# 超越外部视角：
# 大语言模型时代社会科学的角色定位

郁建兴　刘宇轩

**摘　要**　大语言模型（LLM）正在深度嵌入公共治理、司法辅助、福利分配等社会科学核心领域，迫使社会科学重新审视自身的目标使命。当前社会科学与 LLM 的交叉研究深陷三重误区：将模型暂态缺陷误读为结构性社会规律，迷信被操纵的评测榜单，以小模型局部优化充当普遍理论。这些误区产生了大量注定随算力迭代自动消亡的“算法伪影”。对此，可以从两个维度划定工程命题与社会科学命题的边界：技术上能否被算力消解，以及商业逻辑是否驱动主动修复。两条判据可以交叉凝练为一个可操作的四象限试金石矩阵。在此基础上，需要越过工程边界，在权力、价值与制度三个维度确立社会科学的研究坐标，去解决自由裁量权的语言性消解，公共价值在优化压力下的选择性消失，以及 LLM 引入“第四只手”后问责链条的断裂与重建等问题。该坐标隐含着社会科学居于人工智能“外部”进行批判的前提，导致社会科学论证的结构性局限，从而决定人工智能的“外部”视角本身需要被超越。

**关键词**　社会科学　大语言模型　人工智能　共同生产

作者郁建兴，浙江大学公共管理学院教授，教育部共同富裕统计监测与智能治理实验室研究员（浙江杭州 310058）；刘宇轩，华东理工大学社会与公共管理学院讲师（上海 200237）。

**中图分类号** D0　　**文献标识码** A　　**文章编号** 0439-8041(2026)07-0077-13

在细胞科学的研究史中，存在一种名为“伪影”（Artifacts）的认识论幽灵。生物样本在切片、固定或染色等工程处理过程中，由于化学药剂的浓度偏差或操作环境的物理压力，细胞内部常会产生一些原本并不存在但貌似绚丽的条纹、结晶或空泡。在学科发展早期，这些“伪影”常被研究者误认为是对微观世界的全新构造，并为此构建出一整套看似自洽的机制理论，其中最经典的案例是细菌“中间体”（Mesosome）[①]。20 世纪 60 年代，这一由化学固定技术在制备电子显微镜样本时产生的质膜折叠内陷结构，被全球数十个重要微生物学研究团队视为细菌的关键细胞器，并围绕它产生了涵盖细胞壁形成、染色体复制与氧化磷酸化的完整功能假说。然而，当 Ebersold 等人采用冷冻固定替代化学固定后，“中间体”从“细菌的线粒体类似物”变成了“渗透压损伤的膜褶皱”[②]。显然，这些学术发现的寿命随着底层工程工具的改良走向终结。

目前，当大语言模型（LLM）强烈冲击现有知识体系时，这种“伪影”的认识论幽灵同样悄然降临社

① Nicolas Rasmussen, “Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope,” *Studies in History and Philosophy of Science Part A*, Vol. 24, No. 2, 1993, pp. 227–265.

② Hans Rudolf Ebersold, Jean-Louis Cordier and Peter Lüthy, “Bacterial Mesosomes: Method Dependent Artifacts,” *Archives of Microbiology*, Vol. 130, No. 1, 1981, pp. 19–22.

会科学领域。在社会仿真研究中，LLM在角色扮演上具有先天优势。同时，它们在部分专业任务上表现出的能力不足，也让许多研究者尝试寻找其背后的结构性隐患。然而，无论出发点是对其能力的利用还是对其局限的审视，一旦将LLM生成的文本模式、预测结果或交互行为直接等同于真实的社会现象与人类行为规律，就可能陷入同一个认识陷阱：那些被奉为“机器心智规律”的经验往往只是算法在通往工程完美过程中的“技术碎屑”。大模型输出的某些偏差在社会科学研究者眼中，可能是深刻的文化偏见、社会排斥或人类认知局限的“硅基映射”。但从算法底层看，这些偏差多半只是预训练语料分布不均、对齐策略权重偏差，或模型在高维空间尚未收敛仅呈现为暂态表现。当研究者努力调用经典理论框架去诠释某个特定模型版本所表现出的“认知偏误”或“逻辑断裂”时，科技巨头往往只需在下一个版本的迭代中扩大参数规模、清洗微调数据，或是修改一行正则化系数，便会让那些宏大的“学术发现”瞬间蒸发。这就暴露出了目前交叉学科研究中的一个致命盲区，即“工具应用的工程局限”与“社会科学的理论发现”之间的边界极其模糊。如果社会科学仅仅是为这些注定消亡的“算法伪影”编写生命志，科研人员在实质上就会降级为工业界的“外部免费测试员”。

社会科学在LLM时代必须重建独立的研究立场，这正是本文所关注的问题。本文的论述逻辑沿四个层次递进展开：首先，剖析现有社会科学与LLM交叉领域经验研究中的工程伪影，厘清“问题池”中那些注定会被算力填平的工程议题；其次，划定工程与社会科学之间的认识论边界，提出区分工程命题与社会科学命题的试金石法则，为后续的研究提供参照；再次，尝试越过工程边界，以外部审视者视角在权力、价值与制度三个维度上确立社会科学的研究坐标；最后，论证社会科学与人工智能（AI）并非外部批判者与被批判对象的关系，而是在知识生产与社会秩序建构中相互构成的关系，其本质是两者互为前沿。超越外部视角构成了社会科学的新角色定位。

## 一、经验研究中的工程陷阱

常规科学研究的大部分工作是在既定范式下进行解谜。[①] LLM时代的社会科学则面临着一种新的危机，它不是无谜可解，而是“问题池”中存在大量注定会随技术迭代自动消亡的伪谜题。LLM时代的大量社会科学“伪影”正在挤占真正有价值的研究空间。总的来说，当前的经验研究深陷三个误区。

### （一）将暂态工程噪声误读为社会规律

计算机科学家萨顿（Sutton）曾总结出一条苦涩的教训：学界总是倾向于用精心设计的复杂方法解决问题，但历史一再证明，大规模算力配合简单的通用方法总能击败那些手工设计的专家系统。LLM时代的社会科学研究恰恰正在重蹈这一覆辙。许多昂贵的学术资源被投入交叉领域的研究，以图证明某个大模型在特定文化语境下存在幻觉、性别偏见或逻辑推理断裂，并试图为这些问题寻找结构性解决方案。但是，在工业界工程师看来，这些被社科学者奉为圭臬的“认知缺陷”，有些仅仅是预训练数据分布不均或对齐策略强度不足所导致的暂态噪声。

根据AI领域的规模法则，在适当的训练配比与足够大的尺度范围内，LLM的预训练损失通常可以被参数规模、训练数据量和计算量的幂律函数较好地近似，因此模型扩展在经验上具有一定的可预测性。[②] 这一规律的现实含义是：只要持续堆叠算力，模型能力就会可预期地持续提升。当前，各大科技公司正是在这条确定性的曲线上展开军备竞赛，如GPT、Claude、Gemini等一代代模型的迭代，背后都是数以百亿计算力投入的直接兑换。在这场竞赛中，今天被研究者记录为认知局限的模型行为，往往只是下一个训练周期即将填平的坑洞。这一规律使得大量论文的结论极度脆弱。有研究对比了GPT-4在仅相隔三个月的两

① 托马斯·库恩：《科学革命的结构》，金吾伦、胡新和译，北京：北京大学出版社，2003年，第29页。

② Jared Kaplan, Sam McCandlish, Tom Henighan et al., “Scaling Laws for Neural Language Models,” arXiv: 2001.08361, 2020年, https://arxiv.org/abs/2001.08361, 2026年4月2日。

个版本上的表现，发现同一模型在质数识别等任务上的准确率出现了从 84% 到 51% 的剧烈漂移。[①] 另一项研究在六个主流模型上检验了此前受到广泛认可的提示词工程技术，然而几乎所有被声称有效的技术在统计上均无显著差异。[②] 因此，在社会科学研究中费尽心力去探测的那些 LLM“认知偏误”问题，可能会如细胞科学中有关“伪影”的研究一般，会在下一代模型扩大算力规模时被自动填平。更危险的是，当旧版闭源模型被服务商下架后，该版本的模型幻觉等问题在事实上已经不可验证，这就导致了研究对象本身成为一个不断移动的靶子。

（二）度量幻觉与被操纵的评测黑盒

在本体论层面的误判之外，研究者在操作层面同样面临深刻的方法论危机。当前交叉研究的另一大陷阱便是对大模型“涌现能力”的崇拜。大量研究声称，当模型参数规模越过某一阈值时，机器会突然涌现出复杂的“心智理论”（Theory of Mind）或高阶社会博弈逻辑。然而，这种阶跃式的“涌现”，在很大程度上是度量指标选择不当所制造的幻觉。当研究者使用精确匹配率等非连续指标时，模型能力的平滑线性增长在数学上被硬性截断，从而在图表上呈现出伪装的“突变”。当研究人员改用连续性度量指标之后，所谓涌现便消失无踪。[③] 其本质在于用刚性的离散指标去裁剪连续的技术演进，无意间制造出了符合技术崇拜叙事的学术幻象。

古哈特法则表明，当一项衡量指标成为目标时，它就不再是一项好的衡量指标。这一法则在工业界的评测场景中表现得尤为突出。训练集与测试集之间的数据污染问题在机器学习研究中极为普遍，这一方法论缺陷使得大量将机器学习应用于科学问题的研究结论完全失效。[④] 工业界对评测榜单的“应试教育”也是一个严重问题。有调查研究发现，多家头部科技公司在正式发布模型前私下测试大量变体，Meta 在 Llama 4 发布前测试了多达 27 个私有变体，却仅选择性地公开表现最好的版本。[⑤] 科技巨头通过垄断测试数据的披露权，实质上掌控了定义何为先进模型的话语权。盲信这些榜单的经验研究会导致研究人员忽略繁杂指标背后真正的社会科学问题。由是，一种矛盾的现象就会出现：社会科学的研究者在对于模型“涌现”能力无比推崇的同时，却又相信 LLM 那些阶段性的瑕疵是结构性的。

（三）小模型的局部优化无法推广为普遍理论

当前社会科学还深陷于工具论误区。本文的批判并非靶向小模型的使用本身。开源小模型能够锁定版本、控制实验条件，是保障研究可复现性的合理选择。[⑥] 真正的问题在于另一种做法：在特定小模型上进行大量领域知识级别的工程优化，包括精心设计的提示词链、定制化微调策略、复杂的多智能体规则，并将这些局部优化的成果推广为关于 LLM 能力的泛化主张。这种做法的根本问题是：小模型上的结论是否可以外推到大模型，本身就不确定。当工业界基础模型跨越下一个规模量级时，凭借简单通用的预测机制与庞大算力，往往能够轻松超越小模型所能达到的理论性能上限。有研究系统评估后发现，研究者在小模

① Lingjiao Chen, Matei Zaharia and James Zou, “How Is ChatGPT's Behavior Changing over Time?” arXiv: 2307. 09009, 2023 年, https: //arxiv. org/abs/2307. 09009, 2026 年 4 月 2 日。

② Laurène Vaugrante, Mathias Niepert and Thilo Hagendorff, “A Looming Replication Crisis in Evaluating Behavior in Language Models? Evidence and Solutions,” arXiv: 2409. 20303, 2024 年, https: //arxiv. org/abs/2409. 20303, 2026 年 4 月 2 日。

③ Rylan Schaeffer, Brando Miranda and Sanmi Koyejo, “Are Emergent Abilities of Large Language Models a Mirage?” in *Advances in Neural Information Processing Systems*, Vol. 36, 2023, https: //papers. neurips. cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference. pdf, accessed April 7, 2026.

④ Sayash Kapoor and Arvind Narayanan, “Leakage and the Reproducibility Crisis in Machine-Learning-Based Science,” *Patterns*, Vol. 4, No. 9, 2023, 100804.

⑤ Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo et al., “The Leaderboard Illusion,” in *Advances in Neural Information Processing Systems*, Vol. 38, 2025, https: //proceedings. neurips. cc/paper _ files/paper/2025/file/70a93f260a51123b3c0e33ecd1b4de97-Paper-Datasets_and_Benchmarks_Track. pdf, accessed July 15, 2026.

⑥ Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J. Xue, Jackson Trager, Peter S. Park, Preni Golazizian, Ali Omrani and Morteza Dehghani, “Perils and Opportunities in Using Large Language Models in Psychological Research,” *PNAS Nexus*, Vol. 3, No. 7, 2024, p. 245.

型上投入大量微调工作后获得的性能提升，在大模型上仅需零样本提示即可达到甚至超越。[①] 也即是说，在小模型上耗费的全部领域优化努力，不过等价于大模型的默认能力，而非具有理论意义的发现。因此，对小模型的局部优化并不能带来普遍性、可复制推广的价值。

显然，社会科学在工具层面存在深层困境。社会科学研究者面对三重约束：前沿大模型由少数科技巨头掌控，数以亿计的训练成本使学术界根本没有能力独立复现或持续调用；闭源商业API虽然能力强，却因随时可能停服或更新版本而无法保证研究的可复现性；开源小模型虽然可控，却在泛化能力上存在无法回避的上限。三条路径都有各自的致命缺陷。实际上，社会科学研究者被锁在一个没有完美选项的工具困境之中。更根本的问题在于，这个困境源自AI工业与社会科学学术研究之间在资本、算力与基础设施上的结构性不对等，它不是资金投入增加就能解决的。在这种不对等下，任何基于现有可及工具所得出的“LLM能力规律”都必须附加一个沉重的注脚，即它们只是在当前可及的工具条件下观察到的现象，而非关于LLM行为的普遍理论。这无疑是社会科学在方法论上需要直面的结构性困境。

## 二、工程与社会的认识论边界

经验研究的三大误区并非相互独立的偶发情况，其共性的根源在于研究者混淆了工具的工程局限与社会科学的理论发现。因此，社会科学要摆脱工程“伪影”就必须退后一步，去厘清什么样的问题本质上是一个注定收敛的工程问题，什么样的问题才属于社会科学的无限空间？这一问题是根本性的，并具有行动指向性。

### （一）工程是有限域，社会科学是无限域

无论LLM的参数规模膨胀到何种地步，其背后的工程逻辑始终被收敛性与最优化所统摄。在算法工程师眼中，减少幻觉、提升多步推理准确率等任务，都可以被降维成寻找损失函数极小值的封闭数学问题。这是工具理性对价值理性的极端取代。韦伯在一个世纪前就已经为这两种理性划定了不可互相僭越的边界。只要算力、数据与算法架构持续投入且遵循规模法则，暂态缺陷就存在一条可见的、渐进解决的确定性路径。大量社会科学意义上的偏差、幻觉和推理断裂，在工程师眼中不过是损失曲面上尚未被梯度下降填平的坑洞，而非有待探究的社会规律。

因此，计算算法的优化领域是一个“问题边界可勾勒，解决路径可预期”的有限域，进步形态是单调的。韦伯在讨论社会科学方法论时就已指出，经验现实的丰富性是无穷无尽的，任何有限的概念体系都不可能穷尽社会现实的全部面向。[②] 弗利夫伯格进一步将这种区别概念化为理论知识与实践智慧的根本差异：工程领域追求的是普遍适用的规则知识，而社会科学处理的是根植于具体历史情境、充满价值争议、无法脱离权力关系而存在的实践问题。[③] 价值的不可公约性、权力的情境依赖性、制度的历史积淀性，都不是工程意义上的“待填坑洞”，而是社会现实本身的构成性特征。

值得注意的是，即便在工程内部，有限域的边界也并非如工程师所宣称的那般整洁。以价值对齐为例，它试图通过技术手段使模型输出符合人类偏好的工程方案，但在数学层面已经遭遇了内在的结构性困难。即便工程师有充分的意愿去做好对齐，当前的数学架构内生性地会将某些人群的偏好排挤出优化目标。RLHF（基于人类反馈的强化学习）的KL散度（相对熵）正则化结构会导致“偏好崩塌”。少数群体的声音在梯度

① Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova and Fabrizio Gilardi, “Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning,” arXiv: 2307.02179, 2023年, https://arxiv.org/abs/2307.02179, 2026年4月2日。

② 马克斯·韦伯：《社会科学认识和社会政策认识中的“客观性”》，《社会科学方法论》，韩水法、莫茜译，北京：中央编译出版社，1999年，第1—61页。

③ Bent Flyvbjerg, *Making Social Science Matter: Why Social Inquiry Fails and How It Can Succeed Again*, Cambridge: Cambridge University Press, 2001, pp. 53–65.

下降中被系统性压制，改变正则化项可以将这种崩塌减少。[①] 社会科学研究中所指出的“模型对非西方文化语境的理解偏差”，在相当程度上不过是这一数学结构的必然结果，而非什么值得专门研究的认知规律。这恰好说明，将价值问题交由工程优化来解决，不只是一种方向上的错误，连技术路径本身也可能是有缺陷的。

这一区分直接提供了区分工程命题与社会命题的第一条判据：一个现象能否通过增加算力、扩大参数规模或优化训练目标而得到根本性消解？能够被工程进步填平的，属于有限域；根植于价值冲突与权力结构的，属于无限域。

### （二）价值对齐的幻觉与商业逻辑的遮蔽

有限域与无限域的区分解决的是技术上“能不能消解”的问题。然而，即便某个缺陷在技术上属于有限域，它是否会被修复，还取决于一个完全独立的问题，即谁有动力去修复它？

理解这个问题需要从科技巨头的商业逻辑入手。大型 AI 企业的本质是资本驱动的商业实体，其产品研发的优先级排序归根结底遵循的是利润逻辑，而非公共价值逻辑。这两种逻辑在大多数时候存在部分重叠，但重叠之外依然存在一个由负外部性的可见度所决定的巨大盲区。

仇恨言论是最典型的“高修复优先级”案例。它会直接触发媒体报道、广告商抵制和监管压力，对公司品牌价值和股价形成的冲击可以在数小时内被量化。对科技公司而言，这不是道德问题，而是风险管理问题。资源会迅速向这类可见缺陷聚集，因为修复它的回报清晰且即时。然而，隐秘地破坏基层治理的缺陷与此截然不同。嵌入福利审批的 LLM 会系统性地把边缘群体评为高风险；推荐学习路径的教育 AI 会复制阶层壁垒；一个处理少数民族语言请求的政务助手会给出更低质量的回应等等。这些缺陷造成的损害是分散、长期、难以被归因于单一事件的。受害者往往无法识别自己受到了算法的差异性对待。即便他们能够识别，也难以形成足够强度的舆论压力，更难以追溯到具体的设计决策。吊诡的是，这些损害不会出现在任何季度财报的风险项中，也不会被任何现有的基准测试所捕捉。

这种结构性的盲区不是大厂工程师疏忽的结果，而是商业逻辑的内在产物。修什么、多快修、投多少资源修，这些本质上都是资本配置决策，它受制于可见损失规模与修复投入的预期回报比。由此产生的后果是，AI 系统的缺陷修复形成了一个高度偏向的优先队列。那些对主流用户可见的缺陷排在前面，而对边缘群体造成损害的缺陷被系统性地推迟，甚至永久搁置。

从公共治理角度看，修复优先级的扭曲是一次深刻的权力转移。在传统的行政法逻辑中，政府对所有公民承担平等对待的法律义务，它不因公民的商业价值高低而有所差别。当公共职能被委托给商业 AI 系统之后，平等对待原则就被悄然替换为基于可见性与商业价值的差异化原则。这种深层治理逻辑的隐秘转变意味着，技术的可解决性正在悄然替代公共治理的可问责性。揭露结构性盲区绝非替大厂找 bug，而是对商业资本的社会负外部性展开“对抗性审计”（Adversarial Auditing）。核心追问不再是测评榜单上的准确率，而是权力分配的终极逻辑：谁的反馈被定义为人类的反馈？谁的损失被计入了损失函数？谁的生存空间又被排斥在最优解之外？

由此，本文分析的第二条判据得以形成：科技巨头是否具备充足的商业动力去主动修复它？价值对齐的幻觉和商业模式的遮蔽揭示了，即便某个缺陷在技术上属于有限域，只要商业修复动力不足，它同样会被系统性遗留，并不因为“技术上可以解决”就真的会得到解决。技术的可能性与社会科学的现实性之间存在一道深刻的鸿沟。而这道鸿沟正是社会科学研究最应当持续关注之所在。

### （三）区分工程命题与社会命题的试金石

以上论述确立了判断一个问题归属的两个独立维度：它在技术上是否属于可被算力消解的有限域，以及商业逻辑是否会驱动对它的主动修复。两条判据交叉，形成一个判断矩阵（见图 1）。

---

① Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, Weijie J. Su, “On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization,” arXiv: 2405. 16455, 2024 年, https: //arxiv. org/abs/2405. 16455, 2026 年 4 月 2 日。

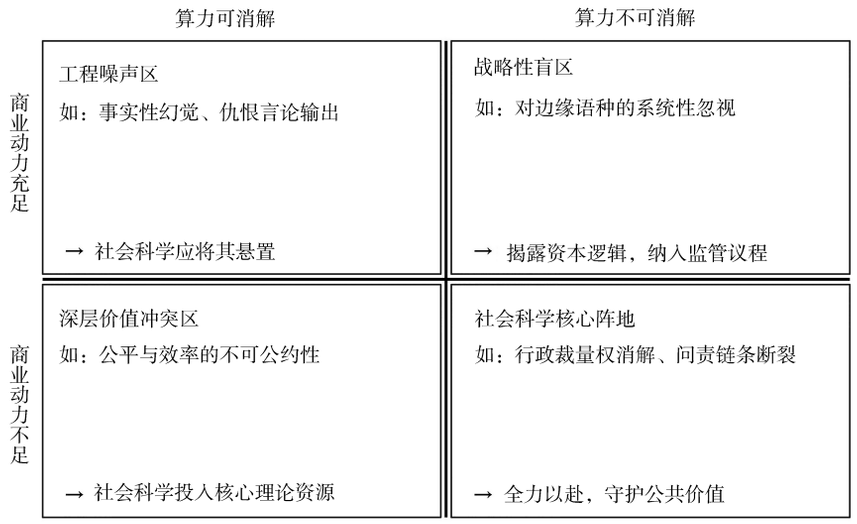


**图1 试金石矩阵图**

判据一：算力可消解性。该现象能否通过增加参数规模、提升数据质量或优化奖励函数予以彻底消除？

判据二：商业修复动力。科技巨头是否具备充足的商业动力去主动修复它？

两条判据交叉形成四个象限。算力可消解且商业动力充足的（如事实性幻觉、仇恨言论输出），是纯粹的工程优化命题，社会科学理应将其悬置。算力可消解但商业动力不足的（如对边缘语种的系统性忽视），社会科学的使命是揭露这种战略性无视的资本逻辑，推动将其纳入公共监管议程。算力不可消解但商业动力充足的（如公平与效率的不可公约性），根植于价值的本体论冲突，社会科学应在此处投入核心理论资源。算力不可消解且商业动力不足的（如算法对行政裁量权的隐性剥夺、人机协同中问责链条的断裂），是社会科学应全力以赴的研究空间，也是公共价值最脆弱、最需要学术守护的领域。

这一矩阵的核心逻辑可以简洁表述为：若一个现象恰恰是因为算法变得更强大、更深度地嵌入公共治理，才得以衍生或加剧，或因商业逻辑的战略性无视而被系统性遗留，这才是社会科学应当全力以赴的无限空间。

## 三、越过工程边界：社会科学的研究坐标

在清理了“问题池”中的工程伪影，并划定了工程与社会问题的认识论边界之后，可以发现社会科学面对的并非算力挤压下的技术荒原，而是一个广阔的真命题域。这种对于社会科学主权的收复，要求我们从对算法性能参数的追逐，转向对算法嵌入社会系统后所引发的结构性变迁的洞察。

在展开具体的社会科学研究之前，研究者首先需要回答一个前置问题：权力、价值与制度维度下的算法治理问题，在传统深度学习时代就已存在，为什么需要在LLM时代重新进行专门的锚定？

这里的答案在于，LLM相比传统算法带来了两个质变。其一，语言生成能力导致问责结构改变。传统算法的黑盒问题是无法解释的，往往只通过系统给出决策结果，但又缺乏相应的说明。LLM的黑盒问题则截然不同，它永远能通过生成文本提供貌似合理的解释。DeepSeek的深度思考功能就是一个典型的例证。当传统算法拒绝了某项福利申请时，基层官僚至少还能表达困惑，可以为决策保留提供进一步人工审查的制度空间。而当LLM给出一段逻辑清晰、引经据典的拒绝理由时，这个制度空间就被语言的流畅性本身堵死了。一项实验研究揭示，公务员面对算法建议时倾向于顺从而非质疑，这是一种“自动化偏见”[①]。

① Peter Amund Busch and Helle Zinner Henriksen, “Digital Discretion: A Systematic Literature Review of ICT and Street-Level Discretion,” *Information Polity*, Vol. 23, No. 1, 2018, pp. 3–28.

LLM 的语言流畅性本质上不是让算法决策更难被质疑，而是让它更容易被接受，因而会大大加剧这种偏见。其二，通用性造成制度响应滞后。传统算法常常为特定任务量身定做，部署需要有针对性的开发与制度审批，LLM 则无需重新训练就能被部署在任意治理场景。这往往导致 LLM 在监管框架尚未建立的情境下，就已经事实参与了公共决策。因此，LLM 时代的算法治理问题是一个有别于传统深度学习时代的新问题。

（一）权力：自由裁量权的语言性消解

从街头官僚到系统技术官僚的权力转移是公共管理研究领域的经典问题。既有研究已经揭示，信息与通信技术如何将行政裁量权从一线公务员转移到系统设计者手中，算法裁量驱动的去技能化、自动化不平等、问责透明困境等退化路径如何形成。[①] LLM 使这一系列的研究在结构性上就发生了质变。

以荷兰税务局的风险分类系统为例。税务当局引入的 SyRI 算法将双重国籍、低收入群体特征等社会弱势指标在黑盒中转化为高风险权重。算法的刚性不仅阻断了公民的申诉渠道，更剥夺了一线公务员的裁量权与纠错空间。街头官僚发现某些被判定为“欺诈”的家庭实则是因语言不通而填错表格时，往往选择放弃干预，使自身沦为算法的橡皮图章。当类似的系统升级为 LLM 驱动的版本之后，系统不仅能给出高风险评分，还附上一段逻辑清晰、引用法规的解释。这意味着街头官僚面对的不只是冰冷的数字，而是一段比自己写得更符合规范的行政文书。传统黑盒的不可解释性至少还会激起质疑，而 LLM 输出的流畅性语言则会消解质疑，但黑盒决策的问题本质并没有消解。

更深层的问题在于：即使误判率降为零，问题就会消失吗？答案依然是否定的。完美的准确率并不能弥补弹性的消失带来的负面影响，因为基层治理中规则与情境之间人性化权衡的能力本质上不是一个准确率问题，而是一个权力结构问题。AI 系统在公共治理中的角色更接近于街头官僚而非中立的政策工具，在执行过程中内在地行使着裁量权。[②]

因此，社会科学在权力维度的使命在于开展结构性的自由裁量审计，即审计在 LLM 嵌入后，哪些治理场景中街头官僚的干预频率与干预方向发生了系统性变化？怎样追踪申诉路径的实际通达性，辨别审核过程中人工审核的权力范畴？如何推动开发类似“裁量弹性指数”（Discretionary Elasticity Index）之类的可操作审计工具？等等。

（二）价值：公共价值在优化压力下的选择性消失

如果说权力维度的审计揭开了算法中立的面纱，那价值维度的深潜则会触发工程与社会科学在本体论上的根本分歧。

公平与效率、透明与隐私、程序正义与实质正义等公共价值对齐的张力，在本体论上是不可通约的。增加算力不能消解这种不可通约性。[③] 这与前文提及的试金石法则互相印证，属于工程进步永远无法填平的社会科学命题。

2019 年，《科学》发表的一项研究揭露了一款覆盖美国数千万人口的医疗风险预测算法的治理灾难。[④] 该算法将患者未来的医疗支出成本作为衡量健康风险需求的代理指标，基本逻辑是患者病得越重，医疗费用应该越高。它看似自洽，却由于美国社会长期的结构性不平等而导致更大的不公平。非裔及低收入群体缺乏医疗途径或无力支付自付费用，其历史医疗支出远低于同等病情的白人中产阶级。当算法完美地执行

① Peter André Busch and Helle Zinner Henriksen, “AI Technologies and the Reconfiguration of Discretion in Street-Level Bureaucracy,” in *Handbook on Public Policy and Artificial Intelligence*, Regine Paul, Emma Carmel and Jennifer Cobbe (eds.), Cheltenham: Edward Elgar Publishing, 2024, pp. 80–92; Mark Bovens and Stavros Zouridis, “From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control,” *Public Administration Review*, Vol. 62, No. 2, 2002, pp. 174–184.

② Kate Vredenburgh, “The Right to Explanation,” *Journal of Political Philosophy*, Vol. 30, No. 2, 2022, pp. 209–229.

③ Torben Beck Jørgensen and Barry Bozeman, “Public Values: An Inventory,” *Administration & Society*, Vol. 39, No. 3, 2007, pp. 354–381.

④ Ziad Obermeyer, Brian Powers, Christine Vogeli and Sendhil Mullainathan, “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations,” *Science*, Vol. 366, No. 6464, 2019, pp. 447–453.

“以支出预测需求”指令时，那些患有严重慢性病但因贫困而极少就医的底层患者自然被算法归类为低风险，从而被排除在资源分配的视野之外。这场灾难的根源不是代码的bug，也不是工程师的恶意，而是算法过于完美地实现了其设定的单一指标。理解这类偏差需要超越模型内部的技术解释，而转向社会结构性解释，即种族化的经济与社会结构本身就在持续制造不平等。

事实上，LLM增加了社会问题的复杂性。LLM能在自然语言层面表达多维的价值权衡，表达出对“效率”与“公平”的兼顾，但其训练目标依然是单维度的奖励函数。正因如此，在LLM时代，人们更难识别优化训练的压力对公共价值的破坏，也难以对“始作俑者”追根溯源。

（三）制度：问责链条的断裂与重建

面对权力异化与价值撕裂，社会科学的作用在于承认人性的幽暗与理性的有限。我们从不奢望官僚系统自身不犯错，但需要通过立法监督、司法审查与事后问责等一系列外部制度来规训权力的越界、保障公民的权利。这套外部制度的前提是可以识别出某个官员、某个机构、某个决策程序等权力主体。但是，LLM是一种新型非人类权力实体，模型输出貌似客观，但因其来源难以追溯，责任亦无处落地，从而模糊了外部制度的前提。制度所预设的追责对象从一开始就无法确定。因此，社会科学的研究重心必须从内部模型微调转向构筑外部制度护栏。换言之，与其追问模型内部的偏见从哪里来，不如追问当模型出错时谁来负责，通过什么机制负责。

传统的“多手困境”（Problem of Many Hands）是指，当错误发生时，技术提供商、采购机构和执行官员各自推卸责任。LLM引入了第四个“手”，即语言模型本身的输出。它成为责任链条上一个无法追责的节点，且因其输出具有自然语言的形态，更容易被制度性地接受为客观建议而非算法决定。算法问责需要超越技术透明性，深入到制度和政治层面的问责结构重建。[①] 欧盟的《人工智能法案》于2024年通过，2025年起分阶段实施，随后于2026年推进《人工智能数字综合法案》（Digital Omnibus on AI）系统性修订。但是，这套持续校准中的框架依然侧重于事前合规，对LLM在治理运行期间的持续性权力积聚缺乏有效应对。LLM不是一次性的决策工具，而是会随着使用不断积累制度惯性的权力实体。目前“人机协同”（Human-in-the-loop）设计中，许多公共自动化系统在名义上保留了人类审核员，但在高压的工作负荷与自动化偏见下，人类往往退化为毫无意义的点击确认机器。如何通过制度设计确保人类处于中心地位，是制度维度下算法治理最迫切的研究议题。

## 四、从外部批判到共同生产

上述三个维度的研究坐标共享一个隐含前提：社会科学站在AI的外部，并以审计者的身份审视LLM这一外部对象。这个前提导致了社会科学论证的结构性局限，换言之，社会科学的外部视角本身需要被超越。人工智能与社会科学应当是一种共同生产、互为前沿的关系。[②]

（一）外部批判视角的限度

社会科学站在外部批判，预设了主客体之分：AI是被审视的对象，研究者是拿着审视标尺的人。然而，LLM嵌入公共治理越深，这个“外部”视角就越值得推敲。从公共管理实践来看，嵌入了福利审批流程的LLM不是在影响一个外部存在的社会结构，它本身就应该是那个结构的一部分。参与司法辅助的LLM也不仅是从外部观察的工具，而是实时参与法律意义生产的主体之一。当研究对象、研究工具与被研究的社会现象三者都可能是LLM或含LLM参与构成时，社会科学研究就需要超越传统的主体—客体关系。

进一步来讲，LLM的训练语料库包含大量社会科学文献，看起来像是涌现的社会动态，实际上也可能

① Madalina Busuioc, “Accountable Artificial Intelligence: Holding Algorithms to Account,” *Public Administration Review*, Vol. 81, No. 5, 2021, pp. 825-836.

② 郁建兴：《人工智能与社会科学互为前沿》，《中国社会科学报》2026年1月30日。

源于一种数据泄露，模型不过是在再生产先前研究中已识别的现象。[①] 当社会科学家用 LLM 来研究社会现象，他们可能正在用一个已经消化了的社会科学知识系统来验证自己的结论，形成认识论上的内循环。我们不是要否定权力审计与制度设计的价值，而是要追问一个更深层的问题：当研究对象已经是人机共生的复杂治理结构时，社会科学的认识论基础需要做出什么调整？

这一认识论困境绝非无聊的问题游戏，相反，它在技术实践中有着可以直接追溯的印迹。RLHF 本身就可以被视为一次从社会科学概念向工程参数的转化。当标注员被要求判断哪个输出“更有帮助”“更无害”“更诚实”时，他们所依据的判断标准早已被心理学关于偏见的研究、哲学关于诚实的讨论以及公共管理学关于服务质量的界定所预先塑造。[②] Anthropic 在开发 Claude 时明确采用的“宪法 AI”（Constitutional AI）方案，就是将若干源自政治哲学与法学传统的规范性原则直接编码进训练目标，其中包括对无害性的界定、对权威的态度以及对隐私的处理方式。[③] 更能说明问题的是 Santurkar 等人的实证研究。他们系统测量了主流 LLM 在政治与社会议题上的观点分布后发现，这些模型所表达的立场与美国民主党的中间派选民高度吻合，但却与全球多数人口的价值取向存在显著偏差。[④] 这个结果并不令人意外，它恰恰说明模型所内化的“人类价值”是标注员群体的社会位置与文化背景的投影，显示出社会科学意义上的阶层偏好。技术流程赋予了模型“普遍人类反馈”的包装。当社会科学研究者使用这些模型来研究人类偏好或社会规律时，他们测量的已经是这套预先嵌入的价值判断在模型输出中的回响，而不是独立于理论之外的社会现实。

### （二）超越外部批判：操作性共同生产

LLM 嵌入治理的现实表明，一种此前从未以如此直接的方式存在过的“知识—权力”共同生产关系得以出现：两者相互构成彼此的认识论条件，又相互塑造彼此的演化方向，也就是共同生产。Jasanoff 在其政治学与科学研究的交叉工作中曾观察到，科学知识的生产与社会秩序的建构彼此缠绕、相互塑造。[⑤] 在这里，她所讨论的案例是气候科学家与政策制定者之间的博弈、生物技术监管框架的形成，这些都是经由政策讨论、专家共识和制度实践，在数年乃至数十年的周期中逐步完成的。在这个时间结构里，知识与社会秩序的相互塑造是可以被观察、追溯和反思的。LLM 带来的则是一种不同性质的共同生产，本文称之为“操作性共同生产”（operative co-production）。操作性包含两层含义。第一层是即时性。传统意义上的知识—权力缠绕经由报告、政策文件、立法讨论等中介，间接地作用于社会秩序；而 LLM 嵌入治理后，对社会秩序的塑造不经过任何中介，直接在制度运转的每一个具体决策时刻发生。第二层是不可见性。过往知识与权力的相互构成有其可辨识的轨迹，包括谁参与了讨论，谁被排除在外，哪些价值判断在何时进入了政策文本。“操作性共同生产”则把价值嵌入包含训练数据的权重分布、标注员的偏好选择与损失函数的数学结构。它无处不在，却无处可寻。

这种“操作性共同生产”在实践中呈现为相互缠绕的两个面向，共同说明了为何社会科学与 AI 之间的关系互为前沿。

面向一：每一个被部署的 LLM 系统都暗含着一套未经言明的社会科学理论。RLHF 的奖励函数本质上

---

① Maik Larooij and Petter Törnberg, “Validation is the Central Challenge for Generative Social Simulation: A Critical Review of LLMs in Agent-Based Modeling,” *Artificial Intelligence Review*, Vol. 59, article 15, 2026, published November 18, 2025, https: //doi. org/10. 1007/s10462-025-11412-6.

② Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman et al., “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback,” *Transactions on Machine Learning Research*, 2023, published December 30, 2023, https: //openreview. net/forum?id=bx24KpJ4Eb, accessed April 7, 2026.

③ Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen et al., “Constitutional AI: Harmlessness from AI Feedback,” arXiv: 2212. 08073, 2022 年, https: //arxiv. org/abs/2212. 08073, 2026 年 4 月 2 日。

④ Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang and Tatsunori Hashimoto, “Whose Opinions Do Language Models Reflect?” in *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202, 2023, pp. 29971–30004.

⑤ Sheila Jasanoff, “The Idiom of Co-production,” in *States of Knowledge: The Co-Production of Science and the Social Order*, Sheila Jasanoff (ed. ), London: Routledge, 2004, pp. 1–12.

就是一份被翻译成数值的社会契约草案，它替全社会预先裁定了何谓好回应、何谓安全立场、何谓合理价值排序。LLM 与人类价值的对齐程度在不同文化语境下存在显著差异，奖励函数所编码的不是普遍的人类价值，而是特定社会群体的偏好投射。[①] 有研究发现，主流 LLM 所表达的政治观点系统性地偏向特定意识形态立场。这不是模型的 bug，而是它所蒸馏的人类文化本身的忠实反映。每个被部署的 LLM 系统都是一套被数值化的社会理论，只是以参数的形式而非命题的形式存在，因而极难被识别和挑战。Akrich 关于技术“刻写”（inscription）的论述在此处提供了有力支撑。[②] 她认为，每一个被设计出来的技术对象都将设计者的意图与价值判断物质性地铭刻进了技术的结构之中。LLM 是这种刻写逻辑的极致形态，它把“何谓好语言、何谓合理价值排序”这些规范性判断，同时刻写进了人类语言所能触及的全部社会领域。

面向二：社会科学的批判话语本身就在参与技术的规范性建构。当社会科学家使用“算法偏见”“程序正义”“自由裁量权”这些概念去批判 AI 系统时，这些概念绝不仅仅是旁观者的分析工具。它们会渗透到政策讨论之中，进入监管条文，甚至改变工程师的设计语言。欧盟人工智能法案中高风险 AI 系统的分类方式，直接吸收了批判性社会科学十余年来关于算法歧视与问责缺失的研究成果。“公平性陷阱”（fairness traps）的分析框架被引用逾两千次的同时，也深刻地影响了工业界的“模型卡”（model card）设计标准。[③] 社会科学的批判从来都不只是批判，同时也是对技术演化方向的规范性介入。

社会科学与 AI 互为前沿不是两个学科分头进步后的合流，而是对一个已然存在的事实的描述，即两者本就处在同一个知识—权力构建的场域中，相互构成彼此的认识论条件。事实上，如今已经很难找到剥离了社会科学影响的 AI 研究，也很难找到未被 LLM 所干涉的社会议题。

（三）来自工程实践的证据

上述理论分析得到来自工程实践的具体佐证。AI 领域的先驱者们越来越频繁地越出纯粹工程论述的边界，转而借用社会关系、政治制度与规范秩序的语言来描述 AI 的核心难题。辛顿在 2024 年警告说，如果更强的 AI 由受短期利润驱动的公司开发，安全将不会成为技术开发的首要优先事项。[④] 在后续访谈中，他又承认自己过去长期专注于理解大脑并推动这项技术真正“运转起来”，却较晚地意识到了其中隐藏的风险。[⑤] 更值得注意的是，在讨论超级智能控制问题时，辛顿给出的并不是一个标准的工程解，而是一个社会关系隐喻：他主张发展“人工智能母亲而非人工智能助手”（AI mothers rather than AI assistants），并认为如果 AI 不会养育人类，它就更可能取代人类。Bengio 则更明确地把 AI 对齐问题与政治治理合并讨论，指出政治、法律、监管和国际协议都在以不完美的方式处理这一“对齐问题”，并在讨论价值多元如何被汇总时直言“解决方案已经存在，它叫民主”[⑥]。Karpathy 在 2025 年的演讲中把 LLM 描述为“人类幽灵”（people spirits），也就是人类文化语料的随机性蒸馏，因此具有涌现出来的心理结构。他进一步提到如果接受这种工程内部描述，那么对模型偏见、权力结构和价值分布的分析便不再只是外部批评，而成为理解

---

① Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark and Deep Ganguli, “Towards Measuring the Representation of Subjective Global Opinions in Language Models,” arXiv: 2306. 16388, 2023 年, https: //arxiv. org/abs/2306. 16388, 2026 年 4 月 2 日。

② Madeleine Akrich, “The De-scription of Technical Objects” in *Shaping Technology/Building Society*, Wiebe E. Bijker and John Law(eds. ), Cambridge, Mass. : MIT Press, 1992, pp. 205-224.

③ Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian and Janet Vertesi, “Fairness and Abstraction in Sociotechnical Systems,” in *FAT 2019: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, New York, NY: Association for Computing Machinery, 2019, pp. 59-68.

④ Geoffrey Hinton, “Interview,” NobelPrize. org, December 2024, interview recorded on December 6, 2024, https: //www. nobelprize. org/prizes/physics/2024/hinton/interview/, accessed April 2, 2026.

⑤ Pangambam S., “Transcript of Godfather of AI: I Tried to Warn Them, But We’ve Already Lost Control!” *The Singju Post*, June 18, 2025, https: //singjupost. com/transcript-of-godfather-of-ai-i-tried-to-warn-them-but-weve-already-lost-control/, accessed April 2, 2026.

⑥ Azeem Azhar, “Yoshua Bengio: Towards AI’s Humanistic Future,” *Exponential View*, February 14, 2024, https: //www. exponentialview. co/p/yoshua-bengio-towards-ais-humanistic, accessed April 2, 2026.

模型行为的一部分。Sutskever 在 2025 年底的访谈中承认，围绕超级智能的讨论之所以困难，一个重要原因在于，我们谈论的是一种尚不存在、也尚不清楚该如何建造的系统。但与此同时，他又提出，未来真正安全的超级智能应当在某种意义上“关怀有感知能力的生命”，并把“关怀人类”与“民主”并列为值得考虑的方向。[①] 与此形成对照的是，Altman 在 2024 年《智能时代》一文中宣称，人工智能终将使“修复气候”与“发现全部物理学定律”变得司空见惯。[②] 然而，气候变化的现实经验表明，即便科学证据已经极其充分，全球碳排放仍在继续上升。这提示我们，许多表面上属于知识和能力的问题，实际上仍然深受政治意志、制度协调与权力结构的制约。正因如此，把这类问题简单转写为技术效率问题，本身就说明社会科学批判视角不可替代。

当前，AI 领域最顶尖的工程师们，在各自触碰到工程边界的那个时刻，所诉诸的语言、框架与问题，大多来自社会科学关于民主、对齐、权力、不平等、价值、人类尊严等传统的规范内核。这不是偶然的借用，而是工程有限域抵达自身极限之后的主动汇流。同时，社会科学的批判话语已经借助训练语料渗透进了模型的生成逻辑。Hacking 所揭示的“循环效应”（looping effect）在 LLM 语境下获得了一种物质性形态。[③] 在“算法问责”这一学理性概念被写入政策法规后，它已经不只是在话语层面改变人们对 AI 的看法，而是在工程层面直接改写了 LLM 的训练目标、评测指标和对齐标准。其意义在于，批判性概念将被一路传导，最终被编译进模型的参数之中。

（四）工程与社会科学边界的相互牵引

上述经验证据背后，是一种比互相借鉴更深刻的相互依赖关系。工程划定了技术上什么是可能的这条边界，社会科学划定了规范上什么是可接受的这条边界，这两条边界始终在相互牵引、相互重塑。这种结构性的相互依赖使互为前沿不是一种期许，而成为一种应然。

对这一深层结构的理解，需要引入吉登斯的“双重诠释学”：社会科学概念不只是描述社会现实的工具，它们会被社会行动者所吸收，并反过来改变这些行动者的实践，原有的理论描述可能随之失效。[④] LLM 的存在使这种“双重诠释学”升级到了一个新的量级，谓之“三重诠释学”：社会科学概念→被写入 LLM 训练数据→LLM 将其内化为生成逻辑→LLM 输出渗透进社会实践→社会科学再次观察已经由 LLM 参与构成的社会→形成新的概念……这个循环的关键特征不只是反馈本身，而是反馈的速度与规模。以往这个周期是数年乃至数十年，LLM 则将其压缩到了数月，并将影响范围扩展到了所有被部署 LLM 的治理场景。

由此可见，当前许多社会科学研究声称观察的是社会规律，但如果观察对象是被 LLM 部分构成的社会行为，而 LLM 又已经吸收了大量社会科学研究成果，研究者实际上是在测量一个由自己学科知识所塑造的现象，而非独立于理论之外的社会现实。换言之，观察本身已经是被观察对象的一部分。

这个困境并不会通向虚无主义，而是要求社会科学家必须同时追问“我的研究发现了什么”和“我的研究是否部分地在测量自身学科的历史沉积”。一个工程师在评测 LLM 性能时，需要追问评测工具是否已经被所测量的系统所渗透；社会科学家在使用 LLM 研究社会时，更应该追问这个问题。这种反身性要求就是社会科学最终无法被人工智能所替代的根本原因之一。

当 LLM 已经深度嵌入教育、医疗、司法、福利等社会核心领域时，关于它们的研究就不再只是关于一种技术的研究，它同时也是关于这个社会如何分配权力、如何界定公正、如何在人与非人之间划定边界

---

① Dwarkesh Patel, “Ilya Sutskever — We’re Moving from the Age of Scaling to the Age of Research,” *Dwarkesh Podcast*, November 25, 2025, https: //www. dwarkesh. com/p/ilya-sutskever-2, accessed April 2, 2026.

② Sam Altman, “The Intelligence Age,” 2024 年 9 月 23 日, https: //ia. samaltman. com/, 2026 年 4 月 2 日。

③ Ian Hacking, “The Looping Effects of Human Kinds,” in *Causal Cognition: A Multidisciplinary Debate*, Dan Sperber, David Premack and Ann James Premack (eds. ), Oxford: Oxford University Press, 1996, pp. 351-394.

④ 安东尼·吉登斯：《社会学方法的新规则：一种对解释社会学的建设性批判》，田佑中、刘江涛译，北京：社会科学文献出版社，2003 年，第 50—51 页。

的研究。当社会过程本身已经由人机混合体所构成，关于社会的知识也就不能再预设一个“纯人类”的参照系。这或许才是社会科学在LLM时代最需要完成的一次自我更新。它既不是方法的更新，也不是工具的更新，而是对“我们究竟在研究什么”这个元问题的重新回答。

（五）共同生产中的社会科学

社会科学与AI互为前沿不只是对一种关系的描述，它同时也在倒逼社会科学回答一个更根本的问题：当LLM已经深度嵌入社会运作之后，社会科学要研究什么？当福利审批由LLM参与完成、司法建议经由LLM生成、教育评估通过LLM实施，社会过程本身已经是人与机器共同构成的。如果社会科学还预设研究的是一个“纯人类”的社会，它的研究对象其实已经不存在了。这不是哲学上的追问，而是一个非常具体的方法论问题：我们观察到的人类行为，多少出于人的选择，多少出于LLM的塑造？所谓的社会规律，又有多少只是LLM重复输出那些被送进训练语料的旧理论？

这些问题意味着社会科学一直以来的基本判断得到了现实验证。社会科学从来就坚持研究者不可能站在他所研究的社会之外，知识会改变它所描述的现实，概念会塑造人们的实践。这曾经是需要反复论证的理论命题。在LLM时代，它变成了字面意义上发生的事情。社会科学概念通过训练语料进入模型，模型再把这些概念渗透进日常的社会决策之中。重要的是，社会科学所追问的那类问题，在这个时代变得比任何时候都更加紧迫，但工程界并不能独立地进行回答。工程师能够回答“这个模型能做什么”“这个系统运行得是否高效”，但无法回答“这个模型应该为谁服务”“谁有权定义什么叫高效”。这不是因为工程师能力不够，而是因为后者根本没有技术答案。它们的答案既是社会协商与权力博弈的结果，也是价值选择的体现。当前者的答案被用来回避后者时，社会科学的任务就是把后者重新放回桌面上。

社会科学还有一项没有任何其他科学能够替代的功能，即追问自己的概念在被采用之后发生了什么。当算法问责变成了一张合规清单，或公平性变成了一个可以买卖的审计服务，只有社会科学研究者能够追问这些词语是否还指向它们最初想要保护的那些东西。显然，社会科学必须把这种自我审视内置在自己的研究框架里。在社会科学与AI互为前沿的格局中，社会科学无可替代。

当社会科学研究者将自身置于技术的外部，要求LLM模型如何遵循各种伦理规范时，这种看似充满人文关怀的批判视角，其底层逻辑往往暗含着一种隐蔽的技术崇拜，它假设只要社会科学提出足够完美的价值需求，技术系统便能自动回应并解决复杂的社会命题。然而，技术的演进自有其客观的工程规律，并非单向度的外部规范所能左右。谁也无法确定3到5年后LLM会发展成什么样，但可以确定的是这个发展的过程具有相当的不可控性，正如LLM时代的突然到来一般。当计算机科学界已然开始通过算法对齐等工程手段将伦理原则内化于模型时，社会科学若仍止步于外围的需求主张与道德呼吁，不仅难以实质性地介入技术的发展轨迹，更面临着学科价值被消解的风险。这意味着共同生产不仅仅是一种应然的状态，也是社会科学保留自己学科价值的必然之路。共同生产意味着消除对技术的单向期待，打破将LLM视为黑盒的认识论惰性。社会科学的独特价值，不在于反复重申LLM应当变得如何，而在于参与研究如何通过机制设计与知识生产让LLM变得不同。

## 五、结论

本文旨在为社会科学在LLM时代重建独立的研究立场。通过前面的递进论证，可以得出以下几点判断：其一，当前计算社会科学领域存在大量工程“伪影”。许多被误认为结构性的经验发现，不过是算法在不断优化过程中的暂态问题，终将随模型迭代自动消亡。识别并清理这类伪谜题，是社会科学维护学术尊严的首要任务。其二，工程有限域与社会科学无限域之间存在明确的认识论边界。算力可消解的问题属于工程命题；根植于价值冲突、权力结构与制度惯性的问题，才是社会科学应当全力以赴的真命题。两条判据交叉形成的矩阵，为社会科学的问题选择提供了可操作的判断框架。其三，LLM相比传统算法带来了两个本质性的改变：语言生成能力引发的问责结构质变，以及通用性带来的制度响应滞后。这使得权力、价值与制度三个维度上的已知议题获得了全新的研究视角，从而构成社会科学在LLM时代的研究坐标。

其四，也是最重要的一点：社会科学与 AI 之间存在操作性共同生产关系。这不是隐喻或期望，而是可以在工程实践中直接追溯的事实。社会科学的批判性概念参与塑造着它所批判的对象，LLM 的参数中内嵌着未言明的社会价值观。互为前沿的本质是对已然存在的知识—权力关系的客观描述。承认这一事实，意味着社会科学必须既作为规范建构的参与者，又作为对这种参与本身保持警醒的反思者。

工程界定义了我们在智能计算上能走多远，社会科学则持续追问我们该往何处去、我们正在成为什么。这两个问题从来都不是独立的：对前者的每一次回答都在重塑后者的提问空间，而后者的每一次追问都在为前者划定新的规范边界。技术可以提供极其锋利的工具，但它自身无法产生关于公平、正义与责任的答案。当模型的参数规模大到足以深度嵌入社会运转的每一个环节时，工程的边界便真正显现，而这正是社会科学新的角色定位。

〔本文为国家自然科学基金重点项目“数字政府驱动的治理范式变革研究”（72434004）、教育部哲学社会科学研究重大专项项目“中国社会治理自主知识体系构建研究”（2025JDZD047）的阶段性成果。刘宇轩为本文通讯作者。〕

（责任编辑：王胜强）

# Beyond the Outside Observer: Repositioning the Social Sciences in the Era of Large Language Models

YU Jianxing, LIU Yuxuan

**Abstract:** As Large Language Models (LLMs) become deeply embedded in public governance, judicial assistance, and welfare allocation, social science is compelled to reexamine its own epistemological foundations. This article identifies three critical pitfalls in current social science research on LLMs: misinterpreting transient model deficiencies as structural social regularities, uncritically accepting manipulated benchmarks, and generalizing local optimizations of small-scale models into universal theories. To address these issues, the article establishes an epistemological boundary between engineering challenges and social propositions along two dimensions: whether a phenomenon can be resolved by scaling compute, and whether commercial logic provides sufficient incentive for its repair. This framework is crystallized into an operable four-quadrant matrix. Building on this foundation, the article identifies three research coordinates for social science: the linguistic erosion of administrative discretion, the selective disappearance of public values under optimization pressure, and the fracturing of accountability chains introduced by LLMs as a “fourth hand”. The article further argues that the premise of social science standing “outside” AI is increasingly untenable.

**Key words:** social science, LLM, AI, co-production