Metacognition Biases Information Seeking in Assessing Ambiguous News

Valentin Guigon, Marie Claire Villeval, Jean-Claude Dreher

Metacognition Biases Information Seeking in Assessing Ambiguous News

Valentin Guigon[1,2], Marie Claire Villeval[2,3,#] and Jean-Claude Dreher[1,#]

[1]Neuroeconomics lab, Institut des Sciences Cognitives Marc Jeannerod (ISCMJ), CNRS UMR 5229 and Université Claude Bernard Lyon 1, 69 Bd Pinel, 69500 Bron, France.

[2]CNRS, Université Lumière Lyon 2, Université Jean-Monnet Saint-Etienne, emlyon business school, GATE, 35 Rue Raulin, 69007, Lyon, France.

[3] IZA, Bonn, Germany.

September 29th, 2024

*: Corresponding author: dreher@isc.cnrs.fr

#: Equal contribution

Abstract

How do we judge the veracity of ambiguous news, and does our metacognitive ability in making such judgments guide our decisions to seek additional information? The mechanisms by which individuals assess the veracity of ambiguous news and subsequently decide whether to acquire extra information to resolve ambiguity remain unclear. In a controlled experiment, participants evaluated non-partisan, ambiguous news and decided whether to seek more information. Individuals' confidence in their veracity judgments did not predict actual accuracy, showing limited metacognitive ability when facing ambiguous news. Despite this, confidence in one's judgment was the primary driver of the demand for additional information about the news. Lower confidence predicted a stronger desire for extra information, regardless of the veracity judgment. Two key news characteristics (imprecision and polarization) led individuals to confidently misinterpret both true and fake news. News imprecision and news tendency to polarize opinions increased the likelihood of misjudgment, highlighting individuals' vulnerability to ambiguity. Structural equation modeling revealed that the demand for disambiguating information, driven by uncalibrated metacognition, may become increasingly ineffective as individuals are drawn in by the ambiguity of the news. Our results underscore the importance of metacognitive abilities in mediating the relationship between assessing ambiguous information and the decision to seek or avoid further information. These findings suggest that interventions aimed at reducing ambiguity in news content and training individuals to calibrate their confidence in judgments could be effective strategies for combating misinformation.

Main Text

Introduction

The unprecedented growth of the internet and social media platforms has been accompanied both by an abundance of content and by the spread of misinformation[1–4]. Misinformation, characterized by false, inaccurate or misleading information, yields devastating consequences at the societal level, fueling polarization and fostering resistance to crucial initiatives such as climate action and vaccination efforts[5–7]. Contrary to disinformation, misinformation does not need to be created deliberately to mislead. The inherent imprecision of misinformation frequently blurs the perceived boundaries between true or false information. Moreover, its capacity to create an illusion of consensus can inadvertently undermine individuals' ability to discern information authenticity. The appearance of credibility stemming from unified perspectives often obstructs critical evaluation, leaving individuals susceptible to unwittingly compromising their assessment of information authenticity[8]. On the contrary, content that appears to divide opinions may create an illusion of unlikeliness, encouraging people to dismiss the information. These two characteristics, imprecision and polarization, increase ambiguity[9,10], amplifying the difficulty to discern between true and false information. Individuals not only face challenges when having to evaluate the veracity of the information they are exposed to, they also struggle to effectively search for extra information to verify social and political claims, sources, and evidence[11]. A number of strategies have been

proposed to prevent the spread of misinformation[12–14], including fact-checking, directing attention to accuracy[15–17], censorship, encouraging more selective sharing by individuals[18,19] or capping the number of others to whom messages can be forwarded[20]. Yet, fact-checking at the speed and scale of today's platforms is often impractical for private companies or government agencies[20]. An alternative approach could involve targeting individuals themselves and focusing on enhancing their abilities to assess the veracity estimation[21,22].

Here, we were interested in understanding the cognitive mechanisms and the relationships between individuals' judgment about the veracity of ambiguous news (news for which the probability of being true is unknown) they are exposed to, the confidence in such judgment, and the willingness to seek additional information to better assess news veracity.

The willingness to gather extra information rather than sticking to one's current knowledge may depend critically on the subjective confidence in one's assessment. This has been demonstrated in the domain of perceptual decision making[23,24] but remains to be studied regarding real news. In perceptual decision tasks, confidence in one's judgment accuracy plays a direct and determinant role in one's willingness to sample more evidence to update one's beliefs[24–27]. In those tasks, subjective confidence in one's judgment accuracy correlates closely with objective accuracy[28–32]. However, in real-world situations, such as assessing the veracity of media news, the role of confidence in information search remains unclear, as is the relationship between subjective confidence and objective accuracy in the assessment of news veracity. Because decision accuracy and confidence are typically highly correlated[24,33,34], it is difficult to identify whether confidence causally influences the demand for extra information. Crucially, to provide such evidence concerning real news, one needs to demonstrate that confidence in one's judgment predicts the demand for extra information, while controlling for objective performance accuracy in judging news as true or false. To that purpose, we designed our experiment using ambiguous news contents, deliberately leading to performance accuracy at chance level when assessing the veracity of these news. Understanding these relationships requires experimentalists to meticulously select news and rigourously control for their level of ambiguity.

Our focus was on news reflecting ambiguous information that could be either false or true, with varying levels of perceived ambiguity. For example, 'Some well-known athletic committees still deliberately maintain certain sports-related discriminations,', 'In 2018, greenhouse gas emissions decreased by 4% in France for the first time', and 'The most common voting system in Europe is the "preferential vote," which involves ranking candidates in order of preference'. Importantly, these news were sourced from agents with no intention to deceive, thereby ruling out a deliberate willingness to propagate fake news. Examples of naïve agents endorsing and spreading false inaccurate information with no intention to deceive abound[35]. Recent research report that only a very small percentage of people purposely endorse sharing misinformation online[36]. We designed an incentivized within-subject experiment in which non-ego relevant news varied in content imprecision and propensity to polarize opinions. Specifically, these news only concerned non-partisan news, that is, news unrelated to political parties. We introduced variations of ambiguity in stimuli at the time of designing the task. Ambiguity was subjectively assessed by a separate sample of 55 participants during a pre-testing. Participants were presented with a set of brief news about ecology, democracy and social justice taken from the press that could be either true or false. Participants had to evaluate the veracity of each brief news and report their confidence in their

judgment on a continuous scale, using a probability elicitation incentivized method. Then, participants had to decide on whether acquiring or not additional information about this news (to be received after the task was performed), and report their willingness-to-pay to have their information-seeking choice implemented (**Fig. 1**). Importantly, this latter procedure ensured that the willingness to acquire or not acquire extra information is equally balanced. This is unlike previous procedures used in the perceptual decision making domain for which acquiring extra information was costly but for which not acquiring extra information was free (e.g., [25]).

We investigated the relationship between objective accuracy in judging news veracity and confidence in this judgment, controlling for the role of news ambiguity in the judgments of news as true or false. Specifically, we first tested whether confidence in one's judgment about news veracity predicts the success in judging news veracity. Under varying ambiguity, we anticipated uncalibrated metacognition, with participants' confidence uncorrelated with actual success. We then tested whether such confidence in one's judgments drives the demand for extra information about these news. In line with the perceptual decision making literature on confidence-based information-seeking[24,25,37], we predicted that the demand for extra information should increase when confidence in one's veracity assessment is at the lowest. In our context, confidence rather than beliefs is anticipated to play a pivotal role. Finally, we performed a moderated mediation analysis to reveal the relationships between the mechanisms underlying judgments of news veracity and the mechanisms underlying the demand for extra information about the news.

---------- Insert Figure 1 about here ----------

Materials and Methods

Participants

269 participants with no history of neurological or psychiatric disorders participated in this online experiment run on Testable.org. Data were collected in two waves. A first one took place with 80 participants in November 2020. A second one with 189 participants spanned from December 2021 to January 2022. Except for additional questions in the final questionnaire, there were no differences in the experimental design between the two waves. Participants to the experiment were students from business school and engineering schools regularly registered in the GATE-Lab pool of experimental subjects, at the University of Lyon, France. They were paid on average $15.92, including a $9 show-up fee, for an experiment that lasted 46 minutes on average. In total, two participants were excluded from the analyses due to outlying response times ("RT") during news evaluation (one subject: RT = 51.79 ± 26.35; one subject: RT = 1.93 ± 1.31) compared to the mean response time (14.41 ± 8.44). Nine participants were excluded because they did not complete the final questionnaire. In total, 258 participants, aged 18 to 34 years, were included in the statistical analyses (127 males, mean age ± SD = 21.9 ± 2.78).

The study was not preregistered. The study was approved by an internal ethics review board and complied with the European data protection regulation (GDPR). Informed consent was obtained from all subjects prior to participation.

Task and Design

To select our stimuli, we set-up a pre-test of every stimulus with independent raters and kept the stimuli that best fitted our criteria (mean agreement=4.66, SD=1.44) (see Supplementary II). The resulting dataset had an average success rate, as calculated per stimulus, of 51.58% (SD=20.24%). The most difficult stimulus to evaluate had a 6.92% success rate and the easiest stimulus to evaluate had a 93.75% success rate. We computed the scores of news content imprecision (M=5.28, SD=1.32) and propensity to polarize (M=6.41, SD=1.62) using Intra-Class Correlations, with 11 raters and 42 heterogeneous samples. The recommended number of raters ranges between 3, with at least 30 heterogeneous samples[38], and 20[39]. Overall, our procedure closely follows the practical guide of Pennycook and colleagues for behavioral research on fake news and misinformation[40]. In addition, we ran a sentiment analysis on all stimuli, separating for true and false news. Out of the 96 stimuli, 93.75% of the news were predominantly categorized as emotionally neutral. The distribution of false information, rated as negative by the sentiment analysis, was broader compared to the distribution of true information. Conversely, we found the opposite trend for information rated as positive (see Supplementary VI.1, Fig. S12).

Individuals' worldviews have been shown to explain what they believe to be true[41]. As a proxy for beliefs, we adapted measures of social distance between individuals to the relationships individuals may maintain with organizations. In our context, a closer social distance would mean more involvement in the concerns related to the themes. To do that, as in [42], we measured closeness and liking. Unlike for the latter, as it may be difficult to interpret in the context of a human being's relationship to an organization, we replaced similarity with familiarity with the organization. This approach was based on the premise that the more an individual is involved in the concerns related to a particular theme, the more that individual will be aware of the different actors operating within that theme. To have a proxy of such prior beliefs we instructed participants in the first part of the experiment to rate various political organizations that were related to the different news domains. We selected 12 organizations active in the domains of ecology, democracy or social justice. Each organization was described by a 1000-character (±20%) statement taken from the organization websites, with minimal manipulation of the original website content. Participants indicated with six responses their liking, familiarity and closeness of values concerning organizations in direct connection with the topics of the news., on a scale from 0 to 7 (Supplementary III.1). For each topic, we selected two organizations aligned with concerns related to the news, and two organizations misaligned with them (See Supplementary IV). Organizations were presented to participants in a randomized order.

We computed the participants' adhesion to each organization (as a proxy of the knowledge of the domain) by aggregating their six responses in a score that was normalized on a scale from 0 to 100. The higher the score, the more likely the participant was to adhere to the organization and be knowledgeable about its domain of activity. After rating the organizations, participants read the instructions on the task.

The second part of the experiment involved a two-stage task (Supplementary III.2). The first stage included the veracity judgment task. Participants were divided into two groups that received

48 different stimuli each. Each of the 48 trials started with a fixation cross on the screen (Fig. 1). Then, a brief news, either true or false, was displayed. Participants were asked to report what was, in their opinion, the number of chances out of 100 that this news was true or false. Their response revealed their degree of confidence in their judgment. To respond, participants moved a slider either to the left (False) or to the right (True). The slider started at -100 on the left side and ended at +100 on the right side. Participants could not respond with 0. Thus, each move in a direction incremented their degree of confidence by 1%. The elicitation of probabilities was incentivized, following the Karni procedure[43]. This elicitation method is incentive-compatible, considered relatively easy to understand[44] and frequently used in economics[45,46]. With this elicitation method, truthful reporting is the unique dominant strategy. We adopted an approach used in previous experiments[47,48] that employs a narrative involving robots to explain the process to participants (Supplementary III.2). Participants were informed that, after the experiment, we would randomly draw eight trials and reward correct veracity judgments in these trials. To be more specific, for each selected trial, considering the participants have reported their confidence $\mu \in [1{,}100]$ regarding their judgment of the news veracity, the elicitation mechanism selected a random number r from a uniform distribution on $[1{,}100]$. If $\mu \geq r$, the participants earned a payoff $\beta \coloneqq 50\ ECU$ if their veracity judgment was correct and a payoff $\beta \coloneqq 0\ ECU$ if their veracity judgment was incorrect. If $\mu < r$, the payoff was determined by a lottery $(r, 50\ ECU, 0\ ECU)$. Participants understood that if r exceeded their reported confidence μ, the outcome would rely on the lottery; otherwise, their own judgment would be used. Participants were informed that truthful reporting was in their best interest. Each correct veracity judgment in this context earned a reward of 50 ECU, with 100 ECU valued at $2. In the narrative, r corresponded to the accuracy level of a robot randomly drawn from a pool of 100 robots.

The second stage corresponded to the elicitation of the demand to receive extra information. After validating their veracity judgment and while their screen was still displaying the brief news, participants were asked to choose between receiving or not additional information related to the same news after the completion of the experiment. Finally, they had to report how much they were willing to pay, between 0 and 25 ECU of their 200 ECU initial endowment, to have their decision implemented (*i.e.*, to receive or not receive further information), using the Becker–DeGroot–Marschak (BDM) procedure[49]. We kept the cost constant between reception choices to compare them while controlling for scaling or anchoring effects. This aligns with literature showing that people may value ignorance and are even willing to pay not to receive information. In the case participants opted for more information, regardless of whether the information was true or false, they were eligible for receiving a debunk article investigating the content of the brief news in details. Debunk articles were taken from the French fake news debunk platforms *Les Décodeurs du Monde*, *AFP Factcheck* and *Libération Checknews* from the period 2017-2020. The additional information was sent by email to the participants after the experiment. Participants were informed about the mechanism and it was always possible to avoid receiving information for sure by paying the maximum amount. If participants did not pay enough to avoid receiving more information, participants received more information about the news, contrary to their decision. All these aspects were made common knowledge before participants made their choices. This choice of design was motivated by the findings in the literature that people may wish not to receive information that decreases the ambiguity they would rather maintain (i.e., valuation of ignorance[50–55]).

At the end of the experiment, we randomly selected eight trials among the 48. For each selected trial, if the participant's Willingness-To-Pay (WTP) was equal or above a randomly selected price between 0 and 25 (each price had an equal probability to be drawn), the program deducted the randomly selected price from his or her 200 ECU endowment and his or her decision was implemented. If the WTP was lower than the price, no deduction was operated and the option the participant did not choose was implemented. Using a bidding mechanism such as the Willingness-to-Pay instead of relying on stated preferences, such as ordinal ranking (i.e., simply choosing to receive or not to receive), reduces the likelihood that participants did not provide sincere responses.

After reading the instructions on the task, the participants filled in a comprehension questionnaire about these instructions (Supplementary III.3).

At the end of the experiment, participants had to fill in several questionnaires allowing us to measure notably their exposition to information and their degree of curiosity (see Supplementary V). Epistemic curiosity may respond to the desire to stimulate positive feelings of intellectual interest or the desire to reduce undesirable states of information deprivation[56]. To check the relationship between veracity assessment, the demand for further information and epistemic curiosity, we administered the Litman questionnaire of Epistemic Curiosity[56]. Participants in the second wave of data collection answered additional questions about their perceived share of fake news circulating on Internet and social media. The objective was to check for a potential relationship between distrust in channels of information and veracity estimations. Sex (biological attribute assigned at birth) and nationality data was collected by Testable.org as self-report measures. Gender data (shaped by social and cultural circumstances) and ethnicity data was not collected.

Data Analysis

Data was analyzed with custom code on MATLAB R2020b, R 3.4.1 and Python 3.11.5. Data met the assumptions of the statistical tests. Detailed checks and validations are provided in the Supplementary material.

We computed power for first-wave (N=79) data and simulated power for sample sizes up to 250 participants. We employed Mixed Linear Models (MLMs) of the confidence hypothesis, controlled for the veracity judgment and the interaction of news veracity with news theme. With $\alpha = .05$, the observed fixed effect of confidence on information-seeking choices ($\beta = -0.15$) replicates findings from the literature on confidence-based information-seeking[24] and yields a power = .99. For an estimated fixed effect twice lower ($\beta = -0.72$), a simulated N = 150 approximates a power = .99. For an estimated fixed effect three times lower ($\beta = -0.48$), a simulated N = 200 approximates a power = .99. We fixed a total sample size of N=250 to adequately test study hypotheses and included 5-10% additional participants to account for potential outliers and dropouts (Supplementary VI.2).

After collecting data from the second wave, Bayesian analyses were conducted, modeling responses using beta-binomial or normal distributions with non-informative Jeffreys priors. Participant behavior consistency across groups and sessions was confirmed, leading to data pooling (Supplementary VI.3 & VI.4).

To control for objective performance accuracy in veracity judgments, we compared the success proportion in estimating veracity against a random distribution using a logistic function within a Bayesian framework. Our null hypothesis assumes a distribution of behaviors equivalent to randomness. We tested the probability of success at $p = .5$ and computed a Bayes factor to compare $p = .5$ and not $p = .5$. We defined a logistic function with priors for *lambda* = 0.5 and *rscale* = 0.5, iterating 10,000 times. Although this rscale is considered a medium value, it represents a tight distribution around the mean in our case. We also computed a logistic function with *rscale* = 1.5 for a wider distribution.

We tested our hypotheses of participants' behavior using repeated measures MLMs. We modelled success in estimating veracity (correct or incorrect), veracity judgment (true or false), confidence (level per trial), demand for more information (choice to receive or not), and Willingness-To-Pay (ECUs amount per trial). The random structure of our MLMs included random effects for participants. Registering to the experiment required respecting our inclusion criteria. However, we failed to make reporting age, sex and education mandatory when fulfilling the socio-demographics fields at the beginning of the experiment. In total, three participants did not report their age, four did not report their sex, and 28 did not report their education. When accounting for the socio-demographics, we excluded 30 participants from the models.

We also tested Bayesian hypotheses of success in estimating veracity through separate Bayesian multilevel linear models (Table 1), aligning with models formulated for null hypothesis significance testing. Each model included the variables of interest, a simplified random structure (subject random effects) to save computation time and weakly informative priors (see Supplementary I.2). Models were compared using information criteria, particularly the Widely Applicable Information Criterion (WAIC), which measures predictive accuracy for a new dataset and penalizes models based on their parameter count. Bayesian stacking was employed to average Bayesian predictive distributions, with model weights derived from their information criteria performance, indicating their probability of being the best in terms of out-of-sample prediction[57].

Finally, we ran a multiple moderated mediation model (see Figure 5). We used a single model using bootstrapping to evaluate the significance of indirect effects across varying levels of the mediator and moderators. News content imprecision and propensity to polarize were the predictor variables, with veracity judgment moderating and confidence mediating their effects. Reception choice was the outcome variable. Confirmatory factor analysis ensured measurement adequacy and all factor loadings except news content propensity to polarize exceeded 0.6, while composite reliability and average variance extracted surpassed recommended thresholds (0.7 and 0.5, respectively)[58].

Results

Judgment of news veracity and success rate

We first confirmed that performance accuracy was at chance level when judging the veracity of news. Participants' average success (i.e., correctly judging a true news as true and correctly judging a false news as false) rate was of 51.6% (SD=6.7)% (Supplementary I.1, Fig. S1). A comparison of the performances with a random distribution within a Bayesian framework confirmed that

performances were at chance level. Modelling random responses with a logistic function $\lambda \sim logistic(\lambda_0, rscale)$ with priors λ_0= 0.5 and rscale =1.5, the Bayes Factor (BF) favored the null hypothesis of chance-level by a factor of about 0.3. This factor is considered the low boundary of a moderate evidence[59], however posteriors probabilities fell within the range of [49-53]% success, centered around 51.5%. (Supplementary I.1, Fig. S3 & Table S2).

Next, we examined in what conditions participants' successes deviated from chance level. Although judgment of ambiguous news veracity was equivalent to chance, participants performed better with true news (M=64%, SD=11.9%) than false news (M=39.1, SD=12%). Lowest accuracy was for democracy-related news (48.6 ± 11.5), with a slightly higher accuracy for news related to ecology (M=52.2%, SD=11.3%) and social justice (M=53.8%, SD=11.8%) (See Supplementary I.1, Table S3). Binomial Mixed Linear Models (MLMs) showed that participants predicted true news significantly more accurately than false news (odds-ratio = 2.77, p<0.001), with the veracity of news interacting with its theme (p<0.001). Democracy-related news had significantly lowest accuracy compared to ecology and social justice (odds-ratio respectively at 1.19 and 1.26, p=0.0013 and p<.001) (see Fig. 2A). Effects remained highly significant (p<0.005) after controlling for socio-demographics, veracity judgment, and confidence (Supplementary I.1, Table S4).

Such relatively higher ability to assess true news accurately can be explained by a general tendency to declare information as true (M=59.5%, SD=10.6%), with slightly more true news declared as true (M=60.9, SD=12.7%) than false news (M=58.2%, SD=3.1%) (See Supplementary I.1, Table S5). Analyses supported that the veracity judgment predicted veracity perception (p=0.003), withstanding the inclusion of control variables (Supplementary I.1, Table S4). Interestingly, binomial MLMs of veracity judgment revealed that participants were especially more likely to judge as true ecology-related (prob. = 0.678, SD=.02) and social justice-related news (prob. = 0.637, SD=.02) than democracy-related news (odds-ratios respectively 1.71, p<.001, and 1.43, p<0.001) (see Fig. 2B).

---------- Insert Figure 2 about here ----------

Our analysis of the relationship between accuracy in judging news veracity and confidence in this judgment showed that participants' confidence did not significantly predict actual success nor veracity judgments (all p>.05).

Results instead demonstrated that individuals' responses were primarily influenced by news ambiguity, specifically both the imprecision and the polarization of news content, potentially leading to a perception of falsity (Supplementary I.1, Fig. S4). We modelled success in veracity judgment with MLMs incorporating imprecision and polarization predictors in interaction with news veracity (Supplementary I.1, Table S7). Note that the news content imprecision and propensity to polarize (from 0 to 10) were obtained from ratings of a group of subjects (n=55) independent from the actual participants in the experiment (see Methods). The interaction effect of each predictor had a highly significant effect on the success of veracity judgment likelihood (both p<0.001). Specifically, success in judging true news increased when their content imprecision and propensity to polarize were at their *minimum* (minimum/maximum, imprecision odds-ratio = 1.82; polarization odds-ratio = 2.17) Conversely, for false news, success increased with *maximal* imprecision (minimum/maximum odds-ratio = 0.53, p<0.001) and *maximal* propensity to polarize

(minimum/maximum odds-ratio = 0.22, p<0.001). Furthermore, MLMs of veracity judgments showed that the likelihood of judging news as true decreased with increased imprecision (p<0.001, odds-ratio = 0.78) and the propensity to polarize (p<0.001, odds-ratio = 0.64). The effects in all models withstood the inclusion of socio-demographics, veracity-theme interaction, and confidence (Supplementary I.1, Table S7).

Finally, we found that alignment of beliefs with news concerns, distrust in experts and socio-demographics had no significant effect on the accuracy of veracity judgments. Using MLMs (see Method; Supplementary I.1, Fig. S5, Table S8 & S9), response times, likely reflecting cognitive reflection, showed a positive effect on judgment accuracy (p=.007, odds-ratio = 1.07), albeit not robust to the inclusion of other factors. We used Bayesian inference hypothesis testing to support these findings. Comparing Bayesian versions of the regression models (see Method, Supplementary I.2, Fig. S6-S9), the winning model featured interaction terms between news veracity and both news content imprecision and propensity to polarize (see Table 1). Overall, individuals' accuracy deviated from chance level in reaction to variations in news ambiguity. Precision and apparent consensus about news content were interpreted as a signal of veracity, while imprecision and apparent polarization were seen as signals of falsity. Note that we found no significant difference between true and false news either in terms of imprecision (M== 5.53, SD=1.24 vs. M=5.17, SD=1.25, $ranksum$, p=0.09), or in terms of polarization (mean ± SD= 6.61 ± 1.62 vs. 6.22 ± 1.62, $ranksum$, p=0.3).

Uncalibrated metacognitive sense of confidence

To further investigate the relationship between confidence and accuracy in estimating veracity, we examined participants' calibration, that is their ability to accurately estimate the chances that the news is true or false (see Supplementary I.3, Table S10). The confidence-accuracy calibration reflects, for given veracity judgments (the news is evaluated as true or false), the relationship between the continuous scale of confidence ([1,100]) and the binary outcome (true or false). This calibration indexes the extent to which confidence in one's judgment predicts the accuracy of this judgment. A perfect calibration is characterized by a linear confidence-accuracy function with 100% accuracy for 100% confidence, 90% accuracy for 90% confidence, etc. We sorted the individual confidence-accuracy relationships into ten bins and represented an area of well-calibrated estimation that spanned 10% (see Fig. 3).

We expected that participants' confidence would be non-calibrated and uncorrelated with actual success in estimating the veracity of ambiguous news. As the plot shows, participants' accuracy in estimating veracity was independent from their confidence in their estimation. Participants were neither well-calibrated, nor ill-calibrated for estimating probabilities. Values above the diagonal signal underconfidence (individuals have a higher proportion of correct guesses than their reported level of confidence) while values below the diagonal reveal overconfidence (individuals have a lower proportion of correct guesses than their reported level of confidence). Fig.3 shows that underconfidence dominates for degrees of confidence below 50% whereas overconfidence dominates for degrees of confidence above 50%. Underconfidence dominates for true news whereas overconfidence dominates for false news, while news veracity judgment did not affect the relationships between confidence and success levels (Supplementary I.3, Fig. S10).

---------- Insert Figure 3 about here ----------

To understand the determinants of confidence during estimation of news veracity, we examined the sources of variability using MLMs of alignment of beliefs with concerns related to the news, socio-demographics and response times. Moreover, we examined models predicting effects of imprecision and polarization predictors on confidence in veracity judgments.

Alignment of beliefs with news concerns and socio-demographics showed no significant effects on confidence (Supplementary I.3, Table S11), while effects of response times were highly significant and negative (p<0.001). Importantly, the interaction of both ambiguity predictors with the judgments of news as true decreased confidence (p<0.001), even after including control variables (Supplementary I.3, Table S12). Sex also revealed higher confidence levels for males than females ($p<0.$001).

Confidence ratings were reliably affected by news content ambiguity, reflecting higher confidence that a news is true under low ambiguity and higher confidence that a news is false under high ambiguity. Comparing confidence levels between judgments of true and false news across three different levels of content imprecision and propensity to polarize revealed a significant effect of these variables on confidence. Even after the inclusion of control variables, confidence was higher for judgments of the news as *true* when imprecision was lowest (t ratio = 3.85, p<.001) and median (t ratio = 2.60, p=.0092). In contrast, confidence was not significantly different for judgments of the news as *false* than for judgments of the news as true when imprecision was at its highest level (t ratio = -1.84, p=.065). Conversely, confidence was higher for judgments of the news as *true* when the news content propensity to polarize was at its lowest level (z ratio = 8.61, p<.001), but higher for judgment of the news as *false* when polarization was highest (z ratio = -8.34, p<.001). Finally, confidence was not significantly different between the two types of judgments for a median polarization level (p=.57). These findings support the use of imprecision and polarization as signals of falsity, influencing veracity estimation.

Demand and avoidance of extra information

Next, we analyzed the demand for or avoidance of extra information about news that might resolve ambiguity. We predicted that despite the lack of calibration (i.e., a low degree of fit between confidence in news veracity judgment and the actual accuracy), individuals would use their metacognitive sense of confidence to decide whether or not to demand extra information about the news. Hence, we expected confidence to primarily explain the demand for extra information, particularly when confidence was low. First, we present participants' reception choices and subsequent Willingness-To-Pay (WTP). Then, we explore linear relationships between confidence and reception choices/WTP. To test our hypothesis, we estimated separate MLMs with variables capturing main and interaction effects of participant confidence and news veracity judgment. The dependent variables were the binomial demand for more information or the continuous WTP. Post-hoc comparisons were conducted on estimated marginal means.

82.9% of participants demanded extra information at least once, with an average frequency of 42.29% (SD=31.9%). Choice of extra information did not significantly differ between news themes (Kruskal Wallis Chi square = 4.39, p = .11, df = 2; democracy: M=41.04%, SD=33.16%; ecology: M=43.27%, SD=33.67%; social justice: M=42.56%, SD=33.09%; see Supplementary I.4, Table S13). Participants chose to receive extra information 42.51% (SD=32.44%) of the time when news were judged as false and 42.07% (SD=32.14%) of the time when news were judged as true. Bayesian modeling of reception choices between judgments (Jeffreys priors: $\alpha = 0.5$, $\beta = 0.5$) revealed a negligible difference (delta = 0.23, 95% Credible Interval [-0.008, 0.012]), indicating similar demand for extra information regardless of veracity judgments.

Participants exhibited a higher willingness-to-pay (WTP) for receiving extra information (M=7.07 ECU, SD=4.96 ECU) compared to not receiving it (M=5.75 ECU, SD=5.69 ECU) (see Fig. 4; see Supplementary I.4, Table S14). Bayesian models of WTP for receiving and not receiving extra information (Jeffreys priors: $\mu = 0$, $\sigma = 1$ from half-Cauchy distribution) showed that participants were willing to pay more to receive it than to avoid it (delta = 1.327, 95% Credible Interval [-2.302, -0.344]).

As predicted, confidence explains the demand for extra information ($p<0.001$, odds-ratio = 0.59), with a significant negative interaction with veracity judgment ($p<0.001$). These effects remained significant even after incorporating controls such as the interaction of news veracity and theme and socio-demographics (Supplementary I.4, Table S15). The results show that the probability of demanding extra information is not affected by news content ambiguity (i.e., imprecision and propensity to polarize) (see Fig. 4A) while it decreases as confidence in one's judgment increases. Specifically, the decrease is more pronounced when the news is judged as false (minimum/maximum confidence; judgment as false, odds-ratio = 6.41; judgment as true, odds-ratio = 2.59) (see Fig. 4B). A regression analysis of WTP further supported these findings, revealing a significant interaction between confidence and the demand for information ($p<0.001$) (see Fig. 4C), holding up against the inclusion of control variables (Supplementary I.4, Table S16). According to this model, the effect size of confidence (*minimum – maximum* confidence levels) on the WTP when participants opted not to receive extra information was -1.74 ($p<0.001$), whereas the effect size for the WTP to receive extra information was only -0.13 ($p=0.69$) (see Fig. 4D).

The alignment of beliefs with news concerns from only two organizations predicted reception choices while we found no evidence for effects of sociodemographics, response times, distrust or ambiguity on decisions to seek information that might resolve ambiguity about the news.

To sum up, there is a significant inverse relationship between the demand for extra information about the news and confidence in one's judgment about news veracity. Moreover, this relationship is stronger for the news that participants judged as false. Supporting these findings, participants are also willing to pay more to not receive more information about what they think they already know.

---------- Insert Figure 4 about here ----------

A moderated mediation analysis further extended the role of confidence in the estimation of ambiguous news veracity (Table 2, Fig. 5 ; Supplementary I.5, Fig. S11). Confidence had a unique

direct effect on the outcome reception choice (standardized interaction β = -0.15, Z = -13.96, p < .001). Its effect was specifically a mediator effect, whereby the ambiguity of news, that is, news content imprecision and news content propensity to polarize, had an indirect effect on the reception choices through the confidence (imprecision: standardized interaction β = -0.06, Z = -3.93, p < .001; polarization: standardized interaction β = 0.11, Z = 6.32, p < .001). Veracity judgment played a role by moderating the effect on the news content imprecision to confidence path (standardized interaction β = 0.1, Z = 2.43, p = .015) as well as the effect on the news content propensity to polarize to confidence path (standardized interaction β = -0.36, Z = -8.34, p < .001). This analysis shows that the uncalibrated metacognition operating during the evaluation of true and false news induces a demand for disambiguating information that is increasingly ineffective as individuals are lured by the ambiguity of the news.

---------- Insert Figure 5 about here ----------

Discussion

Headlines in the real world often do not overtly appear true or false, but instead fall into an ambiguous gray area, which makes them more difficult to evaluate. Using a novel experimental design, we carefully selected non-partisan and non-ego relevant news that offer various levels of content imprecision and polarization. The study was designed to address the complexity of evaluating ambiguous news, particularly within the context of misinformation. By varying the ambiguity of objectively verified true and false headlines, we controlled for subjective biases and ensured a range of cognitive responses. This approach allowed us to study metacognitive processes in a non-partisan context, avoiding reliance on extreme, obvious headlines, or emotionally charged content. Although 93.75% of the stimuli were predominantly categorized as neutral by our sentiment analysis, we observed a slight skew in emotional valence. False information exhibited a broader range of negative sentiment ratings, while true information showed a wider range of positive ratings. This subtle difference aligns with literature suggesting that false information often evokes more negative emotions[1]. Emotions may promote belief in fake news[60]. Despite this, the overall sentiment distribution remained largely neutral, indicating that the skew did not impact the neutrality of the materials.

Participants' accuracy in assessing news veracity hovering at chance level confirmed that we manipulated news with ambiguous contents, thereby allowing us to disentangle the effects of confidence from the effects of objective performance accuracy. We focused on news about ecology, democracy and social justice whose utility was mainly cognitive[51]. That is, we chose news that could help individuals to form more accurate beliefs about the state of the world, and that would neither threaten their identity nor affect their perception of how others would see them. A sentiment analysis confirmed the neutrality of the stimuli emotional valence (Supplementary VI.1, Fig. S12). The reason was to restrict as much as possible distortions in the demand for extra information that would result from motivated reasoning to protect one's image or identity.

How do individuals judge the veracity of ambiguous news? Participants' confidence did not predict their actual accuracy, and they systematically overestimated the prevalence of true news in

the task. This inclination, also known as truth bias[61,62], could stem from the automatic acceptance of statements and the cognitive strain associated with reevaluating previously acknowledged information[63]. It may also be that individuals are inclined to regard information as correct if it is deemed "good enough", avoiding a costly in-depth analysis[64,65]. An alternative perspective suggests that evolution has shaped human communication towards truthfulness, with altruism and trust as norms to ensure cooperation[66]. For instance, children tend to initially trust social partners[67]. Moreover, some defend that there is a prevailing inclination toward intuitive honesty among humans[68], leading individuals to anticipate a higher frequency of true statements in the information they encounter. It may also be that participants' held a baseline assumption that information is true, given the prevalence of true information people encounter[69–71].

While truthful communication is essential, signals must also convey useful information in the presence of ambiguity. Epistemic vigilance[72] has been proposed as an evolutionary tool, encouraging individuals to critically assess the veracity of statements. Our study reveals that participants consider ambiguity dimensions like content imprecision and polarizing tendencies. Higher imprecision and propensity to polarize increased the likelihood of individuals mistakenly declaring news as false with confidence. Caution should be exerted when drawing conclusions from the measures of ambiguity. ICC values indicated moderate reliability of raters on the scoring of news content imprecision and good reliability for the propensity to polarize. However, our results are consistent with previous research showing that individuals disproportionately prefer information that would provide a sense of certainty[73]. The imprecision in information content may signal unreliability, as it provides less clarity in the verifiability of the assertion whereas in the face of conflicting information, content polarization may signal untrustworthiness. Ambiguous content could hinder coordination and impose cognitive strains, leading individuals to preferentially identify such content and avoid it as an epistemic strategy for truth-seeking. The prominence of these dimensions, especially in comparison to alignment with beliefs or distrust toward experts, is consistent with the fact that we manipulated news with a primary emphasis on cognitive utility.

Participants' metacognitive abilities were uncorrelated with success in estimating news veracity and we observed that their confidence-accuracy calibration was flat (Fig. 4). The news stimuli have been chosen to ensure performance at chance on average, with news that ranged in evaluation difficulty between very easy and very hard. Confidence usually strongly correlates with objective accuracy in perceptual decision tasks or adaptive behavior[24,28,29]. However, the relationship between one's accuracy of judgment and one's confidence about judgment is known to vary greatly with task difficulty, whereby confidence is decreasingly predicting accuracy as difficulty increases[74–77]. This could be the case when the false bit of a news is not the central idea but a peripheral idea of the news. The dissociation that we observed between confidence and actual success rate suggests a pattern specific to ambiguous news, in contrast with perceptual information, with individuals struggling to gauge their level of knowledge when confronted with potential misinformation. In tasks with perceptual information, the state of the world is directly accessible and potentially identifiable with enough time to accumulate evidence. In the case of textual information, prior knowledge could theoretically aid in the evaluation of stimuli. However, active engagement and motivation to consider prior knowledge are necessary for effective evaluation[5]. Such processes do not occur routinely during comprehension of textual stimuli. A key factor that

may influence whether individuals engage in careful evaluation is their beliefs about their own susceptibility to misinformation, pointing to an additional metacognitive dimension [78].

Crucially, although individuals held an inaccurate perception of their own knowledge, this metacognitive sense of confidence was the most decisive dimension that guided information-seeking behavior in our experiment. Participants were willing to pay more to not receiving more information about news that they estimated they already knew to be false. These results suggest that the decision to seek additional information likely stems from the expected benefit of this additional information in terms of subsequent cognition and reduction of ambiguity about the state of the world. This key finding presumably reflects that individuals use ambiguity – reflected in their confidence in their judgment – to choose whether to gather more evidence[24–27,37,79]. To fully contextualize our findings, it is important to acknowledge a limitation of the design. Participants could choose to ignore the post-task email or not read the additional information it might contain, rather than paying to avoid receiving additional information. This behavior could not be controlled and stems from the manipulation of written news media. If the reception of information had been endogeneous to the task, verifying that it was read would have required conditioning rewards on responses to questions about the news. However, this would have shifted the utility of the information from cognitive to instrumental, which would be adversarial to testing our hypotheses.

The present study provides empirical evidence indicating the challenges individuals face in distinguishing true from false ambiguous news, often confusing precise or consensual information with truth. Our novel findings underscore the prime role of metacognitive abilities in mediating the relationship between ambiguous information assessment and the demand or avoidance of extra information. Individuals misjudge what they know but they also seek to receive information according to what they know. As a consequence, they misidentify shortfalls in their knowledge, preventing them from filling the gaps. Individuals lacking awareness of their susceptibility to inaccurate information may fail to engage the correct evaluation strategies[78]. This demonstrates that individuals are not only at risk of receiving undetected false information but also inefficiently explore their environment, potentially spreading false information upon sharing it[35]. While previous literature suggests that people share false information due to a lack of attention to accuracy[16,17], our study suggests that their search for information to reduce ambiguity is driven by misplaced confidence in their veracity judgment. The structural equation modeling suggests that this search is increasingly ineffective as individuals are lured by the ambiguity of news. This findings are all the more important as our societies are facing major challenges with the extremely fast technical development of generative AI and the spread of deepfakes that will make the identification of veracity more and more difficult in the immediate future. These findings demonstrating the importance of meta-cognition in the assessment of the veracity of ambiguous news and in the search for information is very consistent with recent research showing a pivotal role of metacognition in belief updating in sensitive domains, such as in politically contested domains[80].

Our results highlight potential interventions and modifications to social media features that complement existing approaches for addressing misinformation and detecting truth. They call for testing within education and media literacy programs[81] approaches targeting individuals' ability to estimate veracity and to engage in self-motivated extra information seeking[78]. This includes exploring methods that encourage people to estimate their confidence in news content and validate it against evidence to increase awareness[22]. It includes as well training with specific search

heuristics[11,82–85] and probability calibration exercises to help people improving their assessment of their own knowledge and their need for further information-seeking[83,84,86]. These interventions complement news content moderation, signaling of trustworthiness, and changes in the incentive structure of media platforms,[12,13,87,88] aiming both to decrease motivations to share content that receives high social reward at the cost of accuracy and to increase accuracy motivation[17,36].

Disclosure statement

The authors confirm that all conditions, data collection procedures, data exclusion procedures and sample size determination procedures have been reported. All collected data and questionnaires have been reported either in the manuscript or in the supplementary materials. The instructions to the task and the questionnaires, translated into English, are all available in the supplementary materials. Exposition to information has been collected in the post-task questionnaire but not analysed due to being outside the scope of this manuscript.

Data availability

All raw and processed data used for the main analyses and supplementary information are freely accessible in .csv format via OSF: https://osf.io/436pq/.

Code availability

The custom code used to produce the results are freely accessible format via OSF: https://osf.io/436pq/. All analyses were carried out with MATLAB version R2020b, Python version 3.11.5 and R version 4.3.1.

References

1. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science (80-.).* **359**, 1146–1151 (2018).

2. van der Linden, S. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat. Med.* **28**, 460–467 (2022).

3. Cinelli, M. *et al.* The echo chamber effect on social media. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

4. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, 1–7 (2020).

5. Rapp, D. N. & Salovich, N. A. Can't We Just Disregard Fake News? The Consequences of Exposure to Inaccurate Information. *Policy Insights from Behav. Brain Sci.* **5**, 232–239

(2018).

6. Tsfati, Y. *et al.* Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. *Ann. Int. Commun. Assoc.* **44**, 157–173 (2020).

7. Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends Cogn. Sci.* **25**, 913–916 (2021).

8. Brady, W. J., Crockett, M. J. & Van Bavel, J. J. The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspect. Psychol. Sci.* **15**, 978–1010 (2020).

9. Baillon, A., Cabantous, L. & Wakker, P. P. Aggregating imprecise or conflicting beliefs: An experimental investigation using modern ambiguity theories. *J. Risk Uncertain.* **44**, 115–147 (2012).

10. Pushkarskaya, H., Smithson, M., Joseph, J. E., Corbly, C. & Levy, I. Neural Correlates of Decision-Making Under Ambiguity and Conflict. *Front. Behav. Neurosci.* **9**, 1–15 (2015).

11. McGrew, S., Breakstone, J., Ortega, T., Smith, M. & Wineburg, S. Can Students Evaluate Online Sources? Learning From Assessments of Civic Online Reasoning. *Theory Res. Soc. Educ.* **46**, 165–193 (2018).

12. Bak-Coleman, J. B. *et al.* Combining interventions to reduce the spread of viral misinformation. *Nat. Hum. Behav.* (2022). doi:10.1038/s41562-022-01388-6

13. Globig, L. K., Holtz, N. & Sharot, T. Changing the incentive structure of social media platforms to halt the spread of misinformation. *Elife* **12**, 1–23 (2023).

14. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).

15. Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychol. Sci. Public Interes.* **21**, 103–156 (2020).

16. Pennycook, G. & Rand, D. G. Nudging social media sharing towards accuracy. *Ann. Am. Acad. Pol. Soc. Sci.* **700**, 152–164 (2022).

17. Capraro, V. & Celadin, T. “I Think This News Is Accurate”: Endorsing Accuracy Decreases the Sharing of Fake News and Increases the Sharing of Real News. *Personal. Soc. Psychol. Bull.* **49**, 1635–1645 (2023).

18. Guess, A. M. & Munger, K. Digital Literacy and Online Political Behavior. *Polit. Sci. Res. Methods* **11**, 110–128 (2023).

19. Traberg, C. S., Roozenbeek, J. & van der Linden, S. Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *Ann. Am. Acad. Pol. Soc. Sci.* **700**, 136–151 (2022).

20. Jackson, M. O., Malladi, S. & McAdams, D. Learning through the grapevine and the impact of the breadth and depth of social networks. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).

21. Lichtenstein, S. & Fischhoff, B. Training for calibration. *Organ. Behav. Hum. Perform.* **26**, 149–171 (1980).

22. Callender, A. A., Franco-Watkins, A. M. & Roberts, A. S. Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition Learn.* **11**, 215–235 (2016).

23. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: Distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).

24. Desender, K., Boldt, A. & Yeung, N. Subjective Confidence Predicts Information Seeking in Decision Making. *Psychol. Sci.* **29**, 761–778 (2018).

25. Schulz, L., Fleming, S. M. & Dayan, P. Metacognitive computations for information search: Confidence in control. *Psychol. Rev.* **130**, 604–639 (2023).

26. Desender, K., Murphy, P., Boldt, A., Verguts, T. & Yeung, N. A postdecisional neural marker of confidence predicts information-seeking in decision-making. *J. Neurosci.* **39**, 3309–3319 (2019).

27. Balsdon, T., Wyart, V. & Mamassian, P. Confidence controls perceptual evidence accumulation. *Nat. Commun.* **11**, (2020).

28. Boldt, A. & Yeung, N. Shared neural markers of decision confidence and error detection. *J. Neurosci.* **35**, 3478–3484 (2015).

29. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. R. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science (80-.).* **336**, 670 (2010).

30. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).

31. Maniscalco, B., Peters, M. A. K. & Lau, H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, Psychophys.* **78**, 923–937 (2016).

32. Pasquali, A., Timmermans, B. & Cleeremans, A. Know thyself: Metacognitive networks and measures of consciousness. *Cognition* **117**, 182–190 (2010).

33. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 1–9 (2014).

34. Meyniel, F., Schlunegger, D. & Dehaene, S. The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Comput. Biol.* **11**, 1–25 (2015).

35. Serra-Garcia, M. & Gneezy, U. Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies. *Am. Econ. Rev.* **111**, 3160–3183 (2021).

36. Pennycook, G. *et al.* Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).

37. Boldt, A., Blundell, C. & De Martino, B. Confidence modulates exploration and exploitation in value-based learning. *Neurosci. Conscious.* **2019**, 1–12 (2019).

38. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **15**, 155–163 (2016).

39. Mondal, D., Vanbelle, S., Candel, M. J. J. M. & Cassese, A. Review of sample size determination methods for the intraclass correlation coefficient in the one-way analysis of variance model. **33**, 532–553 (2024).

40. Pennycook, G., Binnendyk, J., Newton, C. & Rand, D. G. A practical guide to doing behavioral research on fake news and misinformation. *Collabra Psychol.* **7**, 1–13 (2021).

41. Tappin, B. M., Pennycook, G. & Rand, D. G. Bayesian or biased? Analytic thinking and political belief updating. *Cognition* **204**, 104375 (2020).

42. Liviatan, I., Trope, Y. & Liberman, N. Interpersonal similarity as a social distance dimension: Implications for perception of others' actions. **44**, 1256–1269 (2008).

43. Karni, E. A Mechanism for Eliciting Probabilities. *Econometrica* **77**, 603–606 (2009).

44. Li, S. Obviously Strategy-Proof Mechanisms. **107**, 3257–3287 (2017).

45. Charness, G., Gneezy, U. & Rasocha, V. Experimental methods: Eliciting beliefs. *J. Econ. Behav. Organ.* **189**, 234–256 (2021).

46. Schotter, A. & Trevino, I. Belief Elicitation in the Laboratory. *Annu. Rev. Econom.* **6**, 103–128 (2014).

47. Coffman, K. B. Evidence on self-stereotyping and the contribution of ideas. *Q. J. Econ.* **129**, 1625–1660 (2014).

48. Möbius, M. M., Niederle, M., Niehaus, P. & Rosenblat, T. S. Managing Self-Confidence : Theory and Experimental Evidence. (2022).

49. Becker, G. M., DeGroot, M. H. & Marschak, J. Measuring utility by a single-response sequential method. *Behav. Sci.* 226–232 (1964). doi:10.1002/bs.3830090304

50. Charpentier, C. J., Bromberg-Martin, E. S. & Sharot, T. Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7255–E7264 (2018).

51. Sharot, T. & Sunstein, C. R. How people decide what they want to know. *Nat. Hum. Behav.* **4**, 14–19 (2020).

52. Hertwig, R. & Engel, C. Homo Ignorans: Deliberately Choosing Not to Know. *Perspect. Psychol. Sci.* **11**, 359–372 (2016).

53. Persoskie, A., Ferrer, R. A. & Klein, W. M. P. Association of cancer worry and perceived risk with doctor avoidance: an analysis of information avoidance in a nationally representative US sample. *J. Behav. Med.* **37**, 977–987 (2014).

54. Kobayashi, K. & Hsu, M. Common neural code for reward and information value. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 13061–13066 (2019).

55. Golman, R., Hagmann, D. & Loewenstein, G. Information avoidance. *J. Econ. Lit.* **55**, 96–135 (2017).

56. Litman, J. A. Interest and deprivation factors of epistemic curiosity. *Pers. Individ. Dif.* **44**, 1585–1595 (2008).

57. Anderson, D. R. & Burnham, K. P. Avoiding Pitfalls When Using Information-Theoretic Methods. *J. Wildl. Manage.* **66**, 912–918 (2002).

58. Fornell, C. & Larcker, D. F. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *J. Mark. Res.* **18**, 39 (1981).

59. Lee, M. D. & Wagenmakers, E.-J. *Bayesian Cognitive Modeling : A Practical Course*. (2014).

60. Martel, C., Pennycook, G. & Rand, D. G. Reliance on emotion promotes belief in fake news. *Cogn. Res. Princ. Implic.* (2020). doi:10.1186/s41235-020-00252-3

61. Clementson, D. E. Truth Bias and Partisan Bias in Political Deception Detection. *J. Lang. Soc. Psychol.* **37**, 407–430 (2018).

62. Brashier, N. M. & Marsh, E. J. Judging Truth. *Annu. Rev. Psychol.* **71**, 499–515 (2020).

63. Gilbert, D. T., Krull, D. S. & Malone, P. S. Unbelieving the Unbelievable: Some Problems in the Rejection of False Information. *J. Pers. Soc. Psychol.* **59**, 601–613 (1990).

64. Reder, L. M. & Kusbit, G. W. Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *J. Mem. Lang.* **30**, 385–406 (1991).

65. Ferreira, F., Bailey, K. G. D. & Ferraro, V. Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* **11**, 11–15 (2002).

66. Tomasello, M. The Ontogenetic Foundations of Epistemic Norms. *Episteme* **17**, 301–315 (2020).

67. Stengelin, R., Grueneisen, S. & Tomasello, M. Why should I trust you? Investigating young children's spontaneous mistrust in potential deceivers. *Cogn. Dev.* **48**, 146–154 (2018).

68. Capraro, V., Schulz, J. & Rand, D. G. Time pressure and honesty in a deception game. *J. Behav. Exp. Econ.* **79**, 93–99 (2019).

69. Marsh, E. J., Cantor, A. D. & M. Brashier, N. Believing that Humans Swallow Spiders in Their Sleep: False Beliefs as Side Effects of the Processes that Support Accurate Knowledge. *Psychol. Learn. Motiv. - Adv. Res. Theory* **64**, 93–132 (2016).

70. Pennycook, G., Allan, J., Nathaniel, C., Derek, B. & Fugelsang, K. J. A. On the reception and detection of pseudo-profound bullshit Gordon. **10**, 549–563 (2015).

71. Acerbi, A., Altay, S. & Mercier, H. Research note : Fighting misinformation or fighting for information ? *Harvard Kennedy Sch. Misinformation Rev.* **3**, 1–15 (2022).

72. Sperber, D. *et al.* Epistemic vigilance. *Mind Lang.* **25**, 359–393 (2010).

73. Ambuehl, S. & Li, S. Belief updating and the demand for information. *Games Econ. Behav.* **109**, 21–39 (2018).

74. Moore, D. A. & Healy, P. J. The Trouble With Overconfidence. *Psychol. Rev.* **115**, 502–517 (2008).

75. Boldt, A., Gardelle, V. De & Yeung, N. The Impact of Evidence Reliability on Sensitivity and Bias in Decision Confidence. *J. Exp. Psychol. Hum. Percept. Perform.* **43**, 1520–1531

(2017).

76. Weber, N. & Brewer, N. Confidence – Accuracy Calibration in Absolute and Relative Face Recognition Judgments. **10**, 156–172 (2004).

77. Moore, D. A. & Schatz, D. The three faces of overconfidence. *Soc. Personal. Psychol. Compass* 1–12 (2017). doi:10.1111/spc3.12331

78. Salovich, N. A. & Rapp, D. N. Misinformed and unaware? Metacognition and the influence of inaccurate information. *J. Exp. Psychol. Learn. Mem. Cogn.* **47**, 608–624 (2020).

79. Lee, D. K. L. & Ramazan, O. Fact-Checking of Health Information: The Effect of Media Literacy, Metacognition and Health Information Exposure. *J. Health Commun.* **26**, 491–500 (2021).

80. Fischer, H. & Fleming, S. Why metacognition matters in politically contested domains. *Trends Cogn. Sci.* **28**, 783–785 (2024).

81. McGrew, S. Teaching lateral reading: Interventions to help people read like fact checkers. *Curr. Opin. Psychol.* **55**, 101737 (2024).

82. Atanasov, P., Witkowski, J., Ungar, L., Mellers, B. & Tetlock, P. Small steps to accuracy: Incremental belief updaters are better forecasters. *Organ. Behav. Hum. Decis. Process.* **160**, 19–35 (2020).

83. Chang, W., Chen, E., Mellers, B. & Tetlock, P. Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgm. Decis. Mak.* **11**, 509–526 (2016).

84. Mellers, B. *et al.* Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychol. Sci.* **25**, 1106–1115 (2014).

85. Donovan, A. M. & Rapp, D. N. Look it up: Online search reduces the problematic effects of exposures to inaccuracies. *Mem. Cogn.* **48**, 1128–1145 (2020).

86. Moore, D. A. *et al.* Confidence calibration in a multiyear geopolitical forecasting competition. *Manage. Sci.* **63**, 3552–3565 (2017).

87. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. *Sci. Adv.* **8**, 1–12 (2022).

88. Celadin, T., Capraro, V., Pennycook, G. & Rand, D. G. Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts. *J. Online Trust Saf.* **1**, 1–20 (2023).

Acknowledgments

This research has benefited from the financial support of IDEXLYON from Université de Lyon (project INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and of

the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche to JCD (ANR-21-CE37-0032 and ANR-24-CE37-4261), and by MITI 2020 CNRS to JCD and MCV. We thank Pr Edmund Derrington for critically reading and correcting English in the draft of the manuscript.

Author information

Authors and Affiliations

CNRS, Neuroeconomics lab, ISCMJ and Université Claude Bernard Lyon 1, Lyon, France.
V. Guigon, J.-C. Dreher
CNRS, Université Lumière Lyon 2, Université Jean-Monnet Saint-Etienne, emlyon business school, GATE, 35 Rue Raulin, 69007, Lyon, France**.**
V. Guigon, M. C. Villeval
IZA, Bonn, Germany.
M. C. Villeval

Contributions

CRediT author statement: Valentin Guigon: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Marie Claire Villeval**: Conceptualization, Methodology, Validation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jean-Claude Dreher**: Conceptualization, Methodology, Validation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Corresponding author

Correspondence to J.C. Dreher: dreher@isc.cnrs.fr

Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Research Transparency Statement

Funding: This research has benefited from the financial support of IDEXLYON from Université de Lyon (project INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche to JCD (ANR-21-CE37-0032, ANR-24-CE37-4261) and financial support from the CNRS through the interdisciplinary programs MITI CNRS (MITI-2020-247719). Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research complies with the Declaration of Helsinki (2013), aside from the requirement to preregister human subjects research, and received approval from an internal ethics review board. This research complied with the European data protection regulation (GDPR). Informed consent was obtained from all subjects prior to participation. Preregistration: The study was not preregistered. Materials: All study materials are publicly available (https://osf.io/436pq/?view_only=1292b9f54f7d41a08f6e7274876ff6ae). Data: All primary data are publicly available (https://osf.io/436pq/?view_only=1292b9f54f7d41a08f6e7274876ff6ae). Analysis scripts: All analysis scripts are publicly available (https://osf.io/436pq/?view_only=1292b9f54f7d41a08f6e7274876ff6ae).

Figures

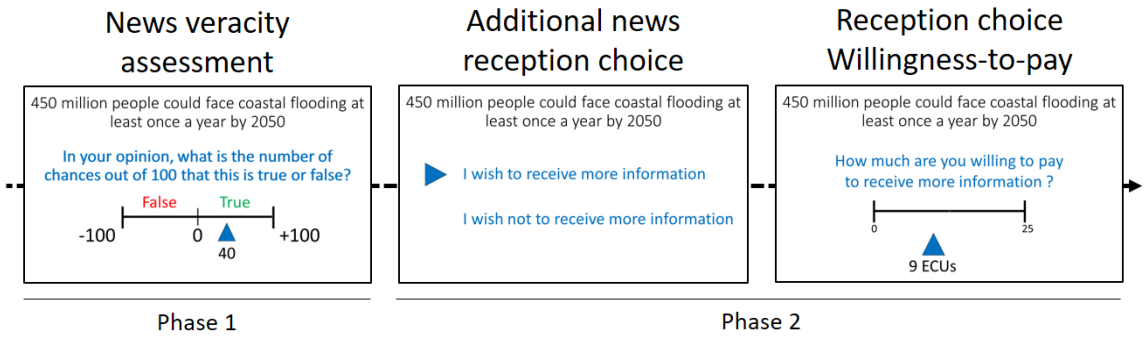

Fig. 1. Description of the task. Participants read a brief news and were incentivized to report the probability that the news was true or false, allowing us to assess both veracity judgment and confidence in one's judgment. A correct evaluation of news veracity (i.e., true news judged as true and false news judged as false) was worth 50 ECU while an incorrect evaluation was worth 0 ECU (eight trials out of forty-eight were selected at random to be paid). Next, participants had to choose between receiving or avoiding receiving more information about the news. Given their choice, they had to indicate how much they were willing to pay (from 0 to 25 ECU) to have this choice implemented (endowment= 200 ECU, eight trials chosen at random to be implemented). A Becker–DeGroot–Marschak (BDM) procedure determined whether their choice would be, or not,

implemented, and at which price, depending on their bid. This procedure ensured that both the demand to receive and the demand to avoid receiving extra information were costly.

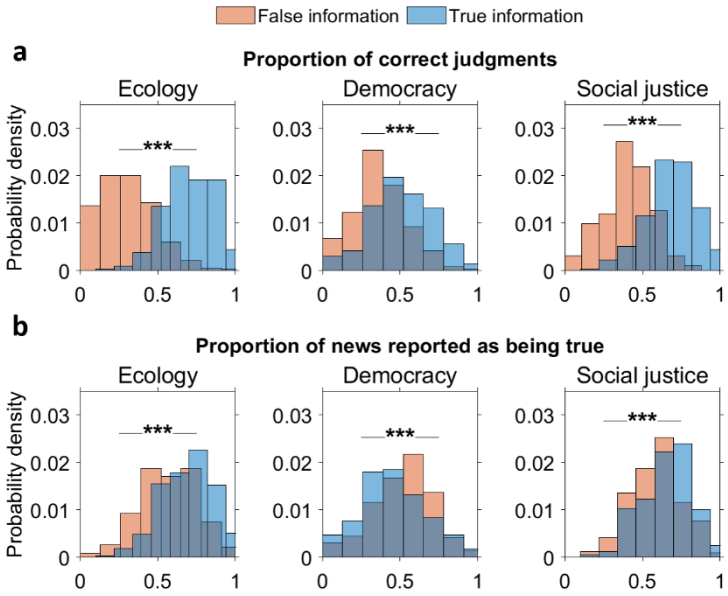

Fig. 2. Distributions of success and of veracity judgment. a) Probability densities of correct veracity judgment (*i.e.,* proportion of false news judged as false and of true news judged as true) are displayed separated by news themes (ecology, democracy, social justice) and news veracity (true or false). Individuals were better at evaluating a news that was true than a news that was false. The likelihood of success was higher for news that were actually true. **b)** Probability densities of news reported as being true are displayed separated by news themes (ecology, democracy, social justice) and news veracity (true or false). There were more news judged as true than false (i.e., Probability Density function skewed to the right), reflecting a bias toward judging news as true, with the exception of democracy-related news.

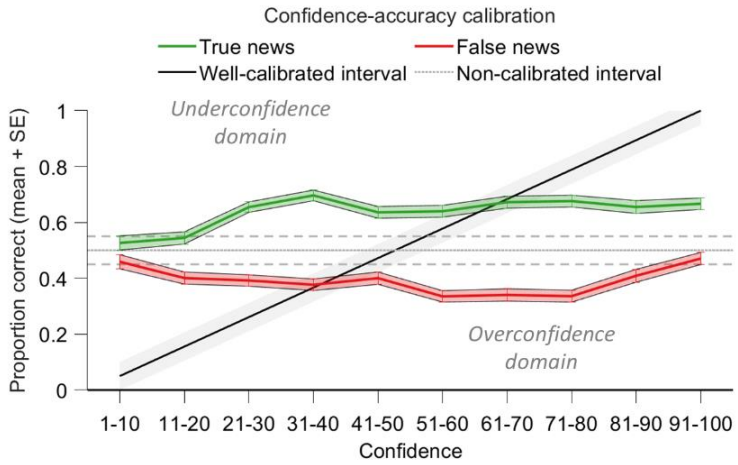

Fig. 3. Calibration analysis (i.e., degree of fit between a person's judgment of performance and his or her actual performance). Participants' metacognition was not calibrated for estimating the probability of news veracity. The confidence-accuracy calibration plot displays the participants' accuracy in estimating probabilities that their judgment was correct as a function of their confidence level. Well-calibrated estimated probabilities would intersect with confidence degrees in the grey area, meaning, for example, that a 0-20 % confidence degree predicts a 0-20 % accuracy in evaluating the news veracity. The plot shows that overall, the proportion of accurate veracity estimations did not increase nor decrease with confidence. Furthermore, the plot emphasizes that accuracy is higher for true news than false ones (the green curve always lies above the red one). Underconfidence dominates for true news whereas overconfidence dominates for false news.

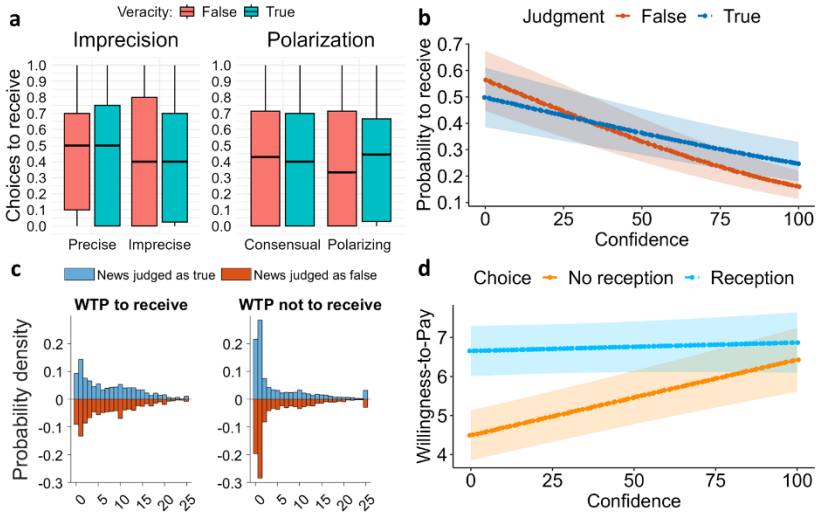

Fig. 4. The likelihood of choosing to receive extra information decreased as confidence in news veracity judgment increased. a) Panels show no decrease in the probability to acquire extra information as imprecision or polarization increases. **b)** The probability to be willing to receive

extra information about the news decreases as the confidence in one's judgment about news veracity increases. This decrease is steeper for news judged as false as compared to those judged as true. **c)** The WTP (max: 25 EUC) was higher for the choices to receive extra information than for the choices to not receive it. **d)** The WTP to receive extra information about the news was not affected by the degree of confidence in one's judgment about news veracity, whereas the WTP to avoid receiving extra information about the news increased with the degree of confidence in judgment about news veracity.

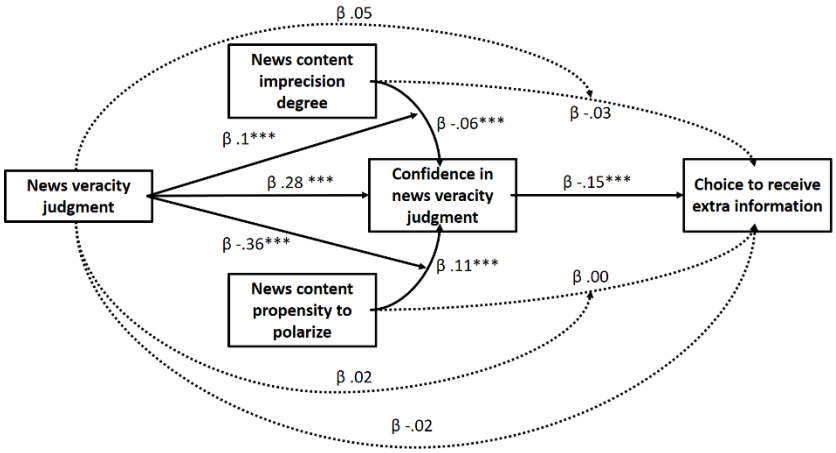

Fig. 5. Mediation effect of confidence, moderated by news content imprecision and propensity to polarize, predict the demand for extra information. News content imprecision and news content propensity to polarize, conditional on the veracity judgment, have indirect effects on reception choices (*i.e.*, decision to acquire extra information about the news) via the confidence in the veracity evaluation. Indirect effects are represented with dotted lines; direct effects are represented with solid lines. The coefficients are standardized. Notes: *p < .05. **p < .01. ***p < .001.

Tables

Table 1: The models of news content imprecision and news content polarization explain best participants' success in assessing the news veracity. The table displays model comparisons ordered by WAIC. The best model has the lowest WAIC, showing best out-of-sample capacity, and higher weight, showing best prediction of in-sample data. Models are described in the Data Analysis section of the Methods.

Model comparison with WAIC

Models	ΔWAIC	ΔSE	WAIC	SE WAIC	pWAIC	Weight
News content imprecision & polarization	0.00	0.00	16890.63	33.54	10.30	1
News content polarization	-30.22	8.06	16951.08	29.88	8.27	0
News content imprecision	-105.50	14.44	17101.64	17.33	8.59	0
Response times	-125.99	16.96	17142.62	11.55	8.86	0
Non-informative beta response	-132.22	16.72	17155.06	9.99	7.36	0
Beliefs alignment	-134.52	16.37	17159.68	7.00	6.25	0
Subject random-effect	-134.56	16.37	17159.75	6.69	6.31	0
Veracity judgment	-135.01	16.34	17160.65	7.26	7.30	0

Table 2: The effect of news content imprecision and polarization, moderated by the veracity judgment, on the reception choices is mediated by the confidence in the veracity judgment. The table displays a moderated mediation model. Confidence is the single variable with a direct effect on the reception choices. Crucially, the confidence mediates the effect of news content imprecision and news content propensity to polarize, conditional on the veracity judgment, on the reception choices. The model is described in Data Analysis section of the Methods. Notes: *b* = unstandardized coefficients. *b** = standardized coefficients. CI = confidence interval.

Outcome	Predictor	*SE*	*Z*	*p*	*b*	95% CI (*b*)	*b**	95% CI (*b**)
Reception	Imprecision	0.01	-1.79	.073	-0.03	[-0.06, 0.00]	-0.03	[-0.07, 0.00]
Reception	Polarisation	0.01	0.23	.815	0.00	[-0.02, 0.03]	0.00	[-0.03, 0.04]
Reception	Confidence	0.00	-13.96	< .001***	-0.01	[-0.01, -0.00]	-0.15	[-0.18, -0.13]
Reception	Judgment	0.11	-0.43	.666	-0.05	[-0.28, 0.16]	-0.02	[-0.13, 0.08]
Reception	Imprecision × Judgment	0.02	0.93	.353	0.02	[-0.02, 0.06]	0.05	[-0.05, 0.15]
Reception	Polarisation × Judgment	0.02	0.38	.707	0.01	[-0.03, 0.04]	0.02	[-0.08, 0.12]
Confidence	Imprecision	0.32	-3.93	< .001***	-1.25	[-1.89, -0.62]	-0.06	[-0.09, -0.03]
Confidence	Polarisation	0.30	6.32	< .001***	1.87	[1.31, 2.44]	0.11	[0.07, 0.14]
Confidence	Judgment	2.56	6.21	< .001***	15.92	[10.87, 20.92]	0.28	[0.19, 0.37]
Confidence	Imprecision × Judgment	0.43	2.43	.015*	1.05	[0.20, 1.95]	0.10	[0.02, 0.19]
Confidence	Polarisation × Judgment	0.37	-8.34	< .001***	-3.08	[-3.81, -2.36]	-0.36	[-0.45, -0.28]