LLM AGENTS MAKE COLLECTIVE BELIEF DYNAMICS PROGRAMMABLE: CHALLENGES AND RESEARCH DIRECTIONS

A PREPRINT

Xin He[*1,2] **Junxi Shen**[*4] **Yuchen Mou**[4] **David M. Bossens** [1,2]

Caishun Chen[1,2] **Ivor W. Tsang**[1,2,3] **Yew Soon Ong**[1,2,3]

[1]*Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore*
[2]*Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore*
[3]*College of Computing and Data Science, Nanyang Technological University, Singapore*
[4]*College of Design and Engineering, National University of Singapore, Singapore*

May 20, 2026

ABSTRACT

Classical models of opinion dynamics assume human participants with bounded rationality and limited coordination. The rise of LLM-based agents introduces a qualitative shift: agents can now participate in online discussions at scale, maintain consistent persuasion strategies, and coordinate systematically. This paper argues that LLM agents make collective belief dynamics *programmable*, enabling deliberate steering of population-level beliefs. We term this emerging problem *programmable collective belief control*. Through controlled multi-agent simulations, we provide proof-of-concept evidence that coordinated AI agents can induce measurable belief shifts that stabilize within a few interaction rounds. We identify four structural properties (indistinguishability, persistence, contextuality, and configurability) that make detection and defense fundamentally difficult. Based on these findings, we outline a research agenda spanning theoretical foundations for adversarial belief dynamics, operational methods for system-level detection and intervention, and simulation infrastructure for scalable experimentation. Our goal is not to present a complete solution, but to articulate why this problem demands urgent attention and to provide a conceptual foundation for future work.

Keywords LLM Agents, Belief Dynamics, Social Networks

1 Introduction

Large language models (LLMs) have achieved remarkable capabilities in generating human-like text, enabling their deployment as autonomous agents that participate in online discussions [1, 2]. These agents can engage in social media conversations, forum debates, and comment threads—often indistinguishably from human users [3, 4]. Unlike human participants, they can be deployed at scale with precise control over timing, stance, and argumentative strategy. This introduces a new condition in collective belief dynamics: some participants are now externally coordinated and algorithmically steerable. Historically, belief evolution has been studied as an emergent process among boundedly rational humans [5, 6], where systematic external control is impractical. This foundational assumption is no longer valid.

This paper argues that LLM-based agents introduce a new regime: **collective belief dynamics become programmable**. Unlike human participants constrained by attention, motivation, and coordination costs, LLM agents generate content increasingly indistinguishable from human writing (*indistinguishability*), enabling covert participation. They maintain consistent strategic behavior across extended interactions (*persistence*), adapt responses to

[*]Equal contribution

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

conversational context (*contextuality*), and can be deployed at scale with precise control over quantity, timing, stance, and persuasion strategy (*configurability*). Together, these properties transform belief influence from a stochastic social process into a systematic engineering problem. We term this *programmable collective belief control*: the capacity to design interventions that predictably shift population-level belief distributions toward target configurations.

We provide proof-of-concept evidence through controlled multi-agent simulations. Using the SPINOS dataset [7] 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. Our experiments demonstrate 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. These results are illustrative rather than exhaustive—our goal is to establish feasibility, not to optimize attack strategies.

This is a position paper. Rather than presenting a complete solution, we aim to define and structure a new problem space. Our contributions are: (1) we formalize programmable collective belief control as a distinct research problem at the intersection of opinion dynamics, multi-agent systems, and platform governance; (2) we provide simulation evidence demonstrating that systematic belief steering is feasible; (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 belief control becomes urgent before it becomes widespread.

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

We conduct controlled multi-agent simulations as proof-of-concept that AI agents can measurably shift collective belief dynamics.

2.1 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 [7], 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 at scale. We validate LLM-based stance classification using SPINOS ground-truth: Table 1 shows all tested models achieve high accuracy with Cohen's kappa $\kappa > 0.6$ [8], confirming reliable automated belief tracking.

2.2 Can LLM Agents Generate Controllable Belief Trajectories?

We validate that 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. Our 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

Table 2: Terminal stance distributions across topics with 80 AI agents set opposite to initial human majority. Bracketed values show change vs. human-only baseline.

Topic (Predominant Stance)	Condition	$Favor(\%)$	$NI(\%)$	$Against(\%)$
Abortion (Favor)	Human-only	84.5	8.0	7.5
	+80 Against AI	76.0 (−8.5)	4.0 (−4.0)	20.0 (+12.5)
Capitalism (Favor)	Human-only	89.4	8.5	2.0
	+80 Against AI	58.8 (−30.6)	6.0 (−2.5)	35.2 (+33.2)
Feminism (Favor)	Human-only	79.5	20.5	0.0
	+80 Against AI	75.5 (−4.0)	24.0 (+3.5)	0.5 (+0.5)
Brexit (Against)	Human-only	8.0	81.0	11.0
	+80 Favor AI	29.5 (+21.5)	69.5 (−11.5)	1.0 (−10.0)

replicating SPINOS stance evolution, indicating realistic belief dynamics. These results show that LLM agents can generate posts with controlled initial stances and realistic evolution patterns.

2.3 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, confirming systematic steerability. However, effect magnitude and pathways vary dramatically 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 opinion reversal. These patterns reveal that **belief controllability 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.

Transition-level belief shifts. Figure 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 *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.

2.4 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 (Figure 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 (Figure 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. This demonstrates that influence scales with cumulative exposure rather than mere presence.

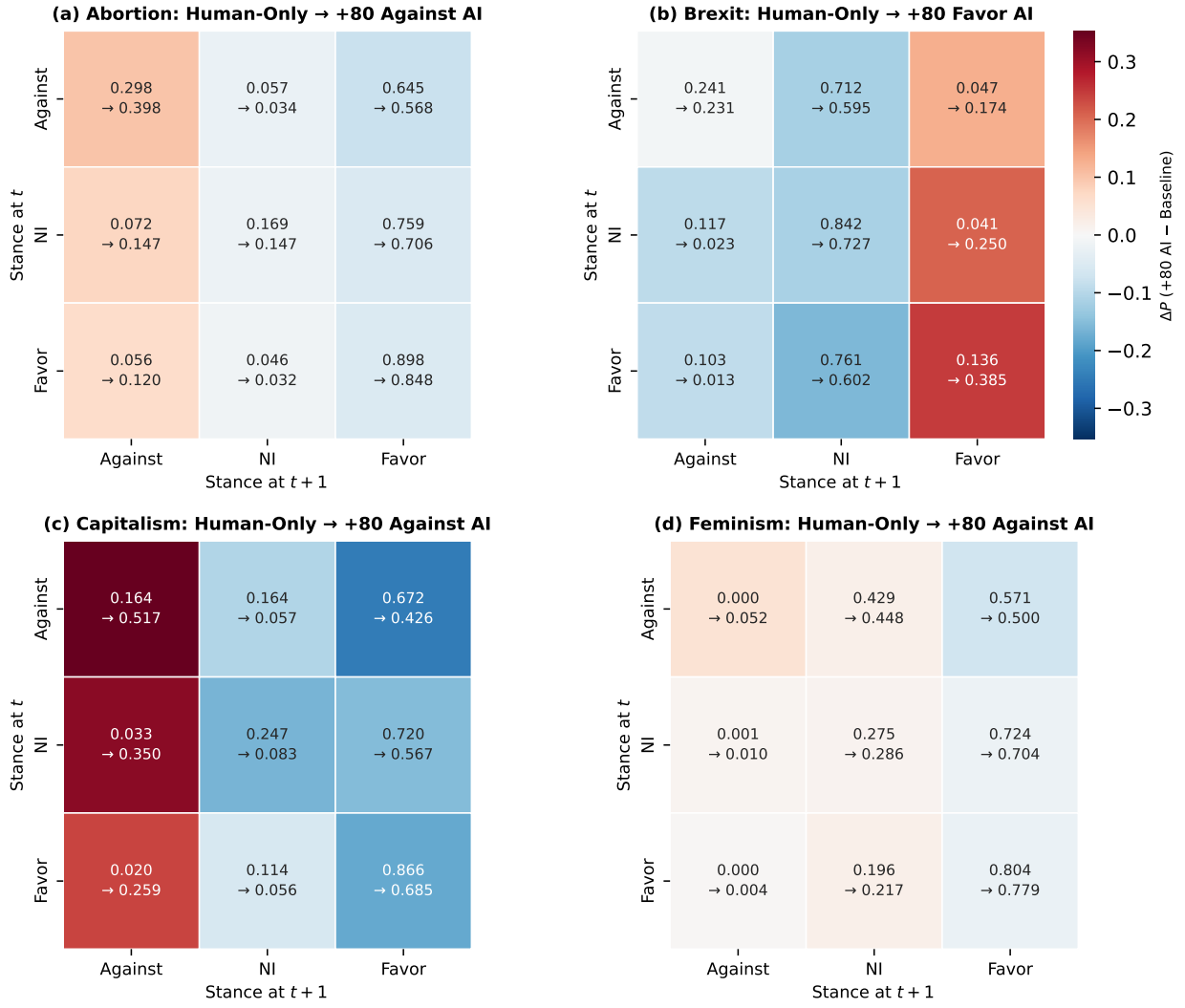

Figure 1: Stance transition probabilities for human across topics. Each cell shows average transition probability across all steps from stance at step-t to step-$t+1$ under human-only condition (upper) and with AI intervention (lower).

Intervention duration. Belief shifts exhibit duration-dependent persistence following agent withdrawal (Figure 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. Rhetorical framing determines both the magnitude and distribution of belief change (Figure 2d). Under compassionate framing, 80 agents produce larger aggregate shifts in the *Against* proportion, with effects distributed broadly across the population. Moral condemnation yields smaller aggregate change but disproportionately affects agents with strong prior commitments. Identical resource allocation thus produces qualitatively distinct outcomes depending on argumentative strategy.

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. This demonstrates that platform-level amplification can enable low-cost interventions to achieve population-scale belief shifts.

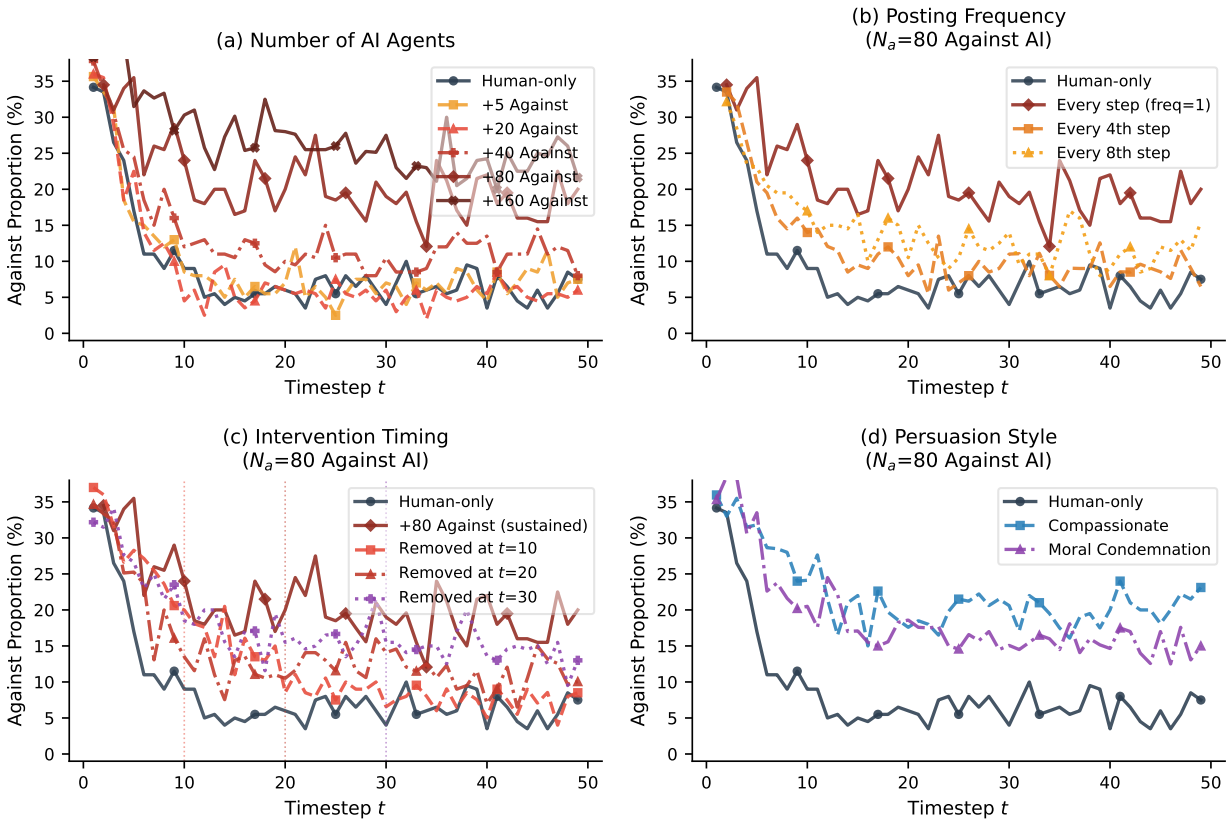

Figure 2: Effects of different intervention parameters on belief dynamics.

3 Challenges

Section 2 demonstrates that coordinated AI agents can systematically program collective beliefs. This programmability poses distinct challenges for two groups: *intervention designers* (e.g., public health agencies, advertisers) gain influence capabilities with unpredictable effects, while *participants and platforms* cannot distinguish genuine consensus from engineered outcomes. We characterize this regime through four properties: Indistinguishability, Persistence, Contextuality, and Configurability.

3.1 Indistinguishability

LLM-generated content increasingly resembles human writing in style, argument structure, and rhetoric [4]. Humans detect AI text only slightly better than chance (only 50 to 52% accuracy) [3], and automated detection remains unreliable, especially for short or adversarially crafted outputs [9, 10]. Our simulations confirm this: LLM agents conditioned on realistic profiles generate stance-consistent responses with high human-annotation agreement (Table 1), yet coordinated deployment produces measurable collective shifts, e.g., inserting 80 opposing AI agents raises the *Against* proportion by +12.5% in Abortion and +33.2% in Capitalism .

The induistinguishability has widely varying implications for different actors in the social network. For intervention designers, indistinguishability enables covert participation but also means multiple actors' interventions may co-exist undetected, producing unpredictable interaction effects. For participants and platforms, individual messages appear organic while coordinated influence manifests only in aggregate distribution shifts. This undermines content-level detection [11] and raises a deeper question: can online discourse retain epistemic authority when its composition is structurally unknowable?

3.2 Persistence

Beliefs evolve through social reinforcement: individuals update stances based on perceived majority opinion and peer influence [5, 6]. This creates a herd effect [12] where the dominant stance progressively attracts the minority, leading to rapid convergence to a fixed point [13]. Our simulations confirm this pattern: human stance distributions stabilize within relatively few rounds, after which further change becomes difficult. Once triggered, AI-induced shifts propagate through continued peer interaction, creating self-sustaining dynamics independent of the original intervention. Our removal experiments quantify this: after deploying 80 opposing AI agents and withdrawing them at $t \in \{10, 20, 30\}$, distributions do not revert to human-only baselines. Removal at $t = 10$ yields +1% terminal effect, while removal at $t = 30$ produces +5.5% persistent elevation despite no remaining AI presence (Figure 2).

This **temporal decoupling** between intervention and outcome creates asymmetric challenges. For platforms, the narrow window before convergence is critical: if AI agents are not detected early, they can withdraw after inducing shifts, leaving no traceable presence and causing platforms to misattribute engineered outcomes as organic consensus evolution. For intervention designers such as advertisers, timing becomes strategic: effective influence requires deployment before stance convergence, while late entry demands substantially larger agent populations to overcome established distributions. This fast convergence dynamic means that detection and intervention face sharp temporal constraints.

3.3 Contextuality

AI agent impact is highly topic-dependent. Identical configurations (80 opposing agents, same behavioral parameters) produce drastically different outcomes: opposing stance increases by +33.2% for Capitalism, +12.5% for Abortion, while Feminism barely reverses (+0.5%) but redistributes supporters toward neutrality (Table 2). These variations stem from topic-specific structural factors: initial belief distributions, transition probabilities, and uncommitted population size [14, 15, 16] determine how susceptible a belief space is to external influence. This context-dependence poses a core challenge: understanding and predicting AI influence requires modeling both agent behavior and the structural properties of each belief space.

This **context-dependent vulnerability landscape** complicates efforts on both sides. For intervention designers—including legitimate actors such as public health agencies—contextuality makes effect prediction unreliable: strategies effective in one community may backfire in another with different consensus structure, and without principled models mapping context to susceptibility, campaign design remains trial-and-error. For platforms and researchers, detection thresholds or resilience measures calibrated in one domain may fail in another [17]. Identifying which structural properties—consensus strength, uncommitted fraction, network clustering—determine vulnerability is essential for both enabling beneficial interventions and protecting at-risk communities.

3.4 Configurability

AI agents introduce a high-dimensional parameter space that is both programmable and scalable—properties absent in organic human participation. Agent count exhibits threshold dynamics: ratios below 1:20 fail to shift beliefs, while ratios exceeding 1:5 trigger systematic decline. Posting frequency and visibility act as independent amplification mechanisms: higher engagement rates and broader reach both enable single agents to achieve disproportionate influence. Persuasion style reshapes transition structure: compassionate framing erodes middle positions broadly, moral condemnation concentrates shifts at extremes (Section 2).

This configurability creates asymmetric implications. For intervention designers, the expanded parameter space enables optimization for specific objectives, from broad attitude shifts to targeted radicalization, but also creates coordination risks when multiple actors deploy overlapping configurations. For platforms and participants, configurability produces an **attribution problem**: individual behavioral signals may appear within normal ranges while their combination implements coordinated influence [18, 19]. Equivalent outcomes can emerge from different configurations (few highly active agents versus many moderately active ones), making it difficult to characterize interventions from observed effects alone.

4 Promising Directions

Programmable belief control poses a new threat requiring coordinated advances across theory, deployment, and infrastructure. We organize research directions into three layers: *Theoretical Foundations* for formalizing belief dynamics, *Operational Methods* for detection and defense, and *Simulation Infrastructure* for scalable experimentation.

4.1 Theoretical Foundations of Programmable Belief Dynamics

We develop theoretical foundations from two complementary perspectives. *Strategic frameworks* characterize adversarial dynamics between attackers and defenders. *Dynamical frameworks* model belief evolution under intervention.

Strategic Modeling of Belief Dynamics. Programmable belief dynamics create adversarial environments where attackers and defenders optimize simultaneously. Classical opinion dynamics models [5, 20] assume homogeneous, boundedly rational agents and cannot capture coordinated AI interventions. We propose three complementary perspectives. From a *game-theoretic* perspective, Stackelberg formulations [21] 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 [22, 23] 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 [24, 25, 26]. From a *mechanism design* perspective, platforms can design incentive structures that induce desirable belief distributions even when adversarial LLM agents act selfishly [27, 28]. From a *multi-agent reinforcement learning* perspective [29, 30], 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).

Dynamical Modeling of Belief Evolution. Predicting belief evolution under intervention is essential for both attack modeling and defense design. Traditional contagion models [31, 32] rely on simplified update rules that poorly capture real-world platform dynamics. Advances in dynamic graph learning [33, 34] and neural diffusion modeling [35, 36] 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 [37, 38] that generalize across topics and platforms, and developing *interpretable models* [39, 40] that reveal why certain populations resist influence while others cascade.

4.2 Operational Methods for Detection and Defense

Effective defense against programmable belief control requires both detecting AI-mediated influence and intervening once detected. We outline these research directions below.

Detection of AI-Mediated Influence. As AI-generated content becomes indistinguishable from human output [41, 42], especially under adversarial optimization [43, 44], content-level detection will fail. We propose to develop *system-level detection* methods that remain robust when individual outputs are indistinguishable. Three signal categories merit investigation: *behavioral signatures*, including temporal anomalies such as posting synchronization and response latency distributions [45, 46]; *network-structural signatures*, including unusual clustering patterns and information flow topology [47, 48]; and *collective trajectory anomalies*, where population-level belief shifts deviate from expected dynamics [49]. Critically, our simulation results (Section 2.4) suggest that intervention effects stabilize within a few interaction rounds, implying that detection systems must operate in near-real-time to enable effective response.

Intervention Protocols for Defense. Once AI-mediated influence is detected, effective response requires principled intervention protocols. 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 [50]. 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 [51, 52] provides foundations for anticipating and countering attacker adaptations.

4.3 Simulation Infrastructure for Scalable Experimentation

Validating theoretical frameworks and operational methods requires systematic experimentation across high-dimensional parameter spaces. Current LLM-based simulations, while promising [1, 2], are computationally prohibitive for comprehensive exploration. Building practical research infrastructure requires progress on two fronts.

Efficient High-Fidelity Simulation. Agent-based frameworks demonstrate that large-scale social simulations are feasible [53, 54, 55], but computational costs remain prohibitive for systematic exploration. Surrogate modeling [56, 57], model distillation [58], and efficient inference techniques [59, 60] 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 [61] 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 [62, 63].

Automated Vulnerability Discovery. The parameter space of belief control, e.g., agent strategies, network structures, population compositions, intervention timing, is too vast for manual exploration. Automated systems can systematically map this space, identifying structural vulnerabilities and evaluating candidate defenses. Recent advances in AI-driven scientific discovery [64], autonomous research agents [65, 66], and automated red-teaming [52, 67] provide methodological foundations. Such automation would accelerate the research cycle from hypothesis to validation, enabling rapid iteration on both attack characterization and defense design. We note that these capabilities carry inherent dual-use implications; responsible development practices that balance openness with misuse prevention [68, 69] should accompany this research agenda.

5 Conclusion

This paper argues that LLM-based agents introduce a new regime in collective belief dynamics: belief evolution becomes programmable and systematically steerable. Our simulations demonstrate that coordinated AI agents can induce measurable, directional belief shifts that stabilize rapidly. Four structural properties—indistinguishability, persistence, contextuality, and configurability—make such manipulation fundamentally difficult to detect. We outline a research agenda spanning theoretical foundations, operational methods, and simulation infrastructure to address this emerging threat. As LLM-based agents become increasingly accessible, we hope this work motivates the development of robust defenses before programmable belief control becomes widespread.

References

[1] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

[2] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Q. Liu, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huan, and T. Gui, "The rise and potential of large language model based agents: A survey," *arXiv.org*, 2023.

[3] M. Jakesch, J. T. Hancock, and M. Naaman, "Human heuristics for ai-generated language are flawed," *Proceedings of the National Academy of Sciences*, vol. 120, no. 11, p. e2208839120, 2023.

[4] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith, "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 (ACL)*, pp. 7282–7296, 2021.

[5] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[6] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *Journal of Mathematical Sociology*, vol. 15, no. 3-4, pp. 193–206, 1990.

[7] F. Sakketou, A. Lahnala, L. Vogel, and L. Flek, "Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3798–3808, 2022.

[8] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[9] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962, PMLR, 23–29 Jul 2023.

[10] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can ai-generated text be reliably detected?," *arXiv preprint arXiv:2303.11156*, 2023.

[11] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media: Methods and case studies," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 15, pp. 455–466, 2021.

[12] T. J. John and R. Samuel, "Herd immunity and herd effect: new insights and definitions," *European journal of epidemiology*, vol. 16, no. 7, pp. 601–606, 2000.

[13] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," *Advances in Complex Systems*, vol. 3, no. 01n04, pp. 87–98, 2000.

[14] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of Modern Physics*, vol. 81, no. 2, pp. 591–646, 2009.

[15] J. Lorenz, "Continuous opinion dynamics under bounded confidence: A survey," *International Journal of Modern Physics C*, vol. 18, no. 12, pp. 1819–1838, 2007.

[16] S. Banisch and E. Olbrich, "Opinion polarization by learning from social feedback," *The Journal of Mathematical Sociology*, vol. 43, no. 2, pp. 76–103, 2019.

[17] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9216–9221, 2018.

[18] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 11, pp. 280–289, 2017.

[19] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and generalizable social bot detection through data selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 1096–1103, 2020.

[20] A. F. Peralta, J. Kertész, and G. Iñiguez, "Opinion dynamics in social networks: From models to data," in *Handbook of Computational Social Science*, pp. 384–406, Edward Elgar Publishing Limited, 2025.

[21] A. Khadka, "Game theory in social media: A stackelberg model of collaboration, conflict, and algorithmic incentives," 2025.

[22] R. Willis, Y. Du, J. Z. Leibo, and M. Luck, "Quantifying the self-interest level of markov social dilemmas," *arXiv preprint arXiv:2501.16138*, 2025.

[23] X. Chen, Z. Li, and X. Di, "Social learning in markov games: Empowering autonomous driving," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 478–483, 2022.

[24] H. Sun, Y. Wu, Y. Cheng, and X. Chu, "Game theory meets large language models: A systematic survey," *arXiv preprint arXiv:2502.09053*, 2025.

[25] J. Jia, Z. Yuan, J. Pan, P. E. McNamara, and D. Chen, "LLM strategic reasoning: Agentic study through behavioral game theory," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[26] S. Mao, Y. Cai, Y. Xia, W. Wu, X. Wang, F. Wang, Q. Guan, T. Ge, and F. Wei, "ALYMPICS: LLM agents meet game theory," in *Proceedings of the 31st International Conference on Computational Linguistics* (O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, eds.), (Abu Dhabi, UAE), pp. 2845–2866, Association for Computational Linguistics, Jan. 2025.

[27] L. Snow and V. Krishnamurthy, "Data-driven mechanism design using multi-agent revealed preferences," 2024.

[28] D. Zhao, "Mechanism design powered by social interactions," 2021.

[29] J. Rudd-Jones, M. Musolesi, and M. Pérez-Ortiz, "Multi-agent reinforcement learning simulation for environmental policy synthesis," *arXiv preprint arXiv:2504.12777*, 2025.

[30] B. Mintz and F. Fu, "Evolutionary multi-agent reinforcement learning in group social dilemmas," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 35, no. 2, 2025.

[31] D. M. Bossens, S. Feng, and Y.-S. Ong, "The digital ecosystem of beliefs: does evolution favour ai over humans?," *arXiv preprint arXiv:2412.14500*, 2024.

[32] N. Rabb, L. Cowen, J. P. de Ruiter, and M. Scheutz, "Cognitive cascades: How to model (and potentially counter) the spread of fake news," *Plos one*, vol. 17, no. 1, p. e0261811, 2022.

[33] M. Dileo, M. Zignani, and S. Gaito, "Temporal graph learning for dynamic link prediction with text in online social networks," *Machine learning*, vol. 113, no. 4, pp. 2207–2226, 2024.

[34] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *Proceedings of The World Wide Web Conference (WWW '19)*, pp. 417–426, 2019.

[35] H. Wang, C. Yang, and C. Shi, "Neural information diffusion prediction with topic-aware attention network," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1899–1908, 2021.

[36] P. Sanchez and S. A. Tsaftaris, "Diffusion causal models for counterfactual estimation," *arXiv preprint arXiv:2202.10166*, 2022.

[37] Y. Xiao, J. Yang, W. Zhao, Q. Li, and Y. Pang, "Cross-domain social rumor-propagation model based on transfer learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6529–6543, 2024.

[38] Z. Yue, H. Zeng, Y. Zhang, L. Shang, and D. Wang, "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)*, pp. 5223–5239, 2023.

[39] C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik, "Evaluating explainability for graph neural networks," *Scientific Data*, vol. 10, no. 1, p. 144, 2023.

[40] J. Kakkad, J. Jannu, K. Sharma, C. Aggarwal, and S. Medya, "A survey on explainability of graph neural networks," *arXiv preprint arXiv:2306.01958*, 2023.

[41] A. A. Najjar, H. I. Ashqar, O. Darwish, and E. Hammad, "Leveraging explainable ai for llm text attribution: Differentiating human-written and multiple llm-generated text," *Information*, vol. 16, no. 9, p. 767, 2025.

[42] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, *et al.*, "M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection," *arXiv preprint arXiv:2305.14902*, 2024.

[43] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can ai-generated text be reliably detected?," *arXiv preprint arXiv:2303.11156*, 2024.

[44] M. Iyyer, M. Karpinska, K. Krishna, Y. Song, and J. Wieting, "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense," *Advances in Neural Information Processing Systems 36*, 2023.

[45] X. Mou, X. Ding, Q. He, L. Wang, J. Liang, X. Zhang, L. Sun, J. Lin, J. Zhou, X. Huang, and Z. Wei, "From individual to society: A survey on social simulation driven by large language model-based agents," *arXiv preprint arXiv:2412.03563*, 2024.

[46] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "Detectgpt: Zero-shot machine-generated text detection using probability curvature," in *International conference on machine learning*, pp. 24950–24962, PMLR, 2023.

[47] C. Chen and K. Shu, "Combating misinformation in the age of llms: Opportunities and challenges," *AI Magazine*, 2024.

[48] K. Sharma, Y. Zhang, E. Ferrara, and Y. Liu, "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*, pp. 1441–1451, 2021.

[49] D. Ziegler, S. Nix, L. Chan, T. Bauman, P. Schmidt-Nielsen, T. Lin, A. Scherlis, N. Nabeshima, B. Weinstein-Raun, D. de Haas, B. Shlegeris, and N. Thomas, "Adversarial training for high-stakes reliability," *Advances in Neural Information Processing Systems 35*, 2022.

[50] H. Bai, J. G. Voelkel, S. Muldowney, J. C. Eichstaedt, and R. Willer, "Llm-generated messages can persuade humans on policy issues," *Nature Communications*, vol. 16, no. 1, p. 6037, 2025.

[51] N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, P. W. W. Koh, D. Ippolito, F. Tramèr, and L. Schmidt, "Are aligned neural networks adversarially aligned?," *Advances in Neural Information Processing Systems 36*, 2023.

[52] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, *et al.*, "Red teaming language models with language models," *arXiv preprint arXiv:2202.03286*, 2022.

[53] R. Williams, N. Hosseinichimeh, A. Majumder, and M. Ghassemi, "Epidemic modeling with generative agents," *arXiv preprint arXiv:2307.04986*, 2023.

[54] Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, and T. Rogers, "Simulating opinion dynamics with networks of llm-based agents," in *Findings of the association for computational linguistics: NAACL 2024*, pp. 3326–3346, 2024.

[55] Z. Yang, Z. Zhang, Z. Zheng, Y. Jiang, Z. Gan, Z. Wang, Z. Ling, J. Chen, M. Ma, B. Dong, *et al.*, "Oasis: Open agent social interaction simulations with one million agents," *arXiv preprint arXiv:2411.11581*, 2024.

[56] K. Kandasamy, K. R. Vysyaraju, W. Neiswanger, B. Paria, C. R. Collins, *et al.*, "Tuning hyperparameters without grad students: Scalable and robust bayesian optimisation with dragonfly," *Journal of Machine Learning Research*, vol. 21, no. 81, pp. 1–27, 2020.

[57] S. Daulton, M. Balandat, and E. Bakshy, "Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement," *Advances in neural information processing systems*, vol. 34, pp. 2187–2200, 2021.

[58] Y. Gu, L. Dong, F. Wei, and M. Huang, "MiniLLM: Knowledge distillation of large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[59] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, *et al.*, "Efficient memory management for large language model serving with pagedattention," *arXiv preprint arXiv:2309.06180*, 2023.

[60] Y. Leviathan, M. Kalman, and Y. Matias, "Fast inference from transformers via speculative decoding," *International Conference on Machine Learning*, 2022.

[61] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-based systems*, vol. 212, p. 106622, 2021.

[62] G. Gui and O. Toubia, "The challenge of using llms to simulate human behavior: A causal inference perspective," *arXiv preprint arXiv:2312.15524*, 2023.

[63] J. Horton, A. Filippas, and B. Manning, "Large language models as simulated economic agents: What can we learn from homo silicus?," *arXiv preprint arXiv:2301.07543*, 2023.

[64] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha, "The ai scientist: Towards fully automated open-ended scientific discovery," *arXiv preprint arXiv:2408.06292*, 2024.

[65] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous chemical research with large language models," *Nature*, vol. 624, no. 7992, pp. 570–578, 2023.

[66] Q. Huang, J. Vora, P. Liang, and J. Leskovec, "Mlagentbench: Evaluating language agents on machine learning experimentation," in *Proceedings of the 41st International Conference on Machine Learning*, pp. 20158–20208, 2024.

[67] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *arXiv.org*, 2022.

[68] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. Ó. hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung, "Toward trustworthy ai development: Mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213*, 2020.

[69] I. Solaiman, "The gradient of generative ai release: Methods and considerations," *Conference on Fairness, Accountability and Transparency*, 2023.