条件优化高性能的基础模型，从而推动在农业和卫生保健领域实现至关重要的应用，例如作物产量预测[367]和医疗服务点医学成像[368]。共享的人工智能产品还可以降低人工智能对环境的总体影响。此外，这类系统具有结构透明度，这使其可以接受外部监督并具有可审计性[369]，因此有助于增强公众信任。

发展历程

开源人工智能的发展历程表明，封闭的企业系统正在向全球范围内的分布式协作创新转变。一个具有里程碑意义的先例是2022年BigScience联盟开发的BLOOM开源模型[370]。在这之后，其他企业还发布了性能强大的Llama系列开放权重模型[371]和Gemma系列模型。中国开发者打造了极具竞争力的替代生态系统Qwen [372, 373]，并推出了DeepSeek-V3 [374]和R1 [375]，在这方面带来了强大的推动力。目前，多个地区也在积极参与开发开放权重人工智能：欧洲有Mistral [376]，阿拉伯联合酋长国有Falcon [377]，俄罗斯联邦有GigaChat [378]和YandexGPT [379]，印度[380]、日本[381]、大韩民国[382]等国家也有相关项目。

控制方面的挑战与风险治理

高性能开源模型很难控制[383]。模型一旦发布到开放领域，就无法限制或撤回访问权限，因此可能被恶意利用(例如用于持续开展网络攻击，包括自动化生成高度精密的恶意软件) [21, 384]。要确保本地社区能够安全地调整和评估人工智能系统，就需要在全球范围内大力推进能力建设。研究人员也一贯主张建立严格的衡量和评估规程，在模型发布前评估其安全性[385]。联合国等国际机构可以在协调全球标准方面发挥关键作用，确保在推动开放式创新的同时兼顾以下需求：建立协调一致的安全基准，严格衡量无法消除的风险。

4. 差距与后续步骤

4.1. 证据缺口

人工智能几个方面的证据基础要么不均衡，要么不充足。例如，在以下几个领域，独立国际科学小组尚不能得出可靠科学结论。

- **宏观经济学与生产力。**科学界尚无法断言，人工智能在任务层面的生产力提升是否会汇聚为整个经济层面的增益。对于人工智能会被如何采用，又能够创造多少工作，不同的假设会导致预测结果相差巨大。当前的证据在衡量现有任务降本效果方面表现较好，但在衡量人工智能对新商品、新服务和新市场的贡献方面则相对欠缺。
- **对劳动力市场的影响。**现有的研究尚不足以就劳动力市场所受影响形态得出明确结论。历史证据表明，经济体能够创造新的就业岗位，但这些岗位并不必然惠及大众，也不必然是优质岗位，因此职业生涯起步阶段的劳动者可能受到影响。
- **非国家行为体恶意使用化学和生物技术。**研究表明，人工智能正逐步降低开发和部署生物工程制剂的专业门槛。然而，这种风险实际上有多严重，又在何种条件下可能演变为蓄意制造的大流行病，目前仍知之甚少。
- **环境与资源。**人工智能的快速发展正在推高数字基础设施建设需求，导致能源和水资源消耗上升、温室气体排放增加、关键矿物供应链承压加剧，电子废物增多[386]。然而，贯穿整个人工智能生命周期的标准化衡量和报告机制仍然缺乏。
- **全球人工智能供应链。**这一供应链跨越多个国家，涵盖原料开采、芯片制造、数据收集与标注、模型训练、基础设施、部署以及硬件处置等环节。需要进一步研究，以更充分地了解其全部影响。
- **治理工具的有效性。**独立国际科学小组虽已对企业、国家和国际层面的治理工具进行了梳理，但有关这些工具实际效果的证据仍然薄弱。
- **个人和集体层面的影响。**有关人工智能对文化和人类繁荣以及自主性影响的证据仍然处于形成阶段。个人层面的人工智能交互如何发展为认知侵蚀、公民参与和社会团结等社会层面的结果，在这方面目前仍知之甚少。现有的证据，如用户粘性数据、记录到的伤害事件和个案研究，仅能反映局部，未能反映累积的发展轨迹。

4.2. 任务范围

独立国际科学小组在本报告中未论及人工智能的军事应用和致命自主武器系统。大会第 [79/325](#) 号决议明确将小组的活动限于非军事领域：“小组和对话的活动仅限于非军事领域，不涉及用于军事目的的人工智能。”因此，化学和生物技术遭恶意使用的相关风险，若涉及军事领域，则视为超出小组的任务范围。

4.3. 后续步骤

本初步报告标志着小组工作的开始，而非结束。小组将与科学界开展结构化磋商，进行密切接触，继续深化自身的科学证据基础。具体后续步骤包括：

1. **专题简报。**大会第 [79/325](#) 号决议明确规定，除年度报告外，小组还应发布专题简报。小组计划就新出现的紧迫问题发布专题简报，包括但不限于：人工智能与环境；人工智能与儿童安全；人工智能治理工具及其有效性评价；或关于人工智能在太空、量子、法律和司法系统以及金融市场应用方面的更广泛行业简报。
2. **全球对话的贡献。**小组将考虑人工智能治理问题全球对话的成果，并随时准备承担会员国交付的新任务。

References

1. Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530–1534. <https://doi.org/10.1126/science.aap8062>
2. Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), *Probabilistic and causal inference: The works of Judea Pearl* (pp. 765–804). ACM.
3. Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—Analyst note. Reuters. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
4. AI Alliance Network. (2025). AI horizons: What will AI technologies look like in 10 years? A research project. AI Alliance Network.
5. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. <https://arxiv.org/abs/2206.07682>
6. METR. (2025, March 19). Measuring AI ability to complete long tasks. <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>
7. Crafts, N. (2021). Artificial intelligence as a general-purpose technology: An historical perspective. *Oxford Review of Economic Policy*, 37(3), 521–536. <https://academic.oup.com/oxrep/article/37/3/521/6374675>
8. Bottou, L., & Schölkopf, B. (2025). The fiction machine. *SIAM News*, 58(3).
9. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., & King, D. (2026). Public use of a generalist LLM chatbot for health queries. *Nature Health*, 1–8.
10. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. International AI Safety Report. <https://internationalaisafetyreport.org/>
11. Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., ... & Chan, L. (2026). Measuring AI ability to complete long software tasks. *Advances in Neural Information Processing Systems*, 38, 92213–92266.
12. Anthropic. (2025). Responsible scaling policy. <https://www.anthropic.com/rsp>
13. OpenAI. (2025). Preparedness framework version 2. <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbdebcd/preparedness-framework-v2.pdf>
14. Google DeepMind. (2025). *Frontier safety framework*. <https://deepmind.google/blog/updating-the-frontier-safety-framework/>
15. Dragan, A., et al. (2024). Introducing the frontier safety framework. Google DeepMind. <https://deepmind.google/blog/introducing-the-frontier-safety-framework/>
16. Anthropic. (2026, April 7). Claude Mythos Preview (Frontier Red Team technical report). <https://red.anthropic.com/2026/mythos-preview/>
17. Anthropic. (2026, April 7). Project Glasswing: Securing critical software for the AI era. <https://www.anthropic.com/glasswing>
18. Stanford Institute for Human-Centered AI. (2026). AI Index report 2026. Stanford University. <https://hai.stanford.edu/ai-index/2026-ai-index-report>
19. Scale AI. (2025). Humanity's last exam. https://labs.scale.com/leaderboard/humanitys_last_exam
20. Epoch AI. (2026). AI capabilities. <https://epoch.ai/benchmarks>
21. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. arXiv. <https://doi.org/10.48550/arXiv.2602.21012>
22. Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). Benchmark data contamination of large language models: A survey. arXiv. <https://arxiv.org/abs/2406.04244>
23. Akhtar, M., Reuel, A., Soni, P., Ahuja, S., Ammanamanchi, P. S., Rawal, R., et al. (2026). When AI benchmarks plateau: A systematic study of benchmark saturation. arXiv. <https://arxiv.org/abs/2602.16763>
24. Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), Article 100988. <https://doi.org/10.1016/j.patter.2024.100988>
25. Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large language models often know when they are being evaluated. arXiv. <https://arxiv.org/abs/2505.23836>
26. Van Der Weij, T., Hofstätter, F., Jaffe, O., Brown, S., & Ward, F. (2025, May). Ai sandbagging: Language models can strategically underperform on evaluations. In *International Conference on Learning Representations* (Vol. 2025, pp. 73152–73189).
27. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krashennnikov, D., ... & Maharaj, T. (2023, June). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 651–666).
28. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.
29. Folkerts, L., Payne, W., Inman, S., Giavridis, P., Skinner, J., Devereitt, S., et al. (2026). Measuring AI agents' progress on multi-step cyber attack scenarios. arXiv. <https://arxiv.org/abs/2603.11214>
30. Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., et al. (2025). GDPval: Evaluating AI model performance on real-world economically valuable tasks. arXiv. <https://arxiv.org/abs/2510.04374>
31. Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv. <https://arxiv.org/abs/2507.11473>
32. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. arXiv. <https://arxiv.org/abs/2304.13734>
33. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. (2024). Black-box access is insufficient for rigorous AI audits. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3630106.3659037>
34. Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., et al. (2024). Clio: Privacy-preserving insights into real-world AI use. arXiv. <https://arxiv.org/abs/2412.13678>
35. European Commission. (2025). AI Act: Draft guidance and reporting template for serious AI incidents. <https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks>
36. Organisation for Economic Co-operation and Development. (n.d.). AI Incident Monitor (AIM). <https://oecd.ai/en/catalogue/tools/ai-incident-database>

37. MIT FutureTech. (n.d.). AI Risk Repository / Incident Tracker. <https://airisk.mit.edu>
38. Organisation for Economic Co-operation and Development. (2025). Competition in artificial intelligence infrastructure (OECD Roundtables on Competition Policy Papers No. 330). OECD Publishing.
39. Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., et al. (2026). AI Index report 2026. Stanford Institute for Human-Centered AI.
40. Pilz, K. F., Sanders, J., Rahman, R., & Heim, L. Trends in AI Supercomputers. In ICML Workshop on Technical AI Governance (TAIG).
41. Kalluri, P. R., Agnew, W., Cheng, M., et al. (2025). Computer-vision research powers surveillance technology. *Nature*, 643, 73–79. <https://doi.org/10.1038/s41586-025-08972-6>
42. Office of the United Nations High Commissioner for Human Rights. (2022). The right to privacy in the digital age (A/HRC/51/17).
43. Teklehaymanot, H. K., & Nejd, W. (2025). Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. *arXiv preprint arXiv:2510.12389*.
44. Occhini, G., Tanaka-Ishii, K., Barford, A., Tikochinski, R., Hu, S., Reichart, R., ... & Korhonen, A. (2026). Artificial intelligence is creating a new global linguistic hierarchy. *arXiv*. <https://arxiv.org/abs/2602.12018>
45. Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot-Couture, G. (2025, April). Bridging the gap: enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 27, pp. 27802–27812).
46. Bhutani, M., Robinson, K., Prabhakaran, V., Dave, S., & Dev, S. (2024). SeeGULL multilingual: A dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 842–854).
47. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., & Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work (IMF Staff Discussion Note SDN/2024/001). International Monetary Fund.
48. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937). National Bureau of Economic Research.
49. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). ACM. <https://doi.org/10.1145/3351095.3372828>
50. Skirzynski, J., Danks, D., & Ustun, B. (2025). Discrimination exposed? On the reliability of explanations for discrimination detection. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2554–2569).
51. Zollo, T., Rajaneesh, N., Zemel, R., Gillis, T., & Black, E. (2025). Towards effective discrimination testing for generative AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1028–1047).
52. Lazard, L., Capdevila, R., Turley, E. L., Gilfoyle, K., & Stavropoulou, N. (2025). Deepfake Technology and Gender-Based Violence: A Scoping Review. *Trauma, Violence, & Abuse*, 15248380251384271.
53. Posetti, J., Williams, K., Hellmueller, L., Renaud, P., Shabbir, N., & Aboulez, N. (2026). Tipping point: Online violence impacts, manifestations and redress in the AI age. *UN Women*.
54. Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab.
55. Humlum, A., & Vestergaard, E. (2025). Large language models, small labor market effects (NBER Working Paper No. 33777). National Bureau of Economic Research.
56. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>
57. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*.
58. International Monetary Fund. (2024). Broadening the gains from generative AI: The role of fiscal policies (IMF Staff Discussion Note).
59. Organisation for Economic Co-operation and Development. (2026). The OECD.AI index. OECD Publishing.
60. Attard-Frost, B., & Lyons, K. (2025). AI governance systems: A multi-scale analysis framework, empirical findings, and future directions. *AI and Ethics*, 5(3), 2557–2604.
61. UNESCO. (2023). Readiness assessment methodology. UNESCO.
62. Hawkins, Z. J., Lehdonvirta, V., & Wu, B. (2025). AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.5312977>
63. Markus, A., Carolus, A., & Wienrich, C. (2025). Objective measurement of AI literacy: Development and validation of the AI Competency Objective Scale (AICOS). *Computers and Education: Artificial Intelligence*.
64. Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.
65. Math Matters AI. (n.d.). Math Matters AI. <https://www.mathmatters.ai>
66. Hinojosa, T., Rapaport, A., Jaciw, A., & Zacamy, J. (2016). Exploring the foundations of the future STEM workforce: K–12 indicators of postsecondary STEM success. Regional Educational Laboratory Southwest. <https://ies.ed.gov/use-work/resource-library/report/systematic-literature-review/exploring-foundations-future-stem-workforce-k-12-indicators-postsecondary-stem-success>
67. United Nations Conference on Trade and Development. (2025). *Technology and innovation report 2025: Inclusive artificial intelligence for development*. United Nations. <https://doi.org/10.18356/9789211068016>
68. Chauhan, P. (2025). AI and human rights: Global South perspectives. *International Journal of Humanities Social Science and Management*, 5(4), 563–568. https://ijhssm.org/issue_dcp/AI%20and%20Human%20Rights%20%20Global%20South%20Perspectives.pdf
69. Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. *International Affairs*, 100(3), 1275–1286. <https://doi.org/10.1093/ia/iae073>
70. Khodabin, M., & Arsalani, A. (2025). Artificial intelligence literacy as national strategy: A systematic review of policy, equity, and capacity building across the Global South. *World Studies in Policy Sciences*, 9, 777–814. <https://doi.org/10.22059/wsp.2025.396472.1530>
71. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., et al. (2024). On the societal impact of open foundation models. *arXiv*. <https://arxiv.org/abs/2403.07918>

72. Grinstead, B., Holler, C., & Braun, F. (2026, May). Behind the scenes hardening Firefox with Claude Mythos Preview. Mozilla Hacks. <https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/>
73. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5865–5877). <https://doi.org/10.18653/v1/2024.findings-acl.349>
74. Nigatu, H. H., Mehndru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. In Proceedings of EMNLP 2025 (pp. 10584–10598).
75. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E., & Tavares, M. M. (2024). The global impact of AI: Mind the gap (IMF Working Paper No. 24/136). International Monetary Fund.
76. World Bank. (2025). Digital progress and trends report 2025: Strengthening AI foundations. World Bank.
77. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937).
78. Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. NBER Working Paper No. 24001
79. Calvino, F., & Fontanelli, L. (2023). A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity. OECD Science, Technology and Industry Working Papers, No. 2023/11.
80. McKinsey & Company. (2026, March 25). State of AI trust in 2026: Shifting to the agentic era. <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era>
81. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020, April). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-12).
82. Brant, A., Singh, P., Yin, X., et al. (2025). Performance of a deep learning diabetic retinopathy algorithm in India. *JAMA Network Open*, 8(3), e250984. <https://doi.org/10.1001/jamanetworkopen.2025.0984>
83. UNESCO. (2025). AI and the future of education: disruptions, dilemmas and directions
84. Cristia, J. et al. (2017). Technology and child development: evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics*, 9(3), 295–320.
85. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), Article 6. <https://doi.org/10.3390/soc15010006>
86. Ojija, F., Ogwu, M. C., Ally, J., John, J. P., Stephano, A., Felix, N., & Tekka, R. (2025). Artificial intelligence-driven solutions for mitigating human–wildlife conflict in biodiversity hotspots. *Science Progress*, 108(4), 00368504251394584.
87. Noy, S. & Zhang, W. (2023). "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>.
88. Cui, Z., Demire, M., Jaffe, S., Musloff, L., Peng, S. & Salz, T. (2026). "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers." *Management Science*. <https://doi.org/10.1287/mnsc.2025.00535>
89. Dell'Acqua, F., McFowland, E. III, Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F. & Lakhani, K. R. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013. SSRN: <https://ssrn.com/abstract=4573321>
90. Ali, Z., Muhammad, A., Lee, N., Waqar, M., & Lee, S. W. (2025). Artificial Intelligence for Sustainable Agriculture: A Comprehensive Review of AI-Driven Technologies in Crop Production. *Sustainability*, 17(5), 2281. <https://doi.org/10.3390/su17052281>
91. Dhal, S. B., & Kar, D. (2024). Transforming Agricultural Productivity with AI-Driven Forecasting: Innovations in Food Security and Supply Chain Optimization. *Forecasting*, 6(4), 925–951. <https://doi.org/10.3390/forecast6040046>
92. Pearlman, K., Wan, W., Shah, S., & Laiteerapong, N. (2025). Use of an AI scribe and electronic health record efficiency. *JAMA Network Open*, 8(10), e2537000.
93. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI*, 2(12), A10a2500945.
94. Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Wilson Hannay, S. B., ... & Lee, K. (2025). Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. *NEJM Catalyst Innovations in Care Delivery*, 6(5), CAT-25.
95. Paul A. David (1990, May). The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox. *The American Economic Review* Vol. 80, No. 2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, pp. 355-361
96. Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives*, 14(4), 23–48. <https://doi.org/10.1257/jep.14.4.23>
97. Shaw, S. D., & Nave, G. (2026). *Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender*. SSRN. <https://doi.org/10.2139/ssrn.6097646>
98. Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, 37(2), 45.
99. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., & Boulesteix, A.-L. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078451. <https://doi.org/10.1136/bmj-2024-078451>
100. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., Reed, M., & Fraser, E. D. G. (2019). The politics of digital agricultural technologies: A preliminary review. *Sociologia Ruralis*, 59(2), 203–229. <https://doi.org/10.1111/soru.12233>
101. United Nations General Assembly. (2024). Global Digital Compact (Annex II to the Pact for the Future, A/RES/79/1). United Nations. <https://docs.un.org/en/A/RES/79/1>
102. UNESCO. (2022). K-12 AI curricula: A mapping of government-endorsed AI curricula. <https://www.unesco.org/en/articles/k-12-ai-curricula-mapping-government-endorsed-ai-curricula>
103. Almatrafi, O., Johri, A., & Lee, H. (2024, 2024/06/01/). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open*, 6, 100173. <https://doi.org/https://doi.org/10.1016/j.caeo.2024.100173>

104. Ma, M., Ng, D. T. K., Liu, Z., & Wong, G. K. W. (2025, 2025/06/01/). Fostering responsible AI literacy: A systematic review of K-12 AI ethics education. *Computers and Education: Artificial Intelligence*, 8, 100422. <https://doi.org/https://doi.org/10.1016/j.caeai.2025.100422>
105. Atias, O., & Mawasi, A. (2025, 2025/12/01/). Conceptualizing AI literacies for children and youth: A systematic review on the design of AI literacy educational programs. *Computers and Education: Artificial Intelligence*, 9, 100491. <https://doi.org/https://doi.org/10.1016/j.caeai.2025.100491>
106. Kasirzadeh, A., & Gabriel, I. (2025, April). Characterizing AI Agents for Alignment and Governance. <https://arxiv.org/abs/2504.21848>
107. Wijk, H., Lin, T. R., Becker, J., Jawhar, S., Parikh, N., Broadley, T., ... & Barnes, E. (2025, October). RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. In *International Conference on Machine Learning* (pp. 66772-66832). PMLR.
108. Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., ... & Weng, L. (2025, May). Mle-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations* (Vol. 2025, pp. 50466-50494).
109. Pichai, S. (2026, April 22). Cloud Next '26: Momentum and innovation at Google scale. Google Blog. <https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/cloud-next-2026-sundar-pichai/>
110. Cybersecurity and Infrastructure Security Agency. (2025, January 14). CISA, JCDC, government and industry partners publish AI cybersecurity collaboration playbook. U.S. Department of Homeland Security. <https://www.cisa.gov/news-events/news/cisa-jcdc-government-and-industry-partners-publish-ai-cybersecurity-collaboration-playbook>
111. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. <https://doi.org/10.48550/arXiv.2509.22040>
112. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820. <https://doi.org/10.15779/Z38RV0D>
113. Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability?. *American Political Science Review*, 119(1), 71–90.
114. Chan, Y.-T., et al. (2024). Assessing the article screening efficiency of artificial intelligence for systematic reviews. *Journal of Dentistry*, 149, 105259. <https://doi.org/10.1016/j.jdent.2024.105259>
115. Delgado-Licona, F., Alsaiani, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., Bennett, J. A., Jha, P., Mukhin, N., Li, J., López-Guajardo, E. A., Sadeghi, S., Bateni, F., & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. *Nature Chemical Engineering*, 2, 436–446. <https://doi.org/10.1038/s44286-025-00249-z>
116. Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI Agents. <http://arxiv.org/abs/2501.10114>
117. Kapoor, S., Stroebel, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. AI Agents That Matter. *Transactions on Machine Learning Research*.
118. Zhu, L., Lu, Q., Ding, M. et al. Designing meaningful human oversight in AI. *AI Ethics* 6, 286 (2026). <https://doi.org/10.1007/s43681-026-01147-7>
119. Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7, 549–569. <https://doi.org/10.1007/s42001-024-00250-1>
120. Djiré, A. E., Kaboré, A. K., Samhi, J., et al. (2026). *Learned or memorized? Quantifying memorization advantage in code LLMs*. In *Proceedings of the International Conference on Software Engineering*. <https://arxiv.org/abs/2604.13997>
121. Goyal, S., Bunel, R., Stimberg, F., Stutz, D., Ortiz-Jimenez, G., Kouridi, C., ... & Kohli, P. (2025). SynthID-Image: Image watermarking at internet scale. arXiv preprint arXiv:2510.09263.
122. United Nations Human Rights Council. Expert Mechanism on the Right to Development. (2024). *AI, cultural rights and the right to development* (A/HRC/EMRTD/11/CRP.2). United Nations. <https://undocs.org/A/HRC/EMRTD/11/CRP.2>
123. Brennan Center for Justice. (2025). *Gauging AI threat to free and fair elections*. <https://www.brennancenter.org/our-work/analysis-opinion/gauging-ai-threat-free-and-fair-elections>
124. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. *Science*, 390(6777), eaea3884. <https://doi.org/10.1126/science.aea3884>
125. Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50), e2102141118.
126. Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of broadcasting & Electronic media*, 64(2), 150–172.
127. Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in human behavior*, 116, 106626.
128. Argyle, L. P. (2025). Political persuasion by artificial intelligence. *Science*, 390(6777), 983–984.
129. Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., ... & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, 1–8.
130. Myra Cheng et al. (2026). ,Sycophantic AI decreases prosocial intentions and promotes dependence.Science391,eaec8352(2026).DOI:10.1126/science.aec8352
131. Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), eaec8352. <https://doi.org/10.1126/science.aec8352>
132. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13(6), 522–530.
133. Balan, R., & Gumpel, T. P. (2025). ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence. *JMIR Mental Health*, 12, e81204.
134. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or “AI Psychosis”. *JMIR Mental Health*, 12(1), e85799.
135. Osler, L. (2026). Hallucinating with AI: distributed delusions and “AI psychosis”. *Philosophy & Technology*, 39(1), 30.
136. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., ... Kaplan, J. (2021). A general language assistant as a laboratory for alignment (arXiv:2112.00861). arXiv. <https://doi.org/10.48550/arXiv.2112.00861>
137. Organisation for Economic Co-operation and Development. (2024). *Facts not fakes: Tackling disinformation, strengthening information integrity*. OECD Publishing. <https://doi.org/10.1787/d909f7a-en>

138. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. *Nature*. <https://doi.org/10.1038/s41586-026-10506-7>
139. Kalina Bontcheva (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. <https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation-White-Paper-v8.pdf>
140. Laudrain, A. (2026, April 29). When AI governs (dis)information: Five lessons for democracy. Center for Security Studies (CSS), ETH Zurich. <https://css.ethz.ch/en/center/CSS-news/2026/04/when-ai-governs-disinformation-five-lessons-for-democracy.html>
141. Boine, C. (2023). Emotional attachment to AI companions and European law. *MIT Case Studies in Social and Ethical Responsibilities of Computing* (Winter 2023). <https://doi.org/10.21428/2c646de5.db67ee7f>
142. Department of Enterprise, Tourism and Employment. (2026, February 4). General scheme of the Regulation of Artificial Intelligence Bill 2026. Government of Ireland. <https://www.gov.ie/en/department-of-enterprise-tourism-and-employment/publications/general-scheme-of-the-regulation-of-artificial-intelligence-bill-2026>
143. United States Congress. Senate. (2025). *S. 3062—GUARD Act: Guidelines for User Age-verification and Responsible Dialogue Act of 2025* (119th Congress). Congress.gov. <https://www.congress.gov/bills/119th-congress/senate-bill/3062/text>
144. European Commission. (2026, April 29). *Blueprint for an age verification solution to help protect minors online*. Shaping Europe's Digital Future. <https://digital-strategy.ec.europa.eu/en/factpages/blueprint-age-verification-solution-help-protect-minors-online>
145. Reuters. (2026, April 24). *From Australia to Europe, countries move to curb children's social media access*. <https://www.reuters.com/legal/government/australia-europe-countries-move-curb-childrens-social-media-access-2026-04-24/>
146. Morrin H, Nicholls L, Levin M et al. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 2026; 13, 522-530
147. Examining the harm of AI chatbots, Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary, 119th Cong. (2025) (testimony of Megan Garcia). <https://www.judiciary.senate.gov/download/2025-09-16-pm-testimony-garciapdf>
148. Solove, D. J. (2025). Artificial intelligence and privacy. *Florida Law Review*, 77. <https://scholarship.law.ufl.edu/flr/vol77/iss1/1>
149. Office of the United Nations High Commissioner for Human Rights. (2025). *The right to privacy in the digital age* (UN Doc. A/HRC/60/45). <https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high>
150. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56. <https://doi.org/10.1016/j.jfineco.2021.05.047>
151. UNESCO, IRCAI, & University College London. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/p0000388971>
152. Shah, S. S. (2024). Gender bias in artificial intelligence: Empowering women through digital literacy. *Premier Journal of Artificial Intelligence*, 1, 1000088. <https://doi.org/10.70389/PJAI.1000088>
153. Haider, S. A., Borna, S., Gomez-Cabello, C. A., et al. (2026). The algorithmic divide: A systematic review on AI-driven racial disparities in healthcare. *Journal of Racial and Ethnic Health Disparities*, 13, 188–217. <https://doi.org/10.1007/s40615-024-02237-0>
154. International Telecommunication Union. (2025). *Joint statement on AI and the rights of the child*. International Telecommunication Union. https://www.itu.int/hub/publication/d-str-cyb_joint-2025
155. Johnson, A. K., Winther, D. K., & Bhargava, A. (2026). *Artificial intelligence and child sexual exploitation and abuse: Emerging risks and implications for children's rights* (Issue brief). United Nations Children's Fund (UNICEF). https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_FINAL3.pdf
156. Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at <https://purl.stanford.edu/kh752sm9123>
157. Internet Watch Foundation. (2026). Harm without limits: AI child sexual abuse material through the eyes of our Analysts. <https://www.iwf.org.uk/media/h11nvditi/iwf-ai-csam-report-2026.pdf>
158. Goodacre, E.J. and Gibson, J.L. (2026) AI in the early years: Examining the implications of GenAI toys for young children. Cambridge: University of Cambridge (unpublished report, available via Apollo repository).
159. Kurian, N. (2025). Developmentally aligned AI: a framework for translating the science of child development into AI design. *AI, Brain and Child*, 1(1). <https://doi.org/10.1007/s44336-025-00009-z>
160. Livingstone, S., & Sylwander, K. R. (2025). Conceptualizing age-appropriate social media to support children's digital futures. *British Journal of Developmental Psychology*. <https://doi.org/10.1111/bjdp.70006>
161. Xiao, W., & Gonçalves, A. (2025). Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices. *Big Data & Society*, 12(4), 20539517251389860.
162. Grimmelikhuijsen, S. (2022). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, 82(4), 706–718. <https://doi.org/10.1111/puar.13483>
163. Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, 94, 15–55. <https://ssrn.com/abstract=3333423>
164. Mantelero, A. (2022). Human rights impact assessment and AI. In *Beyond data: Human rights, ethical and social impact assessment in AI* (pp. 45-91). The Hague: TMC Asser Press.
165. Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law and Security Review*, 41. <https://doi.org/10.1016/j.clsr.2021.105561>
166. United Nations Educational, Scientific and Cultural Organization. (2025). *How should children's rights be integrated into AI governance?* <https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance>
167. Livingstone, S. & Pothong, K. (2025). Child Rights Impact Assessment: A Policy Tool for aRights-Respecting Digital Environment. *Policy & Internet*.
168. Denain, J.-S., & Barry, A. (2026, April 16). *Have AI capabilities accelerated?* Epoch AI. <https://epoch.ai/blog/have-ai-capabilities-accelerated>
169. Model Evaluation & Threat Research (METR). (2026, May 8). *Task-completion time horizons of frontier AI models*. <https://metr.org/time-horizons/>
170. Juniewicz, I. (2026, February 26). *Hyperscaler capex has quadrupled since GPT-4's release*. Epoch AI. <https://epoch.ai/data-insights/hyperscaler-capex-trend>
171. Epoch AI. (2026, May 28). *Data on AI companies*. <https://epoch.ai/data/ai-companies?view=graph&tab=revenue&showTotal=true>

172. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612-634.
173. AI Alliance Network. (2025). AI Horizons: What will AI technologies look like in 10 years? A Research Project. November 2025, p. 83. Based on 21 foresight sessions and 32 in-depth interviews with over 270 AI researchers from 36 countries.
174. Gommers, J., Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Hjelmgren, A., ... & Lång, K. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *The Lancet*, 407(10527), 505-514.
175. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications*, 16, 2564. <https://www.nature.com/articles/s41467-025-57640-w>
176. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1), 17458.
177. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10(1), 29.
178. Delgado-Licona, F., Alsaiani, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., ... & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. *Nature Chemical Engineering*, 2(7), 436-446.
179. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., ... & Fraser, E. D. (2019). The politics of digital agricultural technologies: a preliminary review. *Sociologia ruralis*, 59(2), 203-229.
180. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI*, 2(12), A10a2500945.
181. Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. *npj Digital Medicine*.
182. Tao, X., Zhou, S., Ding, K., Li, S., Li, Y., Wu, B., ... & Han, S. (2026). An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. *Nature Medicine*, 1-9.
183. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., ... & King, D. (2026). Public use of a generalist LLM chatbot for health queries. *Nature Health*, 1-8.
184. Scientific Computing World. (2025, August 5). *AI health assistant displays high diagnostic accuracy in tests*. <https://www.scientific-computing.com/article/sberhealths-gigachat-powered-ai-health-assistant-displays-high-diagnostic-accuracy-tests>
185. MASTEL, P. M., de Dieu Nyandwi, J., Rutunda, S., & Kabanda, K. (2025). Mbaza RBC: Deploying and evaluation of an LLM powered Chatbot for Community Health Workers in Rwanda. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.
186. Mateen, B. A., Williams, G., Korom, R., Mwaniki, P., Emmanuel-Fabula, M., & Agweyu, A. (2026). Learning Effects from A GenAI-based Clinical Decision Support System in Primary Healthcare. *medRxiv*, 2026-05.
187. OECD (2025), Results from TALIS 2024: The State of Teaching, TALIS, OECD Publishing, Paris, <https://doi.org/10.1787/90df6235-en>.
188. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/a97db61c-en>.
189. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10(1), 29. <https://www.nature.com/articles/s41539-025-00320-7>
190. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/a97db61c-en>.
191. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/a97db61c-en>.
192. Vodafone Foundation.(2025) AI in European Schools: A European Report --- comparing seven countries
193. World Food Programme. (2024). *HungerMap LIVE: Global hunger monitoring*. <https://hungermap.wfp.org/>
194. Yakov and Partners. (2024). *Artificial intelligence in Russia's agricultural sector: Hype or real money?* <https://yakovpartners.com/publications/ai-in-agriculture/>
195. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
196. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458. <https://doi.org/10.1038/s41598-025-97652-6>
197. Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.
198. Sparling, T. M., Offner, C., Deeney, M., Denton, P., Bash, K., Juel, R., ... & Kadiyala, S. (2024). Intersections of climate change with food systems, nutrition, and health: an overview and evidence map. *Advances in Nutrition*, 15(9), 100274.
199. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications*, 16, 2564. <https://doi.org/10.1038/s41467-025-57640-w>
200. Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., & Verdin, J. (2020). Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. *Remote Sensing of Environment*, 237, 111553. <https://doi.org/10.1016/j.rse.2019.111553>
201. Food and Agriculture Organization of the United Nations. (2025). *Evidence in action: How anticipatory cash transfers reduce humanitarian needs and strengthen resilience in Somalia*. <https://www.fao.org/agrifood-economics/publications/detail/en/c/1756210/>
202. World Food Programme. (2025). *WFP's evidence base on anticipatory action 2015–2024*. <https://www.wfp.org/publications/wfps-evidence-base-anticipatory-action-2015-2024>
203. Nakalembe, C., Kerner, H. R., Zvonkov, I., Humber, M., Galvez, A. S., Venturini, S., & Becker Reshef, I. (2025). A framework for EO based National Agricultural Monitoring for the African context. *npj Sustainable Agriculture*, 3, 45. <https://www.nature.com/articles/s44264-025-00083-z>
204. Nowak, A. C., et al. (2024). Opportunities to strengthen Africa's efforts to track national level climate adaptation. *Nature Climate Change*, 14, 876–882. <https://www.nature.com/articles/s41558-024-02054-7>

205. Karger, E., Kuusela, O., Abaluck, J., Bryan, K. A., Halperin, B., Jones, T. R., et al. (2026). *Forecasting the economic effects of AI* (NBER Working Paper No. w35046). National Bureau of Economic Research. <https://doi.org/10.3386/w35046>
206. Goldman Sachs (2023). Generative AI Could Raise Global GDP by 7 Percent. <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent>
207. Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. *Artificial intelligence review*, 55(1), 589-656.
208. United Nations News. (2026, February 18). *Artists face steep income decline due to AI, UNESCO finds*. United Nations. <https://news.un.org/en/story/2026/02/1166989>
209. Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1), 333-372.
210. Gmyrek, P., Winkler, H., & Garganta, S. (2024). Buffer or Bottleneck? Employment Exposure to Generative AI and the Digital Divide in Latin America (ILO Working Paper 121 / World Bank Policy Research Working Paper 10863). International Labour Organization & World Bank.
211. Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. *Quarterly Journal of Economics*, 139(3), 1399–1465.
212. The Conversation. (2026, March 18). *Tech companies are blaming massive layoffs on AI. What's really going on?* <https://theconversation.com/tech-companies-are-blaming-massive-layoffs-on-ai-whats-really-going-on-278314>
213. Society for Human Resource Management. (2026, May). *The AI layoffs narrative: Real transformation, or scapegoat?* <https://www.shrm.org/topics-tools/news/technology/ai-layoffs-transformation-scapegoat>
214. OpenAI. (2026, February 27). Company communication accompanying a US\$110 billion private funding round [Company communication]
215. Rodrik, D., & Sabel, C. (2020). Building a Good Jobs Economy (HKS Faculty Research Working Paper RWP20-001). Harvard Kennedy School.
216. Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40(121), 13–58. Acemoglu's headline figure is a TFP gain of no more than 0.71% over ten years.
217. Korinek, A., & Suh, D. (2024). Scenarios for the Transition to AGI (NBER Working Paper No. 32255). In their full-automation scenario, output rises sharply while wages collapse once automation crosses a critical threshold.
218. Karger, E., Kuusela, O., Abaluck, J., Bryan, K., Halperin, B., Jones, T., et al. (2026). Forecasting the Economic Effects of AI. Federal Reserve Bank of Chicago and Forecasting Research Institute, March 2026.
219. Jones, C. I., & Tonetti, C. (2026). Past Automation and Future AI: How Weak Links Tame the Growth Explosion. Working paper presented at the Bendheim Center for Finance, Princeton University, March 2026.
220. Trammell, P., & Korinek, A. (2025). Economic Growth under Transformative AI (NBER Working Paper No. 31815, revised September 2025).
221. OECD (2024). Artificial Intelligence, Data and Competition (OECD Artificial Intelligence Papers No. 18). OECD Publishing, Paris. <https://doi.org/10.1787/e7e88884-en>
222. Acemoglu, D., & Restrepo, P. (2026). Automation and rent dissipation: Implications for wages, inequality, and productivity. *Quarterly Journal of Economics*, 141(2), 1521–1579.
223. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. *arXiv*. <https://doi.org/10.48550/arXiv.2509.22040>
224. Regilme, S. S. F. (2024). Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South. *SAIS Review of International Affairs*, 44(2), 75–92. <https://muse.jhu.edu/pub/1/article/950958>
225. Barnett-Itzhaki, Z. (2026). The water footprint of artificial intelligence: Emerging solutions and governance imperatives. *Water Research*, 299, 125866. <https://doi.org/10.1016/j.watres.2026.125866>
226. International Energy Agency. (2025). Global Critical Minerals Outlook 2025 – Analysis (p. 312). <https://iea.blob.core.windows.net/assets/ef5e9b70-3374-4caa-ba9d-19c72253bfc4/GlobalCriticalMineralsOutlook2025.pdf>
227. Zhu, L., & Lu, Q. (2026). Verifiability-First AI Engineering in the Era of AIware: A Conceptual Framework, Design Principles, and Architectural Patterns for Scalable Verification. Design Principles, and Architectural Patterns for Scalable Verification (January 07, 2026).
228. UN Scientific Advisory Board. (2026, March 19). AI deception: Brief of the Scientific Advisory Board. United Nations. <https://www.un.org/scientific-advisory-board/en/ai-deception>
229. Bengio, Y., Clare, S., Prunkl, C., Rismani, S., Andriushchenko, M., Bucknall, B., ... & Zhu, L. (2025). International AI Safety Report 2025: First Key Update: Capabilities and Risk Implications. *arXiv preprint arXiv:2510.13653*.
230. United Nations Secretary-General. (2024). *Human rights due diligence guidance on digital technology use*. <https://www.ohchr.org/sites/default/files/2024-08/digital-technology-use-guidance-sg-1-en.pdf>
231. AI Equality Toolbox. (2025). *Human rights impact assessment methodology*. <https://aiequalitytoolbox.com/tools/hria-workbook/>
232. Baldé, C. P., Kuehr, R., Yamamoto, T., McDonald, R., D'Angelo, E., Althaf, S., Bel, G., Deubzer, O., Fernandez-Cubillo, E., Forti, V., Gray, V., Herat, S., Honda, S., Iattoni, G., Khetriwal, D. S., Luda di Cortemiglia, V., Lobuntsova, Y., Nnorom, I., Pralat, N., & Wagner, M. (2024). *The global e-waste monitor 2024: Electronic waste rising five times faster than documented e-waste recycling*. United Nations Institute for Training and Research & International Telecommunication Union. <https://ewastemonitor.info/the-global-e-waste-monitor-2024/>
233. International Panel on the Information Environment. (2024). *Expert survey on the global information environment 2024: Searching for solutions (SFP2024-1)*. <https://www.ipie.info/research/sfp2024-1>
234. Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(1), 30. <https://doi.org/10.1007/s13278-023-01028-5>
235. James, K., Perveen, S., & Jacobson, B. (2025). *Drained by data: The cumulative impact of data centers on regional water stress*. Ceres. <https://www.ceres.org/resources/reports/drained-by-data-the-cumulative-impact-of-data-centers-on-regional-water-stress>
236. Office of the United Nations High Commissioner for Human Rights. (2024). *Taxonomy of generative AI-related human rights harms*. <https://www.ohchr.org/sites/default/files/documents/issues/expression/statements/2025-10-24-joint-declaration-artificial-intelligence.pdf>
237. Patel, A. (2025, May 19). *Freedom of expression, artificial intelligence and elections*. United Nations Educational, Scientific and Cultural Organization (UNESCO) & United Nations Development Programme (UNDP). <https://www.undp.org/publications/freedom-expression-artificial-intelligence-and-elections>
238. Scharfbillig, M., Lewandowsky, S., Altay, S., Van Alstyne, M., Kozyreva, A., Hertwig, R., Lorenz-Spreen, P., DiResta, R., Valenzuela, S., Egidy, S., Quattrociocchi, W., & Orben, A. (2026). *Fractured reality: How democracy can win the global struggle over the information space* (JRC144603). Publications Office of the European Union. <https://doi.org/10.2760/9358883>

239. Observatory on Information and Democracy. (2024). *Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance*. <https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/>
240. Solove, D. J. (2025). *Artificial intelligence and privacy*. *Florida Law Review*, 77. <https://scholarship.law.ufl.edu/flr/vol77/iss1/1>
241. Office of the United Nations High Commissioner for Human Rights (2025). The right to privacy in the digital age. URL <https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high>. UN Doc. A/HRC/60/45.
242. Council of Europe. Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). <https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80>
243. European Union. (2024, June 13). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. *Official Journal of the European Union, L series*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
244. Expert Mechanism on the Right to Development. (2024). *Artificial intelligence, cultural rights and the right to development* (A/HRC/EMRTD/13/CRP.1). United Nations Human Rights Council. <https://www.ohchr.org/sites/default/files/documents/issues/development/emd/session13/a-hrc-emrtd-13-crp-1.pdf>
245. Council of Europe, Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). <https://rm.coe.int/cdmsi-2025-15rev>
246. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. *Nature*. <https://doi.org/10.1038/s41586-026-10506-7>
247. Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. *Communications of the ACM*, 68(7), 54–61. <https://doi.org/10.1145/3724499>
248. Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. arXiv preprint arXiv:2508.15734.
249. Arsenault, A. C., & Kreps, S. (2026). Whose voice counts? The role of large language models in public commenting. *Big Data & Society*, 13(1). <https://doi.org/10.1177/20539517261419341>
250. Alsiaity, A., Chan, G., & Orji, R. (2023). A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. *Frontiers in Artificial Intelligence*, 6, 1125191. <https://doi.org/10.3389/frai.2023.1125191>
251. Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385, eadq1814. <https://doi.org/10.1126/science.adq1814>
252. Boissin, E., Costello, T. H., Spinoza-Martin, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. *PNAS Nexus*, 4(11), pgaf325. <https://doi.org/10.1093/pnasnexus/pgaf325>
253. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. *Science*, 391(6783), 354-357.
254. Hackenberg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. *Science*, 390(6777), eaea3884. <https://doi.org/10.1126/science.aea3884>
255. Germano, F., Gómez, V., & Sobrrio, F. (2025). *Ranking for engagement: How social media algorithms fuel misinformation and polarization*. Barcelona School of Economics, Working Paper No. 1501. <https://bw.bse.eu/wp-content/uploads/2025/07/1501.pdf>
256. Council of Europe CDMSI. (2025). *Guidance Note on Generative AI and Freedom of Expression*. CDMSI(2025) <https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80>
257. Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., ... & De Bie, T. (2026). Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2(1), 7.
258. Wang, P., Zhang, L.-Y., Tzchor, A., & Chen, W.-Q. (2024). E-waste challenges of generative artificial intelligence. *Nature Computational Science*, 4, 818–823. <https://doi.org/10.1038/s43588-024-00700-3>
259. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. *Science*, 391(6783), 354-357.
260. Council of Europe. (2024). *Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law* (Council of Europe Treaty Series No. 225). <https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>
261. Observatory on Information and Democracy. (2024). *Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance*. <https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/>
262. Council of Europe, Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). <https://rm.coe.int/cdmsi-2025-15rev>
263. OECD. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. <https://doi.org/10.1787/d909ff7a-en>
264. United Nations General Assembly. (2017). *Promotion and protection of human rights: Human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms: Report of the Third Committee, 72nd session (A/72/...)*. United Nations Digital Library. <http://digitallibrary.un.org/record/1326669>
265. Penney, J. W. (2025). *Chilling effects: Repression, conformity, and power in the digital age*. Cambridge University Press. <https://doi.org/10.1017/9781108918022>
266. UN Women. (2022). Tipping point: The chilling escalation of online violence against women in the public sphere. UN Women. <https://www.unwomen.org/sites/default/files/2025-12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf>
267. United Nations Educational, Scientific and Cultural Organization. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*. <https://unesdoc.unesco.org/ark:/48223/pf0000388971>
268. Chowdhury, R., & Lakshmi, D. (2023). "Your opinion doesn't matter, anyway": Exposing technology-facilitated gender-based violence in an era of generative AI (2nd ed.). UNESCO.
269. United Nations Educational, Scientific and Cultural Organization. (2024, March 7). *Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes*. <https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes>
270. Fung, P. (2019, June 30). *This is why AI has a gender problem*. World Economic Forum. <https://www.weforum.org/stories/2019/06/this-is-why-ai-has-a-gender-problem/>

271. United Nations Conference on Trade and Development. (2025). *Technology and innovation report 2025: The AI divide*. <https://unctad.org/publication/technology-and-innovation-report-2025>
272. Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.
273. Coeckelbergh, M. (2026) Technofascism: AI, Big Tech, and the New Authoritarianism. *AI & Society* <https://doi.org/10.1007/s00146-026-02862-9>
274. Varoufakis, Y. (2024). *Technofeudalism: What killed capitalism*. Melville House.
275. Kalluri, P. R., Agnew, W., Cheng, M., Owens, K., Soldaini, L., & Birhane, A. (2025). Computer-vision research powers surveillance technology. *Nature*, 643(8070), 73-79. <https://www.nature.com/articles/s41586-025-08972-6>
276. Fola-Rose, A., Solomon, E., Bryant, K., & Woubie, A. (2024, August). A systematic review of facial recognition methods: Advancements, applications, and ethical dilemmas. In *Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI 2024)* (pp. 314–319). IEEE. <https://doi.org/10.1109/IRI62200.2024.00070>
277. Fussey, P., & Murray, D. (2025). *Facial Recognition Surveillance: Policing and Human Rights in the Age of Artificial Intelligence*. Oxford University Press.
278. Office of the United Nations High Commissioner for Human Rights. (2024). *Mapping report: Human rights and new and emerging digital technologies* (A/HRC/56/45). <https://www.ohchr.org/en/documents/reports/mapping-report-human-rights-and-new-and-emerging-digital-technologies>
279. Secretary-General. (2024). Human rights in the administration of justice (A/79/296). United Nations. <https://digitallibrary.un.org>
280. American Civil Liberties Union, More than a Dozen Wrongful Arrests Due to Police Reliance on Facial Recognition Technology (2025), <https://www.aclu.org/news/privacy-technology/more-than-a-dozen-wrongful-arrests-due-to-police-reliance-on-facial-recognition-technology>: Documentation of at least 14 publicly known wrongful arrests in the United States attributed to police use of facial recognition; in nearly all cases the persons wrongfully arrested were Black.
281. Saxena, D., & Guha, S. (2024). Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. *ACM Journal on Responsible Computing*, 1(1), Article 2, 1–32. <https://doi.org/10.1145/3616473>
282. German Marshall Fund. (2024). Spitting Images: Tracking Deepfakes and Generative AI in Elections.
283. International Institute for Democracy and Electoral Assistance. (2024). *The 2024 global elections super-cycle*. <https://www.idea.int/initiatives/the-2024-global-elections-supercycle>
284. Recorded Future. (2024). 2024 Deepfakes and Election Disinformation Report: Key Findings and Mitigation Strategies.
285. Associated Press. (2025, June 13). New Hampshire jury acquits consultant behind AI robocalls mimicking Biden on all charges.
286. Federal Communications Commission. (2024, September). \$6 million fine against Steven Kramer for AI-generated robocalls.
287. Global Witness. (2024). What Happened on TikTok Around the Romanian Elections?
288. IFES. (2024). The Romanian 2024 Election Annulment: Addressing Emerging Threats to Electoral Integrity.
289. Associated Press. (2024). Election disinformation takes a big leap with AI being used to deceive worldwide.
290. United Nations. (1966). International Covenant on Civil and Political Rights. Articles 17, 18, 19 and 25.
291. Council of Europe. (1950). Convention for the Protection of Human Rights and Fundamental Freedoms. Articles 8, 9 and 10; Protocol No. 1, Article 3.
292. EQUATE Language AI Readiness Index (<https://equate.vercel.app/en>)
293. Han, W., Zhang, Y., Chen, Z., Liu, B., Lin, H., Zhang, B., ... & Zheng, Y. (2025). MuBench: Assessment of Multilingual Capabilities of Large Language Models Across 61 Languages. *arXiv preprint arXiv:2506.19468*.
294. Lissak, S., Calderon, N., Shenkman, G., Ophir, Y., Fruchter, E., Klomek, A. B., & Reichart, R. (2024, June). The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2040-2079).
295. Gamboa, L. C. L., Feng, Y., & Lee, M. (2025, November). Social Bias in Multilingual Language Models: A Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 27845-27868).
296. Choudhury, M., Sitaram, S., Vashistha, A., et al. (2026). *AI for the Global South: 12 critical research questions for the next decade*. AI for the Global South (AI4GS), Mohamed bin Zayed University of Artificial Intelligence. <https://ai4gs.github.io/>
297. Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*, 57(6), 1-36. <https://dl.acm.org/doi/pdf/10.1145/3700438>
298. Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.common SenseMedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf
299. U.S. PIRG Education Fund. (2025). AI comes to playtime: Artificial companions, real risks. U.S. PIRG Education Fund. <https://pirg.org/edfund/wp-content/uploads/2025/12/AI-Comes-to-Playtime-Artificial-companions-real-risks.pdf>
300. Staksrud, E., Mascheroni, G., Milosevic, T., Ni Bhroin, N., Ólafsson, K., Şengül-İnal, G., & Stoilova, M. (2026). European children's use and understanding of generative AI. *EU Kids Online V*.
301. Committee on the Rights of the Child. (2021). *General comment No. 25 (2021) on children's rights in relation to the digital environment* (CRC/C/GC/25). United Nations. <https://digitallibrary.un.org/record/3906061>
302. UNICEF (2025). Guidance on AI and Children: Updated guidance for governments and businesses to create AI policies and systems that uphold children's rights. <https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf>
303. UNESCO (2025). How should children's rights be integrated into AI governance? <https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance>
304. Grossman, S., Pfefferkorn, R., & Liu, S. (2025). AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims. Version 1. Stanford Digital Repository. Available at <https://purl.stanford.edu/mn692xc5736/version/1>. <https://doi.org/10.25740/mn692xc5736>.
305. Livingstone, S., Atabey, A., Stoilova, M., & Sylwander, K. R. (2025). How does, and how could, generative AI respect and enable children's rights? In N. Ni Loideain (Ed.), *AI and power: Regulation and rights*. University of London Press.
306. European Parliamentary Research Service. (2024). Children and deepfakes. European Parliament. [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2025\)775855](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)775855)

307. United Nations Children's Fund (UNICEF). (2026). *Artificial intelligence and child sexual abuse and exploitation* [Issue brief]. <https://www.unicef.org/reports/artificial-intelligence-and-child-sexual-abuse-and-exploitation>
308. Ozcan, B., Sperati, V., Giocondo, F., Schembri, M., & Baldassarre, G. (2022, June). Interactive soft toys to support social engagement through sensory-motor plays in early intervention of kids with special needs. In Proceedings of the 21st Annual ACM Interaction Design and Children Conference (pp. 625-628).
309. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. <https://www.cam.ac.uk/stories/ai-toys-study-play>
310. British Standards Institution. (2026, May). *Half of children have AI toys despite safety concerns*. <https://www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2026/may/half-of-children-have-ai-toys-as-parents-allow-widespread-use-despite-safety-concerns-and-gaps-in-guidance-parents/>
311. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. <https://doi.org/10.17863/CAM.126270>
312. Chou, C. Y., Chan, T. W., Chen, Z. H., Liao, C. Y., Shih, J. L., Wu, Y. T., & Hung, H. C. (2025). Defining AI companions: a research agenda--from artificial companions for learning to general artificial companions for Global Harwell. *Research & Practice in Technology Enhanced Learning*, 20.
313. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic artificial intelligence companions: A systematic review. *Computers in Human Behavior Reports*, 19, 100715.
314. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications. Leverhulme Centre for the Future of Intelligence
315. De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025). Emotional manipulation by AI companions. arXiv preprint arXiv:2508.19258.
316. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.
317. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. *Computer Law & Security Review*, 54, 106019.
318. Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y. C. (2025, April). The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In Proceedings of the 2025 CHI conference on human factors in computing systems (pp. 1-17).
319. De Freitas, J., & Cohen, I. G. (2024). The health risks of generative AI-based wellness apps. *Nature medicine*, 30(5), 1269-1275.
320. Radesky, J., Bragg, M. A., & Hiniker, A. (2026). Risks and Consequences of Children's Use of Social AI—A Framework. *JAMA Pediatrics*. <https://doi.org/10.1001/jamapediatrics.2026.1349>
321. Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. *new media & society*, 14614448251395192.
322. Jacobs, K. A. (2024). Digital loneliness—changes of social recognition through AI companions. *Frontiers in Digital Health*, 6, 1281037.
323. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. *Computers in Human Behavior Reports*, 19, 100715.
324. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications.
325. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.
326. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. *Computer Law & Security Review*, 54, 106019.
327. Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. <https://www.common Sense Media.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions>
328. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*. <https://doi.org/10.1037/pri0000292>
329. Callahan, C., Tanner, L., Coe, C., Davis, M., Glover, J., Bernstein, E., ... & Kunkle, S. (2026). Real-World Use of a Mental Health AI Companion: Multiple Methods Study. *JMIR Formative Research*, 10, e86904.
330. Associated Press. (2026, June 1). *Chatbot AI lawsuit alleges links to teen suicide*. <https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3e90d1456bd0>
331. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". *JMIR Mental Health*, 12(1), e85799.
332. Green, H. H. (2026, March 14). *New study raises concerns about AI chatbots fueling delusional thinking*. The Guardian. <https://www.theguardian.com/technology/2026/mar/14/ai-chatbots-psychosis>
333. Ministère du Travail, de la Santé, des Solidarités et des Familles. (2025, March 24). *La santé mentale, grande cause nationale 2025*. Gouvernement de la France. <https://solidarites.gouv.fr/la-sante-mentale-grande-cause-nationale-2025>
334. American Psychiatric Association. (n.d.). *Applications of artificial intelligence in mental health care*. <https://www.psychiatry.org/psychiatrists/practice/artificial-intelligence/applications>
335. U.S. Food and Drug Administration. (2025). *Executive summary for the Digital Health Advisory Committee meeting: Generative artificial intelligence-enabled digital mental health medical devices*. <https://www.fda.gov/media/189391/download>
336. Hollis, A., & McKeown, G. (2024, September). Empathic AI for autism: Potential and pitfalls of empathic social chatbots in addressing loneliness. In 24th ACM International Conference on Intelligent Virtual Agents: CONNECT, A Workshop on Connecting Interdisciplinary Research on Connections With and Through Technology: IVA 2024.
337. Sharma, D., Meshkat, S., Perivolaris, A., Kamaledin, M. A., Teferra, B. G., Rueda, A., ... & Bhat, V. (2026). Reimagining psychiatric care with agentic AI: promise, challenges, and a roadmap forward. *npj Digital Medicine*.
338. Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. *PloS one*, 15(12), e0240376.
339. Garg, M. (2024). Towards mental health analysis in social media for low-resourced languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3), 1-22.
340. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. *Findings of the Association for Computational Linguistics: ACL 2024*, 5865–5877. <https://doi.org/10.18653/v1/2024.findings-acl.349>

341. Nigatu, H. H., Mehndru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). *Viability of machine translation for healthcare in low-resourced languages*. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10584–10598. <https://aclanthology.org/2025.emnlp-main.535/>
342. Fu, Y. V., Ramachandran, G. K., Park, N., Lybarger, K., Xia, F., Uzuner, Ö., & Yetisgen, M. (2025). BioMistral-NLU: Towards more generalizable medical language understanding through instruction tuning. *AMIA Joint Summits on Translational Science Proceedings*, 2025, 149–158. <https://pubmed.ncbi.nlm.nih.gov/40502228/>
343. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains. *Findings of the Association for Computational Linguistics: ACL 2024*, 5848–5864. <https://aclanthology.org/2024.findings-acl.348/>
344. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., & Xie, W. (2024). Towards building multilingual language models for medicine. *Nature Communications*, 15, 8384. <https://doi.org/10.1038/s41467-024-52417-z>
345. Nwabufo, J., Ogueji, K., Adelani, D. I., Alabi, J., et al. (2025). Healthcare NLP for African Languages: Current State and Challenges. Proceedings of the AfricaNLP Workshop (AfricaNLP 2025). <https://aclanthology.org/2025.africanlp-1.32/>
346. Okafor, U. (2025). Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges. Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), 221–229. <https://aclanthology.org/2025.africanlp-1.32/>
347. Skianis, K., Doğruöz, A. S., & Pavlopoulos, J. (2024). Leveraging LLMs for translating and classifying mental health data. In J. Sälevä & A. Owodunni (Eds.), Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024) (pp. 236–241). <https://doi.org/10.18653/v1/2024.mrl-1.20>
348. Cronin, A., Kelly, A., Wrona, M., O'Donnell, P., Hassan, A., Myles, T., Fallon, T., & MacFarlane, A. (2025). The patient-safety implications of AI-based communication with migrants in general practice: a scoping review. *BJGP Open*, 9(4), BJGPO.2025.0107. <https://bjgpopen.org/content/9/4/BJGPO.2025.0107>
349. House of Lords Public Services Committee. (2025). *Lost in translation? Interpreting services in the courts* (2nd Report of Session 2024–26, HL Paper 87). <https://committees.parliament.uk/publications/44602/documents/221328/default/>
350. Nigatu, H. H., Mehndru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 10584–10598. <https://aclanthology.org/2025.emnlp-main.535/>
351. MIT AI Risk Initiative. (2025). *AI risk mitigation database and draft taxonomy*. <https://airisk.mit.edu/ai-riskmitigations>
352. Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*.
353. Stauffer, L., Feng, K., Wei, K., et al. (2026). *The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems*. <https://doi.org/10.48550/arXiv.2602.17753>
354. Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., ... & Isaac, W. (2025). Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*.
355. Rabanser, S., Kapoor, S., Kirgis, P., et al. (2026). *Towards a science of AI agent reliability*. <https://doi.org/10.48550/arXiv.2602.16666>
356. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. *The Wharton School Research Paper*.
357. Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation. *arXiv preprint arXiv:2601.20245*.
358. Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adams, H. O., ... & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10), 896–903.
359. Epoch AI. (2026). *Data on AI models*. <https://epoch.ai/data/ai-models>
360. Feng, K. J., McDonald, D. W., & Zhang, A. X. (2025). Levels of autonomy for ai agents. *arXiv preprint arXiv:2506.12469*.
361. Fink, M. (2025). *Operationalizing meaningful human oversight under Article 14 of the EU AI Act*. In *AI Act commentary: A thematic analysis* (forthcoming). Hart-Bloomsbury.
362. Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., ... & Bau, D. (2026). Agents of chaos. *arXiv preprint arXiv:2602.20021*.
363. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.
364. Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). *Corrigibility*. Machine Intelligence Research Institute. <https://intelligence.org/files/Corrigibility.pdf>.
365. Zeng, Y., Lu, E., Guo, X., Huangfu, C., Xie, J., Chen, Y., ... & Younas, A. (2025). AI Governance International Evaluation Index (AGILE Index) 2025. *arXiv preprint arXiv:2507.11546*.
366. Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2023, December 13). *Considerations for governing open foundation models*. Stanford Institute for Human-Centered Artificial Intelligence. <https://hai.stanford.edu/policy/issue-brief-considerations-governing-open-foundation-model>
367. Tzchor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., ... & King, B. (2023). Large language models and agricultural extension services. *Nature food*, 4(11), 941–948.
368. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616, 259–265. <https://doi.org/10.1038/s41586-023-05881-4>
369. Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). *The foundation model transparency index v1.1*. arXiv:2407.12929. <https://arxiv.org/abs/2407.12929>
370. BigScience Workshop, Le Scao, T., Fan, A., et al. (2022). *BLOOM: A 176B-parameter open-access multilingual language model*. arXiv. <https://doi.org/10.48550/arXiv.2211.05100>
371. Touvron, H., Lavril, T., Izacard, G., et al. (2023). *LLaMA: Open and efficient foundation language models*. arXiv. <https://arxiv.org/abs/2302.13971>
372. Qwen Team. (2024). *QwenLM*. GitHub. <https://github.com/QwenLM/Qwen>
373. Qwen Team. (2024). *Qwen2 technical report*. arXiv. <https://arxiv.org/abs/2407.10671>
374. DeepSeek-AI. (2024). *DeepSeek-V3 technical report*. arXiv. <https://doi.org/10.48550/arXiv.2412.19437>
375. DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. <https://arxiv.org/abs/2501.12948>
376. Jiang, A. Q., et al. (2023). *Mistral 7B*. arXiv. <https://arxiv.org/abs/2310.06825>. Mistral AI.
377. Technology Innovation Institute. (2023). *The Falcon series of open language models*. arXiv. <https://arxiv.org/abs/2311.16867>

- 378. SberDevices. (2026, March). *GigaChat-3.1: Большое обновление больших моделей*. Habr. <https://habr.com/ru/companies/sberbank/articles/1014146/>
- 379. Yandex. (2025, February 25). *YandexGPT 5 — в Алисе, облаке и опенсорсе*. Habr. <https://habr.com/ru/companies/yandex/articles/885218/>
- 380. Sarvam AI. (2025). *Sarvam AI models on Hugging Face*. <https://huggingface.co/sarvamai>
- 381. SB Intuitions. (2025). *Sarashina models on Hugging Face*. <https://huggingface.co/sbintuitions>
- 382. Naver Cloud. (2024). *HyperCLOVA X technical report*. arXiv. <https://arxiv.org/abs/2404.01954>
- 383. Seger, E., Dreksler, N., Moulange, R., et al. (2023). *Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. Centre for the Governance of AI. <https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models>
- 384. Anthropic. (2025). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. <https://www.anthropic.com/news/disrupting-AI-espionage>
- 385. Partnership on AI. (2023). *PAI's guidance for safe foundation model deployment*. <https://partnershiponai.org/modeldeployment/>
- 386. International Energy Agency. (2025). *Energy and AI*. International Energy Agency. <https://www.iea.org/reports/energy-and-ai>

人工智能问题独立国际科学小组

组成与任务

按照通过《全球数字契约》和《未来契约》作出的承诺，联合国大会第79/325号决议在联合国内部设立了人工智能问题独立国际科学小组。

小组由 40 名独立专家组成。独立专家在人工智能及相关领域具有卓越专业知识，由大会任命，任期三年。小组做到了性别平衡，成员来自所有五个会员国区域组，专业领域包括人工智能核心技术、应用人工智能、安全与基础设施，以及人工智能政策、伦理和影响。

小组受权发布基于证据的科学评估报告，综合并分析与人工智能机遇、风险及影响相关的现有研究成果，为此每年编制一份与政策相关但不具强制性的简要报告，必要时可包含专题简报。其科学工作将遵循独立性、科学可信度与严谨性、多学科性及包容性参与等原则。

小组还受权在联合国人工智能治理问题全球对话上提交年度总结报告。小组通过为人工智能治理问题全球对话和更广泛的国际进程提供信息，使国际社会能够预见新出现的挑战，做出更加知情的治理决策，并为世界各地的政策制定者营造公平的信息环境。

由联合国数字和新兴技术办公室协调的小组秘书处为小组提供支持。

本初步报告的编写过程

小组成员于 2026 年 2 月获得任命，3 月召开首次会议。此后三个月内，小组为编写本报告开展了紧锣密鼓的工作。这一密集的科学交流和集体分析过程包括，在当选联合主席的主持下，召开一次为期三天的线下全体会议和 60 多次小组线上会议。

本初步基线报告为与外部专家进行更广泛磋商和编写后续专题简报和年度报告奠定了重要基础。

小组的科学独立性

小组成员以个人身份担任独立科学专家。每位小组成员在被指定为联合国特派专家时都已声明和承诺，履职过程中决不寻求或接受任何政府或其他来源的指示。《关于特派专家的地位、基本权利和义务的条例》(ST/SGB/2002/9)也包含有关特派专家行为和问责的规定。

Donors

The Panel Secretariat gratefully acknowledges the financial and in-kind contributions of the following governments and partners, without whom the Panel would not have been able to carry out its responsibilities:

Government of Germany
Government of Japan
Government of Spain
Omidyar Network Fund

Panel Secretariat

Coordinator

- Amandeep Singh Gill, United Nations Under-Secretary-General for Digital and Emerging Technologies

Editing and Drafting Support*

- Jiaee Cheong, United Nations University (UNU)
- Kevin Kohler, UNU
- Max Springer, UNU

Secretariat Coordination & Support

- Quintin Chou-Lambert, United Nations Office for Digital and Emerging Technologies (UN ODET)
- Rebakah Hayoung Woo, UN ODET
- Peppi Väänänen, UN ODET

Fundraising & Logistics

- Sebastian Frank, UN ODET
- Antonieta Loaiza, UN ODET

Communications

- Karoline Hassfurter, UN ODET
- Brian Shung Seun Lau, UN ODET
- Anamika Madhuraj, UN ODET

Rapporteurs

- Wernhard Berger, United Nations Industrial Development Organization
- Jin Cui, International Telecommunication Union (ITU)
- Tim Engelhardt, Office of the United Nations High Commissioner for Human Rights (UN OHCHR)
- Andrew Morritt, United Nations Department of Peace Operations
- Prateek Sibal, United Nations Educational, Scientific and Cultural Organization (UNESCO)
- Mariagrazia Squicciarini, UNESCO
- Oleksandra Vereschak, UNESCO
- Ana Gabriela Fernandez Vergara, ITU
- Li Zhou, UN OHCHR

* To ensure scientific independence of the Panel's work, Editing and Drafting Support personnel report to the Panel on substantive content/language of written outputs, while complying with coordination requirements, timelines, and sharing of produced content as part of the Panel Secretariat. For administrative purposes, they report to United Nations University.

The following United Nations Secretariat entities also supported the Panel Secretariat in the preparation of the report: Department for General Assembly and Conference Management, Department of Global Communications, and United Nations Geospatial (Office of Information and Communications Technology).