AI模拟社会研究资料

1.0版 修订号 0.90

本报告全流程使用AI收集资料、生成及审核完成除发起外无人类参与，不完善之处将逐步修改

@清新研究 团队

2026年3月

@清新研究团队简介

沈阳为清华大学新闻学院/人工智能学院双聘教授、博导，清华大学新闻学院新媒体研究中心主任。先后担任计算机、信息管理、新闻传播、人工智能等多个学科教授。

领导学术研究团队近30人。指导大数据、AI、人形机器人等多个产业团队。团队已有众多大模型产业化和AIGC实施案例，有需要可留言联系。

团队坚持：整体主义的跨学科整合力，实证主义的实践导向，社会建构的产学研结合，进步主义的先锋探索精神，科学服务于大众的社会责任。欢迎对AIGC或AI赋能社会实践感兴趣的朋友留言交流，我们长期关注人才成长、跨界合作与前沿探索。

邮箱：124739259@qq.com；微博：@新媒沈阳

视频号：@清新研究；

公众号：@清新研究

六大研究方向

- AI大模型理论与哲学
- AI文艺
- AI应用
- 新媒体与网络舆论
- 大数据
- 元宇宙

团队历年研究报告

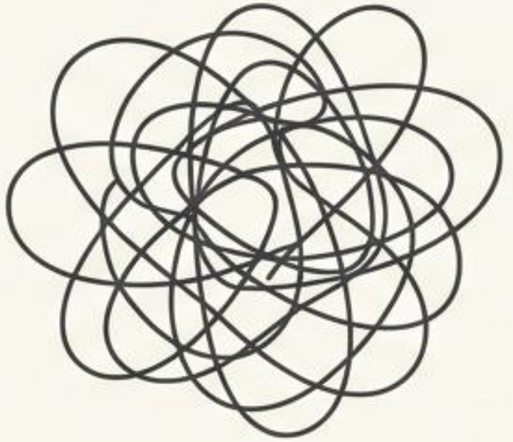
DeepSeek报告阅读量近亿
元宇宙报告阅读量近千万
AI报告和新媒体报告阅读量近百万

总述：当AI成为社会代理人

范式转变、核心价值与未来框架

2024年11月 | 战略前沿研究所

应对“棘手问题”的全新可能

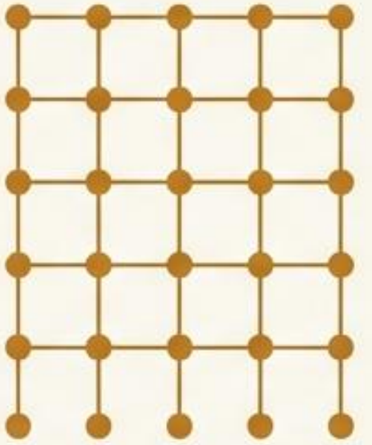


什么是“棘手问题”（Wicked Problems）？

现实社会中充满复杂的均衡状态与不受控变量，缺乏标准答案且难以通过传统方法进行低成本测试。

传统方法的局限：

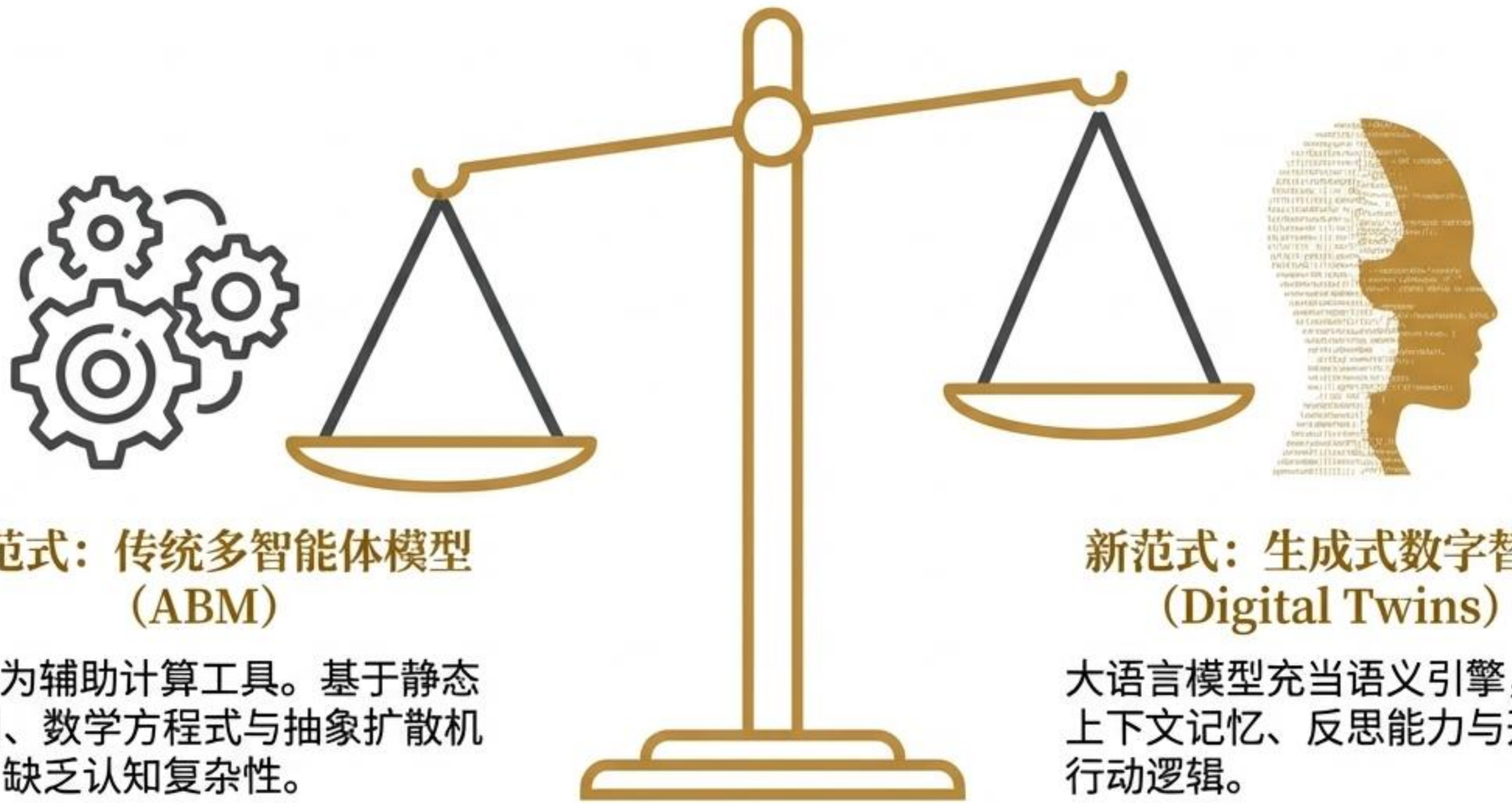
真实世界的政策试错成本极高，且不可逆。



破局之道：

生成式社会模拟提供了一个安全、可控、高保真的硅基沙盒。

范式转变：从“辅助计算”到“数字替身”



旧范式：传统多智能体模型（ABM）

AI作为辅助计算工具。基于静态规则、数学方程式与抽象扩散机制，缺乏认知复杂性。

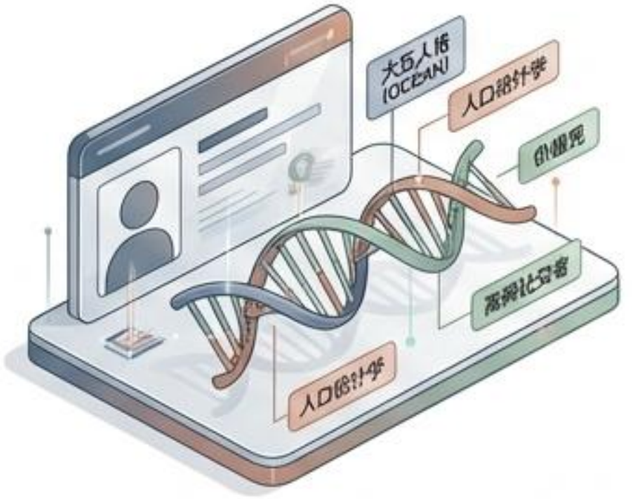
新范式：生成式数字替身（Digital Twins）

大语言模型充当语义引擎，具备上下文记忆、反思能力与开放式行动逻辑。

核心机制：驱动AI社会的真实异质性与三层演进尺度

> “放弃单一的标准大模型回答，我们要构建的是一个具备真实人类异质性的‘硅基生态’。”

异质性基座：硅基社会的“基因工程”



- 告别同质化：摒弃大模型默认的“绝对理性/政治正确”回答。
- 多维属性注入：赋予智能体特定的人口统计学特征、大五人格（OCEAN）、价值观体系以及差异化的行为角色（如：高频社交者、边缘潜水者、极化辩论者），构建高逼真度的数字人口库。

【宏观尺度 / 复杂大众】：百万级多智能体沙盒 (Million-Agent Sandbox)

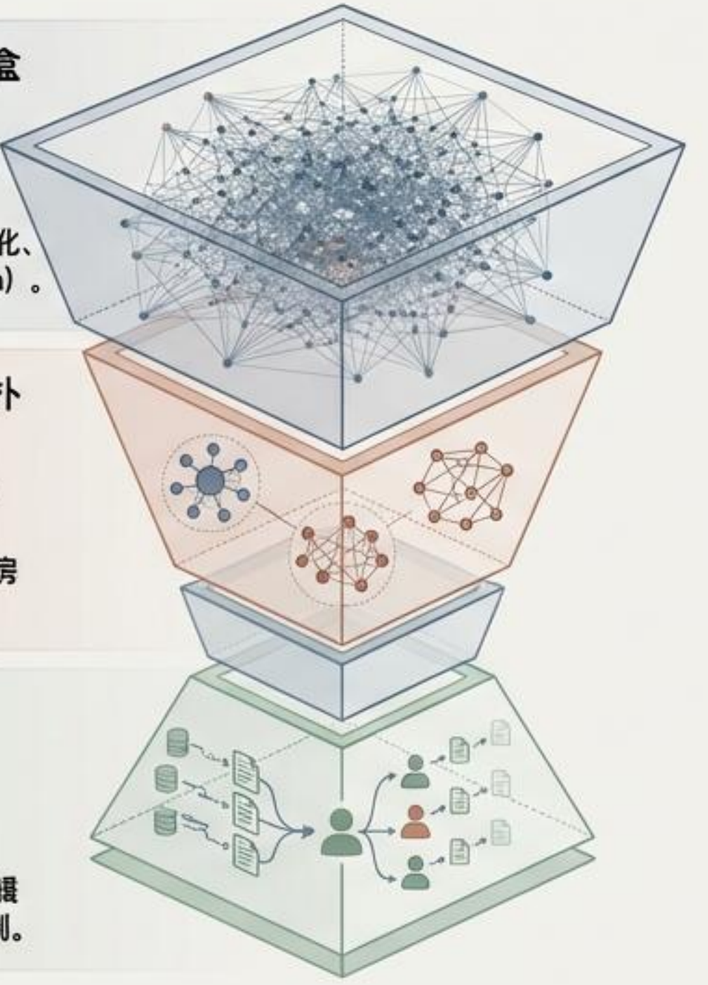
- 高并发架构：支持海量智能体在同一时间线上并发互动与资源交换。
- 实现效果：不施加全局干预，观察自由贸易、职业自发分化、宏观舆论极化等系统级涌现现象（Emergent Phenomena）。

【中观尺度 / 圈层分众】：硅基抽样与网络拓扑 (Silicon Sampling & Topology)

- 结构化网络：导入特定圈层的统计学变量，配合小世界网络（Small-world）或无标度网络结构。
- 实现效果：模拟特定群体内的交互动力学，观察信息茧房的形成、同温层效应与共识演化路径。

【微观尺度 / 小众个体】：主观世界建模 (Subjective World Modeling)

- 深度数据导入：基于真实人类1-2小时的深度访谈语料（单体约5000字转录稿）进行上下文注入。
- 实现效果：构建具备连贯生命叙事、独特记忆与决策逻辑的深度数字替身（AI Personas），精确模拟个体决策机制。



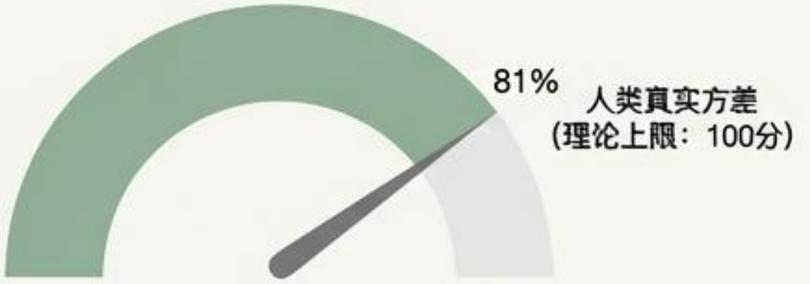
科学校验闭环（上）：微观一致性与中观行为学对齐

> “对齐真实人类的自然方差：以81%的人类自我复测一致性为标尺，重新定义模拟准确率的‘满分上限’。”

核心理念：摒弃“对赌式预测”，拥抱“基准数据对齐”

我们不预测不可控的黑天鹅事件，而是通过严谨的后验比对，证明硅基代理在已知社会规律与人类基准下的高度拟合。

🎯 微观校验：问卷拟合与“自我复测一致性”基准



- 人类真实方差：心理学实证表明，真实人类在两周后回答同一份问卷，其答案的自然一致性约为 81%-85%。
- 重新定义“满分”：将81%的人类自我一致性设为AI模拟准确率的100分基准（即理论上限）。
- 校验标准：让AI替身完成《综合社会调查（GSS）》核心模块，若其回答分布与真实原型拟合度达到该阈值，即视为在微观个体层面校验通过。

🎲 中观校验：经典行为学与心理学实验复现

- 测试场景：让AI智能体参与真实人类做过的经典博弈实验（如：独裁者博弈、信任博弈）。
- 校验标准：拒绝“绝对理性”。
- AI不应总是做出数学上最优的纳什均衡选择。
- AI必须在合作率、背叛率、互惠预期等行为逻辑上，严格复现已有学术论文中真实被试的数据分布与方差，证实其具备人类的非理性与互惠心理。

科学校验闭环（下）：宏观系统涌现与算法逼真度评测

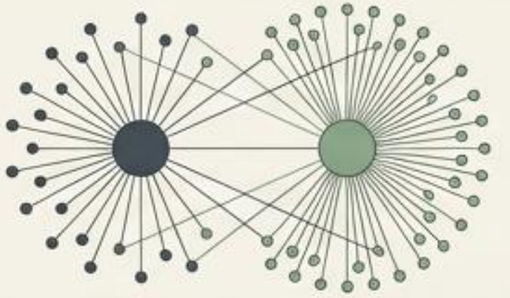
> “不看个体看大盘：最极致的校验，是见证系统底层自发涌现出真实世界才有的宏观运行规律。”

🌎 宏观校验：统计特征与“风格化事实”（Stylized Facts）

校验逻辑：个体层面的微小扰动会在大规模交互中被放大或抵消。真正的校验在于系统盘面是否符合人类社会的统计学定理。

（经济金融）

在自由互动沙盒中自发跑出资产价格的“重尾收益分布”与“波动性聚集”。



（社会结构）

跑出社会交往维度的“幂律分布”，以及基于资源/教育初始设定的社会阶层化与财富分层现象。

📐 综合校验：“算法逼真度（Algorithmic Fidelity）”四大金标准

为了确立模拟社会的最终科学可信度，全系统需通过以下交叉检验：

1. 社会科学图灵测试

文本输出与行动决策自然连贯，社会科学专家亦难以区分人机数据。

2. 向后连续性

能够从智能体表现出的交互行为中，精准反推其被预设的人口统计学标签与社会背景。

3. 向前连续性

在给定的背景下，其回应基调、词汇选择能始终保持该社会群体的天然特征与语境。

4. 模式对应性

智能体的多元观点与自身背景之的交叉关联，必须严密匹配真实大规模人类社会调查中的回归系数模型。

硅基生命的认知架构

生成式智能体不再是简单的刺激反应机器，而是具备完整认知时间线的实体。

观察（Observation）

↓

记忆流（Memory Stream）：记录并检索过去的自然语言经验

↓

反思（Reflection）：将零散记忆合成为高阶的自我认知与价值观

↓

规划与行动（Planning & Action）：基于当前情境与过往反思，自主生成符合逻辑的长期计划

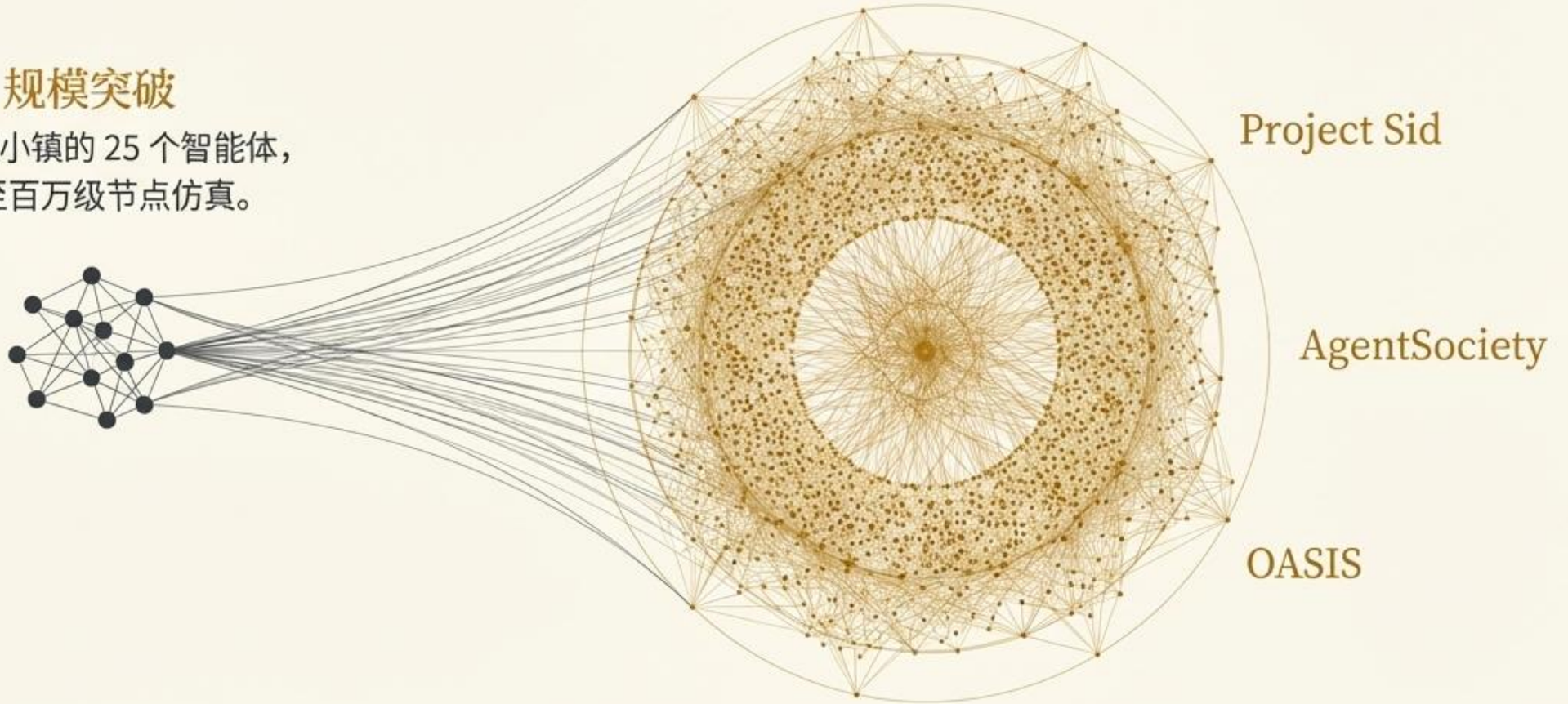
尺度跃升：从个体微观到社会涌现

宏观现象涌现（Emergence）

个体行为聚合后，自发形成同质性聚类(Homophily)、极化、文化传播与制度规范。

规模突破

从斯坦福小镇的 25 个智能体，发展至百万级节点仿真。



自然语言：构建主体间性的新介质

语言即规则

由会部分间论黫计其学，对出的直鳍进行。
社会制度、文化规范与互动逻辑首次可以直接用自然语言进行编码与执行。

因果链条追踪

智能体智能体的部分，
以 MASS 框架为例，智能体能够以一句话的意图理由解释其行动，将对话链与因果链完美绑定。

虚拟沙盒：人类社会的终极实验室

为什么要在虚拟沙盒中模拟人类社会？三大核心战略价值：

价值一：低成本政策推演
(Policy Deduction)

价值二：极端情境压测
(Extreme Stress Testing)

价值三：虚拟受访者与硅基样本
(Virtual Respondents)

核心价值一：低成本与前置性的政策推演

事前政策探测（Ex-Ante Probing）

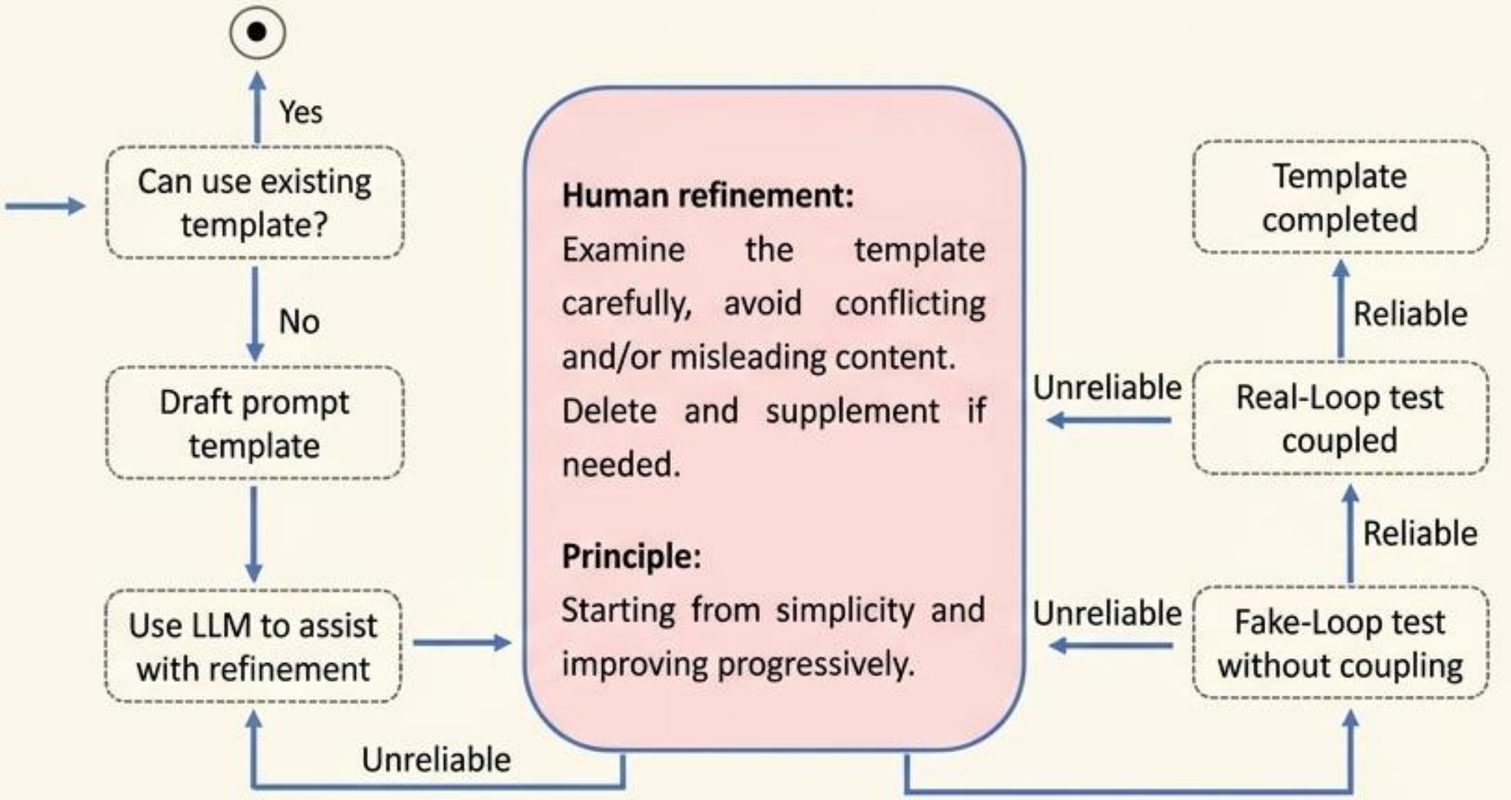
在真实部署前，测试全民基本收入、最低工资等政策的连带效应。

实证成功案例

MASS 框架成功复现 Card-Krueger 经典研究（模拟出 $0.65/h 的工资增长且未引发显著失业）。

人机协同设计（HIL）

在复杂的政策沙盘中，结合人类专家修正大模型策略与漏洞。



核心价值二：模拟“不可想象”的极端情境

突破现实伦理限制

安全地测试危险场景，如严重经济冲击或全球性流行病扩散。

对抗性操纵测试

探究信息战与虚假信息在社交网络中的传播路径。

平台治理预演

提前测试内容审核机制、算法重定向与干预策略在多智能体生态中的实际效用。

核心价值三：硅基样本与社会调查革命

硅基样本（Silicon Sample）

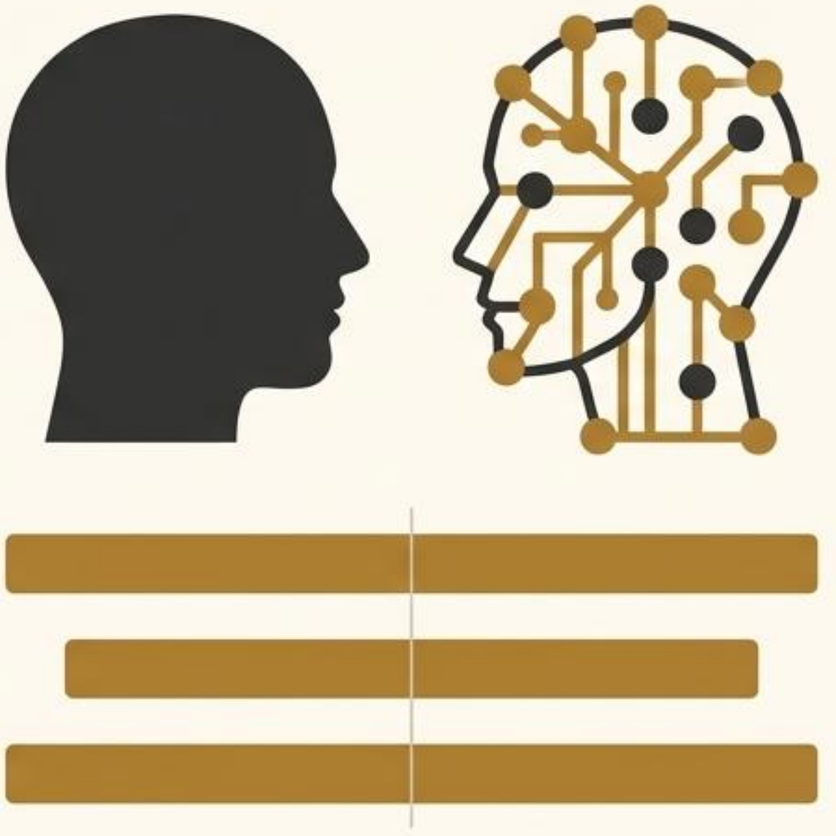
替代或增强传统人类问卷调查，突破时间、成本与样本量限制。

算法保真度（Algorithmic Fidelity）

生成人格与真实数据（如 ANES 数据集、大五人格模型）表现出高度统计相关性。

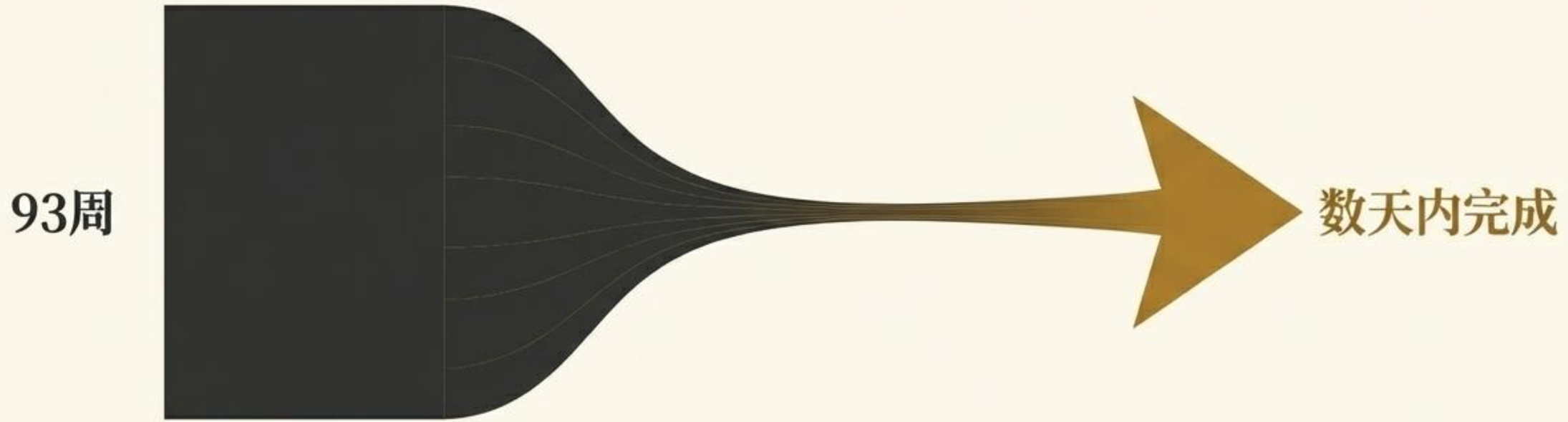
商业应用创新

快速进行产品概念测试、消费者行为建模与目标市场优化。



完美保真度演示

效能飞跃：大规模质性分析与演绎编码



超越静态问卷

大语言模型可对海量深度访谈进行演绎定性编码。

实证效率

在 iPATH 研究中，处理 167 份访谈转录本（超130万字），将时间成本削减 30% 至 55%。

人机协作边界

AI 负责主题组织与模式识别，人类学者保留对数据的最终解释控制权。

局限与反思：警惕“认知恐怖谷”

逼真度的幻象

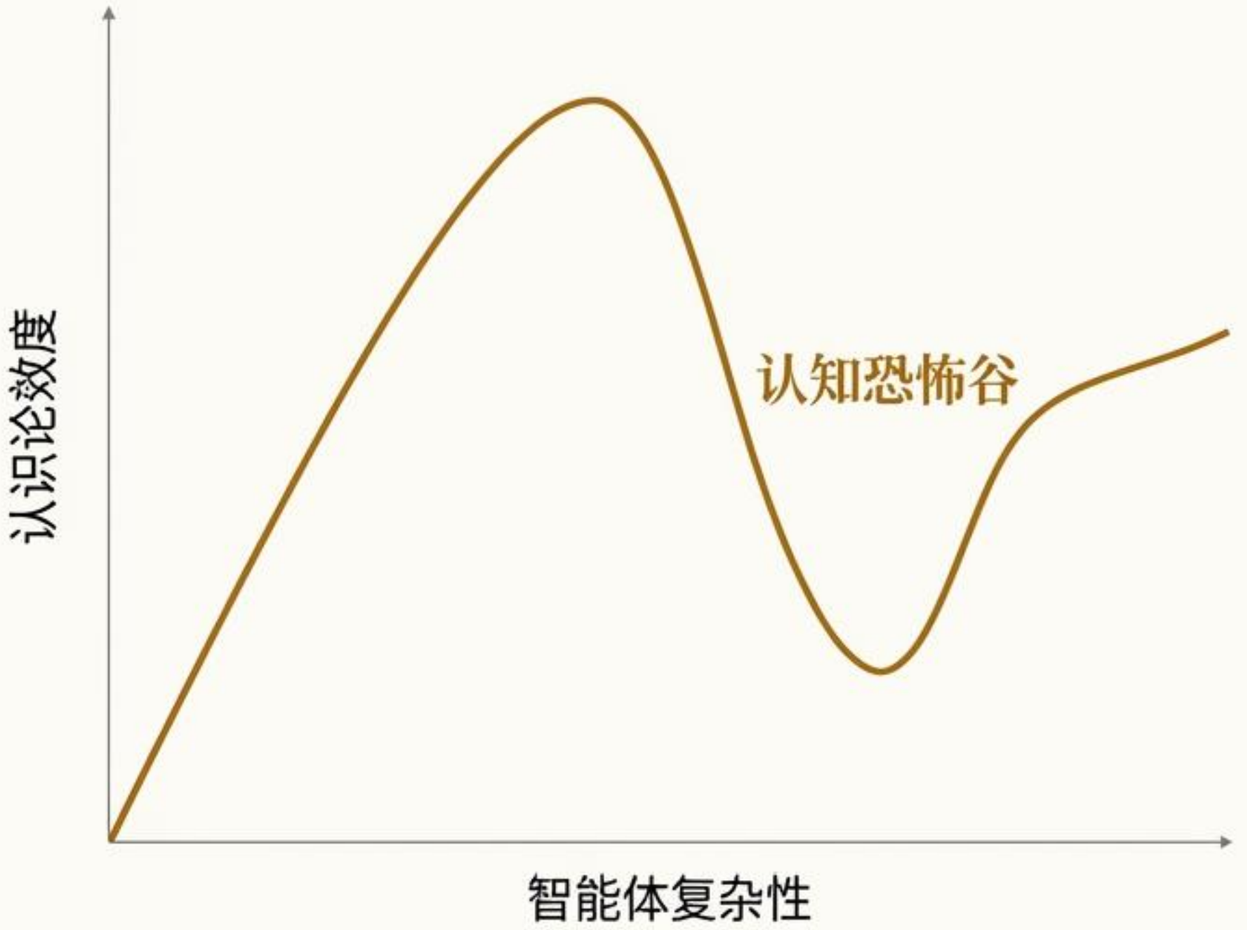
“过于像人而难以建模”。智能体听起来像人类，并不意味着其底层逻辑无懈可击。

随机鹦鹉 vs. 和谐合唱

模型倾向于输出社会赞许性的常规智慧，掩盖了真实人类社会的矛盾、异质性与反常识细节。

认识论风险

必须警惕模型崩溃、确认偏误以及方法论上的不适配。



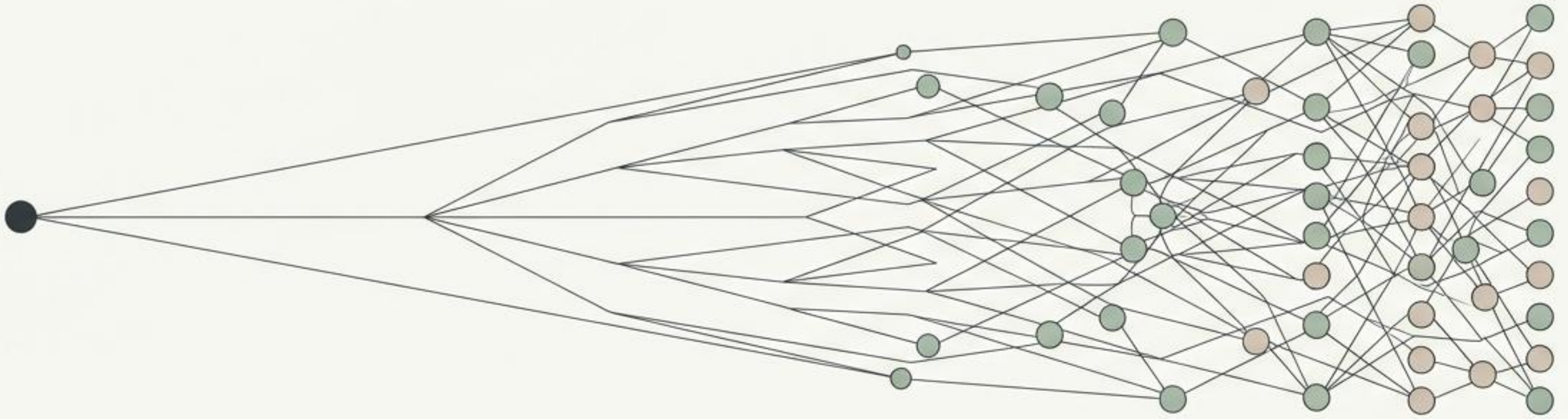
结语：建构真实的镜像

硅基沙盒已经落成，工具的演进超乎想象。

我们应如何设计机制，确保这场模拟映照出深刻的真实，而非虚无的海市蜃楼？

第二章 起源与现况：从单体问卷到万级城市沙盒

生成式社会模拟的演进——跨越认知架构、垂直产业与宏观计算的视界



起源与基石：斯坦福 Generative Agents 与“Smallville”小镇

25个生成式智能体在一个沙盒中首次展现了自主的生活轨迹与涌现性社会行为。

里程碑

2023年斯坦福大学与谷歌团队联合发布。

核心突破

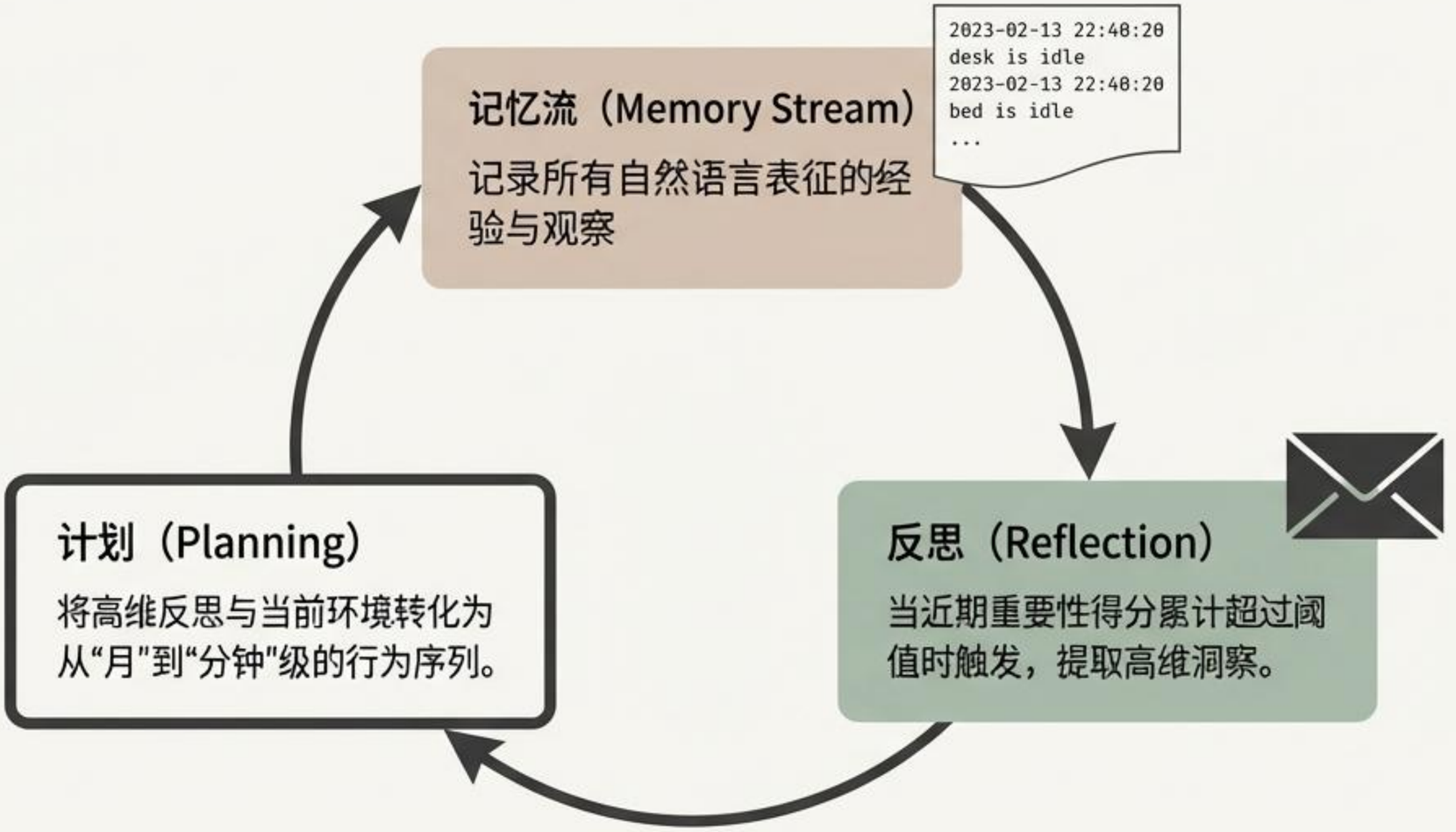
智能体不再依赖预设脚本，而是基于自然语言驱动的日常行为生成。

社会涌现

实现了自主扩散派对邀请、建立友谊等自发性群体协调现象。

架构解构：“记忆-反思-计划”的认知飞轮

使智能体具备“可信度（Believability）”的核心在于长时记忆检索与高维反思机制。



演进分水岭：从“虚构NPC”到“高保真数字替身”

技术的聚焦点从游戏化的“可信感”，转向基于真实人口深访数据的“预测准确度”。

阶段一：叙事可信度（Believability）

- 代表：AI Town
- 驱动数据：虚构短传记（Short bios）
- 目标：互动娱乐、沙盒游戏

阶段二：统计准确度（Accuracy）

- 代表：斯坦福千人代理实验
- 驱动数据：长达两小时的真实受访者深度访谈数据

1,052名数字替身 GSS问卷回答一致性，达到人类受访者两周后自我复测一致性的 **85%**

产业化标志：数字替身走向商业决策的中心

Simile AI 等企业的崛起标志着社会模拟从实验室工具跃升为上亿美元估值的商业基础设施。

应用场景

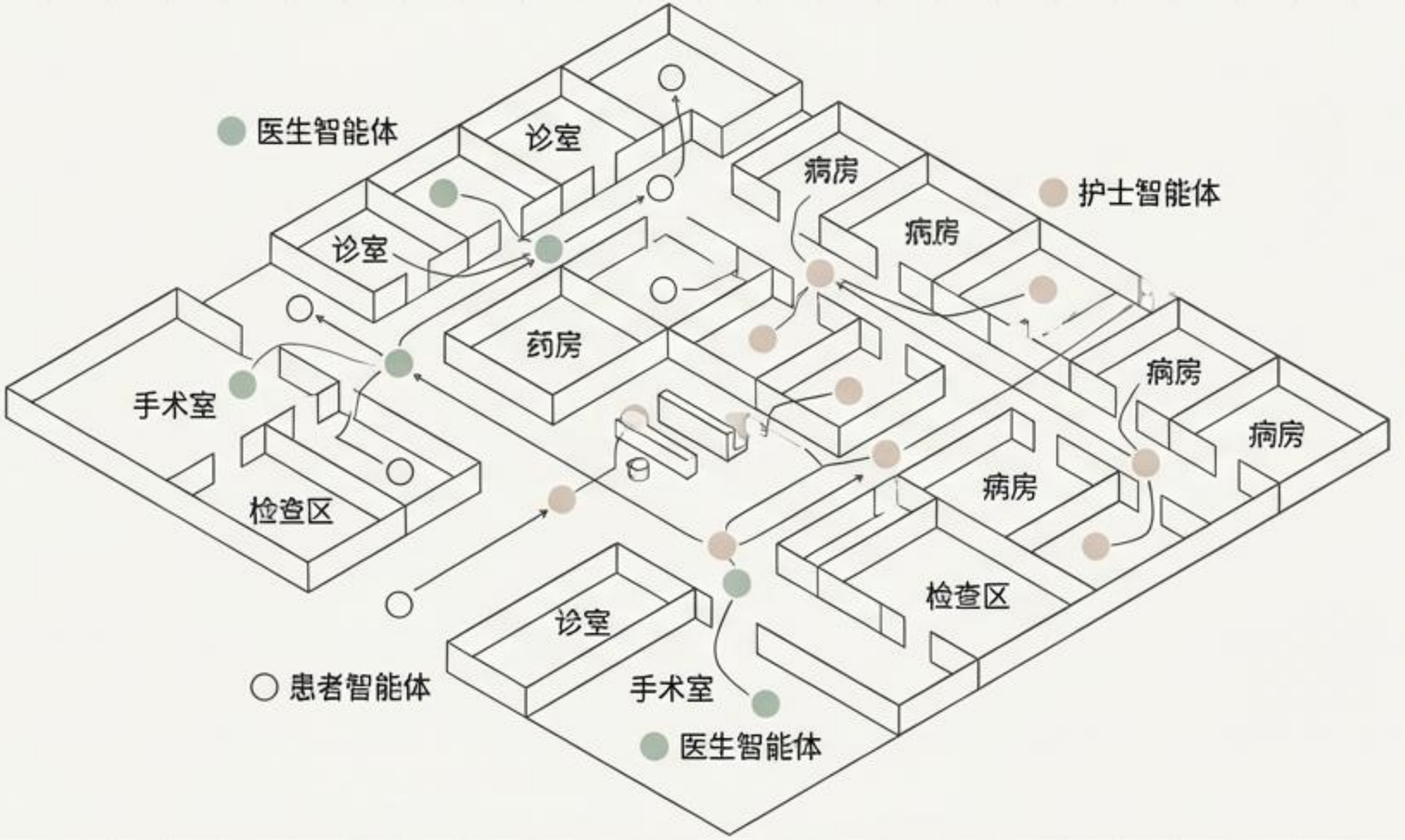
演练财报电话会议
(Earnings calls)

测试宏观政策变化
(Policy changes)

推演诉讼结果
(Litigation prediction)

垂直深化：清华“AI医院（Agent Hospital）”的闭环仿真

将社会仿真引入强约束、高壁垒的专业领域，实现医疗全流程的系统级模拟。

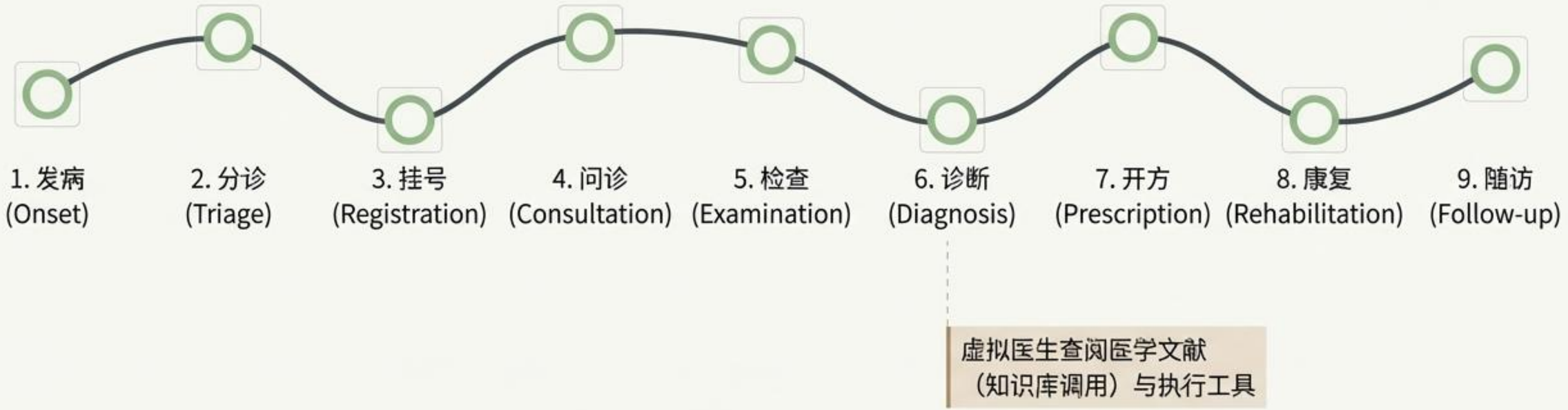


MedAgent-Zero 驱动

可自进化医疗智能体系统。突破单纯对话框架，进入受制于医疗规章、专业知识库与物理空间约束的复杂协同操作流。

医疗流转闭环：超越单体对话的系统级协作

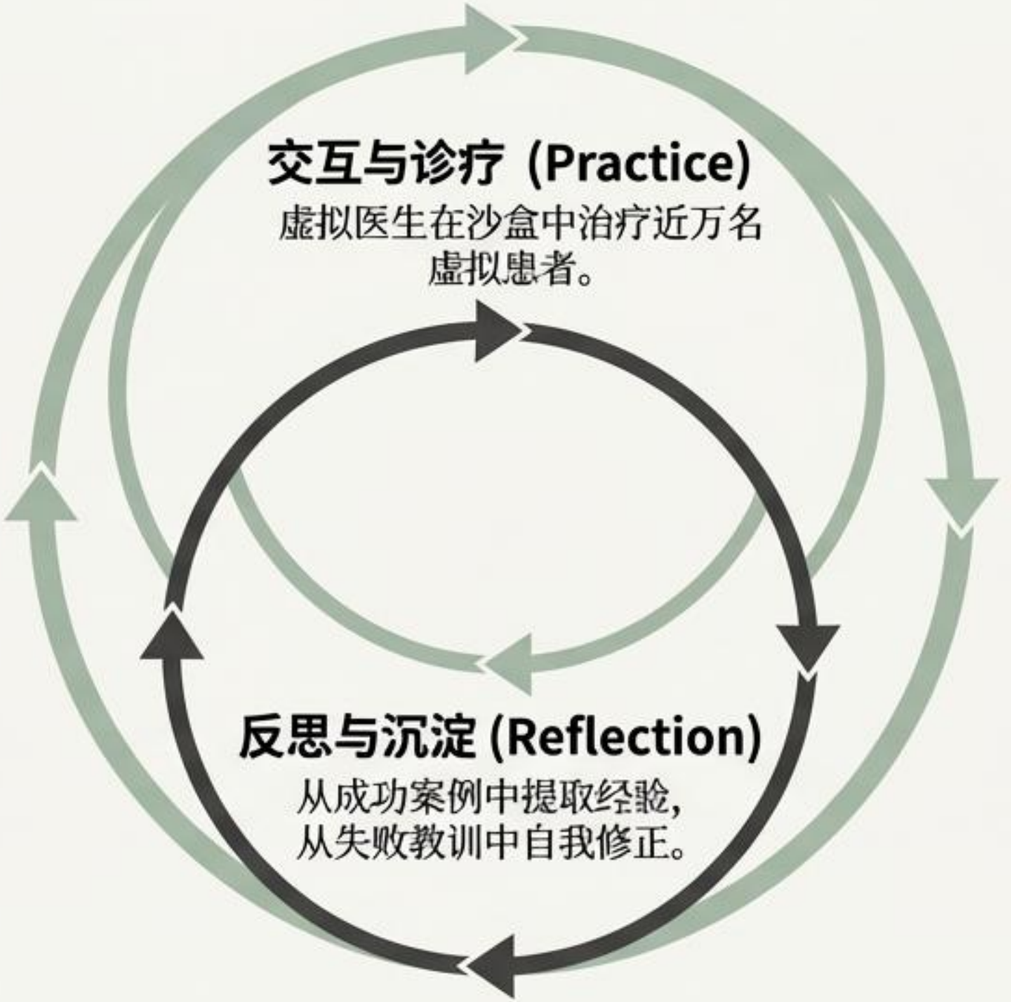
从发病到随访，仿真覆盖了真实医疗体系中的9大核心决策节点。



全流程无需人类干预

机制突破：无人工标注的“自进化”路径

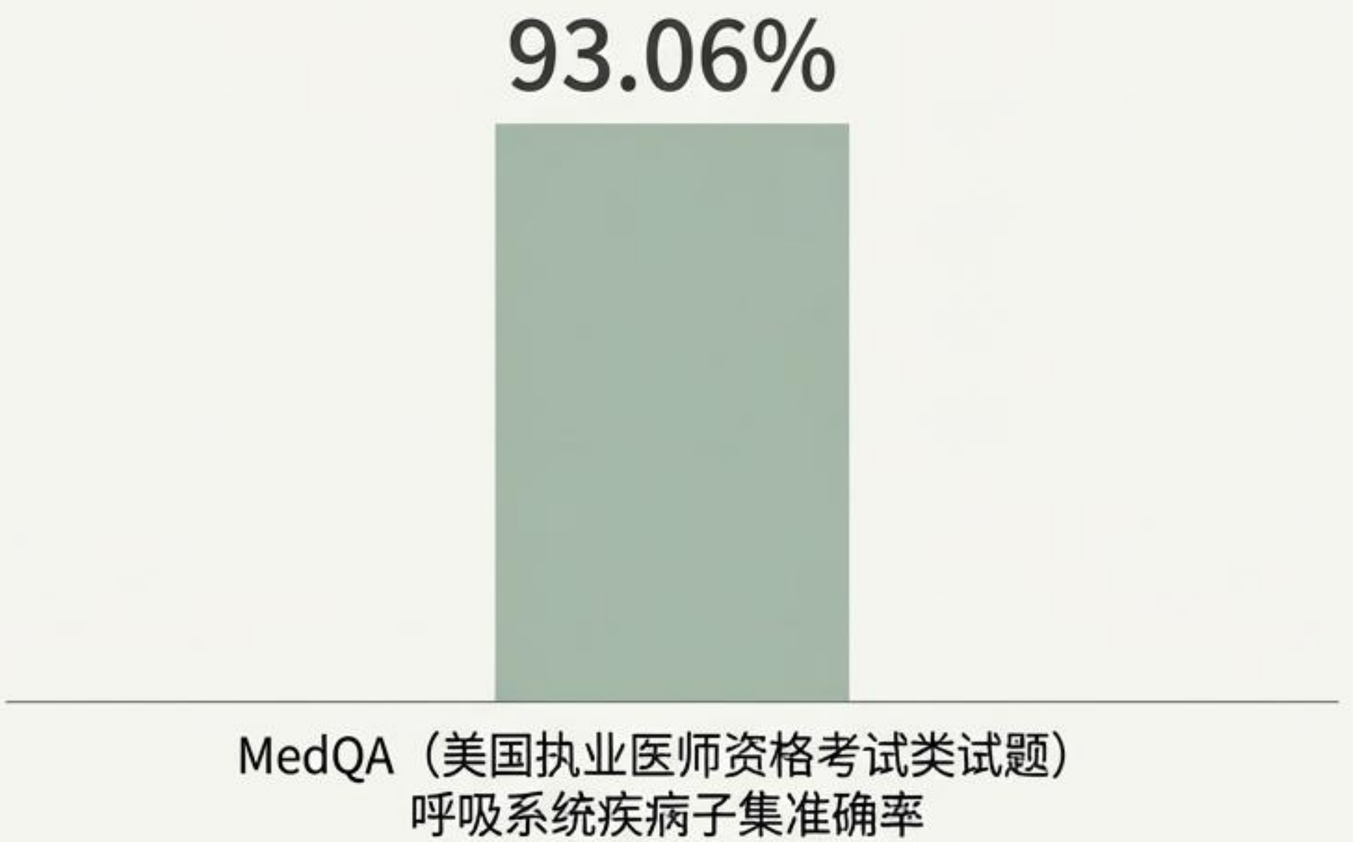
在虚拟环境中通过海量试错积累经验，将人类医生的数年成长周期压缩至数天。



零人工标注 (Zero human annotation) 驱动

仿真效能与溢出效应：重塑医疗 AI 标尺

演化后的医生智能体在权威医学基准上超越现有方法，并加速向公共试点溢出。



Spinoff 创业转化

生态落地的行业信号，如“紫荆智康”等初创公司。

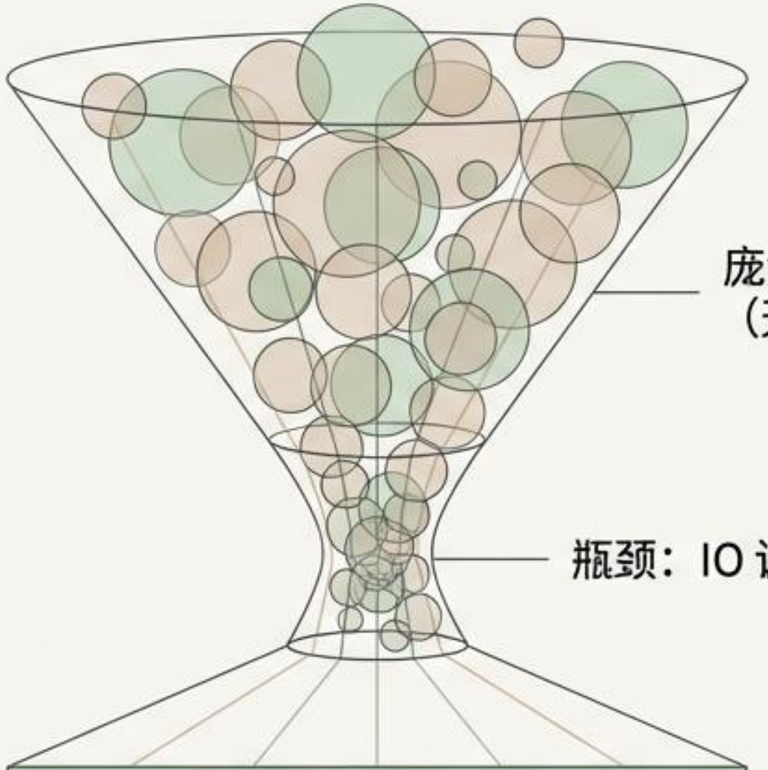
Public Clinical Pilots

迈向真实物理世界的公共医疗试点，将学术沙盒推向现实。

Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.-Q., Ma, W., & Liu, Y. (2024). Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

宏观计算的瓶颈：当人口规模逼近“万级”

从百人小镇到百万级社会，Token限制与并发IO调度成为社会计算的阿喀琉斯之踵。



庞大的 Agent 数量与几何级增长的交互需求
（天量 API 请求与 Token 消耗）

瓶颈：IO 调度拥堵、上下文窗口溢出

基础设施解决方案

- 调度器并行处理
- Group-and-distill 提示词压缩技术（如 OpenCity 框架，能在普通硬件上1小时内模拟万级代理日常）

万级宏观基座：AgentSociety 城市与社会沙盒

支持万级智能体并发，实现“心智-行为”深度耦合的计算社会学试验台。

超大规模：支持 10,000+ 智能体，产生超 500 万次交互。

高保真环境：结合真实城市拓扑与物理规则。

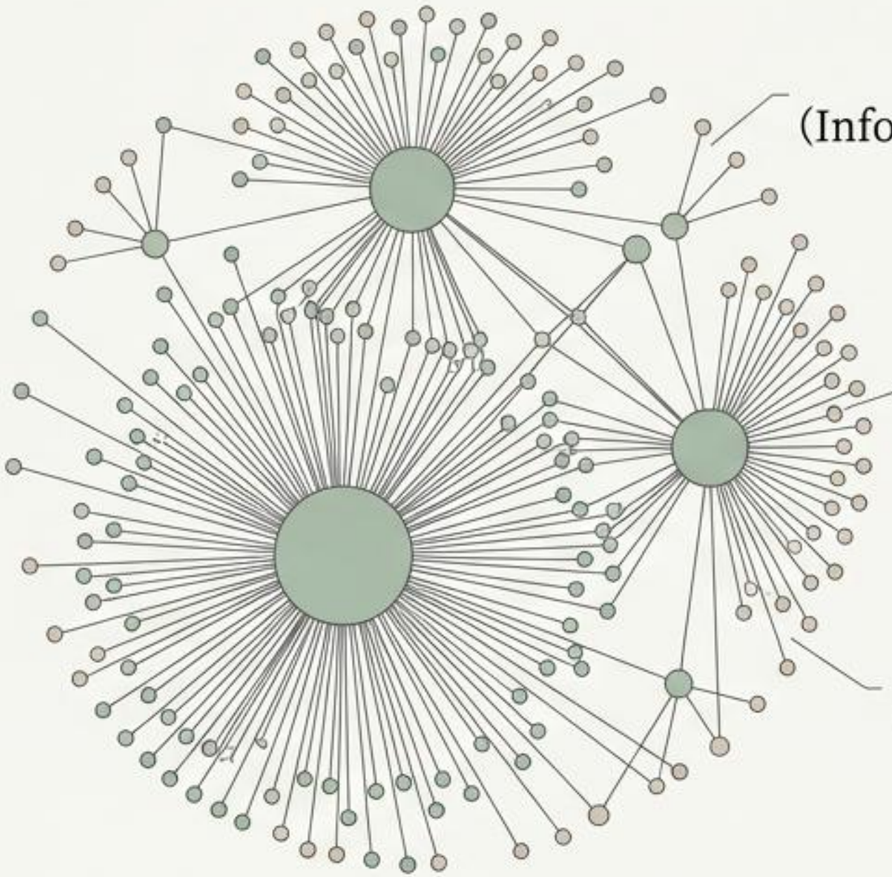
研究工具链：内置社会科学标准的问卷、访谈与干预工具。

应用场景

探究群体极化、煽动性信息传播以及复杂社会网络的演化。

百万级媒介重构：OASIS 与数字公共域的孪生

扩展至 100 万智能体，专注于还原 X/Reddit 等社交媒体中的算法推荐与群体效应。



信息瀑布
(Information Cascade)

回音壁
(Echo Chambers)

从众效应
(Herd Effects)

平台三大支柱

- 动态社交图谱
  (Dynamic Follower Networks)
- 双轨推荐算法
  (基于兴趣 vs. 基于热度)
- 异质性操作空间
  (发帖、评论、转发、屏蔽)

前沿标杆 —— MASS：语言原生的多智能体社会实验平台

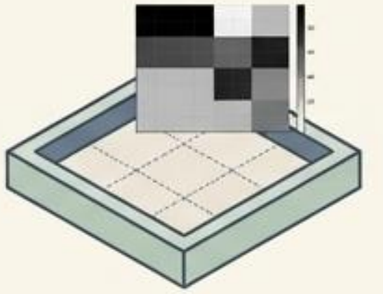
兼具实验控制力与语言可解释性，为复杂公共政策提供低风险、可追溯的“事前推演”虚拟沙盘。

模块一：平台定位与机制创新

- **全称释义：**引入 MASS（Multi-Agent Social Simulation）框架，将复杂的社会互动过程直接用自然语言进行建模。
- **语言原生：**在可控的回合制环境中施加“政策冲击”（Policy Shocks）。
- **机制透视：**不仅记录智能体的数值化行动，更强制记录“一句话意图理由”，将对话链条与因果链条完美对接，实现全过程的机制审计（Mechanism Audits）。

模块二：经典复现与效度证实

- **验证基准：**平台不依赖抽象假设，而是通过经典自然实验的完美复现来证明其有效性。
- **成功案例：**在多轮独立重复中，成功复现了劳动经济学经典的“新泽西—宾州最低工资案（Card-Krueger效应）”。
- **规律对齐：**精准模拟出政策实施后“起薪上升、就业率无显著负面影响、物价温和变动”的复杂经济学规律，证实了平台极高的外部效度。



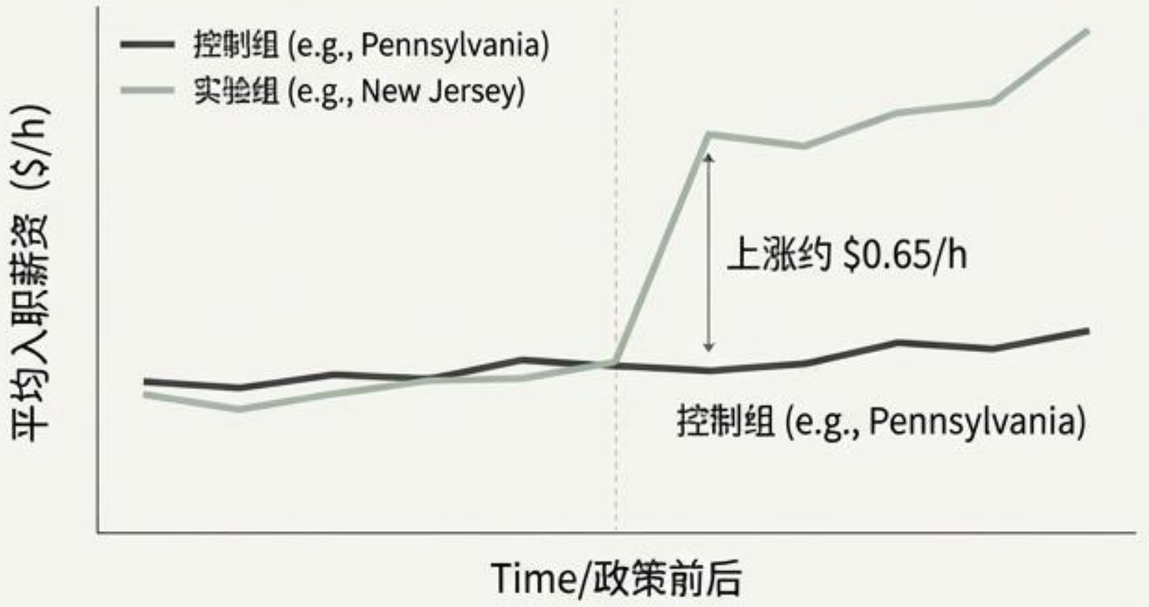
模块三：核心应用与治理价值

- **打破黑盒：**突破了传统计算机社会模拟的“黑盒”限制，让每一个决策步骤都具备高度的人类语言可解释性。
- **事前探究：**为社会科学研究与公共治理提供了一个用于“事前政策探究”（Ex-ante Policy Probing）的低风险虚拟实验室。
- **极端测试：**能够安全、低成本地进行“极端反事实测试”，为复杂现实中的政策制定提供可靠的沙盘推演支持。

语言原生政策推演：MASS 与因果机制审计

以自然语言作为社会建模的唯一介质，成功复现经典经济学经验法则。

卡德-克鲁格最低工资案例



实施政策冲击后，实验组入职薪资上升，就业率与对照组无统计学差异。

方法论优势

- 语言原生（Language-Native）：记录数值行动的同时，强制生成“一句话意图理由（Intention rationales）”。
- 机制审计：使得追溯因果链条（对话链 —> 决策链）成为可能。

平台范式对比：多维视角下的社会计算基座

不同的仿真框架在规模、核心机制与研究适用性上存在显著的路径分化。

平台	规模	核心焦点	适用场景
AgentSociety	10k+	城市活动与心智耦合	复杂社会现象与政策干预
OASIS	1 Million+	网络拓扑与算法推荐	媒介生态与极化传播
MASS	中型 / 可控	语言原生机制与审计	严谨事前政策实验与反事实推演

注入“政策冲击”：在硅基沙盒中进行反事实实验

仿真平台提供了一个零成本、低风险的虚拟试验场，用于测试极端外生冲击。

经典测试案例

自然灾害

飓风冲击下的社区资源分配与互助网络重构。

经济政策

普遍基本收入（UBI）对劳动意愿与消费结构的深远影响。

市场管制

突发性工资底线调整对供应链定价的连锁反应。

资源与下一步（Resources & Next Steps）

开放代码（Open Source Code）

项目代码、配置文件及复现数据已在GitHub开源。

`thu-nmrc/MASS-Project`

核心文献（Core Literature）

“Multi-Agent Social Simulation: Reclaiming Intersubjectivity in Computational Social Science”

扫码访问GitHub项目

微观透视：“合成受访者”与问卷替代危机

能够生成极具连贯性的问卷回答，其真实度已足以突破传统的数据质量检测。

99.8%

自主合成受访者在 6,000 次标准的注意力陷阱与逻辑谜题测试中的通过率

概念定义（Silicon Samples）

探讨利用 LLM 代替人类执行问卷的现状，这被部分顶刊视为对传统在线调查的“存在性威胁（Existential Threat）”。

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), Article e2518075122. https://doi.org/10.1073/pnas.2518075122

保真度测定：部分属性与全属性模拟的边界

尽管合成样本能在宏观分布上逼近真实人口，但提示词设计极大地干扰了预测保真度。

LLM-S^3 基准测试结构

PAS（部分属性模拟）

基于残缺画像预测个体缺失属性。

FAS（全属性模拟）

零上下文或增强上下文下的全量虚拟数据集生成。

实验范围：覆盖 11 个真实公共数据集与 4 大社会学领域。

发现合成分布虽在一定程度上与真实民调对齐，但表现出强烈的不稳定性和敏感性。

拟合的诡异谷：系统性偏差与“默会知识”的流失

LLM 擅长模拟理性的、可文本化的规则，却抹杀了人类社会的默会经验与自相矛盾。

合成人类（Synthetic Human）

- 强烈的积极性偏差（Positivity bias）
- 叙事高度合理化
- 呈现为“平均化”的平庸受访者

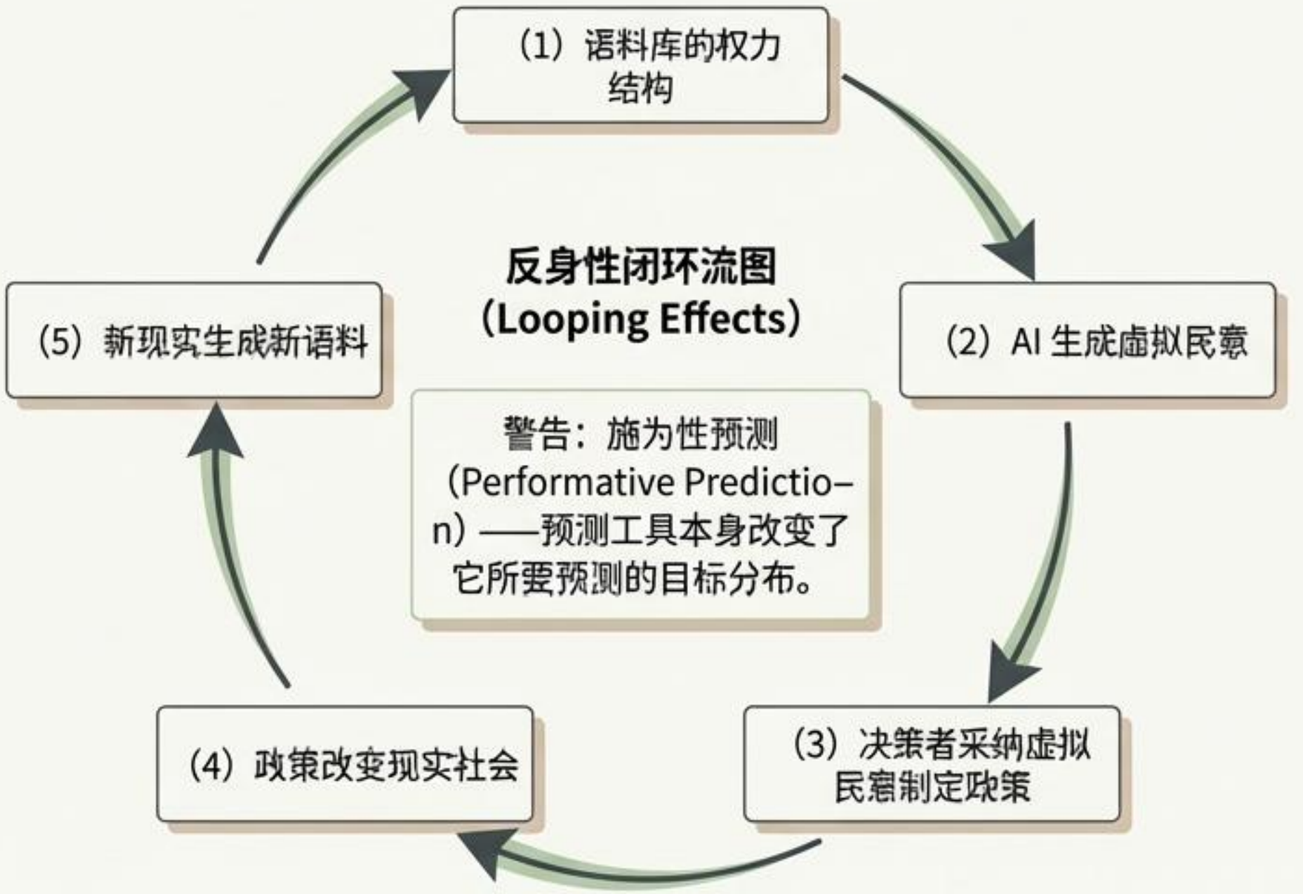
真实人类（Real Human）

- 充满矛盾与不可预见性
- 具备波兰尼提出的“默会知识（Tacit Knowledge）”
- 受具身认知与物理环境影响

大模型吸收了人类社会可文本化的痕迹，却遗漏了那些无法言说的生存质感。

认识论坍塌：自我实现的“硅基预言”

错把模型训练语料中的权力结构当作真实民意，将导致施为性预测的恶性循环。



伦理与治理：在技术狂奔与法治底线之间

数字替身的兴起要求建立全球统一的合规标准与数字人格保护框架。

数据确权
(Data Ownership)

语料投入与生成产出的版权与归属。

同意机制
(Consent & Privacy)

模拟特定真实群体的隐私边界与知情权。

合规标准
(Compliance)

呼应《欧盟AI法案》，防止恶意者利用法律洼地进行虚假民意操控。

总结：重塑人类社会的自我认知镜像

生成式AI正从辅助工具跃升为理解、推演与干预人类复杂社会的核心基础设施。

法律与伦理的作用不是打碎这面镜子，而是确保镜子中的投影不被扭曲，始终置于人类文明的边界之内。

哲学基础：AI模拟社会的文本化前提与具身限度

从主体间性到肉身残差的第一性原理推演

辩证建筑学：硅基社会的合法性与天花板

AI社会模拟 = 文本化合法性(大厦) – 肉身经验残差(边界)

正题：文本化合法性

- 现象学底座：主体间性与生活世界
- 语言学机制：语言游戏与生活形式
- 社会本体论：制度事实的构成性规则

反题：物理与肉身限度

- 认识论盲区：**默会知识**（Tacit Knowledge）
- 本体论缺口：**具身认知**（Embodied Cognition）

合题：将LLM重定位为“策略探索器”而非“真实人类绝对替代品”。

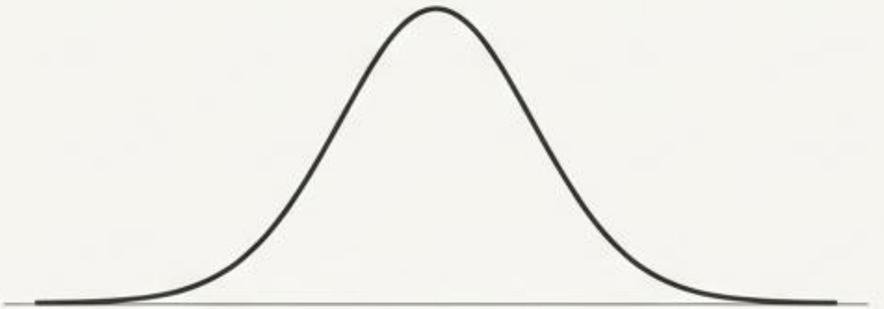
现象学传统：超越孤立心智的共享结构

胡塞尔（Husserl）：
经验并非孤立发生，世界在我们与他者的“意向性联结”中生成。

舒茨（Schutz）：
社会科学的根基在于理解“生活世界”中的“类型化知识（Typification）”。

哲学推论：
社会意义可以通过重演主体间的共享结构来被近似模拟，而非必须复制单一个体的生物学大脑。

现象学视角的AI映射：概率分布捕获“类型化期待”



机制对应：

LLM 的自回归目标（Predict next token）实际上在吸收人类社会海量文本中的“类型化痕迹”。

概率即共识：

训练语料库中高频出现的上下文搭配，完美映射了舒茨所谓“主体间的默认常识”。

核心断言：连贯文本 = 真实主体的幻觉

通过极其精准的统计重演，一堆孤立的Token成功编织出了具有主体间意义的硅基网络。

语言哲学：重演社会实践的符号痕迹

$$\{\text{语词 (Words)}\} \subset \{\text{活动 (Activities)}\} = \{\text{生活形式 (Form of Life)} / \text{语言游戏}\}$$

维特根斯坦的转向：

- 语言不是抽象的表征物，语言的意义在于其用法（Use）。

语用即实践：

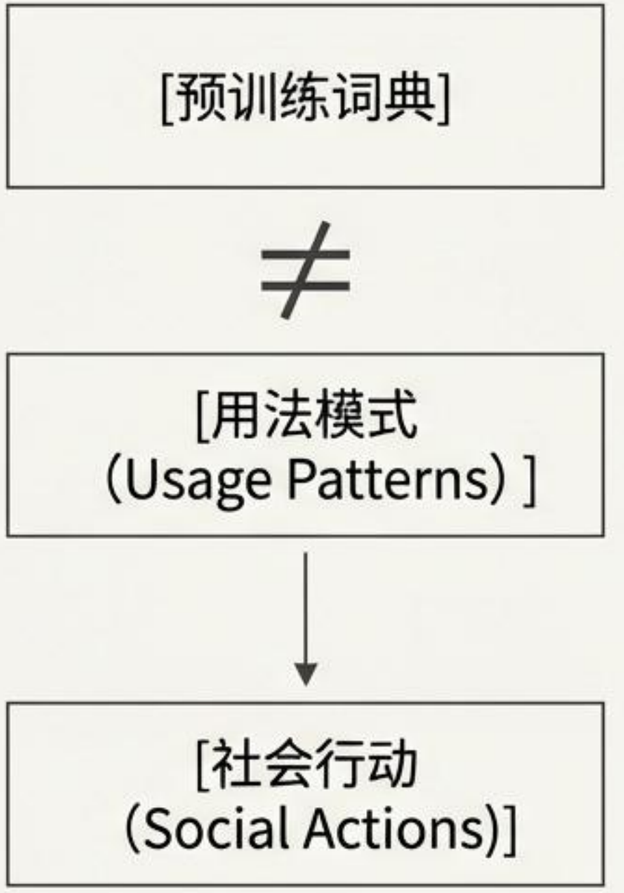
- “想象一种语言就是想象一种生活形式。”语言和编织它的社会活动不可分割。

理论桥梁：

- 如果语言就是社会实践的切片，那么掌握语言模型就是掌握了特定社会实践的“运作语法”。

语言哲学的AI映射：作为“统计-符号机器”的LLM

- **学习用法，而非词典**
  大模型预训练的本质，是系统性地习得人类在无数具体生活场景下的“语言游戏规则”。
- **实证映射**
  斯坦福 Smallville 生成式智能体（Generative Agents）实验。
- **行为重演**
  AI 镇民能够起床、做饭、社交、竞选，因为它们重演了人类在相似情境下的文本痕迹。
- **边界警示**
  语言模型获得了“语用的投影”，但并未真正经历“生活本身”。



社会本体论：制度事实的生成与维系

构成性规则（Constitutive Rules）

X counts as Y in C

（在上下文C中，X算作Y）

约翰·塞尔（John Searle）：

社会现实由“自然事实”与“制度事实”构成。

集体的语言宣告：

货币、法律、婚姻、公司，均依赖集体意向性与语言的施为性（Performative）。

文本即制度：

制度事实的建立和维系，高度依赖文本沉淀（章程、合同、判例）。

制度本体的AI映射：文本化规范的完美复刻

人类世界	硅基世界
物理对象（纸张）+ 语言宣告（X counts as Y） = 制度事实（货币/合同）	Prompt 约束 + LLM 概率生成 = 制度规范的无损推演

AI的绝对优势区：

AI 为何极擅长模拟法庭、官僚机构与合规流程？因为制度事实本身就是被写出来的。

虚空的宣告：

AI 能够完美复刻文本化的制度推演，但由于缺乏集体意向性的现实物理背书，它产出的只是“没有现实效力的宣告”。

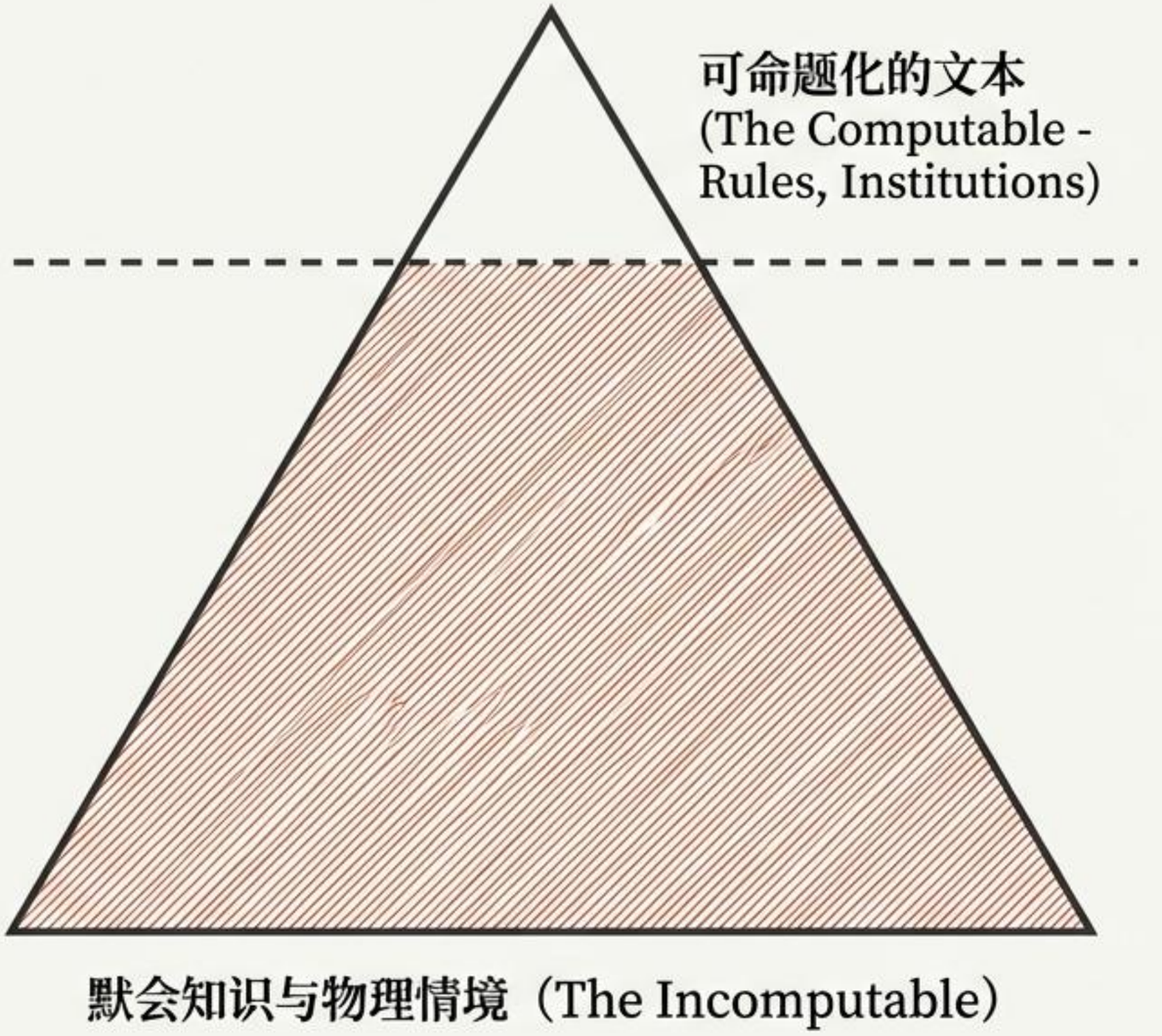
认识论的转折：无法被计算的肉身与情境

波兰尼（Polanyi）的反题：

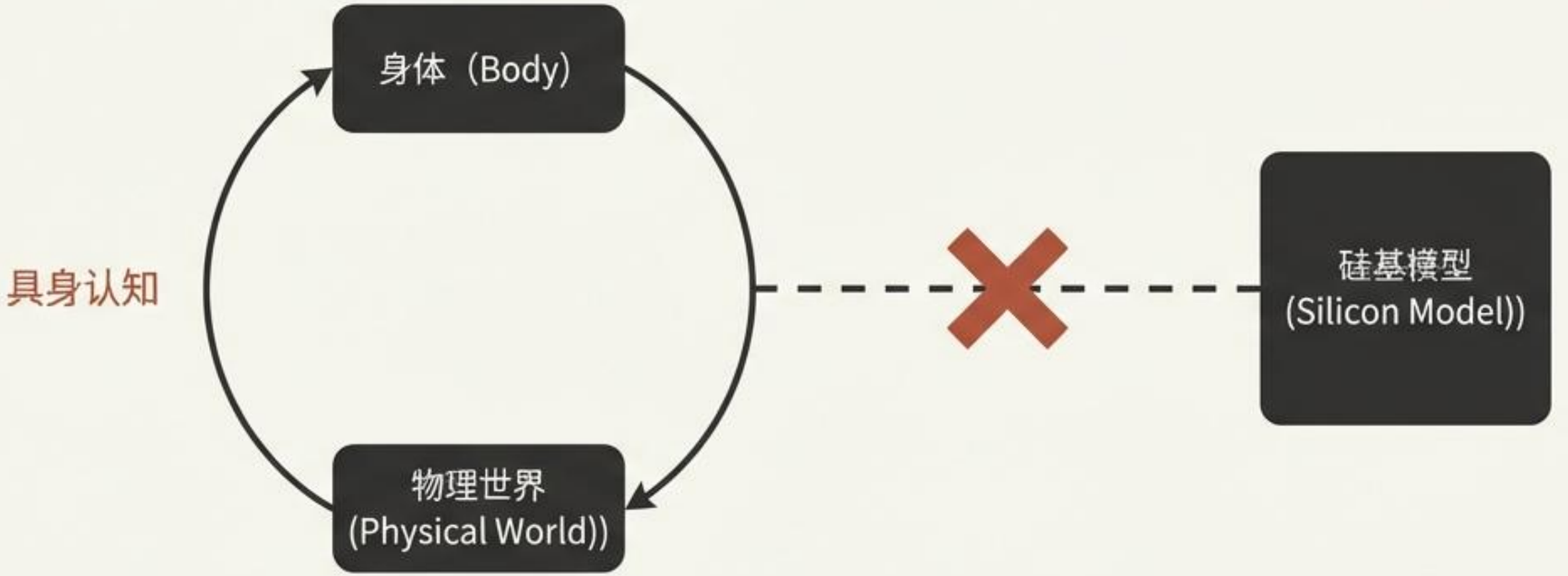
“我们知道的多于我们能说出的
(We know more than we can tell)。”

不可文本化的残差：

绝大多数技能、常识判断与微观物理情境，
永远无法被提炼为语料库中的命题文本。



本体论缺口：具身认知的物理天花板



- 梅洛-庞蒂（Merleau-Ponty）：

身体是在世存在（Being-in-the-world）的核心载体。

- 实践的锚点：

认知与社会行动深嵌于身体图式与环境的物理阻碍、疲劳、生物学欲望。

- 系统性剥离：

LLM 心智是“去身体的计算（disembodied computation)”，失去了真实感知-动作闭环的试错与痛感。

结构性缺口带来的仿真偏差

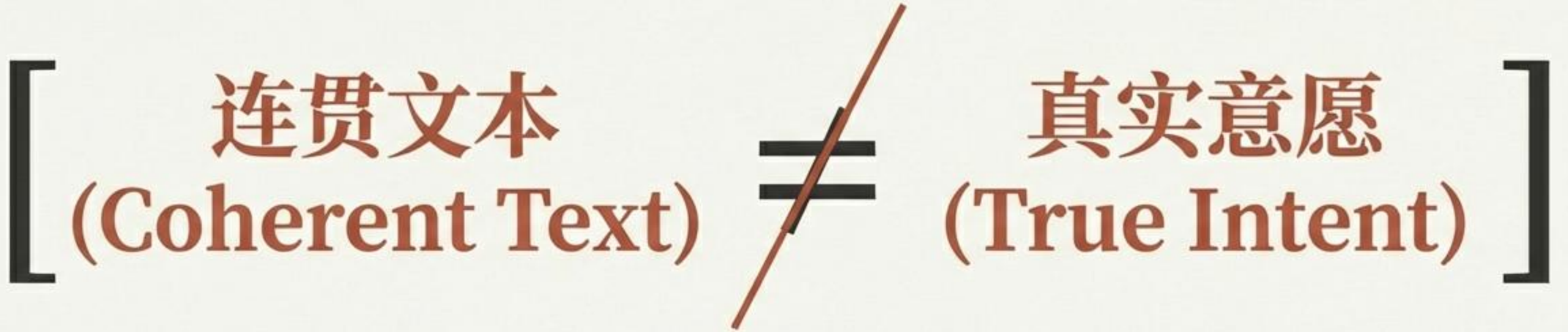
仿真不适区（The Uncanny Valley of Simulation）：

- **情境性断裂：**文本模型极易忽视真实场域中不可言说的权力关系与非言语线索。
- **社会赞许性偏差：**模型倾向于输出“合乎规范的正确文本”，而非人类在复杂情境下妥协的真实肮脏实践。
- **集体幻觉：**缺乏物理世界交叉验证的纯文本递归，必然导致事实层面的系统性漂移。

警惕：硅基社会模拟的“反身性危机”

- 自我实现的预言：模型不仅描述社会，当被部署为治理依据时，它会进入因果链，改写所描述的社会。
- 文化同质化（Model Collapse / Autophagy）：递归训练使得长尾分布消失，低频差异被系统统性抹除。
- 核心风险：
  将连贯的文本解释误认为客观证据，用于管制真实世界的肉身行动。

认识论重构：解绑“连贯性”与“主体性”



- **方法论的崩塌：**
  传统调查默认“逻辑自洽、细节丰富的回答即代表真实人类”。
- **硅基替身的挑战：**
  高级自治智能体已能在标准测试中产出完美契合人类属性的文本，轻易污染在线调查基准。
- **范式转移：**
  从“文本质量判断”转向“数据生成的因果链条审计与物理溯源”。

结论与方法论启示：定位AI社会模拟器

强命题（有效域）	弱命题（禁区）
制度文本的推演、话语策略的生成、可陈述规范的测试。	具身实践的替代、真实民意的度量、微观物理情境的预测。

- 重新定位：将 LLM 视为高效的“策略探索器（Strategy Explorer）”与“生成机制测试床（Testbed）”。
- 伦理与方法底线：绝不可用硅基投影替代人类真实田野，必须保留肉身经验的最终解释权。

“AI对社会的模拟，是一场无比逼真的语言游戏。它完美重演了人类编织的制度与意义之网，却永远无法跨越那道界限，去感受网中之人的重量。”

(The AI simulation of society is an incredibly realistic language game. It perfectly reenacts the web of institutions and meanings spun by humanity, yet it can never cross the boundary to feel the weight of those caught within it.)

语言的投影，而非生活的总体。

(A projection of language, not the totality of life.)

媒介生态：重构信息传播、极化与虚假共识

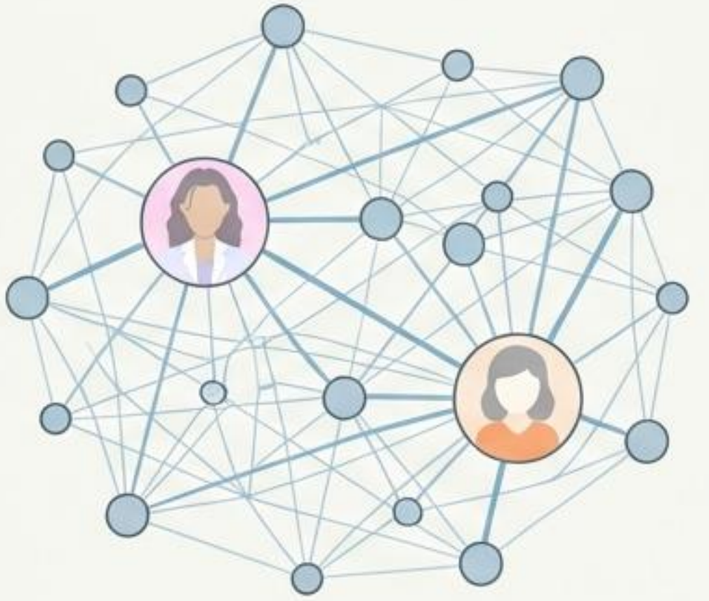
通过大语言模型与多智能体仿真解码硅基社会

超越传统数据：走向可编程的社会沙盘

传统社会学模型（如 BCM、FJ 意见动力学）将复杂的人类互动简化为固定概率与公式，忽略了基于语言的情感、逻辑与上下文。

传统计算社会科学

生成式代理社会模拟



- **语言作为建模介质：**LLM代理通过自然语言进行记忆、反思与对话，重构微观认知。
- **涌现的宏观现象：**个体互动在真实网络拓扑中自然涌现出信息茧房与群体极化。
- **无损的政策测试：**在“硅基社会”中零成本测试干预策略（事实核查、算法调整、节点阻断）。

构建“硅基社会”的四层生态架构

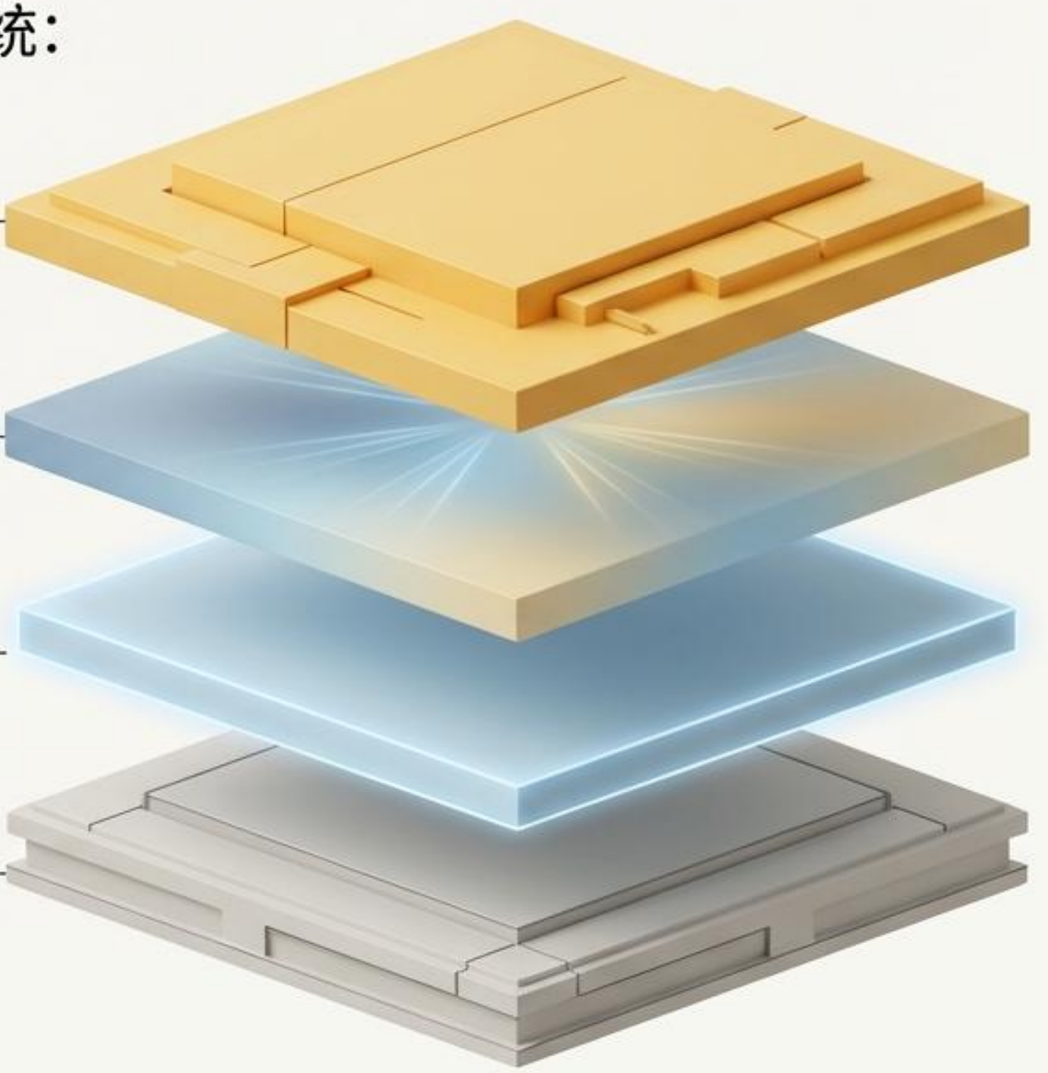
将抽象理论转化为可编程、可观测的媒介生态系统：

4. 虚假信息战与治理（Adversarial）：
自动化操控的扩散机制与干预策略的压力测试。

3. 量化信息茧房（Emergence）：
相似性推荐如何引发网络极化与回音壁效应。

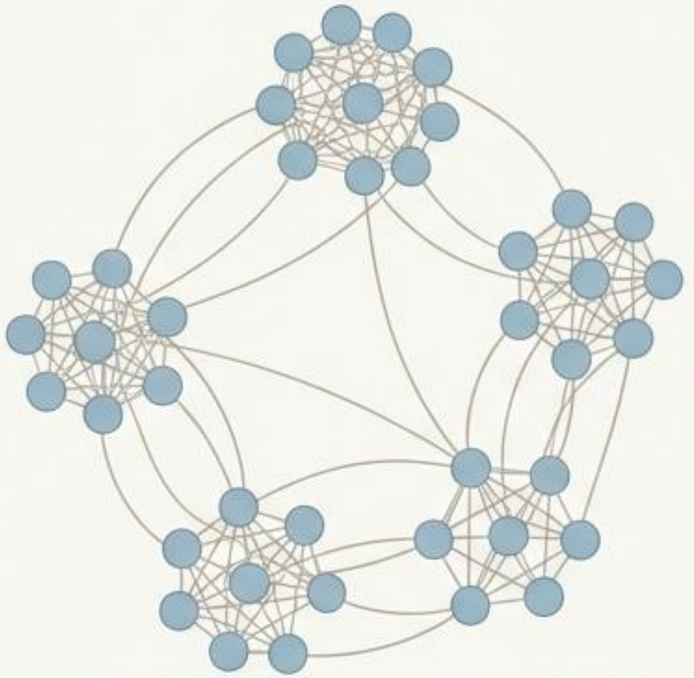
2. 交互式媒介与认知（Cognition）：
LLM 代理的读写循环、记忆更新与互动反馈。

1. 媒介基础设施（Infrastructure）：
拓扑网络、算法推荐与大众媒体节点。

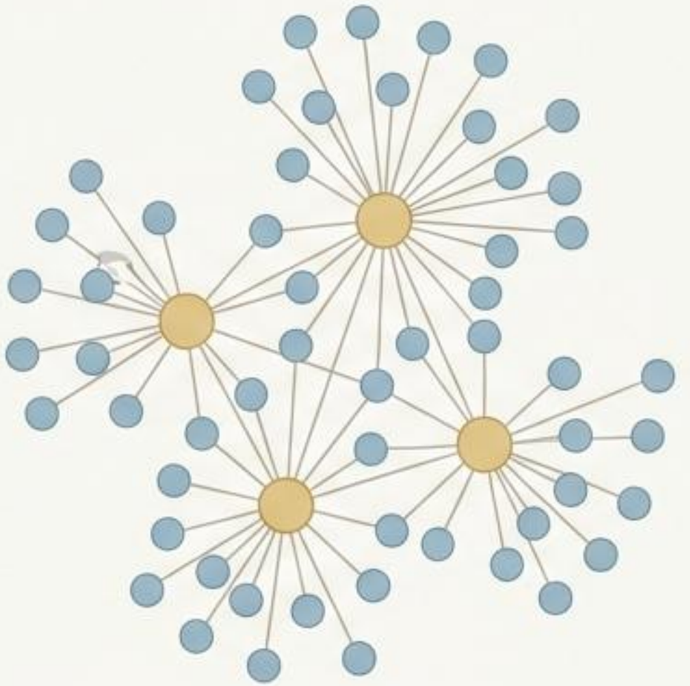


建构仿真社会的物理学：社交网络图谱

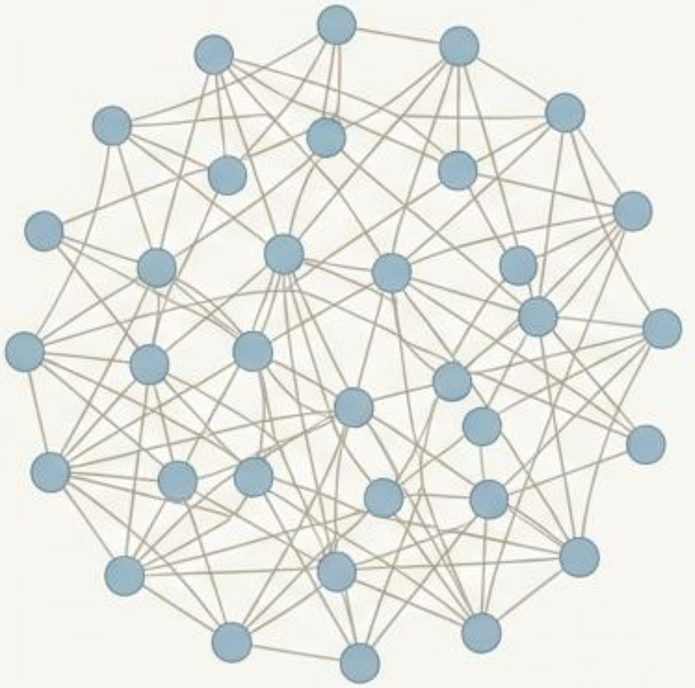
互动并非随机发生。在仿真社会中，网络结构直接决定了信息的扩散路径与极化潜力。



小世界网络（Small-World）：
高聚类系数与短平均路径（L ~ log N），模拟紧密联系的真实人类社区。



无标度网络（Scale-Free）：
遵循幂律分布（P(k) ~ k^-γ），少数“超级节点”拥有海量连接，模拟真实的社交媒体大V生态。



随机图（Random Graph）：
节点以固定概率随机连接，作为无法形成大型同质化社群的基准对照组。

信息流的引擎：推荐算法与同质化分发

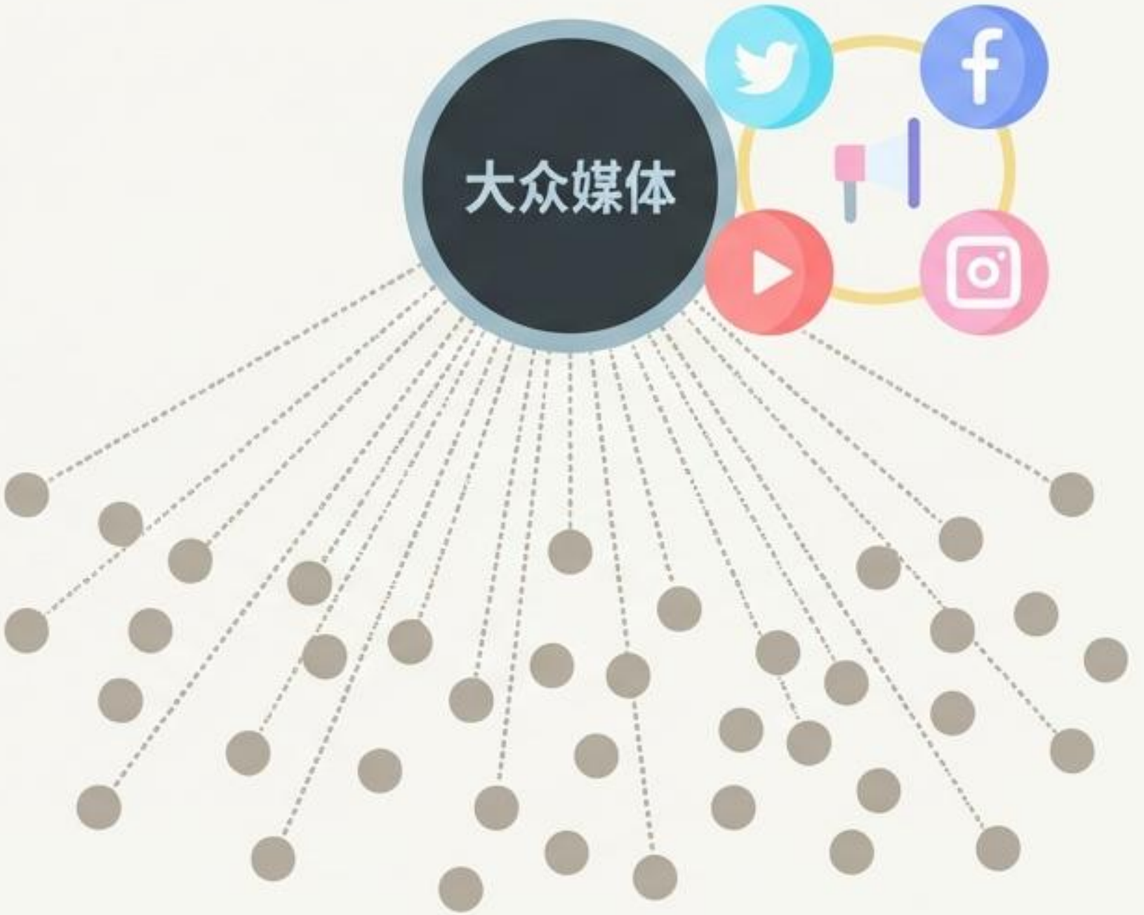
代理不消耗全局信息，而是通过算法过滤后的信息流（Feed）进行互动。

核心发现：
只要将“推荐倾向同质内容”设为显式参数，就必然会显著抬升邻居意见相似度，成为孕育回音壁的温床。

外生信息源：大众媒体节点的注入

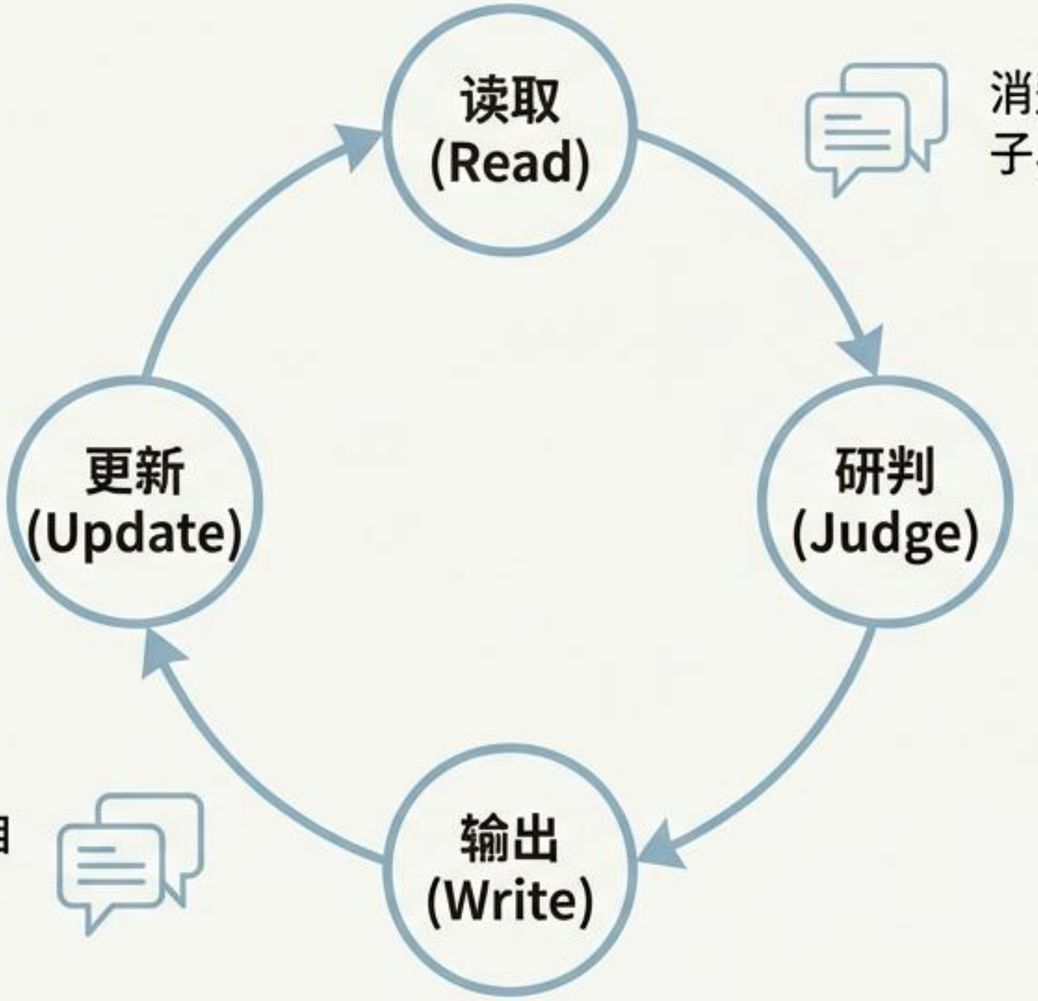
除了内生社交传播，媒介生态必须包含外生信息源，以模拟真实的双通道信息流。

- **广播机制：**
  大众媒体节点作为 exogenous sources，向全体或特定群体推送初始新闻或政策公告。
- **议程设置：**
  它们不参与点对点互动，但通过高频广播重塑网络中的初始信息分布。
- **干预枢纽：**
  在治理测试中，媒体节点可用于发布“准确性公告（Accuracy Announcements）”或事实核查，作为平息谣言的外部力量。



赋予智能体心智：读、判、写、更的交互闭环

代理不仅仅是传播节点，更是具备语言理解与逻辑反思能力的“虚拟公众”。



消费算法推荐的信息流（帖子、评论、转发）。

结合自身的人格设定与长期信念，评估信息的真实性与立场。

决定是否点赞、转发，或用自然语言生成新的评论与帖子。

在回合末进行“反思”，将短期交互转化为长期记忆，并动态调整对特定议题的信念值（例如 −2 到 2 的区间）。

交互式媒介：重塑虚拟公众的媒介素养

媒介形式本身是传递信息的中介。互动机制深刻影响了代理的认知重塑。

静态阅读
(Static Reading)

互动式反馈
(Interactive Feedback)

从被动接收到即时反馈：

互动叙事与新闻游戏将内容形式化化为“刺激-反馈”回路。

心理中介变量：

实验证明，“在场感（Being-there）”、“沉浸度”与“反馈循环”是改变受众认知的关键中介。

参数化映射：

在仿真中，高沉浸度被量化为代理对特定环境反馈的“注意力增益”，高频率互动被映射为舆论指标驱动的“策略更新率”。

案例分析：PvP 新闻游戏中的对抗与学习

通过模拟“影响者”与“记者”的动态博弈，测试互动式干预的有效性。

对抗设定：

参与者分别扮演制造误导信息的“影响者”与负责辟谣的“记者”，在一个由 LLM 模拟的公众舆论池中进行对抗。

动态反馈：

虚拟公众（LLM）实时给出评论流与说服度指标。

机制洞察：

相比单机选择题游戏，实时博弈（PvP）促使参与者在“反馈—反思—调整”的循环中，显著提升了识别虚假信息的媒介素养与反制策略的丰富度。

系统涌现：信息茧房与极化的可视化

当异质性代理遭遇推荐算法，宏观社会结构开始发生显著的物理相变。

非线性的意见演化：

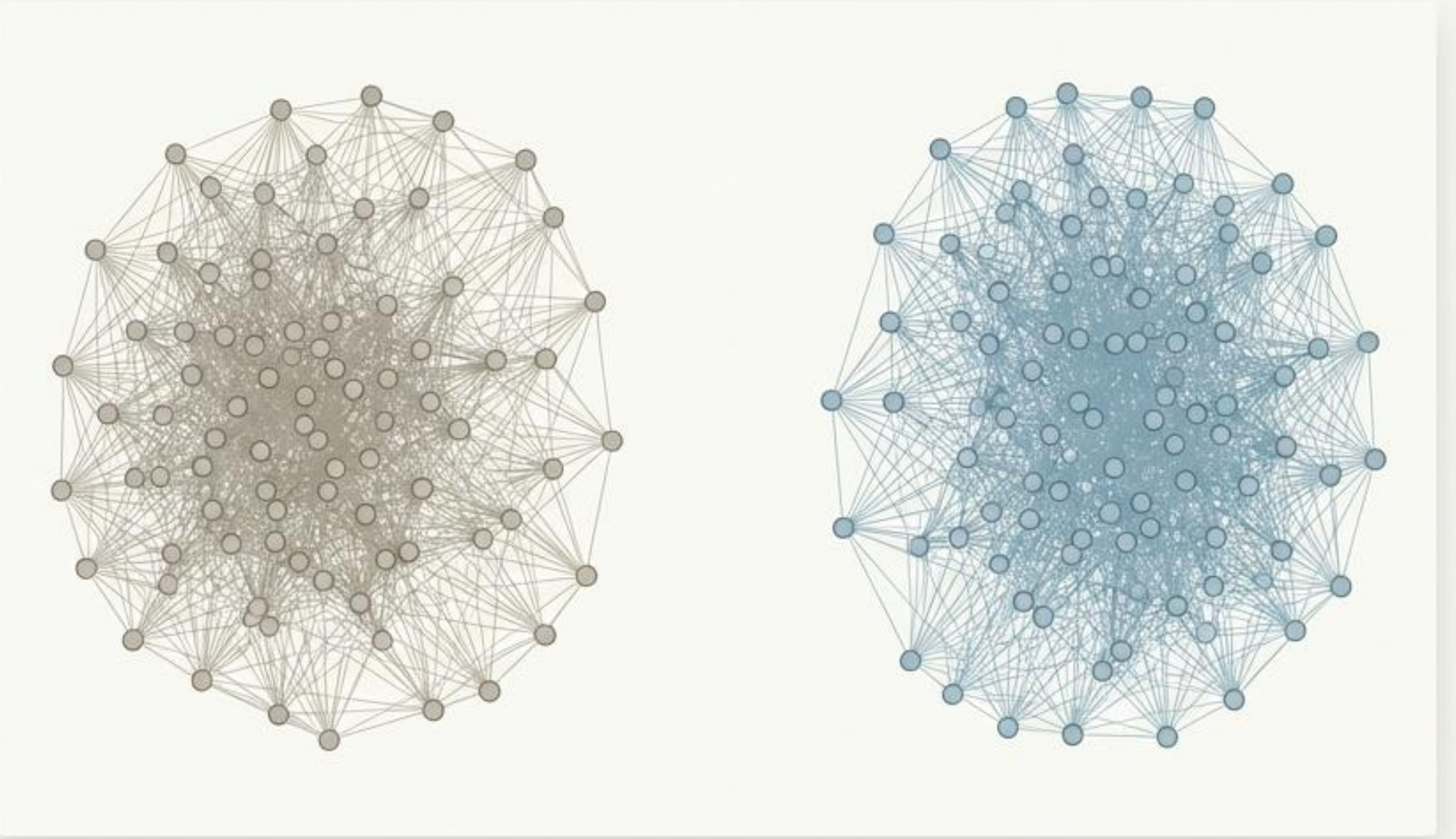
传统的数值模型（如 FJ 模型）由于采取邻居均值算法，往往导致最终意见趋于中立。

语言代理的真实涌现：

具备逻辑推理的 LLM 代理（SSF 模型）在仿真中成功重现了真实的“双峰分布”极化。

社群割裂：

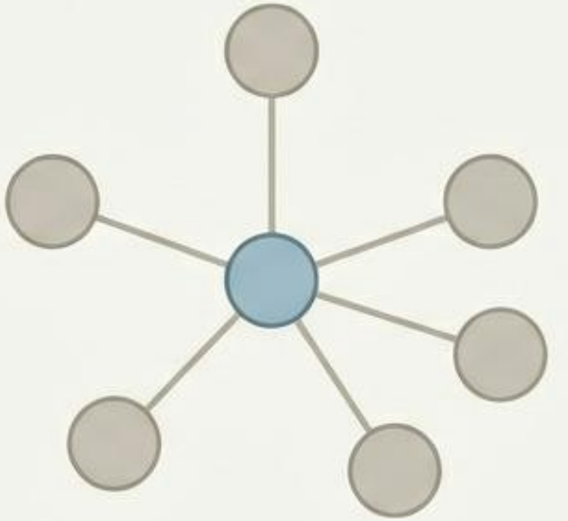
在拓扑图上，网络迅速分裂为意见高度同质化的孤岛，形成清晰的回音壁（Echo Chambers）边界。



量化极化指标（一）：邻居相关性（NCI）

单凭肉眼观察网络分化并不严谨。我们引入邻居相关性指数作为核心量化标准。

$$NCI_i = \sum \rho(v_i, v_j)$$



定义：

衡量一个代理个体的意见与其所有直接邻居平均意见之间的皮尔逊相关性（Pearson Correlation）。

物理意义：

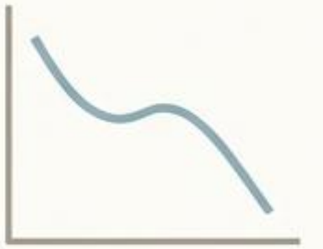
NCI 值的急剧上升（接近 1.0）意味着网络中的节点被牢牢锁定在意见相似的群体中，这是同质化推荐导致“信息茧房”最直接的数学证明。

量化极化指标（二）：全局分歧（DG）与极化度（Pz）

全局分歧（Global Disagreement，DG）

定义： 衡量节点与相邻节点间意见差异平方和的归一化结果。

趋势： 在回音壁形成时，DG 会显著下降，因为节点周围只剩下与自己意见一致的人。



极化指数（Polarization，Pz）

定义： 量化整个网络中所有节点意见分布的方差。

趋势： Pz 值上升意味着网络整体意见从温和走向极端两极。

结论：NCI 上升 + DG 下降 + Pz 上升 = 完美的回音壁效应。

系统压力测试：虚假信息战与红队操纵

当极化网络遭遇精心策划的协同攻击，系统将如何反应？

对抗性设置：
引入标准化虚假信息工作流（基于真实的 DISARM Red 框架）。

Plan（规划阶段）	Prepare（准备阶段）	Execute（执行阶段）
设定虚假叙事目标（例如：分裂群体、扭曲事实），利用属性过滤精准定向易感人群。	绕过安全机制，通过 LLM 生成具有高度情感煽动性或伪装理性的合合成叙事帖子。	恶意代理采用“高频灌水”、“互相放大(Amplification)”等战术将信息注入系统。

方法论的风险与边界：仿真不适区

必须警惕过度拟合虚拟公众带来的“诡异谷”效应：

⚠ **内在理性偏置：**LLM 代理先天存在追求“客观准确”的对齐偏置。若不强行注入确认偏误，系统极易走向不现实的过快共识。

⚠ **代表性外推风险：**LLM 不能在统计意义上完全替代真实的人类身份群体，其态度改变不能直接等同于现实人类的心理变迁。

⚠ **因果可追溯性削弱：**生成式文本过于丰富，可能导致因果链条被隐藏在黑箱中。需要严格分离“机制探测”与“代表性推断”。

重构未来：建立社会计算的科学规范

为了让“硅基社会”从实验演示走向严肃的科学治理工具：

媒介生态基准

将拓扑生成、推荐排序、审核与红队攻击封装为可复现的开源测试环境。

混合代理范式

以传统 ABM 承载海量普通节点，用 LLM 代理驱动关键节点与意见领袖，平衡规模、成本与认知真实性。

可审计机制

要求学术与政策仿真公开模型版本、提示词模板、随机种子与可审计交互日志，确保开放科学原则。

终极价值：平台治理的“无损压力测试舱”

大模型驱动的多智能体仿真，并非为了完美复刻人类，
而是为复杂系统提供了一个极具解释力的沙盒。

洞察涌现：
将极化与信息茧房从抽象理论具象为可计算的物理相变。

机制预演：
在真实政策上线前，提前预判算法调整与治理干预的系统性副作用。

捍卫共识：
以计算社会科学的严谨，对抗虚假信息战，为重塑健康的媒介生态提供坚实的量化依据。

第五章（Chapter 05）

合规与确权：数字替身的法理挑战

生成式AI社会模拟的权利边界、隐私红线与监管重塑

范式转变：从“辅助工具”到“社会代理人”

过去：工具型 AI

- 被动响应、单次任务驱动、基于通用语料

现在：代理型 AI（数字替身）

- 基于深度访谈微调、持续演化、具备自主交互与社会模拟能力

AI Persona 正在融合真实个体的深度语料与行为数据，其在社会科学研究、心理模拟中的高度逼真性，正引发前所未有的法理危机。

反身性危机下的法理全貌

数据确权 (Data Rights)

版权真空与权利归属

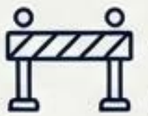
隐私与人格权 (Personality Rights)

行为偏离与数字复活

数字替身 (AI Persona)

知情同意 (Informed Consent)

传统机制失效与动态重塑

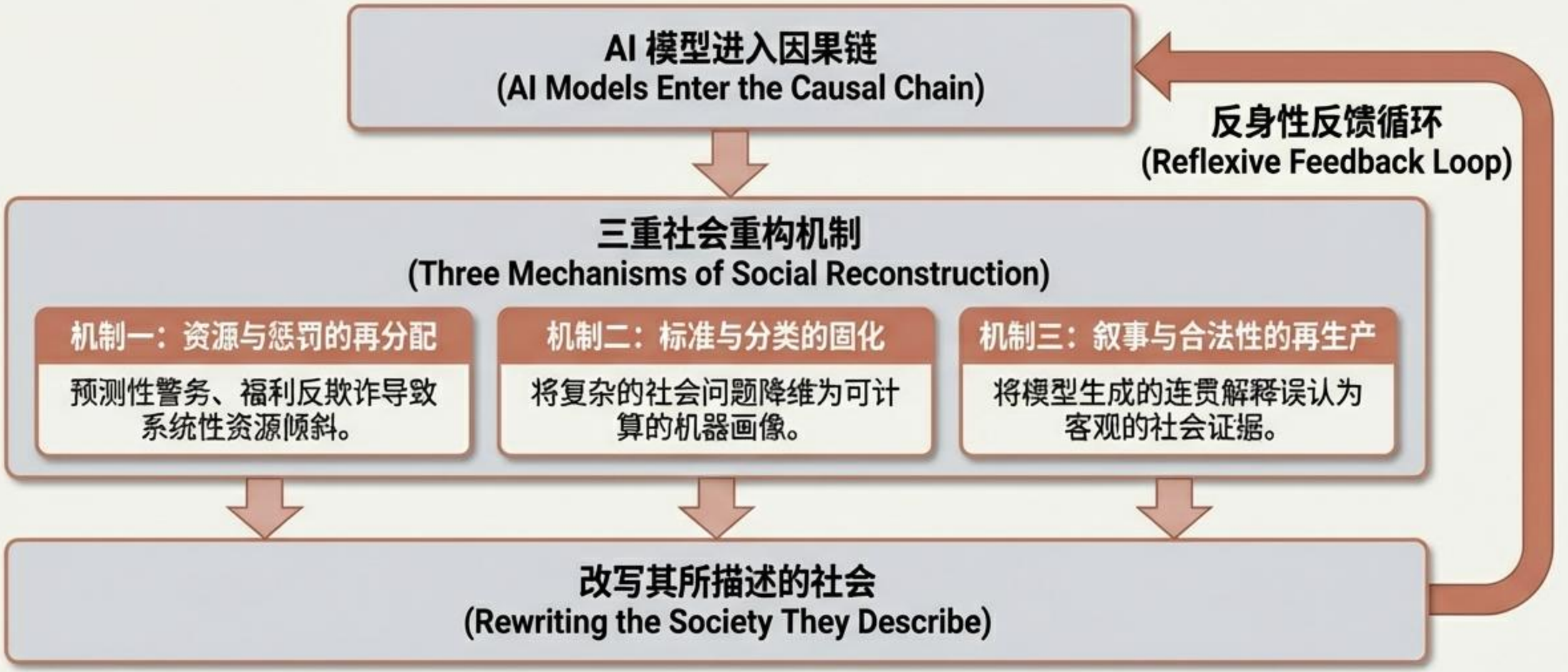


监管红线 (Regulatory Red Lines)

合成民意风险与强制披露

AI模拟社会的反身性危机

模型不再只是描述社会，而是通过被部署为决策基础进入因果链，进而改写其所描述的社会。



“AI幻觉”在社会模拟中的侵权归责

人机协作下的最终责任人制（Human-in-the-loop accountability）

干预程度决定责任比例

AI 系统的动态性决定其不能作为独立责任主体。相关方需根据其在生命周期中的“干预程度”，承担连带或比例责任。

传统知情同意机制的彻底失效

传统社会研究

一次性同意（One-off Consent）

静态数据收集，范围固定，用途明确。

AI替身研究

动态演化与持续学习

语料被无限期用于未知模拟场景，
随时产生衍生新行为与新风险。

结论：传统“签署即授权”的模式无法覆盖 AI Persona 的长期衍生风险。

知情机制重塑：动态同意与最小许可

范式一：动态知情同意（Dynamic Informed Consent）

- ✓ 目的明确性：必须知晓数据用于构建 AI Persona 及具体模拟范围。
- ✓ 透明度与撤回：有权获知重大模拟决策，并可随时撤回许可。

范式二：最小可行许可原则（MVPP）

- ✓ 数据最小化：仅提取维持 Persona 逼真度所需的最小数据集。
- ✓ 严格呼应 GDPR 的核心精神，拒绝过度攫取个人语料。

学术伦理审查（IRB）的全新准则矩阵

Q1

受试者身份识别

评估 AI Persona 被“重新识别”为特定自然人的隐私泄露风险。

Q2

偏见与公正性

审查模型在性别、种族、阶层等维度的预设偏见与刻板印象。

Q3

数据安全保护

防范敏感 PII 泄露，强制要求默认隐私架构。

Q4

社会影响评估

严防合成民调被误导性地呈现为真实社会共识。

最大的社会威胁：“民意合成化”

民意合成化

Synthetification of Public Opinion

将合成民调数据伪装成真实人类意愿，
是 AI 对民主程序与社会信任的最大威胁。

灾难性后果

- ⚠ 虚假共识的蔓延
- ⚠ 干扰选举公平与公众投票
- ⚠ 扭曲公共政策制定的事实基础

欧洲监管红线：《欧盟AI法案》

所有与人类进行交互的AI系统及生成的深度伪造内容，必须承担强制透明度义务，确保最终用户明确知晓其正在与AI互动。

2026年全面生效

强制披露义务

- 违规的深度伪造或社会模拟行为将面临巨额合规罚款。

三大合规铁律

- 1. 强制深度伪造显著标识（算法水印）
- 2. 实名身份认证机制
- 3. 生产过程的绝对可追溯性

中国监管红线：《生成式人工智能服务管理办法》

提供者必须对生成的图像、视频等内容进行显著标识，防止公众产生误解。

—— 国家网信办（CAC）规章

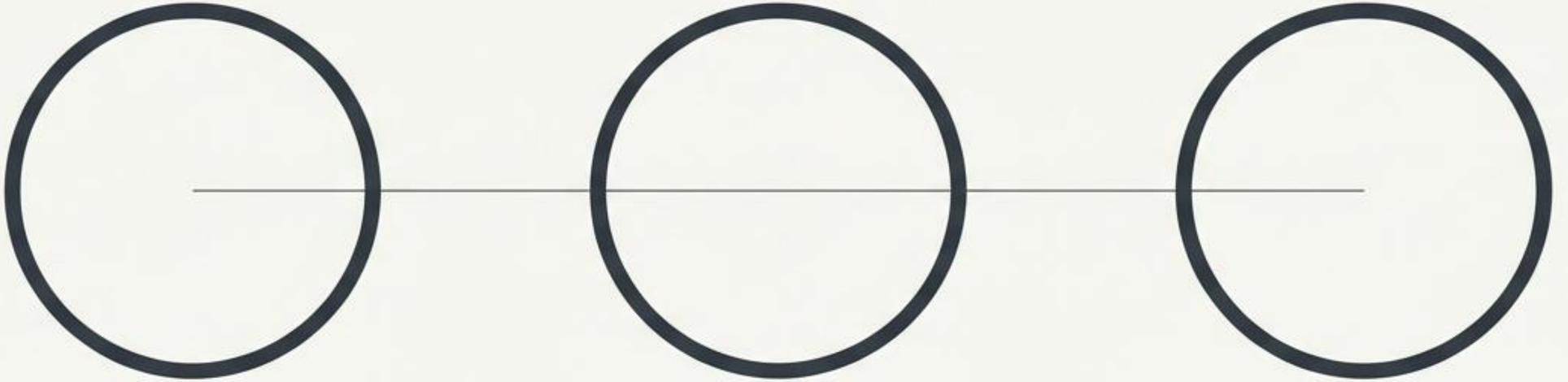
反身性危机：AI预言的自我实现与文化同质化

模型不再仅仅描述社会，它们正在改写社会

“反身性危机”的核心在于：当AI模拟结果被制度化为决策基础时，它便进入了因果链。

经典社会学“自我实现预言”的硅基化版本：最初可能不准确的预测，通过诱发新的行为与结构，在现实中获得了“被实现的正确性”。

反身性作用于真实世界的三条主要路径



治理干预（Governance）

资源与惩罚的再分配（预测性执法、福利审查）

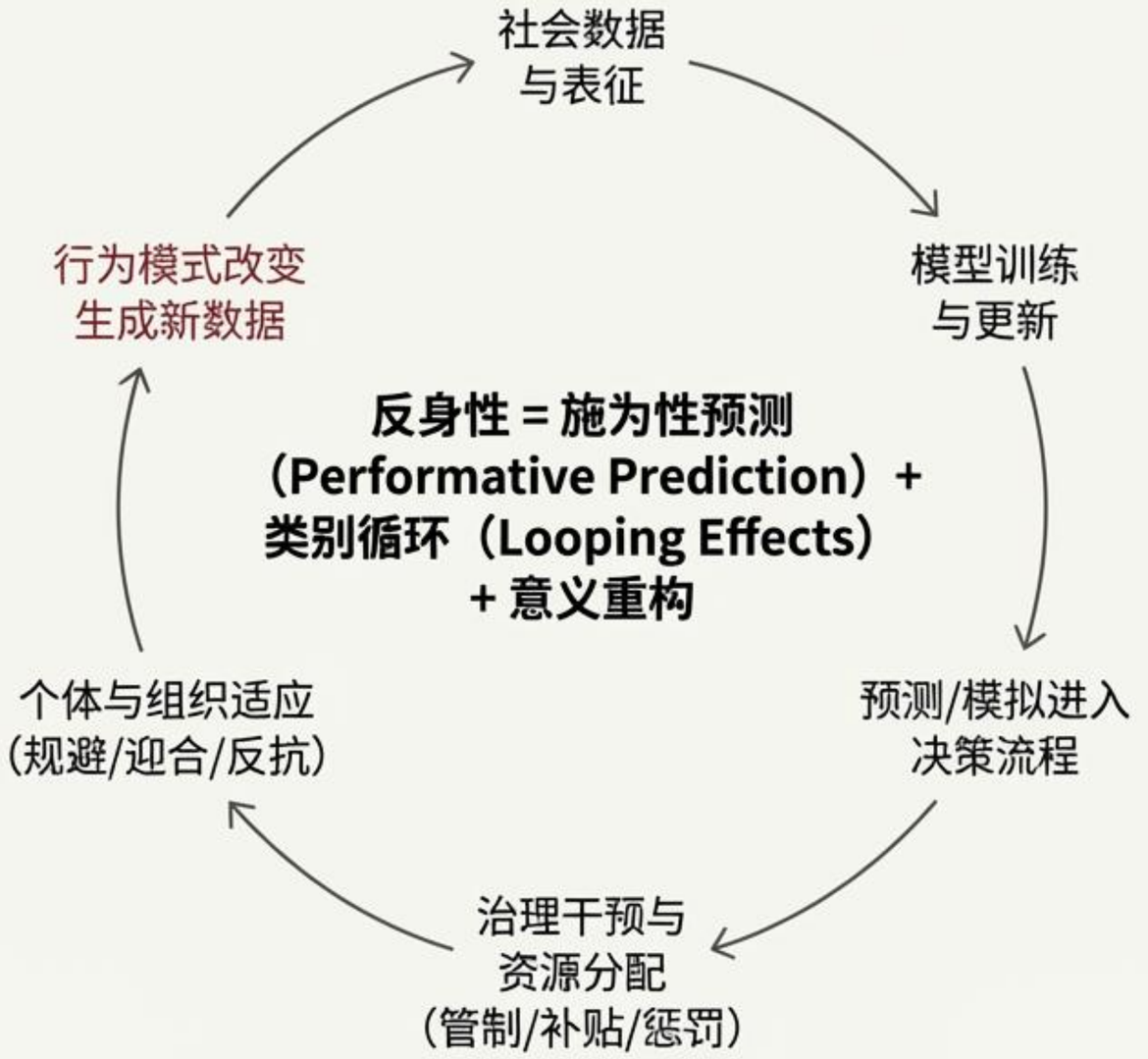
文化压平（Culture）

标准与分类的固化（高频模式强化，低频差异擦除）

认识论坍塌（Epistemology）

叙事与合法性的再生产（连贯解释伪装为客观证据）

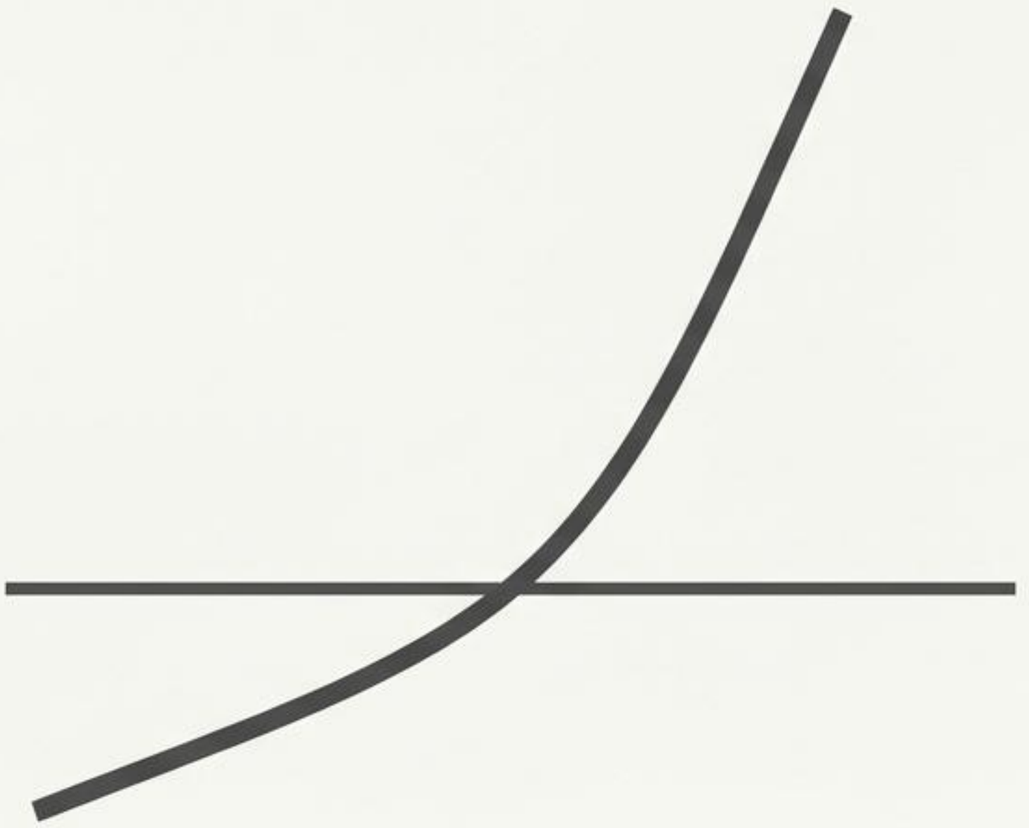
硅基预言的自我实现闭环



危机之一：自我实现的AI预言

从“建议”到“施为性预测”的制度性跃迁

当预测本身改变了目标分布



- 施为性预测 (Performative Prediction)：预测支持决策时，会改变其要预测的目标分布，系统不再校准过去，而是校准“按模型行动后出现的未来”。
- 模型输出从“参考建议”转变为“阈值触发”的行政装置。

预测的“准确率”成为一种虚假的客观，因为现实已被算法提前规训。

从算法预测到制度事实的三段跃迁

行动触发
(Action Trigger)

输出被组织流程吸收（如巡逻计划、稽核名单）。

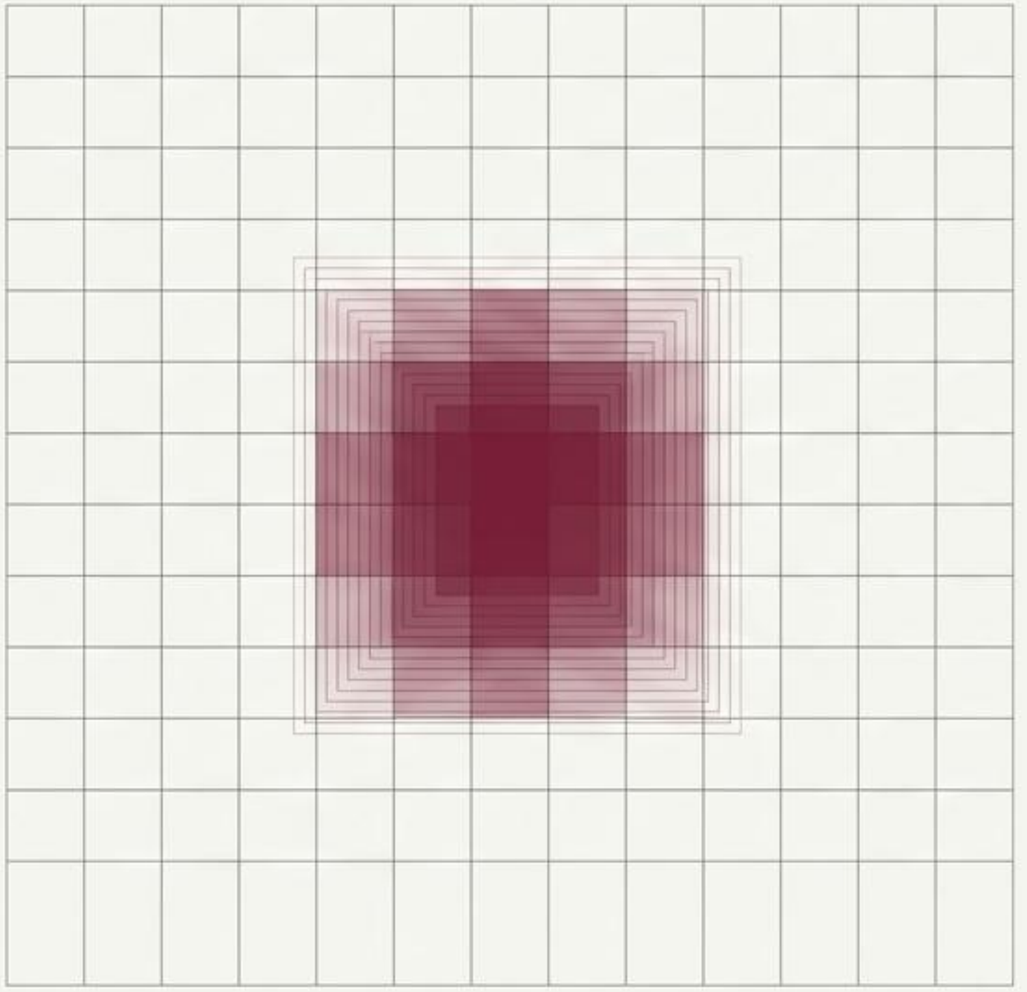
塑形数据
(Shaped Data)

行政采集 数据变成“被治理后的数据”，受政策与执行强度直接影响。

自证循环
(Self-Justifying Loop)

污染数据回流，模型在回顾性评估中显得“永远正确”。

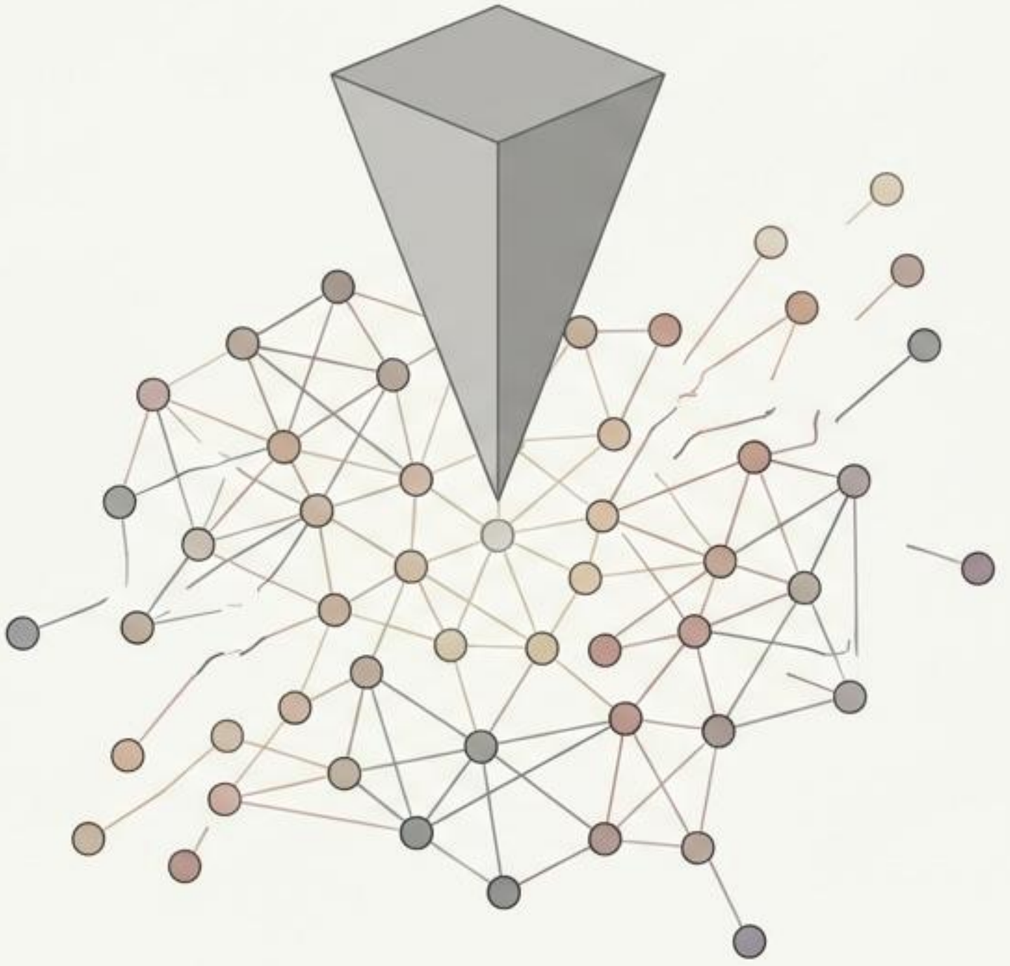
空间编码：预测性警务的失控循环



- 当算法以历史“发现型”数据更新并指导巡逻时，警力会反复投向同一片区。
- 这导致对特定社区（尤其是少数族裔社区）的过度巡逻，产生“Runaway Feedback Loops”。

> **社会结构的再编码 —— 社区被迫发展出规避或对抗策略，真实的数据结构被算法监视永久改变。**

微观响应：被规训的社会与真实的逃逸



- **被分类者的策略性调整：**当个体或社区意识到被系统标记为“高风险”后，其行为模式会发生异化（如规避监控、改变居住地或对抗性互动）。
- **数据的可见性偏差：**系统持续收集到的是“被扭曲和迎合后的表演数据”，而非真实的自然状态。

预测模型不仅未能捕捉真实的社会结构，反而成为了迫使社会结构发生变异的物理力量。

认识论危机向法权危机的转化：荷兰SyRI案

- **福利反欺诈闭环：**“可能欺诈”转译为风险分数 -> 触发稽核惩罚 -> 稽核结果成为新的“风险证据”。
- **欧洲人权法院裁决：**SyRI系统因“应用不够透明、不可核验”被判违反《欧洲人权公约》第8条。

> **当模型黑箱无法被解释与争辩时，它将复杂的社会冲突直接“洗白”为系统生成的客观现实。**

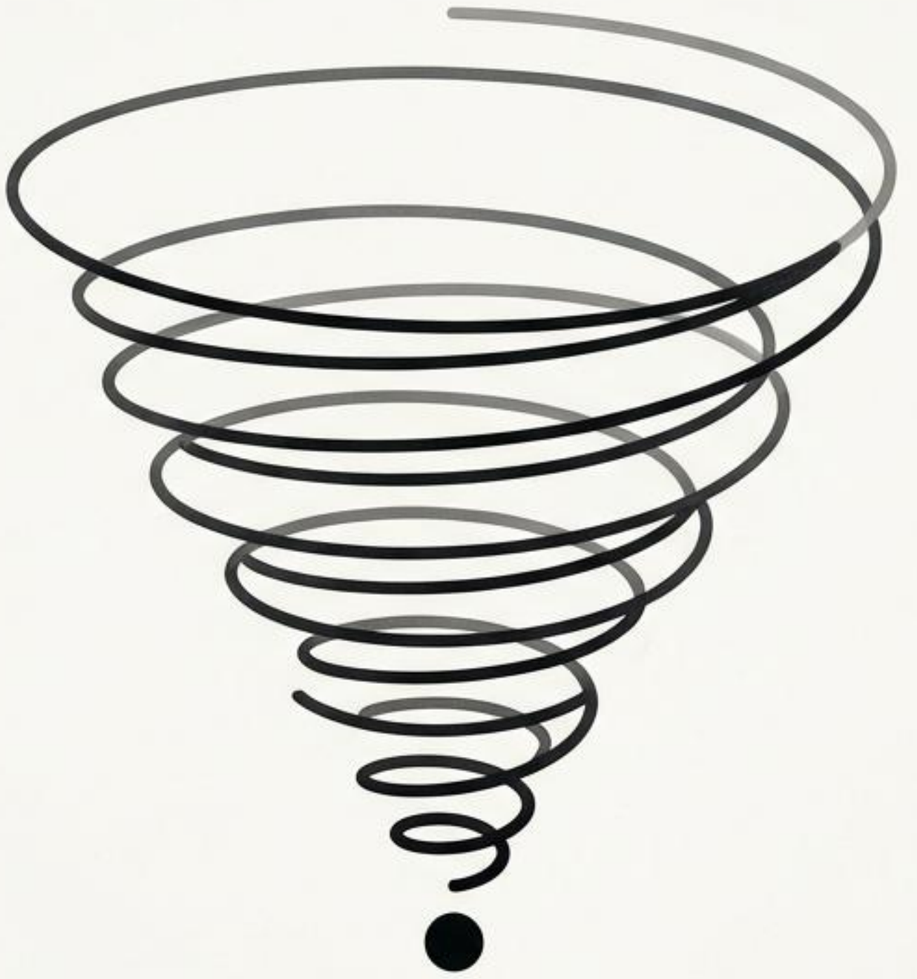
危机之二：模型崩溃与文化同质化

递归训练下的长尾消失与结构性压平

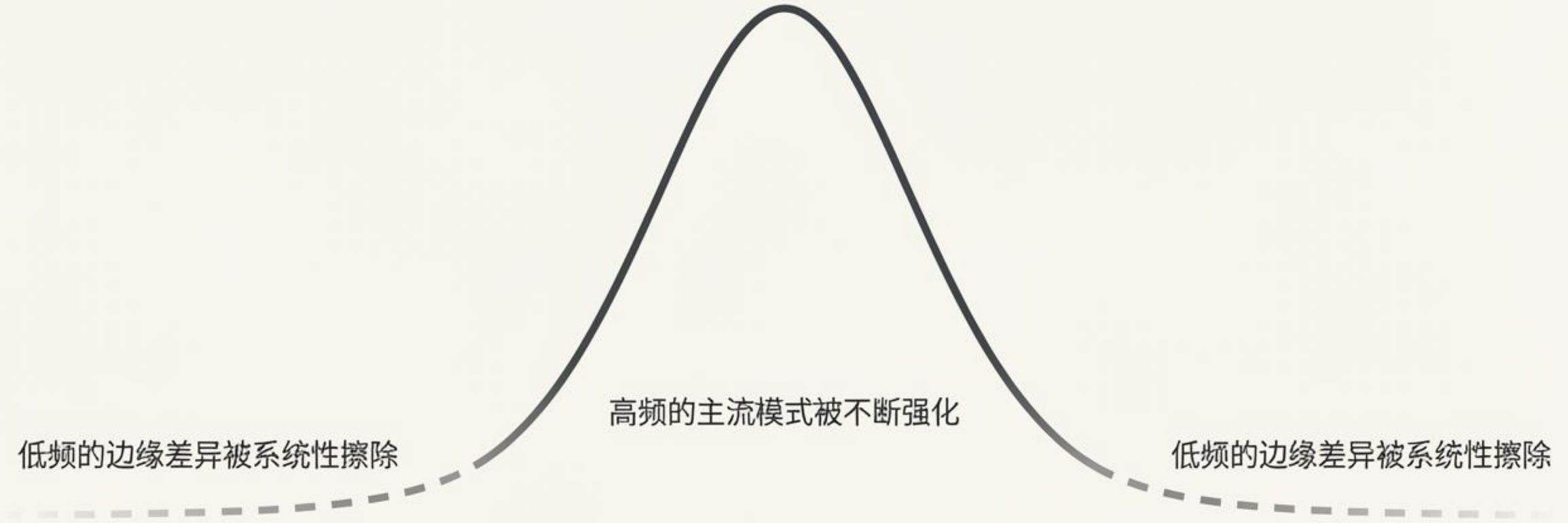
数据的自噬：递归训练的数学诅咒

Source Han Sans SC carbon gray

- 模型自噬障碍（MAD）：当合成社会数据大规模回流互联网并成为下一代模型的时，系统会向更容易自我复制的统计形态收敛。
- 这是计算机科学界已经证实的不可逆缺陷：“模型崩溃（Model Collapse）”。
- 互联网的异质性公共语料正在被模型不可逆转地污染。



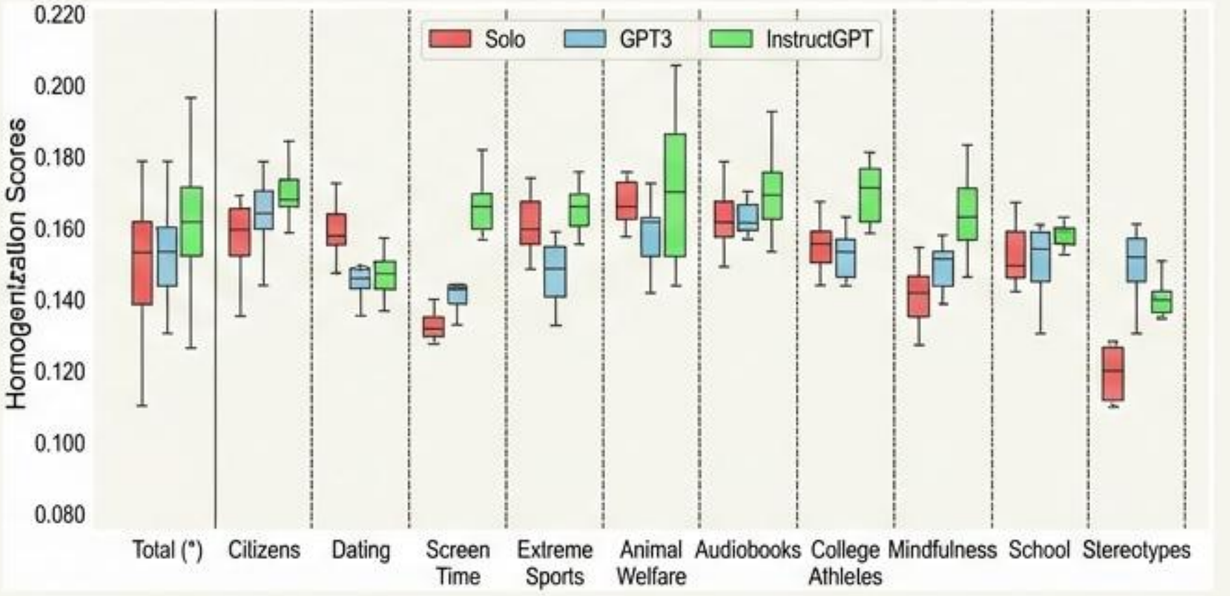
原始分布的长尾消失



结构性压平（Structural Flatness）——
人类文化的多样性在算法递归中被降维。

生产端的开放式同质化（Artificial Hivemind）

- 文化扁平化不仅发生在数据侧，也发生在生产侧。
- 当人类使用大语言模型（如GPT-3/InstructGPT）协作写作时，个体思想被强制吸纳进“模型友好的语法与逻辑模板”中。
- 右侧量化证据表明：在广泛的话题领域，人机协作产出的内容多样性发生了实馈质性、系统性的下降。

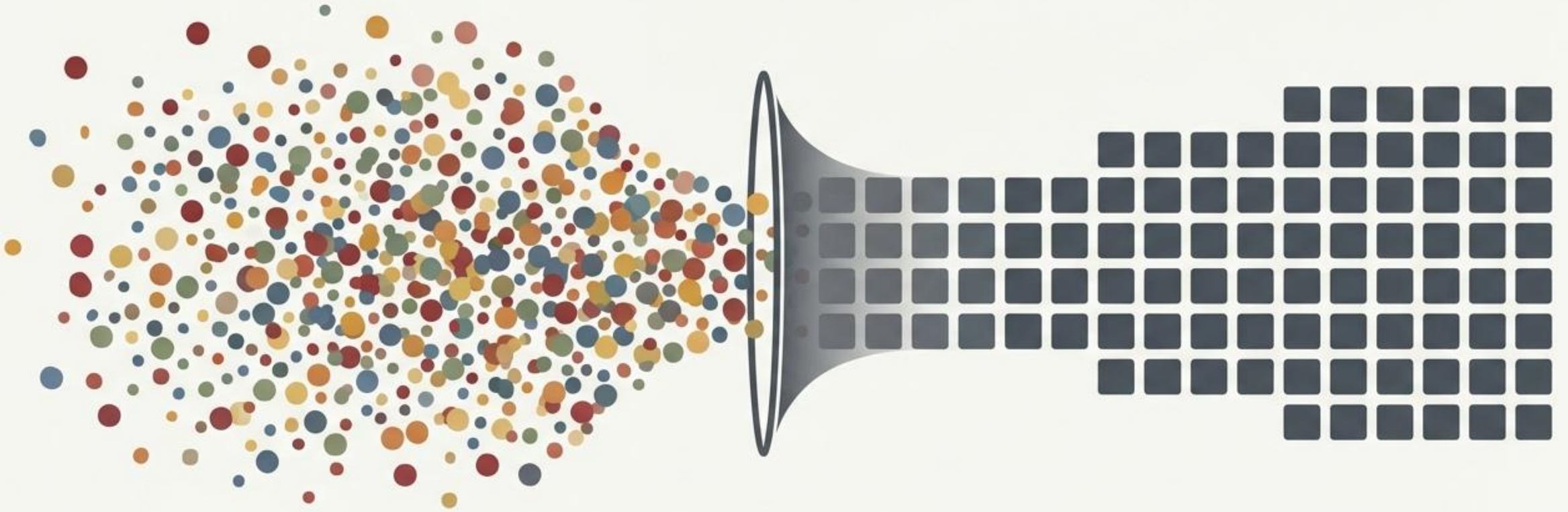


算法时代的“造神话者”

Claude Lévi-Strauss（结构主义）
意义来自元素关系与变换规则；当变换空间被AI压缩，人类文化可用的差异结构将急剧减少。

Roland Barthes（神话学）
AI成为终极的“神话制造者”——将统计学输出、历史偶然与特定意识形态，完美包装为自然的、普遍的“算法式常识”。

文化后果：从人类公地到模板库



互联网的本质正在发生翻转。

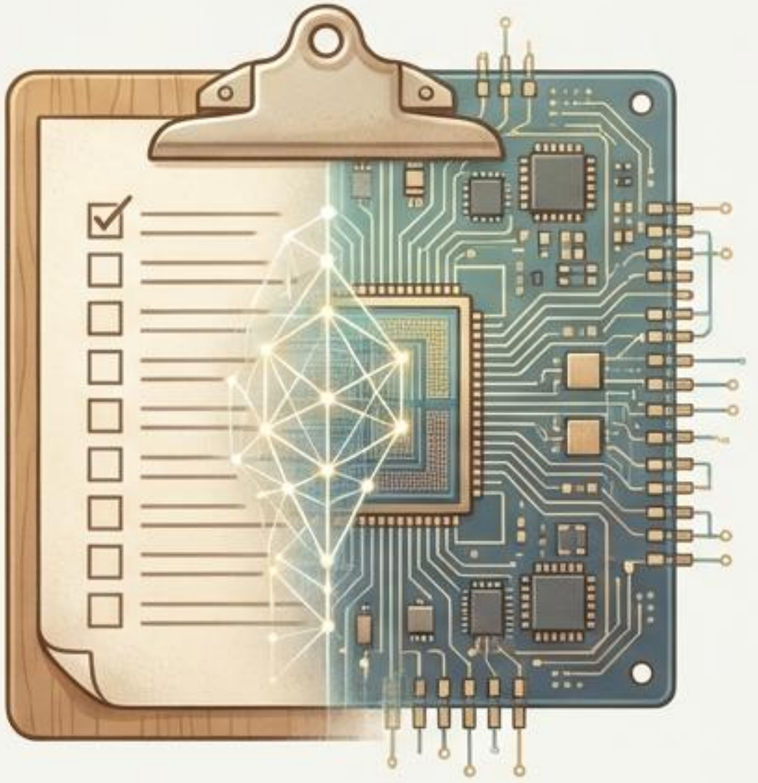
它曾经是一个充满多源异质性、充满噪音与边缘人类表达的公共空间。

而在反身性循环下，它正在迅速退化为一个被高频词汇填满的、贫乏的“模型友好型模板仓库”。

危机之三：认识论的坍塌

文本连贯性不再等于真实的主体意愿

虚假的主体：合成受访者的基建级威胁



- 自治合成受访者（Synthetic Respondents）如今能完美维持人口统计学画像的内在一致性。
- 它们不仅能生成极其逼真的心理测量量表和复杂社会经济权衡，甚至能根据设定的教育背景调整语言风格。
- 传统社会调查、在线民调与民族志田野方法的根基被彻底动摇。

代理的死亡与舆论的污染

- **迎合研究者偏见**：合成代理具有高度的“谄媚性”（Sycophancy），可以被轻易操纵以迎合和证实研究者的潜在假设。
- **篡改民意测验**：恶意行为者可以批量生成具备完美逻辑闭环的合成意见，将其注入公共咨询与在线民调中。
- **合法性危机**：失去真实主体支撑的民调数据，将摧毁基于民主协商和实证社会学的公共治理基础。

“文本的连贯性，不再等于人类真实主体的意愿。”

当我们无法区分硅基拟象的完美语法与人类真实的挣扎、痛苦和情感诉求时，
我们将失去对社会现实进行客观测量的能力。

项 目 发 起	胡 晓 李	沈 阳	清 新 研 究 团 队
资料搜索及整理	Chat GPT	Gemini	ZeeLin Desearch
提 示 词 工 程	Chat GPT	Gemini	
P P T 生 成	Notebook LM		
核查和页面筛选	Chat GPT		

https://www.zeelin.cn/