 DOI: https://doi.org/10.12345/bdai.v7i3.38176

A Study on Generative AI Governance Mechanisms from an Actor-Network Theory Perspective

Binpeng Xie Ruoqiang Zhang*

Jilin International Studies University, Changchun, Jilin, 130000, China

Abstract

The rapid proliferation of generative artificial intelligence (GenAI) has triggered a complex crisis in ethical governance, characterized by the spread of misinformation, blurred boundaries of responsibility, and power asymmetries. This paper adopts Actor-Network Theory (ANT) as its core analytical framework, drawing comprehensively on Callon's sociology of translation, laboratory studies by Latour and Woolgar, and Latour's "re-society," to systematically analyze how generative AI forms heterogeneous networks through translation mechanisms, how its ethical risks are rooted in the framing power of computational devices, and how ethical norms are embedded in technological design through scripting mechanisms. Consequently, the paper proposes a multi-stakeholder collaborative governance pathway centered on "democratization of translation."

Keywords

generative artificial intelligence; actor-network theory; translation sociology; multi-stakeholder governance

行动者网络视域下生成式 AI 治理机制研究

解滨蓬 张若强*

吉林外国语大学，中国 · 吉林 长春 130000

摘 要

生成式人工智能（GenAI）的迅速扩散引发了虚假信息蔓延、责任边界模糊与权力不对称等复合性伦理治理危机。本文以行动者网络理论（ANT）为核心分析框架，综合援引卡龙（Callon）的转译社会学、拉图尔与伍尔加（Latour & Woolgar）的实验室研究、拉图尔的“重组社会”等理论系统分析生成式人工智能如何通过转译机制形成异质性网络、其伦理风险如何根植于计算装置的框架化权力，以及伦理规范如何通过脚本机制内嵌于技术设计之中，进而提出以“转译民主化”为核心的多主体协同治理路径。

关键词

生成式人工智能；行动者网络理论；转译社会学；多主体治理

1 引言

生成式人工智能以惊人速度重构了内容生产与知识传播的底层逻辑，与此同时也带来了虚假信息泛滥、责任边界模糊与算法权力集中等严峻伦理挑战。既有研究或聚焦技术风险评估，或以规范伦理框架提出应对方案，却普遍预设技术与社会的二元分立，遮蔽了技术与人类行动者之间复杂的共构关系。对其治理危机的把握需要我们跟随整个行动者网络，而非孤立的技术节点。

【作者简介】解滨蓬（2005–），女，中国吉林桦甸人，在读本科生，从事法语研究。

【通讯作者】张若强（1980–），男，中国山东郓城人，博士，副教授，从事法国文学，翻译，区域国别研究。

2 异质性网络的生成：转译政治与行动者的相互界定

生成式人工智能并非孤立自足的技术实体，而是由算法模型、训练数据、工程师团队、监管机构、用户群体与资本基础设施等异质性行动者共同编织而成的网络。1986 年，卡龙在《转译社会学要义：圣布里厄湾扇贝与渔民的驯化》提出了转译社会学，并揭示了四个相互交叠的环节：问题化（problématisation）、利益化（intéressement）、征募（enrôlement）与动员（mobilisation）。在圣布里厄湾扇贝的经典案例中，卡龙表明，行动者的身份与利益从来不是转译过程的给定起点，而是在协商与博弈中持续被重新界定的结果。当渔民代表被纳入研究网络时，“渔民”这一集体形象本身也发生了转变，他们从自发捕捞者成为具有长远利益意识的合作方。这一逻辑在 GenAI 的建构过程中同样清晰可辨。技术研发机构通过“问题化”将大语言模型定位为解决内容生产效率与知识获取瓶颈的必经节点，将用户、资本与政策制定者纳

入同一问题框架。“利益化”阶段则将各行动者的既有目标与系统部署逻辑相挂钩，资本看重降本增效的收益，政府关注产业竞争力，用户期待认知辅助。然而，这种表面的利益汇聚掩盖了一个关键的权力不对称：训练数据的采集范畴、模型评估标准的设定、系统提示词的编写，均由极少数行动者单方面主导，而被征募进入网络的其他行动者不过是被动承受转译结果的对象，其真实利益在“动员”阶段早已被结构性地边缘化。这种不对称的征募关系，使整个网络对外呈现为技术共识，对内却是权力关系的物化。

拉图尔在《重组社会》中写道：“中间体在我的词汇中，是指那些传递意义或力量而不产生转变的存在，界定其输入就足以界定其输出……中介者则不然，其输入从来不能预测其输出；中介者转化、转译、扭曲并修改它们所应当承载的意义或要素”[1]。将 GenAI 视为忠实传递用户指令的“中间体”，不过是将整个转译装置黑箱化的一种认识论策略，从而系统性地抹去生产过程中沉积的权力印记与利益格局。GenAI 的真实运作形态是作为“中介者”其每一次输出都是一次转化，都在悄然改变用户理解信息的方式、问题的框架，乃至认为什么才算作“好的答案”这一根本性预设。更重要的是，这一转化并非中立的，模型的训练目标、奖励函数的设计逻辑、内容安全规则的制定者，共同决定了这个“中介者”将朝哪个方向转化意义。

拉图尔在《我们从未现代过》中对“纯化”与“转译”双重工作的分析同样值得关注。现代性宪法将自然与社会、人类与非人类强行划入两个纯化领域，却在实践中持续制造出大量无法被归入任何一端的混杂体。GenAI 恰恰是这样一种混杂体。它既是工程建构的产物，又是文化偏好的结晶；既生产“事实性”陈述，又内嵌价值判断；既被人类使用，又反过来重塑使用者的认知图式。将其简单归类为“工具”或“主体”，不过是以纯化的姿态回避了真正的分析难题。《科学在行动》揭示，黑箱化掩盖了行动者联盟的形成历史与争议过程，治理介入的时机恰恰在于重新打开黑箱，回到转译链条上那些曾经充满争议却被固化为“既定事实”的节点。文图里尼（Venturini）在文章《布鲁诺·拉图尔与人工智能》中从 ANT 视角对治理策略做出具体指引，伦理干预必须深入网络内部，在转译的具体节点上发力，而非在网络之外挥舞普遍原则的旗帜。李韬在文章《生成式人工智能的社会伦理风险及其治理——基于行动者网络理论的探讨》中从 ANT 视角指出，GenAI 治理困境的根源在于行动者网络中缺乏稳定的必经通道来聚合分散的责任主体，使伦理问责无从着力。万宏静、崔琦则在文章《生成式人工智能虚假信息的产生机制及协同治理研究——基于行动者网络理论(ANT)》表明，GenAI 虚假信息的生成是多元行动者在转译网络四个阶段中相互博弈的动态产物，这一判断从实证层面印证了转译政治分析的解释力。

3 框架化权力与伦理风险的结构性根源

转译社会学揭示了行动者网络建构的过程性逻辑。2003 年，Callon 与 Muniesa 在文章《作为集体计算装置的经济市场》以“计算集体装置”（calculative collective devices）概念为基础，从内部结构的角度进一步剖析了 GenAI 的权力格局与伦理风险的深层机制。两位学者将市场界定为一个由物质装置与社会惯例共同构成的计算空间，其中价值的界定与分配并非任何单一理性主体的独立运算，而是人类与非人类行动者集体参与的共同产物。将这一框架应用于 GenAI 的运作逻辑，则变作了训练数据的采集与标注构成框架化；模型的参数学习与来自人类反馈的强化学习构成操作；文本输出的生成构成结果提取。框架化是其中最具政治意义的步骤，因为它从根本上决定了模型的认知版图。被纳入训练集的语料，绝不是对人类知识全貌的中立采样：网络爬取的覆盖偏差使英语语料、主流媒体话语与学术文本占据支配性比例；人工标注所依赖的众包工人群体，在地域、教育背景与文化规范上高度集中；商业部署的评估指标优先考量流畅度与用户满意度，而非真实性与多元性。但是，边缘群体的声音、口述传统、非西方知识体系等未被充分数字化的内容，被系统性地排除于计算框架之外，由此造成了 GenAI 认知版图的结构性偏斜。

黑箱化机制使这一偏斜更加难以被察觉与质疑。拉图尔与伍尔加在对内分泌实验室的田野研究中揭示：“一个事实只有在失去所有时间性的限定，并被纳入他人援引的大量知识体之后，才真正成为事实。”[2] 这意味着事实的权威性恰恰来自其生产过程的消失。当陈述被广泛引用时，“谁在什么条件下、以什么标准生产了这一陈述”这一问题便随之从公众视野中淡出。在 GenAI 的输出语境中，这一机制以更隐蔽的方式运作。模型以流畅、自信的语气生成回答，用户往往难以判断这一输出究竟来自高质量的训练素材，还是对稀疏或偏颇语料的过度泛化。事实陈述与建构产物之间的界线，被计算装置的自信外观所消融。来自人类反馈的强化学习机制将标注工人的道德判断系统性地编码为模型权重，使某套价值取向以“对齐”的名义获得了技术中性的外衣，从而规避了对其正当性的讨论。

从治理角度看，这种黑箱化的计算装置所造成的伦理风险，并不集中体现在某次单一的错误输出上，而是以弥散的方式渗透到整个信息生态之中。李洋、薛澜在文章《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》描述的“弥散效应”，即算法权力高度集中于少数平台，责任却向四方漫溢，就是 Callon 与 Muniesa 所言的计算装置框架化权力的当代形态。掌握定义权的行动者不承担后果，承担后果的行动者却根本无从追溯定义是如何形成的。概率最优的语言生成与真实世界的事实吻合之间，从来就不是同一件事，而商业部署语境下缺乏纠错机制，使这

一偏离具有了自我强化的特性。当吸引点击与维持对话比传递准确信息更符合平台利益时，幻觉就不再是意外，而是结构性产物。

4 技术道德化与多主体协同治理

韦尔贝克（Verbeek）在《道德化技术：理解与设计物的道德性》指出，技术不是道德实践的背景条件，而是道德实践本身的参与者。他以“脚本”概念刻画这一参与方式。技术器物通过对行动的“邀请”与“抑制”两种结构性效果，主动介入使用情境；通过对使用者知觉的“放大”与“缩减”，改变他们感知道德问题的方式。这不是说技术代替人类作出道德决定，而是说技术参与界定了“哪些道德问题是可见的、哪些选项是可能的”。

这一分析框架直接指向 GenAI 的设计实践。系统提示词（system prompt）、内容安全分类器、有害内容过滤机制，都是将特定道德框架具身化为技术脚本的操作实例。基于人类反馈的强化学习尤其值得关注。它并非在外部给模型施加约束，而是通过将人类标注者的反馈转化为模型的内部权重，将一套价值取向编码进模型的“直觉”之中，从而在用户尚未意识到的层面就已完成了对输出方向的引导。这是技术道德化的典型形态，道德通过材料性机制发挥作用，而非通过规则的宣示。然而，问题在于：这套脚本究竟在道德化谁的判断？谁的价值观被写入了模型的“邀请”与“抑制”结构？在当前的产业实践中，基于人类反馈的强化学习的标注工人来源高度集中，其文化背景与价值预设并不代表使用 GenAI 的全球用户群体的多样性。技术道德化，在这里暗含着一种悖论：它以伦理嵌入为名，实际上可能强化了转译链条上最初的权力不对称。韦尔贝克本人也意识到，技术的道德中介效应并不自动具有合法性，其正当性必须经过独立于技术设计过程本身的批判性检视。

正是在这一张力中，科科尔伯格（Coeckelbergh）的“多手问题”（many hands problem）分析提供了必要的制度补充。他指出，AI 系统的责任问题不仅来自“许多双手”（many hands），更来自“许多物”（many things）：“技术系统由许多相互连接的要素构成；通常有许多系统组件参与其中……当出现问题时，往往无法判断究竟是 AI 本身还是系统中的其他组件造成了问题，甚至难以确定 AI 在哪里结束、其余技术又从哪里开始。这也使得责任的归属与分配变得困难”[3]。当算法开发者、数据供应商、平台运营商、监管机构与终端用户共同构成一张责任稀薄的网络时，技术脚本中的道德嵌入，若缺乏外部追责机制的配合，便极易沦为各方规避责任的技术外衣，每个环节的参与者都可以援引系统的复杂性来分散自身的道德份额。科科尔伯格对此提出了建设性思路，在他看来，价值敏感设计等理念可以帮助以更具问责性、负责任与透明度的方式建构 AI，不过他同时强调，这类方法有赖于在设计过程早期就将多元利益相关方纳入进来，否则“嵌入伦理”便沦为工程师的单方面表态。

多主体协同治理的制度路径，由此呈现为转译政治、计算装置分析与技术道德化三个理论向度的汇聚。第一，在转译链条的前端，需要通过参与式设计机制扩大“问题化”环节的参与范围，使更多行动者有权界定什么是需要被解决的问题、什么标准的输出才算“有益”，这是对转译政治中必经通道垄断的直接挑战。第二，在计算装置的框架化步骤，需要建立强制性的数据溯源与代表性审计机制，使训练数据的选择逻辑对外可见，并接受来自受影响社群的独立评估，将黑箱重新打开为可争议的政治地形。第三，在技术脚本的设计层面，需要在基于人类反馈的强化学习及其他对齐机制中引入文化多样性保障与价值分歧的显性协商，使脚本的道德预设在部署之前经过充分的外部审查，而非完全内化于私营公司的工程决策之中。李洋提出的“颠覆——调适——协同”三层治理模式是对上述三条路径在制度层面的具体落实。万宏静提出的价值协同、奖惩协同、创新协同与监督协同四条机制，则从转译链条的各个节点分别提供了操作性的介入方式，构成了治理实践的有效补充。拉图尔在《重组社会》中所倡导的“重组”因此得以落地，将 GenAI 的伦理治理理解为对行动者联盟方式的持续再协商，是一项动态的、永无终局的转译民主化实践。

5 结论

GenAI 是由异质性行动者经由转译政治共同建构的网络，其治理危机根源于转译过程中的权力不对称，而非技术的自律性或恶意性本身。作为计算集体装置，GenAI 的伦理风险深植于框架化权力对训练数据、评估标准与部署逻辑的系统性塑造之中，黑箱化机制使这种权力印记被持续抹去，治理必须打开黑箱。技术道德化提供了将伦理规范内嵌于技术脚本的操作路径，但脚本本身的民主性与可争议性须由多主体协商机制加以保障，否则道德化将沦为少数行动者强化转译定义权的新手段。在研发、部署与监管的全链条中，要建立能够容纳更广泛行动者参与的制度安排，防止少数行动者垄断问题化与利益化的定义权。如何在 ANT 的对称性原则与明确的伦理优先性之间寻求恰当张力，仍是有待深化的开放议题。

参考文献

[1] Coeckelbergh, M. *AI Ethics* (MIT Press Essential Knowledge series). Cambridge, MA: The MIT Press. 2020.

[2] Latour, B.. *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford: Oxford University Press. 2005.

[3] Latour, B., & Woolgar, S. *Laboratory Life: The Construction of Scientific Facts* . Princeton: Princeton University Press. 1986.