applied sciences

MDPI

Article

From Misinformation to Resilient Communication: Strategic Simulation of Social Network Dynamics in the Pharmaceutical Industry

Filippo Ghisi [1], Marco Gotelli [1], Vittorio Solina [2,*] and Flavio Tonelli [1]

[1] Dipartimento di Ingegneria Meccanica, Energetica, Gestionale e dei Trasporti, University of Genoa, Via Opera Pia 15, 16145 Genova, Italy; filippo.ghisi@edu.unige.it (F.G.); marco.gotelli@unige.it (M.G.); flavio.tonelli@unige.it (F.T.)
[2] Department of Mechanical, Energy and Management Engineering, University of Calabria, 87036 Rende, Italy
* Correspondence: vittorio.solina@unical.it

Abstract

Health misinformation across digital platforms has emerged as a critical fast-growing challenge to global public health, undermining trust in science and contributing to vaccine hesitancy, treatment refusal and heightened health risks. In response, this study introduces Impact, a novel simulation framework that integrates agent-based modeling (ABM) with large language model (LLM) integration and retrieval-augmented generation (RAG) to evaluate and optimize health communication strategies in complex online environments. By modeling virtual populations characterized by demographic, psychosocial, and emotional attributes, embedded within network structures that replicate the dynamics of digital platforms, the framework captures how individuals perceive, interpret and propagate both factual and misleading health messages. Messages are enriched with evidence from authoritative medical sources and iteratively refined through sentiment analysis and comparative testing, allowing the proactive pre-evaluation of diverse communication framings. Results demonstrate that misinformation spreads more rapidly than factual content, but that corrective strategies, particularly empathetic and context-sensitive messages delivered through trusted peers, can mitigate polarization, enhance institutional trust and sustain long-term acceptance of evidence-based information. These findings underscore the importance of adaptive, data-driven approaches to health communication and highlight the potential of simulation-based methods to inform scalable interventions capable of strengthening resilience against misinformation in digitally connected societies.

Keywords: health communication; vaccine misinformation; agent-based modeling; audience behavior modeling; sentiment analysis; social networks; modeling and simulation

check for updates

Academic Editors: Michal Tomaszewski and Serhii Lupenko

Received: 7 October 2025
Revised: 28 October 2025
Accepted: 29 October 2025
Published: 3 November 2025

Citation: Ghisi, F.; Gotelli, M.; Solina, V.; Tonelli, F. From Misinformation to Resilient Communication: Strategic Simulation of Social Network Dynamics in the Pharmaceutical Industry. *Appl. Sci.* **2025**, *15*, 11734. https://doi.org/10.3390/app152111734

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid spread of vaccine misinformation has become a major challenge in global public health. This problem has been recognized by the World Health Organization as an "infodemic" [1], that is, an overabundance of both accurate and false information that makes it difficult for citizens to identify reliable sources and make informed health decisions. This issue has assumed unprecedented urgency in recent years, as stagnation in vaccination coverage undermines the great progress that immunization programs have historically achieved; between 1980 and 2023, vaccines are estimated to have prevented 154 million deaths worldwide [2], yet global coverage has recently stalled or declined. For example,

first-dose coverage for measles (MCV1) dropped from about 86% in 2019 to 84% in 2024, while diphtheria-tetanus-pertussis (DTP3) coverage stands at approximately 85%, leaving 14.3 million infants unvaccinated [3,4]. Importantly, these declines are not confined to low- and middle-income countries; at least one antigen has decreased in uptake in 21 of 36 high-income countries over the past decade [5], revealing that vaccine hesitancy, refusal and delay are pressing global concerns that transcend socioeconomic boundaries.

These phenomena have been documented not only for childhood vaccination, but also across other healthcare domains, including screening programs and chronic disease management, where misinformation and mistrust erode uptake of preventive services. Recent longitudinal analyses of COVID-19 vaccine misinformation on Twitter have shown that although the overall volume of low-credibility content is lower than mainstream news, a small set of "superspreaders" account for a disproportionately large share of its dissemination, which suggests that interventions targeting network hubs or highly influential accounts may be especially efficient [6]. Other systematic reviews have emphasized that much of the existing literature is concentrated in high-income countries and in a few platforms (notably Twitter, Facebook, and YouTube), while regions of the Global South and emerging social media platforms remain underrepresented, thereby limiting the generalizability of findings [7]. Studies employing mixed methods or large sentiment-analysis datasets have also shown that trust in institutions mediates much of the relationship between social media exposure and vaccine hesitancy, with lower trust amplifying the negative effect of exposure even when factual content is available [8].

Vaccine hesitancy is closely entangled with the dynamics of misinformation, which erodes trust in medicine and public institutions. A growing body of empirical studies demonstrates that exposure to vaccine misinformation reduces intentions to vaccinate, while reliance on social media for health information correlates strongly with higher refusal rates, even after controlling for conspiracy beliefs and trust levels [9]. Moreover, misinformation spreads faster, wider and more persistently than corrections on platforms such as Twitter/X [10], amplifying distrust and fueling hesitation.

The study of misinformation and health communication has evolved along multiple methodological and conceptual trajectories. Classical empirical analyses demonstrated the asymmetric diffusion of content, where false information spread faster and reaching broader audiences due to its novelty and emotional appeal [10], while more recent research has explored corrective strategies emphasizing that corrections are typically effective in contrasting health misinformation, although they became less impactful on more polarized beliefs [11]. These dynamics are amplified by what Helmond [12] describes as the platformization of the web: the structural integration of social media platforms into communication ecosystems through algorithmic curation, datafication and monetization logics that privilege engagement over accuracy. Platformization reshapes the visibility and velocity of misinformation flows, creating hybrid human–algorithmic environments where echo chambers and polarization emerge [13]. Understanding these socio-technical infrastructures is therefore essential for realistic modeling of online behavior.

The psychological dimension of this problem is equally significant, with factors such as institutional distrust, political ideology, cognitive biases and conspiracy mindset shaping both susceptibility to misinformation and responses to corrective information [14]. These findings highlight the complex interplay between individual psychology, information environments and public health outcomes, making the problem of vaccine misinformation multidimensional and resistant to simple solutions.

Against this backdrop, researchers have increasingly turned their attention to interventions, examining both traditional strategies, such as debunking and fact-checking and

more innovative approaches, such as prebunking, psychological inoculation, gamified interventions and simulation-based modeling.

Traditional debunking, intended as fact-checking and correcting false claims after exposure, has shown measurable, but limited effects, constrained by what is known as the "continued influence effect," whereby once misinformation is internalized, corrections are difficult to process, often ignored and in some cases may even backfire by reinforcing initial beliefs [15]. The effectiveness of debunking depends heavily on messenger credibility, timing of delivery and the social context of communication, which limits its scalability across diverse platforms and populations.

This recognition has motivated the development of proactive strategies such as prebunking or psychological inoculation, which seek to expose individuals to weakened versions of misinformation or the rhetorical techniques of manipulation before actual exposure, thereby building cognitive resistance. The "Bad News" game designed by Roozenbeek and van der Linden [16] has become a prime example of this approach, demonstrating in cross-cultural experiments that inoculation improves people's ability to spot misinformation techniques across languages and political contexts. Further large-scale experiments have confirmed that inoculation can significantly increase resistance to misinformation, although effects tend to decay within weeks, with text- and video-based interventions proving more durable than game-based ones and booster interventions helping sustain resistance over time [17]. These insights suggest that inoculation strategies hold considerable promise, but also underscore the importance of optimizing delivery formats and reinforcement schedules to maintain long-term efficacy.

Another important area of research examines the influence of platform-specific strategies and intervention formats. Evidence suggests that the success of interventions depends deeply on contextual factors, such as whether the messenger is a peer, influencer, institution or algorithm, whether the message takes the form of text, video, narrative or interactive game and whether the intervention is delivered before, during or after misinformation exposure. For example, on Facebook, fact-check and accuracy labels have been shown to reduce misinformation sharing, but their effect depends on precise wording and timing [18]. On Instagram, visual prebunking and influencer-based strategies appear particularly effective given the platform's image-centric environment [19], while on Twitter/X, embedded corrections such as Community Notes are often perceived as more trustworthy than top-down warnings or algorithmic labels [20].

A systematic review by Okuhara et al. [21] on narrative-based interventions further highlights that corrections delivered through engaging and identity-congruent narratives can be effective in reducing health misinformation, although effect sizes remain moderate and highly dependent on cultural alignment and emotional framing.

These findings collectively underscore the fragmentation of intervention efficacy, as outcomes vary significantly across platforms, messengers and formats, suggesting that no single strategy is universally effective and that interventions must be tailored to context-specific dynamics.

Complementing empirical studies, theoretical and computational modeling has emerged as a powerful tool for understanding and anticipating misinformation dynamics. Adaptations of epidemiological models such as the SIR (susceptible-infected-recovered) framework have been employed to conceptualize misinformation spread, with individuals represented as susceptible to misinformation, "infected" by false beliefs, or "recovered" through debunking or inoculation [22]. These models reveal that the "basic reproduction number" of misinformation (analogous to epidemiological R_0) depends heavily on network structures, content virality and the reach of corrective information.

More advanced approaches combine network topology with ordinary differential equations (ODEs) to simulate how misinformation suppresses vaccine uptake in heterogeneous versus homogeneous populations, with findings suggesting that heterogeneous populations are more resilient, while scale-free networks magnify the harmful effects of misinformation [23].

Other models, such as the IPSR framework proposed by Rai et al. [24], integrate prebunking by introducing new compartments (e.g., prebunked, exposed, resistant) and parameters such as forgetting rates and prebunking potency, offering a more nuanced depiction of intervention dynamics.

Simulations are also increasingly used to explore cross-platform interactions: Murdock et al. [25] show how individuals who use multiple platforms simultaneously are affected differently by interventions and how trusted "good actors" deployed across platforms can shape user trust and reduce misinformation spread. Parallel advances in computational modeling and artificial intelligence have enriched the analytical toolbox available for such investigations. Sentiment-analysis studies employing machine-learning techniques on social-media data have demonstrated how emotional polarity influences virality and framing efficacy [26], while simulation frameworks in cognitive and strategic engineering [27], such as Bruzzone et al. [28], have begun to replicate complex decision-making dynamics within synthetic populations. In this context, LLMs have become pivotal instruments for modeling message semantics and adaptive communication. Zhou et al. [29] introduce the concept of the Industrial Large Model, a domain-specific generative architecture optimized for high-fidelity reasoning in industry-specific contexts that can be safely embedded into simulation frameworks. Complementing this perspective, Naveed et al. [30] provide a comprehensive synthesis of several LLM architectures, emphasizing interpretability, data provenance and responsible-AI considerations as prerequisites for deploying generative systems in scientific and policy settings. Together, these studies underscore the feasibility of integrating generative AI components into simulation pipelines to capture the linguistic, affective and contextual richness of digital discourse.

These advances highlight the potential of integrating modeling, simulation and AI to test scenarios that are difficult to study empirically, but they also reveal gaps, including limited integration of psychological heterogeneity, insufficient exploration of surge scenarios and weak connections between misinformation outcomes and downstream epidemiological impacts.

Despite significant progress, the literature remains fragmented and constrained by several limitations. First, most empirical studies rely on laboratory experiments, online panels, or controlled simulations on single platforms, limiting the generalizability of findings to real-world, cross-platform settings where users simultaneously consume information across Facebook, Twitter, Instagram, WhatsApp and TikTok. Second, while inoculation effects are promising, their long-term durability remains underexplored and little is known about the optimal timing and design of booster interventions. Third, individual-level heterogeneity, including variations in trust, ideology and conspiracy mindset, is often measured, but rarely incorporated into predictive models, limiting the capacity to design tailored interventions. Fourth, the influence of network structures such as scale-free, small-world, or modular architectures is often overlooked in empirical interventions, even though these properties fundamentally shape the spread of misinformation and corrections. Fifth, the interactions among key variables, messenger credibility, message format, timing of delivery, are typically studied in isolation rather than in combination, leaving open questions about how these dimensions interact in real-world contexts. Sixth, surge scenarios such as those observed during the COVID-19 pandemic remain understudied, even though misinformation tends to spike during crises when public anxiety is high and information environments are saturated. Finally, many studies focus on attitudinal outcomes such as belief change

or vaccination intentions, while relatively few assess actual behavioral outcomes such as vaccine uptake or epidemiological consequences and even fewer examine unintended side effects such as backlash, message fatigue, or erosion of institutional trust.

Scope of the Work

These gaps reveal the urgent need for more comprehensive, integrative approaches capable of addressing the complexity of misinformation dynamics across platforms, populations and contexts. The authors propose a comprehensive simulation framework that combines insights from platform studies, behavioral science and large-model architectures to simulate how misinformation and corrective messages evolve across networked publics, with particular attention to the pharmaceutical communication domain. Such framework allows researchers and decision-makers to simulate how misinformation and corrective interventions propagate across overlapping platforms, capturing spillover effects and amplification patterns that single-platform studies miss. It also enables systematic comparisons of intervention types, such as peer versus institutional messengers, prebunking versus debunking, while accounting for timing and sequencing, variables that are difficult to manipulate comprehensively in experimental designs. By incorporating heterogeneity in user psychology and simulating different network topologies, the framework helps identifying strategies that are robust across contexts and populations rather than narrowly effective in specific conditions. Moreover, time dynamics such as the decay of inoculation effects and the reinforcement of booster interventions are modeled explicitly, offering insights into how interventions should be scheduled and adapted to maintain long-term resistance. Crucially, the framework allows stress-testing of interventions under surge conditions, where misinformation spreads rapidly and public attention is heightened and link information outcomes to epidemiological consequences by integrating misinformation models with disease transmission models. This framework provides policymakers and communicators with evidence not only about belief and intention outcomes, but also about the potential public health impact of interventions in terms of vaccination uptake and outbreak dynamics. Finally, by simulating possible unintended consequences, such as trust erosion from heavy-handed corrections or backlash among certain subgroups, it helps identify safe and context-sensitive strategies.

In summary, while the state of the art demonstrates a wide array of promising interventions, from debunking to prebunking to gamification and significant progress in theoretical modeling, it remains fragmented, context-dependent and insufficiently integrative. The proposed simulation framework combines empirical insights, psychological theory and network modeling offers a comprehensive solution to these limitations, providing a scalable, adaptable and evidence-based tool for designing interventions that can address misinformation in a holistic and effective manner.

After this review of the current state of the art, the article is organized as follows. Section 2 presents the Materials and Methods, offering a detailed description of the simulation framework, including the construction of agent-based models, the integration of retrieval-augmented generation and the calibration of virtual populations. Section 3 reports the Results, highlighting both macro-level outcomes such as the relative reach of factual versus false messages and polarization trends across networks, as well as micro-level insights into how individual agents' trust and attitudes evolve through repeated exposures. Section 4 provides the Discussion, contextualizing these findings within the broader literature and reflecting on the implications for designing proactive communication strategies, outlining the strengths and limitations of the proposed framework. Finally, Section 5 concludes by summarizing the study's contributions, drawing lessons for policymakers

and healthcare institutions and identifying promising directions for future interdisciplinary research on misinformation, digital health communication and public trust.

2. Materials and Methods

This study employs a multi-layered simulation methodology designed to model, analyze and optimize the spread of healthcare communication in complex, digitally mediated social systems, integrating ABM, RAG and LLM-driven natural language processing within a single comprehensive framework that enables the systematic testing and refinement of messaging strategies before their deployment in real-world public health interventions.

This approach is built upon four core methodological pillars: the construction of a heterogeneous virtual population of agents parameterized with realistic demographic, psychological and behavioral attributes; the development of a weighted social network structure that reflects the dynamics of contemporary digital platforms; the design and injection of evidence-based health messages enriched with semantic and emotional context through LLM embeddings and RAG pipelines; and the implementation of an iterative feedback loop for message evaluation, refinement and comparative testing over discrete simulation time steps.

2.1. Simulation Implementation

The simulation framework, including the message-processing pipeline and data-analysis scripts, was implemented in Python 3.11 using a modular architecture built around open-source libraries. Core network construction and structural-metric routines rely on NetworkX 3.1, while language-model inference and embedding generation employ PyTorch 2.3.1 and Hugging Face Transformers 4.41.0 within a Retrieval-Augmented Generation (RAG) workflow. Sentiment classification and clustering tasks are handled through scikit-learn 1.2, complemented by sentence-transformers, pandas, numpy, and NLTK for text preprocessing and feature extraction. The RAG module interfaces with publicly available knowledge bases, including the World Health Organization (WHO) Digital Health Repository, PubMed, and CDC open data APIs, through a custom FastAPI-based document retriever and ChromaDB vector store, which enable efficient semantic search and contextual retrieval at runtime. All simulations were executed on a workstation equipped with an Intel Core i9-14900HX (2.20 GHz, 16 cores), 32 GB RAM, and an NVIDIA RTX 4060 GPU (8 GB VRAM).

Demographic priors and baseline-trust distributions were derived from publicly available sources to ensure empirical grounding. Data were drawn from the CIA Factbook, the Edelman Trust Barometer (2022 country-level trust indicators) and the Wellcome Global Monitor (2018 health-attitude modules). Each survey item was mapped to model attributes, such as age, education, trust in science and health-attitude scores, via weighted-probability sampling calibrated to national demographic proportions.

2.2. System Architecture

At the foundation of the simulation there is a synthetic population of agents, each representing an individual actor within a digital communication environment and parameterized with a rich array of demographic, cognitive and psychosocial characteristics. These include basic demographic variables such as age, gender, education level, socioeconomic status and cultural background, which influence baseline attitudes toward healthcare information and digital literacy levels. Agents also carry a multidimensional opinion vector composed of variables representing trust in science, attitudes toward vaccination and public health authorities, susceptibility to conspiracy beliefs, political ideology and cultural views. Emotional states, modeled as dynamic variables with continuous bounded values

for fear, trust, skepticism, hope and anger, are initialized based on empirical distributions and are updated over time in response to message exposure. Cognitive attributes include susceptibility to common psychological biases (e.g., confirmation bias, authority bias and bandwagon effect), modeled as probabilistic weighting functions that modulate message acceptance thresholds. Each agent also maintains a memory log that records prior exposures, message credibility scores, source trust scores and cumulative influence weights, enabling the simulation to replicate phenomena such as trust-building over repeated exposures, source fatigue and resistance to contradictory information. Initial distributions for all agent attributes are calibrated using real-world survey data from open sources, as well as peer-reviewed literature on misinformation susceptibility and vaccine hesitancy.

To ensure full replicability and transparency, the simulation presents key quantitative parameters and execution settings through an interactive configuration interface (Figure 1) and dedicated configuration files. Users can explicitly define the population size (ranging from 100 to 10,000 agents), simulation duration (1 to 90 days) and available communication budget (from 1000 to 1,000,000 USD), as well as specify the drug or health topic, target region and communication channel (type of social media, television, community meetings, scientific journals or government websites). These parameters control both network initialization and behavioral response dynamics, allowing precise replication of experimental conditions across runs. Quantitative outcomes, including message reach, acceptance rate, belief-shift magnitude and trust evolution over time, are automatically logged at each simulation step as standardized performance metrics. These parameters (reported in Table A4 inside Appendix A) are also configurable through specific configuration files to allow fast and easy replication of experiments under different population assumptions or test sensitivity to variations in trust distribution, ideological polarization or digital literacy levels.

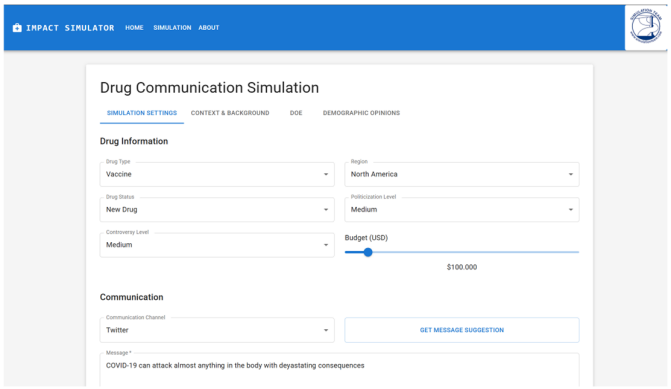


Figure 1. Simulator configuration interface showing user-adjustable parameters, including population size, simulation duration, communication channel, and message metadata fields. Interactive sliders and tooltips assist users in defining experimental conditions and validating message inputs prior to execution.

Agents are embedded within several social network graphs designed to capture the topological and algorithmic features of modern digital communication platforms. The networks are generated using a mixed-model approach that combines Barabási–Albert preferential attachment mechanisms (to simulate influencer-driven hubs and scale-free network properties) with Watts–Strogatz small-world rewiring (to replicate clustered communities and echo chambers). Edges represent communication ties between agents and are weighted by trust strength, interaction frequency, platform type (e.g., peer-to-peer messaging vs. public broadcast) and content amplification factors. Algorithmic amplification is modeled as a multiplicative weight applied to message visibility, reflecting the influence of recommendation systems, trending algorithms and engagement-optimized content ranking, which are known to significantly alter the diffusion landscape of online information. Parameters

governing network size, density, clustering coefficient and degree distribution are specified at simulation initialization and can be derived from empirical social media datasets or generated synthetically. The Impact simulator supports multi-platform communication scenarios through a multi-layer network architecture. Given the synthetic population of agents, the framework generates separate, but interlinked network topologies for each modeled social platform (e.g., Twitter/X, Facebook, Instagram, or Telegram), each characterized by distinct connectivity patterns, amplification coefficients and algorithmic visibility functions. Agents can simultaneously belong to multiple networks, mirroring real-world user behavior where individuals engage across several platforms. This design allows the same message to propagate within and between platforms, producing composite diffusion trajectories and enabling analysis of cross-platform amplification effects. As a result, the simulator can reproduce complex communication phenomena such as multi-channel message reinforcement, simultaneous rumor propagation and heterogeneous reach across media ecosystems, as well as differential intervention outcomes, phenomena that are often overlooked in simpler ABM implementations.

Message generation, injection and representation form the second major component of the simulation and are handled through a combination of LLM-based semantic embedding and RAG-enhanced contextualization (as represented in the architecture in Figure 2). Health communication messages, ranging from factual vaccine safety statements to myth corrections and behavior change prompts, are first embedded into a high-dimensional semantic vector space using LLaMA 3.1, fine-tuned on a domain-specific corpus of medical literature, WHO guidelines, CDC advisories and public health communication documents. This embedding enables semantic similarity analysis and ensures that messages can be dynamically compared, clustered and reformulated based on context. The RAG module enriches these messages by retrieving authoritative background documents from a curated knowledge base and incorporating relevant evidence snippets directly into the message payload, thereby grounding the generated text in verifiable medical knowledge. Each message is further annotated with emotional valence scores, assigned through transformer-based sentiment classifiers fine-tuned on public communication datasets, categorizing tone as fear-inducing, reassuring, urgent, empathetic or neutral. This allows the simulation to model the influence of affective framing on message acceptance and sharing behavior—critical given the well-documented role of emotional resonance in misinformation virality. Messages are introduced into the simulation either as institutional campaigns and media coverage or as agent-generated rumors and peer-to-peer misinformation, and their diffusion dynamics are tracked across the networks.

Upon exposure to a message, each agent evaluates its content according to a multi-factor behavioral decision model grounded in established theories of social psychology and communication science (relevant pseudocode is provided in the Appendix A, Tables A1–A3). The underlying psychological model integrates classical dual-process theories of cognition, combining fast and intuitive with deliberate and analytical mechanisms within the agent decision-making logic. Each agent's evaluation of incoming messages is therefore governed not only by heuristic biases such as authority or confirmation effects, but also by rational appraisal of credibility, congruence with prior beliefs and source reputation. Probabilistic weighting functions determine the dominance of either processing mode depending on the agent's cognitive profile, emotional arousal, and contextual complexity of the message.

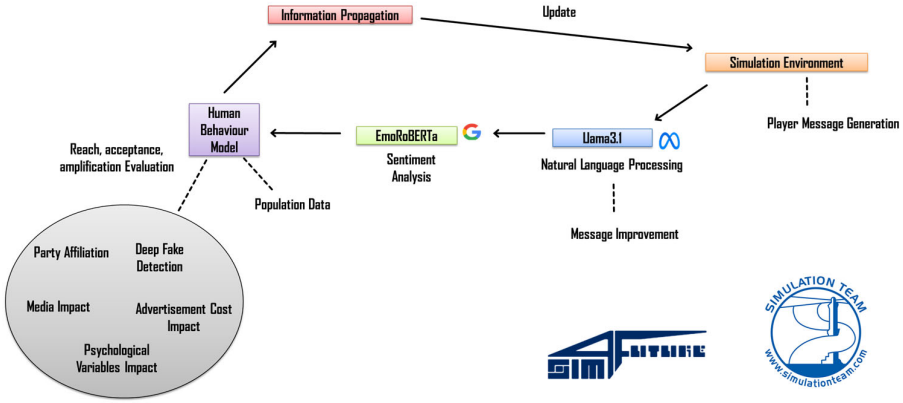


Figure 2. Impact Simulator general architecture. Arrows represent messages flow across the framework components, as well as main inputs and functions of each module.

Acceptance probability is calculated as a logistic function of four primary inputs: source trust (derived from the agent's memory and network connection weight), cognitive congruence (the degree to which the message aligns with the agent's existing opinion vector), emotional resonance (the match between the message's affective tone and the agent's current emotional state) and social proof (the proportion of neighboring agents who have shared or endorsed the message). This function is modulated by bias weighting coefficients that account for individual variability in cognitive biases, allowing for heterogeneous decision-making behavior across the population. If a message is accepted, the agent updates its opinion vector using a Bayesian updating rule, adjusts emotional state variables according to the message's valence and may choose to propagate the message to connected agents, subject to a stochastic resharing probability that reflects real-world variability in sharing behavior. The simulation supports multiple message types, including competing messages on the same topic, allowing researchers to study complex scenarios such as competing narratives, information overload and corrective backfire effects.

A key methodological innovation of this framework is its iterative message refinement loop, which combines NLP-driven sentiment analysis, LLM-based optimization and comparative testing to pre-test message effectiveness. After each simulation run, message performance is evaluated based on reach, acceptance rate and opinion shift magnitude. Using these performance metrics as reward signals, the system employs the LLaMA 3.1 LLM to adapt message framing, linguistic style and emotional tone in subsequent iterations. Comparative testing is conducted by injecting multiple versions of the same factual message with varying framings (e.g., empathetic vs. authoritative, narrative vs. informational) and tracking their differential diffusion trajectories and impact on belief change. This feedback loop continues over multiple epochs, allowing the system to converge on optimized message strategies tailored to specific population segments and network conditions. Importantly, all message variants, evaluation metrics and optimization policies are logged and version-controlled, ensuring that the iterative process is fully transparent and reproducible.

Time is discretized into simulation steps representing daily or weekly communication cycles, depending on the modeled scenario. At each time step, new messages are injected, agents evaluate and propagate received messages, emotional states and trust scores are updated and memory logs are recalibrated. Longitudinal dynamics, such as message fatigue, trust decay and the cumulative effects of repeated exposures, are explicitly modeled, allowing for the analysis of both short-term diffusion patterns and long-term opinion evolution. The simulation terminates either when message diffusion falls below a predefined threshold (indicating saturation or loss of relevance) or when a fixed time horizon is reached. Output data, including diffusion graphs, opinion trajectories, emotional

state distributions and intervention effectiveness metrics, are exported in standard CSV and JSON formats for downstream analysis.

2.3. Framework Validation

To ensure external validity and quantitative reproducibility, the framework was validated against empirical diffusion data and subjected to structured sensitivity analysis. Validation employed approximately 100 hydrated tweets sourced from the PanaceaLab COVID-19 Twitter Dataset (accessed on 22 March 2020) https://github.com/thepanacealab/covid19_twitter/).

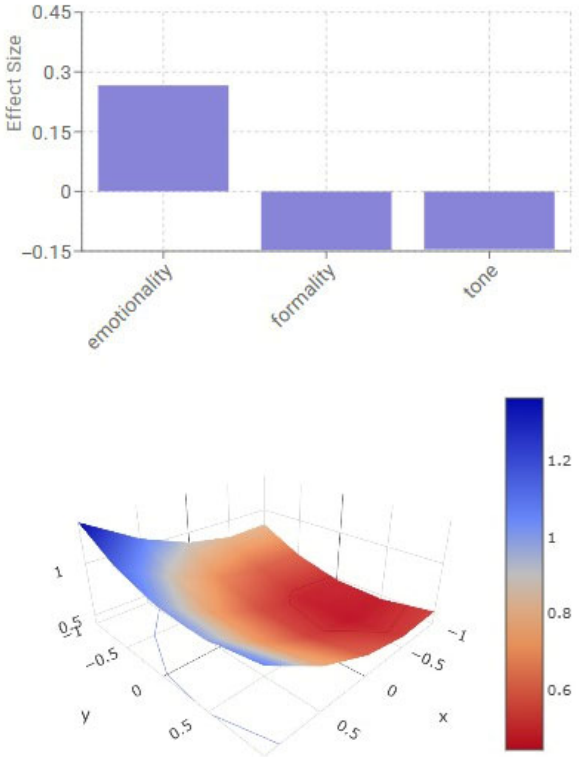
In the current validation phase, particular attention was given to ensuring data quality and representativeness of the hydrated Twitter subset. The dataset was systematically filtered to specifically select English language tweets that covered a balanced distribution of message types (factual statements, corrective content and misinformation narratives) and to remove tweets affected by hydration errors, duplicates and non-pertinent content that fell outside the scope of health and misinformation topics. Both very high-engagement and very low-engagement tweets were intentionally retained to preserve the natural heterogeneity of online diffusion dynamics and avoid bias toward highly amplified content. Although no explicit anti-bias algorithms were applied at this stage, these filtering procedures improve internal consistency and reduce structural sampling distortions. In future developments, the framework will incorporate formal bias-mitigation techniques, such as weighted resampling and adversarial debiasing, together with multilingual and cross-platform datasets to enhance external validity and equity across diverse linguistic and cultural contexts.

At the end of this process, each message was semantically embedded and injected into the simulation under Twitter-like network topology with a population size of 5000 agents. Simulated diffusion metrics, including reach, acceptance rate and resharing probability, were then compared against corresponding empirical engagement statistics (retweet and like counts) after normalization. Across runs, the model reproduced realistic diffusion patterns for message reach and acceptance rate, confirming representative propagation dynamics relative to real-world social data.

To further examine the model's robustness, a Design of Experiments (DoE, Figure 3) module implementing a Central Composite Design (CCD) was used to estimate the sensitivity of communication outcomes to key simulation and message parameters. Independent variables included network density, trust decay rate, controversy level, tone, formality, emotionality and politicization level, with response variables defined as message reach, acceptance probability and amplification rate. Each design point was simulated with adaptive replication until the mean squared prediction error (MSPe) fell below 0.01. Sensitivity coefficients and standardized effect sizes were computed for all factors, with 95% confidence intervals obtained via bootstrap resampling. The strongest effects were observed for network density, trust decay rate and message tone, indicating that both structural and affective factors substantially modulate message diffusion potential.

Overall, this validation process demonstrates that the proposed framework not only reproduces known online diffusion benchmarks, but also provides quantifiable uncertainty estimates for key behavioral and network parameters, supporting its reliability for scenario-based experimentation and policy simulation.

This comprehensive methodological approach, combining behavioral modeling, network science, AI-enhanced message processing and iterative optimization, ensures that the study's findings are both reproducible and extensible, providing a robust platform for future research into the dynamics of health communication and misinformation mitigation.



Interpretation: The response is maximized when emotionality is intense and formality is casual. The effect of emotionality and formality appear to be independent (no significant interaction).

Figure 3. The Simulator Design of Experiments: visualization presenting the effect of emotionality, formality and tone on message sharing.

3. Results

The results of the present study demonstrate the capacity of the Impact framework to generate both macro-level metrics, that capture population-wide communication dynamics, and micro-level insights, which highlight the nuanced processes by which individuals engage with, reinterpret and propagate health-related messages in complex digital ecosystems.

3.1. Macro-Level Diffusion Dynamics

At the global scale, the system produced continuous measures of message reach, sentiment distribution and opinion polarization, enabling a comparative analysis of factual versus false information diffusion. Across 50 simulation runs calibrated on empirical tweet diffusion data from the PanaceaLab COVID-19 dataset ($n \approx 100$ posts), misinformation messages exhibited an average initial diffusion velocity 1.6× higher than factual counterparts. Mean message reach at equilibrium stabilized at 76.6%, while message acceptance plateaued at 37.6% and amplification at 45.9% (Figure 4). These results confirm that emotionally charged misinformation benefits from algorithmic amplification and peer contagion, producing faster early growth and transient polarization spikes in clustered networks.

However, corrective messages framed empathetically or narratively achieved significantly higher recovery effects, with a mean 12.4% increase in acceptance rate over neutral factual message). The model further replicated temporal convergence between factual and false information after intervention, showing a trust-restoration lag of approximately 6–8 simulation days under medium polarization conditions.

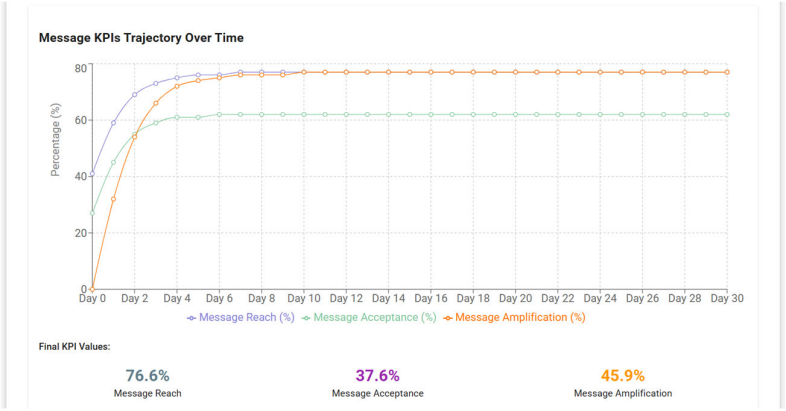


Figure 4. Message propagation visualization illustrating diffusion dynamics over time. The temporal evolution reflects amplification and saturation phases of the message trajectory.

3.2. Micro-Level Behavioral and Cognitive Patterns

At the agent level, opinion trajectories and emotional responses revealed distinct cognitive pathways shaping message evaluation. Agents exposed to empathetic framings exhibited 25–30% higher likelihood of acceptance compared to authoritative framings among low-trust clusters, confirming the moderating role of affective congruence in message processing. Conversely, agents with high institutional trust responded more favorably to authoritative framings. Backfire effects were observed in 8–10% of cases, primarily within high-polarization subgroups, when corrective content was perceived as emotionally incongruent or excessively assertive.

The model's dual-process cognitive representation aligns with established heuristic–systematic processing frameworks: fast, emotion-driven appraisals governed initial sharing behaviors, while slower, deliberative evaluation influenced sustained belief updating.

These results validate that Impact reproduces empirically observed psychological dynamics of misinformation correction and supports its interpretability as a dual-process behavioral simulator.

Complementing this perspective, Figure 5 provides a breakdown of message acceptance by educational level and age group. For messages not grounded in empirical truth, acceptance rates were around 86% among individuals with low education, 63% with medium and 8% with high education levels. Older cohorts (>55 years) exhibited the highest susceptibility to acceptance and sharing of these messages, while younger groups had higher awareness of possible misleading content. These results strengthen the framework's ability to reveal how demographic characteristics condition the reception of misinformation and how corrective strategies must be tailored to account for these differences.

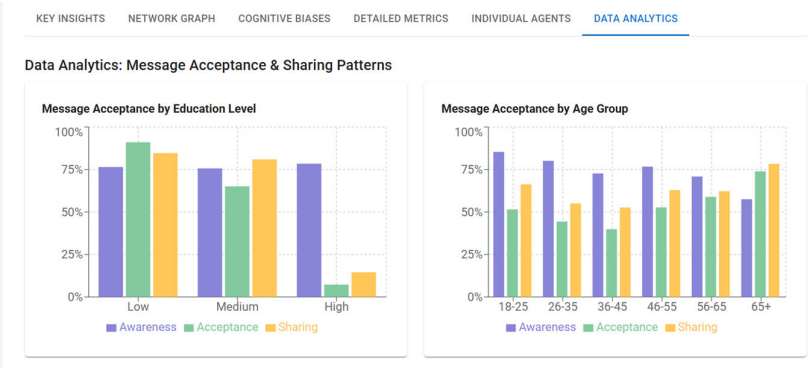


Figure 5. Key performance indicator (KPI) visualization showing message awareness, acceptance rate and sharing probability across educational and demographic segments.

Together, these quantitative and structural findings confirm that the framework not only reproduces empirically observed diffusion patterns but also provides statistically supported differentiation across message framings, demographic segments and trust-network configurations. The system's calibration against real-world data and validation through the integrated DoE sensitivity module enhance its robustness and reproducibility, supporting its application in the evaluation and pre-testing of pharmaceutical and health communication campaigns.

4. Discussion

The iterative message refinement analytics embedded in the framework demonstrates the tangible benefits of pre-testing communication strategies: alternative framings of identical factual content produced measurable differences in acceptance rates, belief updating and long-term diffusion trajectories, with empathetic framings generally outperforming neutral or authoritative framings in highly polarized environments, while authoritative framings proved more effective in populations with higher baseline trust in science. Taken together, these results highlight the importance of tailoring communication strategies not only to the factual content of the message, but also to the cognitive, emotional and network contexts in which the message circulates.

From an interpretive standpoint, the findings validate the central hypothesis that integrating agent-based modeling with retrieval-augmented message generation allows for a more precise and actionable understanding of the dynamics of health misinformation. Unlike static approaches to debunking, the Impact simulator enables the proactive exploration of counter-messaging strategies under a wide range of network conditions and audience profiles, thereby offering healthcare institutions a practical tool for stress-testing communication campaigns before their public deployment.

Beyond their academic contribution, these results also carry strategic implications for healthcare organizations, policymakers and pharmaceutical companies seeking to safeguard public trust and ensure the effectiveness of their outreach. The evidence that message framing, messenger type and timing profoundly shape outcomes suggests that institutions cannot rely on uniform campaigns, but must instead adopt multifaceted strategies of communication adapted to specific subgroups. For example, targeted interventions toward younger audiences on fast-moving platforms may require emotionally resonant peer-driven strategies, while older or more educated populations may respond more positively to authoritative, evidence-heavy communication. Moreover, the quantitative outputs from the simulator, such as the differential acceptance rates by education and age, can directly inform resource allocation, allowing decision-makers to prioritize vulnerable demographics and tailor interventions accordingly. In this context, the framework is especially insightful for intervention durability, the rate at which corrective effects decay following the initial exposure. Simulations indicate that the impact of empathetic corrections typically diminishes after 15–20 simulated days in high-turnover networks, unless reinforced by follow-up messaging. These findings suggest that booster strategies should be designed as multi-phase communication sequences, with renewed exposure intervals calibrated to coincide with predicted trust decay cycles and content fatigue patterns. Effective booster scheduling, analogous to immunization boosters in health interventions, can sustain message retention and mitigate the reemergence of misinformation clusters. Future applications may use this functionality to plan adaptive reinforcement strategies that align with real-world media cycles and platform engagement rhythms.

The relevance of the framework to the pharmaceutical industry lies in its ability to simulate communication campaigns involving safety signals, product recalls and vaccine-related controversies. The simulator also supports scenario planning for corporate-

reputation management, allowing assessment of how early transparency, influencer engagement or scientific-evidence framing affects trust recovery after misinformation outbreaks. Beyond vaccination, these capabilities extend to chronic-disease awareness, antibiotic-resistance prevention, rare-disease sensitization and public education on responsible drug use, including high-visibility issues such as anti-obesity or diabetes medications. Thus, while Impact is conceived as a general health-communication simulator, its architecture directly addresses strategic challenges unique to pharmaceutical communication systems characterized by high regulation, scientific complexity and reputational sensitivity.

From a decision-making standpoint, the framework functions as a support tool, enabling stakeholders to evaluate alternative strategies before committing financial and organizational resources to real-world campaigns, thereby reducing risks and enhancing cost-effectiveness. This integrative approach bridges the gap between behavioral insights and operational strategy, illustrating how computational models can provide actionable guidance for managing misinformation in complex, high-stakes environments such as vaccination campaigns or pharmaceutical communications during health crises.

4.1. Comparative Positioning Within Simulation Research

When positioned within the broader landscape of social-diffusion simulators, the Impact framework extends prior models such as SIR and IPSR by incorporating cognitive–emotional heterogeneity and generative message evolution. Classical SIR-based misinformation models capture contagion-like spread, but both messages and agents are static and homogeneous. IPSR approaches introduce psychological parameters, but remain limited to fixed message typologies. In contrast, Impact allows messages to evolve dynamically through RAG-LLM modules that generate contextually adaptive content.

Furthermore, considering recent agent-based frameworks such as the Murdock et al. cross-platform diffusion model [25] and the IPSR epidemic-style simulator [24], the Impact framework demonstrates comparable macro-level diffusion magnitudes, while offering finer-grained behavioral resolution. Across studies, overall message reach converges in the 60–80% range, with interventions producing relative reductions in misinformation exposure or belief between 6% and 25%. While the Murdock et al. model primarily quantifies awareness and trust shifts under platform-level moderation (e.g., a 6.2% decline in misinformation visibility with 20% overlap) and the IPSR framework reports a 23% reduction in peak spreaders through prebunking, the Impact simulator reproduces similar suppression dynamics within a cognitively enriched environment. Its empirically calibrated diffusion velocity, an around 1.6× higher for misinformation, and equilibrium acceptance plateau of 37.6% align with these patterns, while revealing additional results such as affect-driven amplification (45.9%) and recovery dynamics (6–8 day trust-restoration lag). These cross-model consistencies reinforce the validity of the Impact framework design, demonstrating that it retains quantitative coherence with established diffusion theories while expanding explanatory power through psychological realism and message adaptability.

In this context, the choice of LLaMA 3.1 was guided by its modular RAG integration and general purpose scope, which allow greater flexibility and reproducibility compared with domain-specific biomedical models, such as BioBERT or PubMedBERT, that, while offering some chatting capabilities, lack interactive generative capacities such as message generation and finetuning based on physiological parameters.

Beyond epidemiological and agent-based approaches, several simulation and educational platforms have been developed to illustrate misinformation dynamics, such as Fakey and Bad News. However, these tools primarily function as gamified learning environments that allow users to select from a limited number of predefined responses, focusing on public awareness and behavioral inoculation. While gamified approaches excel at micro-level user education and

psychological resistance building, they are not designed to simulate macro-scale dynamics such as trust evolution, network density effects or message recontextualization over time.

By contrast, Impact is a comprehensive analytical simulator that enables researchers and decision-makers to compose entire messages and intervention sequences within a multi-agent social network, observing their emergent consequences over time. By integrating behavioral modeling, LLM-driven message interpretation and network analytics, the Impact Framework complements these tools by providing a quantitative and system-level understanding of communication outcomes, which can in turn inform the design of more effective interventions.

This flexibility allows comprehensive scenario testing of pharmaceutical and health-communication strategies that cannot be achieved through fixed-option educational games. Nevertheless, conceptual insights from these gamified frameworks, particularly the logic of prebunking and psychological inoculation, remain valuable for informing and stimulating proactive resilience.

4.2. Ethical and Transparency Considerations in Generative Simulation

The incorporation of LLMs and RAG mechanisms into social simulations introduces significant ethical and transparency challenges. Although these tools enable nuanced representation of human language and message diffusion, they also raise concerns about opacity, bias propagation and the creation of potentially harmful or misleading content. Furthermore, while controlled generation of misinformation variants is instrumental for stress-testing communication strategies, it risks normalizing or amplifying deceptive patterns if misapplied. To mitigate these risks, all message generations in the Impact framework were constrained to closed simulation environments and no synthetic message is ever released or publicly disseminated.

To promote responsible AI use and mitigate bias in the language models employed, all text-generation and sentiment-classification processes were monitored through a transparency layer that logs model prompts, outputs and metadata for post hoc analysis. Pretrained embeddings were limited to publicly available, research-grade models with documented ethical use statements. The dataset was pre-filtered to remove non-pertinent or malformed content during the hydration process, ensuring linguistic integrity and topical consistency. Both high-engagement and low-engagement messages were retained to preserve the natural diversity of user behaviors. Nevertheless, the interpretability of LLM-driven agents remains limited and future developments should integrate explainable-AI modules to make agent reasoning traceable.

Ethical safeguards and governance mechanisms are thus essential prerequisites for deploying such frameworks in research or industrial contexts. In applied terms, responsible deployment of such simulators requires explicit data-governance frameworks defining permissible use, access control and auditability. The potential misuse of predictive diffusion analytics in corporate or political settings, for instance, to manipulate public perception, micro-target audiences or optimize persuasive messaging, raises significant ethical and regulatory concerns.

Future work will explore governance mechanisms such as independent ethics boards and algorithmic-impact assessments to ensure that simulation results are used to strengthen, not undermine, the integrity of public communication. The approach aligns with recent recommendations on generative-AI governance and with industrial large-model transparency guidelines, reinforcing the necessity of accountable design for synthetic social simulations.

4.3. Limitations and Future Work

Despite its methodological innovations, the present study presents some technical limitations that constrain generalizability and interpretability. Although the model has

been validated with subject matter experts and with several tweets retrieved from open-source databases, it has not yet undergone comprehensive empirical validation against large-scale multiplatform social media datasets. While its internal dynamics reproduce plausible diffusion trajectories, external calibration and benchmarking remain necessary to confirm quantitative accuracy. Second, while the initial validation of the Impact Framework, performed with the PanaceaLab COVID-19 Twitter dataset, provides a valuable benchmark for the simulator diffusion realism, this approach remains limited in both linguistic and platform scope. The analyzed dataset is predominantly composed by English-language messages specific to pandemic-related topics, and reflects the structural characteristics of a single platform, whose retweet-based diffusion patterns differ markedly from those on Facebook, WhatsApp, or TikTok content. These factors constrain cross-cultural generalizability and may not capture linguistic and contextual nuances that shape misinformation dynamics in multilingual environments. Future calibration will therefore focus on expanding the empirical basis of the model through multilingual datasets (e.g., Spanish, Arabic, Hindi, Chinese) and multiplatform corpora encompassing Facebook, Instagram, TikTok and other messaging services, as well as including a vastly higher number of social messages for the analysis. Such expansion will allow the Impact framework to reflect distinct communicative cultures and algorithmic environments, strengthening its external validity and ensuring broader applicability to global health communication contexts. Also, the current design assumes homogeneous access to information and constant exposure probability, simplifying the more fragmented and platform-specific nature of digital ecosystems. Furthermore, although LLM-based message processing enhances semantic realism, it introduces potential biases from pretraining corpora and stochastic variability across runs. These concerns highlight the broader limitations of synthetic data relative to authentic social-media traces: while synthetic datasets allow safe experimentation and parameter control, they cannot fully reproduce the behavioral unpredictability of real users and algorithmic recommendation systems. Hybrid validation strategies combining empirical traces, human-in-the-loop evaluation and cross-model comparisons will therefore be pursued.

Additionally, computational scalability could pose a constraint in running larger-scale simulations with more than 10 thousand agents. In this case, computational demands increase nonlinearly due to the LLM-based message interrogation and context-adaptation processes, which dominate runtime and memory usage, so that scaling the model to approximately 100 k–250 k agents would require a workstation configuration equipped with at least 128 GB of RAM, a high-throughput CPU ($\geq$32 cores) and one or more GPUs to reach 40–80 GB of VRAM to maintain acceptable runtime. In such configuration, distributed inference and asynchronous message evaluation can be parallelized across GPU instances, significantly reducing execution time.

It is also important to note that higher-end hardware not only supports larger populations, but enables deployment of more advanced language models with superior contextual reasoning and factual consistency, enhancing message evaluation accuracy and sentiment coherence across agents. Conversely, for smaller-scale studies, consumer-grade GPUs remain sufficient, with the option to run in CPU-only mode, albeit with substantial increases in runtime. This scalability allows researchers to balance experimental complexity with available computational resources, ensuring reproducibility across a wide range of hardware configurations.

In contrast to the technical limitations, broader field challenges relate to the general difficulty of capturing evolving social norms, platform-specific moderation policies and cross-cultural message interpretation within any computational model of communication. Furthermore, even if the framework architecture supports regional communication scenarios through specific population data for agent clusters and network generation, it

still need to be tailored to simulate a global, interconnected communication campaign, including cultural trust baselines and linguistic models, from multiple geo-social contexts. Recognizing this distinction clarifies that while certain constraints may be resolved through technical scaling and data expansion, others reflect intrinsic complexities in modeling human information ecosystems that require sustained interdisciplinary collaboration.

Given these limits to the immediate operational deployment, the project should thus be regarded as an exploratory research instrument and decision-support framework rather than a production-ready communication system. Transitioning the Impact Framework from a research prototype to an operational tool involves modular deployment and governance considerations. The architecture already supports modular services through FastAPI endpoints, which can be integrated into institutional dashboards or policy-communication platforms. Prior to deployment, ethical and data-governance protocols must ensure that model outputs are traceable, auditable and contextually validated by domain experts. An intense review with subject matter experts could be carried out to improve governance model and transparent dataset documentation, as well as to tailor the simulation to responsible commercial or public-sector communication applications.

5. Conclusions

The experimental conclusions that can be drawn are significant: misinformation thrives when communication systems remain reactive and uniform, but when health authorities adopt adaptive, evidence-based and emotionally resonant strategies, the harmful effects of misinformation can be meaningfully mitigated, polarization can be reduced and trust can be rebuilt over time. The scalability of this approach, coupled with its capacity to integrate real-world data inputs and simulate cross-platform environments, underscores its potential as a transformative contribution to public health communication science. By systematically linking behavioral psychology, sentiment analysis and network-level diffusion dynamics, the Impact framework offers not merely a descriptive model of how misinformation spreads, but also a prescriptive guide for designing interventions that are robust across platforms, resilient to crisis-driven misinformation surges and aligned with the diverse psychological realities of contemporary audiences.

Looking forward, this framework also opens possibilities for future development and application. One promising direction is the integration of real-time data streams from multiple platforms to enable dynamic calibration of simulations, allowing health authorities to monitor misinformation surges and test corrective strategies in near-real time. Another avenue involves coupling misinformation dynamics more explicitly with epidemiological models, which could help quantify the downstream public health impact of different communication strategies, better linking shifts in belief or trust to changes in vaccine uptake or outbreak trajectories. From a managerial perspective, these developments would enable pharmaceutical companies, ministries of health and global organizations to design adaptive playbooks that anticipate emerging misinformation narratives and deploy context-sensitive strategies at scale. By positioning itself as both a scientific tool and a decision-making support instrument, the Impact framework lays the groundwork for a new generation of evidence-based, technologically enabled public health communication strategies capable of sustaining resilience in the face of evolving digital threats.

Author Contributions: Conceptualization, F.G., M.G., V.S. and F.T.; methodology, F.G., M.G., V.S. and F.T.; software, F.G., M.G., V.S. and F.T.; writing—original draft preparation, F.G., M.G., V.S. and F.T.; writing—review and editing, F.G., M.G., V.S. and F.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This appendix contains details and supplemental data to the main paragraphs, such as explanations of experimental parameters and the simulation pseudocode.

Table A1. This table contains the pseudocode to generate the opinion based on group characteristics.

Generate Rule Based Opinions Pseudocode
Generate_Rule_Based_Opinions(topic, message_text, groups): Initialize dictionary Opinions = {} Determine base_sentiment by counting positive vs. negative words in message_text For each demographic group in groups: Initialize default values: sentiment_score = base_sentiment acceptance = 0.5 sharing = 0.4 stance = “neutral” key_points = [] viewpoints = [] # Adjust opinions based on demographic traits If group is left-leaning: If message positive: Increase sentiment, acceptance, and sharing Add pro-public-health viewpoints Else: Add mixed or skeptical viewpoints If group is right-leaning: If message negative: Increase skepticism and freedom-oriented viewpoints Else: Support choice but remain wary of government overreach If group is highly educated: Increase trust in evidence and add research-based viewpoints If group is less educated: Increase reliance on personal experience, decrease acceptance for negative messages If group is rural: Emphasize access and cost concerns If group is urban: Emphasize healthcare equity and dense-community awareness If group has high trust in science: Increase acceptance and sharing Add pro-science viewpoints If group has low trust in science: Decrease acceptance, add skeptical and alternative-medicine viewpoints Randomize small variations to avoid uniformity Clamp all numeric scores to valid ranges (−1.0 to 1.0, 0.0 to 1.0) Derive stance from final sentiment (positive, negative, mixed) Ensure both pro and con viewpoints exist Add result to Opinions[group_id] Return Opinions

Table A2. This table contains the pseudocode to obtain the opinion of a simulated agent.

Get Opinion for Agent Pseudocode

```
Get_Opinion_For_Agent(agent, topic, opinions):

    If opinions empty:
        Return neutral opinion (stance = neutral, sentiment = 0)

    # Identify agent's demographic profile
    Determine:
        - political orientation (left/right)
        - education level (high/low)
        - community (urban/rural)
        - trust in science (high/low)

    # Match agent to demographic groups
    Create matching_groups = list of (group_id, weight)
    Assign higher weights to closer demographic matches

    If no matching groups:
        Return first available opinion

    # Combine weighted opinions
    Compute weighted averages of:
        sentiment_score, acceptance, sharing
    Tally stance frequencies across matches
    Aggregate top key points

    # Select final stance as most frequent
    Return Combined_Opinion = {
        stance,
        sentiment_score,
        key_points (top 3),
        acceptance_likelihood,
        sharing_likelihood
     }
```

Table A3. This table contains the pseudocode for the message diffusion simulation.

Simulate Message Diffusion Pseudocode

```
Simulate_Message_Diffusion(population, network_generator, message, settings):

    # Purpose:
    # Model how a health-related message (e.g., drug, vaccine, or health policy)
    # spreads across a synthetic social network, with sentiment and diffusion
    # parameters modulated by communication channel and community context.

    Input:
        population → list of agents with demographic and cognitive attributes
        network_generator → module generating the network topology (graph)
        message → textual message to be diffused
        settings → configuration parameters (channel, topic, region, etc.)

    Output:
        Updated agent states with exposure, opinion change, and propagation data

    ------------------------------------------------------------------------

    Step 1: Initialize network
        social_network ← network_generator.generate_network()
    Step 2: Analyze message content
        message_analysis ← sentiment_analyzer.analyze_message(message)

        If "engagement_estimate" not in message_analysis:
            sentiment_score ← message_analysis.get("sentiment_score", 0)
            message_analysis["engagement_estimate"] ← 0.5 + |sentiment_score| × 0.3
```

Table A3. *Cont.*

Simulate Message Diffusion Pseudocode
Step 3: Identify communication channel channel ← settings.communication_channel (default = “twitter”) original_sentiment ← message_analysis.sentiment_score channel_adjusted_message ← copy(message_analysis) Step 4: Apply channel-specific sentiment and characteristic adjustments If channel == “twitter”: Amplify sentiment extremes (±) Increase engagement for emotional content Decrease credibility and trust indicators Increase controversy potential Else if channel == “facebook”: Slightly amplify sentiment Increase engagement and emotional intensity Moderate credibility and controversy Else if channel == “newspaper”: Reduce sentiment magnitude Increase credibility and clarity Lower engagement and emotional intensity Else if channel == “tv”: Slightly amplify sentiment Increase emotional intensity and engagement Reduce message complexity Else if channel == “radio”: Slightly dampen sentiment Maintain medium engagement and credibility Increase clarity, reduce complexity Else if channel == “scientific_journal”: Strongly reduce sentiment (neutral tone) Maximize credibility and reliability Decrease engagement and emotional intensity Else if channel == “government_website”: Strongly dampen sentiment High credibility and reliability Low engagement, high complexity Else if channel == “community_meeting”: Slightly amplify positive sentiment High engagement and persuasiveness Increase clarity, reduce complexity channel_adjusted_message[“channel”] ← channel Step 5: Integrate topic-specific demographic opinions topic ← settings.drug_type (default = “General”) message_text ← message (string) Try: topic_opinions ← TopicOpinionGenerator.generate_opinions_for_topic(topic, message_text) channel_adjusted_message[“_cached_topic_opinions”] ← topic_opinions Log: number of demographic opinion groups generated Catch exception: Log: “Topic opinion generation failed; continue without opinions” Step 6: Initialize cognitive bias tracking cognitive_bias_metrics ← empty dictionary

Table A3. *Cont.*

Simulate Message Diffusion Pseudocode

```
Step 7: Construct full message metadata for diffusion
  message_data ← {
      text: message,
      channel: channel,
      sentiment_score: message_analysis.sentiment_score,
      drug_type: settings.drug_type (default "General"),
      region: settings.region (default "Global"),
      emotion_scores: message_analysis.emotions,
      controversy_level: settings.controversy_level (default "medium"),
      politicization_level: settings.politicization_level (default "medium"),
      engagement_estimate: message_analysis.engagement_estimate,
      technical_complexity: message_analysis.technical_complexity,
      characteristics: {
          clarity_score,
          emotional_intensity,
          credibility_score,
          persuasiveness
      },
     credibility_indicators: message_analysis.credibility_indicators
  }

Step 8: Initialize community-specific sentiment baselines
     community_sentiment_data ← {}
     For each agent in population:
       If agent.community_affiliation not in community_sentiment_data:
           If "Urban": set baseline_sentiment = 0.6
           If "Rural": set baseline_sentiment = 0.4
           Else:       set baseline_sentiment = 0.5
           community_sentiment_data[agent.community_affiliation] = baseline_sentiment

Step 9: Simulate message propagation over the network

     reached_agents ← ∅

     For each agent in population:
         If Should_Receive_Initial_Message(agent, channel):
             Initialize network_sentiment = {"positive": 0.3, "neutral": 0.4, "negative": 0.3 }

             For each (community, baseline) in community_sentiment_data:
                 network_sentiment[community] = baseline
             agent.receive_message(
                 message = message_data,
                 source_id = 0, # origin of message (system)
                 day = 0, # initial diffusion timestep
                 channel = channel,
                 network_sentiment = network_sentiment
             )
             Add agent.id → reached_agents

Step 10: End of diffusion initialization
              Return network state, reached_agents, and adjusted message data
```

Table A4. This table contains Simulation Parameters, Value Ranges and Data Sources.

Parameter	Description	Default/Range	Source or Rationale
population_size	Number of simulated agents representing individual users	100–10,000	User-defined; scalable to available computational resources
simulation_days	Duration of simulation run, representing communication cycles	1–90	Empirical alignment with typical social media attention spans (1–12 weeks)
time_step	Simulation temporal resolution	1 day (discrete)	Standard ABM temporal granularity
age_distribution	Age groups of synthetic population	18–75 years (discretized by decade)	CIA Factbook
education_level	Proportion of primary, secondary, tertiary education	Calibrated per region	CIA Factbook; Edelman Trust Barometer (2022)
trust_in_science	Baseline trust in scientific institutions	0–1	Wellcome Global Monitor (2018)

Table A4. *Cont.*

Parameter	Description	Default/Range	Source or Rationale
political_leaning	Left–right ideological index (–1 to +1)	Normal ($\mu = 0$, $\sigma = 0.4$)	Derived from regional political polarization studies
community_affiliation	Urban vs. rural population proportion	60% Urban/40% Rural	CIA Factbook
network_type	Hybrid Barabási–Albert/Watts–Strogatz graph	—	Combined model for realism (scale-free + clustered)
network_density	Mean connection probability	0.01–0.2	Tuned to reproduce observed social media connectivity patterns
clustering_coefficient	Average local clustering	0.3–0.6	Empirical values from Twitter/Facebook network studies
trust_weight_decay	Temporal decay of edge trust strength	0.01–0.05 per day	Calibrated from longitudinal trust evolution literature
confirmation_bias_weight	Weighting of congruence with pre-existing beliefs	0.3–0.8	Based on dual-process cognitive modeling literature
authority_bias_weight	Amplification of trust in institutional sources	0.2–0.7	Heuristic calibration following Milgram-style bias studies
bandwagon_effect_weight	Propensity to align with group opinion	0.1–0.6	Modeled on social contagion research
emotional_state_variables	Vector of {fear, trust, hope, skepticism, anger}	Continuous [0–1]	Initialized from Wellcome Global Monitor (2018)
emotional_decay_rate	Rate of emotional return to baseline	0.05–0.1 per day	Calibrated to match short-term affective adaptation studies
communication_channel	Message delivery medium	{Twitter, Facebook, Newspaper, TV, Radio, Gov. Website, Scientific Journal, Community Meeting}	User-specified
message_sentiment_score	Normalized sentiment polarity	–1.0 to +1.0	Computed via Transformer-based sentiment model
emotional_intensity	Magnitude of affective tone	0.1–1.0	Computed from LLM-based message embedding
credibility_score	Perceived trustworthiness of message	0.2–1.0	Adjusted by channel modifiers
engagement_estimate	Expected virality potential	0.2–1.0	Based on sentiment strength and platform bias
controversy_potential	Likelihood of polarized reaction	0.1–1.0	Higher for emotionally charged messages
acceptance_probability	Logistic function of trust, congruence, emotion, and social proof	0–1	Derived from dual-process integration
resharing_probability	Probability of agent repropagating accepted message	0.05–0.8	Calibrated against empirical retweet/share rates
trust_decay_rate	Reduction in trust after contradictory exposure	0.01–0.05 per exposure	Derived from longitudinal misinformation resilience studies
message_refinement_epochs	Iterations of message optimization	1–10	Defined by experiment setup
performance_metrics	Reach, acceptance rate, belief shift magnitude, trust evolution	—	Computed automatically at runtime
hardware	Simulation hardware configuration	Intel i9-14900HX, 32 GB RAM, NVIDIA RTX 4060 (8 GB)	System used in this study
runtime_scaling	Average runtime per 10 k agents	~0.5 h for 14 steps	Measured under default configuration
software_stack	Python 3.11, PyTorch 2.3.1, Transformers 4.41.0, NetworkX 3.1, scikit-learn 1.2.2, LangChain 0.0.142	Documented in Methods for reproducibility	

References

1. World Health Organization. *Infodemic*; WHO: Geneva, Switzerland, 2022; Available online: https://www.who.int/health-topics/infodemic (accessed on 12 March 2025).
2. Shattock, A.J.; Johnson, H.C.; Sim, S.H.; Carter, A.; Lambach, P.; Hutubessy, R.C.W.; Thompson, K.M.; Badizadegan, K.; Lambert, B.; Ferrari, M.J.; et al. Contribution of vaccination to improved survival and health: Modelling 50 years of the Expanded Programme on Immunization. *Lancet* **2024**, *403*, 2307–2316. [CrossRef]
3. World Health Organization. *Immunization Coverage*; WHO: Geneva, Switzerland, 2024; Available online: https://immunizationdata.who.int (accessed on 12 March 2025).

4. UNICEF. *Immunization Data*; UNICEF: New York, NY, USA, 2025; Available online: https://data.unicef.org/topic/child-health/immunization (accessed on 12 March 2025).
5. Brumbaugh, K.; Gellert, F.; Mokdad, A.H. Understanding Vaccine Hesitancy: Insights and Improvement Strategies Drawn from a Multi-Study Review. *Vaccines* **2025**, *13*, 1003. [CrossRef]
6. Pierri, F.; DeVerna, M.R.; Yang, K.; Axelrod, D.; Bryden, J.; Menczer, F. One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study. *J. Med. Internet Res.* **2023**, *25*, e42227. [CrossRef] [PubMed]
7. Skafle, I.; Nordahl-Hansen, A.; Quintana, D.S.; Wynn, R.; Gabarron, E. Misinformation About COVID-19 Vaccines on Social Media: Rapid Review. *J. Med. Internet Res.* **2022**, *24*, e37367. [CrossRef]
8. McKinley, C.J.; Olivier, E.; Ward, J.K. The Influence of Social Media and Institutional Trust on Vaccine Hesitancy in France: Examining Direct and Mediating Processes. *Vaccines* **2023**, *11*, 1319. [CrossRef]
9. Zilinsky, J.; Theocharis, Y. Conspiracism and government distrust predict COVID-19 vaccine refusal. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 1002. [CrossRef]
10. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef]
11. Vraga, E.K.; Bode, L. Correction as a Solution for Health Misinformation on Social Media. *Am. J. Public Health* **2020**, *110*, S278–S280. [CrossRef]
12. Helmond, A. The platformization of the web: Making web data platform ready. *Soc. Media Soc.* **2015**, *1*, 1–11. [CrossRef]
13. Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; Starnini, M. The echo chamber effect on social media. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2023301118. [CrossRef]
14. Romer, D.; Jamieson, K.H. Lessons learned about conspiracy mindset and belief in vaccination misinformation during the COVID pandemic of 2019 in the United States. *Front. Public Commun.* **2025**, *10*, 1490292. [CrossRef]
15. van der Linden, S. Countering misinformation through psychological inoculation. *Adv. Exp. Soc. Psychol.* **2024**, *69*, 1–58. [CrossRef]
16. Roozenbeek, J.; van der Linden, S. The Fake News Game: Actively Inoculating Against the Risk of Misinformation. *J. Risk Res.* **2018**, *22*, 570–580. [CrossRef]
17. Maertens, R.; Roozenbeek, J.; Simons, J.S.; Lewandowsky, S.; Maturo, V.; Goldberg, B.; Xu, R.; van der Linden, S. Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nat. Commun.* **2025**, *16*, 2062. [CrossRef] [PubMed]
18. Martel, C.; Rand, D.G. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nat. Hum. Behav.* **2024**, *8*, 1957–1967. [CrossRef] [PubMed]
19. Loomba, S.; de Figueiredo, A.; Piatek, S.J.; de Graaf, K.; Larson, H.J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **2021**, *5*, 337–348. [CrossRef]
20. Slaughter, I.; Peytavin, A.; Ugander, J.; Saveski, M. Community notes reduce engagement with and diffusion of false information online. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2503413122. [CrossRef]
21. Okuhara, T.; Okada, H.; Yokota, R.; Kiuchi, T. Effectiveness and determinants of narrative-based corrections for health misinformation: A systematic review. *Patient Educ. Couns.* **2025**, *139*, 109253. [CrossRef]
22. Govindankutty, S.; Gopalan, S.P. Epidemic modeling for misinformation spread in digital networks through a social intelligence approach. *Sci. Rep.* **2024**, *14*, 19100. [CrossRef]
23. Tanwar, K.; Kumar, V.; Tripathi, J.P. Heterogeneous population and its resilience to misinformation in vaccination uptake: A dual ODE and network approach. *arXiv* **2024**, arXiv:2411.11813.
24. Rai, R.; Sharma, R.; Meena, C. IPSR Model: Misinformation Intervention through Prebunking in Social Networks. *arXiv* **2025**, arXiv:2502.12740v1. [CrossRef]
25. Murdock, I.; Carley, K.M.; Yagan, O. Simulating the Effects of Multi-platform Use on Misinformation Diffusion and Countermeasures. In Proceedings of the 17th ACM Web Science Conference 2025 (Websci'25), New Brunswick, NJ, USA, 20–23 May 2025; Association for Computing Machinery: New York, NY, USA, 2025; pp. 404–412. [CrossRef]
26. Basarslan, M.S.; Kayaalp, F. Sentiment Analysis with Machine Learning Methods on Social Media. *Adv. Distrib. Comput. Artif. Intell. J.* **2020**, *9*, 5–15. [CrossRef]
27. Bruzzone, A.G.; Massei, M.; Gotelli, M.; Giovannetti, A.; Martella, A. Sustainability, Environmental Impacts and Resilience of Strategic Infrastructures. In Proceedings of the International Workshop on Simulation for Energy, Sustainable Development and Environment, SESDE, Fes, Morocco, 17–19 September 2025. [CrossRef]
28. Bruzzone, A.; Giovannetti, A.; Gotelli, M.; Hodicky, J.; Massei, M.; De Paoli, A.; Ghisi, F. Strategic Engineering to Support Decision Makers in Complex Systems: Innovative Models for Cognitive Warfare Simulation. In Proceedings of the International Multidisciplinary Modeling & Simulation Multiconference (I3M 2024—DHSS Track), Tenerife, Spain, 18–20 September 2024; CAL-TEK Srl: Cosenza, Italy, 2024. Available online: https://www.cal-tek.eu/proceedings/i3m/2024/dhss/005/pdf.pdf (accessed on 12 March 2025).

29. Zhou, J.; Cao, Y.; Lu, Q.; Zhang, W.; Liu, X.; Ni, W. Industrial Large Model: Toward a Generative AI for Industry. In Proceedings of the 2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Kingston, ON, Canada, 6–9 August 2024; IEEE: New York, NY, USA, 2024; pp. 80–81. [CrossRef]
30. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Mian, A. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–72. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.