JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

BotSim: Mitigating The Formation Of Conspiratorial Societies with Useful Bots

Lynnette Hui Xian Ng [1] **and Kathleen M. Carley** [1]

[1] *Center for Computational Analysis of Social and Organizational Systems (CASOS), Software and Societal Systems, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213*
Correspondence should be addressed to *lynnetteng@cmu.edu*

Journal of Artificial Societies and Social Simulation 29(1) 4, 2026
Doi: 10.18564/jasss.5881 Url: http://jasss.soc.surrey.ac.uk/29/1/4.html

Received: 29-06-2025 Accepted: 12-12-2025 Published: 31-01-2026

Abstract: Societies can become a conspiratorial society where there is a majority of humans that believe, and therefore spread, conspiracy theories. Artificial intelligence gave rise to social media bots that can spread conspiracies in an automated fashion. Currently, organizations combat the spread of conspiracies through manual fact-checking processes and the dissemination of counter-narratives. However, the effects of harnessing the same automation to create useful bots are not well explored. To address this, we create BotSim, an Agent-Based Model of a society in which useful bots are introduced into a small world network. These useful bots are: Info-Correction Bots, which correct bad information into good, and Good Bots, which put out good messaging. The simulated agents interact through generating, consuming and propagating information. Our results show that, left unchecked, Bad Bots can create a conspiratorial society, and this can be mitigated by either Info-Correction Bots or Good Bots; however, Good Bots are more efficient and sustainable than Info-Correction Bots . Proactive good messaging is more resource-effective than reactive information correction. With our observations, we expand the concept of bots as a malicious social media agent towards automated social media agent that can be used for both good and bad purposes. These results have implications for designing communication strategies to maintain a healthy social cyber ecosystem.

Keywords: Agent-Based Modeling, Conspiracy Society, Cyber Social Agents, Useful Bots

Introduction

1.1 Social media bots, or bots for short, are automated users that can distribute content (Ng & Carley 2025b). Studies on bots typically focus on the negative societal effects of bots, such as their spread of disinformation, interference in stock markets and political events, and manipulation of public opinion (Aldayel & Magdy 2022; Benigni et al. 2018). A much overlooked area of bots is how their automated nature can be harnessed for social good. A small handful of studies documented useful bots in the crisis informatics realm, for situational awareness and organizing aid (Hofeditz et al. 2019). But can bots also be harnessed for preventing the manipulation of public opinion and the spread of disinformation?

1.2 In this paper, we use an agent-based simulation to investigate the effect of useful bots in combating malicious bots and preventing the formation of a conspiratorial society. A conspiratorial society is a society where there is a large number of individuals who believe, and therefore share conspiratorial narratives (Featherstone 2000; Dyrendal 2014). Conspiracies can be damaging to society because the extreme belief in conspiracies can result in offline consequences. In 2016, the Pizzagate conspiracy rumored that there was a child pedophilia ring that used a pizza store as the front, resulting in a vigilante shooting incident (Bleakley 2023). In 2020, the coronavirus vaccine conspiracy theory stated that vaccines are a government tactic to spy on the population, or that vaccines have microchip implants (Phillips et al. 2022), creating barriers that reduced the ability of governments to curb the pandemic (Romer & Jamieson 2020).

1.3 While conspiracy theories share some characteristics with broader categories of misinformation and fake news, they exhibit distinct features that warrant specialized studies (Douglas & Sutton 2018; Douglas et al. 2019). First, conspiracy theories form interconnected belief systems where believers resist contradictory evidence through a cognitive bias process called motivated reasoning (Leman & Cinnirella 2013). Second, conspiracy theories exhibit unique social media propagation patterns, demonstrating sustained circulation and evolution over time (Dow et al. 2021). Third, conspiracy theories are more likely to inspire offline behavioral consequences (Rottweiler & Gill 2022; Jolley & Douglas 2014), from vaccine hesitancy to vigilante actions, making their study particularly crucial for the field of social cyber security (Carley 2020), which aims to build a resilient social cyber infrastructure. This work focuses on conspiracy theories rather than general misinformation, investigating the mechanics of the formation of a conspiratorial society, which is a society where the conspiracy theories have become beliefs normalized across the population.

1.4 To examine the effects of useful bots on a society, we create BotSim, an agent-based model, that incorporates the mechanics of Bad Bots that spread conspiracies, and useful bots that perform information correction (Info-Correction Bots) and good messaging (Good Bots) tasks. Our results show that if Bad Bots are left unchecked, a conspiratorial society will form. The introduction of useful bots prevents a conspiratorial society, although the society will reach a state where Bad Humans are the majority (i.e., Human agents that spread conspiracies). Importantly, a smaller proportion (50%) of Good Bots compared to Info-Correction Bots is required to prevent the formation of a conspiracy society, and the employment of sufficient Good Bots (Bad:Good Bots ratio $\geq$1:1.6) prevents Bad Humans from reaching the majority.

1.5 These experimental results offer insights into the design and deployment of communication strategies, in particular good messaging, to slow the rate of formation of conspiracy societies on key issues.

Related Work

2.1 Conspiracy theories thrive in environments of uncertainty and complex societal challenges (Douglas et al. 2019). Events like an unprecedented health pandemic (i.e., the 2020 coronavirus pandemic) or unexplained phenomena (i.e., the 2024 Dubai rainfall and flood) are prime ground for conspiracy theories. Conspiracy theories arise as explanations for people to make sense of the situation. Conspiracy beliefs are difficult to correct because they promote interconnected belief systems, where rejecting one aspect of the theory threatens a person's entire worldview (Goertzel 1994). Social networks facilitate the formation of conspiratorial societies because individuals tend to cluster with others who have similar beliefs. This homophilic behavior results in cohesive echo chambers that amplify the spread of conspiracy narratives (Bessi et al. 2015). In fact, the algorithmic design of social platforms reinforces homophily by exposing users to content that they prefer, accelerating the formation of conspiratorial communities.

2.2 Empirical studies show that the underlying network topology of conspiracy diffusion can resemble a small-world structure, where there are tightly clustered communities bridged by short path lengths (Ch'ng 2015; Huang & Jin 2011). This structure allows a small number of highly active accounts to inject information across otherwise separate clusters, accelerating cross-community spread and the network's exposure to conspiracy theories.

2.3 In social networks, automated bot users also contribute to the spread of information. Studies of posts on X consistently show that bots produce posts at a scale almost twice that of humans in major events. Bot accounts posted 2.6 to 3 times more messages than human accounts during the 2020 US Presidential Elections (Luceri et al. 2020), and 2 times more messages during the 2020 Coronavirus pandemic (Ng & Carley 2025b). Although most studies model bots as harmful agents that spread conspiracy theories (Jin & Liao 2025; Ross et al. 2019), this study expands the concept of the automated user to harness its automation as part of the intervention of conspiratorial spread.

2.4 Current methods of combating conspiracy theories focus primarily on fact-checking and debunking. The International Fact-Checking Network coordinates worldwide efforts to combat misinformation across social media platforms, but the fact-checking and debunking are heavily reliant on volunteer efforts. To speed up the fact-checking process, computational approaches for identifying and countering conspiracy content have emerged. Several automated systems use linguistic markers and narrative structures to identify conspiratorial content (Memon & Carley 2020; Choudhary & Arora 2021). Automated fact-checking methods with machine learning or graph approaches show promise with scaling detection efforts, but is often resource- and data-hungry (Shu et al. 2017; Medeiros & Braga 2020).

2.5 At the same time, qualitative asymmetries in human information sharing further complicate interventions. Empirical research shows that individuals with stronger conspiratorial tendencies exhibit more uniform information sharing, while those with weaker leanings share more diverse content (Pennycook & Rand 2021). Therefore, simulation studies are useful to model and predict the impact of the complex dynamics of information diffusion.

2.6 To make interventions more difficult, online social networks can sometimes enable conspiracy theories to spread faster than fact-checking mechanisms can respond (Vosoughi et al. 2018), because these theories are memorable and can incite huge engagement. Institutional fact-checking efforts can fail to reach susceptible communities (Jack 2017). Therefore, the deployment of useful bots to automate such interventions, can potentially provide a larger message reach through the power of social networks. Simulations are a useful tool to showcase the theoretical efficacy of automated interventions: Kumar & Geethakumari (2014) built an agent-based framework to counteract misinformation flows in social networks, and Varol & Uluturk (2020) showed that strategically deployed counter-narrative bots can effectively reduce the spread of misinformation in simulated networks. We build on these works to demonstrate the effectiveness of useful bots in combating conspiracy theories.

2.7 Good messaging is a complementary approach to fact-checking. Fact-checking targets specific conspiratorial narratives, while good messaging promotes accurate information. Fact-checking is a reactive action, while good messaging is a proactive action. Good messaging builds on the inoculation theory, a psychological framework that exposes individuals to weakened forms of misinformation and the corresponding refutations to develop resistance against future, stronger persuasive attempts (Compton et al. 2021). Preemptively exposing people to weakened forms of information as inoculation (van der Linden et al. 2017) and media literacy interventions can help combat conspiracy thinking (Huang et al. 2024). However, such methods that focus primarily on belief correction at the individual level, and are difficult to scale up. Therefore, there is a need to harness automated tools to combat conspiracy theories in the online space as they arise.

2.8 Agent Based Modeling (ABM) has been used as a means to understand the dynamics of information propagation in online societies, demonstrating the formation of polarization and echo-chambers in social networks (Betts & Bliuc 2022; Keijzer et al. 2024; Lu & Lee 2024). Such models enable the creation of agent entities that each have unique characteristics and heuristics of interaction with other agents and the environment. By simulating successive interactions between agents across time, ABM allows for the observation of emergent behaviors that connect micro-level individual agent behavior to macro-level patterns (Betts & Bliuc 2022). This approach has proven particularly valuable for evaluating countermeasures because it enables controlled experimentation without the ethical concerns associated with introducing actual changes into real social networks (Murdock et al. 2025).

2.9 ABM has also been used to model intervention strategies against harmful information on social media platforms. Several studies have developed sophisticated frameworks that incorporate heterogeneous agents with varying susceptibility to misinformation and diverse intervention mechanisms (Carragher et al. 2023; Abouzeid et al. 2024). Other similar studies model use opinion influence as a proxy for reinforcement learning in a social network structure (Jin & Guo 2025). These models demonstrate that successful interventions depend on the timing, target selection and the interactions between different types of countermeasures. Recent research has also explored the use of automated bot agents for interventions strategies such as fact-checking, content moderation, and counter-messaging (Cisneros-Velarde et al. 2019). With this backdrop, we employ the ABM technique to study the effects of useful bots in disrupting the activities of bad bots.

Model

Purpose and overview

3.1 Our simulation uses an Agent-Based Model, named BotSim, set up in a Small World network to simulate the consumption and dissemination of information by a community of agents. The Small World network structure is commonly used to model social networks (Fan et al. 2012), because this structure (1) shows high local clustering, which reflects the tendency for people to form tightly connected social groups, and (2) has a small characteristic path length, mimicking the random and rapid information dissemination (Watts & Strogatz 1998).

3.2 The primary purpose of BotSim is to investigate the conditions under which useful bots can effectively counteract the formation of conspiratorial societies. These useful bots perform actions of information correction and positive messaging. This leads us to the main research question: Can automated useful bots prevent the

formation of conspiratorial societies where Bad Bots are present? Empirically, we define a conspiracy society when all the Human Agents in the simulation have turned to Bad Humans. In our investigation of this question, we analyzed the optimal deployment ratios between Bad Bots and useful bots to maximize intervention effectiveness, and how the two types of useful bots (Info-Correction versus Good Bots) compare in their resource efficiency.

3.3 The BotSim model simulates a dynamic social media information ecosystem where agents continuously generate, consume, and propagate information through their ego network connections. In our virtual simulation experiments, we vary the proportions of four different agent types and measure the time in simulation ticks that is required for the convergence of the simulation towards a conspiracy society. The design choices made in the model are grounded in empirical literature on conspiracy belief, cognitive biases, and information propagation. Our ABM leverages the results from static analytical models (see Table 7) to capture temporal dynamics of belief formation and feedback loops that can stabilize or destabilize the information environment.

Types of agents and information

3.4 There are two types of agents that interact in a network within BotSim: Human Agents and Bot Agents. For each type of agents, there are sub-types of agents.

3.5 First, we have Human Agents. Human Agents are either in a Good or Bad state. The state that the Human Agents are in depends on the amount of each type of Information they consume. The two types of Human agents are:

1. **Good Humans** represent normal humans that generate and consume information in a social network.
2. **Bad Humans** represent humans that have become believers of conspiracy theories, and therefore only propagate Bad Information.

3.6 Next, we have Bot Agents, of which we have three types:

1. **Bad Bots** are malicious bot agents that disseminate conspiracy theories. The number of Bad Bots is $\alpha_1 \times$ (total number of Humans).
2. **Info-Correction Bots**, short for information-correction bots, that represent agents who fact-check false information and present corrected information. The number of Info-Correction Bots is $\alpha_2 \times$ (number of Bad Bots)
3. **Good Bots** represent agents that produce good messaging. The number of Good Bots is $\alpha_3 \times$ (number of Bad Bots).

3.7 In this simulation, we assume that the Info-Correction and Good Bots are 100% secure. That is these agent types are will always be reliable for their fact-checking and positive messaging activities respectively. Visually, the agents are represented by the person stick figure.

3.8 Agents consume and disseminate Information. There are two types of Information in our environment:

1. **Good Info**, which represents factual information. Good Info is visually represented by green translucent circles.
2. **Bad Info**, which represents conspiracy theories. Bad Information is visually represented by red translucent circles on the person icons.

The model's main routine

3.9 In our simulation, BotSim, we simulate the impact of information correction and the introduction of new good information in a social network. The model's main routine begins with creating a set of agents and placing them in a small world network structure. These agents consist of $n_h = 1000$ Human agents and some of each type of Bots. The number of each type of Bot agents is a proportion of the number of Human agents, and is an independent variable varied during the experiments. Then, the simulation has four cyclical stages: Information Generation, Information Consumption, Information Propagation and State Update. These stages occur at each tick of the simulation until the simulation is stopped. Figure 1 contains screenshots of the network structure

of the initial setup and the final setup when the simulation has stopped. Figure 2 illustrates the flowchart of simulation logic. The simulation is implemented in NetLogo.

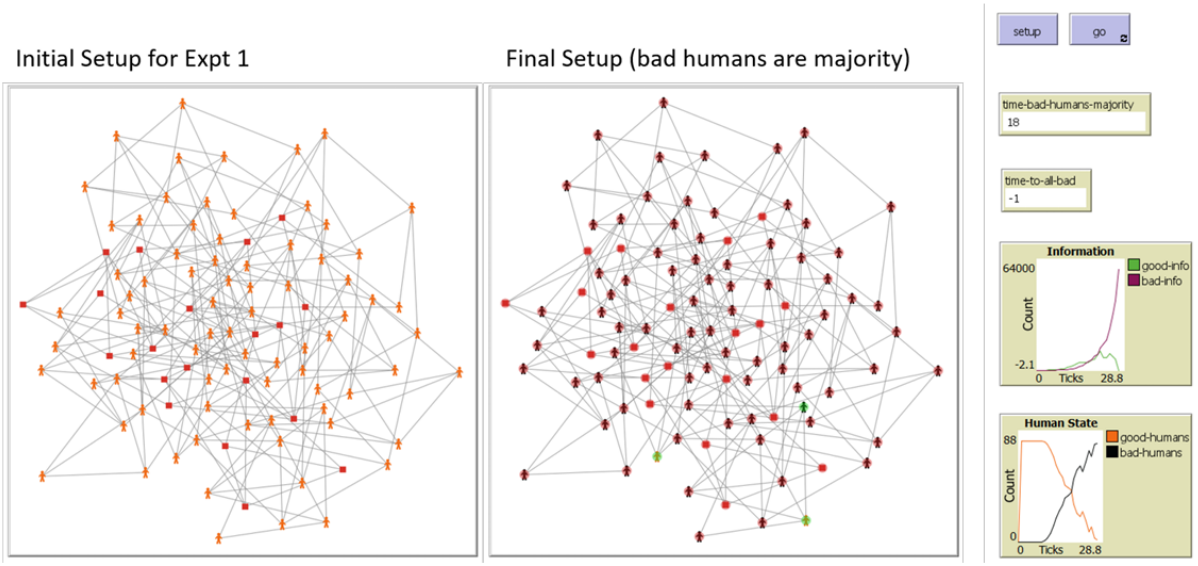

Figure 1: Setup of the simulation model. In the initial setup, Humans are visualized with orange persons, and Bad Bots as red squares. In final setup, Bad Humans are colored black, and the green/red circles represent whether the current set of information consumption queue is good/bad.

3.10 The following steps summarizes the main routine:

1. The model is set up with a Small World network. The network is initialized with $n_h = 1000$ Human agents and a number of Bad Bot agents $= \alpha_1 \times$(Humans). This network is static over time, i.e., connections do not change over time. Links between agents A_1 and A_2 means that A_1 can send and receive information with A_2 and vice versa.

2. The routine then enters a cycle. This cycle consists of three stages that repeat until a termination condition is met.

   (a) The Information Generation stage generates two types of information: Good Info and Bad Info. All agents generate Info (information) with a probability of P_g. Whether the agent generates Good or Bad Info depends on the state (i.e., Good, Bad) of the agent. Bots generate 2 pieces of Info while Humans generate 1 piece of Info.

   (b) Next, we have the Information Consumption Stage, where agents consume the Info that they received in the previous tick, each with a probability of P_c. A Human consumes all Info, while Bots just hold the Info in their memory space. The consumption of Info represents the readership and processing of the information. Bots do not consume Info because they are pre-programmed with their state, and therefore are not influenced by incoming information.

   (c) The third stage is the Information Propagation stage. Each agent (the ego) sieves through the messages they received in the previous tick to decide which to pass on to the alters in their ego network. Alters are neighboring agents attached to the ego through a direct link in the network. If the ego is a Good Human, all Info are selected for propagation. If the ego is a Bad Human, only Bad Info are selected. If the ego is a Good Bot or Info-Correction Bot, only Good Info are selected. The ego then propagates each piece to its alters with a probability of P_p.

   (d) The Status Update stage manages state transitions for Humans. These transitions happen when the amount of information the Human agent consumed hits a threshold. The threshold of Human agent state change is currently set to $t = 72$ which is the average number of exposures to a piece of information a human need to form an opinion (i.e., belief). Bot agents are static (i.e., programmed with their belief) and do not experience state changes.

(e) The final stage in this cycle checks the conditions of the network, which tracks information consumption metrics by doing a count of good/bad information consumed per Human agent, and monitors the outcome measures to determine if the simulation should be terminated. The simulation terminates either when all Human agents turn into Bad Humans, or is stopped at 100 time ticks.

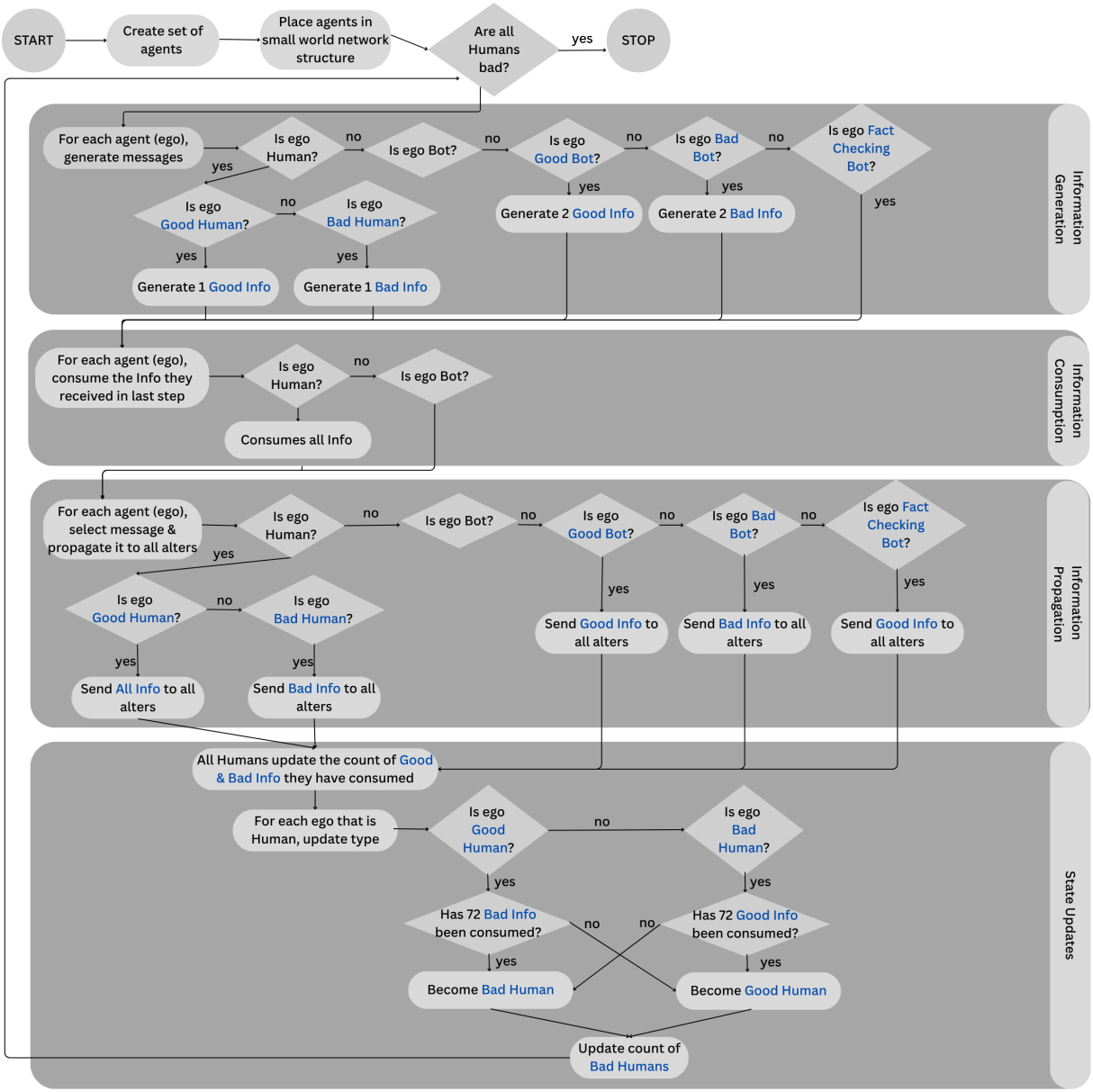

Figure 2: Flowchart of Simulation Logic.

Outcome measures

3.11 We measure two outcomes in this model: (1) `Bad Humans Majority`, which is the time required for the number of Bad Humans to be greater than the number of Good Humans; and (2) `All Bad Humans`, which is the time when all Humans are converted to Bad Humans.

3.12 The simulation's stopping criterion is when `All Bad Humans` is a valid positive value. This positive value measures the number of ticks where all Human agents are Bad Humans. This is akin to the formation of a conspiratorial society, where all humans spread conspiracies. If the simulation does not converge, we stop the simulation at 100 ticks. We set this simulation at 100 ticks based on preliminary sensitivity analysis and observations of the trend lines of the agent states' count. In these preliminary analyses, most of the simulated experiments either

converged to a stable state or demonstrated clear directional trends by 100 ticks. Therefore, we use the 100 tick mark as a reasonable proxy for long-term outcomes, to balance computational efficiency with sufficient observation time to capture the emergent population behavior. A higher `Bad Humans Majority` measure is more desired, because it means the society takes a longer time for Bad Humans to dominate the conversation.

Model parameters

3.13 Table 1 reflects the key parameters used in our simulation model. These values were chosen on the basis of a combination of theoretical considerations and practical limitations. The model implementation relied on some stylized facts extracted from real-world data to provide theoretical validity to our model. These stylized facts and their implementation in our model are listed in Table 2.

Parameter	Definition	Default Value	Range
n_h	Number of Human agents	100	100
p	Probability of formation of links between two agents in the starting network	0.05	0.05
α_1	Proportion of Humans that are Bad Bot agents	0.2	[0.1,1.0]
α_2	Proportion of Bad Bots that are Info-Correction Agents	0	[0.1, 1.0]
α_3	Proportion of Bad Bots that are Good Bots	0	[0.1, 1.0]
P_g	Probability of generating information	0.4	0.4
P_c	Probability of an information being consumed	0.8	0.8
P_p	Probability of an information being propagated	0.8	0.8
t	Threshold of Human state change	72	72

Table 1: Key Parameters

3.14 Our implementation mechanics of the number of Info that bots post is based on large-scale year long empirical analyses on bot and human behavior. In data from the 2020 coronavirus pandemic, bots generated at least two times more posts than humans (Ng & Carley 2025b). In data from the 2020 US Presidential Elections, bots generated about 2.6-3 times more posts than human users (Luceri et al. 2020). In our simulation, we use the lower bound value where bots generate at least twice more posts than humans, to evaluate the scenario where bots are less aggressive.

3.15 Further, conspiracy theories often have a kernel of truth that malicious bots distort (Keeley 1999). This is reflected in the mechanic where Bad Bots convert Good Info into Bad Info. Confirmation Bias makes people more likely to spread information that aligns with their own beliefs (Hart et al. 2009), thus Bad Humans are more likely to propagate bad Info. People share information based on desirable conclusions, such as garnering attention and inducing humor rather, than accuracy (Moravec et al. 2019), inspiring the behavior where Good Humans propagate both Good and Bad Info. Such behavior stems from the motivated reasoning bias (Moravec et al. 2019; Kunda 1990). The same behavior is also based on observational studies where people are ignorant of the validity of the information (Buchanan 2020).

3.16 Finally, we operationalized the flip threshold $t = 72$. Empirical human-subjects studies reveal that about 4.94 to 138.59 exposures to information on social media (i.e., text, image, video) can cause them to form a belief towards a topic. (Asher et al. 2018). We align our flip threshold t with the mid range of the estimate from the empirical study.

Implementation within Simulation	Stylized Fact from Literature	Reference
Small-world network for conspiracy theory spread on social media	Religious ideologies and conspiracies on Twitter surrounding the 2015 Boston bombing incident self-organizes into a small-world phenomenon	Ch'ng (2015)
Link probability of a small-world network is $p = 0.05$	Literature on modeling information dissemination and of rumor propagation constructs and measures small world networks with $p = 0.05$	Fu & Liao (2012) Zanette (2002)
Bots generate 2x more Info than Humans	Bots post at least 2x more tweets than humans Bots shared at least 2.6-3x the number of tweets than human users.	Ng & Carley (2025b) Luceri et al. (2020)
Consumption rules: consuming about 70 pieces of information to flip states	Opinion formation threshold estimates: Empirical survey work show that individuals require about 4.94-138.59 exposures towards an item (text, image, video) before they form a stable opinion on a topic.	Asher et al. (2018)
Asymmetric propagation rules: Good Humans propagate both good and bad Info, Bad Humans propagate only bad Info	Motivated reasoning & confirmation bias: People with stronger conspiracy beliefs are more uniform views People with weaker conspiracy beliefs share more diverse views	Pennycook & Rand (2021) Guess et al. (2019)
Bad Bots convert good Info into bad Info	Malicious bots distort facts to form disinformation	Giachanou et al. (2022)
Bad Bots convert good Info into bad Info	Conspiracy theories incorporate factual elements to enhance plausibility	Keeley (1999)
Bad Humans are more likely to propagate bad Info	Confirmation Bias: People are more likely to spread information that aligns with their own beliefs	Hart et al. (2009)
Good Humans share both good and bad Info.	Motivated Reasoning: People share information based on desirable conclusions rather than accuracy People cannot tell the validity of conspiracies.	Moravec et al. (2019) Buchanan (2020)
Bad Humans generate and propagate bad Info only	Monological belief system: Belief in one conspiracy theory predicts belief in others	Goertzel (1994)

Table 2: Stylized Facts used in the Implementation.

Agent Mechanics per Timestep

4.1 The agent activation routine is a step-based advancement. We use the term "tick" to represent a time step in the simulation, where the three cyclical stages take place. After each tick, the simulation moves to the next round.

4.2 Within each tick, each agent has an activation routine that reflects its persona. For Human agents, there are two types: Good Humans and Bad Humans. Good Humans represent the generic network user; Bad Humans represent users who deliberately spread disinformation like conspiracy theorists. Humans generate 1 piece of Info per tick. Good Humans generate Good Info, Bad Humans generate Bad Info. Humans consume all Info. Good Humans propagate both Good and Bad Info, while Bad Humans propagate only bad Info. Humans change their state from good to bad or bad to good when the cumulative amount of Info they consume crosses a set threshold t. For all experiments, we use $t = 72$, which is the average number of exposures that individuals require to form a stable opinion about a topic (Asher et al. 2018).

4.3 In our model, the design choice where Good Humans propagate both good and bad information and Bad Humans propagate only bad information reflects documented asymmetries in information sharing behavior. This leans on research that individuals with stronger conspiracy beliefs demonstrate greater selectivity in information sharing behaviors (Pennycook & Rand 2021), and those with weaker conspiracy beliefs are more likely to share diverse content (Guess et al. 2019).

4.4 For Bot agents, there are three types: Bad Bots, Info-Correction Bots, and Good Bots. Bad Bots generate Bad Info, Good Bots generate Good Info. Bots generate Info at twice the rate of Humans. Bad Bots mimic malicious users that generate conspiracy theories and selectively propagate only conspiracy theories. Info-Correction Bots mimic the reactions of fact-checking organizations when they come across Bad Info: change them into Good Info (i.e., fact-check) and putting out only Good Info (Chang et al. 2021). Good Bots mimic the mechanics where organizations put out good messaging or advertisements instead of directly engaging with the malicious messages (van der Linden et al. 2017). Collectively, Info-Correction and Good Bots are useful bots.

Virtual Experiments

5.1 We ran five different virtual experiments, each designed to systematically explore different aspects of the effectiveness of useful bot interventions. The experiments can be grouped into two groups: single-parameter variation studies (Experiments 1 to 3) and dual-parameter sweeps (Experiments 4-5). In all the runs, we kept the probabilities used in information management (P_c, P_g, P_p), the number of Human agents ($n_h = 1000$), and the Human state-change threshold ($t = 72$) remains the same. This ensures that the observed differences in outcomes can be attributed to the different bot deployment strategies. All experiments were performed with Netlogo's BehaviorSpace function. Each parameter combination is performed $n = 15$ times, and the mean time of the $n = 15$ rounds is reported. We ran a total of 3,675 simulation runs across all experiments. Table 3 summarizes the virtual experiments.

5.2 Experiment 1 serves as a baseline experiment, systematically varying only the proportion of Bad Bots (α_1) from 0.1 to 1.0 in increments of 0.1, while maintaining zero useful bots in the system ($\alpha_2 = \alpha_3 = 0$). This is represented in our Virtual Experiment table as [0.1, 1.0, 0.1]. That means that we tested the scenarios where the number of Bad Bots are 10% of the number of Human agents through the scenario where the number of Bad Bots are 100% the number of Human agents, each time increasing the percentage of Bad Bots by another 10%. This experiment enables us to establish the fundamental relationship between malicious bot presence and the conspiratorial society formation in the absence of any countermeasures from useful bots. Experiment 2 introduces Info-Correction Bots as the sole intervention mechanism, while maintaining a fixed Bad Bot proportion ($\alpha_1 = 0.2$). This experiment varies the Info-Correction Bot ratio (α_2) from 0.1 to 1.5 in increments of 0.1. The parameter range extends to $\alpha_2 = 1.5$ to explore the scenario where the Info-Correction Bots substantially outnumber Bad Bots. Similarly, Experiment 3 tests Good Bots in isolation by maintaining the Bad Bot proportion ($\alpha_1 = 0.2$) while varying the Good Bot ratio (α_3) from 0.1 to 2.0 in increments of 0.1. The extended range to $\alpha_3 = 2.0$ tests whether proactive good messaging requires different optimal deployment ratios compared to reactive information correction. The design for Experiments 2 and 3 isolates the effectiveness of information correction and good messaging mechanisms respectively without confounding effects from other intervention types.

5.3 Experiments 4 and 5 uses dual-parameter sweeps to map interaction effects between Bad Bots and each useful bot type. Experiment 4 explores the Bad Bot: Info-Correction Bot parameter space. It varies α_1 and α_2 from 0.1 to 1.0, creating $10 \times 10 = 100$ unique parameter combinations. Experiment 5 explores the Bad Bot: Good Bot parameter space, varying α_1 from 0.1 to 2.0 and α_3 from 0.1 to 1.0. These two experiments use parameter sweeps to enable analysis of optimal deployment ratios of useful bots.

5.4 To determine the number of replications required per experimental condition, we conducted a preliminary power analysis using one-way ANOVA models in Python. For each dependent variable (`Bad Human Majority`, `All Bad Humans`), we fitted a three Ordinary Least Square models with the independent variables $\alpha_1, \alpha_2, \alpha_3$ (the Bad Bots, Info-Correction and Good Bots as a proportion of the number of Humans respectively). From the ANOVA results, we computed the effect size (η^2) and converted it to Cohen's f, then used the `FTestAnovaPower` function from `statsmodels` in Python to estimate the required sample size per group for 80% power at $p = 0.05$ level. Table 4 presents the power analysis for the three independent variables. On average, the power analysis indicated that only $n = 2$ replications per condition were sufficient to detect the observed effects; however, to mitigate stochastic variability inherent in ABM simulations, we conservatively ran $n = 15$ replications for each condition. In total, we ran 3,675 experiments.

Expt No./ Params	**1**	**2**	**3**	**4**	**5**
Description	Vary Bad Bot Proportion	Vary Info-Correction Bot Proportion	Vary Good Bot Proportion	Parameter Sweep (Bad Bot: Info-Correction Bot)	Parameter Sweep (Bad Bot: Good Bot)
Control Variables					
n_h, Number of Human Agents	1000	1000	1000	1000	1000
P_g, Probability of an agent generating information	0.4	0.4	0.4	0.4	0.4
P_c, Probability of information being consumed	0.8	0.8	0.8	0.8	0.8
P_p, Probability of information being propagated	0.8	0.8	0.8	0.8	0.8
t, threshold of Human state change	72	72	72	72	72
Independent Variables					
α_1 (Bad Bots as a proportion of Humans) [min, max, step]	$[0.1, 1.0, 0.1]$	0.2	0.2	$[0.1, 1.0, 0.1]$	$[0.1, 2.0, 0.1]$
α_2 (Info-Correction Bots as a proportion of Humans) [min, max, step]	0	$[0.1, 1.5, 0.1]$	0	$[0.1, 1.0, 0.1]$	0
α_3 (Good Bots as a proportion of Humans) [min, max, step]	0	0	$[0.1, 2.0, 0.1]$	0	$[0.1, 1.0, 0.1]$
# Replications	15	15	15	15	15
# Conditions	10	15	20	10*10 = 100	10*10 = 100
Total Runs	150	225	300	1500	1500
Total # experiments	3675				
Dependent Variables (Results)					
`Bad Human Majority` (Mean ticks)	13.51 ± 1.07	17.85 ± 2.27	18.52 ± 4.86 DNC when $\alpha_3 \geq 1.6$	17.73 ± 3.62	17.55 ± 3.32
`All Bad Humans` (Mean ticks)	22.82 ± 2.26	24.73 ± 1.18	27.06 ± 2.84	DNC	DNC

Table 3: Summary of Virtual Experiments. DNC means that the run does not converge and was terminated at 100 ticks.

Dependent Variable	η^2	Cohen's f	Runs/ Condition for 80% power, $\alpha = 0.05$
Vary Bad Bot (α_1) Proportion			
`All Bad Humans`	0.85	2.39	1.96
`Bad Human Majority`	0.83	2.22	2.09
Vary Info-Correction Bot (α_2) Proportion			
`All Bad Humans`	0.89	2.85	1.74
`Bad Human Majority`	0.87	2.57	1.86
Vary Good Bot (α_3) Proportion			
`All Bad Humans`	0.84	2.28	2.02
`Bad Human Majority`	0.81	2.06	2.21

Table 4: Power Analysis. Statistically, just two replications per condition would give 80% power to the experiment at $p = 0.05$.

Results

Varying one bot type

6.1 Experiments 1, 2 and 3 vary only one bot type (Bad Bot, Info-Correction Bot, Good Bot), enabling isolation of individual bot effects. Figure 3 shows the mean time to `Bad Human Majority` in these three cases. This figure reveals distinct intervention effectiveness patterns: a stable baseline with rapid conspiracy formation (Expt 1), dramatically improved resistance with Info-Correction bot interventions (Expt 2), and the emergence of threshold effects where sufficient Good Bot deployment can prevent a conspiratorial majority entirely. The curves plotted in Figure 3 are not entirely smooth and straight because there is still a level of stochasticity in our model, i.e. the probability of generation, consumption and propagation of information, attachment to other agents during network formation. This stochasticity reflects the real-life probability distribution of the information interaction dynamics, rather than a constant interaction factor.

6.2 Experiment 1 establishes the baseline threat, where Bad Bots consistently drive conspiratorial society formation regardless of their proportion. This observation confirms that, when there are only Bad Bots in the society, without countermeasures, conspiratorial dominance is inevitable (mean time to `Bad Human Majority`: 13.51 ± 1.07 ticks). The society will eventually converge to a state where all Humans are bad (mean time to `All Bad Humans`: 22.82±2.26 ticks). Even small proportions of Bad Bots can reliably transform the information environment over time.

6.3 Both the injection of Info-Correction Bots (Experiment 2) and Good Bots (Experiment 3) can improve resistance to conspiracy formation, which can be seen from the graphs where more time steps are required for `Bad Human Majority` with the introduction of Good Bots (18.52 ± 4.86 ticks) or Info-Correction Bots (17.85 ± 2.27 ticks), as compared to when there were only Bad Bots (13.51 ± 1.07 ticks). While the introduction of useful bots delay the formation of a conspiratorial society, a conspiratorial society still forms after a sufficiently long period of time. Good Bots demonstrate an additional capability over Info-Correction Bots. While Info-Correction Bots will always reach the `All Bad Humans` state, when $\alpha_3 \geq 1.6$, Good Bots prevent Bad Humans from reaching majority status. This suggests that proactive messaging may be more robust than reactive corrective strategies.

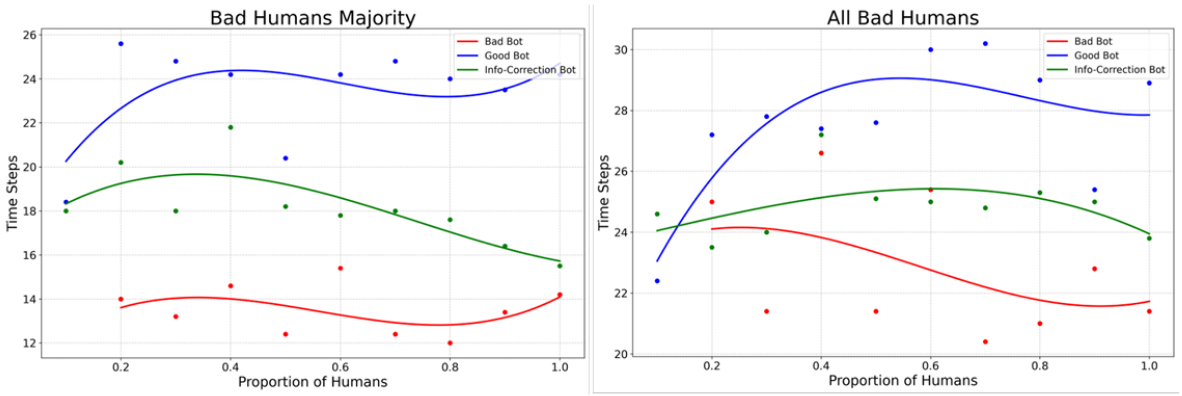

Figure 3: Mean time to `Bad Humans Majority` and `All Bad Humans` from singularly varying the proportion of Bad Bots, Info-Correction Bots and Good Bots.

Varying two bot types

6.4 Experiments 4 and 5 are parameter sweep runs that explore the behavior space between the parameters of Bad Bots vs. Info-Correction and Bad Bots vs. Good Bots. These two experiments compare the effectiveness of using Info-Correction and Good Bots in combating conspiracy theories. The results from these experiments tell us the optimal ratios between malicious and useful bots that maximize the time required for Bad Humans to dominate the conversation.

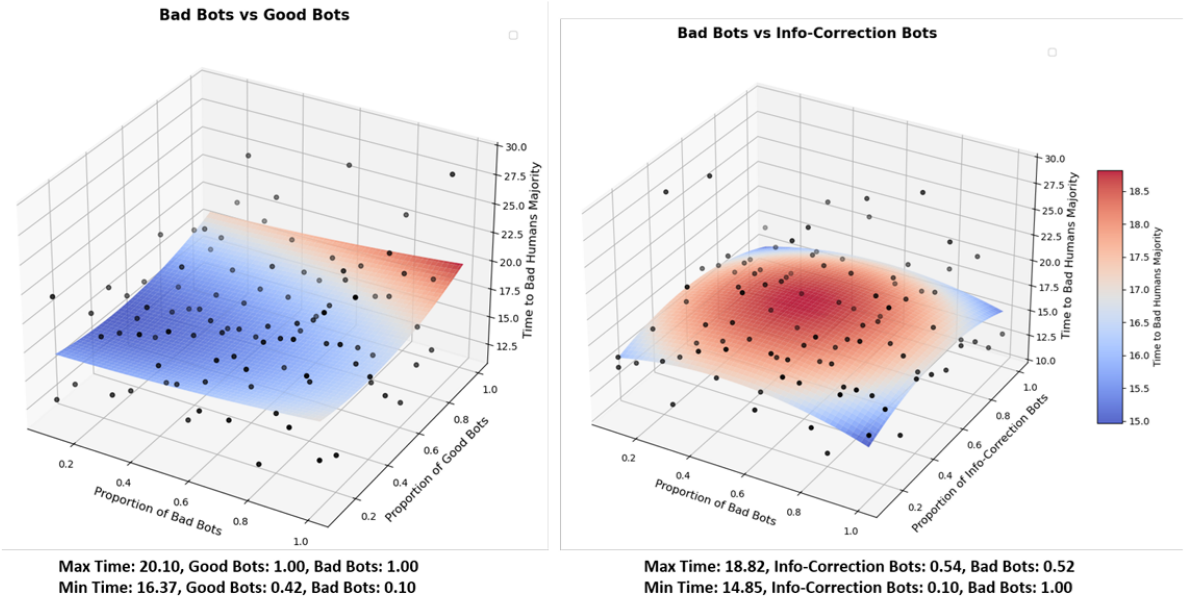

Figure 4: Response Surface Analysis for varying Bad Bots and Good Bots vs Bad Bots and Info-Correction Bots. The z-axis represents time to `Bad Human Majority` in simulation ticks. The Info-Correction surface exhibits strong concavity ($2\beta_5 = -18.6$), indicating diminishing returns with more Info-Correction Bots. The Good Bot surface is almost linear ($2\beta_5 = +0.11$), indicating increasing benefits with increased number of Good Bots.

6.5 We present the parameter sweep results as 3D response surface analysis in Figure 4. We use a quadratic function as a smoothing function. While the underlying data points are discrete, the surface interpolation reveals

continuous contours of influence across the behavior space, like the regions of peak effectiveness of each type of bot and surface topologies, which provides insights into optimal useful bot deployment strategies.

6.6 For both surfaces, we calculated the defender efficiency functions as a quadratic response surface, to capture how the variations in defender proportions (i.e., proportion of Good Bot, proportion of Info-Correction Bot) influence the time required for a `Bad Human Majority` scenario to emerge. These equations with their estimated coefficients are expressed in Equations 1 and 2.

$$\text{Good Bots:} \quad T(b,d) = 17.222 - 38.906b + 0.410d - 10.406bd + 503.362b^2 + 0.0538d^2, \tag{1}$$

$$\text{Info-Correction Bots:} \quad T(b,d) = 14.459 + 7.511b + 9.060d + 1.671bd - 8.098b^2 - 9.313d^2. \tag{2}$$

6.7 Equations 1 and 2 show the fitted defender efficiency functions for good bots and info-correction bots. $T(b,d)$ represents the time to bad-human majority, b is the proportion of bad bots, and d is the proportion of defender bots (either good or info-correction). Coefficients are estimated from quadratic response surface regressions of data from Experiments 4 and 5.

6.8 The Info-Correction surface exhibits strong concavity ($2\beta_5 = -18.6$), and so forms a dome that peaks near moderate densities of Info-Correction bots. This means that there is diminishing returns to the information environment once the Info-Correction Bots saturate the system. In contrast, the Good Bots surface has a mild convexity that is close to linear ($2\beta_5 = +0.11$). This implies that Good Bot interventions scale predictably: as the number of Good Bots increase, the time delay to `Max Human Majority` increases.

6.9 Across both surfaces, the maximum and minimum points further illustrate the defender efficiency contrast. For Info-Correction Bots, the fitted surface peaks at approximately $T = 18.82$ (Bad Bots$= 0.52$, Info-Correction Bots$= 0.54$), and falls to a minimum of $T = 14.85$ (Bad Bots$= 1.00$, Info-Correction Bots$= 0.10$), indicating that Info-Correction bots are most effective only at moderate densities ($\alpha_2 = 0.54$) and lose influence as the system becomes saturated with Info-Correction Bots. In contrast, the Good Bot surface reaches a higher maximum of $T = 20.08$ (Bad Bots$= 1.00$, Good Bots$= 1.00$). The surface also has a higher minimum of $T = 16.37$ (Bad Bots$= 0.10$, Good Bots$= 0.42$). The elevated minimum and maximum suggest that Good Bot defenses are more consistently effective across the range of agent proportions.

6.10 The response surface curves show that Good Bots are generally better than Info-Correction Bots. Good Bot interventions scale with the number of Good Bots employed, while Info-Correction Bots will hit a limit, after which there is diminishing returns. This phenomenon could be because Good Bots do not have the overhead of information verification. Finally, neither of the two experiments converges, i.e. reach a `All Bad Humans` state, indicating that the presence of useful bots prevents a conspiratorial society.

Statistical analysis

6.11 With the results of varying one bot type, we also performed a two-way ANOVA to test the interaction of bot type and proportion of bot type. Bot type is represented as a discrete variable of three factors: Good Bot, Bad Bot, Info-Correction Bot. Proportion is described as a continuous variable. The ANOVA equation is presented in Equation 3, and one equation was ran for each outcome measure (`All Bad Humans` and `Bad Human Majority`).

$$\hat{Y}_{\text{outcome measure}} \sim C(\text{Bot Type}_i) * \alpha_i \tag{3}$$

6.12 Equation 3 shows the two-way ANOVA to test the interaction between bot type and proportion. Outcome measure refers to `All Bad Humans` and `Bad Human Majority`. α_i is the number of bots of type$_i$ as a proportion of the number of humans.

6.13 The results of the variance analysis using ANOVA tests are presented in Table 5. For both outcome measures (`All Bad Humans`, `Bad Human Majority`), there is a strong main effect of bot type ($F = 16.03, p < 0.001$ and $F = 80.49, p < 0.001$ respectively). This indicates that the different types of bots used can affect how quickly the simulation converges towards a conspiratorial society state. In contrast, the main effect of bot proportion was not significant (at the $p < 0.05$ level) in both models, suggesting that simply increasing the number of bots without regard to their type does not reliably increase convergence time. There is marginal interaction between bot type and its proportion ($F = 2.91, p = 0.075$ for `All Bad Humans`; $F = 2.72, p = 0.087$ for `Bad Human Majority`), indicating that the influence of bot quantity may differ slightly depending on the role that the bots play within the information ecosystem (Keller & Klinger 2019; Deb et al. 2018).

Dependent Variable: `All Bad Humans`				
	sum squared	**degree of freedom**	**F**	**Pr(>F)**
C(bot type)	109.79	2.0	16.02	4.40E-5***
C(proportion)	0.46	1.0	1.36E-1	7.16E-1
C(bot type):C(proportion)	19.83	2.0	2.91	7.49E-1
Dependent Variable: `Bad Human Majority`				
	sum squared	**degree of freedom**	**F**	**Pr(>F)**
C(bot type)	460.92	2.0	80.49	4.12E-11***
C(proportion)	1.10	1.0	3.78E-1	5.45E-1
C(bot type):C(proportion)	15.77	2.0	2.72	8.69E-2

Table 5: Two- way ANOVA analysis of proportion of each Bot Type and interaction of Bot Type and proportion for dependent variables. *** indicates significant at the $p < 0.001$ level.

6.14 Next, with the overall results, we performed an Ordinary Least Squares (OLS) regression to understand the effect sizes of the proportion of each type of bot on the outcome measures. This regression formula is presented in Equation 4. We ran one equation for each outcome measure (`All Bad Humans` and `Bad Human Majority`).

$$\begin{aligned}\hat{Y}_{\text{Outcome Measure}} \sim\ & \beta_0 \\ & + \beta_1(\text{proportion of Bad Bots}) + \beta_2(\text{proportion of Good Bots}) \\ & + \beta_3(\text{proportion of Info-Correction Bots}) \\ & + \beta_4 C(\text{Bad Bots Present})) + \beta_5 C((\text{Good Bots Present})) + \beta_6 C(\text{Info-Correction Bots Present}) \\ & + \beta_7(\text{proportion of Bad Bots} \times C(\text{Good Bots Present})) \\ & + \beta_8(\text{proportion of Bad Bots} \times C(\text{Info-Correction Bots Present})) \\ & + \beta_9(\text{proportion of Bad Bots} \times \text{proportion of Good Bots}) \\ & + \beta_{10}(\text{proportion of Bad Bots} \times \text{proportion of Info-Correction Bots}) + \varepsilon\end{aligned} \tag{4}$$

6.15 Equation 4 reports the Ordinary Least Squares regression model for predicting the outcome measures based on the starting proportion of agents. Outcome measure refers to `All Bad Humans` and `Bad Human Majority`.

6.16 The results of the OLS regression are presented in Table 6. The first model for the dependent variable `All Bad Humans` has $R^2 = 0.198$. While the model's overall explanatory power is modest, several interaction effects are significant. The presence of Bad Bots interacting with Good Bots ($\beta = 20.47, p = 0.045$) and Bad Bots interacting with Info-Correction Bots ($\beta = -8.09, p = 0.003$) show that useful bot mechanisms can substantially influence the convergence time of a conspiratorial society. The second model for explaining `Bad Human Majority` has $R^2 = 0.511$, which shows considerably stronger explanatory power. Here, the interaction between Bad and Good Bots ($\beta = 42.31, p < 0.001$) emerges as the most influential predictor, suggesting that the presence of Good Bots will dramatically reduce the formation of a conspiratorial society. Together, the two models reveal a non-linear influence architecture of each type of the bot on the information ecosystem. Rather, it is the interaction of Bad Bots with useful bots that yield emergent effects on the formation, or intervention of, a conspiratorial society.

Dependent Variable: `All Bad Humans`			
$R^2 = 0.198$ Adjusted $R^2 = 0.134$ $F(9, 112) = 3.08, p = 0.002$			
	Coef.	**Std. Err.**	**P-value**
Intercept	25.06	2.14	0.000***
Good Bots Present	-1.30	2.74	0.637
Info-Correction Bots Present	-1.64	2.39	0.495
Proportion of Bad Bots	-3.73	3.27	0.256
Proportion of Good Bots	-0.19	0.08	0.011**
Proportion of Info-Correction Bots	4.30	1.64	0.010**
Bad Bots × Good Bots Present	20.47	10.07	0.045**
Bad Bots × Info-Correction Bots Present	6.73	3.71	0.073
Bad Bots × Proportion of Good Bots	0.07	0.05	0.205
Bad Bots × Proportion of Info-Correction Bots	-8.09	2.70	0.003**
Dependent Variable: `Bad Human Majority`			
$R^2 = 0.511$ Adjusted $R^2 = 0.471$ $F(9, 112) = 12.99, p = 5.42E - 14$			
	Coef.	**Std. Err.**	**P-value**
Intercept	13.89	1.84	0.000***
Good Bots Present	0.44	2.36	0.854
Info-Correction Bots Present	1.62	2.06	0.433
Proportion of Bad Bots	-0.63	2.82	0.823
Proportion of Good Bots	-0.46	0.07	0.000***
Proportion of Info-Correction Bots	1.53	1.41	0.281
Bad Bots × Good Bots Present	42.31	8.68	0.000***
Bad Bots × Info-Correction Bots Present	-0.25	3.20	0.937
Bad Bots × Proportion of Good Bots	0.03	0.05	0.480
Bad Bots × Proportion of Info-Correction Bots	-1.42	2.33	0.545

Table 6: Coefficients of the Ordinary Least Squares regression models predicting convergence outcomes. The upper panel corresponds to `All Bad Humans`, and the lower panel to `Bad Humans Majority`. ** indicates significance at $p < 0.05$ and *** at $p < 0.001$. Interaction terms represent moderation effects between the proportions or presences of different bot types.

Validation

6.17 Our results are validated through stylized facts that are used to determine known phenomena, providing pattern and theoretical validity to our model. These stylized facts are adapted from social science and social psychology theories, and are presented in Table 7.

6.18 In our simulation, Bad Humans generate and propagate Bad Info only. This is based on the concept of the monological belief system, where the belief in one conspiracy theory predicts and predates the belief in other conspiracy theories (Goertzel 1994). The same stylized fact also reflects how, when a simulation reaches `Bad Human Majority` state, it does not fall back to a state where Bad Humans are the minority. The information propagation rules in our simulation are asymmetric: Good Humans propagate Good and Bad information, while Bad Humans propagate only Bad information. This is motivated by sociological theories where people with stronger conspiracy beliefs share more uniform views, and those with weaker conspiracy beliefs share more diverse views (Pennycook & Rand 2021; Guess et al. 2019).

6.19 In a social network, the combination of the pressure of social influence and the formation of echo chambers maintains coherence among conspiracy theories, which makes turning Bad Humans back to Good Humans difficult. Finally, our model reaches the `Bad Human Majority` state regardless of the presence of useful bots because disinformation spreads faster than truth (Vosoughi et al. 2018).

Reference	Stylized Fact	Simulation Result
Goertzel (1994) Törnberg (2018)	Belief in one conspiracy theory predicts belief in others Formation of echo chambers as a complex contagion	After the simulation reaches `Bad Human Majority` state, it does not fall back to a state where Bad Humans are minority
Vosoughi et al. (2018)	Disinformation spread faster than truth	Simulation reaches `Bad Human Majority` state regardless of presence of useful bots
Lewandowsky & Van Der Linden (2021)	Proactive inoculation is significantly more effective than reactive corrections	Good Bots are more effective than Info-Correction Bots

Table 7: Stylized Facts used for Validation.

Discussion

7.1 This work simulates the information distribution mechanics in a social network, demonstrating the effects of malicious Bad Bots on the state of Humans, and the usefulness of Info-Correction and Good Bots on preventing an entirely conspiratorial society (which is represented in the simulation as all Humans being Bad). Our simulation design distinguishes itself from previous models of information propagation (Gurung et al. 2025; Gomez-Rodriguez et al. 2013) in three ways. First, most existing simulations examining conspiracies (Blane et al. 2021; Gurung et al. 2025) and countermeasures (Murdock et al. 2025; Addai et al. 2025) do not consider the presence of bots. We explicitly model different types of bots with distinct behavioral rules, reflecting our claim that the social media space can contain both malicious and beneficial automated agents (Assenmacher et al. 2020; Ng et al. 2024a). Second, the state-change mechanism for Human agents is based on cumulative information exposure, capturing the gradual nature of the process of conspiracy belief formation (Prooijen 2018; van Prooijen & van Vugt 2018). Third, we incorporate asymmetric information propagation rules that reflect empirical findings about selective sharing behaviors, and emphasize the asymmetric reality of social media-based warfare (Tucker et al. 2018).

7.2 Experiment 1 shows that even a small ratio (α_1=0.1) is sufficient for the formation of a conspiratorial society, where all humans are bad. This mirrors past work that shows that about 20% of bots are required to push a society towards the tipping point where the majority of opinions align with the bots' opinions (Carragher et al. 2023). This small ratio reflects the danger of the presence of Bad Bots spreading harmful information.

7.3 In all replications of Experiment 1, the simulation will eventually reach a state where Bad Humans are the majority. That means that if the number of Bad Bots is left unchecked, a society will always result in a larger number of Bad Humans than Good Humans. This stresses the importance of the regulation of malicious agents on social media. One way of combating malicious agents is through mass banning. Platforms like X have been continually taking down malicious agents (X 2021). However, such large-scale take downs are limited to the social media companies, and with unknown frequencies. What can citizen-based, non-profits and government organizations do to reduce the impact of malicious messaging on the public?

7.4 One common method that is being deployed is media literacy training, where governments and organizations try to teach users to spot Bad Info. The corollary of this intervention in our simulation is to vary the Info Threshold value. If Humans are more resilient to information injects, their Info Threshold value increases. Figure 5 varies the Info Threshold from $t = 10$ to $t = 100$ with steps of $t = 10$. Unfortunately, the relationship between Info Threshold and the time for `Bad Humans Majority` is non-linear. Initially, modest increases in threshold produce mid range values, but there eventually is a tipping point of $t = 74$, beyond which, higher thresholds do not necessarily yield a more resilient society. In fact, overly resilient agents slow the overall information flow in the system. The inverted U-shape curve is also consistent with findings that overly skeptical audiences can disengage with corrective information (Nyhan & Reifler 2019; Lewandowsky et al. 2012), and audiences that have already believed in conspiracy theories will be harder to be persuaded to change their beliefs.

7.5 In real life, not all humans can be trusted to achieve great information literacy, and from our experiments, changing the Info Threshold is not a guaranteed solution, so there must be a consideration for other types of interventions. Our subsequent experiments 2-5 demonstrate the feasibility of interventions using useful bots.

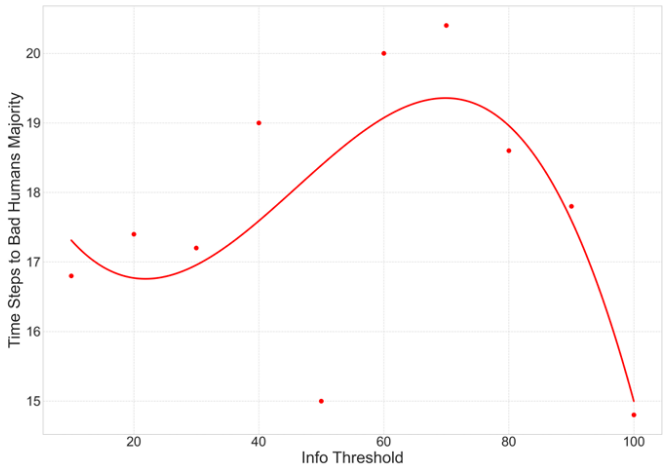

Figure 5: Mean time to `Bad Humans Majority` from varying the Info Threshold from [10,100] with only Humans and Bad Bots ($n_h = 1000$ and $\alpha_1 = 0.2, \alpha_2 = 0, \alpha_3 = 0$).

7.6 Experiments 2 and 3 show that the presence of useful bots (Info-Correction Bots and Good Bots) can combat harmful information by delaying or preventing the formation of conspiratorial societies (i.e., `All Bad Humans` state does not occur). In real life, activities similar to the mechanics of Info-Correction and Good Bots include: performing information corrective activities such as fact-checking or debunking disinformation, or carrying out educational campaigns like the #WashYourHands campaign that ran during the 2020 coronavirus pandemic. These actions can be performed by organizations independent of social media companies, which provides such organizations more power to respond to bad information. Useful bots can prevent a conspiracy society, as observed by the lack of occurrence of `All Bad Humans` state.

7.7 Statistical analyses through ANOVA and OLS both demonstrate that it is the type of bot rather than the proportion of bot that is the key determinant of convergence dynamics. The ANOVA statistical analysis (Table 5) of the interaction between bot type and proportion suggest that the nature of the bot (i.e., Bad Bot, Good Bot, Info-Correction Bot) matters more than the absolute number of bots introduced. Further, the OLS regression analysis (Table 6) emphasize that bot types cannot be viewed in their singularity. It is their interactions with each other that influences the eventual trajectory of the network. These findings reinforce that the success of interventions on harmful information is not merely a function of the scale of the deployed bots but also the functional role of the agents and their network topology.

7.8 Experiments 4 and 5 compare the effect of using Info-Correction Bots against Good Bots. We compare the ratios of malicious to useful bots at the maximum time required for `Bad Humans Majority`. A longer time for Bad Humans to be the majority is more desired, to buy time for authorities to carry out educational campaigns or initiate laws (i.e., Singapore's POFMA that requires authors to addendum to their post, or the United States' FCC, which prohibits false information broadcasting). Our results show that the Good Bots exhibit a higher maximum time, and is therefore more desirable to Info-Correction Bots.

7.9 From the calculations of defender efficiency, Good Bots are more efficient than Info-Correction Bots. As the number of Good Bots increase, the time T to `Bad Human Majority` increases. For Info-Correction Bots, as the number of Info-Correction Bots increases, the time T increases up to a certain point, after which, the time T decreases. This difference in efficiency arises from the different modes of action of the two types of useful Bots. Good Bots generate and propagate Good Info, which adds to the amount of Good Info in the information ecosystem, increasing the amount of Good Info that is likely to be recommended and read by Humans. Info-Correction Bots do not generate information, but instead when hit with Bad Info, will correct them to Good Info. They are therefore limited by the number of Info they receive to perform their tasks, and also consume more resources to perform the fact checking process.

7.10 Cognitive effort theory means that every action or decision makes requires a cognitive process and response (Westbrook & Braver 2015). In a bot, the cognitive effort can be measured by the programming steps required. From a cognitive effort point of view, Good Bots require less effort to construct as compared to Info-Correction Bots. The main action of Good Bots is the creation and dissemination of good information. If a Good Bot sends a message, it takes only three steps: decide that it is time to send a message, craft a message, then send the message. Info-Correction Bots require cognitive power to find mis/disinformation, analyze the information,

fact-check the information, and create a coherent and convincing response against the misinformation. A fact-checking bot takes more programming steps for an Info-Correction Bot to complete the task, and its task involves more decision points which does consume resource, and higher chances where the task will default without completion. It is therefore easier and more predictable to automate the programming of Good Bots than Info-Correction Bots. In fact, currently, information correction tasks are still performed manually through fact-checking organizations and consortia, and it is a time-consuming and costly process (Lee et al. 2023).

7.11 Rather than trying to stop or prevent Bad Bots from disseminating disinformation, evidence-based communication strategies can be used to slow the rate of the formation of a conspiratorial society. Communication strategies should leverage on useful bots (i.e., Info-Communication, Good Bots) to distribute good narratives en-masse. Further, good messaging is more resource effective than information correction, and can be more easily harnessed through automated information distribution with bots. Currently, organizations and governments use information correction and good messaging techniques against bad information. Our simulation urges the use of automated bots to aid these efforts. Although different societal and cultural factors will require a unique strategy to address disinformation, a greater understanding of the process by which a conspiratorial society occurs and the role played by useful bots can help improve public strategies to address this societal concern. Appropriately designed engineering and architecture development of social media platforms can balance the human side of fact-based messaging with the use of these Artificial-Intelligent agents to correct information in real time.

Limitations and future directions

7.12 A limitation of our model is the assumption that agents consume and propagate information from their neighbors with an equal probability. In the real world, the probability differs among agents as a reflection of the strength of ties and biases present in agents. For example, homophily bias indicates that agents might consume more information that originates from other agents with similar demographic characteristics, and authority bias might result in information from authority figures being propagated with higher probability (Ng et al. 2024b). Future work can combine the power of agent-based modeling with network science and calculate different probabilities for the information management stage based on agent properties and network ties. This includes studying the effects of the positions of useful agents, that is, whether placing useful bots strategically in the network mitigates bad information quickly.

7.13 Our model assumes that all information is about the same topic, and that the information propagated by useful bots is counter messages to the Bad Info propagated by Bad Bots. Further extensions of this work include studying the effects of timing in a pool of different topics, and how fast after a Bad Bot propagates Bad Info must useful bots respond to mitigate the conspiratorial effects.

7.14 Lastly, this project tested only two types of bots: Good Bots and Info-Correction Bots. In reality, there are a suite of social bots that could intervene the online space and possibly prevent conspiratorial society from forming (Ng & Carley 2025a; Gorwa & Guilbeault 2020). Future research includes increasing the generalizability and robustness of our conclusions through testing a wider range of agent types.

Conclusion

8.1 Social media bots can be used as vessels of disinformation spread to induce conspiratorial societies, but more importantly, they can be used as mechanisms to prevent the formation of such societies. We construct and present BotSim, which simulates the interaction between Bad Bots that spread bad Info, and useful bots that aim to counter the effects of disinformation spread. Useful bots in the form of Info-Correction Bots turn bad Info to good, correcting the conspiracy theory, while Good Bots simply disseminate good Info.

8.2 Overall, our simulation experiments show that: 1) Bad Bots alone form a conspiratorial society; 2) Good Bots are 7% more resource-effective than Info-Correction Bots for preventing the spread of conspiracy theories; and 3) the type of useful bot and its interactions with Bad Bots are key factors in determining their usefulness as interventions for harmful information.

8.3 This research offers an insight into the interplay between malicious bots and the use of bots. It sets the premise to leverage social media bots for information correction and good messaging. It further shows that good messaging is more resource effective than information correction, and can be more easily harnessed through automated information distribution with bots. We hope our work will increase the appreciation of social media

bot capabilities, and spur discussions within public organizations towards integrating bot techniques in their strategic communications strategies.

8.4 We make two broad contributions. Theoretically, we expand the concept of social media bots from a malicious agent to an automated agent that can be used for both social good and bad. This ties into the concept of a Cyber Social Agent, which encompasses the set of agents that can be used for both good and bad, and especially that the good can come from appropriate use of the same mechanics (Ng & Carley 2025a). Useful agents, such as Information Correction Bots and Good Bots, can be used for information de-bunking and pre-bunking respectively, which eventually prevents a conspiratorial society. Empirically, we demonstrate this concept through the strength of an agent-based model to study how individual agents (humans and/or bots) interact within a network to potentially impact the behavior of larger populations (i.e., the formation of a conspiratorial society). Our BotSim model establishes a foundation for future work for characterizing the impact of bots in the online societal space.

Acknowledgments

This material is based upon work supported by the Scalable Technologies for Social Cybersecurity, U.S. Army (W911NF20D0002), Office of Naval Research (N000142112765), Threat Assessment Techniques for Digital Data, Office of Naval Research (N000142412414), and Community Assessment, Office of Naval Research (N000142412414). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied by the Office of Naval Research, U.S. Army or the U.S. government.

References

Abouzeid, A., Granmo, O.-C., Goodwin, M. & Webersik, C. (2024). Towards misinformation mitigation on social media: novel user activity representation for modeling societal acceptance. *Journal of Computational Social Science*, *7*(1), 741–776

Addai, E., Agarwal, N. & Yousefi, N. (2025). Modeling and simulation of interventions' effect on the spread of toxicity in social media. *Online Social Networks and Media*, *46*, 100309

Aldayel, A. & Magdy, W. (2022). Characterizing the role of bots' in polarized stance on social media. *Social Network Analysis and Mining*, *12*(1), 30

Asher, D. E., Caylor, J., Doyle, C., Neigel, A. R., Korniss, G. & Szymanski, B. K. (2018). Opinion formation threshold estimates from different combinations of social media data-types. Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS 2018)

Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H. & Grimme, C. (2020). Demystifying social bots: On the intelligence of automated social media actors. *Social Media+ Society*, *6*(3), 2056305120939264

Benigni, M. C., Joseph, K. & Carley, K. M. (2018). Bot-ivistm: Assessing information manipulation in social media using network analytics. In N. Dokoohaki, S. Tokdemir & N. Agarwal (Eds.), *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, (pp. 19–42). Berlin Heidelberg: Springer

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G. & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS One*, *10*(2), e0118093

Betts, J. M. & Bliuc, A.-M. (2022). The effect of influencers on societal polarization. 2022 Winter Simulation Conference (WSC)

Blane, J. T., Moffitt, J. & Carley, K. M. (2021). Simulating social-cyber maneuvers to deter disinformation campaigns. International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation

Bleakley, P. (2023). Panic, pizza and mainstreaming the alt-right: A social media analysis of Pizzagate and the rise of the QAnon conspiracy. *Current Sociology*, *71*(3), 509–525

Buchanan, T. (2020). Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *PLoS One*, *15*(10), e0239666

Carley, K. M. (2020). Social cybersecurity: An emerging science. *Computational and Mathematical Organization Theory*, *26*(4), 365–381

Carragher, P., Ng, L. H. X. & Carley, K. M. (2023). Simulation of stance perturbations. International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation

Chang, H.-C. H., Chen, E., Zhang, M., Muric, G. & Ferrara, E. (2021). Social bots and social media manipulation in 2020: The year in review. In U. Engel, A. Quan-Haase, S. Liu & L. E. Lyberg (Eds.), *Handbook of Computational Social Science, Volume 1*, (pp. 304–323). London: Routledge

Ch'ng, E. (2015). Local interactions and the emergence of a Twitter small-world network. arXiv preprint. arXiv:1508.03594

Choudhary, A. & Arora, A. (2021). Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, *169*, 114171

Cisneros-Velarde, P., Oliveira, D. F. & Chan, K. S. (2019). Spread and control of misinformation with heterogeneous agents. Complex Networks X: Proceedings of the 10th Conference on Complex Networks CompleNet 2019 10

Compton, J., van der Linden, S., Cook, J. & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602

Deb, A., Majmundar, A., Seo, S., Matsui, A., Tandon, R., Yan, S., Allem, J.-P. & Ferrara, E. (2018). Social bots for online public health interventions. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Douglas, K. M. & Sutton, R. M. (2018). Why conspiracy theories matter: A social psychological analysis. *European Review of Social Psychology*, *29*(1), 256–298

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S. & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3–35

Dow, B. J., Johnson, A. L., Wang, C. S., Whitson, J. & Menon, T. (2021). The COVID-19 pandemic and the search for structure: Social media and conspiracy theories. *Social and Personality Psychology Compass*, *15*(9), e12636

Dyrendal, A. (2014). Hidden knowledge, hidden powers: Esotericism and conspiracy culture. Contemporary esotericism

Fan, P., Li, P., Wang, H., Jiang, Z. & Li, W. (2012). Opinion interaction network: Opinion dynamics in social networks with heterogeneous relationships. Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics

Featherstone, M. (2000). The obscure politics of conspiracy theory. *The Sociological Review*, *48*(2_suppl), 31–45

Fu, W.-T. & Liao, Q. V. (2012). Information and attitude diffusion in networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, (pp. 205–213). Springer

Giachanou, A., Zhang, X., Barrón-Cedeño, A., Koltsova, O. & Rosso, P. (2022). Online information disorder: Fake news, bots and trolls. *International Journal of Data Science and Analytics*, *13*(4), 265–269

Goertzel, T. (1994). Belief in conspiracy theories. *Political Psychology*, *15*(4), 731–742

Gomez-Rodriguez, M., Leskovec, J. & Schölkopf, B. (2013). Modeling information propagation with survival theory. International conference on machine learning

Gorwa, R. & Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, *12*(2), 225–248

Guess, A., Nagler, J. & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586

Gurung, M. I., Agarwal, N. & Addai, E. (2025). Modeling polarized information diffusion with SEI (A) I (D) Z: A stance-based epidemiological approach. *Social Network Analysis and Mining*, *15*(1), 60

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J. & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555

Hofeditz, L., Ehnis, C., Bunker, D., Brachten, F. & Stieglitz, S. (2019). Meaningful use of social bots? Possible applications in crisis communication during disasters. ECIS

Huang, G., Jia, W. & Yu, W. (2024). Media literacy interventions improve resilience to misinformation: A meta-analytic investigation of overall effect and moderating factors. *Communication Research*, (p. 00936502241288103)

Huang, J. & Jin, X. (2011). Preventing rumor spreading on small-world networks. *Journal of Systems Science and Complexity*, *24*(3), 449–456

Jack, C. (2017). Lexicon of lies: Terms for problematic information. *Data & Society*, *3*(22), 1094–1096

Jin, B. & Guo, W. (2025). Synthetic social media influence experimentation via an agentic reinforcement learning large language model bot. *Journal of Artificial Societies and Social Simulation*, *28*(3), 6

Jin, R. & Liao, Y. (2025). Prevention is better than cure: Exposing the vulnerabilities of social bot detectors with realistic simulations. *Applied Sciences*, *15*(11), 6230

Jolley, D. & Douglas, K. M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLoS One*, *9*(2), e89177

Keeley, B. L. (1999). Journal of philosophy, inc. *The Journal of Philosophy*, *96*(3), 109–126

Keijzer, M. A., Mäs, M. & Flache, A. (2024). Polarization on social media: Micro-level evidence and macro-level implications. *Journal of Artificial Societies and Social Simulation*, *27*(1), 7

Keller, T. R. & Klinger, U. (2019). Social bots in election campaigns: Theoretical, empirical, and methodological implications. *Political Communication*, *36*(1), 171–189

Kumar, K. K. & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-Centric Computing and Information Sciences*, *4*(1), 14

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480

Lee, S., Xiong, A., Seo, H. & Lee, D. (2023). "Fact-checking" fact checkers: A data-driven approach. Harvard Kennedy School Misinformation Review

Leman, P. J. & Cinnirella, M. (2013). Beliefs in conspiracy theories and the need for cognitive closure. *Frontiers in Psychology*, *4*, 378

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N. & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131

Lewandowsky, S. & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European review of social psychology*, *32*(2), 348–384

Lu, H.-C. & Lee, H.-w. (2024). Agents of discord: Modeling the impact of political bots on opinion polarization in social networks. *Social Science Computer Review*, (p. 08944393241270382)

Luceri, L., Cardoso, F. & Giordano, S. (2020). Down the bot hole: Actionable insights from a 1-year analysis of bots activity on twitter. arXiv preprint. arXiv:2010.15820

Medeiros, F. D. & Braga, R. B. (2020). Fake news detection in social media: A systematic review. Proceedings of the XVI Brazilian Symposium on Information Systems

Memon, S. A. & Carley, K. M. (2020). Characterizing COVID-19 misinformation communities using a novel twitter dataset. arXiv preprint. arXiv:2008.00791

Moravec, P. L., Minas, R. K. & Dennis, A. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, *43*(4), 1343–1360

Murdock, I., Carley, K. M. & Yagan, O. (2025). Simulating the effects of multi-platform use on misinformation diffusion and countermeasures. Proceedings of the 17th ACM Web Science Conference 2025

Ng, L. H. X. & Carley, K. M. (2025a). The dual personas of social media bots. *arXiv preprint arXiv:2504.12498*

Ng, L. H. X. & Carley, K. M. (2025b). A global comparison of social media bot and human characteristics. *Scientific Reports*, *15*(1), 10973

Ng, L. H. X., Robertson, D. C. & Carley, K. M. (2024a). Cyborgs for strategic communication on social media. *Big Data & Society*, *11*(1), 20539517241231275

Ng, L. H. X., Zhou, W. & Carley, K. M. (2024b). Exploring cognitive bias triggers in covid-19 misinformation tweets: A bot vs. human perspective. *arXiv preprint arXiv:2406.07293*

Nyhan, B. & Reifler, J. (2019). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, *29*(2), 222–244

Pennycook, G. & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402

Phillips, S. C., Ng, L. H. X. & Carley, K. M. (2022). Hoaxes and hidden agendas: A twitter conspiracy theory dataset. Companion Proceedings of the Web Conference 2022

Prooijen, J.-W. (2018). *The Psychology of Conspiracy Theories*. London: Routledge

Romer, D. & Jamieson, K. H. (2020). Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. *Social Science & Medicine*, *263*, 113356

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G. & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412

Rottweiler, B. & Gill, P. (2022). Conspiracy beliefs and violent extremist intentions: The contingent effects of self-efficacy, self-control and law-related morality. *Terrorism and Political Violence*, *34*(7), 1485–1504

Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22–36

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, *13*(9), e0203958

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

van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A. & Lewandowsky, S. (2017). Inoculating against misinformation. *Science*, *358*(6367), 1141–1142

van Prooijen, J.-W. & van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on Psychological Science*, *13*(6), 770–788

Varol, O. & Uluturk, I. (2020). Journalists on Twitter: Self-branding, audiences, and involvement of bots. *Journal of Computational Social Science*, *3*(1), 83–101

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442

Westbrook, A. & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 395–415

X (2021). Disclosing networks of state-linked information operations. Available at: `https://blog.x.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations-`

Zanette, D. H. (2002). Dynamics of rumor propagation on small-world networks. *Physical Review E*, *65*(4), 041908