WINPEC Working Paper Series No. E2531
February 2026

Fighting Fake News with Peer Feedback: Theory and Experiment

Yasushi Asako Yoshio Kamijo

Daiki Kishishita Masayuki Odora

Waseda INstitute of Political EConomy
Waseda University
Tokyo, Japan

Fighting Fake News with Peer Feedback: Theory and Experiment*

Yasushi Asako† Yoshio Kamijo‡ Daiki Kishishita§ Masayuki Odora¶

February 19, 2026

Abstract

The diffusion of fake news on social media poses a growing challenge to society. This study develops a theoretical model and laboratory experiment to examine whether user-to-user feedback such as likes or negative comments can reduce fake news sharing. We construct a sender-receiver game in which the sender receives a signal from potentially unreliable sources and decides whether to share it, while feedback from the receiver enables learning about information quality over time. The model predicts that (i) peer feedback induces self-selection: individuals with unreliable sources learn to stop sharing, and (ii) fake news spreads more when senders are motivated by reputation rather than accuracy, though this effect is modest. We test these predictions by developing a novel lab experiment based on the ball-and-urn experimental design. Consistent with the theory, feedback reduces fake news sharing, but effects are weaker due to underreaction in belief updating and noisy decisions. Differences across motivational incentives are minimal. These findings highlight both the potential and limits of peer feedback in preventing fake news sharing, offering implications for platforms seeking to curb misinformation through user-to-user feefback.

Keywords: Fake news; Communication; Social learning; Peer feedback; Social media

JEL classification: L82; D83; C92; D72

---

*We thank Yoichi Hizen, Greg Sheen, and participants at the 2025 Annual Meeting of the Japan Public Choice Society, the 2026 Winter Meeting of the Japan Society for Quantitative Political Science, and the International Workshop on Polarization, Disinformation, and Democratic Backsliding in Asia and Beyond as well as seminar participants at Kobe University and Okayama University for their helpful comments. This study was financially supported by JSPS Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research (Grant Number JPJS00123811919).

†Faculty of Political Science and Economics, Waseda University. 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo, Japan. 169-8050. E-mail: yasushi.asako@waseda.jp

‡Faculty of Political Science and Economics, Waseda University. 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo, Japan. 169-8050. E-mail: yoshio.kamijo@gmail.com

§Graduate School of Economics, Hitotsubashi University. 2-1 Naka, Kunitachi, Tokyo, Japan. 186-8601. E-mail: daiki.kishishita@gmail.com

¶Global Education Center, Waseda University, 1-104 Totsukamachi, Shinjuku-ku, Tokyo 169-8050, Japan. Email:odora_masa@fuji.waseda.jp

1 Introduction

The spread of fake news poses a growing threat to societies (Jerit and Zhao, 2020; Nyhan, 2020), particularly through social media platforms such as X and Facebook. According to the Pew Research Center, 23% of the U.S. citizens reported sharing a made-up news story on social media either knowingly or not in 2016 (Barthel, Mitchell and Holcomb, 2016).[1] The social costs of misinformation have made uncovering the mechanisms that suppress its diffusion a central challenge for policymakers and platform designers. This study contributes to this effort by examining whether user-to-user feedback on social media can independently curb the spread of fake news, without requiring platform-level interventions.

A central puzzle is that many individuals who share fake news are not motivated by malice. Instead, they are often genuinely unaware of the unreliability of their own information sources. Individuals who believe or share misinformation tend to systematically overestimate the accuracy of their knowledge, leading them to wrongly believe they can easily distinguish real from fabricated content (e.g., Pennycook and Rand, 2020; Lyons et al., 2021).[2] Media literacy interventions attempt to correct these inflated beliefs by encouraging individuals to verify source credibility, and empirical evidence shows that such interventions can reduce fake news sharing (e.g., Guess et al., 2020; Hameleers, 2022). Yet such interventions require sustained, external, and costly effort, limiting their scalability in large and decentralized information environments.

In contrast, social media platforms contain built-in mechanisms that may naturally correct these inflated beliefs. Users routinely receive likes or negative reactions when they post content, and they observe others' posts even when they do not participate. These interactions provide cues about content accuracy (Eckles, Kizilcec and Bakshy, 2016; Bode and Vraga, 2018; Badrinathan and Chauchard, 2024). If negative feedback—or disagreement with others' posts—leads users to reassess the reliability of their information sources as well as content accuracy, then peer feedback could serve as a self-correcting mechanism for the information environment, operating without external regulation or algorithmic interventions.[3]

However, despite a growing theoretical literature on fake news sharing (Papanastasiou, 2020; Cisternas and Vásquez, 2023; Acemoglu, Ozdaglar and Siderius, 2024; Danenberg and Fudenberg, 2024; Sisak and Denter, 2024; Kranton and McAdams, 2024), we still know little about whether peer feedback can generate dynamic learning about the reliability of one's information sources, and whether such a mechanism is strong enough to curb misinformation.

[1] A similar pattern is observed in other countries: for example, in Japan, approximately 25% of individuals exposed to misinformation reported sharing it through some channel. Notably, about 44% of them shared it online (Ministry of Internal Affairs and Communications , Japan).

[2] Assenza, Cardaci and Huber (2024) show that correcting overconfidence increases demand for fact-checking.

[3] Although not directly about fake news sharing, Deolankar, Fong and Sriram (2025) analyze Reddit data and find that negative feedback leads users to moderate the intensity of their tone in subsequent posts. This finding is consistent with our conjecture. The encouragement effect of positive feedback is also reported. For example, Burtch et al. (2022) find that peer awards induce recipients to make more frequent posts in Reddit. It is also shown that observing negative comments on fake news increases the propensity to engage in correction behaviors (Meiske et al., 2024).

Most existing models analyze incentives to share but do not address how individuals learn that they themselves may be connected to unreliable sources. This paper fills this gap by developing a formal model of dynamic learning from peer feedback and testing its implications through a controlled laboratory experiment.

To this end, we develop a sender-receiver game. In each period, a sender observes an imperfect signal (i.e., a piece of news) about a time-variant unknown state, and decides whether to share it. The receiver independently receives another imperfect signal. If the sender shares the signal—analogous to posting on social media—the receiver responds with either "like" or "dislike," where "dislike" represents negative feedback. Importantly, if the sender does not share, the receiver instead posts their own signal publicly, which serves as indirect feedback from other users. Thus, regardless of the sender's choice, the sender always observes whether the receiver's public signal agrees or disagrees with the sender's private signal. This feature mirrors the visibility of others' content on social media platforms.

A critical feature is that the sender may be exposed to unreliable information sources. The sender can be either high type or low type: low types are more likely to observe a signal negatively correlated with the true state (fake news), while high types obtain more accurate signals. Crucially, senders do not know their own type, reflecting empirical findings that individuals are often unaware of source reliability. The communication game is repeated over time, allowing senders to gradually learn about their type from accumulated feedback.

The model yields three central predictions. First, the sender chooses to share their signal if and only if the number of positive peer feedback responses minus the number of negative ones, denoted by Δ, exceeds a certain threshold. Note that Δ incorporates both direct and indirect feedback.

Second, low-type senders are more likely to receive negative feedback, while high-type senders are more likely to receive positive feedback. Through repeated interactions, low-type senders come to realize that their information sources are likely unreliable and thus cease sharing news. In contrast, high-type senders learn that their sources are more reliable and continue to share information. In this way, feedback from others helps to curb the spread of fake news by inducing self-selection based on source reliability.

Third, empirical research suggests that prompting individuals to consider accuracy reduces the likelihood of sharing fake news (Pennycook et al., 2021). Motivated by this, we compare two types of motivations in the sender's payoff structure: *reputational motivation*, where the payoff depends on peer feedback, and *accuracy motivation*, where the payoff depends on the truthfulness of the shared signal. We find that the model's core predictions hold under both motivational structures. Moreover, the sharing threshold on Δ is slightly higher under accuracy motivation than under reputational motivation, implying that accuracy-focused incentives may more effectively deter fake news sharing. However, the magnitude of this effect is modest, suggesting that the presence of reputational motives is not the primary cause of fake news sharing.

To test these predictions, we conducted a laboratory experiment at Waseda University,

Tokyo, Japan in 2025, using a ball-and-urn framework to implement the sender–receiver game. Participants were randomly matched into sender–receiver pairs in each round for 10 or 20 rounds, with roles and types fixed. High-type senders received signals generated from the true urn, while low-type senders drew from the incorrect urn. The design cleanly implements differences in source reliability. We also elicited senders' beliefs using the BDM mechanism (Becker, DeGroot and Marschak, 1964).

In real-world settings, motivations for sharing fake news are shaped by multiple factors. One such factor is politically motivated reasoning—the psychological tendency to accept information that aligns with one's ideological or partisan beliefs (Jerit and Zhao, 2020). For example, using Twitter data from the United States, Osmundsen et al. (2021) document that individuals who express greater hostility toward political opponents are the most likely to share political misinformation. Our experiment deliberately removes political content in order to isolate the role of informational reliability. By abstracting from political considerations, we provide a clean empirical test of the dynamic learning mechanism relative to a theoretical benchmark. If dynamic learning mechanism does not induce self-selection in this politically neutral environment, this suggests that politically motivated reasoning alone cannot account for the diffusion of misinformation observed in real-world settings.

The results are summarized as follows. First, regardless of the payoff structure, the likelihood of sharing fake news increases with the net number of positive peer feedback responses, denoted by Δ (i.e., the number of positive minus negative responses). However, in contrast to the theoretical prediction, we do not observe a clear threshold behavior. Even when Δ falls below the predicted threshold, more than 25% of participants continue to share the signal. Conversely, when Δ exceeds the threshold but remains at a moderate level, only 50–75% of participants choose to share. This suggests a more gradual behavioral response to peer feedback than theoretically predicted.

Second, feedback induces self-selection: after ten rounds, approximately 75% of high-type senders share, compared to only 40–50% of low types. However, self-selection is weaker than predicted. Most importantly, not only low types over-share but also high types under-share relative to theory. These over-sharing and under-sharing undermine the quality of social media platforms.

Third, sharing is slightly higher under reputational motivation than accuracy motivation when Δ is small, though not statistically significant. At the aggregate level, differences vanish, consistent with the model's prediction that motivation structure has only modest effects.

To understand the quantitative discrepancies between theory and experiment, we examine belief updating. Beliefs generally respond in the predicted direction—positive feedback increases the belief of being high type—but the magnitude is more conservative than Bayesian updating (Phillips and Edwards, 1966). Underreaction is especially strong when senders do not share and subsequently see that the receiver's post matches their signal (i.e., positive but indirect feedback).[4] The literature has pointed out negativity bias: bad information is processed

[4]Two types of positive feedback exist: a "like" when the sender shares (direct feedback), and a "match" when

more thoroughly than good (Baumeister et al., 2001).[5] Our results show that such negative bias appears for indirect feedback rather than direct feedback.

Overall, our findings show both the potential and the limits of peer feedback as a mechanism to curb misinformation. While feedback promotes self-selection and reduces fake news sharing, behavioral frictions limit its effectiveness. These results have several implications for platform design and policy interventions.

First, although the diffusion of fake news is often attributed primarily to sharing by users connected to unreliable sources, our findings indicate that users connected to reliable sources are reluctant to share news, which ultimately undermines the informational quality of social media platforms. This suggests that effective interventions should not only curb misinformation from unreliable sources but also actively encourage sharing by users with reliable sources.

Second, while prior literature suggests that prompting individuals to consider accuracy reduces the likelihood of sharing fake news, our results show that this effect is minimal when receivers are reliable. This implies that accuracy prompts alone are unlikely to substantially curb misinformation and that policies focusing solely on reputational incentives may be insufficient.

Third, both direct feedback (i.e., likes or dislikes) and indirect feedback (i.e., others' posts) play critical roles in shaping sharing behavior. Our results further show that responses to positive indirect feedback are particularly weak, highlighting the potential value of platform designs that make indirect feedback more salient alongside direct feedback.

The remainder of the paper is organized as follows. Below, we summarize the related literature. Section 2 develops a formal model, and Section 3 characterizes the equilibrium of the game and derives theoretical predictions. Section 4 explains the details of our experimental design. Section 5 presents experimental results. Section 6 concludes.

Related literature: The propagation of fake news on social media platforms has drawn growing attention not only in economics (Aridor et al., 2024) but also in psychology and communication studies (Pennycook and Rand, 2021; Van Der Linden, 2022). This study contributes to this interdisciplinary literature by highlighting both the potentials and limits of peer feedback as a mechanism to curb misinformation. We combine a formal theoretical model with a controlled laboratory experiment to isolate how feedback between users shapes the dynamics of fake news sharing.

Specifically, our work connects three strands of research: (i) incentivized experiments examining fake news within strategic communication games, (ii) theoretical models analyzing the incentives underlying misinformation sharing, and (iii) incentivized experiments investigating belief updating about one's ability under feedback.

---

the sender does not share and the receiver's public signal aligns (indirect feedback). Theoretically both have equal informational content given equilibrium strategies, but participants treat them differently. In a sequential social learning model, Çelen, Kariv and Schotter (2010) find that participants behave differently depending on whether they observe their predecessor's action or advice, even though they are informationally equivalent. Although the context is different, our finding echoes with theirs.

[5]To be precise, another strand of literature has found that good news about one's ability is processed more due to self-serving bias (Möbius et al., 2022).

First, our work relates to experimental studies examining the spread of fake news in strategic communication environments (Serra-Garcia and Gneezy, 2021; Thaler, 2021, 2024; Feess, Jost and Ressi, 2024). These studies implement incentivized experiments to understand how individuals evaluate and transmit misinformation. For example, Serra-Garcia and Gneezy (2021) design an experiment in which participants assess whether news is fake and decide whether to share it. They find that participants are overconfident in detecting fake news and are more likely to share false rather than true news. Moreover, receivers tend to trust shared news more than unshared news. Our study contributes to this literature by showing that individuals can learn about the reliability of their information sources through feedback from others. Unlike existing work, our experiment explicitly incorporates an endogenous learning process in a dynamic setting, allowing us to study how feedback shapes fake news sharing over time.

Second, several studies have developed theoretical models analyzing the incentives behind fake news sharing on social media platforms (Papanastasiou, 2020; Acemoglu, Ozdaglar and Siderius, 2024; Cisternas and Vásquez, 2023; Danenberg and Fudenberg, 2024; Sisak and Denter, 2024; Kranton and McAdams, 2024). For instance, Acemoglu, Ozdaglar and Siderius (2024) analyze the case where users have reputational motivation and show that the presence of echo chambers induces fake news sharing as users anticipate positive feedback from like-minded individuals. Our theoretical model enriches this literature by incorporating feedback-based learning, which allows users to update their beliefs about the reliability of their information sources. Moreover, by comparing reputational and accuracy motivations, we highlight how the incentive structure shapes fake news sharing behavior.[6]

Third, our study contributes to the experimental literature on belief updating about one's ability under feedback (Zimmermann, 2020; Möbius et al., 2022; Oprea and Yuksel, 2022; Drobner and Goerg, 2024; Bolte and Fan, 2024; Hagenbach and Saucet, 2025). They examine whether people update belief about their ability such as IQ or task performance in a Bayesian rational way. Our experiment can be viewed as an extension to this setting, where individuals learn about their ability to access accurate information. While most of this literature relies on computer-generated feedback, our design features feedback from another strategic player, the receiver. Notably, in the reputational motivation condition of our experiment, the sender's payoff depends directly on the receiver's response, introducing a richer social learning environment. Although Hagenbach and Saucet (2025) study learning through interactions with others,[7] their setting differs in important ways. In particular, while they consider situations where the sender's payoff depends on the receiver's guess like ours, their senders convey beliefs about the receiver's ability. In contrast, our setting allows the sender to infer their own ability (i.e., their access to accurate information) based on the receiver's reaction.

[6]Sisak and Denter (2024) examine how two distinct types of reputation concerns—desiring to be perceived as talented versus signaling one's worldview—generate different patterns of fake news sharing.

[7]In addition, Oprea and Yuksel (2022) conduct an experiment on learning about an IQ score, where each participant can observe the beliefs of a counterpart who performed similarly on the test.

2 Model

We begin by developing a simple model that allows us to analyze the role of user-to-user feedback in the sharing of fake news on social media platforms. To this end, we consider a two-player communication game between a sender and a receiver that captures key features of communication on social media.

We first present a static model. Subsequently, we extend the model to a dynamic one. The state of the world is given by $\omega \in \{0, 1\}$, where $\Pr(\omega = 0) = \frac{1}{2}$. At the beginning of the game, the sender observes an imperfect signal of ω, denoted by $\sigma^S \in \{0, 1\}$. This can be interpreted as observing a news article on the internet. After observing it, the sender decides whether to share the received signal σ^S on a social media platform or not (i.e., the sender's message is given by $m \in \{\sigma^S, \emptyset\}$, where $\emptyset$ indicates "not sharing").[8] Observing the message, the receiver decides how to respond to the message ($r \in \{l, d\}$), where l represents "like" and d represents "dislike." "Dislike" corresponds to posting a negative comment on the sender's post.[9] Below, we present the details of the model.

Sender: The sender's type is $\theta \in \{h, l\}$, on which the accuracy of the signal depends: for each $k \in \{0, 1\}$,

$$\Pr(\sigma^S = k | \omega = k) = \begin{cases} p \in (\frac{1}{2}, 1) & \text{if } \theta = h \\ 1 - p & \text{if } \theta = l \end{cases}.$$

Whereas the type-h sender is likely to observe the precise information, the type-l sender is likely to observe the incorrect information. Therefore, type h's information sources are reliable, whereas type l's information sources are unreliable (Acemoglu, Ozdaglar and Siderius, 2024). In this sense, type l is susceptible to fake news. The prior probability of θ being h is $q \in (0, 1)$. We assume that neither the sender nor the receiver knows the sender's type.

Regarding the sender's payoff, we consider the following two types. The first type of payoff is based on *reputational motivation*, where the payoff is given by

$$u(m, r) = \begin{cases} 1 & \text{if } m = \sigma^S, r = l \\ -\alpha & \text{if } m = \sigma^S, r = d \\ 0 & \text{if } m = \emptyset \end{cases}.$$

$\alpha \in (0, 1]$ represents the psychological/reputation cost that the receiver posts "dislike" on the sender's message (Acemoglu, Ozdaglar and Siderius, 2024).[10]

[8] The sender is prohibited from sending a message different from the received signal, which is typically assumed in the model of news sharing on social media platforms (e.g., Acemoglu, Ozdaglar and Siderius, 2024) In this sense, the signal is verifiable information; thus, the game is modeled as a disclosure game (Milgrom, 1981). An interpretation is that the sender observes a news article and decides whether to share it or not.

[9] On some platforms, such as YouTube and Reddit, there is a dislike or downvote button.

[10] In practice, posting a "dislike" may be more costly for receivers because expressing disapproval can involve psychological costs, and many platforms lack a dislike button, requiring users to post a comment instead. Suppose

Conversely, the second type of payoff is based on *accuracy motivation* (Papanastasiou, 2020), where the payoff is given by

$$u(m,\omega) = \begin{cases} 1 & \text{if } m = \omega \\ -\alpha & \text{if } m \neq \omega \\ 0 & \text{if } m = \emptyset \end{cases}.$$

For simplicity, we assume that the sender sends $m = \sigma^S$ in an indifferent case. In practice, both types of payoffs are reasonable. Therefore, examining both cases is useful in ensuring the robustness of the theoretical predictions.

Receiver: The receiver observes a signal $\sigma^R \in \{0,1\}$ on the value of ω. It is common knowledge that $\Pr(\sigma^R = k|\omega = k) = p$ for each $k \in \{0,1\}$ (i.e., the receiver is known to be type h). This can be interpreted as a situation where there are a number of receivers, and due to the wisdom of crowds, the receivers are type h as a collective actor, whereas each individual receiver could be type l.[11]

Another possible interpretation that the receiver is of type h is that the receiver acts as a fact-checker who indicates whether the sender's message is correct or incorrect, albeit with a small probability of error. Indeed, various fact-checking organizations provide truthfulness ratings, and empirical evidence shows that being identified as truthful affects sharing behavior (Li and Chang, 2023).

After observing σ^R and m, the receiver decides what response to the sender's message to post when $m \neq \emptyset$ ($r \in \{l,d\}$), and decides to what message to post when $m = \emptyset$ ($r' \in \{0,1\}$). Thus, irrespective of whether the sender posts a message, the sender can observe the receiver's opinion. In real-world social media environments, even users who do not actively post can observe others' opinions. According to the Pew Research Center, approximately 49% of Twitter users in the US are "lurkers" who tweet only infrequently and primarily use the platform to observe what others are saying (Odabaş, 2022). This setting allows us to incorporate such a feature.[12]

The receiver's payoff when $m \neq \emptyset$ is[13]

$$v(m,\omega,r) = \begin{cases} 1 & \text{if } m = \omega \ \&\ r = l \text{ or } m \neq \omega \ \&\ r = d \\ 0 & \text{otherwise} \end{cases}.$$

that receivers express "dislike" toward seemingly fake news only with probability α, and that doing so imposes a cost of -1 on the sender. This alternative setup yields the same expected utility function for the sender.

[11] Indeed, in the aggregate level, people tend to distinguish between true and fake news (Pfänder and Altay, 2025).

[12] If the sender cannot observe the receiver's opinion when $m = \emptyset$, the incentive for experimentation arises in the dynamic model. That is, the sender would post a message for an initial period to obtain information about whether the sender is of high type or not. This setting rules out such a possibility.

[13] Alternatively, we can add an option of not responding to the sender's message ($\emptyset$) to the action space of the receiver, which makes the setting consistent with the sender's setting.

That is, the sender wants to post "like" (resp. "dislike") when the sender's post coincides with (resp. differs from) the state of the world.

On the other hand, the receiver's payoff when $m = \emptyset$ is given by

$$v(m, \omega, r') = \begin{cases} 1 & \text{if } r' = \omega \\ 0 & \text{otherwise} \end{cases}.$$

That is, when the sender does not post anything, the receiver simply tries to post a message that is likely to be matched with the state of the world.

Timing of the game and equilibrium concept: The timing of the game is summarized as follows:

1. Nature chooses $\theta \in \{h, l\}$ and $\omega \in \{0, 1\}$.
2. The sender receives a signal $\sigma^S \in \{0, 1\}$ and decides whether to share the signal or not: $m \in \{\sigma^S, \emptyset\}$.
3. Observing a private signal $\sigma^R \in \{0, 1\}$ and m, the receiver sends a message ($r \in \{l, d\}$ if $m \neq \emptyset$ and $r' \in \{0, 1\}$ if $m = \emptyset$).

We characterize the (pure-strategy) perfect Bayesian equilibrium of this game.

Dynamic model: Later, we extend this static model to a dynamic model ($t = 1, 2, ..., T$). In the dynamic model, the sender's type is time-invariant, whereas the state of the world in each period, ω_t, is independently drawn. Furthermore, at each period, the sender is randomly matched with a new receiver.[14]

3 Equilibrium and Theoretical Predictions

In this section, we characterize the equilibrium of both the static and dynamic games, and subsequently derive the set of theoretical predictions to be tested in the laboratory experiment.

3.1 Equilibrium Characterization

Static equilibrium: We start with analyzing the static model. First, the receiver's equilibrium strategy is derived as follows. The omitted proofs are contained in the Appendix.

Lemma 1. (Receiver's strategy).

(a). *Suppose that $m \neq \emptyset$. Then, $r = l$ if and only if $m = \sigma^R$.*

[14] Whether the receiver can observe the sender's previous activities or responses by the previous receivers does not matter in the equilibrium characterization.

(b). *Suppose that $m = \emptyset$. Then, $r' = 0$ if and only if $\sigma^R = 0$.*

Therefore, the receiver posts "like" if and only if the signal shared by the sender coincides with his or her private signal, σ^R. This is because the receiver's private signal is more likely to be precise than the sender's private signal in that the sender could be type l, but the receiver is known to be type h. Furthermore, when the sender does not share anything, the receiver posts a message that coincides with his or her private signal because the signal is informative about the state of the world.

Having this result in hand, the sender's equilibrium strategy is derived as follows. Note that the sender's strategy depends on whether the sender has reputational motivation or accuracy motivation. Therefore, we present the equilibrium strategy for each type of the payoff structure.

Proposition 1. (Static equilibrium).

Suppose that the sender subjectively believes that she or he is type h with probability $\tilde{q}$.

(a). *Assume that the sender has reputational motivation. Then, the sender chooses $m = \sigma^S$ if and only if*

$$\tilde{q} \geq \bar{q}_R := \frac{1}{(2p-1)^2}\left[\frac{\alpha}{1+\alpha} - 2p(1-p)\right].$$

Otherwise, the sender chooses $m = \emptyset$.

(b). *Assume that the sender has accuracy motivation. Then, the sender chooses $m = \sigma^S$ if and only if*

$$\tilde{q} \geq \bar{q}_A := \frac{p\alpha - (1-p)}{(2p-1)(1+\alpha)}.$$

Otherwise, the sender chooses $m = \emptyset$.

(c). *$\bar{q}_A \geq \bar{q}_R$, where the equality holds when $\alpha = 1$. Therefore, the sender with reputational motive is more likely to share σ^S than the sender with accuracy motive.*

Irrespective of whether the sender is motivated by reputation or accuracy, the equilibrium is characterized by a threshold strategy: the sender shares the received signal if and only if their subjective probability of being type h, $\tilde{q}$, exceeds a certain threshold. When the sender is more likely to be type l, their private signal is likely to be incorrect. When the sender has accuracy motivation, this causes hesitation in posting the signal by definition. Furthermore, when the sender has reputational motivation, this magnifies the concern that the receiver "dislikes" the message, thus causing hesitation in posting the signal. In either case, the sender posts the signal if and only if $\tilde{q}$ is sufficiently large (i.e., the sender's information sources are reliable with a sufficiently large probability). This is (a) and (b) in the above proposition.

Importantly, the threshold value differs based on the sender's payoff structure. Specifically, (c) in the proposition shows $\bar{q}_A \geq \bar{q}_R$, meaning that a sender with reputational motivation is more likely to share the received signal than one with accuracy motivation.

The intuition is as follows. Because $\alpha < 1$ is assumed, $\bar{q}_R$ and $\bar{q}_A$ are less than a half. This means that the sender with $\bar{q}_R$ or $\bar{q}_A$ perceives that his or her private signal is likely to

be wrong. Therefore, for the sender with $\bar{q}_R$ or $\bar{q}_A$, what matters is the case of sharing the incorrect signal rather than the case of sharing the correct signal. Even when sharing the incorrect signal, the cost of sharing it might be small for the sender with reputation motivation because the receiver may also have received the incorrect signal, and the sender might avoid receiving "dislike" feedback. Conversely, with accuracy motivation, sharing the incorrect signal always incurs a cost α. As a result, the cost of sharing the incorrect signal is lower for senders with reputational motivation than for those with accuracy motivation. This explains why a sender motivated by reputation is more likely to share the received signal compared to one motivated by accuracy. Furthermore, this finding aligns with an empirical observation that shifting attention to accuracy enhances the quality of news that people subsequently share (Pennycook et al., 2021).

Learning by the sender: If $\bar{q}_R, \bar{q}_A \notin [0,1]$, this would imply that senders either never share news or always share news, regardless of their belief about being type-h, which is unrealistic. Furthermore, if $q > \bar{q}_R, \bar{q}_A$ does not hold, it implies that no sender shares news at the beginning of the game. Therefore, to focus on meaningful cases, we assume throughout that $\bar{q}_R, \bar{q}_A \in [0,1]$ and $q > \bar{q}_R, \bar{q}_A$.

Assumption 1. (Interior thresholds).
$\bar{q}_R, \bar{q}_A \in [0,1]$ *and* $q > \bar{q}_R, \bar{q}_A$ *hold.*

We extend the above static model to a dynamic model ($t = 1, 2, ..., T$). Let q_t be the subjective probability of being type h for the sender at the beginning of period t and $q_{t+1}(r_t, q_t)$ be the sender's subjective probability of being type h at the beginning of period $t+1$ given that the receiver's response in period t was r_t. Note that $q_1 = q$ holds by definition. Therefore, $q_{t+1}(l, q_t)$ is the updated belief when the receiver posted "like" on the signal shared by the sender and $q_{t+1}(d, q_t)$ is the updated belief when the receiver posted "dislike" on the signal shared by the sender. Similarly, we define $q_{t+1}(r'_t, q_t)$. Note that the sender' learning does not depend on the sender's payoff structure.

When $m_t = \sigma_t^S$, the sender updates the belief to

$$q_{t+1}(r_t = l, q_t) = \frac{q_t[p^2 + (1-p)^2]}{q_t[p^2 + (1-p)^2] + (1-q_t)2p(1-p)};$$

$$q_{t+1}(r_t = d, q_t) = \frac{q_t 2p(1-p)}{q_t 2p(1-p) + (1-q_t)[p^2 + (1-p)^2]}.$$

Therefore,[15]

$$q_{t+1}(l, q_t) > q_t > q_{t+1}(d, q_t).$$

[15]Here,

$$p^2 + (1-p)^2 - 2p(1-p) = 4\left(p - \frac{1}{2}\right)^2 > 0.$$

Furthermore, when the sender does not share anything (i.e., $m_t = \emptyset$), the sender updates the belief to

$$q_{t+1}(r'_t = \sigma_t^S, q_t) = \frac{q_t[p^2 + (1-p)^2]}{q_t[p^2 + (1-p)^2] + (1-q_t)2p(1-p)};$$
$$q_{t+1}(r'_t \neq \sigma_t^S, q_t) = \frac{q_t 2p(1-p)}{q_t 2p(1-p) + (1-q_t)[p^2 + (1-p)^2]}.$$

Therefore, $q_{t+1}(r'_t = \sigma_t^S, q_t) = q_{t+1}(r_t = l, q_t)$ and $q_{t+1}(r'_t \neq \sigma_t^S, q_t) = q_{t+1}(r_t = d, q_t)$. We will refer to $r'_t = \sigma_t^S$ as *match* and $r'_t \neq \sigma_t^S$ as *mismatch*. The above equalities indicate that "like" and "match" (resp. "dislike" and "mimsatch") are informationally equivalent.

Because the belief updating is invariant of the choice of whether to share the received signal or not, the sender's dynamic incentives for experimentation do not exist.

Consequently, we have

$$q_{t+1}(l, q_t) = q_{t+1}(r'_t = \sigma_t^S, q_t) > q_t > q_{t+1}(d, q_t) = q_{t+1}(r'_t \neq \sigma_t^S, q_t).$$

Therefore, when the sender received "dislike" or observed the receiver's message contradicting with the sender's private signal, the sender downwardly updates the belief about own type; that is, the sender becomes less confident on the reliability of his or her information sources.

Dynamic equilibrium: Based on this learning effect, we derive the characterization of the dynamic equilibrium. For this, we introduce a notation on the *net number of positive feedback*: for $t \geq 2$, let

$$\Delta(t) := \underbrace{\sum_{\tau=1}^{t-1} \mathbf{1}_\tau\{r_\tau = l \text{ or } r'_\tau = \sigma_\tau^S\}}_{=:P(t)} - \underbrace{\sum_{\tau=1}^{t-1} \mathbf{1}_\tau\{r_\tau = d \text{ or } r'_\tau \neq \sigma_\tau^S\}}_{=:N(t)}, \tag{1}$$

where $P(t)$ is the number of periods until the beginning of period t such that the sender receives positive feedback, and $N(t)$ is the number of periods until the beginning of period t such that the sender receives negative feedback. For example, suppose that the receiver posted "dislike" in periods 1 and 2, whereas the receiver posted "like" in period 3. Then, $\Delta(4) = 1 - 2 = -1$. As we observed before, "like" and "match" (resp. "dislike" and "mismatch") are informationally equivalent. Therefore, q_t only depends on $\Delta(t)$.

Having this notation in hand, we obtain the following characterization of the equilibrium.

Proposition 2. (Dynamic equilibrium).

(a). *In period t, the sender chooses $m = \sigma_t^S$ if and only if $q_t \geq \bar{q}_R$ (resp. $q_t \geq \bar{q}_A$) when the sender has reputational motivation (resp. accuracy motivation). Otherwise, the sender chooses $m = \emptyset$.*

(b). *In the case of reputational motivation (resp. accuracy motivation), there exists* $\underline{\Delta}_R$ *(resp.* $\underline{\Delta}_A$*) such that the sender chooses* $m_t = \sigma_t^S$ *if and only if* $\Delta(t) \geq \underline{\Delta}_R$ *(resp.* $\Delta(t) \geq \underline{\Delta}_A$*). Furthermore,* $\underline{\Delta}_R \leq \underline{\Delta}_A$*, with equality holding when* $\alpha = 1$*.*

(c). *Regardless of the sender's payoff structure,*

$$\lim_{T\to\infty} \Pr(m_T = \sigma_T^S \mid \text{sender = type } h) = 1; \quad \lim_{T\to\infty} \Pr(m_T = \sigma_T^S \mid \text{sender = type } l) = 0.$$

First, (a) of the proposition indicates that the equilibrium in each period is reduced to the static equilibrium given in Proposition 1.

Second, combining this with the updated beliefs derived above, we obtain (b) of the proposition. It indicates that the sender shares the signal if and only if $\Delta(t)$ exceeds a certain threshold. This is straightforward because the sender becomes more confident on the reliability of his or her information sources as the number of positive feedback increases. Furthermore, the threshold value in the case of reputational motivation is small than or equal to that in the case of accuracy motivation. That is, the sender with reputational motivation is more likely to share news than the sender with accuracy motivation. This is a direct consequence of Proposition 1 (c).

Lastly, (c) of the proposition indicates that feedback from the receiver enables self-selection based on the reliability of information sources. Type-l senders are likely to receive a wrong signal; thus, they are likely to receive negative feedback; thus, in the long run, $\Delta < \underline{\Delta}$ holds and type-l senders stop sharing news for sure. On the other hand, type-h senders are likely to receive a correct signal; thus, they are likely to receive positive feedback; thus, in the long run, $\Delta \geq \underline{\Delta}$ holds and type-h senders share news for sure.

3.2 Theoretical Predictions

Given the above characterization of the equilibrium, we derive theoretical predictions that will be tested through a laboratory experiment. In the experiment, we will use a specific set of parameter values: $p = 0.8$, $\alpha = 0.6$, and $q = 0.6$. Therefore, we present the theoretical predictions based on this set of parameter values.

Under this parameter value, $\bar{q}_R \simeq 0.153$ and $\bar{q}_A \simeq 0.292$. Consequently, $\underline{\Delta}_R = -2$ and $\underline{\Delta}_A = -1$. Therefore, we obtain the following prediction about sharing behavior by combining the particular parameter values with Proposition 2 (b):

Prediction 1. (Peer feedback and sharing behaviors).

Regardless of the sender's payoff structure, the sender is more likely to share the signal as the net number of positive feedback, $\Delta(t)$*, increases. Specifically, sharing behavior is characterized by the threshold property. In the case of reputational motivation, the sender shares the signal if and only if* $\Delta(t) \geq -2$*. In the case of accuracy motivation, the sender shares the signal if and only if* $\Delta(t) \geq -1$*.*

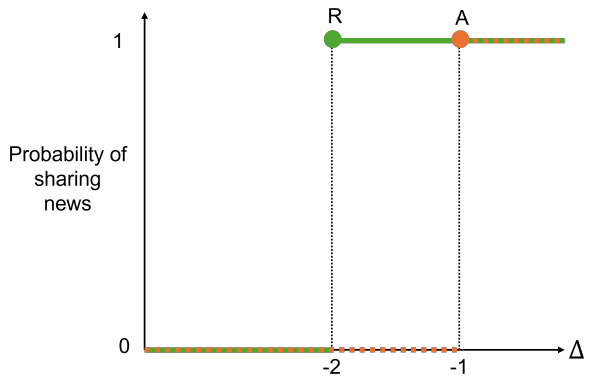

Figure 1: Equilibrium Sharing Behavior

This property is also visualized in Figure 1, where the threshold property certainly holds.

Another important prediction is the self-selection due to peer feedback, highlighted in Proposition 2 (c). Though the proposition derives only the asymptotic property where $T \to \infty$, similar properties can be extended to finite cases. Figure 2 reports the expected probability of sharing the signal among type-h senders and among type-l senders, respectively.[16] While it is not necessarily monotonic, as time goes on, self-selection gradually proceeds compared with the initial period. Therefore, we obtain the following prediction:

Prediction 2. (Self-selection).

Regardless of the sender's payoff structure, self-selection proceeds as time goes on.

This theoretical prediction implies that even without external interventions, such as algorithmic changes by platforms, user-to-user feedback spontaneously prevents the spread of fake news.

The last theoretical prediction is regarding the difference in the sender's payoff structure. As shown in Proposition 2 (c), $\underline{\Delta}_R < \underline{\Delta}_A$, meaning that the sender is more likely to share the signal when motivated by reputation rather than accuracy. This leads to the following prediction:

Prediction 3. (Effect of motivation).

The sender is more likely to send the signal when the sender is motivated by reputation than by accuracy. Specifically, when $\Delta(t) = -2$, the sender shares the signal under reputational motivation but does not do so under accuracy motivation. For any other $\Delta(t) \neq -2$, the sender behaves in the same way regardless of the payoff structure.

An important caveat is that the effect of motivation is only modest. Except for the case

[16] Let J_θ be the probability that type-θ sender observes like or match in each period. Then, this can be regarded as a Bernoulli trial with a success probability of J_θ. Therefore, the probability that $\Delta(t) = \delta$ is given by the probability that the number of success, $P(t)$, is $\frac{t-1+\delta}{2}$ because $\Delta(t) = P(t) - N(t) = P(t) - [t - 1 - P(t)] = 2P(t) - (t - 1)$. This implies that the probability of news sharing of type-θ sender is the probability that the number of successes is greater than or equal to $\frac{t-1+\delta}{2}$ in a Bernoulli trial where the success probability is J_θ. Using this property, we can easily calculate this. Note that we do not report the probability for $t = 1$ because it is obvious that the probability of sharing is one in the initial period.

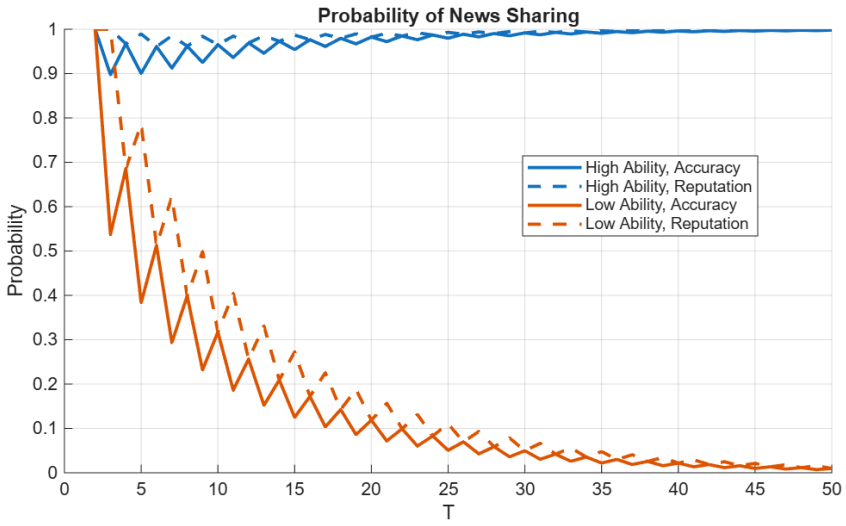

Figure 2: Probability of News Sharing ($p = 0.8, \alpha = 0.6, q = 0.6$)

where $\Delta(t) = -2$, the sender behaves in the same way regardless of the payoff structure (see also Figure 1). Thus, at the aggregate level, the effect of motivation is not so large.[17]

This is confirmed by Figure 2. First, for a type-h sender, the probability of news sharing remains above 0.9 regardless of the payoff structure. Second, for a type-l sender, the difference can be substantial—for example, at $t = 5$ the gap in probabilities is around 40 percentage points. However, this difference quickly converges to zero over time, and it is exactly zero in even periods.[18] Hence, at the aggregate level, the effect of motivation is not particularly large.

4 Experimental Design

We conducted laboratory experiments to test the above hypotheses. They were approved by the Institutional Review Board at Waseda University and pre-registered at AsPredicted (#208,478).[19]

4.1 Settings

The experiments were conducted 14 times at Waseda University, Tokyo, Japan, in 2025 (see Table A.1 in the Online Appendix for details). A total of 295 undergraduate and graduate students from various majors at Waseda University participated. Participants were recruited through the Sona system, which is used exclusively by Waseda students.[20]

Upon arrival, participants were randomly assigned to individual cubicles equipped with computers, ensuring that they could not view others' screens. They then received written instructions (see Online Appendix D), which were read aloud by the computer at the beginning

[17] See Online Appendix B for further discussion.
[18] In an even period, Δ takes only odd values; thus, it is never the case that $\Delta(t) = -2$.
[19] https://aspredicted.org/qb5nj4.pdf
[20] This is a system for participant management. Refer to https://www.sona-systems.com/ for more details.

of the session. All tasks and interactions were computerized, and the experiments were implemented using z-Tree.

Each session involved between 14 and 26 participants, and no one took part in more than one session. Participants were paired with another individual, and either two or four consecutive sessions of the same game were conducted, each consisting of 10 or 20 rounds. The sessions with 20 (respectively, 10) rounds correspond to the game with $T = 20$ (respectively, $T = 10$) in the theoretical model. In each round, every participant made a single decision. Pairs were randomly re-matched at the beginning of each round, ensuring that partners changed each time. Participants were not informed of their partners' identities. Within each pair, one participant was assigned the role of sender and the other of receiver, and these roles remained fixed throughout all sessions. That is, a participant who began as a sender (receiver) remained a sender (receiver) until the end of the experiment.

At the end of the experiments, participants were also asked to take a survey (see Online Appendix F).

4.2 Rules and Decisions

Two urns were used in the experiment—one black and one white—as illustrated in Figure 3. At the beginning of each round, one of the two urns was randomly selected, with each having an equal probability (50%) of being chosen. Each urn contained five balls, either black or white. The black urn contained four black balls and one white ball, whereas the white urn contained four white balls and one black ball. The color of the selected urn corresponds to ω_t in the theoretical model.

Figure 3: Two Urns

Sender's decision: In each round, the sender made the first decision. A ball was drawn from the urn selected at the beginning of the round, and the sender observed its color (either black or white), which corresponds to σ^s in the theoretical model. The receiver, however, could not observe the color of the drawn ball. After observing the ball's color, the sender was asked to send a message to the receiver indicating which urn was likely selected at the beginning of the

round. The sender could only send a message that matched the color of the observed ball. In other words, the sender had two options:

1. Send a message predicting that the selected urn matched the color of the observed ball (i.e., $m = \sigma^S$), or

2. Send no message at all (i.e., $m = \emptyset$).

Each sender was randomly assigned one of two types: type h (referred to as the correct-urn type in the instructions) with probability 0.6, or type l (the incorrect-urn type) with probability 0.4. A type-h sender observed a ball drawn from the urn selected for that round, whereas a type-l sender observed a ball drawn from the other, unselected urn. Thus, $q = 0.6$. A sender's type was determined once at the beginning of each session and remained fixed throughout all rounds of that session. The sender's type was re-randomized across sessions. Neither the sender nor the receiver was informed of the sender's type. After each draw, the ball was returned to the urn, ensuring that the urn compositions remained constant throughout the experiment.

Receiver's decision: After the sender's decision, the receiver made the second decision. First, the receiver learned whether the sender had sent a message and, if so, the content of that message. Then, a ball was drawn from the urn, and the receiver observed its color (i.e., σ^R), which the sender did not observe. The receiver's ball was always drawn from the correct urn selected at the beginning of the round; therefore, receivers were always type h.

If the sender had sent a message, the receiver chose one of the following two options ($r \in \{l, d\}$):

1. Evaluate the sender's message positively ("like"); or

2. Evaluate it negatively ("dislike").

If the sender had not sent a message, the receiver instead predicted which urn had been selected ($r' \in \{0, 1\}$):

1. The black urn; or

2. The white urn.

Belief elicitation: At the end of each round, the sender learned the receiver's choice — either the evaluation or the prediction — but the actual urn selected in that round was never revealed. Afterward, the sender was asked to report their belief about the probability of being type-h. An additional bonus payment was determined based on this reported belief.

The sequence of each session is summarized in Figure 4.

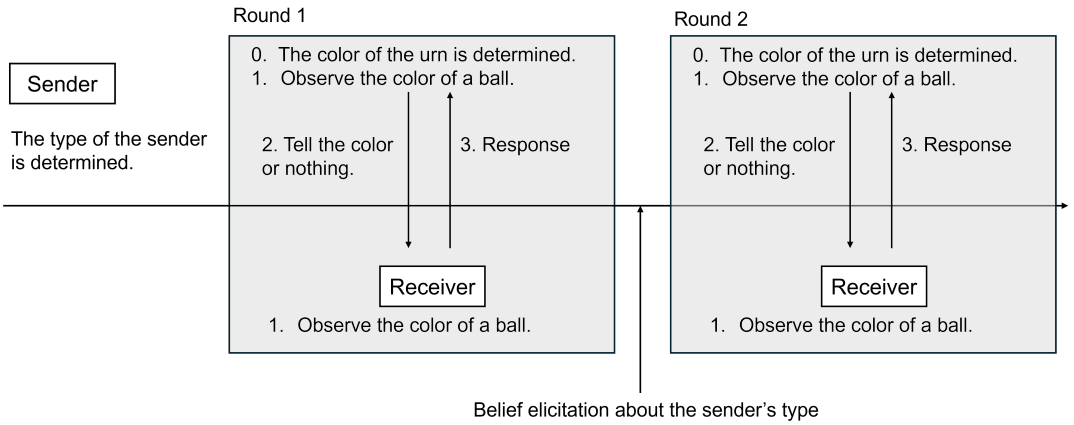

Figure 4: Sequence of Each Session

	10 rounds	20 rounds
Reputation	$R10$	$R20$
Accuracy	$A10$	$A20$

Table 1: Conditions in Experiments

4.3 Treatments

We introduced two treatments to examine the influence of senders' motivations.

1. **Condition R** (reputational motivation): The sender's reward probability depends on the receiver's evaluation, not on the accuracy of the sender's prediction.

2. **Condition A** (accuracy motivation): The sender's reward probability depends on the accuracy of the sender's prediction, not on the receiver's evaluation.

In addition, we varied the number of rounds per session to create two further conditions. Thus, there were four treatments in total: one dimension concerned the sender's motivation, and the other concerned the number of rounds (see Table 1).

1. In the baseline condition, each session consisted of 10 rounds, and four sessions were conducted within a single experiment. We refer to this as $R10$ when the sender had reputational motivation, and $A10$ when the sender had accuracy motivation.

2. In the extended condition, each session consisted of 20 rounds, and two sessions were conducted within each experiment. This treatment was designed to examine long-run outcomes. We refer to this as $R20$ when the sender had reputational motivation, and $A20$ when the sender had accuracy motivation.

4.4 Experimental Rewards

An experimental reward, in addition to the participation fee (2,000 JPY), depended on participants' decision outcomes (During the period in which the experiments were conducted, 1 JPY was approximately equivalent to 0.0065–0.007 USD). To eliminate the influence of individual risk attitudes, the experiment was designed so that participants' decisions affected the probability of receiving a reward rather than its amount. We refer to this measure as the *reward probability*. The method for calculating the reward probability differed between senders and receivers.

Sender's reward probability: In the treatments with reputational motivation ($R10$ and $R20$), the sender's reward probability was 0.8 if the sender sent a message and the receiver evaluated it positively ("like"), and 0.16 if the receiver evaluated it negatively ("dislike"). Whether the sender's prediction was correct or not did not affect the reward probability. If the sender chose not to send a message, the probability of earning a reward was 0.4.

In the treatments with accuracy motivation ($A10$ and $A20$), the sender's reward probability was 0.8 if the sender sent a message and the prediction was correct, and 0.16 if the prediction was incorrect. The receiver's evaluation—whether "like" or "dislike"—did not affect this probability. If the sender chose not to send a message, the probability of earning a reward was 0.4.

These probabilities correspond to set $\alpha = 0.6$ as in the numerical example discussed in Section 3.2.

Because performance in the game affects only the reward probability, and expected utility is linear in probability regardless of risk preferences, this design ensures that participants behave as if they were risk-neutral (Harrison, Martínez-Correa and Swarthout, 2013), which is consistent with our theoretical assumption. This can be also interpreted as a variant of the binarized scoring rule (Hossain and Okui, 2013).

Receiver's reward probability: For receivers, the reward probability depended on whether their choice, either the evaluation of the sender or the prediction of the correct urn, was appropriate. When the sender sent a message, the receiver's reward probability was determined as follows: if the sender's prediction was correct, the reward probability was 0.8 when the receiver chose "like" and 0.2 when the receiver chose "dislike." Conversely, if the sender's prediction was incorrect, the reward probability was 0.2 when the receiver chose "like" and 0.8 when the receiver chose "dislike."

If the sender did not send a message, the receiver's reward probability depended solely on the accuracy of their own prediction: it was 0.8 if the prediction was correct and 0.2 if it was incorrect. Receivers were not informed of their reward probabilities at the end of each round; however, the probabilities for all rounds were displayed at the end of each session.

Payment rounds: In treatments $A10$ and $R10$, one round out of ten in each session was randomly selected as the payment round. In treatments $A20$ and $R20$, two rounds out of twenty were randomly selected as payment rounds. One payment round was randomly selected from rounds 1 to 10, and another from rounds 11 to 20. In both cases, there were four payment rounds in total because four sessions were conducted in treatments $A10$ and $R10$, whereas two sessions were conducted in treatments $A20$ and $R20$. Only the outcomes from these selected rounds were used to determine participants' final payoffs; all other rounds were disregarded. Participants were not informed in advance which rounds would be chosen as payment rounds.

The reward probability assigned to a participant in a payment round determined their chance of earning 200 JPY as the experimental reward. To implement this, a lottery was conducted. A random integer between 1 and 100 was drawn, and the participant received 200 JPY if the number drawn was less than or equal to their reward probability. For example, if a participant's reward probability in the payment round was 0.16, they received 200 JPY if the number drawn was 16 or lower, but not if it was 17 or higher.

Additional bonus for senders: Each participant's final payment consisted of three components: the participation fee, the experimental reward determined by the reward probability in the selected payment rounds, and an additional bonus that applied only to senders. The additional bonus was provided to elicit senders' belief about whether they were type-h. It was determined using the binarized scoring method proposed by Hossain and Okui (2013).

At the end of each round, the sender reported their belief about the probability of being type h. The sender's report in the payment round was used to determine the probability of receiving the additional bonus, following the procedure below:

1. A random integer between 1 and 100 was drawn.
2. If the randomly selected number was greater than or equal to the probability reported by the sender, the sender received a lottery that paid 50 JPY with a probability equal to the selected number.
3. If the randomly selected number was less than the reported probability, the sender received a lottery that paid 50 JPY only if they were a type-h sender.

After determining the lottery type, the lottery was conducted, and the sender received the 50 JPY bonus if the outcome was successful. In the case of the lottery, conditional on being a type-h, payment was made only when the sender was indeed a type h. Unlike other belief elicitation methods, such as the quadratic scoring rule, this method, known as the BDM method, ensures that truth-telling is optimal regardless of a participant's risk preferences (Becker, DeGroot and Marschak, 1964). We explicitly explained the optimality of truthful reporting, along with a proof of this property, in the experimental instructions (see Online Appendix D).

Table 2: Summary of Key Variables

Treatment	Sender Type	Round	Obs.	Share	Belief
$A10$ and $A20$	type l	1-10	1170	0.511	36.942
$A20$	type l	11-20	310	0.358	25.326
$A10$ and $A20$	type h	1-10	1090	0.700	62.080
$A20$	type h	11-20	390	0.762	68.915
$R10$ and $R20$	type l	1-10	930	0.519	43.098
$R20$	type l	11-20	290	0.510	41.824
$R10$ and $R20$	type h	1-10	1290	0.683	64.832
$R20$	type h	11-20	450	0.816	71.496

4.5 Comprehension Test and Practice Session

To ensure that participants fully understood the experiment, a comprehension test (see Online Appendix E) and a practice session consisting of three rounds were administered before the main task. These were designed to help participants grasp the basic rules of decision-making, rather than to discover optimal strategies. Any questions raised by participants were addressed individually until all instructions were clearly understood. Thus, after completing the comprehension test and practice session, it could be reasonably concluded that participants understood the experimental rules.

5 Results

As we presented in Section 4, this experiment aimed to evaluate the decision-making processes of participants by establishing two primary conditions: A (accuracy motivation) and R (reputational moivation), each comprising type h and type l. Decisions were repeated within each condition, and the primary variables of Share (sharing behavior) and Belief (belief updating) were measured throughout.

Section 5.1 presents the descriptive analysis of the data. Section 5.2 then examines sharing behavior in detail, followed by an analysis of belief updating in Section 5.3.

5.1 Descriptive Analysis

We start with descriptive analysis. The summary statistics on the key variables are summarized in Table 2.

Prediction 1 (Peer feedback and sharing behaviors): First, to see whether Prediction 1 holds, we examine the histogram of the proportion of Sharing (Figure 5, left panel). Although the theoretical model predicts that sharing behavior should be homogeneous between participants, the data reveal substantial heterogeneity in individual decision-making. Even among type-h senders, some participants consistently choose not to share, whereas others among

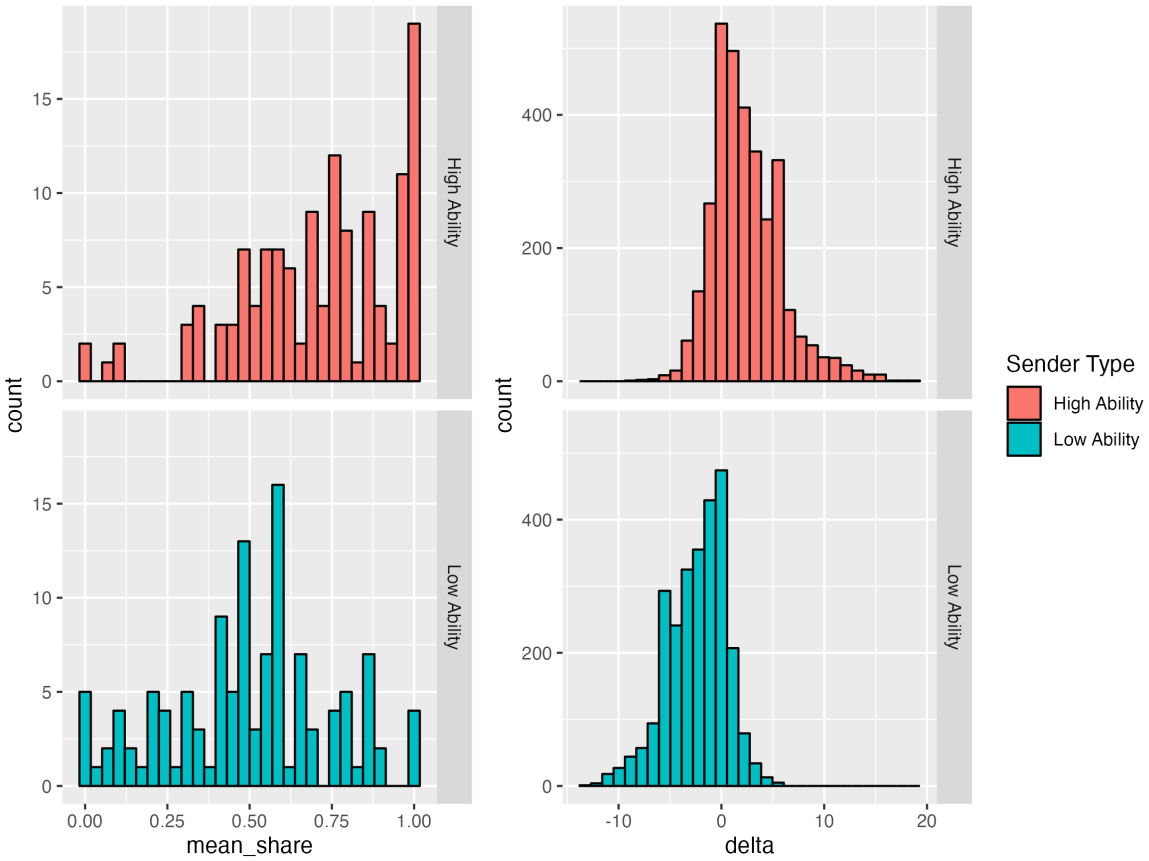

Figure 5: Histogram of Sharing Behavior and Δ

type-l senders always share their signals. This variability may arise from the' uncertainty of participants about their own classification as types h or l.

According to the model, the cumulative feedback variable Δ influences decision-making. Thus, the distribution of Δ should provide further insight (Figure 5, right panel). As anticipated, type-h senders exhibit, on average, higher values of Δ, with most cases exceeding the thresholds of -2 and -1 predicted in the theory. However, the level of sharing is relatively low, suggesting that there may be errors or biases in the process of translating Δ into decision-making. Note that attributing these discrepancies to risk attitudes is not justified because our design ensures that every participants behaves in a risk-neutral manner. Similarly, for type-l senders, approximately half of the cases fall below the threshold, which is consistent with the observed sharing rates of 40-50%.

Prediction 2 (Self-selection): Second, in relation to Prediction 2, when comparing the types h and l in Table 2, Sharing and Belief were found to be higher in the former, aligning with theoretical expectation. However, the observed differences were more minor than anticipated, primarily due to the unexpectedly low levels of Sharing and Belief within type-h senders.

Additionally, within type-h senders, a slight increase in both Sharing and Belief was observed when comparing the first half (1-10) and the second half (11-20). Conversely, within type-l senders, a decrease in both Sharing and Belief was noted between the two halves, corroborating the theoretical implications.

To see these points in detail, Figure 6 illustrates the trends in Sharing over time, with observed values represented by solid lines and theoretical values depicted by dashed lines. Within type-h senders, both Conditions A and R show a pronounced tendency for the observed values to fall below the theoretical predictions. Although there appears to be a narrowing of the gap in the latter half of the trials, this effect is exceedingly limited. Conversely, within type-l senders, the difference between the observed and theoretical values is more minor, with a clear trend of reduction over time; however, the difference still persists. This suggests that participants with lower ability exhibit a more consistent alignment with theoretical expectations regarding Sharing.

In summary, although the diffusion of fake news is often attributed primarily to sharing by users connected to unreliable sources, our findings indicate that users connected to reliable sources (type h) are also reluctant to share news, which ultimately undermines self-selection through peer feedback.

Prediction 3 (Effect of motivation): Lastly, regarding Prediction 3, Table 2 shows no clear differences between Conditions A and R (this point is formally tested in later subsections), although some differences are observable – for example, in the share of type-l senders during periods 11–20. This pattern is consistent with the theoretical prediction that such differences exist but are only modest in magnitude.

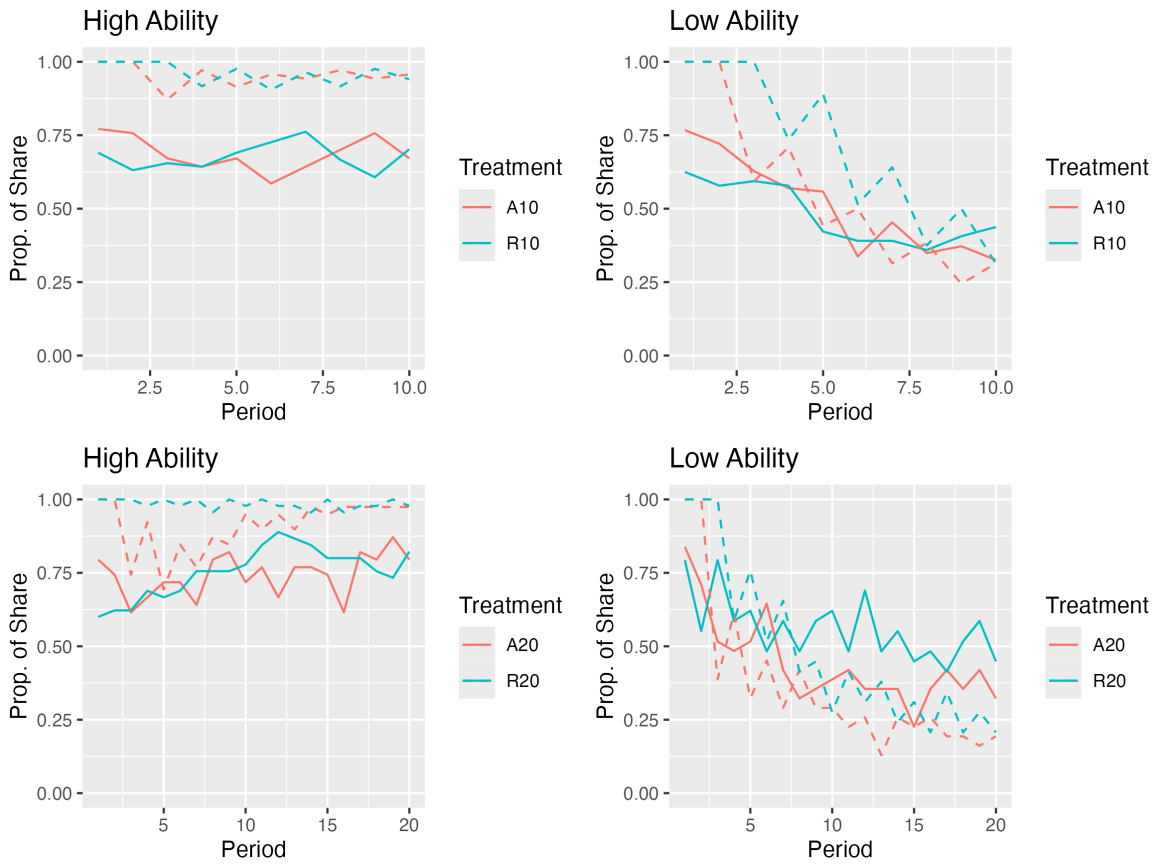

Figure 6: Trend of Sharing Behavior

Condition	Likelihood
Accuracy	0.925
Reputation	0.945

Table 3: Likelihood that Receivers Follow the Equilibrium Strategy

This implies that accuracy prompts alone are unlikely to substantially curb misinformation and that policies focusing solely on reputational incentives may be insufficient.

Receivers' behaviors: While our primary focus is on senders' sharing behavior, it is also informative to examine whether receivers' behavior aligns with the theoretical predictions presented in Lemma 1. Our theory predicts that receivers should follow their private signals: (a) when the sender transmits a message, receivers post a "like" if and only if the color indicated by the message coincides with the color of the ball they observe; and (b) when the sender does not transmit any message, receivers truthfully reveal the color of the ball they observe. This is true regardless of the experimental condition.

Table 3 reports the fraction of receiver behaviors consistent with these theoretical predictions. The results show that receivers' behavior aligns remarkably well with the theory, with compliance exceeding 90%.[21] Therefore, it is unlikely that deviations from equilibrium behavior on the receiver side account for the observed deviations in senders' behavior.

5.2 Sharing Behavior

To rigorously test the empirical patterns observed in the previous subsection on sharing behaviors, we conducted a regression analysis controlling for various factors (Table 4). Given that decision-making occurs repeatedly, we employed a linear mixed-effects model to account for individual effects, order effects, and session effects. The independent variables in this analysis were Treatment, Sender Type, and Δ.

The results of the regression analysis are summarized as follows. First, the coefficient on Treatment is consistently insignificant, regardless of the set of control variables. Second, as noted in the previous subsection, the coefficient on Sender Type is statistically significant (type-h is 20.6 percentage points more likely to share in column (2)), indicating that type-h senders tend to share their signals more than type-l senders.

Theoretically, it is the value of Δ that governs decision-making. To account for this, column (3) includes Δ as an independent variable. The results show that a higher Δ significantly increases the probability of sharing. Moreover, once Δ is included, the estimated coefficient on Sender Type becomes much smaller and is no longer significant at the 5% level. This suggests that the difference in sharing behavior across Sender Types is driven by differences in Δ, as predicted by the theory.

[21] In Condition R, a "dislike" directly reduces the sender's stage-game payoff, which could induce receivers to avoid choosing "dislike" for altruistic reasons. However, our data provide little evidence of such behavior.

Table 4: Regression Analysis of Sharing Behavior

	Dependent variable:			
	Sharing Behavior (Share = 1, Not = 0)			
	(1)	(2)	(3)	(4)
Treatment (R = 1; A = 0)	0.037	0.019	0.007	0.007
	(0.035)	(0.033)	(0.031)	(0.026)
Sender Type (h = 1; l = 0)		0.206	0.028	0.027
		(0.014)	(0.016)	(0.016)
Delta			0.039	0.030
			(0.002)	(0.002)
Match (t-1)				−0.024
				(0.018)
Good (t-1)				0.219
				(0.017)
Bad (t-1)				0.158
				(0.016)
Constant	0.587	0.489	0.578	0.477
	(0.032)	(0.028)	(0.030)	(0.026)
Observations	5,920	5,920	5,920	5,920
Log Likelihood	−3,725.252	−3,622.421	−3,427.361	−3,308.740
Akaike Inf. Crit.	7,460.503	7,256.843	6,868.722	6,637.481
Bayesian Inf. Crit.	7,493.934	7,296.959	6,915.525	6,704.342

Note: Random effects regarding Subjects and Blocks are considered.

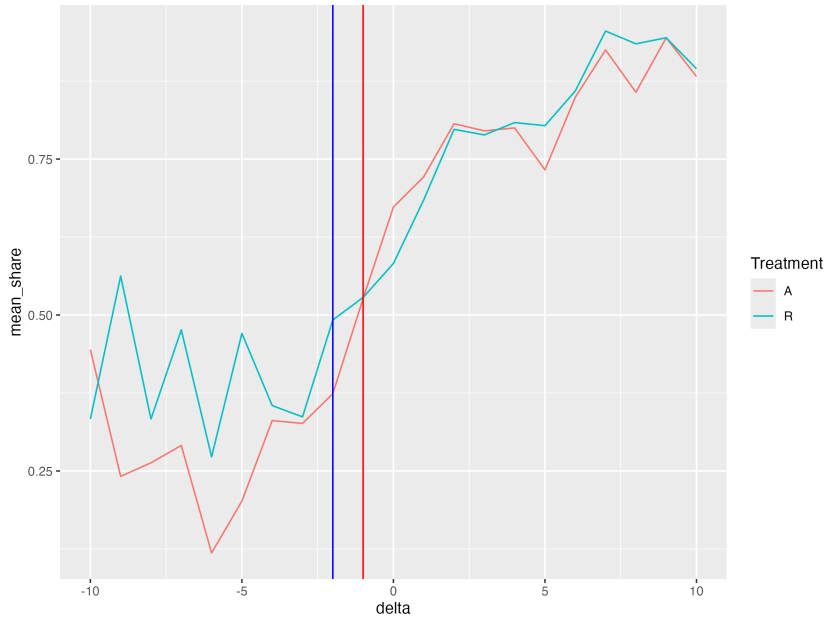

Figure 7: Rate of Sharing Depending on Δ

Finally, to examine whether senders respond to Δ itself or merely to feedback in the previous period, column (4) includes lagged feedback as additional covariates. Although these variables are also statistically significant, the coefficient on Δ changes little and remains statistically significant. This indicates that senders respond to cumulative feedback rather than only to one-period feedback.

To conduct a more detailed empirical analysis of the theoretical relationship between Δ and sharing behavior, we refer to the non-continuous relationship described in Prediction 1 and illustrated in Figure 1, where sharing occurs based on threshold values (-1 for A and -2 for R). The empirical counterpart of this theoretical relationship is shown in Figure 7, which plots the proportion of sharing as a function of Δ. Observations with $\Delta < -10$ or $\Delta > 10$ are excluded due to the small number of cases in those ranges.

When accounting for the inherent noise in individual decision-making, the observed data appear broadly consistent with the theoretical prediction. Specifically, we find a positive relationship between Δ and sharing: larger expected gains are associated with more optimal decisions, albeit with some errors. This pattern aligns closely with the theoretical model, which predicts that greater disparities in expected payoffs lead to more accurate optimization behavior. Moreover, the finding that the proportion of Share converges toward 50% near the threshold is consistent with the idea that expected gain differences are minimal in that region.

Nevertheless, there remains a significant and non-negligible deviation from the theoretical relationship between Δ and the sharing rate. First, even when Δ takes on sufficiently large values, sharing does not necessarily approach one; attributing this solely to random error would be premature. Likewise, when Δ decreases, the sharing rate remains around 20–40%, suggesting that this pattern cannot be explained by random noise alone. Second, there appears to be a difference in sharing behavior between Conditions A and R for small values of Δ,

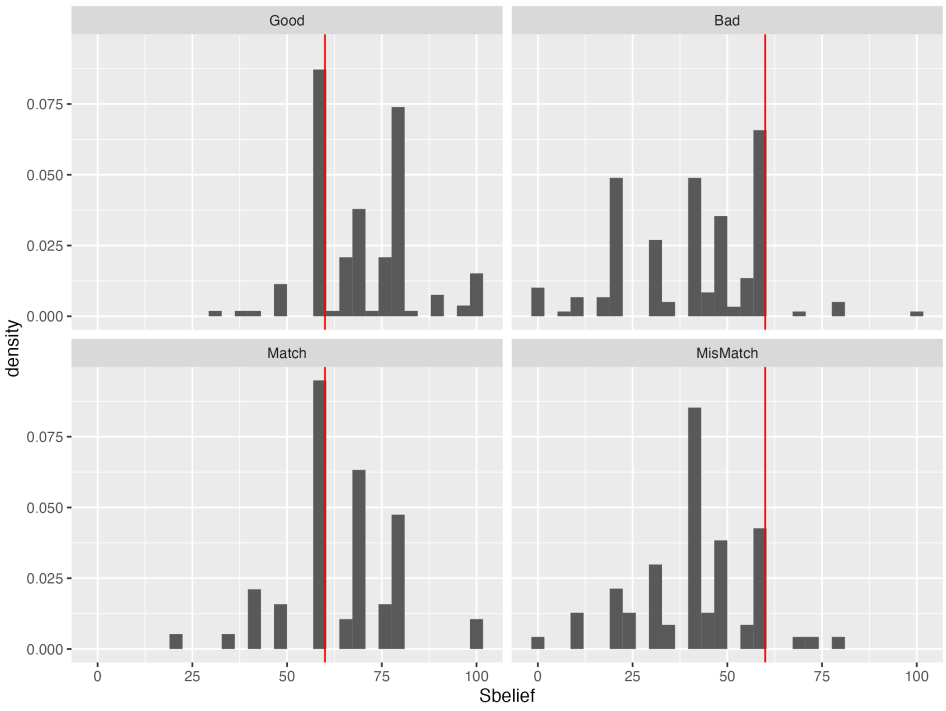

Figure 8: Belief Updating in the 1st Period

although this difference is not statistically significant according to the linear mixed-effects model.

To further investigate these observations, it is necessary to examine how senders update their beliefs in response to feedback about their decisions. The next subsection therefore turns to an analysis of belief formation.

5.3 Belief Updating

To elucidate the characteristics of belief updating, we first examine the beliefs reported after feedback in the initial round (i.e., $\tilde{q}_2$), as shown in Figure 8. The two left panels show histograms after positive feedback (Like or Match), where the updated theoretical value remains constant at 76.1% (0.761), while the two right panels present histograms after negative feedback (Dislike or Mismatch), where the theoretical value is 41.4% (0.414). According to the experimental design, the initial belief was set at 60 ($q = 0.6$).

These figures reveal that although most of the participants update their beliefs in the correct direction, their reported beliefs do not converge to the theoretical values. In three of the four cases—excluding Mismatch—the mode remains at the initial belief of 60, suggesting a degree of stickiness in belief updating. Moreover, some participants even revise their beliefs in the opposite direction, underscoring the complexity and heterogeneity of the belief updating process.

To examine whether belief updating differs across conditions, we estimated a linear mixed-effects model with Belief as the dependent variable and condition, Δ, and other control variables

Table 5: Regression Analysis on Belief Updating

	Dependent variable:			
	Belief			
	(1)	(2)	(3)	(4)
Treatment (R=1; A=0)	4.166	4.526	4.313	4.730
	(1.840)	(1.897)	(1.808)	(1.865)
Delta	4.371	4.373		
	(0.080)	(0.080)		
Gender		1.374		1.594
		(1.922)		(1.890)
CRRA		−2.710		−2.141
		(1.741)		(1.713)
QuizTotal		0.196		0.689
		(1.270)		(1.249)
Economics		−0.743		−1.237
		(2.017)		(1.984)
Num. Match			2.224	2.228
			(0.231)	(0.231)
Num. Good			4.293	4.294
			(0.117)	(0.117)
Num. MisMatch			−5.772	−5.774
			(0.177)	(0.177)
Num. Dislike			−2.805	−2.803
			(0.174)	(0.174)
Constant	49.394	49.342	51.142	49.887
	(1.567)	(3.982)	(1.466)	(3.885)
Observations	5,920	5,920	5,920	5,920
Log Likelihood	−26,928.310	−26,921.040	−26,773.780	−26,766.700
Akaike Inf. Crit.	53,868.630	53,862.070	53,565.570	53,559.400
Bayesian Inf. Crit.	53,908.750	53,928.930	53,625.740	53,646.320

Note: Random effects regarding Subjects and Blocks are considered.

as independent variables. The results, reported in Columns (1) and (2) of Table 5, show a consistent tendency for beliefs to be higher in Condition R. This finding remains robust after including additional control variables.[22]

Furthermore, by decomposing Δ, we examined whether there are differences in the impact of feedback types on Belief. Let us denote the total number of likes so far, that of dislikes, that of matches, and that of mismatches by N_L, N_D, N_M, and N_{MM}, respectively. Then, by its definition,

$$\Delta = N_L - N_D + N_M - N_{MM}$$

(see equation (1)). Consequently, we also estimated a model that included these four variables as independent variables instead of Δ (Columns (3) and (4) in Table 5). The results indicate that the tendency for Belief to be higher in the Reputation condition persists. Additionally, some heterogeneity in the effects of feedback types was observed: the effect of Match appears weaker than that of other feedback types. This suggests that the nature of feedback may influence belief updating in different ways.

Although this suggests heterogeneity in belief updating across feedback types, this interpretation warrants caution. The previous model assumes a linear relationship between belief and feedback, whereas, theoretically, these relationships are non-linear due to the nature of Bayesian updating. Consequently, the coefficients from the prior model may not be valid for direct interpretation. To address this issue, we estimate an alternative model using the logit of belief as the dependent variable. Let B_t denote belief after feedback in round t,[23] and let $\mathbf{1}_t\{F\}$ be an indicator function that takes one if a participant received type F in round t, where $F \in \{L, D, M, MM\}$. As shown in the Appendix A.4, if participants follow Bayesian updating, the difference between the logit of B_t and that of B_{t-1} is linear in $\mathbf{1}_t\{F\}$. Specifically, we have[24]

$$\log\left(\frac{B_t}{100 - B_t}\right) - \log\left(\frac{B_{t-1}}{100 - B_{t-1}}\right) = \beta_L \mathbf{1}_t\{L\} + \beta_D \mathbf{1}_t\{D\} + \beta_M \mathbf{1}_t\{M\} + \beta_{MM} \mathbf{1}_t\{MM\}, \quad (2)$$

where

$$\beta_L = \beta_M = -\beta_D = -\beta_{MM} = \log\left(\frac{p^2 + (1-p)^2}{2p(1-p)}\right) \simeq 0.754.$$

Motivated by this property, we estimated a linear mixed-effects model with the change in the logit of Belief as the dependent variable and $\mathbf{1}_t\{F\}$ as independent variables.

The results are reported in Table 6. First, the estimated effects are generally smaller in

[22] Gender is a dummy variable equal to 1 for female and 0 for male. Observations are omitted when the response is "Other. "Economics is a dummy variable equal to 1 if a subject has taken any course related to economics or game theory. CRRA represents the constant relative risk aversion parameter derived from Question 3 of the post-experiment survey, following the method of Holt and Laury (2002). QuizTotal is the total number of correct answers to Questions 4–6 (ranging from 0 to 3), based on the cognitive reflection test developed by Frederick (2005).

[23] In the theory, this is denoted by $\tilde{q}_{t+1}$. We use not $t+1$ but t because the posterior belief was elicited at the end of round t not the beginning of round $t+1$ in the experiment.

[24] This specification has been widely used to test whether participants follow Bayesian updating in the literature of experimental economics (e.g., Möbius et al., 2022).

Table 6: Regression Analysis on Belief Updating: Log-Likelihood

	Dependent variable:		
	DifL		
	Accuracy	Reputation	Pooled
Like	0.65	0.64	0.64
	(0.07)	(0.08)	(0.05)
Dislike	−0.67	−0.68	−0.67
	(0.10)	(0.10)	(0.07)
Match	0.48	0.48	0.48
	(0.21)	(0.16)	(0.13)
MisMatch	−0.78	−0.60	−0.70
	(0.10)	(0.15)	(0.09)
R^2	0.12	0.12	0.12
Adj. R^2	0.12	0.12	0.12
Num. obs.	2960	2960	5920

Robust standard errors are calculated.

magnitude than the theoretical predictions, indicating that belief updating is more conservative than Bayesian updating (Phillips and Edwards, 1966). Second, in both the Accuracy and Reputation conditions, like and dislike feedback exhibit the symmetric relationship as predicted by theory (i.e., $\beta_L = -\beta_D$); however, an asymmetry emerges between Match and Mismatch. Specifically, the negative effect of Mismatch on belief is stronger than the positive effect of Match. This difference is statistically significant. The literature documents a negativity bias, whereby negative information is processed more thoroughly than positive information (Baumeister et al., 2001). Our results indicate that such negativity bias manifests in responses to indirect feedback (Match vs. Mismatch), rather than direct feedback (Like vs. Dislike). This distinction underscores the importance of differentiating between direct and indirect feedback.

Moreover, the effect of Mismatch on belief reduction is more pronounced in Condition A, which likely contributes to the observed tendency for Belief to be higher in Condition R than in Condition A (as reported in Table 2).

Lastly, Match and Like have the same information value in the equilibrium; therefore $\beta_L = \beta_M$ should hold, but this is not the case.

It should also be noted that biased belief updating alone does not fully account for the deviations from the theoretical predictions. Figure 9 shows how the probability of sharing varies with the subjective belief B_{t-1}, focusing on periods $t \geq 2$. The theory predicts that senders in Condition R should share their signals if and only if $B_{t-1} > 29.2$, whereas senders in Condition A should do so if and only if $B_{t-1} > 15.3$.[25] However, we observe that even participants with $B_{t-1} \leq 10$ share with probability exceeding 0.2, while some participants with

[25]That is, $\bar{q}_R \simeq 0.292$ and $\bar{q}_A \simeq 0.153$.

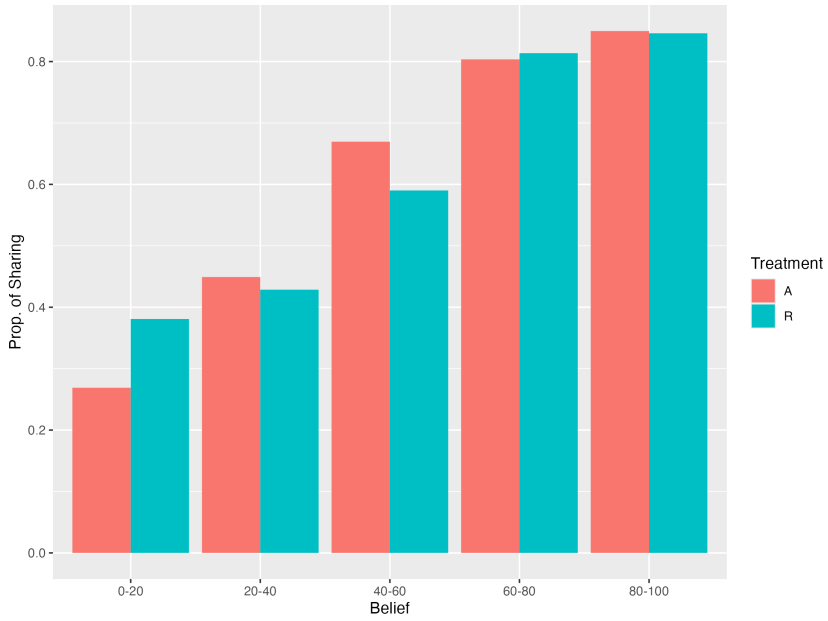

Figure 9: Probability of Sharing and Belief

$B_{t-1} > 90$ fail to share with probability exceeding 0.1. These decision errors, together with biased belief updating, contribute to the deviations from the theoretical predictions.

6 Concluding Remarks

This study examined how peer feedback on social media affects the spread of fake news by combining a formal model with a controlled laboratory experiment. In our model, senders learn about the reliability of their information sources from others' reactions, and such feedback endogenously generates self-selection: individuals with unreliable sources gradually stop sharing misinformation. Furthermore, fake news spreads more when senders are motivated by reputation rather than accuracy, though this effect is modest.

The laboratory evidence broadly supports these predictions. First, feedback reduces fake news sharing and induces behavioral divergence between type-h and l participants. Second, differences across motivational incentives are minimal.

Nevertheless, the effects of peer feedback are quantitatively weaker than predicted. Two deviations account for this gap: belief updating is more conservative than Bayesian learning, and sharing decisions is noisy. These findings suggest that while individuals conceptually learn from peer reactions, their behavioral adjustment remains incomplete and sluggish.

Overall, our results highlight both the potential and limitations of peer feedback as a self-organized corrective force against misinformation. First, our findings suggest that effective platform interventions should go beyond suppressing misinformation from unreliable sources and instead actively encourage sharing by users with reliable sources. Moreover, accuracy prompts alone appear insufficient to curb fake news when receivers are reliable, calling into

question approaches that rely primarily on reputational incentives. Finally, both direct and indirect feedback shape sharing behavior, yet users respond weakly to positive indirect feedback, highlighting the importance of platform designs that make such signals more salient.

A Omitted Proofs

A.1 Proof of Lemma 1

Proof of (a): Suppose that the receiver thinks that the sender is type h with probability $\tilde{q}$. Without loss of generality, suppose that $m = 0$.

(i). Case 1: $\sigma^R = 0$**.**

$$\mathbb{E}[v(0,\omega,l)|\sigma^R = 0] = \frac{\tilde{q}\frac{1}{2}p^2 + (1-\tilde{q})\frac{1}{2}(1-p)p}{\tilde{q}[\frac{1}{2}p^2 + \frac{1}{2}(1-p)^2] + (1-\tilde{q})[\frac{1}{2}(1-p)p + \frac{1}{2}p(1-p)]}.$$

$$\mathbb{E}[v(0,\omega,d)|\sigma^R = 0] = 1 - \frac{\tilde{q}\frac{1}{2}p^2 + (1-\tilde{q})\frac{1}{2}(1-p)p}{\tilde{q}[\frac{1}{2}p^2 + \frac{1}{2}(1-p)^2] + (1-\tilde{q})[\frac{1}{2}(1-p)p + \frac{1}{2}p(1-p)]}.$$

Because

$$\tilde{q}\frac{1}{2}p^2 + (1-\tilde{q})\frac{1}{2}(1-p)p > \tilde{q}\frac{1}{2}(1-p)^2 + (1-\tilde{q})\frac{1}{2}p(1-p) \Leftrightarrow p > 1-p$$

holds,

$$\mathbb{E}[v(0,\omega,l)|\sigma^R = 0] > \mathbb{E}[v(0,\omega,d)|\sigma^R = 0].$$

Therefore, $r = l$ if $m = \sigma_R = 0$.

(ii). Case 2: σ^R **=1.**

$$\mathbb{E}[v(0,\omega,l)|\sigma^R = 1] = \frac{\tilde{q}\frac{1}{2}p(1-p) + (1-\tilde{q})\frac{1}{2}(1-p)^2}{\tilde{q}p(1-p) + (1-\tilde{q})[\frac{1}{2}p^2 + \frac{1}{2}(1-p)^2]}.$$

$$\mathbb{E}[v(0,\omega,d)|\sigma^R = 1] = 1 - \frac{\tilde{q}\frac{1}{2}p(1-p) + (1-\tilde{q})\frac{1}{2}(1-p)^2}{\tilde{q}p(1-p) + (1-\tilde{q})[\frac{1}{2}p^2 + \frac{1}{2}(1-p)^2]}.$$

Because

$$\tilde{q}\frac{1}{2}p(1-p) + (1-\tilde{q})\frac{1}{2}(1-p)^2 > \tilde{q}\frac{1}{2}p(1-p) + (1-\tilde{q})\frac{1}{2}p^2 \Leftrightarrow p > 1-p$$

holds under the assumption that $\tilde{q} < 1$,

$$\mathbb{E}[v(0,\omega,l)|\sigma^R = 1] < \mathbb{E}[v(0,\omega,d)|\sigma^R = 1].$$

Therefore, $r = d$ if $m = 0$ but $\sigma^R = 1$.

Proof of (b): Suppose that the sender does not share the signal, i.e., $m = \emptyset$. Suppose without loss of generality that the receiver privately observes $\sigma^R = 0$, and she assigns a probability of $\tilde{q} \in (0,1)$ to the sender being the high type. It suffices to show $\Pr(\omega = 0 \mid \sigma^R = 0, m = \emptyset) > 1/2$, which is equivalent to

$$\frac{\Pr\left(\sigma^R = 0, m = \emptyset \mid \omega = 0\right)}{\Pr\left(\sigma^R = 0, m = \emptyset \mid \omega = 1\right)} > 1.$$

Intuition is that the magnitude of the reciever's signal is greater than the information contained in $m = \emptyset$. First, let $\pi \coloneqq \Pr(\sigma^S = \omega \mid \omega) = \tilde{q}p + (1-\tilde{q})(1-p)$ denote an expected accuracy of the sender's signal, so $1 - p < \pi < p$. Define $\rho_k \coloneqq \Pr(m = \emptyset \mid \sigma^S = k)$ as the probability that the sender chooses not to share his signal $\sigma^S = k$. Then, we have

$$\begin{aligned}
&\frac{\Pr\left(\sigma^R = 0,\ m = \emptyset \mid \omega = 0\right)}{\Pr\left(\sigma^R = 0,\ m = \emptyset \mid \omega = 1\right)} > 1 \\
\Leftrightarrow\ &\frac{p\left[\pi\rho_0 + (1-\pi)\rho_1\right]}{(1-p)\left[\pi\rho_1 + (1-\pi)\rho_0\right]} > 1 \\
\Leftrightarrow\ &\log\left(\frac{p}{1-p}\right) + \log\left(\frac{\pi\rho_0 + (1-\pi)\rho_1}{\pi\rho_1 + (1-\pi)\rho_0}\right) > 0 \\
\Leftrightarrow\ &\log\left(\frac{p}{1-p}\right) - \log\left(\frac{\pi\rho_1 + (1-\pi)\rho_0}{\pi\rho_0 + (1-\pi)\rho_1}\right) > 0.
\end{aligned}$$

To show this inequality, we resort to the monotonicity of the function $\log\left(x/(1-x)\right)$ on $x \in (0,1)$. Rewrite

$$\frac{\pi\rho_1 + (1-\pi)\rho_0}{\pi\rho_0 + (1-\pi)\rho_1} = \frac{\gamma}{1-\gamma},$$

where

$$\gamma \coloneqq \pi\frac{\rho_1}{\rho_1 + \rho_0} + (1-\pi)\frac{\rho_0}{\rho_1 + \rho_0}.$$

It follows from $\gamma < p$ that the above inequality holds, thus proving (b) of Lemma 1. □

A.2 Proof of Proposition 1

Without loss of generality, suppose that $\sigma^S = 0$.

Proof of (a): On the one hand, when $m = 0$, the sender's expected payoff is

$$\begin{aligned}
\mathbb{E}[u(0,r)|\sigma^S = 0] =&[\tilde{q}(p^2 + (1-p)^2) + (1-\tilde{q})2p(1-p)] \\
&- [\tilde{q}2p(1-p) + (1-\tilde{q})(p^2 + (1-p)^2)]\alpha.
\end{aligned}$$

On the other hand, when $m = \emptyset$, the sender's expected payoff is 0.

Therefore, the sender chooses $m = 0$ if and only if

$$\begin{aligned}
&\tilde{q}(p^2 + (1-p)^2) + (1-\tilde{q})2p(1-p) \geq [\tilde{q}2p(1-p) + (1-\tilde{q})(p^2 + (1-p)^2)]\alpha \\
\Leftrightarrow &\tilde{q}(p^2 + (1-p)^2) + (1-\tilde{q})2p(1-p) \geq \frac{\alpha}{1+\alpha} \\
\Leftrightarrow &\tilde{q} \geq \bar{q}_R := \frac{1}{(2p-1)^2}\left(\frac{\alpha}{1+\alpha} - 2p(1-p)\right).
\end{aligned}$$

Proof of (b): On the one hand, when $m = 0$, the sender's expected payoff is

$$\mathbb{E}[u(0,\omega)|\sigma^S = 0] = [\tilde{q}p + (1-\tilde{q})(1-p)] - [\tilde{q}(1-p) + (1-\tilde{q})p]\alpha.$$

On the other hand, when $m = \emptyset$, the sender's expected payoff is 0.

Therefore, the sender sends $m = \sigma^S$ if and only if

$$\tilde{q} \geq \bar{q}_A := \frac{p\alpha - (1-p)}{(2p-1)(1+\alpha)}.$$

Proof of (c):

$$\bar{q}_A \geq \bar{q}_R \Leftrightarrow -(1-p)(1-\alpha) \leq 0,$$

which always holds.

□

A.3 Proof of Proposition 2

Proof of (c): We focus on the case where the sender is type h with reputational motivation. The remaining cases are analogous and omitted. It suffices to show

$$\lim_{T\to\infty} \Pr\left(\Delta(T) \geq \underline{\Delta}_R \mid \theta = h\right) = 1.$$

For each period τ, let $W_\tau \in \{0, 1\}$ be the indicator that the sender observes "like" or "match." Since the sender is assumed to ge type h, a sequence $(W_\tau)_{\tau\geq 1}$ is i.i.d. Bernoulli with mean

$$\Pr(W_\tau = 1) = p^2 + (1-p)^2,$$

which is strictly greater than $1/2$.

The net number of positive feedback up to the beginning of period T, $\Delta(T)$, is therefore

$$\Delta(T) \;=\; \sum_{\tau=1}^{T-1} W_\tau - \sum_{\tau=1}^{T-1}(1-W_\tau) \;=\; 2\sum_{\tau=1}^{T-1} W_\tau - (T-1).$$

Hence,

$$\Delta(T) \geq \underline{\Delta}_R \quad \Longleftrightarrow \quad \frac{\sum_{\tau=1}^{T-1} W_\tau}{T-1} \;\geq\; \frac{\underline{\Delta}_R}{2(T-1)} + \frac{1}{2},$$

and the right-hand side threshold converges to $1/2$ as $T \to \infty$. Choose small $\varepsilon > 0$ so that for T large enough,

$$\frac{\underline{\Delta}_R}{2(T-1)} + \frac{1}{2} \;<\; \frac{1}{2} + \varepsilon \;<\; p^2 + (1-p)^2 - \varepsilon.$$

Then, for T large enough,

$$\begin{aligned}
\Pr\left(\Delta(T) \;\geq\; \underline{\Delta}_R \mid \text{sender is type } h\right) &= \Pr\left(\frac{\sum_{\tau=1}^{T-1} W_\tau}{T-1} \;\geq\; \frac{\underline{\Delta}_R}{2(T-1)} + \frac{1}{2}\right) \\
&\geq\; \Pr\left(\frac{\sum_{\tau=1}^{T-1} W_\tau}{T-1} \;\geq\; p^2 + (1-p)^2 - \varepsilon\right) \\
&\longrightarrow\; 1,
\end{aligned}$$

where the convergence to 1 follows from the weak Law of Large Numbers. □

A.4 Derivation of Equation (2)

Remember that

$$\begin{aligned}
q_{t+1}(r_t = l, q_t) &= \frac{q_t[p^2 + (1-p)^2]}{q_t[p^2 + (1-p)^2] + (1-q_t)2p(1-p)} \\
\Leftrightarrow \frac{q_{t+1}(r_t = l, q_t)}{1 - q_{t+1}(r_t = l, q_t)} &= \frac{q_t}{1-q_t}\frac{p^2 + (1-p)^2}{2p(1-p)}.
\end{aligned}$$

By taking the log, we have

$$\log\left(\frac{q_{t+1}(r_t = l, q_t)}{1 - q_{t+1}(r_t = l, q_t)}\right) - \log\left(\frac{q_t}{1-q_t}\right) = \log\left(\frac{p^2 + (1-p)^2}{2p(1-p)}\right).$$

Similarly, we have

$$\log\left(\frac{q_{t+1}(r'_t = \sigma^S, q_t)}{1 - q_{t+1}(r'_t = \sigma^S, q_t)}\right) - \log\left(\frac{q_t}{1-q_t}\right) = \log\left(\frac{p^2 + (1-p)^2}{2p(1-p)}\right).$$

$$\log\left(\frac{q_{t+1}(r_t = d, q_t)}{1 - q_{t+1}(r_t = d, q_t)}\right) - \log\left(\frac{q_t}{1-q_t}\right) = -\log\left(\frac{p^2 + (1-p)^2}{2p(1-p)}\right).$$

$$\log\left(\frac{q_{t+1}(r'_t \neq \sigma^S, q_t)}{1 - q_{t+1}(r'_t \neq \sigma^S, q_t)}\right) - \log\left(\frac{q_t}{1-q_t}\right) = -\log\left(\frac{p^2 + (1-p)^2}{2p(1-p)}\right).$$

By applying them, we have Equation (2).[26] □

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2024. "A model of online misinformation." *Review of Economic Studies*, 91(6): 3117–3150.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. 2024. "The economics of social media." *Journal of Economic Literature*, 62(4): 1422–1474.

Assenza, Tiziana, Alberto Cardaci, and Stefanie J Huber. 2024. "Fake news: susceptibility, awareness and solutions." *Working Paper*.

Badrinathan, Sumitra, and Simon Chauchard. 2024. ""I don't think that's true, bro!" social corrections of misinformation in India." *The International Journal of Press/Politics*, 29(2): 394–416.

Barthel, Michael, Amy Mitchell, and Jesse Holcomb. 2016. "Many Americans Believe Fake News Is Sowing Confusion." *https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/*, Pew Research Center, Journalism & Media, Accessed: 2025-10-15.

Baumeister, Roy F, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. "Bad is stronger than good." *Review of General Psychology*, 5(4): 323–370.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak. 1964. "Measuring utility by a single-response sequential method." *Behavioral Science*, 9(3): 226–232.

Bode, Leticia, and Emily K Vraga. 2018. "See something, say something: Correction of global health misinformation on social media." *Health Communication*, 33(9): 1131–1140.

Bolte, Lukas, and Tony Q Fan. 2024. "Motivated mislearning: The case of correlation neglect." *Journal of Economic Behavior & Organization*, 217: 647–663.

Burtch, Gordon, Qinglai He, Yili Hong, and Dokyun Lee. 2022. "How do peer awards motivate creative content? Experimental evidence from Reddit." *Management Science*, 68(5): 3488–3506.

Çelen, Boğaçhan, Shachar Kariv, and Andrew Schotter. 2010. "An experimental test of advice and social learning." *Management Science*, 56(10): 1687–1701.

Cisternas, Gonzalo, and Jorge Vásquez. 2023. "Misinformation in Social Media: The Role of Verification Incentives."

[26] B takes the value from 0 to 100 (not 0-1 scale). Therefore, we use $\log(B/(100-B))$.

Danenberg, Tuval, and Drew Fudenberg. 2024. "Endogenous Attention and the Spread of False News." *arXiv preprint arXiv:2406.11024*.

Deolankar, Varad, Jessica Fong, and S Sriram. 2025. "The Effect of Downvotes on Content Creation: Evidence from Social Media." *Available at SSRN 4522092*.

Drobner, Christoph, and Sebastian J Goerg. 2024. "Motivated belief updating and rationalization of information." *Management Science*, 70(7): 4583–4592.

Eckles, Dean, René F Kizilcec, and Eytan Bakshy. 2016. "Estimating peer effects in networks with peer encouragement designs." *Proceedings of the National Academy of Sciences*, 113(27): 7316–7322.

Feess, Eberhard, Peter J Jost, and Anna Ressi. 2024. "Fake News and the Problem of Disregarding True Messages: Theory and Experimental Evidence." *Available at SSRN 4810742*.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives*, 19: 25—-42.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." *Proceedings of the National Academy of Sciences*, 117(27): 15536–15545.

Hagenbach, Jeanne, and Charlotte Saucet. 2025. "Motivated skepticism." *Review of Economic Studies*, 92(3): 1882–1919.

Hameleers, Michael. 2022. "Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands." *Information, Communication & Society*, 25(1): 110–126.

Harrison, Glenn W, Jimmy Martínez-Correa, and J Todd Swarthout. 2013. "Inducing risk neutral preferences with binary lotteries: A reconsideration." *Journal of Economic Behavior & Organization*, 94: 145–159.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92: 1644–1655.

Hossain, Tanjim, and Ryo Okui. 2013. "The binarized scoring rule." *Review of Economic Studies*, 80(3): 984–1001.

Jerit, Jennifer, and Yangzi Zhao. 2020. "Political misinformation." *Annual Review of Political Science*, 23(1): 77–94.

Kranton, Rachel, and David McAdams. 2024. "Social connectedness and information markets." *American Economic Journal: Microeconomics*, 16(1): 33–62.

Li, Jiexun, and Xiaohui Chang. 2023. "Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media." *Information systems frontiers*, 25(4): 1479–1493.

Lyons, Benjamin A, Jacob M Montgomery, Andrew M Guess, Brendan Nyhan, and Jason Reifler. 2021. "Overconfidence in news judgments is associated with false news susceptibility." *Proceedings of the National Academy of Sciences*, 118(23): e2019527118.

Meiske, Biljana, Amalia Álvarez-Benjumea, Giulia Andrighetto, and Eugenia Polizzi. 2024. "Nudging punishment against sharing of fake news." *European Economic Review*, 168: 104795.

Milgrom, Paul R. 1981. "Good news and bad news: Representation theorems and applications." *The Bell Journal of Economics*, 380–391.

Ministry of Internal Affairs and Communications (Japan). 2025. "Survey on ICT Literacy." *https://www.soumu.go.jp/main_content/001008791.pdf*, Accessed: 2025-10-15.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2022. "Managing self-confidence: Theory and experimental evidence." *Management Science*, 68(11): 7793–7817.

Nyhan, Brendan. 2020. "Facts and myths about misperceptions." *Journal of Economic Perspectives*, 34(3): 220–36.

Odabaş, Meltem. 2022. "5 facts about Twitter 'lurkers'." Pew Research Center. Accessed: 2025-11-14.

Oprea, Ryan, and Sevgi Yuksel. 2022. "Social exchange of motivated beliefs." *Journal of the European Economic Association*, 20(2): 667–699.

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. "Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter." *American Political Science Review*, 115(3): 999–1015.

Papanastasiou, Yiangos. 2020. "Fake news propagation and detection: A sequential model." *Management Science*, 66(5): 1826–1846.

Pennycook, Gordon, and David G Rand. 2020. "Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking." *Journal of Personality*, 88(2): 185–200.

Pennycook, Gordon, and David G Rand. 2021. "The psychology of fake news." *Trends in Cognitive Sciences*, 25(5): 388–402.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. "Shifting attention to accuracy can reduce misinformation online." *Nature*, 592(7855): 590–595.

Pfänder, Jan, and Sacha Altay. 2025. "Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements." *Nature Human Behaviour*, 1–12.

Phillips, Lawrence D, and Ward Edwards. 1966. "Conservatism in a simple probability inference task." *Journal of Experimental Psychology*, 72(3): 346.

Serra-Garcia, Marta, and Uri Gneezy. 2021. "Mistakes, overconfidence, and the effect of sharing on detecting lies." *American Economic Review*, 111(10): 3160–3183.

Sisak, Dana, and Philipp Denter. 2024. "Information Sharing with Social Image Concerns and the Spread of Fake News." *arXiv preprint arXiv:2410.19557*.

Thaler, Michael. 2021. "The supply of motivated beliefs." *arXiv preprint arXiv:2111.06062*, 385.

Thaler, Michael. 2024. "The fake news effect: Experimentally identifying motivated reasoning using trust in news." *American Economic Journal: Microeconomics*, 16(2): 1–38.

Van Der Linden, Sander. 2022. "Misinformation: susceptibility, spread, and interventions to immunize the public." *Nature Medicine*, 28(3): 460–467.

Zimmermann, Florian. 2020. "The dynamics of motivated beliefs." *American Economic Review*, 110(2): 337–363.

Online Appendix for "Fighting Fake News with Peer Feedback: Theory and Evidence" (Not for Publication)

Contents

B Additional Discussion on Theory

In the numerical example used both for deriving the theoretical predictions and for implementing the experiment, we obtained $\bar{\Delta}_A - \bar{\Delta}_R = 1$; in this sense, the difference was minimal.

One might wonder whether this result is specific to the chosen numerical values. To examine this, Figure A.1 plots the value of $\bar{\Delta}_A - \bar{\Delta}_R$ across a range of parameter values. The gray region in the figure indicates cases where Assumption 1 fails to hold; specifically, where $\bar{q}_R < 0$.

A large difference between $\bar{\Delta}_A - \bar{\Delta}_R$ arises when (α, p) lies near the boundary of the gray region, where $\bar{q}_R$ is close to zero. In such cases, senders would share news under reputational motivation even if they believe they are type-h with a probability of only about 0.01, which is unrealistic. In more realistic cases, where $\bar{q}_R$ is not extremely small, $\bar{\Delta}_A - \bar{\Delta}_R$ tends to remain small.

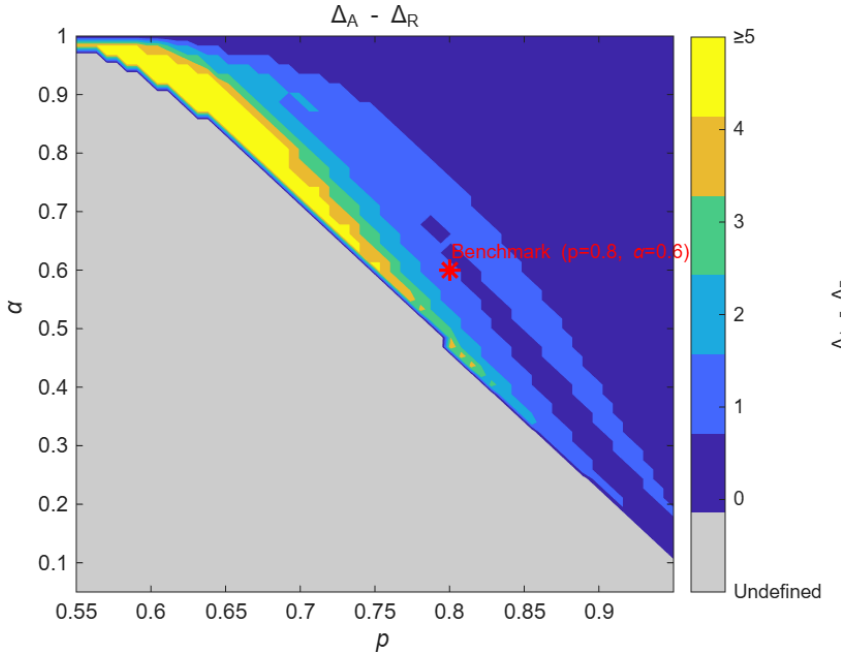

Figure A.1: $\bar{\Delta}_A - \bar{\Delta}_R$ $(q = 0.6)$

C Details of the Implementation of Experiments

The dates of the experimental sessions and the number of participants in each session are reported in Table A.1. Note that one student participated in two sessions (May 28, $A20$, and June 4, A20). We excluded this participant's data, as well as that of their partner, from the second session (June 4, $A20$). Consequently, the final sample consists of 295 participants, although the total number of participants listed in the table is 296.

Table A.1: Dates of Experiments

Date	**Condition**	**Number of participants**	**Number of sessions**
January 20, 2025	$A10$	22	4
January 22, 2025	$R10$	14	4
January 22, 2025	$A10$	14	4
January 24, 2025	$A10$	16	4
January 24, 2025	$R10$	14	4
May 28, 2025	$R20$	26	2
May 28, 2025	$A20$	22	2
June 2, 2025	$A20$	26	2
June 2, 2025	$R20$	26	2
June 4, 2025	$R20$	22	2
June 4, 2025	$A20$	22	2
June 25, 2025	$R10$	22	4
June 25, 2025	$A10$	26	4
July 2, 2025	$R10$	24	4

D Instructions

Thank you very much for participating in today's experiment. During the session, please do not communicate with other participants. You may take notes if you wish. At the end of the experiment, you will receive your payment in cash. All participants will receive a participation fee of 2,000 yen. In addition, you will earn extra rewards based on the decisions you make in the following tasks.

This experiment has [$A10$ and $R10$: four sessions] [$A20$ and $R20$: two sessions], each consisting of [$A10$ and $R10$: 10 rounds] [$A20$ and $R20$: 20 rounds]. In every round, you will decide together with another participant according to the rules explained below. Your partner will be randomly matched at the beginning of each round so that you may have a different partner each time. You will not be told who you are paired with. In each pair, one person will act as the Sender, and the other as the Receiver. Your role will remain the same throughout all four sessions—if you start as a Sender, you will stay a Sender until the end of the experiment; the same applies if you start as a Receiver.

The decision task in this experiment is as follows. As shown in the figure, there are two urns: a black urn and a white urn. Each urn contains five balls, which can be either black or white. In the black urn, there are more black balls than white balls, while in the white urn, there are more white balls than black balls. Specifically:

- The black urn contains four black balls and one white ball.
- The white urn contains four white balls and one black ball.

Black Urn
Four black balls, one white ball

White Urn
Four white balls, one black ball

At the beginning of each round, one of the two urns will be randomly selected. You will not know which urn has been chosen. Each urn is selected with an equal probability of 50%. Please note that the urn is chosen at the start of every round, not at the beginning of each session. Since each session consists of [$A10$ and $R10$: 10 rounds] [$A20$ and $R20$: 20 rounds], the urn will be selected [$A10$ and $R10$: 10 times] [$A20$ and $R20$: 20 times]. After the urn is selected in a round, both the Sender and the Receiver will make their decisions.

Sender's Decision

In each round, the Sender makes the first move. A ball is drawn from the urn selected at the beginning of the round, and the Sender can see the color of that ball (either black or white). The

Receiver does not see the color of the ball drawn. After observing the ball's color, the Sender can send a message to the Receiver about which urn they believe was selected. However, the Sender may only send a message that matches the color of the ball they saw. Specifically, the Sender has two options:

1. Send a message predicting that the urn matching the color of the drawn ball (black or white) was chosen.

2. Send no message about which urn was selected.

In other words, if the Sender observes a black ball, they may send the message "the black urn," but not "the white urn." Likewise, if they draw a white ball, they may send "the white urn was selected," but not the opposite. Thus, the Sender must choose either to send a message consistent with the ball's color or to send no message at all.

However, the ball shown to the Sender is not always drawn from the actual urn selected at the beginning of the round. With a 60% probability, the Sender is a "Correct-urn type", meaning they see a ball drawn from the actual urn chosen for that round. With the remaining 40% probability, the Sender becomes an "Incorrect-urn type," in which case the ball they see is drawn from the other urn (the one not selected for that round).

Whether a Sender is "Correct-urn type" or "Incorrect-urn type" is determined once at the beginning of each session and remains fixed for all [$A10$ and $R10$: 10 rounds] [$A20$ and $R20$: 20 rounds] within that session. For example, suppose the black urn is chosen in Round 1 and the white urn in Round 2. A Correct-urn type will see a ball from the black urn in Round 1 and from the white urn in Round 2. In contrast, an Incorrect-urn type will see a ball from the white urn in Round 1 and from the black urn in Round 2.

In a different session, the Sender's type (correct-urn or Incorrect-urn) may change, as it is re-randomized at the start of each session. Neither the Sender nor the Receiver knows whether the ball shown to the Sender was drawn from the correct urn or not. **After each draw, the ball is returned to the urn,** so the total composition of each urn remains the same throughout the experiment.

Receiver's Decision

In each round, the Receiver makes the second decision. First, the Receiver learns whether the Sender sent a message; if so, the Receiver learns its content. After that, a ball is drawn from an urn in front of the Receiver, and the Receiver can see the color of that ball. The Sender does not know the color of the ball drawn in front of the Receiver. If the Sender did send a message, the Receiver must choose one of the following two options:

1. Evaluate the Sender's message positively ("like")

2. Evaluate the Sender's message negatively ("dislike")

If the Sender did not send a message, the Receiver instead must predict which urn was selected by choosing one of the following two options:

1. The black urn is the correct urn.

2. The white urn is the correct urn.

The ball shown to the Receiver is always drawn from the correct urn selected at the beginning of the round. It is never drawn from the Incorrect urn.

End of Each Round

At the end of each round, the Sender learns the Receiver's choice, regardless of whether a message was sent or not. This means that if the Sender sent a message, they will learn how the Receiver evaluated it. If the Sender did not send a message, they will learn the Receiver's prediction about which urn was selected. However, the actual urn chosen in that round is not revealed to either participant. Afterward, the Sender will be asked to report their belief about the probability that they are a "Correct-urn type." On the screen, the belief reported in the previous round will be displayed for reference. Since the Receiver's behavior in the new round may affect this belief, please update it to reflect your new assessment. In the first round, the displayed belief starts at 60%. Please enter an integer between 0 and 100, representing your subjective probability (in percent) that you are a Correct-urn type. An additional bonus payment will be determined based on this belief report. The reward system is designed so that you are better off reporting your true belief about the probability that you are a Correct-urn type. (Details of the bonus calculation will be explained later.)

The decisions described above for both the Sender and the Receiver constitute one round. This round is repeated 10 times within each session.

Probability of Earning an Experimental Reward

The probability of earning an experimental reward in addition to the participation fee (the reward probability) depends on the outcome of the decisions made during the round. The calculation differs between Senders and Receivers.

[*A*10 and *A*20: **Sender's Reward Probability**

If they send a message and their prediction turns out to be correct, the probability of earning an experimental reward is 80%. If the prediction is wrong, the probability is 16%. The Receiver's evaluation—whether the message is rated "like" or "dislike"—does not affect this probability. Even if the message is rated negatively, a correct prediction still results in an 80% chance of earning the reward, while an incorrect one remains at 16%. If the Sender chooses not to send a message, the probability of earning a reward is 40%.]

[*R*10 and *R*20: **Sender's Reward Probability**

If the Sender sends a message and the Receiver evaluates it positively ("like"), the probability of earning an experimental reward is 80%. If the Receiver evaluates it negatively ("dislike"), the probability

is 16%. Whether the Sender's prediction is correct or not does not matter. Even if the prediction is wrong, a positive evaluation still gives an 80% chance of earning the reward; conversely, even if the prediction is correct, a negative evaluation gives only a 16% chance. If the Sender chooses not to send a message, the probability of earning a reward is 40%.]

Receiver's Reward Probability

The Receiver's probability of earning an experimental reward depends on whether their choice (the evaluation of the Sender or the prediction of the correct urn) is appropriate or not. If the Sender sends a message, the Receiver's bonus probability is determined as follows:

1. When the Sender's prediction is correct:
    - If the Receiver chooses "like," the probability of earning a reward is 80%.
    - If the Receiver chooses "dislike," the probability is 20%.
2. When the Sender's prediction is incorrect:
    - If the Receiver chooses "like," the probability of earning a reward is 20%.
    - If the Receiver chooses "dislike," the probability is 80%.

If the Sender does not send a message, the Receiver's probability of earning a reward depends on the accuracy of their own prediction:

- If the Receiver's prediction is correct, the probability is 80%.
- If the prediction is incorrect, the probability is 20%.

The Receiver will not be informed of these probabilities at the end of each round. However, at the end of the session, the probabilities for all rounds will be displayed.

Payment

Your final payment will consist of three components: the participation fee, the experimental reward based on your reward probability in each round, and an additional bonus that applies only to Senders. Both Senders and Receivers are eligible for the experimental reward, but only Senders can receive the additional bonus. These rewards are determined through a lottery process, as explained below.

Among the [$A10$ and $R10$: 10 rounds] [$A20$ and $R20$: 20 rounds] in each session, [$A10$ and $R10$: one round] [$A20$ and $R20$: two rounds] will be randomly selected as the "payment round." [$A20$ and $R20$: One payment round was randomly selected from rounds 1 to 10, and another from rounds 11 to 20.] Since there are [$A10$ and $R10$: four sessions] [$R20$ and $R20$: sessions], there will be four payment rounds in total. Only the results from these selected payment rounds will be used to determine your final rewards; all other rounds are disregarded. You will not know in advance which round will be chosen as a payment round.

The reward probability assigned to you in a paying round determines your chance of earning 200 yen as the experimental reward. For example, if your reward probability in that round is 80%, you have an 80% chance of receiving 200 yen. To determine whether you receive the 200 yen, a lottery will be conducted. A random integer between 1 and 100 will be drawn. If the number drawn is less than or equal to your reward probability, you will receive 200 yen. If it is greater, you will not receive the payment. For instance, if your reward probability in the paying round is 16%, you will earn 200 yen if the number drawn is 16 or lower, but not if it is 17 or higher.

Additional Bonus for Reporting the Probability of Being a "Correct-urn type"

At the end of each round, the Sender reports their belief about the probability that they are a "Correct-urn type." The additional bonus is determined based on the Sender's response in the paying round, according to the lottery rule described below. After the type of lottery is determined, a draw is conducted, and depending on the result, the Sender may receive an additional bonus of 50 yen.

How the Lottery is Assigned

1. A random integer between 1 and 100 is selected.

2. Suppose the randomly selected number is greater than or equal to the probability reported by the Sender. In that case, the Sender receives a lottery that pays 50 yen with the same probability as the selected number. For example, if the Sender reports 56% and the randomly selected number is 74, they receive a lottery that pays 50 yen with a 74% chance.

3. If the randomly selected number is less than the probability reported by the Sender, the Sender receives a lottery that pays 50 yen only if they are a Correct-urn type.

Once the lottery type is determined, the lottery is conducted, and the Sender receives the additional bonus based on the result. In the case of the "lottery that pays 50 yen only if the Sender is actually a Correct-urn type," the payment is made only if the Sender truly is a Correct-urn type.

Under these rules, the best strategy for the Sender is to report honestly the probability they personally believe reflects their chance of being a Correct-urn type. Reporting a probability higher or lower than one's true belief does not increase the expected reward. If you wish to understand the reasoning behind this, please refer to the "Appendix" at the end of the instruction sheet.

Next Steps

At the end of the experiment, you will be asked to complete a short questionnaire. Before starting the experiment, you will take part in a comprehension test followed by a practice

session. Please use these to make sure you fully understand how the experiment works. Note that the comprehension test will not include any questions about the contents of the appendix.

Appendix

The instruction explained the rule for providing an additional bonus based on the Sender's reported probability of being a "Correct-urn type." The instruction also noted that the optimal strategy for maximizing one's expected reward is to report one's true belief about this probability. Here, this appendix explains why that is the case.

Let us consider the situation in which you, as the Sender, believe that there is a 40% chance that you are a Correct-urn type.

Suppose you report 40% as your belief. If the random number drawn between 1 and 100 is greater than or equal to 40, you will receive a lottery that pays the reward with that number's probability. For example, if the number drawn is 60, you receive a lottery that pays the reward with a 60% chance. In this case, your overall chance of receiving the reward is at least 40%. On the other hand, if the random number drawn is less than 40, you receive a lottery that pays only if you are a Correct-urn type. Since you personally believe that you are a Correct-urn type with a 40% probability, your chance of receiving the reward in this case is 40%.

Now suppose that, even though you believe the probability is 40%, you instead report a higher number—say, 70%. If the random number drawn between 1 and 100 is 70 or higher, you will receive a lottery that pays the reward with that number's probability. If the random number is below 70, you will receive a lottery that pays only if you are a Correct-urn type. As a result, when the random number falls between 41 and 69, your reported probability causes you to lose the advantage you would have had by answering honestly. In those cases, the chance of receiving a reward, which would have been greater than 40% if you had reported truthfully, is now reduced to 40%.

Finally, suppose that although you believe the probability is 40%, you report a lower number—say, 20%. If the random number drawn between 1 and 100 is 20 or higher, you receive a lottery that pays the reward with that number's probability. If the random number is below 20, you receive a lottery that pays only if you are a Correct-urn type. In this case, when the random number falls between 20 and 39, your chance of receiving a reward—which would have been 40% had you reported truthfully—drops to less than 40%.

The same reasoning applies to any other belief, not just 40%. Therefore, it is always in your best interest to report your true belief about the probability that you are a Correct-urn type.

E Comprehension Test

1. Mark T if the statement is true, and F if it is false.

   - The roles of Sender and Receiver do not change across sessions.
   - The correct urn is the same in every round.
   - You are matched with the same partner in every round.
   - Whether the Sender is a "Correct-urn type" does not change within a session, but may change across sessions.
   - The Receiver is always a "Correct-urn type."
   - The Sender is better off reporting a probability higher than their true belief about being a "Correct-urn type."

2. If the Sender sends a message, what affects the Sender's "probability of earning an experimental reward"? Circle one:

   A. The correctness of the message

   B. The Receiver's evaluation

3. The black urn is selected, and the Sender observes a black ball. What message can the Sender send? Circle one:

   A. "The black urn"

   B. "The white urn"

   C. "white urn"

4. The black urn is selected, the Sender observes a black ball, and the Sender sends a message. The Receiver gives the message a "dislike." What are the reward probabilities? Sender: __ % Receiver: __%

F Post-Experiment Survey

In relation to the experiment you just participated in, we would like to ask you a few additional questions. Your participation in this survey will not cause you any disadvantage, so please feel free to answer honestly.

All responses will be processed statistically, and no information that could identify individual participants will be disclosed. After the completion of this experiment, your responses and behavioral data will be stored separately from any personal information necessary for payment. No individual will be identifiable in any data analysis or publication of the results.

We kindly ask for your cooperation in answering all questions carefully.

Q1 What is your gender?

1. Male
2. Female
3. Other

Q2 What is your age? Please enter in half-width (single-byte) numbers below.

Q3 There are two lotteries: Lottery A, which gives either 200 yen or 160 yen, and Lottery B, which gives either 385 yen or 10 yen. The probabilities of each outcome differ across the following cases. For each case, please indicate which lottery you prefer, A or B.

On the one hand, Lottery A guarantees 160 yen, while Lottery B guarantees only 10 yen, so you would likely prefer A. On the other hand, Lottery A guarantees 200 yen and Lottery B guarantees 385 yen, so you would likely prefer B. There should be a point between these two extreme cases where your preference switches from A to B. Please answer while being aware of that switching point.

1. (A) 10% chance of 200 yen, 90% chance of 160 yen vs. (B) 10% chance of 385 yen, 90% chance of 10 yen
2. (A) 20% chance of 200 yen, 80% chance of 160 yen vs. (B) 20% chance of 385 yen, 80% chance of 10 yen
3. (A) 30% chance of 200 yen, 70% chance of 160 yen vs. (B) 30% chance of 385 yen, 70% chance of 10 yen
4. (A) 40% chance of 200 yen, 60% chance of 160 yen vs. (B) 40% chance of 385 yen, 60% chance of 10 yen
5. (A) 50% chance of 200 yen, 50% chance of 160 yen vs. (B) 50% chance of 385 yen, 50% chance of 10 yen

6. (A) 60% chance of 200 yen, 40% chance of 160 yen vs. (B) 60% chance of 385 yen, 40% chance of 10 yen
7. (A) 70% chance of 200 yen, 30% chance of 160 yen vs. (B) 70% chance of 385 yen, 30% chance of 10 yen
8. (A) 80% chance of 200 yen, 20% chance of 160 yen vs. (B) 80% chance of 385 yen, 20% chance of 10 yen
9. (A) 90% chance of 200 yen, 10% chance of 160 yen vs. (B) 90% chance of 385 yen, 10% chance of 10 yen
10. (A) 100% chance of 200 yen, 0% chance of 160 yen vs. (B) 100% chance of 385 yen, 0% chance of 10 yen

Q4 A pond is partially covered with water lilies, and the area covered doubles every day. If it takes 6 days for the entire pond to be completely covered, how many days will it take for the pond to be half covered? Please choose one answer below.

1. 1 day
2. 2 days
3. 3 days
4. 4 days
5. 5 days
6. 6 days
7. I don't know

Q5 Five machines take 5 minutes to make 5 parts. How many minutes will it take 100 machines to make 100 parts? Please choose one answer below.

1. 1 minute
2. 5 minutes
3. 10 minutes
4. 50 minutes
5. 100 minutes
6. 500 minutes
7. I don't know

Q6 A bat and a ball together cost 110 yen. The bat costs 100 yen more than the ball. How much does the ball cost? Please choose one answer below.

1. 1 yen
2. 5 yen
3. 10 yen
4. 50 yen
5. 100 yen
6. 105 yen
7. I don't know

Q7 There is a proverb, "You cannot catch a tiger's cub without entering its den," meaning that one must take risks to achieve great results. On the other hand, another proverb says, "A wise man keeps away from danger," meaning one should avoid risks as much as possible.

Which of these sayings better describes your usual behavior? Please rate your usual behavior on a scale from 0 to 10, where 10 represents complete agreement with the "tiger's den" idea, and 0 represents complete agreement with the "wise man avoids danger" idea.

Q8 Now we would like to ask your impressions of the experiment. Did you understand the explanation of the experiment?

1. Fully understood
2. Mostly understood
3. Neither
4. Did not really understand
5. Did not understand at all

Q9 In the experiment screen, did you understand how to make your choices?

1. Fully understood
2. Mostly understood
3. Neither
4. Did not really understand
5. Did not understand at all

Q10 Was there anything unclear or anything you noticed in the explanation of the experiment? Please write freely below.

Q11 Did anything make you feel doubtful or confused during the experiment? Please write freely below.

Q12 Overall, what kind of strategy did you use throughout the experiment? Please write freely below.