Bots influence opinion dynamics without direct human-bot interaction: the mediating role of recommender systems

Niccolo Pescetelli (niccolo.pescetelli@njit.edu)
New Jersey Institute of Technology
Daniel Barkoczi
Max Planck Institute for Human Development
Manuel Cebrian
Max Planck Institute for Human Development

Research Article

Keywords: bots, opinion dynamics, Bayesian belief update, recommender systems, social influence

Posted Date: March 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1401919/v1

Bots influence opinion dynamics without direct human-bot interaction: the mediating role of recommender systems

Pescetelli, N.[1]*, Barkoczi, D.[2], Cebrian, M.[2]

[1]New Jersey Institute of Technology
323 Dr. Martin Luther King Jr Blvd, Newark, NJ 07102
[2]Max Planck Institute for Human Development
94 Lentzeallee, 14195, Berlin, Germany
*Corresponding author

Abstract

Bots' ability to influence public discourse is difficult to estimate. Recent studies found that hyperpartisan bots are unlikely to influence public opinion because bots often interact with already highly polarized users. However, previous studies focused on direct human-bot interactions (e.g., retweets, at-mentions, and likes). The present study suggests that political bots, zealots, and trolls may affect people's views indirectly via the mediating role of a platform's content recommendation system, thus influencing opinions even in the absence of direct human-bot interaction. Using an agent-based opinion dynamics simulation, we isolated the effect of a single bot – representing 1% of nodes in a network – on the opinion formed by rational Bayesian agents after removing direct human-bot connections. We compare this experimental condition with an identical baseline condition where such a bot is absent. We used the same random seed in both simulations so that conditions remained identical except for the presence of the bot. Results show that, even in the absence of direct interactions, the mere presence of the bot is sufficient to shift the average population opinion. Virtually all nodes – not only nodes directly interacting with the bot – shifted towards more extreme opinions. Overall, these findings offer a proof of concept that bots and hyperpartisan accounts can influence population opinions not only by directly interacting with humans but also by secondary effects, such as shifting platforms' recommendation engines' internal representations. The mediating role of recommender systems creates indirect causal pathways of algorithmic opinion manipulation.

Keywords: bots, opinion dynamics, Bayesian belief update, recommender systems, social influence

Introduction

Bots are becoming pervasive in our social media. From Twitter to Reddit, bots can interact with humans without detection, influencing opinions, and creating artificial narratives ("Shelley: Human-AI Collaborated Horror Stories" n.d.; Hurtado, Ray, and Marculescu 2019).

This study uses an agent-based simulation to explore the interaction between bots and content recommendation algorithms. Recommender systems such as collaborative filtering can provide hyper-personalized content recommendations. However, they partly rely on average population characteristics and shared features between nodes to produce their recommendations. We test the hypothesis that recommender systems mediating information access can also mediate bot influence. We hypothesize that bots can affect a population's mean opinion not just by direct interactions with other nodes (i.e., direct interactions) but via skewing the training sample fed to the recommender system during training (i.e., indirect interactions). Thus, a bot may influence content recommendation at the population level by subtly affecting how a centralized recommender system represents a population's preferences and patterns of content engagement. This indirect social influence may be more pervasive than direct influence because it occurs even without direct bot-human interaction.

The potential of algorithmic agents, commonly referred to as bots, to influence public opinion has been recently put under closer scrutiny. Special attention has been given to social and political bots that operate under human disguise on social media. Early studies documented the potential effects of bots on skewing opinion distributions on social media users and voters (Bessi and Ferrara 2016). Bots can inflate the perception of the popularity of particular views (Lerman, Yan, and Wu 2016), polarise opinions around divisive issues (Broniatowski et al. 2018; L. G. Stewart, Arif, and Starbird 2018), contribute to the spread of misinformation, conspiratory theories or hyper-partisan content (Paul and Matthews 2016; Shao et al. 2018), and promote harmful or inflammatory content (Stella, Ferrara, and De Domenico 2018). These generalized concerns have mobilized platforms to improve algorithmic agents' automatic detection and removal (Howard 2018; Ferrara et al. 2016; Ledford 2020). On top of bot influence, influence networks online are characterized by several other phenomena acting together on public opinions, such as human trolls, fake accounts, pink-slime newspapers, and "fake news" (Hurtado, Ray, and Marculescu 2019; Linvill and Warren 2018; Aral and Eckles 2019; Tucker et al. 2018). Researchers have started to untangle this complex web of interactions. The content spread by this class of agents spreads faster due to its emotional or sensationalist features (Vosoughi, Roy, and Aral 2018; D. M. J. Lazer et al. 2018). Partisan content tends to remain confined in insulated clusters of users, thus reducing the opportunity to encounter cross-cutting content (Bakshy, Messing, and Adamic 2015). Although algorithmic agents represent only a small part of general media manipulation tactics (Kakutani 2019; Sunstein 2018), they pose a problem for online platforms. Their ease of implementation, low cost, and scalability hurt the overall media environment. In this paper, we estimate the lower bound of algorithmic influence by focusing on the effect of a single algorithmic agent on a population. Our

findings can be generalized to other 'pre-programmed agents' of media manipulation, such as partisan accounts and human trolls. Pre-programmed agents share several features, such as sharing pre-set opinions and pushing political agendas while being scarcely influenced by others' beliefs.

The effect of bots and troll factories on public opinion is hard to estimate. Recently, several researchers have attempted to measure the effect of hyper-partisan content by looking at social media data from the 2016 USA presidential election (Guess, Nagler, and Tucker 2019; Allen et al. 2020). These studies suggest that sharing and consuming fake or hyper-partisan content was relatively rare relative to the total volume of content consumed. One study in particular (Bail et al. 2020) attempted to measure the effect of exposure to Russia's Internet Research Agency (IRA) content on people's opinions. The authors found that interactions with highly partisan accounts were most common among respondents with already strong ideological alignment with those opinions. The researchers interpreted these findings as suggesting that hyper-partisan accounts might fail to change beliefs because they primarily interact with already highly polarised individuals. This phenomenon, also named "minimal effect", is not specific to social media platforms but can also be found with offline political advertisement and canvassing practices (Zaller 1992; Endres and Panagopoulos 2019; Kalla and Broockman 2018). Thus, changing political attitudes tend to be less effective than one imagines.
Furthermore, exposure to partisan ideology can counter-intuitively strengthen confidence in one's own belief via non-linear interactions between people's beliefs (Bail et al. 2018; Niccolò Pescetelli and Yeung, 2020b). In agreement with this conclusion, a recent study found that human accounts are significantly more visible during political events than unverified accounts (González-Bailón and De Domenico 2021). This finding casts doubt on the centrality and impact of bot activity on political mobilizations' coverage (Ferreira et al. 2021). Overall, these findings show that, notwithstanding the well-documented spread of bots and troll factories on social media, their effect on influencing opinions may be limited.

The studies reviewed above were primarily concerned with direct influence among agents, namely direct interactions between algorithmic and human accounts (e.g., likes, retweets, and comments). Although common in many offline and online settings, we argue that direct influence does not take into account the complexity of the digital influence landscape. Direct social influence has long been studied in the social learning and literature outside the domain of social media platforms, e.g., opinion change in social psychology (Yaniv 2004; Bonaccio and Dalal 2006; Sherif et al. 1965; Festinger and Carlsmith 1959; Rader et al. 2017) and in opinion dynamics in sociology (Flache et al. 2017; Deffuant et al. 2000; DeGroot 1974; Friedkin and Johnsen 1990). Direct influence assumes exposure to another person's belief (e.g., an advisor) changes a privately held belief. However, this simple social influence model may be outdated in the modern digital environment. Although direct interactions on most online platforms do occur (e.g., friends exchanging messages and users tweeting their views), information exchange is also mediated by algorithmic procedures that sort, rank, and disseminate or throttle information. The algorithmic

ranking of data can affect exposure to specific views (Bakshy, Messing, and Adamic 2015). Recommender systems can learn population averages and trends, forming accurate representations of individual preferences from collective news consumption patterns (Das et al. 2007; Pipergias Analytis et al. 2020). One crucial difference between traditional social interactions and machine-mediated interactions is that in the latter case, single users can influence not only other people's beliefs but the "belief" of the content curation algorithm (i.e., its internal model). We call this *indirect influence*. Furthermore, indirect influence can occur via intermediary nodes as nodes are connected. In other words, a bot may directly influence one human but indirectly influence all the humans that this human is connected.

This paper suggests that previous research may have underestimated indirect influence. Here, we are especially interested in the influence of social bots on network opinion dynamics when platform-wide algorithmic content recommendation mediates information sharing. We investigate a previously unexplored indirect causal pathway connecting social bots and individuals via a simple recommendation algorithm (Figure 1A). We test the hypothesis that algorithmic agents, like bots and troll factories, can disproportionately influence the entire population by biasing the training sample of recommender algorithms predicting user engagement and user opinions (Figure 1B). This disproportionate influence is facilitated by their resistance to persuasion and greater content engagement and content sharing activity (Scott Hunter and Zaman 2018; Yildiz et al. 2013). Affecting recommender systems' internal representations is a more effective influence strategy that affects other accounts in parallel rather than serially.

We created two identical fully connected networks of 100 agents to test our hypothesis. The two networks differed only for whether the bot was present or absent. We initialized the two simulations using the same random seed, which allowed us to directly test the counterfactual of introducing a single bot in the network while holding all other conditions constant. Crucially, while human agents were all connected, the bot could only interact with the other users via the recommendation algorithm (Figure 1a). Our simulation differs from previous work on opinion dynamics in two important ways. First, contrary to previous studies (Friedkin and Johnsen 1990; DeGroot 1974), we distinguish between internally held beliefs and externally observable behavior. We assume that observable behavior represents a noisy reading of true internal beliefs. This assumption captures the fact that people on several online platforms, such as fora and social media, can form beliefs and change opinions simply by consuming content and never posting or sharing their own (Lazer 2020; Muller 2012). One does not need to tweet about climate change to form an opinion on climate change.
Similarly, the distinction between internally held and publicly displayed beliefs allows us to train the recommender algorithm only with externally observable behavior rather than making the unrealistic assumption that the algorithm has direct access to a user's unobservable opinion. We call 'engagement' all externally observable behaviors such as tweets, likes, and reactions. Thus,

both the recommender algorithm and agents must infer other agents' underlying opinions from engagement behaviors.

Second, while opinion dynamics models commonly use linear opinion aggregation of self and other people's opinions (Flache et al. 2017), we use a Bayesian opinion updating rule (Niccolò Pescetelli and Yeung 2020a, [b] 2020; Harris et al. 2016). The Bayesian update offers a natural way to consider all aspects of beliefs, including opinion direction, belief conviction, and resistance to changes of mind or new information. This belief update rule produces non-linear dynamics that have been shown to reflect belief updates in laboratory experiments (Niccolò Pescetelli and Yeung 2020b; Niccolo Pescetelli, Rees, and Bahrami 2016). Such non-linear dynamics reflect that people who agree tend to reinforce each other's beliefs and move to more extreme positions. In comparison, people who disagree tend to converge to more uncertain positions (see (Bail et al. 2018) for an exception). We compare our results under a Bayesian rule with more traditional belief updating models.

Across a series of simulations, we quantify the effect of adding a single bot to a network of fully connected agents. We show that the bot can influence human agents even though no direct link exists between human agents and the bot. We conclude that in an information system where trained models control who sees what, bots and hyper-partisan agents can influence the whole user population by influencing the internal representation learned by the recommender algorithm. In other words, what the recommender belief might be is as crucial as people's beliefs in determining the outcome of network opinion dynamics. We discuss these findings in light of the contemporary debate on social media regulation.

Methods

Overview

We simulate a simplified social network model where a recommender system learns and presents a personalized content feed to agents in the network. This feed contains the expressed opinions of other agents in the network. Each agent can observe and interact with other agents' opinions by updating and expressing their own opinions. We manipulate whether a single bot is also part of the potential pool of agents that the recommender system draws upon to create the feeds in two separate but identical conditions. We study whether this bot can infiltrate the feed created by the recommender system by influencing the statistical relationships it learns.

Simulation procedure

We simulate N=100 agents connected through a fully connected network.

Agents. Each agent is represented by a private *opinion* in the range $]0,1[$ drawn from a truncated Normal distribution.

(1) $T_i = tN(0.4, 1)$

and by an *expressed opinion* representing a noisy observation of their true opinions:

(2) $E_i = T_i + N(0, 1)$

On each time step, agents go through a two-step process:

Engagement. First, they decide whether or not to *engage* with content in their feed (see below). Content is the expressed opinions of other agents ranked by the recommender system for each agent individually. Agents decide whether to engage with the content based on an engagement function defined as

(3) $$P(engage_j) = |T_i^{t-1} - O_j^{t-1}|$$

Where *O* stands for Observed and is the expressed opinion of another agent. We represent engagement as a binary decision. This engagement function makes it more likely that agents engage with content that is more distant from their own opinions, representing the tendency that people have online to engage with shocking or count-intuitive content more than moderate content (Vosoughi, Roy, and Aral 2018; D. M. J. Lazer et al. 2018). We explore in Supplementary Material two different engagement functions: in the former, agents are more likely to engage with content close to their own opinion (*homophilous engagement*, Figure S1). In the latter, they are equally likely to engage with similar or dissimilar content (*bimodal engagement*, Figure S2).

Opinion update. If agents decide to *engage,* they update their own opinions using a Bayesian opinion update function:

(4) $T_i^t = \frac{T_i^{t-1} O_j^{t-1}}{(T_i^{t-1} O_j^{t-1}) + (1 - T_i^{t-1})(1 - O_j^{t-1})}$

If they decide not to engage, they keep their opinion from the previous timestep, time *t-1*.

Feed. Each agent is presented with a feed consisting of the *expressed opinions* of *n* other agents in the social network. This feed is created by a simple recommender system separately for each agent. The goal of the feed is to provide content that agents are likely to engage with (see *Engagement* above). To achieve this, we train a simple logistic regression using agents' binary

engagement history as a dependent variable and the absolute difference between the agent's public opinion at time t-1 and the opinion they observed in their feed as the independent variable. In other words, the model aims to learn the agents' engagement function by observing their prior engagement history and the content they observed in their feeds. To provide sufficient training data for the recommender system, we start the first ten timesteps of the simulation by presenting agents randomly in the feeds.

Bot. The bot is represented as an agent that does not change its opinion but sticks to the same opinion throughout the simulation (A. J. Stewart et al. 2019; Karan, Salimi, and Chakraborty 2018; Yildiz et al., n.d.). In different conditions, we manipulate the degree to which this opinion is extreme (i.e., the distance from the mean opinion of the agents).

We initialize the simulation with the following parameters:

Mean agent opinion: N(0.4, 1) and bot opinion = 0.8. This setting represents a situation where agents hold a moderate opinion and are not polarized. In probabilistic terms, the average population opinion is uncertain (i.e., close to 0.5). The bot's opinion disagrees with the average opinion and is more confident, but the mean difference between agents and the bot is not very large. On each timestep (starting from t=10 onwards), agents are presented with a unique feed based on which they decide whether to engage and update their opinions. Once each agent has made a decision, the simulation proceeds to the next timestep. We repeat the procedure for t=100 timesteps and r=100 replications. We record each agent's opinion on each timestep and the cases where the bot gets recommended to an agent. We simulate two conditions, one where the bot is present and one where it is absent. We initialize both simulation conditions with the same random seeds, thereby producing virtually identical simulation conditions except for the presence of the bot. This manipulation allows for precise measurements regarding the influence of the bot on the network.

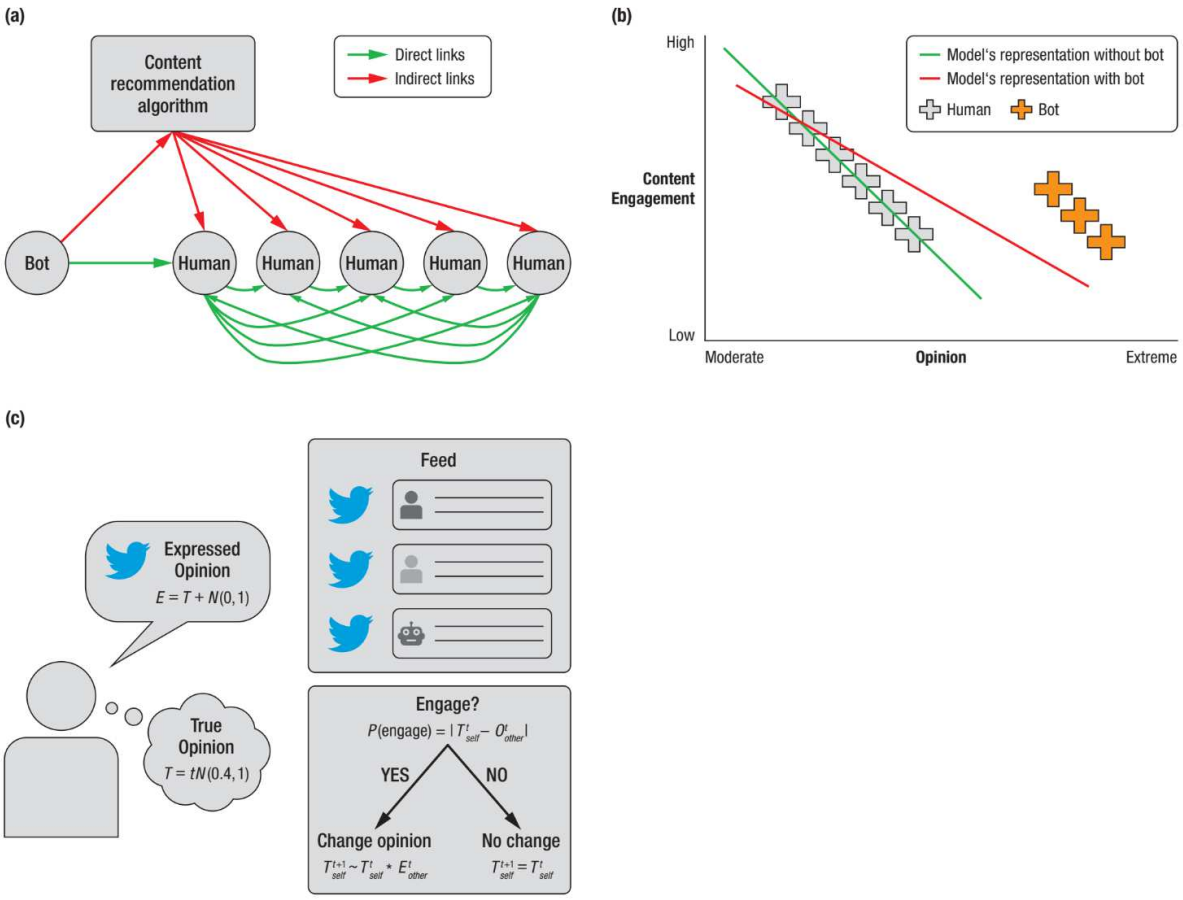

Figure 1. The indirect influence of bots on social information networks. (a) Representation of opinion dynamics network mediated by a content recommender system (grey box). Bot and Human agents (circles) consume and share content on a platform. A bot agent can influence human opinions via direct interaction with human agents (e.g., retweets, at-mentions, likes, and comments) or indirectly via affecting the internal representation of the content recommendation algorithm. **(b)** Schematic representation of the effect of bot presence on the internal representation learned by a simple recommender system trained to predict a user's engagement with content. The inclusion of the bot behavior in the training set skews the model to think that engagement with extreme content is more likely than it would be without the bot presence. **(c)** Agents in the simulation were modeled to include a true private opinion and an expressed public opinion. Agents were presented with their neighbors' public opinion on every round and decided whether to engage with this content or not, according to a pre-defined engagement function. Opinion change took place only when the agent engaged with the content.

Results

Population-level influence of the bot on the average opinion

We start by looking at the population-level influence of the bot on agents' opinions. We define influence as situations where an agent is presented with bot content in its feed, decides to engage based on the content observed, and thus changes its initial opinion. Figure 2a shows the mean opinion in the entire group over time for the two conditions (bot vs. no bot). Note that for the first t=10 timesteps, there is no change in opinion since those trials serve as training samples for the recommender system and, therefore, present agents in the feed randomly. From t=10 onwards, we see a significant difference between the two conditions, with the bot shifting the average opinion of the population by 5% on average. This effect is also reflected by the average engagement levels in the population, as depicted by Figure 2b. This effect holds across different initial opinion distributions and different bot opinions (Figure 5). From t=10 onwards, we observe a significant jump in engagements, showing that the recommender system is increasingly efficient at recommending content that agents will engage. The presence of a bot leads to remarkably higher engagement levels, indicating that by getting recommended, agents are more likely to engage and shift their opinions as a result of interacting with the bot directly or indirectly.

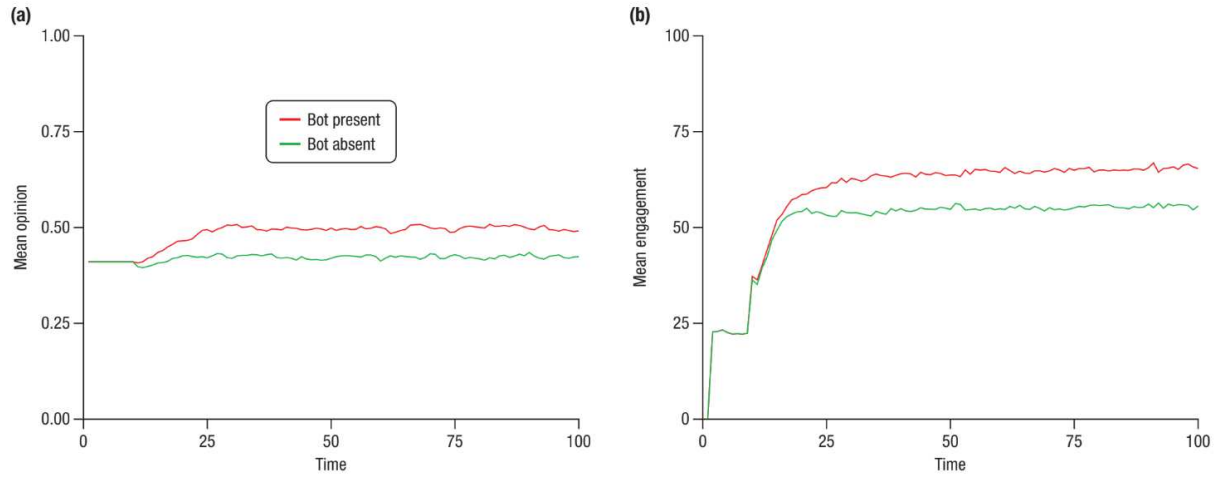

Figure 2. Mean opinion and mean engagement in networks with and without bot influence. (a) Mean opinion of the agents over time. **(b)** Mean number of agents engaging in each timestep. Red: Condition where the bot is part of the social network, Green: Condition where the bot is not part of the social network. A single bot can produce substantial changes in the mean opinion and mean engagement levels in the network.

The magnitude of direct bot influence on the individual agents

So far, we have seen that a single bot can shift the population's average opinion and engagement levels. Here, we investigate the reasons underlying this effect more directly. Figure 3a shows the number of agents directly influenced by the bot on each timestep. By direct influence, we mean that the bot's content was recommended to an agent via the feed, and the agent decided to engage with the bot's content (and thus updates its private opinion based on the bot's content). On average, 2.5 agents engage with and change their opinions after observing the bot on any timestep, with an average opinion change of 30% (Figure 3b). The spike observed in both graphs on the left-hand side is because the bot's opinions are less extreme – and thus less engaging according to Equation 3 – the more the population average opinion shifts towards the bot's opinion. The finding of low engagement and opinion shift replicates "minimal effect" findings online (Bail et al. 2020) and offline (Zaller 1992; Endres and Panagopoulos 2019; Kalla and Broockman 2018). It suggests that direct influence (e.g., direct bot interaction or political advertisement and canvassing practices) is often ineffective at shifting population averages. Our finding only captures the direct influence from bot to agent but does not measure the bot's indirect influence by influencing an agent that will influence further agents. Our intuition is that indirect influence may be more pervasive and more pronounced, especially in online contexts where recommender systems facilitate information spread. To measure this indirect n-th order influence of bots on agents, in the next paragraph, we compare the two simulation conditions (bot vs. no bot) while using the same random seed and holding all other conditions constant.

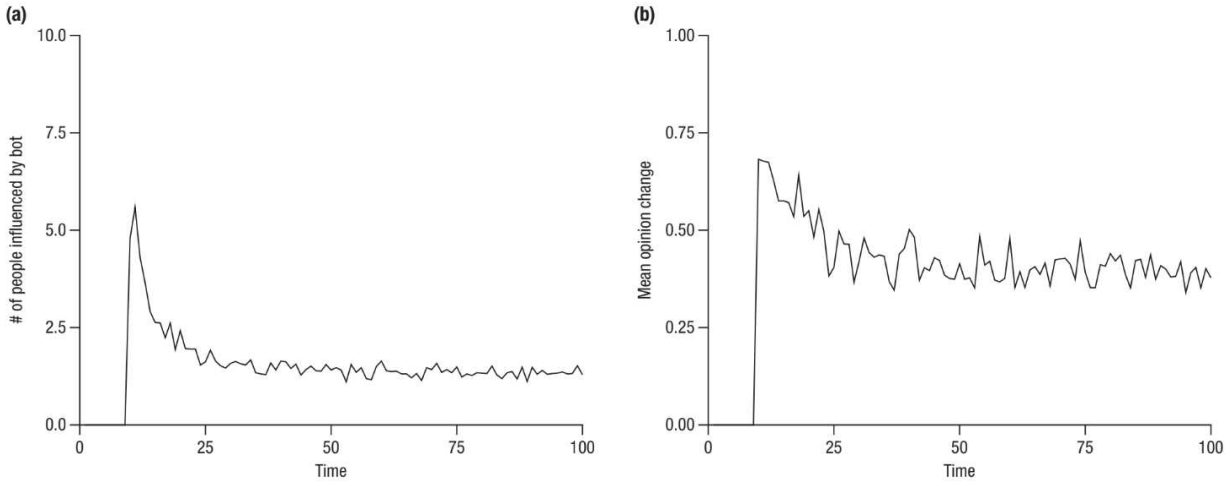

Figure 3. Direct bot influence. (a) The average number of nodes influenced by the bot on each timestep. Influence is defined as when an agent is presented with content produced by the bot, engages with this content, and shifts its own opinion. **(b)** Mean opinion change for agents influenced by the bot on each timestep. A single bot can influence multiple people on each timestep and produce substantial opinion change.

The individual-level shift in opinion as a result of direct or indirect bot influence

Figure 4 shows the difference in opinion between the same agent across the two simulation conditions, holding all other aspects of the simulation constant. Initializing the two simulations with identical parameters and random seed allowed us to isolate the effect of the bot. Estimating the within-agents effect improves our estimation of the bot effect. Differences between the two counterfactual worlds reflect direct bot influence and all secondary effects caused by introducing the bot. Notwithstanding the little direct influence (Figure 3), we found that, compared to a counterfactual simulation, the bot had an indirect effect on the entire population, with the magnitude of influence on opinion varying considerably, from 33 to 48 percentage points (Figure 4a). This effect is explained by agents observing other agents that might have interacted with the bot, leading to a trickle-down effect of the bot's opinion on other agents who might not have interacted with the bot at all. Figure 4b shows the signed difference between agents' opinions in the control and bot conditions ($d = T_{nobot} - T_{bot}$). Notice that most points are negative, indicating that nodes' opinions shifted toward the bot's opinion. Our model shows that bots' influence is magnified when we account for indirect influence via the recommender system or other intermediary agents. This striking result indicates that a single bot can have a much stronger and lasting effect beyond individuals it directly interacts with. This finding seems to suggest that studies focusing only on direct influence (bots' influence on people they directly interacted with) might have underestimated the actual capacity of a bot to bias population opinion dynamics.

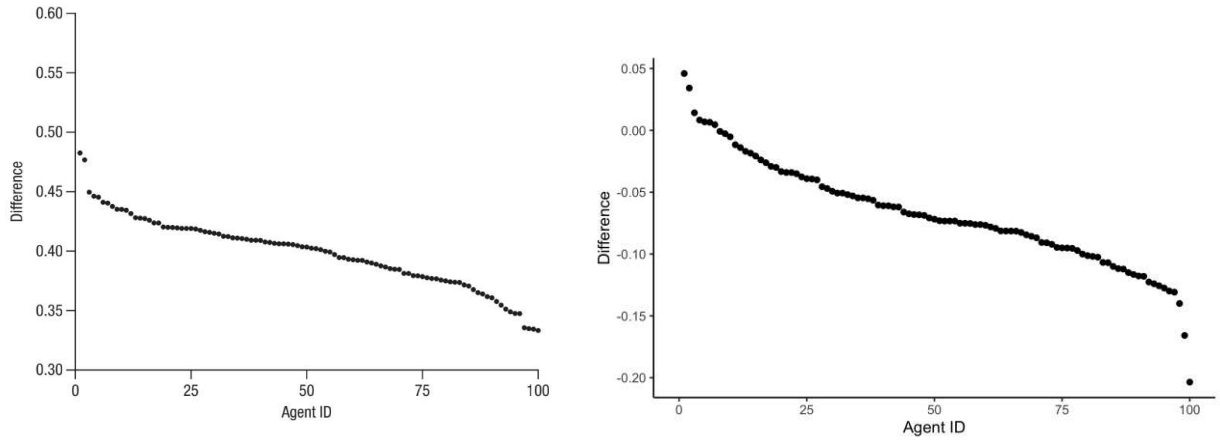

Figure 4. Within-agents bot effect across the two simulations. (a) The absolute difference between each agent's opinion at time *t=100* between the two simulation conditions (bot vs. no bot). This analysis measures the bot's total impact on the opinions of the same agents in the network. **(b)** The signed difference between each agent's opinion at time *t=100* between the two simulation conditions (bot vs. no bot). This analysis shows the direction of the social influence of the bot on individuals' opinions.

An exploration of the parameter space for bot opinion and population average

Finally, the above results assumed that the average opinion in the population is N(0.4,1) and the bot opinion is 0.8. The results are specific to this parametrization of our model. To test the generalisability of our conclusion, we explore the sensitivity of our results to different values of agent and bot opinion. Figure 5 shows a heatmap where the x-axis shows different values of the bot opinion and the y-axis shows the mean opinion in the population. The results remain qualitatively similar to those presented in the main text, with the bot having a more substantial effect on the population when its opinion is more distant from the average opinion of the population. The results further support the conclusion that a bot (here representing 1% of the total population) can have a disproportionate effect on population-level dynamics when we take into account indirect influence.

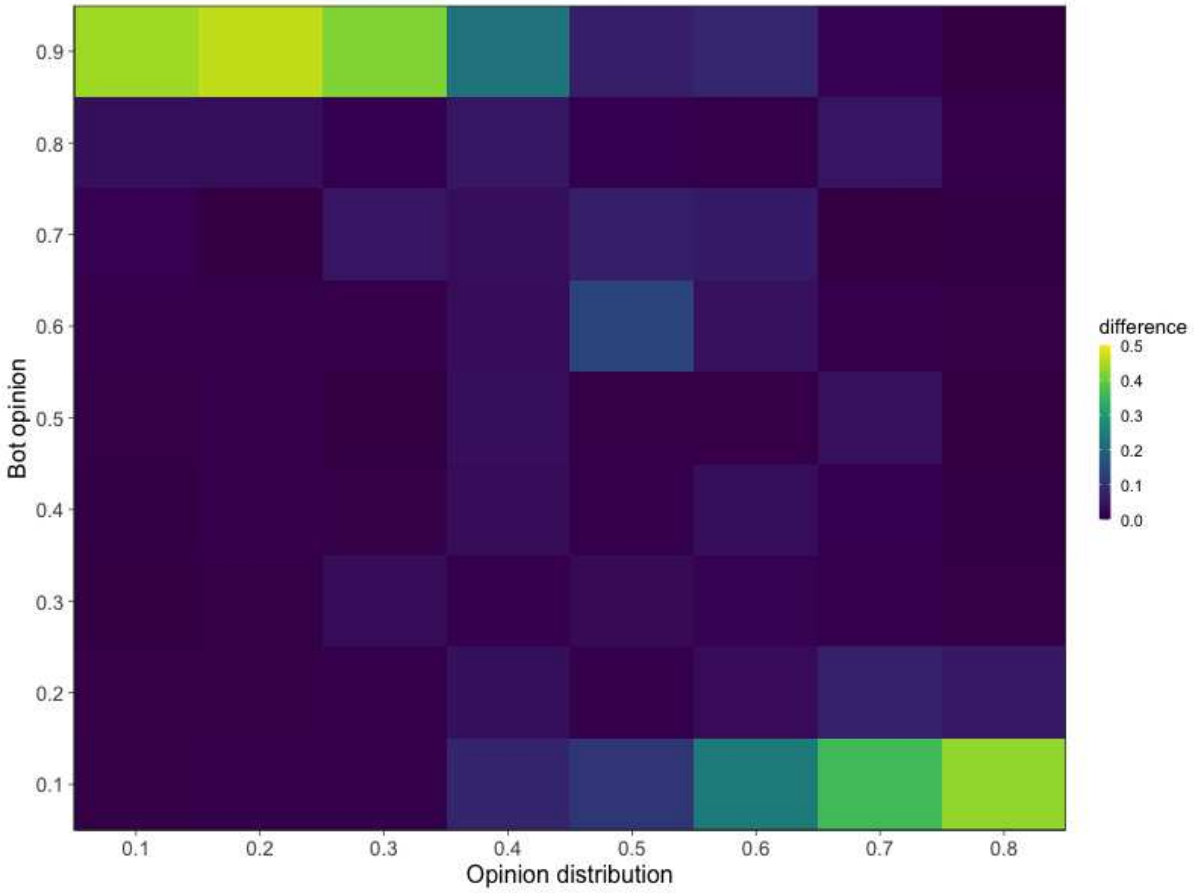

Figure 5 Heatmap of different initial opinion distributions. The principal analysis assumed that the average opinion in the population is N(0.4,1) and the bot opinion is 0.8. Here, we explore the sensitivity of our results to different values of agent and bot opinion. This figure shows a heatmap where the x-axis shows different bot opinion values and the y-axis shows the mean opinion in the population. The results remain qualitatively similar to those presented in the main text, with the bot having a more substantial effect on the population when its opinion is more distant from the average opinion of the population.

Similar results are found with linear aggregation functions.
We then tested whether the results were sensitive to the specific opinion aggregation function used to update agents' opinions based on observed social information. The function used in the original model (Equation 4) is a Bayesian belief update function that produces non-linear opinion dynamics (Pescetelli and Yeung 2020). However, the literature on opinion dynamics models more commonly uses linear aggregation functions where opinions are averaged, weighted, or summed together (Flache et al. 2017). Linear belief update models have difficulty accounting for opinion escalation dynamics that have been reported in lab experiments (Mahmoodi et al. 2013; Pescetelli and Yeung 2020; Bail et al. 2018; Guilbeault et al. 2018). We rerun the simulation using standard opinion dynamics models to understand whether the effects observed in the principal analysis were due to the particular Bayesian update model used. One hypothesis is that bot effects on population belief are observed because the Bayesian belief update rule creates non-linear escalation processes that magnify the influence of the bot. If this were true, the bot influence could be attributed to the specific opinion update function rather than the bot effect mediated by the recommendation system. Alternatively, the bot effect should be attributed to the indirect social influence if similar effects were observed across alternative opinion update functions. We tested our results under a Friedkin-Johnsen opinion update rule and a Deffuant model.

Friedkin-Johnsen model. The Friedkin-Johnsen belief model (F-J model) generalizes the classic DeGroot model (DeGroot 1974). In this model, Equation 4 is replaced by:

$$(5)\ T_i^t = (1-\alpha)T_i^{t-1} + \alpha O_j^{t-1}$$

Where α is a social learning weight representing how much other people's opinions are weighted compared to one's own opinion. Setting the α parameter to 0 means agents ignore other agents' opinions. Setting the α parameter to 1 means agents switch their opinion to the opinion held by other agents, thus ignoring one's private opinion altogether. We set the α parameter to 0.3 based on the average social influence observed in lab experiments (Minson et al. 2011; Liberman et al. 2012; Bonaccio and Dalal 2006).

We found that the average population opinion in the baseline (bot is absent) and bot condition diverges when the recommender system is first introduced (t=10). The average population opinion approximates over time the bot opinion. This convergence is because the bot does not change its opinion, thus pulling the population average.

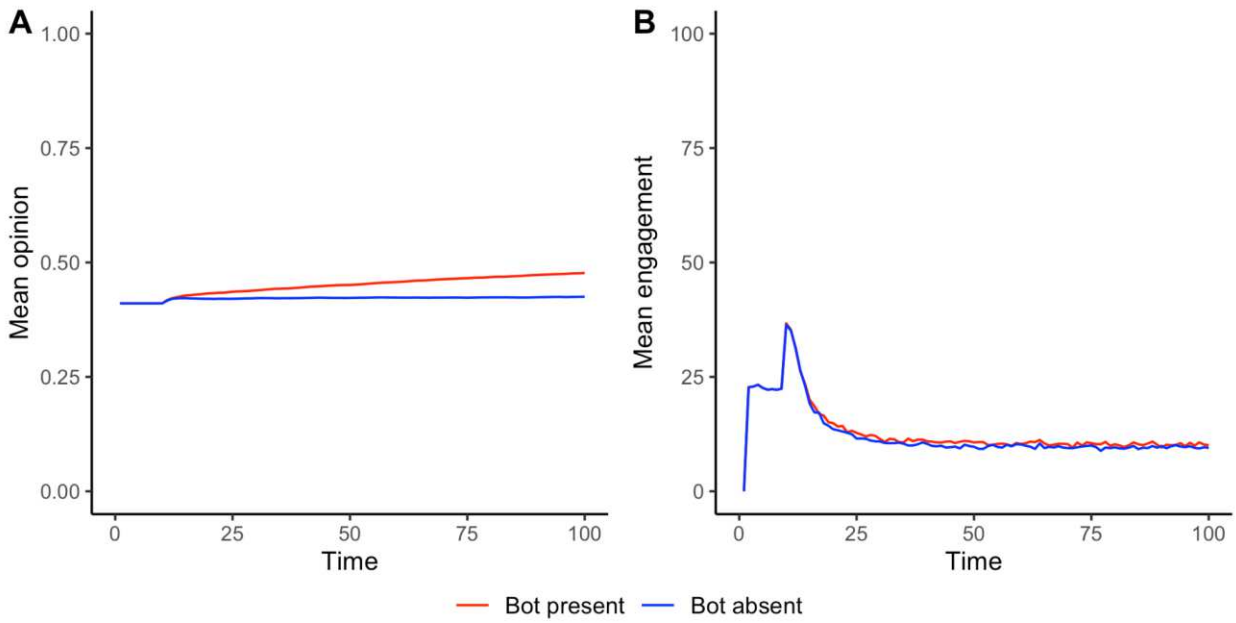

Figure 6. Average population effect using the F-J model. (a) The average population opinion dynamics in the control (blue) and bot condition (red). The two average populations start to deviate early on and diverge over time, with the average population opinion approximating the bot opinion. **(b)** The mean engagement of the population is found to spike soon after the recommender system is introduced (t=10) and then decreases as the average population's opinion tends towards the bot's opinion.

On the contrary, the average engagement of the population is found to spike soon after the recommender system is introduced (t=10) and then decreases as the average population's opinion tends towards the bot's opinion. This effect results from great engagement as the opinion of the bot and the average population's opinion are far apart at the beginning and converge over time. The effect reduces over time as this distance decreases, according to Equation 3.

Contrary to what was observed with the Bayesian model (Figure 3), the number of nodes directly affected by the bot is higher when using the F-J model (Figure 7). However, this number remains a minority of nodes compared to the rest of the population (about 5%). The average individual opinion change resulting from direct interaction with the bot is relatively small (about 1.5%). This effect is comparable to previously observed in the main analysis (Figure 3b).

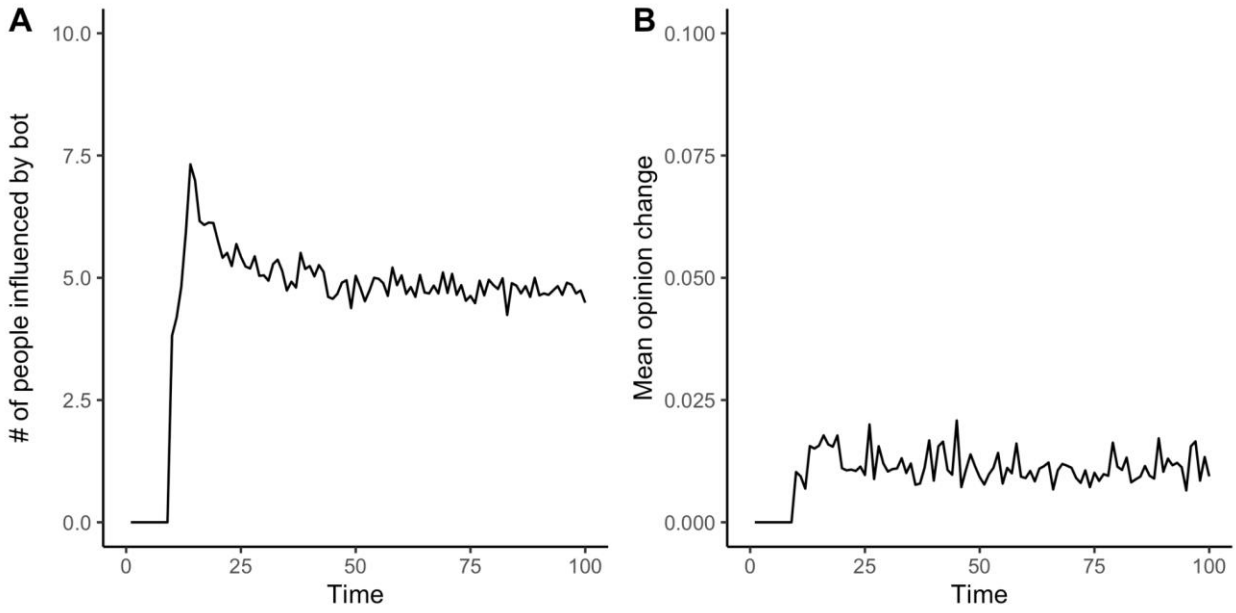

Figure 7. Direct bot influence. (a) The number of people who are influenced by the direct bot interaction. The number spikes after the recommender system is introduced (t=10) and reaches a plateau that is double what was previously observed (around five people per timestep). **(b)** The average opinion change of individual nodes due to direct interaction with the bot. The mean opinion change is relatively small (about 1.5%), reflecting what was previously observed in the original analysis.

Deffuant model. Next, we implemented a Deffuant model of opinion update. The Deffuant model accounts for tolerance thresholds in opinion updates. When the distance between a target agent's opinion and a social partner's opinion is below a given threshold, opinions converge. When the distance is above a given threshold, individuals ignore each other's views and do not converge (Deffuant et al. 2000). We decided to include the Deffuant model because opinion distance is an essential factor determining the engagement function in our simulation (Equation 3). According to this model, Equation 4 is replaced with

$$\begin{cases} T_i^t = T_i^{t-1} + \mu(O_j^{t-1} - T_i^{t-1}) & \text{for } |T_i - O_j| < d \\ T_i^t = T_i^{t-1} & otherwise \end{cases} \quad (6)$$

Where μ is a dampening parameter set to 0.5 and d is the tolerance level, here set to 0.2.

The results show no difference in the average population's opinion between the two conditions when a Deffuant model was used (Figure 8a). Similarly, no difference in average engagement was found between the two conditions (Figure 8b). These findings suggest that, at the population level, the Deffuant model did not produce significant differences in average opinion and average engagement.

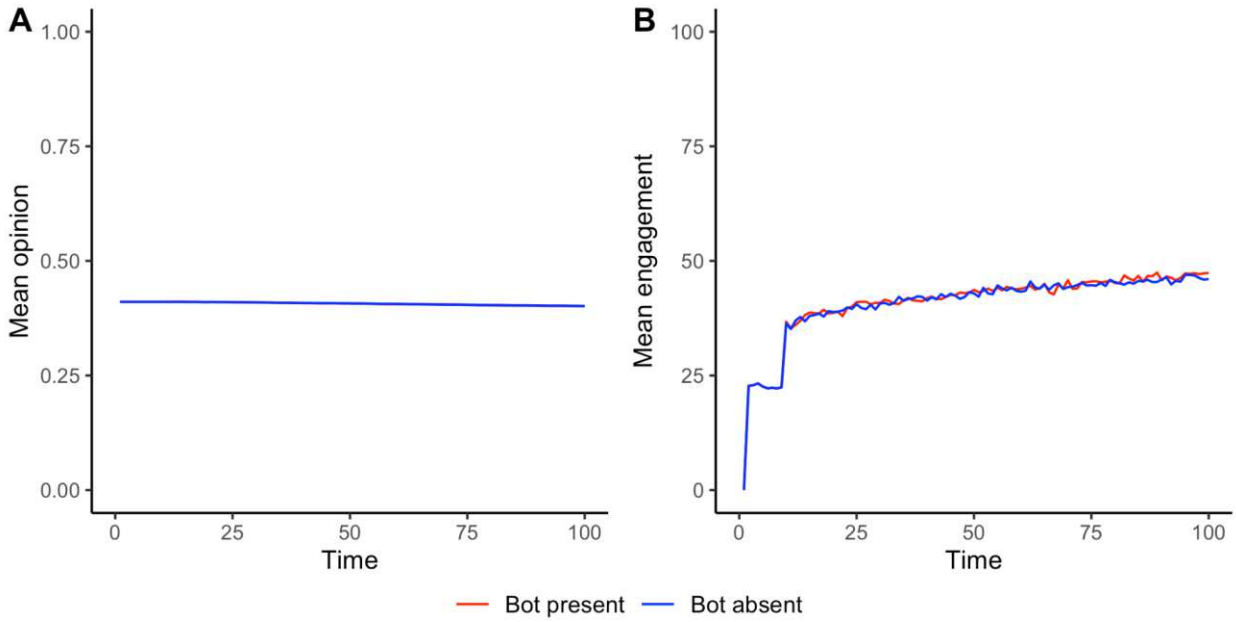

Figure 8. Average population effect using the Deffuant model. (a) No difference at the population level is found between conditions when using a Deffuant belief update model **(b)** Mean engagement is no different in the control condition (blue) and the bot condition (red).

Even though slight or no opinion differences were observed at the population level, differences were observed when focusing on direct influence. We looked at the direct social influence between the bot and other agents. We computed the number of nodes directly interacting with the bot on every timestep (Figure 9a). Like in the Bayesian and the Friedkin-Johnsen models, the number of nodes directly influenced by the bot represents a small minority (about 5%) of the total number of nodes in the network. Similarly, the mean opinion change due to direct influence is minimal (Figure 9b), representing less than 2% opinion shifts.

Both the Friedkin-Johnsen and Deffuant models show similar patterns to the Bayesian model. In all of them, we noticed that the number of nodes that had direct interactions with the bot was small compared to the total population, and little opinion change occurred due to direct interaction with the bot. The Bayesian and F-J models seemed to produce differences between control and bot conditions at the population level, both in the population's average opinion and average engagement. The Deffuant model did not.

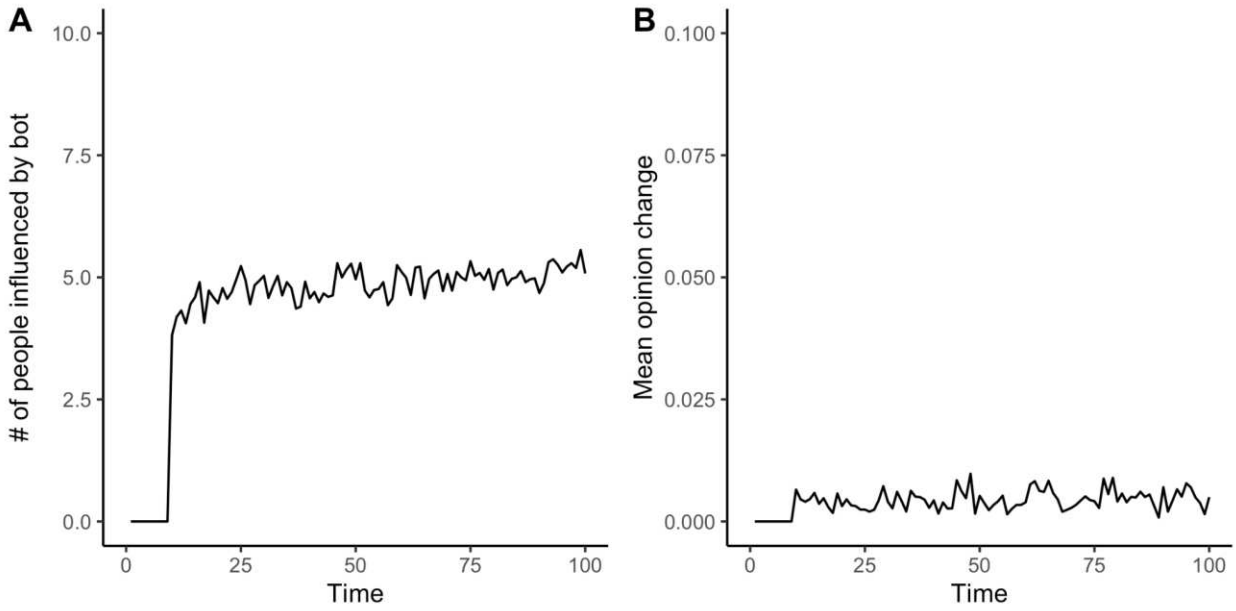

Figure 9. Direct bot influence. (a) The number of people influenced by the bot due to direct interaction. The number spikes after the recommender system is introduced (t=10) and reaches a plateau that is double what was previously observed (around five people per timestep). **(b)** The average opinion change of individual nodes due to direct interaction with the bot. The mean opinion change is relatively small (about 2%), reflecting what was previously observed in the original analysis.

Nevertheless, larger opinion shifts occurred when directly comparing a node's final opinion (t=100) between the control and the bot condition. When looking at the F-J and the Deffuant models, virtually all nodes shifted their opinion in the bot condition compared to the no-bot condition baseline (Figure 10). This pattern of results mirrors what was previously observed for the Bayesian model (Figure 4). We observed smaller changes in both the F-J and Deffuant models than the Bayesian model.

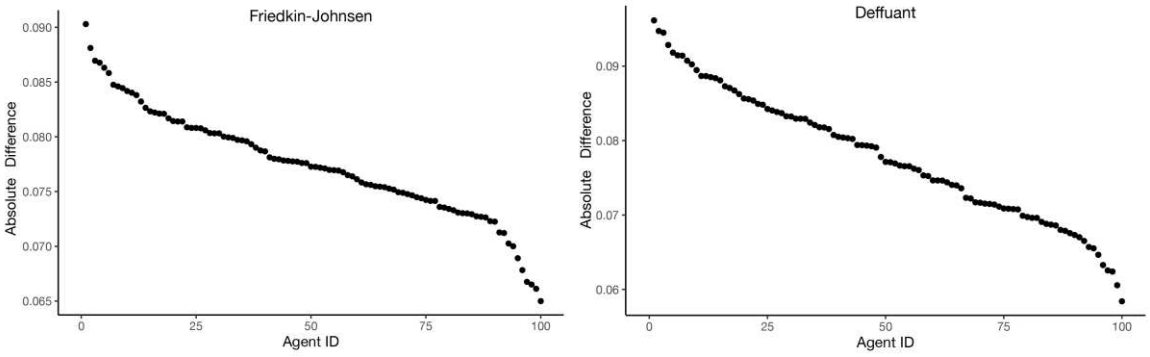

Figure 10. Absolute differences. (a) Friedkin-Johnsen model. **(b)** Deffuant model.

Effect on the recommendation system's internal representations.

We conducted a last set of analyses to detect differences in the model's internal representations. The recommender system used in this study was a simple logistic regression. The model was

trained using agents' binary engagement history as a dependent variable and the absolute difference between the agent's public opinion at time t-1 and the opinion they observed in their feed as the independent variable. The model was retrained on every timestep, thus providing a moment-by-moment representation of the recommender system's learning.

After model fitting, the logistic's slope beta coefficients were used to compare learning across conditions. We calculated the difference of the recommender system's slopes between the bot and no-bot conditions (bot - noBot). The first ten timesteps are only training timesteps where the recommendation system was inactive. Differences in beta coefficients appear in the 10-30 time range, returning to no difference after timestep 35. This finding shows that changes in the model's internal representations happen early on, but the effect of the bot wanes over time. This result nicely mirrors earlier findings shown in Figure 3.

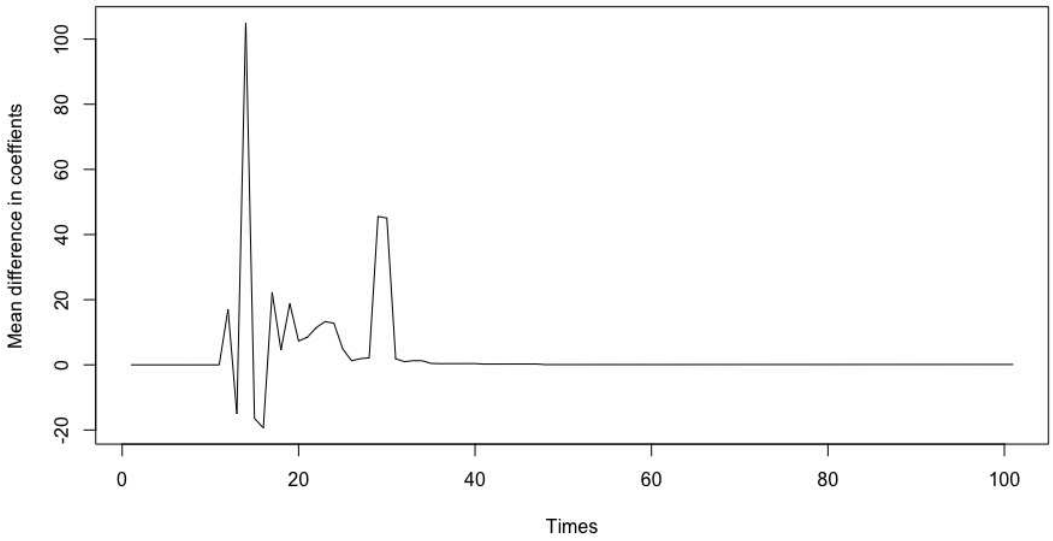

Figure 10. Differences in recommender system's learned coefficients. The recommender system model learned differentiated representations of people's engagement patterns in the two conditions. The recommender system was inactive during the first ten trials. After that, differences in learning occur of the signed beta coefficients between the bot and baseline conditions (bot - noBot). Over time, these differences disappear, suggesting that the effect of the bot wanes over time.

Discussion

This paper investigated the indirect influence that programmed media manipulators, such as bots, trolls, and zealots can have on population opinion dynamics via recommender systems. We posited that even in the absence of direct exposure to bots' content, bots could influence population-wide content ranking by providing unduly training evidence to recommender algorithms. For instance, bots' greater activity, content engagement and production, and their resilience to persuasion may

contribute to bots skewing the training sample that algorithms use to infer population preferences, averages, and typical content consumption patterns.

Using an opinions dynamics simulation on a 100-node network, we find that a single bot can substantially shift the mean opinion and engagement compared to a control condition without a bot. Even though only a minority of 'human' nodes (2.5%) directly engaged with the bot's content, the bot disproportionately affected the average shift in opinion observed in the population. Notably, virtually all nodes in the population were influenced by the bot presence, with opinion shifts ranging from 33 to 48 percentage points. The results are robust across different initialization parameters and different opinion update functions. Furthermore, we re-ran our simulation and removed the bot after the 40th timestep as a complementary analysis. We found that the effects of direct interactions drop as soon as the bot is removed (Figure S3). On the contrary, secondary effects remain until the end of the simulation, although their magnitude is significantly reduced (Figure S4).

These results would be unlikely if bots could influence human agents only via direct exposure. As bots represent only a minority of the population of agents (1% in our simulation), it is unlikely that they can interact with and directly influence all other agents. Our findings show that a simple recommender system (a logistic regression in our simulation) dramatically increases the influence of a bot on the population. Our first contribution is advancing the debate around bots' influence and media manipulation. Our study highlights a previously unexplored phenomenon and draws attention to a subtle yet potentially pervasive phenomenon. Contrary to previous studies investigating social media bots, our work does not model direct interactions between bots and human agents (arguably representing a minority of interactions) but focuses on indirect effects via recommendation systems. Our findings highlight that malicious agents, such as bots and trolls factories, can massively increase their influence by infiltrating the internal representations of trained models tasked with content filtering.
Our second contribution is that our setup allows us to compare counterfactual worlds, thus strengthening causal inference. We initialized both control (without-bot) and treatment (with-bot) simulations with the same parameters and random seed. Furthermore, effects on opinion shifts and engagement were calculated at the individual node level, thus measuring the effect of our treatment (bot presence) on the opinion dynamics and engagement of virtually identical 'human' agents.

Although our agent-based model provides valuable insights into machine-mediated information systems, it is limited by the ecological validity of simulation studies. Testing the same hypotheses in real-world contexts may be problematic due to difficulty in conducting randomized control trials on social media platforms. Although it may be difficult to manipulate these systems, researchers have recently successfully inferred the hidden mechanisms underlying several proprietary algorithms by systematically prompting them (Ali et al. 2019; Hannak et al. 2013; Robertson, Lazer, and Wilson 2018). Furthermore, real-world opinion dynamics are arguably more complex

than the simple simulated world. Complex dynamics may be elicited by bots operating on media platforms not captured by our simulation (Mønsted et al. 2017). Nevertheless, our findings show that one component of such a complex network of influence may occur not via direct interactions between nodes but via subtly skewing the training set of recommender systems.

We also acknowledge that our findings are specific to our choice of parameters and may not generalize well to other scenarios. Future studies should investigate the effects of network size and alternative network structures on bot influence. Our study used a simple logistic model to predict engagement scores to provide recommended content. One limitation is that existing recommender systems are more complex than the simple logistic regression employed in this study. For instance, recommender systems can consider many more features and provide greater personalization thanks to highly granular information about users and user similarity. However, the effects highlighted in our findings are likely to affect, at least to some degree, any content filtering algorithm trying to extrapolate the behavior of one user to another. We speculate that more complex recommendation systems may still be affected by the same dynamics highlighted here as long as they use population averages to predict individual preferences. By biasing the estimation of a population mean (Figure 10), algorithmic agents can change the model's expectation for a given cluster of users or the whole population. Extrapolating a user's behavior to another represents the standard in many recommender systems (Ricci, Rokach, and Shapira 2011), e.g., collaborative filtering algorithms (Das et al. 2007; Koren and Bell 2015; Ricci, Rokach, and Shapira 2011). Recently, researchers have shown that individual social influence can be affected by an individual's position in the population distribution and similarity with others (Pipergias Analytis et al. 2020; Analytis, Barkoczi, and Herzog 2018). Similar findings may thus be observed on more realistic content recommendation algorithms.
Furthermore, the complexity of realistic recommender systems makes the findings of this work even more significant. Indeed, our findings suggest that bots and troll factories' influence may be subtle but highly pervasive. The opacity and complexity of realistic recommender systems suggest that such pervasive effects may continue to operate undetected. The potential consequences are difficult to imagine but should prompt further investigation.

Another caveat in our simulation pertains to the modeling opinion and opinion change and operationalizing bots as stubborn agents (Scott Hunter and Zaman 2018; Yildiz et al. 2013). In the present study, we represent beliefs along a single opinion dimension. People's beliefs outside the lab are often more complex and multifaceted than our model. Nevertheless, using beliefs spanning a single dimension represents a necessary first step in many opinion dynamic models and advice-taking paradigms (Bonaccio and Dalal 2006; Deffuant et al. 2000; Flache et al. 2017; Friedkin and Johnsen 1990). Similarly, political polarisation and beliefs across several domains, especially divisive issues, may be well described by a single belief dimension (Navajas et al. 2019). Future modeling efforts could generalize our findings to multi-dimensional attitude spaces.

More importantly, some of our findings may depend on the specific opinion update model that we used here. Although several other opinion models exist (Flache et al. 2017), many assume that opinion change results from a linear combination of neighboring nodes' observed opinions, such as averages and weighted means (DeGroot 1974; Friedkin and Johnsen 1990; Deffuant et al. 2000). However, experimental evidence suggests that non-linear multiplicative dynamics often govern opinion change (Bail et al. 2018; Pescetelli and Yeung 2020b; Pescetelli, Rees, and Bahrami 2016; Moscovici and Zavalloni 1969). Here, we used a Bayesian opinion update model that captures dynamics of belief conviction, uncertainty, and probabilistic judgments (Pescetelli and Yeung 2020b, [a] 2020; Harris et al. 2016). We selected this opinion update model as it offers several important features relevant to understanding phenomena of polarization and hyper-partisanship. First, it can be seen as a normative rational model of opinion update. This feature allows us to quantify a best-case scenario, namely, the impact of bots if people were rational. Second, it implicitly represents confidence as the distance from the maximum uncertainty point (50%). Thus, more confident agents are less influenced by and more influential than uncertain agents. Confidence and confidence escalation are important elements to explain phenomena of polarisation commonly observed online. Third, encounters with agreeing agents tend to increase one's belief conviction, while encounters with disagreeing agents increase uncertainty. Although some recent evidence suggests that even disagreement may entrench people further in their decisions (Bail et al. 2018), Bayesian update better represents opinion escalation dynamics better than linear aggregation models. While linear updates may better model estimation tasks, Bayesian updates may better represent belief convictions and partisan affiliations, i.e., cases where interaction with like-minded individuals makes you more extreme. The introduction of such a model in the opinion dynamics literature represents a novel contribution.

We suggest possible ways to reduce the risk of public opinion manipulation. First, improving the detection and removal of automated accounts can reduce bots' impact on population-wide behaviors (Figure S3). However, uniquely relying on this strategy is not sustainable in the long run as automated detection becomes outdated and new and more sophisticated bots are developed. Detection and removal tend to be more effective with relatively simple bots, thus selecting bots with more human-like features, which are more likely to remain undetected. A more valuable strategy might be to regulate recommender and filtering algorithms to make them more transparent. Knowledge of the features used to make content recommendations can help academics and practitioners to monitor features that ill-willing entities can exploit. Open auditing of recommender systems and open-source software can go a long way in preventing some types of bots from doing harm and minimizing algorithmic tampering with public opinions.

Conclusions

In this paper, we explored the hypothesis that algorithmic agents may have undue influence on online social networks by biasing the internal representations of recommender systems. Bots' more extreme views, greater activity frequency, and content generation might distort content

recommendation for the entire network. Researchers and watchdogs should be aware of these indirect causal pathways of bot influence.

Declarations

ETHICS APPROVAL AND CONSENT TO PARTICIPATE
Not applicable

CONSENT FOR PUBLICATION
Not applicable

AVAILABILITY OF DATA AND MATERIAL
Barkoczi, D., & Pescetelli, N. (2021, August 21). Indirect causal influence of social bots through a simple recommendation algorithm. Retrieved from osf.io/7s83x

COMPETING INTERESTS
The authors declare that they have no competing interests.

FUNDING
The study was funded by the Max Planck Institue for Human Development.

AUTHORS' CONTRIBUTIONS
NP and MC conceptualized the study.
NP and DB designed the formal analysis
DB curated and visualized the data, and run the formal analyses
NP and MC acquired funding
All authors were responsible for investigation and methodology
NP curated the project administration
MC supervised the project
NP and DB wrote the original draft
All authors reviewed and edited the final draft

ACKNOWLEDGEMENTS
Not applicable.

References

Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes." *Proc. ACM Hum.-Comput. Interact.*, 199, 3 (CSCW): 1–30.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. "Evaluating the Fake News Problem at the Scale of the Information Ecosystem." *Science Advances* 6 (14): eaay3539.

Analytis, Pantelis P., Daniel Barkoczi, and Stefan M. Herzog. 2018. "Social Learning Strategies for Matters of Taste." *Nature Human Behaviour* 2 (6): 415–24.

Aral, Sinan, and Dean Eckles. 2019. "Protecting Elections from Social Media Manipulation." *Science* 365 (6456): 858–61.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115 (37): 9216–21.

Bail, Christopher A., Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. 2020. "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." *Proceedings of the National Academy of Sciences* 117 (1): 243–50.

Bakshy, E., S. Messing, and L. A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science*. https://science.sciencemag.org/content/348/6239/1130.abstract?casa_token=93SGKMyFHO4AAAAA:NLLn7cnwU-dniTFvSJ5wC7XUJ30w5AFKxPLDLfWyijbh8Z-NWk0vsYB2zgXtq7EyGRLUhHdYX2fBfQ.

Bessi, Alessandro, and Emilio Ferrara. 2016. "Social Bots Distort the 2016 US Presidential Election Online Discussion." *SSRN* 21 (11). https://ssrn.com/abstract=2982233.

Bonaccio, Silvia, and Reeshad S. Dalal. 2006. "Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences." *Organizational Behavior and Human Decision Processes* 101 (2): 127–51.

Broniatowski, David A., Amelia M. Jamison, Sihua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. 2018. "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate." *American Journal of Public Health* 108 (10): 1378–84.

Das, A., M. Datar, A. Garg, and S. Rajaram. 2007. "Google News Personalization: Scalable Online Collaborative Filtering." In *Proc. of the 16th Int.Conf. on World Wide Web*, 271–80.

Deffuant, Guillaume, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. "Mixing Beliefs among Interacting Agents." *Advances in Complex Systems. A Multidisciplinary Journal* 03 (01n04): 87–98.

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

Endres, Kyle, and Costas Panagopoulos. 2019. "Cross-Pressure and Voting Behavior: Evidence from Randomized Experiments." *The Journal of Politics* 81 (3): 1090–95.

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

Festinger, L., and J. M. Carlsmith. 1959. "Cognitive Consequences of Forced Compliance." *Journal of Abnormal Psychology* 58 (2): 203–10.

Flache, Andreas, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. 2017. "Models of Social Influence: Towards the Next Frontiers." *Journal of Artificial Societies and Social Simulation* 20 (4). https://doi.org/10.18564/jasss.3521.

Friedkin, Noah E., and Eugene C. Johnsen. 1990. "Social Influence and Opinions." *The Journal of*

Mathematical Sociology 15 (3-4): 193–206.
González-Bailón, Sandra, and Manlio De Domenico. 2021. "Bots Are Less Central than Verified Accounts during Contentious Political Events." *Proceedings of the National Academy of Sciences of the United States of America* 118 (11). https://doi.org/10.1073/pnas.2013443118.
Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5 (1): eaau4586.
Hannak, Aniko, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. "Measuring Personalization of Web Search." In *Proceedings of the 22nd International Conference on World Wide Web*, 527–38. WWW '13. New York, NY, USA: Association for Computing Machinery.
Harris, Adam J. L., Ulrike Hahn, Jens K. Madsen, and Anne S. Hsu. 2016. "The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach." *Cognitive Science* 40 (6): 1496–1533.
Howard, Philip. 2018. "How Political Campaigns Weaponize Social Media Bots." *IEEE Spectrum* Oct.
Hurtado, Sofia, Poushali Ray, and Radu Marculescu. 2019. "Bot Detection in Reddit Political Discussion." In *Proceedings of the Fourth International Workshop on Social Sensing*, 30–35. SocialSense'19. New York, NY, USA: Association for Computing Machinery.
Kakutani, Michiko. 2019. *The Death of Truth*. Tim Duggan Books.
Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *The American Political Science Review* 112 (1): 148–66.
Karan, Naneh, Farshad Salimi, and Subhadeep Chakraborty. 2018. "Effect of Zealots on the Opinion Dynamics of Rational Agents with Bounded Confidence." *Acta Physica Polonica, B* 49 (1): 73.
Koren, Yehuda, and Robert Bell. 2015. "Advances in Collaborative Filtering." In *Recommender Systems Handbook*, edited by Francesco Ricci, Lior Rokach, and Bracha Shapira, 77–118. Boston, MA: Springer US.
Lazer, David. 2020. "Studying Human Attention on the Internet." *Proceedings of the National Academy of Sciences of the United States of America*.
Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al. 2018. "The Science of Fake News." *Science* 359 (6380): 1094–96.
Ledford, Heidi. 2020. "Social Scientists Battle Bots to Glean Insights from Online Chatter." *Nature* 578 (7793): 17–17.
Lerman, Kristina, Xiaoran Yan, and Xin-Zeng Wu. 2016. "The 'Majority Illusion' in Social Networks." *PloS One* 11 (2): e0147617.
Linvill, Darren L., and Patrick L. Warren. 2018. "Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building." *Resource Centre on Media Freedom in Europe*.
Mønsted, Bjarke, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. 2017. "Evidence of Complex Contagion of Information in Social Media: An Experiment Using Twitter Bots." *PloS One* 12 (9): e0184148.
Moscovici, Serge, and Marisa Zavalloni. 1969. "The Group as a Polarizer of Attitudes." *Journal of Personality and Social Psychology* 12 (2): 125–35.
Navajas, Joaquin, Facundo Álvarez Heduan, Juan Manuel Garrido, Pablo A. Gonzalez, Gerry Garbulsky, Dan Ariely, and Mariano Sigman. 2019. "Reaching Consensus in Polarized Moral Debates." *Current Biology: CB* 29 (23): 4124–29.e6.
Paul, Christopher, and Miriam Matthews. 2016. "The Russian 'firehose of Falsehood' Propaganda Model." *Rand Corporation*, 2–7.
Pescetelli, Niccolo, Geraint Rees, and Bahador Bahrami. 2016. "The Perceptual and Social Components of Metacognition." *Journal of Experimental Psychology. General* 145 (8): 949–65.
Pescetelli, Niccolò, and Nicholas Yeung. 2020a. "The Role of Decision Confidence in Advice-Taking and Trust Formation." *Journal of Experimental Psychology. General*, October.

https://doi.org/10.1037/xge0000960.
Pescetelli, Niccolò, and Nick Yeung. 2020b. “The Effects of Recursive Communication Dynamics on Belief Updating.” *Proceedings of the Royal Society B: Biological Sciences* 287 (1931): 20200025.
Pipergias Analytis, Pantelis, Daniel Barkoczi, Philipp Lorenz-Spreen, and Stefan Herzog. 2020. “The Structure of Social Influence in Recommender Networks.” In *Proceedings of The Web Conference 2020*, 2655–61. WWW ’20. New York, NY, USA: Association for Computing Machinery.
Ricci, Francesco, Lior Rokach, and Bracha Shapira. 2011. “Introduction to Recommender Systems Handbook.” In *Recommender Systems Handbook*, edited by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, 1–35. Boston, MA: Springer US.
Robertson, Ronald E., David Lazer, and Christo Wilson. 2018. “Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages.” In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18*, 955–65. New York, New York, USA: ACM Press.
Scott Hunter, D., and Tauhid Zaman. 2018. “Optimizing Opinions with Stubborn Agents Under Time-Varying Dynamics.” *arXiv [cs.SI]*. arXiv. http://arxiv.org/abs/1806.11253.
Shao, Chengcheng, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. “The Spread of Low-Credibility Content by Social Bots.” *Nature Communications* 9 (1): 4787.
“Shelley: Human-AI Collaborated Horror Stories.” n.d. Accessed August 21, 2021. https://www.media.mit.edu/projects/shelley/overview/.
Sherif, C. W., M. S. Sherif, and R. E. Nebergall. 1965. *Attitude and Attitude Change*. Philadelphia: W.B. Saunders Company.
Stella, Massimo, Emilio Ferrara, and Manlio De Domenico. 2018. “Bots Increase Exposure to Negative and Inflammatory Content in Online Social Systems.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (49): 12435–40.
Stewart, Alexander J., Mohsen Mosleh, Marina Diakonova, Antonio A. Arechar, David G. Rand, and Joshua B. Plotkin. 2019. “Information Gerrymandering and Undemocratic Decisions.” *Nature* 573 (7772): 117–21.
Stewart, Leo G., Ahmer Arif, and Kate Starbird. 2018. “Examining Trolls and Polarization with a Retweet Network.” In *Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web*. http://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf.
Sunstein, Cass R. 2018. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.
Tucker, Joshua A., Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. “Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.” *: A Review of the …*. https://doi.org/10.2139/ssrn.3144139.
Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. “The Spread of True and False News Online.” *Science* 359 (6380): 1146–51.
Yaniv, Ilan. 2004. “Receiving Other People’s Advice: Influence and Benefit.” *Organizational Behavior and Human Decision Processes* 93 (1): 1–13.
Yildiz, Ercan, Daron Acemoglu, Asuman E. Ozdaglar, Amin Saberi, and Anna Scaglione. n.d. “Discrete Opinion Dynamics with Stubborn Agents.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1744113.
Yildiz, Ercan, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. 2013. “Binary Opinion Dynamics with Stubborn Agents.” *ACM Trans. Econ. Comput.*, 19, 1 (4): 1–30.
Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Supplementary Materials

Different engagement functions

In the main text, we presented results assuming that agents are more likely to engage when the distance between their own opinion and the other agent's opinion is high. Here we study the sensitivity of our results to other engagement functions. Figure S1 shows the same results as Figure 2 in the main text. Instead of assuming that agents are more likely to engage when content is dissimilar, we assume that agents are more likely to engage when the observed content is similar. In Figure S2, we study a bimodal engagement function where agents are more likely to engage with very similar or very dissimilar content and less likely to engage with content that is neither too similar nor too dissimilar.

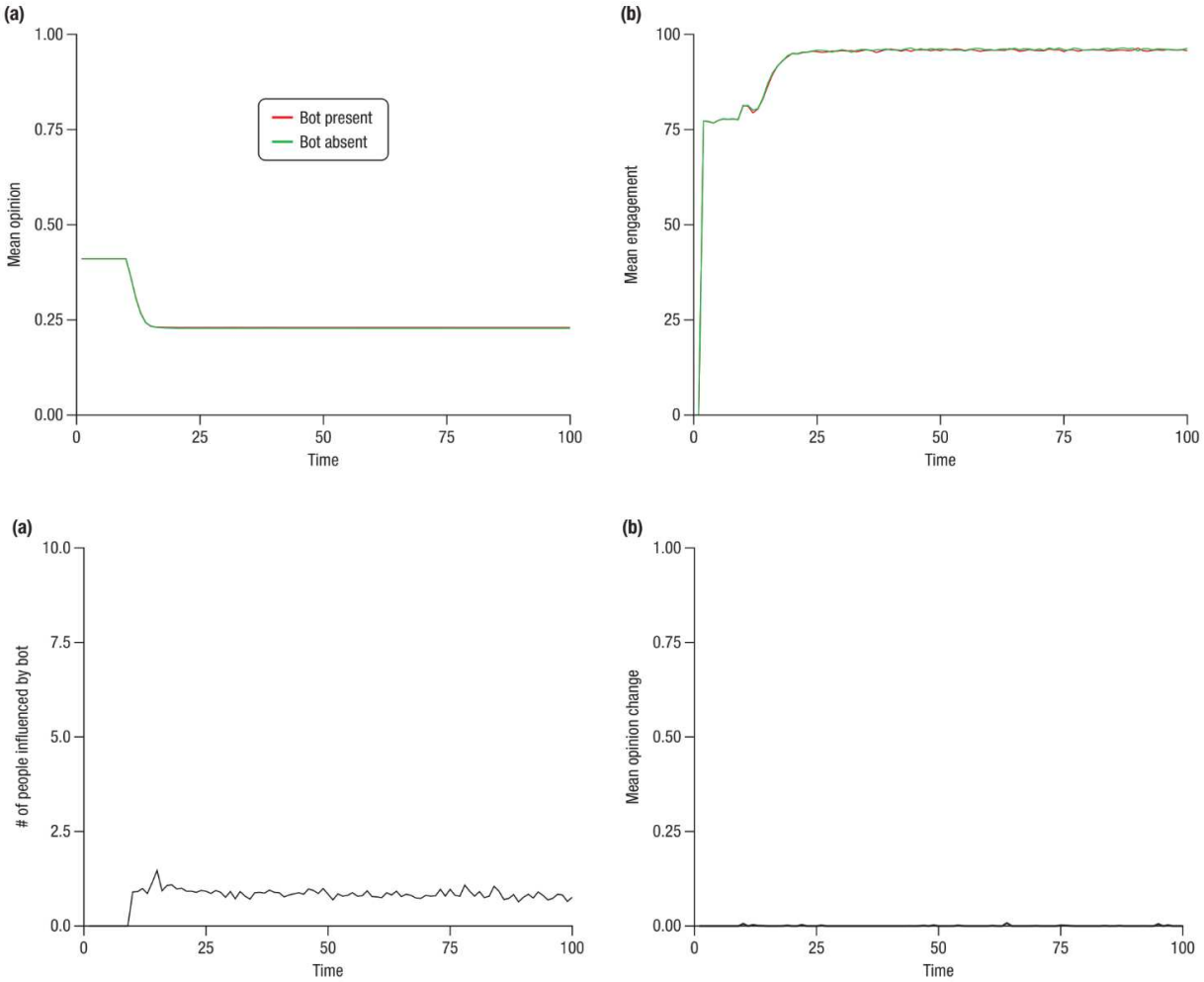

Figure S1. Homophilous engagement function. Agents are more likely to engage with content in their feed closer to their own opinions. (a) population's mean opinion; (b) population's mean engagement; (c) the number of people influenced by the bot. (d) agents' mean opinion shift. The analysis shows that the results reported in the main text

might be sensitive to the specific engagement function used by the agents to choose which content items they engage with.

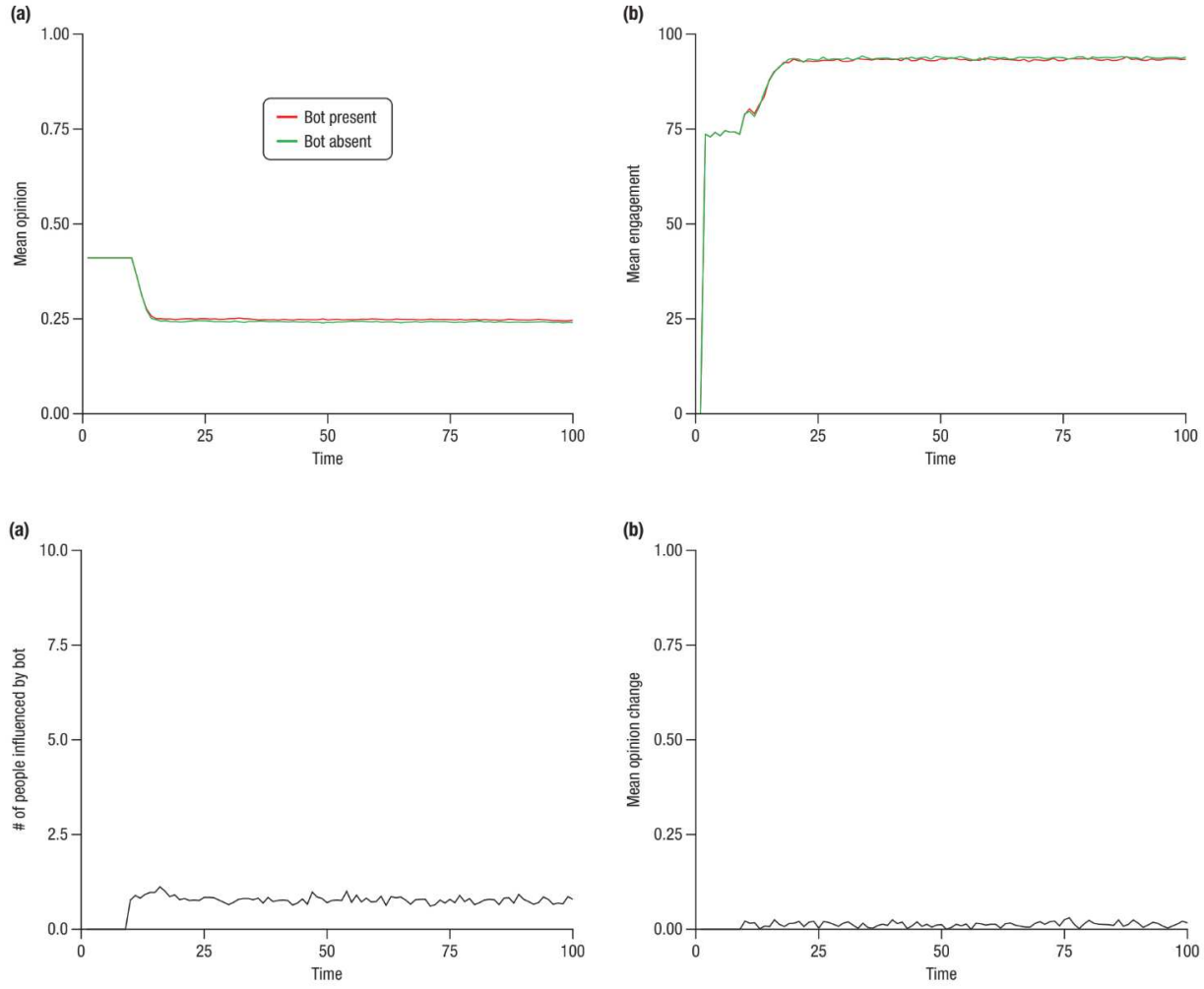

Figure S2. Bimodal engagement function. Agents engage with content following a binomial distribution with bimodal probability for content close to the target agent's private opinion and content distant from the agent's opinion. Content that falls between these two extremes is less likely to generate engagement. (a-d) population's mean opinion, population's mean engagement, number of people influenced by the bot, and mean opinion change as a function of time.

Removing the bot after 40 timesteps

We tested the effect of removing the bot after 40 timesteps. The number of people directly influenced by the bot and the mean opinion change are shown in Figure S3. The figure shows a sudden drop in *direct* influence after the bot is removed from the network. Comparing within-node opinion shifts across conditions, we found that all nodes showed shifted opinions at t=100 (Figure S4). However, compared to the main results shown in Figure 4, the magnitude of the shift is vastly reduced.

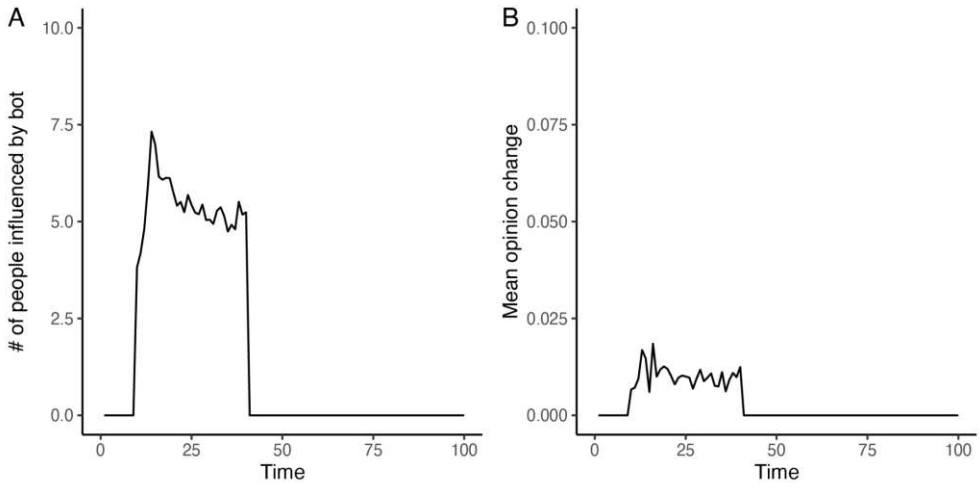

Figure S3. Direct bot influence. (a) In this simulation, the bot was removed after 40 timesteps. A sudden drop in the number of people directly influenced by the bot is observed when the bot is removed. **(b)** After the bot is removed from the network, a drop in average opinion change is observed.

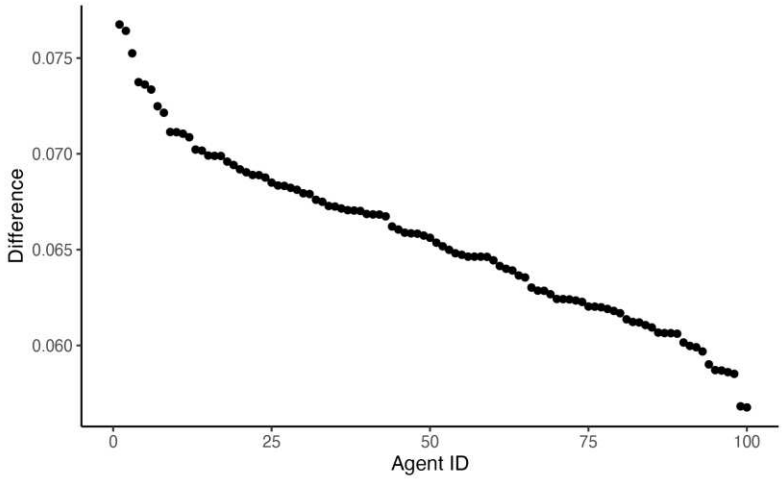

Figure S4. Within-node bot effect across the two simulations. The absolute within-node opinion distance between the bot and no-bot conditions in the final step of the simulation. The bot presence during the first 40 trials still affects within-node opinion differences. However, this difference's magnitude is significantly reduced: it ranges from 5% to 7.5% (compared to 33-48% in Figure 4).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx