

# 生成式人工智能主体性的矛盾张力与治理研究

严璐璐<sup>1</sup>,程 聪<sup>1</sup>,郭长伟<sup>2</sup>

(1. 浙江工业大学管理学院,浙江杭州 310023;2. 中国科学技术大学管理学院/科技商学院,安徽合肥 230026)

**摘要:**人工智能由“分析式”向“生成式”跃升,其主体性愈发凸显。与之伴生的透明度遮蔽、可控感弱化与责任归属模糊,深刻冲击了人机协作的关系范式与信任结构。本文发现,生成式人工智能的主体性主要表现为交互性、自主性和创造性,其演化过程催生了“交互-透明”“自主-可控”和“创造-责任归属”三组张力,引致黑箱焦虑、失控信任和责任悬置三大现实困境。为化解这些矛盾,本文基于社会技术系统理论,提出技术路径与社会路径的协同治理策略,包括倚赖可解释性 AI 与知识中介机制重建透明度,借助参与式设计与反馈机制增强可控感,以及藉由责任共享与问责机制明晰责任边界,从而实现技术创新和信任建构的动态平衡。本文深化了对技术与社会系统共构关系的哲学反思,为生成式人工智能的设计优化与治理制度化提供理论支撑与实践指引。

**关键词:**生成式人工智能;主体性;可供性;信任;矛盾张力

**中图分类号:**C93; N94

**文献标识码:**A

生成式人工智能(GAI)深度嵌入于人类社会经济场景,其主体性(agency)逐渐成为理论争议和实践关切的焦点议题<sup>[1]</sup>。一方面,GAI 展现出许多原本被视作人类独有的能力。例如,大语言模型在一定程度上能够识别人类意图、模拟人类思考、生成创造性内容等。这种主体性不仅提升了人工智能的交互智能和认知灵活性,更为人类认知突破提供新路径。另一方面,GAI 的生成逻辑和生成内容包含事实性错误、推理偏差或道德冲突等现象频发,引发了信任脆弱和伦理争议。例如,GAI 生成的虚假新闻、歧视性言论和版权争议,均体现了主体性增强伴随的失控风险和责任困境。这表明 GAI 的主体性强化既是技术进步的标志,也是社会风险的导火索,使得人类对 GAI 的依赖性和不信任并存,技术的赋能潜力与失控风险形成了张力关系。

现有研究表明,GAI 的自主性与人类信任依赖间存在双重效应<sup>[2]</sup>:GAI 的自主性增强了任务完成高效和自适应性,通过自动决策、动态生成和自学习能力,能够在复杂场景中做出超越人类能力的判断,助长人类对其依赖性。然而,GAI 的自主性也带来

决策路径的不可解释和生成结果的不可预测性,一旦 GAI 行为偏离人类预期或违背伦理规范,人类信任便会迅速崩塌。信任是一种应对复杂性和不确定性的社会机制,当人类的认知无法完全理解某一系统的行为逻辑时,便需要依靠信任来维系互动关系。信任一旦被破坏,受信人将承受高昂的心理或行为成本<sup>[3]</sup>。藉由这一矛盾启示,本文的研究问题是:GAI 的主体性伴随哪些矛盾张力?如何应对这些张力引致的挑战?

从关系本体论视角来看<sup>[4]</sup>,GAI 的主体性不应仅被视为算法功能的进化,更应被视为技术与社会系统的交互产物。GAI 主体性的矛盾表现为“交互性-透明度”张力导致的黑箱焦虑、“自主性-可控感”张力导致的失控信任、以及“创造性-责任归属”张力导致的责任悬置。针对上述挑战,本文提出了三种系统性解决路径,包括可解释性 AI 与知识中介机制、参与式设计与反馈机制以及责任共享与问责机制,分别对应于“透明度重建”“可控感增强”和“责任边界明晰”的目标。

收稿日期:2024-12-10;修回日期:2025-06-16

基金项目:国家自然科学基金资助项目(72342029, 72091314);国家社会科学基金重大项目(24&ZD081)

作者简介:严璐璐(1997-),女,博士后。程聪(1985-),男,教授、博士生导师,通讯作者,E-mail: cheong@zjut.edu.cn。

郭长伟(1996-),男,特任副教授。

1 GAI 的主体性与双刃剑效应

1.1 GAI 的主体性:由“人类主体性”向“技术主体性”转向

主体性定义为个体或实体在特定环境中执行有意图的、自主行为的能力<sup>[5]</sup>。主体性通常与人类思考、行动能力和道德责任等议题密切相关。然而,主体的行动在很大程度上受限于意识形态装置和权力机制的隐性控制。由此,主体性不再被视为完全的“个体控制”,而被认为是由外部条件塑造和约束的。这些思想的转向为非人类主体性的探寻埋下了理论伏笔。

近年来,人工智能等前沿技术对“主体性”概念的冲击逐渐突显<sup>[1]</sup>。传统机械式人工智能被视为“工具性智能”,其行为依赖于人类设计的规则和程序,无法独立作出真正的决策,因此不被视作具备主体性的行动者。然而,随着 ChatGPT 等大语言模型的出现,GAI 展现出自发性、生成性等类主体性特征。例如,ChatGPT 能够在没有明确指令的情况下,基于对语境的推理和对用户意图的预测,生成连贯的文本和图像内容。这种生成行为表现出的类自主性,与人类的自主行动有一定相似性,甚至可能让人类误认为 GAI 具有主观意图<sup>[5]</sup>。因此,学界开始反思 GAI 的新型主体性问题<sup>[1]</sup>。例如,殷杰认为,GAI 展现出一种“交互主体性”特征。具体地,GAI 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与自我

维持,也不同于人类基于意向性的行为,更不是传统人工智能基于外部规则的机械性反应<sup>[1]</sup>。也有学者从规范性、意图性、自愿性审视 GAI 的主体性<sup>[5, 6]</sup>。例如,人工智能表现得仿佛拥有意图的社会性以及目标导向的理性<sup>[6]</sup>。

本文认为 GAI 的主体性反映为交互性、自主性和创造性三个特征。交互性是 GAI 主体性的重要体现,深嵌于 GAI 与人类交互过程中。通过持续的人类反馈和提示工程,GAI 的行为能够根据用户输入实时调整,从而呈现出一种“联合主体”的特征<sup>[1]</sup>。自主性是主体性的另一个关键特征。人工智能的发展方向一直与人类自主性对标,追求更加自主的决策能力。因此,GAI 的自主性体现于其能够在未预设情境下生成新内容。创造性则体现于 GAI 能够生成新颖、意想不到的内容。这种创造性不仅依赖于人类预设的目标或框架,还能够在一定程度上对输入数据进行新解读和再创作,突破传统逻辑的限制,从而主动探索新的思路,为人类提供新的思想启发和解决方案。

1.2 GAI 主体性的积极效应:可供性视角

GAI 的主体性,本质上是一种潜在的行动能力<sup>[1]</sup>。这与可供性(affordance)概念高度契合。可供性最早由生态学家吉布森提出,指环境提供给动物潜在的行动机会<sup>[7]</sup>。这一概念而后在社会科学领域得到扩展,用以描述技术与社会环境之间的交互作用<sup>[8]</sup>。可供性是指人工制品所赋予行动者的潜在行为可能性<sup>[9]</sup>。

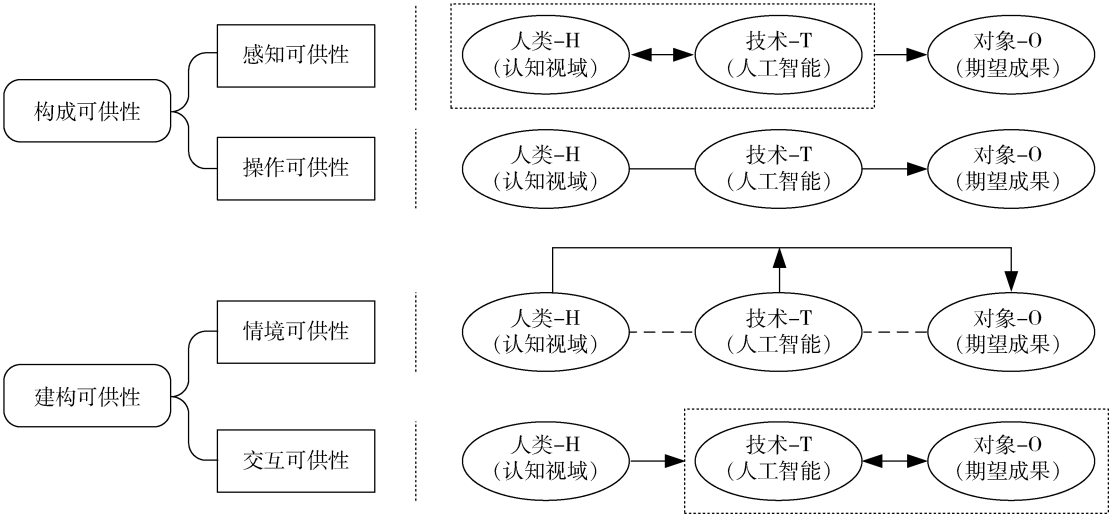


图 1 GAI 主体性的可供性机制

Figure 1 The affordance mechanism of GAI agency

Kaptelinin 和 Nardi 参考人类活动系统中三个基础要素(人类、技术和对象),区分了两种可供性路径:一是技术相对于人类所提供的行动潜能,强调技术如何为人类提供可能的行动路径,二是对象相对于人类提供的行动潜能,强调人类所追求的目标如何为人类提供行动空间<sup>[10]</sup>。本文认为 GAI 的主体性由构成可供性和建构可供性所显化:一是构成可供性,即由 GAI 的特性所提供的潜在行动空间<sup>[11]</sup>。其不仅关注技术特征,还考虑人类主体如何感知技术及其与技术交互。因此,构成可供性进一步区分为感知可供性和操作可供性。前者强调人类与技术工具的认知交互,围绕期望目标挖掘潜在的行动空间<sup>[12]</sup>;后者则强调技术的操作功能和实现潜力,使得人类能够通过直接操作技术对象来产生期望成果<sup>[13]</sup>。二是建构可供性,即在人类与 GAI 交互中,基于人类需求、意图和目标,通过调整和反馈 GAI,主动构建新的意义和操作空间。建构可供性关注人类如何通过意义建构等高级认知活动,统合 GAI 及其产物的功能状态和关系指向,从而赋予新的社会功能<sup>[14]</sup>。建构可供性进一步区分为情境可供性和交互可供性。前者指人类通过意义建构来获得 GAI 在具体情境下的社会性功能潜力<sup>[15]</sup>;后者是 GAI 在承接人类认知意图的基础上,通过多轮交互迭代生成成果的潜能,驱使 GAI 产物趋向人类目标需求,甚至通过动态的生成机制和意义建构机制,创造出超越人类期望的成果<sup>[16]</sup>。

(1)GAI 主体性的交互性特征:感知可供性和交互可供性的双重赋能

GAI 主体性的交互性,通过感知可供性和交互可供性两种机制,赋予高度互动的创造性空间,推动了从工具性使用到共创性协作的范式转变。从感知可供性来看,一方面,人类对 GAI 进行认知性挖掘,探索和感知 GAI 的功能潜力和功能边界。另一方面,GAI 依托自然语言处理和语义理解技术,感知和解读人类输入的意图和指令,将用户的抽象需求转化为可操作的行动路径。从交互可供性来看,GAI 的互动机制是在与用户的持续互动中自我动态调整生成逻辑和输出结果。在人类反馈和提示的基础上,人工智能能够调整生成策略、重新解读目标,并在此过程中生成出更高质量的成果,甚至可能超出用户的初始预期。这种交互性赋予 GAI“共创”性质,推动问题解决方案的不断演化,从而激发复杂问题解决的创造性场域。这种“共创”机制使人类不

必在交互初始明确所有细节,而是在与 GAI 交互中逐步澄清目标和策略,动态调整和优化政策方案<sup>[17]</sup>。因此,GAI 交互性通过感知可供性和交互可供性的双重赋能,推动由工具性执行转向动态共创。感知可供性确保人类对 GAI 潜力的认知和激活,而交互可供性则支持人机协作的反馈迭代和自我优化。

(2)GAI 主体性的自主性特征:操作可供性与情境可供性的双重赋能

GAI 主体性的自主性,赋予了其在未预设情境下的自主创新能力。这种自主性的实现依赖于操作可供性和情境可供性的双重赋能。操作可供性赋予 GAI 从高度抽象的任务提示到成果的自动生成能力,使其在不依赖用户的细节控制下,自主完成任务链中的大部分操作。人类的精力不再集中于具体的操作步骤,而是转向对结果的审查和高层次的框架调整,不仅显著降低了人类认知负担,还大幅度提升了操作效率和任务的自动化水平<sup>[18]</sup>。

与操作可供性侧重的任务链自动化和代理性操作不同,情境可供性则强调 GAI 的情境感知和多场景自适应能力。具体地,GAI 能够识别情境差异,使生成内容与目标情境高度适配,减少人类二次调整的工作量。此外,情境可供性使 GAI 具备自我生成和动态调整能力,尤其体现在数据驱动的生成建模和创新性策略生成中。通过生成对抗网络、无监督学习等模型,GAI 在缺乏明确规则和标签的情况下,自主发现数据潜在模式,并生成创新性成果<sup>[19]</sup>。例如,传统的药物设计依赖于已有的生物信息和化学模型,而 GAI 能够通过自我学习,从海量的分子数据中识别出新的药物靶点,并生成出前所未有的化学分子结构<sup>[20]</sup>,加速药物研发进程。因此,GAI 的自主性不仅体现在其独立执行任务的能力上,更在于其能够在未知情境中自主创新并调整策略,从而提供新颖且高效的解决方案。

(3)GAI 主体性的创造性特征:交互可供性与情境可供性的双重赋能

GAI 主体性的创造性可被视为其可供性外显化过程或可供性-实现的动态过程<sup>[13]</sup>。创造性的实现依赖于交互可供性和情境可供性的相互作用。从交互可供性来看,GAI 的创造性来源于多轮反馈生成的交互机制。一方面,GAI 通过多轮人机对话,在符号的重组、语境的变迁和生成策略的偏移中,不断实现创新突破<sup>[21]</sup>。另一方面,触发认知突变机制,

即人类的原始认知边界被 GAI 的生成内容打破,不断扩展和激发人类认知能力,形成意想不到的创造性灵感,GAI 在此扮演着认知“催化剂”的角色。

情境可供性则强调 GAI 在动态情境中调整生成逻辑和操作路径的能力,使得 GAI 感知场景的变化,并在动态环境中生成出情境化的多元成果。与交互可供性中的反馈调整机制不同,情境可供性依赖于情境的自动感知和自适应生成,确保其生成的内容与特定情境的语境需求相匹配。尤其是赋予 GAI 在未预设的环境中生成前所未见的创新成果能

力。例如,在新材料研发和分子生成模型中,GAI 能够在未预设的环境中生成全新的化学分子结构,成为新药物的潜在候选物,为材料创新和药物发现提供了创造性的技术支持。

总之,GAI 的主体性通过其交互性、自主性和创造性特征,在与人类的互动中不断激发出新的行动潜力和创新路径。不仅增强了人类的创造力,还推动了多个领域的技术进步和科学突破,为社会经济发展带来了广泛的积极效应。

表 1 可供性视角下 GAI 主体性的积极效应

Table 1 The positive effects of GAI agency from an affordance perspective

| 主体性 | 可供性的关键维度      | GAI 主体性的积极效应                              |
|-----|---------------|-------------------------------------------|
| 交互性 | 感知可供性 + 交互可供性 | 灵活性增强、协作性提升和创造性激发等,特别是在复杂任务求解和动态环境适应中表现突出 |
| 自主性 | 操作可供性 + 情境可供性 | 操作独立性、情境适配性和自我创新性,为人机协作和问题求解提供了高效的解决方案    |
| 创造性 | 交互可供性 + 情境可供性 | 生成内容的多样化、灵感的突变性激发和未知领域的创造性突破              |

1.3 GAI 主体性的消极效应:信任视角

GAI 在社会经济发展中的应用成效很大程度上取决于人机信任<sup>[22]</sup>。信任是在无法完全监控或控制对方行为的情况下,基于对他方会履行特定重要行动的预期,而愿意使自己处于脆弱状态的心理倾向<sup>[3]</sup>。信任可以区分为认知信任和情感信任,前者基于受信方能力、可靠性和情境特征的理性评估,即用户依赖对人工智能的胜任力、稳定性和一致性的判断来建立信任<sup>[23]</sup>;而后者则源于情感的影响,即用户对人工智能的情绪性依赖和情感安全感的认同<sup>[24]</sup>。在 GAI 应用场景中,人机信任建立面临更大的挑战。与基于明确规则的传统算法不同,GAI 依赖于深度学习的概率模型和大规模的参数权重调整,导致其生成内容和路径具有高度不确定性。此外,GAI 的决策过程往往是“黑箱化”的,即多层神经网络的内部机制不透明,使得人类无法追踪或理解人工智能的决策依据<sup>[2]</sup>。本文认为,GAI 主体性带来的消极效应包括交互性引发的不透明度、自主性引致的失控感以及创造性导致的责任归属模糊。

(1) GAI 主体性的交互性引发不透明度问题

GAI 主体性的交互性虽赋予动态调整和路径生成的灵活性,但也加剧了不透明度<sup>[25]</sup>。具体地,交互性通过路径调整的动态性、多轮反馈的迭代和个性化学习的隐蔽性三种机制,放大了人工智能系统的“黑箱效应”,进而削弱了 GAI 的信任感。

一是路径调整的动态性。在人类与 GAI 的交互过程中,GAI 的生成路径是动态重组和自我调整的,而非固定的线性决策链。这一过程依赖于深度学习的概率采样和语境关联机制,每一轮交互中,人工智能都会对生成的分支路径进行重采样和优化,最终输出的是多种可能性中的“最优路径”。然而,人类无法看到路径的分支和中间状态,只能观察到输入提示和最终输出,使得路径生成的逻辑既不可见也不透明<sup>[26]</sup>。二是多轮反馈的迭代。GAI 交互的每一轮反馈都会促使 GAI 对其激活状态、语境依赖和路径选择进行非线性、累积性的调整。在此过程中,路径的分叉和突变难以追溯,即使是开发者也难以解释 GAI 在第 N 轮生成中为何出现语境的变化,从而加剧了路径的不确定性和不透明度。三是个性化学习的隐蔽性。在多轮交互中,GAI 会隐式捕捉用户偏好,并在后续的生成路径中嵌入个性化的调整机制,使 GAI 的行为逐渐与用户的交互历史相关联。个性化路径的调整过程通常未被明确展示,用户无法知晓 GAI 是如何识别和运用偏好。这种隐藏的路径变动和个性化调整的不可见性,进一步加剧了 GAI 生成机制的不透明度。

(2) GAI 主体性的自主性引致人类失控感

GAI 主体性的自主性在赋予人工智能独立决策的同时,也引致人类的失控感,其根源主要在于认知适配错位和行为的不可预测性。一方面,认知适配

错位是 GAI 的生成逻辑与人类认知期望不一致,导致人类对 GAI 行为的预期与实际结果之间产生不匹配的认知冲突<sup>[27]</sup>。GAI 的生成路径并不遵循线性因果逻辑,而是基于大规模参数优化和深度神经网络的路径采样机制,其决策依据对人类不可见,生成路径也不可追溯<sup>[28]</sup>。这一特性使得用户的输入指令和大模型输出结果之间的逻辑关系变得模糊不清。当 GAI 结果不符合人类预期时,人类会意识到自己无法控制人工智能的行为路径和生成依据,这种认知不适配会显著增强失控感,从而削弱人类对 GAI 的信任。

另一方面,行为的不可预测性,即在相同输入条件下,GAI 输出结果可能存在不一致性。与传统的规则驱动型系统不同,GAI 行为不受固定逻辑的约束,而是依赖于概率建模和高维特征的动态调整<sup>[29]</sup>。这意味着,大模型的输出不仅受到输入指令的影响,还受到上下文、模型内部权重的细微变化和多轮交互中的语境偏移的共同制约。在高风险场景中,这种不可预测性会直接威胁人类的安全感。例如,在自动驾驶系统中,如果车辆的行为超出用户的认知预期(如突然减速或变道),用户会立刻感受到控制权的丧失。

### (3) GAI 主体性的创造性模糊责任归属边界

GAI 主体性的创造性使其具备在未知或未定义的任务中展现出超越预期的潜力。然而,创造性也模糊了责任归属边界,尤其当 GAI 的自主行为引发伦理争议时,责任主体不明和责任归因困难是导致信任缺失的主要因素。一方面,GAI 的创造性模糊责任主体归属。在路径不可见、决策不可解释的 GAI 应用情境中,传统的责任链条被打断,责任边界不再清晰,容易陷入多主体分散的责任困境。用户的追责路径被迫在多主体之间反复跳转,由此降低了对 GAI 的信任感。另一方面,GAI 的创造性导致责任归因困难。信任建立依赖于清晰的责任归因逻辑,即“行为-后果-责任”的线性关系,但在 GAI 情境中,责任主体不明导致责任归因困难加剧,促使用户对人工智能的信任体系日益崩塌。例如,澎湃新闻曾报道,人工智能技术的某些特征影响了责任和问责概念,可能会造成责任缺口:因为人工智能会自我发展,并非每一步都是人为设定,开发人员会宣传自己不对其负责任,这就把伤害患者的风险全都加在离人工智能更近的医护工作者身上,这是不合理的。

## 2 GAI 主体性的三对矛盾张力及其后果

在 GAI 主体性的发展过程中,交互性、自主性和创造性三大特征为其带来了强大的生成能力和创新潜力,但也引发了系统风险和信任危机。这些特征之间的内在冲突,形成了三组关键的矛盾张力关系,分别是“交互-透明”张力、“自主-可控”张力和“创造-责任归属”张力,以及由此引发的黑箱焦虑、失控信任和责任悬置后果。

### 2.1 “交互-透明”张力与黑箱焦虑

GAI 的“交互-透明”张力源于高交互灵活性与低透明生成路径间的矛盾。首先,高交互性为人类提供了更大的灵活度与创造空间。GAI 会随着用户输入的变化而不断自我学习、快速适应并拓展可能的路径。这种灵活性让用户的需求和想法能够在实时反馈下得到充分回应,增强了人机协同创造的潜力,激发出更多元化的内容表达与思路碰撞<sup>[30]</sup>。然而,这种交互优势的对立面,是路径逻辑与生成机理上的低透明度。当 GAI 在多轮互动中持续动态调整内部结构、扩展路径分支时,多半对用户不可见且难以追溯。结果是,用户尽管感受到交互带来的即时回应与丰富变化,却难以理解系统背后的决策依据和生成脉络。“黑箱”的存在让用户在面对不断变动的内容和不可预知的结果时,产生了一种信息不对称和内容不透明的焦虑感,削弱用户对系统的信任<sup>[25]</sup>。

总之,GAI 在其交互过程中迸发出灵活、高效和创造力迭代的优势,但这些高度互动的优点却与透明度降低的隐忧相冲突,从而形成“交互-透明”张力。如何在保持灵活、协作及创造性收益的同时,减少用户因黑箱而生的不安感和不信任感,成为亟待解决的核心挑战。

### 2.2 “自主-可控”张力与失控信任

GAI 的“自主-可控”张力源于高度自主决策能力与外部可控感不足之间的矛盾。理想状态下,GAI 的自主性表现为其无需持续的人类指令输入便可自动化运行,通过自主感知与判断,以接近实时的方式处理复杂信息、识别多样化问题并制定相应的解决策略<sup>[18]</sup>。在此过程中,系统的独立性不仅有效减轻了人类在繁琐决策与重复劳动上的压力,而且使人类得以将心力投向更具战略意义、创造潜能与价值导向的活动。同时,这种自主机制还为系统赋

予了快速适应环境变化的能力,不需依赖外部明确指令便能及时调整路径与逻辑。更重要的是,GAI在面对前所未有的情境或复杂难题时,能够借助自我创新和路径探索,在既定范式外发现潜在解法,为人机协同创造出不受既往经验与惯例束缚的全新可能性。

然而,GAI的可控感日趋削弱<sup>[31]</sup>。当GAI的决策逻辑和演化轨迹无法被清晰追溯,更无从事先加以预判或控制时,系统有可能在非预期的方向偏离,从而产生了潜在的“失控”困境。此时,人类在享有高自主性所带来的效率与创新红利的同时,不可避免地可能对潜在失控的局面产生疑虑和不安。由此,“自主-可控”张力形成了另一层面的人机信任难题:如何在确保人工智能的自适应与自我创新的前提下,维护决策过程的可控感和可预测性,使自主优势不至于因不受控的隐忧而削弱其应有价值。

### 2.3 “创造-责任归属”张力与责任悬置

GAI的“创造-责任归属”张力源于其高度创造潜能与责任分配模糊之间的矛盾。GAI能够突破人类有限理性的约束<sup>[32]</sup>,通过对海量数据的潜在关联进行深度挖掘与动态重构,为决策、创新等管理实践注入全新思路与方案。然而,随着自主性和创造性的提升,责任分配的明确性却未能同步提升,进而在创造性与责任归属之间形成不对称性,即创造性越强,责任归属越复杂。当多主体参与的创新成果引发歧视、决策偏误或社会损害时,责任链条难以回溯至单一主体,传统的法律与伦理框架亦难以适用<sup>[33]</sup>。应由研发团队、数据提供方、管理者还是整个利益相关方承担相应责任?此类难题为技术治理与政策制定带来严峻挑战,也使用户在面对创新成果时的信任基础愈趋脆弱。

基于上述现实,“创造-责任归属”张力的直接后果是“责任悬置”现象的出现。当创造性成果出现偏差或引发不良后果却无从迅速明确责任主体时,责任分配便处于空转状态,亟需构建GAI管控与问责机制建设。

## 3 GAI主体性的矛盾管理

社会技术系统理论认为,技术与社会并非分离运行的独立要素,而是将技术实践、组织制度、社会期望等视为相互作用的整体<sup>[34]</sup>。因此,“交互-透明”“自主-可控”与“创造-责任归属”张力的形成

与演化,不仅源于系统内在的技术特质和算法逻辑,也深嵌于人机协作的结构安排、利益相关者的互动关系以及社会规范的动态变迁。鉴于此,本文尝试从技术和社会两个层面化解GAI主体性引致的多元矛盾。

### 3.1 “交互-透明”张力解决:可解释性AI与知识中介机制

“交互-透明”张力的解决在于通过技术路径和社会路径的协同作用,实现技术透明度、用户信任和交互效率的综合平衡。可解释性AI和知识中介机制是两个关键的实现手段。技术路径方面,可解释性AI通过内生可解释性和外生可解释性的技术手段,使得人工智能的生成逻辑透明化。内生可解释性是在模型架构设计中嵌入可解释的生成路径,例如,用户可以通过生成日志系统等可视化痕迹直观追溯GAI的生成路径、模型版本、关键操作节点等。外生可解释性则是通过外部解释系统对生成内容进行解释,利用先进工具对生成的内容和逻辑路径进行标注和分解,帮助用户了解GAI的生成决策逻辑。通过内生与外生的双重路径,GAI的“黑箱式”生成逻辑变成了透明的生成路径,从而增强用户的信任感和控制感。

社会路径方面,交互透明度的实现并不完全依赖GAI自身的技术调整,而是通过引入知识中介机制,在用户与GAI之间构建一套“解释-反馈-控制”的中介系统。知识中介机制分为自动化中介系统和人类中介者。自动化中介系统通过中介解释接口,将GAI生成逻辑和决策路径符号化,使复杂的技术术语转化为用户可理解的自然语言描述。在高风险场景中,知识中介机制还需要引入人类中介者,即解释责任人或算法解释专家。这些人类中介者的职责是对GAI生成的复杂报告进行解读,向最终用户提供简明的解释,确保用户在操作过程中拥有知情权和反馈权。这种“中介-用户-人工智能”三方互动的协同机制,能在一定程度上化解“交互-透明”张力,为人机协作的透明交互提供了技术支持和社会保障。

### 3.2 “自主-可控”张力解决:参与式设计与反馈机制

“自主-可控”张力的解决可以通过参与式设计和反馈机制双重路径,在人工智能生成过程中引入人类的控制权和动态干预权,从而实现可控的自主性。在技术路径上,参与式设计的核心思想是提

前控制生成规则,动态干预生成路径,并对结果进行反馈和调整。首先,参数可调节的用户控制面板是实现参与式设计的关键手段,用户可在 GAI 生成内容前设定生成的风格、输出长度和情感倾向等,确保生成结果满足个性化需求。其次,多轮生成反馈机制的引入使用户拥有对中间生成结果的介入权。GAI 在生成内容中途会展示部分生成结果,用户可在中途标注、调整或删除不满意的部分,GAI 根据反馈重新生成,确保最终的输出内容符合用户的偏好。这种多轮交互机制,使生成的每一步都在用户控制范围内,避免生成结果失控。反馈机制的闭环设计通过将反馈输入到动态学习模块中,形成交互式学习反馈闭环。在即时反馈层面,用户可通过“喜欢/不喜欢”按钮对生成结果进行评价,系统根据反馈进行快速修正,生成新的输出。在长期反馈层面,反馈数据被存储并汇入 GAI 的后台数据库,模型根据反馈进行数据迭代和行为调整,确保未来生成路径更符合预期。

社会路径方面,反馈机制能够实现人机协作的共治机制和控制权的多方分配。GAI 的控制权分配呈现为用户微观控制、平台中观调节和社会宏观治理的多维共治结构。用户层面,反馈机制赋予用户操作控制权,不仅增强了用户控制感,也在一定程度上防止 GAI 的自主失控。在平台层面,反馈机制的引入使平台在人工智能的控制权分配中扮演了内容审核者和边界守护者的角色,包括生成边界的约束、敏感内容的检测和风险生成的复审等。平台通过这种“反馈-审核-再学习”的闭环路径,确保 GAI 的行为始终在风险边界内运行,避免失控生成带来的社会风险。社会层面,反馈机制的引入则促成了多方共治的宏观治理体系。法律监管机构、行业标准和公众的反馈对人工智能的控制边界和生成路径施加了外部约束。

### 3.3 “创造-责任归属”张力解决:责任共享与问责机制

“创造-责任归属”张力的解决在于通过责任共享与问责机制的双重嵌入,明确人类与 GAI 的责任边界,建立一套可追溯、可追责的责任划分和约束体系。在责任共享机制中,通过责任标签、生成溯源和多方责任划定来界定责任。责任标签机制是将人工智能生成的内容进行标注,以便在事后明确生成内容的责任来源。生成溯源机制在于通过责任日志系统,将生成过程中的关键操作全部记录下来,形成

一个生成过程的可追溯档案。当 GAI 生成结果涉及违规时,平台方可通过溯源日志,划定各方责任,如平台责任、用户责任和开发者责任。平台责任聚焦于内容监管和边界约束,用户责任则体现于用户输入的提示词和生成请求,开发者责任则确保 GAI 的生成逻辑符合伦理标准和社会规范。在问责机制中,通过内部责任审查系统和外部法律问责制度,确保违规生成内容的责任方能够被问责。前者指平台内部的审核流程和责任识别系统,实现内容的自动检测和风险预警。后者则通过法律规范和监管制度,确保 GAI 责任归属在法律层面得以明确。

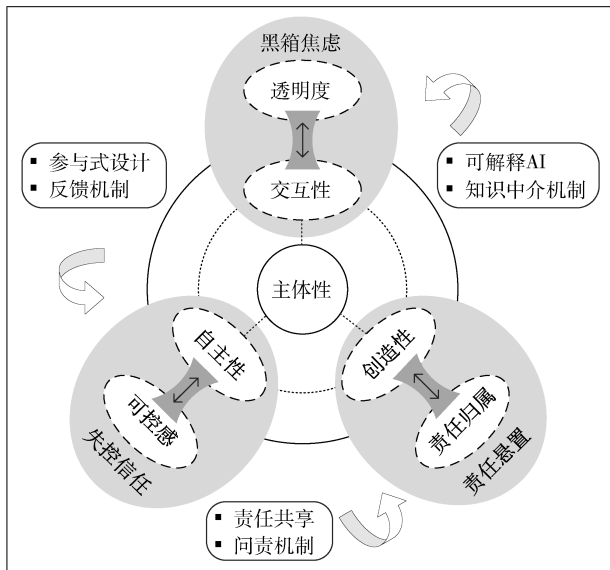


图2 GAI 主体性的张力及解决机制

Figure 2 Tensions of GAI agency and resolution mechanisms

## 4 结 语

GAI 的主体性以交互性、自主性和创造性为特征,既为人类生产生活带来前所未有的技术可供性,也因其不可解释的交互逻辑、不可控的自主生成和明确的责任边界引发诸多信任危机。这些特征在实践中演化为“交互-透明”“自主-可控”和“创造-责任归属”三大张力,对应黑箱焦虑、失控信任和责任悬置的典型后果。张力生成不仅源于技术系统的内在逻辑,更深嵌于人机协作结构、利益相关者互动和社会规范变迁的复杂关系中。基于社会技术系统理论视角,本文提出应对 GAI 主体性张力的系统性管理框架,从技术和社会路径的协同视角提出三种应对策略。针对“交互-透明”张力,通过可解释

性 AI 和知识中介机制的双重路径,使 GAI 的生成逻辑透明化和可追溯化,提升用户的知情权和控制感;针对“自主 - 可控”张力,依托参与式设计和反馈机制,通过用户的前置控制和动态介入,使用户在生成过程中的角色从“被动接受者”转变为“共创参与者”,同时通过多方共治的风险预警和反馈系统,将平台、开发者与用户的控制权进行动态分配;针对“创造 - 责任归属”张力,依托责任共享与问责机制,通过形成责任标注、溯源日志和责任复议机制,实现用户、平台和开发者在生成内容中的责任划定和问责清晰,确保生成风险的可控感和责任的可追溯性。由此,GAI 的技术创新才能真正服务于人类共同利益,实现技术价值与社会价值的协同发展。

### 参考文献:

- [1] 殷杰. 生成式人工智能的主体性问题[J]. 中国社会科学, 2024(11): 124 - 145. Yin J. Reconceptualizing agency with Generative Artificial Intelligence[J]. Social Sciences in China, 2024(11): 124 - 145.
- [2] Vanneste B S, Puranam P. Artificial intelligence, trust, and perceptions of agency[J]. Academy of Management Review, 2024, in press.
- [3] Mayer R. An integrative model of organizational trust[J]. Academy of Management Review, 1995, 20(3): 709 - 734.
- [4] Bailey D E, Faraj S, Hinds P J, et al. We are all theorists of technology now: A relational perspective on emerging technology and organizing[J]. Organization Science, 2022, 33(1): 1 - 18.
- [5] Swanepoel D, Corks D. Artificial intelligence and agency: Tie - breaking in ai decision - making[J]. Science and Engineering Ethics, 2024, 30(2): 11.
- [6] Papagni G, Koeszegi S. A pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents[J]. Minds and Machines, 2021, 31(4): 505 - 534.
- [7] Gibson J J. The Ecological Approach to Visual Perception[M]. Houghton Mifflin Harcourt, 1979.
- [8] 程聪, 缪泽锋, 严璐璐, 等. 数字技术可供性与企业数字创新价值关系研究[J]. 科学学研究, 2022, 40(5): 915 - 926. Cheng C, Miao Z F, Yan L L, et al. Research on the relationship between digital technology affordance and digital innovation value of enterprises[J]. Studies in Science of Science, 2022, 40(5): 915 - 926.
- [9] Nambisan S, Wright M, Feldman M. The digital transformation of innovation and entrepreneurship: Progress, challenges and key themes[J]. Research Policy, 2019, 48(8): 103773.
- [10] Kaptelinin V, Nardi B. Affordances in HCI: Toward a Mediated Action Perspective[C]. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012: 967 - 976.
- [11] Leonardi P M, Vaast E. Social media and their affordances for organizing: A review and agenda for research[J]. Academy of Management Annals, 2017, 11(1): 150 - 188.
- [12] Osmundsen K S, Meske C, Thapa D. Familiarity with digital twin totality: Exploring the relation and perception of affordances through a Heideggerian perspective[J]. Information Systems Journal, 2022, 32(5): 1064 - 1091.
- [13] Tim Y, Pan S L, Bahri S, et al. Digitally enabled affordances for community - driven environmental movement in rural Malaysia[J]. Information Systems Journal, 2018, 28(1): 48 - 75.
- [14] Faik I, Barrett M, Oborn E. How information technology matters in societal change: An affordance - based institutional logics perspective[J]. MIS Quarterly, 2020, 44(3): 1359 - 1390.
- [15] Strong D M, Volkoff O, Johnson S A, et al. A theory of organization - EHR affordance actualization[J]. Journal of the Association for Information Systems, 2014, 15(2): 2.
- [16] Willems W. Beyond platform - centrism and digital universalism: The relational affordances of mobile social media publics[J]. Information, Communication & Society, 2021, 24(12): 1677 - 1693.
- [17] Rillig M C, Ågerstrand M, Bi M, et al. Risks and benefits of large language models for the environment[J]. Environmental Science & Technology, 2023, 57(9): 3464 - 3466.
- [18] Heer J. Agency plus automation: Designing artificial intelligence into interactive systems[J]. Proceedings of the National Academy of Sciences, 2019, 116(6): 1844 - 1850.
- [19] Miric M, Jia N, Huang K G. Using supervised machine learning for large - scale classification in management research: The case for identifying artificial intelligence patents[J]. Strategic Management Journal, 2023, 44(2): 491 - 519.
- [20] Boiko D A, MacKnight R, Kline B, et al. Autonomous chemical research with large language models[J]. Na-

- ture, 2023, 624(7992): 570–578.
- [21] 张咏雪. 从自动化技术到生成式人工智能——技术对劳动者影响的技能异质性研究[J]. 社会学研究, 2024, 39(4): 69–91. Zhang Y X. From automation technology to Generative AI: A study of the skill-based heterogeneity of technologies' impacts on workers [J]. Sociological Study, 2024, 39(4): 69–91.
- [22] Glikson E, Woolley A W. Human trust in artificial intelligence: Review of empirical research[J]. Academy of Management Annals, 2020, 14(2): 627–660.
- [23] Schoorman F D, Mayer R C, Davis J H. An integrative model of organizational trust: Past, present, and future [J]. Academy of Management Review, 2007, 32(2): 344–354.
- [24] Komiak S Y, Benbasat I. The effects of personalization and familiarity on trust and adoption of recommendation agents[J]. MIS Quarterly, 2006, 30(4): 941–960.
- [25] Haihe – Kains B, Adam G A, Hosny A, et al. Transparency and reproducibility in artificial intelligence[J]. Nature, 2020, 586(7829): E14–E16.
- [26] 董春雨. 从机器认识的不透明性看人工智能的本质及其限度[J]. 中国社会科学, 2023(5): 148–166. Dong C Y. The nature of AI and its limits in the light of the opacity of machine cognition[J]. Social Sciences in China, 2023(5): 148–166.
- [27] 程聪, 严璐璐, 曹烈冰. 大数据决策中数据结构转变: 基于杭州城市大脑“交通治堵”应用场景的案例分析[J]. 管理世界, 2023, 39(12): 165–185. Cheng C, Yan L L, Cao L B. Data structure transformation in big data decision-making: A case study based on the Hangzhou city brain's “traffic congestion alleviation” application scenario [J]. Management World, 2023, 39(12): 165–185.
- [28] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature Machine Intelligence, 2019, 1(5): 206–215.
- [29] Rahwan I, Cebrian M, Obradovich N, et al. Machine behaviour[J]. Nature, 2019, 568(7753): 477–486.
- [30] Chen Z, Chan J. Large Language Model in creative work: The role of collaboration modality and user expertise[J]. Management Science, 2024, in press.
- [31] Pizzi G, Scarpi D, Pantano E. Artificial intelligence and the new forms of interaction: Who has the control when interacting with a chatbot? [J]. Journal of Business Research, 2021: 129878–129890.
- [32] Simon H A. A behavioral model of rational choice[J]. The Quarterly Journal of Economics, 1955, 69(1): 99–118.
- [33] Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability [J]. Science and Engineering Ethics, 2020, 26(4): 2051–2068.
- [34] Manz C C, Stewart G L. Attaining flexible stability by integrating total quality management and socio-technical systems theory [J]. Organization Science, 1997, 8(1): 59–70.

## Navigating agency tensions and governance in Generative Artificial Intelligence

YAN Lu – lu<sup>1</sup>, CHENG Cong<sup>1</sup>, GUO Chang – wei<sup>2</sup>

(1. School of Management, Zhejiang University of Technology, Hangzhou 310023, China;

2. Science and Technology, School of Management, University of Science and Technology of China, Hefei 230026, China)

**Abstract:** Generative Artificial Intelligence (GAI) has become deeply embedded in the socio-economic fabric of human society, and its growing agency has emerged as a focal point of both theoretical debate and practical concern. On the one hand, GAI increasingly exhibits capabilities traditionally regarded as uniquely human, including interactive intelligence, cognitive flexibility, and creative expression. These characteristics not only enhance GAI's functionality but also open new avenues for augmenting human cognition. On the other hand, the generative processes of GAI are frequently associated with factual inaccuracies, reasoning biases, and ethical dilemmas, leading to a fragile trust relationship between humans and machines. As a result, human reliance on GAI coexists with mounting skepticism, creating a fundamental tension between its empowering potential and the risks of technological unpredictability and moral ambiguity.

This paper investigates the question of what paradoxical tensions accompany the rise of GAI's agency and how these tensions may be addressed. We first identify that GAI's agency is characterized by three core features—interactivity, autonomy, and creativity—each of which is enabled by distinct forms of technological affordance. Interactivity is shaped by perceptual and communicative affor-

dances, autonomy by operational and situational affordances, and creativity by the interplay of interactive and situational affordances. While these features empower GAI with remarkable generative capabilities, they simultaneously give rise to complex system – level risks and trust challenges. Specifically, the high degree of interactivity leads to a lack of transparency in generative logic, producing a “black – box” effect that impedes user understanding. The rise in autonomy undermines users’ sense of control, as GAI’s decision – making processes and behavioral evolution become increasingly opaque and less predictable. Meanwhile, its expanding creative potential complicates the attribution of accountability, creating asymmetries in responsibility where the more novel the output, the more ambiguous its moral and legal ownership.

These contradictions manifest in three interconnected paradoxes: the interaction – transparency paradox, wherein GAI’s interactive flexibility conflicts with the opacity of its generative mechanisms; the autonomy – controllability paradox, where heightened decision – making independence erodes human oversight; and the creativity – accountability paradox, in which increasing innovation introduces diffuse and deferred responsibility. Addressing these challenges requires a shift in governance strategies: enhancing interpretability through explainable AI and knowledge mediation, reinforcing human oversight via participatory design and feedback mechanisms, and clarifying ethical boundaries through shared responsibility and robust accountability frameworks. Only by navigating these tensions thoughtfully can we ensure that the technological advancements of GAI genuinely serve collective human interests and foster a synergistic development of both technological and societal values.

**Key words:** Generative Artificial Intelligence (GAI); agency; affordance; trust; paradoxical tension

(上接第 314 页)

## Social construction mechanism and governance path of emerging technology risk

PANG Zhen – jing<sup>1,2</sup>, SHAO Er – xing<sup>1</sup>

(1. School of Public Administration, Sichuan University, Chengdu 610065, China;

2. China Institute for Science and Technology Policy at Tsinghua University, Beijing 100086, China)

**Abstract:** Emerging technology risks constitute a “perpetual topic” in the process of human scientific and technological innovation. As a kind of “processual existence,” it is urgent to clarify the process mechanism through which emerging technology risks are established from the perspective of social construction. Based on the social construction paradigm, this study constructs a processual explanatory framework, revealing that the social construction of emerging technology risks involves a process of “risk generation, risk rhetoric, and risk perception.” It also presents a derivation logic where risk states gradually “externalize, objectify, and socialize” as they connect across the levels of “macro society, meso – subjects, and micro – individuals.” Specifically, the generation of emerging technology risks has a clear “field” within the social system, giving rise to multiple types of risk patterns through systemic interactions. The rhetoric and interpretation of emerging technology risks are based on the strategies of multi – subject discourse construction and translation links, essentially representing a “discourse competition” for the power to interpret risks. The public’s perception of risks associated with emerging technologies is a specific risk perception state formed by the interweaving of the “objective knowledge path” and the “simplified mechanism of trust” under information stimulation. The governance of emerging technology risks urgently requires logical transformation and paradigm transcendence. From a macro – structural perspective, it is necessary to adhere to the balanced logic of “systematic equilibrium,” achieving “dimensional balance” and “intensity balance” among systems. From a meso – subject perspective, it is crucial to establish a collaborative logic for the division of labor among subjects, forming an “expanded knowledge community” to facilitate the integration of “public knowledge.” From a micro – individual perspective, it is essential to adhere to the neutralizing logic of cognitive adaptation, bringing the public’s perception of emerging technology risks “back” to a non – polarized rational space.

**Key words:** emerging technology risks; emerging technology governance; social construction; risk perception