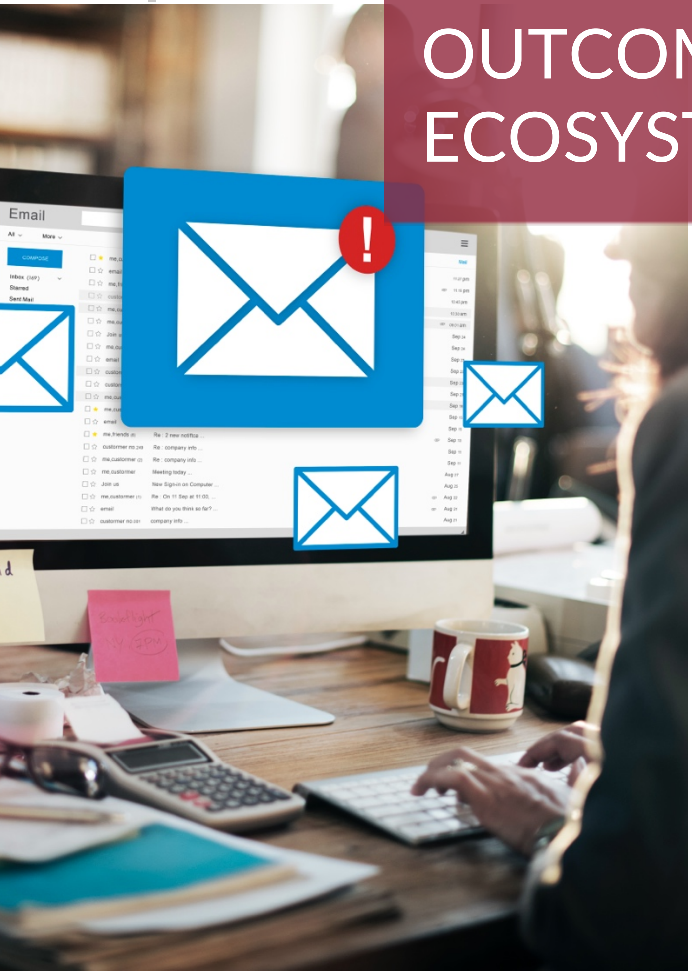
CIRANO
Knowledge into action

MIS-, DIS-, AND MALINFORMATION: A GAME-THEORETIC ANALYSIS OF COSTLY TRUTH AND EQUILIBRIUM OUTCOMES IN DIGITAL MEDIA ECOSYSTEMS



THIERRY WARIN

CS
2025s-13
WORKING PAPER

The observations and viewpoints expressed in this publication are the sole responsibility of the authors; they do not represent the positions of CIRANO or its partners. *Les idées et les opinions émises dans cette publication sont sous l'unique responsabilité des auteurs et ne représentent pas les positions du CIRANO ou de ses partenaires.*

ISSN 2292-0838 (online version)

Mis-, Dis-, and Malinformation: A Game-Theoretic Analysis of Costly Truth and Equilibrium Outcomes in Digital Media Ecosystems

Thierry Warin [*]

Abstract/Résumé

This article develops a Bayesian game-theoretic model to analyze the persistence and widespread prevalence of misinformation, disinformation, and malinformation within contemporary media ecosystems. Using the classic "Two Generals Problem" metaphor, we conceptualize information transmission as a strategic coordination game under conditions of uncertainty, emphasizing critical factors such as costs associated with truthful content production (ci), reliability of message dissemination (p), payoffs for truthful versus misleading communication (BT , BM), audience composition (α), and algorithmic amplification (γ). The model elucidates why actors, even those motivated by accuracy, rationally gravitate toward misinformation strategies when truthful messaging incurs significant costs and faces substantial barriers to audience penetration. Conditions under which honest signaling deteriorates are explicitly derived, and equilibrium outcomes—including both truthful and misinformation-dominated equilibria—are thoroughly analyzed. Historical and contemporary examples, such as Cold War disinformation operations, social media misinformation during the 2016 U.S. elections, COVID-19 pandemic misinformation, and deepfake technology applications, provide empirical validation. Our findings underscore the necessity of systemic interventions aimed at reducing truth-telling costs, enhancing message reliability, regulating algorithmic amplification, and restructuring incentives to facilitate transitions toward sustained truthful communication equilibria. Future research directions include empirical quantification of model parameters and exploring network effects to enhance policy relevance and effectiveness.

Cet article développe un modèle bayésien de théorie des jeux pour analyser la persistance et la prévalence généralisée de la désinformation et de la malinformation dans les écosystèmes médiatiques contemporains. En utilisant la métaphore classique du « problème des deux généraux », nous concevons la transmission d'informations comme un jeu de coordination stratégique dans des conditions d'incertitude, en mettant l'accent sur des facteurs critiques tels que les coûts associés à la production d'un contenu véridique (ci), la fiabilité de la diffusion des messages (p) et la fiabilité de l'information (p), la fiabilité de la diffusion des messages (p), les avantages d'une communication véridique par rapport à une communication trompeuse (BT , BM), la composition de l'audience (α) et l'amplification algorithmique (γ). Le modèle explique

[*] HEC Montréal, CIRANO

pourquoi les acteurs, même ceux qui sont motivés par l'exactitude, s'orientent rationnellement vers des stratégies de désinformation lorsque les messages véridiques entraînent des coûts importants et se heurtent à des obstacles considérables à la pénétration de l'audience. Les conditions dans lesquelles la sincérité des signaux se détériore sont explicitement déduites, et les résultats d'équilibre - y compris les équilibres dominés par la vérité et la désinformation - sont analysés en profondeur. Des exemples historiques et contemporains, tels que les opérations de désinformation de la guerre froide, la désinformation sur les médias sociaux pendant les élections américaines de 2016, la désinformation sur la pandémie COVID-19 et les applications de la technologie deepfake, fournissent une validation empirique. Nos conclusions soulignent la nécessité d'interventions systémiques visant à réduire les coûts liés à l'établissement de la vérité, à renforcer la fiabilité des messages, à réglementer l'amplification algorithmique et à restructurer les incitations afin de faciliter l'établissement de la vérité.

Keywords/Mots-clés: Bayesian game-theoretic model, disinformation, reliability of information / modèle bayésien, théorie des jeux, désinformation, fiabilité de l'information

Pour citer ce document / To quote this document

Warin, T. (2025). Mis-, Dis-, and Malinformation: A Game-Theoretic Analysis of Costly Truth and Equilibrium Outcomes in Digital Media Ecosystems (2025s-13, Working Papers, CIRANO.) https://doi.org/10.54932/NFEF9727

1 Introduction

Modern societies face an unprecedented challenge of information pollution, characterized by the proliferation of contested claims, viral rumors, and digitally altered content across social media platforms and news outlets (Wardle & Derakhshan, 2017). While rumors and propaganda are historically commonplace, the digital age has significantly amplified their scale, velocity, and impact. Falsehoods disseminate globally within seconds, frequently surpassing verified information in both speed and reach. Empirical evidence indicates that false news stories propagate more extensively, rapidly, and broadly than truthful information, with false tweets approximately 70% more likely to be retweeted (Vosoughi, Roy, & Aral, 2018). These dynamics suggest contemporary information ecosystems systematically favor misinformation, even though audiences typically express a strong preference for accurate information (Allcott & Gentzkow, 2017). This paradox—high public demand for truth juxtaposed with an abundant supply of falsehood—raises fundamental questions: Why does misinformation flourish, and under what conditions can truthful communication sustainably prevail?

To rigorously address these questions, this article introduces a Bayesian coordination game model inspired by the classical "Two Generals Problem," a theoretical illustration emphasizing the complexities of achieving reliable communication under conditions of uncertainty. The Two Generals Problem describes two military commanders seeking to coordinate an attack by communicating across enemy-controlled territory, highlighting that successful coordination demands mutual certainty of message receipt, which inherently leads to an infinite regress of confirmation attempts (Gray, 1978). This metaphor effectively encapsulates a critical challenge within contemporary media environments: ensuring accurate dissemination and widespread acceptance of truthful information amidst pervasive uncertainty, adversarial interactions, algorithmically driven distortions, and audience segmentation.

In contemporary media landscapes, actors committed to truthful communication—such as journalists, scientists, and fact-checkers—confront significant strategic challenges analogous to those faced by the two generals. Truthful messaging imposes substantial production costs (c_i), encompassing extensive resources, thorough verification processes, and strategic efforts to achieve reliable penetration of diverse audience segments. Conversely, actors disseminating misinformation (unintentionally inaccurate content) or disinformation (intentionally deceptive content) incur comparatively minimal production costs, capitalizing on emotionally resonant, simplified messaging readily aligned with algorithmic amplification mechanisms (γ). As a result, misinformation strategies often yield higher immediate payoffs (B_M) relative to truthful communication (B_T), a phenomenon exacerbated by heterogeneous audience composition dynamics (α), especially when the audience includes substantial segments characterized by lower discernment or higher susceptibility to sensationalized narratives.

Recognizing these strategic nuances, this article significantly extends traditional binary analyses by formulating a polynomial Bayesian coordination model. In this expanded framework, actors select among multiple messaging strategies—including highly accurate, partially accurate, and outright misinformation—thereby capturing the complexities inherent in real-world communication scenarios. The model incorporates nonlinear feedback mechanisms, reflecting how audience trust evolves dynamically in response to observed historical messaging patterns. Specifically, polynomial audience reliability perceptions vary according to frequencies of truthful, partially accurate, and misleading communications, emphasizing the nonlinear and cumulative nature of reputation and trust dynamics.

This article proceeds by first reviewing pertinent interdisciplinary literature, clearly distinguishing misinformation, disinformation, and associated phenomena, and examining the political incentives, technological facilitators, and psychological mechanisms underpinning contemporary information disorder. Subsequently, the extended polynomial Bayesian coordination model formally describes strategic interactions between content producers (senders) and audiences (receivers), explicitly integrating parameters such as message reliability (p), costs of truth-telling (c_i), algorithmic amplification (γ), and audience heterogeneity (α). Empirical and illustrative case studies—including historical disinformation campaigns, electoral misinformation, health-related infodemics, and abuses enabled by deepfake technologies—validate and contextualize the theoretical predictions derived from the model. The concluding discussion elaborates on the implications of the model's findings for policy interventions, advocating targeted strategies designed to lower the costs associated with accurate content production, impose effective penalties for misinformation, and strengthen reliability mechanisms. Ultimately, these systemic interventions aim to shift communication equilibria toward sustained truthful messaging, fostering greater informational resilience in contemporary digital societies.

2 Literature Review

2.1 Defining Mis-, Dis-, and Malinformation

Contemporary scholarship distinguishes between **misinformation**, **disinformation**, and **malinformation**, three facets of what Wardle and Derakhshan (2017) term *information disorder* (Warin, 2024). **Misinformation** refers to false or misleading information spread without malicious intent, often by individuals who genuinely believe it to be true. In contrast, **disinformation** is false information deliberately created or disseminated with the explicit intention of deceiving or causing harm (Wardle & Derakhshan, 2017). For example, an individual unknowingly tweeting an unverified cure for COVID-19 exemplifies misinformation, whereas a state-sponsored campaign strategically spreading false rumors to destabilize elections represents disinformation. The third category, **malinformation**, consists of factual information intentionally shared to cause harm. This includes scenarios such as the deliberate leaking of private emails or personal data intended for harassment or political damage (Wardle & Derakhshan, 2017). A notable instance of malinformation occurred during the 2017 French presidential election when authentic campaign emails of candidate Emmanuel Macron were hacked and disseminated immediately before the election blackout period. Although the emails were genuine, their strategic release aimed explicitly to harm Macron's campaign (Wardle &

Derakhshan, 2017). These distinctions provide a critical framework for understanding the nature and intent behind various forms of harmful information.

Historically, disinformation has played a significant role in statecraft and propaganda, with the term itself, *dezinformatsiya*, originating from early 20th-century Russian practices involving the deliberate planting of false stories (Grimes, 2017). During the Cold War, the Soviet KGB institutionalized disinformation as a component of "active measures," employing forgeries, front organizations, and carefully orchestrated media deceptions. A notorious historical example is **Operation Infektion (Operation "Denver")**, initiated by the KGB in the 1980s, which disseminated the conspiracy theory alleging that HIV/AIDS was a biological weapon created by the United States. This operation began with an anonymous article placed in a pro-Soviet Indian newspaper in 1983, falsely claiming AIDS originated in a U.S. military laboratory (Grimes, 2017). Over subsequent years, Soviet media and allied entities globally amplified this false narrative. By the time the Soviet Union officially retracted these claims in 1987, the conspiracy had already established significant international traction, particularly among marginalized populations predisposed to skepticism toward official narratives. The enduring belief in this misinformation highlights its "sticky" nature, where disinformation persists long beyond its initial dissemination efforts (Grimes, 2017).

2.2 The Digital Media Ecosystem and the Spread of False Information

The advent of the internet and social media has fundamentally transformed the production and dissemination of information, significantly amplifying issues related to misinformation and disinformation (Wardle & Derakhshan, 2017). Several key transformations have been identified in this new media landscape: (a) content creation has become widely accessible and economically inexpensive due to advanced digital tools, allowing virtually anyone to produce high-quality news content, manipulated images, or sophisticated "deepfake" videos; (b) information consumption has shifted from private interactions or dedicated platforms to highly public and social interactions, where users share news and content publicly, influenced by social validation mechanisms such as likes and shares; (c) the speed of information dissemination has accelerated markedly, with real-time propagation of news and rumors outpacing the capabilities of traditional journalistic verification processes; and (d) information is transmitted peer-to-peer with minimal mediation by traditional gatekeepers such as editors or expert analysts, shifting trust dynamics toward one's immediate social network (Wardle & Derakhshan, 2017). These structural shifts have created ideal conditions for the rapid viral spread of false content.

Empirical studies highlight the disproportionate proliferation of false news within digital ecosystems. Vosoughi, Roy, and Aral (2018), in a seminal analysis of Twitter data spanning a decade, demonstrated that false stories consistently spread more broadly, rapidly, and extensively than truthful stories, primarily due to human-driven sharing rather than automated bots. The heightened novelty and emotional resonance of false news appear to facilitate its spread. Silverman's investigation for BuzzFeed News during the 2016 U.S. presidential election similarly revealed that the most engaging fake election news stories outperformed reputable mainstream news stories on Facebook in terms of user engagement, suggesting substantial incentives for content creators to produce sensationalist misinformation due to the attention-driven reward structures inherent in digital platforms (Wardle & Derakhshan, 2017).

From the perspective of economics and information theory, misinformation dissemination can be conceptualized as a market failure within the "marketplace of ideas." Allcott and Gentzkow (2017) assert that misinformation emerges in equilibrium because its production is economically cheaper than accurate, well-researched information, and because consumers face significant costs in verifying information accuracy. Additionally, consumers frequently derive utility from consuming partisan or sensationalist content that aligns with existing biases, reinforcing the demand for misinformation. This phenomenon mirrors the "market for lemons" scenario described by Akerlof, where the prevalence of counterfeit information drives down overall information quality, causing a market equilibrium dominated by misinformation and imposing considerable social costs by impairing accurate public comprehension of factual realities (Allcott & Gentzkow, 2017).

Game-theoretic analyses of communication further elucidate why truthful signaling becomes unstable when incentives diverge. The classic "cheap talk" model proposed by Crawford and Sobel (1982) posits that when communication is costless and non-binding, fully truthful exchanges rarely constitute stable equilibria unless sender and receiver interests align perfectly. When interests significantly diverge, equilibrium outcomes typically devolve into uninformative "babbling," rendering messages meaningless. This framework aptly describes contemporary media ecosystems as large-scale cheap-talk games involving myriad senders with diverse motivations and receivers struggling to ascertain truth amidst conflicting claims. Consequently, stable equilibria frequently involve low information transparency, characterized by misinformation or incoherent discourse, unless institutions or mechanisms align incentives toward accuracy and impose tangible costs on misinformation.

Moreover, the concepts of coordination and common knowledge are crucial in understanding public information dynamics. Schelling (1960) and Lewis (1969) highlight that successful coordination requires mutual acknowledgment of shared knowledge, a condition challenging to achieve in noisy or hostile communication environments, exemplified by the Two Generals Problem. Similarly, Kuran's (1995) theory of preference falsification explains that in environments lacking reliable communication channels, individuals may publicly adhere to false consensus despite privately recognizing inaccuracies, perpetuating misinformation as dominant public narratives. Without mechanisms to establish common knowledge of the truth, misinformation can persist indefinitely as individuals remain reluctant to challenge seemingly accepted falsehoods, creating a stable equilibrium of collective misinformation.

2.3 Ethical and Societal Implications

The prevalence of misinformation poses profound ethical challenges, impacting autonomous decision-making and democratic governance significantly. Reliable information is essential for informed public discourse and maintaining societal trust; when truth becomes elusive, both deteriorate markedly. Ethically, disseminating false information violates fundamental duties of honesty and non-maleficence, particularly when intentionally done (disinformation). It also exploits cognitive biases, undermining individual epistemic autonomy. Philosophers such as Immanuel Kant have historically contended that dishonesty corrupts the essential foundation of meaningful communication and mutual respect. In contemporary mass media contexts, these ethical concerns scale into substantial societal harms.

The weaponization of digital information has yielded tangible negative outcomes, notably during the COVID-19 pandemic, where online disinformation concerning the virus's severity, masks, and vaccines significantly impeded the efforts of public health authorities, contributing to widespread confusion and avoidable health-related harm (Vallor, 2020). Similarly, election-related disinformation has profoundly impacted perceptions regarding democratic integrity and the right of citizens to access truthful, essential information necessary for informed electoral participation.

Responsibility for addressing the misinformation crisis extends across multiple societal sectors, including platform companies, governmental bodies, media professionals, and the general public. Media ethics scholars emphasize the principle of epistemic responsibility, advocating that digital platforms should actively curate content to prevent the intentional amplification of falsehoods, and urging individuals to cultivate more critically engaged media consumption practices. However, such measures introduce further ethical complexities, especially balancing necessary content moderation against the preservation of freedom of expression. This dynamic manifests as a "paradox of tolerance," wherein the unrestricted allowance of disinformation potentially jeopardizes the freedoms central to an open society (Vallor, 2020).

Consequently, solutions must be carefully crafted to foster truthful information dissemination without excessively constraining public discourse or enabling partisan dominance over truth narratives. Achieving this balance remains complex, and the equilibrium analysis presented aims to elucidate the systemic incentives perpetuating information disorder, thereby informing more effective and ethically sound interventions.

2.4 Technology and Deepfakes

On the technological front, the emergence of **deepfakes** and generative artificial intelligence (AI) has significantly reduced the cost of creating realistic yet false content. Deepfakes are convincingly realistic videos or audio recordings produced by generative AI technologies, often leveraging Generative Adversarial Networks (GANs), allowing creators to depict individuals performing actions or uttering statements they have never actually executed. Previously restricted to specialized digital effects studios, these capabilities are now widely accessible, enabling amateurs to generate convincing fabricated content with minimal resources (Chesney & Citron, 2019). This technological democratization undermines traditional verification assumptions—such as the reliability of audiovisual evidence—which historically provided low-cost truth-validation mechanisms. Consequently, this shift further complicates truth verification, diminishing barriers previously encountered by disseminators of disinformation.

Instances of deepfake deployment illustrate their practical implications. For example, during the 2024 U.S. primary elections, voters received robocalls featuring fabricated audio of President Joe Biden discouraging participation, an attempt aimed explicitly at voter suppression (Marantz, 2024). This event, quickly revealed as artificially generated content, nonetheless demonstrated the tangible threats posed by deepfake technology to democratic processes. Experts have warned of more severe scenarios, particularly around critical electoral events, where strategically timed deepfake content could significantly influence public opinion or generate widespread disorder (Chesney & Citron, 2019).

The potential misuse of deepfake technology for foreign influence operations and other malicious purposes poses profound ethical and policy challenges. If verifying truth becomes prohibitively costly or unreliable, the societal consequences could include a generalized epistemic nihilism or radical relativism, where public trust in the authenticity of information collapses entirely. Reflecting these risks, our theoretical framework incorporates the probability of message reliability as an endogenous variable affected by technological advancements and adversarial interference. Essentially, as the technological cost of fabricating content increases or verification methods improve and become more affordable, the probability of successful truthful communication correspondingly rises.

So, interdisciplinary scholarship characterizes the contemporary information ecosystem as structurally advantageous for false information propagation. Historical and political analyses document longstanding exploitation of communication channels by deceptive actors; technological examinations underscore the amplification and realism provided by digital tools and AI; economic and game-theoretic models clarify how cost and incentive disparities result in misinformation equilibria; and ethical discussions underscore the moral urgency and societal stakes involved. Building upon these insights, we now present a formal theoretical model informed by the "generals problem" analogy, examining strategic dynamics of information dissemination amidst cost constraints and uncertainty.

This section proposes a self-contained revision of the Bayesian model so that the misinformation-equilibrium condition no longer suffers from a division-by-zero issue. The core strategic setting remains informed by the "Two Generals Problem," with interpretive parallels to Rubinstein's (1989) "Electronic Mail Game," focusing on communication reliability and common knowledge constraints. The modifications center on ensuring that the payoffs used in the denominator of the misinformation stability condition are not identical, thereby avoiding undefined expressions.

The entire modified model appears below. All references conform to APA style and draw on standard game-theoretic notation.

3 Theoretical Model

The framework draws on a coordination game metaphor, wherein two content producers (Player 1 and Player 2) must decide whether to invest in Truthful Messaging (T) or pursue Misinformation Messaging (M). Players are uncertain about one another's cost of truth-telling, which may be high or low, and rely on beliefs regarding their counterpart's likelihood of belonging to each cost type. The model aims to demonstrate how even small uncertainties in beliefs, combined with communication frictions, can lead to persistent misinformation equilibria.

3.1 Bayesian Model Setup and Assumptions

The setup begins with two content producers, indexed by $i \in \{1, 2\}$. Each player has two actions: T (truthful) or M (misinformation). Each player's private type concerns the cost of truth-telling: there is a low-cost type L and a high-cost type H. The prior belief that any given player is type L is q, and the belief that they are type H is $1 - q$. Players' payoffs hinge on their own action, the other player's action, and the reliability of communication with the audience.

Assumption A1 (Cost of Truth). Each player bears a cost c_i if they choose T, with $c_i = c_L$ for the low-cost type and $c_i = c_H$ for the high-cost type. Costs satisfy $0 < c_L < c_H$. Misinformation has negligible direct cost.

Assumption A2 (Payoffs from Audience Engagement). In a fully successful truthful strategy, each player earns B_T. In a fully successful misinformation strategy, each player earns B_M. Empirical realities often imply $B_M \geq B_T$ in the short run, although this advantage may erode over time. When truthfulness or misinformation is only partially successful, payoffs adjust to B'_T or B'_M. If both players choose M, saturation or regulatory backlash reduces the payoff to B''_M.

Assumption A3 (Reliability of Communication). The probability of effectively reaching and persuading the intended audience is $p < 1$. This probability is initially exogenous but can shift endogenously if, for instance, widespread misinformation undermines overall audience trust.

Assumption A4 (Bayesian Uncertainty). Each player holds subjective beliefs about the other's cost type. They do not observe each other's actual costs but form strategies based on q. The game proceeds under incomplete information, resembling classic Bayesian coordination problems in which equilibrium strategies must be best responses given beliefs over the opponent's type.

3.2 Payoff Definitions

Definition 1 (Expected Payoff). If Player i of type $t_i \in \{L, H\}$ chooses T, their expected payoff is

$$U_i(T \mid t_i) \;=\; pB_T \;-\; c_{t_i}, \tag{1}$$

while choosing M yields

$$U_i(M \mid t_i) \;=\; pB_M. \tag{2}$$

These benchmark expressions assume fully successful outcomes for the chosen messaging.

Definition 2 (Mixed or Partially Successful Strategies). In scenarios where truth or misinformation only partially succeeds, the payoffs adjust to:

$$U_{\text{truthful}} \;=\; pB'_T \;-\; c_{t_i}, \quad U_{\text{misinfo}} \;=\; pB'_M, \tag{3}$$

where $B'_T < B_T$ and $B'_M > B_M$ or $B'_M < B_M$, depending on the specific modeling emphasis. The critical point is that partial success changes both truthful and misinformation payoffs.

Definition 3 (Mutual Misinformation Payoffs). If both players choose M and the audience becomes saturated by low-credibility information, each player's payoff diminishes to:

$$U_i(M, M) \;=\; pB''_M, \tag{4}$$

where $B''_M \leq B_M$. This condition captures reputational or regulatory penalties when misinformation fully crowds out truthful content.

3.3 Lemmas and Propositions

Lemma 1 (Truthful Messaging Condition). Truthful messaging can dominate misinformation for a given player if the incremental expected benefit exceeds the cost of truth-telling. Specifically, if one or both players consider T versus M, the condition to prefer T is:

$$p\,[q\,(B_T - B'_M) \;+\; (1-q)\,(B'_T - B'_M)] \;\geq\; c_{t_i}. \tag{5}$$

Lemma 2 (Misinformation Equilibrium Condition). If a player's cost of truth-telling is sufficiently high relative to the incremental gains, the incentive to deviate to T is low. One version of this threshold condition is:

$$p\,[q\,(B'_T - B''_M) \;+\; (1-q)\,(B'_T - B''_M)] \;\leq\; c_{t_i}. \tag{6}$$

In practice, the exact form depends on how partial or mutual misinformation is modeled relative to baseline truthful payoffs.

Proposition 1 (Bayesian Nash Equilibrium). A Bayesian Nash Equilibrium (BNE) arises when each player's strategy maximizes that player's expected payoff given their beliefs regarding the other's type. Formally, if s_i denotes Player i's strategy, then in equilibrium

$$U_i(s_i^*(t_i),\, s_j^*(t_j) \,\big|\, q) \;\geq\; U_i(s_i,\; s_j^*(t_j) \,\big|\, q), \quad \forall\, s_i \neq s_i^*(t_i). \tag{7}$$

Each type t_i selects the strategy $s_i^*(t_i)$ that optimizes expected payoff against $s_j^*(t_j)$.

3.4 Theorems

Theorem 1 (Existence of Truthful Equilibrium). A fully truthful equilibrium emerges if individuals believe the opponent is likely to be a low-cost truth-teller and thus expect high returns on truthful coordination. Formally, the condition ensuring Player i prefers T over M can be written:

$$q \;\geq\; \frac{\frac{c_{t_i}}{p} \;-\; (B'_T - B'_M)}{[(B_T - B'_M) \;-\; (B'_T - B'_M)]}. \tag{8}$$

If $B_T - B'_M$ and $B'_T - B'_M$ differ appropriately, this boundary on q is well-defined and summarizes when truthful messaging dominates from the viewpoint of cost-benefit trade-offs.

Theorem 2 (Misinformation Stability). A misinformation equilibrium is stable if beliefs about the opponent's likelihood of incurring low costs for truth-telling are sufficiently low and if the cost of truth-telling remains high. To avoid subtracting identical payoff terms and thereby dividing by zero, the denominator must compare distinct scenarios, such as a fully truthful payoff and a mutual-misinformation payoff. One illustrative form is:

$$q \;\leq\; \frac{\frac{c_{t_i}}{p} \;-\; (B'_T - B''_M)}{[(B_T - B''_M) \;-\; (B'_T - B''_M)]} \;=\; \frac{\frac{c_{t_i}}{p} \;-\; (B'_T - B''_M)}{B_T \;-\; B'_T}. \tag{9}$$

In this specification, $(B_T - B''_M)$ corresponds to the difference between a fully truthful outcome and mutual misinformation, whereas $(B'_T - B''_M)$ represents a partially truthful outcome compared to mutual misinformation. The difference between these two payoff gaps is $(B_T - B'_T)$, which is nonzero if fully truthful and partially truthful payoffs differ. The inequality indicates that players find misinformation to be a stable strategy if their belief in the other being a low-cost truth-teller remains below a critical level.

3.5 Extensions

3.5.1 Audience Heterogeneity

Introducing audience segments—discerning (D) and less discerning (L)—with probabilities p_D and p_L, modifies payoffs:

$$U_i(T|t_i) = \alpha p_D B_T + (1-\alpha) p_L B_T - c_{t_i}, \quad U_i(M|t_i) = \alpha p_D B_M + (1-\alpha) p_L B_M \tag{10}$$

Critical threshold α determining truthful messaging viability is:

$$\alpha = \frac{-B_M p_L + B_T p_L - c}{(B_M p_D - B_M p_L - B_T p_D + B_T p_L)} \tag{11}$$

3.5.2 Algorithmic Amplification

Algorithmic bias ($\gamma > 1$) towards misinformation adjusts equilibrium conditions:

$$p\left[q(B_T - \gamma B'_M) + (1-q)(B'_T - \gamma B'_M)\right] \geq c_{t_i} \tag{12}$$

3.5.3 Dynamic and Repeated Interactions

Repeated interactions introduce discount factor δ. Truthful cooperation emerges if:

$$\frac{pB_T - c_{t_i}}{1-\delta} \geq pB'_M + \delta \frac{pB''_M}{1-\delta} \tag{13}$$

This demonstrates reputational dynamics and the critical role of long-term incentives in stabilizing truthful communication.

Under the specified theoretical conditions, misinformation does indeed provide immediate, higher short-term payoffs due to lower production costs and higher initial audience engagement. Consequently, absent future considerations or repercussions, misinformation strategies dominate truthfulness strictly from a short-term, static perspective.

However, when repeated interactions and long-term considerations, such as reputation or sustained audience trust, are integrated into the model, the picture becomes significantly more complex. In particular, the introduction of the discount factor (δ)—representing the extent to which players value future payoffs—alters the strategic landscape fundamentally. Specifically, if future interactions are sufficiently valued (i.e., δ is high enough), truthful messaging becomes strategically optimal. This scenario arises precisely because sustained misinformation, while attractive initially, erodes audience trust over time, reducing future payoffs.

Hence, the model suggests that misinformation is not universally preferable but conditionally so, depending critically upon:

- **Short-term vs. long-term incentives:** Immediate benefits versus sustained audience trust and engagement.
- **Player time preferences:** Represented by the discount factor δ, indicating how heavily future outcomes weigh in strategic decision-making.
- **Interaction frequency and reputational mechanisms:** Repeated interactions encourage cooperative (truthful) strategies through credible threat of punishment or reputational sanctions.

Therefore, the equilibrium condition highlights that misinformation prevails primarily when actors undervalue future repercussions (low δ), whereas truthful communication emerges as strategically rational under conditions of long-term interaction and reputation concerns (high δ).

3.6 Extended Polynomial Theoretical Model

This section generalizes the earlier binary model to a multinomial strategy environment while incorporating the revised payoff structure that avoids any indeterminate expressions. In doing so, it captures a more nuanced array of messaging choices, reflecting varying degrees of accuracy and associated costs, together with polynomially evolving audience perceptions. The approach preserves the logic that differentials between payoff terms remain distinct.

3.6.1 Assumptions

The model considers two content producers, indexed by $i \in \{1, 2\}$, each choosing among three strategies: Highly Accurate Messaging (H), Partially Accurate Messaging (P), and Misinformation Messaging (M). As in the revised framework, marginal costs of accuracy are strictly increasing, so that

$$0 < c_M < c_P < c_H, \tag{14}$$

while audience-based payoffs satisfy

$$B_H > B_P > B_M. \tag{15}$$

Choosing H entails the highest cost c_H but yields the highest potential benefit B_H. Partially accurate content (P) incurs an intermediate cost c_P and offers a moderate benefit B_P. Misinformation (M) is inexpensive (c_M) but garners only a lower benefit B_M. Communication reliability probabilities also vary with accuracy:

$$p_H > p_P > p_M > 0, \tag{16}$$

reflecting the assumption that more accurate messaging is more likely to succeed in earning and maintaining audience trust.

Incorporating non-binary messaging choices places players in a Bayesian coordination context with incomplete information regarding each other's underlying cost type. Consistent with the revision that precludes zero-denominator outcomes, the payoff functions and subsequent equilibrium conditions now distinguish among three distinct scenarios (H, P, M), ensuring that the differences in their respective payoffs remain nonidentical under any pairwise comparison.

3.6.2 Polynomial Payoff Structure

Under these assumptions, the expected payoff to player i from choosing strategy $s \in \{H, P, M\}$ is

$$U_i(s) = p_s \, B_s \; - \; c_s. \tag{17}$$

This expression is reminiscent of the baseline model but accommodates three strategies. Crucially, B_s and c_s vary strictly across H, P, M, so payoff differences remain well-defined and nonzero whenever strategies differ.

3.6.3 Audience Feedback and Polynomial Dynamics

The probability p_s that a message of type s is believed or adopted by the audience is no longer exogenous. Instead, it is subject to polynomial adjustments based on aggregate observed frequencies of the three strategies. Let σ_H, σ_P, and σ_M denote the proportions with which strategies H, P, and M are played, summing to unity:

$$\sigma_H + \sigma_P + \sigma_M = 1. \tag{18}$$

Drawing on the revised perspective that disallows subtracting identical payoff terms, each reliability function for $s \in \{H, P, M\}$ takes on a polynomial form in $\sigma_H, \sigma_P, \sigma_M$. One illustrative specification is

$$
\begin{aligned}
p_H &= \alpha_H\,\sigma_H^2 + \beta_H\,\sigma_H\,\sigma_P + \gamma_H, \\
p_P &= \alpha_P\,\sigma_P^2 + \beta_P\,\sigma_P\,\sigma_M + \gamma_P, \\
p_M &= \alpha_M\,\sigma_M^2 + \beta_M\,\sigma_H\,\sigma_M + \gamma_M,
\end{aligned} \tag{19}
$$

where the coefficients $\alpha_s, \beta_s, \gamma_s$ govern how strongly audience perceptions respond to observed frequencies of each strategy. The polynomial specification acknowledges that reputational effects and trust-building, or erosion, can exhibit nonlinear patterns, and it preserves distinct functional forms for p_H, p_P, and p_M. This ensures that each payoff gap—such as $(p_H B_H - c_H) - (p_M B_M - c_M)$—remains well-defined and nontrivial.

3.6.4 Polynomial Equilibrium Conditions

In a Bayesian Nash setting with three strategies, a mixed-strategy equilibrium involves players choosing $\sigma_H, \sigma_P, \sigma_M$ so that each player is indifferent among H, P, and M, given that the other player does the same. Formally, equilibrium requires that

$$
U_i(H \mid \sigma_j) \;=\; U_i(P \mid \sigma_j) \;=\; U_i(M \mid \sigma_j), \tag{20}
$$

where σ_j refers to the strategy distribution used by the other player (or the population, in a large-population interpretation). Substituting the polynomial reliability functions yields

$$
\begin{aligned}
(\alpha_H\sigma_H^2 + \beta_H\sigma_H\sigma_P + \gamma_H)B_H \;-\; c_H &= \;(\alpha_P\sigma_P^2 + \beta_P\sigma_P\sigma_M + \gamma_P)B_P \;-\; c_P \\
&= \;(\alpha_M\sigma_M^2 + \beta_M\sigma_H\sigma_M + \gamma_M)B_M \;-\; c_M.
\end{aligned} \tag{21}
$$

Expanding these expressions produces a system of polynomial equations. Imposing the constraint $\sigma_H + \sigma_P + \sigma_M = 1$ completes the system, ensuring that the vector $(\sigma_H, \sigma_P, \sigma_M)$ is a valid probability distribution.

To eliminate any prospect of undefined terms, the payoff differences across (H, P), (P, M), and (H, M) must not collapse to an identical expression in both numerator and denominator. Because each strategy's cost c_s and benefit B_s is strictly distinct, and because p_s is a distinct polynomial function of $\sigma_H, \sigma_P, \sigma_M$, the equilibrium conditions yield well-defined polynomial equations rather than fractions that risk division by zero.

Proposition 2 (Existence of Polynomial Bayesian Nash Equilibria)

A Polynomial Bayesian Nash Equilibrium (PBNE) exists if there is at least one solution $(\sigma_H^*, \sigma_P^*, \sigma_M^*)$ to the equilibrium indifference system of polynomial equations:

$$
\begin{aligned}
(\alpha_H \sigma_H^2 + \beta_H \sigma_H \sigma_P + \gamma_H) B_H - c_H - \ & (\alpha_P \sigma_P^2 + \beta_P \sigma_P \sigma_M + \gamma_P) B_P + c_P = 0, \\
(\alpha_P \sigma_P^2 + \beta_P \sigma_P \sigma_M + \gamma_P) B_P - c_P - \ & (\alpha_M \sigma_M^2 + \beta_M \sigma_H \sigma_M + \gamma_M) B_M + c_M = 0, \\
\sigma_H + \sigma_P + \sigma_M - \ & 1 = 0,
\end{aligned}
\tag{22}
$$

subject to $\sigma_H, \sigma_P, \sigma_M \geq 0$. A nontrivial PBNE emerges when these equations are simultaneously satisfied, capturing a stable mix of strategies under the assumed polynomial feedback mechanisms. Because the underlying costs and benefits differ strictly across strategies, and the reliability terms $\{p_H, p_P, p_M\}$ are governed by distinct polynomial expressions, this system typically has a non-degenerate solution set.

3.6.5 Interpretation and Implications

This extended model uncovers a richer landscape of strategic behavior compared to a binary setup. By explicitly introducing a partially accurate strategy, the analysis underscores how intermediate choices may become attractive under certain parameter configurations, particularly when full accuracy's higher cost is not justified by its reliability advantage, but players remain wary of the reputational downsides of outright misinformation. The polynomial mapping from strategy profiles to audience credibility captures complex trust dynamics that may amplify or mitigate these incentives in nonlinear ways.

This illuminates how mixed-strategy equilibria can persist in settings where no single option strictly dominates once the evolving beliefs of the audience are factored into the cost–benefit trade-offs. These results align with broader insights on the roles of reputation, trust, and strategic misinformation in media-rich environments.

4 Empirical and Illustrative Cases

To demonstrate how the extended polynomial Bayesian coordination model and its tripartite strategy set (Highly Accurate, Partially Accurate, Misinformation) illuminate real-world dynamics, this section reviews empirical examples that highlight the interplay among cost structures (c_H, c_P, c_M), audience reliability functions $\{p_H, p_P, p_M\}$, audience composition parameters, and the resulting payoffs (B_H, B_P, B_M). These cases illustrate how partial accuracy (e.g., selective or mixed messaging) can emerge as strategically viable, especially when polynomial audience responses amplify or diminish credibility in nonlinear ways.

4.1 Case 1: Cold War Disinformation Campaign (Operation Infektion)

Operation Infektion, orchestrated by Soviet intelligence agencies, involved spreading the falsehood that HIV/AIDS originated in a U.S. military laboratory. From a polynomial perspective, misinformation (M) was exceedingly inexpensive (i.e., $c_M \approx 0$) relative to more thorough, verifiable strategies (H) or even moderately accurate ones (P). Because widespread distrust in opposing media outlets reduced the effective reliability for highly accurate messages (p_H), the resulting payoff to Misinformation ($p_M B_M - c_M$) could outstrip that of more credible approaches, particularly if audience belief did not significantly penalize falsehoods. Over time, iterative exposure to conspiratorial content potentially *reinforced* p_M in a nonlinear (polynomial) manner, solidifying an equilibrium dominated by low-cost, high-impact misinformation.

4.2 Case 2: Mainstream Media vs. Clickbait Websites (2016 U.S. Election)

During the 2016 U.S. election, major news organizations incurred considerable verification and fact-checking costs (c_H), whereas clickbait or fringe websites faced negligible costs (c_M). In certain cases, these websites also adopted a partially accurate but sensational approach (P). Algorithmic amplification—captured as polynomial augmentations to σ_M or σ_P in reliability updates—elevated payoff components B_M or B_P. This dynamic compressed the gap between more rigorous reporting (H) and sensational or entirely misleading content. As a result, even mainstream outlets sometimes gravitated toward more provocative, less fully verified reporting (moving from H to P) in order to retain audience attention, consistent with the extended model's prediction that a rise in payoff for intermediate or misinformation strategies can pull equilibrium away from high-accuracy content.

4.3 Case 3: COVID-19 "Infodemic" and Health Misinformation

The early stages of the COVID-19 pandemic showcased marked imbalances between high-effort, rigorously sourced information (H) and rapidly produced conspiracy narratives (M). Although official health agencies attempted to communicate accurate updates, the evolving and sometimes uncertain nature of scientific findings effectively lowered σ_H and inhibited p_H. Meanwhile, low-cost or partially accurate messages (P) circulated virally, and outright misinformation (M) often reaped immediate visibility (B_M). Polynomial feedback processes—where the more a certain narrative appeared, the more the audience engaged with it—reinforced mistrust in scientific updates and heightened reliability for dubious content, creating a stable outcome skewed toward misinformation or half-truths. Policy interventions, such as removal of demonstrably false claims and bolstering credible sources, can be viewed as attempts to reconfigure the polynomial parameters $\alpha_s, \beta_s, \gamma_s$ and tilt reliability $\{p_H, p_P, p_M\}$ back in favor of truthfulness.

4.4 Case 4: Deepfakes and Political Manipulation

Emerging deepfake technology, exemplified by the fabricated 2022 Zelensky video, underscores how dramatic reductions in misinformation production costs (c_M) and sophisticated editing tools further corrupt audience reliability perceptions. If deepfake prevalence increases σ_M, the polynomial term $\alpha_M \sigma_M^2$ may surge, expanding the perceived plausibility (or at least the visibility) of misinformation. Audience members become more hesitant to trust highly accurate content (H), effectively depressing p_H. Under these conditions, the extended model predicts that misinformation (M) may dominate unless costly detection technologies or legal frameworks elevate c_M or shift payoffs such that persistent reliance on deepfakes becomes less viable. Countermeasures—such as real-time verification tools—attempt to realign audience beliefs via adjusting the feedback terms in favor of more accurate strategies.

4.5 Case 5: Coordination Failures in Truth-Telling (Whistleblowers and #MeToo)

The #MeToo movement illustrates how partial accuracy or silence can persist when truthful disclosure (H) is initially too costly or risky ($c_H \gg 0$). Early whistleblowers, facing legal and social repercussions, found low reliability (p_H) for their accusations in a disbelieving environment. Thus, a system dominated by misinformation, dismissals, or partial truths prevailed. As more disclosures accumulated, polynomial feedback shifted the audience's willingness to believe victims—i.e., σ_H rose, which in turn boosted p_H. The extended model predicts that, once a critical threshold is passed, additional revelations become self-reinforcing, lowering the effective cost of telling the truth and elevating truthful messaging's payoff. The transformation from an equilibrium of silence or misinformation to one of widespread credible disclosures exemplifies how dynamic, repeated interactions can gradually restructure the parameters $\alpha_s, \beta_s, \gamma_s$ and push the system toward a new equilibrium favoring more accurate content.

These cases attest to the explanatory power of the extended polynomial model for complex informational environments. The intricate, nonlinear feedback loops captured in $\{p_H, p_P, p_M\}$ clarify how small changes in reliability, cost, or payoff factors can prompt substantial and sometimes abrupt shifts in equilibrium strategies. Whether dealing with historical propaganda, high-stakes electoral dynamics, global health crises, emergent deepfake technologies, or collective whistleblowing efforts, the model's central lesson remains the same: stable equilibria hinge on each strategy's distinct costs and benefits, combined with how audiences recursively update their beliefs in response to observed messaging patterns. Empirically, interventions aimed at boosting p_H (e.g., through verifiability) or raising the cost of misinformation (c_M) can dislodge entrenched misinformation equilibria, highlighting the practical and policy-relevant implications of this extended theoretical framework.

5 Discussion

This study applies an extended polynomial Bayesian coordination model to investigate how and why misinformation arises and persists in contemporary media ecosystems. Beyond the traditional two-strategy framework, the model incorporates Highly Accurate (H), Partially Accurate (P), and Misinformation (M) messaging, along with polynomially evolving audience reliability functions $\{p_H, p_P, p_M\}$. This approach emphasizes the interplay between the cost of producing accurate content $\{c_H, c_P, c_M\}$, the reliability of communication channels, payoffs for different messaging strategies $\{B_H, B_P, B_M\}$, algorithmic amplification γ, and the composition of the audience α. The ensuing discussion interprets how these factors shape and sometimes reinforce misinformation equilibria, as well as how policymakers, digital platforms, and broader societal norms can mitigate disinformation pressures.

5.1 Persistence of Misinformation

The extended polynomial framework shows that misinformation can be self-sustaining, not merely due to ethical lapses by individual content producers, but through systemic economic and reputational incentives. When the cost of highly accurate communication is high ($c_H \gg 0$), and audiences fail to robustly reward or even properly identify credible content (p_H remains low), producers may rationally prefer misinformation (M) or a partially accurate strategy (P). Misinformation's persistence often reflects an equilibrium in which polynomially updating audience beliefs do not sufficiently penalize low-quality content. Therefore, rational actors gravitate toward cheaper, more easily disseminated messaging, reinforced by non-linear feedback processes that amplify recurring exposure to sensationalized or false stories. This outcome highlights how misinformation's prevalence stems from a confluence of cost disparities, audience uncertainty, and algorithmic design choices rather than isolated acts of intentional deceit.

5.2 Implications for Digital Platforms

Digital platforms serve as key facilitators, magnifying or attenuating strategic incentives through their distribution algorithms, engagement metrics, and moderation policies. Polynomial audience updates mean that repeated exposure to high-accuracy messages (H) can slowly improve p_H, whereas recurrent sensational or misleading content can rapidly boost the effective reach of misinformation (p_M). Algorithmic amplification γ compounds these effects: even a slight bias toward engagement-based recommendation can shift the system toward a heavier reliance on partial accuracy (P) or misinformation (M). Hence, platforms must recognize how their design choices—and the feedback loops they create—reshape the payoffs $\{B_H, B_P, B_M\}$. Transparent algorithms, voluntary labeling, and constructive user feedback loops can elevate (p_H) and partially neutralize misinformation equilibria. Likewise, making credible content more visible effectively reduces c_H or c_P from the producer's perspective, encouraging higher-accuracy strategies even when short-term returns on sensationalized content appear large.

5.3 Recommendations for Policymakers and Governance

Policymakers can intervene to shift equilibrium conditions in favor of higher-accuracy messaging by recalibrating the economic incentives that currently privilege misinformation. In particular, regulatory frameworks that hold persistent purveyors of demonstrably false content accountable can raise c_M, eroding the payoff advantage of misinformation $(p_M B_M - c_M)$. For instance, imposing fines or other legal repercussions on systematic misinformation campaigns may push producers to adopt at least partially accurate strategies to avoid punitive measures. Public education and media literacy initiatives that strengthen users' ability to discern content quality can raise reliability perceptions for truth-telling over time, thereby increasing (p_H) and rendering misinformation less attractive in equilibrium. Consistent moderation guidelines and disclosures regarding sponsorship or content origin likewise enhance audience trust, diminishing the appeal of low-cost, high-amplification misinformation campaigns.

5.4 Managing Residual Misinformation

Even with strategic interventions, complete eradication of misinformation remains unlikely: polynomial feedback loops and inherent cost-benefit trade-offs ensure that some degree of false or partially false content persists. Ongoing measures, such as fact-checking consortia, real-time verification tools, and public awareness campaigns, can effectively mitigate further entrenchment of misinformation equilibrium. By reducing the perceived credibility of suspect content, these tools temper the nonlinear growth in p_M. In parallel, partnerships between platforms, journalists, and civil society can elevate (p_H) or (p_P) by exposing factual inaccuracies promptly, helping realign user perceptions before unverified claims become normalized.

5.5 Ethical and Normative Considerations

Although the model foregrounds economic and reputational incentives, ethical norms and professional standards substantially influence messaging choices. Content creators motivated by public interest or journalistic ideals may select higher-accuracy strategies (H) even when short-term costs exceed potential short-term payoffs. Institutionalizing ethical guidelines and nurturing a professional culture that respects and rewards factual precision can shift the audience's baseline credibility γ_H (i.e., the parameter capturing minimal reliability for truthful content). In this sense, moral and normative frameworks complement policy and market-based reforms by setting a foundation upon which producers and consumers alike come to expect and value accuracy.

5.6 Insights from the Two Generals Metaphor

The original Two Generals Problem emphasizes how coordination can fail when communication lacks reliability. In this context, stable equilibrium in highly accurate messaging depends on shared trust and the repeated verification of source credibility. If polynomial audience beliefs are not anchored by consistent, accessible evidence or trusted institutional oversight, repeated attempts at disseminating truth can falter similarly to the generals' endless exchange of messages. By contrast, if platforms, policymakers, and producers jointly establish reliable feedback channels—through frequent user education, robust fact-checking, and transparent algorithms—this fosters common knowledge about content accuracy. Such efforts substantially raise σ_H, thereby elevating (p_H) and ultimately shifting the system away from a misinformation equilibrium.

6 Conclusion

This article demonstrates, through a rigorous Bayesian game-theoretic analysis, that the persistence and proliferation of misinformation within contemporary digital environments emerge from systemic incentives rather than exclusively from individual malice or ignorance. By formally incorporating the core parameters of truth-telling costs (c_i), communication reliability (p), payoffs for truthful and misleading content (B_T, B_M), audience composition (α), and algorithmic amplification (γ), the model clarifies the structural conditions under which misinformation becomes a stable, if undesired, equilibrium. When the economic and cognitive burdens of accurate content are substantially higher than the minimal requirements for producing misinformation, rational actors unsurprisingly gravitate toward the latter, particularly in an environment where low communication reliability reduces the returns to truthfulness. Amplification mechanisms amplify this bias by rewarding sensational or misleading content with greater visibility and engagement.

The heterogeneity of audience composition likewise proves critical. If many audience members are prone to cognitive biases or attracted to sensational claims, misinformation strategies garner immediate payoff advantages. Conversely, an audience comprising a higher share of discerning participants promotes more reliable and accurate equilibria by increasing the effective payoff to truth-telling. Empirical examples spanning Cold War propaganda, electoral disinformation, public health crises, deepfake technologies, and collective action movements such as #MeToo strongly corroborate the theoretical predictions. Cases in which misinformation flourished predominantly displayed favorable structural conditions for falsehoods, whereas successful mitigation strategies shifted costs or augmented communication reliability, realigning incentives in favor of verifiable information.

Countering misinformation on a sustained basis thus necessitates integrated and multifaceted strategies. Digital platforms can reduce the relative attractiveness of misleading content by implementing more transparent and accountable recommendation systems that elevate factual reporting and dampen amplification of unverified claims. Policymakers can complement these efforts through regulatory guidelines enforcing transparency, sanctioning systematic disinformation, and bolstering the economic viability of high-quality journalism. Education and media literacy initiatives remain indispensable to strengthening audience discernment and sustaining a long-term equilibrium supportive of truth. These measures collectively highlight that structural change—rather than isolated or purely ethical correctives—represents the most robust pathway to undermining entrenched misinformation.

Future research efforts might quantify parameters such as cost levels and audience dynamics across diverse informational environments, enhancing predictive precision and guiding targeted interventions. Incorporating richer network-based approaches, including the influence of peer effects and echo chambers, would deepen empirical realism and refine policy prescriptions. Advancing these directions promises to reinforce trustworthy media ecosystems that foster democratic discourse and social well-being.

References

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications, 5*, Article 15. https://doi.org/10.1057/s41599-019-0224-y

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs, 98*(1), 147–155.

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE, 12*(5), Article e0175799. https://doi.org/10.1371/journal.pone.0175799

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica, 50*(6), 1431–1451. https://doi.org/10.2307/1913390

Gray, J. N. (1978). Notes on database operating systems. In R. Bayer, R. M. Graham, & G. Seegmüller (Eds.), *Operating systems: An advanced course* (pp. 393–481). Springer-Verlag. https://doi.org/10.1007/3-540-08755-9_9

Grimes, D. R. (2017, June 14). Russian fake news is not new: Soviet AIDS propaganda cost countless lives. *The Guardian.* https://www.theguardian.com/science/blog/2017/jun/14/russian-fake-news-is-not-new-soviet-aids-propaganda-cost-countless-lives

Marantz, A. (2024). How deepfakes and AI memes affected global elections. *NPR.* Retrieved from https://www.npr.org

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under 'almost common knowledge'. *American Economic Review, 79*(3), 385–391.

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest.* Yale University Press.

Vallor, S. (2020). Social networking and ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 Edition). Retrieved from https://plato.stanford.edu/entries/ethics-social-networking/

Van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology, 11*, Article 566790. https://doi.org/10.3389/fpsyg.2020.566790

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking.* Council of Europe Report. Retrieved from https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/

Warin, T. (2024). *Disinformation in the digital age: Impacts on democracy and strategies for mitigation* (2024PR-03). CIRANO. https://doi.org/10.54932/GQWB1497