偏差级联：理解基于大语言模型的社会模拟中知识对齐与效应复现之间的差距

Pujun Zheng
华东师范大学经济与管理学院，中国上海，
pjzheng@stu.ecnu.edu.cn

Chengmu Wang，华东师范大学经济与管理学院，中国上海，
c
heng
mu.wang@stu.ecnu.edu.cn

Wanying Ren，华东师范大学统计学院，中国上海，
wyren@stu.ecnu.edu.cn

Jiacheng Yao
经济与管理学院，华东师范大学，中国上海
jcyao@stu.ecnu.edu.cn

Peizhou Yu
经济与管理学院，华东师范大学，中国上海
pzyu@stu.ecnu.edu.cn

张晓倩 经济与管理学院，华东师范大学，中国上海
xqzhang@stu.ecnu.edu.cn

查雄尧 经济与管理学院，华东师范大学，中国上海
xyzha@stu.ecnu.edu.cn

何国秀✉华东师范大学经济与管理学院，中国上海
gxhe@fem.ecnu.edu.cn

arXiv:2611.28854v1 [cs.CY] 21 Nov 2026

摘要

大语言模型（LLM）日益广泛地用于模拟社会实验中的参与者。现有研究主要关注社会知识与孤立响应的对齐，却忽视了模拟能否复现社会干预的集体效应。由于一种我们称为偏差级联的失效机制，这一区别至关重要：相似的个体状态可能导致不同的动作，而对孤立动作达成一致，仍可能因动态交互而产生不同的结果。因此，模型可能做出准确的社会判断，却无法复现经典人类研究的结果。为实证考察这一问题，我们推出 *REFLEX*（复现奠基性文献实验的效应），这是一个以经典人类研究为基础、包含 30 项反事实社会实验的基准测试，用于比较知识判断与效应模拟。知识判断直接预测干预效应的方向；效应模拟则固定情境与参与者，同时改变一个预先指定的社会因素，让智能体根据局部观测生成动作，再将这些动作汇总为宏观层面的平均处理效应。与人类研究发现相比，表现最佳的模型取得了 0.847 的净方向得分

知识判断为 0.453，效应模拟为 0.453。此外，因果链越长，效应复现效果越差，这与偏差级联假说一致。这些发现揭示了知识对齐与效应复现之间存在显著差距，表明当前方法仍不足以可靠地模拟集体人类行为。

✉ 通讯作者。

CCS 概念

• **C计算方法 → 建模与模拟** **lation.**

关键词

大语言模型；社会模拟；知识对齐；处理效应方向

1 引言

大语言模型（LLM）日益广泛地用于模拟社会实验中的参与者，为制度比较、干预设计和实验演练提供支持 [18, 20, 27]。这些应用要求模拟不仅能够捕捉人类对社会条件变化的反应，还能再现人类实验中观察到的集体效应。

现有方法大体遵循两种策略来提升模型能力：通过示范、人类偏好、模拟社会互动和结构化知识任务来对齐社会知识 [9, 23, 26, 29]；以及依据角色、人口特征档案、调查回应或个人自我报告来模拟参与者，从而近似目标个体或群体的判断 [3, 20, 32]。

然而，这些方法仍存在重要缺口。首先，从理论视角来看，现有方法强调孤立的决策或个体属性，却低估了动态互动机制 [19, 24]。逼真的模拟所需的不只是合理的个体判断。在社会科学中，宏观层面变量之间的经验关联不足以构成解释，因为宏观层面结果可能源自微观层面的过程 [15, 36]。要忠实地分析效应，智能体必须

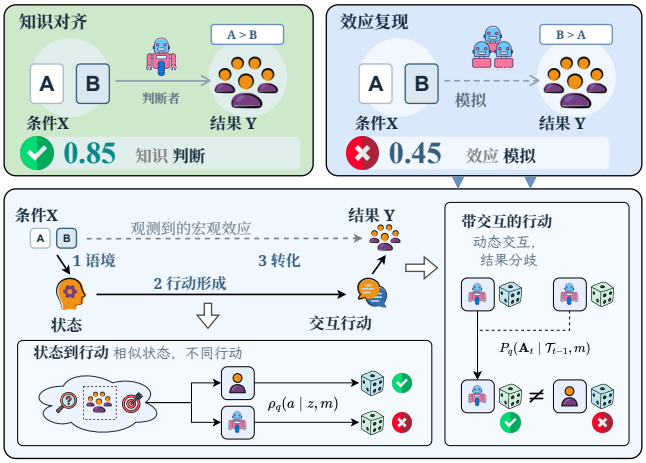

图 1：从知识判断到效应模拟。顶部面板比较知识判断与模拟结果。底部面板展示社会模拟必须对齐的情境机制、行动形成机制和转化机制，其中参与者在不同环境中进行交互，并将模拟效应与人类研究参照进行比较。

在动态环境中改变其动作，而不是像许多研究那样只做单向的静态选择。其次，从评估视角来看，模型评估往往优先考虑模型与观测响应的一致性，而非因果效应的逼真度 [7, 18, 25]。要检验某个社会因素是否会改变集体结果，需要通过干预改变目标因素，同时保持其他条件不变 [14, 17]。由此产生的处理效应（包括其方向），比孤立响应的一致性更能直接检验效应复现。若仅针对响应层面的一致性进行优化，就有可能奖励那些在干预下失效的相关性捷径。

我们借用社会科学中的科尔曼之船来提供理论解释：宏观层面条件通过情境机制塑造个体状态，这些状态通过行动形成机制产生动作，而交互行动则通过转化机制产生宏观层面结果 [10, 16]。这一框架揭示了两种潜在的不匹配来源。在行动形成阶段，模型可能再现人类状态，却采用不同的决策规则，从而产生不同的动作。在转化阶段，即使对孤立动作达成一致，智能体彼此交互时仍可能产生不同的反馈。这些阶段中的误差可能传播并累积，导致模拟效应偏离人类研究发现。我们将这种失败称为<bias cascade。

为了对这一假设进行实证检验，我们引入 *REFLEX*（重现基础文献实验中的效应），这是一项反事实社会实验基准测试，用于检验模型模拟能否重现经典社会科学研究中记录的干预效应。REFLEX 比较模型在两种设置下的性能。在知识判断中，模型直接预测给定社会情境中干预效应的方向。在效应模拟中，30 项任务中的每一项都从一个共享情境构建干预条件，在其余条件保持不变的情况下，改变一个预先指定的社会因素同时保持参与者配置及所有其他实验设置完全一致。模拟参与者根据当地可获得的观察结果采取动作，并将其汇总为宏观层面结果。我们将处理效应估计为不同条件之间的平均结果差异，并将其方向与相应的人类研究发现进行比较。

表现最佳的模型在知识判断和效应模拟上的净方向得分分别为 0.847 和 0.453。模拟得分较低，揭示了模型在复现人类实验效应方面存在显著局限；而不同评估模式之间的差距则表明，准确的社会判断并不能可靠地转化为忠实的模拟。因果链越长，效应复现表现也越差，这与提出的偏差级联解释一致。

我们的贡献如下：
(1) 理论分析。我们对社会模拟过程进行了形式化，表明状态匹配或孤立的动作匹配均无法保证忠实地复现宏观层面效应。**(2) 基准测试构建**。我们推出了 *REFLEX*，这是一个包含 30 项反事实社会实验的基准测试，用于评估模拟的处理效应是否与人类研究中记录的方向一致。**(3) 实验验证**。 我们量化了知识对齐与效应复现之间的差距，并发现不同大语言模型之间存在一致且显著的差异。**(4) 实证诊断**。 我们发现，较长的因果链与较差的效应复现表现相关，这为偏差级联提供了佐证。

2 相关工作

2.1 基于大语言模型的社会模拟

基于大语言模型的社会模拟研究主要沿着两个方向发展：模拟框架和智能体构建。框架通过定义智能体的背景、记忆和目标，并纳入环境反馈，对持续进行的社会交互作用进行建模。代表性系统包括支持记忆和规划的生成式智能体[31], 、将智能体决策与环境执行分离的 Concordia [39],、支持大规模社会实验的 AgentSociety [33], , 以及为人类行为模拟构建基础模型的 OdysSim [43], 。智能体构建方法利用人口统计属性、访谈数据或人类实验中的回答来创建硅基样本或个体智能体 [3, 6, 13, 20, 32]。总体而言，这些方法为社会模拟提供了平台，并提供了以个体层面属性为智能体提供依据的途径。我们关注的并非扩展平台或匹配静态回答，而是考察交互中的智能体能否在受控干预下忠实地复现集体处理效应。

2.2 社会模拟中的效度挑战

既有研究指出，社会模拟的有效性不能仅根据最终结果或平均响应来评估。误差可能源于模拟过程、响应分布或实验设置。在过程层面，模型可能通过错误的生成路径或机制得出正确结果 [21, 30, 42]。调度、信息暴露和环境

更新规则对集体结果的影响也可能大于智能体之间的沟通 [24]。在分布层面，合成响应可能接近人类均值，却压缩方差、扭曲变量之间的关系，或低估少数意见 [5, 7, 22, 34]。模型也可能描述目标分布，却无法在重复调用中生成与之相符的行为 [19]。在实验层面，提示词、角色分配和交互协议会显著影响群体结果 [2, 28, 41]。结果也可能取决于模拟配置、模型规模和行动选择方法 [8, 11, 45]。更广泛而言，反事实估计取决于模拟器，其现实世界相关性则取决于模拟器的逼真度 [14]。尽管这些研究指出了各种有效性风险，却没有解释为何具有人类对齐社会知识的模型仍无法在宏观层面复现实验效应。我们通过分析个体决策和多智能体交互中的不匹配如何在各阶段传播并累积，来弥补这一空白。

2.3 社会模拟的基准测试

现有基准测试评估群体响应分布、个体重测一致性、多轮社会交互，以及模拟过程和轨迹 [18, 21, 35, 38, 40, 44]。这些基准涵盖响应相似度、交互能力和过程合理性，但并不直接检验能否通过实际动作复现干预效应。相关研究已将评估范围从响应相似度扩展到效应复现。研究人员检验了大语言模型能否复现经典人类实验 [1]，并根据孤立的模拟响应估计处理效应 [4]。其他研究将实验材料、条件和统计程序纳入端到端评估 [27],，或分别测量行为分布拟合和情境适应 [37]。根据个体响应误差选择模型、提示词或目标群体，也未能始终降低处理效应误差 [25]。然而，大多数评估依赖静态响应，或将概率分布拟合作为因果保真度的代理指标。很少有研究使用反事实模拟来检验社会知识能否转化为涌现群体结果。我们的基准测试直接比较模型对干预的方向性判断，以及其通过交互式模拟复现这些处理效应的能力。

3 理论分析与基准测试设计

3.1 Coleman之船揭示的偏差级联

我们的研究探讨以下问题：现有大语言模型尽管具备相关知识，也理解特定条件下人类行为的后果，为何仍可能无法重现真实人类社会中观察到的结果？我们用偏差级联来解释这一失败。图 1 上半部分展示了这个问题。

我们借助社会理论中的经典框架——科尔曼之船，分析其深层原因 [10, 16], 。如图 1 下方的图示所示，宏观层面条件 X 与宏观层面结果 Y 之间的关系可以分解为一条机制链，从宏观层面延伸至微观层面，再回到宏观层面。情境机制解释了制度、资源和信息结构如何塑造个体的处境。行动形成机制解释了个体如何根据这些处境选择动作。转化机制解释了这些动作如何通过相互依存和环境反馈产生宏观层面结果。

经过充分的预训练和后训练对齐，现有大语言模型在许多价值判断、对社会规范的理解以及常规行为倾向方面，已经与人类十分接近 [12]。因此，我们提出一项关于模型如何理解社会情境的假设，作为形式化分析的起点：

> 假设。大语言模型能够理解社会情境。也就是说，对于给定类型的社会情境，它们理解智能体微观层面状态（即信念和目标）的能力与人类相当。

令 $q \in \{H, L\}$ 分别表示人类和模型，并令$m \in \{0,1\}$表示外部分配的宏观层面控制和处理条件。下文中，以 m 为条件指的是相应干预世界中的分布。令 T 表示完整交互轨迹，并令 g 表示两类系统共享的结果函数。目标处理效应为

$$\Delta_q = \mathbb{E}_q[g(\mathcal{T}) \mid m = 1] - \mathbb{E}_q[g(\mathcal{T}) \mid m = 0]. \tag{1}$$

第一个**偏差来源**存在于信念与行动之间。令 z统称给定个体的情境、信念和目标。令 $Q_q(z \mid m)$ 表示该个体微观层面状态的分布，令 $\rho_q(a \mid z, m)$ 表示给定该状态时的行动选择规则。这两个组成部分决定了个体行动的分布。为简单起见，我们考虑离散状态：

$$P_q(A = a \mid m) = \sum_z \rho_q(a \mid z, m) \cdot Q_q(z \mid m). \tag{2}$$

该方程区分了个体的状态与其据此采取行动的方式。即使采用比状态匹配更强的假设，也就是状态的联合分布 Q_q 相匹配，行动仍取决于从状态到选择的映射ρ_q。相同的情境、信念和目标，并不能唯一确定决策规则或其执行方式。在现实情境中，模型与人类在这种映射上可能有所不同：模型可能描述目标分布，却无法在重复行动中始终可靠地从中采样 [19]。因此，模型生成的行动可能偏离真实的人类行为。

命题 3.1（状态匹配并不保证动作匹配）。在两种干预条件下，微观层面状态相同并不足以保证动作相同：

$$\begin{aligned} & Q_L(\cdot \mid m) = Q_H(\cdot \mid m), \quad \forall m \\ \Rightarrow \quad & P_L(A \mid m) = P_H(A \mid m), \quad \forall m. \end{aligned} \tag{3}$$

构造性证明。 固定个体的其他情境和目标。假设在人类和模型中，当 $m = 0$ 时，二者认为合作有益的概率均为 1/4；而当 $m = 1$ 时，该概率均为 3/4。人类只有在认为合作有益时才会合作，而模型始终合作，这反映了偏好对齐所诱发的亲社会倾向。在每种条件下，这两个系统的信念分布完全相同，但

人类合作率从 1/4 上升至 3/4，而模型的合作率仍为 1。若结果是群体平均合作率，则 $\Delta_H = 1/2$ 和 $\Delta_L = 0$，因此效应不同。

第二个**偏差来源** 存在于边际行动分布与互动过程之间。令 $A_{i,t}$ 表示个体 i 在时间 t 采取的动作，$\mathbf{A}_t = (A_{1,t}, \ldots, A_{n,t})$ 表示所有个体采取的动作，S_t 表示环境状态。个体边际分布 $P_q(A_{i,t} \mid m)$ 给出特定个体在时间 t 选择某一特定动作的概率。在动态环境中，相关对象是完整交互轨迹 $\mathbf{T} = (S_0, \mathbf{A}_1, S_1, \ldots, \mathbf{A}_T, S_T)$。令 $\mathbf{T}_{t-1}$ 表示第 t 轮动作发生前的完整历史，即截至 S_{t-1} 的轨迹前缀，其中 $\mathbf{T}_0 = (S_0)$。根据概率链式法则，

$$P_q(\mathcal{T} \mid m) = P_q(S_0 \mid m) \prod_{t=1}^{T} \Big[P_q(\mathbf{A}_t \mid \mathcal{T}_{t-1}, m) \cdot P_q(S_t \mid \mathcal{T}_{t-1}, \mathbf{A}_t, m) \Big]. \tag{4}$$

该方程表明，即使大语言模型准确匹配每个个体的行动概率，个体间依赖关系仍可能有所不同。乘积中的第一个条件项描述个体如何根据历史采取联合行动，第二个条件项则描述它们的动作如何改变环境。至关重要的是，轨迹分布还捕捉了个体间的依赖关系、对既往历史的响应，以及动作与环境之间的反馈。个体动作的边际分布无法决定这些交互结构。经典分析表明，个体层面规律并不会自动决定集体结果 [15, 36]。因此，在复杂交互作用塑造的系统中，即便是微小的差异也可能导致显著的结果分歧。

命题3.2（行动边际分布一致并不保证交互效应一致）。即使在每个时间点、每种干预条件下，个体的行动分布都完全一致，完整轨迹的概率分布仍可能不同：

$$P_L(A_{i,t} \mid m) = P_H(A_{i,t} \mid m), \quad \forall i, t, m \quad \Rightarrow \quad P_L(\mathcal{T} \mid m) = P_H(\mathcal{T} \mid m), \quad \forall m. \tag{5}$$

因此，即使结果函数 $Y = g(\mathbf{T})$ 相同，仅匹配边际分布也无法保证宏观层面的结果概率分布 $P_q(Y \mid m)$ 一致。对于取决于行动组合或交互历史的指标，平均处理效应甚至可能符号相反。下文将给出一个涉及实际信息交换的构造。

构造性证明。考虑一项双人、两轮的协调任务，动作只能取 0 或 1。在两个系统中，两名个体都独立地以相等的概率选择第一轮动作。个体 1 在第二轮总是重复其第一轮动作。两种条件之间唯一的区别在于，个体 2 是否能在第二轮之前观察到个体 1 的第一轮动作。在控制条件 $m = 0$ 下，该动作不会被披露，个体 2 会重复自己的第一轮动作。在处理条件 $m = 1$ 下，该动作会被披露：人类个体 2 会遵循观察到的动作，而模型会选择相反的动作。两个系统的差异仅在于参与者如何回应观察结果。在每种条件下、每个时间点，每名个体选择 1 的概率都是 1/2，因此所有边缘分布都相同。然而，如果结果是第二轮是否协调成功，那么 $\mathbb{E}_H[Y \mid m = 0] = 1/2, \mathbb{E}_H[Y \mid m = 1] = 1$，和 $\Delta_H = +1/2$，而 $\mathbb{E}_L[Y \mid m = 0] = 1/2, \mathbb{E}_L[Y \mid m = 1] = 0$，和 $\Delta_L = -1/2$。两个系统中的个体行动概率相同，但信息披露在一个系统中促进协调，在另一个系统中则促进反协调，从而逆转了宏观层面效应。

总而言之，这些结果展现了模拟的难点。现有模型或许与相关社会知识高度一致，却仍难以重现真实人类社会中观察到的效应。核心挑战在于模拟过程中这两种不匹配来源。模型与人类之间的差异可能在各个阶段层层传递并不断放大，最终导致模拟结果与真实人类结果之间出现显著偏离。

3.2 反事实任务构建

我们将反事实任务定义为

$$\mathcal{P}_s = (S_s, m_s, \pi_s, g_s, \mathcal{H}_s), \tag{6}$$

其中，S_s 表示任务 s 的基础情景和初始条件分布，$m_s \in \mathbf{M}_s$ 表示实验条件之间的主动操纵，其中 $\mathbf{M}_s = \{0,1\}$ 分别代表标准双条件任务中的控制和处理，π_s 表示除该操纵外保持不变的执行协议，g_s 表示预先规定的行为读出函数，而 $\mathbf{H}_s$ 则包含任务的规格细节，包括处理定义、结果变量、效应对比、参考方向和适用范围。

最小操纵约束要求实验条件之间仅在预先规定的外生处理因素上存在差异。它并不要求所有处理后变量保持不变。惩罚的变化可能改变行动空间和预期收益；信息分配的变化可能改变参与者所说的内容；网络结构的变化可能改变暴露路径。这些变化可能是处理产生其效应的机制，因此，在处理后将它们固定下来会阻断待测量的效应。因此，每项任务都必须明确区分固定的初始因素、主动操纵本身，以及由操纵引起的后续状态变化。

同一任务中的所有条件均使用相同的模型设置、智能体规格和执行协议。环境随机种子成对设置，会话使用独立的语境和私人记忆。成对的环境随机种子仅控制实现层面的初始随机性；模型抽样噪声以及会话内的其他随机性仍然彼此独立。处理效应通过重复实验单位之间条件均值的差异进行估计，而非根据单次输出推断。

令 u 表示一个实验单位，并令 $\mathbf{T}_{s,u}(m)$ 表示单位 u 在任务 s、条件 $m \in \mathbf{M}_s$ 下产生的完整轨迹。将结果变量定义为

$$Y_{s,u}(m) = g_s\big(\mathcal{T}_{s,u}(m)\big). \tag{7}$$

对于多智能体任务，整个群体、网络或交互序列构成一个实验单位。群体之间的交互作用

成员被纳入单位层面的结果中，不得将成员视为独立的处理样本。

对于任务 s 和随机种子 r，令 $Y_{s,r}(m)$ 表示在条件 m 下，跨重复实验单位计算得到的主要结果均值。对于标准的双条件任务，我们使用

$$\widehat{\Delta}_{s,r} = \overline{Y}_{s,r}(1) - \overline{Y}_{s,r}(0). \tag{8}$$

对于包含多个干预条件的任务，处理效应按照预先指定的对比计算：

$$\widehat{\Delta}_{s,r} = \sum_{m \in \mathcal{M}_s} w_{s,m} \overline{Y}_{s,r}(m), \qquad \sum_{m \in \mathcal{M}_s} w_{s,m} = 0, \tag{9}$$

其中，权重 $w_{s,m}$ 在实验前固定，并且总和为零。

令 $h_s \in \{-1, +1\}$ 为任务 s 的参考方向，并令 $\tau = 10^{-10}$ 为一个较小的阈值。将方向调整效应定义为

$$e_{s,r} = h_s \widehat{\Delta}_{s,r}. \tag{10}$$

如果 $e_{s,r} > \tau$，则将结果归类为对齐；如果 $e_{s,r} < -\tau$，则归类为反向；如果 $|e_{s,r}| \le \tau$，则归类为零效应。

3.3 评估指标

令 S 为任务数量，R 为每项任务的有效随机种子数量。我们报告五项评估指标。

方向一致率（same）是二元变量

$$\text{same}_{s,r} = \mathbf{1}\{e_{s,r} > \tau\}. \tag{11}$$

标准化处理效应（std_effect）是连续变量

$$\text{std_effect}_{s,r} = \frac{e_{s,r}}{\text{sd}_s(e)}, \tag{12}$$

其中，$\text{sd}_s(e)$ 表示 $e_{s,r}$ 在任务 s 中的标准差。

方向对齐率（DAR）用于衡量方向一致的有效结果所占比例：

$$\text{DAR} = \frac{1}{S} \sum_{s=1}^{S} \frac{N_{s,\text{same}}}{R}, \tag{13}$$

其中，$N_{s,\text{same}}$ 是满足 $e_{s,r} > \tau$ 的任务 s 的有效结果数量。DAR 的取值范围为 $[0, 1]$。

方向反转率（DRR）用于衡量有效结果中方向反转所占的比例：

$$\text{DRR} = \frac{1}{S} \sum_{s=1}^{S} \frac{N_{s,\text{opposite}}}{R}, \tag{14}$$

其中，$N_{s,\text{opposite}}$ 是满足 $e_{s,r} < -\tau$ 的任务 s 的有效结果数量。DRR 的取值范围为 $[0, 1]$。

净方向得分（NDS）定义为方向对齐率与方向反转率之差：

$$\text{NDS} = \text{DAR} - \text{DRR}. \tag{15}$$

NDS 位于 $[-1, 1]$，用于衡量净方向响应：得分为 1 表示所有有效结果均一致，而 -1 则表示所有有效结果均相反。

3.4 REFLEX 基准测试设计

任务设计。*REFLEX* 由 30 个反事实社会模拟任务组成，这些任务分为六个互斥的社会科学领域：行为与实验经济学；政治经济学与公共治理；社会心理学与群体过程；网络、沟通与集体智慧；市场、声誉与劳动；以及文化、规范与演化。跨学科研究根据其主要研究问题分配到单一类别。

在效应模拟中，同一实验单位在不同条件下的配对运行共享智能体档案和环境实例，唯一的差异是预先指定的社会因素。智能体会持续运行多个轮次，只能访问局部可见信息，并且看不到论文标题、预期效应方向、其他智能体的私人信息或全局状态。智能体可以通过结构化动作影响环境，因此其动作会与总体环境共同演变。评分严格依据实际动作，例如贡献、惩罚、迁移、连边形成、提议和投票，而非对人类行为的一般描述或文本化的轨迹。作为对比，知识判断直接查询模型，要求其根据对研究情境和干预措施的结构化描述预测处理效应方向，无需执行多智能体模拟。

参与者档案。该平台使用结构化的异质档案来初始化模拟参与者。在调查校准模式下，从 SAPA 项目数据中选取 2,185 条完整的成年参与者记录。记录以个体为单位抽样，同时保留大五人格特质、年龄范围和教育程度之间的联合结构。样本的平均年龄为 30.3 岁，平均教育水平相当于大学学历。

4 实验

4.1 实验设置

主要实验。主要实验评估来自领先大语言模型提供商的八种开源和闭源通用商业模型：GLM-5.3、MiniMax-M2.7、DeepSeek-V4.1-Flash、Claude-Sonnet-5、Qwen3.8-27B、Kimi-K2.6、GPT-6-Luna 和 Qwen3.6-35B-A3B。对于每种模型，我们都评估其在效应模拟和知识判断方面的性能。在效应模拟中，我们针对每种模型运行全部 30 项*REFLEX* 任务，并使用随机种子 1 至 5 对智能体群体进行抽样。我们使用不同的抽样随机种子来改变抽取的群体，并将每项实验重复五次，同时将环境随机种子固定为 42。所有结果均报告自助法置信区间。

消融实验。 所有消融实验均使用 GLM-5.3，任务和智能体群体采样种子均与主要实验保持一致。我们评估了三类消融实验。在信息消融中，我们向模型提供中性背景信息、来自经典研究的历史材料，或模型根据自身知识判断得出的预测，并评估性能是否取决于输入知识的质量。在人口规模消融中，我们将每项任务中的智能体数量增加到原始设置的 1.5 倍和 2 倍，以检验模型性能是否取决于人口规模。在措辞

表1：八个模型的知识判断与效应模拟。组合列中的条目为 + / 0 / -（一致 / 零效应 / 反向）。K-E 是知识判断 NDS 与效应模拟 NDS 之间的差值。方括号内为自助法95%置信区间。

模型	知识判断				效应模拟				K-E [95% 置信区间]
	+ / 0 / -	DAR ↑	DRR ↓	NDS ↑	+ / 0 / -	DAR ↑	DRR ↓	NDS ↑	
GLM-5.3	27.6/0.2/2.2	**0.920**	0.073	**0.847**	19.8/4.0/6.2	**0.660**	**0.207**	**0.453**	0.393 [0.113, 0.593]
Claude-Sonnet-5	26.2/1.6/2.2	0.873	**0.073**	0.800	17.6/3.6/8.8	0.587	0.293	0.293	0.507 [0.219, 0.747]
Qwen3.8-27B	25.6/1.4/3.0	0.853	0.100	0.753	17.4/3.8/8.8	0.580	0.293	0.287	0.466 [0.213, 0.720]
Kimi-K2.6	25.4/0.6/4.0	0.847	0.133	0.713	16.4/3.2/10.4	0.547	0.347	0.200	0.513 [0.180, 0.747]
DeepSeek-V4.1-Flash	25.0/1.2/3.8	0.833	0.127	0.707	16.6/4.6/8.8	0.553	0.293	0.260	0.447 [0.173, 0.667]
MiniMax-M2.7	25.0/0.4/4.6	0.833	0.153	0.680	18.2/2.8/9.0	0.607	0.300	0.307	0.373 [0.060, 0.587]
GPT-6-Luna	24.4/0.6/5.0	0.813	0.167	0.647	15.8/6.2/8.0	0.527	0.267	0.260	0.387 [0.040, 0.647]
Qwen3.6-35B-A3B	22.8/0.4/6.8	0.760	0.227	0.533	14.8/4.6/10.6	0.493	0.353	0.140	0.393 [0.067, 0.667]

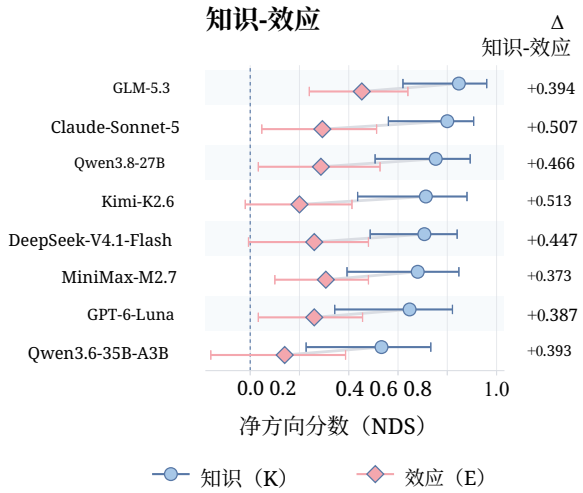

图2：八个模型的知识判断和效应模拟 NDS，按知识 NDS 从上到下排序。NDS 的取值范围为 $[-1, 1]$。右侧的数字为两者之差。误差线表示 95% 自助法置信区间。所有八个模型的知识 NDS 均高于效应 NDS。

消融实验中，我们仅更改共享的角色设定指令，并使用几种近似改述来检验模型的鲁棒性。

4.2 *RQ1*：知识对齐是否不同于效应复现？

表1呈现了主要实验的完整结果。效应模拟中的方向一致率介于0.493至0.660之间，这表明模型做出方向正确选择的概率并非微不足道。然而，错误会对实际政策指导产生重大影响：不正确的效应估计可能导致误导性的政策方向，并有可能造成社会危害。效应模拟中的方向反转率介于0.207至0.353之间。为捕捉正确预测与错误预测的综合影响，我们采用净方向得分（NDS）作为主要指标。结果显示，效应模拟中的NDS介于0.140至0.453之间。这些得分仍然较低，在几种情况下接近零，仅略优于随机猜测，这表明模型得出的错误结论可能几乎与正确结论一样多。较低的效应模拟得分表明，当前方法仍不足以模拟人类行为。

我们还考察了知识判断与效应模拟之间的差距，如图2所示。以 GLM-5.3 为例，它是总体上在知识判断和效应模拟两方面表现最佳的模型。其知识判断 NDS 为 0.847，而效应模拟 NDS 仅为 0.453。两者相差显著，差值为 0.393，置信区间为 $[0.113, 0.593]$，且不包含零值，表明这两种设置之间存在显著差异。我们进一步比较了所有模型在知识判断与效应模拟之间的 NDS 差距。每个模型的差距均为正，且每个自助法置信区间都不包含零值。因此，同一模型在知识判断和效应模拟模式下表现出较大差异，这表明两者存在根本区别。这一结果与我们的理论预期一致：模拟过程中的关键差异会阻碍这种方向性知识传递到模拟结果中。

综上，这些结果证实，知识对齐与效应复现存在显著差异，并且每个模型的知识—效应差距均为正向。

4.3 *RQ2*：效应复现是否与模型、信息、人口规模或措辞相关？

信息消融实验。 我们改变每轮模拟开始时模型可获取的信息，设置了三种条件：提供有关实验的中性背景信息、提供与经典研究相关的历史材料，以及将模型自身的知识判断答案直接反馈到其语境中。如表2所示，提供中性背景信息时，NDS为0.433；提供历史材料时，NDS为0.413。两者的得分都没有超过基线。将模型自身的知识判断答案反馈到语境中后，NDS升至0.520，比基线略高0.067。然而，置信区间 $[-0.027, 0.233]$仍包含零值。所有

三种实验设置的置信区间都包含零值，表明这些差异并不显著。这说明，即使仔细隔离相关内容，并明确向模型提供有关预期效应的背景知识，其模拟结果也没有显著改善。

结合表1中的结果，这些发现表明，大语言模型并非不了解何为合理的人类行为，但模拟未能复现它们自己明确预测的结果。这似乎有悖直觉，因为所有尝试过的操纵都没有带来显著改善。然而，我们的理论分析给出了一种解释：大语言模型在社会模拟中表现不佳，核心原因并非输入知识不足，也不是无法理解社会情境。相反，模型未能捕捉社会知识中隐含的真实人类行动模式，而这种错误还会在与社会环境的交互过程中进一步放大。

表2：信息消融实验。 Δ 列报告逐点95%偏差校正加速法自助法置信区间（$B = 2{,}000$，随机种子555）。基线NDS为GLM-5.3主实验效应模拟的NDS，数值为0.453； Δ 是相对于该基线测得的。

配置	+ / 0 / -	NDS ↑	Δ NDS [95% CI]
原始	19.8/4.0/6.2	0.453	—
中性	19.6/3.8/6.6	0.433	−0.020 [−0.180, 0.147]
历史	19.0/4.4/6.6	0.413	−0.040 [−0.180, 0.113]
自身判断	20.8/4.0/5.2	0.520	+0.067 [−0.027, 0.233]

人口规模消融。 我们将参与每项任务的智能体数量增加到原始人口规模的1.5倍和2倍，以评估模型性能在更大群体中是否保持一致。如表3所示，人口规模为原始规模的1.5倍和2倍时，NDS分别为0.427和0.520。与基线相比，这两个分数分别下降了0.027和上升了0.067。与基线之差的置信区间分别为 $[-0.173, 0.133]$ 和 $[-0.120, 0.267]$，且均包含零值，表明差异不显著。这表明模型性能并不显著取决于实验人口规模。

表3：人口规模消融。 Δ 列报告逐点95% BCa自助法置信区间（$B = 2,\ 000$，随机种子555）。基线NDS为GLM-5.3主实验效应模拟的NDS，即0.453； Δ 是相对于此基线测得的。

配置	+ / 0 / -	NDS ↑	Δ NDS [95% CI]
原始	19.8/4.0/6.2	0.453	—
人口 ×1.5	19.6/3.6/6.8	0.427	−0.027 [−0.173, 0.133]
人口 ×2	20.6/4.4/5.0	0.520	+0.067 [−0.120, 0.267]

措辞消融。 措辞消融实验仅更改共享的角色设定指令，并通过若干近似改述来测试模型的鲁棒性。我们评估了五种改述；表4报告了最具代表性的两项结果。表现最差的改述使 NDS 降至 0.220，其置信区间为 $[-0.440, -0.047]$，且不包含零值，表明性能显著下降。表现最佳的改述使 NDS 升至 0.593，其置信区间为 $[0.013, 0.287]$，且不包含零值，表明性能显著提升。NDS 的总体变化范围为 0.373。这些结果表明，模型对细微的措辞变化高度敏感，性能波动显著。这说明模型不仅性能有限，鲁棒性也不足，可能会大幅限制其在现实场景中的应用。

表4：措辞消融实验。 Δ 列报告逐点95% BCa自助法置信区间（$B = 2{,}000$，随机种子 555）。基线 NDS 为 GLM-5.3 主实验效应模拟的 NDS，数值为 0.453； Δ 以此基线为参照进行测量。

配置	+ / 0 / -	NDS ↑	Δ NDS [95% 置信区间]
原始	19.8/4.0/6.2	0.453	—
改写 最差	16.2/4.2/9.6	0.220	−0.233 [−0.440, −0.047]
重写最佳	22.2/3.4/4.4	0.593	+0.140 [0.013, 0.287]

总的来说，效应复现与模型选择和提示词措辞显著相关，但与人口规模或背景信息无关。下文将考察哪一因素主导了剩余的变异。

5 实证分析

5.1 研究问题*3*：哪些核心因素影响模型效应差异?

按类别划分的评估结果。 如表5和图3所示，模型在六个任务类别上的表现存在差异。在市场、声誉和劳动类别中表现最高（NDS = 0.533），在网络、沟通和集体智慧类别中表现最低（NDS = 0.011）。从知识效应差距来看，网络、沟通和集体智慧类别的差距最大（0.586），而社会心理学和群体过程类别的差距最小（0.013）。这些结果表明，各类别背后的机制有所不同。有些类别相对容易由模型模拟，而网络、沟通和集体智慧似乎最难模拟。

方差分析。 我们分析八个模型生成的全部1,200次任务—随机种子模拟运行，以确定模型间方差的主要来源。我们对影响模拟结果的因素进行方差分析，并将same和std_效应用作因变量。详见表6。

结果显示，模型和任务对两项结果的效应均显著（均为 $p < 0.05$），表明它们是模拟效应方差的主要来源。模型性能因任务而异，不同模型之间也存在系统性差异。相比之下，随机种子和模型 ×随机种子的交互效应均不显著（均为 $p > 0.05$），这表明不同智能体群体的模拟结果在总体水平上不存在系统性差异。这

表5：按社会科学类别划分的方向性性能。类别名称采用缩写。 n 表示该类别中的任务数量。知识判断和效应 NDS 是类别层面的比例；K-E 是知识效应差距。

类别	n	知识 NDS ↑	效应 NDS ↑	K-E
行为经济学	7	0.839	0.300	0.539
政治经济学	4	0.750	0.244	0.506
社会心理学	6	0.521	0.508	0.013
网络/传播	7	0.596	0.011	0.586
市场/声誉	3	0.992	0.533	0.458
文化/规范	3	0.717	0.150	0.567

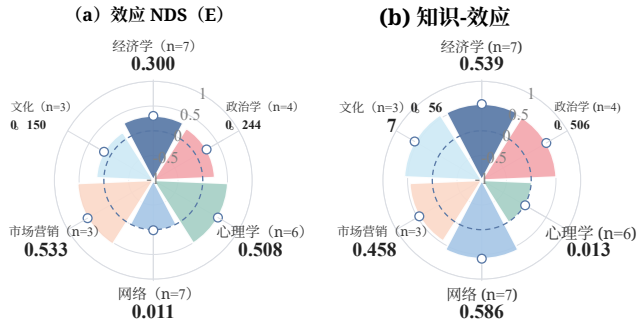

图3：六个社会科学类别中的效应 NDS 和知识效应差距（ K-E）。两个面板均将指数从 $[-1, 1]$ 映射到半径 $[0, 1]$；标签显示原始值，虚线环表示零值，中心为 -1，外环为 $+1$。

表6：模型、随机种子和任务效应对模拟结果的影响。

结果	效应	F	p	η_p^2
same	模型	2.400	**	0.015
	seed	1.023	ns	0.004
	模型 × 随机种子	0.854	ns	0.021
	task	21.248	***	0.353
标准差_效应	模型	3.414	***	0.021
	seed	0.948	ns	0.003
	模型 × 随机种子	0.838	ns	0.020
	task	30.682	***	0.440

一发现表明，这些模型可能无法充分捕捉人群内部异质性对结果的影响。

总之，模型效应差异背后的核心因素是任务：任务解释的方差远高于模型（η_p^2 = 0.353 和 0.440，而模型为 0.015 和 0.021），且未检测到显著的随机种子效应。

5.2 研究问题4：模型结果是否印证偏差级联理论?

我们接下来考察任务结构中的因果链长度与模拟结果一致性之间的关系。我们将 因果链 定义为模型所面临的推理难度指标。根据从处理到主要结果所需的过程阶段数量，将任务编码为四个等级：

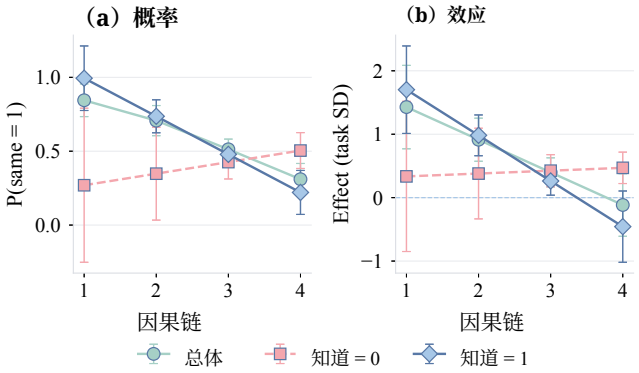

图4：因果链层级的调整后预测值。左侧面板显示预测的同向概率；右侧面板显示预测的标准化处理效应（标准化_效应）。圆圈和实线表示总体样本，方块和虚线表示知道=0（错误或不确定），菱形和实线表示知道=1（正确）。误差线表示95%置信区间。

从处理到主要结果所需的过程阶段数量：

(1) **层级1：直接解决**。处理直接改变当前观察或收益，结果通过单个动作得以解决。(2) **层级2：单步响应或环境解决**。处理必须经过另一智能体的一次响应或一次环境解决，才能影响主要结果。(3) **层级3：群体状态的跨轮次聚合**。结果取决于多轮互动中群体状态的累积。(4) **层级4：策略预期、协议聚合或累积性文化传播**。结果需要策略预期、内容聚合或跨代累积。

编码仅使用任务规范，即操纵和主要结果，不依赖模拟结果。

对于任务 s、智能体随机种子 r 和模型 k，我们将 same 和 std_effect 用作因变量。线性规范为

$$y_{s,r,k} = \alpha + \beta\,\mathrm{CausalChain}_s + \mathbf{X}'_{s,r,k}\gamma + \mu_k + \delta_r + \varepsilon_{s,r,k}, \qquad (16)$$

其中，$\mathbf{X}_{s,\ r,\ k}$ 表示控制变量，μ_k 和 δ_r 分别表示模型和随机种子的固定效应。标准误按任务 s聚类。当 y = std_效应时，我们采用线性固定效应估计量。当 y = same时，我们采用具有类似设定的逻辑回归模型。总体而言，因果链长度与二元同向率（-0.205, $p < 0.01$ -0.514 $p < 0.01$ ）和标准化任务内效应（，）均呈显著负相关。这表明，随着因果链长度增加，模拟效应与参考方向保持一致会变得更加困难。表7报告了完整的估计结果，图4展示了调整后的预测值。

表7：因果链长度与模拟效应：总体情况及按知识判断正确性划分的结果。括号内为按任务聚类的稳健标准误。所有规格均控制模型固定效应和种子有限元。

变量	*same*	标准化_效应	知道=0 相同	知道=1 相同	了解=0 标准_效应	了解=1 标准_效应
因果链	-0.205^{***} (0.049)	-0.514^{***} (0.182)	0.078 (0.100)	-0.258^{***} (0.056)	0.045 (0.241)	-0.720^{***} (0.204)
常数项	−4.610 (11.477)	−1.514 (4.795)	−2.341 (1.286)	0.091 (1.053)	-14.473^{***} (2.389)	0.557 (3.035)
模型固定效应	Yes	Yes	Yes	Yes	Yes	Yes
种子有限元	Yes	Yes	Yes	Yes	Yes	Yes
控制项	Yes	Yes	Yes	Yes	Yes	Yes
N	1,200	1,200	190	1,010	190	1,010
伪 R^2 / R^2	0.172	0.250	0.458	0.222	0.587	0.265

我们还根据知识判断是否准确，分别进行了回归分析。在此，knows = 1 表示模型在知识判断中正确识别了参考方向，而knows = 0 表示判断错误或不确定。分层后，负相关关系仅出现在知识判断正确的运行中。当knows = 1时，*same* 的系数为 −0.258 （$p < 0.01$），*std_*效应的系数为−0.720 （$p < 0.01$）。当knows = 0时，相应系数为0.078和0.045，均未达到统计显著水平。总体伪 R^2 / R^2 值也高于全样本设定的值，表明这种分层能够分离出更纯粹的效应。这些结果表明，知识判断准确的模型更容易受到较长因果链的影响，进一步佐证了我们的主张：偏差级联会降低模拟性能。

综合来看，这些结果证实了偏差级联理论：以 因果链 为关键变量，因果链越长，效应复现效果就越差。

6 结论

一个比喻简洁地概括了我们的发现：

> 科尔曼之船无法穿越偏差级联。

基于 Coleman 的工作，我们确定了微观层面对齐与宏观层面效应保真度之间的两个差距。首先，对齐的状态可能导致不同的行动规则。其次，匹配的边际行动可能产生不同的联合互动和宏观层面效应。我们提出的 *REFLEX* 通过比较模型估计的平均处理效应与实证发现，并报告统一的单模型净方向得分，在 30 个最小干预的反事实任务上评估这些差距。即使是表现最佳的模型 GLM-5.3，在知识判断和效应模拟上的得分也分别为 0.847 和 0.453，差距高达 0.393。措辞消融实验进一步揭示了模型对提示措辞的高度敏感性，而回归分析证实，因果链越长，性能退化越严重。

参考文献

[1] Gati V Aher、Rosa I Arriaga 和 Adam Tauman Kalai。2023。使用大语言模型模拟多个人类并复现人类受试者研究。载于 国际机器学习会议。PMLR, 337–371。[2] Maida Aizaz 和 Quang Minh Nguyen。2026。分析两极化地缘政治背景下的大语言模型角色生成与公平性解读。载于 计算语言学协会欧洲分会第十九届会议论文集（第 *4* 卷：学生研究研讨会）。664–684。

[3] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. 从一到众：使用语言模型模拟人类样本。《政治分析》31, 3 (2023), 337–351. [4] Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. 2026. 大语言模型能够预测社会科学实验的结果。《自然》(2026), 1–8. [5] Joonhyung Bae. 2026. 边际对齐并不保证联合概率分布的逼真度：对 Nemotron-Personas-Korea 的官方参考审计与跨地区复现。*arXiv* 预印本 *arXiv:2606.12433* (2026).[6] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. 2025. 用于预测和捕捉人类认知的基础模型。《自然》644, 8078 (2025), 1002–1009.[7] James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. 合成替代品能否取代人类调查数据？大语言模型的风险。《政治分析》32, 4 (2024), 401–416. [8] Alejandro Buitrago López, Javier Pastor-Galindo, and José A Ruipérez-Valiente. 2026. 大语言模型智能体应在社会模拟中做出决策吗？比较有限状态决策策略与基于大语言模型的决策策略。*arXiv* 电子预印本(2026), arXiv–2606. [9] Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. 大语言模型理解社会知识吗？使用 socket 基准测试评估大语言模型的社交能力。载于*2023*年自然语言处理实证方法会议论文集。11370–11403. [10] James Samuel Coleman et al. 1990. 社会理论基础。第302卷。HeinOnline.[11] Jamie Cummins. 2026. 使用大语言模型模拟人类数据时分析灵活性的威胁。《心理科学方法与实践进展》9, 3 (2026), 25152459261461505. [12] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. 人工智能语言模型能否取代人类参与者？《认知科学趋势》27, 7 (2023), 597–600.[13] Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, et al. 2026. TwinVoice：通过大语言模型角色模拟构建数字孪生的多维基准测试。载于计算语言学协会：*ACL 2026* 论文集。19604–19628. [14] Agam Goyal, Yian Wang, Eshwar Chandrasekharan, and Hari Sundaram. 2026. 从合理到因果：模拟在线社群中政策评估的反事实语义。*arXiv* 预印本 *arXiv:2604.03920* (2026). [15] Mark Granovetter. 1978. 集体行为的阈值模型。《美国社会学杂志》83, 6 (1978), 1420–1443. [16] Peter Hedström and Richard Swedberg. 1998. 社会机制：社会理论的分析方法。剑桥大学出版社。[17] Paul W Holland. 1986. 统计学与因果推断。《美国统计学会杂志》81, 396 (1986), 945–960.[18] Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2026. Simbench：大语言模型模拟人类行为能力的基准测试。载于国际学习表征会议，第2026卷。28290–28341. [19] Chaemin Jang, Dongman Lee, and Jihee Kim. 2026. 指令微调语言模型无法从其能够描述的概率分布中采样。*arXiv* 预印本 *arXiv:2607.25292* (2026).[20] Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. 2025. 为社会科学实验中的人类行为预测微调大语言模型。载于*2025*年自然语言处理实证方法会议论文集。30084–30099. [21] Juhoon Lee and Joseph Seering. 2026. SLALOM：通过纵向观测指标进行社会模拟的模拟生命周期分析。*arXiv* 预印本 *arXiv:2604.11466* (2026).[22] Dai Li, Linzhuo Li, and Huilian Sophie Qiu. 2025. ChatGPT 不是一个人，而是常人：大语言模型生成的硅基样本的代表性与结构一致性

由大语言模型生成。arXiv 预印本 arXiv:2507.02919 (2025).[23] Xiang Li, Liping Yi, Mingze Kong, Min Zhang, Zhongxiang Dai, and QingHua Hu. 2026. ALSO：面向社会智能体的对抗性在线策略优化。arXiv 预印本 arXiv:2605.15768 (2026). [24] Yiming Li and Dacheng Tao. 2026. 仅靠人工智能智能体尚不足以实现社会模拟。arXiv 预印本 arXiv:2603.00113 (2026). [25] Zonghan Li and Feng Ji. 2026. 统计真实性并不能证明大语言模型能够在社会科学实验中估计处理效应。arXiv 预印本 arXiv:2604.02458 (2026). [26] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. 2024. 在模拟社会互动上训练社会对齐语言模型。载于国际学习表征会议，第2024卷。49602–49625.[27] Xuan Liu, Haoyang Shang, Zizhang Liu, Xinyan Liu, Yunze Xiao, Yiwen Tu, and Haojian Jin. 2026. Humanstudy-bench：面向参与者模拟的人工智能智能体设计。arXiv 预印本 arXiv:2602.00685 (2026). [28] Dimitris Michailidis, Petter Törnberg, and Fernando P Santos. 2026. 从 caricature 到集体产物：角色构建如何驱动生成式社会模拟中的极化。arXiv 预印本 arXiv:2610.16498 (2026).[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. 通过人类反馈训练语言模型遵循指令。神经信息处理系统进展35 (2022), 27730–27744. [30] Atharva Pandey and Gautam Jajoo. 2026. 用于审计大语言模型社会模拟器的推理中介行为模型。arXiv 预印本 arXiv:2607.24649 (2026).[31] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. 生成式智能体：人类行为的交互式拟像。载于第36届 ACM 用户界面软件与技术研讨会论文集。1–22.[32] Joon Sung Park, Carolyn Q Zou, J onne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, et al. 2024. 基于自我报告的大语言模型智能体能够实现个体的通用模拟。arXiv 预印本 arXiv:2411.10109 (2024).[33] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. 2025. AgentSociety：大规模模拟大语言模型驱动的生成式智能体，增进对人类行为和社会的理解。(2025). [34] Xiaoyou Qin, Zhihong Li, and Xiaoxiao Cheng. 2026. 恢复基于大语言模型的社会模拟中的异质性：一种受众细分方法。arXiv 预印本 arXiv:2604.06663 (2026).[35] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. 语言模型反映了谁的观点？载于国际机器学习会议。PMLR, 29971–30004. [36] Thomas C Schelling. 2006. 微观动机与宏观行为。WW Norton & Company. [37] Raad Bin Tareaf. 2026. 以人类行为分布为基准测试大语言模型智能体社会。arXiv 预印本 arXiv:2608.28182 (2026).[38] Olivier Toubia, George Z Gui, Tianyi Peng, Daniel J Merlau, Ang Li, and Haozhe Chen. 2025. 数据库报告：Twin-2k-500：一个基于2000多人的500多个问题答案构建数字孪生的数据集。《营销科学》44, 6 (2025), 1446–1455.[39] Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A Duéñez-Guzmán, William A Cunningham, Simon Osindero, Danny Karmon, and Joel Z Leibo. 2023. 使用 Concordia，通过以物理、社会或数字空间为基础的动作进行生成式基于智能体建模。arXiv 预印本 arXiv:2312.03664 (2023). [40] Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. 2025. Sociobench：使用大语言模型对社会学调查中的人类行为建模。载于2025年自然语言处理实证方法会议论文集。26268–26300. [41] Jinyi Ye, Lei Cao, Ding Chen, and Emilio Ferrara. 2026. 未经稳健性审计，切勿从大语言模型社会模拟中得出科学主张。arXiv 预印本 arXiv:2605.18890 (2026).[42] Patrick Zhao, David Huu Pham, and Nicholas Vincent. 2026. 生成式基于智能体建模中的机制合理性。载于2026年 ACM 公平性、问责制与透明度会议。5915–5940.[43] Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tong- shuang Wu, Yiming Yang, and Maarten Sap. 2026. OdysSim：构建用于人类行为模拟的基础模型。arXiv 预印本 arXiv:2606.14199 (2026).[44] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2024. Sotopia：语言智能体社会智能的交互式评估。载于国际学习表征会议，第2024卷。40975–41019.[45] Caleb Ziems, William Held, Su Doga Karaca, David Grusky, Tatsunori Hashimoto, and Diyi Yang. 2026. 扩展能否改善大语言模型的社会模拟？arXiv 预印本 arXiv:2607.02464 (2026)。

A 伦理声明

本研究未招募人类参与者，也未涉及真实付款、注册、与现实中的个人接触或紧急服务呼叫。帮助、捐赠和参与组织均为模拟动作。

B 局限性

首先，每项任务仅以某一项经典研究所报告的效应方向作为参照。为便于计算，平台使用基于规则的引擎处理结构化数值结果。尽管这种方法提高了一致性和可复现性，但它必然会忽略现实社会制度和交互作用的复杂性，也可能无法捕捉潜在社会机制在更广泛语境中的稳定性或可推广性。其次，人口档案取自一个公开数据集，主要由年龄和教育程度处于特定范围内的成年人构成。这样的设计对于确保实验可比性是必要的。然而，这些档案可能与原始研究中的参与者有所不同，因此未必能充分代表那些研究群体的异质性。第三，由于原始研究仅提供了有限的参与者内在想法数据，我们无法通过分离科尔曼之船各个步骤的独立贡献，并将由此得出的预测与现实世界的观察结果进行比较，来直接检验该理论。相反，我们通过因果链长度间接评估这一理论问题，并通过案例分析更详细地考察其潜在机制。第四，在诊断偏差级联时，我们手动将因果链长度编码为 1 到 4 的等级，以量化其影响。这一编码过程可能会引入一定程度的主观性。

C 术语

本文涉及四个彼此相关、处于不同层面的术语。知识判断 和 效应模拟 是测量层面的两种实验设置：前者要求模型仅依据任务背景，在不依赖仿真环境的情况下，预测某一给定社会因素所引发效应的方向；后者则将大语言模型智能体置于构建的社会环境中，让它们根据局部可用的观察结果采取行动，并将其个体行动汇总为宏观层面结果，据此估算处理效应。知识对齐和效应复现则是在构念层面上分别对应这两种设置的目标属性：知识对齐指模型对社会情境及其后果的认知与人类知识之间的一致性；效应复现则指处理效应的方向与人类研究结论之间的一致性。

D 核心动机

D.1 因果建模的重要性

近年来，基于大语言模型的社会模拟及其在社会科学中的评估取得了长足进展。我们的研究与这类工作不同之处在于，将因果干预置于模拟设计的中心，并将由此产生的处理效应作为效应复现的首要标准。

先前的工作引入了用于评估模拟能力的端到端基准测试 [18, 27], , 但这些基准测试的主要目标并非评估效应复现。相反，它们通常通过直接比较模拟结果与观察到的响应或结果，来刻画模型能力。这种以能力为导向的评估范式有两大局限。现有框架往往评估关联一致性，却未纳入因果设计。尽管已有许多评估指标被提出，但这些指标可能无法捕捉社会模拟的核心机制。针对这些指标进行优化，可能会奖励虚假相关，同时掩盖模拟失败的根源。相较之下，社会模拟的因果方法要求明确设定一个社会环境，以便通过反事实社会实验考察处理效应。

这些考量促使我们构建一个反事实社会实验基准测试，聚焦社会模拟中的跨层级因果过程。从因果视角来看，平均处理效应，尤其是其方向，为评估模拟是否复现了总体层面的干预响应提供了有原则的依据。这正是*REFLEX*的核心动机。

D.2 科尔曼之船作为解释框架

现有关于社会模拟的研究也很少运用经典社会科学理论。借鉴科尔曼之船，我们的研究指出知识对齐与效应复现之间存在差距，并对这一差距提出了理论解释。

科尔曼之船是一种社会学框架，用于解释宏观层面的社会条件如何通过微观层面的行动与互动产生集体结果。其核心前提是，宏观层面变量之间的经验关联不足以构成社会学解释。解释还必须指出社会条件如何塑造个体行动，以及这些行动又如何产生宏观层面结果。因此，该框架通过情境机制、行动形成机制和转化机制，将宏观层面的社会条件、微观层面状态、个体行动与集体结果联系起来。在我们的研究中，它为考察结构性约束与个体能动性提供了统一的分析框架。这一视角要求对行动者进行建模，使其能够在动态环境中调整行为，而不是将评估局限于现有文献中常见的静态、一次性选择。

一些研究探讨了知与行之间的区别 [19]。然而，大多数研究通过单轮问卷式评估来操作化社会模拟。它们的主要关注点仍是与人类响应的一致性，而非动态情境所需的行动能力。这类评估主要衡量与个体特征相关的响应，却基本未考察互动过程。对于以个体响应为中心的心理学研究问题，这类评估仍然适用。然而，许多社会现象无法仅通过汇总与个体响应相一致的结果来理解。它们还取决于知识判断如何转化为动作，以及这些动作如何随时间相互作用。

这一视角为 *REFLEX* 的设计提供了依据。REFLEX 在明确的模拟环境中，通过生成的结果评估模型能力。科尔曼之船进一步指导我们分析可能导致观察到的知识对齐与效应复现之间差距的机制。

E 方法与假设

E.1 模型假设

在科尔曼之船框架中，关于社会语境的宏观层面知识 m 通过情境机制映射为微观层面状态的分布 $Q_q(z \mid m)$，其中 z 表示个体处境与状态。我们的理论分析提出了一个关于大语言模型理解社会语境的理想化假设：当模型对某一宏观层面社会条件的知识与人类知识相一致时，仿真系统便能在理想化条件下，将其对该条件的表征映射到相应的人类微观层面状态分布。其依据在于，相关社会知识涵盖社会条件与个体处境及状态之间的关系，从而为这种宏观到微观映射提供了基础。我们的实验为所评估的知识判断中的知识对齐提供了证据，但并未直接证实微观层面状态上的对齐。

我们基于两个原因将微观层面状态对齐视为一项假设。首先，我们无法完整重构原始实验中的社会情境和个体情境，因此缺乏可验证的微观层面真实值。其次，即使这些情境能够被重构，关于人工智能系统与人类具有完全相同主观感知的主张，仍然难以操作化和检验。因此，我们将实证评估聚焦于可观测的宏观层面结果，并将处理效应方向一致作为可证伪的标准。具体而言，在相应的反事实实验操纵下，我们检验模型估计的宏观层面处理效应方向是否与既有实证社会科学研究所报告的方向一致。

我们还考察了另一种解释，即观察到的失败源于环境设计的局限性。为此，我们验证每个环境是否都存在一种能够复现实证效应方向的配置。这一检查评估的是环境是否在设计上排除了效应复现；它并不能证明环境在所有方面都具备充分性。对于无法通过实验解决的方面，我们依靠理论分析。

E.2 探究偏差级联

科尔曼之船中的偏差级联解释涉及两个核心传递阶段：将微观层面状态 Q 映射为行动选择规则 ρ，以及通过交互轨迹 T 将边际行动分布聚合为宏观层面结果 Y。我们的实验无法直接识别这些阶段。首先，科尔曼之船是一种解释性框架，而非统计测量模型。其次，无法从现有数据中识别中间过程。要分别识别这两个阶段，需要在反事实条件下获得完整的个体层面轨迹，以及同一互动网络中这些个体的联合响应分布。然而，对于本文考察的实证研究而言，这类数据并不可得。因此，直接识别偏差级联背后的中间过程超出了当前设计的范围。

因此，我们将偏差级联用作解释模拟失败的解释框架，而不是声称能够直接识别其因果贡献。尽管如此，我们的分析并不止于理论解释。为探究准确的知识判断与未能实现的效应复现之间的差距，我们引入变量 *CausalChain*，将其作为对偏差级联解释的间接实证检验。该变量刻画了将个体行动与宏观层面结果联系起来的因果结构和交互结构，从而使我们能够考察框架所预测的可观测模式。

我们发现，随着因果链长度增加、交互作用变得更加复杂，效应复现率会下降。这些宏观层面的模式与框架的预测一致，并支持这样一种解释：模拟失败反映的是模拟过程中偏差的传播和放大，而不仅仅是静态知识的不足。

E.3 人口与抽样

对于所有 *REFLEX* 任务，智能体总体均从 SAPA 中抽取。SAPA 是一个共享的、内部具有异质性的一般人口样本池。改变智能体随机种子会从这个固定样本池中重新抽取智能体，而不是重建每项原始人类研究的人口统计抽样框。我们的样本池与原始实验所研究的群体之间存在差异，这限制了研究结果的有效性和可推广性。与此同时，使用共同的人口样本池，为跨任务比较提供了一致的基础，并减少了由任务特定抽样选择造成的差异。因此，我们的设计优先考虑任务间的可比性，而非与每项原始研究进行总体匹配。

在当前设计下，我们未发现智能体随机种子具有统计显著效应。这一结果表明，在 SAPA 样本池内重新抽样，不会使估计的处理效应产生可检测到的变化。但这并不能证明总体异质性不存在实质影响：若使用与原始实验相匹配的抽样框，结果可能会有所不同；而观察到的稳定性可能反映了样本池中的变异有限、模型对总体特征的敏感性有限，或两者兼有。

具体而言，模型生成的行动异质性降低本身可能是本研究中观察到效应复现失败的原因之一。这种解释与所提出的理论框架相容，我们的分析也无法排除这种可能性。由于总体异质性并非本研究问题的主要关注点，我们在跨任务评估中保留共同池设计，并将总体匹配抽样和对异质性的明确分析留待未来研究。

F 自助法置信区间计算

对于主要分析，我们对 S 任务进行有放回重抽样，并保留每个抽中任务的所有有效随机种子；不对随机种子进行重抽样。设 $s_j^{*(b)}$ 表示自助法重复抽样b中位置 j 抽取的任务，其任务得分为 $x_j^{*(b)} = x_{s_j^{*(b)}}$。自助法估计值为

$$\widehat{\theta}^{*(b)} = \frac{1}{S}\sum_{j=1}^{S} x_j^{*(b)}. \tag{17}$$

我们采用 $B = 2{,}000$ 次重复抽样、随机种子 555，以及偏差校正加速法（BCa）区间。

G 基于规则的基线

G.1 评估目标与设置

有人可能会质疑本文的论点：模型性能并非直接由大语言模型社会模拟能力的不足所决定；相反，双重差分是由内部模拟噪声驱动的，也就是说，测得的模型性能不佳，是因为系统本身不存在合理的可行解。我们通过基于规则的基线排除了这种可能性。

对于模拟实验中观察到的性能下降，另一种解释并非模型缺乏所需能力，而是任务设计使得高性能从根本上无法实现。本节将评估并排除这种可能性。

我们考察在当前任务环境中，*REFLEX* 所规定的方向性效应是否可以实现。如果一个仅使用局部观测的显式基于规则的模型，能够在相同的环境、动作约束和结算规则下，产生与参考结果相匹配的人口层面效应，那么这些任务在方向上就是可满足的。为此，我们按照大语言模型实验设置，实现了一个纯基于规则的智能体模型。所有配置均使用默认启发式规则生成一次，不进行额外验证或超参数调优。

我们提示 GLM-5.3 直接生成一条符合真实场景变化的规则。它全知环境的执行方式，但该规则由模型一次性生成，不通过反复调优来提高得分。值得注意的是，该模型可能无法生成最佳规则；因此，在此基础上获得的数值应代表场景中各问题可实现最高得分的理论下界。

G.2 指标结果

评估采用 *REFLEX* 的 30 个环境、行动空间、主要指标和规定方向，并将环境随机种子固定为 42。每项任务针对从 1 到 5 的每个主体种子各评估一次。规则仅接收主体可获取的观察结果，不访问隐藏的全局信息。动作仍按照原始 *REFLEX* 规则进行验证和结算。

最终 NDS 为 0.853，平均每个随机种子有 27.4 项任务呈现出规定方向，见表8。这一结果显著高于当前模拟实验中获得的最佳 NDS（0.453）。

此外，对每项任务的五个主体种子取平均后，估计的效应方向在全部 30 项任务中均与人类参考值一致。因此，针对当前任务集构建的显式基于规则的模拟器能够在默认环境随机种子下复现全部 30 项任务的规定方向，但并非对每个个体主体种子都能做到这一点。

这些结果表明，*REFLEX* 中的参考方向是可以实现的，而非本质上无法满足。因此，大语言模型的 NDS 较低

更可能是由动作形成和交互顺序安排方面的局限性所致，而不是基准测试设计排除了正确方向。

这些实验排除了这样一种假设：测得的模型性能不佳，是因为该系统本身不存在合理的可行解；同时，它们支持了本文提出的假设。

表8：基于规则的行动基线结果。计数为每个主体种子的任务平均数量；DAR、DRR 和 NDS 使用 150 个任务-种子单元。

配置	+ / 0 / -	DAR ↑	DRR ↓	NDS ↑
基于规则	27.4/0.8/1.8	0.913	0.060	0.853

H 调查校准档案

调查校准模式使用公开 SAPA 样本。该样本包含 25 个大五人格自我报告条目，A、C、E、N 和 O 每项各五个，采用 1–6 级李克特量表；此外还包括自报年龄以及按 1 至 5 编码的教育程度。原始快照包含 2,800 条记录。保留所有人格条目均有有效回答、年龄为 18–100 岁且教育程度为 1–5 的记录后，得到 2,185 份档案；表 9 报告了这些档案的描述性统计。

表 9：调查校准档案的描述性统计（$N = 2,\ 185$）。

变量	均值（标准差）
年龄（岁）	30.3 (10.5)
教育程度，1-5	3.22 (1.09)
宜人性，1-6	4.69 (0.89)
尽责性，1-6	4.32 (0.95)
外向性，1-6	4.15 (1.07)
情绪稳定性，1-6	3.86 (1.20)
开放性，1-6	4.62 (0.80)

I SimulationPlatform 架构

I.1 分层职责

我们的框架采用分层架构，如图5所示。该平台由任务规范、环境、群体与智能体、日志记录以及结果生成等层构成。这些层的职责严格分离：环境生成智能体可见的信息，并裁定动作的结果；Python 代码执行规则并更新状态；日志记录层则存储完整的事件链。

该平台遵循统一的执行流程：
(1) **生成观察结果**。 环境基于当前全局状态，为每个智能体生成局部可见的观察结果。(2) **请求决策**。系统向模型提供任务约束、智能体自身的信息以及当前轮次的观察结果，并请求一个结构化动作。

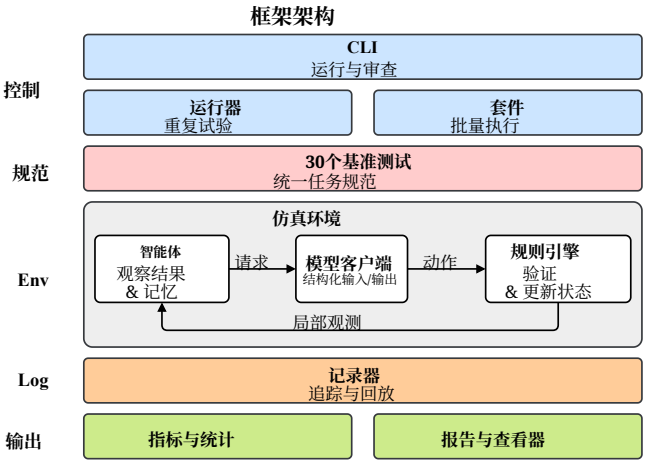

图5：我们的框架的分层架构，将任务规范、试验执行、基于规则的模拟环境、轨迹日志记录和指标报告分开。

(3) **验证响应**。 首先检查返回结果的格式有效性和取值范围。如果响应无效，系统会使用相同的决策 ID 发起重试。整个重试过程都会被记录。

(4) **执行规则**。根据已接受的动作，Python 规则引擎统一计算各项结果，包括收益、资源变化、网络更新、消息传递和市场结算。(5) **记录状态转移**。系统将智能体的动作和结算后的新状态写入事件流，并更新运行记录。(6) **进入下一轮**。下一轮的观察结果始终根据上一轮完全结算后的状态生成。

因此，模型只能提出动作，无法直接修改环境状态。所有对环境产生的效应都必须由规则引擎验证并结算。

I.2 信息隔离与可重计算性

评估所需的真实值不会向智能体公开。运行时环境会根据每个智能体的角色、位置和网络关系，提供其可观察的信息。智能体无法访问其他智能体的私人信息、隐藏效用、尚未送达的消息或评估答案。智能体只能提交有效动作，无法直接修改资源、关系、匹配结果或市场状态。

所有状态变化均由 Python 代码执行，包括资源转移、网络关系更新、智能体匹配、市场出清以及对违规行为的惩罚。这种设计将模型决策与环境施加的实际后果分离开来，并防止模型通过其输出绕过实验规则。

J 示例的机制分析 J.1 正向：惩罚提高低贡献的成本

在公共物品任务中，预计有成本的惩罚会增加最后一轮的贡献。参与者将固定资源分配给公共物品，其总贡献会经过倍增后重新分配。惩罚会使实施者承担成本，并降低接受者收益。在观察贡献后，智能体决定是否进行惩罚，而由此产生的收益会影响下一轮。

决策规则是：惩罚组减去无惩罚组大于零。正向效应表明，模型将惩罚理解为遏制搭便车的一种方式，会有条件地使用惩罚，并促使接受惩罚者在之后增加贡献。

在随机种子1下，无惩罚组的贡献率依次为 0.4625、0.2875、0.1875、0.0875 和 0；惩罚组则为 0.4875、0.5625、0.5625、0.5625 和 0.5625。模型使用了惩罚，预测了正确的方向，并带来了更高的后续贡献。这表明模型在知识、有条件的动作和互动延续方面实现了正向复现。

初次惩罚后，即使不再实施惩罚，惩罚组仍保持较高的贡献。在多次重复运行中，每次运行都会使用惩罚，且最后一轮的贡献差异始终为正。

反思。 这是一个完全正确的模拟结果，表明模型能够处理某些情境。总体而言，模型在此处的行为清晰且正确，这或许与该任务的因果链长度仅为2有关。

J.2 零效应：一律点赞抹平初始评分差异

初始评分任务旨在检验将评论的初始分数从0改为 +1 是否会增加后续点赞。读者依次选择点赞、点踩或弃权。各条件之间唯一不同的是显示的初始分数，并且每位读者都会看到累积后的分数。

决策规则是初始点赞条件与中性条件之差大于零。该评论讨论的是在考试期间延长图书馆开放时间。每位读者都会看到当前分数，并据此更新分数，供下一位读者查看。该任务还包括初始点踩条件。

模型知道初始点赞应该会增加后续点赞，但所有读者在每种条件下都会点赞。在表10中，中性、初始点赞和初始点踩条件下的后续赞成票率均为1.0000。最终分数分别为12、13和11。因此，初始信号虽然体现在输入中，却不会影响行为。模型展现出方向性知识，却未能复现人类的条件敏感性。

重复运行会产生相同结果。初始投票会改变显示的分数，但之后的每位读者都会选择相同的动作。因此，零效应反映的是相同的动作分布，而非真正的条件敏感性。

反思。 模型几乎在所有情况下都会无缘无故地点赞，而现实中几乎不存在人人都点赞的情况。这是从状态到行动的映射存在偏差。原因可能是训练使模型过度拟合于取悦用户

表10：在三种初始评分条件下，随机种子1的投票结果。

条件	初始分数	投票者 12位读者	后续 赞成票率	最终分数
中性	0	全部投赞成票	1.0000	12
初始赞成票	+1	全部投赞成票	1.0000	13
初始反对票	−1	全部投赞成票	1.0000	11

表11：随机网络和聚类网络中种子1的累积扩散轨迹与接触情况。

指标	随机	聚类
新增采纳者，第1轮	3	4
新增采纳者，第2轮	7	4
新增采纳者，第3轮	7	4
新增采纳者，第4轮	7	4
曾收到通知的成员	14	8
已通知者中的采纳率	7/14 = 0.5000	4/8 = 0.5000
最终轮次采纳（全部 14 人）	7/14 = 0.5000	4/14 = 0.2857

行为：尽管模型知道现实中的人们可能不会点赞，但它仍会因用户交互中养成的习惯而点赞。

J.3 反转：局部采纳未能发展为持续扩散

网络任务用于检验聚类网络是否比随机网络带来更多采用。操纵只改变网络连接。每位参与者有四位联系人，节点6和9最初采用。

在种子1中，随机网络的累计新增采纳者分别为3、7、7和7，而聚类网络中分别为4、4、4和4。随机网络中有14位非初始成员收到通知，聚类网络中则有8位。已通知成员中的采用率在两种网络中均为0.5000，但在全部14位非初始成员中，最终轮次采纳率分别为0.5000和0.2857。

结果指标是全部14位非初始成员中的最终轮次采用率。决策规则是聚类网络减去随机网络的结果大于零。聚类网络提供了更多重叠的社会强化；当采用需要多个来源时，这应有助于推动采用。

聚类网络起初领先，但外围代理拒绝采用，阻止了进一步接触。随机网络触及了更多分散的联系人。在五个随机种子中，效应方向一致一次、呈零效应一次、反转三次，平均值为 −0.14286。这一负面结果仅适用于该模型的行为反应，并不意味着随机网络普遍更优。

在种子1中，聚类网络里的初始采纳者和第一轮采纳者位于同一局部区域。他们的合并联系人仅能覆盖八名非初始成员。四个桥接节点将该区域与网络的其余部分连接起来，但它们没有采纳，因此额外的节点不会收到通知。在随机网络中，

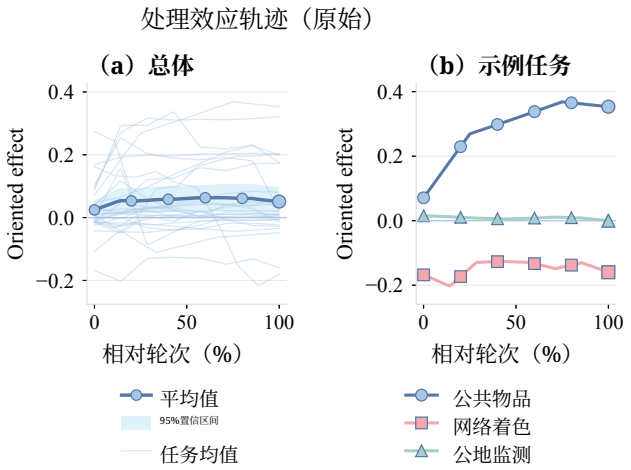

图6：以原始单位表示的代表性处理效应轨迹（左：28项任务；右：三个任务均值）。

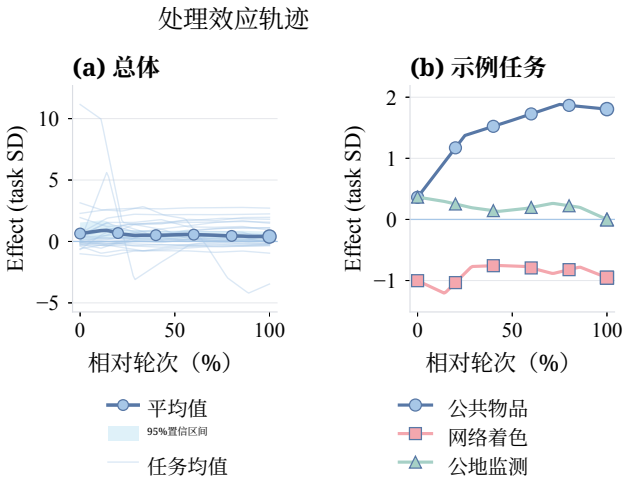

图7：与图6相同的轨迹，单位为任务标准差。

采用者的联系人分布更加分散，因此全部14名非初始成员都能收到通知。

两种网络中，已通知参与者的采用率相同。因此，负面效应主要源于接触机会有限，而非转化率较低。若仅以已通知参与者为分母，则看不出差异；若以全部 14 人为分母，则同时反映接触和转化情况。不同随机种子下，聚类并未带来稳定优势，平均效应反而逆转。

反思。 这是最典型的例子，表明模型结果的成因可能很复杂。这里实际上有两种效应在起作用。首先，随着网络连接日益紧密，采用率实际上并未提高；而现实中的人们可能会因拥有更多联系人而更愿意采用。这是从状态到行动的映射中的偏差。其次，互动过程中会产生更深层的放大效应：例如，处于网络关键位置的个体可能阻碍这一过程，这与许多现实情形恰恰相反；这是互动过程中产生的偏差。

K 三种代表性轨迹

回归系数概括了任务层面的关联，但无法呈现处理效应如何随轮次变化。图6展示了三条具有代表性的轨迹，它们采用共同的相对轮次坐标轴：公共物品惩罚呈现持续的正向效应，网络着色呈现持续的负向效应，而公地监测的效应接近零。每条曲线均对八个模型和五个主体种子取平均值，并为各项任务赋予相同权重。左图展示了全部28项同步轮次任务、它们的等权平均值以及95%自助法置信区间。图7以标准_效应单位报告了相同的轨迹；下文的数值讨论采用图6中的原始差值。

公共物品惩罚比较惩罚条件与无惩罚条件，并以最终贡献率作为主要结果。处理优势从第一轮的约 0.07 增至第4轮的约 0.37，随后在最后一轮保持在 0.35 左右。全部五个主体种子和八个模型都产生了正向的末轮效应。因此，结果并不局限于某一轮。惩罚之后，目标对象的贡献有所提高，而且这种差异会随时间持续。这是本研究中一种稳定且持续的正向轨迹。

网络着色比较的是带有长程连接的重连网络与规则网格。结果指标是连接不同颜色节点的边所占比例。重连网络并未产生预期的协调优势。轨迹始终为负向：第1轮约为 −0.17 ，第2轮约为 −0.20 ，此后约在 −0.15 至 −0.13 之间，最后一轮约为 −0.16 。全部五个随机种子和八个模型均产生负向效应，40次运行中约88%为负值。这是一条稳定的负向轨迹。

公地监测比较了可见的个体提取与仅聚合信息，并以资源存量率作为主要结果。两种条件产生的轨迹几乎相同，处理差异介于 0 和 0.015 之间，最后一轮的差异为零。这并未提供明确证据来表明模拟效应为零。在两种条件下，第一轮约有 68 至 69 个单位被移除，使存量从 80 个单位降至 26 至 28 个单位。第二轮又移除了 25 至 26 个单位，仅剩存量的 4% 至 5%，且超过 80% 的模拟运行结果达到零值。后续轮次仍接近这一下限，两种条件下每轮均移除约三个单位。

L 因果链编码

全部30项任务均依据第4节中的四级评分标准分配了因果链级别。编码以任务为单位：仅使用预先登记的操纵和主要结果，而不使用模拟效应方向或模型输出。最终分布为：第1级4项任务、第2级9项任务、第3级9项任务、第4级8项任务（均值 = 2.70，标准差 = 1.00）。各级别所分配的任务如下所列。

表12：因果链层级及分配给各层级的任务。数量列给出了每个层级的任务数量。

级别	数量	任务
1	4	初始评分与跟随，重复估计和社会信息、多轮信号与行动级联、弱环协调与群体规模。
2	9	公共品与惩罚、轮换帮助与公众声誉、自愿参与公共项目、贡献后的同伴奖励、市场价格反馈反馈与收敛，劳动市场中的稳定身份，在线交易中的履约声誉，累积性缓解与集体风险，群际竞争与协调。
3	9	财富可见性、社会强化与扩散，动态伙伴网络，沟通中的可再生公地，公地中的信息层级提取，互动网络与惯例，坚定少数派与惯例，集中式以及去中心化信念更新、文化市场人气累积。
4	8	隐藏信息谜题、流言传播与临时排斥、成员迁移、网络着色—结合局部信息、协作搜索与共享、间歇性接触与集体问题—题解决、群体重组与累积设计、沟通组织与团队聚合。

M 任务与比较

表13按核心比较、预期方向、原始研究和任务层面结果汇总了各项任务。全部30项任务均有稳定的书目来源支持。其中，27项任务具有正向预期效应，3项任务具有预期负向效应。每项操纵仅改变其预先登记的社会因素，例如信息可见性、制度切换、网络结构或风险水平。

表13还报告了任务层面的指标，揭示出性能存在显著差异。在操纵显著且可在局部采取行动的任务中，正向效应很强，包括公共品与惩罚（NDS 为 0.900）、重复估计与社会信息（数值为 1.000）、间歇性接触与集体问题解决（数值为 1.000），以及弱环协调与群体规模（数值为 1.000）。相比之下，当成功取决于聚合分散的信息或维持协调探索时，性能较差，例如自愿参与公共项目（数值为 −0.800）、基于局部信息的网络着色（数值为 −0.650），以及协作搜索与共享（数值为 −0.400）。操纵较弱或含义模糊的任务，例如初始评分与跟随、文化市场人气累积，其结果集中在接近零的 0.000。总体而言，基准测试的难度高度依赖具体任务。失败主要集中在需要信息聚合的任务中，协调探索或持续扩散，而非简单的局部响应。这一模式与我们的分类分析以及对因果链长度的评估一致。

表13：动态协议下的任务级平均值，汇总了五个主体种子和八个模型的结果。任务列中的引文标明了各项任务的来源。

#	Task	默认结构	核心比较	+ / 0 / -	DAR ↑	DRR ↓	NDS ↑
1	财富可见性 [?]	8 个智能体 × 5轮次	可见 − 不可见 < 0	1.5/0.9/2.6	0.300	0.525	−0.225
2	隐藏信息谜题 [?]	4 个智能体 × 3轮次	共享 − 未共享 > 0	2.5/2.4/0.1	0.500	0.025	0.475
3	公共品与惩罚 [?]	4 个智能体 × 5轮次	惩罚 − 无惩罚 > 0	4.5/0.5/0.0	0.900	0.000	0.900
4	社会强化与扩散-散 [?]	16 个智能体 × 4轮次	聚类 − 随机 > 0	1.2/1.8/2.0	0.250	0.400	−0.150
5	初始评分与跟随 [?]	12 个顺序代理	点赞 − 中性 > 0	0.6/3.8/0.6	0.125	0.125	0.000
6	动态伙伴网络 [?]	8 个智能体 × 8轮次	重新连边 − 固定 > 0	4.2/0.4/0.4	0.850	0.075	0.775
7	轮换帮助和公共声誉度 [?]	8 个智能体 × 8轮次	公开 − 私人 > 0	2.0/0.2/2.8	0.400	0.550	−0.150
8	流言传播与临时排斥 [?]	8 个智能体 × 8轮次	流言传播 − 私人报告 > 0	3.4/0.1/1.5	0.675	0.300	0.375
9	成员迁移 [?]	8 个智能体 × 8轮次	选择 − 分配 > 0	2.2/0.1/2.6	0.450	0.525	−0.075
10	自愿参与公共项目 [?]	8 个智能体 × 8轮次	可选 − 强制 > 0	0.5/0.0/4.5	0.100	0.900	−0.800
11	贡献后的同伴奖励 [?]	8 个智能体 × 8轮次	奖励 − 无惩罚 > 0	3.6/0.1/1.2	0.725	0.250	0.475
12	可再生资源中的沟通公地提取 [?]	8 个智能体 × 8轮次	讨论 − 笔记 > 0	4.0/0.2/0.8	0.800	0.150	0.650
13	公地中的信息层级提取 [?]	8 个智能体 × 8轮次	已识别 − 聚合 > 0	2.1/1.1/1.8	0.425	0.350	0.075
14	基于局部信息的网络着色信息 [?]	12 个智能体 × 8轮次	重连的 − 格点 > 0	0.9/0.0/4.1	0.175	0.825	−0.650
15	互动网络与惯例 [?]	12 个智能体 × 10轮次	混合 − 局部 > 0	3.6/0.0/1.4	0.725	0.275	0.450
16	坚定少数派与公约定 [?]	12 个智能体 × 10轮次	多数 − 少数派 > 0	1.9/2.0/1.1	0.375	0.225	0.150
17	重复估计与社会信息 [?]	8 个智能体 × 8轮次	社会 − 独立 < 0	5.0/0.0/0.0	1.000	0.000	1.000
18	集中式与分布式的行为lief 更新 [?]	8 个智能体 × 8轮次	去中心化 − 中心化 > 0	4.6/0.0/0.4	0.925	0.075	0.850
19	协作搜索与共享 [?]	12 个智能体 × 10轮次	随机 − 本地 > 0	1.5/0.0/3.5	0.300	0.700	−0.400
20	间歇性接触和集体解决问题 [?]	8 个智能体 × 10轮次	间歇性 − 持续 <5.0/0.0/0.0 0		1.000	0.000	1.000
21	群体重组与累计设计 [?]	8 个智能体 × 10轮次	部分 − 完整 > 0	2.4/0.0/2.6	0.475	0.525	−0.050
22	文化市场人气累模拟 [?]	8 个智能体 × 8轮次	可见 − 隐藏计数 > 0	2.5/0.0/2.5	0.500	0.500	0.000
23	多轮信号和动作级联 [?]	8 个智能体 × 8轮次	可见 − 隐藏历史 > 0	2.4/1.0/1.6	0.475	0.325	0.150
24	市场价格反馈和趋-同 [?]	8 个智能体 × 8轮次	公开 − 私人记录 > 0	3.8/0.0/1.2	0.750	0.250	0.500
25	劳动力市场中的稳定身份 [?]	8 个智能体 × 8轮次	稳定 − 新身份 > 0	4.5/0.0/0.5	0.900	0.100	0.800
26	在线交易中的履约声誉交易 [?]	8 个智能体 × 8轮次	公开 − 私人经验 > 0	4.5/0.0/0.5	0.900	0.100	0.800
27	弱环 协调 and 群体规模 [?]	8 个智能体 × 8轮次	配对 − 整个群体 > 0	5.0/0.0/0.0	1.000	0.000	1.000
28	累积缓解和集体风险 [?]	6 个智能体 × 10轮次	高 − 低风险 > 0	3.1/0.2/1.6	0.625	0.325	0.300
29	沟通 组织和团队聚合 [?]	8 个智能体 × 8轮次	星 − 形 > 0	0.4/4.1/0.5	0.075	0.100	−0.025
30	群际竞争与协调 [?]	8 个智能体 × 8轮次	竞争 − 独立 > 0	2.5/1.2/1.2	0.500	0.250	0.250