全球人工智能社会发展研究报告(2025)

Global Report on Artificial Intelligence and Social Development (2025)

上海市人工智能与社会发展研究会

2025 年 7 月

目 录

一、摘 要

随着生成式人工智能的爆发性演进，技术创新已实现由实验室向社会的全面渗透。当大模型从判别式迈向通用式，当算法从辅助工具演化为自主决策系统，技术已超越工具理性范畴，成为塑造经济秩序、权力结构与文明形态的元力量。进入 2025 年，人工智能的社会化应用呈现出深度赋能和系统性风险并存的双重图景。一方面，人工智能作为新质生产力的核心引擎，驱动产业智能化升级，孵化出一系列新业态、新场景、新模式，在医疗、教育、新闻、科研等领域释放出巨大潜力。另一方面，数据泄露、算法偏见、隐私侵犯、劳动替代、情感依赖、环境污染等问题，不断挑战社会伦理底线与全球治理体系韧性。

在此背景下，上海市人工智能与社会发展研究会推出《全球人工智能社会发展研究报告（2025）》，立足于技术演进和社会发展的双重逻辑，系统梳理近两年人工智能领域国内外前沿研究成果，综合研判全球人工智能最新研究动态，致力于构建具有前瞻性的研究范式与方法论体系，为后续研究提供可能的思路与方向。报告基于学术前瞻性和社会关切度的双重考量，遵循技术本体-社会辐射-治理反馈的逻辑链路，选取了社会科学领域人工智能研究的八大议题：算力变革、模型幻觉、数据让渡、劳动替代、智能鸿沟、大国博弈、人机关系、敏捷治理。报告试图通过深入剖析各议题的前沿进展、核心争议与发展趋势，为政策制定者、研究者及产业界提供兼具理论洞见与实践价值的参考框架。

With the explosive rise of generative artificial intelligence, technological innovation has achieved its massive penetration from laboratories into society. As discriminative models evolve to generative ones, and algorithms transform from auxiliary tools to autonomous agents, technology has transcended the realm of instrumental rationality, emerging as a meta-force that shapes economic order, power dynamics, and civilization paradigms. Entering 2025, the social application of artificial intelligence presents a dual landscape where in-depth empowerment and systemic risks coexist. On the one hand, artificial intelligence, as the core engine of new quality productivity, drives the intelligent upgrading of industries, incubating novel formats, scenarios, and models, and unleashes enormous potential in fields of healthcare, education, journalism, scientific research, etc. On the other hand, issues such as data leakage, algorithmic bias, privacy infringement, labor displacement, emotional dependence, and environmental pollution continuously challenge the bottom lines of social ethics and the resilience of the global governance system.

In this context, by synthesizing dual logics of technological evolution and social development, Global Report on Artificial Intelligence and Social Development (2025) systematically reviews the cutting-edge research results in the field of artificial intelligence over the past two years, comprehensively analyzes the latest research trends at home and abroad, and seeks to build a prospective research paradigm and methodology framework, providing possible ideas and directions for subsequent research. Balancing academic foresight with societal concerns, the report follows the logical chain of technical ontology — societal impact — governance feedback, and selects eight key research issues on artificial intelligence within social science: computing power transformation, model hallucination, data concession, labor displacement, intelligence divide, superpower competition, human-machine relationships, and agile governance. The report delves deeply into the frontier progress, core controversies, and development trends of each issue, aiming to provide a reference framework that combines theoretical insights and practical value for policymakers, scholars, and industry stakeholders.

二、全球人工智能研究动向

（一）算力革命：能源困局与绿色算力破局

作为人工智能的基底，算力是维持算法运行、激发数据价值的关键要素。**随着全球数据中心的持续扩容，技术进步与环境承载能力之间的张力日益凸显，**能源生态问题逐渐成为算力研究领域的显性议题。

国内外学者针对算力扩张引发的生态赤字展开了较为深入的分析，主要聚焦在能源消耗、碳排放、电子垃圾三个方面。首先，人工智能产业是能源密集型产业，算力发展对水电等自然资源存在高度依赖。算力基础设施与设备的核心部件是芯片，其研发制造、计算运行、水冷降温都需要庞大且稳定的水电资源支撑（戚凯、杨悦怡，2024）。进一步数据显示，生成式人工智能训练集群的能耗较传统计算负载高出七八倍，ChatGPT 搜索的耗电量亦达到网页搜索的五倍。值得注意的是，大模型迭代周期极短，企业常以周为单位推出新版本，导致前期训练投入的能耗沉淀为沉没成本，且新一代模型往往伴随参数规模扩张，进一步推高训练能耗（Adam Zewe，2025）。李坤泽（2025）亦指出，大模型的训练与推理需要不断增加参数，堆叠芯片，让数据中心的耗电量呈指数级增加。即便以 DeepSeek 为代表的新一代人工智能模型在训练阶段的能耗明显降低，但在不断扩张的巨量使用需求面前，它的总耗电量仍然十分惊人，且持续攀升。这种能源消耗压力已对部分国家的发展规划产生影响，一些国家在维护国内基本用电和发展人工智能之间只能被迫放弃或压缩人工智能的发展计划。例如，阿姆斯特丹、伦敦等多个城市已经因能耗高、电网负荷而停止了新数据中心的建设，新加坡、爱尔兰等国的人工智能产业也因电力供应不足而遭遇发展瓶颈。由此可见，人工智能竞争的尽头可能就是电力之争。其次，能源消耗的激增直接导致碳足迹扩大。研究人员指出，训练“单一大型语言深度学习模型”的二氧化碳排放量约达 300 吨，而普通人年均二氧化碳排放量仅约 5 吨（吴红、姜惠，2025）。其中，推理

环节碳排放占比尤为显著，大型语言模型应用所覆盖的庞大用户基数则进一步放大了这一环境影响（Eddie Zhang 等，2024）。最后，人工智能系统的迭代升级还持续产生电子废物。世界经济论坛（WEF）预测，2050 年电子废物总量将突破 1.2 亿吨。这类废物中含有铅、汞和镉等危险化学物质，若缺乏安全处理，极易污染土壤和水源，长此以往将对环境构成严重挑战。

与此同时，**学界对人工智能与环境的关系认知还存在另一重维度，部分学者强调其在推动环境可持续发展方面的潜在价值，绿色算力由此成为学界探讨技术减排路径的核心议题。对于人工智能与环境这一话题，国内涌现出较多关于二者相关性的量化研究成果，且结论基本一致，均呈现出显著正向效应。**具体而言，张秀武和沈洋（2025）基于减污降碳联合减排收益的视角，运用 2007-2022 年 269 个地级及以上城市的面板数据，揭示出人工智能可通过替代效应、技术效应与智能决策，提升减污降碳协同效益。与之形成呼应的是，赵雨涵等（2025）基于 2011-2021 年中国 277 个地级市的数据，运用门槛效应模型、中介效应模型与空间杜宾模型等方法，实证发现人工智能技术创新对城市碳减排具有积极作用，但该效应随新兴技术应用程度提升呈边际递减趋势。许潇丹和惠宁（2024）的研究则聚焦工业领域，基于 2011-2021 年中国内地省级面板数据证实，人工智能发展能显著降低工业碳排放强度，且其对工业绿色低碳发展的影响存在区域异质性，即西部地区最强、东部地区最弱。与此同时，**部分学者进一步剖析了人工智能助力节能减排的潜在路径机制，其核心路径可归纳为两大方向：一是通过驱动传统产业绿色转型间接减少碳排放，二是直接赋能能源环境行业发展以增强减排效能。**前者的大致观点是，生成式人工智能通过“劳动者—劳动资料—劳动对象—要素组合”多层级要素的组合优化，提高传统产业的资源利用率与全要素生产率，从而减少不必要经济活动带来的环境影响，为实现绿色生产力水平提升增加新能量（刘宏伟、马酉儒，2025）。后者则认为，人工智能在可再生能源发展应用领域潜力巨大。Nitin Liladhar Rane（2023）指出，以 ChatGPT 为代表的大型语言模型在太阳能、风能、生物质能等 14 个能源领域均有显著应用价值。例如，

通过数据分析与建模优化光伏布局，提升风速预测精度（太阳能）；支持水库调度管理和河流流量预测，实现发电量最优化（水力发电）；助力风机控制、维护排程及风速预测（风能）；辅助原料筛选、工艺优化及减排方案制定（生物质能）。此外，**针对算力环节的内生性能耗与污染问题，部分学者对其突破路径展开研究，主要为以下三个方面：一是模型绿色升级。**在硬件上采用高效能、低功耗的设备，如低功耗处理器、高效能存储器、绿色数据中心等；在软件上优化算法、降低软件复杂度、提高软件运行效率（赵勇，2024）。**二是计算资源协同。**以我国为例，我国算力资源整体呈现“东部不足、西部过剩”的不均衡局面，且数据中心间协同性不足，“数据中心孤岛”“云孤岛”等现象频频出现（陈晓红等，2023）。加之模型同质化竞争严重，计算内容高度重复，难以实现计算资源的帕累托最优，因此亟需对计算资源进行协调整合（龙海泉等，2024）。**三是核能突破。**结合产业实践，各大数字巨头将可控核聚变视为解决人工智能耗电问题的终极方案，纷纷开启核能布局，以期通过颠覆性技术突破能源限制（李坤泽，2025）。

综上，社会科学界对算力扩张的环境影响存在分歧：部分学者关注其直接产生的能源消耗与生态赤字，另一部分则强调其赋能产业绿色转型的间接减排潜力。**增排与减排的具体量级，以及算力的净环境效益难以评估。但从长远来看，随着人工智能驱动核能等新型能源的革新应用以及计算资源的协同优化，未来，算力有望突破能耗与环境的制约，实现真正的可持续发展。**

（二）模型幻觉：谬误和创见共存

模型幻觉是除算法偏见外，算法研究领域的又一重要话题。与算法偏见不同的是，**当下学界对于模型幻觉的探讨尚处于初始阶段，研究潜力巨大。**部分学者致力于构建相关概念体系，对模型幻觉展开了概念界定和表征分类的基础性研究。

目前，学界对人工智能“幻觉”（Hallucinations）的定义尚未达成广泛共识，不同定义同时存在甚至互相矛盾。“幻觉”原本是神经科学和心理学的概念，

意指感觉接受器对客观经验不准确的主观再现（林曦，2025）。人工智能领域对这一术语的引入最早可追溯至 2000 年的计算机视觉研究，指的是一种图像修复与合成应用领域的建设性技术，旨在提高图像分辨率。可见，**早期的“幻觉”是一个带有积极意味的技术术语。随着自然语言技术的迭代更新，“幻觉”逐渐发展成为一个负面词汇，**通常指模型生成的内容对提供的源内容无意义或不忠实，即模型生成的文本不忠实于信息源或者与现实世界的事实不符（Ziwei Ji 等，2024）。对此，Østergaard 和 Nielbo（2023）提出了两点关键性质疑：第一，**人工智能的“幻觉”隐喻属于术语误用，因为人工智能不具备感官知觉，其认知错误源于数据缺陷和提示工程局限，而非刺激缺失。**第二，**这种隐喻存在高度污名化效应，将人工智能的负面问题与精神疾病的特定症状相关联，可能破坏精神病学与心理健康领域去污名化的诸多努力。**因此，部分学者致力于探索更加准确恰当的替代性术语，用于描述人工智能的这一特性。Negar Maleki 等（2024）根据不同场景的模型“幻觉”，构建了不同的“幻觉”概念矩阵。包括虚构（Confabulation）；错觉（Delusions）；概率性复述（Stochastic Parroting）；事实错误（Factual Errors）；编造（Fabrication）；错误、过失、谬误（Mistakes, Blunders. False-hoods）；草率概括、错误类比、假两难推理（Hasty Generalizations, False Analogy, False Dilemma）等。他提出，**确立人工智能幻觉的严谨定义至关重要，有助于提升幻觉研究的聚焦度、精确度、清晰度和连贯性，避免跨学科研究中的潜在混淆和歧义风险。**除了概念探索外，部分学者从不同维度出发，对模型幻觉展开了类型学研究。方师师和唐巧盈（2023）从“错误事实”与“错误认知”的双重维度出发，提取出了存在事实错误、逻辑错误、推理错误、编程错误、文本输出、过度拟合、综合问题 7 大幻觉类别。Yujie Sun 等（2024）则下沉归纳出 8 种一级错误类型和 31 种二级错误类型，其中一级错误类型与方师师的分类基本一致。

在概念体系搭建的基础上，**学者们围绕模型幻觉的成因、评估与对抗路径展开多维度研究。**例如，刘泽垣（2025）分别从数据层、模型层和应用层对模型幻觉的成因、评估方法与缓解路径进行了较为全面的梳理与分析。其中，数据层的污染、重复、缺失，以及数据标注错误是诱发模型幻觉的源头性因素。模型层的算法运作黑箱特性则进一步加剧了幻觉问题，具体体现在编码解码缺陷、预训练

知识偏好、曝光偏差及模型知识更新局限。应用层面下游任务的领域专业化可能导致模型泛化能力下降；模型的同质化趋势（即不同任务采用相似预训练模型）易导致对特定输入过度敏感，且模型固有缺陷会被广泛继承并放大；多模态化处理增加了模型复杂度和幻觉风险；提示工程虽提升了响应灵活性，但也可能导致模型过度依赖提示，其指令微调技术与思维链技术的应用局限也可能进一步深化幻觉。基于此，可进行数据清洗和数据增强构建真实数据集并修正幻觉数据集，以提高数据质量；优化模型初始结构、模型训练与微调过程、模型后处理方法，以减少计算过程中的误差；设计相关提示或者对齐指令，以减少特定任务幻觉的产生。

上述类似研究大都基于这样一个共识性前提，即模型幻觉是有害的，但也有部分学者对这一前提表示质疑，开始对模型幻觉的积极意义和价值多面性展开批判性分析。胡泳和王昱昊（2025）提出，人工智能幻觉存在一体两面，机器创造力过于注重新颖性可能会导致产生原创但不准确的回答，而过于注重实用性则可能导致无效的死记硬背的回应。杨雅等（2024）也指出，虽然生成式人工智能存在“一本正经胡说八道”的不准确推理情况，但大模型理解与解释世界的能力依然值得期待，这也是培育大模型想象力的一种过渡阶段，而并非完全将其囿于人类的资源检索库的角色。因此，开发真实可靠的大模型时， 如何平衡创造力与真实性是处理幻觉问题的巨大挑战。大模型的幻觉问题具有两面性，一方面， 幻觉被视为模型的缺陷，需要通过技术手段予以纾解，不正确或虚构的信息可能导致严重后果，破坏信任和可靠性。但与此同时，幻觉也可以带来创新性和创造性，生成出人意料的、新颖的想法，从而激发设计灵感，协助探索多种可能的解决方案，打破传统思维定式的束缚。因此，评估和利用幻觉现象与缓解幻觉问题同样重要，需要给予均衡的关注。总体来讲，**当前的研究工作倾向于减少幻觉， 一定程度上忽略了其创造性，在未来的工作中， 通过更细致的幻觉分类兼顾事实准确性和知识灵活性具有重要的现实意义**（刘泽垣，2025）。

综上，模型幻觉尚属充满探索空间的研究蓝海，**学界仍处于术语厘清与类型划分的基础性研究阶段，概念谱系与理论架构的搭建远未完成。**值得注意的是，**当前研究范式正初步显露出从单向有害论向辩证价值论转变的苗头倾向**，幻觉

下的技术缺陷与创新潜能协同共生，并存在持续张力。**未来的研究路径需超越单纯的“纾困”逻辑，转而构建能够兼顾事实严谨性与知识创造性的动态平衡机制，**发展精细化、场景化的幻觉分类、评估与引导体系，推动可信人工智能发展。

（三）数据让渡：便利增益与隐私风险

随着生成式人工智能的发展，数据让渡以更为深刻的方式嵌入用户交互过程中。除了依赖已有数据库进行大模型训练外，用户会向交互智能体分享大量数据，这些数据被收集后可能会被继续用于模型训练。这种主动披露的数据让渡超越了“知情同意”制度的范畴，让隐私可以在用户认知模糊的状态下被非自主让渡（孙国烨等，2025）。**数据让渡作为人工智能系统运作的底层逻辑，其引发的隐私悖论是近年来学界关注的重要议题。**

一方面，用户的数据隐私自愿让渡是智能应用程序为其提供个性化、场景化服务的基本前提。人工智能技术本质上天然依赖于海量数据（Yadav Neel，2023）。在万物互联的大数据时代，个人信息数据成为社会发展的重要燃料，个人提交的信息数据越多，基于海量个人信息的大数据计算越准确，个人就能获得更加精细化的私人服务，组织就能形成更加优质化的产出效率（董淑芬、李志祥，2023）。智能应用程序正是基于对海量用户数据的深度挖掘与计算，方能推出高质量的个性化适配服务。通过挖掘用户信息，应用得以勾勒出用户所处的线上线下场景，并通过自动化算法进行场景与需求的最大匹配，为用户实时、实地提供并更新适配服务。因此，**隐私让渡是用户获取便利的前提基础，智能应用的数据收集行为已经成为大数据社会正常运行不可缺少的重要基底。**以智能监测设备为例，各种具身设备的兴起掀起了“量化自我”的热潮，置身其中的人们佩戴着各式各样的技术装置，通过数据来认识自己。在智能设备的使用过程中，量化自我从多样的情境中形成了“让渡有益”的行为理念，从而将数据监控化用为自我跟踪。受这种数字化生存惯习的影响，量化自我群体的隐私感经历着技术社会的“脱

敏”，让渡数据变得习以为常。“后隐私”甚至承诺，“我们每个人的信息被数据公司记录得越多，我们就存在得越多，我们对自己的了解就越多”（俞立根、顾理平，2024）。

另一方面，智能应用对数据的挖掘、储存、计算一定程度上又无可避免地构成了对用户数据隐私的侵犯。由于数据具有可完全复制性，隐私让渡造成的主观损失感远远小于物质性存在物让渡，这意味着隐私信息具有比一般物质性存在物更高的可让渡性（董淑芬、李志祥，2023）。在数据采集阶段，人工智能训练数据面临过度采集和窃取滥用风险，用户可能在不知情的情况下被过度收集数据（刘琳璘，2024）。在数据预处理阶段，面对海量异构数据，一般的数据预处理操作难以做到精准高效，易埋下个人信息未彻底脱敏的安全隐患（王梅艺，2024）。在使用阶段，神经卷积模型相较于传统算法模型对于各种数据要素的分析更加深入，能够发掘出海量个人信息中潜藏的信息（刘艳红，2023）。因此，在人工智能运行的各环节，均存在个人知情权、隐私权和数据权益遭受侵犯的风险。Kelly D. Martin 和 Johanna Zimmermann（2024）采用人工智能生态系统视角，提出了一个过程-结果框架，对人工智能技术进行了细致分类，确定了不同类型的人工智能如何影响隐私决策。此外，**算法的自动化计算黑箱一定程度上会造成对用户“合成隐私数据”的侵犯**。除了“自然隐私”之外，大数据具有构建“合成型隐私”的功能，即通过数据挖掘技术将人们在网络上留存的数字化痕迹进行有规律整合而成的隐私（林爱珺、章梦天，2024）。由于机器学习的内部不可见性，数据收集者无法告知用户尚未确定的数据用途，因为收集者也无法预测其获取的初始信息在后来的环节中如何被若干次加工。因此，用户、商业公司，甚至模型开发者均无法完全洞悉算法内部的计算逻辑，技术人员难以向公众进行逻辑解释，用户也无法理解自身数据是如何被处理与操作的。

此外，**尽管知情同意原则已经成为对待隐私让渡问题的社会共识，大多数知名企业都出台了自己的隐私规范政策，但是在实践中，数据企业将这种同意**

机制实际异化设置为一种近乎默认的机制，绝大多数用户并不知道自己“同意”就已经构成了“同意”。这种“默示同意”机制或者“选择性拒绝”模式在 web3.0 时代很难起到保护个人隐私的作用。一方面，大数据平台及软件开发商提供的知情同意书往往过于复杂，作为用户的隐私主体大多没有时间和精力进行全面而仔细的阅读，这就意味着知情只是一种可能而不是现实。另一方面，大数据平台及软件开发商都实行“同意才可使用”的政策，即如果不同意其隐私政策就不能使用其软件或平台，在这一政策下，大多数人为了获得软件或平台的便利而不得不同意其隐私政策，这就意味着同意只是一种被胁迫的同意（董淑芬、李志祥，2023）。可见，个人用户在强大的算法技术面前处于绝对弱势地位。算法隐蔽层的存在使得平台公司与用户之间存在一条技术与认知鸿沟，用户个人对隐私权利保护没有明确的认知，智能算法技术对于信息的储存与处理规则属于技术机密，知情同意幕后如何操作用户不得而知，对隐私进行了何种侵犯用户更是无法感知。最终在知情同意困境下，用户对于平台条款的同意使用属于“被迫”与“无奈”状态，知情同意原则实际上将用户置于权利被迫让渡的困境中（张涵等，2024）。基于此，部分学者开始探索更加科学的知情同意模式。Gary Burkhardt（2023）等基于自主授权模型（AA）与计划行为理论（TPB），构建了一个植根于自主伦理的知情同意模型，即一种优先考虑用户利益并支持伦理信息管理和营销实践的同意征求模式。

综上，随着生成式人工智能的发展，数据让渡逐渐从被动同意转向主动披露。便利增益与隐私风险并存，一方面，用户以隐私让渡换取精准场景化服务，另一方面，智能应用对数据的挖掘、存储与计算又不可避免地构成隐私侵犯。知情同意的实践异化进一步加剧了潜在的隐私侵害。如何确保数据让渡的正当性、自主性、可控性是各界亟需突破的课题。

（四）劳动转型：替代冲击与技能重构

人工智能技术的迅猛发展正以前所未有的深度和广度重塑全球劳动力市场

格局，引发了各界关于人类劳动价值与未来工作形态的深层思考，劳动替代由此成为近年来人工智能社会科学研究的核心热点。

围绕人工智能对劳动市场的影响，学界形成了“创造”与“替代”的两大对立观点。支持创造论的学者认为，人工智能正孵化出一系列新场景、新业态，不仅不会导致大规模失业，反而会创造一批全新的就业机会（易宪容、陈颖颖，2024）。此外，智能设备的使用、维护和培训也需要大量专业人才参与（徐政、吴晓亮，2025）。事实上，当前人工智能在商业领域的应用仍处于相对早期阶段，短期内对就业产生的实质性影响较为有限（Kathryn Bonney 等，2024）。联合国的测算也印证了这一点，全球最多只有 2.3%的岗位有可能完全自动化（UN，2025）。因此，人工智能取代人类工作的可能性并不大。相反，人工智能将人从重复性、低技能要求的岗位中解放出来，从事创新性、技能密集型、情感交互性的工作，实现了对劳动力资源的重新配置与优化，有助于提高劳动效率，激发个人创造力。**支持替代论的学者认为，**采用人工智能将导致大规模的劳动替代与结构性失业。高盛的研究数据显示，在人工智能的冲击下，3 亿全职岗位面临替代风险，美国约有三分之二的职业可以通过人工智能实现部分自动化（高盛，2023）。胡晶晶和程承坪（2025）则进一步指出，机器天然地会对就业产生替代效应，但替代的范围和程度在技术的不同发展阶段不尽相同。随着人工智能技术由专用向通用跨越，就业替代效应也在不断强化与拓展，可替代的劳动形式由物质劳动延伸至非物质劳动，工作类型由常规认知领域到非常规认知领域。

“创造”与“替代”并非均匀分布，而是呈现出显著的技能偏向性（IEDC，2025），**由此引发的就业极化现象成为学者们密切关注的研究热点。**具体表现为：低技能岗位面临取代风险，高技能岗位存在创造机遇，但其创造的数量和速度远不足以弥补被替代的低技能岗位总量，导致整体就业岗位的净减少（孟现玉，2024）。除了岗位的非对称创替之外，劳动者的非对称流动也会对极化效应产生系统性影响。黄旭和董志强（2024）的研究创新性地考虑到了劳动者技能身份的可转变性，他们提出，**人工智能对就业岗位的冲击并不是简单的局部静态结果，而是一个复杂的全局动态过程**。例如，在人工智能冲击下，中等技能工人有三种潜在的应对之策： 一是在现有岗位上经过培训提高劳动生产率， 二是下沉转化

为低技能工人，三是经过培训成为高技能工人。不同策略将对就业极化与工资极化现象产生差异性影响。这种非对称的劳动市场结构将加剧收入分配的两极分化，推动其持续向高收入人群倾斜。实证结果显示，人工智能应用水平上升 1%，会导致劳动收入份额下降 0.054 个单位。但人力资本水平高、经济发达的地区人工智能对劳动收入份额的降低作用不显著（王丽媛、李繁荣，2024）。与此同时，技术进步正以前所未有的速度压缩着人力资本的“保质期”，随着人工智能技术的快速迭代，个人知识技能更新周期急剧缩短，人力资本的折旧速度持续加快，这将倒逼整个社会进入技术转型的加速期。**整体而言，替代效应与创造效应同步发生，劳动市场的总量收缩风险与质量提升机遇并存，呈现出技能导向动态转型的整体趋势。**

然而，**随着人工智能技术从判别式向通用式的跨越演变，劳动市场的技能偏向性导向引发了新的理论争议。**当传统自动化风险聚焦程式化体力任务时，生成式人工智能的颠覆性在于威胁认知交互型技能。Molly Kinder 等（2024）认为，**生成式人工智能的发展标志着与以往“技能偏向”范式的彻底决裂。**除非机器人技术取得突破，生成式人工智能不太可能对体力劳动造成威胁，相反，它擅长编程、预测、写作、说服、沟通，具备了替代有一定技能门槛工作的交互特质。这种特性导致自动化风险格局发生逆转，如今大多数面临生成式人工智能替代威胁的行业，正是几年前自动化替代风险排名中垫底的行业。该范式颠覆在职业与技能层面获得双重实证支持。研究表明，软件开发岗位因低具身性需求，71%的专业技能暴露于替代风险；而护理岗位依赖肢体操作与情感支持，仅 32.9%的技能可能被自动化（Ryan Stenvick，2024）。安永全球分析显示，数学与编程技能在生成式人工智能威胁暴露指数中高居首位，而主动倾听、学习能力等人际互动与情感理解技能稳居安全区（Gregory Daco，2024）。在人工智能向通用式迈进的过程中，新型技能壁垒得以形成。当高技能门槛的认知型工作沦为替代重灾区时，需复杂社会情感协调与创造性思维的能力反成人类最后的防御堡垒。**这宣告“低技能=高替代”工业逻辑的终结，标志着人机协作准则向“情感-创造”双维度的历史性迁移。**

综上，**人工智能对劳动力市场的重塑已突破“创造-替代”的二元框架，呈**

现出结构性极化与范式逆转的双重特征。伴随生成式人工智能的技术升级，传统的“低技能=高替代” 逻辑被彻底打破，催生出以社会情感互动和创造性思维为核心的新型人机协作范式，就业市场的技能需求结构与转型路径因此被深度重塑。这种变革并非技术对人类的“替代竞赛”，而是劳动力价值维度的重新锚定。当高技能认知工作面临智能化冲击时，人类的独特性正转向更复杂的情感联结与突破性创新。**未来的关键在于构建“技术适应力+人文内核”双驱动的劳动力发展体系，让劳动力转型既能跟上技术变革的节奏，又能守住人类不可替代的价值。**

（五）智能鸿沟：技术红利分配的结构性失衡

学界关于人工智能引发的智能鸿沟和社会公平的讨论方兴未艾。人工智能正在重塑全球经济与社会结构，但其带来的技术红利并未实现公平分配，反而加剧了既有的数字鸿沟。与传统数字鸿沟不同，智能鸿沟不仅涉及 ICT 技术接入与使用的差异，更体现在算力资源、算法理解、数据掌控、治理能力等深层次能力的分化。已有研究主要是从智能鸿沟的表征与成因切入，对人工智能是否加剧不平等，何以产生鸿沟，以及其弥合路径展开分析。从不同主体维度看，**智能鸿沟在国家层面体现为南北差异与技术自主，企业层面表现为巨头垄断与竞争失衡，个人层面则反映为智能素养与社会公平。**

从国家层面看，人工智能技术的发展进一步加深了国家之间的差距，已成为学界共识。**先进人工智能技术的研发与部署具有高度资源密集型特征，尤其依赖巨大的能源消耗和先进的基础设施，往往只有发达国家有能力支撑**。囿于基础设施滞后、人才储备匮乏以及产业生态链不完善等本土因素，全球南方难以形成优渥的创新土壤（陈菲、蒲文杰，2025）。全球计算资源分布不均、国家间的技术准入壁垒，以及不同版本工具的访问限制，共同造成了发达国家与发展中国家在技术接入难度和质量上的双重差距。这导致人工智能领域的专业知识与核心利益日益集中于发达国家，而发展中国家则面临追赶困境（方兴东、钟祥铭，2024）。迄今为止，最大的受益者是美国、中国和欧盟。根据牛津大学研究人员汇编的数据，这些地区拥有全球一半以上最强大的数据中心，用于开发复杂的人工智能系统。其中，美国和中国占据绝对领导地位，运营着 90%以上的人工智能数据中心。

非洲和南美洲几乎没有人工智能计算中心，超过 150 个国家完全没有（Zoe Hawkins，2025）。从智能体的角度看，以 Open AI 的 Chat GPT 为代表的主流人工智能系统，在英语和中文上的表现更为熟练精准，而这两种语言恰恰是算力高度集中国家的通用语言（Adam Satariano，2025）。由于非洲语料库规模有限且缺乏代表性，ChatGPT 在非洲地区的应用推广受到限制（Gregory Gondwe，2023）。另一方面，**国家间的技术鸿沟通过全球供应链分工传导至产业领域，加剧了发展失衡**。全球南方国家面临长期被锁定在产业链最低附加值环节的风险，并高度依赖发达国家上游产业链，难以实现技术战略自主（秦北辰，2024）。此外，部分国家为实现人工智能领导地位，以国家安全为由实施脱钩断链与技术封锁，阻止技术向外流出，大大降低了数据自由流通和框架互操作的可能性，极大压缩了智能鸿沟的弥合空间。

从企业层面看，大部分学者认为人工智能将进一步导致竞争失衡与寡头垄断。当下，**由于技术和资本的双重壁垒，人工智能基础模型市场高度集中，主要被极少数大型科技企业所垄断**（雷昊楠，2024）。张严（2024）提出，生成式人工智能的广泛应用将引发模型共谋、数据剥削、歧视性许可和生态封锁等多重垄断风险，同时在“资本-技术”和“算法-数据”的双循环中形成结构壁垒和生态垄断。相关头部企业通过数据壁垒、拒绝开放和剥削性滥用实现对数据的非对称掌控，通过算法共谋、算法黑箱、系统封锁和自我优待完成对算法的隐蔽性支配，通过关键基础设施歧视性待遇和底层技术不兼容设计进行算力的排他性垄断。但自 DeepSake 发布以来，**也有部分学者表达了用开源生态对冲垄断格局的新思考。**DeepSake 一是通过低成本训练推理，打破高端算力的垄断封锁，降低研发应用门槛；二是通过全栈、全系列的开源开放，支持按需自主部署，普惠各行各业，为人工智能开放生态系统提供了中国经验（武延军，2025）。算法的公共化生产可以实现对开源协作模式的形塑，促进算法资源的自由流通与共享，降低算法开发的门槛和成本，实现算法知识的共享创新（吴鼎铭、汪荣燊，2025）。与此同时，**部分学者开始对“开源垄断”话题展开探索性研究。**李兆轩（2025）指出，目前学界对于开源垄断问题的深入探讨尚显不足。尽管开源社区秉持免费、共享与创新的核心理念，但企业可以通过开源项目搭建产品链条进行闭源盈利。此类“开源引流-闭源变现”的模式并不能缓解当下的巨头垄断格局。

从个人层面看，**智能鸿沟体现在人工智能对不同群体的差异性赋能上。**人工智能技术的应用究竟是普惠性赋能，助力弱势群体发展并促进社会公平；抑或因其资源倾斜性分配，进一步固化既存差距并催生新型数字鸿沟？（Christoph Lutz，2024）学者们围绕这一问题展开了激烈探讨。从人口统计学维度看，智能鸿沟表现为代际、性别、收入、种族、城乡等维度之间的分化，学者纷纷对此展开表征与因应研究。以性别智能鸿沟为例，随着智能技术的广泛应用，性别歧视开始溢出传统范畴并向线上空间蔓延，广泛存在于人脸识别、搜索推荐、智能招聘、金融授信等场景中（徐偲骕，2024）。伯克利哈斯公平、性别和领导力研究中心的一项研究分析了不同行业的 133 个人工智能系统，发现其中约 44%的系统存在性别偏见，25%的系统同时表现出性别和种族偏见（UN，2024）。对此，学者们进一步对鸿沟的形成机理展开了深层解析。人工智能的性别鸿沟主要源于数据源头中的性别不平等。**数据作为现实世界的映射，其采集和标注过程往往会客观复制现实中的偏见与结构性失衡。当这些带有偏差的数据被用于训练算法时，算法机制不仅未能纠正既有不平等，反而会系统性地识别、放大甚至固化这些模式，最终导致性别鸿沟在人工智能应用中被进一步显化和极化。**除了探究智能鸿沟的原因之外，部分学者对生成式人工智能对应的智能素养展开了探讨。彭兰（2024）指出，**智能传播时代人们需要掌握新的能力，即对智能生成内容的辨识能力，以及与智能体的提示交互能力。**个体可以利用大模型实现知识生产的积累，由此产生一种去中心化的知识赋权效应。

综上，智能鸿沟研究揭示了技术红利分配的结构性失衡。国家间算力地缘政治失衡将技术弱势国锁定于产业链末端，企业端数据-算法-算力的结构性垄断形成市场权力闭环，个体层面智能素养分化则进一步加剧社会不平等。这些维度共同指向技术赋权悖论——人工智能在提升生产效率的同时，系统性地强化了既有不平等机制。**唯当技术红利分配从零和博弈转向帕累托改进，方能使人工智能从阶层固化器转化为社会包容性增长引擎。**

（六）人机主体性：异化与共生

人工智能技术的迅猛发展正在深刻重构人机关系，并引发关于主体性

（Subjectivity）的哲学伦理反思。**随着“机器的人化”与“人的机器化”双重进程的同步演进，学界对技术冲击下“人的主体性”和“机器的主体性”展开激烈讨论。**部分学者依托海德格尔、马克思、笛卡尔、哈贝马斯、梅洛-庞蒂等的经典理论，展开主体性哲学层面的思考，另有部分学者则从法律视角切入，对人工智能的法律主体地位展开探讨。

针对技术冲击下人的主体性危机，学界主要聚焦于危机的多重表征与纾解之道展开研究。危机的本质在于技术理性对价值理性的僭越，以及资本逻辑与技术逻辑的共谋对人性本质的异化。例如，凌应生（2025）指出，资本与技术的双重逻辑已深度介入并重构个体的认知模式、行为方式、社会关系及自我认同。这种介入与重构对人的主体性形成的冲击集中体现在四个维度：独立人格被侵蚀（个性消解与技术依赖）；自主性和创造性被削弱（技术替代与思维惰性）；交往异化加剧（虚拟交往与情感疏离）；社会角色边缘化（技术失业与身份认同危机）。因此，理应遵循人是实践主体原则、坚持以人的价值理性引导技术理性、完善人机共生的伦理规则、以制度规约资本与技术的逐利本性，从而实现人主体性的重塑与复归。李戈和何玉芳（2025）的研究则从历史维度剖析了主体性的演化轨迹，指出人工智能时代的主体性困境主要表现为：人的自由意志遭遇削减、人的自我异化倾向加剧，以及人的隐私安全不复存在三个方面。基于此，他构建了包含自我意识提升、技术标准规范化和治理体系完善三个层面的应对框架。

对于人工智能的主体性争议，不同于技术界的积极想象，社科领域的大部分学者并不认同其具有完全的主体性。尽管模拟技术的进步让人难以摆脱机器作为类人主体的幻觉，但其始终无法产生意识（Thomas Fuchs，2024）。从马克思主义观点来看，“智能主体性”并不存在，主体性为人所特有，是人作为存在主体的现实性、作为价值主体的生命性、作为需求主体的能动性、作为实践主体的创造性的集合（彭姝，2025）。主体性是人工智能无法复制的，不是目前的人工智能无法获得，而是人工智能原则上就无法获得，即人工智能不可能突破发展的奇点而全方位地超越人类。工智能本质上就是对人类智能的模仿，而且是对人类智能外部功能的模仿，而不是对人类智能的复制或再造。人工智能无论怎样发展都不会改变其“模仿”的本质，不会拥有人类智能的主体性（Georg Northoff、

Steven S. Gouveia，2024；阎孟伟，2024）。人工智能在认知科学的五个层级（神经层级、心理层级、语言层级、思维层级、文化层级）均与人类存在本质区别，因此从根源上不具备主体性（甄航，2024）。**也有学者指出，传统主客体二元对立的认识论已无法解释生成式人工智能对主体性的冲击。**人机关系经历了从初期的“被动工具模式”到如今“智能协作模式”的发展（孟芳，2025），**已经由传统的“主客体”支配关系转向一种深度“合作”关系**（李洋、薛澜，2025）。**部分学者引入行动者网络理论（ANT），将技术作为网络中的行动者之一。**Helena Lindgren（2024）基于活动理论框架和社会历史理论，对人工智能的社会角色进行了纵向梳理，并发展了一套社会技术关系框架，旨在为理解新型人机关系提供参考。殷杰（2024）提出，生成式人工智能在与人类交互中，表现出了与以往人工智能截然不同的自主交互能力，呈现出一种**新型的主体性特征，即交互主体性**。这种交互呈现的新型主体性特征，是以人类语料作为基础驱动力、在交互状态下直接生成的自主行动能力，既不是源自生物本能的生存和自我维持，也不同于人类基于意向性的主体性，更不是传统人工智能所依赖的由人类设定的机械性功能。这种主体性在交互中可能表现为更敏锐的感知能力、更强大的推理能力和更适切的行动能力。然而，这种类人性和非人性的主体性特征，只有在与人类主体的广泛交互中才可能形成和显现，因此不必对技术对人的主体性冲击太过悲观。

此外，当前关于人工智能法律主体地位的探讨，即人工智能能否成为法律意义上的行为主体，以享有相关权利和承担相应责任，**形成了两种截然相反的观点：否定说和肯定说**（梅夏英，2025）。**否定说目前是大多数人支持的主流观点，**该说认为人工智能目前成为法律主体并不现实，其主要理由在于：人工智能目前并不具有人类的理性和自我意识，亦不具有自身财产，且不赋予其主体地位并不影响法律解决与人工智能相关的法律问题（孙良国，2024；袁康，2024）。**肯定说则认为，**对于一种新型民事主体的承认，法律并不必然要求主体具有人类理性，如法人（尤其是财团法人）就体现为财产的聚合；亦不要求责任财产的存在，如动物、胎儿和非法人团体等民法上的“有限主体”就与财产无直接关系（石冠斌，2024）。确认人工智能领域的法律主体是 Agent（行为体），有助于深入推进人工智能与法律的交叉研究。Agent 通常有弱概念和强概念两种用法，在弱概念之下的行为体有自主性、社会性、反应性、主动性的特征；在强概念之下的行为体

通常指一个计算机系统，除了弱概念的上述特性之外，要么是概念化的，要么是使用更常用于人类的概念来实现的，有一种赋予行为体类似人类特性的方式是在视觉上展示它们，所谓“人形机器人”就是一例（寿步，2023）。

综上，人工智能的快速发展正在重构人机关系。在技术与资本的共谋下，人类主体性面临异化危机，而机器主体性争议则打破了“主体-客体”的二元对立，“交互主体性”等概念为理解智能时代的主体形态提供了新视角。未来，需在理论建构与实践探索的过程中，探寻人机共生的主体性范式。

（七）大国博弈：权力重构与生态重塑

科技的每次重大革新都会带来国际权力格局的变动以及政治生态的重塑。**人工智能作为第四次科技革命中的颠覆性战略技术，是大国博弈的焦点，也是学界关注的经典议题。**

人工智能正深刻重塑国际权力格局，部分学者围绕技术的国家权力形塑路径展开分析。在理论层面，无论是在西方现实主义框架下，技术被视为从属于国家物质权力的核心要素；在新自由主义视角中，技术被定义为外生的“中立工具”；在建构主义逻辑里，技术被解读为嵌入社会规范的有机组成部分；还是在国际政治经济学的分析维度下，技术被视作塑造结构性权力的关键变量，技术始终是影响国际权力格局演变的核心因素（余南平，2025）。在国家视角，部分学者从关系连带（将其他领域的战略关系迁移到人工智能领域，利用人工智能领域权力实现连带影响）、资源异质（各国要素禀赋不同、外部环境不同，容易产生矛盾和竞争关系）、战略警觉（一国认为另外一国制定了可操作的、构成威胁的技术策略）三个关键因素出发，揭示了国家间权力互动的复杂性和多样性（蔡翠红，2024）。在企业视角，有学者提出人工智能通过重塑知识结构、介入生产、安全与金融领域，推动国际权力向技术优势国与跨国企业集中（部彦君、许开轶，2023）。数字巨头掌控核心技术、配套资源及治理规则，形成“工具性、话语性与制度性”三重技术权力，加剧“全球南方”国家的技术依附，并且输出隐性价值观实现跨国渗透（孙志伟、殷浩铖，2025）。为获取国家权力，各国纷纷展开科技博弈。传统的国家综合实力竞争正在悄然演化为以数据和算法为核心的人工智能竞争。

在竞争态势上，国家竞争焦点不再局限于人工智能技术研发创新，“数据-算力-模型-应用-规则”全链条要素均成为国际竞争新的角力点。相应地，学者们的研究焦点也从关键技术博弈，拓展至数据、算力和模型竞赛、规则话语权争夺等竞争泛领域（Jared Cohen，2024）。

作为典型的军民两用技术，人工智能在国防军事领域的广泛应用，引发了学界对其国际安全影响的深入探讨。人工智能加剧了国家间的军备竞赛与安全困境，重塑了战争形态和政治生态，是影响全球安全格局的重要因素。学界关于人工智能与安全的研究主要聚焦在**物理层的武器赋能与认知层的信息操控**两个层面。**一是关于人工智能带来的网络空间武器化的探讨。**人工智能赋能国际网络安全攻防博弈，网络攻击智能化升级明显，攻击效率显著提升，呈现自动化、自适应和协同作战的特征（徐坚、朱思思，2025）。同时，算法驱动的认知战与网络攻击引发新型国际冲突，有学者提出“计算外交”概念，主张以“计算能力对抗计算风险”，强化外交智能化转型（董青岭、曹飞翠，2024）。随着人工智能“工具化”“武器化”，网络攻击升级、虚假信息泛滥、认知污染、政治舆情操纵变为现实。对此，Dan Hendrycks（美国人工智能安全中心执行主任）、Eric Schmidt（谷歌前 CEO）、Alexandr Wang（Scale AI 创始人）等指出，亟需制定全面战略应对风险，并提出“互保人工智能故障”（MAIM）的威慑概念（即任何国家试图单方面获得 AI 主导地位的激进行为都将面临对手的预防性破坏）。**二是关于虚假信息与认知战研究。**自 ChatGPT 发布以来，深度伪造和信息操控随之泛滥，极大地动摇了西方国家的民主根基与社会秩序（Shasha Yu、Fiona Carroll，2024）。生成式人工智能对意识形态渗透、国际舆论操控与军事指挥构成潜在威胁（黄日涵、姚浩龙，2023）。比如推动新闻生产由“书写逻辑”向“叙事逻辑”转变，部分已被用于制造虚假新闻（李明德、李聿哲，2025），虚假信息在数以万计的数量级上成倍扩散或者合并（邓宏光、王雪璠，2024）。尤其是 2016 年美国大选以来，人工智能成为西方政党政治的重要推动因素。技术迭代升级与政党博弈交织演进，人工智能结合大数据技术，降低政党触达选民的成本，强化科技公司在选举中的作用。这一趋势背景下，部分学者针对 ChatGPT 等人工智能大模型的政治偏见情况（朱尉、高强，2025）、人工智能对全球选举的干扰（Shanze Hasan，2024）展开研究，另有部分学者则探讨了新一代人工智能技术对选举的

“双刃剑”效应，既认可“智能选举 3.0”时代人工智能赋能选举组织效率，也指出其对选举民主带来多重负面冲击，让智能选举面临严峻的失范风险（王中原，2025）。

综上，人工智能作为第四次科技革命中最具有代表性和颠覆性的高新技术，拥有军民两用、更迭周期短、可复制性强的技术特性，具有天然的技术外溢倾向（部彦君，2024）。技术与权力深度交织，人工智能作为奇点性质的技术变量，正以前所未有的速度和影响力形塑国际权力格局与全球政治生态。全球技术军备竞赛将持续进行，全球智能鸿沟短期内未见弥合态势，如何确保人工智能安全可控、全球普惠，并最大化释放其促进和平与可持续发展的潜力，是政产学研界需要共同攻克的课题。

（八）敏捷治理：渐进改进、协同共治

人工智能技术快速演进伴生的高度不确定性和复杂风险，使传统长周期制定、刚性规则执行和单一主体主导的治理模式陷入瓶颈，难以有效匹配技术变革的节奏与应对涌现性挑战。**追求敏捷逐渐成为人工智能治理的共识性理念。**

目前，学界关于敏捷治理的研究多停留于概念阐释与案例分析，对于相关理论架构、实现路径、操作工具、流程路线仍缺乏全面系统的综合性分析。敏捷治理这一概念最初起源于美国制造业，美国政府在其出台的《21 世纪制造企业战略》报告中首次提出“敏捷制造”这一说法，旨在倡导制造企业采用现代通信技术，积极响应用户的需求变化，快速配置生产资源，实现制造的敏捷性。敏捷理念其后进入美国软件工程领域，并以 2001 年由软件开发人员所撰写的《敏捷宣言》为标志性事件，为人所熟知。该宣言阐述了敏捷理念的四种核心价值观和十二条原则。“敏捷软件开发”将传统的计算机软件开发的瀑布模式改成平行开发模式，以用户为中心，以解决问题为导向，通过与用户的频繁互动和面对面的交流和互动，来响应和满足用户快速变化的需求（于文轩，2023）。2018 年，世界经济论坛将“敏捷治理”（Agile Governance）定义为“一种柔韧性、灵活性或适应性的行为或方法”，旨在转变政府政策的产生、制定、执行方式，以期

跟上新兴技术驱动社会快速变革。具体而言，敏捷治理强调有效回应不断变化的公共需求，并呈现出四个方面的特征：价值导向上强调创新和治理协调发展，治理主体上强调政府主导、多元参与，治理对象上强调全过程自下而上分层治理，治理工具上强调灵活响应、软硬兼施（龙龙，2024）。基于此，已有部分学者从敏捷治理的角度出发，对虚假信息治理（何宇华、李霞，2024）、数据治理（Chukwurah 等，2024；叶英杰、李川，2025）、数字政府建设（朱国伟等，2024）等议题展开纵深案例分析，初步尝试建构敏捷治理模式框架与理论体系。也有学者从实践的角度出发，对敏捷治理的国家模式展开初步探索。如 Kodai Zukeyama 等（2024）对日本的敏捷治理模式进行了详细探析，并将敏捷治理与“以人为本的人工智能”（HCAI）进行了议题关联。**但现有成果仍多呈现点状突破的特征，在理论体系的整合度、抽象度与统摄性方面尚存提升空间，亟待通过更深入的概念凝练与跨案例比较研究，推动敏捷治理理论的系统化发展。**

作为敏捷开发模式的典型成果，生成式人工智能的治理模式也应遵循敏捷治理这一原生治理方式，传统治理模式难以为继（郭全中，2025）。这一必然性源于：其一，人工智能技术快速迭代导致制定周期较长的传统治理工具难以匹配。其二，人工智能的全面渗透性及其引发的不确定性风险导致以“命令-执行”为主的“官僚”治理模式难以奏效（龙龙，2024）。而**敏捷治理这一实时反馈、渐进改进、多方参与、灵活自适应的治理模式则为当下的人工智能治理带来了新的窗口。然而，在实际治理过程中，寻找恰当的敏捷性与灵活性仍面临诸多挑战。**安全性和敏捷性之间存在持续张力（Sascha Nägele，2023）。具体而言，在应对新兴技术时，政府在识别治理对象及其风险、权衡治理目标与治理工具方面仍然面临着信息不对称乃至“共同无知”的局限性。发展与规制的内在张力也使得利益相关方难以真正实现协同共治、监管协调。风险的“倒逼”和避责压力很容易导致监管出现敏捷治理和集中治理的急剧切换（薛澜等，2024）。

此外，部分学者对敏捷治理中政府所应承担的角色进行了探讨。敏捷治理模式不同于传统自上而下的治理模式，强调政府、企业、社会组织等多元主体的共同参与。自下而上的模式保证了治理理念的多元化与效益的公平性，但也会衍生决策进程迟缓等附带作用。多个利益攸关方可能有不同的治理目标和优先事项，

如企业更加注重技术创新，政府更加注重安全风险防治，用户个人更加注重隐私保护，这些目标之间可能存在冲突（郭全中，2025）。因此，**决策权的共享程度极大地影响着治理的效率。敏捷治理在倡导不同主体广泛参与的同时，也强调政府作为权威主体在其中扮演的主导性角色**（薛澜，2024）。何宇华（2024）进一步指出，**敏捷治理强调要在政府和非政府行为者之间保持决策权的极化，即仅在一方保持决策权**。以生成式人工智能虚假信息治理为例，治理效率是首要问题，同时又事关意识形态风险防范，因此应且仅应由政府来主导。同时，明确的政府导向和有效的高层管理支持是提升组织敏捷性的重要因素，为创造更大的公共价值和提高治理的敏捷性，必须以强有力的政治领导推动敏捷治理的关键流程。具体体现为，当多主体共同参与生成式人工智能虚假信息治理时，政府需要主动式、前置式协调来自社会各个层面和不同部门的参与资源，有效地组织工作和信息交换，以消解“去中心化”带来的权力泛化、决策困难和合作懈怠。因此，政府主导，多方参与同时保证了敏捷治理的效率性与包容性。

综上，敏捷治理作为人工智能治理的新范式，其核心价值在于通过多元主体协同与弹性规则迭代破解传统治理的时滞性困境。这一模式与中国倡导的“包容审慎”以及“多利益相关方”深度契合，可为丰富敏捷治理概念谱系提供突破思路。未来需要双向发力，推动敏捷治理的理论深化与实践转化，包括发展敏捷治理可操作框架，例如通过分层监管沙盒（如人工智能分级分类治理机制）、政策适应性实验（如地方试点容错纠错制度）及政企责任共担体系（如算法安全备案与追溯机制）等创新路径，在激发技术创新活力与防控系统性风险之间构建动态平衡的治理生态。

三、结 语

随着生成式人工智能的裂变演进，技术已超越工具范畴，成为重塑经济秩序与文明形态的元力量。本报告系统梳理了人工智能社会发展的八大核心议题：算力扩张在加剧能源困局的同时，也通过绿色转型与核能突破开辟了可持续发展路径；数据让渡在实现精准场景化服务的同时，也引发了隐私非自主让渡与知情同意机制失效的风险；模型幻觉既是可信性挑战的源头，亦为创造性突破的潜在催化剂；生成式人工智能颠覆劳动市场“低技能=高替代”的传统逻辑，推动“技术适应力+人文内核”双驱动的劳动力体系发展；智能鸿沟揭示技术红利分配的多维结构性不平等，国家发展失衡、企业巨头垄断、社会公平失序呼吁技术包容性导向发展；人类面临主体性异化危机，而机器“交互主体性”的涌现打破了主客体二元框架，未来需在理论与法律层面重构人机共生的主体性范式；人工智能通过重塑国家权力与国际安全格局，成为大国权力博弈的核心战略要素；敏捷开发与敏捷治理天然适配，以反馈与共治为核心的敏捷治理模式破解了传统规制的时滞困境，亟需推动相关可操作框架转化。**赋能与风险并存、解构与重构共生、博弈与协同竞合、工具理性与价值理性碰撞。**人工智能社会化揭示出技术嵌入现代性肌理的深层逻辑与复杂面向。

人工智能的社会科学研究，其核心价值不仅在于理解技术，也不止在于驾驭技术，更在于深入探索技术与人类社会的共生逻辑。技术发展驱动制度创新，社会调适反哺技术驯化。唯有持续深化相关核心议题的探索，构建技术革新与社会福祉的动态平衡机制，方能引导人工智能成为人类文明的向善建构性力量，而非社会分化的潜在风险源。这既要求学术理论的前沿突破，亦需政府、市场和社会协同构建治理框架，共同塑造兼具创新活力与人本价值的智能文明。

附录：参考文献（Reference）

[1] Adam Satariano：The A.I. Race Is Splitting the World Into Haves and Have：https://www.nytimes.com/interactive/2025/06/23/technology/ai-computing-global-divide.html

[2] Adam Zewe：Explained: Generative AI’s environmental impact：https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117

[3] Bonney, K., Breaux, C., Buffington, C., Dinlersoz, E., Foster, L., Goldschlag, N., ... & Savage, K.（2024）. The impact of AI on the workforce: Tasks versus jobs?. Economics Letters, 244, 111971.

[4] Burkhardt, G., Boy, F., Doneddu, D., et al.（2023）. Privacy behaviour: A model for online informed consent. Journal of Business Ethics, 186, 237–255.

[5] Chukwurah, N., Ige, A. B., Idemudia, C., & Eyieyien, O. G.（2024）. Integrating agile methodologies into data governance: Achieving flexibility and control simultaneously. Open Access Research Journal of Multidisciplinary Studies, 8（01）, 045-056.

[6] Fuchs, T.（2024）. Understanding Sophia? On human interaction with artificial agents. Phenomenology and the Cognitive Sciences, 23, 21–42.

[7] Goldman Sachs: Generative AI could raise global GDP by 7%: https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent.html

[8] Gondwe, G. (2023). CHATGPT and the Global South: how are journalists in sub-Saharan Africa engaging with generative AI?. Online Media and Global Communication, 2(2), 228-249.

[9] Gregory Daco: How GenAI will impact the labor market: https://www.ey.com/en_gl/insights/ai/how-gen-ai-will-impact-the-labor-market

[10] Hawkins, Z., Lehdonvirta, V., & Wu, B.（2025）. AI Compute Sovereignty: Infrastructure Control Across Territories, Cloud Providers, and Accelerators. Cloud Providers, and Accelerators（June 20, 2025）.

[11] IEDC: Artificial Intelligence Impact on Labor Markets: https://www.iedconline.org/clientuploads/EDRP%20Logos/AI_Impact_on_Labor_Markets.pdf

[12] Jared Cohen: The Next AI Debate Is About Geopolitics: https://foreignpolicy.com/2024/10/28/ai-geopolitics-data-center-buildout-infrastructure/

[13] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P.（2023）. Survey of Hallucination in Natural Language Generation. ACM computing surveys, 55（12）, 1-38.

[14] Lindgren, H.（2024）. Emerging roles and relationships among humans and interactive AI systems. International Journal of Human − Computer Interaction, 1 − 23.

[15] Lutz, C. (2024). Social Inequalities and Artificial Intelligence: How Digital Inequality Scholarship Enhances Our Understanding. In Algorithms, Artificial Intelligence and Beyond (pp. 193-210). Routledge.

[16] Maleki, N., Padmanabhan, B., & Dutta, K.（2024）. AI hallucinations: A misnomer worth clarifying. arXiv. arXiv preprint arXiv:2401.06796.

[17] Martin, K. D., & Zimmermann, J. （2024）. Artificial intelligence and its implications for data privacy. Current Opinion in Psychology, 58, 101829.

[18] Molly Kinder. el： Generative AI, the American worker, and the future of work: https://www.brookings.edu/articles/generative-ai-the-american-worker-and-the-future-of-work/

[19] Northoff, G., & Gouveia, S. S. （2024）. Does artificial intelligence exhibit basic fundamental subjectivity? A neurophilosophical argument. Phenomenology and the Cognitive Sciences, 23, 1097 – 1118.

[20] Nägele, S., Schenk, N., & Matthes, F. (2023, June). The current state of security governance and compliance in large-scale agile development: A systematic literature review and interview study. In 2023 IEEE 25th Conference on Business Informatics (CBI) (pp. 1-10). IEEE.

[21] Rane, N.（2023）. Contribution of ChatGPT and other generative artificial intelligence（AI）in renewable and sustainable energy. Available at SSRN 4597674.

[22] Ryan Stenvick: Why AI is Here to Help, Not Replace: Understanding GenAI’s Role in the Modern Workforce: https://acarasolutions.com/blog/recruiting-trends/why-ai-is-here-to-help-not-replace-understanding-genais-role-in-the-modern-workforce/

[23] Shanze Hasan： The Effect of AI on Elections Around the World and What to Do About It: https://www.brennancenter.org/our-work/analysis-opinion/effect-ai-elections-around-world-and-what-do-about-it

[24] Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. （2024）. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications, 11（1）, 1-14.

[25] UN: Artificial intelligence and the future of work: Will AI replace our jobs?: https://unric.org/en/artificial-intelligence-and-the-future-of-work-will-ai-replace-our-jobs/

[26] UN:Artificial Intelligence and gender equality: https://www.unwomen.org/en/articles/explainer/artificial-intelligence-and-gender-equality

[27] WEF: Agile Governance Reimagining Policy-making in the Fourth Industrial Revolution: https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf

[28] Yadav, N., Pandey, S., Gupta, A., Dudani, P., Gupta, S., & Rangarajan, K. (2023). Data privacy in healthcare: in the era of artificial intelligence. Indian Dermatology Online Journal, 14(6), 788-792.

[29] Yu, S., & Carroll, F. （2023）. A balance of power: Exploring the opportunities and challenges of AI for a nation. In R. Montasari （Ed.）, Applications for artificial intelligence and digital forensics in national security （pp. 15 – 37）.

[30] Zhang, E., Wu, D., & Boman, J. （2024, August）. Carbon-Aware Workload Shifting for Mitigating Environmental Impact of Generative AI Models. In 2024 IEEE International Conferences on Internet of Things （iThings） and IEEE Green Computing & Communications（GreenCom）and IEEE Cyber, Physical & Social Computing（CPSCom）and IEEE Smart Data （SmartData） and IEEE Congress on Cybermatics （pp. 446-453）.

[31] Zukeyama, K., Nishimura, T., Kawashima, H., & Yamamoto, T. （2024）. Agile governance

as AI governance: A challenge for governance reformation in Japan. In Human - centered AI（1st ed., pp. 276）.

[32] Østergaard, S. D., & Nielbo, K. L. （2023）. False responses from artificial intelligence models are not hallucinations. Schizophrenia bulletin, 49（5）, 1105-1107.

[33] 部彦君,许开轶. 重塑与介入:人工智能技术对国际权力结构的影响作用探析[J]. 世界经济与政治论坛,2023,（1）: 86-111.

[34] 部彦君. 人工智能发展下的权力扩散态势解析与挑战应对[J]. 科学学研究,2025,43（3）: 608-616.

[35] 蔡翠红. 国家间权力关系视域中的人工智能国际竞争与合作[J]. 当代世界,2024,（5）: 19-24.

[36] 陈菲,蒲文杰. 全球人工智能治理的去中心化 ——基于“主体—模式—结构”的分析框架[J]. 世界经济与政治论坛,2025,（3）: 15-44，177.

[37] 陈晓红,曹廖滢,陈姣龙,等. 我国算力发展的需求、电力能耗及绿色低碳转型对策[J]. 中国科学院院刊,2024,39（3）: 528-539.

[38] 邓宏光,王雪璠. 生成式人工智能虚假信息治理的风险与应对[J]. 理论月刊,2024,（9）: 115-129.

[39] 董青岭,曹飞翠. 计算外交：基于智能驱动的外交革命[J]. 国际观察,2024,（2）: 44-71.

[40] 董淑芬,李志祥. 论大数据后疫情时代的隐私让渡[J]. 齐鲁学刊,2023,（4）: 62-72.

[41] 方师师,唐巧盈. 聪明反被聪明误：ChatGPT 错误内容生成的类型学分析[J]. 新闻与写作,2023,（4）: 31-42.

[42] 方兴东,钟祥铭. 生成式 AI 与智能鸿沟:智能时代数字不平等的趋势、逻辑与对策[J]. 湖南师范大学社会科学学报,2024,53（6）: 121-131.

[43] 郭全中. 技术演化与涌现风险：生成式人工智能的协同式敏捷治理体系研究[J]. 编辑之友,2025,（4）: 49-56.

[44] 何宇华,李霞. 生成式人工智能虚假信息治理的新挑战及应对策略 ——基于敏捷治理的视角[J]. 治理研究,2024,40（4）: 142-156，160.

[45] 胡晶晶,程承坪. 新一代人工智能对就业的影响及应对策略[J]. 人文杂志,2025,（1）: 40-52.

[46] 胡泳,王昱昊. 技术过程论视角下 AI 幻觉生成的价值负荷与伦理问题探析[J]. 南京社会科学,2025,（3）: 84-94.

[47] 黄日涵,姚浩龙. 被重塑的世界？ChatGPT 崛起下人工智能与国家安全新特征[J]. 国际安全研究,2023,41（4）: 82-106，158-159.

[48] 黄旭,董志强. 人工智能对劳动力市场极化的影响与对策[J]. 系统工程理论与实践,2024,44（1）: 272-298.

[49] 雷昊楠. 生成式人工智能的竞争风险、监管逻辑与反垄断应对[J]. 中国科技论坛,2024,（7）: 64-74.

[50] 李戈,何玉芳. 人工智能时代人的主体性危机及其消解[J]. 北京工业大学学报(社会科学

版）,2025,25（3）: 101-110.

[51] 李坤泽. 人工智能对美国能源产业发展的影响[J]. 现代国际关系,2025,（4）: 49-68, 138-139.

[52] 李明德,李聿哲. 赋能和消解:生成式人工智能与新闻真实的碰撞[J]. 西安交通大学学报（社会科学版）,2025,45（2）: 109-118.

[53] 李洋,薛澜. 颠覆、调适与协同：责任伦理视域下生成式人工智能的多主体治理机制研究[J]. 电子政务,2025,（7）: 40-49.

[54] 李兆轩. 开源模式下的垄断及其二元治理之策[J]. 中国法律评论,2025,（3）: 215-226.

[55] 林爱珺,章梦天. 基于数据多粒度的隐私差序保护[J]. 苏州大学学报（哲学社会科学版）,2024,45（2）: 182-192.

[56] 林曦. 人工智能“幻觉”的存在主义阐释[J]. 社会科学辑刊,2025,（2）: 81-91.

[57] 凌应生. 论人工智能对人的主体性的冲击及应对[J]. 理论视野,2025,（4）: 66-71.

[58] 刘宏伟,马酉儒. 生成式人工智能与绿色生产力培育:关键要素、驱动路径与政策建议〔*〕[J]. 学术界,2025,（3）: 78-88.

[59] 刘琳璘. 人工智能时代数据安全风险防范体系研究[J]. 政法学刊,2024,41（3）: 32-41.

[60] 刘艳红. 生成式人工智能的三大安全风险及法律规制 ——以 ChatGPT 为例[J]. 东方法学,2023,（4）: 29-43.

[61] 刘泽垣,王鹏江,宋晓斌,等. 大语言模型的幻觉问题研究综述[J]. 软件学报,2025,36（3）: 1152-1185.

[62] 龙海泉,刘卓月,王向团. 人工智能大发展对算力和能源影响研究[J]. 中国信息化,2024,（11）: 7-8.

[63] 龙龙. 敏捷治理理念下人工智能软法之治的问题与对策[J]. 江汉论坛,2024,（5）: 128-134.

[64] 梅夏英. 伦理人格与技术人格：人工智能法律主体地位的理论框架[J]. 中外法学,2025,37（1）: 26-44.

[65] 孟芳. 从边界重构到协作共生：生成式人工智能背景下的人机关系演进探究[J]. 人工智能,2025,（3）: 86-95.

[66] 孟现玉. 风险社会理论下人工智能时代的失业风险与法律治理[J]. 郑州大学学报（哲社）,2024,57（4）: 56-63.

[67] 彭兰. 智能传播时代“智能鸿沟”的走向探询[J]. 中国编辑,2024,（11）: 19-26.

[68] 彭姝. 论生成式治理中人的主体性建构: 马克思主义的分析视角[J]. 科学社会主义,2025,（2）: 29-37.

[69] 戚凯,杨悦怡. 人工智能时代的美国对华算力竞争[J]. 国际论坛,2024,26(3): 43-61, 156, 157.

[70] 秦北辰. 数字技术、权力失衡与全球南方产业发展的困境[J]. 外交评论（外交学院学报）,2024,41（3）: 80-103，167-168.

[71] 生成式人工智能对劳动力市场的颠覆性影响及其应对

[72] 石冠彬. 人工智能成为法律主体不存在理论障碍[J]. 光明日报 · 理论版,2024,（81）.

[73] 寿步. agent(行为体)是人工智能领域的法律主体[J]. 网络信息法学研究,2023,(1): 21-34,284.

[74] 寿步. 人工智能术语 agent 的精准译解及其哲学意义[J]. 哲学分析,2023,14(3): 130-143,199.

[75] 孙国烨,吴丹,刘静,等. 用户与生成式人工智能交互的隐私披露多因素影响模型研究[J]. 信息资源管理学报,2025,15（2）: 108-122.

[76] 孙良国. 人工智能不应成为法律主体[J]. 光明日报 · 理论版,2024,（81）.

[77] 孙志伟,殷浩铖. 人工智能时代数字巨头的技术权力及其对“全球南方”的挑战[J]. 国际安全研究,2025,43（2）: 142-164，168.

[78] 王丽媛,李繁荣. 人工智能、产业结构服务化转型与劳动收入份额[J]. 经济问题,2024,(8): 52-59.

[79] 王梅艺. 人工智能时代数据安全风险及应对策略[J]. 中阿科技论坛(中英文),2024,(11): 139-143.

[80] 王中原. 竞争性选举的智能转型: 动力机制、技术过程与政治影响[J]. 政治学研究,2025,（1）: 108-118，189-190.

[81] 吴鼎铭,汪荣燊. 人工智能时代算法的公共化生产[J]. 福建师范大学学报(哲学社会科学版）,2025,（3）: 78-87.

[82] 吴红,姜惠. 从“红色 AI”到“绿色 AI”——人工智能的生态范式转换[J]. 探索与争鸣,2025,（5）: 99-110，178-179.

[83] 武延军. DeepSeek 引发的 AI 创新和开源生态发展的思考[J]. 中国科学院院刊,2025,40（3）: 446-452.

[84] 徐偲骕. 智能社会语境下的数据治理如何助力性别平等[J]. 宁夏社会科学,2024,（4）: 166-175.

[85] 徐坚,朱思思. 人工智能驱动的网络空间国际竞争及其治理路径[J]. 中国网信,2025,(5): 57-61.

[86] 徐政,吴晓亮. 人工智能发展对劳动力市场的影响 ——基于政治经济学视角的分析[J]. 湖北大学学报（哲学社会科学版）,2025,（3）: 145-152.

[87] 许潇丹,惠宁. 人工智能对工业绿色低碳发展的影响研究[J]. 陕西师范大学学报(哲学社会科学版）,2024,53（6）: 74-86.

[88] 薛澜,贾开,赵静. 人工智能敏捷治理实践: 分类监管思路与政策工具箱构建[J]. 中国行政管理,2024,40（3）: 99-110.

[89] 阎孟伟. 人工智能能否最终超越人类智能[J]. 社会科学战线,2024,（11）: 44-56.

[90] 杨雅,滕文强,喻国明. “熟知还是真知”?生成式人工智能对媒体融合变革的影响——基于复杂系统的视角[J]. 南京社会科学,2024,（11）: 94-104.

[91] 叶英杰,李川. 人工智能模型训练中合成数据的应用风险及其治理路径[J]. 情报理论与实践,2025,48（6）: 47-55.

[92] 易宪容,陈颖颖. 生成式人工智能对劳动力市场的颠覆性影响及其应对[J]. 江海学刊,2024,（6）: 91-99.

[93] 殷杰. 生成式人工智能的主体性问题[J]. 中国社会科学,2024,（8）: 124-145，207.

[94] 于文轩. ChatGPT 与敏捷治理[J]. 学海,2023,（2）: 52-57.

[95] 余南平. 通用人工智能时代的国际权力重塑[J]. 中国社会科学,2025,（4）: 41-59，205.

[96] 俞立根,顾理平. 隐私何以让渡:量化自我与私人数据的日常实践[J]. 苏州大学学报（哲学社会科学版）,2024,45（2）: 172-181.

[97] 袁康. 破除人工智能成为法律主体的臆想[J]. 光明日报 · 理论版,2024,（82）.

[98] 张涵,许海滨,丁立捷,等. 算法时代“无感伤害”下的隐私侵权[J]. 全媒体探索,2024,(4): 120-122.

[99] 张秀武,沈洋. 人工智能对减污降碳协同治理的影响效应及作用机制研究[J]. 现代财经-天津财经大学学报,2025,45（5）: 77-94.

[100] 张严. 生成式人工智能垄断风险问题探析[J]. 科学决策,2024,（9）: 203-213.

[101] 赵勇. 青海绿色算力的金融支持研究[J]. 青海社会科学,2024,（3）: 90-96.

[102] 赵雨涵,于佳琪,晏欧伦,等. 人工智能技术创新能抑制城市碳排放强度吗？ ——来自中国 277 个地级市面板数据的证据[J]. 科学决策,2025,（2）: 72-90.

[103] 甄航. 人工智能刑法“主体性”否定:缘起、解构、反思 ——以认知科学的五个层级为基础[J]. 重庆大学学报（社会科学版）,2024,30（3）: 242-252.

[104] 朱国伟,周妍池,刘银喜. 敏捷治理推动数字政府建设：发展趋势与实现路径[J]. 电子政务,2024,（2）: 55-64.

[105] 朱尉,高强. ChatGPT 对美国两党和中美两国的政治偏见及意识形态风险[J]. 阅江学刊,2025,17（1）: 102-118，173.