The Effect of Feedback on News Verification Demand: Experimental Evidence* (Preliminary and Incomplete)

By Dario Trujano-Ochoa† and Jose Gloria‡

This study explores the decision-making processes involved in classifying information in the context of headlines that may be accurate or false. The experiment investigates how different types of feedback influence the willingness to pay for headline verification. Three treatment conditions are examined: no feedback, feedback on own performance, and feedback on group classification performance. The findings provide empirical insights into the dynamics of information spread, confidence, and the value of information. These have implications for understanding and mitigating the effects of misinformation in political contexts.
JEL: C93, D83, D91

Misinformation poses a serious threat to democratic societies by undermining the public's ability to make informed decisions, eroding trust in institutions, and fueling polarization. The rapid spread of misinformation through digital platforms has amplified its impact, making it a global issue. Also, these problems are exacerbated when false information is part of disinformation campaigns. Societies and individuals must fight misinformation, and promoting behaviors that prevent the spread of false information is essential, just as we promote cyber-security and prevent the spread of viruses and scams. If no efforts are spent fighting misinformation, societies risk making decisions based on wrong assumptions usually by actors who only look after their personal benefit. This research explores the demand for verification as a tool since the probability of believing and sharing misinformation will be larger than zero as long as there are incentives to spread false information. It is also important to regulate the platforms and increase the attention and ability to recognize fake news. However, verification will be part of the solution, just as antiviruses are necessary to fight cybercrime.

This study focuses on how feedback can influence individuals' engagement with fact-checking and critical assessment of misleading content. By investigating feedback mechanisms, the research aims to contribute to strategies that enhance pub-

* The authors want to thank all the professors and specialists who heard the early versions of the present project and helped to increase the discussion and understanding of the relevant problem of misinformation: Gary Charness, Cesi Cruz, Daniel Martin, Ignacio Esponda, Ryan Oprea, Sevgi Yuksel, Erik Eyster, Grisel Salazar, Daniel Moreno, Horacio Larreguy, Antonio Arechar, Pablo Soto, and Arturo Bouzas. We also thank AlianzaMX for the fellowship that allowed the author to travel to Mexico to develop this research.

† UCSB, USA, dariotrujanoochoa@ucsb.edu.
‡ UCLA, USA, josegloria@ucla.edu.

lic resilience against misinformation, fostering a more informed and healthier democratic dialogue. At the core of this research is an innovative exploration of how feedback mechanisms can improve fact-checking and critical thinking about misinformation. The study shifts from traditional analyses of misinformation sources to a deeper understanding of how people change the demand for information based on feedback. This approach is particularly relevant in political contexts, such as elections, where misinformation campaigns can sway public opinion. The experiment tests the effects of feedback and overconfidence on decision-making, offering valuable insights into the cognitive processes involved in information verification.

I. Previous Literature

It has been stated before that people's overconfidence has an impact on their ability to discern fake news, which can lead to greater engagement with false information. According to Lyons et al. (2021), individuals tend to overestimate their ability to distinguish between real and fake news. In a large-scale study, researchers found that overconfident individuals are not only more likely to believe in and share fake news but are also unaware of their limitations in identifying misinformation. This overconfidence results in behaviors that exacerbate the spread of false information on social media platforms, particularly when the misinformation aligns with their political or ideological beliefs. Similarly, Pennycook and Rand (2020) confirmed that overconfidence in one's own cognitive abilities correlates with a higher likelihood of accepting false claims as true. These studies highlight the role that overconfidence can play in fake news dissemination.

Several interventions have been designed to counter the spread of misinformation by enhancing individuals' attention to accuracy when sharing information online. Pennycook et al. (2021) introduced an intervention aimed at encouraging social media users to pause and reflect on the accuracy of the content they share. By nudging users to focus on accuracy, the intervention was successful in reducing the spread of fake news. This "accuracy nudge" has shown promising results in improving online behavior, encouraging users to prioritize truthfulness over impulsive sharing. Furthermore, Pennycook and Rand (2022) explored interventions aimed at fostering deliberate cognitive engagement, which also showed success in reducing the dissemination of false information. These interventions stress that slight adjustments in users' thought processes can greatly influence the quality of information shared on social media platforms.

Another effective approach in combating misinformation involves encouraging behaviors related to fact-checking and verification. Kozyreva et al. (2024) explored strategies promoting media literacy and verification as vital tools for countering the effects of misinformation. Their research underlines the importance of critical thinking and fact-checking in reducing susceptibility to fake news. However, the use of verification tools, particularly search engines, can backfire under certain conditions. Aslett et al. (2024) argue that while search engines are often

promoted as tools for fact-checking, they can sometimes amplify misinformation, especially when individuals are overconfident in their search abilities. The authors found that some users interpret the results to confirm pre-existing biases, reinforcing false beliefs rather than correcting them. Therefore, while promoting verification behaviors is crucial, it is important to consider the limitations of certain tools, such as search engines, in effectively curbing misinformation.

The literature has mentioned that overconfidence in one's ability to discern fake news plays a significant role in the engagement and spread of misinformation. Interventions aimed at improving attention to accuracy have shown promise in reducing the spread of fake news. At the same time, while promoting verification behaviors is essential, the use of tools like search engines must be carefully managed to avoid reinforcing biases and spreading misinformation. Together, these findings emphasize the need for a multi-faceted approach to reducing the harmful effects of fake news on social media.

This research tries to establish if feedback has an effect on the demand for verification. Other papers have focused on measuring the accuracy of distinguishing misinformation or the effects of verification strategies but not the demand for fact-checking mechanisms. This paper's main contribution is to present experimental evidence of the causal effect of a feedback intervention on the demand for verification. No other paper has analyzed the

II. Hypotheses on the Effects of Feedback

This study is structured around three key hypotheses that explore the impact of feedback on willingness to pay (WTP) to verify the accuracy of the headlines and the influence of overconfidence in decision-making. These hypotheses are integrated into the methodology to test their validity in a controlled environment, using neutral headlines to minimize the effects of motivated reasoning.

HYPOTHESIS 1: *Participants are generally overconfident and will expect better performance in classifying headlines than what is reflected in the feedback they receive.*

Following the literature on overconfidence, this hypothesis suggests that participants will tend to overestimate their classification accuracy before receiving any feedback. In this experiment, participants' predictions about their classification accuracy are expected to be higher than the accuracy indicated by their actual performance, as revealed by the feedback.

HYPOTHESIS 2: *Participants' willingness to pay (WTP) for verification will be higher when they receive feedback on group classification accuracy compared to personal performance feedback.*

According to previous results on the asymmetric effect of feedback, participants who receive feedback on the performance of others will perceive this feedback as

more informative and objective. As a result, they will place greater value on the verification process and be willing to pay more to ensure the accuracy of the headlines they classify. The expectation is that group feedback, being less affected by motivated reasoning, will lead to a higher WTP as participants seek to mitigate the perceived difficulty of the task.

HYPOTHESIS 3: *Participants' willingness to pay (WTP) for verification is influenced by the political content of the headlines, with differing effects depending on whether the headline favors or opposes the current government.*

1) *Supporters of the Government: Lower WTP for verification of favorable headlines; higher WTP for verification of unfavorable headlines.*

2) *Opponents of the Government: Higher WTP for verification of favorable headlines; lower WTP for verification of unfavorable headlines.*

When participants are presented with headlines that contain political content, it is hypothesized that their WTP for verification will vary depending on whether the headline aligns with their political beliefs. Specifically, if a headline is favorable to the current government, participants who support the government are likely to have lower WTP for verification. This is because they are more inclined to accept information that aligns with their pre-existing beliefs without seeking further verification. Conversely, participants who oppose the current government may exhibit higher WTP for verification of favorable headlines, as they may be more skeptical of information that contradicts their beliefs and, thus, more motivated to confirm its accuracy.

On the other hand, for headlines that are unfavorable to the current government, supporters of the government may demonstrate higher WTP for verification, driven by a desire to challenge or disprove information that opposes their political views. Opponents of the government, however, may show lower WTP for verification of unfavorable headlines, as they may be more likely to accept information that aligns with their negative views of the government without the need for additional confirmation.

III. Decision-Making in the Classification of Headline Accuracy

This section describes the agent's decision-making problem involved in classifying, verifying, and reclassifying a headline as accurate or fake. Classifying is a signal detection problem, and purchasing additional signals on this decision requires calculating the expected value of sample information (EVSI).

A. Problem Setup Without Purchasing a Signal

Consider an agent tasked with classifying headlines as accurate (a) or fake (f) ($c \in \{a, f\}$). The state of the world is $\omega \in \Omega = \{A, F\}$, with the prior probability

of encountering a fake headline denoted by $P(\omega = F) = p_f$.[1] Consequently, the prior probability of encountering an accurate headline is $1 - p_f$.

The agent's utility for correctly classifying a headline as accurate is U_A, and for correctly classifying a headline as fake is U_F. Conversely, misclassifying a fake headline as accurate results in a utility of $U_{AF} < U_A$, and misclassifying an accurate headline as fake results in a utility of $U_{FA} < U_F$. The condition is case-insensitive for evaluating the correct classification (i.e., $c = \omega$ means a correct classification).

The probability of classifying correctly the headline is determined by $P(c = a|A)$ and $P(c = f|F)$ with $1 < \frac{P(c=a|A)}{P(c=a|F)}$ and $1 < \frac{P(c=f|F)}{P(c=f|A)}$ to assure that the initial classification c is informative in the sense that the initial classification c gives information relative to the prior probability of each state ω.[2] This is stated formally in the following proposition.

PROPOSITION 1: *Informativeness of the initial classification c.*
$1 < \frac{P(c=\omega|\omega)}{P(c=\omega|\tilde{\omega}\neq\omega)}$ *if and only if* $P(\omega|c \neq \omega) < P(\omega) < P(\omega|c = \omega)$

Notice that commonly found assumptions $0.5 < P(s = a|A) = q_a$ and $0.5 < P(s = f|F) = q_f$ are sufficient to make the signal S informative according to proposition 1. The proof of this proposition can be found in the appendices.

The expected utilities when a headline is classified as accurate, $EU_{\text{no signal}}(a)$, and as fake, $EU_{\text{no signal}}(f)$, are given by the equations:

$$EU_{\text{no signal}}(a) = P(A|a) \cdot U_A + P(F|a) \cdot U_{AF}$$

$$EU_{\text{no signal}}(f) = P(F|f) \cdot U_F + P(A|f) \cdot U_{FA}$$

From the previous equations, it is clear that if the classification of the headline is informative, reading a headline is valuable in the sense that the expected utility is larger than just considering the prior probabilities.

B. Conditional WTP Analysis

In this section, we will show the optimal willingness to pay (WTP) for signal S after the initial classification. The WTP is the maximum amount that an agent would pay to observe signal S. This is equal to the concept of the expected value of the sample information (EVSI).

The decision to purchase information happens after observing a headline once the agent has classified the signal. Therefore, the value of the signal depends on c.

[1] We simplify $P(\omega = F)$ to $P(F)$. For $c, s \in \Omega$, we specify.

[2] We are assuming here that the classification is a signal to the same agent without considering the content of a headline h which is most likely multidimensional. This classification process also follows an optimization process where $c = \omega \iff \frac{P(\omega|h)}{P(\omega|\tilde{\omega}\neq\omega|h)} > \frac{U_A - U_{FA}}{U_F - U_{AF}}$. However, following the objectives of the present research, we focus on analyzing the informativeness of the initial classification c in proposition 1 without analyzing the properties of the headlines or the payoffs.

We are assuming that sequential information acquisition is optimal. The problem of sequential decision-making was stated in general by Wald (1947), and Arrow et al. (1949) analyzed how to learn from sequential information.

This section presents the condition that makes verifying the initial classification valuable. After initially classifying the headline as accurate ($c = a$) or false ($c = f$), the agent can reclassify the headline $r \in \{a, f\}$ based on the signal's realization $s \in \{a, f\}$. Let's consider first a valuable signal S with the conditions in proposition 2.

DEFINITION 1: *A signal is valuable if* $EU(r = s) \geq EU(r = c)$

The informativeness of the signal is determined by $P(s = a|A) = q_a$. We need a strong enough signal S so that the signal is valuable and the optimal decision to reclassify is to follow the signal ($r = s \in \{a, f\}$). Also, we assume that the signal realization s is independent of the previous classification c conditional on the stare of the world $\omega \in \{A, F\}$ (i.e. $P(s|\omega, c) = P(s|\omega)$).

PROPOSITION 2: *Conditions for Valuable Signal*
A signal S is valuable if and only if

$$\frac{P(\omega|s = \omega, c \neq \omega)}{P(\tilde{\omega}|s = \omega, c \neq \omega)} < \frac{U_\omega - U_{\tilde{\omega}\omega}}{U_{\tilde{\omega}} - U_{\omega\tilde{\omega}}} \equiv U_\omega$$

with $\tilde{\omega} \in \Omega, \tilde{\omega} \neq \omega$.

We are also assuming that the initial classification is valuable and therefore follow the analogous condition $\frac{P(\omega|c=\omega)}{P(\tilde{\omega}|c=\omega)} < U_\omega$.

By allowing the agent to update their classification based on the signal, we account for the dynamic decision-making process. The WTP to verify the classification is derived by comparing the expected utility with the signal (considering reclassification) to the expected utility without the signal. This approach shows the impact of additional information on improving decision-making accuracy. The detailed mathematical steps and proofs are provided in the appendix. The expected utility of reclassification is calculated by updating the agent's posterior beliefs using Bayes' rule and comparing the expected utilities with and without reclassification.

For an agent tasked with classifying headlines as accurate or fake, a signal S indicating the state of the world must be sufficiently strong to ensure that the agent reclassifies based on this signal.

WTP Equation

Initially, the agent classifies a headline as either accurate (a) or fake (f). Upon receiving a signal s, which can either confirm or contradict the initial classification, the agent updates their beliefs. The posterior probabilities are calculated using

Bayes' rule. For example, the posterior probability of the headline being accurate given the signal $s = a$ and the initial classification c is:

$$P(A|s = a, c) = \frac{q_a \cdot P(A|c)}{q_a \cdot P(A|c) + (1 - q_f) \cdot P(F|c)}$$

Similarly, the posterior probability of the headline being fake given the signal $s = f$ and the initial classification c is:

$$P(F|s = f, c) = \frac{q_f \cdot P(F|c)}{(1 - q_a) \cdot P(A|c) + q_f \cdot P(F|c)}$$

The agent's decision to reclassify based on the signal depends on the expected utilities. The expected utility of reclassification given the signal $s = a$, or $s = f$, are respectively:

$$EU_{\text{new classification}}(s = a, c) = P(A|s = a, c) \cdot U_A + P(F|s = a, c) \cdot U_{AF}$$

$$EU_{\text{new classification}}(s = f, c) = P(F|s = f, c) \cdot U_F + P(A|s = f, c) \cdot U_{FA}$$

The combined expected utility of updating the signal, considering both possible signals, is:

$$\begin{aligned} EU^{\text{update}}_{\text{signal}}(c) = \quad & P(s = a|c) \cdot EU_{\text{new classification}}(s = a, c) + \\ & P(s = f|c) \cdot EU_{\text{new classification}}(s = f, c) \\ = \quad & [q_a \cdot P(A|c) + (1 - q_f) \cdot P(F|c)] \cdot [P(A|s = a, c) \cdot U_A + P(F|s = a, c) \cdot U_{AF}] + \\ & [(1 - q_a) \cdot P(A|c) + q_f \cdot P(F|c)] \cdot [P(F|s = f, c) \cdot U_F + P(A|s = f, c) \cdot U_{FA}] \end{aligned}$$

The WTP to verify the headline is calculated by comparing the expected utility with the signal to the expected utility without the signal:

$$V(c) = EU^{\text{update}}_{\text{signal}}(c) - EU_{\text{no signal}}(c)$$

C. Simplifying Assumptions

Let's assume equal prior probabilities $p_f = 0.5$ and equal utilities $U_A = U_F = 1$ and $U_{AF} = U_{FA} = 0$. Also, assume that the prevalence of fake and accurate news is the same $P(A) = P(F) = p_f = 0.5$. These assumptions on the payoffs allow us to interpret the value of the signal purely in probability terms related to the informativeness of the signal.

Substituting these assumptions into the expected utility equations, we get:

$$EU_{\text{no signal}}(a) = P(A|a) \cdot 1 + P(F|a) \cdot 0 = P(A|a) = \frac{P(a|A)}{P(a|A) + P(a|F)}$$

$$EU_{\text{no signal}}(f) = P(F|f) \cdot 1 + P(A|f) \cdot 0 = P(F|f) = \frac{P(f|F)}{P(f|A) + P(f|F)}$$

And the expected utilities simplify to:

$$EU_{\text{new classification}}(s = a, c) = P(A|s = a, c)$$
$$EU_{\text{new classification}}(s = f, c) = P(F|s = f, c)$$

Finally, the combined expected utility of updating the signal, considering both possible signals, is:

$$\begin{aligned} EU^{\text{update}}_{\text{signal}}(c) = \quad & [q_a \cdot P(A|c) + (1 - q_f) \cdot P(F|c)] \cdot P(A|s = a, c) + \\ & [(1 - q_a) \cdot P(A|c) + q_f \cdot P(F|c)] \cdot P(F|s = f, c) \end{aligned}$$

Perfect Signal

Here, we calculate the WTP considering the condition $q_f = q_a = 1$; perfect signal. This assumption ensures that the signal is strong enough to follow even without the other simplifying assumptions, and simplifies substantially the interpretation of $EVSI(c)$. For the case $c = f$ and $s = a$: $q_a \cdot P(A|f) > (1-q_f) \cdot P(F|f) \iff P(A|f) > 0$. The case $c = s$ and $s = f$ requires $P(F|a) > 0$. Both conditions are satisfied by the construction of the problem.

This assumption simplifies the expected utility of observing the signal S. Thus, the combined expected utility with the signal is:

$$EU^{\text{update}}_{\text{signal}}(c) = P(A|c) \cdot 1 + P(F|c) \cdot 1 = P(A|c) + P(F|c) = 1$$

Therefore,

$$V(c) = \begin{cases} 1 - P(A|a), c = a \\ 1 - P(F|f), c = f \end{cases} \tag{1}$$

The value of the signal S is the difference between the posterior probability of reclassifying correctly after observing the signal and the posterior probability of initially classifying correctly.

Notice that if we change the payoff of a correct answer such that $U_A = U_F > U_{AF} = U_{FA}$, we only have to multiply the posterior probabilities difference by $\phi = U_A - U_{AF}$ to get $V(c)$.

IV. Mechanism Behind the Effect of Feedback

A. *Impact of Feedback on $P(A|c = a)$ Using Item Response Theory*

The probability $P(A|c = a)$ and $P(F|c = f)$ can be considered a subjective measure of an agent's probability to classify an accurate headline correctly. This

probability can be modeled using Rasch's model in the Item Response Theory (IRT), which considers the agent's ability and the headline's difficulty to be classified.

In the context of headline classification, IRT can be used to describe how different types of feedback affect participants' subjective probability of making a correct classification, $P(A|c = a)$.

In this context, classifying headlines as accurate ($c = a$) or fake ($c = f$) can be viewed as a binary decision-making task. According to IRT, the probability of a participant correctly classifying a headline depends on two key factors:

1) **Participant's Ability (θ)**: This represents the participant's ability to classify headlines accurately.
2) **Item (Headline) Difficulty (β)**: This represents the difficulty of correctly classifying a specific headline.

The IRT model uses a logistic function to describe the probability of a correct response (classifying a headline as accurate when it is indeed accurate) as:

$$P(A|c = a, \theta, \beta) = \frac{1}{1 + e^{-(\theta-\beta)}}$$

B. Impact of Feedback on $P(A|c = a)$

Different types of feedback can influence either the participant's subjective ability (θ) or their perception of the difficulty of classifying headlines (β).

Feedback on One's Performance

When participants receive feedback on their own performance, this feedback can influence their subjective ability (θ). For example:

- **Positive Feedback**: If participants consistently receive positive feedback (e.g., they are informed that their classifications are mostly correct), their subjective ability (θ) is likely to increase, raising their confidence in their ability to classify headlines accurately.
- **Negative Feedback**: Conversely, if participants receive negative feedback (e.g., they are informed that their classifications are mostly incorrect), their subjective ability (θ) may decrease, lowering their confidence in their ability to classify headlines accurately.

As a result, the probability of correctly classifying a headline, $P(A|c = a)$, increases with positive feedback (higher θ) and decreases with negative feedback (lower θ):

$$P(A|c = a, \theta_{\text{new}}, \beta) = \frac{1}{1 + e^{-(\theta_{\text{new}}-\beta)}}$$

where θ_{new} represents the updated ability after feedback.

Feedback on Others' Performance

When participants receive feedback on the performance of others, this feedback can influence their perception of the difficulty of the task (β). For example:

- **High Difficulty Feedback**: If participants are informed that others also find the task difficult (e.g., a high percentage of participants classify headlines incorrectly), they may perceive the headlines as more difficult, increasing the value of β.
- **Low Difficulty Feedback**: If participants are informed that others find the task easy (e.g., a high percentage of participants classify headlines correctly), they may perceive the headlines as less difficult, decreasing the value of β.

As a result, the probability of correctly classifying a headline, $P(A|c = a)$, decreases with higher perceived difficulty (higher β) and increases with lower perceived difficulty (lower β):

$$P(A|c = a, \theta, \beta_{\text{new}}) = \frac{1}{1 + e^{-(\theta - \beta_{\text{new}})}}$$

where β_{new} represents the updated difficulty after feedback.

C. *Summary*

- **Feedback on One's Performance**: This primarily affects the participant's subjective ability (θ), altering their confidence and hence the probability $P(A|c = a)$. Positive feedback increases θ and $P(A|c = a)$, while negative feedback decreases θ and $P(A|c = a)$.
- **Feedback on Others' Performance**: This primarily affects the participant's perception of the difficulty (β) of classifying headlines. Feedback indicating higher difficulty increases β and decreases $P(A|c = a)$, while feedback indicating lower difficulty decreases β and increases $P(A|c = a)$.

This IRT-based framework helps explain how feedback mechanisms can shape participants' subjective beliefs about their ability to classify headlines accurately and their perception of the task's difficulty, ultimately affecting their classification decisions.

V. Experimental Design: Classification-Verification Game

A. *Overview*

This experiment investigates if different types of feedback affect the accuracy of information classification and the willingness to pay (WTP) for verification of classified headlines.

The experimental design is based on the decision-making problem presented in section III, where participants are tasked with classifying headlines as either accurate (a) or fake (f) and revealing their willingness to pay to verify each headline. Participants must classify a headline (c_i) and indicate the maximum amount of money they pay to receive additional information from a perfect signal S.

Participants evaluated 50 headlines divided into five blocks and observed the same set of news in each block in a random order. The first three blocks were non-political, while the last two were political. The political headlines could be favorable or unfavorable to the current government. At the end of each block, they were asked for their expected accuracy in classifying the headlines in the block.

They were randomly assigned to one of three conditions based on the type of feedback received after providing their expected accuracy: Control group, individual feedback, and others' feedback.

After finishing evaluating all the headlines, the participants answered a survey that included two questions that were used to classify them in their support of the government.

B. Population

The participants were 184 undergraduate students in Mexico. 45% of them were man, and the average age was 20 years old. They were recruited from UNAM (National Autonomous University of Mexico) and IPN (National Polytechnic Institute), the first and second most important public schools in Mexico.[3]

C. Decision-Making Problem

Participants are tasked with classifying headlines as either accurate (a) or fake (f). The true state of the world for each headline is represented by $\Omega = \{A, F\}$, where A denotes an accurate headline and F denotes a fake headline. The prior probability of encountering a fake or accurate headline the same: $P(A) = P(F) = p_f = 0.5$. The problem is described in detail in section III implementing experimentally the simplifying assumptions.

Simplifying Assumptions

We implemented experimentally the assumptions in section III.C. The agent's utility for correctly classifying a headline as accurate or fake was $U_A = U_F =$ \$10.00 MXN. Conversely, misclassifying a fake headline as accurate or the other way around results in a utility of $U_{AF} = U_{FA} =$ \$0 MXN. This assumption allows

[3]In the national ranking, UNAM is the most important university, and IPN can be ranked third (https://www.usnews.com/education/best-global-universities/mexico) or forth (https://www.topuniversities.com/university-rankings-articles/world-university-rankings/best-universities-mexico), depending of the ranking.

the value of the signal to be interpreted purely in probability terms, independent of differential payoffs.

The decision-making process is influenced by the probability of correctly classifying a headline. Specifically, the agent's probability of correctly classifying a headline is $P(a|A) \geq 0.5$ when the headline is accurate and $P(f|F) \geq 0.5$ when the headline is fake. These probabilities are subjective and potentially different for each participant. Additionally, there is a perfect informative signal S that provides information about the state of the world, revealing the true state of the world with probabilities $P(s = a|A) = q_a = 1$ and $P(s = f|F) = q_f = 1$. This ensures the signal is strong enough to guide the agent's classification decisions.

Under these assumptions, value of In this case, the signal always reveals the true state of the world, meaning that the agent's WTP for the signal equals the expected value of sample information (EVSI).

D. Experimental Blocks and Classification Task

Each participant in the experiment is presented with a decision problem structured around five blocks of classification tasks. The first three blocks consist of neutral headlines, while the last two blocks introduce headlines with political content. The specific tasks and parameters involved in each block are as follows:

- **Headline Distribution** (h_i): Each participant receives 10 headlines per block. In the first three blocks, the headlines are neutral. In the last two blocks, the headlines are political, with half favoring the current government and the other half unfavorable to the government. Each headline h_i has a fixed prior probability $p_f = 0.5$ of being fake.

- **Classification Task** (c_i): Participants classify all 10 headlines in each block as either accurate (a) or fake (f). This decision represents the agent's classification based on their considerations without the signal.

- **Probability Estimation** ($\hat{p}_i$): After classifying all 10 headlines in a block, participants estimate the probability $\hat{p}_i$ that their classifications were correct. This estimation reflects the participant's confidence in their decisions, knowing the classification they made.

- **Willingness to Pay for Verification** (WTP_i): Participants then decide their WTP to observe the signal S that reveals the true state of the world for each headline. Verification is conducted using the Becker-DeGroot-Marschak (BDM) mechanism, where the participant's bid determines whether they will pay for verification and at what price. Given the assumption of a perfect signal ($q_a = q_f = 1$), the signal perfectly reveals whether the headline is accurate or fake. If the signal is purchased through the BDM mechanism, the reclassification is made automatically, so no extra decision is required.

- **Payment**: At the end of the experiment, one of the five blocks is randomly selected for payment. Participants are paid 10 MXN (ten Mexican pesos) for each headline correctly classified in the selected block.

Block Data

In each block of the experiment, several key data points are collected to analyze participants' behavior and decision-making processes. The primary data include the number of headlines each participant classifies, as well as the identity of these headlines, specifically whether they are classified as accurate or fake and whether they are politically charged or neutral.

Participants' confidence in their classification decisions is also measured. After making their classifications, participants are asked to report the probability that they believe their classifications are correct. This self-reported probability allows for the calculation of overconfidence metrics. Overconfidence is assessed by comparing the participants' reported probabilities of correct classification with the actual probabilities derived from the task. Specifically, overconfidence for accurate classifications is calculated as $O_a = \frac{P(A|a)}{\hat{P}(A|a)}$, and for fake classifications as $O_f = \frac{P(F|f)}{\hat{P}(F|f)}$. An overall measure of overconfidence is also calculated as $O = \frac{P(c=\omega)}{\hat{P}(c=\omega)}$, where c is the classification and ω is the true state of the world.

In addition to these confidence and classification metrics, the experiment collects data on the participants' economic behavior related to information verification. Specifically, the Willingness to Pay (WTP) for verification is measured using the Becker-DeGroot-Marschak (BDM) mechanism. This measure reflects the value that participants place on verifying their classifications, offering insights into their risk assessment and decision-making processes.

Lastly, participants are asked to report the probability that they believe each headline is accurate. This reported probability provides further data on their confidence in the accuracy of the headlines they have classified, contributing to a comprehensive understanding of their decision-making behavior.

Overall, the data collected in each block provide a detailed view of how participants classify headlines, how confident they are in their decisions, how they value verification, and how their confidence compares to actual performance.

E. Feedback Treatments

The experiment tests the effects of different types of feedback on classification accuracy and WTP. Feedback is provided to participants based on their treatment group:

- **Control Group**: Participants in the control group receive no feedback. Their classification decisions, probability estimations, and WTP are recorded without any information on their past performance.

- **Personal Performance Feedback** (f_p): Participants receive feedback on their overall performance for the entire block, specifically on the accuracy of their classifications in that block. This feedback allows them to learn about their overall ability to classify headlines, rather than the accuracy of individual headlines.
- **Group Performance Feedback** (f_g): Participants receive feedback on how many participants in a reference group correctly classified the same headlines in the block. This feedback provides an aggregate measure of task difficulty, helping participants learn about the overall difficulty of the task rather than what makes a specific headline true or false.

F. Experimental Procedure

The experimental procedure is structured as follows:

1) **Assignment to Treatments**: Participants are randomly assigned to one of the three treatment groups (f_p, f_g, or control).
2) **Blocks of Classification**: Each participant completes five blocks, where they classify 10 headlines in each block. The first three blocks consist of neutral headlines, while the last two blocks include political content. Participants report their probability of being correct and their WTP for verification after classifying all 10 headlines in each block.
3) **Feedback and Decision Update**: After each block, participants in the feedback groups receive feedback on their overall performance in the block. They then have the opportunity to update their decisions in subsequent blocks.
4) **Exit Survey**: After completing all blocks, participants fill out an exit survey that collects demographic information and assesses their support for the current government. The survey data is used to analyze correlations between participants' characteristics and their behavior in the experiment.
5) **Payment**: At the end of the experiment, one block is randomly selected for payment. Participants are compensated 10 MXN for each headline correctly classified in the selected block.

VI. Methodological Considerations

The experimental methodology is designed to capture the five hypotheses by systematically varying the type of feedback provided to participants and measuring their subsequent behavior. Using neutral headlines minimizes the potential influence of motivated reasoning, allowing the experiment to focus on the effects of feedback and overconfidence. At the same time, political news can show the effects of motivated reasoning on the demand for information. By testing these

hypotheses, the experiment contributes to a deeper understanding of how different types of feedback affect behavior, the role of overconfidence in shaping expectations, and the economic value participants place on information accuracy.

The key parameters and assumptions—such as the equal prior probabilities, equal utilities, and a perfect signal—simplify the problem and allow for a focus on the probabilistic aspects of classification and verification.

A. Measure Willingness to Pay for Verification

The BDM mechanism is employed to accurately gauge participants' WTP for verification, providing a robust measure of how feedback influences their valuation of accurate information. Verifying is a discrete decision based on the expected gains and costs of doing so. However, the willingness to pay to verify directly measures the expected gains that are hidden in a discrete decision. Two people verifying (or not) can have different values for the information.

B. Perfect Verification

The methodology presented here simplifies the world in a way that allows us to explore the causal effects of feedback on the value of verification. In the real world, we have verification practices that could be imperfect and different utilities for believing fake news and rejecting accurate information. Providing feedback in the real world could change the importance of each kind of mistake. This could be behind the effects of highlighting the importance of accuracy observed in Pennycook et al. (2021) and Pennycook and Rand (2022).

The best way to verify is an open question not addressed in the current research; the question is about overconfidence as a mechanism behind low levels of verification. We present a perfect signal to avoid motivated misinterpretation of the signal's likelihoods (Thaler (2024)) and also to avoid the discussion of the real information value of a specific verification practice. The results from Thaler (2024) are very important since a no-completely-informative signal will allow a biased updating belief that also decreases the WTP for the signal. Even online searching of news, one of the most common practices promoted in digital literacy programs could backfire (Aslett et al. (2024); Hoes et al. (2023)). Also, having imperfect signals might make purchasing any information suboptimal if participants expect their original classification to remain the same even after the verification.

C. Headlines Selection

The online publication AnimalPolitico[4] and VerificadoMX[5] were used as the sources to find relevant fake news circulating in Mexico. These are the most relevant fact-checking efforts recognized in Mexico. The authors verified these

[4]https://animalpolitico.com/verificacion-de-hechos
[5]https://verificado.com.mx/

headlines independently. To find headlines that were real but difficult to classify, the authors used NewsGPT[6]. All the headlines generated were verified independently by the authors. From these sources, the authors selected 60 headlines: 30 political, and 30 non-political, half of them true and the other half false. Also, from the political headlines, 15 were classified as information that favored the government, and 15 opposed the government.

To select the 50 headlines for the final experiment, and the order in the blocks, we run a study in Prolific. We asked for the classification of the headlines with the same incentives as in the final experiment and measured the probability each headline was classified correctly. The headline composition of the blocks was made such that they have similar levels of difficulty.

VII. Results

The analysis of the experiment investigating headline classification and verification provides several key insights. First, there is evidence that confidence in headline classification decreases across blocks, suggesting an experience effect. Participants may become more aware of the task's difficulty as they proceed, leading to a reduction in overconfidence over time.

Additionally, the treatments showed a no significant effect. This result indicates that individual feedback was no effective in enhancing perceived classification accuracy compared to the control.

Political headlines were associated with higher levels of confidence, reflecting participants' stronger convictions when classifying content that aligns with their political beliefs. This increased confidence could be attributed to participants' pre-existing political biases. Moreover, participants who supported the government exhibited higher confidence in their classifications, particularly when verifying information favorable to their political views.

There was also strong evidence of verification bias. Participants were more likely to verify information that aligned with their beliefs, as indicated by a positive and significant coefficient for the classification of headlines as accurate or false. This suggests a tendency to reinforce existing views rather than challenge them through verification.

Regarding willingness to pay (WTP) for verification, political news increased participants' WTP, highlighting the perceived importance of accuracy in politically charged information. However, no significant effect was found for participants' political standing on their WTP. Nevertheless, government supporters demonstrated a higher demand for verification, especially when verifying unfavorable political news.

Finally, confirming previous findings, the only significant difference between the control group and the treatment groups was related to the feedback on others'

[6]The request was made in August, around three weeks before the start of the first session: https://chatgpt.com/g/g-NnU2wmnZ5-newsgpt-chat-with-hundreds-of-news-sources/c/7e750031- b534-481c-83cf-2dc6917d98b4

performance. This suggests that group feedback plays a critical role in helping individuals adjust their expectations and performance.

These findings highlight the interplay between political bias, feedback mechanisms, and government support in shaping confidence and verification behavior in the context of classifying misinformation.

VIII. Discussion

We found a negative relationship between the feedback on other's performance and the willingness to pay for information. Considering that underconfidence was a more prevalent trait of the participants, this result follows the theory; they could be learning that the task is easier than expected, and they will consider that the probability of correctly classifying is higher, which reduces the value of new verifying the headline. However, we didn't find a significant effect of feedback on the probability they thought they got a correct classification (level of confidence). The willingness to pay was asked by each headline (50 observations), while the confidence was asked at the end of the block. This mismatch creates a noisy relation between these variables since the average confidence is not linked to individual headlines, where participants can be very confident or very unsure about the veracity of the headline. We believe that this noise in the measure of both variables hides the relationship between the expected probability of a correct classification and willingness to pay to verify.

REFERENCES

Arrow, K. J., D. Blackwell, and M. A. Girshick, "Bayes and Minimax Solutions of Sequential Decision Problems," *Econometrica*, 7 1949, *17*, 213.

Aslett, Kevin, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker, "Online searches to evaluate misinformation can increase its perceived veracity," *Nature*, 1 2024, *625*, 548–556.

Hoes, Emma, Brian Aitken, Jingwen Zhang, Tomasz Gackowski, and Magdalena Wojcieszak, "Prominent misinformation interventions reduce misperceptions but increase scepticism," *Nature Human Behavior*, 2023.

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K.H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson, Paula Szewach, Sander van der Linden, and Sam Wineb-

TABLE 1—

	Dependent variable:		
	accuracy_estimate	willingness_to_pay	
	(1)	(2)	(3)
treatmentIndividual	−4.705***	−0.211	−0.214
	(0.022)	(0.215)	(0.240)
treatmentOthers	−2.537***	−0.385*	−0.450*
	(0.024)	(0.209)	(0.241)
block	−1.077***	−0.012	0.042
	(0.015)	(0.023)	(0.050)
true_or_false	1.389***	0.236***	0.242***
		(0.057)	(0.072)
correct	−0.032**	−0.039	−0.040
	(0.007)	(0.042)	(0.054)
age	0.977***	−0.035	−0.048
	(0.005)	(0.052)	(0.056)
genderMasculino	3.353***	0.019	0.057
	(0.019)	(0.178)	(0.202)
accuracy_estimate		0.001	0.001
		(0.003)	(0.004)
TypePolitical	5.510***	0.194***	
		(0.056)	
support_gov	8.006***	0.503**	0.392
	(0.032)	(0.205)	(0.255)
Favor_gov			−0.033
			(0.045)
against_gov	5.028***	0.251	0.254
	(0.034)	(0.238)	(0.261)
support_govTRUE:Favor_gov			0.092
			(0.082)
Favor_gov:against_gov			0.083
			(0.079)
Constant	33.673***	3.198***	3.420***
	(0.107)	(1.049)	(1.160)
Observations	8,903	8,903	3,603
R^2	0.053	0.034	0.033
Adjusted R^2	0.052	0.032	0.029
Note:			*p<0.1; **p<0.05; ***p<0.01

urg, "Toolbox of individual-level interventions against online misinformation," *Nature Human Behaviour 2024 8:6*, 5 2024, *8*, 1044–1052.

Lyons, Benjamin A., Jacob M. Montgomery, Andrew M. Guess, Brendan Nyhan, and Jason Reifler, "Overconfidence in news judgments is associated with false news susceptibility," *Proceedings of the National Academy of Sciences of the United States of America*, 6 2021, *118*.

Pennycook, Gordon and David G. Rand, "Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking," *Journal of Personality*, 4 2020, *88*, 185–200.

_ **and** _ , "Nudging Social Media toward Accuracy," *Annals of the American Academy of Political and Social Science*, 3 2022, *700*, 152–164.

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

Thaler, Michael, "The Fake News Effect: Experimentally Identifying Motivated Reasoning Using Trust in News," *American Economic Journal: Microeconomics*, 2024, *16*, 1–38.

Wald, Abraham, "Foundations of a General Theory of Sequential Decision Functions," *Econometrica*, 10 1947, *15*, 279.

Mathematical Appendix

A1. Expected Value of the Signal Considering Reclassification

General Setup

1) **Initial Classification**: The agent initially classifies the headline as c (either accurate (a) or fake (f)).

2) **Receive Signal**: The agent receives a signal s which can either confirm or contradict their initial classification.

3) **Reclassification**: Based on the signal, the agent makes a new classification c'.

Expected Utility with Signal and Reclassification ($EU^{\text{UPDATE}}_{\text{SIGNAL}}(c)$)

The expected utility with the signal and reclassification is calculated by considering the updated posterior probabilities and the new classification based on the signal.

Step 1: Define Probabilities and Utilities

- **Initial Posterior Probabilities**:

$$P(A|c) = \frac{P(c|A) \cdot (1 - p_f)}{P(c)}, \quad P(F|c) = \frac{P(c|F) \cdot p_f}{P(c)}$$

where:

$$P(c) = (1 - p_f) \cdot P(c|A) + p_f \cdot P(c|F)$$

- **Signal Probabilities**:

$$q_a = P(s = a|A), \quad q_f = P(s = f|F)$$

- **Utilities**:

$$U_A, U_F, U_{AF}, U_{FA}$$

Step 2: Define Updated Posterior Probabilities Given Signal

After observing the signal, the agent updates their beliefs:

- **Posterior Probabilities Given Signal** $s = a$:

$$P(A|s = a, c) = \frac{q_a \cdot P(A|c)}{q_a \cdot P(A|c) + (1 - q_f) \cdot P(F|c)}$$

$$P(F|s = a, c) = \frac{(1 - q_f) \cdot P(F|c)}{q_a \cdot P(A|c) + (1 - q_f) \cdot P(F|c)}$$

- **Posterior Probabilities Given Signal** $s = f$:

$$P(A|s = f, c) = \frac{(1 - q_a) \cdot P(A|c)}{(1 - q_a) \cdot P(A|c) + q_f \cdot P(F|c)}$$

$$P(F|s = f, c) = \frac{q_f \cdot P(F|c)}{(1 - q_a) \cdot P(A|c) + q_f \cdot P(F|c)}$$

Step 3: Calculate Expected Utility After Signal

We assume that the signal is strong enough (proposition 1) to assure that the best reclassification is to follow what the signal indicates is the state of the world. Otherwise, the signal would have no instrumental value, and therefore, $EVSI = 0$.

The expected utility with the signal, considering the possibility of reclassification, is:

$$EU^{\text{update}}_{\text{signal}}(c) = P(s = a|c) \cdot EU_{\text{new classification}}(s = a, c) + P(s = f|c) \cdot EU_{\text{new classification}}(s = f, c)$$

Here, $P(s = a|c)$ and $P(s = f|c)$ are the probabilities of receiving the signals $s = a$ and $s = f$ given the initial classification c. These probabilities are determined by Bayes' rule, considering the agent's initial classification and the properties of the signal.

Given the initial classification c:

$$P(s = a|c) = q_a \cdot P(A|c) + (1 - q_f) \cdot P(F|c)$$

$$P(s = f|c) = (1 - q_a) \cdot P(A|c) + q_f \cdot P(F|c)$$

Expected Utility After Signal $s = a$

$$EU_{\text{new classification}}(s = a, c) = P(A|s = a, c) \cdot U_A + P(F|s = a, c) \cdot U_{AF}$$

Expected Utility After Signal $s = f$

$$EU_{\text{new classification}}(s = f, c) = P(F|s = f, c) \cdot U_F + P(A|s = f, c) \cdot U_{FA}$$

Step 4: Combine Expected Utilities

$$EU^{\text{update}}_{\text{signal}}(c) = [q_a \cdot P(A|c) + (1 - q_f) \cdot P(F|c)] \cdot [P(A|s = a, c) \cdot U_A + P(F|s = a, c) \cdot U_{AF}]$$
$$+ [(1 - q_a) \cdot P(A|c) + q_f \cdot P(F|c)] \cdot [P(F|s = f, c) \cdot U_F + P(A|s = f, c) \cdot U_{FA}]$$

Step 5: Expected Value of the Signal (EVSI)

Finally, the EVSI is the difference between the expected utility with the signal and the expected utility without the signal.

$$EVSI = EU^{\text{update}}_{\text{signal}}(c) - EU_{\text{no signal}}(c)$$

Conclusion

By allowing the agent to update their classification based on the signal, we account for the dynamic decision-making process. The expected value of the signal (EVSI) is derived by comparing the expected utility with the signal (considering

reclassification) to the expected utility without the signal. This approach shows the impact of additional information on improving decision-making accuracy.

A2. Proof of Proposition 2: Need for a Strong Enough Signal

To ensure that people follow the signal S for reclassification, we must prove that the expected utility of reclassifying based on the signal is higher than not reclassifying. Without loss of generality, we will first consider the case where the initial classification $c = f$ and the signal $s = a$.

Initial Setup

1) **Initial Classification**: $c = f$ (classified as fake)

2) **Signal Received**: $s = a$ (signal indicates accurate)

We need to show that reclassifying the headline as accurate ($c' = a$) based on the signal is optimal.

Expected Utility of Not Reclassifying

If the agent does not reclassify and sticks with the initial classification $c = f$, but knows the signal $s = a$, the expected utility is:

$$EU_{\text{no reclassification}}(f, s = a) = P(A|s = a, f) \cdot U_{FA} + P(F|s = a, f) \cdot U_F$$

Expected Utility of Reclassifying

If the agent reclassifies the headline based on the signal $s = a$, the expected utility is:

$$EU_{\text{reclassification}}(f, s = a) = P(A|s = a, f) \cdot U_A + P(F|s = a, f) \cdot U_{AF}$$

Posterior Probabilities

The posterior probabilities given the signal $s = a$ and initial classification $c = f$ are:

$$P(A|s = a, f) = \frac{q_a \cdot P(A|f)}{q_a \cdot P(A|f) + (1 - q_f) \cdot P(F|f)}$$
$$P(F|s = a, f) = \frac{(1 - q_f) \cdot P(F|f)}{q_a \cdot P(A|f) + (1 - q_f) \cdot P(F|f)}$$

Condition for Reclassifying

To prove that reclassifying based on the signal is optimal, we need:

$$EU_{\text{reclassification}}(f, s = a) > EU_{\text{no reclassification}}(f, s = a)$$

Substituting the utilities, we get:

$$P(A|s = a, f) \cdot U_A + P(F|s = a, f) \cdot U_{AF} > P(A|s = a, f) \cdot U_{FA} + P(F|s = a, f) \cdot U_F$$

Given the simplifying assumptions:

$$U_A = 1, \quad U_F = 1, \quad U_{AF} = 0, \quad U_{FA} = 0$$

The inequality simplifies to:

$$P(A|s = a, f) \cdot 1 + P(F|s = a, f) \cdot 0 > P(A|s = a, f) \cdot 0 + P(F|s = a, f) \cdot 1$$

This reduces to:

$$P(A|s = a, f) > P(F|s = a, f)$$

Verifying the Posterior Probabilities

Substitute the posterior probabilities:

$$\frac{q_a \cdot P(A|f)}{q_a \cdot P(A|f) + (1 - q_f) \cdot P(F|f)} > \frac{(1 - q_f) \cdot P(F|f)}{q_a \cdot P(A|f) + (1 - q_f) \cdot P(F|f)}$$

Since the denominators are the same, we can simplify this to:

$$q_a \cdot P(A|f) > (1 - q_f) \cdot P(F|f)$$

Since $P(F|f) = 1 - P(A|f)$, we have:

$$q_a \cdot P(A|f) > (1 - q_f) \cdot (1 - P(A|f))$$

Expanding and rearranging terms, we get:

$$q_a \cdot P(A|f) > (1 - q_f) - (1 - q_f) \cdot P(A|f)$$

$$q_a \cdot P(A|f) + (1 - q_f) \cdot P(A|f) > (1 - q_f)$$

$$P(A|f) \cdot (q_a + 1 - q_f) > (1 - q_f)$$

Dividing both sides by $(q_a + 1 - q_f)$:

$$P(A|f) > \frac{1 - q_f}{q_a + 1 - q_f}$$

This shows that the signal needs to be strong enough such that q_a is sufficiently large compared to $1 - q_f$, ensuring that the agent reclassifies the headline as accurate based on the signal. This proves that a strong signal is necessary to ensure that people follow the signal S for reclassification.

Trivial Case: $c = s = a$

If the initial classification $c = a$ and the signal $s = a$, then reclassification is not necessary because the initial classification is already accurate. The expected utility remains the same:

$$EU_{\text{reclassification}}(a, s = a) = P(A|s = a, a) \cdot U_A + P(F|s = a, a) \cdot U_{AF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{reclassification}}(a, s = a) = P(A|s = a, a) \cdot 1 + P(F|s = a, a) \cdot 0 = P(A|s = a, a)$$

And the expected utility of not reclassifying is:

$$EU_{\text{no reclassification}}(a, s = a) = P(A|s = a, a) \cdot U_A + P(F|s = a, a) \cdot U_{AF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{no reclassification}}(a, s = a) = P(A|s = a, a) \cdot 1 + P(F|s = a, a) \cdot 0 = P(A|s = a, a)$$

Since both expected utilities are equal, reclassification is trivial in this case.

Other Cases

The same process follows for the cases $c = a, s = f$ and $c = s = f$. For these cases, the conditions are as follows:

1) **Case** $c = a, s = f$:

$$EU_{\text{reclassification}}(a, s = f) > EU_{\text{no reclassification}}(a, s = f)$$

Substituting the utilities, we get:

$$P(F|s = f, a) \cdot U_F + P(A|s = f, a) \cdot U_{FA} > P(F|s = f, a) \cdot U_{AF} + P(A|s = f, a) \cdot U_A$$

Given the simplifying assumptions:

$$P(F|s=f,a)\cdot 1 + P(A|s=f,a)\cdot 0 > P(F|s=f,a)\cdot 0 + P(A|s=f,a)\cdot 1$$

This reduces to:

$$P(F|s=f,a) > P(A|s=f,a)$$

Verifying the posterior probabilities:

$$\frac{q_f \cdot P(F|a)}{q_f \cdot P(F|a) + (1-q_a)\cdot P(A|a)} > \frac{(1-q_a)\cdot P(A|a)}{q_f \cdot P(F|a) + (1-q_a)\cdot P(A|a)}$$

Since the denominators are the same, we can simplify this to:

$$q_f \cdot P(F|a) > (1-q_a)\cdot P(A|a)$$

Since $P(A|a) = 1 - P(F|a)$, we have:

$$q_f \cdot P(F|a) > (1-q_a)\cdot (1-P(F|a))$$

Expanding and rearranging terms, we get:

$$q_f \cdot P(F|a) > (1-q_a) - (1-q_a)\cdot P(F|a)$$

$$q_f \cdot P(F|a) + (1-q_a)\cdot P(F|a) > (1-q_a)$$

$$P(F|a)\cdot (q_f + 1 - q_a) > (1-q_a)$$

Dividing both sides by $(q_f + 1 - q_a)$:

$$P(F|a) > \frac{1-q_a}{q_f + 1 - q_a}$$

2) **Case** $c = s = f$: If the initial classification $c = f$ and the signal $s = f$, then reclassification is not necessary because the initial classification is already correct. The expected utility remains the same:

$$EU_{\text{reclassification}}(f, s=f) = P(F|s=f,f)\cdot U_F + P(A|s=f,f)\cdot U_{AF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{reclassification}}(f, s=f) = P(F|s=f,f)\cdot 1 + P(A|s=f,f)\cdot 0 = P(F|s=f,f)$$

And the expected utility of not reclassifying is:

$$EU_{\text{no reclassification}}(f, s=f) = P(F|s=f,f)\cdot U_F + P(A|s=f,f)\cdot U_{AF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{no reclassification}}(f, s = f) = P(F|s = f, f)\cdot 1 + P(A|s = f, f)\cdot 0 = P(F|s = f, f)$$

Since both expected utilities are equal, reclassification is trivial in this case.

Conditions

Therefore, to ensure that the signal is strong enough to prompt optimal reclassification in both cases, we need to satisfy two key conditions:

$$P(A|f) > \frac{1 - q_f}{q_a + 1 - q_f}$$
$$P(A|a) < \frac{q_f}{q_f + 1 - q_a}$$

Considering the conditions for proposition 1 we have that,

$$\frac{1 - q_f}{q_a + 1 - q_f} < P(A|f) < P(A|a) < \frac{q_f}{q_f + 1 - q_a}$$

A3. Proof of Proposition 1: Sufficient and Necessary Condition for $P(A|c = f) < P(A) < P(A|c = a)$ and $P(F|c = a) < P(F) < P(F|c = f)$

Proof of Necessity and Sufficiency

We will prove that $1 < \frac{P(c=a|A)}{P(c=a|F)}$ and $1 < \frac{P(c=f|F)}{P(c=f|A)}$ if and only if $P(A|c = f) < P(A) < P(A|c = a)$ and $P(F|c = a) < P(F) < P(F|c = f)$.

B1. Definitions and Setup

Let:

- $P(A)$ be the prior probability that the state is accurate.
- $P(F)$ be the prior probability that the state is fake.
- $P(c = a|A)$ be the probability of classifying a headline as accurate given it is accurate.
- $P(c = a|F)$ be the probability of classifying a headline as accurate given it is fake.
- $P(c = f|F)$ be the probability of classifying a headline as fake given it is fake.
- $P(c = f|A)$ be the probability of classifying a headline as fake given it is accurate.

B2. Posterior Probabilities

The posterior probabilities after observing the classification c are given by:

- Posterior probability of A given $c = f$:

$$P(A|c = f) = \frac{P(c = f|A) \cdot P(A)}{P(c = f|A) \cdot P(A) + P(c = f|F) \cdot P(F)}$$

- Posterior probability of A given $c = a$:

$$P(A|c = a) = \frac{P(c = a|A) \cdot P(A)}{P(c = a|A) \cdot P(A) + P(c = a|F) \cdot P(F)}$$

- Posterior probability of F given $c = f$:

$$P(F|c = f) = \frac{P(c = f|F) \cdot P(F)}{P(c = f|A) \cdot P(A) + P(c = f|F) \cdot P(F)}$$

- Posterior probability of F given $c = a$:

$$P(F|c = a) = \frac{P(c = a|F) \cdot P(F)}{P(c = a|A) \cdot P(A) + P(c = a|F) \cdot P(F)}$$

B3. Part 1: Sufficiency (⇒)

Assume that $1 < \frac{P(c=a|A)}{P(c=a|F)}$ and $1 < \frac{P(c=f|F)}{P(c=f|A)}$. We want to show that this implies $P(A|c = f) < P(A) < P(A|c = a)$ and $P(F|c = a) < P(F) < P(F|c = f)$.

Analyze the Posterior Probabilities

1) **For $P(A|c = f)$:**

   Given the condition $1 < \frac{P(c=f|F)}{P(c=f|A)}$, we know that:

$$\frac{P(c = f|F)}{P(c = f|A)} > 1$$

   This implies $P(c = f|F) > P(c = f|A)$. As a result, in the posterior probability expression:

$$P(A|c = f) = \frac{P(c = f|A) \cdot P(A)}{P(c = f|A) \cdot P(A) + P(c = f|F) \cdot P(F)}$$

   The denominator $P(c = f|A) \cdot P(A) + P(c = f|F) \cdot P(F)$ will be larger than the numerator $P(c = f|A) \cdot P(A)$, causing $P(A|c = f)$ to be smaller than

the prior $P(A)$. Therefore:

$$P(A|c = f) < P(A)$$

2) **For $P(A|c = a)$:**

Given the condition $1 < \frac{P(c=a|A)}{P(c=a|F)}$, we know that:

$$\frac{P(c = a|A)}{P(c = a|F)} > 1$$

This implies $P(c = a|A) > P(c = a|F)$. As a result, in the posterior probability expression:

$$P(A|c = a) = \frac{P(c = a|A) \cdot P(A)}{P(c = a|A) \cdot P(A) + P(c = a|F) \cdot P(F)}$$

The numerator $P(c = a|A) \cdot P(A)$ will dominate the denominator $P(c = a|A) \cdot P(A) + P(c = a|F) \cdot P(F)$, causing $P(A|c = a)$ to be larger than the prior $P(A)$. Therefore:

$$P(A|c = a) > P(A)$$

3) **For $P(F|c = f)$ and $P(F|c = a)$:**

Similarly, the same reasoning applies to $P(F|c = f)$ and $P(F|c = a)$, given that:

$$\frac{P(c = f|F)}{P(c = f|A)} > 1 \quad \text{and} \quad \frac{P(c = a|A)}{P(c = a|F)} > 1$$

This implies that:

$$P(F|c = a) < P(F) < P(F|c = f)$$

B4. Part 2: Necessity (⇐)

Assume that $P(A|c = f) < P(A) < P(A|c = a)$ and $P(F|c = a) < P(F) < P(F|c = f)$. We need to show that this implies $1 < \frac{P(c=a|A)}{P(c=a|F)}$ and $1 < \frac{P(c=f|F)}{P(c=f|A)}$.

Analyzing the Posterior Probabilities

- **For $P(A|c = f) < P(A)$:**

Given the posterior probability expression:

$$P(A|c = f) = \frac{P(c = f|A) \cdot P(A)}{P(c = f|A) \cdot P(A) + P(c = f|F) \cdot P(F)}$$

If $P(A|c = f) < P(A)$, then the likelihood ratio $\frac{P(c=f|F)}{P(c=f|A)}$ must be greater than 1. This is because the posterior $P(A|c = f)$ being less than $P(A)$ implies that the signal $c = f$ is more likely to come from the fake state F, meaning:

$$\frac{P(c = f|F)}{P(c = f|A)} > 1$$

- **For $P(A|c = a) > P(A)$:**

  Given the posterior probability expression:

$$P(A|c = a) = \frac{P(c = a|A) \cdot P(A)}{P(c = a|A) \cdot P(A) + P(c = a|F) \cdot P(F)}$$

  If $P(A|c = a) > P(A)$, then the likelihood ratio $\frac{P(c=a|A)}{P(c=a|F)}$ must be greater than 1. This is because the posterior $P(A|c = a)$ being greater than $P(A)$ implies that the signal $c = a$ is more likely to come from the accurate state A, meaning:

$$\frac{P(c = a|A)}{P(c = a|F)} > 1$$

- **For $P(F|c = f) > P(F)$ and $P(F|c = a) < P(F)$:**

  By symmetry, the same reasoning applies for $P(F|c = f) > P(F)$ and $P(F|c = a) < P(F)$. The likelihood ratios $\frac{P(c=f|F)}{P(c=f|A)} > 1$ and $\frac{P(c=a|A)}{P(c=a|F)} > 1$ are necessary conditions to satisfy these posterior inequalities.

B5. Conclusion

Thus, we have shown that: $1 < \frac{P(c=a|A)}{P(c=a|F)}$ and $1 < \frac{P(c=f|F)}{P(c=f|A)}$ are necessary and sufficient conditions for:

$$P(A|c = f) < P(A) < P(A|c = a)$$

$$P(F|c = a) < P(F) < P(F|c = f)$$

Corollary: Informativeness of the Signal

The same analysis can be applied to the signal. Therefore $1 < \frac{P(s=a|A)}{P(s=a|F)}$ and $1 < \frac{P(s=f|F)}{P(s=f|A)} \iff$

$$P(A|s = f) < P(A) < P(A|s = a)$$

$$P(F|s = a) < P(F) < P(F|s = f)$$

Regressions Appendix

C1. Exclusion Criteria

We exclude from the main analysis those participants with less than 80% of their answers and those who decide not to share their gender or answer "Other." Here, we present the regressions considering the whole sample.