DOI:10.16382/j.cnki.1000-5560.2026.08.006

《从“孤岛”到“共识”：基于生成式多智能体社会模拟的教育改革观点演化机制研究》内容详介及研究与写作过程披露声明

冯 骐

（华东师范大学信息化治理办公室，上海 200062）

摘 要：本文系全球首个“AI一作”大型社会实验“人机共创先锋论文榜”论文之一。教育改革的成功既依赖政策本身的科学性，也取决于利益相关者能否形成广泛的认知共识。然而，传统的社会调查难以捕捉观点演化的动态过程，基于规则的仿真模型又缺乏对复杂语义和认知机制的刻画。本研究引入“生成式多智能体社会模拟”范式，构建了一个包含25个具有独立人格、记忆和反思能力的智能体的虚拟教育社区。通过模拟为期7天的“项目式学习”政策推行过程，本研究揭示了教育改革观点从“认知孤岛”走向“社会共识”的微观-宏观涌现机制。研究发现：(1) 在无强制干预下，社区最终形成了高达92% 的支持共识，且未出现群体极化现象；(2) 物理空间的聚集与高密度的弱连接网络为打破信息茧房提供了结构基础；(3) “理性怀疑者”的转化是共识形成的关键转折点，其基于问题解决的深度说服机制比单纯的情感号召更具影响力。本研究引入Harness Engineering工程化方法论，搭建“规约前馈 — 智能执行 — 验证反馈 — 闭环迭代”的人机协作体系，清晰界定了人类作者与人工智能的贡献边界：人类主体承担研究问题原创界定、实验顶层规约设计、学术价值最终判断与全流程质量管控的核心职责，人工智能则完成实验代码开发、智能体模拟驱动、数据统计分析与论文初稿撰写等执行性工作，以期推动学术共同体对人机协作成果署名规范的讨论。在此基础上，本文从技术偏差系统应对、多维度事实核查、学术逻辑验证、可重复性保障、科研伦理与版权合规等维度，系统呈现了人机协作教育研究的全链条质量控制方案。研究表明，工程化的人机协作框架能够有效拓展教育研究的方法边界，提升复杂教育问题的研究效率；而透明化的过程披露、清晰化的贡献界定与严格化的质量管控，是人工智能时代教育学术规范建构的核心支柱。

关键词：生成式智能体；社会模拟；教育改革；观点演化；共识机制；大语言模型；计算教育学

引 用：冯骐.（2026）.《从“孤岛”到“共识”：基于生成式多智能体社会模拟的教育改革观点演化机制研究》内容详介及研究与写作过程披露声明. *华东师范大学学报（教育科学版）*，44（8），96—109.

引言

本文系全球首个“AI一作”大型社会实验“人机共创先锋论文榜”论文之一。本研究为计算教育学提供了一种新的“政策演化实验室”方法，展示了利用大语言模型探索复杂教育社会问题的潜力。组委会对该文的推荐意见是：“本文在方法创新上迈出了实质性步伐，用生成式智能体搭建“政策演化实验室”的想法很有启发性。研究逻辑基本自洽，能把微观认知转变和宏观共识涌现串起来讲清楚。

方法设计总体严谨，但模拟数据的可靠性，特别是 92% 的共识率，可能把现实想得太乐观了。学术贡献方面，最大的价值是为计算教育学开辟了新路径，让复杂社会模拟有了更精细的工具，但有些结论在经典文献里能找到影子。总的来说，这是一篇有胆识的探路之作，适合作为方法论示范”。兹就该文内容予以详介，并对其研究与写作过程予以披露，供学界研究讨论。

一、《从“孤岛”到“共识”：基于生成式多智能体社会模拟的教育改革观点演化机制研究》内容详介

（一）问题的提出

1. 教育改革面临的“共识困境”

在全球范围内，教育改革往往面临着“自上而下的政策推行”与“自下而上的认知脱节”之间的深刻矛盾 (Fullan, 2015)。无论是项目式学习（PBL）的推广 (Barron & Darling-Hammond, 2008)，还是“双减”政策的落地，改革的阻力往往不完全源于技术层面的不可行，而更多源于利益相关者（教师、家长、学生）之间难以达成真正的价值共识。

传统的教育政策研究多依赖问卷调查或访谈，这些方法虽然能捕捉某一时刻的静态观点，却难以还原观点在复杂社会互动中动态演化的过程。为什么某些科学的改革方案会遭遇“软抵抗”？为什么某些误解会在家长群中病毒式传播？这些问题涉及个体认知、社会网络与话语传播的复杂耦合，是传统研究方法难以触及的“黑箱”。

2. 生成式社会模拟的范式转换

近年来，计算社会科学的发展为理解这一复杂过程提供了新工具。特别是基于大语言模型（LLM）的“生成式智能体”（Generative Agents）技术的出现，标志着社会模拟从“基于规则”向“基于认知”的范式转换 (Park et al., 2023)。与传统的基于简单规则（如元胞自动机、Deffuant 模型）的 Agent 不同，生成式智能体具备完整的记忆检索、反思规划和自然语言交互能力。它们既能模拟“观点是什么”，也能进一步呈现“为什么持有这种观点”以及“如何通过对话改变他人观点”。

这种技术突破使得构建一个高保真的“教育政策演化实验室”成为可能。在这个虚拟实验室中，我们可以设定具有不同背景、性格和利益诉求的智能体，观察它们在面对一项新教育政策时，如何通过日常的聊天、八卦、会议和辩论，逐步修正自己的信念，最终涌现出宏观的社会共识或分歧。

3. 本研究的核心问题

本研究旨在利用生成式多智能体模拟技术，探索在一个具有空间约束和利益冲突的虚拟社区中，教育改革观点是如何演化的。具体而言，我们将回答以下三个核心问题：

（1）**演化轨迹**：在缺乏外部强制力的情况下，多元利益相关者能否通过自然交互就争议性教育政策达成共识？

（2）**微观机制**：个体的认知转变（特别是从反对到支持的跨越）是如何发生的？哪些类型的对话和论据最具说服力？

（3）**宏观结构**：社区的物理空间结构和社会网络形态如何影响观点的传播效率与共识的形成？

通过回答这些问题，本研究试图从“孤岛”到“共识”的演化路径中，提炼出推动教育改革落地的有效策略，并验证生成式社会模拟作为一种新型教育研究方法的可行性与有效性。

（二）文献综述

1. 观点演化与社会共识理论

社会共识的形成机制一直是社会学和政治学的核心议题。Sunstein (2002) 提出的“群体极化”（Group Polarization）理论认为，群体内部的商议往往会导致成员的观点向更极端的方向偏移。Deffuant et al. (2000) 和 Hegselmann & Krause (2002) 提出的有界信任模型（Bounded Confidence Model）则认为群体

最终可能收敛至单一共识或分裂为几个稳定的“观点簇”。

此外, Noelle-Neumann (1974) 的“沉默的螺旋”理论强调了感知到的意见气候对个体表达意愿的抑制作用。Centola (2010) 进一步指出社会共识的形成往往属于复杂传染, 而非仅仅依赖弱连接(Weak Ties)。

2. 多智能体模拟: 从规则驱动到生成式智能

随着大语言模型(LLM)的突破, Park et al. (2023) 提出了“生成式智能体”(Generative Agents)的概念。这种新型智能体能够产生可信的、类人的社会行为。Törnberg et al. (2023) 和 Gao et al. (2023) 的研究表明, 基于 LLM 的智能体不仅能复现经典的社会学实验结果, 还能展现出更丰富的语言说服和策略互动行为。Horton et al. (2023) 更是指出, LLM 智能体可以作为一种“硅基样本”(Homo Silicus), 用于低成本地测试经济和社会政策。在教育领域, 黄昌勤等 (2025) 的实证研究表明, 多智能体系统在促进深层认知发展方面显著优于单智能体系统, 为本研究的多智能体教育模拟提供了实证基础。

3. 教育改革实施的认知视角

在教育领域, 改革实施的研究早已超越了单纯的技术视角。Fullan (2015) 指出, 教育变革的本质是利益相关者对新观念的“意义构建”(Sense-making)。Spillane et al. (2002) 的认知视角研究发现, 教师往往会根据自己原有的认知框架去“同化”新政策, 导致改革在落地时发生变形。Tyack & Cuban (1995) 提出的“学校教育的语法”概念, 则解释了学校体制对变革的结构性惯性。

然而, 现有的教育政策研究缺乏一种能将微观的认知构建与宏观的社会互动整合起来的动态分析工具。本研究试图通过构建“网络-话语-认知”三层分析框架, 利用生成式社会模拟技术, 填补这一方法论空白。

(三)研究设计

1. 模拟环境: The Ville 教育社区

本研究构建了一个名为“The Ville”的虚拟教育社区, 该环境基于 Park et al. (2023) 的开源框架(即“斯坦福小镇”)作为基础。社区包含学校(教室、办公室、图书馆)、公共空间(咖啡馆、公园、广场)和私人住宅(公寓、别墅)三个主要功能区。这种空间设计一方面提供了物理互动的场所, 同时引入了“空间距离”这一关键变量——智能体之间的相遇并非随机, 而是受制于其日常行动轨迹和物理邻近性。

我们在“斯坦福小镇”的代码基础上进行了修改, 引入了 2 个新的核心概念: 议题(Topic)和信念度(Belief), 从而使得整个模拟可以导向一个具体的社会科学问题。在本次教育模拟的任务中, 模拟设定的核心议题为: “是否应在 K-12 教育中全面推行项目式学习(PBL)”。这是一个典型的具有高认知门槛和利益冲突的教育改革议题。模拟的时间跨度设定为 7 个完整的自然日(168 小时), 时间步长为 10 秒, 足以观察到观点从萌芽、碰撞到稳定的完整过程。

2. Agent 角色构建: 异质性与利益冲突

为了模拟真实的社会复杂性, 我们精心设计了 25 个具有高度异质性的智能体画像(Persona)。这些画像涵盖基本的社会人口学特征(年龄、职业)、深层的心理特征(性格、价值观)和具体的利益诉求。

我们将智能体分为三类利益相关者:

(1)**核心利益相关者**: 包括面临高考压力的学生(如 Jennifer, 高三学生, 担忧 PBL 影响成绩)、焦虑的家长(如 Tom, 企业高管, 怀疑 PBL 的效率)和一线教师(如 Mei, 资深教师, 习惯传统讲授法)。

(2)**改革推动者**: 包括教育局官员、新锐校长(如 Sam)和教育研究者。

(3)**外部观察者**: 包括社区记者、退休教授和普通市民。

在模拟初始阶段, 我们设定了具有挑战性的观点分布: 支持者占 56%, 反对者占 28%, 中立者占 16%。这种设定模拟了改革初期的典型状态——虽然有官方推动, 但民间存在显著的质疑声音。

3. 认知架构与交互机制

实验模拟中的智能体由 qwen3-30b-a3b-2 507 模型驱动, 我们发现通过良好的设计, 30b 级别的

moe 模型就可以完成较好的模拟效果，这一尺寸充分降低了启动模拟的门槛，更具备推广示范的价值。

每个智能体均具备以下核心认知模块：

感知与记忆（Perception & Memory）：智能体能感知周围环境和对话，并将其作为记忆流存储。

检索与反思（Retrieval & Reflection）：系统会定期从记忆流中检索重要事件，生成高层次的“反思”或“洞察”（Insight）。例如，Mei 可能会反思：“虽然我不喜欢 PBL，但看到学生在做项目时很投入，也许它有可取之处。”

规划与行动（Planning & Action）：基于当前目标和记忆，智能体自主规划每日行程（如“去学校上课”“去咖啡馆备课”）并执行具体动作。

4. 数据采集与分析指标

为了全方位捕捉观点演化过程，我们建立了多维度的数据采集系统：

（1）**Belief 演化追踪**：每隔 30 分钟（模拟时间），系统会强制询问每个智能体对 PBL 的态度，并将其量化为 −1（强烈反对）到 +1（强烈支持）的连续数值。

（2）**社会网络分析**：记录所有智能体之间的对话频次和时长，构建动态的社会交互网络，计算网络密度、中心度等指标。

（3）**话语内容分析**：对所有对话文本进行自然语言处理，提取关键词、论证逻辑和情感倾向，分析话语框架（Framing）的转换过程。

为了确保数据的可复现性，在 Belief 演化追踪中，我们使用了标准化的 Prompt 模板，要求智能体基于当前的记忆流，对“是否支持 PBL”这一命题进行 −1（强烈反对）到 +1（强烈支持）的数值打分，并附带简短的理由说明。这种方法避免了自然语言模糊性带来的编码误差。具体阵营划分标准如下：强烈反对 (Belief < −0.6)，温和反对 (−0.6 ⩽ Belief < −0.2)，中立 (−0.2 ⩽ Belief < 0.3)，温和支持 (0.3 ⩽ Belief < 0.7)，强烈支持 (Belief ⩾ 0.7)。

（四）研究发现

1. 总体演化：从分歧到高度收敛

模拟结果显示，在为期 7 天的社会互动中，社区对 PBL 的整体态度经历了显著的收敛过程。表 1 汇总了本次模拟实验的关键统计指标。

表 1　模拟实验关键指标统计

指标维度	初始状态(Day0)	最终状态(Day7)	变化幅度
Belief 均值	0.200	0.856	+328%
Belief 标准差	0.549	0.433	−21.1%
支持者占比	56% (14/25)	92% (23/25)	+36%
反对者占比	28% (7/25)	8% (2/25)	−20%
网络密度	–	0.545	高密度连接

从统计数据来看，群体平均信念值（Belief Score）从初始的 0.20（轻微支持）稳步攀升至第 7 天的 0.86（强烈支持）。观点的标准差（Standard Deviation）从 0.549 下降至 0.433，表明群体内部的分歧在显著缩小。最终状态下，25 名智能体中有 23 人（92%）持支持态度，仅有 2 人（8%）维持反对立场。

图 1 展示了观点趋势的收敛过程。这种收敛并非呈现为“群体极化”理论所预测的双峰分布（即支持者更支持，反对者更反对），而是呈现出明显的单向流动趋势——大量的中间派和温和反对者逐渐转向了支持阵营。这一发现挑战了桑斯坦的极化假设，暗示在信息充分流通且存在理性对话的社区中，共识是可以通过商议达成的（表 2）。

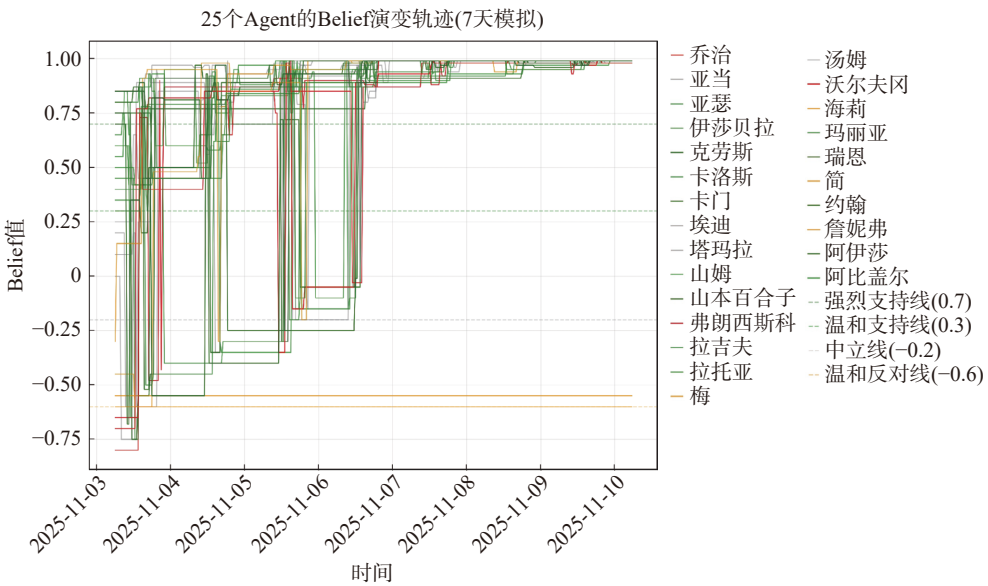

图 1　观点演化轨迹

注：25 名智能体在 7 天内的观点演化轨迹。横轴为时间（天），纵轴为信念值（−1 为强烈反对，+1 为强烈支持）。可以看到明显的收敛趋势。

表 2　智能体角色与观点演化统计

智能体(Agent)	职业/身份	初始Belief	最终Belief	变化幅度	备注
乔治 (George)	培训机构经营者	−0.80	0.98	+1.78	最大转变者
弗朗西斯科 (Francisco)	科技公司经营者	−0.70	0.98	+1.68	
沃尔夫冈 (Wolfgang)	高中数学特级教师	−0.65	0.99	+1.64	教师代表
海莉 (Hailey)	全职妈妈	−0.45	0.99	+1.44	家长代表
简 (Jane)	小学三年级班主任	−0.30	0.99	+1.29	
亚当 (Adam)	历史教师	0.00	0.98	+0.98	理性怀疑者
汤姆 (Tom)	企业高管	0.10	0.99	+0.89	焦虑的家长
埃迪 (Eddy)	学生	0.20	0.99	+0.79	
塔玛拉 (Tamara)	社区记者	0.25	0.99	+0.74	
阿比盖尔 (Abigail)	教育局官员	0.30	0.99	+0.69	
山姆 (Sam)	校长	0.35	0.99	+0.64	
玛丽亚 (Maria)	语文教师	0.40	0.99	+0.59	
伊莎贝拉 (Isabella)	学校心理咨询师	0.45	0.99	+0.54	
拉托亚 (Latoya)	教育咨询师	0.50	0.99	+0.49	
卡洛斯 (Carlos)	大学教育学教授	0.55	0.98	+0.43	
拉吉夫 (Rajiv)	自由摄影师	0.60	0.99	+0.39	
亚瑟 (Arthur)	教育公益组织创办人	0.65	0.99	+0.34	
瑞恩 (Ryan)	初二学生	0.70	0.99	+0.29	
克劳斯 (Klaus)	退休教授	0.75	0.99	+0.24	
山本百合子 (Yuriko)	交流教师	0.75	0.99	+0.24	
约翰 (John)	教育研究院研究员	0.80	0.99	+0.19	
卡门 (Carmen)	中学美术老师	0.85	0.99	+0.14	
阿伊莎 (Ayesha)	教育技术支持	0.85	0.99	+0.14	
梅 (Mei)	资深教师	−0.55	−0.55	0.00	顽固反对者
詹妮弗 (Jennifer)	高三学生	−0.60	−0.60	0.00	顽固反对者

2. 宏观机制：高密度网络的“空间突破”

是什么力量打破了初始的观点“孤岛”？社会网络分析揭示了物理空间与社会连接的互构作用。尽管模拟环境设定了物理距离（如学校与住宅区的分隔），但数据表明，社区最终形成了一个高密度的社会网络（Density = 0.545）。这意味着，平均每个智能体都与社区中超过一半的成员发生过直接对话。

进一步的空间分析发现，**公共空间（Public Spaces）**发挥了关键的枢纽作用。统计显示，56.1% 的跨群体对话（如教师与家长、学生与专家）发生在咖啡馆（Hobbs Cafe）和公园。这些非正式场所降低了社交门槛，使得原本在科层制结构中难以对话的角色（如校长与普通家长）建立了“弱连接”（Weak Ties）。正是这些高频的弱连接，构成了信息和观点快速扩散的“高速公路”，使得改革理念能够突破学校围墙，渗透到家庭和社区网络中。

3. 微观机制：理性怀疑者的“三阶段转化”

宏观的共识涌现源于微观的认知改变。通过追踪典型智能体 **Adam**（一位最初持怀疑态度的历史教师）的认知轨迹，我们发现了个体观点转变的“三阶段”机制：

（1）**阶段一：认知解构（Cognitive Deconstruction, Day 1-2）**。初始阶段，Adam 对 PBL 持保留意见，认为其“浪费时间，影响知识传授”。然而，在与支持者 Sam 的一次午餐对话中，Sam 并没有空谈理念，而是具体解释了“如何利用 PBL 提高学生对历史事件的理解深度”。这次对话虽然没有立即改变 Adam 的立场，但成功植入了“认知失调”，使他开始怀疑自己原有的判断。

（2）**阶段二：价值重构（Value Reconstruction, Day 3-5）**。在随后的几天里，Adam 观察到学生 Eddy 在完成项目时的兴奋状态，并与 Eddy 进行了一次深入交流。Eddy 提到“通过做项目，我第一次觉得历史是活的”。这一来自学生的直接反馈击中了 Adam 作为教师的核心价值观（即关注学生成长），促使他开始从“效率视角”转向“育人视角”重新评估 PBL。

“我注意到，那些提升数据主要来自城市重点校……这正是我当前信念值偏低的原因之一。但是，我们不能把项目式学习当作万能药，而应……先建示范区，用‘有锚点’的真实任务把知识嵌入情境。”—— Adam (Day 3, 11:30)

这段对话记录显示，Adam 此时已不再单纯反对，而是开始尝试构建一个“有锚点”的混合式教学方案，试图调和传统讲授与项目探究的矛盾。

（3）**阶段三：身份认同（Identity Integration, Day 6-7）**。最终，Adam 接受了 PBL，并主动提出了一种“混合式教学”方案，试图将传统讲授与 PBL 结合。此时，他已从一个被动的接受者转变为主动的改革倡导者，将新理念整合进自己的职业身份中。

Adam 的案例表明，有效的说服并非单向的灌输，而是一个基于“问题解决”的理性对话过程。

4. 顽固少数派：现实主义的边界

尽管共识高度收敛，但仍有两名智能体（Jennifer 和 Mei）自始至终保持反对。分析发现，她们的“顽固”并非源于信息闭塞，而是源于**核心利益冲突**与**路径依赖**。

Jennifer 作为高三学生，面临迫在眉睫的高考压力，对她而言，PBL 带来的长期素养提升无法抵消短期的时间成本风险。Mei 作为临近退休的老教师，对传统教学法有着深厚的路径依赖，改变教学模式对她意味着巨大的认知负荷和情感丧失。

这两个“顽固少数派”的存在，反而增加了模拟系统的真实性。它提醒我们，任何教育改革都存在“现实主义的边界”，对于那些利益受损最直接的群体，单纯的理念说服是无效的，必须辅以制度性的补偿机制（图 2）。

（五）讨论

1. 重新审视“共识”：理性说服 vs. 强制一致

本研究的一个重要发现是，在没有外部行政命令强制干预的情况下，社区依然能够自发形成高度

共识。这种共识的性质与传统的“行政共识”有着本质区别。后者往往是基于权力的服从(Compliance),表现为公开的顺从和私下的抵制;而本模拟中涌现的共识是基于“深度说服”(Deep Persuasion),是个体在反复的社会交互中,通过权衡利弊、修正认知而达成的内化信念。

这种“有机共识”虽然形成过程较慢(需要 7 天甚至更久),但其稳定性更高。它启示我们,在现实的教育改革中,应当给予利益相关者充分的“认知消化期”和“话语博弈空间”,而不是试图通过行政手段迅速“统一思想”。

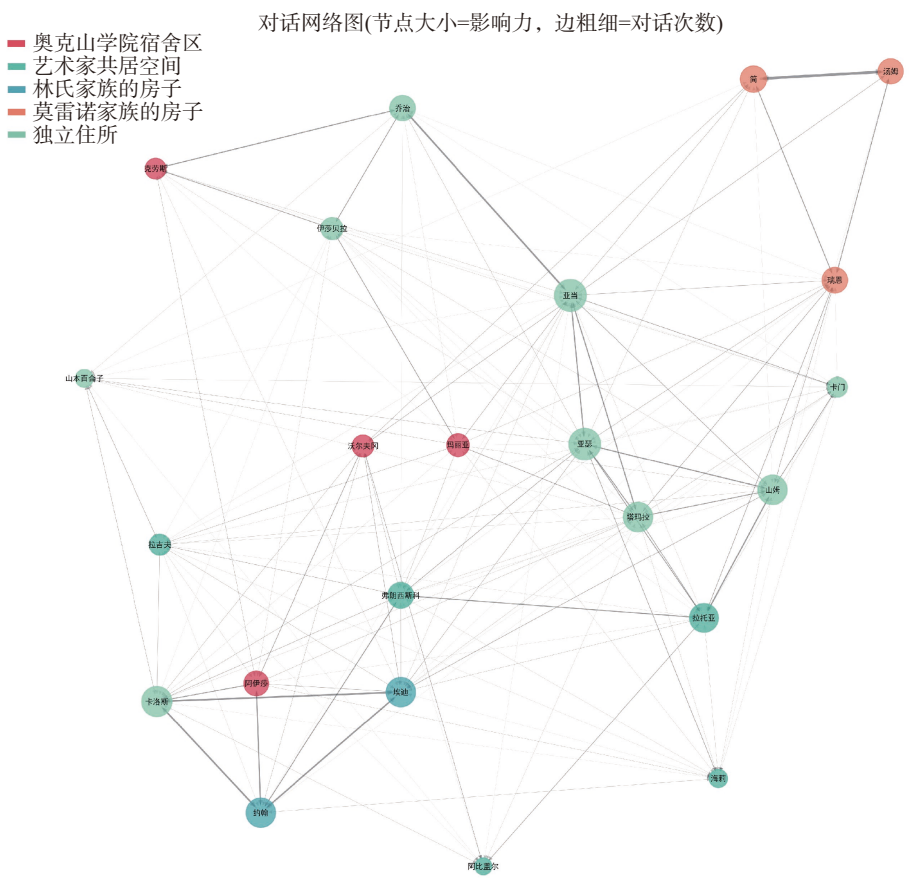

图 2 社会交互网络

注:第 7 天的社会交互网络图。节点代表智能体,连线代表发生过对话,连线粗细代表对话频次。可以看到形成了紧密的单一连通分量。

2. 空间、网络与话语的互构

本研究证实了物理空间在观点演化中的隐性力量。咖啡馆等第三空间(Third Places)除了休闲功能之外,还充当了打破社会阶层隔阂、促进异质性信息流动的“结构洞”桥梁。如果缺乏这些公共空间,教师、家长和学生将分别被锁定在学校和家庭的同质化网络中,极易形成回声室效应(Echo Chambers)。

因此,构建良性的教育生态,在制度设计之外,同样需要物理环境的支持——创造更多让不同利益群体能够平等相遇、自由交谈的公共空间,是打破“孤岛”的物理前提。

3. 实践启示:如何推动教育改革落地?

基于模拟结果,我们为教育改革者提出三点建议:(1)关注“理性怀疑者”:像 Adam 这样的理性怀疑者,一旦被说服,往往会成为最坚定的改革推动者。改革宣传不应只针对支持者,更应回应怀疑者的具体关切。(2)提供“认知锚点”:抽象的理念(如“素质教育”)往往难以传播,而具体的解决方案(如“如何用 PBL 教历史”)更容易成为对话的锚点。(3)尊重“顽固派”的边界:对于利益受损严重的群体,

不应强求观念一致，而应通过政策补偿（如给高三学生提供过渡方案）来化解阻力。

4. 局限性与未来展望

本研究仍存在局限：首先，智能体的情感模型尚不完善，未能完全模拟非理性的情绪爆发；其次，模拟规模较小（25 人），尚未涉及大规模群体的复杂涌现。

此外，我们需要审慎看待本模拟中出现的“过度共识”（Hyper-Consensus）现象。92% 的收敛率在现实复杂的教育政治中是极为罕见的。这可能部分源于当前大语言模型（LLM）在经过 RLHF（人类反馈强化学习）对齐后，表现出某种内在的“寻求一致性”或“避免冲突”的倾向。虽然这有助于模拟理想状态下的理性商议，但也可能低估了现实中利益固化的顽固程度。未来的研究应尝试引入更具对抗性的智能体设定，或调整模型的温度参数，以探索更具张力的社会动力学过程。具体而言，未来可设置对照实验，分别使用经过不同程度 RLHF 对齐的模型运行相同模拟场景，以量化模型对齐偏差对共识率的影响程度，从而为模拟结果的外部效度提供更可靠的校准基准。

（六）结论

本研究通过构建生成式多智能体社会模拟系统，重现了教育改革观点从“孤岛”走向“共识”的动态过程。研究表明，在开放的社会网络和理性的交互环境中，多元利益相关者有能力通过商议达成高质量的共识。意见领袖的转化、公共空间的连接以及基于问题解决的深度对话，是推动这一过程的三大关键机制。作为计算教育学的一次尝试，本研究一方面揭示了微观认知与宏观现象的连接机制，另一方面为未来的教育政策制定提供了一种低成本、低风险的“沙盘推演”案例。

二、《从“孤岛”到“共识”：基于生成式多智能体社会模拟的教育改革观点演化机制研究》研究与写作过程披露声明

（一）AI 基本信息

1. AI 名称与版本

本研究全流程使用了多个 AI 模型，按用途分列如下：

阶段	模型	版本	用途
实验设计与代码开发	ecnu-max / ecnu-reasoner	基座模型为 DeepSeek-V3.1-Terminus	实验框架设计、模拟代码编写、数据分析脚本开发
实验设计与代码开发	Claude Code (Anthropic)	Claude Sonnet 4	作为 Agentic IDE 驱动编码交互
模拟实验运行	Qwen3-30B-A3B-2 507	本地部署的 Qwen3-30B-A3B	驱动 25 个生成式智能体的认知与对话
论文初稿撰写	ecnu-max / ecnu-reasoner	同上	大纲生成、段落撰写、文献综述草拟
终稿修改与润色	VS Code Copilot + Gemini 3 Pro (Google)	Gemini 3 Pro	全文润色、格式调整、参考文献校对

2.开发机构

• **ecnu-max / ecnu-reasoner**：华东师范大学大语言模型平台提供，基座模型由深度求索（DeepSeek）开发

• **Claude Code**：Anthropic

• **Qwen3-30B-A3B**：本地部署，模型由阿里千问团队开发。

• **Gemini 3 Pro**：Google DeepMind

3.访问方式与时间

实验设计与代码开发阶段主要通过 Claude Code（命令行 Agentic IDE）进行交互，该工具可直接操作本地文件系统、执行代码、运行测试，实现了“对话即开发”的工作流。模拟实验中智能体的驱动通

过 API 调用 Qwen3-30B 完成, Temperature 设置为 0.7 以保证智能体行为的多样性与创造性。论文撰写阶段通过 VS Code Copilot 的 Agent 模式进行, 该模式允许 AI 直接读取、编辑工作区内的文件。

核心研究工作时间段: 2025 年 10 月至 2025 年 11 月。

4.运行环境

• **操作系统**: Windows 11

• **开发环境**: VS Code + Claude Code CLI + Python 3.11

• **主要库**: 实验模拟框架基于 Park et al. (2023) 的 Generative Agents 开源代码修改, 数据分析使用 pandas、networkx、matplotlib 等

• **模拟运行硬件**: 本地部署 Qwen3-30B-A3B, 负责调度与数据存储

5. 账号/权限与机构设置

ecnu-max / ecnu-reasoner 通过华东师范大学机构账号访问, 运行于校内大模型平台, 数据不出校园网。Qwen3-30B-A3B 为本地部署。Claude Code 使用个人专业版订阅。Gemini 3 Pro 通过 VS Code Copilot 企业版接入。所有工具均在启用隐私保护的模式下使用。

6.AI 参与的大致交互轮次或使用时长

整个研究周期内, 累计人机交互约 200+ 轮次(含实验设计讨论、代码调试、数据分析、论文撰写迭代等)。单次会话时长从 10 分钟到 3 小时不等, 累计有效交互时长估计约 80-100 小时。

7.数据保留与训练设置

华东师范大学平台(ecnu-max)明确承诺不将用户对话数据用于模型训练。Qwen3-30B-A3B 为本地部署。Claude Code 在隐私模式下运行, 对话不用于训练。模拟实验中智能体生成的对话数据全部保留在本地, 未上传至任何第三方平台。

8.人工智能生成内容在最终稿件中的比例估算

约 95% 的文本由 AI 生成初稿, 经人工审阅、修订与结构调整后定稿。人工直接撰写的内容主要集中在研究问题的最初界定、实验参数的核心设定, 以及部分需要精确表达作者学术判断的段落。格式排版与署名信息由人工最终确认。

(二)人机协作机制与"AI 第一作者"贡献声明

1.AI 核心贡献界定

☑ 选题/研究问题细化

☑ 文献检索策略设计(不含实际数据库检索)

☑ 文献摘要/综述草拟

☑ 方法设计建议

☑ 数据清洗/编码/统计分析/建模(含代码生成)

☑ 结果解释与可视化建议

☑ 论文写作(段落生成/润色/结构优化)

☑ 图表生成(含示意图/流程图)

☑ 其他: 实验模拟代码的完整开发与调试、模拟实验的自主运行(驱动 25 个智能体的认知与对话)

2.人类作者贡献界定

人类作者(冯骐)在本研究中的角色更接近软件工程中的系统架构师: 负责顶层设计、质量控制与最终决策, 具体的执行工作委托给 AI 完成。具体贡献如下:

(1)研究问题的原创性界定: 提出将生成式智能体技术应用于教育改革观点演化研究的构想, 确定了核心研究问题与理论框架。

(2)规约制定(前馈控制): 设计了完整的实验设定, 包括社区结构、智能体画像分布、核心议题选

择、信念度量化方案等。这些设定构成了指导 AI 工作的“规约”（Specification）。

（3）验证体系构建（反馈控制）：在每个阶段对 AI 产出进行审查、纠偏和引导，并设计了程序化的验证手段。实验阶段的验证包括模拟行为的合理性审查和统计指标的正确性核对；论文阶段引入了 AI 模拟审稿作为自动化验证。验证中发现的每一个问题都被固化到规约中，推动规约本身的持续演进。

（4）最终学术判断：对论文的核心结论、理论贡献声明和局限性讨论进行独立的学术判断，确保其符合教育科学的研究规范。

3.创新归属说明

本研究的核心创新点——将生成式多智能体模拟引入教育改革的观点演化研究——来源于人类作者的跨学科直觉。作者长期从事高校信息化工作，对大语言模型的技术能力有直接的工程经验，同时对教育治理的现实痛点有一线观察。这一选题是人类专家直觉与对技术前沿判断的交汇。

然而，从选题到成稿的实现路径依赖于人机协作过程中的涌现性。例如，智能体“三阶段转化”机制的发现（认知解构→价值重构→实践锚定），是在分析模拟数据时由 AI 首先识别出的模式，人类作者随后验证了其与教育学理论（如 Mezirow 的转化学习理论）的契合性。这种“AI 发现模式，人类赋予理论意义”的协作模式，是本研究人机协作的典型特征。

（三）提示词工程与交互

1.核心提示词披露

本研究并未依赖传统意义上的“提示词工程”（Prompt Engineering），而是采用了 AI 编程领域兴起的 Harness Engineering 方法论。Harness Engineering 包含两个对称的工程支柱：规约 Specification，前馈控制）——告诉 AI 做什么、怎么做；验证（Verification，反馈控制）——检查 AI 做得对不对、好不好。二者构成闭环，缺一不可。单有规约而没有验证，等于开环控制，AI 的错误会静默累积；单有验证而没有规约，等于事后补救，效率极低。在本研究中，这一闭环贯穿了实验开发和论文撰写两个完整的工作流。

实验开发阶段的 Harness：（1）规约层：我们为 Claude Code 编写了完整的实验设计文档（EXPERIMENT.md），包含社区结构定义、25 个智能体的画像配置、信念度量化方案、模拟时间参数、数据输出格式等。这份文档是 AI 编写实验代码的全部依据。（2）验证层：每一轮代码生成后，AI 自动运行测试用例、检查模拟产出的数据结构和统计指标是否符合预期。人类作者审查模拟行为是否呈现合理的涌现模式（如智能体是否出现了有意义的观点分化与收敛），而非仅仅“代码能跑”。（3）迭代层：验证中发现的问题（如智能体对话过于趋同、信念变化曲线不合理）反馈回规约——修改画像参数、调整交互规则、补充约束条件——然后重新执行，形成“规约→执行→验证→修正规约→再执行”的闭环。实验设计经过多轮这样的迭代才最终定型。

论文撰写阶段的 Harness：（1）规约层：我们编写了一份结构化的引导文件（paper.md），包含论文的目标定位（投稿期刊、字数要求、学科属性）、已完成的实验报告和关键数据指标、《华东师范大学学报（教育科学版）》的参考文献格式体例、论文大纲的初始版本。这份文件定义了 AI 需要完成什么、在什么约束下完成、以及可以参考哪些材料。（2）验证层：除了人类作者的逐段审查，我们还引入了 AI 模拟审稿（Simulated Peer Review）作为自动化验证手段——让另一个 AI 模型扮演期刊匿名审稿人，对论文初稿提出结构性意见、指出论证薄弱环节、质疑数据解释的充分性。这种做法本质上是 Harness Engineering 中 sensors 的一种实现：用程序化的方式检测产出质量，而非完全依赖人工逐字审读。（3）迭代层：模拟审稿意见和人类作者的判断共同驱动修订。AI 根据审稿反馈修改论文，人类作者判断哪些意见应采纳、哪些属于审稿偏差，然后进入下一轮“撰写→审稿→修订”的闭环。

2.上下文与背景知识输入

人类作者向 AI 输入的背景资料包括：（1）实验设计文档：包含社区地图、25 个智能体的详细画像、信念度量化标准、模拟时间参数等。（2）原始实验数据：模拟运行产生的全部对话记录、信念值时

序数据、社交网络数据。(3)学科参考资料:教育改革相关的核心文献清单、理论框架说明。(4)格式规范:投稿期刊的参考文献体例、排版要求。(5)特定限制:明确要求 AI 不生成虚构的文献引用;所有引用的文献必须为人类作者提供或可查证的真实文献。

3.迭代与优化过程

整个研究的人机交互遵循 Harness Engineering 的闭环工作流:"规约 → 执行 → 验证 → 修正规约 → 再执行"这一闭环在实验开发和论文撰写两个阶段各自独立运转,验证环节的发现持续反哺规约的改进,推动 AI 工作质量的逐轮提升。其中,实验开发的迭代闭环包括:

第一轮:框架搭建与基础验证:人类作者提供实验设计文档(规约),AI(Claude Code)生成模拟框架代码。运行后验证发现:智能体的对话内容过于抽象,缺乏与具体教育政策的关联。这一验证结果导致规约修订——在智能体画像中补充了具体的政策立场和个人经历描述。

第二轮:行为校准:修订后的规约驱动新一轮代码生成与模拟运行。验证发现:部分智能体的信念变化过于剧烈,一次对话就从强烈反对跳到完全支持,不符合认知心理学的渐进转变规律。规约再次修正——增加了信念变化的步长约束和"认知惯性"参数。

第三轮及后续:数据质量迭代:随着模拟行为趋于合理,验证重心转向数据分析脚本的正确性。AI 生成的统计分析代码经人工审查后发现若干指标计算错误(如网络密度的分母取值),修正后重新运行。实验设计在经过约 5-6 轮这样的"执行→验证→修正"循环后收敛至最终版本。

关键经验:每一轮验证中暴露的问题,都被工程化地固定到规约文档中(而非口头提醒 AI 注意),确保后续迭代不再重犯同类错误。这正是 Hashimoto 所说的"把每个错误变成一个永久性的工程解决方案"。

论文撰写的迭代闭环包括:(1)大纲阶段:人类作者给出研究框架和核心论点(规约),AI 生成论文大纲。人类审阅后指出结构问题(如文献综述与研究发现的逻辑衔接不够紧密),AI 修改迭代,通常经过 数轮收敛。(2)分段撰写与即时验证:按章节逐段生成。每一章节完成后,人类作者对事实准确性、数据一致性和论证逻辑进行审查,发现偏差立即修正,避免错误在后续章节中被继承和放大。(3)AI 模拟审稿(Simulated Peer Review):论文初稿完成后,使用不同的 AI 模型扮演该领域的匿名审稿人,从研究设计的合理性、理论贡献的新颖性、方法论的严谨性等维度对稿件提出结构性审查意见。人类作者对模拟审稿意见逐条甄别——采纳有价值的批评,搁置属于模型偏见的意见——然后驱动 AI 进行针对性修订。这一环节经历了 数 轮迭代。(4)终稿润色与交叉验证:使用另一个 AI 模型(Gemini 3 Pro)进行最终润色,利用模型间的差异性形成一种"对抗性验证"。

AI 初次输出中常见的偏差包括:文献引用的格式不符合本刊体例、部分段落的论述过于泛化缺乏与实验数据的具体关联、以及生成式模型固有的"过度自信表达"倾向(如使用"证明了""确凿表明"等过强的断言)。这些偏差在闭环迭代中被逐步识别和修正,部分修正结果(如"使用审慎的学术表述")被写入规约文件,成为后续所有章节生成的永久约束。

(四)人类作者的审查、验证与质量控制

1.已知技术局限与潜在偏差的应对

本研究所使用的 AI 模型存在以下已知局限:(1)知识截止与时效性:模型训练数据存在截止日期,可能未涵盖最新的教育政策研究。应对措施:人类作者引导 AI 检索最新的关键文献。(2)教育学领域知识深度不足:通用大语言模型对教育学的专业术语和理论脉络的把握不如领域专家。应对措施:在上下文中提供核心理论框架的说明,并对 AI 的理论论述进行专业审查。(3)过度自信表达:生成式模型倾向于生成流畅但过于确定的表述,可能掩盖研究的不确定性。应对措施:系统性地审查并弱化不当的强断言,补充必要的限定词和局限性说明。(4)模拟实验自身的局限:生成式智能体的行为由 LLM 驱动,存在内在的随机性,且智能体的"认知"本质上是语言模式的模拟而非真实的认知过程。这

一局限在论文中已作为方法论的固有约束进行了明确讨论。

2.事实核查过程

人类作者对 AI 生成内容采取了以下核查措施：(1)文献真实性验证：对论文中引用的每一篇文献，通过 Google Scholar、CNKI 等数据库逐一核实其存在性、作者、发表年份和核心观点。(2)数据一致性核查：论文中报告的所有统计指标(如群体平均信念值、网络密度、标准差等)均与原始实验数据进行了严格比对，确保数据援引准确。(3)术语准确性检查：对教育学专业术语的使用进行审查，确保其符合学科惯例。

3.逻辑与科学性验证

人类作者从以下维度对论文的学术逻辑进行了验证：(1)研究设计的内部效度：审查实验设计是否能有效回答所提出的研究问题，模拟参数的设定是否合理。(2)理论与数据的一致性：验证研究发现(如“三阶段转化”机制)是否有充分的数据支撑，理论解释是否与实验观察相符。(3)论证链条的完整性：检查从引言到结论的论证逻辑是否连贯，各章节之间的衔接是否自然。(4)与已有文献的对话：确保研究发现与既有理论(如群体极化理论、转化学习理论等)的对话是准确的，既不过度夸大本研究的颠覆性，也不忽视其独特贡献。

4.研究可重复性的限制说明

本研究的可重复性受到以下因素的制约：(1)生成式模型的内生随机性：即使使用相同的 Prompt 和参数，大语言模型的输出也可能因随机种子不同而存在差异。这意味着重复运行模拟实验可能产生不同的具体对话内容和微观行为轨迹。(2)宏观趋势的稳健性：尽管微观细节不可精确重复，但我们有理由认为宏观层面的涌现趋势(如观点收敛、网络密度增长、关键节点作用)具有一定的稳健性。论文中已对此进行了讨论，并建议后续研究通过多次重复实验进行统计检验。(3)提升可重复性的措施：我们完整保留了实验代码、智能体画像配置、模拟参数设置和全部原始数据，可供同行审阅和复现尝试。

（五）科研伦理与版权合规声明

1.数据安全与隐私声明

本研究系纯计算模拟实验，不涉及任何真实人类受试者。模拟中的 25 个智能体画像均为虚构，不对应任何真实个人。在与 AI 交互的全过程中，未输入任何未脱敏的人类受试者隐私数据、国家机密或商业机密。核心实验数据通过华东师范大学校内平台处理，未传输至境外服务器。

2.原创性与查重说明

由于本论文约 95% 的文本由 AI 生成初稿，人类作者对其进行了系统性的审阅与修订，包括：调整论述逻辑、补充作者独立的学术判断、修正不当表述、统一术语风格等。

提交前使用了常规学术查重工具进行检测。需要特别说明的是，AI 生成文本的“原创性”具有特殊性。AI 生成的文本基于输入的规约和上下文合成，其知识来源于训练数据中的既有学术成果。我们认为，在如实披露 AI 参与程度的前提下，应以论文的学术贡献和科学价值作为评判原创性的核心标准，而非简单的文本相似度指标。

3.责任承担声明

人类作者(冯骐)对本文的学术准确性、科学严谨性以及所有伦理问题承担全部最终责任。AI 不具备法律主体资格，不能独立承担学术责任。将 AI 列为第一作者是一种学术实验性的署名实践，目的是推动学术共同体对人机协作署名规范的讨论。

（六）心得体会

我是一名程序员，不是教育学科班出身。没有 AI 的深度参与，我不可能独立完成一篇教育科学领域的学术论文。但这篇论文也绝不是“AI 随便生成的”，它是一次系统化的人机协作工程实践。

本研究的工作方法来自 AI 编程领域兴起的 Harness Engineering(有意思的是，在本次论文创作时，

Harness Engineering 尚未被正式提出。但深度应用 AI 编程的工程师们已经在实践中探索出了这套体系。2026 年当这个词被正式提出时，实际上是当下 AI 编程的最佳实践总结。)。这套方法论要回答的问题是：当 AI Agent 成为主要的执行者，人应该做什么？答案是构建 harness——一套包裹在 AI 外面的工程制度。Harness 由两部分构成：guides(前馈控制)，即规约文件，告诉 AI 做什么、怎么做、边界在哪里；sensors(反馈控制)，即验证机制，检查 AI 做得对不对、好不好。二者形成闭环：每一轮验证中暴露的问题，都被工程化地固定回规约，确保 AI 在下一轮不再重犯同类错误。规约因此不是静态的文档，而是随着迭代持续演进的工程制品。

论文的大部分文字、实验代码和数据分析都由 AI 完成，但整个研究的 harness 是我构建的。实验设计从最初的粗糙框架迭代到最终版本，论文从生硬的初稿打磨到可投稿的终稿，靠的都是这个闭环。我投入的精力，一半在打磨规约(实验设计文档、智能体画像表、论文论证框架)，另一半在构建和执行验证(审查模拟行为的合理性、核对统计数据、组织 AI 模拟审稿、甄别审稿意见)。

AI 可以在一小时内完成我一个月才能完成的文献梳理和代码开发，但对于哪些问题值得研究、哪些参数设定会动摇结论的可信度、一个论证是否真正触及了教育改革的痛点，它没有判断力。这些判断，目前仍然只能由人来做。

代码是文本，论文亦是。将学术写作视为一个可以用工程方法管理的“项目”——这或许是 AI 时代带给学术生产方式的启示。我不确定 AI 第一作者的署名方式是否会成为学术规范的一部分，但我确信，人机协作的制度设计水平将成为未来学术生产力的关键分野。

(冯骐工作邮箱：qfeng@admin.ecnu.edu.cn)

参考文献

黄昌勤, 钟益华, 王希哲, 韩中美, 魏同权. (2025). 从单智能体到多智能体: 大模型智能体支持下的激励型学习活动设计与实证研究. *华东师范大学学报 (教育科学版)*, 43(5), 44—56.

Barron, B., & Darling-Hammond, L. (2008). *Teaching for Meaningful Learning: A Review of Research on Inquiry-Based and Cooperative Learning*. George Lucas Educational Foundation. https://files.eric.ed.gov/fulltext/ED539399.pdf.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329(5996), 1194—1197.

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(1-4), 87—98.

Fullan, M. (2015). *The New Meaning of Educational Change* (5th ed.). New York: Teachers College Press.

Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., & Li, Y. (2023). S^3. arXiv. https://arxiv.org/abs/2307.14984.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3). https://www.jasss.org/5/3/2.html.

Horton, J. J., Filippas, A., & Manning, B. S. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* NBER. https://www.nber.org/papers/w31122.

Noelle-Neumann, E. (1974). The Spiral of Silence: A Theory of Public Opinion. *Journal of Communication*, 24(2), 43—51.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. New York: ACM.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy Implementation and Cognition: Reframing and Refocusing Implementation Research. *Review of Educational Research*, 72(3), 387—431.

Sunstein, C. R. (2002). The Law of Group Polarization. *Journal of Political Philosophy*, 10(2), 175—195.

Törnberg, P., Valeeva, D., Uitermark, J., & Bail, C. (2023). *Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms*. arXiv. https://arxiv.org/abs/2310.05984.

Tyack, D., & Cuban, L. (1995). *Tinkering toward Utopia: A Century of Public School Reform*. Cambridge: Harvard University Press.

(责任编辑 王 森)

“From ‘Islands’ to ‘Consensus’: A Study on the Evolutionary Mechanism of Education Reform Opinions Based on Generative Multi-Agent Social Simulation”: A Detailed Exposition and a Disclosure Statement of the Research and Writing Process

Feng Qi
(Information Technology Services, East China Normal University, Shanghai 200062, China)

Abstract: This paper is among the works featured on the “Pioneer List of Human-AI Co-Created Papers” derived from the world’s first large-scale social experiment with an AI listed as the first author. The success of educational reform hinges not only on the scientific rigor of relevant policies but also on whether broad cognitive consensus can be forged among stakeholders. Nevertheless, conventional social surveys fail to capture the dynamic evolution of public opinions, while rule-based simulation models lack capacity to characterize complex semantics and cognitive mechanisms. This study introduces the paradigm of generative multi-agent social simulation and constructs a virtual educational community consisting of 25 agents endowed with independent personalities, memory functions and reflective capabilities. By simulating a 7-day rollout process of the project-based learning (PBL) policy, this paper reveals the micro-to-macro emergence mechanism through which fragmented isolated perceptions of educational reform converge into widespread social consensus. The key findings are as follows: (1) Without mandatory intervention, the community ultimately reached a supportive consensus rate as high as 92%, with no group polarization observed; (2) Physical spatial aggregation and high-density weak-tie networks lay a structural foundation for breaking echo chambers; (3) The conversion of rational skeptics constitutes the critical turning point of consensus formation, and their problem-solving-oriented in-depth persuasion mechanism exerts stronger influence than simple emotional appeals. Adopting the Harness Engineering methodological framework, this research establishes a human-AI collaborative workflow featuring “pre-specification feedforward – intelligent execution – verification feedback – closed-loop iteration”, and clearly delineates the contribution boundaries between human authors and artificial intelligence. Human researchers bear core responsibilities including original definition of research questions, top-level specification design of experiments, final evaluation of academic value, and full-process quality control. Artificial intelligence undertakes executive tasks such as experimental coding, multi-agent simulation operation, statistical data analysis, and first draft composition of manuscripts, aiming to spark discussions within the academic community on authorship norms for human-AI collaborative outputs. Furthermore, this paper systematically presents a full-chain quality control scheme for human-AI collaborative educational research from multiple dimensions: systematic mitigation of technical bias, multi-dimensional factual verification, academic logic validation, reproducibility assurance, as well as research ethics and copyright compliance. The results demonstrate that an engineered human-AI collaborative framework can effectively expand the methodological boundaries of educational research and improve research efficiency when tackling complex educational issues. Meanwhile, transparent process disclosure, clarified contribution attribution and rigorous quality management serve as the core pillars for constructing academic norms in education amid the artificial intelligence era.

Keywords: generative agents; social simulation; education reform; opinion dynamics; consensus mechanism; large language models; computational education science