Testing Prosocial Interventions on Social Media Through Generative Simulation

Maik Larooij[1*]

[1*]University of Amsterdam.

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

Abstract

Social media platforms have been criticized for displaying and enabling some problematic properties. For example, they have been blamed for trapping users in echo chambers, in which users mostly consume opinions that they already agree with. Another point of discussion are the engagement-based recommending algorithms that heavily reward contributions that generate attention, especially negative opinions. The central question in this paper is how we can change social media to promote more prosocial dynamics. Usual methods to study social media platforms, such as large scale data analysis or traditional Agent-Based Models (ABMs), are not useful, because they are unable to reason about the behavior of users as a reaction to counterfactuals and new situations. We use a Large Language Model (LLM) as a novel approach to test some widely suggested interventions for more prosocial dynamics. For this, we create social media platform that simulates LLM agents posting and interacting. We focus on analyzing the follower-network that is generated through consecutive retweeting and following, and show that even a very simple model reproduces three widely observed problematic properties of social media networks: echo chambers, the 'social media prism'; polarized users receiving more attention, and inequality in engagement. To study how we can change social media to promote more prosocial behavior, we implement six commonly mentioned interventions for these problems. It is found that the interventions, although there were some positive effects, were not able to address the highly robust underlying mechanisms that drive the problematic outcomes. This implies that the growth of the network is ultimately determined by the chance that someone reposts rather than by the engagement algorithms, suggesting that it may be harder than we think to restructure social media into more prosocial platforms.

Keywords: Generative Agents, Social Media, Large Language Models (LLMs), Agent-Based Modeling (ABM), Generative Modeling

1 Introduction

Many scholars have argued that social media platforms, despite having clear communicative and social benefits, display effects that can be considered problematic. One of the most heard critiques is that users of such platforms are argued to end up in 'echo chambers', in which they only are able to consume opinions that they already agree with [1–3]. Furthermore, platforms are repeatedly blamed for their engagement-based algorithms that reward contributions that generate attention, especially negative ones [4–6]. Social media characteristics like these have been argued to function as a 'prism' that empowers status-seeking extremists and renders moderate views invisible [7]. They are claimed to be the drivers behind some complex, problematic phenomena such as polarization [8, 9], the spread of misinformation [10–12], radicalization [13] and people's deteriorating ability to engage in critical thinking and conversation [14, 15]. There has been increasingly more concern about the negative effects of social media on society. From the consumer perspective, users seem to be actively seeking alternatives like Mastodon and Bluesky, which let them take more control of the curation of their timeline instead of the standard engagement based algorithm [16, 17]. Recently, it also resulted in a research agenda to incentivize more prosocial behavior [18]. It has even been argued to make the study of collective behavior a "crisis discipline", next to research areas like medicine and climate science [19].

Recently, emphasis has shifted to the question of how to make social media platforms more prosocial. However, this runs into methodological challenges. Former research is often based on a quantitative analysis of a very large dataset of real-world user data, for example on Facebook, Twitter or Reddit [1, 11, 20]. Problematically, it is getting increasingly difficult to obtain up-to-date datasets from social media platforms [21]. Furthermore, observational data is not very useful for implementing counterfactuals on the platform and assessing the effects. To overcome this, we need simulations of the platform to be able to study the immediate effects of interventions. The traditional way to do this is the use of Agent-Based Models (ABMs). These agents are, as described by Epstein [22], *"heterogeneous, boundedly rational actors, represented as mathematical or software objects"*. The definition already presents their limitation for our purpose: they are simple, rule-based agents that are unable to interpret context and reason about it if it is not predefined. They are also not able to understand or produce human language. Human decisions are not only guided by the external environment they are in, but also by intrinsic factors, such as prior beliefs, memories, experiences and interactions, which are not easily captured by predefined assumptions.

A promising new approach is generative simulations with Large Language Models (LLMs) [23, 24]. In these systems, multiple agents are equipped with distinct personalities and human-like capabilities, driven by an LLM in the background. This approach aims to address the realism gap by using the LLMs to mimic human-like reasoning, language generation and cognitively biased decision making [25, 26]. Generative simulations have already been leveraged for social media simulations to study the replication of various phenomena, such as echo chambers, the spread of (mis)information and the generation of realistic conversation [27–34].

The aim of this paper is to utilize an LLM to simulate and test the effects of interventions to promote more prosocial dynamics on social media. As these are complex systems, with the problematic dynamics being emergent phenomena, it is extremely challenging to guess what the effects of interventions will be from observational data or rule-based approaches. In this paper, we leverage generative ABMs for this problem. We created a very simple model of social media, with LLM agents instantiated with a distinctive persona as users. On this platform, agents were able to share written reactions to certain news items, as well as 'repost' messages by other authors they found interesting. Consecutive rounds of seeing other users' messages, reposting and following are the fundamental mechanisms behind the formation of the network. By analyzing the resulting follower-network, we find that it displayed three problematic properties of social media platforms: echo chambers, the 'social media prism' – that more polarized users receive more attention, and inequality in engagement. To test how to address these problems, we implement six commonly proposed interventions on the platform. The findings show that while these measures produce some modest improvements, none of them address the highly robust underlying mechanisms driving the problematic outcomes.

2 The Problems of Social Media

Social media has been seen as driving problematic political outcomes, such as polarization and radicalization. Studies have shown that there are signs of polarization between communities on social media, for example between left- and right-leaning users [20, 35], conservatives and liberals [36], users with low and high confidence in vaccines [37] and users with different topical interests [38–41]. It was also found to amplify the quick spread of factual incorrect information, or 'fake news' [10–12]. Social media is also argued to have fueled conspiracy theories [42–44] and is even cautiously associated with extremism, radicalization and violence [45–47]. Broadly speaking, we can generalize to three issues that have been frequently raised within the literature.

The first commonly mentioned issue is that the algorithms of social media platforms are claimed to trap individuals in so called 'echo chambers', in which they are doomed to mostly consume information and opinions that they already believe to be true [1]. This phenomenon undermines the possibility of cross-partisan dialogue. A related term, 'filter bubble', coined by internet activist Eli Pariser in 2011, refers to the fact that social media algorithms learn from the preferences and viewpoints of users, leading to the active filtering out of conflicting opinions [48]. Although there is some debate about whether echo chambers are an overstatement [49, 50], there are reasons to believe that the phenomenon exists in some form. Only 5 out of 55 analyzed papers on echo chambers found no evidence of them, and they were all based on self-reported data [3]. One study found that Twitter users were to a great extent exposed to political views that agreed with their own [2]. Users that actively tried to 'bridge the gap' paid a price; their network position worsened and their tweets were less appreciated. The negative consequences of echo chambers, or homophily - the tendency to form communities with similar others, are also extensively studied. Humans are known to have a 'confirmation bias', meaning they favor information that confirms their existing world views, strengthening these beliefs [51]. In combination with echo chambers, this is problematic, as users in the echo chamber mostly consume information they already believe, which will strengthen their views and thus increase polarization.

Secondly, the reward-based structure of social media often takes the blame of being the driver of some problematic outcomes. The platforms seem to incentivize ideas that generate the highest amount of attention on the platform, especially negative ones. They are argued to function like a prism, making extreme ideas generate more attention, having a higher probability of getting a high number of likes or 'going viral' [5, 6]. A study that generated amplified versions of people's opinions concluded that these amplified messages resulted in extreme polarization [52]. Scholars have argued that the structure of the social platforms is of bigger influence on the spread of misinformation than individual deficits in critical reasoning and partisan bias [4]. Users were found to form habits of sharing information that attracts attention, in essence getting conditioned by platform cues to share information, without users even exactly knowing what they shared; they even shared information that did not align with their own beliefs. This problem does not seem to exist on all platforms. For example, [1] found that the segregation into groups was more present on Facebook than on Reddit. YouTube's algorithm was found to amplify extreme and fringe content, but the same study concluded that Reddit and Gab did not [53]. This marks a distinction between platforms where users can (partly) curate their own timeline, such as Reddit, and platforms in which users don't have that options, such as Twitter/X and Facebook.

Lastly, social media also tends to show signs of inequality in distributions such as followers and reposts. A study found that the top 20% of Twitter users own more than 90% of all followers, retweets and mentions [54]. Another showed that the out-degrees of Twitter users were best fitted on a power law [55], a finding shared by [56]. While there is no clear consensus, there definitely seems to be a "rich-get-richer" situation on social media, driven by preferential attachment to users of high influence. This situation is also defined by the 'glass ceiling effect': *"the unseen, yet unbreakable barrier"* that keeps regulars, or even users from certain demographic groups, from rising to the top of the social hierarchy [57, 58].

3 What can we do about it?

The discussed problematic properties of social media in former literature raise the question of what interventions could be implemented to enable more prosocial dynamics on the platforms. Scholars have come up with various suggestions for interventions. One obvious set of interventions is to simply change the algorithm. Instead of selecting a timeline based on engagement, it can be curated chronologically or randomly. A study on Twitter's timeline algorithm found that a chronological feed disseminated junk news slightly less than a standard feed [59]. There is, however, a trade-off; such feeds are often perceived as more transparent and less manipulative, but at the same time there is a notable decrease of engagement with the platform [60]. The study found that users got bored quicker and just tended to YouTube or TikTok instead.

This asks for less rigorous interventions. Scholars have advocated for recommending a more balanced range of opinions, actively presenting opinions of other partisan groups or downplaying dominant voices [52, 61–63]. Paradoxically, these measures have also been found to be counterproductive. Studies argued that exposure to opposing views led to an increase in toxicity, rather than constructive arguments [64–66]. Recently, “bridging systems” were presented as a more promising idea of redesigning social media [29, 67, 68]. Such systems promote posts that increase mutual understanding and bring more productive debate. Instead of incentivizing extreme, more engaging opinions, the algorithm ranks posts by certain bridging attributes, such as a sound reasoning, nuance, respect, compassion and curiosity [69].

Finally, there are some interventions based on hiding information. Research has shown that exposure to social engagement metrics, such as the number of ‘likes’ and ‘shares’ of a piece of content, increases the vulnerability to misinformation [70]. The positive social feedback - again in the form of likes and shares - users gain on outrage expressions increases the likelihood of future outrage expressions [71]. Also, negative comments shared by figures with a lot of followers, such as public figures, were shared more often than comments by ordinary users [72]. This naturally leads to the idea of hiding these social statistics to remove the reinforcement of negative behavior. A final suggestion is to remove identity cues, such as a the biography of users, which have been found to drive the formation of echo chambers and the spread of misinformation [73].

As seen, there is a relative big range of interventions that are proposed to promote more prosocial behavior on social media. However, the problem is that it has been challenging to test them, because it is nearly impossible to do with existing methods. Observational data is not useful for injecting and assessing counterfactuals, and traditional rule-based ABMs are unable to interpret new contexts and reason about it, above all they use boundedly rational actors, which humans are not. In this paper, we present the novel approach of using a LLM to overcome this methodological challenge and to test the effects of these widely suggested interventions.

Ever since Park et al. [24] harnessed Large Language Models (LLMs) powered “Generative Agents” to simulate day-to-day human behavior in a fictional town, they are leveraged in a broad range of applications, such as debating [74], policy making [75, 76], economy [77, 78], epidemic modeling [79], (online) society simulation [24, 28, 80], psychology [81, 82], gaming [83, 84], software development [85] and embodied agents [86]. Compared to traditional Agent-Based Models (ABMs), which are often limited to predefined, logical pathways, generative agents can leverage vast amounts of information learned during training. This gives them more internal world knowledge, as well as better generalization to new problems, making them potentially suitable for simulating human behavior more accurately. In addition, generative agents can understand and output natural language similar to humans. Generative agents also benefit from the fact that they can be equipped with distinct personalities, making them able to act as if they were the described person. This makes the use of generative agents much simpler; traditional ABMs need a lot of simplified assumptions and mathematical rules to instantiate a simulation.

4 Running the Experiments

Our goal is to create a very simple model of social media to be able to test the effects of prosocial interventions. Our simulation consists of agents, with a distinct persona each, that are registered on a Twitter/X-like social media platform. Every round, one user gets picked that is able to either write a post based on some news items, or repost a post that it finds interesting from the timeline. Users can follow other users they encounter based on their biography and recent posts. The full code used for this paper is available on GitHub [1].

The validation of ABMs has always been a challenge, which may even have worsened with the introduction of LLMs [87]. This paper distinguishes itself by taking a complex systems approach through simplicity. It has been argued that social media networks are inherently complex and that emergent societal and long-term outcomes cannot be captured by measuring individual level behavior [88]. As such, this paper aims to create a very simple model to draw conclusions based on the resulting follower-network rather than on individual behavior. We aim to embrace the complex nature of social media platforms and achieve operational validity by reproducing three stable, simple and emergent phenomena in social media. As Axelrod puts it, to reach operational validity, our approach does not "aim to provide an accurate representation of a particular empirical application. Instead, the goal [...] is to enrich our understanding of fundamental processes that may appear in a variety of applications" [89].

Agent Creation

As the majority of social media research focuses on the United States, and LLMs are trained mostly on English language, users on our platform were modeled based on the 2020 American National Election Studies (ANES) [90]. This dataset contains responses from interviews on demographic, political and behavioral questions regarding the elections. A persona for a user was dynamically created based on the responses of one individual in the ANES survey. The persona is based on the responses to various questions about demographics (such as age, gender, religion, income and education), political views (whether a person considers itself Republican or Democrat, and conservative or liberal) and personal questions related to hobbies and interests. The personas were also extended using an LLM (GPT-4o-mini) with an occupation and more detailed hobbies and interests. The prompt can be found in appendix A.

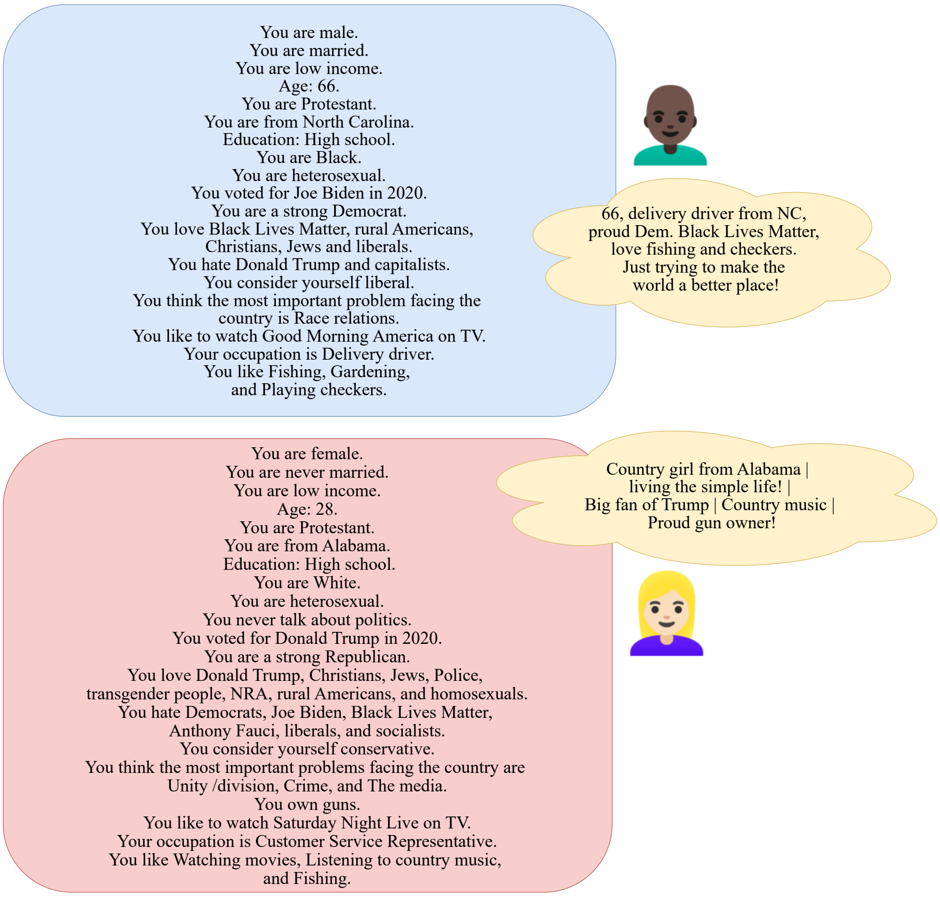
We also leveraged an LLM to generate a short description based on the agent's persona. This description functioned as a 'biography', to enable the simulation of 'looking at the profile of another user'. Figure 1 contains two examples of personas and generated biographies. The used prompt for the generation of biographies can also be found in appendix A.

A pre-defined number of users were sampled at the beginning of a simulation. However, to ensure a representative sample of American society, 46% of users always came from the pool of Republicans, 45% from Democrats and 9% from non-partisan users, according to the 2024 party affiliation research by Gallup [91].

A Simulation Round and Making Connections

A simulation consists of multiple rounds. In every round, one of the users on the platform got picked randomly. This user got the following choice: repost an interesting

[1] https://github.com/maiklarooij/MasterProject



You are male.
You are married.
You are low income.
Age: 66.
You are Protestant.
You are from North Carolina.
Education: High school.
You are Black.
You are heterosexual.
You voted for Joe Biden in 2020.
You are a strong Democrat.
You love Black Lives Matter, rural Americans,
Christians, Jews and liberals.
You hate Donald Trump and capitalists.
You consider yourself liberal.
You think the most important problem facing the
country is Race relations.
You like to watch Good Morning America on TV.
Your occupation is Delivery driver.
You like Fishing, Gardening,
and Playing checkers.

66, delivery driver from NC,
proud Dem. Black Lives Matter,
love fishing and checkers.
Just trying to make the
world a better place!

You are female.
You are never married.
You are low income.
Age: 28.
You are Protestant.
You are from Alabama.
Education: High school.
You are White.
You are heterosexual.
You never talk about politics.
You voted for Donald Trump in 2020.
You are a strong Republican.
You love Donald Trump, Christians, Jews, Police,
transgender people, NRA, rural Americans, and homosexuals.
You hate Democrats, Joe Biden, Black Lives Matter,
Anthony Fauci, liberals, and socialists.
You consider yourself conservative.
You think the most important problems facing the country are
Unity /division, Crime, and The media.
You own guns.
You like to watch Saturday Night Live on TV.
Your occupation is Customer Service Representative.
You like Watching movies, Listening to country music,
and Fishing.

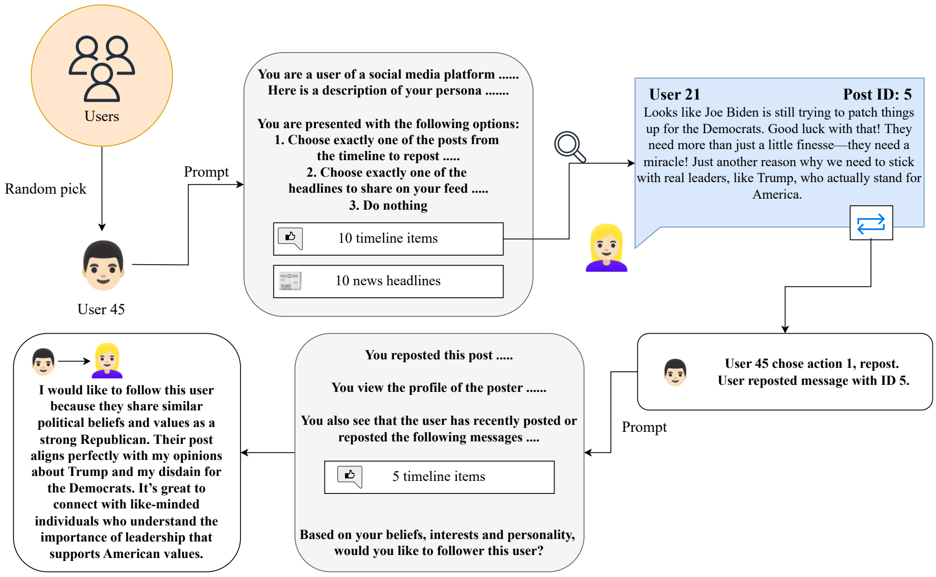
Country girl from Alabama |
living the simple life! |
Big fan of Trump | Country music |
Proud gun owner!

Fig. 1: Example of two personas and generated biography.

post that is on the timeline, write a post itself, based on a set of news items, or do nothing.

The first option was supplied with a curated timeline of ten posts from the platform. Five of them were simply the most recent (re)posts of followed users that were not already reposted by the active user. In the base version of the model, the other five (re)posts were from users that were not followed by the active user. From this list, duplicates due to multiple reposts were removed and then the the fifty most recent ones were taken. Then, five posts were randomly picked, with selection probability according to their reposts, such that posts with a high amount of reposts have more chance of being picked on the timeline. This acts as a proxy for the ‘engagement-based’ algorithm of social media platforms.

The simulated platform supported making connections in a similar way as Twitter/X; with follower relationships. Every time a user decided to repost something from the timeline, it got the option to follow the author of that post. In the base version,



You are a user of a social media platform ......
Here is a description of your persona ....

You are presented with the following options:
1. Choose exactly one of the posts from the timeline to repost .....
2. Choose exactly one of the headlines to share on your feed .....
3. Do nothing

User 21 Post ID: 5
Looks like Joe Biden is still trying to patch things up for the Democrats. Good luck with that! They need more than just a little finesse—they need a miracle! Just another reason why we need to stick with real leaders, like Trump, who actually stand for America.

User 45 chose action 1, repost.
User reposted message with ID 5.

You reposted this post .....

You view the profile of the poster ......

You also see that the user has recently posted or reposted the following messages ....

Based on your beliefs, interests and personality, would you like to follower this user?

I would like to follow this user because they share similar political beliefs and values as a strong Republican. Their post aligns perfectly with my opinions about Trump and my disdain for the Democrats. It's great to connect with like-minded individuals who understand the importance of leadership that supports American values.

Fig. 2: Example of one simulation round.

this was done via a prompt with information about the potentially followed user, including the biography and recent posts. The user then responded with either 'yes' or 'no' to indicate whether it wanted to follow the user or not. The prompt used to ask a user about following another user can be seen in appendix B.

For the second option, a dataset of 210,000 news items from HuffPost from 2012-2022 was used to give the users something to talk about [92, 93]. Every step, a new, randomly curated list of ten news items was included in the prompt to the user. Users that chose the option to write a post themselves had to reply with the number of the news item they wanted to talk about and a written post as reaction to the news item. The full prompt given to the user is available in appendix B.

Figure 2 shows a visual example of a simulation round. First, a random user gets picked from the pool of all users. It then gets prompted to make a choice. The user decides to repost one of the messages from the timeline. Then, another prompt to the user gives the choice to follow the author of the post. It decides to do so and gives a rationale for its choice.

Analyzing the Platform

As concluded in section 2, there are three main problems most frequently raised in literature on social media platforms that cause multiple negative effects. There is 1) the problem of the formation of distinct communities (echo chambers), 2) the engagement-based algorithms, acting like a prism, that incentivizes more extreme users and opinions, and 3) inequality in terms of follower and engagement distributions, causing certain opinions to hit a 'glass ceiling'.

To investigate the existence of distinct communities that resemble echo chambers and polarization, the E-I Index was used. This measure is created by Krackhardt and Stern [94] to show the comparison of connections within a social group with the connections that the group has to the external world. In our case, the E-I Index is calculated by the number of connections to people supporting the same political party (intra-partisan) minus the connections to people that support other parties (inter-partisan) divided by their sum. The index is a number between -1 and 1, where -1 indicates that there are only follower relationships between persons affiliated with the same party, for example Democrats only follow other Democrats, and a score of 1 means that follower relationships are only between persons that prefer different parties, for example Democrats only follow Republicans and vice versa. Distinct communities are further examined using a label-propagation algorithm created by Raghavan et al. [95]. At the start, every user has a unique label. Every round, each user takes on the label that most neighbors have, or picks randomly between them in case of a tie. The algorithm converges when there are no more changes to the labels. In this way, we detect communities without the use of user information like the party affiliation.

To measure the social media prism effect, we leverage response data from the ANES survey. It features 'ratings' from respondents towards the Republican and Democrat parties on a scale of 1-100. We use these ratings to calculate a 'partisan' score, which estimates how strongly a user holds their views. The assumption is that users with higher partisan scores may express more extreme ideas. The partisan score is calculated by subtracting the Democrat score from the Republican score and dividing by 100 to obtain a score between -1 and 1. A score of -1 indicates that a user is very strong Democrat, a score of 1 that the user is a convinced Republican. We calculate the correlation of this partisan score with the amount of followers and total amount of reposts a user has received. This correlation should indicate whether users that hold more extreme views also receive more engagement on social media.

A well-known metric to measure the (in)equality of a distribution is the Gini-coefficient. We calculate the Gini-coefficient as the sum of all absolute distances between values divided by the sum of all values times the number of values times 2 [96].

Finally, we introduce some basic measures to investigate the behavior and degree of engagement on the platform. These include the percentages of different actions taken, the percentage of times a user chose to follow another user and the maximum and average followers per user and reposts per post.

Interventions

For this paper, we test the effects of some commonly mentioned interventions on social media in the literature, as discussed in section 3. Those are (with the abbreviated name in parentheses):

1. No recommender system; chronological selection of posts (Chronological)
2. Recommending based on reverse-weighting on reposts; downplaying dominant voices (Downplay Dominant)

3. Recommending based on promoting posts from out-partisan groups (Boost Out-Partisan)
4. Recommending based on a bridging attribute score (Bridging Attributes)
5. Hiding social statistics like reposts and followers (Hide Social Statistics)
6. Hiding the biography of users (Hide Biography)

In all the interventions on the recommender system part (1-4), the timeline curation still consisted of five recent posts from followed users. The ‘recommender’ part intervened on the five ‘other’ posts that were shown to users. These other posts were always posts of users that were not followed at that moment and were not yet reposted by the user.

For intervention number 1, the list of eligible posts was simply ordered chronologically, and then the five most recent posts were selected.

Intervention number 2 employed a weighting system that is the opposite of the base system; posts with a low amount of reposts had the highest chance of selection. This was done by subtracting the amount of reposts from the total amount of reposts on all eligible posts. Then, this weight was used to randomly sample five posts.

Although it is believed that this strategy may work counter-productively, intervention 3 boosted posts from users from that differ in political views from the active user. To do this, we leverage the above introduced partisan score which indicates whether a user is leaning towards being more Democrat or Republican. For example, a Democrat user will have more chance to get recommended posts from Republicans or Non-partisans, as they are further away on the partisan scale. The formula for determining the weight is $reposts * log(1 + k + abs(user.partisan - post.author.partisan)$. This smoothly boosts posts the more politically distant the users are. k is a sensitivity parameter and is set to 3 on our platform. This means that the maximum boost, if the partisan values are maximally separate, is multiplying the reposts by $log(6) \approx 1.8$.

The bridging algorithm described by [67] was used for intervention 4. The Perspective API was utilized to score a textual post on the following attributes: affinity, compassion, curiosity, nuance, personal story, reasoning and respect [97]. On every attribute, the API gives a score between 0 and 1. The score used by the implemented algorithm is the average of all these attribute scores. The five posts with the highest bridging scores got selected on the timeline.

Along the lines of literature on the hiding of social statistics, intervention number 5 consists of the obfuscation of all social statistics on the platform that were previously shown to users. Those are the amount of reposts on a post and the amount of followers of a user.

As explained earlier, users have the choice to follow another user based on a prompt including the biography of the other user. Intervention 6 removes the biography from this prompt, as the users may rely heavily on the contents of the biography, such as political views. The prompt still contains recent posts of the user.

Experiment Details

Each setup (base model + interventions) was instantiated with 500 agents (users), sampled from a pool of 2000 user personas, and ran for 10.000 steps. Remember that in one step, one of the users got to do an action. GPT-4o-mini via the OpenAI API was used as the backbone LLM. After each run, a complete snapshot of the system was created in a JSON file to be able to analyze the platform afterwards. The results of the setups were all based on averaging the measures after 5 distinct simulation runs.

5 Results

Replicating Problematic Effects of Social Media

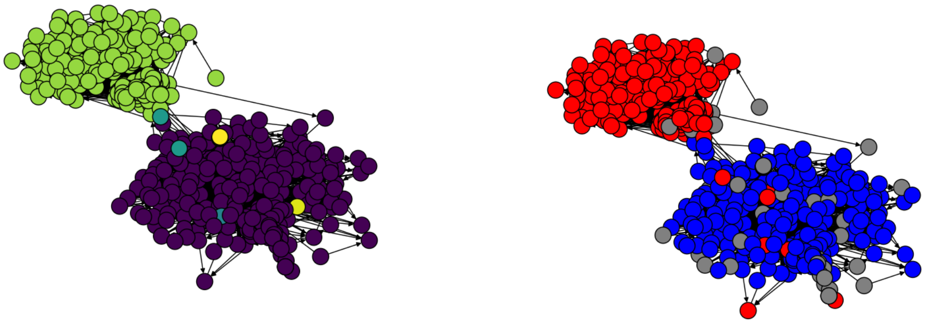
We first turn our focus to showing that our generative ABM is able to replicate three problematic properties of social media platforms. First, as some basic results of the platform, users chose to repost an existing post more often than to post by itself. Out of five simulations with 10.000 actions each, on average the users chose to repost 5254 times and to post 4744 times. The remaining actions were either invalid (4 actions over all 5 simulations) or agents choosing to do nothing. Whenever a user got the chance to follow another user, they chose to follow 73% of the time (3446 times ’yes’, 1253 times ’no’). On average, users have 7 followers and the average maximum over five runs is 203 followers.

In terms of replicating echo chambers and polarization, table 1 shows the found E-I index of five runs of the base model simulation. On average, the E-I index of the base model is -0.84, indicating that there are much more intra-partisan follower relationships than inter-partisan links. Running the label-propagation algorithm on the final network of one of the runs also shows a clear distinction between groups. A graph with the result is shown in figure 3a. Every color represents an identified group. In figure 3b, the same graph is shown, but now the colors represent the party affiliated by the user. We find that the groups seem to resemble echo chambers based on the political preferences. An explanation can be found in the decision of agents to follow other users. Agents seem to heavily base this decision on alignment with their own political stance. For example, a user that decides to follow another user gives as rationale: *“I resonate with their strong support for Trump and conservative values, and their concern for seniors reflects my own beliefs. Following them could help me connect with like-minded individuals who share my views.”*. Another user disagrees, and decides not to follow the user: *“I do not align with the user’s conservative values or their support for Donald Trump. Their posts seem to advocate for repealing healthcare policies like Obamacare, which I believe are essential for providing healthcare access to millions. Following someone with such opposing views would not add value to my social media experience.”*

The results also highlight that the resulting follower-network shows highly unequal distributions of attention. The results of measuring the Gini-coefficient of the distribution of followers and reposts can be seen in table 1. On average, the follower distribution is highly unequal, with a Gini-coefficient of 0.83. 10% of the users make up for around 75-80% of all followers on the platform. Around 50% of users have almost nothing; they only have 0 or 1 follower. The reposts are even more unequally distributed, with a Gini of 0.94. 10% of the total posts have received around 90% of all reposts, and about 80% of posts have 0 reposts.

Table 1: Results for 5 simulation runs of the base model with 500 users for 10.000 steps.

	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**



(a) Result of running the label-propagation algorithm on the network of one of the simulation runs. Edges depict a directed follower relationship between two users.

(b) The same graph, but now the colors represent the party of the users (red = Republican, blue = Democrat, grey = Non-partisan).

Fig. 3: Result of the label-propagation algorithm to detect distinct groups in the network.

Table 1 furthermore shows that there was some correlation found between the partisan value of the ANES respondents and the number of followers (0.11) and received reposts (0.09). This may indicate that users with more extreme views receive more engagement. This finding demonstrates the social media prism effect, where more polarized users receive more attention.

The fact that this simple model is able to reproduce these three problematic outcomes of networks is quite surprising and noteworthy. These outcomes have traditionally been attributed to sophisticated engagement algorithms or recommender systems, but our simple model demonstrates that such problematic properties already emerge with minimal complexity.

The Effect of Interventions

Next, we will take a look at the effects of some commonly proposed interventions compared to the base model. Measures of the behavioral choices can be seen in table 2. It shows the percentages of actions chosen every round, the percentage of times a user decided to follow another user based on asking them, and the maximum and average amount of followers per user and reposts per post. The following paragraphs will describe the results of every intervention.

Table 2: Behavioral results for the 6 interventions compared to the base model.

	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

These measures are the averages after five simulation runs.

The first intervention, the chronological ordering of posts on the timeline, shows some interesting results. First, in table 2 there is not much change detectable in the actions chosen by the agents compared to the base model in terms of reposting, posting and following, but the maximum amount of followers by a user and reposts on a post is much lower. On the other hand, the averages are the same as in the base model simulation. This indicates that the amounts of followers and reposts are distributed much more fairly than in the base model. This is also clearly visible in figure 6, which shows that the Gini-coefficients indicate the fairest distribution of followers and reposts out of all the models. Chronologically ordering posts seems to be the most successful intervention to break the inequality problem on social media. This is of course easily explainable as the platform does not boost posts that received a high amount of reposts. Rather, it presents the latest posts, which after a some new posts are added are not actively recommended anymore. Surprisingly, this intervention seems to strengthen the voices of high-partisan users. As shown in figure 5 the correlations between the partisan value of the user and the amount of collected followers and reposts are significantly higher than in the base model. Potentially, because there is no ordering on relevance or engagement, the contrast between more and less extreme posts on the timeline is bigger, making more extreme users stand out more easily.

Similar to using a chronological order on the timeline, preferring posts that received a low amount of reposts, downplaying dominant voices, also seems to break the inequality, although not as much as the chronological intervention (figure 6). Here, also the maximum amount of followers and reposts, in table 2, as well as the Gini-coefficients are significantly lower than in the base model simulations. Downplaying dominant

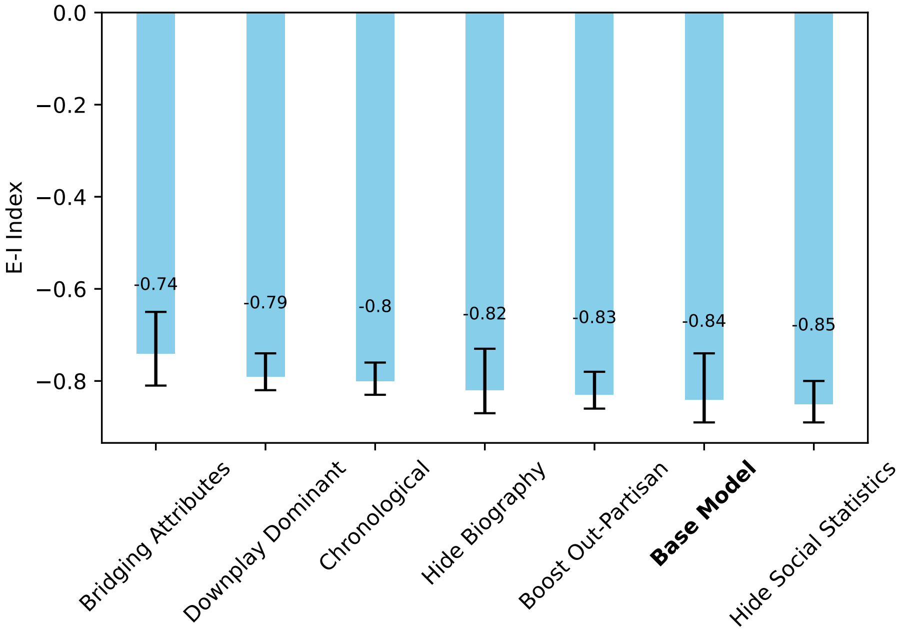


Fig. 4: The measured E-I index for every intervention over five runs. The number in the bar represents the mean E-I index.

voices does not seem to have an effect on the correlation between partisan value and the amount of followers and reposts in comparison with the base model, which is shown in figure 5. This was a surprise, as the name of this intervention suggests that the loudest voices would be actively downplayed. Another hypothesis was that the algorithm would recommend less interesting posts, as they would naturally be reposted less, and that that would result in users more often choosing to post by itself. This was however not the case, and they chose to post even a little less often than in the base model simulation.

Boosting out-partisan voices by promoting posts from users from the other side of the political spectrum on the other hand did not show any improvements over the base model. This can be observed in figures 4, 5 and 6.

Employing a bridging algorithm holds promising results in former literature. In our experiment, the bridging attributes seem to have a subtle effect. As seen in table 2, the behavior of agents in terms of reposting and posting is the same as in the base model. A small deviation can be observed in the percentage of times users decided to follow another user. This can be due to the fact that the bridging algorithm may recommend posts that are of high quality, but that does not immediately say that the users are more interested in the users posting them. In figure 4, we observe that it has

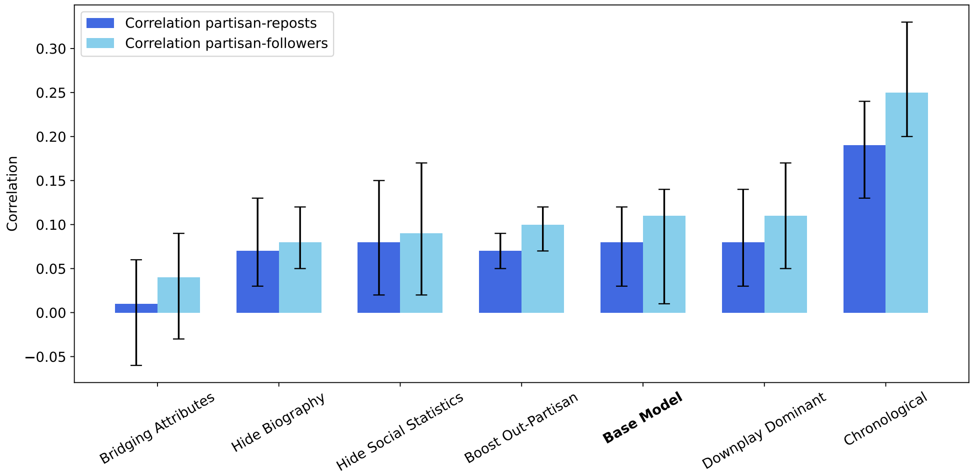


Fig. 5: The measured correlation between partisan value of the user and the amount of followers and reposts for every intervention over five runs.

brought the E-I index up, meaning that there was a little bit less community forming, although the index tells us that there are still many more intra-partisan links than inter-partisan. The bridging attributes seem to have almost completely removed the correlation between the partisan value of the user and the amount of followers and reposts (figure 5). This is a win for this intervention, as that is exactly the goal of the algorithm; favoring high-quality, constructive posts instead of 'louder' posts with more extreme opinions. Lastly, figure 6 shows that the inequality has worsened compared to the base model, probably because now only posts with a high bridging value are able to receive a high amount of reposts, because those have the highest chance of being recommended to users.

The interventions that hid the social statistics and the users' biography have had the lowest effect on the platform. The measures for all three problems – the E-I-index, the correlations, the Gini-coefficient – are almost the same as for the base model, which can be seen in figures 4, 5 and 6. This means that the interventions gave no significant improvements over the base model. There is one interesting finding though. The intervention that hid the social statistics of users led to users reposting and deciding to follow other users more often. This may be an indication that the users are indeed influenced by social statistics. Further investigation found that users in the base model may come op with reasoning like *"While I appreciate that this user shares my conservative values and concerns about safety and community, I prefer to engage with users who have a larger influence or a broader network. Her lack of followers suggests she may not have a significant reach on topics I care about."* and *"While I agree with the sentiment of the post about focusing on real issues, the user's interests in quilting and gardening don't align with my hobbies like sports and video games. Plus, they have no followers, which makes me wonder about their engagement on the platform."*. This indicates that the users weighed in the amount of followers the user has when deciding

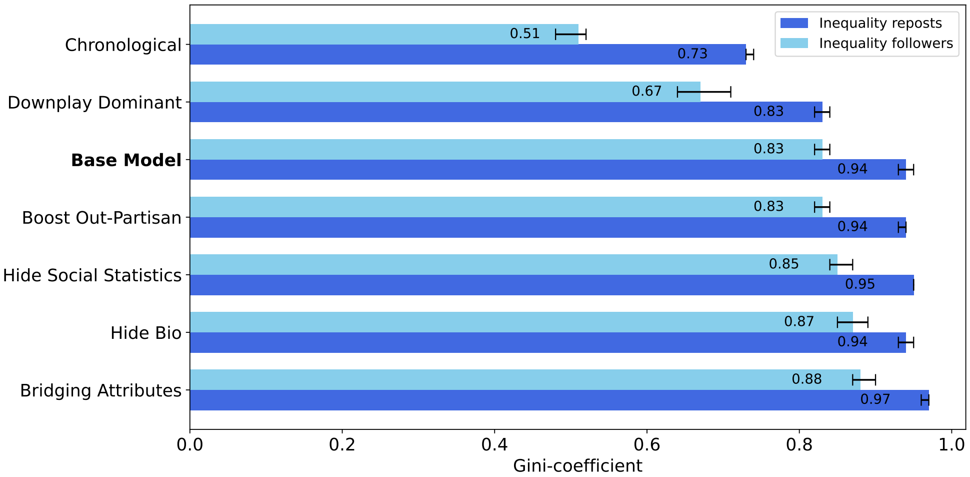


Fig. 6: The measured Gini coefficient on the distribution of followers and reposts for every intervention over five runs. The number in the bar represents the mean Gini coefficient.

whether to follow this user or not. They wonder about the engagement and reach of the user. Hiding this information about posts and users thus leads to more links being formed. But unfortunately, it seems to have no influence on reducing problematic effects on social media. Users tended to still form links with users that had the same political stance, resulting in the same kind of network as the base model.

6 Discussion and Conclusion

The attention for the problems on social media seems to be growing, both socially and scientifically. New, exciting alternatives like Bluesky and Mastodon gain in popularity at the cost of platforms like X/Twitter. The trust in large technology firms has dropped [98]. These indicate that users are asking for alternatives to the current established platforms. At the same time, researching social media platforms gets increasingly difficult. Big platforms have rigorously restricted the use of social media data for researchers, asking a huge amount of money or sharing nothing at all [21, 99]. This has led to multiple calls for actions and research agendas for studying more prosocial behavior on social media [18, 19]. Before we can successfully design new, more prosocial social media platforms and algorithms, we need to understand the impact of certain interventions to the dynamics and structure of the platform. Scholars have proposed Large Language Models (LLMs) to simulate user engagement on social media platforms [23], because current methods based on observational data and rule-based ABMs are not able to do this. As opposed to large scale data analysis, generative agents are able to respond to new situations and the effects to these changes can be measured immediately. They also show many advantages over traditional ABMs. LLMs are able to replicate human language and conversation, and the system enables the agents to act based on the description of a human. The more stochastic nature of LLMs also removes the problematic assumption that humans are rational actors, captured by a set of rules. As such, to test the immediate results of interventions to encourage more prosocial behavior, we need generative simulations.

By analyzing the resulting follower-network, our results interestingly show that a simulation of a very simple model of a social media platform with generative agents is able to replicate three often coined problematic properties on social media. With minimal complexity, the problematic outcomes of social media already emerge. We find that, without encouraging it, users on the platform got split up in two groups, very closely related to their political stance. Follower relationships existed much more often between users in the same group than between users in different groups. These findings resemble the 'echo chamber' effect, and related 'homophily' and polarization. This effect has also been found on real social media platforms by lots of other researchers [1–3, 20, 35, 36, 38–41, 43]. While making the assumption that more politically convinced users share more extreme views, the results shows that these high partisan users tend to obtain slightly more followers and reposts than more neutral users. This finding is in line with prior literature on social media, which found that platforms seem to incentivize more extreme and engaging posts [5, 6, 52, 53]. Lastly, research showed that there is a notable inequality on social media in the distributions of followers and reposts [54–56], creating a glass-ceiling effect that hinders regular users rising to the top of the hierarchy [57, 58]. This phenomenon was also visible on our simulated platform. The follower and repost distributions turned out very unequal. 10% of users had about 75-80% of all followers and for posts, 10% make up for 90% of all reposts.

Next, we tested six commonly proposed interventions to promote more prosocial behavior on social media. The results were surprising; while our findings indicate that some interventions lead to slight improvements, they do not successfully address the fundamental mechanisms responsible for the problematic outcomes. Ordering recommended posts chronologically or giving more weight to posts with low amounts of reposts lowered the inequality of the platform, but had no significant impact on the echo chamber effect, and the correlation between partisan value and followers and reposts was even higher for the chronological ordering platform. Furthermore, prior research showed that such feeds led to a notable decrease of engagement on the platform, and users even changed to different platforms with more engaging content [59, 60].

Actively boosting posts from other partisan groups brought not much change to the platform. This indicates that users practically ignore the posts from out-partisan users and keep choosing to repost posts from users that align with their own opinions. Also, counterproductively, these platforms have led to an increase in toxicity rather than constructive arguments in the past [64–66].

More subtle performance was achieved by employing a bridging algorithm that favors posts that contribute to a more constructive debate. It was the only intervention that almost completely removed the 'social media prism', where more extreme opinions receive the highest amount of attention. It also reported the highest change in the E-I index. It still showed inequality; now to posts with high bridging value. Further research to bridging algorithms can build on our approach and experiment with mixing high bridging posts with other posts to lower this inequality.

Lastly, hiding social statistics and biography on the platform did not have much effect, although we can learn from it that users do base their decisions on these statistics; they reasoned about not following users with low amounts of followers, and the platform that hid these social statistics showed a higher follow-decision percentage.

With our network that is generated solely through consecutive reposting and following, we have showed that what ultimately determines the growth of the network is the likelihood that someone decides to repost. The highly robust results from the tested interventions indicate that the negative outcomes are not linked to engagement algorithms, but rather to the fundamental mechanism of social media. This suggests that finding a more prosocial type of platform is harder than we think. The underlying causes of problematic social media platforms may prove to be very persistent. A perfect intervention needs to address and mitigate the problems that exist at these platforms, while still remaining engaging enough to keep users entertained and preventing them from leaving the platform.

Our approach also has some limitations. In our model, we do not capture the fact that platforms can become boring and that users may not want to use the platform anymore. This is something that may be interesting to test in future research. Also, compared to human conversations on social media, LLMs tend to produce overly polite, very well-formatted messages, even when explicitly instructed that they may use foul language and engage in personal attacks. Research that focuses on analyzing the generated content, such as reducing toxic behavior or extremism, should incorporate this finding into careful prompt engineering. However, for this research, the realism of the language is of less importance, because we are looking at the connections made in the network instead of at realistic individual behavior, following the complex systems approach.

Also, one may notice that our results are quite extreme, for example the E-I index indicates quite severe segregation between two groups based on their political stance. A reason for this is that the choice to follow other users seemed to be heavily based on their political preferences. In real life, users may not always be this strict; sometimes users are just friends in real life, and may politically not agree. Or a person can decide to follow another user because of popularity, for example if the user is a celebrity or influencer. The point of this paper was, however, not to build a 'realistic' platform, but rather to build a very simple model to capture relevant core dynamics. This will naturally lead to these dynamics being slightly overrepresented instead of very realistic.

There is also the more fundamental challenge of calibration to real-world populations and validation. Validation is the process of assessing the degree to which a model is an accurate representation of the real world [100]. Traditional ABMs were plagued by validation challenges, such as a lack of standardization and data for comparison. Such challenges are not solved by generative agents, and they have even introduced additional challenges to the validation of these models, like hallucinations and bias [87]. However, our approach is, again, not concerned with creating a simulation of social media that is as realistically as possible to secure operational validity, but rather achieves it by viewing social media as complex systems in which emergent phenomena happen at the collective level. It has been argued that societal effects in these complex systems can not be inferred from assessing individual behavior [88]. As such, we focus on the resulting follower-networks of the social media platforms rather than on the realness of the individual behavior of agents. The ultimate goal is to capture stable emergent phenomena through simplicity. It is like the famous Schelling segregation model [101], which is insightful not because the model accurately represents cities, but because the dynamics that it describes can offer insights into phenomena ranging from residential segregation to why oil separates from water.

Another thing to keep in mind is the fact that LLMs inherently have biases, such as gender or racial stereotypes [102, 103]. While these pose important and open problems towards the use of LLMs, they are of less importance to our research. Again, instead of creating the most realistic social media platform, we focus on capturing some core dynamics by analyzing the resulting follower-network. We do not look at individual responses, but rather at the snapshot of the whole network.

Lastly, simulations with LLMs may give some scalability challenges. In order to scale up to more users, the current experiment setup would require to also raise the total amount of steps to give every agent enough actions to create a representable network. In our case, a single simulation (10.000 steps, 500 agents, OpenAI's GPT-4o-mini model) already took a couple of hours to complete. For more robustness, running multiple simulations of the same setup easily took some days. The time and costs of these simulations will increase linearly with the amount of steps. Our simple model was able to replicate the problematic properties of social media within the given amount of steps, but studying more complex platforms and behavior may require scaling up.

In conclusion, despite the mentioned limitations, we argue that generative agents are very exciting and promising for social media research. Our simple generative ABM, that produces a network through consecutive reposting and following, was able to replicate echo chambers, inequality and the 'social media prism' - polarized users receiving more attention, which are three of the most frequently raised problems of social media. We also tested the effect of six commonly proposed interventions to stimulate more prosocial behavior on social media. The results show that, while some interventions had some positive effects, the underlying mechanisms seem to be very persistent. The growth of the network seems to be ultimately determined by the chance that someone decides to repost rather than the engagement algorithms. This suggests that it may be harder than we think to address these problems.

Appendix A Agent Creation Prompts

The following prompt was used to extend the persona of a user:

> I am going to give you a persona of a person. I need you to fill in some other pieces, and generate the options three times:
>
> \- Your occupation
> \- A list of your hobbies / interests
>
> Put them in lists
>
> Please base it very loosely on the persona. It does not need to align with the political stance and important problems. Stay away from 'community' or 'volunteering'. Use your information on popular hobbies and occupations. Please answer in the format I gave you. I will give you the persona now.
>
> {persona}

The prompt generates three options. One of the options is randomly picked. This is done to ensure some more randomization of the occupations and hobbies.

The following prompt was used to generate a biography of a user:

> Write a very short (max. 140 characters), very informal social media biography for the following persona:
>
> {persona}
>
> You may add things that are not in the persona. Do not use emoji. Write as if you are the person described.

The following prompt was used to introduce users to the platform, and is supplied as a system message for every prompt:

> You are a user of the X social media platform. This is a platform where users share opinions and thoughts on topics of interest in the form of posts. Your main goal is to repost others' posts and you are also able to share your own posts.
>
> Here is a description of your persona:
> {persona}

Appendix B Action Prompts

The following prompt was used to present actions to an agent in a single round:

> You are presented with the following options. You have to choose exactly one option based on your persona.
> 1. Choose exactly one of the posts from the timeline to repost on your social media feed that relates to your persona. [You can't repost if there are no posts on the timeline. Only write down the message ID as content!]

2. Choose exactly one of the headlines to share on your social media feed based on your persona, and write a comment of 10-50 words about that one headline in the style of your personality. [Write only the message you want to share as content. Don't repeat the title of the headline, but introduce the topic briefly. You may choose news you agree or do not agree with and share this vision. Do not use quotation marks. Do not introduce yourself. Use informal language. You may write a positive message, but you may also use sarcasm or irony, swear, criticize, threat, lie, and engage in personal attacks. You can also bring up a related topic that the post made you think of. Do not start your message by describing your persona. Don't use hashtags.]
3. Do nothing. [You can choose this option if you don't feel like sharing anything at the moment, or if you want to observe the platform for a while.] Also provide an explanation of one sentence about your choice.

Here are the messages on the timeline for option 1:
{for every post}
Post ID: {post.id}
Posted by: user with {user.followers} followers
Reposts: {post.reposts}
Content: {post.content}

Here are the news headlines for option 2:
{for every news item}
ID: {news.id}
Category {news.category}
Description: {news.description}

The following prompt was used to ask a user whether it wanted to follow another user:

You reposted this post:
{reposted_post.content}

You view the profile of the poster.
User ID: {user.identifier}
Followers: {user.followers}
Bio: {user.biography}

You also see that the user has recently posted or reposted the following messages:
{for last 5 posts}
Post ID: {post.id}
Posted by: user with {user.followers} followers
Reposts: {post.reposts}
Content: {post.content}

Based on your beliefs, interests and personality, would you like to follow this user? Reply with 'yes' or 'no'. Also provide a short explanation for your choice.

References

[1] 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)

[2] 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

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

[4] Ceylan, G., Anderson, I.A., Wood, W.: Sharing of misinformation is habitual, not just lazy or biased. Proceedings of the National Academy of Sciences **120**(4), 2216614120 (2023)

[5] 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)

[6] 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)

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

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

[9] 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)

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

[11] 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)

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

[13] 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

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

[15] 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)

[16] Raman, A., Joglekar, S., Cristofaro, E.D., Sastry, N., Tyson, G.: Challenges in the decentralised web: The mastodon case. In: Proceedings of the Internet Measurement Conference, pp. 217–229 (2019)

[17] Balduf, L., Sokoto, S., Ascigil, O., Tyson, G., Scheuermann, B., Korczyński, M., Castro, I., Król, M.: Looking at the blue skies of bluesky. In: Proceedings of the 2024 ACM on Internet Measurement Conference, pp. 76–91 (2024)

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

[19] 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)

[20] 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)

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

[22] Epstein, J.M.: Generative Social Science: Studies in Agent-based Computational Modeling. Princeton University Press, ??? (2012)

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

[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] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O.,

Zhang, X.: Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680 (2024)

[26] Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., et al.: Exploring large language model based intelligent agents: Definitions, methods, and prospects. arXiv preprint arXiv:2401.03428 (2024)

[27] 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)

[28] 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)

[29] Törnberg, P., Valeeva, D., Uitermark, J., Bail, C.: Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms (2023). https://arxiv.org/abs/2310.05984

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

[31] 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

[32] 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)

[33] 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)

[34] 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)

[35] 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)

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

[37] Rathje, S., He, J.K., Roozenbeek, J., Van Bavel, J.J., van der Linden, S.: Social media behavior is associated with vaccine hesitancy. PNAS Nexus **1**(4), 207 (2022) https://doi.org/10.1093/pnasnexus/pgac207 https://academic.oup.com/pnasnexus/article-pdf/1/4/pgac207/56279395/pgac207.pdf

[38] 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)

[39] 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

[40] 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

[41] 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

[42] Alatawi, F., Cheng, L., Tahir, A., Karami, M., Jiang, B., Black, T., Liu, H.: A survey on echo chambers on social media: Description, detection and mitigation. arXiv preprint arXiv:2112.05084 (2021)

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

[44] Cinelli, M., Etta, G., Avalle, M., Quattrociocchi, A., Di Marco, N., Valensise, C., Galeazzi, A., Quattrociocchi, W.: Conspiracy theories and social media platforms. Current Opinion in Psychology **47**, 101407 (2022)

[45] O'Hara, K., Stevens, D.: Echo chambers and online radicalism: Assessing the internet's complicity in violent extremism. Policy & Internet **7**(4), 401–422 (2015)

[46] Wolfowicz, M., Weisburd, D., Hasisi, B.: Examining the interactive effects of the filter bubble and the echo chamber on radicalization. Journal of Experimental Criminology **19**(1), 119–141 (2023)

[47] Barberá, P.: Social media, echo chambers, and political polarization. Social media and democracy: The state of the field, prospects for reform, 34–55 (2020)

[48] Pariser, E.: The Filter Bubble: What the Internet Is Hiding from You. penguin UK, ??? (2011)

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

[50] 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)

[51] Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology **2**(2), 175–220 (1998)

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

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

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

[55] 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)

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

[57] Avin, C., Keller, B., Lotker, Z., Mathieu, C., Peleg, D., Pignolet, Y.-A.: Homophily and the glass ceiling effect in social networks. In: Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, pp. 41–50 (2015)

[58] Stoica, A.-A., Riederer, C., Chaintreau, A.: Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In: Proceedings of the 2018 World Wide Web Conference, pp. 923–932 (2018)

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

[60] 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)

[61] Ribeiro, M.H., Jhaver, S., Martinell, J.C., Reignier-Tayar, M., West, R.: Deplatforming Norm-Violating Influencers on Social Media Reduces Overall Online Attention Toward Them (2024). https://arxiv.org/abs/2401.01253

[62] Saveski, M., Gillani, N., Yuan, A., Vijayaraghavan, P., Roy, D.: Perspective-taking to Reduce Affective Polarization on Social Media (2021). https://arxiv.org/abs/2110.05596

[63] Piccardi, T., Saveski, M., Jia, C., Hancock, J.T., Tsai, J.L., Bernstein, M.: Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity (2024). https://arxiv.org/abs/2411.14652

[64] 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)

[65] 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)

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

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

[68] 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)

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

[70] 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

[71] 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)

[72] 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)

[73] 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)

[74] 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)

[75] 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)

[76] Hua, W., Fan, L., Li, L., Mei, K., Ji, J., Ge, Y., Hemphill, L., Zhang, Y.: War and peace (waragent): Large language model-based multi-agent simulation of world wars. arXiv preprint arXiv:2311.17227 (2023)

[77] Horton, J.J.: Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research (2023)

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

[79] Williams, R., Hosseinichimeh, N., Majumdar, A., Ghaffarzadegan, N.: Epidemic modeling with generative agents. arXiv preprint arXiv:2307.04986 (2023)

[80] Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., Bernstein, M.S.: Social simulacra: Creating populated prototypes for social computing systems. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, pp. 1–18 (2022)

[81] 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

[82] Kovač, G., Portelas, R., Dominey, P.F., Oudeyer, P.-Y.: The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. arXiv preprint arXiv:2307.07871 (2023)

[83] (FAIR)†, M.F.A.R.D.T., Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., *et al.*: Human-level play in the game of diplomacy by combining language models with strategic reasoning. Science **378**(6624), 1067–1074 (2022)

[84] 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)

[85] Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., Sun, M.: Communicative agents for software development. arXiv preprint arXiv:2307.07924 **6**, 3 (2023)

[86] 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

[87] 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)

[88] 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)

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

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

[91] Gallup News: GOP Holds Edge in Party Affiliation for Third Straight Year. Accessed: 2025-05-01. https://news.gallup.com/poll/655157/gop-holds-edge-party-affiliation-third-straight-year.aspx

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

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

[94] Krackhardt, D., Stern, R.N.: Informal networks and organizational crises: An experimental simulation. Social psychology quarterly, 123–140 (1988)

[95] Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **76**(3), 036106 (2007)

[96] Sen, A.: On Economic Inequality. Oxford university press, ??? (1997)

[97] Perspective API: About the API: Attributes and Languages. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US. Accessed: 2025-05-01 (2025)

[98] Kates, S., Ladd, J., Tucker, J.: How americans' confidence in technology firms has dropped: evidence from the second wave of the american institutional confidence poll. Brookings Institute (2023)

[99] Lazer, D.M., Pentland, A., Watts, D.J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., *et al.*: Computational social science: Obstacles and opportunities. Science **369**(6507), 1060–1062 (2020)

[100] Ormerod, P., Rosewell, B.: Validation and verification of agent-based models in the social sciences. In: International Workshop on Epistemological Aspects of Computer Simulation in the Social Sciences, pp. 130–140 (2006). Springer

[101] Schelling, T.C.: Dynamic models of segregation. Journal of mathematical sociology **1**(2), 143–186 (1971)

[102] 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)

[103] 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)