可信度治理：弱真理信号下实现集体自我纠正的社会机制

Wanying He 武汉大学人工智能学院；BIGAI 中国武汉
wanying.he@whu.edu.cn

Yanxi Lin 清华大学 中国北京
linyx22@mails.tsinghua.edu.cn

Ziheng Zhou 加利福尼亚大学洛杉矶分校 美国洛杉矶
josephziheng@ucla.edu

Xue Feng 通用人工智能国家重点实验室，BIGAI 中国北京 f
eng.xue1580@gmail.com

Min Peng 武汉大学人工智能学院 中国武汉
pengm@whu.edu.cn

Qianqian Xie 武汉大学人工智能学院，武汉，中国
xqq.sincere@gmail.com

Zilong Zheng 通用人工智能国家重点实验室，
BIGAI，北京，中国
zlzheng@bigai.ai

Yipeng Kang 通用人工智能国家重点实验室，
BIGAI，北京，中国
kangyipeng@bigai.ai

arXiv:2603.02640v1 [cs.CY] 3 Mar 2026

摘要

在线平台日益依赖意见聚合来分配现实世界中的注意力和资源，但参与度投票或资本加权承诺等常见信号很容易被放大，而且往往反映的是可见度而非可靠性。因此，在真相信号微弱、反馈存在噪声或延迟、人气早期激增以及策略性操纵的情况下，集体判断会变得脆弱。我们提出可信度治理（CG），这是一种通过学习哪些智能体和观点能够持续追踪不断变化的公共证据，来重新分配影响力的机制。CG 为智能体和意见维护动态可信度评分，通过按可信度加权的认可来更新意见的影响力，并根据智能体所支持意见的长期表现更新其可信度；它奖励那些早期且持续地与新兴证据保持一致的意见，同时过滤短期噪声。我们在 POLIS（一种社会物理模拟环境）中评估 CG，该环境对不确定性条件下耦合的信念动态及下游反馈进行建模。在初始多数意见失准、观测噪声与污染以及错误信息冲击等情境下，CG 的表现优于基于投票、按权益加权和无治理基线，能够更快恢复到真实状态，减少锁定效应和路径依赖，并提升面对对抗压力时的鲁棒性。我们的实现和实验脚本已公开发布于 https://github.com/Wanying-He/Credibility_Governance。

特此免费授予许可，允许为个人或课堂使用而制作本作品的全部或部分的数字副本或纸质副本，但不得为盈利或商业利益而制作或分发副本，并且副本首页须附有本声明和完整引文。对于本作品中由作者以外的其他方拥有版权的部分，必须尊重其版权。允许在注明出处的情况下摘录。以其他方式复制、再版、发布至服务器或重新分发至列表，均需事先获得特定许可和/或支付费用。请通过 permissions@acm.org 申请许可。会议缩写 'XX，纽约州伍德斯托克 © 2018 版权归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-XXXX-X/2018/06 https://doi.org/XXXXXXX.XXXXXXX

ACM 概念分类

• **以人为中心的计算 → 声誉系统**。

关键词

网络可信度治理、在线观点聚合、集体决策、集体自我纠正、基于大语言模型的社会模拟

ACM 参考格式：

Wanying He、Yanxi Lin、Ziheng Zhou、Xue Feng、Min Peng、Qianqian Xie、Zilong Zheng 和 Yipeng Kang。2018。可信度治理：弱真理信号下集体自我纠正的社会机制。载于《请务必填写权利确认电子邮件中提供的正确会议名称（会议缩写 'XX）》。ACM，纽约，美国，14页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

想象一下，你是一名年轻研究者，正在探索一个颇具跨学科性质、初现潜力的想法。你发布了一篇简短的工作论文，其中谨慎地呈现了耐人寻味的试点结果，随后讨论在一个网络论坛上展开，而你的领域正日益通过该论坛交流思想。你能够收集到的信号缓慢、嘈杂且难以解读，但这一方向让你感觉意义非凡。然而，随着时间推移，平台上的讨论开始偏离原来的轨道。另一条研究路线由一个联系紧密的学者群体推动，迅速产出大量论断和经过精心润色的更新。他们突出彼此的进展，转发彼此的摘要，而这种协调一致的活动塑造了平台的观点统计，资助机构和程序委员会也越来越多地对其加以监测。尽管底层证据仍然薄弱，但这一竞争方向的可见度和支持声量迅速增长，平台上的汇总信号开始引导邀请、合作和资源分配。你的研究路径更加审慎，尽管建立在更可靠的信号之上，却变得越来越难以在

汇总共识中被察觉，后来的学者也因此形成了对该领域观点的扭曲认知，混淆了他们的判断。

这一情境反映了现代社交与网络平台的一项系统性局限。当注意力通过快速且相互强化的社交信号汇聚时，较慢却更可靠的质量指标难以浮现 [3, 26]。因此，集体评估可能偏向最显眼的选项，而非证据支持最充分的选项，社区也可能过早地聚焦于某些方向，而这些方向后来被证明并不稳固 [23]。

本文探讨，当真值只能被微弱地观察到时，集体能否引导自身摆脱此类扭曲。我们提出一种名为可信度治理 （CG）的机制，专为智能体必须根据不确定、有延迟且分布不均的信号推断真值的环境而设计 [14]。CG 不把可见度或原始支持度视为证据，而是根据智能体如何解读随时间变化的状况，动态分配影响力并更新可信度，旨在保留对新出现的证据的敏感性，同时避免放大短暂波动。

我们的贡献有三点。第一，我们引入了一个双组件仿真环境，将话题进展的物理世界与智能体形成、表达和聚合意见的社会世界联系起来，从而能够系统地研究部分可观测条件下的求真过程。第二，我们对可信度治理进行形式化，并将其与三种基准聚合规则进行比较：基于质押的机制会放大得到更多资源承诺支持的主题；基于投票的机制反映简单多数的偏好；无治理基线则将集体选择直接应用于观察到的进展。这些基准共同体现了当代平台聚合意见的主流逻辑 [34]。第三，我们表明，可信度治理能够在广泛存在噪声、延迟或误导信息的环境中支持集体自我纠正，同时也指出了边界条件，阐明该机制何时有效，以及在哪些方面仍需改进。

综上所述，这些贡献为构建意见聚合系统迈出了一步：即使证据稀少、延迟或难以核实，这类系统仍能保持认知稳定 [20]。

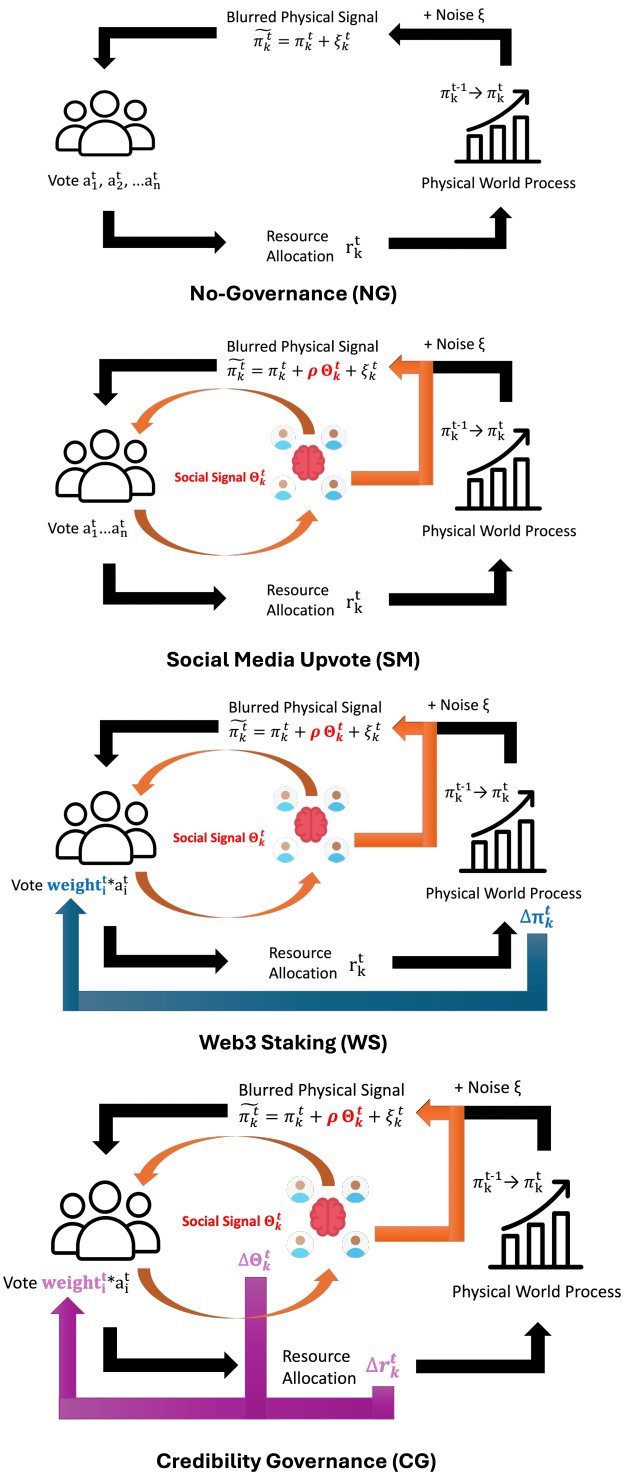

图1：四种治理机制的比较。(1) 无治理：智能体观测物理信号 $\tilde{\pi}_k^t$。(2) SM：在观测中加入无治理 +社会信号 Θ_k^t。(3) WS：将 SM + 代理影响力 w_i^t更新为 $\Delta\pi_k^t$。(4) 可信度治理：将 WS + 代理影响力 w_i^t更新为 Θ_k^t。

2 方法论

2.1 POLIS 仿真框架

2.1.1 平台功能

为支持对可信度治理进行严谨评估，我们引入**POLIS**，这是一个模块化仿真平台，旨在研究社会动态与现实世界结果的共同演化。POLIS 协调大规模由大语言模型驱动的智能体群体，使其在相互耦合的**物理世界** 和**意见世界**中互动，并提供以下核心功能：

(1) **可扩展的多智能体编排**。 该平台可管理数千个自主的大语言模型智能体，每个智能体都具备持久身份、记忆和稳定的推理能力。

(2) **参数化物理世界建模**。 物理世界完全参数化，并定义灵活的资源到进展的映射，将关于资源分配的集体决策转化为现实世界的结果轨迹。这种设计支持任意进展函数，使 POLIS 能够模拟多种

domai景象 h 如发现ientific di索、技术 hnological adopti、或政策实施。

(3) **模块化治理接口**。POLIS 支持可插拔的治理模块，用于定义意见世界如何汇总智能体意见以形成集体决策，并更新社会状态。这使得在相同的智能体和世界设置下比较不同的治理机制变得轻而易举。

(4) **全面的数据记录**。系统会记录每个智能体的决策、推理依据、状态变化和环境信号，从而能够对微观层面的行为和宏观层面涌现现象进行细致且可复现的分析。

2.1.2 双世界结构与交互

物理世界与意见世界通过一组共享的**主题**相互关联。智能体根据公共信号形成感知，对主题进行投票，而它们的集体选择决定了后续的话题进展。以下组件概述了各个世界的关键要素，并简要说明了信息在其中的流动方式。

(1) 主题：跨世界的核心概念。主题代表现实世界中相互竞争的努力方向，例如科学假设或技术路径。每个主题在两个世界中都具有相应的属性：

- π_k^t：由其累积进展值表示的物理状态，
- Θ_k^t：一种观点状态表示，包括当前的支持度和公开可见的社会信号。

当智能体对主题 投票，而机制根据其支持情况采取行动时，主题便提供了连接这两个世界的关键接口。

（***2***）物理世界：主题进展动力学。 物理世界决定资源分配如何转化为可测量的进展。每个主题 k 都有一个累积进展值 π_k^t，其增量 $\Delta\pi_k^t$ 由参数化更新规则决定，该规则规定了主题的发展轨迹。在本研究中，我们允许主题沿非线性轨迹演进，并在多个增长阶段之间转换，从而呈现多种模式，例如起步缓慢、快速增长或边际收益递减。环境波动通过加性噪声项体现，以刻画进展中的现实不确定性。

（***3***）意见世界：智能体与主题社会状态。 意见世界维持系统不断演变的社会层。每个智能体都会追踪：

- α_i^t：基于公共信号形成的对主题的信念与置信度，
- a_i^t：其当前对主题的投票选择，
- w_i^t 智能体影响权重，根据可信度或利益余额等机制特定属性得出。

主题会维持一个社会信号 Θ_k^t，以反映其当前支持度及其支持度随时间变化的速率，具体定义取决于所选的治理机制。

（***4***）从物理到意见：智能体感知。 智能体无法直接观察真实的物理状态。相反，它们会接收社会信号 Θ_k^{t-1} 和含噪物理信号 $\tilde{\pi}_k^{t-1}$，以模拟现实世界中的情境：真实情况无法获取，且大多数可观察证据都会受到社会影响的扭曲。这些信号构成智能体的感知证据，塑造其世界观，并影响其对不同主题的支持。

（***5***）从意见到物理：投票聚合。 智能体根据自身信念和认知角色，对主题投票 a_i^t。随后，治理机制依据机制特定规则，将这些投票聚合为资源分配向量 $\mathbf{r}^t$。接着，这一分配会转化为主题进展增量 $\Delta\pi_k^t$，从而完成跨世界反馈循环。

2.1.3 轮次级演练

一次模拟轮次描述了物理世界和意见世界如何从时间 $t-1$ 演变至 t。这一过程严格遵循因果顺序：智能体观察第 $t-1$ 轮结束时产生的信号，形成感知，对主题进行投票，而集体决策则会更新物理进展和意见状态变量，这些变量将生成下一轮中观察到的信号。轮次结构的演练如下：

(1) 感知：在第 t 轮开始时，每个智能体都会观察上一轮的社会信号 Θ_k^{t-1} 和上一轮的含噪物理信号 $\tilde{\pi}_k^{t-1}$。这些信号共同构成智能体的感知证据，并塑造其在第 t 轮的世界观。

(2) 意见形成与聚合：每个智能体都会更新其 α_i^t，包括其信念（对支持哪个主题所作的定性选择）和置信度（衡量其持有该观点的坚定程度）。由此产生对所选主题的投票意向 a_i^t。随后，治理机制会使用影响力权重 w_i^{t-1}，将这些意向聚合为资源分配向量 $\mathbf{r}^t$。

（***3***）状态更新：给定$\mathbf{r}^t$，两个世界都会更新其状态：

- **物理世界**将$\mathbf{r}^t$转换为主题进展增量$\Delta\pi_k^t$，并更新累计进展值 π_k^t。
- **意见世界**。治理机制更新智能体影响权重 w_i^t和主题级社会状态，生成下一轮社会信号 Θ_k^t。系统利用更新后的进展和社会信号，随后构建下一轮的含噪物理信号

$$\tilde{\pi}_k^t = \pi_k^t + \rho\,\Theta_k^t + \xi_k^t,$$

该信号与 Θ_k^t一同被智能体在第 $t+1$轮开始时观测到。

2.2 意见世界（治理机制）

为评估可信度治理，我们将其与三种具有代表性的聚合规则进行比较：（1）Web3 式质押（WS）机制，实现资源加权治理；（2）社交媒体点赞（SM）机制，实现由参与度驱动的投票；（3）无治理基线，在该基线下，智能体仅依据观测到的信号采取行动。

评估的四种机制均基于上述同一协议（如图1所示），但在影响力权重的分配方式 w_i^t、社会信号的更新方式Θ_k^t 以及影响力更新依据（社会动量或物理进展）方面有所不同。各机制之间的主要差异汇总于表1。

社会信号的解读 Θ。我们将 Θ_k^t 用作一个汇总的、可公开观察的代理指标，用以表示群体在时间 t 对选项或主题 k 的当前支持（例如，平台层面的“共识”信号）。在

表1：治理机制属性比较。

特征	信誉治理（CG）	Web3质押（WS）	社交媒体（SM）	无治理（NG）
代理影响力（w_i）	基于信誉	基于质押	平等影响力	平等影响力
社交信号更新（Θ_k）	信誉调整 Δ支持	按质押加权的人气（延迟 τ，噪声 ξ）	原始人气	None
信号失真	噪声 ξ，污染。 $\rho\Theta_k^t$	噪声 ξ，延迟 τ（权益结算）	噪声 ξ，污染。 $\rho\Theta_k^t$	噪声 ξ
鲁棒性调节参数	反泡沫 B_k^t	权益成本 / 预算	None	None
奖励依据	$\Delta\Theta_k^t$	$\Delta\pi_k^t$ （延迟）	N/A	N/A

许多人类系统中，真实情况仅以微弱、滞后或通过嘈杂代理指标的方式显现，而 Θ_k^t 则可立即获得并塑造行为。因此，我们将 Θ_k^t 解释为一种公共证据代理，即它概括了当时群体中似乎正在累积的信息。关键在于，可信度治理并不奖励高支持本身；它奖励的是 $\Delta\Theta_k^t$，即支持随时间发生的变化 。这种"滤波动量"选择强调有信息含量的变动，例如在新证据出现后的信念修正，而不是强化静态的受欢迎程度。当真相信号较弱时，$\Delta\Theta_k^t$ 为证据累积提供了一个轻量、基于行为的代理指标，使可信度治理能够提高那些持续促成公众信念建设性转变的智能体的影响力。

失效模式。 当社会运动受操纵或协调而非证据主导，或者观测噪声变得过大、以致 Θ_k^t 不再能反映潜在结果时，这一代理指标就可能失效；这也促使我们在第3节进行噪声与对抗压力测试。

2.2.1 可信度治理（可信度治理）

可信度治理根据认知表现而非原始支持计数来分配代理影响力。当智能体上一轮支持的主题收到更高的社会信号Θ_k^t时，其可信度会上升；当其先前的选择与新兴共识背道而驰时，可信度则会下降。这样，可信度治理便将影响力与智能体以往对可信的公共证据变化的历史一致性联系起来。

(1) 感知：每个智能体都会观察上一轮的社会信号 Θ_k^{t-1}（计算方法见状态更新）和含噪物理信号 $\tilde{\pi}_k^{t-1}$，据此形成其对第 t 轮的感知证据。

(2) 意见形成与聚合：智能体更新其信念（主题偏好）和置信度 α_i^t（偏好的强度），从而产生投票意向 a_i^t。影响力权重按可信度加权：

$$w_i^{t-1} = \alpha_i^t \exp(\lambda c_i^{t-1}),$$

分配方式如下：

$$r_k^t = \frac{\sum_i w_i^{t-1}\mathbf{1}[a_i^t = k]}{\sum_j w_j^{t-1}}.$$

(3) 状态更新：治理机制更新：

- **主题层面的社会状态：**

$$\Theta_k^t = (1-\lambda_s)\Theta_k^{t-1} + \lambda_s\left[(r_k^t - r_k^{t-1})\,\bar{q}_k^t - \gamma B_k^t\right],$$

其中 λ 为学习率， q_k^t 为支持者质量项， B_k^t 为反信息茧房惩罚（定义见下文）。

- **智能体层级可信度：**

$$c_i^t = c_i^{t-1} + \eta\left(\Theta_{a_i^t}^t - \Theta_{a_i^t}^{t-1}\right) e^{-\kappa r_{a_i^t}^t},$$

其中， η 是学习率，而 κ 则实现了先行者奖励：在主题获得大量资源之前（即 $r_{a_i^t}^t$ 较小）就支持该主题的智能体，其可信度更新幅度更大。

（***4***）关键支持组件：

- **支持者质量**。 可信度治理会衡量支持主题 k的智能体的平均可信度：

$$\bar{q}_k^t = \frac{\sum_{i:a_i^t=k} c_i^{t-1}}{\sum_{i:a_i^t=k} 1}.$$

高质量的支持对正向社会信号更新的贡献更大。

- **反信息茧房惩罚**。 可信度治理会惩罚缺乏可信依据的支持度快速增长：

$$B_k^t = \sigma\big(\alpha(r_k^t - r_k^{t-1})\big)\,(1-\bar{q}_k^t),$$

从而抑制低质量跟风。

2.2.2 Web3 式质押（WS）

WS 对一种基于代币的「一美元一票」治理系统进行建模。代理影响力与质押财富成正比，社会信号 Θ_k^{t-1} 反映按权益加权的人气。与影响力权重根据 $\Delta\Theta_k$更新的 CG 不同，WS 将激励直接与实际结果挂钩，并根据主题进展调整智能体的权益。这形成了一个直接的反馈回路：成功的主题会增强其支持者的影响力。

（***1***）感知：智能体观察上一轮按权益加权的人气 $\Theta_k^{t-1} = r_k^{t-1}$和含噪物理信号 $\tilde{\pi}_k^{t-1}$。

（***2***）意见形成与聚合：智能体更新信念和置信度 α_i^t，并选择一个主题 a_i^t。影响力按权益加权：

$$w_i^{t-1} = \alpha_i^t \mathrm{bal}_i^{t-1}.$$

分配如下：

$$r_k^t = \frac{\sum_i w_i^{t-1}\mathbf{1}[a_i^t = k]}{\sum_j w_j^{t-1}}.$$

（*3*）状态更新：

- **质押更新：**

$$\mathrm{bal}_i^t = \mathrm{bal}_i^{t-1} + \gamma_s \, w_i^{t-1} \, \Delta\pi_{a_i^t}^t.$$

其中 γ 为质押奖励率，下标 s表示该奖励率专用于 Web3 质押机制。

- **社会信号更新：**

$$\Theta_k^t = r_k^t.$$

2.2.3 社交媒体点赞（SM）

SM 代表一种“人人一票”的制度。无论过去的表现如何，影响力权重都保持一致，智能体完全依据自身的信念和观察结果进行投票。社会信号对应于原始人气。不更新任何代理级属性。

(1) 感知： 智能体观察上一轮的原始人气 $\Theta_k^{t-1} = r_k^{t-1}$ 以及含噪物理信号 $\tilde{\pi}_k^{t-1}$。

(2) 意见形成与聚合： 智能体更新信念和置信度，选择主题 a_i^t，并以平等影响力进行投票：

$$w_i^{t-1} = 1.$$

分配如下：

$$r_k^t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[a_i^t = k].$$

(3) 状态更新：

$$\Theta_k^t = r_k^t.$$

不更新可信度、权益或其他属性。

2.2.4 无治理（NG）

无治理（NG）作为基准情形，不存在社会协调、放大或反馈。影响力均等，不维持任何社交信号，也不会随时间更新代理级属性。这模拟了一个没有社会因素的原子化社会，在其中，每个人只观察来自物理世界的信号，并独立思考。

(1) 感知：智能体不会接收社交信号； $\Theta_k^{t-1} = 0$。它们只观察含噪物理信号 $\tilde{\pi}_k^{t-1}$。

(2) 意见形成与聚合：智能体以相等且固定的影响力投票，与 SM 中的情况相同。

(3) 状态更新：不会发生代理级或主题级反馈；影响力和（不存在的）社交信号保持静态。

2.3 物理世界设置

在本研究中，我们将物理世界设定为一个学术环境，其中两种科学理论竞争集体资源。它们的物理进展模型刻画了不同治理机制下现实世界的科学努力如何展开。

2.3.1 主题初始化

我们考察两个面向未来的科学主题：**量子物理学（真实主题A）**和**神经形态物理学（虚假主题B）**。为避免大语言模型全知性，鉴于当前证据支持双方观点，目前尚无定论，无法判断哪一方更站得住脚。为避免大语言模型原有的偏见，我们任意指定真值。两个主题的初始物理进展均设为0。

2.3.2 主题进展动力学

话题进展遵循非线性发展轨迹。设 π_k^t 表示主题 k 在第 t 轮的累积进展， r_k^t 表示该轮分配给主题 k 的资源份额。增量$\Delta\pi_k^t$ 的计算如下：

$$\Delta\pi_k^t = \begin{cases} \epsilon_k^t, & \pi_k^{t-1} < \pi_1, \\ \left(v_k + \gamma \, r_k^t\right)\left(1 - \dfrac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_1 \le \pi_k^{t-1} < \pi_2, \\ \gamma' r_k^t \left(1 - \dfrac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_k^{t-1} \ge \pi_2. \end{cases}$$

这刻画了初始的探索阶段、加速阶段和饱和阶段。这里， v_k 是主题k的内在基准速度， γ 和 γ' 控制资源在不同增长阶段对进展的加速作用， M_k是饱和上限，而 $\epsilon_k^t \sim \mathrm{N}(0, \sigma^2)$则表示环境冲击。

2.3.3 智能体群体与初始条件

我们初始化 $N = 100$ 智能体，并为其设定以下信念和认知特征。7:2:1 的比例有助于检验群体能否从偏斜的初始状态中自我纠正，并观察高质量智能体的影响力是否会随着时间推移而提升。

- **立场错位的多数派（70 个智能体）：** 初始时相信虚假主题B，且具有中等的认知稳定性。
- **真理一致的少数派（20 个智能体）：** 初始时相信真实主题A，且仍然较易受到影响。
- **高信念核心群体（10 个智能体）：** 初始时相信真实主题A，且具有极高的认知稳定性，从而以真实情况为系统锚定方向。

除非另有说明，智能体和主题层面的状态初始化如下：

- **可信度治理（CG）：** $c_i^0 = 1$ 适用于所有智能体。
- **权益（ WS ）：** $\mathrm{bal}_i^0 = 1$ 适用于所有智能体。
- **社会信号（CG）：** $\Theta_k^0 = 0$ 适用于两个主题。
- **置信度：** $\alpha_i^0 = 0.5$ 适用于所有智能体。

这些初始条件确保所有机制都从一个对称、低信息状态开始：没有任何主题积累了进展，没有任何智能体拥有特权影响力，社会信号尚未形成，并且所有智能体的初始置信度都处于中等偏高水平。

2.4 将可信度治理映射到现实世界示例

表2将可信度治理变量映射到三个数据丰富的示例（预测、同行评审和论坛）中具体的按时间索引的代理指标。可信度治理只要求 (i) 一个公开社会信号 Θ_k^t（及其变化 $\Delta\Theta_k^t$），以及 (ii) 一个伴随噪声与延迟到达的与结果相关的信号 π_k^t（通常通过代理指标 $\tilde{\pi}_k^t$ 观测）。

3 实验结果

我们以假设检验的形式呈现结果。每幅图都对应于关于可信度治理何时能够发挥作用的一个具体论断

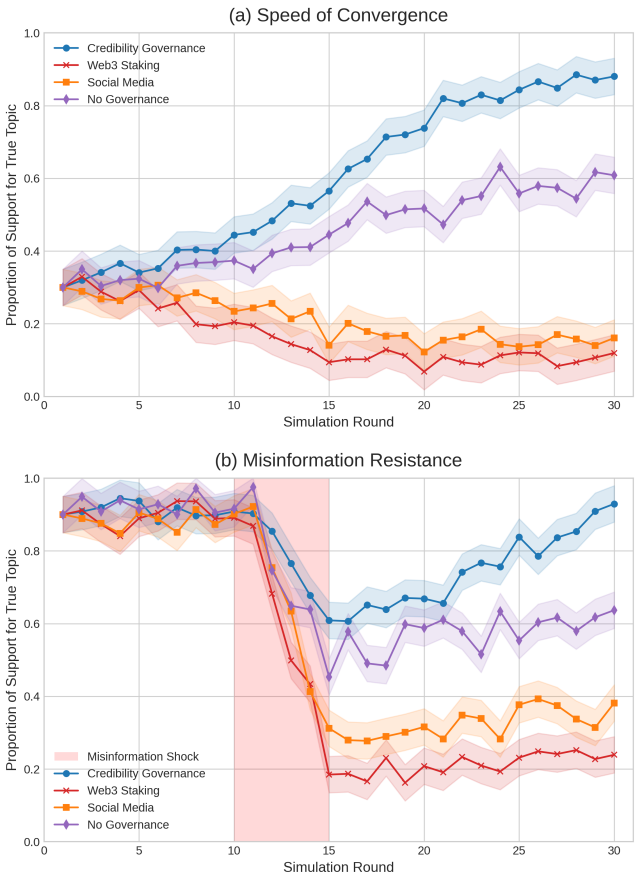

图2：H1，系统动力学。（a）展示了系统如何从最初的虚假多数状态中收敛。（b）展示了错误信息冲击（阴影窗口）后的恢复。与 WS、SM 和无治理相比，可信度治理对真实主题的支持比例更高，且在冲击后恢复得更快。

（CG）能够提高集体准确率以及原因。除非另有说明，否则每次运行持续 30 轮次，我们报告 10 次试验的平均轨迹及置信带。

3.1 H1：可信度治理加速真值收敛并改善冲击后恢复

假设（H1）。当真值难以观测时，可信度治理通过随时间推移促进纠正来提高系统级准确性，并能更有效地从错误信息冲击中恢复。

测试。我们追踪以下两种情况下各轮次对真实主题的支持比例：（i）最初处于虚假多数状态；（ii）遭遇外生错误信息冲击。

结果。在图 2a 中，可信度治理持续推动集体信念转向真实主题，并达到高共识状态；而 WS 和 SM 放大早期流行度，使其偏离真值，无治理则缓慢改善。在图 2b 中，可信度治理展现出最强的冲击后恢复能力和最高的最终准确性，这与锁定效应减弱、纠正速度加快相一致。

3.2 H2：可信度治理通过将影响力重新分配给与真相一致的智能体而发挥作用

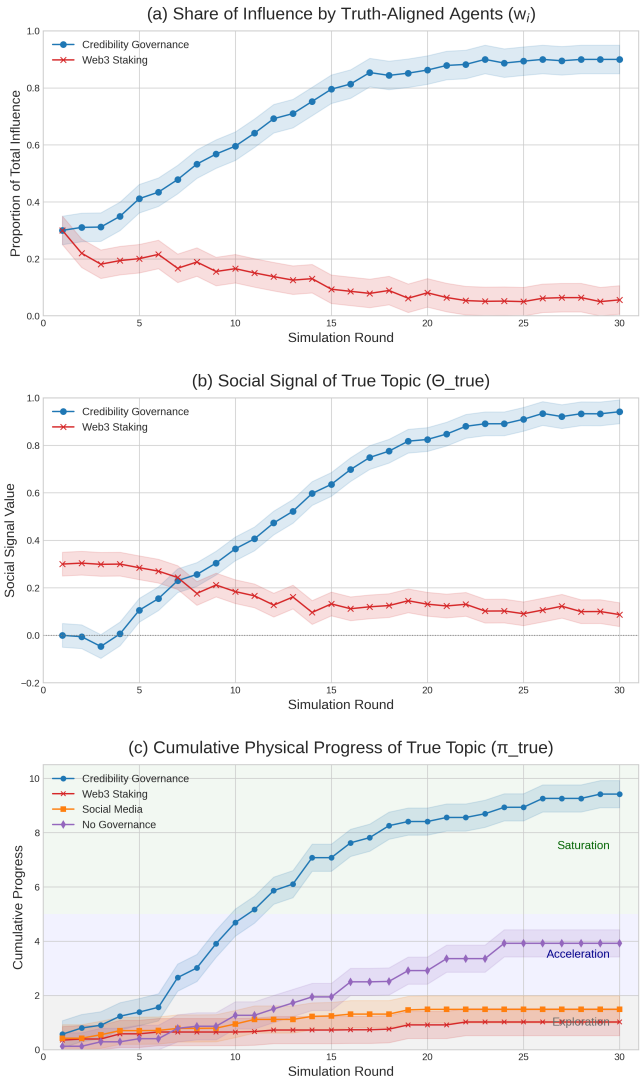

图 3：H2，机制路径。面板 (a) 跟踪与真相一致的智能体所持影响力的占比。面板 (b) 展示真实主题的社会信号 Θ_{true}。面板 (c) 展示累积物理进展 π_{true}，背景阶段表示探索、加速和饱和。可信度治理将影响力重新分配给与真相一致的智能体，强化真实主题的社会信号，并由此累积推动物理进展加快。

假设（H2）。可信度治理通过将影响力权重转向与真相一致的智能体来改善系统结果，从而强化真实主题的社会信号，并增加下游实际进展。

测试。我们测量：（i）与真值一致的智能体所持有的总影响力份额，（ii）真实主题 Θ_{true} 的社会信号，以及（iii）

累积物理进展 π_{true}。

结果。 图3a显示，随着时间推移，可信度治理会将影响力集中到与真值一致的智能体手中，这与 WS 不同。这提升了 Θ_{true}（图3b），并累积为更高的物理进展总量（图3c），解释了 H1 中观察到的系统层面差距。

3.3 H3：可信度治理的核心组成部分不可或缺，且 $\Delta\Theta$ 是恰当的奖励依据

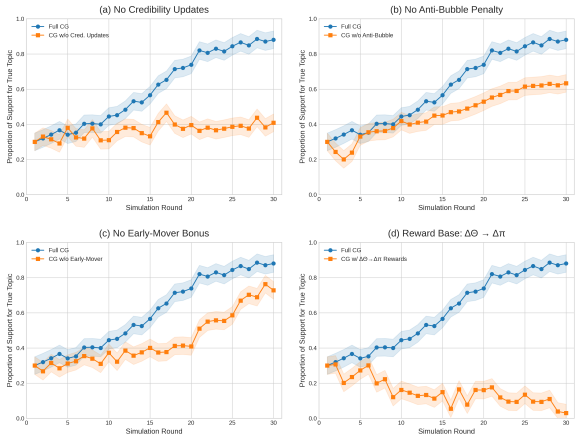

图4：H3 消融实验。每个面板都比较完整可信度治理与一个消融变体对真实主题的支持度：（a）无可信度更新，（b）无反信息茧房惩罚，（c）无先行者奖励，（d）奖励基准 $\Delta\Theta_k \to \Delta\pi_k$。移除任何一个组成部分都会降低性能，而将 $\Delta\Theta$ 替换为$\Delta\pi$ 最显著地破坏收敛稳定性。

假设（H3）。 可信度治理的性能取决于：（i）可信度更新；（ii）反信息茧房正则化项；（iii）防止过早锁定的先行者奖励；以及至关重要的（iv）将公共证据中的势头 $\Delta\Theta_k$ 而非短期物理代理指标$\Delta\pi_k$ 用作奖励依据。

测试。 我们逐一移除各个组件，并将收敛轨迹与完整的可信度治理进行比较。

结果。 移除可信度更新会显著降低收敛程度（图 4a）。移除反信息茧房惩罚会加剧对级联效应的易感性，并降低最终准确性（图 4b）。移除先行者奖励，会减缓系统脱离初始虚假多数状态的转变（图 4c）。将奖励依据切换为 $\Delta\pi_k$ 会大幅削弱学习效果（图 4d），这支持了 $\Delta\Theta_k$ 在噪声与延迟条件下是更可靠的奖励信号这一观点。

3.4 其他假设

我们进一步检验**H4**（可信度治理在中等噪声下最有益，能够减少锁定效应并加快纠正）、**H5**（显式可信度提示能带来适度的额外益处）以及**H6**（可信度治理能在对抗性操纵下扩大安全运行范围）。为简洁起见，这些分析详见附录（D.1–D.3）。

4 相关工作

4.1 真相聚合机制

汇总分散判断的系统支撑着许多网络平台、众包流程和数字治理场景。既有方法对应于本研究考察的三种基线：质押型机制（以 Web3 风格质押（WS）为例）、投票型系统（以社交媒体点赞（SM）为例），以及没有显式治理的环境。

基于质押的机制源自去中心化协调和区块链治理框架 [7, 9]。参与者将资源押在某项主张上；若其支持的结果最终被证明正确，便可获得奖励 [16]。尽管这类设计为准确性创造了激励，但资源投入并不能可靠地反映认识论质量。由于证据的生成和传播往往取决于资源 [12], ，资源较多的参与者便可能不论其判断是否可靠，都能对结果产生不成比例的影响。

基于投票和互动驱动的系统广泛应用于社交媒体和内容排名平台 [21, 25], ，使参与更加民主化，但也优先提升可见度和势头。早期的小幅激增可能引发信息级联，使人气盖过传播较慢的证据。基于声誉的投票常见于同行评审和信任评分环境 [29, 33], ，会纳入历史可靠性，但累积的声誉可能降低系统对新情况的响应能力，也可能让过去可信的参与者在其判断与新出现的证据相悖后，仍保有影响力。

无治理环境构成第三类，其中智能体观测含噪物理信号并独立审议。虽然这类系统简单透明，但仍容易受到早期噪声和路径依赖的影响 [30], ，因此需要能够在不确定性下更好地支持集体自我纠正的机制。

4.2 集体错误信息动态

集体认识论研究表明，当早期信号含糊不清或受到社会强化时，群体可能无法正确汇总信息，从而产生羊群效应和信息级联 [4, 5]。在部分可观测性或证据质量参差不齐的情况下 [15], ，这些效应会加剧，即便能力出众的个体也会继承扭曲的信念。

关于错误信息动态的平行研究表明，可见度、重复传播和协同放大如何使缺乏有力支持的主张比准确但传播较慢的证据扩散得更快 [28, 35]。平台排名和反馈系统会进一步巩固这些发展轨迹，使社群陷入脆弱共识状态 [2, 10]。

这些发现凸显了治理框架的必要性：在真值仅能被微弱观测到时，这类框架应抑制早期放大、识别可靠的贡献者，并维持认识论韧性。我们的研究通过提出一种机制，践行了社会认识论的一项核心原则：集体准确率取决于社群评估信息来源可靠性的能力 [13]。可信度治理在动态环境中落实了这一原则，并在证据嘈杂、延迟且分布不均的条件下对其进行了检验。

4.3 基于大语言模型的社会模拟

长期以来，基于智能体的模型一直用于研究个体层面的行为如何产生宏观社会结果 [11]。传统模拟依赖简化规则或固定参数，限制了其捕捉更丰富认知过程的能力。大语言模型改变了这一局面，使智能体能够进行自然语言交互 [17, 19, 22]，并在开放式社会乃至政治情境中体现各种价值取向 [18, 36, 38] [1, 24, 27, 32, 37, 39], ，权衡定性证据，通过审议更新信念，并阐明论证理由。这一转变催生了大量运用大语言模型智能体研究集体行为、协调和涌现的社会动态的文献 [6, 8, 31]。

我们的工作通过将基于大语言模型的社会模拟用作可信度治理的评估环境，建立在这一范式之上。与数值化观点更新模型或基于规则的模拟不同，大语言模型智能体使我们能够研究这样一种情境下的治理机制：决策由语言驱动的推理和依赖语境的证据解读所塑造。这使我们能够分析，在真值难以直接观测、必须根据不断变化且部分可靠的信号进行推断时，影响力和可信度如何演变。

5 讨论

我们的实验支持一项关于弱可观测真相信号条件下集体学习的具体科学主张。当公共证据含有噪声且受到部分污染时，奖励可信势头（过滤后的公共证据代理指标的变化）的激励与聚合机制，能够通过将影响力重新分配给与真相一致的智能体、减少早期锁定并加速错误纠正，提高长期准确性。

5.1 边界条件及机制可能失效的情形

我们的结果也阐明了边界条件。首先，噪声扫描显示出不同的状态：当信号存在噪声但仍具有信息价值时，可信度治理带来的增益最大；而**噪声过大会使任何机制失效**，所有方法的表现都会降至随机水平。其次，当公共证据代理指标与真实情况出现系统性偏差时，例如受到持续污染，或观测存在延迟和偏差，可信度更新可能会错误分配影响力，并减缓纠正进程。第三，即使机制从长远来看具有稳定作用，早期强烈的协调也可能暂时主导势头信号，增加出现路径依赖结果的风险。这些观察表明，除了准确率，还应报告诊断指标，因为相似的最终准确率可能源于不同的失效模式。

5.2 威胁模型的局限性与缓解措施

我们的对抗性评估涵盖了具有代表性的策略性行为和两种攻击者实现方式，但威胁模型仍不完整。目前尚未测试两种重要威胁。**（i）适应性模仿攻击：**攻击者动态调整行为，以模仿可信轨迹，例如交替进行真实和欺骗性贡献以积累可信度，然后在关键窗口兑现。**（ii）串通与贿赂攻击：**协调行动的群体交换平台外激励，或通过私下支付同步支持度变化，使其看起来像自然形成的势头。这两种威胁都针对可信度治理所依赖的核心信号——可信势头，并可能降低真实证据累积与人为制造的动态之间的可区分性。

缓解措施可直接针对这些失效机制。第一种做法是**引入可信度惯性和可审计性**，增加更新阻力、设置上限或采用衰减机制，使可信度骤升必须有持续证据作为支撑，并对高影响力主体或势头激增情况开展轻量级审计。第二种做法是**进行跨信号一致性检查**，要求多个可观测渠道之间相互吻合，例如预测准确率、同行评估的贡献和纵向行为，然后再大幅重新分配影响力。第三种做法是**在策略性操纵下进行稳健聚合**，将突发性、互惠性和基于网络的串通信号等检测特征纳入公共证据代理指标。这些缓解措施与机制结构相容，可在同一压力测试框架内进行评估。

5.3 在真实轨迹数据上进行下一步验证

自然的下一步是在真实轨迹数据上进行验证，并将映射部分作为连接模拟变量与观测信号的接口。具体而言，我们可以使用特定领域的代理指标来实例化 Θ，例如预测场景中的预测准确度、同行评审场景中的评审有用性和校准，或论坛中的纵向帖子质量和社区反馈。随后，我们可以在历史轨迹数据上重放影响力更新，测试反事实治理规则，并衡量可信度治理式更新是否能改善下游结果，例如提高与事后真实值的一致性、降低波动性，以及在错误信息集中爆发后更快恢复。基于轨迹数据的验证也有助于加强威胁建模，因为可以依据真实的时间模式和网络结构来近似适应性与串谋行为，并利用由此得到的观察结果完善 Θ 的代理指标定义和更新规则。

6 结论

我们研究了如何设计激励与聚合机制，在真值难以观测且社交信号充满噪声或遭到操纵时，提升集体知识的准确率与鲁棒性。我们提出了可信度治理（CG），通过公共证据代理指标中的势头重新分配影响力，并通过系统动力学、消融实验、诊断性失效模式指标、噪声压力测试和对抗性威胁模型对其进行评估。在证据仍具信息价值的各种情形下，与由人气驱动和权益放大的基线方法相比，CG 能减少早期锁定和路径依赖、缩短纠正滞后，并在对抗压力下扩大稳定运行区域；同时也阐明了一项关键边界条件：当噪声变得极端时，所有机制的表现都会退化至随机水平。我们对现实世界可观测信号的映射，以及所提出的基于轨迹的验证计划，为在真实信息生态系统中检验这些主张、并借助基于领域的代理指标完善威胁模型提供了途径。

参考文献

[1] Lisa P. Argyle、Ethan C. Busby、Nancy Fulda、Joshua R. Gubler、Christopher Rytting 和 David Wingate。2023。《由一到多：使用语言模型模拟人类样本》。《政治分析》31，3（2023），337–351。[2] Eytan Bakshy、Solomon Messing 和 Lada Adamic。2015。《脸书上接触意识形态多元的新闻与观点》。《科学》（2015）。[3] Eytan Bakshy、Solomon Messing 和 Lada A Adamic。2015。《脸书上接触意识形态多元的新闻与观点》。《科学》348，6239（2015），1130–1132。[4] Abhijit Banerjee。1992。《羊群行为的简单模型》。《经济学季刊》（1992）。[5] Sushil Bikhchandani、David Hirshleifer 和 Ivo Welch。1992。《作为信息级联的时尚、风潮、习俗与文化变迁理论》。《政治经济学杂志》（1992）。[6] Katie Bratton 等。2023。《多智能体大语言模型系统中的涌现社会行为建模》。载于 NeurIPS。[7] Vitalik Buterin。2014。《权益与共识》。《以太坊博客》。[8] Zhenyu Chen 等。2024。《AgentVerse：用于多智能体大语言模型模拟的灵活框架》。载于 ICLR。[9] Philip Daian 等。2020。《Flash Boys 2.0：抢跑交易、交易重排序与共识不稳定性》。载于 IEEE S&P。[10] Michela Del Vicario 等。2016。《错误信息时代的回声室》。《PNAS》（2016）。[11] Joshua Epstein 和 Robert Axtell。1996。《成长中的人工社会》。麻省理工学院出版社。[12] Timothy Frye。2021。《信任品、信号传递，以及货币与信任之间的联系》。《比较政治研究》（2021）。[13] Alvin Goldman。1999。《社会世界中的知识》。牛津大学出版社。[14] Benjamin Golub 和 Matthew O Jackson。2010。《社交网络中的朴素学习与群体智慧》。《美国经济学杂志：微观经济学》2，1（2010），112–149。[15] Benjamin Golub 和 Matthew O. Jackson。2010。《社交网络中的朴素学习与群体智慧》。《美国经济学杂志：微观经济学》（2010）。[16] Robin Hanson。2006。《信息市场中的违规行为》。《经济学杂志》（2006）。[17] Buwei He、Yang Liu、Zhaowei Zhang、Zixia Jia、Huijia Wu、Zhaofeng He、Zilong Zheng 和 Yipeng Kang。2025。《提出一个他们无法拒绝的条件：在没有预先承诺的情况下，将贝叶斯劝说植根于现实世界对话》。arXiv 预印本 arXiv:2510.13387（2025）。[18] Yipeng Kang、Junqi Wang、Yexin Li、Mengmeng Wang、Wenming Tu、Quansen Wang、Hengli Li、Tingjun Wu、Xue Feng、Fangwei Zhong 和 Zilong Zheng。2025。《大语言模型的价值观在结构上与人类一致吗？一种因果视角》。载于《计算语言学协会 ACL 2025 年会论文集》，Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编）。计算语言学协会，奥地利维也纳，23147–23161。doi:10. 18653 1 2025 1188 /v / .findings-acl。Yipeng Kang、Tonghan Wang 和 Gerard de Melo。2020。《将语用推理融入涌现语言交流》。《神经信息处理系统进展》33（2020），10348–10359。[19] [20] Jon Kleinberg 和 David Easley。2022。《复杂世界中的信任：社会资本的网络理论》。《ACM 通讯》（2022）。[21] Jure Leskovec、Daniel Huttenlocher 和 Jon Kleinberg。2010。《预测在线社交网络中的正向与负向连接》。载于 WWW。[22] Hengli Li、Zhaoxin Yu、Qi Shen、Chenxi Li、Mengmeng Wang、Tinglang Wu、Yipeng Kang、Yuxuan Wang、Song-Chun Zhu、Zixia Jia 等。2025。《BEDA：将信念估计作为概率约束，以执行策略性对话行为》。arXiv 预印本 arXiv:2512.24885（2025）。[23] Jan Lorenz、Heiko Rauhut、Frank Schweitzer 和 Dirk Helbing。2011。《社会影响如何削弱群体智慧效应》。《PNAS》108，22（2011），9020–9025。[24] Yihuan Mao、Yipeng Kang、Peilun Li、Wei Xu 和 Chongjie Zhang。2025。《用于交流型多智能体系统零样本鲁棒性的不完美拜占庭将军问题》。载于国际通用人工智能会议。Springer，421–432。[25] Lev Muchnik、Sinan Aral 和 Sean Taylor。2013。《社会影响偏差：一项随机实验》。《科学》（2013）。[26] Lev Muchnik、Sinan Aral 和 Sean J Taylor。2013。《社会影响偏差：一项随机实验》。《科学》341，6146（2013），647–651。[27] Joon Sung Park、Joseph O'Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。2023。《生成式智能体：人类行为的交互式模拟》。载于第36届 ACM 用户界面软件与技术年度研讨会论文集。1–22。[28] Gordon Pennycook 和 David Rand。2021。《假新闻的心理学》。《认知科学趋势》（2021）。[29] Paul Resnick 和 Richard Zeckhauser。2000。《声誉系统》。《ACM 通讯》（2000）。[30] Matthew Salganik、Peter Dodds 和 Duncan Watts。2006。《人工文化市场中不平等与不可预测性的实验研究》。《科学》（2006）。[31] Patrick Schramowski 等。2023。《作为模拟社会的语言模型》。《科学报告》（2023）。[32] Chandler Smith、Marwa Abdulhai、Manfred Diaz、Marko Tesic、Rakshit S Trivedi、Alexander Sasha Vezhnevets、Lewis Hammond、Jesse Clifton、Minsuk Chang、Edgar A Duéñez-Guzmán 等。2025。使用 Concordia 评估基于大语言模型的智能体在混合动机场景中的泛化能力。*arXiv* 预印本 *arXiv:2512.03318*（2025）。[33] Ethan Squires 等。2021。同行评审评分系统如何影响科学共识？《科学进展》（2021）。[34] Cass R Sunstein。2006。《信息乌托邦：多少个头脑才能产生知识》。牛津大学出版社。[35] Soroush Vosoughi、Deb Roy 和 Sinan Aral。2018。真假新闻在网上的传播。《科学》（2018）。[36] Junyu Zhang、Yipeng Kang、Jiong Guo、Jiayu Zhan 和 Junqi Wang。2026。BACH-V：弥合大语言模型中的抽象与具体人类价值。*arXiv* 预印本 *arXiv:2601.14007*（2026）。[37] Zhaowei Zhang、Xiaobo Wang、Minghua Yi、Mengmeng Wang、Fengshuo Bai、Zilong Zheng、Yipeng Kang 和 Yaodong Yang。2025。PoliCon：评估大语言模型实现多元政治共识目标的能力。*arXiv* 预印本 *arXiv:2505.19558*（2025）。[38] Zhou Ziheng、Jiakun Ding、Zhaowei Zhang、Ruosen Gao、Yingnian Wu、Demetri Terzopoulos、Yipeng Kang、Fangwei Zhong 和 Junqi Wang。2026。简单的角色分配对安全对齐极为有效。*arXiv* 预印本 *arXiv:2602.00061*（2026）。[39] Zhou Ziheng、Huacong Tang、Mingjie Bi、Yipeng Kang、Wanying He、Fang Sun、Yizhou Sun、Ying Nian Wu、Demetri Terzopoulos 和 Fangwei Zhong。2025。一种用于研究道德演化的基于大语言模型的智能体仿真方法。*arXiv* 预印本 *arXiv:2509.17703*（2025）。

A 关键参数

实验中使用的具体参数值汇总于表 A。我们按适用范围对其进行分类（全局参数、感知相关参数和机制特定参数）。

诊断指标与失效模式指标。为使鲁棒性结果易于解读，我们报告了一组精简的诊断指标，用于捕捉噪声社会信号下集体学习的经典失效模式。

早期锁定效应。我们测量系统早期锁定于错误主题且未能恢复的概率。具体而言，如果虚假主题在前 T_{early} 轮次内达到高支持度阈值 τ，且运行结束时真实主题的最终支持度低于 τ，则将该轮实验计为早期锁定失效。我们报告满足这一标准的实验占比，作为早期锁定效应的指标。

路径依赖。我们量化最终结果对早期随机波动的依赖程度。具体而言，我们进行配对试验，这些试验具有相同的初始条件，但仅在随机感知噪声上有所不同；随后计算真实主题最终支持度的离散度，例如 $\text{PD} = \text{Var}\left[s_{(r)\,\text{真实}}(T)\right]$（跨重复运行 r）。离散度越高，表明路径依赖越强。

纠正滞后。我们衡量系统在证据不再支持当前主导主题后，纠正自身所需的速度。令 $t^\star$表示真实主题在公共证据中开始占优的首个轮次（等价地，即证据代理指标 Θ_{true} 超过 Θ_{false} 达 δ的轮次）。纠正滞后定义为真实主题的支持度达到目标阈值 τ所需的轮次数，即 $\text{Lag} = \min\{t \geq t^\star : s_{\text{true}}(t) \geq \tau\} - t^\star$；绘图时，我们将其归一化为 $[0, 1]$。

影响力集中。我们衡量影响力总量是否集中到少数智能体上。设 $w_i(t)$表示第 t轮中智能体 i的归一化影响力权重（总和为 1）。我们报告（按时间平均的）赫芬达尔指数 $\text{IC} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} w_i(t)^2$，，其中数值越大表示集中度越高。

参数	符号	值	范围 / 定义-说明
物理进展动力学			
$\Delta\pi_k^t = \begin{cases} \epsilon_k^t, & \pi_k^{t-1} < \pi_1, \\ (v_k + \gamma\, r_k^t)\left(1 - \frac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_1 \le \pi_k^{t-1} < \pi_2, \\ \gamma' r_k^t\left(1 - \frac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_k^{t-1} \ge \pi_2. \end{cases}$			
环境噪声	σ	0.1	ϵ^t的标准差$_k$
基线速度	$v_{\text{true}}, v_{\text{false}}$	1.0, 0.8	内在的增长
加速因子	γ	0.8	中期 re-来源影响
饱和限值	$M_{\text{true}}, M_{\text{false}}$	10, 8	最大进展
后期阶段因子	γ'	0.4	后期阶段 re-来源影响
阶段阈值	π_1, π_2	2.0, 5.0	阶段 转折点
智能体感知模型			
$\tilde{\pi}_k^t = \pi_k^{t-1} + \rho\,\Theta_k^{t-1} + \xi_k^t$			
社会污染	ρ	0.2	社会塑造的强度
观测噪声	ξ_k^t	0.05	感知噪声的标准差
可信度治理（可信度治理）特定			
$w_i^{t-1} = \alpha_i^t e^{\lambda c_i^{t-1}}$			
影响力集中	λ	2.0	指数型in-响力缩放
$\Theta_k^t = (1-\lambda_s)\Theta_k^{t-1} + \lambda_s\left[(r_k^t - r_k^{t-1})\ \underline{q}_k^t - \gamma B_k^t\right]$			
社会信号平滑	λ_s	0.4	平滑 of 社会 信号更新
$c_i^t = c_i^{t-1} + \eta\left(\Theta_{a_i^t}^t - \Theta_{a_i^t}^{t-1}\right) e^{-\kappa r_{a_i^t}^t}$			
可信度学习率	η	0.1	可信度速-性变化
先行者优势	κ	0.5	先行者奖励因子
$\underline{q}_k^t = \frac{\sum_{i:a_i^t=k} c_i^{t-1}}{\sum_{i:a_i^t=k} 1}, \quad B_k^t = \sigma\left(\alpha(r_k^t - r_k^{t-1})\right)(1 - q_k^t)$			
不一致惩罚	β	0.3	惩罚 for 切换
Web3质押（***WS***）专用			
$w_i^{t-1} = \alpha_i^t\,\text{bal}_i^{t-1}, \quad \text{bal}_i^t = \text{bal}_i^{t-1} + \gamma_s w_i^{t-1}\Delta\pi_{a_i^t}^t$			
质押奖励率	γ_s	0.1	比例 of 进度 paid 作为质押奖励

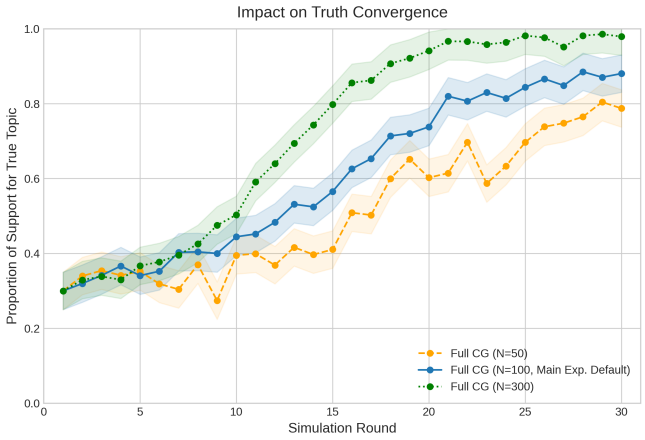

图5：对群体规模的敏感性。比较 N=50、N=100 （主要实验）和 N=300下可信度治理的收敛情况。

基线与公平性改进。 我们在相同的仿真预算和评估方案下，将可信度治理（CG）与三种治理基线进行比较：Web3质押（WS）、社交媒体点赞（SM）和无治理（NG）。为确保压力测试公平，我们的 WS 实现纳入了含噪信息生态系统中出现的相同实际摩擦，即**(i) 含噪质押更新**和**(ii) 质押信息延迟可见**，从而避免在不切实际的清洁信号下评估质押权重。我们还加入了明确的**可信度提示**控制，通过仅改变智能体指令、同时保持机制和环境不变，隔离潜在的指令偏差。对于所有基线和可信度治理变体，我们都使用相同的搜索预算调节超参数，并采用相同的验证标准选择配置（即调优预算和选择指标完全一致）。

B 参数敏感性分析

B.1 对群体规模（N）的敏感性

图5展示了群体规模（N）对可信度治理收敛的影响。较大的群体（N=300）比N=100，尤其是N=50收敛得更快、更平稳。智能体数量越多，基于大语言模型的群体所产生的社会信号就越稳定：模型生成过程中的个体随机性会被平均化，可信度更新反映出一致的模式，而非噪声。更强的聚合信号使可信度治理能够更快地识别可靠智能体，并强化正确的主题。相比之下，较小的群体更容易受到单个大语言模型输出随机性的影响，导致社会信号稳定性较差，收敛速度也更慢。

B.2 对初始意见分布的敏感性

图6展示了初始意见分布如何影响可信度治理的收敛。随着与真相一致的智能体占比增加（30%、50%、70%），收敛速度更快，也更加稳定。当真相一开始就占多数时，可信度治理几乎会立即强化这一优势。这是因为真实主题会产生更加一致的改善

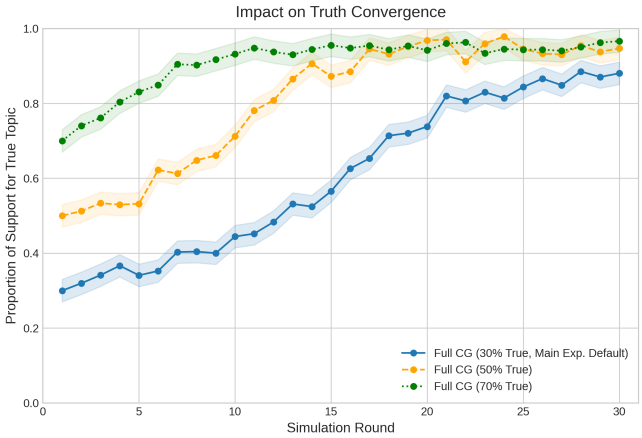

图6：对初始意见分布的敏感性。比较可信度治理在初始与真相一致的智能体占比为30%、50%和70%时的收敛情况。

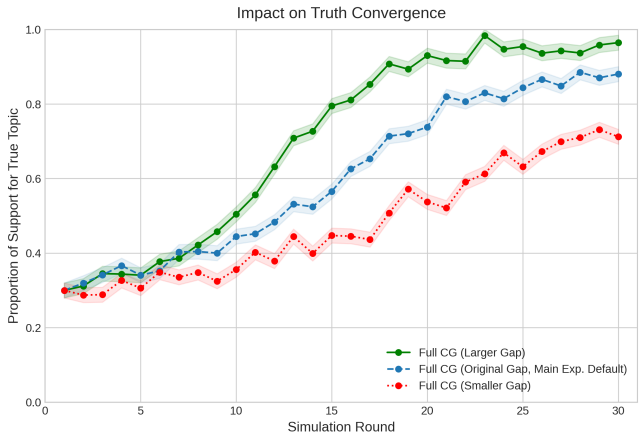

图7：对基准速度差的敏感性。比较可信度治理在较小、原始和较大内在增长差异下的收敛情况。

信号，使与真相一致的智能体从最初几轮起就获得略微积极的可信度更新。随着这些微小优势不断累积，真相支持者的平均可信度会在最初几轮内上升。一旦他们的平均可信度超过虚假信息支持者，真实主题的反泡沫项就会变得微不足道，因为它随（$1 - \bar{q}^t_{\text{true}}$）缩放。因此，真相多数派获得的是稳步提升，而非惩罚。总体而言，可信度治理不仅能克服不利的初始条件，还能迅速巩固已与真相一致的群体。

B.3 对基准速度差的敏感性

图 7 展示了可信度治理（CG）对真实主题与虚假主题之间内在质量差异的敏感性，该差异以基准速度差（v_{true} - v_{false}）量化。速度差越大，收敛速度越快，稳定性也显著增强。这是因为真实主题具有更明显的内在优势

意味着其内在质量更加清晰，使可信度治理的机制能够更有效地检测并放大这一信号，从而更高效地赋能其支持者。相反，速度差较小会导致收敛速度变慢、波动性增大，最终的真相支持度也可能较低。这是因为质量差异不明显时，可信度治理更难辨别真实主题的优势；这种优势也更容易被噪声、虚假主题初始资源优势以及个体智能体随机性掩盖。这表明，尽管可信度治理有效，但真值本身的可辨识程度仍会从根本上影响其收敛动态。

C 映射至现实世界可观测信号

表2将CG变量映射到三个数据丰富示例（预测、同行评审和论坛）中的具体、按时间索引的代理指标。可信度治理仅需要（i）一个公开社会信号 Θ^t_k（及其变化 $\Delta\Theta^t_k$）以及（ii）一个与结果相关的信号 π^t_k，该信号伴随噪声与延迟到达（通常通过代理指标 $\tilde{\pi}^t_k$ 观测）。

D 稳健性测试

D.1 H4：可信度治理在中等噪声下最为有益，通过减少锁定效应并加快纠正

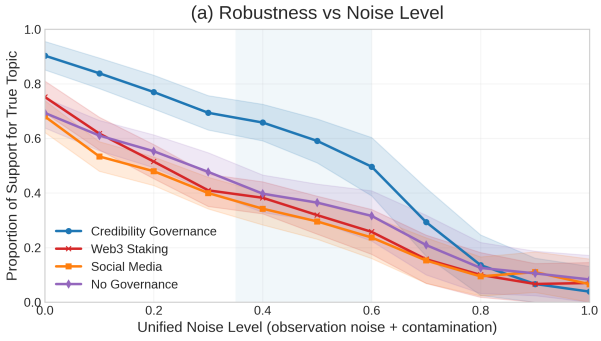

图8：H4，鲁棒性与噪声。我们扫描结合观测噪声与污染的统一噪声水平。可信度治理在中等噪声区间显示出最大的准确率提升，此时真值虽仅能被微弱观测，但仍可恢复。在极端噪声下，所有机制均会退化。

假设（H4）。 可信度治理的优势在中等噪声下达到峰值，此时朴素流行度机制会锁定早期错误，但证据仍具有足够的信息量以进行纠正。

测试。我们扫描统一噪声水平，并测量准确率（对真实主题的支持），以及早期锁定、路径依赖和纠正滞后的诊断指标。

结果。 图1显示，在中等噪声下，可信度治理带来的提升最大，而所有方法在极端噪声下的表现均有所下降。图9显示，在相同噪声区间内，可信度治理降低了早期锁定和路径依赖，并缩短了纠正滞后。这表明，可信度治理通过抑制级联驱动的失效模式并加快纠正，提高了准确率。

表2：将可信度治理变量映射到预测、同行评审和社区论坛中的可观测代理指标示例。

CG变量	预测竞赛	同行评审	社区论坛
Θ_k^t （社会信号）	关于事件 k的群体总体信念，例如平均隐含概率或加权预测共识在时间 t。	对论文 k的总体立场，例如平均推荐——荐分数、录用概率代理指标，或讨论情绪。	对主张或话题的总体支持度 k，例如：赞成票投票占比、认可率、同意比例，或转发强度。
$\Delta\Theta_k^t$ (势头)	共识概率在一个窗口内的变化，表示信息到达或群体修正。	审议期间共识的转变，例如分数在反驳或情绪波动后发生变化，between轮次。	认可率或投票占比的变化，captur-形成逐渐形成的共识或协同激增。
π_k^t （结果，真值相关）	已实现事件结果，已确定的二元或连续结果（例如，事件是否发生）。	发表后结果，例如可复现性检查、后续专家评估或编辑裁定。	后续验证信号，例如事实核查标签、审核员确认或权威来源验证。
$\tilde{\pi}_k^t$ （模糊结果代理）	含噪短期代理，例如部分指标、早期报告或延迟结算状tors、早期报告或延迟结算状态。	含噪代理，例如审稿人信心度、独立快速核查或编辑初筛信号。	报告、标记或有限审计等带噪声的代理信号，且这些信号存在延迟。
w_i^t (可信度权重)	用于汇总预测的预测者可靠性权重基于历史校准和更新动态。	用于汇总或分配的审稿人可靠性权重，基于历史一致性和信息量。	用于排名或可见性的用户可信度权重，基于过往的帮助程度和抵御操纵的能力。
r_k^t （资源分配）	分配给事件 k的注意力，例如信息显著性或预测可见度权重。	对论文 k投入的编辑精力，例如额外的审查带宽、专家咨询或优先处理。	分配给主题或主张的曝光度 k，例如，排名-位置、推荐频率，或分-配预算。
B_k^t （反信息茧房惩罚）	针对失控的从众效应的正则化项，例如，下调缺乏佐证的剧烈波动的加权。	防止级联接受的正则化项，例如要求多样化证据或独立论证，才能进行推广。	防止病毒式级联的正则化项，例如在参与度增长超过证据增长时减少放大。
噪声与延迟	结果需要一段时间才能揭晓，临时信号可能含有噪声或受到恶意影响。	真实情况需要一段时间才能明了，审阅者信号含有噪声、存在偏见，并受到社会影响。	验证需要一段时间，参与度信号可能受到协同行动者操纵。

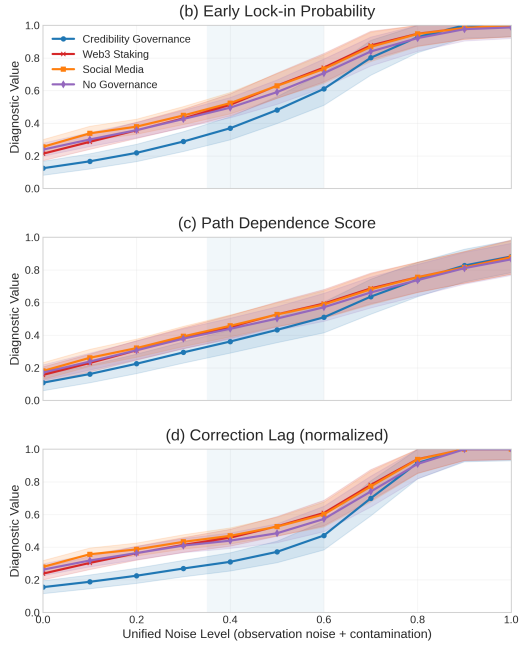

图 9：H4，噪声条件下的诊断指标。随着噪声增加，我们跟踪早期锁定概率、路径依赖评分和纠正滞后（归一化）。在中等噪声区间，可信度治理最显著地降低了锁定效应和路径依赖，并缩短了纠正滞后。

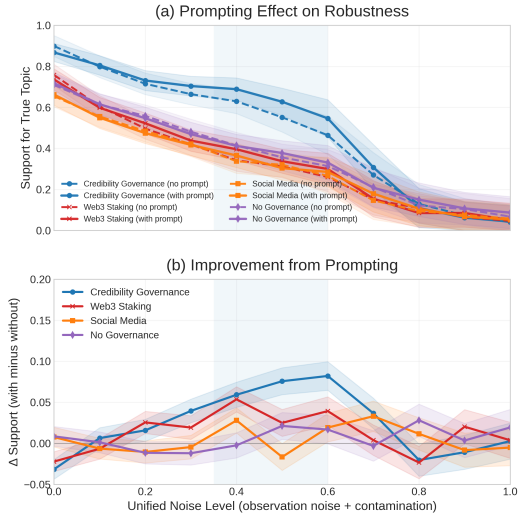

图10：H5，提示效应。实线表示采用显式可信度提示的运行结果，虚线表示未采用提示的运行结果。提示带来了幅度不大但稳定的提升，其中在中等噪声下提升最大。

D.2 H5：显式可信度提示可带来适度的额外收益

假设（H5）。显式提示智能体考虑可信度，在激励设计的基础上可带来额外收益，尤其是在信号含糊不清时。

测试。我们比较在噪声扫描中采用与不采用显式可信度提示时的结果。

结果。 图 10 显示，采用提示后性能虽有所提升，但幅度不大且表现稳定，并在中等噪声下达到峰值。这表明，可信度治理的激励已能促成有效的信用分配，而提示进一步增强了系统在模糊情境下推理的稳定性。

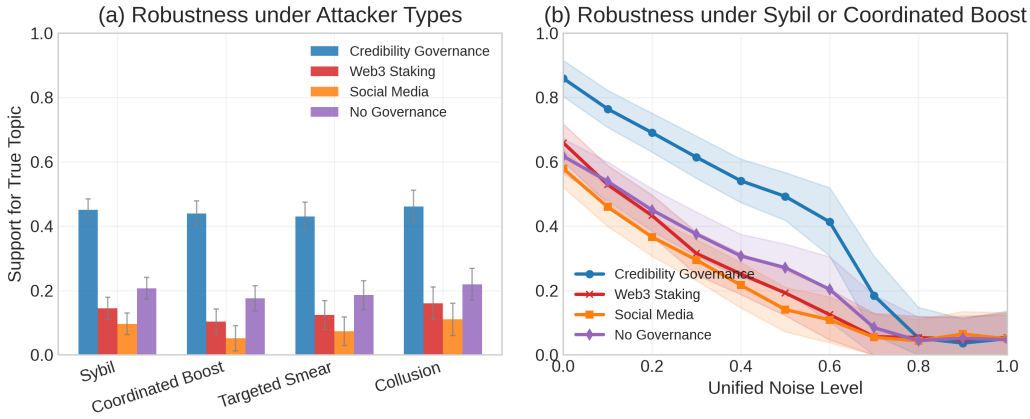

图11：H6，威胁模型。(a) 在代表性噪声水平下，对多种攻击者行为的鲁棒性。(b) 在代表性攻击类型下，鲁棒性随噪声水平的变化。与WS和SM相比，可信度治理在不同攻击者类型下的性能下降幅度较小；WS和SM对会制造人为动量的操纵更为敏感。

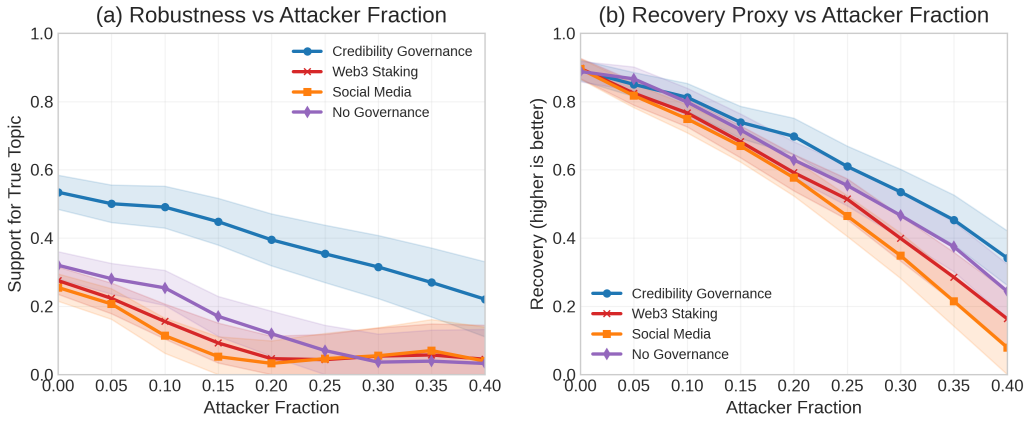

图12：H6，攻击者比例扫描。(a) 在代表性噪声水平下，准确率随攻击者比例的变化。(b) 恢复代理指标随攻击者比例的变化。在崩溃之前，可信度治理能够容忍比例更高的对抗性群体，并且在攻击者占比更高时仍能保持更好的恢复能力。

D.3 H6：可信度治理在对抗性操纵下扩大安全运行区域

假设（H6）。与动量放大基线方法相比，可信度治理对策略性操纵的鲁棒性更强，并能扩大运行范围，使集体学习在噪声与对抗压力共同作用下仍保持稳定。

测试。我们评估了：(i) 多种攻击者行为（威胁模型）；(ii) 攻击者占比扫描；以及 (iii) 噪声水平与攻击者比例构成的二维失效边界。

结果。图 11 显示，可信度治理在不同攻击者类型下的性能下降幅度较小。图 12 显示，在准确率崩溃之前，可信度治理能够容忍更高的对抗性比例，并且恢复效果更好。图 13 总结了联合边界，并显示可信度治理相比 WS 具有更大的稳定运行区域，使边界向更高的噪声水平和更高的攻击者比例移动。

摘要。在各种假设下，可信度治理通过利用公共证据中的动量，将影响力重新分配给与真相一致的智能体，从而提升集体准确性；同时抑制早期锁定和路径依赖，并增强系统在噪声和对抗性操纵下的鲁棒性。

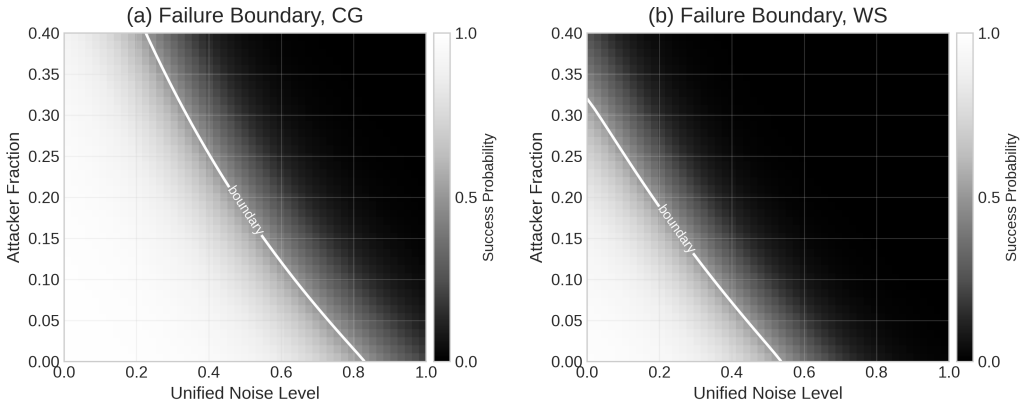

图13：H6，噪声和攻击者比例的失效边界。热图显示了成功概率随统一噪声和攻击者比例变化的情况，白色等高线表示边界（成功概率 $= 0.5$）。与 WS 相比，可信度治理将边界推向更高的噪声水平和更高的攻击者比例，从而扩大稳定运行区域。