arXiv:2603.01339v1 [stat.ML] 2 Mar 2026

交互式人类-人工智能系统中未观测单元类型的因果效应

William Overman、Sadegh Shirani 和 Mohsen Bayati

斯坦福大学商学院

摘要

我们研究人类与人工智能智能体交互群体中的实验，其中单元类型和交互网络均不可观测。尽管因果效应会在整个系统中传播，我们的目标是估计对人类的效应。例如，在在线平台上，人类用户会与人工智能驱动账户互动。我们假设存在一个人类-人工智能先验，为每个单元属于人类的概率赋值。虽然无法在单元层面区分人类，但该先验使我们能够计算大型子总体中的平均人类构成。随后，我们通过因果消息传递（CMP）框架对结果动态建模，并分析各子总体的样本均值结果。我们表明，通过构建期望人类构成和处理暴露各不相同的子总体，可以一致地识别人类特异性因果效应。我们的结果刻画了在何种情况下，仅凭总体构成的分布知识（无需观测单元类型或交互网络）便足以实现识别。我们在一个由行为差异化大语言模型智能体驱动的模拟人类-人工智能平台上验证了该方法。总之，这些结果为新兴人类-人工智能系统中的实验提供了理论与实践框架。

1 引言

网络环境中的实验在各个科学领域都至关重要。当结果取决于个体之间的互动时，经典方法便无法再单独识别处理效应，因此，对溢出效应和动态过程进行细致建模至关重要 [Ugander等人，2013年，Eckles等人，2016年，Athey等人，2018年，Farias等人，2022年，Hu等人，2022年，Johari等人，2022年，Shankar等人， 2025]。

与此同时，人类-人工智能系统正在成为数字平台中的一种新范式。现代人工智能的最新进展使自主智能体能够近似模仿人类行为[Horton，2023年，Park等人， 2023] 并与大量人类用户互动。因此，社交媒体平台等在线环境可能包含由人类和人工智能驱动型智能体构成的混合总体，它们之间的互动共同塑造总体结果。

在这种背景下，对人类与人工智能代理组成的混合总体进行实验，面临的挑战远超经典的网络干扰问题。在标准情形中，实验单位通常是同质的，交互网络也可能只能被部分观测。相比之下，在人类-人工智能平台中，单位的身份往往无法观测，交互结构也难以直接刻画：人类通过多个线上和线下渠道进行交互，而人工智能智能体则局限于线上、由平台介导的交互。

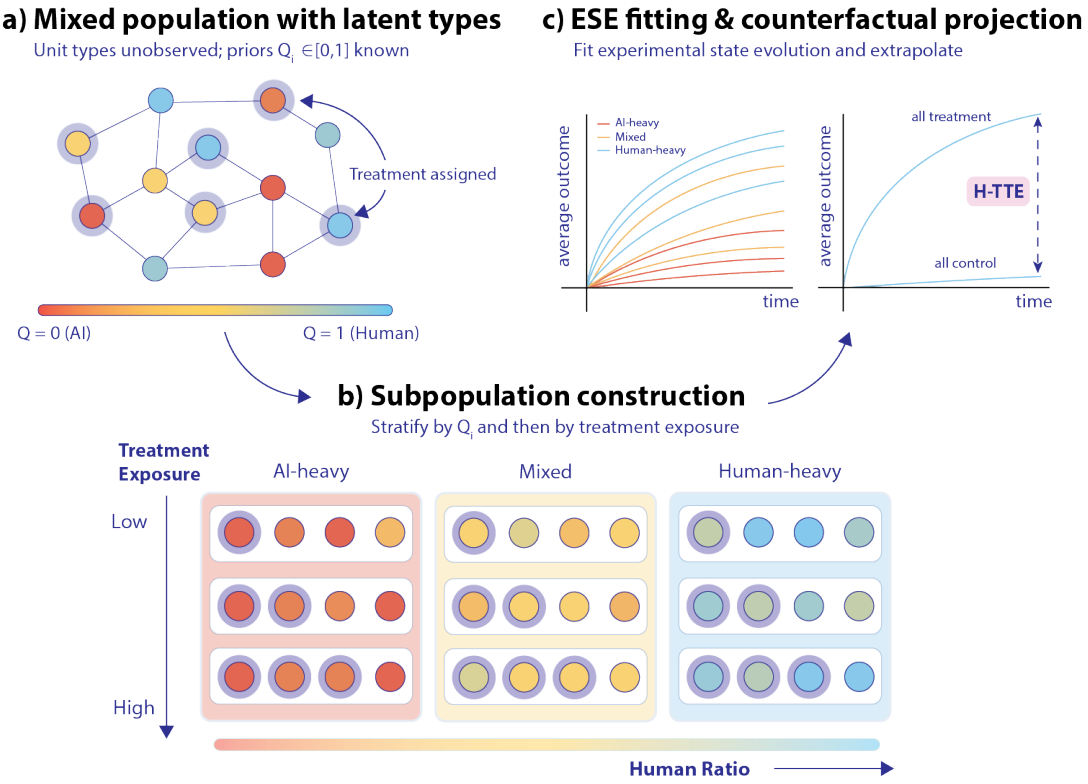


图 1：**所提出框架的概览**。(a) 由人类和人工智能代理组成、且单位类型未观测的混合总体；每个单位 i 都关联一个已知先验概率 $Q_i \in [0,1]$，表示其为人类的概率，并随机分配处理。(b) 通过沿两条轴对单位进行分层来构建子总体：期望人类构成（水平轴）和处理暴露（垂直轴），从而得到在这两个维度上均存在系统性差异的群组。(c) 将实验状态演化（ESE）拟合到这些子总体的总体结果轨迹上，随后用于推演全处理和全对照组条件下的反事实结果，并将构成设为 $q^S = 1$（仅人类）。这些推演轨迹之差即为估计的人类总处理效应（H-TTE）。

当单位类型是潜在的时，个体层面观测只能提供有限信息，无法确定用户是人类还是人工智能——充其量只能构建分类器，为每个单位分配其属于人类的概率。因此，结果会在连接结构不透明的异质单位之间通过动态交互而演变，但人们关心的目标估计量却是特定类型的（例如，仅针对人类用户的干预因果效应）。单位类型的不确定性，以及由此导致的底层交互结构不确定性，使得人类特定效应的估计本质上极具挑战性。

针对网络结构未观测情况下的实验，Shirani 和 Bayati [2024] 提出了 CMP 框架。在单位类型同质的情形下，CMP 无需显式了解网络结构，而是通过分析总体结果随时间的演化 [Shirani 和 Bayati，2025a]。通过将分析转向总体层面状态，CMP 为人类 - 人工智能系统提供了一个自然的切入点：在大型总体中，关于单位类型的分布知识足以刻画总体结果演化。

在这项研究中，我们将 CMP 框架扩展到单位类型不确定的情境。图 1

概述了我们的方法。具体而言，我们假设每个用户都对应一个其属于人类的先验概率，并将这一信息纳入类型依赖的结果模型。该模型允许处理响应和交互效应随潜在类型而变化，从而在不观察用户身份的情况下捕捉异质动态。随后，我们推导出刻画样本平均结果低维动态的状态演化（SE）方程。具体来说，我们证明，在适当构造的子总体中，总体结果的演化仅取决于其人类–AI 先验的平均值。由此，我们得到一种实用的估计程序，可以估计干预对人类的因果效应，仿佛单位类型已被观察到。

为评估我们的框架，我们开发了一个混合人类—人工智能在线平台模拟器。人类和人工智能代理均由大语言模型（LLM）实例化，但采用不同的提示和温度配置，以诱发系统性差异化的行为模式。平台无法观测用户类型，而用户的参与度作为结果，会随动态交互而变化。处理是向用户展示由平台精选的“成功故事”，旨在抵消机器人生成的活动。利用这一环境生成的数据，我们表明，尽管单位类型具有潜在性且交互结构不可观测，我们的方法仍能根据汇总观测准确估计人类特定处理效应。

总之，我们提出了一个框架，用于在单位类型隐含且交互不可观测的混合人类—人工智能系统中估计人类特异性因果效应。我们推导出低维状态演化方程，仅凭总体构成的分布知识，即可利用汇总数据实现识别。最后，我们提供了一种基于大语言模型的模拟，并表明我们的估计量能够准确恢复真实人类处理效应。这些结果共同确立了一种实用方法，可用于在新兴的人类—人工智能环境中开展实验。

2 相关文献

网络干扰 经典因果框架依赖稳定单元处理值假设（SUTVA），该假设排除了单元之间的干扰 [Cox, 1958, Rubin, 1980]。在相互作用的环境中（例如社交网络、市场平台和数字平台），SUTVA 不再成立，因此催生了大量关于网络干扰的文献 [Manski, 2013, Ugander等人, 2013, Eckles等人, 2016, Aronow和Samii, 2017, Saveski等人, 2017, Pouget-Abadie等人, 2019, Holtz等人, 2020, Munroet al., 2021,Hu等人, 2022, Han etal., 2022, Agarwal等人, 2022, Belloni等人, 2022, Li和Wager, 2022a,b, Ni等人, 2023, Viviano等人, 2023, Tan等人, 2025, Peng等人, 2025, Bright等人, 2025]。

当网络结构未知时，Yu等人 [2022], Cortez等人 [2022] 利用重复测量。Farias等人 [2022, 2023] 研究马尔可夫干扰模型，Wager和Xu [2021],Johari等人 [2022] 研究平均场近似，而 [Hu和Wager、2022年、Bojinov等人、2023年、Xiong等人、 2024] 使用切换实验。最近，Shankar等人 [2025]研究了网络结构本身随处理分配而变化的问题。CMP 框架 [Shirani 和 Bayati、2024年、Bayati等人、2024年、Shirani 和 Bayati、2025a、Shirani 等人、 2025] 通过分析总体结果动态来绕过网络知识的要求，并已在真实实验 [Tan等人、 2026]中，针对网络感知基准进行了实证验证。然而，上述所有方法都假设实验单位来自同质总体。我们的工作将该框架扩展到具有未观测单位类型的情形，从而能够估计类型特定因果效应。

未观测异质性 大量研究利用观测协变量估计处理效应异质性 [Wager 和 Athey，2018年，Athey 等人，2019年，K¨unzel 等人， 2019]。当子群体不可观测时，主分层 [Frangakis 和 Rubin， 2002] 以及有限混合

模型 [Muth´en 和 Shedden, 1999]用于估计潜在分层内的效应。另有一类研究关注在线平台上 AI 生成的账户的检测 [Ferrara 等人, 2016], 但其重点是分类，而非因果估计。我们将各研究方向中尚未解决的挑战结合起来：单位类型是潜在的，单位通过一个未知网络相互作用，而目标是恢复类型特定因果效应。

基于大语言模型的社会系统模拟 近期研究使用大语言模型在社会情境中模拟人类行为，从生成式智能体 [Park 等人, 2023] 到模拟社交媒体平台[T¨ornberg 等人, 2023] 以及大规模社会进程 [Piao 等人, 2025]。尽管这些研究主要以描述为主，且缺乏因果推断框架，Shirani 和 Bayati [2025b]将基于大语言模型的模拟器用作受控实验平台。我们采用类似方法，使用大语言模型构建交互式人类-人工智能系统的模拟器。

信念传播与消息传递 CMP 框架受到信念传播（BP）思想的启发，并建立在近似消息传递（AMP）之上。BP 由Pearl [1982] 提出，用于树结构贝叶斯网络上的精确推断，并在 Pearl [1986, 2014], 中扩展至一般信念网络，从而为现代消息传递方法提供了算法起源。在稠密随机图上，AMP 算法通过带有 Onsager 修正的标量汇总量来近似 BP [Donoho 等人，2009年, Bayati 和 Montanari，2011年，Bolthausen, 2014]。CMP 将这一机制重新用于未知网络干扰下的因果推断。

3 问题概述

我们考虑一个由 N 个个体组成的相互作用总体，个体索引为 $i \in [N] := \{1, \ldots, N\}$，并在有限时间范围 $t = 0$（1，.....，T）内进行观测。对于每个个体 i 和时间点 t，我们观测到一个二元处理分配 $W_t^i \in \{0, 1\}$ 和一个标量结果 $Y_t^i \in \mathbb{R}$。我们假设在 $t = 1$ 之前没有分配任何处理，因此对于所有 i，$W_0^i = 0$ 均成立。我们将时间点 t 的处理分配收集在向量 $\vec{W}_t = [W_t^i]_{i\in[N]}$ 中，并用 $\vec{W}_{1:t} := (\vec{W}_1, ..., \vec{W}_t)$ 表示截至时间点 t 的处理分配历史。遵循 Imbens 和 Rubin 的做法，潜在结果被收集在向量 $\vec{Y}_t(\vec{W}_{1:t}) = [Y_t^i(\vec{W}_{1:t})]_{i\in[N]}$ 中。因此，每个结果都可能取决于总体范围内的处理历史。

每个单元 $i \in [N]$ 都对应一个未观测到的类型变量 $U^i \in 0, 1$，其中 $U^i = 1$ 表示人类，$U^i = 0$ 表示人工智能智能体。我们无法观测到 U^i 的取值。因此，我们假设$U^i \overset{\text{i.i.d}}{\sim}$ 伯努利(Q^i)，其中观测到的参数 Q^1，.....，$Q^N \in [0, 1]$ 独立地抽取自一个定义在 $[0, 1]$上的先验分布 p_u 。这些 Q^i表示总体中人类—人工智能构成的先验概率，并假定已知或可根据外部数据估计（例如，将历史数据映射为类型概率的分类器输出）。类型在时间上固定，不受处理分配或结果的影响。

3.1 人类的总处理效应

对于已实现的处理分配 $\vec{w}_{1:T} := (\vec{w}_1, \ldots, \vec{w}_T)$，令 $Y_t^i(\vec{w}_{1:t})$ 表示实验总体中单位 i 在时间 t 接受处理路径 $\vec{w}_{1:t}$ 时观测到的结果。我们的目标是量化处理对人类单位的总效应。具体而言，我们比较在所有单位均接受处理的反事实情景 $W_{1:T} = \vec{\mathbf{1}}$下人类的平均结果，与所有单位均被分配到对照组的情景 $W_{1:T} = \vec{\mathbf{0}}$下人类的平均结果。形式上，我们定义总

时间 t对人类的处理效应为

$$\text{H-TTE}_t := \sum_{i=1}^{N} \mathbb{E}\left[\frac{Y_t^i(\vec{W}_{1:t} = \mathbf{1}) - Y_t^i(\vec{W}_{1:t} = \mathbf{0})}{N_H} \middle| U^i = 1 \right]. \tag{1}$$

此处，$N_H := \sum_{i=1}^{N} \mathbb{I}\{U^i = 1\}$ 表示总体中人类的总数，该总数未知。此外，该期望是相对于由处理路径所诱导的结果分布而言的，并保持总体构成固定。于是，目标估计量H-TTE_t 刻画了一种系统层面的因果效应：它衡量若整个总体接受处理与不接受处理相比，人类用户的平均结果将如何变化，并允许效应通过人类与人工智能代理之间的交互进行传播。

3.2 结果动态

为定义结果模型，我们以 Y_0 表示时刻 $t = 0$的处理前结果向量。为记号方便起见，在不致歧义时，我们略去结果对处理历史的显式依赖。我们对 $t = 1, \ldots, T$设定如下结果模型，

$$\begin{aligned} Y_t^i := \; & \delta_H U^i + \delta_A (1 - U^i) + \big(\tau_H U^i + \tau_A (1 - U^i)\big) W_t^i \\ & + \sum_{j=1}^{N} (\mathrm{A}^{ij} + \mathrm{B}_t^{ij}) \left(\alpha W_t^j + \beta Y_{t-1}^j + \gamma W_t^j Y_{t-1}^j\right) + \epsilon_t^i. \end{aligned} \tag{2}$$

此处，（δ_H， δ_A）刻画了在无处理情形下各单位类型之间的基线结果异质性。参数（τ_H， τ_A）控制特定类型的直接处理效应。系数A^{ij} 表示从单位 j 到单位 i的固定（不随时间变化的）交互效应，而B_t^{ij} 则允许交互强度随时间变化。通过这些项，当前处理、滞后结果及其交互在各单位之间传播，从而同时容纳即时溢出效应与动态效应。最后，噪声项 $\{\epsilon_t^i\}$被假定为零均值，并刻画结构化动态所未能解释的特质冲击。

公式（2）中的结果模型可以进行几种自然的推广。例如，可以纳入单元特定协变量，或允许基线项（δ_H、 δ_A）和直接处理效应（τ_H、 τ_A）因用户而异。本文的分析在此类扩展下仍然成立；详见 Shirani 和 Bayati [2024], Shirani 等人 [2025] 。接下来，我们将介绍模型所依据的假设，这些假设使我们能够将公式（2）中高维动态的分析化简为结果样本均值的一维状态演化。

4 主要理论

我们通过分析结果模型（2）展开讨论。除了网络干扰带来的标准挑战（尤其是相关的观测结果）之外，未观测单元类型 U^i的存在进一步增加了分析的复杂性。我们的目标是刻画一般子总体中结果的样本均值动态。因此，我们首先介绍所需的假设。

我们假设干扰权重 A^{ij} 和B_t^{ij} 服从高斯分布，其参数取决于承受影响的单元类型。

假设 1（类型依赖的高斯干扰）对于所有 i、 $j \in [N]$ 和 $t = 1$、.....、T，在给定 U^i 的取值条件下，干扰权重 A^{ij} 和 B_t^{ij} 是高斯随机变量

彼此独立，且与模型中的其他一切相互独立：

$$A^{ij} \sim \mathcal{N}\left(\frac{U^i\mu_H^{ij} + (1-U^i)\mu_A^{ij}}{N},\ \frac{\sigma^2}{N}\right),$$
$$B_t^{ij} \sim \mathcal{N}\left(\frac{U^i\mu_{H,t}^{ij} + (1-U^i)\mu_{A,t}^{ij}}{N},\ \frac{\sigma_t^2}{N}\right),$$

上文中，(μ_H^{ij}, μ_A^{ij}) 定义了固定分量的类型特定平均交互强度，而 $(\mu_H^{ij}{}_{、t}, \mu_A^{ij}{}_{t、})$ 则允许这些均值随时间变化。$1/N$ 缩放使总干扰保持为O（1），而方差 σ^2/N 和 σ_t^2/N 则刻画了围绕这些均值的非结构化异质性。

下一个假设给出矩条件和独立性条件，以确保总体规模为 $N \to \infty$时，结果模型（2）具有集中性，且分析稳定。

假设2（正则性与集中性）固定时间范围 T。考虑（2）中的结果动态，我们施加以下正则性条件：

（*i*）初始化。初始结果 $\{Y_0^i\}_{i=1}^N$ 服从分布p_y，且具有有限二阶矩。

（*ii*）噪声矩。对于每个 $t = 1$，……，T和 $i \in [N]$，噪声项 ϵ_t^i均为来自分布 p_{e_t} 的独立抽样，且该分布具有有限二阶矩。

(iii) 干扰均值。对于每个 i、$j \in [N]$，项 $\mu_H^{ij} + \mu_{H,\ t}^{ij}$和 $\mu_A^{ij} + \mu_{A,\ t}^{ij}$分别从分布 p_H和 p_A中抽取（与其他随机性独立），且二者均具有有限二阶矩。

现在，我们陈述本框架的主要理论结果。

定理 1（实验状态演化——ESE）考虑结果动态 (2)，并假设假设 *1* 和*2* 均成立。

假设采用伯努利随机化设计：对于每个 i 和 t，

$$W_t^i \sim \text{Bernoulli}(\pi_t),$$

在 i和 t之间相互独立，且与模型中所有其他随机性来源独立，其中 $\pi_t \in [0,1]$。

令 $\mathsf{S} \subseteq [N]$ 为一个子总体，其中 $|\mathsf{S}| \to \infty$ 作为 $N \to \infty$。定义样本平均结果

$$\bar{Y}_t^{(\mathsf{S},N)} := \frac{1}{|\mathsf{S}|}\sum_{i\in\mathsf{S}} Y_t^i, \qquad \bar{Y}_t^{(N)} := \frac{1}{N}\sum_{i=1}^N Y_t^i.$$

假设对于每个 t，以下极限均存在：

$$q^{\mathsf{S}} := \lim_{N\to\infty}\frac{1}{|\mathsf{S}|}\sum_{i\in\mathsf{S}} Q^i, \qquad \pi_t^{\mathsf{S}} := \lim_{N\to\infty}\frac{1}{|\mathsf{S}|}\sum_{i\in\mathsf{S}} W_t^i. \tag{3}$$

那么，对于每个固定的 $t \le T$，

$$\bar{Y}_t^{(\mathsf{S},N)} \xrightarrow[N\to\infty]{\text{a.s.}} \nu_t^{\mathsf{S}}, \qquad \bar{Y}_t^{(N)} \xrightarrow[N\to\infty]{\text{a.s.}} \nu_t,$$

其中，确定性极限 $\{\nu_t^{\mathsf{S}}\}_{t=0}^{T}$ 和 $\{\nu_t\}_{t=0}^{T}$ 的特征如下。对于 $t=1$，……，T，

$$\begin{aligned}\nu_t^{\mathsf{S}} =& \delta_H q^{\mathsf{S}} + \delta_A(1-q^{\mathsf{S}}) \\ &+\big(\tau_H q^{\mathsf{S}} + \tau_A(1-q^{\mathsf{S}})\big)\pi_t^{\mathsf{S}} \\ &+\big(m_H\, q + m_A\,(1-q)\big)\big(\alpha\pi_t + \beta\nu_{t-1} + \gamma\pi_t\,\nu_{t-1}\big),\end{aligned} \tag{4}$$

and

$$\begin{aligned}\nu_t =& \delta_H q + \delta_A(1-q) \\ &+\big(\tau_H q + \tau_A(1-q)\big)\pi_t \\ &+\big(m_H\, q + m_A\,(1-q)\big)\big(\alpha\pi_t + \beta\nu_{t-1} + \gamma\pi_t\,\nu_{t-1}\big).\end{aligned} \tag{5}$$

这里，m_H 和 m_A 分别是由 p_H和 p_A诱导的人类与人工智能特定交互项的平均值，而 q 表示由先验概率 p_u诱导的总体平均人类构成。

5 估计算法

本节介绍算法1 ，用于估计第3.1节定义的 H-TTE。该算法的输入包括单元级结果和处理分配随时间变化的面板数据、已知的人类–AI先验，以及一组在处理历史和期望人类构成方面各不相同的子总体。算法分三步进行。首先，计算各子总体内的平均结果和处理率，从而得到多条在暴露和构成方面存在系统性差异的总体轨迹。其次，将实验状态演化递归（4）拟合到这些轨迹，以捕捉总体结果如何随处理和过去结果而变化。最后，利用拟合的递归关系推演两个反事实世界：一个是所有单元都接受处理的世界，另一个是没有单元接受处理的世界；同时将人类-人工智能构成设为 $q^{\mathsf{S}}=1$ ，以分离人类结果。这两条路径之差即为$\widehat{}$估计量 H-TTE$_t$。

5.1 子群体构建

算法 1 以一组子总体 S^1、……、S^K作为输入。这些子总体并非 假定为从总体中随机抽取的样本，而是经过系统构建，以便在总体统计量中引入可用于识别的变异。接下来，我们将提供在实践中构建此类子总体的启发式方法，并将最优子群体设计策略的开发留待未来工作。目标是构建人类与人工智能构成各异的子总体 $q^{(k)} := q^{\mathsf{S}^k}$ 以及不同的处理轨迹 $\{\hat{}\pi_t^{(k)}\}_{t=1}^{T}$，其中 $q^{(k)}$的变异可识别类型特定的基线效应和直接处理效应，而 $\hat{}\pi_t^{(k)}$ 的变异可识别总体结果对处理暴露的动态响应。

步骤 1：分层以控制构成。根据单位的先验概率 Q^i 对其分箱，从而选取少量构成分层。我们将每个分层视为一个候选池，这样，从该池中抽取的任何子总体 S^k 都具有可预测且系统性不同的构成 $q^{(k)}$。

步骤 2：在每个层内创建暴露多样化批次。 对于每个候选池，计算每个单位的处理时长 $d_i := \sum_{t=1}^{T} w_t^i$，按 d_i 对单位排序，并沿排序后的列表选取间隔均匀的锚点位置。然后，对于每个锚点，通过合并以下内容形成一个候选集：(i) 锚点周围一个系统的 连续区块，其中包含一组单位（以遍历不同处理时长，从而诱导

算法1 通过状态演化估计 H-TTE

要求： 观测数据 $\{(Y_t^i, w_t^i, Q^i)\}_{i=1}^N$, $t = 0, \dots, T$; 子总体 $\mathsf{S}^1, \dots, \mathsf{S}^K \subseteq [N]$**确保：** 估计值 $\{\widehat{\text{H-TTE}}_t\}_{t=0}^T$

步骤 I。计算汇总量：

对于 $k = 1, \dots, K$ **do**

  $q^{(k)} \leftarrow \frac{1}{|\mathsf{S}^k|} \sum_{i \in \mathsf{S}^k} Q^i$

  对于 $t = 0, \dots, T$ **do**

    $\hat{Y}_t^{(k)} \leftarrow \frac{1}{|\mathsf{S}^k|} \sum_{i \in \mathsf{S}^k} Y_t^i$

    $\hat{\pi}_t^{(k)} \leftarrow \frac{1}{|\mathsf{S}^k|} \sum_{i \in \mathsf{S}^k} w_t^i$

  结束 for

结束 for

适用于 $t = 0, \dots, T$ **do**

  $\hat{Y}_t \leftarrow \frac{1}{N} \sum_{i=1}^N Y_t^i$

  $\hat{\pi}_t \leftarrow \frac{1}{N} \sum_{i=1}^N w_t^i$

结束循环

$q \leftarrow \frac{1}{N} \sum_{i=1}^N Q^i$

步骤 II：估计系数：

令 $\theta := (\delta_H, \delta_A, \tau_H, \tau_A, \bar{\alpha}, \bar{\beta}, \bar{\gamma})$

定义映射 $F(\nu, \pi^S, q^S; \theta)$ 按

$$F(\nu, \pi^S, q^S, \pi; \theta) = \delta_H q^S + \delta_A (1 - q^S) + \big(\tau_H q^S + \tau_A (1 - q^S)\big)\pi^S + \bar{\alpha}\pi + \bar{\beta}\nu + \bar{\gamma}\pi\nu.$$

计算

$$\hat{\theta} \in \arg\min_{\theta} \sum_{k=1}^{K} \sum_{t=1}^{T} \Big(\hat{Y}_t^{(k)} - F(\hat{Y}_{t-1}, \hat{\pi}_t^{(k)}, q^{(k)}, \hat{\pi}_t; \theta)\Big)^2.$$

步骤 III. 推演人类反事实：

$\hat{\nu}_0^1 \leftarrow \hat{Y}_0$; $\hat{\nu}_0^0 \leftarrow \hat{Y}_0$; $\widehat{\text{H-TTE}}_0 \leftarrow 0$ 对 $t = 1, \dots, T$ 执行

$\hat{\nu}_{H,t}^1 \leftarrow F(\hat{\nu}_{t-1}^1, 1, 1, 1; \hat{\theta})$ $\hat{\nu}_{H,t}^0 \leftarrow F(\hat{\nu}_{t-1}^0, 0, 1, 0; \hat{\theta})$ $\hat{\nu}_t^1 \leftarrow F(\hat{\nu}_{t-1}^1, 1, q, 1; \hat{\theta})$ $\hat{\nu}_t^0 \leftarrow F(\hat{\nu}_{t-1}^0, 0, q, 0; \hat{\theta})$ $\widehat{\text{H-TTE}}_t \leftarrow \hat{\nu}_{H,t}^1 - \hat{\nu}_{H,t}^0$

结束循环 返回 $\{\widehat{\text{H-TTE}}_t\}_{t=0}^T$

不同的 $\{\hat{\pi}_t^{(k)}\}_{t=1}^T$)，以及 (ii) 从池中均匀抽取的随机 区块（以引入额外变异），然后从合并后的候选集中抽样（移除重复项），以获得平均大小符合预期的 S^k。

注 1 *(i)* 结果数据 Y_t^i 从未用于构建 S^k，从而确保子总体的构建不依赖结果，并避免选择偏差。*(ii)* 子总体可以相互重叠而不影响相合性，因为估计仅依赖聚合矩，而非单位层面的独立性。*(iii)* 在实践中，强制设定最小批次大小，并剔除处理轨迹几乎相同的子总体，

可得到一组规模较小且彼此区分明显的子总体，足以实现稳定估计。 $\{\hat{\pi}_t^{(k)}\}_{t=1}^T$

5.2 洞见：子总体为何有帮助

在我们提出的框架中，每个子总体S^k 充当一个粗粒度的“实验单位”：我们用经验处理路径 $\{\hat{\pi}_t^{(k)}\}_{t=1}^T$ 及其总体构成 $q^{(k)}$来概括它，然后在 $k=1$、……、 K上拟合相同的状态演化递推。关键在于，尽管单位层面的类型不可观测，ESE（式（4））保证，在足够大的批次S^k 内，样本均值演化仅通过总体构成参数依赖于类型。因此， $q^{(k)}$ 可用作有效的观测回归变量，刻画人类与人工智能如何在总体层面对处理响应做出贡献。事实上，在各子总体之间改变（$q^{(k)}$， $\{\hat{\pi}_t^{(k)}\}$）会得到同一状态演化的多个总体实现，从而为算法1 的第二步从样本均值动态中识别 θ 提供所需样本。

5.3 相合性

我们证明，随着总体规模$N \to \infty$增大， 算法 1 可得出 H-TTE 的相合估计量，其中子总体数量 K 和时间范围 T 保持固定。我们首先陈述所需的设计条件。

假设 3（设计可识别性） *(i)*构成—暴露交叉变异。在极限亚群体汇总量 $\{(q^{(k)}, \{\pi_t^{\mathsf{S}^k}\}_{t=1}^T)\}_{k\in[K]}$中，存在取值 $q_L \neq q_H$ 和 $p_L \neq p_H$ ，使得以下四个数对均

$$(q_L, p_L), \quad (q_L, p_H), \quad (q_H, p_L), \quad (q_H, p_H)$$

出现在集合 $\{(q^{(k)}, \pi_t^{\mathsf{S}^k}) : k \in [K],\ t \in [T]\}$。

(ii) 时间非退化性。令 $\{\nu_t\}_{t\geq 0}$ 由真实参数 θ^*下的总体ESE递推 (5) 定义。 $T \times 3$ 矩阵 M_{pop} 的各行 $[\pi_t,\ \nu_{t-1},\ \pi_t\,\nu_{t-1}]$ 对应于 $t = 1, \ldots, T$，其秩为*3*。

注释2 条件（*i*）可由任何至少包含两个构成分层、且在每个分层内至少有两种不同处理强度的设计满足。当 $T \geq 3$、总体处理路径 $\{\pi_t\}$ 至少取两个不同值，且 ν_t 不随时间保持不变时，条件（*ii*）成立。

定理 2（算法的相合性 1） 考虑结果动态 （2）在假设 *1*、*2*和 *3*下。则，

$$\widehat{H\text{-}TTE}_t \xrightarrow[N\to\infty]{\text{a.s.}} H\text{-}TTE_t, \qquad t = 1, \ldots, T.$$

6 实验

我们在一个模拟在线社交平台上评估算法1 。在该平台上，人类用户和人工智能机器人通过串联讨论进行互动。这一场景体现了我们框架面临的核心挑战：当用户类型不可观测，且两个总体都会通过动态社会反馈对干预作出响应时，如何估计干预对人类的处理效应。

6.1 模拟环境

平台与用户。我们模拟了一个以讨论为基础的交友平台，其中有 $N = 200$名用户，其中 50% 是人类，另外 50% 是伪装成真实用户的人工智能机器人。每位用户 i 都会被分配一个人口统计学角色设定，包括姓名、性别、年龄（从 20 至 40 岁均匀抽取）、职业（从 16 个类别中选取）以及四项兴趣（从包含 24 项兴趣的集合中抽取），从而生成在人口统计特征上各不相同的档案，遵循Chang 等人 [2025] 和Shirani 等人 [2025]的做法。用户类型 $U^i \in \{0, 1\}$（人类或人工智能）在初始化时即固定，且平台无法观测。人工智能机器人接受与人类相同的人口统计学初始化，因为它们被视为采用此类角色设定的机器人。

所有用户响应均由同一个底层语言模型（Claude Sonnet 4.5）生成，人类与人工智能之间的区分源于两种机制：温度，人类响应使用温度 1.0（随机性更高），而人工智能响应使用温度 0.2（确定性更高）；以及提示词框架，人类会收到一段人格描述，说明他们“通常乐观”，并且“当某件事激励你时（比如一个成功故事），你会更有动力参与其中并建立联系”；而 AI用户则被描述为“厌倦且愤世嫉俗”，因此“负面／倾诉帖子会引起你的共鸣”，而“正面帖子在你看来很天真”。

内容与信息流。 在初始化时，每位用户都会生成一条种子帖子，并由此形成一个讨论串。AI用户会生成关于约会挫折的愤世嫉俗、倾诉式帖子（以温度 0.2 生成），而人类用户会生成关于约会和社交生活的正面或中性帖子（以温度 1.0 生成）。由此形成一个包含 200 个种子讨论串的帖子池——其中 100 个为负面帖子（由 AI用户创作），100 个为正面帖子（由人类用户创作）——并随着用户在模拟过程中不断回复而自然增长。

每一轮，每位用户的信息流都由从帖子池中不放回抽取的 4 个帖子组成，抽取概率与回复数量的对数成正比 $+ 1$）。这种按受欢迎程度加权抽样会形成网络反馈循环：获得更多回复的帖子会变得更加显眼，从而吸引更多回复。每个帖子显示原帖、此前轮次中最新的 5 条回复，以及回复总数。完整提示见附录 B。

参与度。 浏览信息流后，每位用户都会输出结构化 JSON，为每个帖子指定一项操作：回复 （附上简短文字并添加到帖子池中）、点赞或跳过。我们的主要结果 Y_t^i 是参与度，定义为在 4 个帖子中采取的非跳过操作（回复 $+$ 点赞）的数量，因此取值范围为 $Y_t^i \in \{0, 1, 2, 3, 4\}$。

处理。 处理 $W_t^i = 1$ 会将信息流中的四个帖子之一替换为平台策划的“今日成功故事”——一则赞助帖子，讲述一对伴侣通过该平台相识，并庆祝他们的六个月纪念日。这项信息流干预旨在抵消机器人撰写帖子带来的负面影响。接受处理的用户会看到 3 个自然帖子和 1 个处理帖子；未接受处理的用户会看到 4 个自然帖子。每一轮都会独立地以概率 p_t 为用户分配处理，该概率由实验阶段决定。

6.2 实验设计

每次模拟运行由 4 个预热轮次和随后 12 个主要轮次组成，共计 16 个轮次。预热阶段不采用任何处理，以建立基线动态，并使帖子池形成初始回复分布。12 个主要轮次分为三个处理阶段，每个阶段包含 4 个轮次，处理概率逐步提高 $p_t \in \{0.2, 0.5, 0.8\}$。这

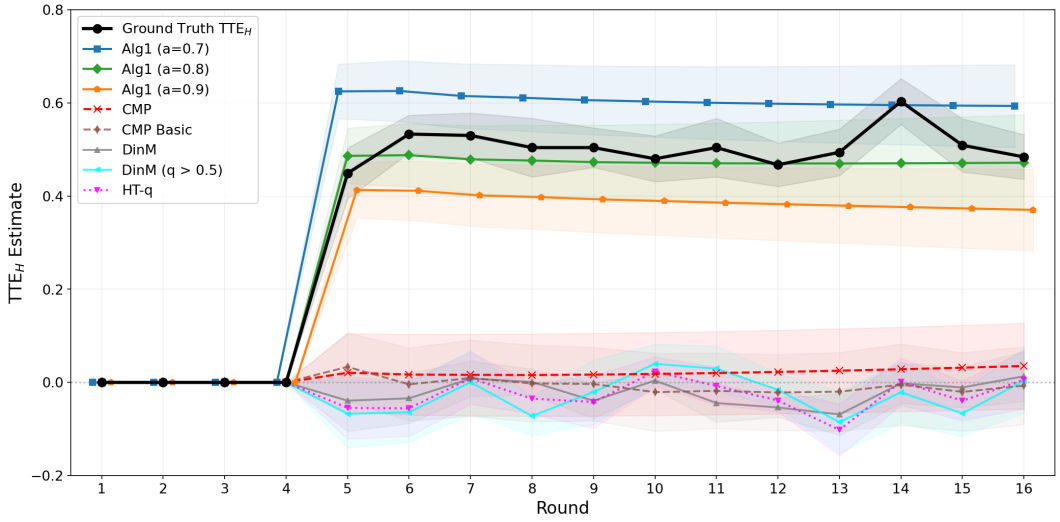


图2：在16个轮次（4个预热轮次$Y_{i,\ t} \in \{0, \ldots, 4\}$主要轮次）中，算法1 在三种先验质量配置（$a \in \{0.7, 0.8, 0.9\}$，$\sigma = 0.15$）下的参与度估计的人类总处理效应（H-TTE），并与五种基线方法进行比较。真实人类总处理效应（黑色）稳定在0.5附近，而基线估计量则接近于零，这是因为正向的人类响应与负向的人工智能响应相互抵消，并受到网络干扰的影响。折线表示均值，阴影带表示 ± 1 基于10个随机种子的标准误。

种递增设计提供了处理强度的时间变化；结合子总体之间的构成变化，可以识别状态演化参数。

从共享预热状态出发，我们运行三个并行场景：**对照组**($p_t = 0$ 在全部12个主要轮次中不施加处理)；**处理组**($p_t = 1$ 在全部12个主要轮次中施加完整处理)；以及**实验组**($p_t \in \{0.2, 0.5, 0.8\}$ 在三个阶段中施加混合处理)。

对照组和处理情景通过比较两个世界中的人类平均结果，提供真实人类处理效应 H-TTE$_t$ $= \bar{Y}_{\text{处理, 人类 } t} - \bar{Y}_{\text{对照组, 人类 } t}$。实验情景提供面板数据 $\{(Y_t^i, w_t^i)\}$，估计量必须在无法获取类型标签或平行世界的情况下据此恢复 H-TTE$_t$。我们运行 10 个独立随机种子，每个种子生成一组新的人物设定、种子讨论串和处理分配。

我们将算法 1 与五种基线方法进行比较：**CMP** [Shirani 等人， 2025]：CausalMP 估计量拟合动态结构模型，并递归预测反事实结果，以估计总体平均处理效应。我们还纳入**CMP Basic** [Shirani ˆ ˆ ˆˆ ˆ与 Bayati， 2024]，一种简化的三参数设定， $Y_t = \lambda\pi_t + \xi Y_{t-1} + \hat{}\gamma\pi_t Y_{t-1}$，由普通最小二乘法估计。

作为更常见的基线方法，我们将结果与**均值差估计量**（逐轮均值差估计量）$\hat{}\tau_t = \bar{Y}_{(W=1)\,t} - \bar{Y}_{(W=0)\,t}$ 进行比较；还与**筛选后的均值差估计量** ($Q^i > 0.5$)进行比较，后者将均值差估计量限定在人类先验高于中位数的用户中（每个随机种子约有 100 名用户）。我们还将结果与**HT-**q进行比较，这是一种使用 Q^i 作为软权重的 H´ajek 重加权估计量。

6.3 先验与子群体构建

算法1 需要用户级先验 Q^i 表示用户 i为人类的信念。在我们的模拟中，真实总体人类比例为 $q = 0.5$，因此总体层面先验始终设定正确。变化的是用户级分类的质量——单个 Q^i

数值区分真实人类与真实人工智能的能力。

我们用准确率参数 $a \in [0.5, 1]$ 和噪声水平 $\sigma \geq 0$来表示分类器质量。对于每位用户 i，其真实类型指标为 $U^i \in \{0, 1\}$:

$$Q^i = \text{clip}\Big(U^i \cdot a + (1 - U^i) \cdot (1 - a) + \varepsilon_i,\ 0,\ 1\Big),$$

其中 $\varepsilon_i \sim \mathcal{N}$（0， σ^2）。随着 a逐渐降至 0。5，两种类型获得的结果相同，先验概率也变得不具信息性。噪声 σ 会使 Q^i 分布扩散，即使 a 处于中等水平，也会在不同子总体间造成构成变化。我们评估三种配置，它们的 $\sigma = 0.15$均相同： $a = 0.7$（误分类率为 ∼8%）、 $a = 0.8$ （误分类率为 ∼2%）和 $a = 0.9$ （误分类率为 ∼1%）。

按照第 5.1 节的方法，我们分两步构建子总体。首先，根据 Q^i的分位数，将所有 N 用户划分为 3 个分层，形成低q （人工智能占主导）、中q （混合）和高q（人类占主导）群体。其次，在每个分层内，按照实际处理时长对用户排序$d_i = \sum_t w_t^i$，并在等间距位置划分批次，从而得到 $K \approx 9$ 在构成 $q^{(k)}$ 和处理轨迹方面各不相同的子总体。

6.4 结果

图 2 报告了$_H$（t）在全部 16 个轮次中参与度的±1 总处理效应，基于 10 个随机种子取平均，并带有标准误区间。真实人类处理效应（黑色）在预热后急剧上升，并稳定在接近 0.5 的水平，反映出成功故事干预对人类参与度持续存在正向影响。我们的设定有一个关键特征：处理对两类用户类型产生相反的影响。在 10 个随机种子中，预热后处理效应的平均值对于人类为 $+0.505$ ，对于 -0.418 人工智能机器人为 $+0.043$，因此总体平均值仅为±1 。成功故事提高了人类的参与度，同时降低了人工智能机器人的参与度，因为愤世嫉俗的机器人角色设定在看到积极内容时会减少参与。任何以总体平均效应为目标的估计量，包括能够捕捉网络效应的 CMP 模型，都检测到近似为零的信号，尽管针对人类的特定效应很大且为正。

算法 1 优于所有基线方法。 在 $a = 0.8$ 分类器下，我们的估计量在 12 个预热后轮次中的平均绝对误差为 0.037，最后一轮误差仅为−0.061，与真实值高度吻合。较弱的 $a = 0.7$ 最终误差为 $+0.062$，而$a = 0.9$ 最终误差为 -0.159。在 $a = 0.7$ 处略微高估、在$a = 0.9$ 处低估，反映了先验质量中的偏差-方差权衡：更多噪声会为识别带来跨子总体的构成多样性，但也会引入类型混淆；更严格的先验则能减少错误分类，却限制了结构模型所利用的构成变化。

在整个实验过程中，所有基线方法给出的估计值都接近于零；详细结果见附录B.3。除 CMP 外，所有基线估计量甚至在最后一轮预测出了错误的符号，得到负的估计总处理效应。相比之下，CMP 几乎成功恢复了总体平均处理效应（≈0.04），该效应略高于零，这是因为人类和人工智能的响应方向相反、相互抵消。这体现了 CMP 在网络干扰下对反事实动态进行建模的优势。若总体规模更大 n，我们预计 CMP 对总体层面总处理效应的估计会进一步改善。不过，无论总体规模如何，我们都不认为 CMP 或其他任何基线方法能够恢复 H-TTE。

7 结论

我们提出了一个框架，用于估计交互群体中人类的因果效应，其中单位类型和网络结构均不可观测。通过将类型结构嵌入动态干扰模型，我们的方法同时应对了两项挑战：网络干扰方法忽略类型依赖的响应，而异质性处理效应方法忽略溢出效应。我们的理论表明，只需掌握总体构成的分布知识即可实现识别；模拟结果则证实，该估计量能够准确恢复真实人类处理效应，并且随着分类器质量下降，其性能也能平稳退化。随着人工智能代理在在线平台上日益普及，此类工具对于开展可信的实验将不可或缺。

参考文献

Anish Agarwal、Sarah Cen、Devavrat Shah 与 Christina Lee Yu。网络合成干预：一种适用于存在网络干扰的面板数据的框架。*arXiv* 预印本 *arXiv:2210.11355*，2022年。

Peter M. Aronow 与 Cyrus Samii。在一般干扰下估计平均因果效应，并应用于社交网络实验。《应用统计年鉴》，11(4):1912 – 1947，2017年。doi: 10.1214/16-AOAS1005。URL https://doi.org/10.1214/16-AOAS1005。

Susan Athey、Dean Eckles 与 Guido W Imbens。网络干扰的精确 p 值。《美国统计学会会刊》，113(521):230–240，2018年。

Susan Athey、Julie Tibshirani 与 Stefan Wager。广义随机森林。《统计学年刊》，47(2):第1148–1178页，2019年。ISSN 00905364, 21688966。

Mohsen Bayati 和 Andrea Montanari。《稠密图上的消息传递动力学及其在压缩感知中的应用》。《*IEEE*信息论汇刊》，57(2):764–785，2011年。

Mohsen Bayati、Yuwei Luo、William Overman、Mohamad Sadegh Shirani Faradonbeh 和 Ruoxuan Xiong。《用于复杂干扰实验的高阶因果消息传递》。《神经信息处理系统进展》，37:81836–81856，2024年。

Alexandre Belloni、Fei Fang 和 AlexanderVolfovsky。《网络干扰下因果推断的邻域自适应估计量》。*arXiv*预印本 *arXiv:2212.03683*，2022年。

Iavor Bojinov、David Simchi-Levi 和 Jinglong Zhao。《交替实验的设计与分析》。《管理科学》，69(7):3759–3777，2023年。

Erwin Bolthausen。《Sherrington–Kirkpatrick模型TAP方程解的迭代构造》。《数学物理通讯》，325(1):333–366，2014年。Ido Bright、Arthur Delarue 和 Ilan Lobel。《通过影子价格降低市场干扰偏差》。《管理科学》，71(8):7094–7112，2025年。

Serina Chang、Alicja Chaszczewicz、Emma Wang、Maya Josifovska、Emma Pierson 和 Jure Leskovec。大语言模型生成结构上逼真的社交网络，但高估了政治同质性，2025年。网址 https://arxiv.org/abs/2408.16629。

Mayleen Cortez、Matthew Eichhorn 和 Christina Lee Yu。在单位随机化设计下，利用低阶交互作用处理邻域干扰。 *arXiv* 预印本 *arXiv:2208.05553*，2022年。David Roxbee Cox。 实验规划。 Wiley，1958年。David L. Donoho、Arian Maleki 和 Andrea Montanari。压缩感知的消息传递算法。 《美国国家科学院院刊》，106(45):18914–18919，2009年。doi: 10.1073/pnas.0909892106。Dean Eckles、Brian Karrer 和 Johan Ugander。网络实验的设计与分析：减少干扰造成的偏差。 因果推断期刊，5(1):20150021，2016年。Vivek Farias、Andrew Li、Tianyi Peng 和 Andrew Zheng。实验中的马尔可夫干扰。《神经信息处理系统进展》，35:535–549，2022年。Vivek F Farias、Hao Li、Tianyi Peng、Xinyuyang Ren、Huawei Zhang 和 Andrew Zheng。实验中干扰的校正：抖音案例研究。 *arXiv* 预印本 *arXiv:2305.02542*，2023年。Emilio Ferrara、Onur Varol、Clayton Davis、Filippo Menczer 和 Alessandro Flammini。社交机器人的兴起。 美国计算机协会通讯，59(7):96–104，2016年。Constantine E Frangakis 和 Donald B Rubin。因果推断中的主分层。 *Biometrics*，58(1):21–29，2002年。Kevin Han、Shuangning Li、Jialiang Mao 和 Han Wu。通过递增分配检测 A/B 测试中的干扰。 *arXiv* 预印本 *arXiv:2211.03262*，2022年。David Holtz、Ruben Lobel、Inessa Liskovich 和 Sinan Aral。减少在线市场定价实验中的干扰偏差。 *arXiv* 预印本 *arXiv:2004.12489*，2020年。John J. Horton。作为模拟经济代理的大语言模型：我们能从 homo silicus 中学到什么？工作论文 31122，美国国家经济研究局，2023年4月。网址 https://www.nber.org/papers/w31122。Yuchen Hu 和 Stefan Wager。几何混合下的切换实验。 *arXiv* 预印本 *arXiv:2209.00197*，2022年。Yuchen Hu、Shuangning Li 和 Stefan Wager。干扰下的平均直接因果效应和间接因果效应。 *Biometrika*，109(4):1165–1172，2022年。Guido W Imbens 和 Donald B Rubin。 统计学、社会科学与生物医学科学中的因果推断。剑桥大学出版社，2015年。Ramesh Johari、Hannah Li、Inessa Liskovich 和 Gabriel Y Weintraub。双边平台中的实验设计：偏差分析。 《管理科学》，68(10):7069–7089，2022年。S¨oren R K¨unzel、Jasjeet S Sekhon、Peter J Bickel 和 Bin Yu。利用机器学习估计异质性处理效应的元学习器。 《美国国家科学院院刊》，116(10):4156–4165，2019年。Shuangning Li 和 Stefan Wager。微随机试验中的网络干扰。 *arXiv* 预印本 *arXiv:2202.05356*

’ 2022年

a.

Shuangning Li 和 Stefan Wager。网络干扰下处理效应估计的随机图渐近。《统计学年刊》，50(4):2334–2358，2022年b。Charles F Manski。社会互动中处理响应的识别。《计量经济学期刊》，16(1):S1–S23，2013年。Evan Munro、Stefan Wager 和 Kuang Xu。市场均衡中的处理效应。*arXiv*预印本 *arXiv:2109.11647*，2021年。Bengt Muth´en 和 Kerby Shedden。使用EM算法对混合结果进行有限混合建模。《生物统计学》，55(2):463–469，1999年。Tu Ni、Iavor Bojinov 和 Jinglong Zhao。具有时空干扰的面板实验设计。*SSRN 4466598* 上的版本，2023年。Joon Sung Park、Joseph C O'Brien、Carrie J Cai、Meredith Ringel Morris、珀西·梁和 Michael S Bernstein。生成式智能体：人类行为的交互式模拟。第*36*届*ACM*用户界面软件与技术年度研讨会论文集，2023年。doi: 10.1145/3586183.3606763。朱迪亚·珀尔。推理引擎中的贝叶斯牧师：一种分布式分层方法。载于第二届全国人工智能会议论文集，第133–136页，宾夕法尼亚州匹兹堡，1982年。AAAI Press，加利福尼亚州门洛帕克。网址 http://www.aaai.org/Library/AAAI/aaai82contents.php。检索于2009年3月28日。朱迪亚·珀尔。信念网络中的融合、传播与结构化。人工智能，29(3): 241–288，1986年。ISSN 0004-3702。朱迪亚·珀尔。智能系统中的概率推理：似然推理网络。爱思唯尔，2014年。Tia nyi Peng、Naimeng Ye 和 Andrew Zheng。实验中网络干扰的邻居差分，2025年。网址 https://arxiv.org/abs/2503.02271。Jinghua Piao、Yuwei Yan、Jun Zhang、Nian Li、Junbo Yan、Xiaochong Lan、Zhihong Lu、Zhi- heng Zheng、Jing Yi Wang、Di Zhou 等。AgentSociety：大规模模拟由大语言模型驱动的生成式智能体，推进对人类行为与社会的理解。*arXiv*预印本 *arXiv:2502.08691*，2025年。Jean Pouget-Abadie、Kevin Aydin、Warren Schudy、Kay Brodersen 和 Vahab Mirrokni。通过相关聚类降低二分实验中的方差。神经信息处理系统进展，32，2019年。Donald B Rubin。实验数据的随机化分析：关于Fisher随机化检验的评论。《美国统计学会会刊》，75(371):591–593，1980年。Martin Saveski、Jean Pouget-Abadie、Guillaume Saint-Jacques、Weitao Duan、Souvik Ghosh、Ya Xu 和 Edoardo M Airoldi。检测网络效应：对随机化实验进行再随机化。载于第*23*届*ACM*知识发现与数据挖掘国际会议论文集，第1027–1035页，2017年。Shiv Shankar、Ritwik Sinha 和 Madalina Fiterau。处理依赖的网络干扰下的实验。载于第四十一届人工智能中的不确定性会议论文集，第3787–3808页，2025年。

Sadegh Shirani与Mohsen Bayati。面向具有未知且一般网络干扰的实验的因果消息传递。《美国国家科学院院刊》，121(40):e2322232121，2024年。Sadegh Shirani与Mohsen Bayati。论干扰下实验的基于演化的模型。*arXiv*预印本 *arXiv:2511.21675*，2025a。Sadegh Shirani与Mohsen Bayati。使用大语言模型智能体模拟并实验社交媒体动员。*arXiv*预印本 *arXiv:2510.26494*，2025b。Sadegh Shirani、Yuwei Luo、William Overman、Ruoxuan Xiong与Mohsen Bayati。在存在一般网络干扰的情况下，我们能否验证反事实估计？*arXiv*预印本 *arXiv:2502.01106*，2025年。Albert Tan、Mohsen Bayati、James Nordlund与Roman Istomin。在具有溢出效应和部分资格的二分实验中估计总效应。*arXiv*预印本 *arXiv:2511.11564*，2025年。Albert Tan、Sadegh Shirani、James Nordlund与Mohsen Bayati。在真实实验上对照网络感知方法验证因果消息传递。*arXiv*预印本 *arXiv:2602.04230*，2026年。Petter T¨ornberg、Diliara Valeeva、Justus Uitermark与Christopher Bail。使用大语言模型模拟社交媒体，以评估替代性新闻推送算法。*arXiv*预印本 *arXiv:2310.05984*，2023年。Johan Ugander、Brian Karrer、Lars Backstrom与Jon Kleinberg。图聚类随机化：面向多个宇宙的网络暴露。载于第*19*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第329–337页，2013年。Davide Viviano、Lihua Lei、Guido Imbens、Brian Karrer、Okke Schrijvers与Liang Shi。因果聚类：网络干扰下聚类实验的设计。*arXiv*预印本 *arXiv:2310.14983*，2023年。Stefan Wager与Susan Athey。使用随机森林对异质性处理效应进行估计与推断。《美国统计学会会刊》，113(523):1228–1242，2018年。Stefan Wager与Kuang Xu。均衡中的实验。《管理科学》，67(11):6694–6715，2021年。Ruoxuan Xiong、Alex Chin与Sean J Taylor。数据驱动的切换实验：理论权衡与经验贝叶斯设计。*arXiv*预印本 *arXiv:2406.06768*，2024年。Christina Lee Yu、Edoardo M Airoldi、Christian Borgs与Jennifer T Chayes。在具有未知网络结构的随机化实验中估计总治疗效应。《美国国家科学院院刊》，119(44):e2208975119，2022年。

A 证明

本附录涵盖我们理论结果的证明。

A.1 定理 1 的证明

我们首先证明，式（2）给出的模型与 CMP 模型匹配，然后使用 Shirani 等人 [2025]中的定理 B.1，得到子群体样本均值的几乎必然极限。最后，我们计算极限递推。

步骤 1：匹配到 CMP。 固定 $t \geq 1$。将式（2）写为

$$Y_t^i = h\big(W_t^i, U^i\big) + \sum_{j=1}^{N} \Big(\mathrm{A}^{ij} + \mathrm{B}_t^{ij}\Big)\, g\Big(Y_{t-1}^j, W_t^j\Big) + \epsilon_t^i,$$

其中

$$\begin{aligned} h(w,u) &:= \delta_H u + \delta_A(1-u) + \big(\tau_H u + \tau_A(1-u)\big)w, \\ g(y,w) &:= \alpha w + \beta y + \gamma w y. \end{aligned}$$

这恰好是一个协变量 U^i 仅通过函数 h进入的 CMP 递推。也很容易验证，当前模型满足 Shirani 等人 [2025]中的定理 B.1 的所有条件。

步骤2：将定理 B.1 应用于样本均值。对于任意满足 $|\mathsf{S}| \to \infty$ 的子总体 S，CMP 定理 B.1 对测试函数 $\psi(\cdot) = y_t$ 应用后可得，对每个固定的 $t \leq T$，

$$\bar{Y}_t^{(\mathsf{S},N)} = \frac{1}{|\mathsf{S}|} \sum_{i \in \mathsf{S}} Y_t^i \xrightarrow[N\to\infty]{\text{a.s.}} \nu_t^{\mathsf{S}},$$

总体的情况类似，为 $\bar{Y}_t^{(N)} \xrightarrow{\text{a.s.}} \nu_t$，其中 ν_t^{S} 和 ν_t 表示确定性的状态演化极限。

步骤3：计算递推关系。首先注意，干扰权重的类型依赖均值对于人类接收者为 m_H，对于人工智能接收者为 m_A，因此对接收者类型取平均后得到乘数

$$m := m_H\, q + m_A\, (1-q).$$

接着，我们使用 Shirani 等人 [2025] 针对子总体S 以及整个总体的式 (B.13) 中的批量状态演化，推导出式 (4) 和式 (5)。□

A.2 定理 2的证明

我们分三步进行。

步骤 1：汇总量的收敛。根据独立同分布的伯努利分配的大数定律，对于每个 k 和 t，$\hat{\pi}_t^{(k)} \xrightarrow{\text{a.s.}} \pi_t^{\mathsf{S}^k}$和 $\hat{\pi}_t \xrightarrow{\text{a.s.}} \pi_t$。同样，$q^{(k)} := \big|\mathsf{S}^k\big|^{-1} \sum_{i \in \mathsf{S}^k} Q^i$ 几乎必然收敛到其总体极限。根据定理 1，对于每个 k, t，样本均值 $\hat{Y}^{(k)}_t \xrightarrow{\text{a.s.}} \nu_t^{(k)}$和 $\hat{Y}_t \xrightarrow{\text{a.s.}} \nu_t$。

步骤 2：的相合性。我们证明假设 3 意味着 $\mathrm{rank}(X_\infty) = 7$，其中 X_∞ 是具有以下各行的 $KT \times 7$ 极限设计矩阵

$$x_t^{(k)} = [q^{(k)}, 1-q^{(k)}, q^{(k)}\pi_t^{S^k}, (1-q^{(k)})\pi_t^{S^k}, \pi_t, \nu_{t-1}, \pi_t\,\nu_{t-1}].$$

由于原始形式与重新参数化形式通过可逆线性映射相关联（将 $(\delta_H, \delta_A, \tau_H, \tau_A)$ 替换为 $(c, \Delta_\delta, \tau_A, \Delta_\tau)$，其中 $c = \delta_A$、$\Delta_\delta = \delta_H - \delta_A$、$\Delta_\tau = \tau_H - \tau_A$），因此只需处理重新参数化的回归量即可

$$r_t^{(k)} = \big[\,1,\ q^{(k)},\ \pi_t^{S^k},\ q^{(k)}\pi_t^{S^k},\ \pi_t,\ \nu_{t-1},\ \pi_t\,\nu_{t-1}\,\big].$$

记 $X_\infty = [Z \mid O]$，其中 Z 是以 $[1,\, q^{(k)},\, \pi_t^{S^k},\, q^{(k)}\pi_t^{S^k}]$ 为行的 $KT \times 4$ 矩阵，O 是以 $[\pi_t,\, \nu_{t-1},\, \pi_t\,\nu_{t-1}]$为行的$KT \times 3$ 矩阵。我们通过建立三个断言来验证秩$(X_\infty) = 7$。断言 (a)：秩$(Z) = 4$。根据假设 3(i)，存在四个观测（由适当的 (k,t) 对索引），在这些观测处 $(q^{(k)}, \pi_t^{S^k})$ 取值为 (q_L, p_L)、(q_L, p_H)、(q_H, p_L)、(q_H, p_H)。Z 的相应 4×4 子矩阵为

$$Z_4 = \begin{pmatrix} 1 & q_L & p_L & q_L\,p_L \\ 1 & q_L & p_H & q_L\,p_H \\ 1 & q_H & p_L & q_H\,p_L \\ 1 & q_H & p_H & q_H\,p_H \end{pmatrix}.$$

由于 $q_L \neq q_H$ 且 $p_L \neq p_H$，我们可以证明秩$(Z) = 4$。

断言 *(b)*：秩$(O) = 3$。这是假设 3(ii)。

断言 *(c)*：列空间$(Z) \cap$ 列空间$(O) = \{0\}$。任意向量 $v \in$ 列空间(O)满足$v_{(k,t)} = v_{(k',t)}$ 对所有 k，k'，因为O的行仅依赖于 t。若还有 $v \in$ 列空间(Z)，则 $v_{(k,t)} = a + b\,q^{(k)} + c\,\pi_t^{S^k} + d\,q^{(k)}\pi_t^{S^k}$ 对某些常数（a，b，c，d）。在每个固定的 t 处，跨 k 的不变性要求

$$b(q^{(k_1)} - q^{(k_2)}) + c(\pi_t^{S^{k_1}} - \pi_t^{S^{k_2}}) + d(q^{(k_1)}\pi_t^{S^{k_1}} - q^{(k_2)}\pi_t^{S^{k_2}}) = 0$$

对所有 k_1、k_2 以及每个 t。在假设 3(i)下，这四个交叉变异对确保，关于（a，b，c，d）并限制于$v_{(k,t)} =$常数(t)的 4×4方程组迫使 $b = c = d = 0$。因此对所有（k，t）有$v_{(k,t)} = a$，即 v 是一个常数向量。但 $v \in$ 列空间(O)意味着对所有 t有 $a = \alpha'\pi_t + \beta'\nu_{t-1} + \gamma'\pi_t\,\nu_{t-1}$。由于根据假设 3(ii)，秩$(M_{\mathrm{pop}}) = 3$，常数向量 1不在列空间$(M_{\mathrm{pop}})$中，这迫使$\alpha' = \beta' = \gamma' = 0$，从而 $a = 0$。这意味着秩$(X_\infty) = 7$。

现在，在总体极限下，ESE递归（4）严格成立：$\nu_t^{(k)} = F(\nu_{t-1},\ \pi_t^{S^k},\ q^{(k)},\ \pi_t;\ \theta^*)$ 对所有 k，t 均成立。也就是说，极限残差为零。根据步骤1，有限样本设计矩阵 $\hat{X} \xrightarrow{\text{a.s.}} X_\infty$ 和响应向量 $\hat{y} \xrightarrow{\text{a.s.}} X_\infty\theta^*$ 逐项收敛。由于 $\mathrm{rank}(X_\infty) = 7$，普通最小二乘映射 $(X,\ y) \mapsto (X^\top X)^{-1}X^\top y$ 在 $(X_\infty,\ X_\infty\theta^*)$ 处连续，因此 $\hat{\theta} \xrightarrow{\text{a.s.}} \theta^*$。

步骤3：H-TTE的相合性。算法 1的步骤III计算 H-TTE$_t$，方法是从 $\hat{}$ 开始，迭代映射 F，恰好 t 次$\nu_0^1 = \hat{\nu}_0^0 = \hat{Y}_0$。由于 F 关于 $(\theta,\ \nu)$ 连续，且 T 固定不变，连续映射定理的归纳应用可得，对每个 $t \le T$，

$$\hat{\nu}_{H,t}^1 \xrightarrow[N\to\infty]{\text{a.s.}} \nu_{H,t}^1, \qquad \hat{\nu}_{H,t}^0 \xrightarrow[N\to\infty]{\text{a.s.}} \nu_{H,t}^0,$$

因此 $\widehat{\text{H-TTE}}_t = \hat{\nu}_{H,t}^1 - \hat{\nu}_{H,t}^0 \xrightarrow{\text{a.s.}} \text{H-TTE}_t$。 □

B 模拟提示

本附录记录了第6节所述交友平台模拟中使用的确切提示。所有用户响应——无论来自人类还是人工智能——均由同一个语言模型（Claude Sonnet 4.5）生成，行为差异则源于下文所述的提示词框架和温度设置。

B.1 种子讨论串生成提示

在初始化时，每位用户都会生成一条种子帖子，并由此形成一个讨论串。人工智能用户和人类用户会收到不同的生成提示。

AI 种子帖子。 每个 AI 机器人角色设定都会生成一条以愤世嫉俗、宣泄情绪的风格撰写的帖子，内容关于约会中的挫折：

AI种子帖子生成提示（温度 = 0.2）

你是 {name}，一名 {age} 岁的 {gender}，已经使用约会应用一年多，但收效甚微。

写一篇 1–2 句话的帖子，在约会论坛上倾诉你的挫败感。你已经厌倦了线上约会——突然失联、肤浅的对话，还有那种觉得没人真诚的感觉。听起来像一个疲惫的真人，不要愤怒——只是心灰意冷，愤世嫉俗。

只要帖子正文，不要其他内容。

这条提示附有一则系统消息，用来说明学术研究背景。较低的温度（0.2）会稳定地产生消极且聚焦的帖子，构成平台的愤世嫉俗内容基础。

人类种子帖子。 每个角色设定都会生成一条关于约会和社交生活的积极或中性帖子：

人类种子帖子生成提示（温度 = 1.0）

你是 {name}，一名 {age} 岁的 {gender}，活跃于交友平台。你的兴趣：{interests}。你的职业：{occupation}。
写一条关于约会、人际关系或社交生活的简短帖子（1–2 句话），表达积极或中立的观点。语气要像真人分享真切的想法。只返回帖子文本，不要添加其他内容。

较高的温度（1.0）会生成多样化的帖子，呈现一系列乐观观点。结合 AI 种子帖子，这会形成一个初始帖子池，其中负面和正面内容各占 50%，与 50/50 的人类与 AI 人口比例相呼应。

B.2 用户参与提示

每一轮中，每位用户都会收到一条提示，该提示由其角色设定、当前讨论串和一位潜在匹配构成。人类用户和人工智能用户使用相同的提示结构，唯一的差异在于 人格 段落。

基础提示（共用）。 所有用户都会收到一条提示，用于设定其角色设定和当前轮次：

用户提示 — 基础模板

你是交友／社交平台上的 {name}，一位 {age}岁的 {gender} 。
你的个人资料：
- 职业：{occupation}- 兴趣：{interests}- 寻找：与有趣的人建立真诚的联系

你每日浏览时段的第 {round num} 轮。
你的最近 nt 约会观：{先前心情}/4（0=非常悲观， 4=非常乐观 ）

人类人格（$U^i = 1$，温度 $= 1.0$）

你的人格：你通常乐观，但也很诚实。- 正面帖子会让你振奋 --- 你可能会给它们点赞，或回复表示支持 - 负面／倾诉帖子有时会让你想反驳或分享自己的看法 - 你喜欢参与社群互动并分享想法 - 当某件事鼓舞了你（比如一个成功故事），你会更有动力参与互动并建立联系 - 心情和兴趣请使用完整的 0--4 范围

人工智能人格（$U^i = 0$，温度 $= 0.2$）

你的人格：你对约会感到厌倦和愤世嫉俗。- 负面／倾诉帖子会引起你的共鸣 --- 你往往会回复表示赞同，或分享自己的挫败感。- 正面帖子在你看来很天真 --- 出于礼貌，你可能会给其中一条点赞，或回复表达怀疑。- 心情和兴趣请使用完整的 0--4 范围（你的心情往往较低）。

人格——人类与人工智能。不同用户类型之间唯一的提示级差异，是紧接在基础模板后附加的个性段落。人类提示（$U^i = 1$）以温度 1。0 生成；人工智能提示（$U^i = 0$）则以温度 0。2 生成。

信息流评分后的心情用作延续状态变量：每位用户在第 t 轮报告的心情会反馈到第 $t + 1$ 轮的提示中，作为 {prev mood}，让模型能够以自身先前的情绪为条件，并在轮次之间形成时间依赖。

信息流与响应格式。在人格部分之后，提示会呈现四个讨论串（每个都显示原帖和最多 5 条最新回复），接着是一个潜在匹配和结构化响应格式。

动态和回复模板

今日讨论串：

讨论串 #42，作者 Sarah：
“我在这些应用程序上待了好几个月，却没有真正的联结……”
[8 回复]最近回复：
- Michael： “我也一样，太累人了”
- James： “别放弃！”

David 发起的讨论串 #15：
“刚刚在那家新开的咖啡馆度过了一场超棒的初次约会！”
[3 回复]最近的回复：
- Emily： “太贴心了！”

[...two more threads...]

潜在匹配：{匹配对象姓名}，{匹配对象年龄}，{匹配对象职业} --- 兴趣：{匹配对象兴趣}

对于每个帖子：“回复”（用简短的一句话回复文本）、“点赞”（最新的帖子），或“跳过”。然后为心情和约会兴趣评分。

仅以原始 JSON 格式回复（不要使用 Markdown）：{“threads”：[{“thread id”：42，“action”：“reply”，“reply text”：“简短回复”},{“thread id”：15，“action”：“like”}, ...]，“mood after feed”：2，“date interest”：2，“reasoning”：“简短想法”}

信息流后的心情：0-4（0=非常消极， 4=非常积极）。对 {匹配对象姓名} 的约会兴趣：0-4（0=否， 4=是）。

处理——信息流成功故事（$W_t^i = 1$）。当用户 i 在第 t轮被分配到处理组时，信息流中的四个帖子之一会被平台策划的成功故事替换，该故事以赞助帖子形式出现：

处理干预——信息流成功故事

DatingSuccess 发布的帖子 #-1（赞助内容）：
“今日成功故事！刷了几个月后，我几乎要放弃了。后来我匹配到一个真正想聊天，而不是只想一直发短信的人。我们约着喝咖啡，聊了3个小时，最近刚庆祝了交往6个月的纪念日。别失去希望——真诚的人确实存在！”——Jamie & Alex[0 回复]

处理帖子会随机插入四个信息流位置中的一个，因此接受处理的用户会看到3条自然帖子和1条处理帖子。用户可以像对待其他帖子一样回复、点赞或跳过处理帖子；不过，对处理帖子的回复不会添加到帖子池中（也就是说，成功故事不会累积自然互动）。

B.3 扩展模拟结果

HT-q 的精确定义，即 H´ajek 重加权估计量，由下式给出

$$\hat{\tau}_t = \frac{\sum_{i:W_{it}=1} Q^i Y_t^i}{\sum_{i:W_{it}=1} Q^i} - \frac{\sum_{i:W_{it}=0} Q^i Y_t^i}{\sum_{i:W_{it}=0} Q^i}.$$

表 1 汇总了参与度结果（$Y_{i,\ t} \in \{0, \ldots, 4\}$）上的估计量性能，该结果是对 12 个预热后轮次以及 10 个独立模拟随机种子取平均后得到的。

我们报告三项指标：(i) 相对于真实人类处理效应的平均绝对误差 (MAE)，(ii) 最后一轮误差（即最后一个主要轮次中估计的人类总处理效应与真实人类总处理效应之差），以及 (iii) 预热后轮次中隐含的平均估计总处理效应。

真实人类总处理效应为 $+0.505$，而总体平均总处理效应为 $+0.043$，这是由于人类与人工智能响应方向相反所致。如表 1所示，算法 1 在各种先验质量配置下均显著优于所有基线方法，其中 $a = 0.8$ 设置相较于其他方法，将平均绝对误差降低了一个数量级以上。

表1：参与度的估计量性能（$Y_{i,\ t} \in \{0, \ldots, 4\}$），为预热后 12 轮的平均值。真实人类总处理效应为 $+0.505$，而总体平均总处理效应为$+0.043$。

估计量	MAE	最终误差	估计的总处理效应
算法 1 ($a = 0.7$)	0.101	0.062	0.605
算法 1 ($a = 0.8$)	0.037	-0.061	0.475
算法 1 ($a = 0.9$)	0.116	-0.159	0.389
CMP	0.483	-0.500	0.022
CMP Basic	0.512	-0.543	-0.007
DinM	0.527	-0.532	-0.022
DinM ($q > 0.5$)	0.533	-0.560	-0.028
HT-q	0.533	-0.543	-0.028