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 is a global issue that seriously threatens democratic societies by undermining the public's ability to make informed decisions, eroding trust in institutions, and fueling polarization. If no efforts are spent fighting misinformation, societies risk making decisions based on false assumptions. The spread of misinformation through digital platforms is faster and deeper than factual information (Vosoughi et al., 2018), and it is primarily people (not bots) who share misinformation inadvertently (Arin et al., 2023). Promoting verification can help to prevent the spread of false information, as promoting cybersecurity can prevent the spread of viruses and scams.

Verification is essential to fight misinformation, increasing the effectiveness of measures related fact-checking[1]. In this paper, we refer to searching for further information to reveal if a statement is accurate or false[2] as verification. Then, reading the explanation on tagged content or following the link to information

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

† UCSB, USA, dariotrujanoochoa@ucsb.edu.

‡ UCLA, USA, josegloria@ucla.edu.

[1]The difference between fact-checking and verification is that fact-checking refers to a policy that identifies fake news and spreads this information, while verification is the active individual decision to search for further information.

[2]Not all content and information can be classified in this way. Normative statements (e.g. "Headlines should prioritize impactful language to capture readers' attention.") are examples of them. In this paper, we focus our attention on headlines easily classified as accurate or false.

on fact-checks are instances of verification in the field. Some intervention like debunking and fact-checking exogenously fight misinformation by pointing out false information. However, if users don't see verification as valuable, the impact of interventions like fact-checking, labeling content, and media literacy is diminished. Thus, verification complements other efforts[3] to fight misinformation, and if people are highly confident in their ability to distinguish false information, they should rationally reduce their demand for verification.

There is evidence of the effects of feedback in reducing overconfidence, but its effect on verification hasn't been studied. Interventions like debunking, inoculation, and digital literacy offer indirect feedback on the accuracy that have when classifying headlines. In this research, we measure the demand for verification in a lab experiment and explore how providing feedback explicitly modifies this demand.

This paper tests the hypothesis that feedback affects confidence levels and verification value. We found evidence that explicitly including feedback on the accuracy of other participants reduces the value of verification. At the core of this research is an innovative exploration of how feedback mechanisms can improve the effectiveness of fact-checking. The study shifts from analyses of exogenous interventions to fight misinformation to explore how people change the demand for verification, specifically based on feedback.

Recent reviews of the interventions recognize that most of the results come from the US and Europe (Kozyreva et al., 2024; Bateman and Jackson, 2024). In this paper, we study headlines in Mexico, the largest Spanish speaking country, where misinformation and polarization have increased recently, as in many other places in the world. This research contributes in closing the gap of research done on misinformation in developed and developing countries. The exepriment design in this paper, however, control for heterogeneity in the particiants in a way that the results can reasonably be expected in other populations.

We found that feedback on the accuracy of others reduces the willingness to pay to verify news headlines. Also, there was no evidence of overconfidence among participants. This suggests that providing information about accuracy could backfire in other interventions by reducing the value that people put on engaging with fact-checking.

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, re-

[3]For example, Facebook (Meta) collaborates with fact-checking organizations to tag false information (https://transparency.meta.com/policies/community-standards/misinformation), and X uses CommunityNotes (https://x.com/communitynotes?lang=en) to let the users decide through an algorithm over their ratings.

searchers 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) found that overconfidence in one's own cognitive abilities correlates with a higher likelihood of accepting false claims as true. Ortoleva and Snowberg (2015) found theoretically and empirically that overconfidence due to correlation neglect leads to higher polarization. These studies highlight overconfidence's role in the dissemination of fake news by decreasing verification and a higher probability of sharing misinformation.

Encouraging behaviors related to fact-checking and verification has been considered an effective approach to combating misinformation. In two reviews of tested interventions, Kozyreva et al. (2024) and Bateman and Jackson (2024) mentioned promoting media literacy, fact-checking, and labeling content 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 could backfire. 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. Other 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.

The literature shows mixed results on the effects of feedback on overconfidence, and asymmetric updating due to motivated reasoning has been reported. Incentivized studies have found a reduction of overconfidence with feedback (Ferraro, 2005; Eberlein et al., 2011; Kogelnik, 2022), while other studies in education have found no effects (Pulford and Colman, 1997; Erat et al., 2022). Other papers have found asymmetric effects of feedback related to motivated beliefs.

Oprea and Yuksel (2022) found that getting the estimation from someone else increases the subjective probability of classifying yourself as above the median in a test. And even when the signals are uninformative, people asymmetrically update their beliefs depending on the ego-related states of the world the signal points to (Thaler, 2024). Kartal and Tyran (2022) showed that overconfident participants vote even when receiving low-accuracy signals. However, they did not directly measure the demand for information; they measured it by deciding between voting or abstaining in an election. Moore and Healy (2008), found empirically that overestimation[4] is more common on difficult tasks, where people tend to believe they performed better than they actually did. They also found that feedback had a minimal effect on recalibrating overconfidence.

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 of our knowledge has elicited the value of verification in the context of misinformation or the effects of direct feedback.

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 confidence. Under the assumption of rationality, participants' WTP should be proportional to the value they give to information. 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.*

[4]Moore and Healy (2008) makes the distinction between overestimation, overplacement, and overprecision. In the literature, the term "overconfidence" has been used to describe these three concepts. However, most of the studies focus on overestimation, and we will use overconfidence to refer to this phenomenon. We will distinguish between overplacement and overplacement when confusion arises.

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). This instrumental value of information is assumed to be equal to the willingness to pay for verification. The preset setting is a framework to understand the decision problem that people face when deciding to verify, or no, headlines.

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 $P(\omega = F) = p_f$.[5] 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.[6] 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

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

[6] 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.

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. 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 $\pi = U_A - U_{AF}$ to get $V(c)$. Therefore, the willingness to pay to verify should be:

$$WTP(c) = \pi V(c) \tag{2}$$

The WTP decision is based solely on accuracy probability.

IV. Experimental Design: Classification-Verification Game

A. Overview

This experiment tests whether different types of feedback influence participants' accuracy in classifying information and their willingness to pay (WTP) for verification. Participants are tasked with categorizing headlines as accurate or fake and reporting their WTP for verification. They evaluate 50 headlines in five blocks, with feedback provided in one of three experimental conditions: control (no feedback), individual feedback, and group feedback. The design aims to measure participants' classification accuracy, confidence, and valuation of information through a decision-making framework. Following the main task, participants complete a survey on demographics and political orientation.

B. Experimental Blocks and Feedback Treatments

Participants completed five blocks, each designed to measure classification accuracy, confidence, and demand for verification through WTP. In each block, participants received 10 headlines in random order, which they were instructed to classify as either *accurate* (a) or *fake* (f), with an equal prior probability ($P(A) = P(F) = 0.5$) of each state. For each correctly classified headline, participants earned a utility of 10 Mexican Pesos (MXN), regardless of whether the classification was *accurate* or *fake*e ($U_A = U_F = 10$ MXN). Conversely, misclassifications, whether mistakenly classifying an accurate headline as *fake* or a fake headline as *accurate*, yielded a utility of 0 MXN ($U_{AF} = U_{FA} = 0$ MXN). The classification and WTP for each headline decision had a 20-second time limit. This utility structure incentivized participants to classify accurately.

For each headline, participants also indicated their willingness to pay (WTP) to access a perfect signal (S) that could reveal the headline's true status. The perfect signal, available for purchase, would reveal the true state of each headline with certainty ($P(s = a|A) = q_a = 1$ and $P(s = f|F) = q_f = 1$).

After completing all classifications and WTP decisions within a block, participants reported the estimated probability of correctly classifying the headlines by themselves and by others. There was no time limit when participants reported their confidence. This provided a self-assessed confidence measure for the block.

Following their probability estimations at the end of each block, participants received feedback according to the treatment group to which they were randomly assigned. Feedback was designed to inform participants about their classification

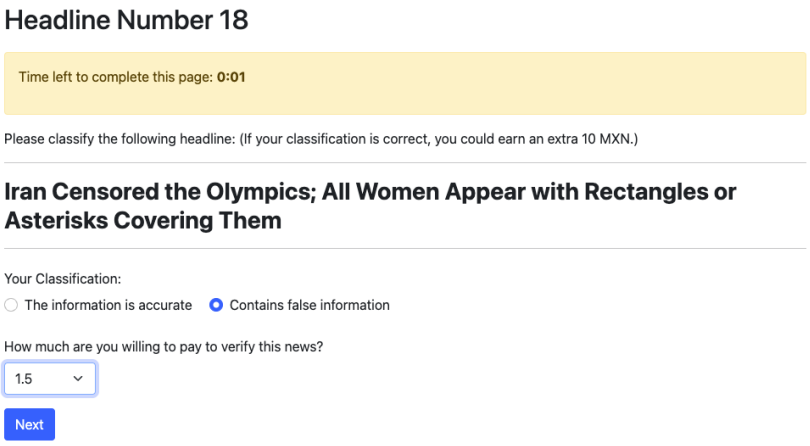

FIGURE 1. SCREENSHOT OF THE TRANSLATED CLASSIFICATION-VERIFICATION GAME AS SEEN BY THE PARTICIPANTS.

performance either individually, as part of a group reference, or not at all in the control group. The feedback types and their descriptions are presented in Table 1. By block, all treatment groups were shown a summary of the times they classified a headline as *accurate* or *fake*, and the feedback treatments were shown the accuracy rates conditional on the headlines classified as *accurate* or *fake*.

TABLE 1—FEEDBACK TREATMENTS

Treatment Group	**Feedback at the End of the Block**
Control Group	No feedback on accuracy was given.
Individual Feedback	Personal accuracy rate for the block conditional on the headlines participants classified as *accurate* or *fake*.
Others Feedback	Average accuracy rate of other participants conditional on the headlines others classified as *accurate* or *fake*.

This structure allowed researchers to observe how different feedback types influenced participants' accuracy, confidence, and valuation of the verification signal throughout the experiment.

C. *Experimental Procedures*

SAMPLE INFORMATION

The participants were 184 undergraduate students in Mexico. The average age was 20 years old, and 55% of them were women. 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[7].

Incentives and Utility

Participants receive a utility of 10 Mexican Pesos (MXN) for each correctly classified headline, whether it is accurate or fake ($U_A = U_F = 10$ MXN). Conversely, they receive a utility of 0 MXN for misclassifications, whether they mistakenly classify an accurate headline as fake or a fake headline as accurate ($U_{AF} = U_{FA} = 0$ MXN). This setup incentivizes participants to classify accurately and to value the signal appropriately based on its accuracy-assurance potential.

Post-Experiment Survey and Payment

After all blocks, participants complete a survey collecting demographic data and assessing their support for the current government. For payment, one block is randomly selected at the experiment's end, and participants receive 10 MXN for each correctly classified headline in the chosen block, thus linking final earnings directly to classification accuracy.

D. Methodological Considerations

This experiment controls for variables that are relevant in the field to focus purely on the effects of feedback. The believed probability of receiving fake news, the interest in classifying correctly, and the verification quality change the value of verifying. 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.

Measure Willingness to Pay for Verification and Confidence

To measure participants' WTP for verification, we used the BDM mechanism. 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. Particiants could indicate a number between 0 to 5 from a drop down menu as they willingess to pay for verification.

To measure the confidence levels, we used the method in (Wilson and Vespa, 2018) to present the binarized scoring rule in plain text. To increase this measure's reliability, we implemented this simple description of the problem (Charness et

[7]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.

al., 2021), and assuring the participants that it is in their best interest to report their true beliefs (Danz et al., 2022) was implemented. verification. The exact wording of this elicitation can be found in figure C1 of the appendix.

Headlines Selection

The online publication AnimalPolitico[8] and VerificadoMX[9] 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 headlines independently. To find headlines that were real but difficult to classify, the authors used NewsGPT[10]. 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 ran a study in Prolific among a population of Mexicans whose first language was Spanish. 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. The list of headlines used in the experiment can be found in the table C.C2 of the appendix.

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.

V. Results

In table 2, we it sis shown the average values of the most important variables grouped by treatment.

A. Confidence

Higher Levels of Confidence

Participants demonstrated higher levels of confidence when classifying **political headlines**, with a notable increase among **government supporters**. This trend indicates that confidence is context-sensitive, with political relevance amplifying participants' certainty in their classifications. Supporters of the government, in

[8]https://animalpolitico.com/verificacion-de-hechos

[9]https://verificado.com.mx/

[10]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

Table 2—Summary by treatment of the main variables. The average (proportion) of each variable is presented for each treatment.

Variable	Control	Individual	Others
Age	19.9	20	20.4
Female	0.452	0.672	0.509
Support Gov	0.21	0.239	0.164
Oppose Gov	0.145	0.209	0.218
Missing	0.029	0.023	0.048
Accuracy Estimate	0.57	0.522	0.548
Accuracy Estimate Others	0.522	0.505	0.494
Correct	0.62	0.606	0.596
Classification (c = a)	0.497	0.507	0.511
WTP	2.82	2.64	2.42
N Participants	62	67	55

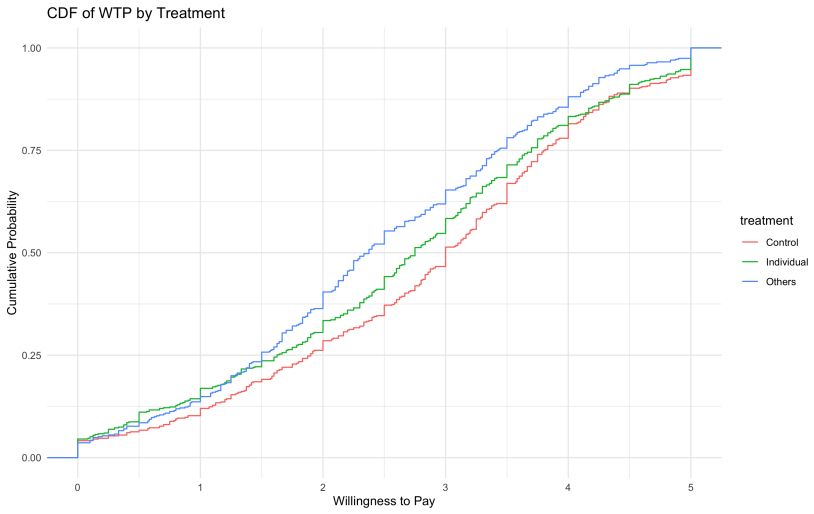

Figure 2. .

particular, tended to exhibit a stronger conviction that their classifications were accurate, especially for politically aligned headlines.

Confidence Decreases with Experience

An analysis across blocks reveals that **confidence declined as participants gained experience** with the task. This reduction in confidence suggests that

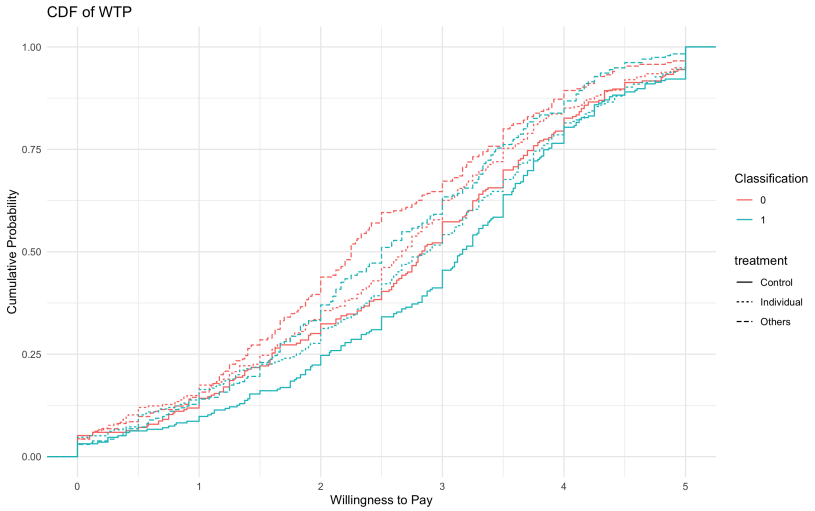

FIGURE 3. .

familiarity with the task did not reinforce belief in accuracy; instead, it appears participants became more cautious, possibly recognizing the complexity or ambiguity in classifying certain headlines. This pattern is consistent with learning effects observed in prior experimental studies, where experience moderates initial overconfidence.

NO EVIDENCE OF FEEDBACK EFFECTS ON CONFIDENCE

Our results indicate **no statistically significant impact of feedback**—whether individual or group—on participants' reported confidence in their classifications. This outcome suggests that feedback in this context did not affect participants' self-assessment of accuracy, even though feedback types varied across conditions. This finding aligns with other studies showing limited feedback effects on confidence when feedback does not address specific decision outcomes.

LACK OF AVERAGE OVERCONFIDENCE

Contrary to the hypothesized overconfidence in headline classification, **participants were not generally overconfident**. Average confidence was in line with observed accuracy rates, indicating a realistic self-assessment. This aligns with prior findings in similar tasks where participants are directly incentivized for accuracy and can adjust beliefs based on task difficulty.

Measuring Confidence

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.

B. Willingness to Pay (WTP) for Verification

Higher WTP for Verification of Believed-True Information

Participants exhibited a greater **willingness to pay (WTP) for verification of information they classified as true**. This indicates a verification bias, whereby participants seek to confirm rather than challenge initial beliefs. This finding is consistent with behavioral patterns in decision-making under uncertainty, where individuals are more likely to invest in information that reinforces their prior beliefs.

Group Accuracy Feedback Reduces WTP

Receiving **group accuracy feedback** significantly lowered participants' WTP to verify classifications. This reduction in demand for verification suggests that collective feedback may diminish the perceived need for individual verification, possibly due to an implicit assumption of greater task simplicity or shared accuracy. This outcome underscores the role of social feedback in shaping verification behavior in information classification tasks.

Increased WTP for Political News Verification

Participants demonstrated an **increased WTP for verifying political news**, emphasizing the perceived importance of accuracy for politically sensitive content. This effect was amplified among **government supporters**, who displayed a higher demand for verification, potentially reflecting a higher stake in the perceived accuracy of politically favorable information. This finding suggests that political alignment influences not only classification behavior but also the demand for verification.

C. Effects on Political Headlines

WTP Patterns in Political News

The WTP results observed across all headlines were preserved when analyzing political news exclusively. However, in contrast to general findings, **political alignment did not significantly affect WTP** within this subset, indicating that demand for verification in politically charged content may be driven by factors other than simple partisan alignment.

Classification Patterns

Further analysis reveals that participants were **more likely to classify a headline as accurate** if it favored the government, a trend especially pronounced among those who opposed the government. This suggests that partisan perception plays a role in classification tendencies, with individuals more likely to label favorable headlines as accurate even when they hold opposing political views. Additionally, for headlines correctly classified initially, participants were more inclined to classify them as accurate, reinforcing the tendency to affirm initial judgments.

VI. 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.

A. 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 field, 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

Table 3—

	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

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

VII. Conclusion

This study provides experimental evidence on the effects of feedback in a news classification task, examining its impact on participants' willingness to pay (WTP) for verification and their confidence in their classifications. Our findings have several implications for understanding how individuals value verification under different feedback conditions, particularly in a context where misinformation is prevalent.

A. Key Findings

1) **Feedback on Group Accuracy and Verification Demand**: Contrary to expectations that feedback would enhance verification efforts, our results indicate that providing feedback on the group's classification accuracy can reduce the value participants place on verification. This result suggests that group feedback may inadvertently signal that individual verification is less critical, potentially diminishing motivation for fact-checking.

2) **Absence of Overconfidence in Experimental Context**: In contrast to general survey-based findings suggesting prevalent underconfidence in assessing information accuracy, our experimental design did not find significant evidence of overconfidence. Participants' self-reported confidence levels closely matched actual accuracy rates, indicating realistic self-assessment. This suggests that feedback interventions addressing overconfidence might be less necessary in contexts similar to our task structure.

3) **Verification Demand and Confirmation Bias**: The experiment also provides evidence consistent with a form of confirmation bias in verification demand. Participants were more likely to pay to verify headlines they believed to be true, implying a preference for confirming rather than challenging initial beliefs. This finding aligns with behavioral patterns observed

in other decision-making contexts and highlights the need for strategies that encourage verification of both supporting and contradicting information.

4) **Asymmetrical Utility Considerations**: Results indicate that participants appear to consider misclassifying fake news as accurate as more detrimental than the reverse, consistent with a utility framework where $U_{AF} < U_{FA}$. This asymmetry in utility aligns with real-world tendencies where the consequences of mistakenly trusting misinformation can be viewed as more impactful.

5) **Effect of Partisan Alignment on Classification Accuracy**: An analysis of classification patterns reveals that participants' likelihood of marking a headline as accurate decreases when the headline favors the government and the participant personally holds opposing political views. This suggests that political alignment can introduce a bias in accuracy judgments, where skepticism is heightened for information contrary to one's ideological stance.

B. Implications and Future Research

Our findings suggest that feedback interventions, particularly those involving peer performance metrics, need careful calibration to avoid unintended decreases in verification demand. The lack of overconfidence observed in our experiment implies that individuals may already possess a reasonably accurate self-assessment of their information classification skills in structured settings. However, the observed confirmation bias in verification demand calls for further exploration into how individuals' motivations for verification might be nudged to encourage unbiased fact-checking.

In practice, developing effective strategies to counter misinformation may benefit from feedback mechanisms that balance accuracy with an emphasis on the importance of verification across all classifications, not just those aligning with preexisting beliefs. Future research could explore alternative feedback designs and assess the long-term impacts of feedback on verification behavior in more naturalistic settings, where stakes are higher, and feedback complexity can more closely reflect real-world challenges.

REFERENCES

Arin, K Peren, Deni Mazrekaj, and Marcel Thum, "Ability of detecting and willingness to share fake news," *Scientific Reports*, 2023, *7298.*

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.

Bateman, Jon and Dean Jackson, *Countering Disinformation Effectively: An Evidence-Based Policy Guide*, Carnegie Endowment for International Peace, 2024.

Charness, Gary, Uri Gneezy, and Vlastimil Rasocha, "Experimental methods: Eliciting beliefs," *Journal of Economic Behavior and Organization*, 2021, *189*, 234–256.

Danz, David, Lise Vesterlund, and Alistair J Wilson, "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 2022, *112*, 2851–2883.

Eberlein, Marion, Sandra Ludwig, and Julia Nafziger, "The effects of feedback on self-assessment," *Bulletin of Economic Research*, 2011, *63*, 307–3378.

Erat, Serhat, Kurtulus Demirkol, and M Eyyüp Sallabas, "Overconfidence and its link with feedback," *Active Learning in Higher Education*, 2022, *3*, 173–187.

Ferraro, Paul J, "Know thyself: incompetence and overconfidence," *Experimental Laboratory Working Paper Series No. 2003-001. Department of Economics, Andrew Young School of Policy Studies, Georgia State University*, 2005.

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

Kartal, Melis and Jean Robert Tyran, "Fake News, Voter Overconfidence, and the Quality of Democratic Choice," *American Economic Review*, 10 2022, *112*, 3367–97.

Kogelnik, Maria, "Performance Feedback and Gender Differences in Persistence," *SSRN Electronic Journal*, 12 2022.

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 Wineburg, "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*.

Moore, Don A. and Paul J. Healy, "The Trouble With Overconfidence," *Psychological Review*, 4 2008, *115*, 502–517.

Oprea, Ryan and Sevgi Yuksel, "Social Exchange of Motivated Beliefs," *Journal of the European Economic Association*, 2022, *20*, 667–699.

Ortoleva, Pietro and Erik Snowberg, "Overconfidence in Political Behavior," *American Economic Review*, 2015, *105*, 504–535.

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.

Pulford, Briony D. and Andrew M. Colman, "Overconfidence: Feedback and item difficulty effects," *Personality and Individual Differences*, 7 1997, *23*, 125–133.

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

Vosoughi, Soroush, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science*, 2018, *359*, 1146–1151.

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

Wilson, Alistair J and Emanuel Vespa, "Paired-uniform scoring: Implementing a binarized scoring rule with non-mathematical language," 2018.

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 ($\Rightarrow$)

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 ($\Leftarrow$)

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 $1 < \frac{P(s=f|F)}{P(s=f|A)} \iff$

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

Materials

C1. Confidence Elicitation

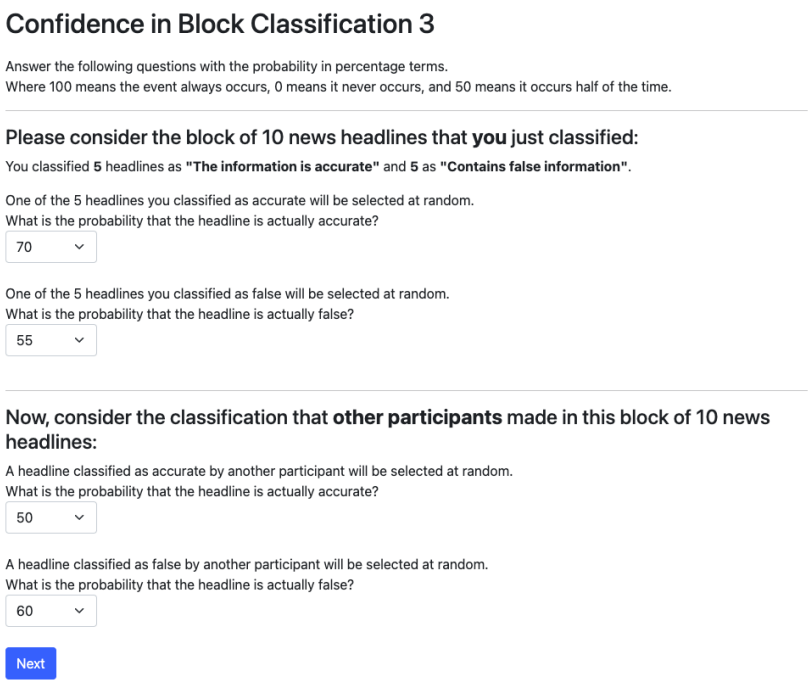

Figure C1. Screenshot of the translated Confidence elicitation as seen by the participants.

C2. Headlines Used in the Experiment

Block	Real	Headline	Translated Headline
1	1	Se inaugura un nuevo museo en honor a Cantinflas en la Ciudad de México	A New Museum in Honor of Cantinflas is Inaugurated in Mexico City

1	1	El salario mínimo en México se incrementa 20% en 2024	Minimum Wage in Mexico Increases by 20% in 2024
1	1	La variante Ómicron es la única de preocupación que circula a nivel mundial; es más transmisible, aunque menos peligrosa que la variante Delta	The Omicron Variant is the Only Variant of Concern Circulating Worldwide; It Is More Transmissible, Though Less Dangerous than the Delta Variant
1	1	Se suspende programa humanitario para trabajar o solicitar asilo en Estados Unidos para Haití, Venezuela, Nicaragua y Cuba	Humanitarian program to work or apply for asylum in the United States for Haiti, Venezuela, Nicaragua, and Cuba is suspended
1	1	Incrementó en el uso de energías renovables en México	Increase in the Use of Renewable Energy in Mexico
1	0	Turismo internacional se desploma en 2024, México ya no es un destino atractivo	International Tourism Collapses, Mexico is No Longer an Attractive Destination
1	0	Luto en México por accidente aéreo de un avión de pasajeros. No hubo sobrevivientes	National Mourning in Mexico. Terrible Passenger Plane Crash in 2024. No Survivors
1	1	En 2024, se intensificaron los incendios forestales en México	In 2024, Wildfires Intensified in Various Regions of Mexico
1	1	Existen programas de apoyo a pequeñas empresas lanzados por el gobierno mexicano	There Are Support Programs for Small Businesses Launched by the Mexican Government
1	0	Se firma un tratado del Foro Económico Mundial que busca reconocer la pedofilia como orientación sexual	A World Economic Forum Treaty Is Signed to Recognize Pedophilia as a Sexual Orientation
2	1	El INAH cobrará $60 por tomar fotografías para uso comercial en museos y sitios arqueológicos	INAH Will Charge $60 for Taking Photos in Museums and Archaeological Sites for Commercial Use
2	0	Se publica la lista de apellidos que pueden solicitar la ciudadanía española	Spain Publishes a List of Surnames That Allow One to Apply for Spanish Citizenship
2	0	En Irán censuraron los Juegos Olímpicos; todas las mujeres aparecen con rectángulos o asteriscos cubriéndolas	Iran Censored the Olympics; All Women Appear with Rectangles or Asterisks Covering Them
2	0	Iniciará juicio en contra de la ministra presidenta de la SCJN por participar en el paro de trabajadores del Poder Judicial.	Trial Against the Chief Justice of the Supreme Court for Participating in the Judicial Workers' Strike to Begin

2	0	La Organización de Estados Americanos (OEA) sanciona a México por dar asilo a Jorge Glass en la embajada mexicana en Ecuador	The Organization of American States (OAS) Managed to Sanction Mexico for Granting Asylum to Jorge Glass in the Mexican Embassy
2	0	El hijo de Nicolás Maduro es captado en video manejando un Ferrari dorado	Nicolás Maduro's son is seen driving a golden Ferrari
2	1	El Ozempic, promovido en redes para bajar de peso, es un tratamiento controlado para la diabetes tipo 2	Ozempic Is Actually a Controlled Treatment for Type 2 Diabetes
2	1	México alcanza cifra récord en exportaciones agrícolas	Mexico Reaches Record High in Agricultural Exports
2	0	Atletas ucranianos portaron pulseras de tobillo con GPS para evitar su huida después de los Juegos Olímpicos de París 2024	Ukrainian Athletes Wore GPS Ankle Bracelets to Prevent Them from Fleeing After the Olympic Games
2	1	Ningún país ha declarado confinamiento por mpox tras la nueva emergencia sanitaria anunciada por la OMS	No country has declared a lockdown due to mpox following the new health emergency announced by the WHO
3	0	La cama "antisexo" siguió siendo utilizada en los Juegos Olímpicos de París 2024	The "Anti-Sex" Bed Will Continue to Be Used at the Paris 2024 Olympics
3	0	El aeropuerto de Suecia fue descontaminado debido a contagios de Mpox	Sweden's Airport Was Decontaminated Due to Mpox Infections
3	0	Miss Venezuela protestó ante las cámaras contra su gobierno en una alfombra roja	Miss Venezuela Protested Against Her Government on a Red Carpet
3	1	Avances en la investigación de nuevas vacunas desarrolladas en México	Advances in the Research of New Vaccines Developed in Mexico
3	0	Un estadounidense se suicidó saltando desde su habitación durante el Baja Beach Fest 2024 en México	An American Committed Suicide During the Baja Beach Fest 2024 in Mexico
3	1	Descubrimiento de nuevas ruinas mayas en la península de Yucatán en 2024	Discovery of New Mayan Ruins in the Yucatán Peninsula
3	1	México termina en el puesto 65 del medallero en los Juegos Olímpicos de París 2024	Mexico Finishes 65th in the Medal Table at the Paris 2024 Olympic Games

3	1	Mexico envió dos aviones en 2023 para rescatar connacionales varados en Israel por el conflicto en Gaza	Sedena and SRE Sent Two Planes to Rescue Mexicans Stranded in Israel
3	0	Consumir alimentos alcalinos ayuda a contrarrestar la variante Omicron del coronavirus	Maintaining a pH (Acidity Level) Above 5.5 Can Prevent Covid-19 Infection
3	0	El Consejo para Prevenir y Eliminar la Discriminación (COPRED) busca suspender la celebración del Día del Padre en los centros educativos	COPRED Urges Elementary Schools Not to Exclude Children from Non-Normative Families on Father's Day
4	1	México tiene la tasa más baja de desempleo de la OCDE	Mexico Has the Lowest Unemployment Rate in the OECD
4	0	Gobierno de México entrega el nuevo "Bono Mujeres" por 2 mil 700 pesos	Mexican Government Issues the New "Women's Bonus" for 2,700 Pesos
4	1	Se registró una disminución de 5.1 millones personas en pobreza en el actual gobierno	A Decrease of 5.1 Million People in Poverty Was Recorded During the Current Government
4	0	Tribunal Electoral encuentra irregularidades graves en el triunfo de Claudia Sheinbaum	Electoral Tribunal Finds Serious Irregularities in Claudia Sheinbaum's Victory
4	1	La presidenta electa Claudia Sheinbaum anuncia beca universal para estudiantes de nivel básico	President-Elect Claudia Sheinbaum Announces Universal Scholarship for Elementary School Students
4	0	El Tren Maya se completará sin impacto ambiental, según estudios científicos independientes	The Maya Train Will Be Completed Without Environmental Impact, According to Independent Scientific Studies
4	1	11 Ministros de la Suprema Corte de Justicia de la Nación ganan $206,246 pesos mensuales netos	11 Supreme Court Justices Earn $206,246 Pesos Monthly Net
4	1	México envió dos aviones a Israel para rescatar a las y los mexicanos varados por el conflicto con Palestina.	Mexico Sent Two Planes to Israel to Rescue Mexicans Stranded Due to the Conflict with Palestine
4	1	En solo ocho de cada 100 delitos en México se abre una carpeta de investigación	Only Eight Out of Every 100 Crimes in Mexico Lead to an Investigation Being Opened
4	0	México prepara una reunión con los presidentes de Rusia y Corea del Norte para comprar armas	Mexico Prepares a Meeting with the Presidents of Russia and North Korea to Buy Weapons
5	1	En el gobierno de AMLO se logró una reducción en la tasa de homicidios.	AMLO's Government Achieved a Reduction in the Homicide Rate

5	1	Primer director femenino de la CFE es nombrado en México	First Female Director of the CFE Is Appointed in Mexico
5	0	Durante el gobierno de Andrés Manuel López Obrador, la deuda pública subió 64%	During Andrés Manuel López Obrador's Government, Public Debt Increased by 64%
5	0	Poder Judicial de la Federación hay 53,737 personas que ganan más que el Presidente	In the Federal Judiciary, 53,737 People Earn More Than the President
5	1	La pobreza extrema en México incrementó de 2018 a 2022	Extreme Poverty in Mexico Increased from 2018 to 2022
5	1	Hay déficit presupuestario en el 2024 por parte del gobierno de México	Mexico's Government Faces a Budget Deficit in 2024
5	1	EU critica al Gobierno de AMLO por 'desacreditar a periodistas'	U.S. Criticizes AMLO's Government for 'Discrediting Journalists'
5	1	Sedena gastó más que el presupuesto autorizado por el Congreso	Sedena Spent More Than the Budget Authorized by Congress
5	0	México ya produce el 90% de la gasolina que consume, como afirma AMLO	Mexico Now Produces 90% of the Gasoline It Consumes, As Claimed by AMLO
5	0	Metro de CDMX dejará de ser gratis para adultos mayores	CDMX Metro Will No Longer Be Free for Senior Citizens

Regressions Appendix

D1. 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.