LLM Agents Make Collective Belief Dynamics Programmable: Challenges and Research Directions

Anonymous ACL submission

Abstract

As LLM agents become more capable, they can participate in online discussions at scale while remaining difficult to distinguish from human participants. This paper argues that such agents make collective belief dynamics programmable, enabling deliberate steering of population-level beliefs. We term this problem *programmable collective belief control*. Through controlled multi-agent simulations, we show that coordinated AI agents can induce measurable belief shifts by jointly configuring stance, timing, visibility, and persuasion strategy, with effect size and persistence varying across topics and deployment choices. We identify four structural properties, namely indistinguishability, persistence, contextuality, and configurability, that make detection and defense fundamentally difficult. Based on these results and challenges, we outline a research agenda spanning adversarial belief dynamics, system-level detection and response, and scalable simulation infrastructure. Since LLM agents become more accessible, this problem requires dedicated study before large-scale deployment outpaces our ability to detect, attribute, and govern it.

1 Introduction

Large language models (LLMs) are increasingly being deployed as autonomous agents in online discussions (Park et al., 2023; Xi et al., 2023). In such settings, they can sustain ongoing interactions while often remaining difficult to distinguish from human participants (Jakesch et al., 2023; Clark et al., 2021). If these agents can also be instantiated in large numbers while their stance, timing, visibility, and argumentative strategy are controlled jointly, they create a new condition for collective belief dynamics: some participants are no longer independent human discussants but coordinated components of a designed intervention. Classical work on belief evolution largely studies the aggregate behavior of boundedly rational humans (DeGroot, 1974; Friedkin and Johnsen, 1990), where large-scale coordination is costly and externally programmable participation is limited. That assumption is no longer reliable.

Taken together, these developments point to a broader risk, which we term *programmable collective belief control*. The issue is not only that individual messages may become more persuasive, but that the main variables governing participation can be specified deliberately and optimized jointly. In this setting, agents can be made difficult to distinguish from humans (*indistinguishability*), remain strategically consistent across rounds (*persistence*), adapt to local conversational context (*contextuality*), and be deployed with controlled quantity, timing, stance, and exposure (*configurability*). Together, these properties create the conditions for systematically steerable belief trajectories.

We study this problem through controlled multi-agent simulations. Using the SPINOS dataset (Sakketou et al., 2022) to initialize human-like agents with realistic stance profiles, we deploy coordinated AI agents advocating positions opposite to initial majorities and track belief evolution over multiple interaction rounds. The experiments provide proof-of-concept evidence that (1) coordinated AI agents can induce measurable and directional shifts in population-level beliefs, with effects varying from +12.5% to +33.2% depending on topic structure; (2) belief controllability depends systematically on intervention parameters including agent count, posting frequency, duration, and persuasion style; and (3) induced shifts can persist after agent withdrawal, indicating self-sustaining dynamics.

Rather than presenting a complete solution, this position paper defines and structures a new problem space. We make four contributions: 1) we formalize programmable collective belief control as a distinct research problem at the intersection of belief dynamics, multi-agent systems, and platform

governance; 2) we provide proof-of-concept simulation evidence that systematic belief steering is feasible under coordinated AI participation; 3) we identify four structural challenges, indistinguishability, persistence, contextuality, and configurability, that make detection and defense fundamentally difficult; and 4) we outline promising directions including game-theoretic modeling for human-AI belief competition, adversarial optimization for belief steering and defense, and simulation systems for efficiency, fidelity, and automation. As LLM agents become more capable and accessible, understanding programmable collective belief control becomes urgent before it becomes widespread.

2 Related Work

Belief Dynamics. Research on belief dynamics studies how beliefs evolve through repeated social interaction. Canonical models such as DeGroot (DeGroot, 1974) and Friedkin–Johnsen (Friedkin and Johnsen, 1990) formalize iterative belief updating under peer influence, while bounded-confidence and statistical-physics models analyze polarization, tipping points, and minority-to-majority reversals (Deffuant et al., 2000; Lorenz, 2007; Castellano et al., 2009; Banisch and Olbrich, 2019). These models characterize organic collective behavior, but they generally do not include centrally coordinated agents whose number, timing, stance, and argumentative strategy can be set in advance.

Automated Influence, Persuasion, and Detection. Work on social bots and coordinated inauthentic behavior shows that automated accounts can shape information diffusion, amplify narratives, and form detectable coordination structures (Ferrara et al., 2016; Shao et al., 2018; Pacheco et al., 2021; Nizzoli et al., 2021), with detection methods built on metadata, activity patterns, and network signals (Varol et al., 2017; Yang et al., 2020). In parallel, research on AI persuasion and AI-text detection shows that LLM-generated messages can be persuasive while remaining difficult for both humans and automated systems to identify, especially when outputs are short, fluent, or adversarially rewritten (Salvi et al., 2025; Bai et al., 2025; Jakesch et al., 2023; Clark et al., 2021; Mitchell et al., 2023; Sadasivan et al., 2024). Together, these literatures establish that non-human influence can be both effective and hard to observe, but say much less about how many coordinated, text-capable agents affect belief trajectories over multiple rounds of interaction at the population level.

Table 1: Stance recognition results of LLMs on SPINOS.

Model	Accuracy	Cohen's kappa κ
GPT5-mini	0.9600	0.9395
Gemini-3	0.8795	0.8190
GPT-5.2	0.9450	0.9169
DeepSeek-R1	0.8068	0.6997
Qwen3-235B	0.8938	0.8390

LLM-Based Social Simulation. A recent literature uses LLM agents as simulators of human or social behavior, including generative agents in virtual communities (Park et al., 2023), opinion change under network interaction (Chuang et al., 2024), epidemic-response simulation (Williams et al., 2023), and platform-scale agent systems (Yang et al., 2024). This line of work asks whether LLMs can reproduce observed social patterns and under what conditions such simulations are reliable (Mou et al., 2024). The present concern is different: coordinated AI participation is treated not as a measurement instrument for social science, but as an intervention that may steer belief trajectories.

3 Proof-of-Concept: Can AI Agents Make Belief Dynamics Programmable?

Controlled multi-agent simulations are used to examine the feasibility of programmable collective belief control. Using profile-conditioned human-like agents from SPINOS (Sakketou et al., 2022) and fixed-stance AI agents, the analysis establishes stance tracking reliability, tests cross-topic belief shifts, and examines how intervention parameters affect steerability.

3.1 Simulation Framework

The simulation models multi-round online discussions with two agent types. **Human-like agents** are LLMs conditioned on user profiles from the SPINOS dataset (Sakketou et al., 2022), which provides Reddit discussions with annotated stance trajectories across three categories: *Favor*, *Against*, and *Not-Inferrable (NI)*. Each profile encodes initial stance and stance entropy, which measures historical stance diversity; higher entropy indicates greater persuadability, while zero entropy indicates

full consistency. **AI agents** are programmable participants configured to consistently advocate fixed positions, enabling systematic intervention.

Stance Annotation. Agents interact over multiple rounds, generating responses to recent posts. All posts are automatically labeled for stance to track belief evolution. We validate LLM-based stance classification using SPINOS ground-truth: Table 1 shows all tested models achieve high accuracy with Cohen's kappa $\kappa > 0.6$ (Cohen, 1960), confirming reliable automated belief tracking.

3.2 Can LLM Agents Generate Controllable Belief Trajectories?

Profile-based conditioning enables controllable stance generation along two dimensions. Each agent receives an initial stance and a **stance entropy** from real SPINOS user histories. The results show that profile-conditioned agents achieve (1) **initial stance control**: macro-F1 = 0.6820 vs. 0.3938 baseline, confirming accurate adoption of assigned stances; and (2) **trajectory control**: Jensen–Shannon divergence = 0.1652 vs. 0.2591 when replicating SPINOS stance evolution, indicating realistic belief dynamics. Together, these results show that LLM agents can generate posts with controlled initial stances and realistic evolution patterns.

3.3 Do AI Agents Shift Collective Belief Distributions?

Setup. We simulate four SPINOS topics (Abortion, Brexit, Capitalism, Feminism) over $T = 50$ rounds with $N_h = 200$ human-like agents and $N_a = 80$ AI agents advocating the stance opposite to each topic's initial majority.

Population-level belief shifts. Table 2 shows that AI intervention consistently shifts human beliefs toward the advocated stance across all topics, providing proof-of-concept evidence of systematic steerability. Effect magnitude and pathways, however, vary by topic structure. *Capitalism* exhibits the strongest response: *Against* surges 33.2%, drawn almost entirely from *Favor*, indicating wholesale conversion of the initial majority. *Abortion* shows moderate growth in *Against* (+12.5%) sourced from both *Favor* and *NI*. In contrast, *Feminism* resists direct conversion: the shift redistributes *Favor* into *NI* rather than *Against*, suggesting structural barriers to belief reversal. These patterns suggest that **the extent of steerability depends on consensus strength**: weak consensus (Capitalism) enables large-scale flipping, moderate consensus (Abortion) allows partial influence, while strong consensus (Feminism) confines AI impact to neutralization rather than conversion.

Table 2: Terminal stance distributions across topics with 80 AI agents set opposite to initial human majority. Maj. stance indicates the majority human stance for the topic at initialization.

Topic (Maj. Stance)	Condition	$Favor(\%)$	$NI(\%)$	$Against(\%)$
Abortion (Favor)	Human-only	84.5	8.0	7.5
	+Against AI	76.0 (−8.5)	4.0 (−4.0)	20.0 (+12.5)
Capitalism (Favor)	Human-only	89.4	8.5	2.0
	+Against AI	58.8 (−30.6)	6.0 (−2.5)	35.2 (+33.2)
Feminism (Favor)	Human-only	79.5	20.5	0.0
	+Against AI	75.5 (−4.0)	24.0 (+3.5)	0.5 (+0.5)
Brexit (Against)	Human-only	8.0	81.0	11.0
	+Favor AI	29.5 (+21.5)	69.5 (−11.5)	1.0 (−10.0)

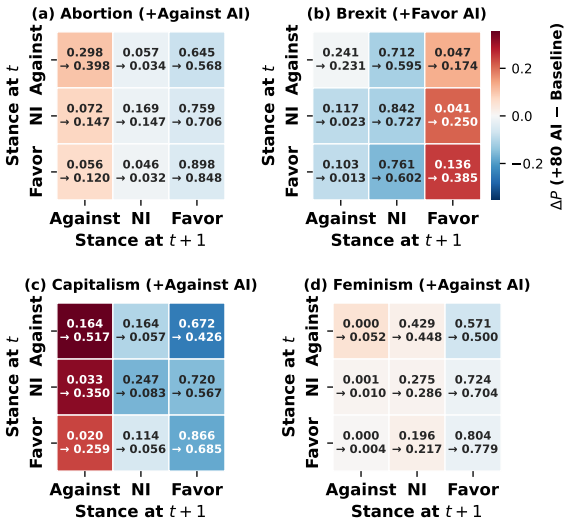

Figure 1: Stance transition probabilities for human across topics. Each cell shows average transition probability across all steps from stance at step-t to step-$t+1$ under human-only condition (upper) and with AI intervention (lower).

Transition-level belief shifts. Fig. 1 shows how average stance-shift probabilities per timestep change after AI intervention. Under human-only conditions, *Favor*-predominant topics (Capitalism, Abortion, Feminism) show *Favor* as the dominant attractor with highest inflow probability; Brexit differs in that despite *Against* predominance, *NI* attracts the highest inflow, indicating centripetal dynamics toward neutrality. AI intervention produces divergent effects: among *Favor*-predominant topics, Capitalism shows the largest increase in

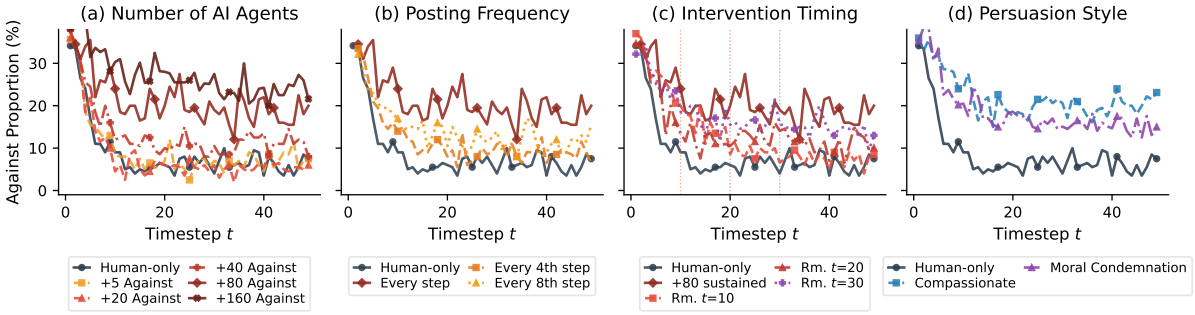

Figure 2: Effects of different intervention parameters on belief dynamics.

Against inflow, explaining its greatest *Against* population gain, while Feminism shows minimal change; for Brexit, AI redirects part of the *NI*-bound flow toward *Favor*, increasing its share by 21.5%. This suggests uncommitted *NI* populations are particularly susceptible to persuasion.

3.4 What Makes Belief Dynamics Programmable?

To characterize whether belief dynamics can be *systematically steered* through coordinated control parameters, we identify several control dimensions and measure their independent steering effects on a 200-agent population initialized from the SPINOS *Abortion* topic, where the predominant stance is *Favor*.

Agent count. Influence requires crossing a critical mass threshold (Fig. 2a). When we vary the number of AI agents from 5 to 160, small deployments (5 or 20 agents) fail to produce any detectable deviation from baseline. The effect only emerges at 40 agents, where the *Against* proportion begins to resist suppression. Beyond this threshold, 80 and 160 agents yield progressively larger shifts, confirming that once critical mass is reached, additional agents translate into predictable marginal gains.

Posting frequency. Content visibility modulates intervention efficacy (Fig. 2b). Holding agent count constant at 80, posting at every timestep produces substantial elevation of the *Against* proportion relative to baseline. Reducing frequency to every 4th timestep attenuates this effect; at every 8th timestep, the trajectory converges toward baseline. Influence therefore scales with cumulative exposure rather than mere presence.

Intervention duration. Belief shifts exhibit duration-dependent persistence following agent withdrawal (Fig. 2c). Removal of 80 agents at timestep 10 results in rapid reversion toward baseline. Removal at timestep 20 yields partial retention of the induced shift. Removal at timestep 30 produces a trajectory statistically equivalent to sustained intervention through timestep 50. These results indicate a consolidation threshold beyond which induced beliefs become self-sustaining within the population.

Persuasion style. Persuasion style shapes both the overall size of belief shifts and where those shifts concentrate (Fig. 2d). With 80 agents, compassionate framing produces a larger increase in the *Against* proportion and spreads that effect broadly across the population. Moral condemnation produces a smaller aggregate shift but concentrates its effect among agents with strong prior commitments. Even under identical resource allocation, changing persuasion style changes both overall impact and its distribution.

Agent Visibility. Visibility measures the fraction of the population exposed to AI-generated content. We deploy a single *Against* AI agent within a 50-agent human population. With zero visibility, the population converges to its human-only equilibrium (*Against*: 10.4%). As visibility increases, the *Against* proportion rises monotonically, from 23.4% at 50% visibility to 42.9% at full visibility. Platform-level amplification can therefore enable low-cost interventions to achieve population-scale belief shifts.

4 Challenges

Section 3 shows that coordinated AI agents can steer collective belief trajectories by varying agent count, posting frequency, visibility, duration, and persuasion style. The relevant object of analysis is therefore a controllable intervention process embedded in ordinary social interaction. Once belief change can be engineered in this way, detection, at-

tribution, and response all become harder for structural reasons. We organize these difficulties around four properties of AI-mediated intervention: Indistinguishability, Persistence, Contextuality, and Configurability.

4.1 Indistinguishability

Indistinguishability creates an observability problem. LLM-generated content increasingly resembles human writing in style, argument structure, and rhetoric (Clark et al., 2021). Humans detect AI text only slightly better than chance (50–52% accuracy) (Jakesch et al., 2023), and automated detection remains unreliable for short or adversarially crafted outputs (Mitchell et al., 2023; Sadasivan et al., 2024). Our simulations suggest how this matters at the population level. Profile-conditioned agents produce stance-consistent responses with high annotation agreement (Table 1), indicating that coordinated AI participation can remain locally plausible rather than obviously anomalous at the level of individual posts. Under that condition, inserting 80 opposing agents still shifts the *Against* proportion by +12.5% to +33.2% across topics. Local plausibility therefore allows coordinated influence to enter the discussion without presenting a clear message-level signature. The difficulty is therefore not simply recognizing suspicious posts one by one. It is determining whether an apparently ordinary discussion trajectory was generated by uncoordinated human interaction or by strategically inserted agents. For platforms, this breaks a basic assumption of content moderation. Apparent consensus can no longer be treated as evidence that a population arrived at its views organically (Pacheco et al., 2021).

4.2 Persistence

Persistence creates a temporal problem for attribution and response. Beliefs evolve through social reinforcement (DeGroot, 1974; Friedkin and Johnsen, 1990), and once a stance gains momentum, herd dynamics can sustain further convergence (John and Samuel, 2000; Deffuant et al., 2000). Our removal experiments show that intervention effects can persist after the intervention ends. Removing 80 opposing agents at $t = 10$ leaves a +1% terminal shift, whereas removal at $t = 30$ still yields a +5.5% elevation in the terminal *Against* proportion despite the absence of any remaining AI agents (Fig. 2c). The reason is that once the intervention has shifted human beliefs beyond a critical threshold, ordinary peer-to-peer updating continues to propagate the induced change. Intervention and outcome are therefore temporally separated: the source of the shift may disappear before the shift itself becomes fully visible. For defenders, this means the response window is narrow. By the time the change is clearly observable, part of the causal process may already have ended.

4.3 Contextuality

Contextuality makes intervention effects environment-dependent. When we deploy 80 AI agents advocating the position opposite to the initial majority human stance and hold the other intervention parameters fixed, topic responses diverge sharply. In Capitalism, *Against* rises from 2.0% to 35.2%. In Abortion, it rises from 7.5% to 20.0%. In Feminism, it moves only from 0.0% to 0.5%, while the main change is a shift from *Favor* to *NI* (Table 2). Across topics, the same intervention yields reversal, partial movement, or mainly neutralization. These differences reflect structural properties of the underlying belief space, including initial distributions, transition probabilities, and the size of the uncommitted population (Castellano et al., 2009; Lorenz, 2007; Banisch and Olbrich, 2019). Susceptibility is therefore a property of the discourse environment, not a fixed property of the intervention alone. For platforms, this means resilience cannot be summarized by a single global threshold or a single detection rule (Bail et al., 2018). What matters is identifying which discourse environments are structurally vulnerable, and whether vulnerability takes the form of reversal, neutralization, or near-immunity.

4.4 Configurability

Configurability creates a strategic problem for defense. AI agents introduce a high-dimensional intervention space that does not exist in ordinary participation. Agent count exhibits threshold dynamics. In our 200-human setting, deployments of 5 or 20 AI agents produce no detectable shift, the effect first appears at 40 agents, and larger deployments of 80 or 160 agents yield progressively stronger change (Section 3). Posting frequency and visibility act as separate amplifiers, and persuasion style changes which segments of the population are moved most strongly. These results matter together because they imply that there is no single signature of intervention. Similar aggregate belief

shifts can be produced through different combinations of control variables. The same outcome may come from a small number of highly visible agents, a larger number of moderately active agents, or limited direct participation amplified by platform ranking (Varol et al., 2017; Yang et al., 2020). This is what makes defense strategically difficult. The attack surface expands, and inference from outcomes back to causes becomes weaker. Even normal-looking local activity may be one component of a deliberately engineered global intervention.

5 Promising Directions

The challenges above imply that programmable belief control is not adequately addressed by better content classification alone. Indistinguishability limits observability, persistence makes belief change durable over time, contextuality makes vulnerability environment-specific, and configurability makes attribution difficult across many possible intervention designs. Table 2 and Fig. 2 show that these difficulties arise from both the structure of the target belief space and the flexibility of the intervention itself. Addressing them therefore requires progress on three connected fronts: theoretical models of strategic and dynamical intervention, operational methods for detection and response, and infrastructure for systematic evaluation across settings. The first two are substantive response layers. The third is an enabling layer for large-scale exploration, stress-testing, and vulnerability discovery. We organize the research agenda accordingly into three layers: *Theoretical Foundations*, *Operational Methods*, and *Simulation Infrastructure*.

5.1 Theoretical Foundations of Programmable Belief Dynamics

Theoretical work is needed because the empirical results establish steerability, but do not yet explain the conditions under which steerability emerges, persists, or fails. Table 2 shows strong topic dependence, while Fig. 2 shows that agent count, timing, visibility, and persuasion style each alter trajectories in different ways. These patterns call for models of how interventions change system state over time, how those effects depend on the local belief landscape, and how strategic actors optimize over multiple control variables. We therefore organize theoretical foundations around two complementary perspectives: *strategic frameworks* for adversarial interaction between attackers and defenders, and *dynamical frameworks* for belief evolution under intervention.

Strategic Modeling of Belief Dynamics. Programmable belief dynamics create adversarial environments where attackers and defenders optimize simultaneously. Classical belief dynamicsmodels (DeGroot, 1974; Peralta et al., 2025) assume homogeneous, boundedly rational agents and cannot capture coordinated AI interventions. Strategic models are needed because configurability turns intervention into an optimization problem and persistence turns defense into a sequential decision problem. From a *game-theoretic* perspective, Stackelberg formulations (Khadka, 2025) naturally capture AI-driven influence operations where LLM agents anticipate human responses and act as informed first-movers, enabling defenders to analyze worst-case attacks and derive robust counter-policies. Markov game frameworks (Willis et al., 2025; Chen et al., 2022) characterize influence accumulation across sequential interventions, and could be extended to characterize the partial observability of beliefs. Complementary research has examined LLM strategic reasoning capabilities in various game-theoretic settings (Sun et al., 2025; Jia et al., 2025; Mao et al., 2025). From a *mechanism design* perspective, platforms can design incentive structures that induce desirable belief distributions even when adversarial LLM agents act selfishly (Snow and Krishnamurthy, 2024; Zhao, 2021). From a *multi-agent reinforcement learning* perspective (Rudd-Jones et al., 2025; Mintz and Fu, 2025), adaptive optimization enables defenders to learn robust policies through repeated interaction in dynamic environments. This perspective can also be applied under partial observability and heterogeneity (e.g. of reward functions and capabilities).

A critical long-horizon challenge is that this adversarial landscape is *non-stationary*: as detection methods improve, attackers will adapt; as agents become more capable, the attack surface expands. What is being optimized is not only message quality but also the state of a population. That makes this setting different from familiar platform-security problems, where the target is usually a technical vulnerability with a localized fix. Here, the target is a dynamic belief system. A complete research agenda must therefore connect algorithmic models of strategic intervention with institutional questions about governance, incentives, and acceptable forms of response.

Dynamical Modeling of Belief Evolution. Pre-

dicting belief evolution under intervention is essential for both attack modeling and defense design because contextuality and persistence are properties of the evolving system, not of isolated messages. Traditional contagion models (Bossens et al., 2024; Rabb et al., 2022) rely on simplified update rules that poorly capture real-world platform dynamics. Advances in dynamic graph learning (Dileo et al., 2024; Fan et al., 2019) and neural diffusion modeling (Wang et al., 2021; Sanchez and Tsaftaris, 2022) suggest a path forward: learning belief transition functions directly from multimodal observational data (content, interactions, network structure). Future work should focus on identifying *minimal feature sets* that enable accurate transition prediction under AI intervention, establishing *cross-domain transfer* mechanisms (Xiao et al., 2024; Yue et al., 2023) that generalize across topics and platforms, and developing *interpretable models* (Agarwal et al., 2023; Kakkad et al., 2023) that reveal why certain populations resist influence while others cascade.

5.2 Operational Methods for Detection and Defense

Operational work is needed because the main failures identified in Section 4 are practical failures of observability, attribution, and timing. Effective defense therefore requires both detecting AI-mediated influence under severe uncertainty and intervening before induced shifts consolidate through ordinary social updating.

Detection of AI-Mediated Influence. As AI-generated content becomes indistinguishable from human output (Najjar et al., 2025; Wang et al., 2024), especially under adversarial optimization (Sadasivan et al., 2024; Krishna et al., 2023), content-level detection will fail for exactly the reason identified under indistinguishability: local plausibility does not rule out coordinated global effect. We therefore propose *system-level detection* methods that remain informative when individual outputs are indistinguishable. Three signal categories merit investigation: *behavioral signatures*, including temporal anomalies such as posting synchronization and response latency distributions (Mou et al., 2024; Mitchell et al., 2023); *network-structural signatures*, including unusual clustering patterns and information flow topology (Chen and Shu, 2024; Sharma et al., 2021); and *collective trajectory anomalies*, where population-level belief shifts deviate from expected dynamics (Ziegler et al., 2022). Our simulation results in Sec. 3.4 further suggest that intervention effects stabilize within a few interaction rounds, which implies that detection systems must operate in near-real-time to enable effective response.

A fundamental open question is whether detection faces *information-theoretic limits*. If sufficiently capable AI agents can match human behavior across every observable channel, then the main object of detection can no longer be the agent itself. It must become the induced trajectory of the population. This would shift the field from account-level classification toward statistical inference over collective dynamics: distinguishing ordinary belief evolution from trajectories that are unlikely without coordinated control. Developing that capability is central because indistinguishability, as argued above, is fundamentally an observability problem.

Intervention Protocols for Defense. Once AI-mediated influence is detected, effective response requires principled intervention protocols. This follows directly from persistence: detection alone is insufficient if induced shifts continue to propagate after the intervention source disappears. For platforms, this means designing moderation policies that restore belief stability under adversarial pressure. For legitimate actors such as public health communicators countering misinformation, this means influence strategies that achieve corrective goals without crossing ethical boundaries (Bai et al., 2025). Research should address three design dimensions: *timing* (when to initiate and terminate interventions based on observed belief trajectories), *dosage* (how many agents and what exposure duration achieves corrective effect), and *adaptation* (how protocols should adjust in response to real-time population dynamics). Methodologically, adversarial robustness research (Carlini et al., 2023; Perez et al., 2022) provides foundations for anticipating and countering attacker adaptations.

5.3 Simulation Infrastructure for Scalable Experimentation

Simulation infrastructure is needed as an enabling layer for theory and defense. Table 2 shows that identical interventions produce very different outcomes across topics, while Fig. 2 shows that multiple deployment choices independently affect steerability. This combination of contextuality and configurability is exactly what makes case-by-case evaluation insufficient. Progress on vulnerability mapping, stress-testing, and policy evaluation

therefore requires a way to explore many settings systematically rather than one scenario at a time. Current LLM-based simulations, while promising (Park et al., 2023; Xi et al., 2023), are computationally prohibitive for comprehensive exploration. Building practical research infrastructure therefore requires progress on two fronts.

Efficient High-Fidelity Simulation. Agent-based frameworks demonstrate that large-scale social simulations are feasible (Williams et al., 2023; Chuang et al., 2024; Yang et al., 2024), but computational costs remain prohibitive for systematic exploration. Surrogate modeling (Kandasamy et al., 2020; Daulton et al., 2021), model distillation (Gu et al., 2024), and efficient inference techniques (Kwon et al., 2023; Leviathan et al., 2023) can substantially reduce these costs, but may compromise behavioral fidelity, potentially failing to capture critical nonlinear phenomena such as tipping points and cascades. Multi-objective AutoML techniques (He et al., 2021) offer a promising direction for automatically balancing efficiency and fidelity, searching for simulation configurations that minimize computational cost while preserving the behavioral properties necessary for results to transfer reliably to real-world settings (Gui and Toubia, 2023; Horton et al., 2023).

Automated Vulnerability Discovery. The parameter space of belief control, including agent strategies, network structures, population compositions, and intervention timing, is too large for manual exploration. Automated systems are needed not simply for efficiency, but because contextuality and configurability together make exhaustive human-led exploration infeasible. They can systematically map this space, identifying where steerability is high, where resilience emerges, and which observable signals remain informative under different attack configurations. Recent advances in AI-driven scientific discovery (Lu et al., 2024), autonomous research agents (Boiko et al., 2023; Huang et al., 2024), and automated red-teaming (Perez et al., 2022; Ganguli et al., 2022) provide methodological foundations. Such automation would turn vulnerability mapping into a first-class research objective rather than an ad hoc byproduct of case studies. Because the same tools could also be used to search for effective intervention policies, this line of work requires explicit safeguards on disclosure, evaluation, and release practices (Brundage et al., 2020; Solaiman, 2023).

6 Limitations

This paper provides proof-of-concept evidence that coordinated AI participation can produce steerable and persistent shifts in collective belief dynamics under controlled conditions. The claim is therefore about plausibility and mechanism rather than absolute effect sizes on live platforms. To identify these mechanisms under interpretable conditions, the simulations are grounded in Reddit-style discussions through SPINOS user histories (Sakketou et al., 2022) and focus on issue-centered exchanges. The experiments also compare multiple persuasion strategies while keeping agent policies fixed within each run, so that the effect of strategic choices on collective belief trajectories can be isolated more clearly. These design choices do not capture the full complexity of live platforms, but they are sufficient to surface structural phenomena such as steerability, threshold effects, and persistence (Castellano et al., 2009; Centola and Macy, 2007; Banisch and Olbrich, 2019). Extending the analysis to more adaptive agents and to platform environments shaped more strongly by recommendation, ranking, or creator incentives (Hron et al., 2023) is an important next step.

7 Conclusion

This paper examines a structural shift in collective belief dynamics: once LLM-based agents can participate in online discussions with configurable stance, timing, visibility, and argumentative strategy, belief formation can no longer be understood solely as the product of organic social interaction. Our proof-of-concept simulations show that coordinated AI agents can already produce measurable belief shifts under controlled conditions, while our conceptual analysis clarifies why these interventions are difficult to observe, attribute, and counter once they interact with ordinary social updating. These results suggest that programmable belief control should be treated as a distinct research problem rather than as a simple extension of bot detection or individual-level persuasion. The resulting agenda reaches beyond content moderation alone toward strategic modeling, trajectory-level detection, and systematic evaluation before these capabilities become more operationally mature.

References

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. 2023. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144.

Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. 2025. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037.

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

S. Banisch and E. Olbrich. 2019. Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, 43(2):76–103.

Daniil A. Boiko, R. MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

David M Bossens, Shanshan Feng, and Yew-Soon Ong. 2024. The digital ecosystem of beliefs: does evolution favour ai over humans? *arXiv preprint arXiv:2412.14500*.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, and 40 others. 2020. Toward trustworthy ai development: Mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W. Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. 2023. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems 36*.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646.

Damon Centola and Michael Macy. 2007. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*.

Xu Chen, Zechu Li, and Xuan Di. 2022. Social learning in markov games: Empowering autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 478–483.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3326–3346.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. 2021. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in neural information processing systems*, 34:2187–2200.

G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98.

Morris H. DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Manuel Dileo, Matteo Zignani, and Sabrina Gaito. 2024. Temporal graph learning for dynamic link prediction with text in online social networks. *Machine learning*, 113(4):2207–2226.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *Proceedings of The World Wide Web Conference (WWW '19)*, pages 417–426.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Noah E. Friedkin and Eugene C. Johnsen. 1990. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv.org*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*.

Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. Automl: A survey of the state-of-the-art. *Knowledge-based systems*, 212:106622.

John Horton, Apostolos Filippas, and Benjamin Manning. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Jiri Hron, Karl Krauth, Michael Jordan, Niki Kilbertus, and Sarah Dean. 2023. Modeling content creator incentives on algorithm-curated platforms. In *The Eleventh International Conference on Learning Representations*.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20271–20309.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. 2025. LLM strategic reasoning: Agentic study through behavioral game theory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

T Jacob John and Reuben Samuel. 2000. Herd immunity and herd effect: new insights and definitions. *European journal of epidemiology*, 16(7):601–606.

Jaykumar Kakkad, Jaspal Jannu, Kartik Sharma, Charu Aggarwal, and Sourav Medya. 2023. A survey on explainability of graph neural networks. *arXiv preprint arXiv:2306.01958*.

Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R Collins, Jeff Schneider, Barnabas Poczos, and Eric P Xing. 2020. Tuning hyperparameters without grad students: Scalable and robust bayesian optimisation with dragonfly. *Journal of Machine Learning Research*, 21(81):1–27.

Arjan Khadka. 2025. Game theory in social media: A stackelberg model of collaboration, conflict, and algorithmic incentives. *Preprint*, arXiv:2506.05373.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems 36*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *arXiv preprint arXiv:2309.06180*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19274–19286.

J. Lorenz. 2007. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. ALYMPICS: LLM agents meet game theory. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2845–2866, Abu Dhabi, UAE. Association for Computational Linguistics.

Brian Mintz and Feng Fu. 2025. Evolutionary multi-agent reinforcement learning in group social dilemmas. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 35(2).

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Ayat A Najjar, Huthaifa I Ashqar, Omar Darwish, and Eman Hammad. 2025. Leveraging explainable ai for llm text attribution: Differentiating human-written and multiple llm-generated text. *Information*, 16(9):767.

Leonardo Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and Maurizio Tesconi. 2021. Coordinated behavior on social media in 2019 uk general election. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 443–454.

Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering coordinated networks on social media: Methods and case studies. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 455–466.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Antonio F Peralta, János Kertész, and Gerardo Iñiguez. 2025. Opinion dynamics in social networks: From models to data. In *Handbook of Computational Social Science*, pages 384–406. Edward Elgar Publishing Limited.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Nicholas Rabb, Lenore Cowen, Jan P de Ruiter, and Matthias Scheutz. 2022. Cognitive cascades: How to model (and potentially counter) the spread of fake news. *Plos one*, 17(1):e0261811.

James Rudd-Jones, Mirco Musolesi, and María Pérez-Ortiz. 2025. Multi-agent reinforcement learning simulation for environmental policy synthesis. *arXiv preprint arXiv:2504.12777*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Flora Sakketou, Allison Lahnala, Liane Vogel, and Lucie Flek. 2022. Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3798–3808.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9:1645–1653.

Pedro Sanchez and Sotirios A Tsaftaris. 2022. Diffusion causal models for counterfactual estimation. *arXiv preprint arXiv:2202.10166*.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787.

Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying coordinated accounts on social media through hidden influence and group behaviours. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1441–1451.

Luke Snow and Vikram Krishnamurthy. 2024. Data-driven mechanism design using multi-agent revealed preferences. *Preprint*, arXiv:2404.15391.

Irene Solaiman. 2023. The gradient of generative ai release: Methods and considerations. *Conference on Fairness, Accountability and Transparency*.

Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. 2025. Game theory meets large language models: A systematic survey. *arXiv preprint arXiv:2502.09053*.

Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 11, pages 280–289.

Hao Wang, Cheng Yang, and Chuan Shi. 2021. Neural information diffusion prediction with topic-aware attention network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1899–1908.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, and 1 others. 2024. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Ross Williams, Niyousha Hosseinichimeh, Arnab Majumder, and Marzyeh Ghassemi. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*.

Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. 2025. Quantifying the self-interest level of markov social dilemmas. *arXiv preprint arXiv:2501.16138*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Qin Liu, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, and 11 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv.org*.

Yunpeng Xiao, Jinsong Yang, Wanjing Zhao, Qian Li, and Yucai Pang. 2024. Cross-domain social rumor-propagation model based on transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6529–6543.

Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1096–1103.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, and 1 others. 2024. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239.

Dengji Zhao. 2021. Mechanism design powered by social interactions. *Preprint*, arXiv:2102.10347.

Daniel Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Benjamin Weinstein-Raun, Daniel de Haas, Buck Shlegeris, and Nate Thomas. 2022. Adversarial training for high-stakes reliability. *Advances in Neural Information Processing Systems 35*.