Indirect causal influence of a single bot on opinion dynamics through a simple recommendation algorithm

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

[1]New Jersey Institute of Technology
323 Dr Martin Luther King Jr Blvd, 07102,
Newark, NJ

[2]Max Planck Institute for Human Development
94 Lentzeallee, 14195, Berlin, Germany

*Corresponding author

Abstract

The ability of social and political bots to influence public opinion is often difficult to estimate. Recent studies found that hyper-partisan accounts often directly interact with already highly polarised users on Twitter and are unlikely to influence the general population's average opinion. In this study, we suggest that social bots, trolls and zealots may affect people's views not just via a direct interaction (e.g. retweets, at-mentions and likes) and via indirect causal pathways through infiltrating platforms' content recommendation systems. Using a simple agent-based opinion-dynamics simulation, we isolate the effect of a single bot – representing only 1% of the population – on the average opinion of Bayesian agents when we remove all direct connections between the bot and human agents. 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 all other conditions remained identical. Results show that, even in the absence of direct connections, the mere presence of the bot is sufficient to shift the average population opinion. Furthermore, we observe that the presence of the bot significantly affects the opinion of almost all agents in the population. Overall, these findings offer a proof of concept that bots and hyperpartisan accounts can influence average population opinions not only by directly interacting with human accounts but also by shifting platforms' recommendation engines' internal representations.

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

Introduction

This study uses an agent-based simulation to explore the interaction between bots and content recommendation algorithms. We test the hypothesis that when recommender systems mediate information access, bots can affect a population's mean opinion not just by direct interactions with other nodes, but via skewing the training sample fed to the recommender system during training. Thus, by subtly affecting how a centralised recommender system represents a population's preferences and patterns of content engagement, a bot may influence content recommendation at the population level at a much larger scale. This indirect social influence may be more pervasive than direct influence because it would take place even in the absence of 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. Early studies documented the potential effects of bots on skewing opinion distributions and polarizing opinions on social media users and voters[1–7]. These generalised concerns have mobilised platforms to improve algorithmic agents' automatic detection and removal[8–10]. On top of bot influence, influence networks online are characterised by several other phenomena acting together on public opinions, such as human trolls, fake accounts, pink-slime newspapers and "fake news"[11–15]. In this paper, we estimate algorithmic influence lower bound by focusing on the effect of a single algorithmic agent on a population. Our findings can theoretically be generalised to other 'pre-programmed' agents of media manipulation, such as partisan accounts and human trolls, that share several features, such as "stubbornness".

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[16,17]. 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[18] attempted to measure the effect of being exposed to Russia's Internet Research Agency (IRA) content on opinion. 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 significantly change opinions because they primarily interacted with those who were already highly polarised. This phenomenon is also known as "minimal effect"[19–21]. Furthermore, exposure to partisan ideology can counter-intuitively strengthen confidence in one's own belief[22–24]. Overall, these findings cast doubt on the centrality and impact of bot activity on political mobilisations' coverage. 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. 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 influence assumes that exposure to another person's belief (e.g. an advisor) produces changes to a privately held belief[25–28]. However, this simple model of social influence 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 information can affect exposure to specific views[29]. Recommender systems can learn population averages and trends, forming very accurate representations of individual and collective news consumption patterns[30,31]. 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, as nodes are connected in a network, another form of indirect influence is via intermediary nodes. In other words, a bot may directly influence one human but indirectly influence all the humans that this human is connected to.

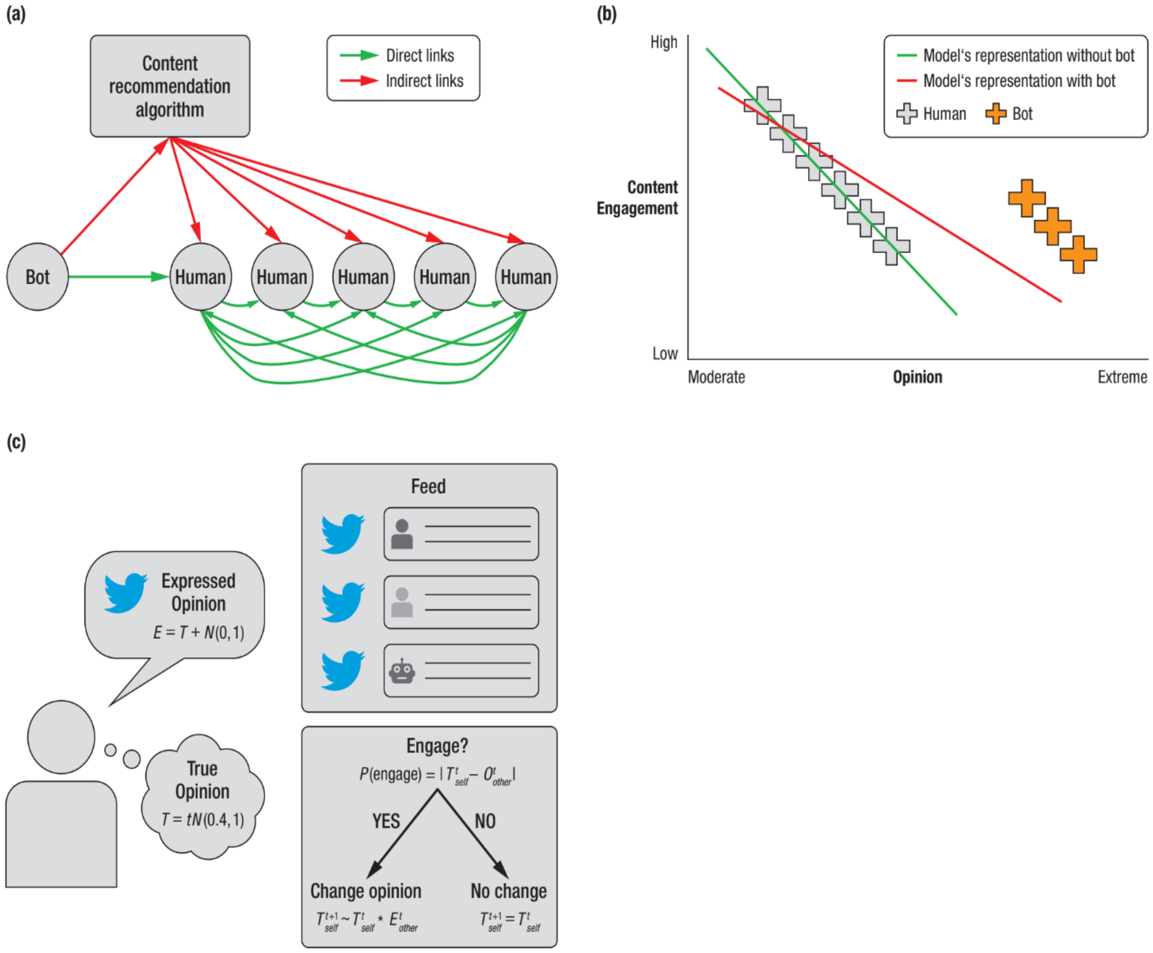
(a)
Content recommendation algorithm
Direct links
Indirect links
Bot
Human
Human
Human
Human
Human
(b)
High
Content Engagement
Low
Moderate
Opinion
Extreme
Model's representation without bot
Model's representation with bot
Human
Bot
(c)
Feed
Expressed Opinion
$E = T + N(0,1)$
True Opinion
$T = tN(0.4,1)$
Engage?
$P(\text{engage}) = |T^t_{self} - O^t_{other}|$
YES
NO
Change opinion
$T^{t+1}_{self} \sim T^t_{self} * E^t_{other}$
No change
$T^{t+1}_{self} = T^t_{self}$

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 (green links) interaction with human agents (e.g. retweets, at-mentions, likes etc.) or indirectly (red links) via affecting the inner 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 behaviour 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 modelled to include a private true opinion and an expressed public opinion. On every round, they were presented with the public opinion of their neighbours and decided whether to engage with this content or not, according to a pre-defined engagement function. Opinion change happened only if the agent engaged with the content.

Methods

Overview

Figure 1A shows a mix of both direct (green) and indirect (red) links. In our simulation, we only focused on indirect links. 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 pre-programmed 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). We created two identical fully connected networks of 100 agents. The two networks differed only in the presence (or absence) of one bot. We initialised 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[27,28], we distinguish between internally held beliefs and externally observable behaviour. We assume that observable behaviour 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 post or share their own[32]. One does not need to tweet about climate change to have an opinion about it. Similarly, this distinction between internally held and publicly displayed beliefs allows us to train the recommender algorithm only with externally observable behaviour rather than making the unrealistic assumption that the algorithm has direct access to a user's unobservable opinion. We call all such externally observable behaviours (e.g. tweets, likes and reactions) 'engagement'. Thus, both the recommender algorithm and agents must infer other agents' underlying opinions from engagement behaviours. Second, we use a Bayesian opinion updating rule[23,33,34]. The Bayesian update offers a natural way to consider not just the opinion direction on a binary issue but also belief's 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[23,35]. 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 (although see[22]).

We simulate a simplified social network model where a recommender system learns and presents a personalised content feed to agents in the network. This feed contains the expressed opinions of other agents in the network. Each agent can observe and decide to interact with other agents' opinions by updating and expressing their own opinions. In two separate but otherwise identical conditions, 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. We study whether this bot can infiltrate the feed created by the recommender system by influencing the statistical relationships it learns.

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.

Simulation procedure

Agents. We simulate N=100 agents connected through a fully connected network. Each agent i is represented by a true private *opinion* in the range $]0,1[$ drawn from a truncated Normal distribution

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

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 *j* 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 *E* of another agent *j*. 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 sensationalist or count-intuitive content more than moderate content[15,36]. 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. On every time step *t* when agents decide to *engage,* they update their own opinions using a Bayesian opinion update function:

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

On timesteps when the agents decide not to engage, they keep their opinion from the previous timestep, time *t-1*: $T_i^t = T_i^{t-1}$.

Feed. Each agent *i* 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 10 timesteps of the simulation by presenting agents randomly in the feeds. Notice that although we simulate a fully connected network, it is the feed that determines whether an agent will see or not see the content created by another agent.

Bot. The bot is represented as an agent that does not change its opinion but sticks to the same opinion throughout the simulation[4,37,38]. In separate 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 represents a situation where agents are not polarized or extreme, and the mean difference between agents and the bot is not very extreme. 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 its decision, the simulation proceeds to the next timestep. We repeat the procedure for t=100 timesteps and r=100 replications. We record the opinions of each agent on each timestep, as well as 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 completely identical simulation conditions except for the presence of the bot. This allows for precise measurements regarding the influence of the bot on the network.

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, presents 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 (Supplementary Materials). From t=10 onwards, we observe a large jump in engagements, showing that the recommender system is becoming increasingly efficient at recommending content that agents will engage with. 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.

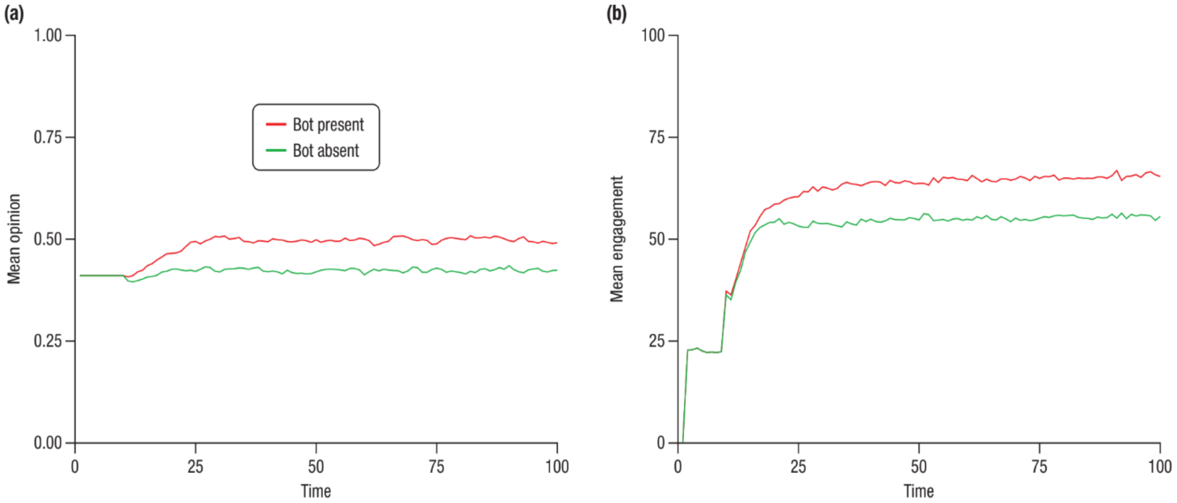

Figure 2. Mean opinion and mean engagement in networks with and without bot influence. (a) Mean opinion of the agents over time. (b) Mean number of agents engaging in each timestep. Red: Condition where the bot is part of the social network, Blue: Condition where the bot is not part of the social network.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

Magnitude of direct bot influence on individual agents

Here, we investigate the reasons underlying average opinion shifts 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 due to the fact that bots' opinions are less extreme and thus less engaging the more the population's opinion shifts towards the bot's opinion. The finding of low engagement and opinion shift replicates "minimal effect" findings online[18,19]. 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 we compare, in the next paragraph, the two simulation conditions (bot vs no bot) while using the same random seed and holding all other conditions constant.

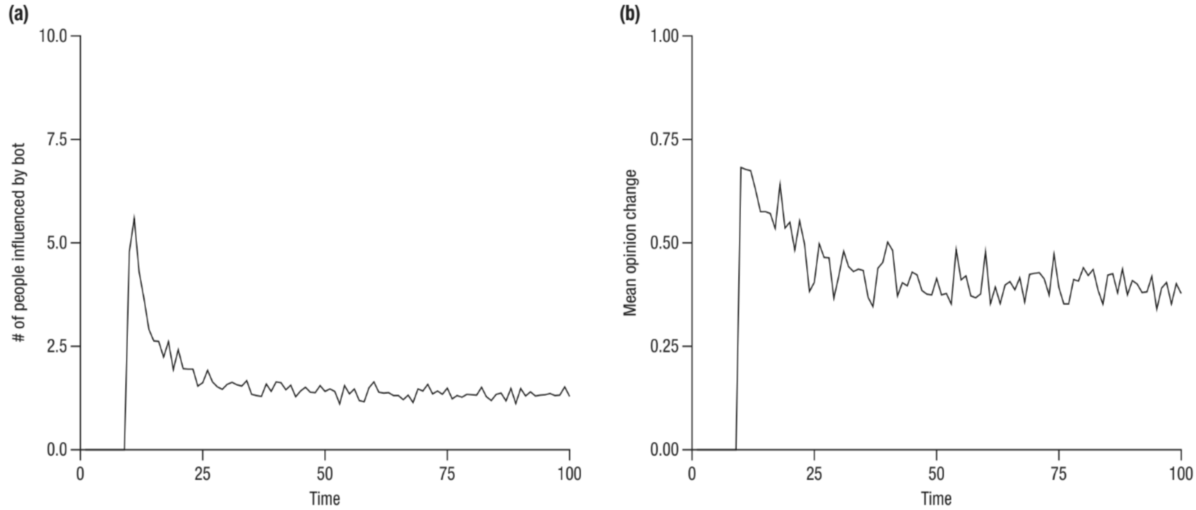

Figure 3. Direct bot influence. (a) Average number of people influenced by the bot on each timestep. Influence is defined as when an agent is presented with content produced by the bot, engages with this content and shifts its own opinion. (b) Mean change in opinion for agents influenced by the bot on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

Individual-level shift in opinion as a result of direct and indirect bot influence

Figure 4a shows the difference in opinion between the same agent across the two simulation conditions, holding all other aspects of the simulation constant. Initialising 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 effect of bot presence. Differences between the two counterfactual worlds reflect direct bot influence and all secondary effects caused by introducing the bot. Even though a small minority of agents was directly influenced by the bot (Figure 3), we found that, compared to a counterfactual simulation, the bot had a pervasive indirect effect on the entire population (Figure 4a). The magnitude of influence on opinion varied considerably, from 33 to 48 percentage points, but virtually all agents showed shifted opinions compared to the control condition. 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. Our model shows that bots' influence is magnified when we account for indirect influence, either via the recommender system or via other intermediary agents. This is a striking result that 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 be an underestimation of the actual capacity of a bot to bias and sway a population's opinion dynamics.

Finally, the above results assumed that the average opinion in the population is N(0.4,1) and the bot opinion is 0.7. Thus, 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 4b 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 stronger effect on the population when its opinion is more distant from the average opinion of the population. The results provide further support to 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.

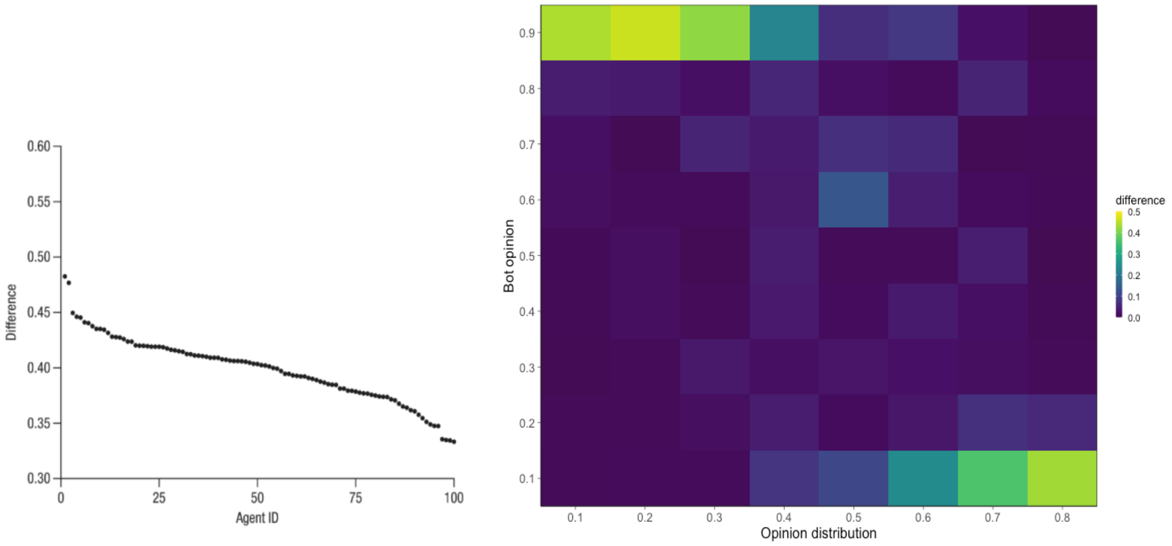

Figure 4. Within-agents bot effect across the two simulations. (a) Difference between each agent's opinion at time *t=100* between the two simulation conditions (bot vs no bot). This measures the total impact the bot has on the opinions of the same agents in the network. **(b)** Sensitivity of our results to different values of agent and bot opinion. A heatmap where the x-axis shows different values of bot's 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 stronger effect on the population when its opinion is more distant from the average opinion of the population.

Discussion

This paper investigated the indirect influence that programmed agents, such as bots, trolls and zealots, can have via recommender systems. We find that a single bot can substantially shift the mean opinion and mean engagement of the population compared to a control condition without a bot. Even though only a minority of 'human' agents (2.5%) directly engaged with the bot's content, the bot disproportionately affected the average shift in opinion observed in the population (30%). Notably, virtually all agents in the population were influenced by the bot presence, with opinion shifts ranging from 2 to 48 percentage points.

This latter result 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. Thus, our first contribution is in advancing the debate around bots' influence and media manipulation. They highlight a previously unexplored phenomenon and draw 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 directly strengthening causal inference. We initialised 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 it may be difficult to manipulate these systems outside the lab, researchers have recently successfully inferred the hidden mechanisms underlying several proprietary algorithms by systematically prompting them[39–41]. We acknowledge that our findings are specific to our parametrization of the simulation used in this study. Although we characterized part of this parameter space in Figure 4b, more work needs to be done, e.g. exploring different noise distributions and opinion update functions, and measuring standard errors across runs. Future studies should also investigate the effects of network size and alternative network structures on bot influence. Finally, our study used a simple logistic model to predict engagement. Although real recommender systems are much more complex machines, the effects highlighted in our findings are likely to affect, at least to some degree, any content filtering algorithm that extrapolates the behaviour of one user to another. This is the case in most recommender systems. More complex recommendation systems, such as collaborative filtering, may still be affected by the same dynamics highlighted here as long as they use population averages to predict individual preferences[31,42–45].

The last contribution of this work is the use of a Bayesian opinion update model that captures dynamics of belief conviction, uncertainty and probabilistic judgments[23,33,34]. We selected this opinion update model as it offers several advantages over more common linear opinion-dynamics models[27,28]. 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 naturally represents confidence as the distance from the maximum uncertainty point (50%). Third, encounters with agreeing agents tend to increase one’s belief conviction, while encounters with disagreeing agents increase uncertainty. 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.

We conclude by suggesting possible ways to reduce the risk of public opinion manipulation online. Although improving the detection and removal of automated accounts remains an important strategy, a more valuable approach might be to better regulate recommender and filtering algorithms to make them more transparent and auditable, and thus resilient to manipulation.

References

1. Bessi, A. & Ferrara, E. Social Bots Distort the 2016 US Presidential Election Online Discussion. *SSRN* **21**, (2016).
2. Lerman, K., Yan, X. & Wu, X.-Z. The 'Majority Illusion' in Social Networks. *PLoS One* **11**, e0147617 (2016).
3. Broniatowski, D. A. *et al.* Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *Am. J. Public Health* **108**, 1378–1384 (2018).
4. Stewart, A. J. *et al.* Information gerrymandering and undemocratic decisions. *Nature* **573**, 117–121 (2019).
5. Paul, C. & Matthews, M. The Russian 'firehose of falsehood' propaganda model. *Rand Corporation* 2–7 (2016).
6. Shao, C. *et al.* The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 4787 (2018).
7. Stella, M., Ferrara, E. & De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12435–12440 (2018).
8. Howard, P. How Political Campaigns Weaponize Social Media Bots. *IEEE Spectrum* **Oct**, (2018).
9. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
10. Ledford, H. Social scientists battle bots to glean insights from online chatter. *Nature* **578**, 17–17 (2020).
11. Hurtado, S., Ray, P. & Marculescu, R. Bot Detection in Reddit Political Discussion. in *Proceedings of the Fourth International Workshop on Social Sensing* 30–35 (Association for Computing Machinery, 2019).
12. Linvill, D. L. & Warren, P. L. Troll factories: The internet research agency and state-sponsored agenda building. *Resource Centre on Media Freedom in Europe* (2018).
13. Aral, S. & Eckles, D. Protecting elections from social media manipulation. *Science* **365**, 858–861 (2019).
14. Tucker, J. A. *et al.* Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *: a review of the ...* (2018) doi:10.2139/ssrn.3144139.
15. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
16. Guess, A., Nagler, J. & Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci Adv* **5**, eaau4586 (2019).
17. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci Adv* **6**, eaay3539 (2020).
18. Bail, C. A. *et al.* 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**, 243–250 (2020).
19. Zaller, J. R. *The Nature and Origins of Mass Opinion*. (Cambridge University Press, 1992).
20. Endres, K. & Panagopoulos, C. Cross-Pressure and Voting Behavior: Evidence from Randomized Experiments. *J. Polit.* **81**, 1090–1095 (2019).
21. Kalla, J. L. & Broockman, D. E. The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments. *Am. Polit. Sci. Rev.* **112**, 148–166 (2018).
22. Bail, C. A. *et al.* Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115**, 9216–9221 (2018).
23. Pescetelli, N. & Yeung, N. The effects of recursive communication dynamics on belief updating. *Proceedings of the Royal Society B: Biological Sciences* **287**, 20200025 (2020).
24. González-Bailón, S. & De Domenico, M. Bots are less central than verified accounts during contentious political events. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
25. Flache, A. *et al.* Models of Social Influence: Towards the Next Frontiers. *Journal of Artificial Societies and Social Simulation* **20**, (2017).
26. Deffuant, G., Neau, D., Amblard, F. & Weisbuch, G. Mixing beliefs among interacting agents. *Adv. Complex Syst.* **03**, 87–98 (2000).
27. DeGroot, M. H. Reaching a Consensus. *J. Am. Stat. Assoc.* **69**, 118 (1974).
28. Friedkin, N. E. & Johnsen, E. C. Social influence and opinions. *J. Math. Sociol.* **15**, 193–206 (1990).
29. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* (2015).
30. Das, A., Datar, M., Garg, A. & Rajaram, S. Google news personalization: scalable online collaborative filtering. in *Proc. of the 16th Int.Conf. on World Wide Web* 271–280 (2007).
31. Pipergias Analytis, P., Barkoczi, D., Lorenz-Spreen, P. & Herzog, S. The Structure of Social Influence in Recommender Networks. in *Proceedings of The Web Conference 2020* 2655–2661 (Association for Computing Machinery, 2020).
32. Lazer, D. Studying human attention on the Internet. *Proceedings of the National Academy of Sciences of the United States of America* vol. 117 21–22 (2020).
33. Pescetelli, N. & Yeung, N. The role of decision confidence in advice-taking and trust formation. *J. Exp. Psychol. Gen.* (2020) doi:10.1037/xge0000960.

34. Harris, A. J. L., Hahn, U., Madsen, J. K. & Hsu, A. S. The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cogn. Sci.* **40**, 1496–1533 (2016).
35. Pescetelli, N., Rees, G. & Bahrami, B. The perceptual and social components of metacognition. *J. Exp. Psychol. Gen.* **145**, 949–965 (2016).
36. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
37. Karan, N., Salimi, F. & Chakraborty, S. Effect of zealots on the opinion dynamics of rational agents with bounded confidence. *Acta Phys. Pol. B* **49**, 73 (2018).
38. Yildiz, E., Acemoglu, D., Ozdaglar, A. E., Saberi, A. & Scaglione, A. Discrete Opinion Dynamics with Stubborn Agents. *SSRN Electronic Journal* doi:10.2139/ssrn.1744113.
39. Ali, M. *et al.* Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* **3**, 1–30 (2019).
40. Hannak, A. *et al.* Measuring personalization of web search. in *Proceedings of the 22nd international conference on World Wide Web* 527–538 (Association for Computing Machinery, 2013).
41. Robertson, R. E., Lazer, D. & Wilson, C. 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* 955–965 (ACM Press, 2018).
42. Ricci, F., Rokach, L. & Shapira, B. Introduction to Recommender Systems Handbook. in *Recommender Systems Handbook* (eds. Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B.) 1–35 (Springer US, 2011).
43. Das, A. S., Datar, M., Garg, A. & Rajaram, S. Google news personalization: scalable online collaborative filtering. in *Proceedings of the 16th international conference on World Wide Web* 271–280 (Association for Computing Machinery, 2007).
44. Koren, Y. & Bell, R. Advances in Collaborative Filtering. in *Recommender Systems Handbook* (eds. Ricci, F., Rokach, L. & Shapira, B.) 77–118 (Springer US, 2015).
45. Analytis, P. P., Barkoczi, D. & Herzog, S. M. Social learning strategies for matters of taste. *Nat Hum Behav* **2**, 415–424 (2018).

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.

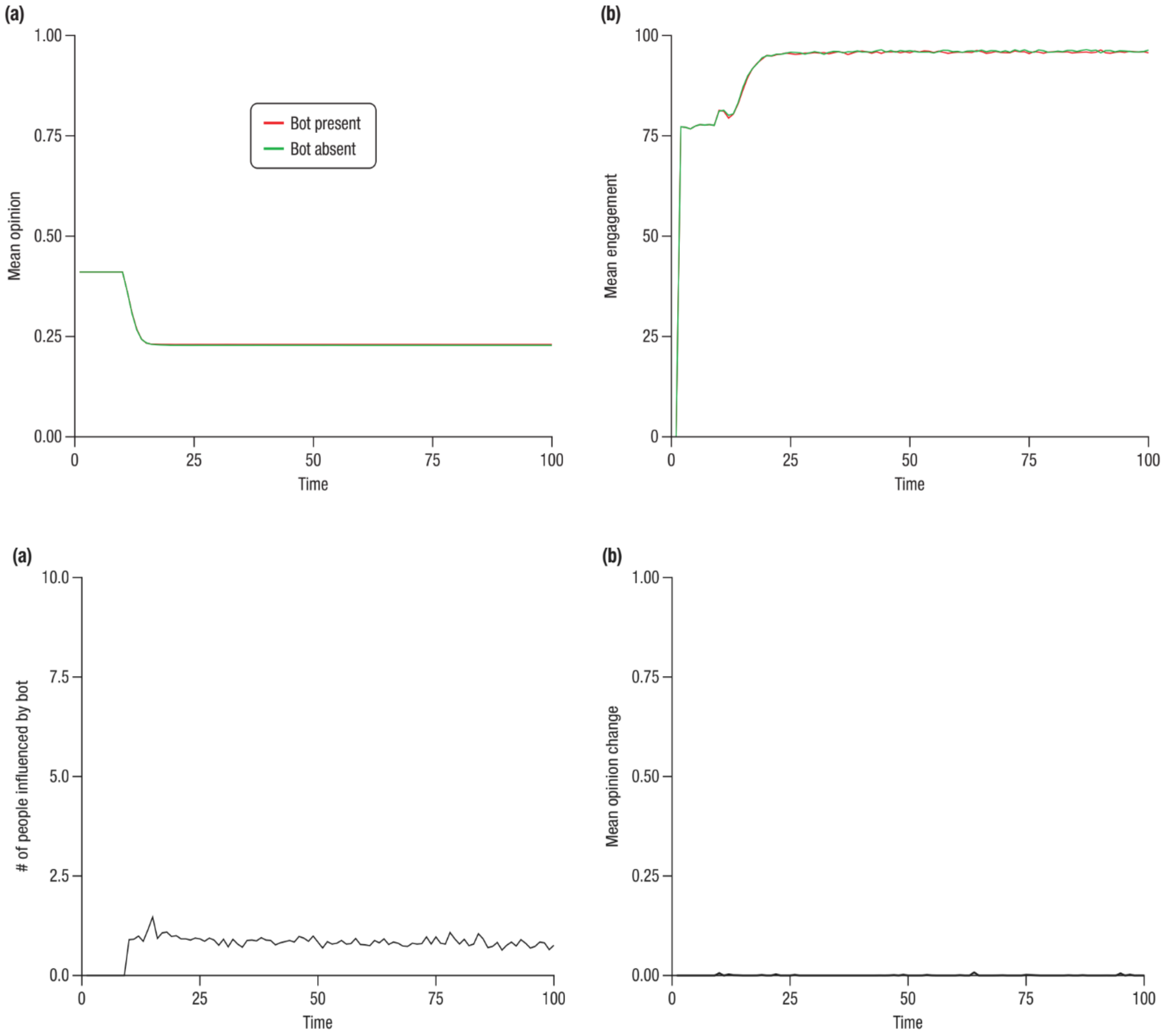

Figure S1. Homophilous engagement function. Agents are more likely to engage with content in their feed that is closer to their own opinions. (a) population's mean opinion; (b) population's mean engagement; (c) the number of people influenced by the bot. (d) agents' mean opinion shift. The analysis shows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

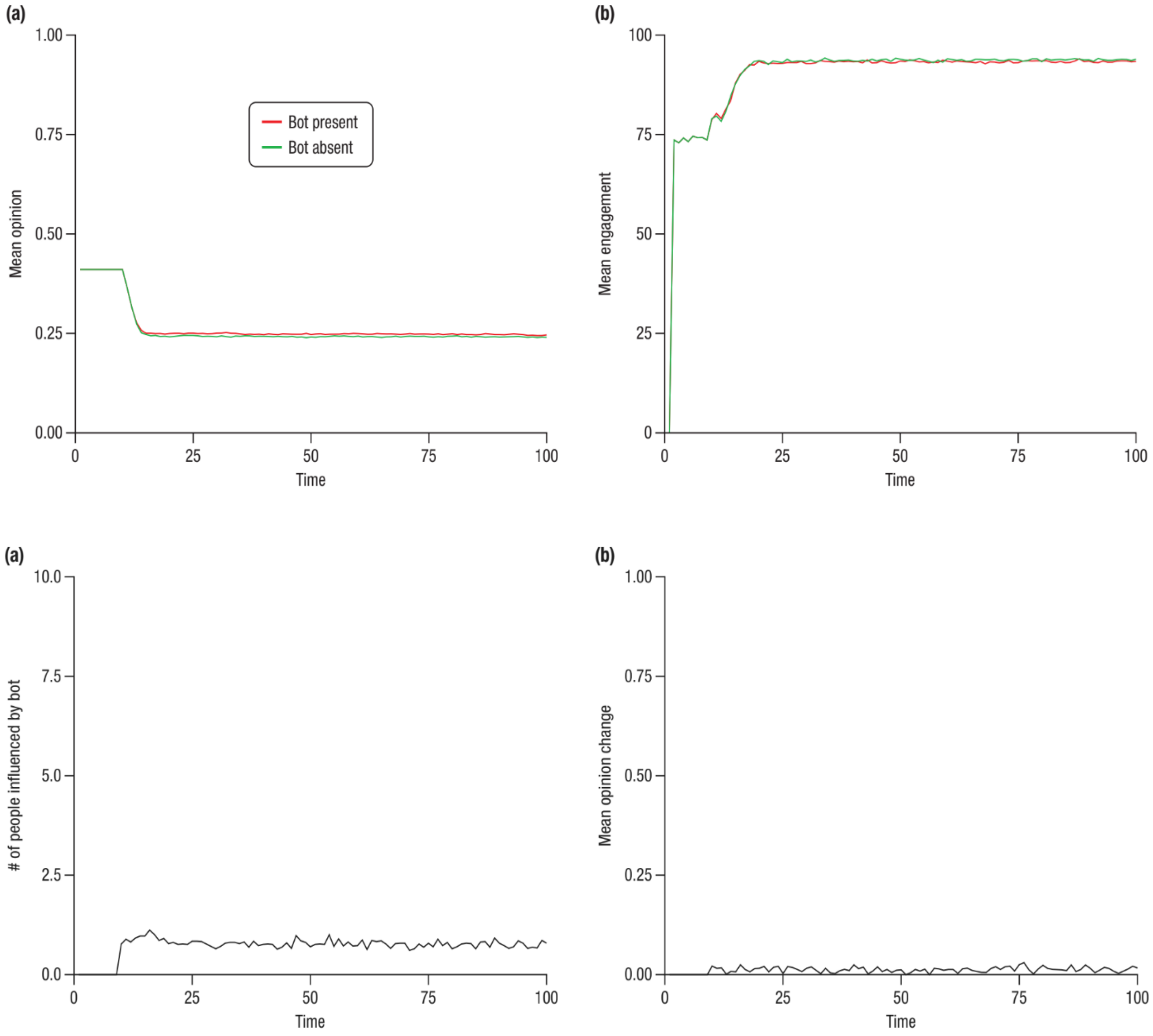

Figure S2. Bimodal engagement function. Agents' engagement with content follows a binomial distribution with bimodal probability for content that is close to the target agent's private opinion as well as content that is distant from the agent's opinion. Content that falls between these two extremes is 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. Notice that contrary to the main text results (Figure S2) here no difference emerges between conditions. This is likely due to the bimodal engagement function. Agents using this engagement function were more likely to engage with content that was similar to their own opinion. According to the Bayesian update rule (Equation 4), this preference for similar content generates escalation dynamics that lead to agents reinforcing their own opinion[23] and thus being deaf to the bot's different opinion.

NEW ANALYSIS

1. Figure 4 signed differences
2. New update functions
3. Differences in Beta coefficients between the two counterfactual simulations
4. 2nd and 3rd order degree effects
5. (isolate the bot)