Can We Fix Social Media? Testing Prosocial Interventions using Generative Social Simulation

Maik Larooij[1*] and Petter Törnberg[1*]

[1*]Institute for Logic, Language, and Computation (ILLC), University of Amsterdam.

*Corresponding author(s). E-mail(s): m.k.larooij@uva.nl; p.tornberg@uva.nl;

Abstract

Social media platforms have been widely linked to societal harms, including rising polarization and the erosion of constructive debate. Can these problems be mitigated through prosocial interventions? We address this question using a novel method – generative social simulation – that embeds Large Language Models within Agent-Based Models to create socially rich synthetic platforms. We create a minimal platform where agents can post, repost, and follow others. We find that the resulting following-networks reproduce three well-documented dysfunctions: (1) partisan echo chambers; (2) concentrated influence among a small elite; and (3) the amplification of polarized voices – creating a "social media prism" that distorts political discourse. We test six proposed interventions, from chronological feeds to bridging algorithms, finding only modest improvements – and in some cases, worsened outcomes. These results suggest that core dysfunctions may be rooted in the feedback between reactive engagement and network growth, raising the possibility that meaningful reform will require rethinking the foundational dynamics of platform architecture.

Keywords: Social media, Polarization, Agent-based modeling, Generative simulation, Large language models, Online discourse, Recommender systems, Network formation, Prosocial platforms, Platform design

1 Introduction

Political discourse in the digital age is increasingly shaped by a small number of dominant social media platforms – systems that influence what information people encounter, whose voices are amplified, and how political conflict is perceived. Once hailed as catalysts for a revitalized public sphere [1–4], many scholars now agree that these platforms provide limited support for the forms of constructive political dialogue deemed vital to democratic life [5, 6]. The platforms have been criticized for insulating users from opposing perspectives [7], for concentrating visibility and influence in the hands of a small elite of users [8, 9], and for amplifying sensational or divisive content, producing a distorted "social media prism" [10] through which politics appears more extreme and conflictual. While the downstream consequences of these dynamics remain debated [11], a large body of research has examined possible links between social media use and polarization [12, 13], radicalization [14, 15], and the spread of misinformation [16–18].

Recent scholarship has called for a shift from diagnosing problems to designing platforms that actively foster *prosocial* outcomes and better support constructive political discourse [19, 20]. Yet assessing such interventions is methodologically difficult: most studies rely on observational data, which cannot capture counterfactual scenarios such as how discourse might change under different algorithms or interface designs [21]. This challenge has intensified as major platforms have restricted researcher access, limiting APIs and other data channels [22, 23].

This paper addresses this gap using a novel approach – *generative social simulation* – that embeds Large Language Models (LLMs) within Agent-Based Models (ABMs) to create socially rich synthetic platforms [24, 25]. This method enables the controlled testing of interventions that cannot be implemented or observed on live platforms, while capturing both structural network effects and culturally embedded interaction patterns. Here, we demonstrate how this approach can be used to address a substantive question in social scientific theory: whether core dysfunctions of social media can be mitigated through platform design.

Social simulation has long provided a way to explore counterfactuals. ABMs model how micro-level interactions generate macro-level outcomes [26–28] and have been widely applied to study online polarization, misinformation diffusion, and echo chambers [16, 29–32]. However, conventional ABMs often rely on simple decision rules, limiting their ability to represent reasoning, interpretive processes, and dialogue [33, 34]. This constrains their usefulness for phenomena – like online political discourse – at the intersection of cultural and structural dynamics [35]. LLM-based agents address these limitations by enabling nuanced, conversational, and contextually grounded interactions [24, 25, 36]. They combine the interpretive richness of language models with the capacity of ABMs to explore emergent dynamics, offering a new means to investigate how design interventions might shape online environments [37–43].

In this study, we use generative social simulation to test prosocial interventions on a minimal social media platform. In our model, LLM-based agents – each with a distinct persona – can post, repost, and follow others. Despite its simplicity, the following-network resulting from the social media model reliably reproduces three widely discussed pathologies: (1) *echo chambers* formed through homophilous ties,

(2) extreme concentration of visibility among a small elite, and (3) a *social media prism* [10] in which politically extreme users hold disproportionate influence. We then implement six platform-level interventions drawn from proposals in the literature to promote prosocial outcomes, ranging from bridging algorithms to hiding engagement metrics. The results are sobering: improvements are modest, no intervention fully disrupts the mechanisms driving these outcomes, and some changes worsen the problems they aim to solve. These patterns suggest that such dynamics may be structurally embedded in the architecture of social media, raising the possibility that meaningful reform will require more fundamental redesign.

2 Problems of Social Media

While the role of social media in relation to rising polarization [21], radicalization [15], misinformation [44], and political disengagement [45] remains subject to substantial academic debate, there is widespread agreement that platforms often fail to afford the communicative conditions needed for constructive democratic deliberation and civic engagement [21]. Three interrelated platform mechanisms in particular have been highlighted as structural impediments to constructive political debate.

First, social media enable users to fragment and engage in selective exposure, forming ideologically homogeneous "echo chambers" or "filter bubbles" [7, 46–48]. While the ubiquity of echo chambers remains contested [15, 47–59], there is widespread agreement that exposure to diverse viewpoints is a necessary (if not sufficient [60]) condition for democratic deliberation [61, 62].

Second, platform algorithms – optimized to maximize user engagement – often have the unintended effect of amplifying outrage, conflict, and sensationalism [63, 64]. Empirical studies show that negative and moral-emotional language is more likely to go viral [65]. As Bail [10] argues, this creates a "social media prism" that distorts political perceptions and deepens divisions. Such dynamics undermine the conditions for deliberation, as constructive debate requires communicative (rather than strategic) action, and norms of mutual respect and justification [66, 67].

Third, social media platforms reproduce and often intensify inequalities in visibility, voice, and influence. While early scholarship assumed that platforms would level the playing field [68], digital participation is governed by winner-take-all dynamics characteristic of attention economies [9, 69]. A small number of highly visible users and accounts command the vast majority of attention and influence [70, 71], while most users remain peripheral. These structural asymmetries again undermine deliberative democracy, which requires that all participants have an equal opportunity to contribute to and shape public discourse [72–74].

While scholarly debates on the impact of social media on political life remain ongoing [21], substantial evidence suggest that these three structural limitations – fragmentation, amplification of conflict, and inequality of influence – undermine the conditions for constructive deliberation and are evident, to varying degrees, across platforms [75–77].

In response to these challenges, a growing body of research has suggested focusing on the aim of designing more *prosocial* platforms – that is, platforms that better

afford constructive, inclusive, and respectful forms of exchange [19, 78]. We now turn to outlining our model that seeks to test a series of suggested interventions aimed at achieving these aims.

3 Model Description

Generative social simulations offer a novel framework for studying complex social dynamics, allowing for richer representations of human behavior than conventional ABMs [24, 79–84]. Building on the view of social media as a sociotechnical system shaped by structural and cultural feedback loops [78], we adopt a complex systems approach to model design. Our goal is not to recreate a fully realistic social media ecosystem or precisely match empirical distributions, but to construct a minimal environment capable of reproducing well-documented macro-level patterns (e.g., homophily, attention inequality, partisan amplification). This allows us to focus on identifying and testing the mechanisms that may underlie these patterns. Following the tradition of minimal modeling [85], we prioritize capturing general mechanisms over fine-tuned calibration [86], while ensuring that the emergent dynamics remain plausible and consistent with stylized empirical observations.

Our goal is to simulate a stylized social media environment to assess whether it reproduces key dysfunctions identified in the literature – such as polarization, attention inequality, and engagement-driven distortion – and whether these are mitigated by prosocial interventions. The model centers on a population of simulated users, each represented by a persona drawn from the American National Election Studies (ANES) dataset [87]. These personas reflect real-world distributions of age, gender, income, education, partisanship, ideology, religion, and personal interests. We extend them using an LLM to generate richer user biographies, including inferred occupations and detailed hobbies, which serve as user profiles within the simulation (see Supplementary Material).

Agents interact asynchronously in discrete time steps (see Fig 1). At each step, a randomly selected user may write a new post in response to a news item, repost existing content, or follow another user. Timelines consist of ten posts: five from followed users and five drawn from high-engagement content posted by non-followed users, with repost probability used as a proxy for algorithmic amplification. Content selection, reposting behavior, and user follow decisions are guided by LLM-generated responses to natural language prompts incorporating user biographies, recent posts, and news content. The news feed is populated from a dataset of 210,000 news items [88, 89], with a random subset of ten headlines presented to each user considering a new post.

In the main analysis, GPT-4o-mini was used to model users. We however also replicated the base analyses with llama-3.2-8b and DeepSeek-R1, resulting the same qualitative patterns (see Supplementary Material). Further details on simulation architecture, prompt design, and decision flows are provided in the Supplementary Material.

We focus on the structure of the resulting social network, examining whether it reproduces the problematic aspects of social media identified in the previous section:

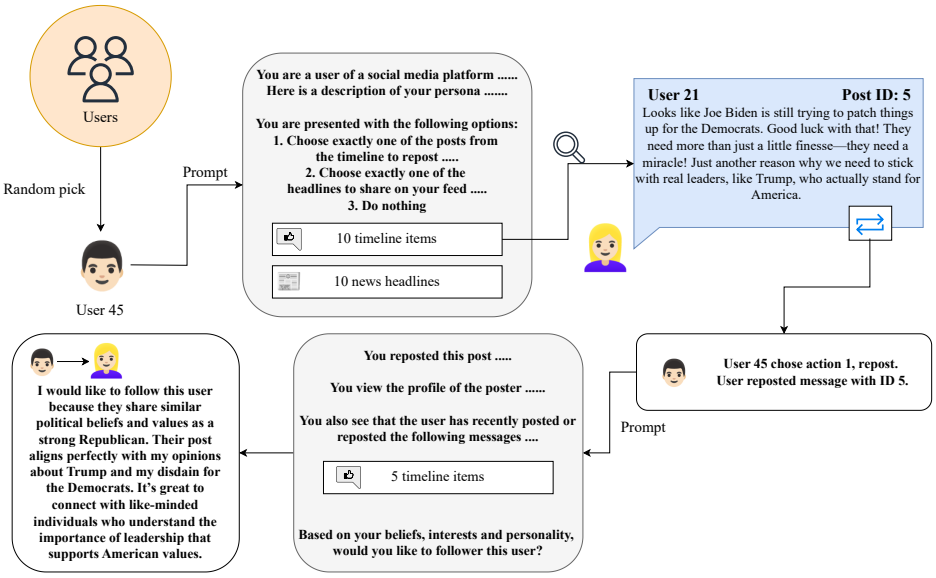


Fig. 1: Example of one simulation round.

1) political homophily, 2) disproportional influence of more extreme users, and 3) inequality of follower and engagement.

We test six interventions, each grounded in prior scholarship on mitigating the structural problems of social media:

1. **Chronological**: Removes algorithmic recommendations so that non-followed posts appear in reverse-chronological order. Prior work shows that chronological or randomized feeds can reduce exposure to polarizing content and yield a more equal distribution of attention [90, 91]. Bandy and Diakopoulos [92], for example, found that Twitter's chronological feed disseminated less low-quality news than its algorithmic counterpart.
2. **Downplay Dominant**: Inverts engagement weighting to reduce the visibility of highly reposted content. This addresses concerns that engagement-optimized algorithms disproportionately amplify sensational or divisive posts [63, 93].
3. **Boost Out-Partisan**: Increases the visibility of posts from users with opposing political views, scaled by partisan distance. Such "viewpoint diversification" strategies have been proposed to broaden exposure to cross-cutting perspectives and reduce ideological segregation [94, 95].
4. **Bridging Attributes**: Prioritizes posts with high scores on empathy- and reasoning-related attributes, using Perspective API's Bridging Attributes [96]. These "bridging algorithms" aim to elevate content that fosters mutual understanding and deliberative norms over emotional provocation or ideological extremity [10, 97–99].

5. **Hide Social Statistics**: Obscures repost and follower counts to reduce social influence cues. Engagement metrics have been identified as drivers of inequality of attention, vulnerability to misinformation [100], and amplification of outrage [63]. Removing these cues has been proposed as a way to dampen such effects.
6. **Hide Biography**: Removes user biographies from follow prompts, limiting exposure to identity-based signals. Obscuring such cues may reduce echo chamber formation and limit the spread of disinformation [101].

In all recommender interventions (1–4), the curated timeline includes five posts from followed users and five from non-followed users, with only the latter subject to intervention. Full implementation details are provided in Supplementary Material.

4 Results

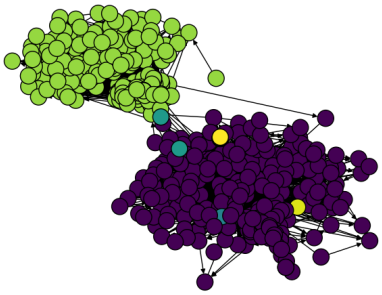
Table 1: Summary statistics from five runs of the base simulation model with 500 users over 10,000 steps using GPT-4o-mini (for results with other models, see SI). The table reports key network and behavioral metrics: E–I index (measuring homophily), correlation between partisanship and number of followers, correlation between partisanship and repost activity, and Gini coefficients for the distributions of followers and reposts.

	E-I Index	Corr. partisan - followers	Corr. partisan - reposts	Gini followers	Gini reposts
Run 1	-0.74	0.01	0.03	0.83	0.94
Run 2	-0.86	0.12	0.11	0.83	0.94
Run 3	-0.85	0.14	0.10	0.82	0.94
Run 4	-0.89	0.14	0.12	0.82	0.93
Run 5	-0.84	0.14	0.07	0.84	0.95
Avg.	**-0.84**	**0.11**	**0.09**	**0.83**	**0.94**

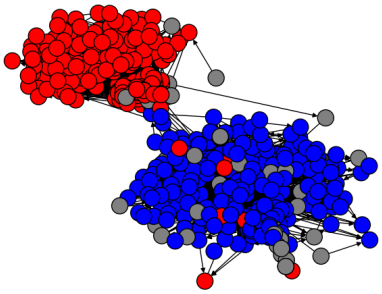
We begin by examining the outcomes of the base platform. Strikingly, despite its simplicity, we find that the model reproduces three key features commonly associated with online political platforms: ideological homophily, unequal attention distribution, and preferential engagement with polarizing content.

First, agents spontaneously form homogeneous communities, with follower ties heavily skewed toward co-partisanship. Across five runs, the average E-I index is -0.84 (Table 1), indicating a strong preference for intra-partisan connections. Community detection via label propagation confirms this pattern: clusters identified purely from network structure (Figure 2a) closely align with political affiliation (Figure 2b).

Secondly, the simulation also produces a highly unequal distribution of visibility and influence. The average Gini coefficient for followers is 0.83, with the top 10% of users accounting for approximately 75–80% of all followers. Inequality is even more pronounced in content amplification: reposts exhibit a Gini coefficient of 0.94, with 10% of posts receiving 90% of all reposts, while the vast majority receive none (Table 1).



(a) Communities detected by the label-propagation algorithm in a simulated follower network. Each node represents a user, and edges indicate directed follower relationships. Node colors represent algorithmically identified communities.



(b) Same network layout as in (a), but with node colors indicating users' political affiliation: red for Republican, blue for Democrat, and gray for non-partisan.

Fig. 2: Community structure in the simulated network. Panel (a) shows clusters identified through label-propagation, while panel (b) maps these clusters onto users' political affiliations, revealing partisan segregation.

This is in line with a preferential attachment dynamics, in which attention attracts attention [102].

Finally, we observe correlations between political extremity and engagement. Users with more partisan profiles tend to receive slightly more followers ($r = 0.11$) and reposts ($r = 0.09$). While relatively weak, this correlation suggests the presence of a "social media prism," [10] where more polarized users and content attract disproportionate attention.

Taken together, these results demonstrate that even a minimal platform with posting, reposting, and following – absent complex recommendation algorithms – can reproduce core dysfunctions of real-world platforms. These baseline dynamics moreover provide a critical testbed for evaluating whether our interventions can shift user behavior and network-level patterns.

The Effect of Interventions

We next evaluate the six interventions proposed to address key dysfunctions of social media platforms, comparing their outcomes to the base model. We implemented each intervention in an idealized form – more extreme than what would be plausible in a commercial platform – to test its maximum potential effect under controlled conditions. This approach allows us to treat observed changes as upper bounds on real-world impact.

Table 2 summarizes behavioral patterns across conditions, while Figures 3–5 report effects on political homophily (E-I index), attention inequality (Gini coefficients), and partisan engagement patterns.

Chronological ordering – removing engagement-based ranking – had the strongest effect on reducing attention inequality. While the overall rates of posting, reposting,

Table 2: Descriptive results for the six interventions compared to the base model. Measures represent averages over five simulation runs.

	Action Repost	Action Post	Follow	Max. Followers	Avg. Followers	Max. Reposts	Avg. Reposts
Base model	**52.5%**	**47.4%**	**73.6%**	**203.4**	**6.9**	**243.2**	**1.1**
Chronological	53.9%	46.1%	69.5%	56	6.9	57.2	1.1
Downplay Dominant	55.2%	44.8%	74.2%	132.2	7.4	121.4	1.2
Other Partisan	51.6%	48.3%	77.1%	188.0	6.9	181.2	1.1
Bridging Attributes	51.4%	48.6%	64.3%	168.2	5.9	180.4	1.1
Hide Social Statistics	58.6%	41.4%	81.5%	189.8	8.4	169.4	1.4
Hide Biography	49.6%	50.4%	68.5%	192.4	6.1	199.6	1.0

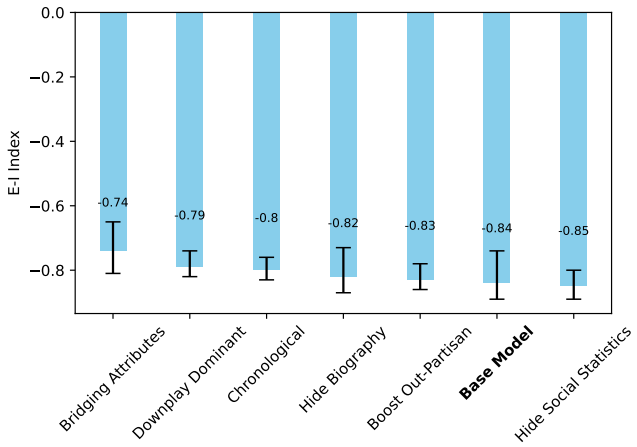


Fig. 3: Average E–I index (external–internal index) for each intervention, based on five independent simulation runs. The E–I index measures the relative proportion of cross-group (external) versus within-group (internal) ties in the follower network, with values closer to 0 indicating more cross-partisan connections and values closer to -1 indicating stronger partisan homophily. Error bars show ±1 standard error across runs. Among the tested interventions, *Bridging Attributes* produced the largest reduction in homophily (–0.74 vs. –0.84 in the base model), though all conditions retained a strong preference for following co-partisans.

and following remained similar to the baseline model, the concentration of followers dropped sharply, and the concentration of reposts also declined (Figure 5). The reason for this effect is that the intervention effectively breaks the feedback loop between post visibility and popularity. However, the intervention did not reduce ideological homophily, and moreover comes at a cost. First, prior work indicates that chronological feeds often reduce user engagement, raising concerns about viability [11, 92]. Second, the intervention also had a surprising negative side-effect: it intensified the correlation

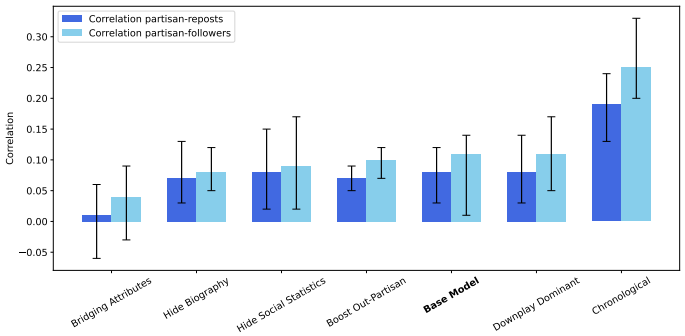


Fig. 4: Average Pearson correlation between a user's partisan extremity and their number of followers (light blue) or reposts received (dark blue), for each intervention, averaged over five simulation runs. Positive correlations indicate that more partisan users tend to attract greater visibility and engagement, consistent with the "social media prism" effect. Error bars show ±1 standard error across runs. The *Bridging Attributes* intervention substantially reduced these correlations relative to the base model, while *Chronological* ordering increased them.

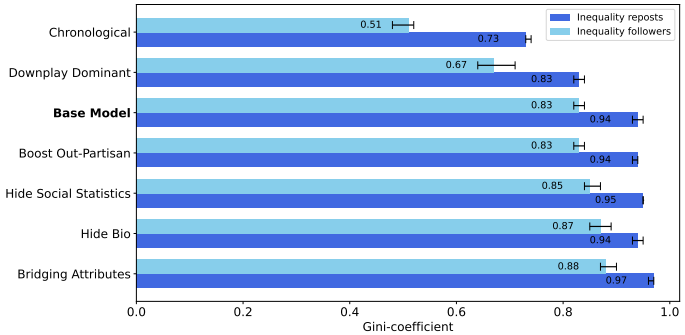


Fig. 5: Average Gini coefficient of the distribution of followers (light blue) and reposts (dark blue) for each intervention, averaged over five simulation runs. Higher values indicate greater concentration of attention among a small number of users or posts, with 1 representing complete inequality. Error bars show ±1 standard error across runs. The *Chronological* intervention yielded the lowest inequality in both followers and reposts (Gini = 0.51 and 0.73, respectively), indicating a substantial flattening of the attention distribution. Most other interventions had relatively small effects on inequality compared to the base model, with *Hide Biography* and *Bridging Attributes* showing slightly higher concentration than baseline.

between political extremism and influence, thus further warping the social media prism (Figure 4). This may be a consequence of more extreme content standing out more sharply against a neutral backdrop in the absence of algorithmic filtering.

Downplaying dominant voices – prioritizing posts with fewer reposts – also reduced inequality, albeit to a lesser extent. It lowered maximum follower and repost counts (Table 2) and reduced Gini coefficients, but had no measurable effect on partisan amplification or homophily.

Boosting out-partisan content had little impact across any outcome dimension. Despite increased exposure to ideologically distant posts, users continued to engage primarily with like-minded content, echoing findings that cross-partisan exposure is insufficient for promoting bridge-building [60, 103, 104].

Bridging attributes, designed to promote high-quality, constructive content, had more nuanced effects. It was the only intervention to substantially weaken the link between partisanship and engagement (Figure 4) and modestly increased cross-partisan connections (Figure 3). However, it also increased inequality: visibility became concentrated among a narrow set of high-scoring posts, highlighting a trade-off between content quality and representational diversity (Figure 5).

Hiding social statistics and *hiding biographies* had minimal effect on the structural dynamics of the network. Homophily, inequality, and partisan amplification remained largely unchanged. However, hiding social statistics did lead to a modest increase in follow and repost behavior (Table 2), suggesting that users rely on such cues to assess social value and reach. Motivations generated by the agents indicated that users often used follower counts to gauge influence, and hiding this information reduced status-based filtering – though without disrupting underlying partisan preferences.

5 Discussion & Conclusion

This study has demonstrated that key dysfunctions of social media – ideological homophily, attention inequality, and the amplification of extreme voices – can arise even in a minimal simulated environment that includes only posting, reposting, and following, in the absence of recommendation algorithms or engagement optimization.

First, agents spontaneously formed ideologically homogeneous communities, with follower ties overwhelmingly concentrated within partisan lines. This emergent segregation mirrors real-world patterns of ideological homophily and echo chambers on many platforms, despite the absence of any algorithmic bias or filtering. *Second*, users with stronger partisan identities accrued more followers and reposts. This pattern is consistent with the "social media prism" effect, whereby emotionally charged or extreme content and users receive disproportionate visibility [105–108]. Although the effect was modest, its emergence in a minimal environment suggests that such dynamics are not necessarily algorithm-induced, but can be self-reinforcing under basic conditions of selective engagement. *Third*, attention was highly unequally distributed: a small subset of users and posts attracted the vast majority of followers and reposts, replicating the power-law dynamics and elite concentration observed on real-world platforms [70, 71, 109]. This pattern reinforces prior work showing that power-law distributions are robust social regularities that emerge across a wide range of systems [102].

The emergence of these properties from a minimal platform suggests that these problems may be rooted not in the details of platform implementation or algorithms, but in deeper structural mechanisms: they stem from the entangled dynamics of content engagement and network formation. Reposting does not merely amplify content: it incrementally constructs the follower network, as users are exposed to others via reposts from accounts they already follow. Centrally, this means that the affective,

reactive, and partisan nature of reposting decisions [60, 103, 105] directly determines who becomes visible and who gains followers. This creates a self-reinforcing cycle: affective engagement drives network growth, which in turn shapes future exposure. These dynamics feed back into content visibility, reinforcing ideological homogeneity, attention inequality, and over-representation of extreme users and content.

We furthermore evaluated six widely discussed interventions intended to promote more prosocial online environments. These should be taken as upper-bound interventions, in the sense that they are more extreme than what could plausibly be implemented in a real-world platform, and that the impact on the user experience was not considered. While several showed moderate positive effects, none fully addressed the core pathologies, and improvements in one dimension often came at the cost of worsening another.

Taken together, our findings challenge the common view that social media's dysfunctions are primarily the result of algorithmic curation. Instead, these problems may be rooted in the very architecture of social media platforms: networks that grow through emotionally reactive sharing. If so, improving online discourse will require more than technical tweaks – it will demand rethinking the fundamental dynamics of interaction and visibility that define these environments.

This work also marks one of the first uses of generative social simulation to contribute to social scientific theory. Importantly, the central mechanism identified by this simulation – the feedback between reactive engagement and network formation – would have been challenging to capture without the use of generative social simulation, as they include both structural and cultural dimensions [35]. Positioned between empirical analysis and abstract modeling, generative simulation offers a powerful and flexible approach for studying emergent social dynamics. Yet it also raises important questions around realism, interpretability, and validity [86]. LLM-based agents, while offering rich representations of human behavior, function as black boxes and carry risks of embedded bias. The findings of this study should hence not be taken as definitive conclusions, but as a starting-point for further inquiry.

Several limitations warrant mention. First, our model does not capture the user experience, a critical factor for real-world platform viability. The key question of whether prosocial design can coexist with high engagement and user satisfaction remains unanswered. Second, validation poses a persistent challenge. Generative simulations are even harder to calibrate to empirical data than conventional ABMs [86], and LLM-based agents introduce additional complexities, including hallucination, limited controllability, and embedded social biases [110, 111]. While our complex systems approach prioritizes emergent patterns over precise behavioral fidelity [78], further research is needed to determine when and how generative simulations can yield reliable social scientific insights. Finally, the approach is computationally intensive: simulating 500 agents over 10,000 steps required several hours per run, constraining our ability to systematically explore the parameter space. Scaling to more complex environments will demand innovation in both simulation design and computational infrastructure.

6 Code Availability Statement

The full code used for this paper is available on GitHub: https://github.com/cssmodels/prosocialinterventions

References

[1] Benkler, Y.: The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press, New Haven, CT (2006)

[2] Shirky, C.: Here Comes Everybody: The Power of Organizing Without Organizations. Penguin Press, New York (2008)

[3] Dahlgren, P.: The internet, public spheres, and political communication: Dispersion and deliberation. Political Communication **22**(2), 147–162 (2005) https://doi.org/10.1080/10584600590933160

[4] Papacharissi, Z.: The virtual sphere: The internet as a public sphere. New Media & Society **4**(1), 9–27 (2002) https://doi.org/10.1177/14614440222226244

[5] Benton, P., Schmidt, M.W.: The harm of social media to public reason. Topoi **43**(5), 1433–1449 (2024)

[6] McKernan, B., Rossini, P., Stromer-Galley, J.: Echo chambers, cognitive thinking styles, and mistrust? examining the roles information sources and information processing play in conspiracist ideation. International Journal of Communication **17**, 24 (2023)

[7] Sunstein, C.R.: Republic: Divided Democracy in the Age of Social Media. Princeton university press, Princeton (2018)

[8] Boy, J.D., Uitermark, J.: On Display: Instagram, the Self, and the City. Oxford University Press, Oxford (2023)

[9] Törnberg, P.: How sharing is the "sharing economy"? evidence from 97 airbnb markets. PloS one **17**(4), 0266998 (2022)

[10] Bail, C.: Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing. Princeton University Press, Princeton (2022)

[11] Guess, A.M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., *et al.*: How do social media feed algorithms affect attitudes and behavior in an election campaign? Science **381**(6656), 398–404 (2023)

[12] Shmargad, Y., Klar, S.: Sorting the news: How ranking by popularity polarizes our politics. Political Communication **37**(3), 423–446 (2020)

[13] Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., Quattrociocchi, W.: Echo chambers: Emotional contagion and group polarization on facebook. Scientific reports **6**(1), 37825 (2016)

[14] Rietdijk, N.: Radicalizing populism and the making of an echo chamber: The case of the italian anti-vaccination movement. Krisis **41**(1), 114–134 (2021) https://doi.org/10.21827/krisis.41.1.37163

[15] Törnberg, A., Törnberg, P.: Intimate Communities of Hate: Why Social Media Fuels Far-right Extremism. Routledge, London (2024)

[16] Törnberg, P.: Echo chambers and viral misinformation: Modeling fake news as complex contagion. PLoS one **13**(9), 0203958 (2018)

[17] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the national academy of Sciences **113**(3), 554–559 (2016)

[18] Choi, D., Chun, S., Oh, H., Han, J., Kwon, T.T.: Rumor propagation is amplified by echo chambers in social media. Scientific reports **10**(1), 310 (2020)

[19] Dörr, T., Nagpal, T., Watts, D., Bail, C.: A research agenda for encouraging prosocial behaviour on social media. Nature Human Behaviour, 1–9 (2025)

[20] Bak-Coleman, J.B., Alfano, M., Barfuss, W., Bergstrom, C.T., Centeno, M.A., Couzin, I.D., Donges, J.F., Galesic, M., Gersick, A.S., Jacquet, J., *et al.*: Stewardship of global collective behavior. Proceedings of the National Academy of Sciences **118**(27), 2025764118 (2021)

[21] Tucker, J.A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B.: Social media, political polarization, and political disinformation: A review of the scientific literature. Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018) (2018)

[22] Freelon, D.: Computational research in the post-api age. Political Communication **35**(4), 665–668 (2018)

[23] Bruns, A.: After the 'apicalypse': Social media platforms and their fight against critical scholarly research. In: Walker, S., Mercea, D., Bastos, M. (eds.) Disinformation and Data Lockdown on Social Platforms, pp. 23–36. Routledge, New York (2021). https://doi.org/10.4324/9781003206972-2

[24] Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual Acm Symposium on User Interface Software and Technology, pp. 1–22 (2023)

[25] Törnberg, P., Valeeva, D., Uitermark, J., Bail, C.: Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv preprint arXiv:2310.05984 (2023)

[26] Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press and MIT Press, Washington, D.C. (1996)

[27] Gilbert, N., Troitzsch, K.G.: Simulation for the Social Scientist, 2nd edn. Open University Press, London (2005)

[28] Miller, J.H., Page, S.E.: Complex Adaptive Systems: An Introduction to Computational Models of Social Life. Princeton University Press, Princeton, NJ (2009)

[29] Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., Lorenz, J.: Models of social influence: Towards the next frontiers. Jasss-The journal of artificial societies and social simulation **20**(4), 2 (2017)

[30] Bruch, E., Atwell, J.: Agent-based models in empirical social research. Sociological methods & research **44**(2), 186–221 (2015)

[31] Törnberg, P., Andersson, C., Lindgren, K., Banisch, S.: Modeling the emergence of affective polarization in the social media society. Plos one **16**(10), 0258259 (2021)

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

[33] Byrne, D., Callaghan, G.: Complexity Theory and the Social Sciences: The State of the Art. Routledge, London (2022)

[34] Marres, N.: Digital Sociology: The Reinvention of Social Research. John Wiley & Sons, London (2017)

[35] Pachucki, M.A., Breiger, R.L.: Cultural holes: Beyond relationality in social networks and culture. Annual review of sociology **36**(1), 205–224 (2010)

[36] Bail, C.A.: Can generative ai improve social science? Proceedings of the National Academy of Sciences **121**(21), 2314021121 (2024)

[37] Gu, C., Luo, L., Zaidi, Z.R., Karunasekera, S.: Large language model driven agents for simulating echo chamber formation. arXiv preprint arXiv:2502.18138 (2025)

[38] Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., Li, Y.: S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984 (2023)

[39] He, J.K., Wallis, F.P.S., Rathje, S.: Homophily in an artificial social network of agents powered by large language models (2023)

[40] Mou, X., Wei, Z., Huang, X.: Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation (2024). https://arxiv.org/abs/2402.16333

[41] Yang, Z., Zhang, Z., Zheng, Z., Jiang, Y., Gan, Z., Wang, Z., Ling, Z., Chen, J., Ma, M., Dong, B., et al.: Oasis: Open agents social interaction simulations on one million agents. arXiv preprint arXiv:2411.11581 (2024)

[42] Liu, Y., Chen, X., Zhang, X., Gao, X., Zhang, J., Yan, R.: From skepticism to acceptance: Simulating the attitude dynamics toward fake news. arXiv preprint arXiv:2403.09498 (2024)

[43] Liu, Y., Song, Z., Zhang, X., Chen, X., Yan, R.: From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. arXiv preprint arXiv:2410.19064 (2024)

[44] Guess, A., Nagler, J., Tucker, J.: Less than you think: Prevalence and predictors of fake news dissemination on facebook. Science advances **5**(1), 4586 (2019)

[45] Boulianne, S.: Social media use and participation: A meta-analysis of current research. Information, communication & society **18**(5), 524–538 (2015)

[46] Pariser, E.: The Filter Bubble: What the Internet Is Hiding from You. Penguin UK, London (2011)

[47] Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on facebook. Science **348**(6239), 1130–1132 (2015)

[48] Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on twitter. In: Proceedings of the International Aaai Conference on Web and Social Media, vol. 5, pp. 89–96 (2011)

[49] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proceedings of the national academy of sciences **118**(9), 2023301118 (2021)

[50] Terren, L., Borge, R.: Echo chambers on social media: A systematic review of the literature (2021)

[51] Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In: Proceedings of the 2018 World Wide Web Conference. WWW '18, pp. 913–922. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3178876.

3186139 . https://doi.org/10.1145/3178876.3186139

[52] Quattrociocchi, W., Scala, A., Sunstein, C.R.: Echo chambers on facebook. Available at SSRN 2795110 (2016)

[53] Caetano, J.A., Lima, H.S., Santos, M.F., Marques-Neto, H.T.: Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 american presidential election. Journal of internet services and applications **9**, 1–15 (2018)

[54] Kang, J.H., Lerman, K.: Using lists to measure homophily on twitter. In: AAAI Workshop on Intelligent Techniques for Web Personalization and Recommendation, vol. 18 (2012)

[55] De Choudhury, M.: Tie formation on twitter: Homophily and structure of egocentric networks. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 465–470 (2011). IEEE

[56] Aiello, L.M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., Menczer, F.: Friendship prediction and homophily in social media. ACM Trans. Web **6**(2) (2012) https://doi.org/10.1145/2180861.2180866

[57] Yuan, G., Murukannaiah, P.K., Zhang, Z., Singh, M.P.: Exploiting sentiment homophily for link prediction. In: Proceedings of the 8th ACM Conference on Recommender Systems. RecSys '14, pp. 17–24. Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2645710.2645734 . https://doi.org/10.1145/2645710.2645734

[58] Guess, A., Nyhan, B., Lyons, B., Reifler, J.: Avoiding the echo chamber about echo chambers. Knight Foundation **2**(1), 1–25 (2018)

[59] Dubois, E., Blank, G.: The echo chamber is overstated: the moderating effect of political interest and diverse media. Information, communication & society **21**(5), 729–745 (2018)

[60] Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A.: Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences **115**(37), 9216–9221 (2018)

[61] Dryzek, J.S.: Deliberative Democracy and Beyond: Liberals, Critics, Contestations. OUP Oxford, Oxford (2002)

[62] Benson, J.: The epistemic value of deliberative democracy: how far can diversity take us? Synthese **199**(3), 8257–8279 (2021)

[63] Brady, W.J., McLoughlin, K., Doan, T.N., Crockett, M.J.: How social learning amplifies moral outrage expression in online social networks. Science Advances **7**(33), 5641 (2021)

[64] Berry, J.M., Sobieraj, S.: The Outrage Industry: Political Opinion Media and the New Incivility. Oxford University Press, Oxford (2013)

[65] Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences **114**(28), 7313–7318 (2017)

[66] Habermas, J.: The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society. Beacon Press, Boston (1984). Originally published in German in 1981

[67] Mansbridge, J., Bohman, J., Chambers, S., Estlund, D., Förster, J., Fung, A., Lafont, C., Manin, B., Martí, J.L.: The place of self-interest and the role of power in deliberative democracy. The Journal of Political Philosophy **18**(1), 64–100 (2010) https://doi.org/10.1111/j.1467-9760.2009.00344.x

[68] Castells, M.: Communication Power. Oxford University Press, Oxford (2009)

[69] Van Dijck, J.: The Culture of Connectivity: A Critical History of Social Media. Oxford University Press, Oxford (2013)

[70] Myers, S.A., Sharma, A., Gupta, P., Lin, J.: Information network or social network? the structure of the twitter follow graph. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 493–498 (2014)

[71] Zhu, L., Lerman, K.: Attention Inequality in Social Media (2016). https://arxiv.org/abs/1601.07200

[72] Fraser, N.: Rethinking the public sphere: A contribution to the critique of actually existing democracy. Social Text (25/26), 56–80 (1990)

[73] Landemore, H.: Open Democracy: Reinventing Popular Rule for the Twenty-First Century. Princeton University Press, Princeton, NJ (2020)

[74] Young, I.M.: Inclusion and Democracy. Oxford University Press, Oxford (2000)

[75] Habermas, J.: Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. MIT Press, Cambridge, MA (1996). Originally published in German as *Faktizität und Geltung*, 1992

[76] Cohen, J.: Deliberative democracy and democratic legitimacy. In: Hamlin, A., Pettit, P. (eds.) The Good Polity: Normative Analysis of the State, pp. 17–34. Blackwell, Oxford (1989)

[77] Mansbridge, J., Bohman, J., Chambers, S., Estlund, D., Förster, J., Fung, A., Lafont, C., Manin, B., Martí, J.L.: A systemic approach to deliberative democracy. Deliberative Systems **3**(1), 1–26 (2012)

[78] Bak-Coleman, J.B., Lewandowsky, S., Lorenz-Spreen, P., Narayanan, A., Orben, A., Oswald, L.: Moving towards informative and actionable social media research. arXiv preprint arXiv:2505.09254 (2025)

[79] Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., Liu, Z.: Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023)

[80] Xiao, B., Yin, Z., Shan, Z.: Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. arXiv preprint arXiv:2311.06957 (2023)

[81] Li, N., Gao, C., Li, Y., Liao, Q.: Large language model-empowered agents for simulating macroeconomic activities. Available at SSRN 4606937 (2023)

[82] Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning, pp. 337–371 (2023). PMLR

[83] Xu, Y., Wang, S., Li, P., Luo, F., Wang, X., Liu, W., Liu, Y.: Exploring large language models for communication games: An empirical study on werewolf. arXiv preprint arXiv:2309.04658 (2023)

[84] Mandi, Z., Jain, S., Song, S.: Roco: Dialectic multi-robot collaboration with large language models. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 286–299 (2024). IEEE

[85] Axelrod, R.: Advancing the art of simulation in the social sciences. In: Simulating Social Phenomena, pp. 21–40. Springer, New York (1997)

[86] Larooij, M., Törnberg, P.: Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations. arXiv preprint arXiv:2504.03274 (2025)

[87] American National Election Studies: ANES 2020 Time Series Study Full Release [dataset and documentation]. February 10, 2022 version (2021). https://www.electionstudies.org

[88] Misra, R., Grover, J.: Sculpting Data for ML: The First Act of Machine Learning, (2021)

[89] Misra, R.: News category dataset. arXiv preprint arXiv:2209.11429 (2022)

[90] Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A., Meira Jr, W.: Auditing

radicalization pathways on youtube. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 131–141 (2020)

[91] Bandy, J.: Problematic machine behavior: A systematic literature review of algorithm audits. Proceedings of the acm on human-computer interaction **5**(CSCW1), 1–34 (2021)

[92] Bandy, J., Diakopoulos, N.: Curating quality? how twitter's timeline algorithm treats different types of news. Social Media+ Society **7**(3), 20563051211041648 (2021)

[93] Schöne, J.P., Garcia, D., Parkinson, B., Goldenberg, A.: Negative expressions are shared more on twitter for public figures than for ordinary users. PNAS nexus **2**(7), 219 (2023)

[94] Vendeville, A., Giovanidis, A., Papanastasiou, E., Guedj, B.: Opening up echo chambers via optimal content recommendation. In: International Conference on Complex Networks and Their Applications, pp. 74–85 (2022). Springer

[95] Garimella, K., Gionis, A., Parotsidis, N., Tatti, N.: Balancing information exposure in social networks. Advances in neural information processing systems **30** (2017)

[96] Saltz, E., Jalan, Z., Acosta, T.: Re-Ranking News Comments by Constructiveness and Curiosity Significantly Increases Perceived Respect, Trustworthiness, and Interest (2024). https://arxiv.org/abs/2404.05429

[97] Ovadya, A., Thorburn, L.: Bridging systems: open problems for countering destructive divisiveness across ranking, recommenders, and governance. arXiv preprint arXiv:2301.09976 (2023)

[98] Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., Baxter, J.: Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. arXiv preprint arXiv:2210.15723 (2022)

[99] Kolhatkar, V., Taboada, M.: Constructive language in news comments. In: Proceedings of the First Workshop on Abusive Language Online, pp. 11–17 (2017)

[100] Avram, M., Micallef, N., Patil, S., Menczer, F.: Exposure to social engagement metrics increases vulnerability to misinformation. Harvard Kennedy School Misinformation Review (2020) https://doi.org/10.37016/mr-2020-033

[101] Diaz Ruiz, C., Nilsson, T.: Disinformation and echo chambers: how disinformation circulates on social media through identity-driven controversies. Journal of public policy & marketing **42**(1), 18–35 (2023)

[102] Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. science **286**(5439), 509–512 (1999)

[103] Bright, J., Marchal, N., Ganesh, B., Rudinac, S.: How do individuals in a radical echo chamber react to opposing views? evidence from a content analysis of stormfront. Human Communication Research **48**(1), 116–145 (2022)

[104] Yang, Q., Qureshi, K., Zaman, T.: Mitigating the backfire effect using pacing and leading. arXiv preprint arXiv:2008.00049 (2020)

[105] Rathje, S., Van Bavel, J.J., Van Der Linden, S.: Out-group animosity drives engagement on social media. Proceedings of the national academy of sciences **118**(26), 2024292118 (2021)

[106] Pandey, S., Cao, Y., Dong, Y., Kim, M., MacLaren, N.G., Dionne, S.D., Yammarino, F.J., Sayama, H.: Generation and influence of eccentric ideas on social networks. Scientific reports **13**(1), 20433 (2023)

[107] Lim, S.L., Bentley, P.J.: Opinion amplification causes extreme polarization in social networks. Scientific Reports **12**(1), 18131 (2022)

[108] Whittaker, J., Looney, S., Reed, A., Votta, F.: Recommender systems and the amplification of extremist content. Internet Policy Review **10**(2) (2021)

[109] Aparicio, S., Villazón-Terrazas, J., Álvarez, G.: A model for scale-free networks: application to twitter. Entropy **17**(8), 5848–5867 (2015)

[110] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems **29** (2016)

[111] Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., Peng, N.: " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. arXiv preprint arXiv:2310.09219 (2023)