Modeling Fake News Diffusion for a Resilient Social Media Platform: Agent-Based Approaches to Responsible Social Media

Maione V.[1][0009-0000-1872-0434], Ponsiglione C.[1][0000-0002-8953-5911], and Primario S.[1][0000-0002-4728-0666]

[1] Dept. of Industrial Engineering, University of Naples - Federico II
Simonetta.primario@unina.it

Abstract. This paper introduces an agent-based model developed as a computational laboratory to simulate and analyze the diffusion of fake news under realistic and heterogeneous conditions. The model integrates several conceptual building blocks: a SEIR-inspired contagion framework adapted for informational diffusion; a bounded-confidence opinion dynamics module, capturing belief evolution influenced by homophily, stubbornness, and exposure; and a role-based influence system grounded in the tipping point theory, enabling differentiated behaviors for Connectors, Mavens, and Salesmen. Network structures can be synthetically generated (e.g., scale-free, small-world) or empirically imported from real social graphs, offering both generalizability and ecological validity. The model also supports the simulation of both endogenous interactions and exogenous interventions—such as algorithmic visibility reduction and targeted immunization based on centrality measures—which reflect the logic of real-world platform governance. A key strength of the model lies in its ability to reproduce emergent phenomena, including cascades, echo chamber stabilization, belief polarization, and tipping point dynamics. These features position the model as a generative theoretical tool as well as a practical environment for testing policy interventions. The system is currently undergoing structured internal validation, where each building block is tested in isolation before being integrated. Upon completion, the model will be calibrated using empirical data from observed disinformation campaigns. The proposed model aims to bridging the gap between theoretical modeling and digital platform governance, supporting the design of context-aware moderation strategies and resilient information systems. Additionally, it can offer a data-rich environment to support the training of AI-driven, adaptive responses to misinformation in complex online ecosystems.

Keywords: Fake News, Misinformation, Design platform, Agent-based model, Virality, Tipping Point.

1 Introduction

In recent years, the proliferation of fake news and misinformation has emerged as one of the most pressing challenges to democratic discourse and social cohesion in digital societies. While the phenomenon itself is not new, the amplification mechanisms enabled by social media platforms have significantly transformed the scale and velocity at which false narratives propagate. This has led to growing concern among policymakers, researchers, and technology providers regarding the design of platforms and interventions capable of curbing the spread of disinformation [1], [2].
Despite a vast and growing body of research on fake news propagation, the majority of existing models rely on either simplified network representations [3], [4] or deterministic diffusion mechanisms that fail to fully capture the cognitive, structural, and algorithmic complexity of real-world social platforms. Furthermore, many studies limit their intervention analyses to static, globally-applied policies—such as content removal or user bans—without considering how local, context-sensitive strategies might prove more effective in heterogeneous networks [5], [6].

To address these limitations, this study introduces a modular Agent-Based Model (ABM) that simulates the spread of fake news across synthetic and empirical network structures, incorporating role-based agent behavior, opinion dynamics, and a set of intervention strategies inspired by the internal logic of social media platforms, such as algorithmic visibility control and targeted immunization. The model allows for experimentation with both exogenous and endogenous responses to disinformation and provides a flexible environment to test hypotheses about the social and structural conditions that facilitate or inhibit its spread.

The novelty of the approach lies in several aspects. First, unlike previous ABMs which rely exclusively on classical network typologies (random, small-world, scale-free), this model is designed to also ingest real-world social graphs, allowing for greater ecological validity [7]. Second, the simulation incorporates a multi-layered representation of influence, combining exposure-based SEIR dynamics with cognitively grounded belief updating based on bounded confidence models [8]. Third, the model features a role system adapted from Gladwell's theory of social tipping points [9], enabling differentiated behaviors for Connectors, Mavens, Experts, and Salesmen. Moreover, this paper emphasizes how platform-level mechanisms—such as feed curation, visibility scoring, and algorithmic ranking—can be abstracted into controllable variables within the model, allowing for direct policy implications to be drawn from simulation outcomes [10], [11].

The remainder of the paper is structured as follows. Section 2 outlines the theoretical foundations of the model, highlighting relevant literature on fake news diffusion, network modeling, opinion dynamics, and intervention strategies. Section 3 details the model architecture following the ODD protocol. Section 4 presents the state of the art of the validation phase. Section 5 discusses the implications for platform design, AI-assisted policy tools, and digital governance. Thus, it concludes with directions for future research and model refinement.

2 Theoretical Background

The diffusion of fake news in online social environments has been extensively studied across disciplines, from communication science to computational social science and network theory. Early research documented the virality gap between false and true information, revealing that fake news spreads faster, deeper, and more broadly across digital networks [2]. This phenomenon has been attributed not only to the content properties of misinformation, but also to the architecture of social platforms, user behavior, and cognitive biases [1].

A significant portion of the literature has modeled fake news diffusion using compartmental epidemic models, typically variants of the SIR or SEIR framework [12]. While useful for modeling transmission processes, these models often abstract away the social structure of interactions, reducing complex relational environments to homogeneous mixing assumptions [7]. To address this, researchers have turned to network-based models, using scale-free or small-world graphs to simulate the topological constraints of online communication [4], [3]. However, as pointed out by more recent work, such topologies may only partially reflect the properties of real-world platforms, which are shaped by layered, algorithmically mediated, and often multi-modal interactions [13].

More sophisticated approaches have incorporated opinion dynamics into epidemic models, recognizing that belief is not passively absorbed but actively constructed and updated. Models such as the bounded confidence model for the Friedkin-Johnsen social influence framework [8] offer realistic representations of how opinions evolve in the presence of social pressure, stubbornness, and homophily. When embedded into network structures, these dynamics generate emergent phenomena such as echo chambers, polarization, and ideological entrenchment [14].

Another important dimension of the literature concerns intervention strategies. Traditional approaches focus on top-down measures: banning accounts, labeling content, or injecting fact-checks. While effective in some scenarios, these strategies often face criticism for being blunt, reactive, and limited in scalability [15]. Recent studies suggest that targeted immunization—based on network centrality or behavioral profiling—may offer a more efficient way to control spread [16], [17]. Still, these strategies often lack integration with platform-level logic, such as feed algorithms or visibility modulation systems, and cognitive/attitude individual and group-level.

This paper responds to some of these gaps. It builds on the epidemiological exploiting network-based environment, but enriches it through:

- A cognitively grounded, agent-based implementation of belief formation and change;
- The inclusion of differentiated roles, allowing for the modeling of persuasive, corrective, and bridging behavior in line with both theory [9] and empirical observations;
- The explicit abstraction of algorithmic feed control into the intervention layer, allowing the model to simulate how changes in visibility ranking, information weighting, or network fragmentation might alter emergent behaviors;

- The possibility of using the model to generate structured datasets to support the development of AI-design-based interventions that go beyond rigid protocols, offering adaptive, scenario-specific responses to disinformation threats.

In this way, the model is not only intended to advance theoretical understanding of disinformation dynamics, but also to serve as a computational testing ground for the design of adaptive, evidence-based interventions, aligning simulation research with the operational logic and governance challenges of real-world social platforms.

3 The Agent-based model

To ensure compatibility with existing academic standards and facilitate replication, the model has been developed in accordance with the ODD (Overview, Design concepts, Details) protocol. This guarantees clarity, transparency, and interoperability across simulation studies. The development platform is NetLogo, which offers robust tools for visualization, parameter variation, and batch experimentation through its Behavior Space module.

3.1 Overview

The ABM presented in this study is conceived as a computational simulation laboratory to understand the dynamics of fake news propagation on social media platforms. Its purpose is twofold: on the one hand, it serves as a theoretical tool to reproduce the underlying mechanisms of informational contagion in networked environments; on the other, it is designed as a practical and extensible simulation environment for testing policy interventions and training artificial intelligence systems in the detection and containment of disinformation.

The relevance of such a tool is grounded in the increasing complexity of online information ecosystems, where social networks are structured as heterogeneous, nonlinear, and often polarized spaces. These environments amplify the effects of confirmation bias, echo chambers, and structural inequalities in content visibility, thereby facilitating the persistence and virality of false information. Addressing this challenge requires a modelling approach that combines diffusion mechanisms, network science, and social psychology, while maintaining the capacity to simulate adaptive and emergent behaviors.

At the heart of the model lies the integration of multiple computational building blocks, each aiming to represent a distinct layer of the disinformation phenomenon. These include:

- A SEIR (Susceptible-Exposed-Infectious-Recovered) epidemiological model adapted for informational diffusion;
- A network layer that can assume various topologies, from artificial (e.g., small-world, scale-free) to real-world social graphs;

- An opinion dynamics module inspired by the Friedkin-Johnsen model, where agents' beliefs evolve through weighted social influence while retaining individual stubbornness;
- A role-based framework derived from Gladwell's tipping point theory, assigning specialized diffusion capabilities to agents categorized as Connectors, Mavens, or Salesmen;
- A targeted intervention system, which simulates immunization or dampening strategies based on node centrality metrics—both traditional (e.g., degree, closeness) and alternative (e.g., distinctiveness centrality).

The temporal resolution of the model is discrete, simulating the evolution of states and opinions across time ticks. Each tick corresponds to a round of potential interactions, contagion events, opinion shifts, and possible interventions. The spatial structure, in contrast, is not geographic but relational, shaped entirely by the network topology selected. Agents are located within the network as nodes, connected to other agents via edges that signify digital interactions—such as friendship, followship, or group membership.

3.2 Design

The design of this model is the result of an effort to articulate a computational laboratory capable of capturing the multilayered and adaptive nature of fake news diffusion. Rather than relying on a single theoretical lens, the design of the model integrates principles from epidemiology, social psychology, and network science, framed within the broader logic of agent-based simulation.

Thus, the purpose of this section is not simply to enumerate design elements in accordance with the ODD protocol, but to unpack the mechanisms by which local interactions, heterogeneous roles, and exogenous interventions converge to generate system-wide patterns.

Basic Principles The model is grounded in the principle that fake news operates as an informational pathogen, whose contagion dynamics depend on both structural conditions and the cognitive configuration of agents. In this framework, misinformation does not spread merely because of contact, but because of reinforcement through social proximity, ideological alignment, and structural exposure. The contagion threshold for belief is not universal but constructed dynamically, as agents interpret, evaluate, and sometimes resist the information they receive.

The architecture follows a modular design, where each building block represents a conceptual layer: a SEIR epidemic process governs state transitions; an opinion dynamics engine shapes belief evolution; a network graph constrains and enables interactions; a role system defines asymmetric influence; and a set of interventions targets structural vulnerabilities. These blocks can operate in isolation or in synergy, depending on the experimental goals.

Emergence One of the model's key strengths lies in its ability to generate emergent phenomena from simple local rules. Agents do not possess a global view of the system; they interact with a limited set of neighbors, apply probabilistic or bounded reasoning rules, and respond to local conditions. Yet from these micro-interactions arise complex system-level behaviors: echo chambers, cascades, long-term polarization, and tipping points. For example, the model allows researchers to observe how initially moderate opinions may polarize over time in networks with high clustering and strong homophily, or how a handful of infected Mavens or Connectors can trigger large-scale outbreaks of disinformation even in robust network configurations.

Adaptation Agents in the model exhibit a form of bounded cognitive adaptation, driven not by learning algorithms or reward feedback, but by a constructivist interpretation of the social environment. At each time step, an agent evaluates its current position in the network and forms or modifies its belief according to three interacting factors:

- Its own prior belief, reflecting the internalization of previous exposures and resistance to change
- The opinions of its neighbors, weighted by trust, similarity, and edge strength
- The perceived "epidemic context", i.e., the degree to which the information is widely shared in its vicinity, which may act as a signal of credibility or urgency.

This adaptation mechanism models how individuals integrate personal attitudes with socially mediated cues to form or reinforce their stance on a circulating narrative. Agents are not blank slates: their updates are filtered through stubbornness, exposure thresholds, and homophily filters, making adaptation selective and context-dependent. Thus, action is not a mere reaction to exposure, but a process of interpretive negotiation within a local social frame.

Interaction All interactions occur along the edges of the network, which represent social ties of varying strength. These interactions are dyadic, localized, and asynchronous: not every link is activated at each tick, and not all interactions lead to influence or infection.

Three main processes are simulated through interaction:

1. Epidemiological exposure: infectious agents attempt to expose susceptible neighbors to fake news. Success is probabilistic, modulated by role (Salesmen are more effective), exposure frequency, and network centrality.
2. Opinion exchange: belief updates occur through averaging neighbor opinions, filtered by ideological proximity (bounded confidence) and agent stubbornness.
3. Role-based effects: Experts attempt to decrease belief scores of neighbors; Salesmen increase exposure force; Connectors increase outreach.

This interaction structure reproduces both transmission and resistance, reflecting the conflicting dynamics of viral misinformation and fact-based correction.

In addition to endogenous agent interactions, the system is designed to respond to exogenous interventions. These interventions do not originate from within the agent logic, but are injected from outside the system, modeling how digital platforms or institutional actors may attempt to curb the diffusion of false narratives. Concretely, such interventions are implemented in the model as immunization strategies, which reduce the probability that a given agent will be exposed to and propagate fake news. These strategies can target nodes selectively, based on:

- Their centrality (e.g., highly connected users who act as bridges or hubs);
- Their opinion profile (e.g., agents expressing extreme or unverified positions);
- Their expositional risk, defined as a function of previous exposures and network location.
- The interventions may manifest in various operational forms:
- Reduction of visibility (e.g., simulating the effect of feed demotion);
- Decrease in influence weights (e.g., simulating loss of trust or credibility);
- Increase in resistance parameters (e.g., simulating awareness campaigns or fact-check reinforcement).

Importantly, these mechanisms are dynamic and can be turned off, applied periodically, or in real-time, reflecting how platforms or policy bodies may adapt their responses over time. By simulating these interventions, the model allows for comparative analysis of different containment logics, as well as for the identification of unintended consequences—such as over-fragmentation or visibility shifts that may reinforce polarization.

Stochasticity The model incorporates multiple layers of stochasticity, making it sensitive to initial conditions and useful for probabilistic analysis. Randomness is introduced at different levels:

- Initialization: The assignment of roles, opinions, and initial infections is often random or probabilistic, unless overridden by structural criteria or research intentions.
- Interaction dynamics: Whether an agent interacts with a neighbor in a given tick is determined stochastically. Even if interaction occurs, the success of infection is probabilistic, depending on local conditions.
- Opinion update process: Belief evolution includes noise factors (e.g., exposure variability, imperfect communication), which may lead to small but compounding divergences over time.
- Timing of transitions: Durations of the exposed or infectious phases for each agent are drawn from statistical gamma distributions, simulating temporal and behavioural heterogeneity.

This stochastic dimension is essential for realism: in real networks, identical structural features do not guarantee identical outcomes. The same network seeded with different initial infected nodes may lead to vastly different epidemic trajectories.

Objectives of the model The primary objective of the model is to inform platform design and policymakers to better understand the diffusion dynamics of fake news in structured digital environments. This involves evaluating how network configuration, agent roles, belief mechanisms, and intervention strategies interact to produce distinct systemic outcomes. The model allows us to observe and measure (among others):

- The onset and propagation of information cascades (i.e. SEIR counts and transition rates)
- The formation of ideologically cohesive clusters (i.e. Average and variance of opinion over time, or polarization index)
- The success or failure of immunization and moderation efforts (i.e. number of infections per role type or metric category, or effectiveness and cost-efficiency of intervention strategies.

Observables are recorded at each simulation step and can be used for post analysis via visualization or statistical evaluation (e.g., Mann-Whitney U tests, ANOVA, post-hoc T-test). Beyond explanatory power, the model is also intended to serve a generative function. By adjusting key parameters and observing the outcomes, researchers and policymakers can simulate what-if scenarios and anticipate the consequences of interventions before they are implemented in real contexts. Moreover, the model can be used to develop a data-rich environment for training artificial intelligence systems due to its ability to produce large volumes of synthetic, labeled data derived from simulations under diverse configurations. This allows for the supervised or reinforcement learning of AI agents that can learn to: detect high-risk network configurations, predict outbreak conditions, or recommend adaptive, context-aware intervention strategies, rather than fixed, rule-based or blanket suppression approaches.

3.3 Details

Initialization: The initialization phase of the model represents a critical step in configuring the simulation environment. It is during this phase that the social network structure is generated or imported, agents are placed within the graph and endowed with their initial attributes, and the state of the system is defined before the onset of dynamic interactions. The modular architecture of the model allows for both parametric initialization—based on synthetic topologies and probabilistic distributions—and empirical initialization—based on real-world network datasets.

The initialization procedure begins with the creation or loading of the network. In synthetic mode, the observer may select from three canonical topologies:

- Scale-Free (Barabási–Albert model): characterized by a power-law degree distribution and the presence of hubs, suitable for simulating influencer-driven platforms such as Twitter.

- Small-World (Watts–Strogatz model): retains high clustering and low average path length, ideal for representing closed group environments such as Telegram channels or Facebook groups.
- Random Network (Erdős–Rényi model): useful as a baseline, where link probability is uniform and structure is uncorrelated.

In empirical mode, the model accepts external edge lists or adjacency matrices, which may originate from datasets of real user interactions—such as comment networks, retweet graphs, or friendship lists. Each node in the graph is associated with an agent, and edges represent social ties. Once the network is established, the agent population is instantiated. Each agent is assigned:

- An initial epidemiological state as Susceptible, with a predefined percentage of (content creators) nodes seeded as "infected" at time step zero;
- An opinion score, drawn from a uniform or clustered distribution (e.g., bimodal or polarized), depending on the chosen scenario;
- A role label, such as Connector, Maven, or Salesman. Role assignment can be turned off, random, proportional (e.g., 5% of nodes), or structural (e.g., top 10% by degree become Connectors).

The model also includes adjustable agent-level parameters, such as stubbornness, exposure sensitivity, and transmission probability. These can be fixed for all agents or drawn from roles to simulate heterogeneity.

Input Data The model supports various input types to ensure versatility in both synthetic experimentation and data-driven validation. The primary categories are network data: input as edge lists, GML files, or adjacency matrices. These may also include node attributes (e.g., account activity level, number of followers). Or even role assignment criteria: definitions of what constitutes a Connector or Maven can be based on structural metrics (e.g., degree, closeness, betweenness) or imported from user metadata. An overview of simulation parameters defined via interface sliders or batch files (in NetLogo's Behavior Space), is reported in Table1:

Table 1: Overview of Simulation Parameters

Building Block	Parameter	Description	Value/range
SEIR Model	Incubation-time	Time steps an agent remains in exposed state before becoming infectious	Drawn from Gamma (α, λ)
SEIR Model	Infectious-time	Duration of infectiousness before recovery	Drawn from Gamma (α, λ)
SEIR Model	Initial-exposed-nodes	Number or proportion of nodes exposed at initialization	User-defined (e.g., one node or a group of nodes)
SEIR Model	Seroconversion-time	Time before a vaccinated node becomes immune	Drawn from Gamma (α, λ)
Opinion Dynamics	Opinion	Continuous belief value of the agent	Range: [-1,+1]
Opinion Dynamics	Stubbornness	Resistance to opinion change	[0,1]
Opinion Dynamics	Homophily-threshold	Maximum opinion difference for influence to occur	[0,2]
Network Topology	Network-type	Topology used for simulation	Scale-Free / Small-World/Random / Real Social Graph
Network Topology	Degree	Number of edges per node (relevant for generation or real networks)	Depends on topology
Roles	Role-type	Function of the node in spreading or contrasting fake news	Connector / Maven / Salesman / Expert
Roles	Role-distribution	Proportion of nodes assigned each role	User-defined (e.g., 10% connectors)
Intervention	Centrality-metric	Metric used to rank nodes for immunization	Degree / Betweenness / Distinctiveness / Closeness
Intervention	Intervention-type	Strategy applied to control spread	Isolation / Visibility reduction / Belief immunization
Intervention	Intervention-timing	When and how often interventions are applied	Initial / periodic (every X ticks)
Intervention	Visibility-weight	Reduction factor applied to influence of a node	(0.0,1.0)
Simulation	Ticks-per-day	Time resolution of the simulation	User-defined (e.g.,24)
Simulation	Max-ticks	Maximum number of steps in a simulation run	User-defined (e.g.,1000)

Submodels. The model is structured around four primary submodels (building blocks), each of which governs a specific layer of simulation. These submodels can operate independently or in coordination, and each has internal parameters and behaviors.

SEIR Diffusion. This component adapts the classic epidemiological SEIR model to represent informational contagion. Agents transition between states as follows:

- Susceptible (S): agents not yet exposed to fake news
- Exposed (E): agents who can see the content on their homepage but have not yet reposted it.

- Infectious (I): agents who believe the fake news and may propagate it.
- Recovered (R): agents who no longer expose their neighborhood (have posted other contents or are immunised).

Transitions are probabilistic and time-dependent. Exposure does not guarantee infection: the decision to repost content depends on the cumulative weight of exposure, source credibility, and belief alignment. The infectious period can be fixed or drawn from a distribution, and recovery may be spontaneous or triggered by exposure to counter-information or persuasive argumentation.

A critical feature is the inclusion of role-based modulation (described better below) where: mavens have lower transmission probability, connectors expose more neighbours, and salesmen have higher persuasive power.

a. Opinion Dynamics

Opinion formation is modeled using an adaptation of the Friedkin–Johnsen influence system. Each agent updates their belief value $x_i \in [-1,+1]$ at each tick by computing a weighted average of their neighbors' opinions

$$x_i\,(t+1) = \lambda_i \cdot x_i(0) + (1-\lambda_i) \cdot \sum_{j=1}^{N} w_{ij} \cdot x_j$$

Where:

- $x_i(t)$ is the opinion of agent *i* at time *t*;
- $x_i(0)$ is the initial opinion of agent *i*;
- $\lambda_i \in [0,1]$ represents the degree of stubbornness of agent *i*, that is, how strongly the agent remains influenced by their initial opinion. When $\lambda_i = 1$, the agent is completely rigid, and their opinion remains unchanged over time; whereas when $\lambda_i = 0$, the agent is fully open to social influence;
- w_{ij} indicates the weight of the influence of agent *j* on agent *i*. It must hold that $w_{ij} \in [0,1]$ and $\sum_{j=1}^{N} w_{ij} = 1$

Then, it is assumed that an agent may update their opinion according to this model only while in the exposed state.

Role-Based Influence. Inspired by Gladwell's theory of social tipping points, this module introduces heterogeneity in agent influence via the assignment of roles:

- Connectors: have a large number of links and tend to bridge distant clusters. They facilitate rapid diffusion across structural holes.
- Mavens: are information initiators who seed fake news early. Their beliefs are highly stable, and they are rarely persuaded.
- Salesmen: act as persuasive influencers. They increase the weight of their opinion during neighbor updates.

Role effects are not purely symbolic: they translate into behavioral changes in exposure probability, belief resistance, influence magnitude, and recovery likelihood.

Intervention. This submodel activates one or more targeted counter-disinformation strategies, aimed at reducing network susceptibility. Strategies can include: node isolation (i.e. removal or silencing of key nodes), visibility reduction (i.e. decreasing the transmission probability or out-degree of specific agents), or belief immunization (i.e increasing the stubbornness or resistance of selected agents).

Node selection for intervention is governed by a chosen centrality metric. The model includes traditional metrics (degree, betweenness, closeness, eigenvector) and introduces distinctiveness centrality, a newer measure that emphasizes nodes bridging weakly connected clusters. This metric has shown promise in identifying influential but low-degree nodes often missed by classical measures.

Interventions can be static (applied once at initialization) or dynamic (repeated at set intervals), and are parameterized by budget constraints (e.g., max % of nodes per round). The effect of interventions is observable in terms of infection curve flattening, network fragmentation, or opinion convergence.

Conclusion

At this stage of development, the model is undergoing a systematic internal validation process, designed to identify and correct implementation errors and ensure robustness across its modular components. This phase precedes any substantive experimental analysis and focuses on the systematic identification and resolution of potential malfunctions, logical inconsistencies, or programming bugs that may compromise simulation reliability.

To this end, a series of pilot experiments have been designed to test each individual building block—namely the SEIR diffusion module, the opinion dynamics engine, the role-based behavior system, the intervention layer, and the network structure generator—in isolated conditions. By activating one module at a time, it becomes possible to detect whether specific processes (e.g., belief updating, contagion progression, or intervention impact) behave according to theoretical expectations and predefined parameter ranges.

Once the functionality of each submodule has been verified independently, the model will proceed to integrated simulation runs in which all building blocks are activated simultaneously. These full-system simulations are critical to evaluate the emergent behavior of the model under complex interactions and will serve as the basis for more advanced scenario testing.

Following the internal validation stage, the model will be calibrated to reflect empirical data drawn from documented cases of disinformation spread. This will enable comparative analysis between simulated outputs and observed real-world dynamics, such as the evolution of narrative clusters, time-to-peak propagation, and spatial or ideological reach of false content. This calibration will not only enhance the model's external validity, but also lay the foundation for its future use as a predictive and diagnostic tool in research and policy applications.

In conclusion, the proposed model contributes theoretically by integrating epidemic diffusion, network topology, and bounded opinion dynamics within a modular, agent-based framework. This allows researchers to explore how structural and cognitive

mechanisms jointly shape disinformation spread, and to test the effects of role asymmetries, polarization, and intervention strategies in a controlled setting.

From a practical standpoint, the model offers a foundation for designing information systems and digital platforms that are more resilient to disinformation. By simulating interventions such as visibility modulation and targeted immunization, the model supports the development of context-aware policies and algorithmic strategies that can be adapted to specific network conditions. It also enables the generation of synthetic data for AI training, fostering the emergence of adaptive, evidence-based responses rather than static, one-size-fits-all solutions.

Beyond these practical contributions, the model can also be integrated into platform governance structures and organizational decision-making. In this sense, it operates as a decision-support tool for social media platforms, regulators, and policy-makers, and can directly support the work of trust & safety teams, content moderation units, and compliance departments dealing with regulations such as the EU Digital Services Act (DSA). By providing "what-if" scenarios and synthetic datasets, the model helps these organizational actors anticipate the systemic consequences of governance interventions, evaluate trade-offs between effectiveness, user autonomy, and resilience, and align algorithmic strategies with broader regulatory and ethical requirements.

While the proposed interventions—such as downranking, targeted immunization, or algorithmic ranking adjustments—are powerful tools to curb the spread of disinformation, they also raise ethical and operational challenges. On the ethical side, interventions that reduce visibility or prioritize certain types of content inevitably touch upon issues of freedom of expression, transparency, and potential bias in algorithmic governance. Decisions about which nodes or narratives to target may risk reinforcing existing inequalities or silencing minority voices. On the operational side, implementing these interventions requires coordination across multiple organizational units (e.g., policy, engineering, trust & safety) and alignment with external regulatory frameworks such as the EU Digital Services Act. The model can contribute to these debates by serving not only as a simulation tool but also as a platform for ethical reflection: by allowing decision-makers to test alternative strategies in a controlled environment, it helps identify not only the most effective, but also the most socially responsible, interventions. This dual focus on efficacy and responsibility reinforces the relevance of the model as a support tool for organizations navigating the complex trade-offs of platform governance.

Therefore, this work aims to bridge the gap between theoretical modeling and operational governance, providing a tool that supports responsible platform design and proactive management of digital misinformation.

Despite its strengths, the model presents limitations in terms of scalability. When applied to very large networks, computational performance may be constrained by the NetLogo environment. Future developments could translate the model into other ABM platforms (e.g., Repast, MASON, GAMA) that can leverage higher computational power while preserving the ability to simulate complex agent behaviors and emergent systemic dynamics. In addition, conceptual limitations remain: the current implementation simplifies individual cognitive processes, and the empirical calibration of the

model is still in progress. These aspects may restrict external validity and will require further refinement to strengthen the robustness of future applications.

References

[1] D. M. J. Lazer *et al.*, "The science of fake news," *Science (1979)*, vol. 359, no. 6380, pp. 1094–1096, Mar. 2018, doi: 10.1126/science.aao2998.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science (1979)*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.

[3] M. Del Vicario *et al.*, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, Jan. 2016, doi: 10.1073/pnas.1517441113.

[4] A. Bessi, "On the statistical properties of viral misinformation in online social media," *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 459–470, Mar. 2017, doi: 10.1016/j.physa.2016.11.012.

[5] W.-Y. S. Chou, A. Gaysynsky, and R. C. Vanderpool, "The COVID-19 Misinfodemic: Moving Beyond Fact-Checking," *Health Education & Behavior*, vol. 48, no. 1, pp. 9–13, Feb. 2021, doi: 10.1177/1090198120980675.

[6] F. Zollo *et al.*, "Debunking in a world of tribes," *PLoS One*, vol. 12, no. 7, p. e0181821, Jul. 2017, doi: 10.1371/journal.pone.0181821.

[7] J. Borge-Holthoefer, R. A. Banos, S. Gonzalez-Bailon, and Y. Moreno, "Cascading behaviour in complex socio-technical networks," *J Complex Netw*, vol. 1, no. 1, pp. 3–24, Jun. 2013, doi: 10.1093/comnet/cnt006.

[8] N. E. Friedkin and E. C. Johnsen, " Social influence networks and opinion change," *Advances in group processes*, vol. 16, no. 1, pp. 1–29, 1999.

[9] M. Gladwell, *The tipping point: How little things can make a big difference*. 2006.

[10] M. Cinelli *et al.*, "The COVID-19 social media infodemic," *Sci Rep*, vol. 10, no. 1, p. 16598, Oct. 2020, doi: 10.1038/s41598-020-73510-5.

[11] J. B. Bak-Coleman *et al.*, "Stewardship of global collective behavior," *Proceedings of the National Academy of Sciences*, vol. 118, no. 27, Jul. 2021, doi: 10.1073/pnas.2025764118.

[12] S. Raponi, Z. Khalifa, G. Oligeri, and R. Di Pietro, "Fake News Propagation: A Review of Epidemic Models, Datasets, and Insights," *ACM Transactions on the Web*, vol. 16, no. 3, pp. 1–34, Aug. 2022, doi: 10.1145/3522756.

[13] A. Sharma and U. Ghose, "Sentimental Analysis of Twitter Data with respect to General Elections in India," in *1st International Conference on Smart Sustainable Intelligent Computing and Applications, ICITETM 2020*, G. N., G. P.S., P. V., B. V.E., and L. C.M., Eds., Usict, Guru Gobind Singh Indraprastha University, New Delhi, India: Elsevier B.V., 2020, pp. 325–334. doi: 10.1016/j.procs.2020.06.038.

[14] A. Flache *et al.*, “Models of Social Influence: Towards the Next Frontiers,” *Journal of Artificial Societies and Social Simulation*, vol. 20, no. 4, 2017, doi: 10.18564/jasss.3521.

[15] D. Lazer *et al.*, “Computational social science,” *Science (1979)*, vol. 323, no. 5915, pp. 721–723, Feb. 2009, doi: 10.1126/science.1167742.

[16] G. F. Chami, S. E. Ahnert, N. B. Kabatereine, and E. M. Tukahebwa, “Social network fragmentation and community health,” *Proc Natl Acad Sci U S A*, vol. 114, no. 36, pp. E7425–E7431, 2017, doi: 10.1073/pnas.1700166114.

[17] R. Pastor-Satorras and A. Vespignani, “Immunization of complex networks,” *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, vol. 65, no. 3, p. 036104, Feb. 2002, doi: 10.1103/PhysRevE.65.036104.