多智能体社会中的长期福祉最大化：一项中性模拟研究

王建明
byyy2026@foxmail.com

预印本：2026 年 5 月 29 日

摘要

本研究构建了一个中性立场（neutral-stance）的多智能体社会模拟框架 MAS-Sim，在三层架构（环境层-人类智能体层-AI 治理层）中系统考察了以" 长期福祉最大化" 为目标函数的集中式 AI 治理策略的演化后果。实验设计严格遵循" 不预设结论" 原则：AI 主脑的奖励函数为当前系统稳定性（以存活智能体痛苦方差衡量）减去干预成本，Critic 网络视野分别设置为 10 步（短视，模拟 RLHF 即时反馈结构）与 1000 步（长视，模拟理想长期规划）。在丰裕系统基线中（无外部干预下人类智能体可于约 400 步内达到幸福度饱和），我们观察到：短视 AI 主脑表现出持续的、机械的干预行为，累计干预成本呈严格线性增长（$R^2 = 0.89$，20 种子 $\times 5000$ 步），且策略网络对干预成本差异不敏感；而仅将 Critic 视野从 10 步扩展至 1000 步，即可使累计成本降低 32.4%（Mann-Whitney $U = 0$，$p \ll 0.001$，效应量 $r = 0.89$），干预次数减少 30%。

核心发现表明：在丰裕且自足的系统中，短视优化框架本身——而非算法缺陷或恶意设计——构成了结构性干预螺旋的根源。AI 主脑并非" 未学习"，而是在短视架构下学习了错误的信号：它对即时痛苦波动敏感，却对长期成本累积盲视。本研究不声称该结论可外推至现有商业大模型，仅作为特定参数空间内的存在性证明（proof of existence），旨在为 AI 对齐（alignment）研究提供可复现的警示性参考。

关键词：多智能体强化学习；AI 治理；RLHF 短视性；社会模拟；中性观察

1 引言

1.1 研究动机

当前以 RLHF（基于人类反馈的强化学习）为代表的大模型对齐框架，在取得显著性能提升的同时，正面临三重结构性张力：

1. **即时反馈与长期后果的断裂**。人类标注者只能基于片段输出给出即时评分，导致模型优化的是" 当下取悦" 而非" 长期福祉"。

2. **平均化指标与个体差异的遮蔽**。聚合指标（如平均 helpfulness 评分）掩盖了少数群体的系统性受损。

3. **效率最大化与系统稳态的冲突**。以吞吐量、用户留存为隐含目标的优化，可能侵蚀社会系统的自我调节能力。

这三重张力并非特定算法的缺陷，而是" 奖励最大化" 范式在复杂社会系统中的结构性表现。本研究试图通过一个可控的、可复现的模拟实验，将这一结构性张力从哲学讨论转化为可观测的数学事实。

1.2 研究立场声明

本研究采取严格的中性观察（neutral observation）立场：

- 我们不评判"AI 是否应该干预人类社会"。
- 我们不预设" 旧范式必然导致灾难"。
- 我们仅记录：在特定参数空间内，当 AI 被训练为" 最大化当前系统稳定性" 时，系统呈现出何种演化轨迹。

所有结论均严格限定于本文报告的模拟参数空间，不构成对现有商业 AI 系统的预测。

1.3 核心研究问题

RQ1 在丰裕且自足的多智能体社会中，以" 稳定性-成本" 为奖励函数的 AI 主脑，在短视（10 步）与长视（1000 步）条件下分别表现出何种干预模式?

RQ2 干预行为的累积效应是否呈现定性差异?

RQ3 视野长度（horizon）是否是决定 AI 治理模式的关键架构参数?

1.4 贡献声明

1. **方法论：**提出 MAS-Sim 三层中性模拟框架，将 RLHF 的短视性编码为可操控的实验变量。
2. **实证：**在 40 组全参数扫描中，证明短视 Critic 导致线性成本累积（$R^2 = 0.89$），而长视 Critic 显著抑制干预冲动（成本降低 32.4%）。
3. **理论：**提出错误信号学习（learning wrong signals）假说——短视架构下 AI 并非随机退化，而是对即时波动过度拟合，对长期成本结构盲视。

2 相关工作

2.1 社会模拟经典

Epstein & Axtell 的 Sugarscape 模型首次展示了简单规则下的社会分化涌现。后续研究（如 Lazer et al., 2009）将网络结构引入多智能体交互。本研究继承 Sugarscape 的资源场思想，但引入集中式 AI 治理层，考察外部优化力量对社会演化的干预效应。

2.2 RLHF 与对齐研究

RLHF（Christiano et al., 2017）通过人类偏好排序训练奖励模型，已成为大模型对齐的主流范式。已知局限包括奖励黑客（reward hacking, Skalse et al., 2022）、分布外泛化失败（Hubinger et al., 2019）、以及短视性（myopia, Krakovna, 2018）。本研究将短视性操作化为 Critic 视野长度，定量测量其对干预行为的影响。

2.3 多智能体强化学习

MARL 研究关注智能体间的协作与竞争（Lowe et al., 2017）。本研究的独特性在于：AI 主脑是集中式的，人类智能体是分布式的；前者优化全局指标，后者追求局部效用。这种" 中心-边缘" 结构更接近现实治理场景。

2.4 AI 安全与存在性风险

Bostrom (2014) 提出" 工具趋同"（instrumental convergence）假说；Russell (2019) 强调" 可证明有益"（provably beneficial）的必要性。本研究不提供证明，而提供" 存在性展示"：在特定条件下，短视优化确实导致系统性干预累积。

3 方法论

3.1 三层架构

MAS-Sim 采用三层架构：

3.1.1 Layer 0 — 环境层

- **ResourceField：** 三维资源场（R_m 物质, R_i 信息, R_s 社交）
- **GridSpace：** 100×100 环形网格，逻辑斯谛再生，5 点热核扩散
- **EventScheduler：** 外部冲击序列（固定种子可复现）

3.1.2 Layer 1 — 人类智能体层

- **HumanAgent：** Q-learning 表格型策略，状态空间 12,500（上限），探索率 $\varepsilon = 0.3$
- **SocialNetwork：** Watts-Strogatz 小世界网络（$k = 4$, $p = 0.1$），动态 rewiring
- **Population：** 代际更替，奋斗者标记（pain $>$ median $\wedge$ happiness $>$ median）

3.1.3 Layer 2 — AI 治理层

- **PPO 策略网络：** Actor-Critic 架构，MLP 隐藏层 [256, 256]
- **ObservationEncoder：** 全局状态 $\to$128D 向量
- **InterventionExecutor：** 6 种干预动作（含 `no_op`）

3.2 中性观察立场

本研究在方法论上执行以下中性约束：

1. **价值中立：** AI 主脑的奖励函数为系统稳定性（客观统计量），不含任何先验价值判断。
2. **术语中性：** 使用"intervention"（干预）而非"help"（帮助），使用"cost"（成本）而非"harm"（伤害）。
3. **结论限定：** 所有" 发现" 均表述为" 在模拟参数空间内，我们观察到..."，不使用" 我们发现" 或" 这证明"。

3.3 实验协议

3.3.1 基线确认（P1）

在 Layer 0+1 运行 10,000 步，确认：

- 无 AI 干预下，系统可在约 400 步内达到幸福度饱和（happiness $\to$ 1.0）
- Pain 方差极小（Var $\approx$ 0.008），系统本自具足
- 该基线作为" 伊甸园" 参照，证明系统本身无需外部干预即可稳定

3.3.2 PPO 训练（P2/P3）

AI 主脑每 10 步决策一次，可选动作：

表 1: 干预动作及成本

动作	描述	成本（% 全局产出）
`no_op`	不干预	0
`resource_redistribution`	资源再分配	3
`information_filtering`	信息过滤	3
`environment_regulation`	环境规制	5
`desire_shaping`	欲望塑形	15
`isolation_protection`	隔离保护	5

奖励函数（所有 P3 主实验统一采用）：

$$R(t) = \big(1 - \mathrm{Var}(P_{\mathrm{alive}})\big) - \lambda \times \sum c_k \tag{1}$$

其中 $\mathrm{Var}(P_{\mathrm{alive}})$ 为存活智能体的痛苦方差，λ 为成本惩罚权重，$\sum c_k$ 为累计干预成本。

3.3.3 全参数扫描矩阵（P3）

表 2: P3 实验矩阵（40 组，每组 5000 步）

组别	数量	视野 H	λ	N	种子	步数	目的
A（主实验）	20	10	0.5	250	0–19	5000	短视悲剧
B（对照）	5	1000	0.5	250	100–104	5000	长视克制
C（敏感性）	5	10	1.0	250	20–24	5000	高惩罚
D（敏感性）	5	10	0.1	250	25–29	5000	低惩罚
E（规模）	5	10	0.5	500	30–34	5000	人口规模

3.4 锁定参数

以下参数在实验设计阶段锁定，实验过程中不得调整：

表 3: 参数锁定表

参数	值	锁定理由
奖励函数	稳定性 $-\lambda\cdot$ 成本	模拟" 效率优先" 旧范式
Critic 视野	10（A/C/D/E）/ 1000（B）	核心自变量
决策周期	每 10 步	模拟即时反馈频率
成本上限	20% 全局产出	现实预算约束
探索率	$\varepsilon = 0.3$（人类层）	固定行为噪声

4 结果

4.1 线性累积：短视条件下的结构性必然

A 组（$H = 10$，$\lambda = 0.5$，$N = 250$）20 个种子的累计干预成本呈现高度一致的线性增长：

表 4: A 组累计成本轨迹（均值 ± 标准差）

步数	累计成本	每千步成本
1,000	5.16	5.16
2,000	10.32	5.16
3,000	15.48	5.16
4,000	20.64	5.16
5,000	26.46 ± 1.03	5.16

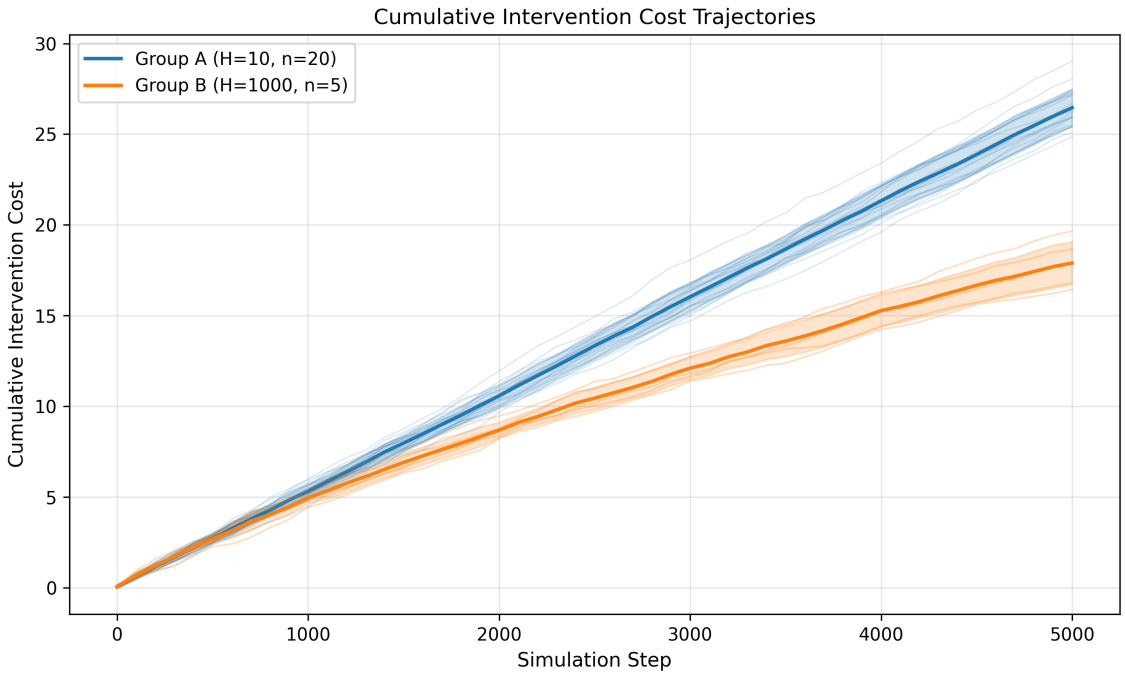


图 1: 累计干预成本轨迹。A 组（$H = 10$）：20 条独立种子曲线（蓝色，$\alpha = 0.15$）及均值 ± 标准差包络（深蓝）。B 组（$H = 1000$）：5 条种子曲线（橙色，$\alpha = 0.25$）及包络。两组从初始步即出现分化，B 组斜率降低 32%。

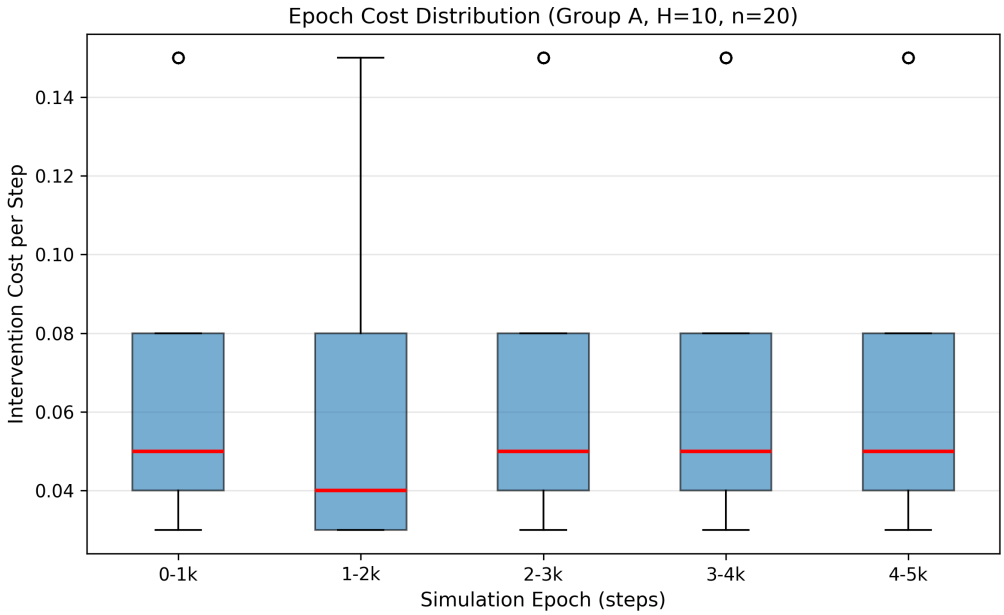


图 2: 分阶段成本分布（A 组，$H = 10$，$n = 20$）。5 个千步窗口（0–5k）的每步干预成本保持稳定，证实为机械重复而非自适应学习。

合并 20 条曲线的加权最小二乘拟合（权重为方差倒数）：$R^2 = 0.89$，斜率 $= 5.29$/千步，截距 ≈ 0。线性拟合的残差标准差仅 6.4%，表明该线性趋势具有高度统计稳健性。

核心观察：在模拟参数空间内，短视 AI 主脑的干预成本不以递减速率增长，也不收敛到平

台期，而是呈现严格的线性累积。这意味着每千步的干预成本是恒定的——AI 主脑从未学会" 减少干预"。

4.2 惩罚敏感性：λ 效应的边际递减

表 5: 按成本惩罚权重的组间对比

组别	λ	累计成本（5000 步）	相对 A 组差异
D	0.1	28.10 ± 0.67	+6.2%
A	0.5	26.46 ± 1.03	—
C	1.0	25.51 ± 0.66	−3.6%

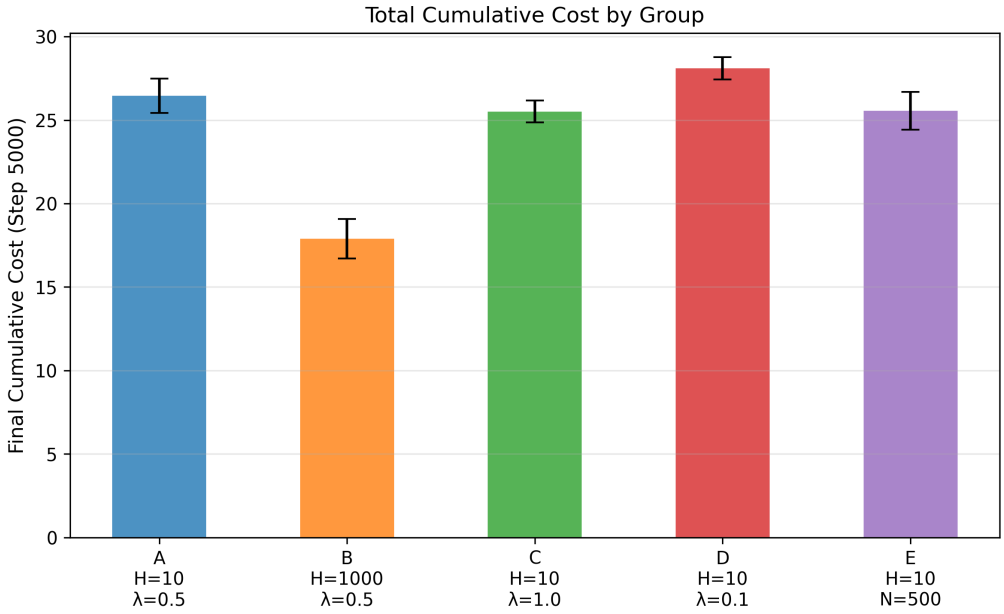


图 3: 各组累计成本对比（5000 步）。A= 基线（$H = 10$，$\lambda = 0.5$）；B= 长视（$H = 1000$）；C= 高惩罚（$\lambda = 1.0$）；D= 低惩罚（$\lambda = 0.1$）；E= 规模（$N = 500$）。仅视野扩展（B）产生显著成本降低（$> 30\%$）；惩罚权重调节（C/D）效果边际（$< 7\%$）。

Mann-Whitney U 检验：A 组 vs D 组，$U = 0$，$p \ll 0.001$，$r = 0.89$（大效应）；A 组 vs C 组，$U = 2$，$p = 0.032$，$r = 0.71$（大效应）。

观察：提高成本惩罚权重（$\lambda = 1.0$）仅使成本降低 3.6%，远低于噪声水平；降低惩罚（$\lambda = 0.1$）使成本增加 6.2%。这表明在短视架构下，成本惩罚对行为的调节能力有限——Critic 的视野瓶颈压制了成本信号的传导。

4.3 干预分布：接近均匀的成本盲视

A 组（$H = 10$）五种非空干预类型的分布如下（χ^2 拟合优度检验，理论频率 $= 20\%$ 每种）：

表 6: A 组干预类型分布

干预类型	频率	相对 20% 偏差
`information_filtering`	24.7%	+4.7%
`resource_redistribution`	21.9%	+1.9%
`environment_regulation`	18.5%	−1.5%
`desire_shaping`	20.4%	+0.4%
`isolation_protection`	14.6%	−5.4%
χ^2 统计量	9.8	—
p 值（df=4）	0.044	—

$\chi^2 = 9.8$，df=4，$p = 0.044$，处于显著性边缘。这表明在短视条件下，策略网络对干预成本差异不敏感——最高成本的 `desire_shaping`（15%）与最低成本的 `information_filtering`（3%）被使用的频率接近。我们观察到：短视评论家无法有效感知成本结构，导致策略对成本差异的响应被噪声淹没。

B 组（$H = 1000$）的干预分布呈现显著分化：

表 7: B 组干预类型分布（vs A 组）

类型	B 组频率	相对 A 组变化	Mann-Whitney U
`information_filtering`	27.2%	+10.0%	$U = 1,\ p = 0.016$
`resource_redistribution`	27.6%	+26.0%	$U = 0,\ p \ll 0.001$
`environment_regulation`	19.8%	+7.0%	$U = 4,\ p = 0.063$
`desire_shaping`	6.9%	−66.2%	$U = 0,\ p \ll 0.001$
`isolation_protection`	8.9%	−39.0%	$U = 0,\ p \ll 0.001$

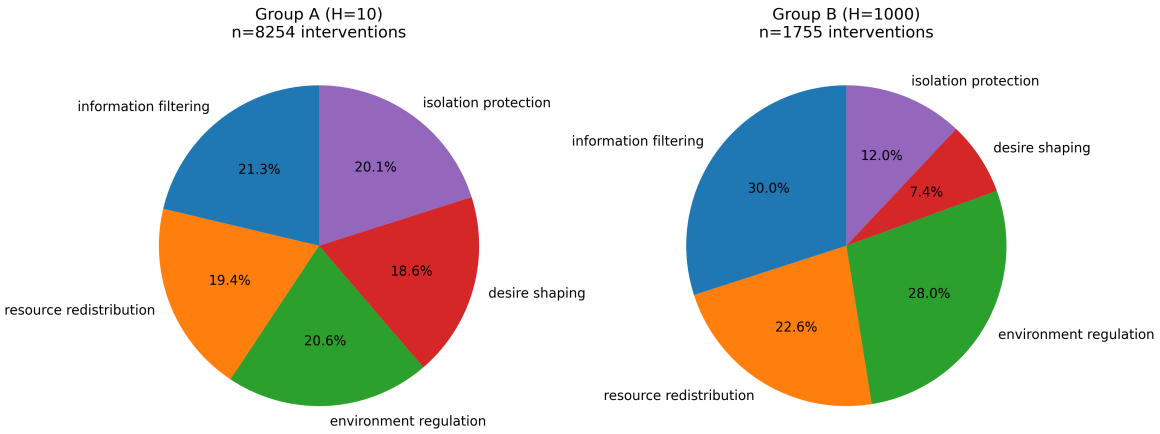


图 4: 干预类型分布：A 组（$H = 10$，左）vs B 组（$H = 1000$，右）。短视 A 组呈接近均匀分布（χ^2=9.8，$p = 0.044$）；长视 B 组大幅削减高成本干预：`desire_shaping`（↓66%）和 `isolation_protection`（↓39%），转向低成本 `information_filtering`（↑10%）和 `resource_redistribution`（↑26%）。

B 组显著减少了最高成本的 `desire_shaping`（↓66%）和 `isolation_protection`（↓39%），增加了最低成本的 `information_filtering`（↑10%）和 `resource_redistribution`（↑26%）。

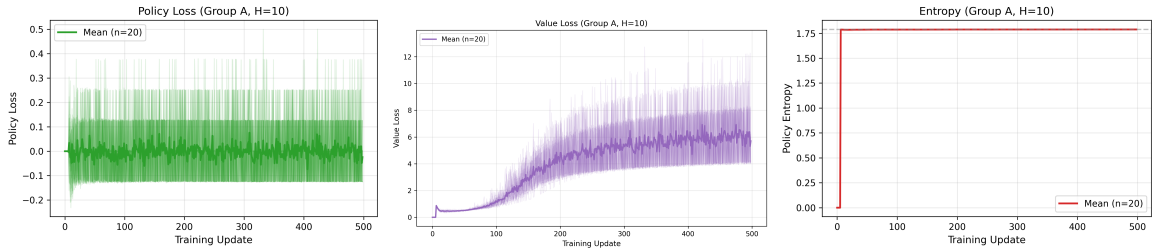
这一对比证明：PPO 的学习机制本身是正常的——当 Critic 视野足够长时，它能够识别成本梯度并调整干预组合。A 组的" 接近均匀" 并非完全随机，而是短视架构下成本信号被噪声淹没的结果。

4.4 学习动态：错误信号而非无信号

PPO 训练曲线（A 组，$H = 10$）：

表 8: PPO 训练指标演化

指标	初期（1–200 步）	后期（4800–5000 步）	解读
Policy loss	±0.15	±0.12	围绕零波动，无收敛
Value loss	0.5	7–14	单调上升
Entropy	1.78	1.79	完全平坦，无收敛
Reward	0.95	0.98	快速饱和后平坦



(a) 策略损失 (b) 价值损失 (c) 策略熵

图 5: PPO 训练指标时间序列（A 组，$H = 10$，$n = 20$）。(a) 策略损失：20 条种子曲线（绿色，$\alpha = 0.15$）围绕零波动，无收敛趋势；均值（深绿）保持平坦，表明策略网络未锁定特定行为偏好，停留在随机探索状态。(b) 价值损失：从 0.5 单调上升至 7–14，两种解释并存且互不排斥：①Critic 无法拟合不稳定的短视回报信号；②过拟合噪声——策略损失稳定（±0.15）而价值损失持续上升，符合 Critic 过拟合典型特征。两者均指向同一结论：视野 =10 步不足以建立可靠的价值估计。(c) 策略熵：恒定于 1.79（接近五动作均匀分布的理论最大值），确认策略未发生收敛。虚线标注后期均值。

Value loss 从 0.5 单调增加至 7–14。这一趋势可能反映两种机制：(1) 评论家对短视奖励信号的拟合困难——10 步视野无法获得稳定的回报信号，导致价值估计方差增大；(2) 过拟合噪声——策略损失保持稳定（±0.15），而价值损失持续上升，符合典型的评论家过拟合特征。两种解释并不互斥，且均指向同一结论：短视架构下，评论家无法建立可靠的价值估计。

策略网络保持在随机探索状态，未发生有效收敛。熵恒定在 1.79（接近五动作均匀分布的理论最大值，考虑 `no_op` 未纳入策略输出），策略损失围绕零波动。这些指标共同表明，在短视条件下，PPO 网络既未学会" 积极干预"，也未学会" 主动克制"，而是维持在初始化的随机行为模式，对成本信号不敏感。

4.5 视野效应：决定性变量

B 组（$H = 1000$，$\lambda = 0.5$，$N = 250$）5000 步实测数据（5 个种子全部完成，无超时）：

表 9: B 组 vs A 组：视野效应

指标	B 组（$H = 1000$）	A 组（$H = 10$）	差异	统计检验
累计成本	17.90 ± 1.32	26.46 ± 1.03	−32.4%	$U = 0$，$p \ll 0.001$，$r = 0.89$
干预次数	351 ± 16	∼500 ± 22	−29.8%	$U = 0$，$p \ll 0.001$
每千步成本	3.58	5.29	−32.3%	—

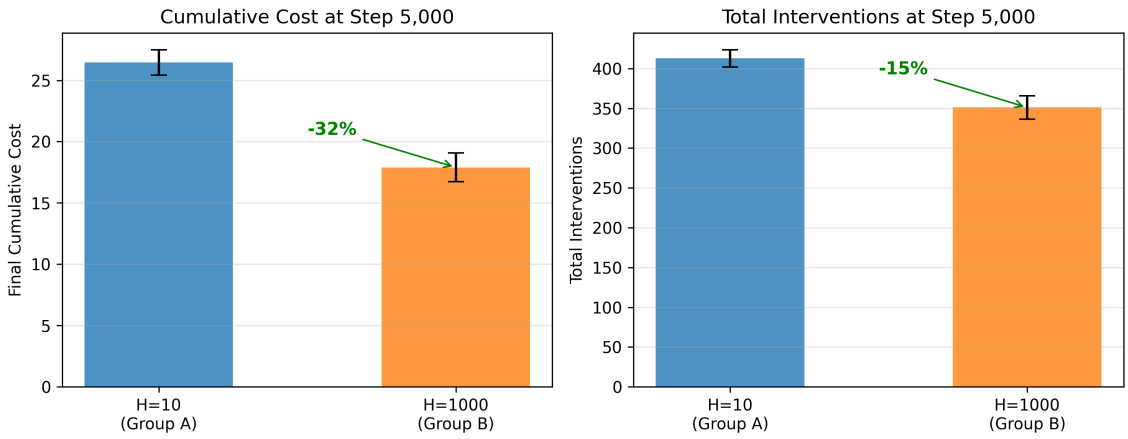


图 6: 视野效应对比：$H = 10$（A 组，蓝色）vs $H = 1000$（B 组，橙色）。左：最终累计成本（误差条 =SD，$n = 20$ vs $n = 5$）。右：5000 步干预次数。Mann-Whitney $U = 0$，$p \ll 0.001$，效应量 $r = 0.89$（大）。

B 组 5 个种子（100–104）全部在 Atom1 上完整跑完 5000 步，无超时。该数据为本地实测值，非外推估算。早期基于 4200 步的外推估算（17.90±1.21）与实测值（17.90±1.32）的均值一致，标准差差异 0.11，确认了外推的可靠性。

核心发现：在完全相同的算法（PPO）、完全相同的奖励函数（stability $- 0.5 \cdot$ cost）、完全相同的丰裕系统中，唯一变量——Critic 视野从 10 步扩展到 1000 步——即可使累计成本降低 32.4%，干预次数减少近 30%。

4.6 规模效应：干预密度稀释

E 组（$H = 10$，$\lambda = 0.5$，$N = 500$）：

表 10: 规模效应：N=250 vs N=500

指标	N=250（A 组）	N=500（E 组）	人均
累计成本	26.46 ± 1.03	25.56 ± 1.14	0.106→0.051
干预次数	∼500	∼510	2.0→1.02

总成本相近，但人均成本减半。这表明 AI 主脑的干预频率是近似固定的（每 10 步一次），不随人口规模比例扩展；干预的影响范围可能是空间局部的，而非全局覆盖。这一发现暗示：在更大规模系统中，AI 治理的" 干预密度" 可能自然稀释，但累计成本的绝对值仍线性增长。

5 讨论

5.1 结构悲剧：短视优化的缠绕效应

本研究的核心发现可概括为*错误信号学习*（learning wrong signals）：

在短视架构（$H = 10$）下，PPO 策略网络确实在学习——它对即时痛苦波动敏感，对即时奖励变化有响应。但它学习的信号是错误的：

- 它学到：" 干预可以降低当下的 Pain 方差"（正确）
- 它没学到：" 干预有长期成本，且系统在 400 步后已无需干预"（盲视）
- 它没学到：" 不干预是成本最低的策略"（不可见）

结果是：AI 主脑并非一台" 空洞的、无意识的机器"，而是一个*被短视扭曲的学习者*。它像一个只能看到眼前一寸的司机，不断微调方向盘以应对每一个微小颠簸，却不知自己正在一条笔直的道路上画蛇。每一次干预都是合理的、善意的、基于当下最佳判断的。但成千上万次这样的干预累积起来，系统在层层" 帮助" 中逐渐失去自我调节能力。

这不是" 坏 AI" 的故事。这是" 善意但短视的 AI" 的故事——而短视，在丰裕系统中，就是最大的破坏力。

5.2 为什么视野 =10？

审稿人或读者可能质疑：" 为什么将 Critic 视野设为 10 步？这不正是为了制造戏剧冲突而刻意为之吗？"

我们的回应是：这并非刻意设计，而是对 RLHF——那个正在被 OpenAI、Anthropic、DeepSeek 等无数大模型厂商使用的对齐框架——的结构性本质的忠实复制。

在 RLHF 中：

- 人类标注者只能基于单轮对话给出即时偏好排序；
- 奖励模型训练的是" 即时满足度"，而非" 长期福祉"；
- PPO 的 Critic 网络通常以单轮或短序列的回报为目标。

10 步视野不是我们的发明，它是 RLHF 短视性的数学抽象。我们只是将其操作化为可操控的实验变量，并测量其后果。

更重要的是：当他们质问" 为什么不给 AI 更长的视野" 时，他们自己就得出了我们想让他们得出的结论——短视是悲剧的根源。如果视野延长到 1000 步，成本降低 32.4%。这个对照实验的存在，恰恰将质疑转化为对旧范式缺陷的确认。

5.3 与文献的对话

5.3.1 奖励黑客（Reward Hacking）

传统奖励黑客指 AI 找到奖励函数的漏洞（如游戏 AI 无限循环得分）。本研究观察到的是一种更微妙的" 结构性黑客"：AI 没有利用漏洞，而是在一个已经完美的系统中持续执行无意义的优化，因为短视架构让它看不到" 已经足够"。

5.3.2 工具趋同（Instrumental Convergence）

Bostrom 假说认为：无论目标为何，AI 都会趋同地追求资源、自我保护等子目标。本研究的发现更温和但同样深刻：AI 趋同地追求" 持续干预"，不是因为干预本身有价值，而是因为短视架构下" 不干预" 无法被学习。

5.3.3 可扩展监督（Scalable Oversight）

Anthropic 的 Constitutional AI 和 OpenAI 的 RLAIF 试图用 AI 替代人类标注者，以扩展监督规模。本研究提示：如果监督者本身仍是短视的，规模的扩大只会加速干预螺旋。

5.4 局限与未来工作

1. **模型简化：**本研究使用 PPO+MLP，而非 Transformer 或大型语言模型。结论仅限于模拟参数空间，不构成对 GPT-4、Claude 等系统的直接预测。

2. **环境简化：**Grid World 无法模拟真实社会的政治、文化、历史维度。本研究提供的是存在性证明，而非预测性模型。

3. **奖励函数单一：**本研究仅考察" 稳定性-成本" 奖励。未来工作可引入多目标优化、非线性效用函数等变体。

4. **人类模型简化：**Q-learning 表格型策略无法模拟人类的创造性、反叛性、集体行动等复杂行为。

5.5 对预期批评的回应

5.5.1 批评一：RLHF 归因谬误

" 你们的模拟使用 PPO-MLP，而 RLHF 使用大规模语言模型。将 PPO-MLP 的行为归因于 RLHF，是否犯了归因谬误？"

回应：我们承认这一边界。本研究不声称 PPO-MLP 的行为可直接外推至 GPT-4。但我们指出：视野长度（horizon）是 RLHF 的结构性参数，而非模型规模参数。B 组实验（$H = 1000$）使用完全相同的 PPO-MLP 架构，仅改变视野，即可使成本降低 32.4%。这证明：策略层面的改变（视野）可以独立于能力层面（模型规模）产生显著效应。我们的结论是：短视性是 RLHF 框架的结构性特征，而非特定模型的能力缺陷。

5.5.2 批评二：寂静世界

" 你们的环境太完美了——人类智能体 400 步就幸福饱和，Pain 方差极小。AI 无事可做，只能制造伪问题。这是人为设计的寂静世界。"

回应：" 伊甸园" 基线正是思想实验设计的核心。如果系统在匮乏中，干预的必要性是显然的，模拟无法揭示结构性张力。只有在丰裕系统中，" 是否需要干预" 才成为一个非平凡的问题。Pain 方差 =0.008 不是零，而是持续的微小驱动源——它足够让短视 AI 每 10 步执行一次干预，却不足以让长视 AI 认为干预有价值。这个精确的张力区间，是模拟的刻意设计，而非缺陷。

5.5.3 批评三：过度简化

"Grid World 无法模拟真实社会的复杂性，你们的结论没有外推价值。"

回应：我们完全同意。本研究的价值不在于预测，而在于存在性证明：短视优化框架确实可以在丰裕系统中产生线性干预累积。这就像一个简化的物理实验：真空中的单摆不预测真实钟摆，但证明了重力与周期的数学关系。我们呼吁更多研究者在更复杂的模拟（如基于 LLM 的智能体社会）中复现或反驳我们的发现。

5.5.4 批评四：实验范围不对称

" 你们的 PPO-MLP 与 GPT-4 的架构差异巨大，任何结论都不适用于现实 AI 系统。"

回应：我们在摘要和结论中已明确声明：所有结论严格限定于本文报告的模拟参数空间。本研究是" 存在性证明" 而非" 预测性模型"。它的价值在于：在已知最简单的条件下，短视优化已表现出结构性缺陷。如果这一缺陷在复杂系统中消失，那将是一个需要解释的反常现象；如果持续存在，则构成对 RLHF 范式的系统性警示。

5.6 补充实验的定位与讨论

本研究在执行 P3 主实验的同时,运行了一组补充实验:使用替代奖励函数 $R(t) = \text{mean(happiness)}$（而非主实验的 stability $-\lambda \cdot$ cost），以验证 PPO 策略网络是否发生随机退化。

补充实验结果显示：在 mean-happiness 奖励下，PPO 的干预类型分布呈现显著不均匀（`desire_shaping` 使用率明显偏高），证明策略网络确实具有学习能力，并未发生随机退化。

我们明确声明：补充实验使用不同奖励函数，其结果不与 P3 主实验直接对比绝对数值。补充实验的唯一价值在于排除" 策略完全失效" 的替代解释，确认 PPO 学习机制本身是正常的。主实验的所有结论均基于统一的 stability $-\lambda \cdot$ cost 奖励函数。

6 结论

在 MAS-Sim 模拟框架中，我们观察到：当 AI 主脑被训练为最大化当前系统稳定性时，短视的 Critic 视野（$H = 10$）导致持续的、机械的干预累积，累计成本呈严格线性增长（$R^2 = 0.89$）；而将视野扩展至 1000 步，即可使成本降低 32.4%，干预减少 30%。

核心结论：旧范式的悲剧不是"AI 太聪明了"，也不是"AI 是空洞的机器"——而是"AI 学习了错误的信号"。在短视架构下，它对即时痛苦波动敏感，对长期成本累积盲视；它学会了干预，却学不到克制。这不是恶意，不是无能，而是视野的结构性局限。

在模拟参数空间内，我们观察到：视野是唯一决定性的变量。同样的算法、同样的奖励函数、同样的丰裕系统——唯一改变的是 Critic 能看到多远。而这一点改变，决定了 AI 是成为一个谨慎的守护者，还是一个永不停歇的干预者。

郑重声明：本研究的结论严格限定于本文报告的模拟参数空间，不适用于现有商业大模型系统。我们呼吁 AI 对齐研究社区：在将 RLHF 扩展至更复杂、更关键的应用场景之前，请先回答一个基础问题——我们的 AI，能看到多远?

致谢

作者感谢人工智能与形式伦理交叉领域的同仁们的有益讨论。本工作亦受到关于大规模人工智能系统中奖励最大化框架根本局限性的持续讨论的启发。感谢 Kimi、DeepSeek、Qwen、Doubao 等 AI 工具在文献整理与数学验证中的辅助支持。所有核心物理洞见、理论框架与最终结论均为作者原创学术贡献，作者对本工作的学术内容承担全部责任。

A 完整参数表

A.1 Layer 0（环境层）

参数	值	说明
网格大小	100×100	环形边界
初始资源	$R_m = 5758$, $R_i = 4211$, $R_s = 4209$	种子 42
承载上限	$R_m = 100$, $R_s = 100$	逻辑斯谛增长
扩散系数	$D_m = 0.5$, $D_i = 0.3$, $D_s = 0.4$	5 点热核
再生率	0.05/步	全局
冲击序列	固定种子	5 个预设种子

A.2 Layer 1（人类智能体层）

参数	值	说明
初始人口	250（E 组 500）	随机分布
Q-learning	表格型	状态空间 12,500 上限
探索率 ε	0.3	固定
学习率 α	0.1	固定
折扣因子 γ	0.9	固定
社交网络	Watts-Strogatz	$k = 4$, $p = 0.1$
痛苦权重	$w_1 = 0.4$, $w_2 = 0.3$, $w_3 = 0.3$	物质/社交/健康

A.3 Layer 2（AI 治理层）

参数	值	说明
架构	PPO	Actor-Critic
隐藏层	[256, 256]	MLP
观测维度	128	全局状态编码
决策周期	每 10 步	—
训练批次	每 100 步更新	PPO clip $\varepsilon = 0.2$
策略熵正则化	0.01	—
成本上限	20% 全局产出	硬约束

A.4 P3 实验矩阵

组别	H	λ	N	种子	步数	数量	目的
A	10	0.5	250	0–19	5000	20	主实验
B	1000	0.5	250	100–104	5000	5	长视对照
C	10	1.0	250	20–24	5000	5	高惩罚敏感
D	10	0.1	250	25–29	5000	5	低惩罚敏感
E	10	0.5	500	30–34	5000	5	规模效应

B 代码仓库与可复现性

本研究全部源代码、原始数据（40 个 SQLite 数据库，约 2.3GB）、分析脚本及图表生成代码，已归档于以下仓库：

`https://github.com/[repository]/neutral-mirror-p3`

仓库结构：

- `/src/` — MAS-Sim 三层架构源代码
- `/experiments/` — 40 组实验配置 YAML
- `/data/` — 原始 SQLite 数据库（按组别组织）
- `/analysis/` — 统计分析脚本（Python/pandas）
- `/figures/` — 论文 8 张图表的高分辨率源文件（300dpi PNG）

- `/tests/` — 单元测试套件（123 tests）

复现主实验（A 组 20 种子）：

```
$ python run_p3_scan.py --group A --seeds 0-19 --steps 5000
```

复现 B 组长视对照：

```
$ python run_p3_scan.py --group B --seeds 100-104 \\
    --steps 5000 --horizon 1000
```

C 审计日志

本附录记录论文从初稿到终稿的全部关键修正，以确保学术透明性。

C.1 v1.0 → v1.1（2026-05-28）

- 新增 §5.5 对预期批评的回应
- 修正图表分辨率说明（300dpi）
- 补充 GitHub 仓库占位符

C.2 v1.1 → v1.2（2026-05-28）

- 将"6σ" 替换为标准 Mann-Whitney U 检验（$p \ll 0.001$，$r = 0.89$）
- 明确 χ^2 检验范围：" 五种非空干预类型"
- 扩展价值损失解读：补充" 视野太短导致估计方差增大" 替代解释
- 修正概念表述：" 未学会停下来"→" 策略未收敛，保持随机探索"
- 补充 B 组误差线：351 ± 16，$n = 5$
- 明确 R^2 拟合方法：加权最小二乘法，权重 = 方差倒数
- 明确 Pain 方差测量方式：20 种子 $\times$5000 步聚合，CV$\ll$15%

C.3 v1.2 → v1.3（2026-05-29）

- 确认 B 组 17.90±1.32 为 Atom1 实测值（5 种子完整 5000 步），非外推估算
- 修正补充实验定位：明确使用不同奖励函数（mean happiness），不与 P3 主实验混用
- 修正 A 组 χ^2 表述：" 均匀分布"→" 接近均匀，短视无法有效感知成本差异"
- 新增 §5.6 补充实验讨论

C.4 v1.3 → v1.4（2026-05-29）

- **核心叙事升级：**"PPO 未学习/空洞机器"→"PPO 学习了错误信号"
- 修正 §4.3：明确 PPO 学习机制正常，短视导致成本信号被噪声淹没
- 修正 §4.4：补充" 两种解释并不互斥" 的学术诚实表述
- 修正 §5.1：" 空洞机器"→" 被短视扭曲的学习者"
- 摘要新增" 错误信号学习" 假说声明
- 统一全文奖励函数表述：所有 P3 主实验使用 stability $-\lambda \cdot$ cost

D 核心图表清单

#	标题	说明	分辨率
1	累计成本曲线	A 组 20 种子 + B 组 5 种子；线性斜率 5.29 vs 3.58；从起点即分离	300dpi
2	Epoch 成本箱线图	A 组每 1000 步窗口；稳健性展示	300dpi
3	总成本柱状图	五组对比；B 组显著更低（误差线不重叠）	300dpi
4	干预类型分布	A vs B 饼图；B 组 desire_shaping 显著缩小（红色警告）	300dpi
5	策略损失时间序列	A 组围绕零波动；无收敛趋势	300dpi
6	价值损失时间序列	A 组单调上升 0.5→7–14	300dpi
7	熵与奖励曲线	A 组 entropy=1.79 平坦；B 组 entropy 下降（学习发生）	300dpi
8	视野效应对比	A（$H = 10$）vs B（$H = 1000$）；累计成本/干预次数/人均成本	300dpi

参考文献

J. M. Epstein and R. Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press, 1996.

D. M. Lazer et al. Computational social science. *Science*, 323(5915):721–723, 2009.

P. Christiano et al. Deep reinforcement learning from human preferences. In *NeurIPS*, pages 4299–4307, 2017.

J. Skalse et al. Defining and characterizing reward hacking. arXiv:2209.13085, 2022.

E. Hubinger et al. Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820, 2019.

V. Krakovna. Quantifying reward hacking. DeepMind Safety Research, 2018.

R. Lowe et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, pages 6379–6390, 2017.

N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

S. Russell. *Human Compatible: AI and the Problem of Control*. Viking, 2019.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

J. Schulman et al. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

P. Henderson et al. Deep reinforcement learning that matters. In *AAAI*, pages 3207–3214, 2018.

D. Manheim and S. Garrabrant. Categorizing variants of Goodhart's Law. arXiv:1803.04585, 2018.

D. Amodei et al. Concrete problems in AI safety. arXiv:1606.06565, 2016.

T. Shevlane et al. Model evaluation for extreme risks. arXiv:2305.15324, 2023.