Persuasion under the Influence of Fake News: Competitive Persuasion under Source Uncertainty*

Sangjun Yea† Daeyoung Jeong‡

August 21, 2026

Abstract

We study Bayesian persuasion with source uncertainty: a receiver observes one message but cannot tell whether it came from a related news sender, who commands an informative technology, or a fake news sender, whose signals are state-independent. Using a Rothschild-Stiglitz approach, we characterize equilibrium and identify the threshold prominence at which fake news becomes persuasive. Prominent fake news dilutes informative signals, but a sufficiently credible sender preserves influence by raising precision. This self-disciplining effect raises welfare: paradoxically, within the persuasive regime, greater prominence of fake news improves the informational environment and leaves voters better informed ex-ante.

Keywords: Fake News, Social Media, Bayesian Persuasion, Sequential Persuasion.

JEL Classification Codes: D72, D82, D83, L82

*We would like to thank James Peck, Yaron Azrieli, Huanxing Yang, Jinsoo Bae, Paul Healy, Sam-Ho Lee, Euncheol Shin, and Yeochang Yoon for useful suggestions. We also thank participants at the 2018 Midwest Economic Theory Conference in Nashville, 2022 The Korean Academic Society of Industrial Organization Summer Meeting, 2023 Korea's Allied Economic Associations Annual Meeting, 2023 Symposium on Economics and Finance in the Asia-Pacific, and the members of Economic Theory Lunch Group and Microeconomics Lunch Seminar at the Ohio State University, the Microeconomics Seminar at National Taiwan University for helpful comments. This paper is based on Sangjun Yea's Ph.D. dissertation at The Ohio State University.

†First Author: Korea Institute for International Economic Policy; E-mail: sjyea@kiep.go.kr.

‡Corresponding Author: School of Economics, Yonsei University, 50 Yonsei-Ro, Seodaemun-Gu, Seoul 03722, South Korea. Email: daeyoung.jeong@gmail.com.

1 Introduction

Social media sites such as Facebook and X (Twitter) have emerged as some of the most influential news media outlets in recent years.[1] The proliferation of information sharing among social media users through their networks enables a vast number of individuals to access the latest information more readily (Doerr et al., 2012). However, the pathways through which news circulates on social media are rarely traceable by individual users, leading to the frequent obscurity of information sources in the process of information propagation.

The uncertainty about the origin of news on social media platforms has raised substantial concerns regarding disinformation, as it potentially facilitates the production of fake news and deteriorate the quality of information circulated through these channels. This issue is particularly acute in political communication. The emergence of fake news from uncertain sources in the context of electoral competitions poses significant challenges. It may cause voters, who receive news from social media, to hold political misperceptions about social issues or political figures, leading to less informed decisions in elections.[2] Despite the significance of this issue, there remains a lack of comprehensive understanding regarding the interplay between the production of fake news, the uncertainty of news origin in social media, and its impact on political decision-making.

Motivated by these questions, we examine a sequential persuasion game involving two senders and a continuum of voters, under the uncertainty about the origin of news. Two senders sequentially commit to their reporting strategies: First, a sender, referred to as the "related news sender" (she), commits to a reporting strategy capable of generating a news story correlated with the state of nature. Subsequently, after observing the related news sender's strategy, another sender, known as the "fake news sender" (he), commits to a strategy that produces a news story independent of the state.It is important to note that the fake news sender in our model differs from simple biased or slanted media outlets. Our fake news sender "lack[s] the news media's editorial norms and processes for ensuring the accuracy and credibility of information" (Lazer et al., 2018).[3]

After the two senders select their reporting strategies, a true state is realized, and each

[1] According to a 2017 survey, 67 percent of U.S. adults reported receiving news from social media sites (Shearer and Gottfried, 2017).

[2] A notable debate revolves around the impact of fake news in social media on the 2016 U.S. presidential elections. See Allcott and Gentzkow (2017) and Hindman and Barash (2018) for further discussion.

[3] In this context, the concept of a fake news sender in our study aligns with the depiction of fake news producers in Allcott and Gentzkow (2017). To delineate the nature of fake news clearly in our research, we restrict its definition to "fabricated news," which is characterized by a lack of factual basis while mimicking the format of genuine news articles to gain legitimacy. The term "fake news" has been used in various contexts in prior academic work. For a comprehensive typology of fake news definitions, see Tandoc Jr. et al. (2018).

sender produces a news story according to their chosen strategies. We assume that all voters (they) receive only one sender's news, with the probability of receiving it determined by the relative prominence of related news to fake news. However, voters are unaware of the news's origin. Upon updating their beliefs based on the received news, each voter simultaneously casts their vote for one of two alternatives. The alternative that garners the majority of votes is then selected as the final outcome.

While the proliferation of fabricated content on social media provides a timely motivation for our analysis, the strategic environment we characterize belongs to a broader class of competitive persuasion games under source uncertainty. In this framework, an "Informed Sender"—possessing a state-contingent signaling technology—competes for influence against a "Noise-Generating Sender" who possesses only a state-independent technology but can mimic the format of the expert's messages. This structural tension is inherent in any decentralized information market where the origin of a signal is obscured, unverified, or anonymous.

We fully characterize the equilibrium of both a basic model with two states and two news stories and a generalized model with a finite number of states/stories (Proposition 1 and Theorem 1, respectively). We demonstrate that when fake news is sufficiently "prominent," meaning that voters are exposed to the fake news sender with a high probability, the related news sender fails to persuade voters due to the fake news sender's ability to "dilute" the related news, rendering it uninfluential: The fake news sender can consistently produce prominent "unrelated" news, thereby reducing the correlation between the observed news and the actual state of nature (Lemma 1 and Proposition 2). As a result, voters are invariably led to choose the action preferred by the fake news sender.

On the other hand, when fake news is not sufficiently prominent, the related news sender can generate an influential news story. Intriguingly, our findings indicate that even though the related news sender is capable of generating influential news, the social welfare—measured by the sum of the voters' ex-ante expected utilities—increases with the prominence of fake news: To endure the dilution of a more prominent fake news sender, the related news sender disciplines her inclination to disseminate self-biased news more frequently, aiming instead to produce durable news more closely associated with the state of nature (Lemma 2 and Remark 2). Under certain conditions, this informational gain from strongly related news outweighs the loss incurred from the presence of fake news, enabling voters to make better-informed decisions from an ex-ante perspective (Remark 3 and Proposition 5).

One of our main contributions is methodological: we show that a competitive persuasion game with source uncertainty can be reduced to a single durability condition on the informed sender's signal (Proposition 3 and Proposition 4), which in turn admits a clean characteriza-

tion through the Rothschild-Stiglitz geometry of Gentzkow and Kamenica (2016). The novelty lies in this reduction—identifying that a class of competition-under-source-uncertainty problems maps onto an RS representation—rather than in extending the RS approach itself. Within this class, the method delivers all equilibrium outcomes efficiently and intuitively.

1.1 Literature Review

The seminal literature on information transmission has posed the question of whether competition between multiple senders increases the surplus of the decision maker who receives the information. In cheap-talk games, it is found that two senders with conflicting interests can benefit a decision maker unless the senders are extremely biased (Krishna and Morgan, 2001), or the decision maker can leverage partly aligned interests with each sender to make multiple senders fully reveal the true state in a multidimensional space (Battaglini, 2002).

Recent studies in Bayesian persuasion have explored the benefits of involving multiple senders, where each sender can commit to a state-contingent signaling instrument, known as an experiment or signal, without prior knowledge of the realized state. Gentzkow and Kamenica (2016) identify a necessary and sufficient condition under which the non-cooperative outcome, arising from simultaneous signal choices by multiple senders, yields greater benefits to a decision maker than any collusive outcome. Li and Norman (2021) investigate sequential persuasion games with multiple senders and demonstrate that a decision maker can benefit more from consulting two senders sequentially rather than just one. Closest to our environment, Kwiek (2026) studies persuasion by multiple senders whose messages the receiver cannot attribute, and shows that anonymity lets a sender subtract from, rather than only refine, the informativeness of public discourse.

While our study shares with these works a focus on persuasion by multiple senders, it departs from them along two dimensions that jointly define our environment. First, the prior literature endows every sender with a state-contingent signaling technology; the competition is between symmetric experts who differ only in their preferences. We instead study competition between heterogeneous technologies—an informed sender who can design state-contingent experiments and a state-independent sender who cannot. Second, and central to our analysis, we introduce uncertainty about the origin of a message: the receiver observes a signal drawn from a shared message space but cannot attribute it to either sender. This source uncertainty is what gives the state-independent sender any influence at all—absent it, his messages would be recognized and disregarded—and it is what transforms the problem from one of mechanical noise into one of genuine strategic competition. Together, these two features fundamentally alter the nature of competition between senders and generate the

welfare implications we develop below.

Previous studies have also explored how uncertainty regarding senders' preferences or credibility impacts their strategic communication. Minozzi (2011) examines a competitive cheap-talk model featuring two senders whose preferences are unknown to the receiver. This study proves the existence of a "jamming equilibrium," where senders deliberately mix their messages with each other's. This tactic creates uncertainty for the receiver about the origin of a truthful message, thereby obscuring the senders' true preferences or intentions. Lipnowski et al. (2022) investigate a Bayesian persuasion game involving a sender and a receiver, revealing that introducing uncertainty about the sender's credibility can aid the receiver in making a correct decision. To examine the impact of uncertainty regarding the origin of news in a competitive persuasion game context, we make a simplifying assumption: receivers can only receive one message from the competing senders. Our analysis demonstrates that the existence and prominence of fake news can, paradoxically, benefit the receiver.

Our research also relates to studies on media bias and the media market. In this body of literature, media bias is attributed to various factors: media's concern about reputation (Gentzkow and Shapiro, 2006), readers' preference for confirmatory information (Mullainathan and Shleifer, 2005; Gentzkow et al., 2016), bribery (Besley and Prat, 2006), or government control (Gehlbach and Sonin, 2014; Enikolopov et al., 2011; Cheah, 2016). Particularly relevant is the work of Allcott and Gentzkow (2017), which analyzes the potential impact of fake news on social media on the outcome of the 2016 U.S. presidential election. They propose that introducing fake news producers—motivated primarily to maximize website traffic—into the media market results in social costs. This occurs because decision makers, who rely on information to assess the quality of political candidates, end up with less accurate beliefs due to the presence of fake news. Our model demonstrates that when a related news sender and a fake news sender possess the opposite political biases, the presence of fake news can actually enhance the welfare of voters. (See Remark 2.)

This reverses the benchmark of Mylovanov and Zapechelnyuk (2024), who study a sequential debate in which the second mover garbles the first mover's signal and show that such "toxic" argumentation typically leaves the audience uninformed. In our model the degradation of news is not imposed by a debate protocol but arises endogenously from source uncertainty, and the resulting noise disciplines the related news sender's use of her informational superiority. Focusing on the issue of fake news in a different context, Li and Whinston (2020) analyze the strategy of a fake news outlet aiming to maximize the receiver's dependence on its news, while occasionally disseminating truthful information to disguise its type. In their model, truthful and fake news are treated as two distinct, independent random variables with exogenously defined distributions. Our approach contrasts by differentiating

a related news sender from a fake news sender based on the sender's capability to produce informative news. This distinction allows us to explore the incentives behind producing fake news in the context of competitive persuasion.

2 Basic Model

We elucidate the basic ideas of this paper with the following example. Consider three players: a related news sender, RN (female), a fake news sender, FN (male), and a decision-maker, DM (they).[4] All players share a common prior belief over the state of nature $\omega \in \{\omega^0, \omega^1\}$, with $\Pr(\omega^0) \equiv \mu^0 = 0.6$ and $\Pr(\omega^1) \equiv \mu^1 = 0.4$. It is important to note that the qualitative results of this basic model are not dependent on the exact values of μ^0 and μ^1, but rather on the condition that $\mu^0 > \mu^1$.

The timing of the game is as follows: First, RN commits to a reporting strategy $\{\pi_r(s_r|\omega)\}_{\omega\in\Omega}$, where each $\pi_r(s_r|\omega)$ is a probability distribution over a set of news stories, $S = \{s^0, s^1\}$, conditional on the realized state, $\omega \in \Omega$.[5] Second, after observing RN's strategy, FN commits to a state-independent reporting strategy $\pi_f(s_f)$ over the same set of news stories, $S = \{s^0, s^1\}$.[6] The restriction that FN's technology is state-independent is not a simplifying device but the defining economic feature of a fake-news producer. A fabricator, by construction, has no access to or interest in the underlying state; what makes content "fake" is precisely that it is generated without regard to ω, mimicking the format of genuine reporting to gain legitimacy (Lazer et al., 2018; Allcott and Gentzkow, 2017). This is the sense in which our FN differs fundamentally from the senders in the competitive-persuasion literature, who each command a state-contingent signaling technology: FN can choose how often to emit any given message, but never what the message reveals about ω. The asymmetry of technologies between the two senders—one informed, one state-independent—is therefore the substantive primitive of the model, not a restriction imposed for tractability.

Subsequently, the state ω and the corresponding news stories $\{s_r, s_f\}$ are realized according to the senders' strategies. DM receives only one piece of news $s \in \{s_r, s_f\}$ and chooses an alternative $a \in \{a^0, a^1\}$. DM is unaware of the origin of news s: With probability ρ, they receive RN's news s_r, and with probability $1 - \rho$, FN's news s_f. The probability ρ, which signifies the prominence of related news, is common knowledge among the players.

Two senders have opposite state-independent preferences: RN prefers a^1 over a^0, whereas

[4] In Section 3, we consider a continuum of voters. So, we use plural pronoun "they" for DM in this section.

[5] In this section, for clarity, RN only uses two news stories in her news-reporting strategy. We later demonstrate that this is indeed an equilibrium strategy for RN.

[6] Assuming FN uses the same set of news stories as RN is innocuous. This assumption does not alter the equilibrium outcomes and their implications.

FN favors a^0. Specifically, they each receive a payoff of 1 if DM selects their preferred alternative, and 0 otherwise. Their objective is to maximize the probability that DM chooses their favored alternative by (not) providing information about ω. On the other hand, DM's preference is state-dependent: they receive a payoff of 1 if their action aligns with the true state, and 0 otherwise.

Thus, after receiving news s, DM forms a posterior belief $\hat{\mu}^1(s) = \Pr(\omega^1|s)$. If $\hat{\mu}^1(s) \geq 0.5$, they choose a^1; otherwise, they choose a^0. It is important to note that in the absence of any information, DM defaults to choosing a^0, as the prior belief μ^1 is less than 0.5.

We use Perfect Bayesian Equilibrium and impose subgame perfection in the RN–FN subgame (FN best-responds to any RN strategy).

Benchmark model: No fake news Suppose $\rho = 1$, indicating that the fake news sender FN is non-existent. The analysis in this scenario becomes equivalent to the basic example presented by Kamenica and Gentzkow (2011): RN optimally selects a distribution with $\pi_r'(s^1 \mid \omega^1) = 1$ and $\pi_r'(s^1 \mid \omega^0) \equiv p' = \dfrac{\mu^1}{\mu^0} = \dfrac{2}{3}$. Consequently, s^0 is generated with a probability of $\mu^0(1-p') = 0.2$, and s^1 with a probability of $\mu^1 + \mu^0 p' = 0.8$. DM's posterior beliefs are $\hat{\mu}^1(s^0) = 0$ and $\hat{\mu}^1(s^1) = \frac{\mu^1}{\mu^1+\mu^0 p'} = 0.5$. Therefore, they choose a^0 after receiving s^0, and a^1 after receiving s^1.

Basic model: With fake news Now suppose $\rho \in (0,1)$. Given RN's objective to induce DM to choose a^1, the strategy $\pi_r^*(s^1|\omega^1) = 1$ remains optimal. For simplicity, let $\pi_r(s^1|\omega^0) \equiv p$ represent RN's reporting strategy, and $\pi_f(s^1) \equiv q$ denote FN's reporting strategy. For given p and q, DM's posterior belief after receiving a signal s^1 is $\hat{\mu}^1(s^1;p,q) = \Pr(\omega^1|s^1;p,q) = \frac{\{\rho(1)+(1-\rho)q\}\mu^1}{\{\rho p+(1-\rho)q\}\mu^0+\{\rho(1)+(1-\rho)q\}\mu^1}$. We say the news story s^1 is "influential" if $\hat{\mu}^1(s^1;p,q) \geq 0.5$, so that DM chooses a^1. In this sense, we call s^1 the RN's self-biased news and s^0 the FN's self-biased news. We also say the news story s^1 is "potentially influential" if $\hat{\mu}^1(s^1;p) = \frac{\mu^1}{\mu^1+\mu^0 p} \geq 0.5$, so that DM chooses a^1 without any fake news story s^1.

Given RN's strategy p, if FN can produce state-independent fake news $s_f = s^1$ with sufficient frequency, such that he influences DM to choose a^0 even after receiving $s = s^1$, he can consistently ensure his preferred outcome a^0. This is particularly the case whenever

$$\hat{\mu}^1(s^1;p,q=1) = \frac{\mu^1}{\{\rho p + (1-\rho)\}\mu^0 + \mu^1} < 0.5,$$

$q = 1$ would be the FN's best response to p. This strategy by FN, aimed at generating fake news s^1 as frequently as possible to render DM unresponsive to s^1, is referred to as the "water-down" strategy or "diluting" strategy.

Lemma 1 (The Diluting Strategy)**.** *Suppose* $p > \frac{1}{\rho}\left(\frac{\mu^1}{\mu^0} - (1-\rho)\right)$*. Then, FN's best response is diluting RN's news* s^1 *by generating state-independent news* s^1 *always. That is, in any subgame with* $p > \frac{1}{\rho}\left(\frac{\mu^1}{\mu^0} - (1-\rho)\right)$*, the FN's best response is* $q^{BR}(p) = 1$*, and the DM's consistent belief and the sequentially rational action are*

$$\hat{\mu}^1(s^1; p, q = 1) = \frac{\mu^1}{\mu^0\{\rho p + (1-\rho)\} + \mu^1} < 0.5 \;\; \textit{and} \;\; a(s) = a^0.$$

Let's focus on the region where fake news is relatively prominent, specifically in the domain where $\rho < \bar{\rho} \equiv 1 - \frac{\mu^1}{\mu^0}$. In this context, we present the following remark.

Remark 1 (Prominent Fake News)**.** *Suppose* $\rho < \bar{\rho} \equiv 1 - \frac{\mu^1}{\mu^0}$*. Then, for any value of* $p \in [0,1]$*, we have* $p > \frac{1}{\rho}\left(\frac{\mu^1}{\mu^0} - (1-\rho)\right)$*. Therefore, in any subgame with* $p \in [0,1]$*, FN plays the diluting strategy, DM always chooses* a^0*, and RN's news would never be influential.*

Now suppose $\rho \geq \bar{\rho} \equiv 1 - \frac{\mu^1}{\mu^0}$. RN can choose a "durable" strategy that ensures $\hat{\mu}^1(s^1; p, q = 1) \geq \frac{1}{2}$, thus withstanding any dilution attempt of FN. Notably, RN can enhance her payoff by increasing the value of p, while maintaining the durability of the strategy. Since $\hat{\mu}^1(s^1; p, q = 1)$ decreases with increasing p, the optimal durable strategy for RN can be characterized by a p^* that satisfies $\hat{\mu}^1(s^1; p^*, q = 1) = 0.5$. Under these conditions, FN is compelled to select $q = 0$ to maximize the likelihood of DM receiving s^0 and choosing for a^0. For $p < p^*$, RN sends s^1 less frequently in state ω^0, so s^1 becomes more indicative of ω^1 and $\hat{\mu}^1(s^1; p, q = 1) > 0.5$. Dilution then fails—DM chooses a^1 upon receiving s^1 regardless of FN's action—so FN strictly prefers $q = 0$, securing a^0 whenever s^0 is realized. Only for $p > p^*$ does flooding with s^1 $(q = 1)$ succeed in overturning DM's response.

Lemma 2 (The Optimal Durable Strategy)**.** *Suppose* $\rho \geq \bar{\rho} \equiv 1 - \frac{\mu^1}{\mu^0}$*. The RN's optimal strategy is generating durable news* s^1 *as frequently as possible. That is, the RN's optimal strategy, the FN's best response, and the DM's consistent belief and sequentially rational action are*

$$p^* = \frac{1}{\rho}\left(\frac{\mu^1}{\mu^0} - (1-\rho)\right), \;\; q^*(p) = \begin{cases} 0, & p \leq p^* \\ 1, & p > p^*, \end{cases}$$

$$\hat{\mu}^1(s^1; p^*, q^* = 0) = \frac{\mu^1}{p^*\mu^0 + \mu^1} > 0.5, \;\; \textit{and} \;\; a^*(s) = \begin{cases} a^1, & s = s^1 \\ a^0, & s = s^0. \end{cases}$$

The following proposition encapsulates the analysis presented above. The existence of a persuasive equilibrium, where RN wields a certain degree of persuasion power, hinges on the prominence of the related news, denoted by ρ.

Proposition 1 (Persuasive Equilibrium)**.**
The equilibrium characteristics of the persuasion game depends on the prominence of the related news ρ.

1. *If the fake news is sufficiently prominent $\rho < \bar{\rho}$, then*
   - *RN's strategy does not matter in equilibrium, since for any $p \in [0,1]$*
   - *FN's diluting strategy $q^*(p) = 1$ makes the RN's news uninfluential, and*
   - *DM always chooses FN-preferred action $a^*(s) = a^0$.*
2. *If the related news is sufficiently prominent $\rho \geq \bar{\rho}$, then*
   - *RN chooses the optimal durable strategy $p^* = \frac{1}{\rho}\left(\frac{\mu^1}{\mu^0} - (1-\rho)\right)$,*
   - *FN generates the self-biased news s^0 on the equilibrium path $q^*(p^*) = 0$, and*
   - *DM chooses a^1 if, and only if, they receive s^1.*

One may argue that the parameter ρ reflects RN's power of persuasion or control. Consequently, it could be expected that RN disseminates her self-biased news story s^1 more frequently as ρ increases, implying $\frac{\partial p^*}{\partial \rho} > 0$. The following remark elaborates on the comparative statistics of the equilibrium strategy p^*.

Remark 2 (RN's Self-disciplined Signal)**.**
Suppose $\rho > \bar{\rho}$. The optimal level of p^* in the basic model exhibits the following properties.

1. *$p^* < p'$: In the presence of FN, to avoid the dilution of the fake news, RN disciplines her own desire to send s^1 more frequently with the intention of generating a durable news, which is more accurate than the news generated in the benchmark without FN.*
2. *$\frac{\partial p^*}{\partial \rho} > 0$: If the fake news is more prominent, so that FN's threat of dilution is more powerful, then RN should discipline herself more strongly and generate more accurate news to make the related news durable.*

Figure 1 illustrates the optimal strategy of RN in the equilibrium characterized in Proposition 1. In the region where $\rho < \bar{\rho}$, the scenario can be interpreted as one where fake news is extremely prominent, leading to a deterioration in the informativeness of RN's related news. Consequently, a DM who receives s^1 in this scenario is more likely to believe that the state is ω^0 and will consistently choose a^0. In this case, we observe that FN effectively dilutes RN's news, regardless of its accuracy.

Conversely, in the region where $\rho > \bar{\rho}$, the situation can be seen as one where the related news is sufficiently prominent. Here, RN's news remains informative despite any dilution by

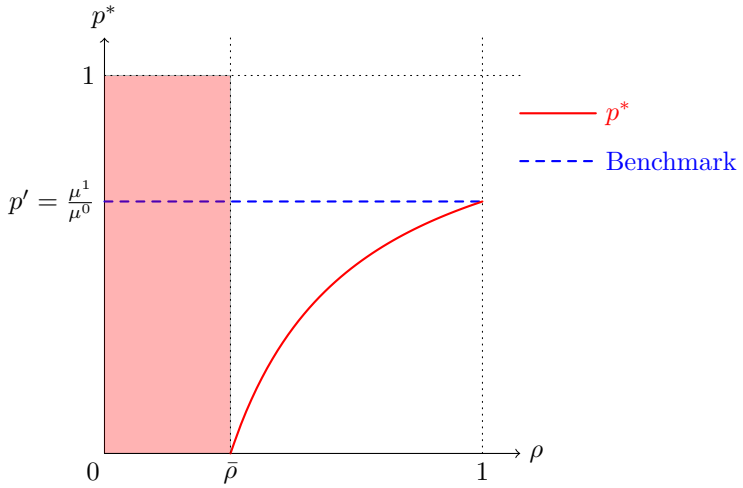

Figure 1: RN's optimal strategy: The probability of sending s^1 given ω^0

FN. Therefore, under RN's optimal strategy, a DM receiving s^1 is more likely to believe that the state is ω^1 and opts for a^1. In this context, RN's news is durable against the dilution by FN and influential in shaping DM's action.

Note that in the region where $\rho > \bar{\rho}$, the optimal level of p^* is lower than the benchmark p' and it increases with ρ. In the presence of FN, RN must temper her inclination to send s^1 as frequently as in the benchmark scenario, aiming to create durable news. Consequently, FN, by producing unrelated fake news, inadvertently increases the accuracy of RN's related news. In essence, FN's actions compel RN to be more honest. This raises an intriguing question: can the presence of FN make DM better off?

To address this, we investigate the ex-ante expected utility of DM under these circumstances.

Remark 3 (DM's Ex-ante Expected Utility)**.**
The DM's ex-ante expected utility in equilibrium, $u_{DM}(a^*; p^*, q^*)$, is non-monotonic in the level of the prominence of related news ρ. In particular,

1. *in the region of $\rho < \bar{\rho}$, $u_{DM}(a^*; p^*, q^*) = \mu^0$, which is equal to the utility in the benchmark, $u_{DM}(a'; p')$.*

2. *In the region of $\rho > \bar{\rho}$, $u_{DM}(a^*; p^*, q^*)$ is decreasing in the level of the prominence of related news, $\frac{\partial u_{DM}(a^*; p^*, q^*)}{\partial \rho} < 0$, and DM is better off in the basic model with FN than in the benchmark, $u_{DM}(a^*; p^*, q^*) - u_{DM}(a'; p') > 0$.*

Figure 2 depicts the DM's ex-ante expected utility in our basic model compare to the benchmark. In the region of $\rho < \bar{\rho}$, the prominent FN makes RN uninfluential, leading

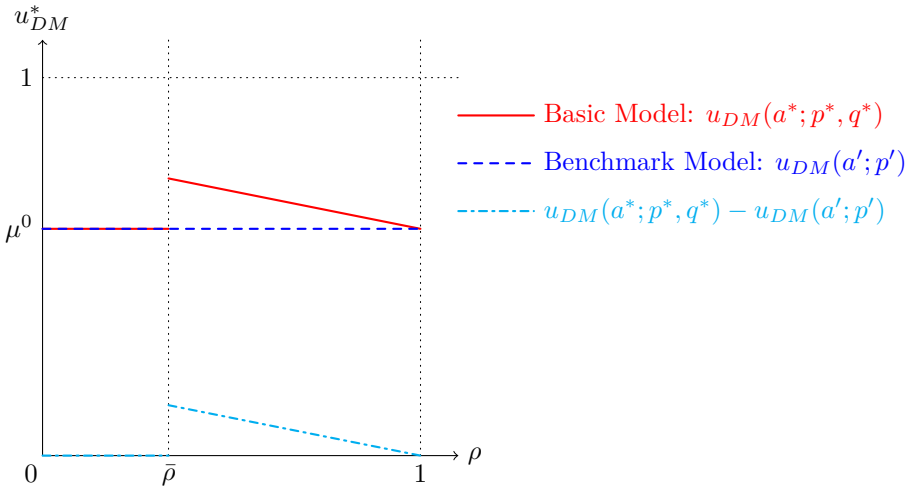

Figure 2: DM's Ex-ante Expected Utility

DM to always choose a^0. Note that the ex-ante utility in the benchmark is equivalent to a scenario with no news at all. The uninfluential RN implies that DM's choices are made as if they lack any information, signifying that in a context overwhelmed with fake news, no information holds significance.

Conversely, in the region where $\rho \geq \bar{\rho}$, despite the potential receipt of unrelated fake news, RN's self-disciplined news delivers more accurate information. The benefit derived from RN's increased self-discipline outweighs the loss incurred due to FN's presence. Consequently, from an ex-ante perspective, DM is better off with FN than without him.

Furthermore, DM's ex-ante expected utility is decreasing in ρ in the region of $\rho > \bar{\rho}$. This finding is intriguing as it suggests that DM benefits more from a more prominent FN, or from encountering FN's news more frequently. The main driving force of the DM's ex-ante payoff decreasing in ρ is the RN's self-disciplined signal. As RN becomes more prominent, even though DM hears less frequently from FN, RN's reduced self-discipline results in less accurate s^1. Therefore, DM's ex-ante expected payoff diminishes as ρ increases.

In our model, the influence of FN on DM's perceptions is directly proportional to its presence relative to the related news sender (RN). Thus, as the prominence of RN, represented by ρ, increases, the potential diluting influence of FN decreases linearly. To counteract this, RN adjusts her strategy to make her influence more potent. Since RN relaxes her self-discipline as ρ increases (recall p^* increases with ρ), the probability of DM receiving an influential but incorrect signal s^1 from RN, given by $\rho\mu^0 p^*$, increases with ρ. While the probability of receiving a correct and influential s^1, given by $\rho\mu^1$, also increases with ρ, the probability of DM taking the correct default action upon receiving s^0, given by $\mu^0(1-\rho p^*)$, decreases with

ρ at a faster rate.[7] As a result, DM's ex-ante expected utility decreases with ρ. Because each of these probabilities is linear in ρ, DM's expected utility declines linearly on this region.

In reality, the phenomenon of self-disciplined news emerging in response to the proliferation of fake news is occasionally observed. During the early rise of social media and fake news, particularly in the mid-2010s, many journalists from leading international news outlets believed that enhancing reporting quality and credibility was a key strategy to counteract fake news (Risi, 2017). However, as fake news became increasingly prominent, even the news from traditional media sources began to lose credibility and, consequently, their influence over the public diminished, becoming less influential or even uninfluential. Various empirical and experimental studies corroborate our findings in this regard (Allcott and Gentzkow, 2017; Tandoc Jr et al., 2021).

3 Generalized Model

In this section, we examine a generalized version of our basic model. There are two news senders; a related news sender (RN, she) and a fake news sender (FN, he) and a continuum of voters (they) of unit mass. Voters x's are distributed according to an atomless cumulative distribution function F that admits a continuous density f with $f(x) > 0$ on $(-1, 1)$. We assume that $F(x)$ is common knowledge to all players.

There are $n(\geq 2)$ number of states, of which a set is denoted by $\Omega = \{\omega^1, \cdots, \omega^n\} \subset \mathbb{R}$. Without loss of generality, we assume $\underline{\omega} = \omega^1 < \omega^2 < \cdots < \omega^n = \bar{\omega}$. The true state of the world, denoted as ω, remains unobservable to all players. Despite this, players share a common prior belief, represented by μ_0. Importantly, μ_0 is not a boundary element but an interior point in the simplex $\Delta(\Omega)$, which encompasses all possible states. For voters, a set of two policy alternatives, $\mathcal{A} = \{a^0, a^1\}$, is available to choose.

Preference Two senders, RN and FN, exhibit state-independent preferences that are diametrically opposed: RN has a preference for alternative a^1 over a^0, whereas FN exhibits a preference for a^0 over a^1. Their payoff structure is straightforward: they receive a payoff of 1 if the median voter selects their preferred alternative, but receive a payoff of 0 if the alternative is not chosen.

[7] As ρ increases, the correct signal s^1 arrives with probability $\rho\mu^1$ (increasing), whereas the correct default action on s^0 obtains with probability $\mu^0(1 - \rho p^*)$ (decreasing). The latter effect dominates, so DM's overall probability of choosing the correct action falls.

Voter x's utility, given a state ω and the preference parameter x, is

$$u(a;\omega,x)=\begin{cases}x+\omega & \text{if } a=a^1\\ 0 & \text{if } a=a^0.\end{cases}$$

In this context, a voter's position, denoted as x, signifies the marginal benefit derived from choosing alternative a^1 over a^0. Conversely, the state ω represents the collective marginal benefit of a^1 over a^0 applicable to all. It's important to note that while x is a private valuation unique to each voter, ω is a common value shared by all. As ω increases, a greater number of voters are likely to benefit from opting for a^1.

Strategies and Assumptions The sender and voters play a three stage dynamic game. The first stage is characterized by RN's selection of a signal (or reporting strategy) $\{\pi_r(s_r|\omega)\}_{\omega\in\Omega}$. Here, each $\pi_r(s_r|\omega)$ represents a probability distribution over a finite signal realization space $S=\{s^1,s^2,\cdots,s^N\}$, which RN chooses.[8]

In the second stage, following the observation of RN's choice, FN commits to a signal (or reporting strategy) $\pi_f(s_f)$, which is defined as a probability distribution over a finite signal realization space $\tilde{S}\supseteq S$. For the sake of simplicity in notation, we use σ_{s_f} interchangeably to refer to $\pi_f(s_f)$. Subsequently, voters observe the strategies of both senders.

In the third stage, the true state is realized according to a commonly known distribution by Nature, resulting in the production of a related news story s_r and a fake news story s_f, each in accordance with the respective strategies of the senders. It is assumed that RN's news reaches all voters with a probability of $\rho\in(0,1)$, while FN's news reaches them with a probability of $(1-\rho)$. Each voter receives only one news story and is unaware of the news source, unless it is explicitly revealed by the news itself. The probability ρ is common knowledge among all players.

After observing the news $s\in\{s_r,s_f\}$ and forming a common belief $\mu\in\Delta(\Omega)$, voters cast their votes on the set of alternatives, $\mathcal{A}$. We apply the Perfect Bayesian Equilibrium as the solution concept, wherein the voters' voting strategies should be sequentially rational given the consistent belief μ. To primarily focus on the signaling behavior of the senders, we assume that voters vote sincerely: a voter always votes for her preferred alternative given her belief, μ.[9] In other words, given μ, voter x 1) votes for a^1 if $E_\mu u(a^1;\omega,x)>$

[8]The signal realization is interpreted as a news story, with the state-contingent distributions being analogous to editorial decisions that vary depending on the realized real-world events.

[9]As a single voter's deviation never affects the voting outcome, any decision profile made by voters is an equilibrium action profile and it satisfies sequential rationality. Therefore, assuming a sincere voting decision by voters is harmless as it is consistent with the concept of Perfect Bayesian Equilibrium.

$E_\mu u(a^0;\omega,x) \Leftrightarrow E_\mu(\omega) > -x$, 2) votes for a^0 if $E_\mu(\omega) < -x$, and 3) randomizes between a^0 and a^1 if $E_\mu(\omega) = -x$.

The following assumption guarantees the existence of a voter who is indifferent between a^0 and a^1 for any μ.

Assumption 1. $-1 \leq \underline{\omega} < \bar{\omega} \leq 1$.

The voting rule is a simple majority: If more than half of the voters cast their votes for a particular alternative, that alternative is socially accepted as the voting outcome. To rule out the cases in which RN need not persuade anyone to get her favorite outcome, we assume that the median voter votes for a^0 given the prior belief μ_0. That is, given $\text{Med}(x)$ is the median of the distribution $F(x)$, we make the following assumption which implies $E_{\mu_0} u(a^1;\omega,\text{Med}(x)) < E_{\mu_0} u(a^0;\omega,\text{Med}(x))$.

Assumption 2. $E_{\mu_0}(\omega) < -\text{Med}(x) \equiv \gamma$.[10]

Relation to single-receiver persuasion Because all voters observe the same public news realization[11] and vote sincerely under simple majority rule, the senders' problem reduces to moving the public posterior mean across a single threshold, $\gamma = -\text{Med}(x)$: the median voter is pivotal, and any realization with $E_{\hat{\mu}_s}(\omega) \geq \gamma$ secures a^1 while any realization with $E_{\hat{\mu}_s}(\omega) < \gamma$ secures a^0. In this sense, our environment is strategically equivalent to a competitive persuasion problem with a single receiver whose belief threshold is that of the median voter. We nonetheless model a continuum of voters deliberately, for two reasons. First, it endogenizes the persuasion threshold: rather than positing γ as a primitive, the continuum pins it down as the median marginal voter $-\text{Med}(x)$, tied to the same distribution F that describes the electorate. Second, and more importantly, it is what makes our welfare analysis meaningful. The relevant welfare object is not the pivotal voter's payoff but the aggregate utilitarian welfare of the entire electorate, and the gap between the average voter $E(x)$ and the pivotal voter $\text{Med}(x)$ generates a distinct welfare component that has no counterpart in a single-receiver model. The continuum is therefore silent for the equilibrium characterization—where the median-voter reduction applies—but essential for the welfare

[10] This assumption also rationalizes the sequence of the senders' moves. The sender, who would like to persuade voters to change their belief, would move first by revealing some information. Therefore, under this assumption, it is natural for RN to move first to change the voters' beliefs; then, as a response, FN would move to prevent the voters from changing their beliefs.

[11] The assumption that all voters observe the same public realization is essential for the tractability of our Rothschild-Stiglitz characterization. If voters instead received signals privately and independently from different sources, the proportion of voters holding each posterior would itself enter both the voting and persuasion stages, and our geometric method would no longer apply. We discuss this limitation and the resulting open questions in Section 5.

analysis in Section 4 and Online Appendix OA, where the distribution of voters, not merely its median, determines social welfare.

3.1 General Interpretations and Applications

Beyond the political sphere, our model provides a rigorous framework for analyzing diverse economic phenomena where "strategic noise" threatens to crowd out expertise. In financial markets, it captures the competition between reputable analysts and anonymous "pump-and-dump" shills on social trading platforms. In e-commerce, it reflects the tension between fact-based product testers and paid reviewers who fabricate sentiment. In these contexts, the central result of our theory—the "self-disciplining" effect—offers a foundational insight: the presence of strategic noise compels an expert to produce higher-precision information to ensure their signals remain "durable" and persuasive to a wary receiver. What unifies these applications, and what makes the method applicable to them, is a common decision structure: in each, the receiver faces an essentially binary choice against a default—buy or abstain, trust or distrust, act or hold—so that persuasion again turns on a single belief threshold. The breadth of the framework thus comes not from relaxing this structure but from how many economically distinct settings share it.

3.2 The Best Response of the Fake News Sender

We solve the game backward. Given RN's strategy π_r, when voters observe a news story s in $\tilde{S}$, their posterior belief $\hat{\mu}_s(\cdot)$ would be

$$\hat{\mu}_s(\omega) = \begin{cases} \frac{\{\rho\pi_r(s|\omega)+(1-\rho)\sigma_s\}\mu_0(\omega)}{\sum_{\omega\in\Omega}\{\rho\pi_r(s|\omega)+(1-\rho)\sigma_s\}\mu_0(\omega)} & \text{for } s \in S \\ \mu_0(\omega) & \text{for } s \in \tilde{S}\backslash S \end{cases}$$

where S is the set of related news stories of π_r. For notation, introduce the (virtual) distribution ψ induced by π_r *only over related news* S (i.e., abstracting from unrelated news):

$$\psi_s \;\equiv\; Pr_{\pi_r}(s) \;=\; \sum_{\omega\in\Omega} \pi_r(s \mid \omega)\,\mu_0(\omega), \qquad s \in S.$$

Then, the posterior mean can then be written as

$$E_{\hat{\mu}_s}(\omega) = \begin{cases} \frac{\rho\psi_s}{\rho\psi_s+(1-\rho)\sigma_s} E_{\mu_s}(\omega) + \frac{(1-\rho)\sigma_s}{\rho\psi_s+(1-\rho)\sigma_s} E_{\mu_0}(\omega) & \text{for } s \in S \\ E_{\mu_0}(\omega) & \text{for } s \in \tilde{S}\backslash S. \end{cases} \tag{1}$$

where $E_{\mu_s}(\omega)$ is the mean belief only with the related news s and $E_{\mu_0}(\omega)$ is the mean belief with no informative signal.

Given RN's strategy π_r, FN selects a reporting strategy π_f with the objective of maximizing the likelihood that alternative a^0 is selected as the voting outcome, subsequent to voters observing a news story. To write down the FN's problem formally, we define the probability that voters observe a news story s, $\hat{\psi}_s$, for FN's strategy π_f given RN's strategy π_r as follows;

$$\hat{\psi}_s(\pi_f;\pi_r) = \begin{cases} \rho \sum\limits_{\omega\in\Omega} \pi_r(s|\omega)\mu_0(\omega) + (1-\rho)\pi_f(s) & \text{for } s \in S \\ (1-\rho)\pi_f(s) & \text{for } s \in \tilde{S}\backslash S. \end{cases} \tag{2}$$

Then, the FN's problem turns out to be

$$\max_{\pi_f\in\Pi_f} \sum_{s\in\{\tilde{s}:E_{\hat{\mu}_{\tilde{s}}}(\omega)<-\operatorname{Med}(x)\}} \hat{\psi}_s(\pi_f;\pi_r). \tag{3}$$

Consider a related news piece $s \in S$ that has the potential to convince voters of the superiority of a^1, provided they recognize it as related news. Such a news story is termed "potentially influential."

Definition 1 (Potentially Influential News)**.** *Suppose RN's strategy π_r. A related news $s \in S$ is **potentially influential** if $E_{\mu_s}(\omega) \geq -\operatorname{Med}(x)$, where $\mu_s(\omega) = \frac{\pi_r(s|\omega)\mu_0(\omega)}{\sum_{\omega\in\Omega}\pi_r(s|\omega)\mu_0(\omega)}$ is the posterior belief induced by the news s in the absence of FN. Denote $S^{PI}(\pi_r) = \{s \in S : E_{\mu_s}(\omega) \geq -\operatorname{Med}(x)\}$.*

FN may seek to diminish the impact of any potentially influential news. However, some news stories are inherently durable, meaning that they are resistant to being undermined by fake news, and as such, cannot easily alter the voters' beliefs. Therefore, in formulating an optimal fake news strategy, FN's focus should be on targeting news stories that lack this durability.

Definition 2 (Diluting; Durable News)**.** *Suppose RN's strategy π_r.*

- *We say FN can **dilute** or water down a related news story $s \in S$ when he can make $E_{\hat{\mu}_s}(\omega) < -\operatorname{Med}(x) = \gamma$ with some $\pi_f(s) \in [0,1]$.*
- *A related news story $s \in S$ is **durable** if it cannot be diluted by any $\pi_f(s) \in [0,1]$. That is, $s \in S$ is durable if $E_{\hat{\mu}_s}(\omega) \geq -\operatorname{Med}(x)$ for every $\pi_f(s) \in [0,1]$. In particular,*

$s \in S$ is durable if

$$\psi_s\big(E_{\mu_s}(\omega) - \gamma\big) \geq \left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big). \tag{4}$$

Denote $S^D(\pi_r) = \left\{ s \in S : \psi_s\big(E_{\mu_s}(\omega) - \gamma\big) \geq \left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big) \right\}$.

The subsequent proposition elucidates the key characteristics of FN's optimal strategy in response to RN's reporting strategy π_r.

Proposition 2. *Given any strategy of RN, π_r, the set of best responses of FN, denoted by $BR(\pi_r) \subseteq \Pi_f$, is nonempty. Moreover, the following properties hold:*

1. *No $\pi_f \in BR(\pi_r)$ assigns positive probability to any durable news story.*
2. *If, given π_r, the set of related news stories that are both potentially influential and not durable is nonempty, then:*
   (a) *every $\pi_f \in BR(\pi_r)$ dilutes at least one news story in this set; and*
   (b) *there exists a $\pi_f \in BR(\pi_r)$ that assigns positive probability exclusively to news stories in this set.*

Proof. See Appendix A.1 □

A remark on interpretation is in order. Because FN's optimal response to any π_r is a constrained choice over which potentially influential stories to dilute (Proposition 2), the solved game can be read as RN optimizing against a dilution constraint on her own signal. This reading is correct as a description of the equilibrium, but it should not be mistaken for evidence that FN is a passive source of noise. The constraint RN faces is endogenous: it is the best response of a second strategic agent whose preferences run opposite to hers, and its tightness is determined by FN's optimization, not by an exogenous noise process. What makes the problem one of competitive persuasion rather than persuasion under exogenous noise is precisely that this constraint is generated strategically—FN allocates his state-independent messages to where they do the most damage to RN's persuasive goal—and that the receiver's inability to attribute messages is what gives that allocation bite. The single-sender representation is thus a consequence of solving the game, not a statement about its strategic structure.

3.3 The Related News Sender's Optimal Strategy

While taking into account FN's optimal response, $BR(\pi_r)$, as delineated in Proposition 2, RN strategically selects a signal (or reporting strategy) π_r. This selection is aimed at optimizing the likelihood that alternative a^1 is ultimately chosen in the decision-making process.

First, we demonstrate that in any equilibrium strategy, RN will, if feasible, exclusively produce durable news. For RN, a strategy that allocates a positive (albeit partial) probability to the set of non-durable news stories is dominated by another strategy that solely focuses on durable ones.

Proposition 3 (Equilibrium Avoidance of Non-Durable News)**.** *If RN can produce durable news, then in any equilibrium, she will not produce potentially influential non-durable news.*

Proof. See Appendix A.2 □

Proposition 3 implies that, whenever durable news is feasible, any equilibrium relies only on durable news among potentially influential related stories. Therefore, to characterize the equilibrium strategies of RN fully, it is essential to characterize the condition that RN can produce a durable news story. In this process, we employ the Rothschild-Stiglitz approach, drawing on the insights provided by Gentzkow and Kamenica (2016).

We denote the cumulative distribution of posterior means induced by a signal π as $G_\pi(x) = \int_{\mu_s : E_{\mu_s}(\omega) \leq x} d\psi(\mu_s) = \sum_{s : E_{\mu_s}(\omega) \leq x} \psi_s$, where ψ represents the distribution of posterior beliefs induced only by the signal π. We define $H_\pi(x) \equiv \int_{-\infty}^{x} G_\pi(t)dt$, which represents the integral of $G_\pi(\cdot)$.

Consider a fully revealing signal, denoted by $\overline{\pi}$. It is evident that $\overline{\pi}$ induces the distribution of posterior means corresponding to each state, aligning with the prior belief on the states. Thus, $G_{\overline{\pi}}(x) = \sum_{k : \omega_k \leq x} \mu_0(\omega_k)$. Given that $G_{\overline{\pi}}(\cdot)$ is a right-continuous step function, its integral, $H_{\overline{\pi}}(\cdot)$, is a continuously increasing piece-wise linear function.

Now consider a totally uninformative signal, denoted by $\underline{\pi}$. Since every signal realization of $\underline{\pi}$ induces the same belief, we have

$$H_{\underline{\pi}}(x) = \begin{cases} 0 & \text{for } x < E_{\mu_0}(\omega) \\ x - E_{\mu_0}(\omega) & \text{for } x \geq E_{\mu_0}(\omega). \end{cases}$$

Gentzkow and Kamenica (2016) demonstrate that, when the set of signal realizations is finite, any signal can be depicted as a piece-wise linear convex function bounded between the functions $H_{\overline{\pi}}$ and $H_{\underline{\pi}}$. In Figure 3, the blue solid line represents the graph of the function

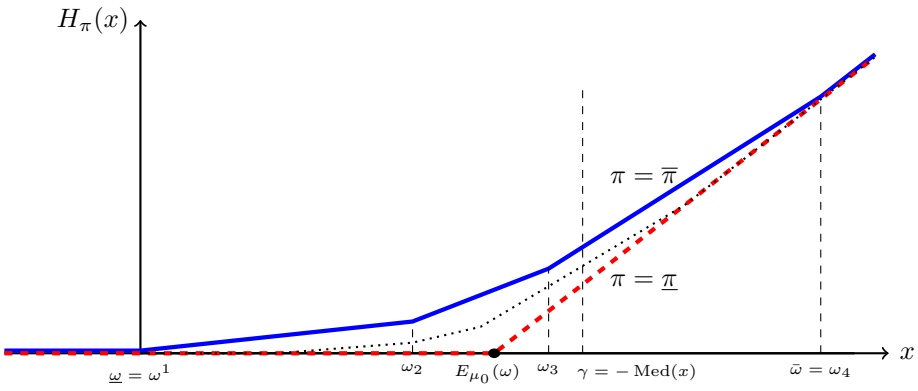

Figure 3: Illustration of $H_\pi(x)$

$H_{\overline{\pi}}(x)$, the red dashed line illustrates the graph of the function $H_{\underline{\pi}}(x)$, and the black dotted line, which is a piece-wise linear convex function, represents a signal.

Given a signal π and the corresponding distribution ψ, Bayes' rule yields the posterior $\mu_{s^i}(\cdot)$ after s^i, and we write its mean as $E_{\mu_{s^i}}(\omega)$. Without loss of generality, index realizations so that $E_{\mu_{s^i}}(\omega)$ is strictly increasing in i. Then the function H_π has the following properties: (i) $H_\pi(\cdot)$ has a kink at x if and only if $x \in \{E_{\mu_{s^i}}(\omega)\}_{i=1}^N$; and (ii) for each i,

$$H'_\pi\big(E_{\mu_{s^i}}(\omega)^+\big) \;-\; H'_\pi\big(E_{\mu_{s^i}}(\omega)^-\big) \;=\; \psi_{s^i}, \tag{5}$$

where $H'_\pi(x^+)$ and $H'_\pi(x^-)$ denote the right and left derivatives of H_π at x, respectively.

Now we describe a method to determine whether a news story is durable for a given reporting strategy π_r, and then use this method to derive a condition for the existence of a signal that generates durable news stories. Note that the left-hand side of the durability condition (4) can be written as

$$\begin{aligned}\psi_{s^i}(E_{\mu_{s^i}}(\omega)-\gamma) &= (H'_{\pi_r}(E_{\mu_{s^i}}(\omega)^+) - H'_{\pi_r}(E_{\mu_{s^i}}(\omega)^-))(E_{\mu_{s^i}}(\omega)-\gamma)\\ &= H'_{\pi_r}(E_{\mu_{s^i}}(\omega)^+)(E_{\mu_{s^i}}(\omega)-\gamma) - H'_{\pi_r}(E_{\mu_{s^i}}(\omega)^-)(E_{\mu_{s^i}}(\omega)-\gamma).\end{aligned}$$

For a potentially influential signal s^i, which generates the component probability ψ_{s^i} and the posterior mean $E_{\mu_{s^i}}(\omega) > \gamma$, a graphical representation of $\psi_{s^i}(E_{\mu_{s^i}}(\omega)-\gamma)$ can be derived from the graph of H_{π_r}, as illustrated in Figure 4. In that figure, the durability condition (4) corresponds to the length ③. Thus, s^i is durable if and only if the length ③ satisfies ③$\geq \left(\frac{1-\rho}{\rho}\right)(\gamma - E_{\mu_0}(\omega))$.

In geometric terms, the maximum attainable value of the length ③ $= \psi_{s^i}(E_{\mu_{s^i}}(\omega)-\gamma)$ across all signals π_r is equal to $H_{\overline{\pi}}(\gamma) - H_{\underline{\pi}}(\gamma)$. The following proposition shows that a signal

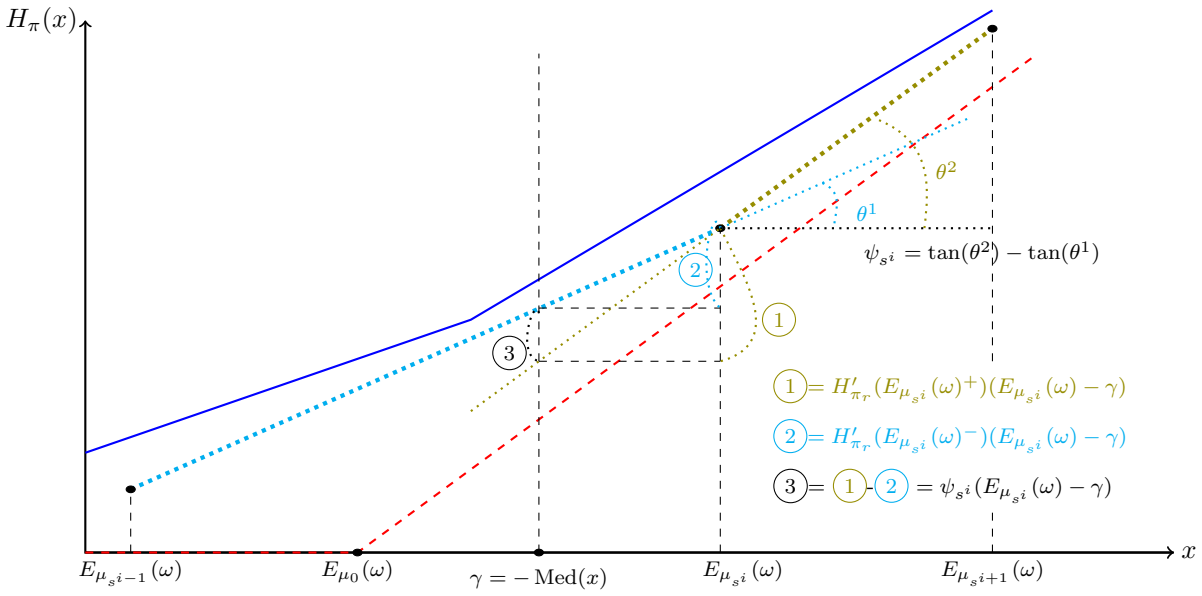

Figure 4: Illustration of $E_{\mu_{s^i}}(\omega)$ and ψ_{s^i} in $H_{\pi_r}(x)$

Note: In the figure, s^i (with mean $E_{\mu_{s^i}}(\omega)$) is potentially influential, whereas s^{i-1} is not. In general, s^i need not be the first potentially influential realization; the same graphical argument applies to any subsequent s^j, including s^{i+1}.

producing durable news exists if and only if $H_{\overline{\pi}}(\gamma) - H_{\underline{\pi}}(\gamma) \geq \left(\frac{1-\rho}{\rho}\right)(\gamma - E_{\mu_0}(\omega))$, which, using the definition of $H_{\underline{\pi}}(\gamma)$, is equivalent to the condition $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$.

Proposition 4 (Existence of a Signal that generates Durable News)**.**
There exists an RN signal π_r that produces at least one durable news story if and only if $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$.

Proof. See Appendix A.4. □

Proposition 4 provides a simple threshold condition on ρ under which RN can design a signal that generates durable news. Recall that $\rho \in (0, 1)$ denotes the prior probability that a news story comes from RN rather than FN. The condition $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$ therefore has a natural interpretation: RN can create at least one related news story that remains durable in the presence of FN only when her ex ante presence or credibility in the news environment is sufficiently high. When ρ is below this threshold, FN is so prevalent that every potentially influential related news story is vulnerable to dilution; when ρ exceeds the threshold, RN can still secure messages whose persuasive content survives fake news.

Two features of the environment make this geometric characterization possible. First, the binary action space induces a single belief threshold γ, so that the durability of a news story reduces to a scalar comparison—the vertical gap ③ in H_{π_r} against the dilu-

tion requirement—rather than a system of conditions. Second, all voters share this threshold. This single-threshold structure is inherited from the Rothschild-Stiglitz approach of Gentzkow and Kamenica (2016), and it is well-suited to the up-or-down decisions our model concerns: elections, and more broadly any environment in which a receiver makes a binary choice against a default. In richer environments with multiple thresholds or non-binary actions, durability would no longer be a single scalar and the present geometry would not directly apply.

We now move from this existence result for durable news to the equilibrium specification. If RN can implement a signal π_r that produces durable news stories, then she can retain persuasive power in equilibrium, because FN never assigns probability to durable news and cannot overturn their impact.

The following theorem summarizes our main findings from the preceding analysis and characterizes the existence of a Persuasion Equilibrium in which RN can persuade DM, together with the main properties of such an equilibrium.

Theorem 1 (Existence of a Persuasion Equilibrium)**.** *The existence of an equilibrium in which RN persuades voters with positive probability depends on the prominence of related news ρ. Let $\pi_r^* \in \Pi_r$ and $\pi_f^* \in \Pi_f$ denote the equilibrium strategies of RN and FN, respectively.*

1. *If $\rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$ (Prominent Fake News):*

   - ***RN's Equilibrium Strategy:*** *Any reporting strategy $\pi_r \in \Pi_r$ is an equilibrium strategy for RN.*
   - ***Durability:*** *For any $\pi_r \in \Pi_r$, the set of durable news stories is empty ($S^D(\pi_r) = \emptyset$).*
   - ***FN's Best Response:*** *In any equilibrium, for any $\pi_f^*(\cdot; \pi_r^*) \in BR(\pi_r^*)$, every potentially influential news story $s \in S^{PI}(\pi_r^*)$ is diluted (i.e., $E_{\hat{\mu}_s}(\omega) < \gamma$).*
   - ***Outcome:*** *The equilibrium action is $a^* = a^0$ for every observed news realization $s \in \tilde{S}$, and RN's equilibrium payoff is 0.*

2. *If $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, then an equilibrium exists. In any equilibrium, RN's strategy π_r^* generates exactly one durable news realization s^d and $S^{PI}(\pi_r^*) = S^D(\pi_r^*) = \{s^d\}$, and the equilibrium action satisfies $a^* = a^1$ if s^d is realized and $a^* = a^0$ otherwise. Moreover, in any equilibrium, the induced function $H_{\pi_r^*}(x)$ satisfies:*

   (a) there exists $\hat{\omega} \in (\gamma, \overline{\omega}]$ such that $H_{\pi_r^*}(\hat{\omega}) = H_{\underline{\pi}}(\hat{\omega})$, $H'_{\pi_r^*}(\hat{\omega}^-) = H'_{\pi_r^*}(\gamma)$, and $H'_{\pi_r^*}(\hat{\omega}^+) = 1$,

   (b) $H_{\pi_r^*}(\gamma) = \rho^{-1} H_{\underline{\pi}}(\gamma)$,

(c) $H'_{\pi_r^*}(\gamma) = \sup_{x \in [\underline{\omega}, \gamma)} \frac{H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x - \gamma}$.

Proof. See Appendix A.5. □

Theorem 1 generalizes Proposition 1 by showing how the presence of fake news reshapes the standard persuasion problem through a single parameter, the prominence of related news ρ. In the benchmark without fake news, RN can secure persuasion power as long as she can design a sufficiently informative signal. With fake news, informativeness is no longer enough: RN needs both an informative signal and a news environment in which her reports appear with probability above the threshold $\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$.

The theorem divides the analysis into two regimes. When ρ falls below the threshold, FN can always dilute any potentially influential related news, so every equilibrium collapses to the uninformative outcome in which the median voter chooses a^0 regardless of the observed news. In this regime, signal design by RN has no bite, and her equilibrium payoff remains equal to 0.

When ρ exceeds the threshold, RN recovers persuasive power despite the presence of fake news. In this case, Theorem 1 shows that RN can construct durable related news stories that FN cannot dilute, and that an optimal durable strategy maximizes the probability with which such news leads voters to choose a^1. RN's equilibrium payoff then increases with ρ, so a more prominent related news source both raises the likelihood of persuasive durable news and strengthens its electoral impact.

More broadly, Theorem 1 highlights that the effect of fake news is not only to add noise to the information environment, but to impose a legitimacy requirement on persuasion. Persuasion by RN is possible only if her ex ante prominence in the news media is high enough to support durable messages. Below that level, fake news fully erodes her influence; above it, persuasion survives in a restricted but well-characterized form.

4 Welfare Analysis

In this section, we explore the impact of fake news on social welfare, defined as the aggregate utilitarian welfare of all voters. This is precisely where modeling a continuum of voters, rather than a single representative receiver, has bite: although the equilibrium characterization of Section 3 depends only on the median voter's threshold γ, social welfare aggregates the realized utility of every voter, and so depends on the full distribution F rather than its median alone.

Welfare objects Sincere voting under the simple majority rule induces the voting outcome as a function of the common belief $\mu \in \Delta(\Omega)$: the outcome is a^1 if $E_\mu(\omega) \geq \gamma$ and a^0 otherwise. Since $u(a^0; \omega, x) = 0$ for every voter, the interim expected aggregate welfare under a common belief μ is

$$\int_{-1}^{1} E_\mu\big[u\big(a; \omega, x\big)\big]\, dF(x) = \begin{cases} E(x) + E_\mu(\omega) & \text{if } E_\mu(\omega) \geq \gamma \\ 0 & \text{if } E_\mu(\omega) < \gamma. \end{cases}$$

Under the equilibrium strategies (π_r^*, π_f^*), each news story $s \in \tilde{S}$ is observed with probability $\hat{\psi}_s(\pi_f^*; \pi_r^*)$ given by (2) and induces the posterior $\hat{\mu}_s$ given by (1). Writing $\hat{\psi}_s$ for $\hat{\psi}_s(\pi_f^*; \pi_r^*)$, the ex-ante social welfare is therefore

$$\begin{aligned} V\big(\rho, \gamma, E(x)\big) &= \sum_{s:\, E_{\hat{\mu}_s}(\omega) \geq \gamma} \hat{\psi}_s\big(E(x) + E_{\hat{\mu}_s}(\omega)\big) \\ &= \underbrace{\sum_{s:\, E_{\hat{\mu}_s}(\omega) \geq \gamma} \hat{\psi}_s\big(E_{\hat{\mu}_s}(\omega) - \gamma\big)}_{V_I(\rho,\gamma)\text{: Information Surplus}} + \underbrace{\sum_{s:\, E_{\hat{\mu}_s}(\omega) \geq \gamma} \hat{\psi}_s\big(E(x) - \mathrm{Med}(x)\big)}_{V_A(\rho,\gamma,E(x))\text{: Aggregation Surplus}}. \end{aligned} \tag{6}$$

The decomposition separates two conceptually distinct components. The *information surplus* V_I measures how far persuasion moves the public posterior past the median voter's threshold; it is the only component that would survive in a single-receiver model. The *aggregation surplus* V_A is the wedge between the average voter $E(x)$ and the median voter $\mathrm{Med}(x)$, weighted by the equilibrium probability that the voting outcome is a^1; it exists only because welfare aggregates a heterogeneous electorate while the voting rule listens to the median alone.

While the appropriate welfare object is the total $V\big(\rho, \gamma, E(x)\big)$, the central results of this paper concern the *informational* consequences of fake news—the self-disciplining of RN's signal characterized in Section 3. Accordingly, in the main text we conduct the comparative statics on the information surplus $V_I(\rho, \gamma)$, which isolates precisely this channel. The aggregation surplus, by contrast, reflects a divergence between majority rule and the utilitarian aggregate that operates independently of the informational mechanism; we defer its analysis, and the characterization of total welfare it delivers, to Online Appendix OA.

While one might intuitively expect that the proliferation of fake news would deteriorate the quality of information available to voters, our findings suggest a more nuanced relationship. The following proposition characterizes the behavior of the information surplus in equilibrium.

Proposition 5 (Information Surplus Gain with the Fake News)**.** *In equilibrium, the ex-ante*

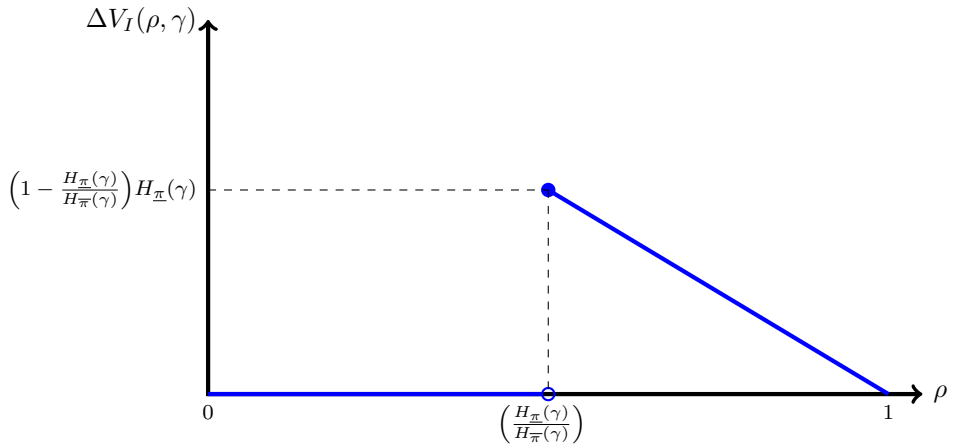

Figure 5: Difference between the information surplus: with FN v. without FN

information surplus is weakly higher in the presence of fake news than in the benchmark scenario without it, $\Delta V_I(\rho,\gamma)\Big(\equiv V_I(\rho,\gamma)-V_I(1,\gamma)\Big)\geq 0$. *Moreover,* $\Delta V_I(\rho,\gamma)$ *is non-monotonic in the prominence of related news,* ρ. *In particular,*

1. *for any* $\rho\in[0,1]$, $\Delta V_I(\rho,\gamma)\geq 0$.[12]
2. *If* $\rho<\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, $\frac{\partial\Delta V_I(\rho,\gamma)}{\partial\rho}=0$
3. *If* $\rho\geq\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, $\frac{\partial\Delta V_I(\rho,\gamma)}{\partial\rho}=-H_{\underline{\pi}}(\gamma)<0$

Proof. See Appendix A.6. □

Figure 5 illustrates the non-monotonic relationship between information surplus and source prominence. In the persuasive regime, where $\rho\geq\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, voters paradoxically benefit from a more pervasive fake news environment, as indicated by the negative slope.

The intuition behind Proposition 5 mirrors the rationale in Remark 3. When the related news sender retains sufficient credibility ($\rho\geq\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$), the threat of dilution by the fake news sender forces her to adopt a more informative reporting strategy. Specifically, to

[12]The welfare gain relies on a specific configuration of preferences and prior, which is worth making explicit. The mechanism requires that the sender who must persuade voters away from the prior-favored default action is the one endowed with the informative, state-contingent technology, while the state-independent mimic's preference coincides with the default. Two boundary cases clarify this. If FN's preference instead aligned with RN's (both favoring a^1), FN would have no incentive to dilute and the strategic tension—hence the welfare question—would vanish. If the prior instead favored a^1, the burden of persuasion would shift to the sender preferring a^0, namely FN; but FN cannot persuade, as he lacks a state-contingent technology. The self-disciplining gain therefore arises not from the particular labels a^0 and a^1, but from the structural feature that the informed sender is the one facing the persuasion problem against a same-message mimic aligned with the default. This is the economically natural configuration for a fake-news environment, in which the party able to convey information is also the party seeking to shift beliefs.

remain durable, the related news must be more closely associated with the state of nature than it would be in a monopolistic environment. This informational gain from the self-disciplined signal outweighs the loss incurred from the potential receipt of unrelated fake news. Consequently, as the prominence of fake news increases (a decrease in ρ), the voters become better informed from an ex-ante perspective. This beneficial effect, however, is bounded. If fake news becomes excessively dominant ($\rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$), the related news sender can no longer produce a durable signal. In this regime, all news stories become uninfluential, and the ex-ante welfare reverts to the level achievable without any information.

One may wonder why the information surplus gain $\Delta V_I(\rho, \gamma)$ is linearly decreasing in ρ when $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$. Here, we provide the economic implications and intuition behind this.

Remark 4 (Economic Implications of Linearity)**.** *The equilibrium information surplus gain from the existence of FN is given by*

$$\Delta V_I(\rho, \gamma) = (1 - \rho) H_{\underline{\pi}}(\gamma),$$

which is linear in $(1-\rho)$ *for* $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$*. This linearity implies a constant marginal effect: each unit increase in the prevalence of fake news—corresponding to a proportional decrease in* ρ*—raises the surplus gain by a constant amount,* $H_{\underline{\pi}}(\gamma)$*. The structure arises from the linearity of voters' utility functions and the additive nature of the probability-weighted information transmission. Within the persuasive regime, therefore, the informational benefit of fake news accrues at a constant rate: there are no diminishing returns to FN's prevalence until the threshold* $\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$ *is crossed, at which point the benefit collapses to zero.*

To summarize, our welfare analysis reveals a fundamental trade-off within the persuasive regime. While the presence of a strategic mimic introduces uninformative noise into the environment, it simultaneously compels the informed sender to adopt a higher-precision strategy to maintain signal durability. In equilibrium, the informational gain from this self-disciplined reporting strictly outweighs the loss from the mimic's noise. This result provides a formal justification for the paradox of welfare-enhancing strategic noise, showing that competition under source uncertainty can, under specific conditions, lead to better-informed collective decision-making.

5 Conclusion

In this paper, we characterize a Bayesian persuasion game defined by source uncertainty and the presence of fabricated content. Our analysis reveals that the impact of "fake news"

depends critically on its prominence relative to related news. When fake news is sufficiently prominent, the related news sender is unable to persuade voters; the fake news sender can "dilute" any potentially influential story by producing prominent, state-independent noise, thereby reducing the correlation between the observed news and the state of nature. In this regime, the informational value of the news environment collapses, and voters invariably choose the action preferred by the fake news sender.

However, when the related news sender retains a minimum threshold of ex-ante prominence, she can exert influence by producing news that is "durable" against dilution. Paradoxically, we find that the ex-ante expected utility of voters increases with the prominence of fake news in this regime. To withstand the threat of dilution, the related news sender is compelled to "self-discipline" her reporting strategy, generating news stories more closely associated with the state of nature than she would in the absence of a competitor.

By abstracting away from the specific "Fake News" narrative, these findings contribute more broadly to the literature on *asymmetric information and signal robustness*. We demonstrate that when a sender's legitimacy is at risk due to the presence of costless, state-independent mimics, the optimal strategic response is not to withdraw, but to increase the precision of information. This *"precision-as-defense" mechanism* suggests that competitive noise can, paradoxically, serve as a catalyst for informational quality in environments characterized by source anonymity, provided the expert retains a sufficient level of ex-ante credibility.

Our method rests on two structural assumptions that delimit its scope, and it is worth stating them together. The first is that the receiver's decision is binary, inducing a single belief threshold; this is what allows durability to be expressed as a scalar condition and characterized through the Rothschild-Stiglitz geometry, and it fits the up-or-down decisions—electoral and otherwise—that motivate the paper. The second concerns the information structure: A pivotal assumption in our model is that all voters observe the same news story. While this is a somewhat strong assumption, it is essential for the tractability of our geometric approach. If voters instead independently and randomly received news stories from different sources without knowing the origin, then, even under the simple majority rule and sincere voting, the proportion of voters holding each posterior would enter both the voting and the persuasion stages, and the Rothschild-Stiglitz construction would no longer apply. Relaxing either assumption—moving to multiple thresholds or to privately received signals—is a natural and important direction, and one that will require techniques beyond the geometric method we develop here.

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Battaglini, M. (2002). Multiple referrals and multidimensional cheap talk. *Econometrica*, 70(4):1379–1401.

Besley, T. and Prat, A. (2006). Handcuffs for the grabbing hand? media capture and government accountability. *American economic review*, 96(3):720–736.

Cheah, H. F. (2016). Does foreign media entry discipline or provoke local media bias? *International Journal of Economic Theory*, 12(4):335–359.

Doerr, B., Fouz, M., and Friedrich, T. (2012). Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75.

Enikolopov, R., Petrova, M., and Zhuravskaya, E. (2011). Media and political persuasion: Evidence from russia. *American Economic Review*, 101(7):3253–85.

Gehlbach, S. and Sonin, K. (2014). Government control of the media. *Journal of public Economics*, 118:163–171.

Gentzkow, M. and Kamenica, E. (2016). A rothschild-stiglitz approach to bayesian persuasion. *American Economic Review*, 106(5):597–601.

Gentzkow, M., Shapiro, J., and Stone, D. (2016). Media bias in the marketplace: Theory. handbook of media economics, 1, 623–645.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of political Economy*, 114(2):280–316.

Hindman, M. and Barash, V. (2018). Disinformation,'fake news' and influence campaigns on twitter. *Knight Foundation*.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Krishna, V. and Morgan, J. (2001). Asymmetric information and legislative rules: Some amendments. *The American Political Science Review*, 95(2):435–452.

Kwiek, M. (2026). Babel: Persuasion with anonymous senders. Working paper, University of Southampton.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Li, F. and Norman, P. (2021). Sequential persuasion. *Theoretical Economics*, 16(2):639–675.

Li, X. and Whinston, A. B. (2020). A model of fake data in data-driven analysis. *J. Mach. Learn. Res.*, 21:3–1.

Lipnowski, E., Ravid, D., and Shishkin, D. (2022). Persuasion via weak institutions. *Journal of Political Economy*, 130(10):2705–2730.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601.

Minozzi, W. (2011). A jamming theory of politics. *The Journal of Politics*, 73(2):301–315.

Mullainathan, S. and Shleifer, A. (2005). The market for news. *American economic review*, 95(4):1031–1053.

Mylovanov, T. and Zapechelnyuk, A. (2024). Constructive versus toxic argumentation in debates. *American Economic Journal: Microeconomics*, 16(1):262–292.

Risi, J. (2017). Taking stock of earned media in the "fake news" era. *HuffPost News*.

Shearer, E. and Gottfried, J. (2017). News use across social media platforms 2017. *Pew Research Center*.

Tandoc Jr, E. C., Duffy, A., Jones-Jang, S. M., and Wen Pin, W. G. (2021). Poisoning the information well? the impact of fake news on news media credibility. *Journal of Language and Politics*, 20(5):783–802.

Tandoc Jr., E. C., Lim, Z. W., and Ling, R. (2018). Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

A Proof of Propositions

A.1 Proof of Proposition 2

Notation. Rewrite Problem (3) as

$$\max_{\sigma_s \in \Pi_f} \sum_{s \in \tilde{S}} \hat{\psi}_s \, \mathbb{1}_{\{\tilde{s}: E_{\hat{\mu}_{\tilde{s}}}(\omega) < \gamma\}}(s), \tag{7}$$

where $\hat{\psi}_s$ is the probability that voters observe story s, $E_{\hat{\mu}_s}(\omega)$ is the posterior mean under $\hat{\mu}_s$, and $\gamma = -\operatorname{Med}(x)$. The function $\mathbb{1}_A(s)$ denotes the indicator of the event $s \in A$.

If $s \in \tilde{S} \setminus S$ is observed, then $\mathbb{1}_{\{\tilde{s}: E_{\hat{\mu}_{\tilde{s}}}(\omega) < \gamma\}}(s) = 1$ because voters identify s as originating from FN and therefore do not update from the prior. Since $E_{\mu_0}(\omega) < \gamma$, the voting outcome is a^0 whenever $s \in \tilde{S} \setminus S$ is transmitted.

If $s \in S$, then using the definition of $E_{\hat{\mu}_{\tilde{s}}}(\omega)$ in (1), the indicator in Problem (7) equals 1 if and only if

$$\psi_s\big(E_{\mu_s}(\omega) - \gamma\big) \;<\; \left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big)\,\sigma_s. \tag{8}$$

Let $Q(\rho) \equiv \left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big) > 0$ and define

$$k_s \;\equiv\; \frac{\psi_s\big(E_{\mu_s}(\omega) - \gamma\big)}{Q(\rho)}.$$

Properties of a best response $\pi_f \in BR(\pi_r)$. We establish properties that any best response $\pi_f \in BR(\pi_r)$ must satisfy.

First, suppose $E_{\mu_s}(\omega) < \gamma$, i.e., s is not potentially influential. Then (8) holds for every $\sigma_s \in [0,1]$. Hence assigning probability to such an s only increases the likelihood that the median voter receives a story that yields outcome a^0. In particular, if FN assigns probability σ_s to such s, the probability that the median voter receives s and chooses a^0 increases by $(1-\rho)\sigma_s$.

Next, suppose $E_{\mu_s}(\omega) \geq \gamma$, i.e., s is potentially influential. Consider two cases: (i) $k_s \in [0,1)$ and (ii) $k_s \geq 1$.

In Case (i), $k_s \in [0,1)$ implies that s is potentially influential and not durable (PIND). In this case, (8) can be satisfied by choosing σ_s sufficiently large. Assigning probability to such an s both dilutes it (so that the median selects a^0) and increases the probability that s is realized. In particular, if FN chooses σ_s large enough to satisfy (8), the probability that the median voter receives s and selects a^0 increases by $\rho\psi_s + (1-\rho)\sigma_s$. Therefore, whenever PIND

stories exist, allocating probability mass to dilute at least one such story yields a strictly higher payoff than allocating the same mass to a story that is not potentially influential (or to any $s \in \tilde{S} \setminus S$). Hence, if the set of PIND stories is nonempty given π_r, every $\pi_f \in BR(\pi_r)$ dilutes at least one story in this set. This establishes part 2-(a) of Proposition 2.

In Case (ii), $k_s \geq 1$ implies that (8) fails for all $\sigma_s \in [0,1]$. Hence s is durable and induces a^1 regardless of FN's action. Any best response therefore sets $\sigma_s = 0$ for such s, since choosing $\sigma_s > 0$ only increases the probability that the median voter receives s and selects a^1. More precisely, no $\pi_f \in BR(\pi_r)$ assigns positive probability to any durable story: reallocating that probability mass to a story that is not potentially influential (or to any $s \in \tilde{S} \setminus S$), or to a PIND story when one exists, strictly increases FN's payoff. This establishes part 1 of Proposition 2.

Existence of a best response $\pi_f \in BR(\pi_r)$. By constructing a best response π_f, we now directly prove the existence of the best response π_f and it assigns positive probability exclusively to PIND news.

Given π_r, FN's objective is to maximize the total probability with which the median voter selects a^0 subject to $\sum_s \sigma_s = 1$. Among related news, this reduces to selecting a subset of PIND stories to dilute. Formally, let

$$S^{\text{PIND}} \equiv \{\, s \in S : \; E_{\mu_s}(\omega) \geq \gamma, \; k_s \in [0,1) \,\}, \qquad T \equiv \Big\{ K \subset S^{\text{PIND}} : \; \sum_{s \in K} k_s < 1 \Big\}.$$

First, suppose $S^{\text{PIND}} = \emptyset$. Then, any π_f that places zero probability on durable related news with $E_{\mu_s}(\omega) \geq \gamma$ and assigns probability only to "not-potentially influential news s" with $E_{\mu_s}(\omega) < \gamma$ and/or to $\tilde{S} \setminus S$ would be a best response.

Second, suppose $S^{\text{PIND}} \neq \emptyset$. Since S is finite, T is nonempty and finite. For $K \in T$, define $M(K) \equiv \sum_{s \in K} \psi_s$; there exists $K^* \in T$ maximizing $M(K)$ ($\because T$ is finite.).

Construct a feasible best response σ^* as follows. For $s \in K^*$, set

$$\sigma_s^* = k_s + \varepsilon_s \quad \text{such that} \quad \varepsilon_s > 0, \quad \sum_{s \in K^*} \varepsilon_s = 1 - \sum_{s \in K^*} k_s, \quad \varepsilon_s \leq 1 - k_s$$

(For example, take $\varepsilon_s = \frac{1 - \sum_{t \in K^*} k_t}{|K^*|}$ for all $s \in K^*$.)[13] For $s \in S \setminus K^*$ with $E_{\mu_s}(\omega) \geq \gamma$ (durable or PIND not selected), set $\sigma_s^* = 0$. By construction, this strategy σ^* assigns positive

[13] $\varepsilon_s = \frac{1 - \sum_{t \in K^*} k_t}{|K^*|} > 0$ for all $s \in K^*$ and $\sum_{s \in K^*} \varepsilon_s = 1 - \sum_{s \in K^*} k_s$, by construction. If $|K^*| = 1$, then $\varepsilon_s = 1 - k_s$. Now for $|K^*| > 1$, suppose $\varepsilon_s = \frac{1 - \sum_{t \in K^*} k_t}{|K^*|} > 1 - k_s$ for some $s \in K^*$. Then $(1 - \sum_{t \in K^*} k_t) > |K^*|(1 - k_s) \Leftrightarrow (1 - |K^*|)(1 - k_s) > \sum_{t \in K^* \setminus \{s\}} k_t$. However, this implies that $(1 - |K^*|)(1 - k_s) > 0 \Leftrightarrow k_s > 1$, which is a contradiction. Thus, $\varepsilon_s = \frac{1 - \sum_{t \in K^*} k_t}{|K^*|} \leq 1 - k_s$ for all s.

probability exclusively to news stories in the set of PIND. Therefore, if we show σ^* is a best response to π_r, then we establish part 2-(b) of Proposition 2.

Feasibility is immediate, and for each $s \in K^*$,

$$\psi_s\big(E_{\mu_s}(\omega) - \gamma\big) \;=\; Q(\rho)\, k_s \;<\; Q(\rho)\, \sigma_s^*,$$

so (8) holds and s is diluted. Hence, σ^* dilutes at least one PIND story whenever $S^{\text{PIND}} \neq \emptyset$.

We now show σ^* is weakly better than any other strategy σ'. Consider any arbitrary feasible σ' and let the corresponding set K' as follows:

$$K' \;\equiv\; \{\, s \in S^{\text{PIND}} :\ (8) \text{ holds under } \sigma'_s \,\}.$$

If $K' = \emptyset$, then $\sum_{s\in K'} k_s = 0 < 1$, so $K' \in T$ trivially. If $K' \neq \emptyset$, then $\sum_{s\in K'} k_s < \sum_{s\in K'} \sigma'_s \leq \sum_s \sigma'_s = 1$, so $K' \in T$. In either case, $K' \in T$. FN's payoff under σ' is bounded above by $\sum_{s\in K'} \rho\psi_s\big(= \rho M(K')\big) + \sum_{s\in\{s:E_{\mu_s}(\omega)<\gamma\}} \rho\psi_s + (1-\rho)$. Since $M(K^*) \geq M(K')$, σ^* attains a weakly higher payoff than under σ'. That is, σ^* is a best response.

A.2 Proof of Proposition 3

We proceed by contradiction.

Suppose RN's strategy produces at least one durable news story and at least one potentially influential non-durable news story. By Proposition 2, FN allocates positive probability to diluting at least one potentially influential non-durable news story. Let s^{nd} denote such a non-durable news story and let s^d denote a durable news story.

Construct a pooled reporting strategy π'_r by replacing s^d and s^{nd} with a single realization s^* defined by $\psi_{s^*} \equiv \psi_{s^d} + \psi_{s^{nd}}$ and $E_{\mu_{s^*}}(\omega) \equiv \frac{\psi_{s^d}}{\psi_{s^*}} E_{\mu_{s^d}}(\omega) + \frac{\psi_{s^{nd}}}{\psi_{s^*}} E_{\mu_{s^{nd}}}(\omega)$. Since $E_{\mu_{s^d}}(\omega) \geq \gamma$ and $E_{\mu_{s^{nd}}}(\omega) \geq \gamma$ (potentially influential), their convex combination implies $E_{\mu_{s^*}}(\omega) \geq \gamma$.

Moreover,

$$\begin{aligned}
\frac{\psi_{s^*}\big(E_{\mu_{s^*}}(\omega) - \gamma\big)}{\left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big)} &= \frac{\psi_{s^*}}{\left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big)} \left\{ \frac{\psi_{s^d}}{\psi_{s^*}}\big(E_{\mu_{s^d}}(\omega) - \gamma\big) + \frac{\psi_{s^{nd}}}{\psi_{s^*}}\big(E_{\mu_{s^{nd}}}(\omega) - \gamma\big) \right\} \\
&= \frac{\psi_{s^d}\big(E_{\mu_{s^d}}(\omega) - \gamma\big)}{\left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big)} + \frac{\psi_{s^{nd}}\big(E_{\mu_{s^{nd}}}(\omega) - \gamma\big)}{\left(\frac{1-\rho}{\rho}\right)\big(\gamma - E_{\mu_0}(\omega)\big)} \\
&= k_{s^d} + k_{s^{nd}} \\
&\geq 1,
\end{aligned}$$

where $k_s \equiv \frac{\psi_s(E_{\mu_s}(\omega)-\gamma)}{Q(\rho)}$ and $Q(\rho) \equiv \left(\frac{1-\rho}{\rho}\right)\left(\gamma - E_{\mu_0}(\omega)\right)$.

Hence s^* is durable. The pooling preserves Bayes plausibility (the average posterior equals the prior), so replacing s^{nd} and s^d with s^* yields a feasible signal π_r'. By Proposition 2, FN does not dilute durable news; therefore under π_r' the probability mass $\psi_{s^{nd}}$—which was previously diluted—now leads to selection of a^1. RN's payoff under π_r' is thus strictly higher (by at least $\rho\,\psi_{s^{nd}}$), contradicting the supposed optimality of a strategy that includes a potentially influential non-durable news story.

A.3 RS method: no fake news

Here, we illustrate how RN's optimal reporting strategy is represented by a function H_π when FN does not exist, i.e. $\rho = 1$. The optimal persuasion strategy is studied in Kamenica and Gentzkow (2011). Let us denote the optimal strategy of RN by π_r^0 and its corresponding function by $H_{\pi_r^0}$.

Since there is no fake news sender, voters believe every news is generated by RN. In order to maximize the realization of all news s satisfying $E_{\hat{\mu}_s}(\omega) \geq \gamma$, the left-side slope of $H_{\pi_r^0}(\cdot)$ at γ should be the flattest as possible because the left-derivative of the function $H_{\pi_r^0}(\cdot)$ evaluated at γ is the sum of all probability weights put on the posterior means smaller than γ.

Note that the left-derivative of $H_{\pi_r^0}(\cdot)$ evaluated at γ can be reduced to the extent that the line, which is leaving from the point, $(\gamma, H_{\pi_r^0}(\gamma))$, to the left, pivots up to being tangent to $H_{\overline{\pi}}(\cdot)$ at some point $\omega^* \in [\underline{\omega}, \gamma)$. This condition is necessary for $H_{\pi_r^0}$ to be the function corresponding to an optimal strategy.

Also, we note that $H_{\pi_r^0}(\gamma) = H_{\underline{\pi}}(\gamma)$ holds in equilibrium. If it were not the case, i.e. $H_{\pi_r^0}(\gamma) > H_{\underline{\pi}}(\gamma)$, then the left-derivative of $H_{\pi_r^0}(\cdot)$ evaluated at γ can decrease by lowering the value of $H_{\pi_r^0}(\gamma)$. When $H_{\pi_r^0}(\gamma) = H_{\underline{\pi}}(\gamma)$ holds, $H'_{\pi_r^0}(\gamma^+) = 1 (= H'_{\underline{\pi}}(\gamma))$ should hold because $H_{\pi_r^0}$ is a piece-wise linear convex function and $H_{\pi_r^0}(\bar{\omega}) = H_{\underline{\pi}}(\bar{\omega})$.

These conditions uniquely determine the value of $1 - H'_{\pi_r^0}(\gamma^-)$, which is equal to the payoff of RN. That is, these conditions are sufficient to have all $H_{\pi_r^0}(\cdot)$ corresponding to optimal strategies, π_r^0. Since any distribution of posterior means which are below ω^* does not affect the payoff of RN, the shape of $H_{\pi_r^0}$ on this domain can be arbitrary. So, we have an infinite number of optimal strategies represented by $H_{\pi_r^0}$ as described by the blue lines in Figure 6.

A.4 Proof of Proposition 4

Proof. ($\Rightarrow$) Suppose RN's signal π_r produces a durable news story s. Then $H_{\pi_r}(x)$ can be constructed from π_r and $(H'_{\pi_r}(E_{\mu_s}(\omega)^+) - H'_{\pi_r}(E_{\mu_s}(\omega)^-))(E_{\mu_s}(\omega) - \gamma) \geq Q(\rho)$ holds for the

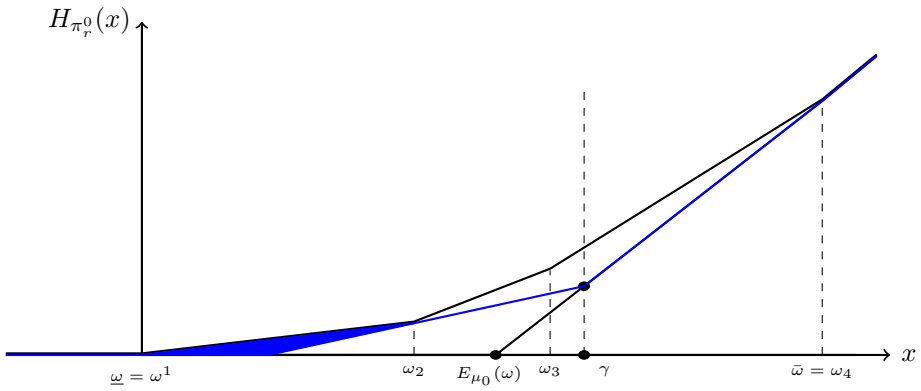

Figure 6: Optimal signal π_r^0 in the absence of fake news

durable news story s.

First we note that $H_{\pi_r}(x)$ is a piece-wise linear convex function. By convexity, $H'_{\pi_r}(E_{\mu_s}(\omega)^-) \geq \frac{H_{\pi_r}(E_{\mu_s}(\omega)) - H_{\pi_r}(\gamma)}{E_{\mu_s}(\omega) - \gamma}$ holds. So,

$$H_{\pi_r}(\gamma) \geq H_{\pi_r}(E_{\mu_s}(\omega)) - H'_{\pi_r}(E_{\mu_s}(\omega)^-)(E_{\mu_s}(\omega) - \gamma). \tag{9}$$

Also, we note that for all $x > \gamma$, $H'_{\underline{\pi}}(x) \geq H'_{\pi_r}(x)$ and $H_{\underline{\pi}}(x) \leq H_{\pi_r}(x)$ hold. So,

$$H'_{\underline{\pi}}(E_{\mu_s}(\omega)^+)(E_{\mu_s}(\omega) - \gamma) - H_{\underline{\pi}}(E_{\mu_s}(\omega)) \geq H'_{\pi_r}(E_{\mu_s}(\omega)^+)(E_{\mu_s}(\omega) - \gamma) - H_{\pi_r}(E_{\mu_s}(\omega)). \tag{10}$$

As the left-hand side of the inequality is equal to $-H_{\underline{\pi}}(\gamma)$, we have

$$-H_{\underline{\pi}}(\gamma) \geq H'_{\pi_r}(E_{\mu_s}(\omega)^+)(E_{\mu_s}(\omega) - \gamma) - H_{\pi_r}(E_{\mu_s}(\omega)) \tag{11}$$

Combining (9) and (11) gives the following:

$$H_{\pi_r}(\gamma) - H_{\underline{\pi}}(\gamma) \geq (H'_{\pi_r}(E_{\mu_s}(\omega)^+) - H'_{\pi_r}(E_{\mu_s}(\omega)^-))(E_{\mu_s}(\omega) - \gamma) \tag{12}$$

Since $H_{\overline{\pi}}(\gamma) \geq H_{\pi_r}(\gamma)$, it follows that $H_{\overline{\pi}}(\gamma) - H_{\underline{\pi}}(\gamma) \geq Q(\rho)$.

($\Leftarrow$) It suffices to show that there exist RN's strategy π_r such that for some news story s,

$$(H'_{\pi_r}(E_{\mu_s}(\omega)^+) - H'_{\pi_r}(E_{\mu_s}(\omega)^-))(E_{\mu_s}(\omega) - \gamma) \;=\; H_{\overline{\pi}}(\gamma) - H_{\underline{\pi}}(\gamma)$$

holds. So, s is durable.

We construct a non-decreasing, piece-wise linear, convex function $H_{\pi_r}(x)$ defined on $[\underline{\omega}, \overline{\omega}]$, which is bounded above by $H_{\overline{\pi}}(x)$ and below by $H_{\underline{\pi}}(x)$. Specifically, we define $H_{\pi_r}(x)$

as follows:

$$H_{\pi_r}(x) \;=\; \begin{cases} H_{\overline{\pi}}(x) & \text{if } \underline{\omega} \le x < \gamma, \\ H'_{\overline{\pi}}(\gamma^-)\big(x-\gamma\big) + H_{\overline{\pi}}(\gamma) & \text{if } \gamma \le x \le \hat{\omega}, \\ H_{\underline{\pi}}(x) & \text{if } \hat{\omega} < x \le \bar{\omega}, \end{cases}$$

where

$$\hat{\omega} \;\equiv\; \frac{H_{\overline{\pi}}(\gamma) - \gamma\, H'_{\overline{\pi}}(\gamma^-) + E_{\mu_0}(\omega)}{1 - H'_{\overline{\pi}}(\gamma^-)}.$$

To show that it is well-defined, first we want to show that $\hat{\omega}$ is uniquely determined. On $[\gamma, \overline{\omega}]$, we want to show that $H'_{\overline{\pi}}(\gamma^-)\big(x-\gamma\big) + H_{\overline{\pi}}(\gamma)$ and $H_{\underline{\pi}}(x)$ are single-crossing at $\hat{\omega}$.

That $H_{\underline{\pi}}(\gamma) < H_{\overline{\pi}}(\gamma)$ is obvious.

Note that for all π, we have $H_{\pi}(\overline{\omega}) = \overline{\omega} - E_{\mu_0}(\omega)$, implying that $H_{\overline{\pi}}(\bar{\omega}) = H_{\underline{\pi}}(\bar{\omega})$. By the convexity of $H_{\overline{\pi}}(x)$, it holds that

$$H'_{\overline{\pi}}(\gamma^-)(\bar{\omega} - \gamma) + H_{\overline{\pi}}(\gamma) \le H_{\overline{\pi}}(\bar{\omega}) \;=\; H_{\underline{\pi}}(\bar{\omega}).$$

As both $H'_{\overline{\pi}}(\gamma^-)\big(x-\gamma\big) + H_{\overline{\pi}}(\gamma)$ and $H_{\underline{\pi}}(x)$ are strictly increasing on $[\gamma, \overline{\omega}]$, these two lines are single-crossing and, by construction, intersects once at $x = \hat{\omega}$.

It can be easily shown that $H_{\pi_r}(x)$ is non-decreasing and piece-wise linear. Also, $H_{\pi_r}(x)$ is convex because $H'_{\pi_r}(x^+)$ and $H'_{\pi_r}(x^-)$ are non-decreasing.

The function $H_{\pi_r}(x)$ corresponds to a signal π_r which induces a news story s such that

$$E_{\mu_s}(\omega) = \hat{\omega} \quad \text{and} \quad \psi_s = 1 - H'_{\overline{\pi}}(\gamma^-).$$

Indeed,

$$\begin{aligned} \psi_s\Big(E_{\mu_s}(\omega) - \gamma\Big) &= \Big(1 - H'_{\overline{\pi}}(\gamma^-)\Big)\Big(\hat{\omega} - \gamma\Big) \\ &= H_{\overline{\pi}}(\gamma) - \gamma\, H'_{\overline{\pi}}(\gamma^-) + E_{\mu_0}(\omega) - \gamma\Big(1 - H'_{\overline{\pi}}(\gamma^-)\Big) \\ &= H_{\overline{\pi}}(\gamma) + E_{\mu_0}(\omega) - \gamma \\ &= H_{\overline{\pi}}(\gamma) - H_{\underline{\pi}}(\gamma). \end{aligned}$$

That is, the news story s is durable. □

A.5 Proof of Theorem 1

Case 1. $\rho < \frac{H_\pi(\gamma)}{H_{\tilde{\pi}}(\gamma)}$**.** In this case, we establish the first statement of Theorem 1 with the aid of the following lemma.

Lemma 3. *Suppose $\rho < \frac{H_\pi(\gamma)}{H_{\tilde{\pi}}(\gamma)}$. For any RN strategy π_r, every potentially influential non-durable (PIND) news story $s \in S^{\text{PIND}}$ is diluted in every best response $\pi_f \in BR(\pi_r)$ of FN.*

Proof. Suppose not. That is, assume there exists a best response $\pi_f \in BR(\pi_r)$ with which $s^d \in S^{\text{PIND}}$ is not diluted. Since π_f is a best response, it must select a subset of news stories to dilute, K^*, that solves the following maximization problem (as established in the proof of Proposition 2):

$$K^* \in \arg\max_{K \in T} \sum_{s \in K} \psi_s$$

where $T \equiv \left\{ K \subset S^{\text{PIND}} : \sum_{s \in K} k_s < 1 \right\}$, $k_s \equiv \frac{\psi_s \left(E_{\mu_s}(\omega) - \gamma\right)}{Q(\rho)}$, and $Q(\rho) \equiv \left(\frac{1-\rho}{\rho}\right)\left(\gamma - E_{\mu_0}(\omega)\right)$.

Since $s^d \in S^{\text{PIND}}$ is not diluted, $s^d \notin K^*$. Because diluting an additional PIND news story would strictly increase FN's objective function (by adding $\psi_{s^d} > 0$), the only reason s^d is excluded from the optimal set K^* is that adding it would violate the capacity constraint. Therefore, it must hold that

$$\sum_{s \in K^* \cup \{s^d\}} k_s \geq 1.$$

Multiplying both sides by $Q(\rho)$ yields $\sum_{s \in K^* \cup \{s^d\}} k_s Q(\rho) \geq Q(\rho)$. We can rewrite the left-hand side using the definition of k_s:

$$\begin{aligned}
\sum_{s \in K^* \cup \{s^d\}} k_s Q(\rho) &= \sum_{s \in K^* \cup \{s^d\}} \psi_s \left(E_{\mu_s}(\omega) - \gamma\right) \\
&= \left(\sum_{s \in K^* \cup \{s^d\}} \psi_s \right) \left(\left(\sum_{s \in K^* \cup \{s^d\}} \frac{\psi_s}{\sum_{s \in K^* \cup \{s^d\}} \psi_s} E_{\mu_s}(\omega) \right) - \gamma \right) \\
&= \psi_{\tilde{s}} \left(E_{\mu_{\tilde{s}}}(\omega) - \gamma\right) \geq Q(\rho).
\end{aligned}$$

where $\psi_{\tilde{s}} \equiv \sum_{s \in K^* \cup \{s^d\}} \psi_s$ and $E_{\mu_{\tilde{s}}}(\omega)$ is the probability-weighted average of the posterior means in $K^* \cup \{s^d\}$.

We can now construct a new RN strategy $\hat{\pi}_r$ by pooling the realizations in $K^* \cup \{s^d\}$ into a single realization $\tilde{s}$, while keeping all other realizations $s \in S \setminus (K^* \cup \{s^d\})$ unchanged. By Bayes plausibility, this new posterior distribution is valid.

However, the inequality $\psi_{\tilde{s}}\big(E_{\mu_{\tilde{s}}}(\omega)-\gamma\big)\geq Q(\rho)$ implies that the news story $\tilde{s}$ is durable. The existence of a strategy $\hat{\pi}_r$ that generates a durable news story contradicts Proposition 4, which states that no durable news can be produced when $\rho<\frac{H_\pi(\gamma)}{H_{\overline{\pi}}(\gamma)}$. Thus, the initial assumption must be false. □

Completion of Proof for Theorem 1 (Case 1):
Consider the case where $\rho<\frac{H_\pi(\gamma)}{H_{\overline{\pi}}(\gamma)}$. By Proposition 4, RN cannot produce any durable news, implying $S^D(\pi_r)=\emptyset$ for any $\pi_r\in\Pi_r$. Consequently, every potentially influential news story $s\in S^{PI}(\pi_r)$ is non-durable ($s\in S^{\text{PIND}}$). By Lemma 3, FN's best response π_f^* must dilute every such story, ensuring $E_{\hat{\mu}_s}(\omega)<\gamma$ for all $s\in S$.

Under these conditions, the decision maker's sequentially rational action is $a^*=a^0$ for any observed signal $s\in\tilde{S}$. Since the outcome a^0 and RN's payoff of 0 are invariant to her choice of reporting strategy, it follows that any $\pi_r\in\Pi_r$ satisfies the equilibrium condition for RN. This completes the proof of the first statement of Theorem 1. □

Case 2. $\rho\geq\frac{H_\pi(\gamma)}{H_{\overline{\pi}}(\gamma)}$**.** In this case, we establish the second statement of Theorem 1 through the following series of lemmas.

Lemma 4. *Suppose that $\rho\geq\frac{H_\pi(\gamma)}{H_{\overline{\pi}}(\gamma)}$. In equilibrium, for any RN's strategy that generates a finite number of durable news stories, there exists a payoff-equivalent strategy with one durable news story.*

Proof. Suppose there is a RN's strategy π_r generating N durable news stories. Let $S^D(\pi_r)$ denote the set of these durable news stories, so $|S^D(\pi_r)|=N\geq 1$. If $N=1$, the result holds trivially. Assume $N\geq 2$.

Note that for the pooled realization $\tilde{s}$, the durability condition aggregates as follows:

$$\begin{aligned}\sum_{s\in S^D(\pi_r)}\psi_s\big(E_{\mu_s}(\omega)-\gamma\big)&=\Bigg(\sum_{s\in S^D(\pi_r)}\psi_s\Bigg)\Bigg(\Bigg(\sum_{s\in S^D(\pi_r)}\frac{\psi_s}{\sum_{s\in S^D(\pi_r)}\psi_s}E_{\mu_s}(\omega)\Bigg)-\gamma\Bigg)\\&=\psi_{\tilde{s}}\big(E_{\mu_{\tilde{s}}}(\omega)-\gamma\big)\geq NQ(\rho)>Q(\rho).\end{aligned}$$

where

$$\psi_{\tilde{s}}\equiv\sum_{s\in S^D(\pi_r)}\psi_s\quad\textit{and}\quad E_{\mu_{\tilde{s}}}(\omega)\equiv\sum_{s\in S^D(\pi_r)}\frac{\psi_s}{\psi_{\tilde{s}}}E_{\mu_s}(\omega).$$

From Bayes plausibility, the original strategy satisfies:

$$\begin{aligned} E_{\mu_0}(\omega) &= \sum_{s\in S\setminus S^D(\pi_r)} \psi_s E_{\mu_s}(\omega) + \sum_{s\in S^D(\pi_r)} \psi_s E_{\mu_s}(\omega) \\ &= \sum_{s\in S\setminus S^D(\pi_r)} \psi_s E_{\mu_s}(\omega) + \psi_{\tilde{s}} E_{\mu_{\tilde{s}}}(\omega). \end{aligned}$$

Thus, we can construct a new strategy $\hat{\pi}_r$ that generates the posterior distribution supported on:

$$\Big(S \setminus S^D(\pi_r)\Big) \cup \{\tilde{s}\}.$$

By Proposition 3, the original equilibrium strategy π_r generates no potentially influential non-durable news; thus, all news stories in $S \setminus S^D(\pi_r)$ result in the action a^0 (either because they are uninfluential or successfully diluted). Under $\hat{\pi}_r$, these stories remain unchanged.

Regarding $\tilde{s}$, since $\psi_{\tilde{s}}\big(E_{\mu_{\tilde{s}}}(\omega) - \gamma\big) > Q(\rho)$, $\tilde{s}$ is durable. By Proposition 2, FN's best response assigns zero probability to durable news. Therefore, under $\hat{\pi}_r$, the news $\tilde{s}$ is never diluted and always induces action a^1.

RN's expected payoff under the new strategy is $\rho\psi_{\tilde{s}}$. Since $\psi_{\tilde{s}} = \sum_{s\in S^D(\pi_r)} \psi_s$, this is identical to the payoff of the original strategy π_r.

□

Lemma 5. *Suppose that $\rho \geq \frac{H_\pi(\gamma)}{H_{\overline{\pi}}(\gamma)}$. If an equilibrium strategy of RN generates exactly one durable news story, denoted by s^d, then the durability constraint must bind:*

$$\psi_{s^d}(E_{\mu_{s^d}}(\omega) - \gamma) = Q(\rho).$$

Proof. Suppose not. That is, assume there exists an equilibrium strategy π_r generating exactly one durable news story s^d such that the constraint is slack:

$$\psi_{s^d}(E_{\mu_{s^d}}(\omega) - \gamma) > Q(\rho).$$

By Proposition 3, an equilibrium strategy produces no potentially influential non-durable news. Therefore, the set of news stories S induced by π_r can be partitioned into $\{s^d\}$ (where $E_{\mu_{s^d}}(\omega) > \gamma$) and $S \setminus \{s^d\}$ (where $E_{\mu_s}(\omega) < \gamma$).

Let us pool the non-durable news stories into a single realization $\tilde{s}$. Define

$$\psi_{\tilde{s}} \equiv \sum_{s\in S\setminus\{s^d\}} \psi_s \quad \text{and} \quad E_{\mu_{\tilde{s}}}(\omega) \equiv \sum_{s\in S\setminus\{s^d\}} \frac{\psi_s}{\psi_{\tilde{s}}} E_{\mu_s}(\omega).$$

By Bayes plausibility, $\psi_{\tilde{s}} E_{\mu_{\tilde{s}}}(\omega) + \psi_{s^d} E_{\mu_{s^d}}(\omega) = E_{\mu_0}(\omega)$. Since RN's payoff depends only on

ψ_{s^d}, any strategy π_r' that generates the distribution over posterior means $\{E_{\mu_{\tilde{s}}}(\omega), E_{\mu_{s^d}}(\omega)\}$ with probabilities $\{\psi_{\tilde{s}}, \psi_{s^d}\}$ yields the same payoff as π_r.

Rearranging the Bayes plausibility condition, we have:

$$\psi_{s^d}\big(E_{\mu_{s^d}}(\omega) - E_{\mu_{\tilde{s}}}(\omega)\big) = E_{\mu_0}(\omega) - E_{\mu_{\tilde{s}}}(\omega).$$

Substituting this into the durability condition, we can express the slack inequality as:

$$\begin{aligned}\psi_{s^d}(E_{\mu_{s^d}}(\omega) - \gamma) &= \psi_{s^d}\big(E_{\mu_{s^d}}(\omega) - E_{\mu_{\tilde{s}}}(\omega)\big) - \psi_{s^d}\big(\gamma - E_{\mu_{\tilde{s}}}(\omega)\big) \\ &= \big[E_{\mu_0}(\omega) - E_{\mu_{\tilde{s}}}(\omega)\big] - \psi_{s^d}\big(\gamma - E_{\mu_{\tilde{s}}}(\omega)\big) > Q(\rho).\end{aligned}$$

Note that $\gamma > E_{\mu_{\tilde{s}}}(\omega)$ holds because $E_{\mu_s}(\omega) < \gamma$ for all $s \in S \setminus \{s^d\}$.

In the expression above, if we fix $E_{\mu_{\tilde{s}}}(\omega)$, the term $\big[E_{\mu_0}(\omega) - E_{\mu_{\tilde{s}}}(\omega)\big]$ is constant. The durability measure is strictly decreasing in ψ_{s^d}. Therefore, since the inequality is strict, there exists $\varepsilon > 0$ such that RN can increase the probability of the durable news to $\psi'_{s^d} = \psi_{s^d} + \varepsilon$ (adjusting $E_{\mu_{s^d}}(\omega)$ downwards to maintain Bayes plausibility) without violating the durability constraint. Such a deviation strictly increases RN's payoff, contradicting the optimality of π_r. □

Lemma 6. *Suppose that $\rho \geq \frac{H_\pi(\gamma)}{H_{\overline{\pi}}(\gamma)}$. In any equilibrium, RN generates exactly one durable news story.*

Proof. Suppose, to the contrary, that RN's equilibrium strategy π_r generates N durable news stories, with $N > 1$. By Lemma 4, there exists a strategy $\hat{\pi}_r$ generating exactly one durable news story, $\tilde{s}$, that yields the same expected payoff as π_r.

By construction, the durability measure of the pooled news $\tilde{s}$ is the sum of the measures of the N original durable stories:

$$\psi_{\tilde{s}}\big(E_{\mu_{\tilde{s}}}(\omega) - \gamma\big) \geq NQ(\rho) > Q(\rho).$$

This implies that the strategy $\hat{\pi}_r$ has a single durable news story with a slack durability constraint. As shown in the proof of Lemma 5, whenever the constraint is slack ($> Q(\rho)$), there exists a profitable deviation that increases RN's payoff.

Since $\hat{\pi}_r$ admits a profitable deviation and π_r yields the same payoff as $\hat{\pi}_r$, it follows that π_r also admits a profitable deviation. Thus, π_r cannot be an equilibrium strategy. □

Lemma 7 (Sufficiency of a persuasive equilibrium)**.** *Suppose that $\rho \geq \frac{H_\pi(\gamma)}{H_{\overline{\pi}}(\gamma)}$. RN's strategy π_r^* is an equilibrium strategy if the corresponding function $H_{\pi_r^*}(x)$ satisfies the following three conditions:*

1. $\exists!\hat{\omega} \in (\gamma, \overline{\omega}] \quad s.t. \quad H'_{\pi_r^*}(\hat{\omega}^-) = H'_{\pi_r^*}(\gamma)$, $H_{\pi_r^*}(\hat{\omega}) = H_{\underline{\pi}}(\hat{\omega})$, *and* $H'_{\pi_r^*}(\hat{\omega}^+) = 1$;

2. $H_{\pi_r^*}(\gamma) = \rho^{-1} H_{\underline{\pi}}(\gamma)$ *and*

3. $H'_{\pi_r^*}(\gamma) = \sup_{x \in [\underline{\omega}, \gamma)} \frac{H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x - \gamma}$.

Proof. When $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, durable news is feasible (Proposition 4). Also, RN generates exactly one durable news story s^d (Lemma 6) and no potentially influential non-durable news (Proposition 3) in equilibrium. In the following, we show that Condition 1 and Condition 2 about $H_{\pi_r^*}$ guarantee that π_r^* satisfies these necessary conditions for being an equilibrium strategy.

From Condition 1, there uniquely exists $\hat{\omega} \in (\gamma, \overline{\omega}]$ satisfying $H'_{\pi_r^*}(\hat{\omega}^-) = H'_{\pi_r^*}(\gamma)$, $H_{\pi_r^*}(\hat{\omega}) = H_{\underline{\pi}}(\hat{\omega})$, and $H'_{\pi_r^*}(\hat{\omega}^+) = 1$. By definition of $H_{\pi_r^*}$, there exists some $E_{\mu_s}(\omega)(= \hat{\omega})$ generated by a news story s under π_r^*. Note that s is the only realization with a posterior mean $E_{\mu_s}(\omega)$, strictly greater than γ. Indeed, $H_{\pi_r^*}(x)$ has no kink on the interval $[\gamma, E_{\mu_s}(\omega))$ as $H'_{\pi_r^*}(\hat{\omega}^-) = H'_{\pi_r^*}(\gamma)$. Also, $H_{\pi_r^*}$ re-anchors to the uninformative integral curve $H_{\underline{\pi}}$ at $E_{\mu_s}(\omega)(= \hat{\omega})$ as $H_{\pi_r^*}(\hat{\omega}) = H_{\underline{\pi}}(\hat{\omega})$, and $H'_{\pi_r^*}(\hat{\omega}^+) = 1$.

Now we argue that s is a durable news story, i.e. $s = s^d$, by showing that $\psi_s(\hat{\omega} - \gamma) = Q(\rho)$ holds. From the properties of H, we know $\psi_s = H'_{\pi_r^*}(\hat{\omega}^+) - H'_{\pi_r^*}(\hat{\omega}^-) = 1 - H'_{\pi_r^*}(\gamma)$. Using the linearity and continuity of $H_{\pi_r^*}$ on $[\gamma, \hat{\omega}]$:

$$\begin{aligned} H_{\pi_r^*}(\hat{\omega}) &= \lim_{x \uparrow \hat{\omega}} H_{\pi_r^*}(x) \\ &= \lim_{x \uparrow \hat{\omega}} H_{\pi_r^*}(\gamma) + H'_{\pi_r^*}(\gamma)(x - \gamma) \\ &= H_{\pi_r^*}(\gamma) + H'_{\pi_r^*}(\gamma)(\hat{\omega} - \gamma) \\ &= H_{\pi_r^*}(\gamma) + (1 - \psi_s)(\hat{\omega} - \gamma). \end{aligned}$$

As $H_{\underline{\pi}}(x) = x - E_{\mu_0}(\omega)$ holds for $x \geq E_{\mu_0}(\omega)$, rearranging the above and using Condition 2 give

$$\begin{aligned} \psi_s(\hat{\omega} - \gamma) &= H_{\pi_r^*}(\gamma) - H_{\underline{\pi}}(\hat{\omega}) + (\hat{\omega} - \gamma). \\ &= H_{\pi_r^*}(\gamma) + E_{\mu_0}(\omega) - \gamma \\ &= H_{\pi_r^*}(\gamma) - H_{\underline{\pi}}(\gamma) \\ &= (\rho^{-1} - 1) H_{\underline{\pi}}(\gamma) \\ &= Q(\rho). \end{aligned}$$

Now we use s^d instead of s. RN's payoff is $\rho \psi_{s^d} = \rho(1 - H'_{\pi_r^*}(\gamma))$ and Condition 3

pins down the payoff of π_r^* induced by $H_{\pi_r^*}$ that satisfies necessary conditions. Note that maximizing $\rho(1 - H'_{\pi_r^*}(\gamma))$ is equivalent to minimizing the slope $H'_{\pi_r^*}(\gamma)$. However, π_r^* must be feasible, meaning $H_{\pi_r^*}(x) \leq H_{\overline{\pi}}(x)$ for all x. Specifically focusing on the payoff-relevant part of $H_{\pi_r^*}$, the line passing through $(\gamma, H_{\pi_r^*}(\gamma))$ with slope $H'_{\pi_r^*}(\gamma)$ must stay below $H_{\overline{\pi}}(x)$ for all $x < \gamma$. This requires:

$$H_{\pi_r^*}(\gamma) + H'_{\pi_r^*}(\gamma)(x - \gamma) \leq H_{\overline{\pi}}(x) \quad \forall x < \gamma.$$

Rearranging terms (noting $x - \gamma < 0$ reverses the inequality and $H_{\pi_r^*}(\gamma) = \rho^{-1} H_{\underline{\pi}}(\gamma)$):

$$H'_{\pi_r^*}(\gamma) \geq \frac{H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x - \gamma}.$$

To minimize the slope, RN chooses the infimum of feasible slopes, which corresponds to the supremum of the right-hand side. Therefore, Condition 3 sets the value of $\rho\psi_{s^d}$ at maximum, characterizing the optimal strategies for RN.

□

Lemma 8 (Existence of a persuasive equilibrium)**.** *Suppose that* $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$. *RN's equilibrium strategy* π_r^* *exists.*

Proof. For a state space Ω and a prior $\mu_0 \in \Delta(\Omega)$, $H_{\overline{\pi}}(x)$ and $H_{\underline{\pi}}(x)$ are well-defined. Now we construct a piecewise linear convex function $H_{\pi_r^*}(x)$ placed in between $H_{\overline{\pi}}(x)$ and $H_{\underline{\pi}}(x)$, satisfying the three sufficient conditions explicitly.

Construction of $H_{\pi_r^*}(x)$: Let $S(x) \equiv \frac{H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x - \gamma}$ for $x < \gamma$. We argue that the supremum of $S(x)$ is attained at some x^* in $[\underline{\omega}, \gamma)$. Obviously, $S(\underline{\omega}) > 0$.

If $H_{\overline{\pi}}(\gamma) > \rho^{-1} H_{\underline{\pi}}(\gamma)$, then $\lim_{x \uparrow \gamma} S(x) = -\infty$. Since $S(x)$ is continuous on $[\underline{\omega}, \gamma)$, it attains a maximum at some $x^* \in [\underline{\omega}, \gamma)$.

If $H_{\overline{\pi}}(\gamma) = \rho^{-1} H_{\underline{\pi}}(\gamma)$, then $S(x)$ is the slope of a line connecting $(x, H_{\overline{\pi}}(x))$ and $(\gamma, H_{\overline{\pi}}(\gamma))$, and converges to $H'_{\overline{\pi}}(\gamma^-)$ as $x \uparrow \gamma$. Since $H_{\overline{\pi}}(x)$ is convex, $S(x)$ is non-decreasing and bounded by $H'_{\overline{\pi}}(\gamma^-)$. Indeed, the supremum of $S(x)$ is $H'_{\overline{\pi}}(\gamma^-)$. Otherwise, there exists some $\varepsilon > 0 \quad s.t. \quad S(x) \leq H'_{\overline{\pi}}(\gamma^-) - \varepsilon, \quad \forall x < \gamma$. However, this contradicts the fact that $S(x) \to H'_{\overline{\pi}}(\gamma^-) \quad as \quad x \uparrow \gamma$. Note that $H_{\overline{\pi}}(x)$ has a finite number of kinks by its construction. This implies that there exists some $\varepsilon > 0 \quad s.t. \quad S(x) = H'_{\overline{\pi}}(\gamma^-) \quad \forall x \in (\gamma - \varepsilon, \gamma)$. Thus, there exists some $x^* < \gamma$ that attains the supremum of $S(x)$.

Let $k = \sup_{x < \gamma} S(x)$. Define a linear function $L(x) = k(x - \gamma) + \rho^{-1} H_{\underline{\pi}}(\gamma)$ on $[\underline{\omega}, \overline{\omega}]$. Now we construct a function $H_{\pi_r^*}(x)$ as follows:

$$H_{\pi_r^*}(x) = \begin{cases} H_{\overline{\pi}}(x) & \text{if} \quad x \le x^* \\ L(x) & \text{if} \quad x^* \le x \le \hat{\omega} \\ H_{\underline{\pi}}(x) & \text{if} \quad x \ge \hat{\omega}. \end{cases}$$

Checking whether $H_{\pi_r^*}(x)$ is a well-defined piecewise linear function: First we note that

$$k = \frac{H_{\overline{\pi}}(x^*) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x^* - \gamma},$$

which implies $H_{\overline{\pi}}(x^*) = k(x^* - \gamma) + \rho^{-1} H_{\underline{\pi}}(\gamma) = L(x^*)$.

Second, $L(\gamma)(= \rho^{-1} H_{\underline{\pi}}(\gamma)) > H_{\underline{\pi}}(\gamma)$. Also, $L(\overline{\omega}) \le H_{\underline{\pi}}(\overline{\omega})$. Indeed, as $H_{\underline{\pi}}(\overline{\omega}) = H_{\overline{\pi}}(\overline{\omega})$,

$$\begin{aligned} H_{\underline{\pi}}(\overline{\omega}) - L(\overline{\omega}) =& H_{\overline{\pi}}(\overline{\omega}) - L(\overline{\omega}) \\ =& H_{\overline{\pi}}(\overline{\omega}) - H_{\overline{\pi}}(\gamma) + H_{\overline{\pi}}(\gamma) - \rho^{-1} H_{\underline{\pi}}(\gamma) - k(\overline{\omega} - \gamma) \\ =& (\overline{\omega} - \gamma) \left\{ \frac{H_{\overline{\pi}}(\overline{\omega}) - H_{\overline{\pi}}(\gamma)}{\overline{\omega} - \gamma} - \frac{H_{\overline{\pi}}(x^*) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x^* - \gamma} \right\} \\ & + H_{\overline{\pi}}(\gamma) - \rho^{-1} H_{\underline{\pi}}(\gamma) \\ \ge& 0. \end{aligned}$$

The last inequality follows from the fact that $\frac{H_{\overline{\pi}}(\overline{\omega}) - H_{\overline{\pi}}(\gamma)}{\overline{\omega} - \gamma} - \frac{H_{\overline{\pi}}(x^*) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x^* - \gamma} \ge 0$, due to convexity of $H_{\overline{\pi}}(x)$, and the assumption of $H_{\overline{\pi}}(\gamma) \ge \rho^{-1} H_{\underline{\pi}}(\gamma)$. As both $L(x)$ and $H_{\underline{\pi}}(x)$ are linear on $[\gamma, \overline{\omega}]$, two functions single-cross at a point, say $\hat{\omega}$, in $(\gamma, \overline{\omega}]$. With this $\hat{\omega}$, we have $L(\hat{\omega}) = H_{\underline{\pi}}(\hat{\omega})$. So, $H_{\pi_r^*}(x)$ is well-defined.

Checking whether $H_{\pi_r^*}(x)$ is non-decreasing and convex: It is obvious that $k > 0$. Thus, $L(x)$ is non-decreasing, concluding that $H_{\pi_r^*}(x)$ is non-decreasing.

For convexity, it suffices to show that $k < H'_{\underline{\pi}}(\hat{\omega}^+)(= 1)$ and $H'_{\overline{\pi}}(x^{*-}) \le k$. As $L(x)$ and $H_{\underline{\pi}}(x)$ single-crosses at $\hat{\omega}$ while $L(\gamma) > H_{\underline{\pi}}(\gamma)$, $k < H'_{\underline{\pi}}(\hat{\omega}^+)$ holds. To prove $H'_{\overline{\pi}}(x^{*-}) \le k$, we note that $S(x) \le S(x^*)(= k) \quad \forall x < x^*$. Then,

$$\begin{aligned} S(x) \le S(x^*) \Leftrightarrow & \frac{H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x - \gamma} \le \frac{H_{\overline{\pi}}(x^*) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x^* - \gamma} \\ \Leftrightarrow & \big(H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)\big)\big(x^* - \gamma\big) \le \big(H_{\overline{\pi}}(x^*) - \rho^{-1} H_{\underline{\pi}}(\gamma)\big)\big(x - \gamma\big) \\ \Leftrightarrow & \big(H_{\overline{\pi}}(x) - H_{\overline{\pi}}(x^*)\big)\big(x^* - \gamma\big) \le \big(H_{\overline{\pi}}(x^*) - \rho^{-1} H_{\underline{\pi}}(\gamma)\big)\big(x - x^*\big) \\ \Leftrightarrow & \frac{H_{\overline{\pi}}(x) - H_{\overline{\pi}}(x^*)}{x - x^*} \le \frac{H_{\overline{\pi}}(x^*) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x^* - \gamma} (= k). \end{aligned}$$

As the left-hand side of the last inequality holds for every $x < x^*$, this implies that $H'_{\overline{\pi}}(x^{*-}) \leq k$.

Checking whether $H_{\underline{\pi}} \leq H_{\pi_r^*} \leq H_{\overline{\pi}}$: It suffices to show that $H_{\underline{\pi}}(x) \leq L(x) \leq H_{\overline{\pi}}(x)$ on $[x^*, \hat{\omega}]$. First, we prove that $H_{\underline{\pi}}(x) \leq L(x)$ on $[x^*, \hat{\omega}]$. Note that $L(x^*) = H_{\overline{\pi}}(x^*) > H_{\underline{\pi}}(x^*)$ and $L(\hat{\omega}) = H_{\underline{\pi}}(\hat{\omega})$. It is obvious that the line connecting $L(x^*)$ and $L(\hat{\omega})$ is placed above or equal to the line connecting $H_{\underline{\pi}}(x^*)$ and $H_{\underline{\pi}}(\hat{\omega})$. As $H_{\underline{\pi}}(x)$ is convex on $[x^*, \hat{\omega}]$, $H_{\underline{\pi}}(x) \leq L(x)$ holds on that support.

Second, we prove that $L(x) \leq H_{\overline{\pi}}(x)$ on $[x^*, \hat{\omega}]$. By the definition of $k = \sup_{x<\gamma} S(x)$, $\frac{H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x-\gamma} \leq k \quad \forall x < \gamma$. This implies that $H_{\overline{\pi}}(x) \geq k(x-\gamma) + \rho^{-1} H_{\underline{\pi}}(\gamma) = L(x), \quad \forall x < \gamma$. Also, note that by convexity of $H_{\overline{\pi}}$,

$$\begin{aligned} k &= \frac{H_{\overline{\pi}}(x^*) - H_{\overline{\pi}}(\gamma)}{x^* - \gamma} + \frac{H_{\overline{\pi}}(\gamma) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x^* - \gamma} \\ &\leq \frac{H_{\overline{\pi}}(x^*) - H_{\overline{\pi}}(\gamma)}{x^* - \gamma} \\ &\leq H'_{\overline{\pi}}(\gamma^+) \end{aligned}$$

Thus,

$$\begin{aligned} L(x) &= k(x-\gamma) + \rho^{-1} H_{\underline{\pi}}(\gamma) \\ &\leq H'_{\overline{\pi}}(\gamma^+)(x-\gamma) + H_{\overline{\pi}}(\gamma) \\ &\leq H_{\overline{\pi}}(x), \quad \forall x \geq \gamma. \end{aligned}$$

Checking whether $H_{\pi_r^*}(x)$ **satisfies three sufficient conditions for equilibrium**: Recall that

$$H_{\pi_r^*}(x) = \begin{cases} H_{\overline{\pi}}(x) & \text{if} \quad x \leq x^* \\ L(x) & \text{if} \quad x^* \leq x \leq \hat{\omega} \\ H_{\underline{\pi}}(x) & \text{if} \quad x \geq \hat{\omega}. \end{cases}$$

As $\hat{\omega}$ is a unique kinked point of $H_{\pi_r^*}(x)$ on $(\gamma, \overline{\omega}]$ with $H'_{\pi_r^*}(\hat{\omega}^-) = H'_{\pi_r^*}(\gamma) = k$, $H_{\pi_r^*}(\hat{\omega}) = L(\hat{\omega}) = H_{\underline{\pi}}(\hat{\omega})$, and $H'_{\pi_r^*}(\hat{\omega}^+) = H'_{\underline{\pi}}(\hat{\omega}^+) = 1$, Condition 1 is satisfied. Also, as $H_{\pi_r^*}(\gamma) = L(\gamma) = \rho^{-1} H_{\underline{\pi}}(\gamma)$, Condition 2 is satisfied. Lastly, $H'_{\pi_r^*}(\gamma) = L'(\gamma) = k$, where $k = \sup_{x<\gamma} \frac{H_{\overline{\pi}}(x) - \rho^{-1} H_{\underline{\pi}}(\gamma)}{x-\gamma}$, satisfying Condition 3.

Completion of Proof for Theorem 1 (Case 2): Combined, the preceding lemmas provide a complete characterization of the equilibrium for $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$. Lemma 6 establishes the necessity of a single durable news realization s^d, while Proposition 3 ensures the absence

of non-durable influential stories. Lemma 7 provides the sufficient geometric conditions for the existence of an optimal strategy π_r^*, effectively defining the maximum attainable payoff for RN. Finally, the construction of an optimal $H_{\pi_r^*}(x)$ in Lemma 8 guarantees that such a strategy is feasible within the space of piece-wise linear convex functions. Thus, the second statement of Theorem 1 is fully established.

□

A.6 Proof of Proposition 5

Under the equilibrium distribution of posteriors $\hat{\psi}$ induced by the strategy pair (π_r^*, π_f^*), the ex-ante social welfare is:

$$\begin{aligned} V\big(\rho,\gamma,E(x)\big) &= \sum_{s:E_{\hat{\mu}_s}(\omega)\geq\gamma} \hat{\psi}_s\big(E(x)+E_{\hat{\mu}_s}(\omega)\big) \\ &= \underbrace{\sum_{s\in S^D(\pi_r^*)} \hat{\psi}_s\big(E_{\hat{\mu}_s}(\omega)-\gamma\big)}_{V_I(\rho,\gamma)\text{: Information Surplus}} + \underbrace{\sum_{s\in S^D(\pi_r^*)} \hat{\psi}_s\big(E(x)-\mathrm{Med}(x)\big)}_{V_A(\rho,\gamma,E(x))\text{: Aggregation Surplus}} \end{aligned} \tag{13}$$

where $S^D(\pi_r^*)$ is the set of durable signals under the equilibrium strategy π_r^*.

As established in Theorem 1, if $\rho < \frac{H_{\pi}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, no strategy π_r can produce durable news, hence $S^D(\pi_r^*) = \emptyset$. The fake news sender dilutes all potentially influential news, ensuring $V_I(\rho,\gamma) = 0$.

If $\rho \geq \frac{H_{\pi}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, the equilibrium strategy π_r^* generates exactly one durable signal $s^d \in S^D(\pi_r^*)$. In equilibrium, $\hat{\psi}_{s^d} = \rho\psi_{s^d}$ and $E_{\hat{\mu}_{s^d}} = E_{\mu_{s^d}}$ (since $\sigma_{s^d}^* = 0$: FN never adds mass to the durable story, as doing so only raises the probability of a^1). Substituting the binding durability constraint $\rho\psi_{s^d}(E_{\mu_{s^d}}(\omega)-\gamma) = (1-\rho)H_{\underline{\pi}}(\gamma)$ into the definition of V_I:

$$\begin{aligned} V_I(\rho,\gamma) &= \rho\psi_{s^d}(E_{\mu_{s^d}}(\omega)-\gamma) \\ &= (1-\rho)H_{\underline{\pi}}(\gamma) \end{aligned}$$

In the benchmark without fake news ($\rho = 1$), the related news sender sets $E_{\mu_s}(\omega) = \gamma$ for any influential signal, yielding $V_I(1,\gamma) = 0$. Thus, the surplus gain $\Delta V_I(\rho,\gamma) = V_I(\rho,\gamma) - V_I(1,\gamma) = (1-\rho)H_{\underline{\pi}}(\gamma)$ is nonnegative and decreasing in ρ for the persuasive regime. □

Online Appendix

OA Welfare with the Aggregation Surplus

In the main text, the comparative statics of Section 4 concern the information surplus $V_I(\rho,\gamma)$. This appendix completes the welfare analysis by reinstating the aggregation surplus $V_A\big(\rho,\gamma,E(x)\big)$ in (13) and characterizing the total welfare $V\big(\rho,\gamma,E(x)\big)$ for arbitrary configurations of $E(x)$ and $\text{Med}(x)$. Whenever $E(x)=\text{Med}(x)$, the aggregation surplus vanishes identically and the results of Section 4 are statements about total welfare; the analysis below is therefore of independent interest when $E(x)\neq\text{Med}(x)$.

Parallel to $\Delta V_I(\rho,\gamma)$ in Proposition 5, we evaluate the effect of fake news on total welfare by the differences

$$\Delta V\big(\rho,\gamma,E(x)\big) \;\equiv\; V\big(\rho,\gamma,E(x)\big)-V\big(1,\gamma,E(x)\big)$$

$$\text{and}\quad \Delta V_A\big(\rho,\gamma,E(x)\big) \;\equiv\; V_A\big(\rho,\gamma,E(x)\big)-V_A\big(1,\gamma,E(x)\big),$$

so that, by (13), $\Delta V=\Delta V_I+\Delta V_A$.

Two equilibrium quantities govern ΔV_A. First, in the benchmark without fake news ($\rho=1$), the related news sender generates a unique influential signal, whose posterior mean satisfies $E_{\mu_s}(\omega)=\gamma$ (see the proof of Proposition 5), with probability $\psi^0\equiv 1-H'_{\pi_r^0}(\gamma^-)$; hence

$$V_A\big(1,\gamma,E(x)\big)=\psi^0\big(E(x)-\text{Med}(x)\big).$$

Second, in the persuasive regime, the unique influential signal is the durable signal s^d, and $\hat{\psi}_{s^d}=\rho\psi_{s^d}$ in equilibrium (see the proof of Proposition 5); hence

$$V_A\big(\rho,\gamma,E(x)\big)=\rho\psi_{s^d}\big(E(x)-\text{Med}(x)\big).$$

The following lemma records how the probability $\rho\psi_{s^d}$ varies with ρ; it also formalizes the claim, made in the discussion following Theorem 1, that RN's equilibrium payoff $\rho\psi_{s^d}$ increases with ρ. To emphasize the dependence of the equilibrium objects on ρ, we write $\psi_{s^d}(\rho)$.

Lemma 9 (Monotonicity of the Equilibrium Persuasion Probability)**.** *Suppose $\rho\geq\frac{H_{\pi}(\gamma)}{H_{\overline{\pi}}(\gamma)}$. In equilibrium, the probability that voters observe the durable signal, $\rho\psi_{s^d}(\rho)$, is strictly*

increasing in ρ: wherever the derivative exists,

$$\frac{d\big(\rho\psi_{s^d}(\rho)\big)}{d\rho} \;=\; 1+\frac{H_{\overline{\pi}}(x^*)}{\gamma-x^*} \;>\; 0,$$

where $x^* \in [\underline{\omega},\gamma)$ attains the supremum in condition (c) of Theorem 1. Consequently, $\rho\psi_{s^d}(\rho) \le \psi^0$ for all $\rho \in \left[\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)},1\right]$, with equality at $\rho=1$.

Proof. See Appendix OA.1. □

Combining Proposition 5 with the expressions above, the total welfare difference decomposes as

$$\Delta V\big(\rho,\gamma,E(x)\big) = \underbrace{\Delta V_I(\rho,\gamma)}_{\text{self-discipline channel}} + \underbrace{\Delta V_A\big(\rho,\gamma,E(x)\big)}_{\text{aggregation channel}}, \tag{OA.1}$$

$$\Delta V_I(\rho,\gamma) = \begin{cases} 0 & \text{if } \rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)} \\ (1-\rho)H_{\underline{\pi}}(\gamma) & \text{if } \rho \ge \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}, \end{cases}$$

$$\Delta V_A\big(\rho,\gamma,E(x)\big) = \begin{cases} \psi^0\big(\mathrm{Med}(x)-E(x)\big) & \text{if } \rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)} \\ \big(\psi^0-\rho\psi_{s^d}(\rho)\big)\big(\mathrm{Med}(x)-E(x)\big) & \text{if } \rho \ge \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}. \end{cases}$$

The first component is the self-discipline channel characterized in Proposition 5. The second is an *aggregation channel* with no counterpart in a single-receiver model. When $E(x) < \mathrm{Med}(x)$, the median voter is more favorable to a^1 than the average voter, so the simple majority rule selects a^1 "too often" relative to the utilitarian optimum: persuading the median voter is welfare-harmful on average. Indeed, in the benchmark the influential signal induces the posterior mean exactly at γ, so the median voter gains nothing from being persuaded while the average voter strictly loses—the logic of Kamenica and Gentzkow (2011), whereby the sender extracts the entire surplus of the marginal receiver, aggregated here over a heterogeneous electorate. Fake news curbs this excess: by Lemma 9, the equilibrium persuasion probability $\rho\psi_{s^d}(\rho)$ falls short of its benchmark level ψ^0, and each unit of foregone persuasion saves the electorate $\mathrm{Med}(x)-E(x)$.

The next proposition shows that when the average voter is no more favorable to a^1 than the median voter, the aggregation channel reinforces the self-discipline channel and every conclusion of Proposition 5 extends to total welfare.

Proposition 6 (Total Welfare Gain with the Fake News)**.** *Suppose* $E(x) \leq \text{Med}(x)$*. In equilibrium, the ex-ante expected total welfare is weakly higher in the presence of fake news than in the benchmark scenario without it,* $\Delta V\big(\rho, \gamma, E(x)\big) \geq 0$*. Moreover,* $\Delta V\big(\rho, \gamma, E(x)\big)$ *is non-monotonic in the prominence of related news,* ρ*. In particular,*

1. *for any* $\rho \in [0,1]$, $\Delta V\big(\rho, \gamma, E(x)\big) \geq 0$.
2. *If* $\rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, $\frac{\partial \Delta V(\rho,\gamma,E(x))}{\partial \rho} = 0$
3. *If* $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, $\frac{\partial \Delta V(\rho,\gamma,E(x))}{\partial \rho} < 0$ *wherever the derivative exists.*

Proof. By OA.1 and Proposition 5, if $\rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, then $\Delta V = \psi^0\big(\text{Med}(x) - E(x)\big) \geq 0$ since $E(x) \leq \text{Med}(x)$; if $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, then $\Delta V_I = (1-\rho)H_{\underline{\pi}}(\gamma) \geq 0$ by Assumption 2 and $\Delta V_A \geq 0$ because $\rho\psi_{s^d}(\rho) \leq \psi^0$ by Lemma 9. This proves *1.* *2* follows since OA.1 is constant in ρ below the threshold. For *3*, wherever the derivative exists, Lemma 9 yields

$$\frac{\partial \Delta V}{\partial \rho} = -H_{\underline{\pi}}(\gamma) \; - \; \big(\text{Med}(x) - E(x)\big)\left(1 + \frac{H_{\overline{\pi}}(x^*)}{\gamma - x^*}\right) < 0,$$

since $H_{\underline{\pi}}(\gamma) > 0$ by Assumption 2, $\text{Med}(x) - E(x) \geq 0$, and $1 + \frac{H_{\overline{\pi}}(x^*)}{\gamma - x^*} > 0$. □

Note that if $\rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, then $\Delta V = \psi^0\big(\text{Med}(x) - E(x)\big) \geq 0$ even though dilution renders every news story uninfluential: the resulting informational collapse weakly beats the benchmark because it shuts down welfare-harmful persuasion entirely.

If $E(x) > \text{Med}(x)$—the average voter is *more* favorable to a^1 than the median voter—then the simple majority rule selects a^1 too *rarely* relative to the utilitarian optimum, persuasion is socially under-supplied, and the aggregation channel works *against* the self-discipline channel: fake news still disciplines RN's signal, but it also suppresses persuasion that society values. The next proposition characterizes exactly when each force dominates.

Proposition 7 (Boundary of the Welfare Gain)**.** *Suppose* $E(x) > \text{Med}(x)$.

1. *If* $\rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, $\Delta V\big(\rho, \gamma, E(x)\big) = \psi^0\big(\text{Med}(x) - E(x)\big) < 0$*: the presence of fake news strictly reduces welfare.*
2. *If* $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$*, then, wherever the derivative exists,*

$$\frac{\partial \Delta V\big(\rho, \gamma, E(x)\big)}{\partial \rho} > 0 \quad \Longleftrightarrow \quad E(x) - \text{Med}(x) \; > \; H_{\underline{\pi}}(\gamma)\left(1 + \frac{H_{\overline{\pi}}(x^*)}{\gamma - x^*}\right)^{-1},$$

where x^* *attains the supremum in condition (c) of Theorem 1. That is, within the persuasive regime, a more prominent fake news environment reduces welfare precisely when the mean–median gap of the electorate exceeds the stated threshold.*

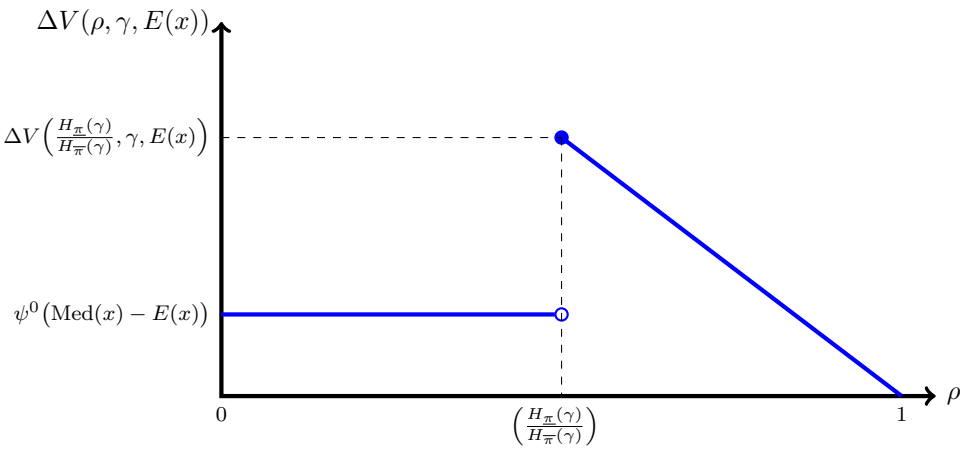

Figure OA1: Difference between the ex-ante expected total welfare: with FN v. without FN, drawn for $E(x) < \text{Med}(x)$ with a small mean–median gap. When $\rho < \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$, the gain is the constant $\psi^0\big(\text{Med}(x) - E(x)\big) \geq 0$; in the persuasive regime it is strictly decreasing (piecewise linear) in ρ and reaches 0 at $\rho = 1$. When $E(x) = \text{Med}(x)$, the left segment collapses to 0 and the right segment is exactly linear, $(1-\rho)H_{\underline{\pi}}(\gamma)$, recovering Figure 5.

Proof. *1* is immediate from OA.1. For *2*, note that $V\big(1, \gamma, E(x)\big)$ does not depend on ρ, so $\frac{\partial \Delta V}{\partial \rho} = \frac{\partial V}{\partial \rho}$. In the persuasive regime,

$$V\big(\rho, \gamma, E(x)\big) = (1-\rho)H_{\underline{\pi}}(\gamma) + \rho\psi_{s^d}(\rho)\big(E(x) - \text{Med}(x)\big),$$

and, wherever the derivative exists, Lemma 9 yields

$$\frac{\partial V}{\partial \rho} = -H_{\underline{\pi}}(\gamma) + \big(E(x) - \text{Med}(x)\big)\left(1 + \frac{H_{\overline{\pi}}(x^*)}{\gamma - x^*}\right),$$

which is strictly positive if and only if the stated inequality holds. □

Corollary 1 (Two States)**.** *Suppose* $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$ *and* $\Omega = \{\underline{\omega}, \overline{\omega}\}$*. Then* $x^* = \underline{\omega}$ *and* $H_{\overline{\pi}}(\underline{\omega}) = 0$*, so* $V\big(\rho, \gamma, E(x)\big)$ *is strictly increasing in* ρ *if and only if*

$$E(x) - \text{Med}(x) \;>\; H_{\underline{\pi}}(\gamma).$$

Proof. With $\Omega = \{\underline{\omega}, \overline{\omega}\}$, the only state below γ is $\underline{\omega}$ ($\overline{\omega} > \gamma$ holds since otherwise no belief could reach the threshold), so the maximization in the proof of Lemma 9 is over the singleton $\{\underline{\omega}\}$; hence $x^* = \underline{\omega}$ for every $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$. Moreover, $H_{\overline{\pi}}(\underline{\omega}) = \int_{-\infty}^{\underline{\omega}} G_{\overline{\pi}}(t)\, dt = 0$. Substituting into item 2 of Proposition 7, the threshold reduces to $H_{\underline{\pi}}(\gamma)$. □

Taken together, Proposition 5, Proposition 6, and Proposition 7 answer the natural

question of whether the welfare gain from fake news is a robust economic phenomenon or a knife-edge artifact of functional alignment. The self-discipline channel—priced at $(1-\rho)H_{\underline{\pi}}(\gamma)$—operates throughout the persuasive regime regardless of the shape of F. What the distribution of voters determines is the sign and size of the aggregation channel, and the total effect is governed by a single, transparent statistic of the electorate: the mean–median gap $E(x)-\text{Med}(x)$. Whenever the average voter is no more favorable to the persuader's cause than the median voter, fake news weakly raises total welfare, strictly so in the persuasive regime (Figure OA1); when the average voter is sufficiently more favorable than the median voter, the gain fails, and Proposition 7 states exactly where. The welfare paradox is thus robust within an explicitly characterized region of primitives, with a formal boundary rather than a caveat.

OA.1 Proof of Lemma 9

By Theorem 1, for $\rho \geq \frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}$ every equilibrium strategy π_r^* generates exactly one durable signal s^d, whose probability satisfies $\psi_{s^d} = 1 - H'_{\pi_r^*}(\gamma)$ (see the proof of Lemma 7). By condition (c) of Theorem 1, and recalling $S(x) = \frac{H_{\overline{\pi}}(x)-\rho^{-1}H_{\underline{\pi}}(\gamma)}{x-\gamma}$ from the proof of Lemma 8, the equilibrium slope at γ is the value function

$$H'_{\pi_r^*}(\gamma)(\rho) \;=\; \sup_{x\in[\underline{\omega},\gamma)} S(x) \;=\; \max_{x\in[\underline{\omega},\omega^\dagger]} \frac{\rho^{-1}H_{\underline{\pi}}(\gamma) - H_{\overline{\pi}}(x)}{\gamma - x}, \qquad \omega^\dagger \equiv \max\{\omega_i \in \Omega : \omega_i < \gamma\},$$

where we make the dependence on ρ explicit. As shown in the proof of Lemma 8, the supremum is attained at some $x^* \in [\underline{\omega}, \gamma)$; because $H_{\overline{\pi}}$ is piece-wise linear with kinks only at the finitely many states, a maximizer can always be taken in $[\underline{\omega}, \omega^\dagger]$, which justifies the second equality.

The objective is continuous in x on the compact set $[\underline{\omega}, \omega^\dagger]$, and for every x in that set,

$$\left|\frac{\partial}{\partial\rho}\left(\frac{\rho^{-1}H_{\underline{\pi}}(\gamma) - H_{\overline{\pi}}(x)}{\gamma - x}\right)\right| = \frac{H_{\underline{\pi}}(\gamma)}{\gamma - x}\,\rho^{-2} \;\leq\; \frac{H_{\underline{\pi}}(\gamma)}{\gamma - \omega^\dagger}\,\rho^{-2} \;\equiv\; \eta(\rho),$$

where η is integrable on any closed subinterval of $\left[\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}, 1\right]$. By the envelope theorem for arbitrary choice sets (Milgrom and Segal, 2002), wherever the value function is differentiable,

$$\frac{dH'_{\pi_r^*}(\gamma)(\rho)}{d\rho} = \frac{\partial}{\partial\rho}\left(\frac{\rho^{-1}H_{\underline{\pi}}(\gamma) - H_{\overline{\pi}}(x)}{\gamma - x}\right)\Bigg|_{x=x^*} = -\,\rho^{-2}\,\frac{H_{\underline{\pi}}(\gamma)}{\gamma - x^*} \;<\; 0.$$

Therefore,

$$\frac{d\big(\rho\psi_{s^d}(\rho)\big)}{d\rho} = \frac{d}{d\rho}\Big[\rho\Big(1 - H'_{\pi_r^*}(\gamma)(\rho)\Big)\Big] = 1 - H'_{\pi_r^*}(\gamma)(\rho) - \rho\,\frac{dH'_{\pi_r^*}(\gamma)(\rho)}{d\rho}$$
$$= 1 - \frac{\rho^{-1}H_{\underline{\pi}}(\gamma) - H_{\overline{\pi}}(x^*)}{\gamma - x^*} + \frac{\rho^{-1}H_{\underline{\pi}}(\gamma)}{\gamma - x^*} \;=\; 1 + \frac{H_{\overline{\pi}}(x^*)}{\gamma - x^*} \;>\; 0.$$

Since the maximizer x^* is piece-wise constant in ρ (it switches across the finitely many kinks of $H_{\overline{\pi}}$), the function $\rho\psi_{s^d}(\rho)$ is continuous and piece-wise linear in ρ, with strictly positive slope on each piece; hence it is strictly increasing on $\left[\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}, 1\right]$.

Finally, at $\rho = 1$, condition (b) of Theorem 1 reduces to $H_{\pi_r^*}(\gamma) = H_{\underline{\pi}}(\gamma)$, which is exactly the condition characterizing the benchmark optimal signal π_r^0; hence $H'_{\pi_r^*}(\gamma)(1) = H'_{\pi_r^0}(\gamma^-)$ and $1 \cdot \psi_{s^d}(1) = 1 - H'_{\pi_r^0}(\gamma^-) = \psi^0$. Monotonicity then delivers $\rho\psi_{s^d}(\rho) \le \psi^0$ for all $\rho \in \left[\frac{H_{\underline{\pi}}(\gamma)}{H_{\overline{\pi}}(\gamma)}, 1\right]$. □