Trust, but Verify:
a Theory of Selective Source Verification

Stefania Merone

November 15, 2026

Abstract

When should individuals verify the reliability of an information source? I consider a Bayesian decision-maker who receives a signal from a source of uncertain quality and can pay a cost to learn its accuracy. I show that verification is only valuable for an intermediate range of prior beliefs where uncertainty about source reliability remains sufficiently high and there is still substantial scope for belief updating. More importantly, verification is selective: when all sources are reliable, sources of signals that contradict prior beliefs are systematically more valuable to verify, implying that asymmetric verification can arise from fully rational behavior. I then characterize how this result changes when one source type is unreliable. Finally, I explore departures from Bayesian updating, and show that motivated reasoning and confirmatory bias are not only unnecessary to explain the verification patterns they are commonly associated with, but may even attenuate or reverse them.

Keywords— belief updating, confirmation bias, costly information acquisition, motivated reasoning, source verification

1 Introduction

Individuals usually obtain information through secondary or tertiary sources that draw on multiple references and synthesize different pieces of information into a coherent discussion. This broad category encompasses both emerging technologies, such as AI overviews and LLM responses, and more traditional sources, including online and offline encyclopedias, monographs, surveys, newspaper articles, and documentaries. However, these sources often draw on information of varying reliability: for instance, AI overviews incorporate content from webpages that differ in their levels of expertise and editorial scrutiny, including several open-content platforms (Chapekis et al., 2026); similarly, investigative reports combine official documents with anonymous tips, eyewitness accounts, and expert interviews. To address the credibility concerns this feature may cause, original sources are often *traceable*, allowing users to identify the source behind each individual statement, as in the case of AI overviews: in these cases, individuals can still benefit from a synthesized, easy-to-read answer, while retaining the option to assess the trustworthiness of the underlying sources when needed. Source verification is a crucial aspect of dealing with uncertain information, as unverified exposure to unreliable sources can distort beliefs and ultimately affect a wide range of economic outcomes such as consumption choices, voting behavior, and financial decisions.

Despite the availability of this option, evidence suggests that people rarely try to check the credibility of original sources. For instance, Wikipedia, one of the most visited websites and arguably the most popular online encyclopedia, has an overall citation engagement rate of just 0.29% (Piccardi et al., 2020). Similarly, a recent analysis of nearly 69,000 Google searches conducted by 900 U.S. adults found that only 1% of searches displaying Google AI Overviews resulted in a click on a link within the summary (Chapekis et al., 2026); also, a comparable pattern has emerged from data attributed to OpenAI, indicating an average click-through rate (CTR) of less than 1%.[1] In-

[1]This measure was reported by the co-founder and CTO of Draft&Goal in this LinkedIn post.

dustry analyses also corroborate this trend, reporting CTRs on cited links in conversational search typically well below 1% (Garanko, 2025; Beus, 2026; Law, 2026). This result persists despite original sources usually being just one click away, even in high-stakes contexts such as when seeking medical information (Hu et al., 2026).

It is important to clarify how source verification, or *source-checking*, differs from the related but distinct practice of *fact-checking*: while the former involves assessing the precision of a source, the latter consists in verifying the correctness of the actual signal received. Although both valuable, these practices are not always both feasible: in particular, fact-checking requires access to a fully revealing source — such as a news agency, an official statistic, or a government record — against which a claim can be checked. In many cases, like the ones mentioned above, this requirement is unlikely to be satisfied; nonetheless, some sources may still be able to make better guesses than others. Thus, source verification becomes the more appropriate practice in these settings.

To address this phenomenon, this paper studies a decision problem where a Bayesian learner receives a report from a potentially unreliable source and optimally chooses whether to verify its precision. My model is deliberately simple: a decision-maker (DM) holds a prior μ over a binary state $\theta \in \{0, 1\}$ and observes a binary signal $s \in \{0, 1\}$ generated by one of two source types: a good source, whose signal matches the true state with probability $q > \frac{1}{2}$, or a bad source, whose signal is correct with probability $p < q$. The source is bad with probability $\varepsilon \in (0, 1)$. After observing the signal, the DM can *verify the source*: she can choose to pay a cost $K > 0$ to learn the source type.

I show that, for the DM to be willing to verify a source, three conditions — each empirically testable with appropriate elicitation methods — must hold simultaneously. First, the cost of verification must be sufficiently low. Second, the DM must give the source the benefit of the doubt: after observing the realized signal, she must remain sufficiently uncertain about the source type. Third, learning the source type must have a sufficiently large impact on her posterior beliefs about

the state, leaving enough scope for belief updating. These conditions are naturally more likely to hold for intermediate priors: indeed, I show that, for a given information environment $\mathcal{I} = \{q, p, \varepsilon\}$, verification cost K, and signal realization s, the set of prior beliefs for which verification is optimal is either empty or a proper subinterval of $[0, 1]$ (Proposition 1). This result mirrors usual findings from models of information acquisition in which the precision of a source is known and the DM decides whether to acquire a signal (Arrow et al., 1949; Keppo et al., 2008).

Additionally, I characterize *which* sources are more valuable to verify, conditional on their reports: since the DM assesses reliability of a source partly based on how its signal aligns with her prior beliefs, the value of verification is asymmetric across signal realizations: all else equal, a signal that confirms the prior makes the source appear more reliable. As a result, when even bad sources are still reliable (i.e., $p \geq \frac{1}{2}$), sources reporting signals that contradict the DM's prior beliefs are always more valuable to verify (Theorem 1).

Instead, when one source is unreliable (i.e. $p < \frac{1}{2}$), the DM does not know if the source can be trusted; so, she is uncertain about how to interpret its signal and, consequently, in which direction to update her beliefs. Then, the ranking of sources of confirmatory and disconfirmatory signals in terms of their value of verification becomes ambiguous, as it depends on both the information environment and the DM's prior beliefs. In particular, I show that this ranking is governed by the signs of two thresholds both measuring the trade-off between the good source's competence and the bad source's deceptiveness, albeit on different scales: when the thresholds have the same sign, the source of one type of signals is always more valuable to verify, regardless of the DM's prior beliefs; whether this is the confirmatory or disconfirmatory report depends on their common sign (Theorem 2). When the thresholds have opposite signs, instead, the ranking depends on the DM's prior beliefs and reverses between intermediate and extreme priors (Theorem 3).

Despite the choice of standard assumptions, including belief updating under Bayes rule, one

of my model's central predictions is apparently behavioral: under suitable conditions, DMs are more likely to verify the sources of signals that contradict their prior beliefs. Although asymmetric response to information is often attributed to biases like motivated reasoning or confirmation bias, my analysis shows that not only neither is necessary for such pattern to emerge, but also that introducing these frictions into the model actually yields somewhat counterintuitive predictions. In Proposition 2, I show that — contrary to the common idea that motivated reasoning should reduce verification of favorable signals while increasing it for unfavorable ones — when a DM exhibits this bias but is initially skeptical about source precision, receiving a signal aligned with the preferred conclusion can increase her uncertainty about the source type, thereby making verification more valuable. On a similar note, Proposition 4 shows that confirmation bias can make DMs appear more impartial in their verification choices by preventing them from recognizing a fraction of disconfirmatory signals as such. Taken together, these results show that behavioral biases are not only unnecessary to explain the verification patterns they are commonly associated with, but may even attenuate or reverse them.

Literature. Formally, this paper contributes to the literature on costly information acquisition (Stigler, 1961; Grossman and Stiglitz, 1980), and, more specifically, to the growing strand on costly verification (Townsend, 1979; Gale and Hellwig, 1985; Ben-Porath et al., 2014). Most models of costly information acquisition, however, assume that, to learn about an unknown state, a decision maker can acquire a costly signal from a source of known precision; then, she chooses to do so if the expected benefit from updating her beliefs outweighs the cost of observing the signal realization. This framework is common across several areas of the literature, including cheap talk (Argenziano et al., 2016; Di Pei, 2015; Barreda, 2024), auctions (Persico, 2000; Shi, 2012), and information design (Gentzkow and Kamenica, 2014). In contrast to this standard assumption, I assume that the signal

realization is freely observed and that the decision maker can pay to learn the source's precision, instead. This change seems particularly natural for settings where individuals are easily exposed to information, such as when scrolling through social media, watching the news, or conducting online research, and must assess the reliability of potentially dubious sources. Also, I focus on a decision-making problem, abstracting away from any strategic interactions. Importantly, this approach also changes the timing of the information acquisition decision: while standard models assume that information acquisition takes place *ex ante*, i.e. before the signal realization, I adopt an *ex post* approach in which the decision is conditional on the observed signal.

This assumption brings my framework closer to the literature on costly verification; however, I depart from standard models by considering a different object of verification. Specifically, these models typically assume that the DM can verify the information itself: upon receiving a signal, she can pay a cost to establish whether its content is truthful. Such verification is often modeled as a technology that allows the DM to learn the underlying state with certainty. While this assumption is natural in settings such as verifiable disclosure (Milgrom, 1981; Grossman, 1981), fake news (Cisternas and Vásquez, 2022), or persuasion with external feedback (Gentzkow and Shapiro, 2006; Godes, 2026), it may be less suitable for questions whose answers are inherently uncertain, such as the effectiveness of a new treatment, the consequences of a policy, or the outcome of an upcoming election; in these cases, a fully revealing source may not be available, even though some sources are more trustworthy than others. Accordingly, rather than allowing the DM to verify the truthfulness of a signal, I assume that she can assess the reliability of its source by paying to learn its precision.

Topically, my paper is also related to the literature on asymmetric responses to information; however, it departs from it in two main ways. First, while most studies attribute this phenomenon to behavioral frictions, such as motivated reasoning (Eil and Rao, 2011; Thaler, 2024) or confirmation bias (Eil and Rao, 2011; Charness and Dave, 2017), I adopt a rational perspective, abstracting away

from behavioral considerations. Second, the economics and psychology research typically examines asymmetric responses to information in terms of asymmetric signal processing and posterior belief updating, either about the source credibility (Lord et al., 1979; Koehler, 1993; Munro and Ditto, 1997; Thaler, 2024) or about the underlying state (Eil and Rao, 2011; Charness and Dave, 2017; Coutts, 2019; Möbius et al., 2022). There are a few exceptions: for instance, Eil and Rao (2011) also examine decisions regarding additional information acquisition. A richer pool of forms of asymmetric responses to information, including asymmetries in information acquisition (Taber and Lodge, 2006), attention (Garrett, 2009), and exposure (Iyengar and Hahn, 2009), is analyzed in the political science literature. In this paper, I study a distinct form of response to information that, to the best of my knowledge, has not been examined in either strand: source verification.

The remainder of this paper proceeds as follows. Section 2 describes the theoretical framework. Section 3 provides an overview of the results. Section 4 extends the theory to incorporate motivated reasoning and confirmation bias. Section 5 concludes.

2 Theoretical Framework

2.1 Environment

I study a decision problem where a Bayesian learner receives a report from a potentially unreliable source, and optimally chooses whether to verify its credibility. The unknown state of the world is $\theta \in \{0, 1\}$, with prior belief $\mu = \Pr(\theta = 1)$. A decision maker (henceforth, DM) receives a signal $s \in \{0, 1\}$ on the state, and knows that it has been generated by one of two possible types of sources, $r \in \{B, G\}$, with the probability of the source being of type B equal to $\varepsilon \in (0, 1)$.

In this context, r defines the degree of *reliability* of the source, i.e., the probability that their signal is going to match the state. Sources of type G are *reliable*, meaning that $\Pr(s = \theta \mid r = G) =$

$q > \frac{1}{2}$: these sources can be thought of as experts, who can still be wrong but are competent enough to make informed guesses. This assumption is particularly natural in settings where the underlying state cannot be known with certainty: for instance, during the early stages of the COVID-19 pandemic in 2020, the effectiveness of vaccines was still uncertain; nevertheless, scientific expertise and experimental trials — despite being subject to sampling variability — enabled scientists to make more informed assessments than the general public. By contrast, sources of type B are assumed to be less precise: $\Pr(s = \theta \mid r = B) = p < q$. Then, two different regimes can arise: either types B are merely *less reliable* (i.e., $p \geq \frac{1}{2}$), or they are *unreliable* (i.e., $p < \frac{1}{2}$). While the former still provide useful information, albeit less precise than that provided by experts, the latter usually generate misleading signals: in the pandemic example, less reliable sources can be thought of as general practitioners who, despite having a medical background, were not directly involved in COVID-19 research. On the other hand, unreliable sources can be associated with some members of the anti-vaccine community who used traditional and social media to spread misinformation about vaccines (Carpiano et al., 2023). Importantly, unreliability is not equivalent to *uninformativeness*: while reliable sources are informative in the sense that their signals are more likely to match the underlying state than to contradict it, unreliable sources are informative in the opposite direction, as their signals are more likely to contradict the state than to match it: in other words, while informativeness requires signals to be correlated with the underlying state, reliability (unreliability) requires this correlation to be strictly positive (negative). Indeed, if the DM knew with certainty that a source was unreliable (i.e. $p < \frac{1}{2}$), she could simply invert its signal and treat it as coming from a reliable source with precision $1 - p > \frac{1}{2}$; thus, the informativeness of a source depends not only on the probability that its signal matches the state, but also on whether the DM knows how to interpret it. In particular, when reliability is uncertain, the DM may be unable to distinguish between signals that should be taken at face value and those that should be inverted.

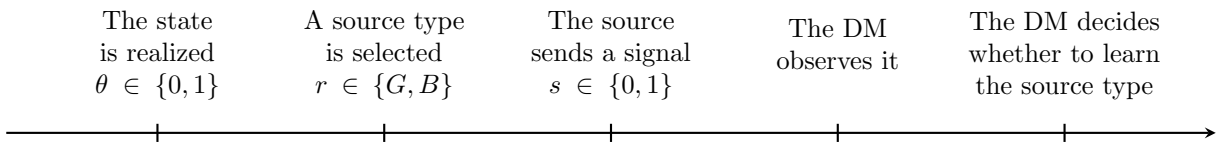


Figure 1: Timeline of the decision problem.

Together, source precisions and probabilities define the *information environment* of the decision problem $\mathcal{I} = (q, p, \varepsilon)$, with the set of admissible information environments $\mathbb{I}$ being defined as:

$$\mathbb{I} = \left\{ (q,p,\varepsilon) \in (0,1)^3 : q > \frac{1}{2} \ \wedge \ p < q \right\}$$

Importantly, I assume that $\mathcal{I}$ is known to the DM. Although restrictive, this assumption is useful because it allows me to disentangle the effects of uncertainty about source types from those of uncertainty about the information environment. Moreover, it can be relaxed by interpreting $\mathcal{I}$ as the DM's subjective assessment of the information environment, rather than its true value: indeed, what ultimately matters for the DM's verification choices is her perceived information environment, regardless of whether her beliefs are correct. How these beliefs are formed, however, is not directly relevant to the current analysis.

2.2 Verification and Updating

After observing signal s, the DM has the option to pay a verification cost $K > 0$ to reveal from which source type it has been generated. I remain agnostic about the nature of this cost: it may reflect the time spent on verification, the cognitive effort involved, the additional resources required for the process, or some combination thereof. A few concrete examples are provided below:

Example 1: AI Assistants and Information Retrieval. An user asks an AI assistant whether a medical treatment is effective ($\theta = 1$) or not ($\theta = 0$). With some probability $(1 - \varepsilon)$,

the assistant correctly summarizes the scientific evidence available online. Since the evidence itself is imperfect, the answer is informative but not necessarily correct ($\frac{1}{2} < q < 1$). However, with complementary probability ε, it will instead retrieve information from non-expert online discussions (e.g. Reddit threads), where aggregate user experience is still positively correlated with the truth ($p \geq \frac{1}{2}$), but less reliable than medical studies ($p < q$). Then, the DM can decide whether to spend time and effort (K) to scan the original source and try to understand whether it is a credible website.

Example 2: Interns and Surgery Decisions. A hospital patient asks the doctor on shift whether she should take elective surgery within the next month ($\theta = 1$) or not ($\theta = 0$). With some probability $(1 - \varepsilon)$, the doctor is an expert in the field ($\frac{1}{2} < q < 1$). However, with complementary probability ε, they are just an intern; so, they still completed medical studies ($p \geq \frac{1}{2}$), but have less practical experience ($p < q$). Then, the patient can decide to spend time and effort (K) to look at their CV online to assess their former experience and infer their degree of trustworthiness.

Example 3: Military Strategy and Corruption. A commander receives an intelligence report about whether an adversary will attack on the left ($\theta = 1$) or on the right ($\theta = 0$). With probability $(1 - \varepsilon)$, the report is based on genuine troop movements and communications; so, it identifies the correct direction of attack with probability $\frac{1}{2} < q < 1$. With complementary probability ε, however, the enemy is trying to deceive the opponent — for example, using decoys or misleading communications. Because such attempts are imperfect, the resulting report is more likely to be wrong than right, but not always so: $0 < p < \frac{1}{2}$. Then, the commander can decide whether to hire extra spies (K) to further investigate on the authenticity of the signal.

Example 4: False Friends and Pieces of Advice. Alice is trying to decide she should get a haircut ($\theta = 1$) or not ($\theta = 0$). Being unsure, she asks Bella, her hairstylist. With probability $(1 - \varepsilon)$, Bella is sincere, and tells her her honest opinion, which could still be wrong but comes nevertheless from a professional ($\frac{1}{2} < q < 1$). However, with complementary probability ε, Bella is still mad at Alice for their last argument and wants her to look bad; so, she reports the opposite of what she truly thinks. Then, before following Bella's piece of advice, Alice can decide whether to buy a present (K) for a friend, Edward, to convince him to investigate whether their last argument has truly been resolved.

Importantly, I assume the verification cost is a fixed, exogenous parameter. This assumption has two implications. First, the cost K does not depend on the signal realization s: this restriction is mostly imposed for simplicity, as signal-dependent verification costs could be accommodated within the framework without fundamentally altering the structure of the decision problem, given that verification choices take place after observing the signal (see Figure 1). Second, the DM cannot influence the magnitude of the verification cost by adjusting the resources devoted to verification. In a more general framework, the DM could choose not only whether to verify, but also how *intensively* to do so; this would introduce an additional endogenous choice, and bring the model closer to the rational inattention literature (see Sims, 2003; 2006). While potentially interesting, characterizing the optimal intensity of verification is beyond the scope of this paper. Furthermore, I assume that the verification cost K is known to the DM. As with the information environment $\mathcal{I}$, this assumption allows me to isolate the effects of uncertainty about source types from other types of uncertainty — in this case, uncertainty about verification costs — and can be relaxed by interpreting K as the DM's subjective assessment of the cost, rather than its true value. The same considerations discussed above regarding the role and formation of such beliefs apply here.

In this setting, verification is *perfect*: by paying the cost K, the DM learn the source type r for

sure. This assumption is particularly plausible in environments where source types are observable, or there exist ways to distinguish between them. For instance, in Example 1, source reliability can be assessed by looking at some features of the website including domain, authors, publication date, URL, layout, and so on.[2] Similarly, in Example 2, source reliability can be assessed through objective data such as graduation mark, specialization, years of experience within the field, and this information is observable for the DM (provided that a CV is available). Conversely, when assessing source reliability primarily relies on subjective judgments or imperfect indicators — such as in Examples 3 and 4 — the assumption of perfect verification becomes more restrictive.

If the DM discovers the source type r, she updates her beliefs to the Bayesian posterior $\mu'_{r,s}$; otherwise, she updates to the posterior of the corresponding mixture model:[3]

$$\mu'_{M,s} = \rho_s \mu'_{G,s} + (1 - \rho_s)\mu'_{B,s} \tag{1}$$

where $\rho_s \coloneqq \Pr(r = G \mid s, \mu)$ denotes the posterior probability that the source type is G, given the realized signal. Importantly, this belief is also updated through Bayes' rule; hence, it depends not only on the observed signal realization, but also on whether such report confirms prior beliefs or not: in particular, a source of type G is more likely to generate reports that agree with the true state; so, if the DM initially considers $\theta = 1$ ($\theta = 0$) to be more likely, a signal equal to 1 (equal to 0) increases the perceived reliability of the corresponding source:

Lemma 1 *For every $\mu \in (0,1)$, $\rho_1 \geq \rho_0 \Leftrightarrow \mu \geq \frac{1}{2}$, with equality at $\mu = \frac{1}{2}$.*

[2]These features belong to the list of questions included in the "CRAAP test", a test to evaluate the nature and value of the information that can be found on and offline (Meriam Library, California State University, Chico, 2010).

[3]In particular, by Bayes rule:

$$\mu'_{G,1} = \frac{\mu q}{\mu q + (1-\mu)(1-q)},\ \mu'_{B,1} = \frac{\mu p}{\mu p + (1-\mu)(1-p)},\ \mu'_{G,0} = \frac{\mu(1-q)}{\mu(1-q) + (1-\mu)q},\ \mu'_{B,0} = \frac{\mu(1-p)}{\mu(1-p) + (1-\mu)p}$$

Importantly, conditional on the realization of the signal, the DM's decision problem reduces to a choice between two information structures, one of which Blackwell-dominates the other (Blackwell, 1951). To see why, observe that if the DM verifies the signal source, her posterior belief is $\mu'_{G,s}$ with probability ρ_s and $\mu'_{B,s}$ with probability $1-\rho_s$; by contrast, if she does not verify, her posterior belief is $\mu'_{M,s}$ with certainty. The two information structures induce the same expected posterior belief, but the former generates a strictly positive variance, whereas the latter is degenerate: thus, the former is a mean-preserving spread of the latter, implying that it is unambiguously more informative in the Blackwell sense.

2.3 Belief Utility

I assume the DM is an expected utility maximizer who values holding accurate beliefs on the state θ, i.e. she wants to discover the truth. I formalize this goal by assuming her belief utility is represented by a quadratic function:

$$u(\mu;\theta) = -(\mu-\theta)^2$$

This is a strictly proper scoring rule: holding fixed the true state θ, utility is the highest when the reported belief is as accurate as possible.[4] Importantly, the quadratic assumption is adopted for the sake of exposition, but results are robust for a large class of proper scoring rules.[5] Then, the expected utility from holding belief μ is:

$$V(\mu) = -Var(\theta)$$

[4]This assumption can be interpreted as the reduced form of a decision problem in which the DM's beliefs determine her optimal action.

[5]In particular, results hold for any scoring rule from the Beta family (Buja et al., 2005). Beta scoring rules are characterized by a weighting function equal to:

$$\omega(t) = t^{\alpha-1}(1-t)^{\beta-1} \quad \text{for some } \alpha > -1, \beta > -1$$

The quadratic case corresponds to $\alpha = \beta = 1$, while the logarithmic score is obtained by setting $\alpha = \beta = 0$.

Notably, since belief utility is a proper scoring rule, $V(\mu)$ is convex in posterior beliefs; thus, it is maximized when beliefs are degenerate ($\mu = 0$ or $\mu = 1$), and minimized at complete uncertainty ($\mu = 0.5$). It follows that potential utility gains from acquiring additional information are naturally largest for intermediate beliefs, and vanish as beliefs become increasingly extreme.

3 Optimal Verification

Under these assumptions, the DM decides whether to verify the signal source by comparing her expected utility from posterior beliefs with and without verification. Specifically, verification is optimal if and only if:

$$\rho_s V(\mu'_{G,s}) + (1 - \rho_s) V(\mu'_{B,s}) - V(\mu'_{M,s}) \geq K \tag{2}$$

Importantly, the left-hand side of Equation 2 is ensured to be positive because the information structure corresponding to verification is always more informative than the one induced by non-verification; formally, this is captured by the convexity of $V(\mu)$ (Blackwell, 1951). However, higher informativeness does not always imply that the DM will verify the signal source in this context because verification is expensive. To understand which features of the problem drive verification choices, it is useful to observe this equivalent formulation of Equation 2:

$$\Delta_s(\mu) := \underbrace{\rho_s(1 - \rho_s)}_{\substack{\text{Variance of source} \\ \text{reliability posterior}}} \underbrace{(\mu'_{G,s} - \mu'_{B,s})^2}_{\substack{\text{Squared belief} \\ \text{distance}}} \geq K \tag{3}$$

In words, Equation 3 requires the expected gain in belief accuracy — or *value of verification* — to outweigh the verification cost.[6] It is easy to see that, in order for this condition to be met, not only

[6] With a logarithmic utility function, the verification condition becomes:

$$\rho_s \, D_{KL}(\mu_G \,||\, \mu_M) + (1 - \rho_s) \, D_{KL}(\mu_B \,||\, \mu_M) \geq K \tag{4}$$

where $D_{KL}(p \,||\, q)$ denotes the Kullback-Leibler divergence between distributions p and q: by revealing source type r, verification modifies the DM's belief on θ; so, the value of verification is the expected KL

the verification cost cannot be too high, but also the DM must give the source the benefit of the doubt, and resolution of uncertainty must materially affect the DM's posterior beliefs.[7] Crucially, these three requirements must be satisfied *at the same time*. To see why, suppose the DM is uncertain about the source type (i.e., ρ_s is somewhat intermediate), but her Bayesian posteriors conditional on the two source types are nearly identical: in this case, verification would barely affect her posterior beliefs, making it potentially not worthwhile even when verification costs are low. Similarly, when the Bayesian posteriors conditional on the two source types differ substantially, but the posterior on the source type ρ_s is close to either 0 or 1, the attenuated posterior $\mu'_{M,s}$ is already very similar to the optimal posterior under verification; so, verification would again have little effect on the DM's posterior beliefs, making it unlikely to compensate the verification cost.

A consequence of this condition is that, since both uncertainty about the source and the effect of resolving it depend on the DM's prior beliefs, so does the value of verification; hence, for a given signal, the same source may be optimally verified by some DMs, but ignored by others.

Definition 1 *For every $K > 0$ and $s \in \{0, 1\}$, the* verification region $\mathcal{V}_s^K \equiv \{\mu \in (0, 1) : \Delta_s(\mu) \geq K\}$ *is the set of all priors for which the expected gain from verification exceeds its cost.*

The following proposition characterizes the verification region, showing that it is nonempty only for sufficiently low verification costs, and consists of an interval of intermediate priors.

---

divergence between the posterior when r is known and the one when it is not. Importantly, Equation 4 can be rearranged so that its left-hand side corresponds to the conditional mutual information between θ and r, evaluated at the realized signal, i.e. $I(\theta; r|s)$. Conditioning on s is crucial, as — before any report arrives — the source type is uninformative about the state (i.e. $I(\theta; r) = 0$): knowing if a source is good or not *a priori* says nothing about the state because source quality is only relevant through the delivered message.

[7] In other words, to be willing to verify the source of the signal, the DM must reply affirmatively to both these questions: (i) am I really unsure about this source type? and (ii) even if I were, would that matter for my final opinion on the state?

Proposition 1 *There exists a unique $\kappa > 0$ such that, for each $s \in \{0,1\}$:*

$$\mathcal{V}^{\mathcal{I}}_{s,K} = \begin{cases} \varnothing & \text{if } K > \kappa, \\ [\underline{\mu}^K_s, \bar{\mu}^K_s], & \text{if } K \leq \kappa, \end{cases}$$

for some $0 < \underline{\mu}^K_s \leq \bar{\mu}^K_s < 1$.

Importantly, the cost cutoff κ is independent of the realized signal: loosely speaking, it captures the maximum gain in belief accuracy that verification can attain across signal realizations, which only depends on the underlying environment $\mathcal{I}$.[8]

While the maximum value of verification is independent of the signal realization, its value for a given prior belief need not be. The following lemma identifies conditions under which this signal-independence is preserved:

Lemma 2 *Verification is equally valuable across signals for degenerate or symmetric priors:*

$$\mu \in \left\{0, \tfrac{1}{2}, 1\right\} \Rightarrow \Delta_1(\mu) = \Delta_0(\mu)$$

This result follows from two observations: when the prior is degenerate ($\mu = 0$ or $\mu = 1$), the DM is already certain about the state; so, verification cannot affect her posterior beliefs. On the other hand, when the prior is symmetric ($\mu = \frac{1}{2}$), neither signal realization can be more consistent with her prior beliefs; hence, they both induce the same posterior assessment of source reliability. Away from these cases, however, the alignment of the signal realization with the DM's prior is informative about the source type, breaking this symmetry.

Definition 2 *For every prior $\mu \in [0,1] \setminus \{\frac{1}{2}\}$, a signal s is* confirmatory *if it supports current beliefs; i.e. either $s = 1$ and $\mu > 0.5$ or $s = 0$ and $\mu < 0.5$. Otherwise, it is* disconfirmatory.

[8] This symmetry follows from the binary structure of the model, together with the choice of a symmetric scoring rule.

As noted above, this distinction is not merely semantic: the value of verification of a source depends on whether its signal confirms or contradicts the DM's prior beliefs because such alignment has diagnostic power on the source type. The following three theorems formalize this intuition by comparing the value of verifying sources of confirmatory and disconfirmatory signals under reliable versus less reliable sources (Theorem 1), and reliable versus unreliable ones (Theorem 2 and Theorem 3).

Theorem 1 *If $p \geq \frac{1}{2}$, disconfirmatory signals are always more valuable to verify:*

$$\forall\, \mu \in (0,1), \quad \operatorname{sign}\left(\mu - \tfrac{1}{2}\right) = \operatorname{sign}[\Delta_0(\mu) - \Delta_1(\mu)]$$

Corollary 1.1 *If $p \geq \frac{1}{2}$, then for any verification cost $K > 0$:*

(i) If the source of a confirmatory signal is verified, so is the one of a disconfirmatory:

$$\text{For any } \mu > \tfrac{1}{2},\ \ \mu \in \mathcal{V}_1^K \ \Rightarrow\ \mu \in \mathcal{V}_0^K \qquad (\text{and symmetrically for } \mu < \tfrac{1}{2})$$

(ii) A source is mostly verified by DMs whose prior disagrees with the realized signal:

$$\mathcal{V}_1^K \neq \varnothing \ \Rightarrow\ \underline{\mu}_1^K + \bar{\mu}_1^K < 1 \qquad \text{and} \qquad \mathcal{V}_0^K \neq \varnothing \ \Rightarrow\ \underline{\mu}_0^K + \bar{\mu}_0^K > 1$$

The intuition behind Theorem 1 is fairly simple. When both sources are reliable, a disconfirmatory signal has two effects: first, it lowers the DM's assessment of source reliability, as a signal contradicting her prior beliefs is more likely to originate from the less reliable source; second, it shifts her beliefs away from the initially favored state, potentially increasing uncertainty about the state and the scope for further updating. Together, these effects make verification always more valuable for disconfirmatory signals: since they are less likely to realize under either sources, their interpretation becomes more sensitive to the source type that generated them.

Corollary 1.1 translates this mechanism into observable behavior: since sources of disconfirmatory signals have a higher value of verification, any DM willing to verify the source of a confirmatory

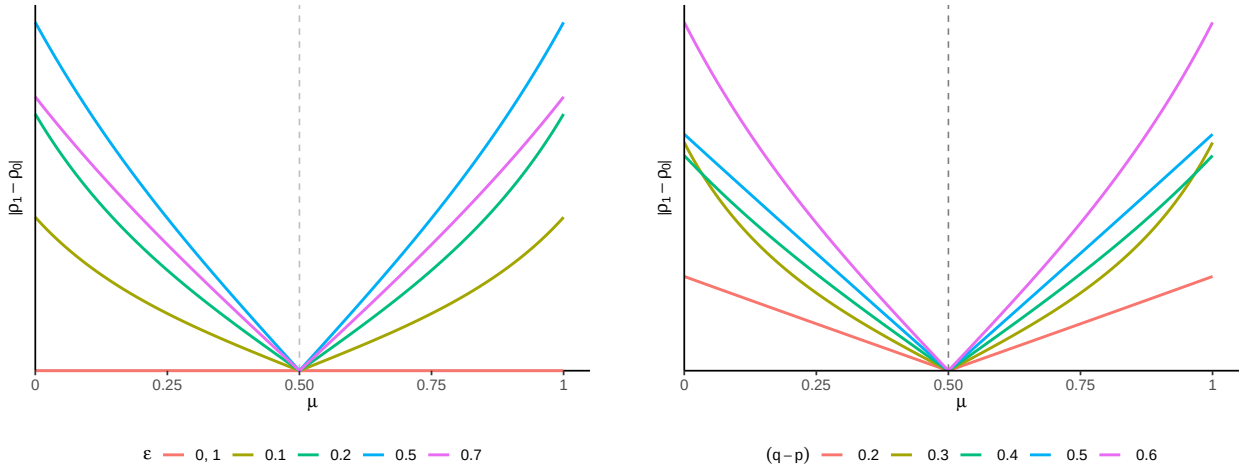
signal will also find it worthwhile to verify a disconfirmatory one. Consequently, for a given cost K, the verification region is skewed toward beliefs that contradict the observed signal: relatively pessimistic (optimistic) DMs verify sources of signals equal to 1 (equal to 0) over a wider range of priors than their optimistic (pessimistic) counterparts. Thus, selective source verification naturally emerges from Bayesian updating if verifying source reliability is costly.

This logic may no longer apply when one source is unreliable: if the DM does not know if the source is reliable, then she also does not know how to extract information from the signal, and — consequently — in which direction she should update her beliefs. To see why, suppose the DM observes a disconfirmatory signal: then, her beliefs should be revised *away* from the initially favored state if the source is reliable, but *toward* it otherwise. Conversely, a confirmatory signal prescribes to update poster beliefs toward the initially favored state if the source is reliable, but away if not.

In this case, the relative value of verifying positive and negative signals (i.e. $\Delta_1 - \Delta_0$), and hence the ranking of confirmatory and disconfirmatory information, is ambiguous a priori, as it depends on both the information environment and prior beliefs. To see why, remember that, by Equation 3, the value of verification $\Delta_s(\mu)$ is a function of both the posterior belief about the source type ρ_s, and the distance in possible state posteriors $(\mu'_{G,s} - \mu'_{B,s})^2$. Panel 2 shows (the absolute value of) the difference in posterior beliefs about the source type for each signal, $\mid \rho_1 - \rho_0 \mid$, as a function of prior beliefs μ, highlighting how it depends on the ex-ante probability ε of a type B (Figure 2a) and signals precisions q, p (Figure 2b). In particular, Figure 2a highlights how the gap is equal to zero when either there is only one type of sources in the environment ($\varepsilon = 0$ or $\varepsilon = 1$), and hence no uncertainty, or when the prior is completely uninformative ($\mu = \frac{1}{2}$), as the signal realization has no diagnostic power on the source type in this case. Otherwise, the gap in posterior beliefs about the source type is strictly increasing in the distance of ε from $\frac{1}{2}$: all else equal, the

higher the uncertainty on the source type, the more the DM is going to infer it from the realized signal. This phenomenon is stronger for more extreme priors about the state: the more confident the DM initially is, the more she is going to judge sources from their signals. Figure 2b depicts how the gap in posterior beliefs about the source type depends on the precision gap $(q - p)$: unless the prior is too extreme, higher precision gaps imply higher gaps in posterior beliefs about the source type, as the DM correctly considers into account that the sources are more informative. As before, the phenomenon is usually exacerbated at extreme priors, even if it can be overruled by the effect of the alignment (disalignment) of signals with prior beliefs in some information environments.

Figure 2: Absolute difference in posterior beliefs about whether the source type is G across signal realizations, $\mid \rho_1 - \rho_0 \mid$, as a function of prior beliefs about the state μ.

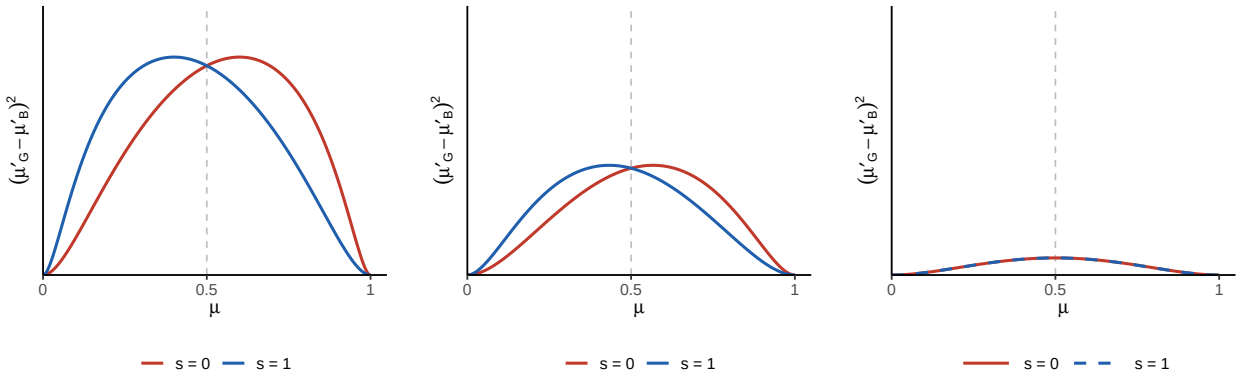


(a) $\mid \rho_1 - \rho_0 \mid$ as a function of μ for $(q, p) = (0.8, 0.6)$ and $\varepsilon \in \{0, 0.1, 0.2, 0.5, 0.7, 1\}$. The gap is null at $\varepsilon = 0$, $\varepsilon = 1$, or $\mu = \frac{1}{2}$. Otherwise, the higher the uncertainty about the source type, the more the DM infers source precision from the realized signal. This phenomenon is stronger for more extreme priors.

(b) $\mid \rho_1 - \rho_0 \mid$ as a function of μ at $\varepsilon = 0.3$ and $(q, p) \in \{(0.4, 0.6), (0.6, 0.9), (.04, 0.8), (0.2, 0.7), (0.3, 0.9)\}$. Higher precision gaps $(q - p)$ usually imply higher gaps in posterior beliefs. The phenomenon is either exacerbated or reversed at extreme priors (i.e. the curves cross for $q - p = 0.3$ and $q - p = 0.4$).

Panels 3 and 4 show how the (squared) gap in posterior beliefs about the state differs across signal realizations and information environments, respectively. Figures 3a, 3b, 3c highlights how the gap is symmetric around $\mu = \frac{1}{2}$ across signal realizations: when the prior is completely agnostic about the state, the signals are equivalent, as they can't neither confirm nor disconfirm prior knowledge (see Lemma 2). Similarly, both the gaps are jointly equal to zero at degenerate priors

Figure 3: Squared gap in posterior beliefs about the state across source precisions, $(\mu'_{G,s} - \mu'_{B,s})^2$, as a function of prior beliefs μ, for signal realizations $s = 0$ (in red) and $s = 1$ (in blue). The gaps are symmetric around $\mu = \frac{1}{2}$ across signals and jointly equal to zero at $\mu = 0$ or $\mu = 1$.



(a) $(\mu'_{G,s} - \mu'_{B,s})^2$ as a function of μ for $(q, p) = (0.9, 0.2)$ across signal realizations. Disconfirmatory signals induce higher dispersion of posterior beliefs.

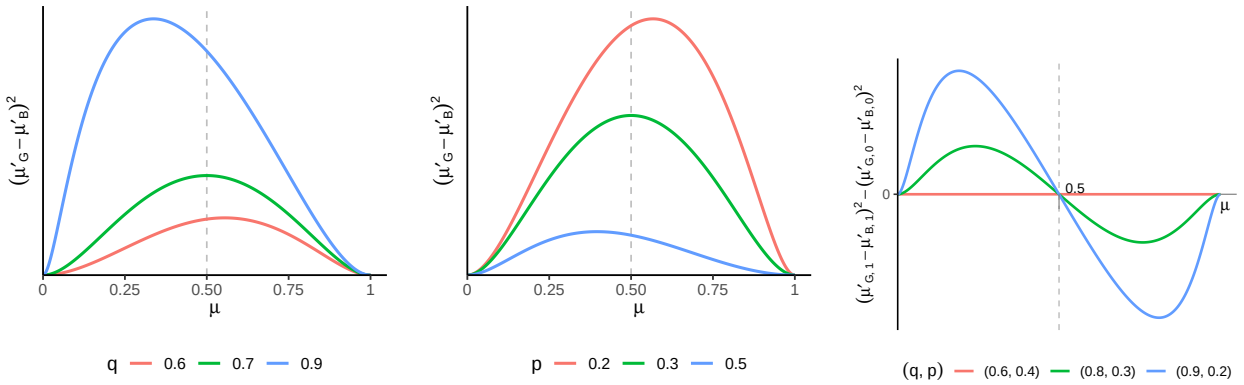
(b) $(\mu'_{G,s} - \mu'_{B,s})^2$ as a function of μ for $(q, p) = (0.8, 0.3)$across signal realizations. Disconfirmatory signals induce higher dispersion of posterior beliefs.

(c) $(\mu'_{G,s} - \mu'_{B,s})^2$ as a function of μ for $(q, p) = (0.6, 0.4)$ across signal realizations. If $q = 1 - p$, $s = 1$ with precision q is equivalent to $s = 0$ with precision p.

(i.e. $\mu = 0$ or $\mu = 1$): in this case, any posterior will always be equal to the prior, regardless of the source type. In all other cases, however, the gaps are usually different, unless $q = 1 - p$ (Figure 3c): in this case, receiving a signal $s = 1$ with precision q is equivalent to receive a signal $s = 0$ with precision p; so, the posterior beliefs for each source type will be the same. Figures 3a and 3b also show that disconfirmatory signals always induce higher dispersion of posterior beliefs; however, this doesn't necessarily translate in higher values of verification when one source type is unreliable ($p < \frac{1}{2}$) as the DM is unable to tell whether signals should be read at face value or not.

Figure 4a plots the gap in posterior beliefs about the state for $s = 1$ at different precision levels q of type G, for a given precision p of the other type. It is easy to see that higher precision of the reliable source makes the gap unambiguously higher, although the magnitude of the change varies with prior beliefs. Figure 4b shows a symmetric result for different precision levels p of type G, while keeping the precision q of type G fixed. Finally, Figure 4c plots the difference between the squared posterior gaps induced by different signal realizations: notably, both the magnitude and sign of this difference can vary substantially as source precisions and prior beliefs change.

Figure 4: Squared gap in posterior beliefs about the state across source precisions, $(\mu'_{G,s} - \mu'_{B,s})^2$, as a function of prior beliefs μ, for a signal realization $s = 1$ and different source precisions.



(a) $(\mu'_{G,s} - \mu'_{B,s})^2$ as a function of μ for $p = 0.3$ and $q \in \{0.6, 0.7, 0.9\}$. Higher precision of the good source makes the gap unambiguously higher. The magnitude of the change varies with the prior.

(b) $(\mu'_{G,s} - \mu'_{B,s})^2$ as a function of μ for $q = 0.7$ and $p \in \{0.2, 0.3, 0.5\}$. Higher precision of the bad source makes the gap unambiguously smaller. The magnitude of the change varies with the prior.

(c) Difference in $(\mu'_{G,s} - \mu'_{B,s})^2$ across signal realizations as a function of μ for $(q, p) \in \{(0.6, 0.4), (0.8, 0.3), (0.9, 0.2)\}$. Both magnitude and sign of this gap vary with source precisions and the prior.

Taken together, these phenomena help explain why the sign of $(\Delta_1 - \Delta_0)$ is sensitive to the specific configuration of the information environment and prior beliefs. Then, to gain further intuition about what determines this sign, and hence whether confirmatory or disconfirmatory signals are more valuable to verify, let me start by considering prior beliefs in a neighborhood of $\mu = \frac{1}{2}$. At this prior, both channels — the one operating through posterior beliefs about source reliability and the other through the distance between possible state posteriors —- vanish, implying that $\Delta_1(\frac{1}{2}) - \Delta_0(\frac{1}{2}) = 0$ (see Lemma 2); consequently, the relative value of verification in a neighborhood of $\mu = \frac{1}{2}$ is governed, to a first-order approximation, by its response to a marginal change in the prior. Then, I can expand $(\Delta_1(\mu) - \Delta_0(\mu))$ around $\mu = \frac{1}{2}$ to show that in a sufficiently small neighborhood of $\mu = \frac{1}{2}$:

$$\text{sign}[\Delta_1(\mu) - \Delta_0(\mu)] = \text{sign}\left\{\left(\mu - \tfrac{1}{2}\right)\left[\left(\frac{d\Delta_1(\mu)}{d\mu}\bigg|_{\mu=\frac{1}{2}}\right) - \left(\frac{d\Delta_0(\mu)}{d\mu}\bigg|_{\mu=\frac{1}{2}}\right)\right]\right\} \tag{5}$$

provided that at least one derivative is nonzero. For the sake of exposition, let $\Delta_C(\mu)$ and $\Delta_D(\mu)$ denote the value of verification of a confirmatory and a disconfirmatory signal, respectively. Also,

let $\delta > 0$ be an infinitesimal constant, and $\tau_N^{\mathcal{T}} = \big(3-2\varepsilon\big)\big(q-\frac{1}{2}\big) + \big(1+2\varepsilon\big)\big(p-\frac{1}{2}\big)$. Then, it is easy to show that Equation 5 implies:

$$\forall\, \mu \in \big(\tfrac{1}{2}-\delta, \tfrac{1}{2}+\delta\big), \quad \Delta_D(\mu) \geq \Delta_C(\mu) \;\Leftrightarrow\; \tau_N^{\mathcal{T}} \geq 0$$

with equality whenever $\tau_N^{\mathcal{T}} = 0$. Strictly speaking, $\tau_N^{\mathcal{T}}$ is a weighted average of the informativeness of the two possible sources, with relative weights being defined by the probability of occurrence of different types. Additionally, as I assumed $p < \frac{1}{2}$, it can be read as a measure of the trade-off between type G's *competence* and type B's *deceptiveness*; so, the condition above states that, for prior beliefs close to $\mu = \frac{1}{2}$, the DM will find more valuable to verify the source of disconfirmatory signals if and only if competence dominates over deceptiveness, even accounting for the probability of occurrence of a type B and the extent of its unreliability. More subtly, $\tau_N^{\mathcal{T}}$ measures the informational quality of the two sources relatively to an uninformative signal. To see why, note that $\tau_N^{\mathcal{T}}$ can be restated as:

$$\tau_N^{\mathcal{T}} = 4\big(m-\tfrac{1}{2}\big) - \big(q-p\big)$$

where $m = (1-\varepsilon)q + \varepsilon p$. So, $\tau_N^{\mathcal{T}}$ measures the aggregate informativeness of an unverified signal (taken at face value), penalized by the reliability gap: for a given expected precision, the higher the wedge between source precisions, the more the accuracy of the signal depends on the source type.

A similar discussion can be made for priors close to 0 or 1. When beliefs are degenerate, the gap in posterior beliefs vanishes regardless of the signal realization and the source type (i.e. $\mu'_{G,1}(1) - \mu'_{B,1}(1) = \mu'_{G,0}(1) - \mu'_{B,0}(1) = \mu'_{G,1}(0) - \mu'_{B,1}(0) = \mu'_{G,0}(0) - \mu'_{B,0}(0) = 0$): this follows from the fact that — unlike in the case of an uninformative prior — posterior beliefs coincide with the prior regardless of the signal observed by the DM, or the source that generated it. Thus, for DMs with degenerate priors, the gap in value of verification $(\Delta_1 - \Delta_0)$ is zero not only because

the signal realization is irrelevant, but also because so is the source itself: as a result, the relative value of verification in neighborhoods of $\mu = 0$ and $\mu = 1$ is governed by its response to a marginal change in the prior only to a second-order approximation. Formally, by expanding $[\Delta_1(\mu) - \Delta_0(\mu)]$ around $\mu = 0$, it is possible to show that for beliefs in a sufficiently small neighborhood of $\mu = 0$:

$$\text{sign}\left[\Delta_1(\mu) - \Delta_0(\mu)\right] = \text{sign}\left\{ \left.\frac{d^2\Delta_1(\mu)}{d\mu^2}\right|_{\mu=0} - \left.\frac{d^2\Delta_0(\mu)}{d\mu^2}\right|_{\mu=0} \right\} \tag{6}$$

An equivalent condition can be found by expanding the gap around $\mu = 1$. Now, let $\tau_E^{\mathcal{I}} = qp[(1-\varepsilon)q + \varepsilon p]^2 - (1-q)(1-p)[1-(1-\varepsilon)q - \varepsilon p]^2$. Then, it can be showed that Equation 6 (and its counterpart for $\mu = 1$) implies:

$$\forall\, \mu \in (0,\delta) \cup (1-\delta, 1), \quad \Delta_D(\mu) \geq \Delta_C(\mu) \;\Leftrightarrow\; \tau_E^{\mathcal{I}} \geq 0$$

with equality whenever $\tau_E^{\mathcal{I}} = 0$. Importantly, also $\tau_E^{\mathcal{I}}$ can be interpreted as a measure of the trade-off between competence and deceptiveness, although on a non-linear scale; so, the same type of reasoning drives verification choices at both uninformative and degenerate priors. To see why, note that:

$$\text{sign}(\tau_E^{\mathcal{I}}) = \text{sign}\left[\underbrace{\ln\left(\frac{m}{1-m}\right) - \ln\left(\frac{1-m}{m}\right)}_{\substack{\textit{Ex-ante}\text{ informativeness} \\ \text{of an unverified signal}}} + \underbrace{\ln\left(\frac{q}{1-q}\right) - \ln\left(\frac{1-p}{p}\right)}_{\substack{\text{Net reliability} \\ \text{of an unverified signal}}} \right] \tag{7}$$

The first term on the right-hand side of Equation 7 corresponds to the swing in beliefs across the two possible signal realizations; so, it measures how much the signal realization is informative for beliefs about the state θ.[9] The second term, instead, compares source G's odds of matching the state to source B's odds of contradicting it, both on a logarithmic scale; hence, it can be seen as an *ex-ante* net measure of trustworthiness of a signal that captures the relative strength of these odds on a multiplicative scale. As a result, both thresholds determining the ranking of confirmatory and disconfirmatory signals at uninformative and degenerate priors compare the competence of the

[9] Note that $\ln\left(\frac{m}{1-m}\right) - \ln\left(\frac{1-m}{m}\right) = \ln\left(\frac{\Pr(\theta=1|s=1)}{\Pr(\theta=0|s=1)}\right) - \ln\left(\frac{\Pr(\theta=0|s=1)}{\Pr(\theta=0|s=0)}\right)$.

reliable source with the deceptiveness of the unreliable one, altough they do so on different scales.

Crucially, these comparisons also extend beyond the cases of uninformative or extreme priors: as matter of fact, the signs of $\tau_N^{\mathcal{I}}$ and $\tau_E^{\mathcal{I}}$ also jointly determine the ranking of confirmatory and disconfirmatory signals at all other priors. In what follows, I will say that source precisions are *balanced* on the linear (multiplicative) scale if $\tau_N^{\mathcal{I}} = 0$ ($\tau_E^{\mathcal{I}} = 0$), high precision (i.e. *competence*) *dominates* on the linear (multiplicative) scale if $\tau_N^{\mathcal{I}} > 0$ ($\tau_E^{\mathcal{I}} > 0$), low precision (i.e. *deceptiveness*) dominates otherwise. Whenever source precisions are balanced on at least one scale, or the same type of precision dominates on both scales, the ranking of confirmatory and disconfirmatory signals is uniquely determined by the dominating force, independently of the DM's prior. This result is formalized in Theorem 2:

Theorem 2 *Assume* $p < \frac{1}{2}$ *and* $\tau_N^{\mathcal{I}}\tau_E^{\mathcal{I}} \geq 0$. *Then,* $\forall\, \mu \in (0,1) \setminus \{\frac{1}{2}\}$*:*

1. *If high precision dominates on both scales, disconfirmatory signals are more valuable to verify:*

$$\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}} \geq 0 \quad \wedge \quad (\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}}) \neq (0,0) \quad \Rightarrow \quad \Delta_D(\mu) > \Delta_C(\mu)$$

2. *If low precision dominates on both scales, confirmatory signals are more valuable to verify:*

$$\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}} \leq 0 \quad \wedge \quad (\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}}) \neq (0,0) \quad \Rightarrow \quad \Delta_D(\mu) < \Delta_C(\mu)$$

3. *If precisions are balanced on both scales, both types of signals are equally valuable to verify:*

$$\tau_N^{\mathcal{I}} = \tau_E^{\mathcal{I}} = 0 \quad \Rightarrow \quad \Delta_D(\mu) = \Delta_C(\mu)$$

Instead, when the two scales disagree, the ranking in terms of value of verification of sources of confirmatory and disconfirmatory signals depends on the DM's prior; in particular, the ranking reverses between DMs holding intermediate beliefs (i.e. closer to $\frac{1}{2}$) and those holding more extreme beliefs (i.e. closer to 0 or 1), where the threshold distinguishing intermediate from extreme beliefs is determined by the information environment $\mathcal{I}$. This is formalized by Theorem 3:

Theorem 3 *Assume* $p < \frac{1}{2}$ *and* $\tau_N^{\mathcal{I}}\tau_E^{\mathcal{I}} < 0$. *Then,* $\forall\, \mu \in (0,1) \setminus \{\frac{1}{2}\}$*:*

1. *If high precision only dominates on a linear scale, disconfirmatory signals are more valuable to verify if and only if beliefs are not too extreme:*

$$\tau_E^{\mathcal{I}} < 0 < \tau_N^{\mathcal{I}} \quad \Rightarrow \quad \Delta_D(\mu) \geq \Delta_C(\mu) \;\Leftrightarrow\; \left|\mu - \tfrac{1}{2}\right| \leq \tfrac{1}{2}\sqrt{\tfrac{\tau_N^{\mathcal{I}}}{\tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}}}}$$

2. *If high precision only dominates on a multiplicative scale, confirmatory signals are more valuable to verify if and only if beliefs are not too extreme:*

$$\tau_N^{\mathcal{I}} < 0 < \tau_E^{\mathcal{I}} \quad \Rightarrow \quad \Delta_D(\mu) \leq \Delta_C(\mu) \;\Leftrightarrow\; \left|\mu - \tfrac{1}{2}\right| \leq \tfrac{1}{2}\sqrt{\tfrac{\tau_N^{\mathcal{I}}}{\tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}}}}$$

In both cases, equality holds if and only if a belief threshold is hit with equality.

Together, Theorem 2 and Theorem 3 define five cases, that exhaustively describe the possible sign configurations of two thresholds $(\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}})$ (see Table 1). These cases are summarized by Figure 5, which — for each prior μ — depicts whether the DM will find more valuable to verify the source of a disconfirmatory signal (in red), a confirmatory one (in blue), or will be indifferent between the two (purple lines).

	$\tau_N^{\mathcal{I}} > 0$	$\tau_N^{\mathcal{I}} = 0$	$\tau_N^{\mathcal{I}} < 0$
$\tau_E^{\mathcal{I}} > 0$	Theorem 2, Case 1	Theorem 2, Case 1	Theorem 3, Case 2
$\tau_E^{\mathcal{I}} = 0$	Theorem 2, Case 1	Theorem 2, Case 3	Theorem 2, Case 2
$\tau_E^{\mathcal{I}} < 0$	Theorem 3, Case 1	Theorem 2, Case 2	Theorem 2, Case 2

Table 1: Possible sign configurations of the thresholds' pair $(\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}})$. Each cell identifies the theorem and corresponding case covering that configuration.

In the proof, I show that the difference in verification value across signal realizations for a given prior μ can be reformulated as:

$$(2\mu - 1)[\tau_N^{\mathcal{I}}\mu(1-\mu) + \tau_E^{\mathcal{I}}(2\mu - 1)^2]$$

This expression suggests that the relative importance of the thresholds $\tau_N^{\mathcal{I}}$ and $\tau_E^{\mathcal{I}}$ in determining the source of which signal is more valuable to verify depends on how extreme beliefs are: this follows

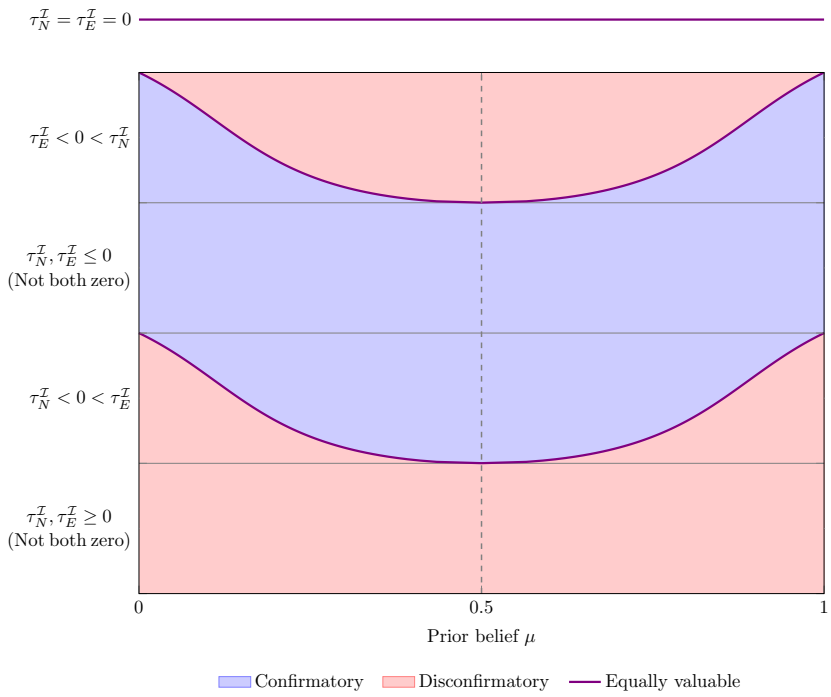


Figure 5: For each prior μ, the color of the area denotes whether the DM finds more valuable to verify the source of a disconfirmatory signal (in red), a confirmatory one (in blue), or both (purple line). Each horizontal band, together with the purple horizontal line at the top, represents one case of Theorem 2 and Theorem 3.

from the fact that $\mu(1-\mu)$ is maximized at $\mu = \frac{1}{2}$ and vanishes at $\mu \in \{0,1\}$, while $(2\mu-1)^2$ exhibits the opposite pattern. When the thresholds have the same sign (or when at least one of them is zero), the gap in verification value will have the same sign regardless of the DM's prior: this is the case of when source precisions are balanced on at least one scale, or the same type of precision dominates on both scale. When, instead, thresholds have different signs (i.e. high precision dominates the low one only on one scale but not on the other), the gap in verification value will change sign at some interior prior belief; so, DMs with intermediate and extreme priors will rank the source verification value of confirmatory and disconfirmatory signals in opposite ways.

4 Departures from Bayesian Updating

In Section 3, I showed that selective source verification can arise under Bayesian updating. This contrasts with a large part of the literature that interprets asymmetric responses to information as evidence of behavioral biases such as confirmation bias (Wason, 1960; Rabin and Schrag, 1999) or motivated reasoning (Kunda, 1990; Epley and Gilovich, 2016): according to these theories, individuals depart from rationality to favor their preferred or prior beliefs. More precisely, both phenomena can be seen as cognitive processes in which individuals unconsciously interpret and evaluate information in ways that support conclusions they are inclined to accept; however, there is a fundamental difference in what guides their interpretation of new evidence: their existing beliefs in the case of confirmation bias; what they would like to be true in the case of motivated reasoning. For instance, consider a person who would like to think she is smart, but currently believes she is not: then, when presented with new evidence about her intelligence, she will tend to interpret it in favor of the hypothesis that she is indeed smart if she engages in motivated reasoning, in contrast otherwise. Thus, although the two phenomena are closely related, they are not equivalent.

The benchmark developed in this paper shows that, at least within the domain of source verification, neither of the two is necessary to observe asymmetric choices; nonetheless, it seems natural to wonder whether the result survives to the introduction of these behavioral distortions. For this reason, the remainder of this section explores how to incorporate confirmation bias (Subsection 4.2) and motivated reasoning (Subsection 4.1) within the framework, and how such behavioral distortions modify the asymmetry in verification choices across signals realizations observed in the Bayesian benchmark.

4.1 Motivated Reasoning

Motivated reasoning is the tendency to seek, interpret, or recall information in ways that support a desired conclusion (Kunda, 1990). Experimental findings in both psychology and economics provide extensive evidence of this phenomenon across a range of domains, such as self-perception (Eil and Rao, 2011; Möbius et al., 2022), financial outcomes (Seybert and Bloomfield, 2009; Mayraz, 2011), political views (Taber and Lodge, 2006), and climate change (Druckman and McGrath, 2019). In particular, research has repeatedly highlighted its connection with selective behaviors, including selective attention (Luo and Zhao, 2019) and selective exposure (Hastall and Knobloch-Westerwick, 2013; Knobloch-Westerwick et al., 2013). Building on this literature, I investigate a distinct form of selective treatment of information: the decision of whether to verify the source of a piece of news.

To investigate this mechanism, I consider a setting in which the DM has a preference for one state of the world to be true: this captures situations in which individuals derive utility from optimistic beliefs, or — more generally — have a stake in a specific outcome being realized. For instance, a patient may hope that a medical treatment is effective, or a student may wish they are smarter than average. Then, signals can be classified as follows:

Definition 3 *A signal s is* favorable *if it matches the DM's preferred state,* unfavorable *otherwise.*

To formalize this intuition, I assume without loss of generality that the DM prefers state $\theta = 1$; so, a favorable signal corresponds to a positive signal $s = 1$. Then, motivated reasoning can be represented by the following log-odds distortion:

$$\log\left(\frac{\tilde{\rho}_s}{1-\tilde{\rho}_s}\right) = \log\left(\frac{\rho_s}{1-\rho_s}\right) + \eta \cdot m(s) \tag{8}$$

where $\tilde{\rho}_s$ denotes the *perceived* posterior probability that the source is good conditional on signal s, $m(s) = 1$ if $s = 1$, -1 otherwise, and $\eta \geq 0$ captures the strength of motivated reasoning: as η increases, favorable signals are perceived as increasingly more reliable than under Bayesian updating, while the opposite holds for unfavorable ones. In particular, when $\eta = 0$, the model reduces to the Bayesian benchmark in Section 2.

This log-odds specification has two useful properties. First, additive shifts in log-odds translate into multiplicative changes in odds while ensuring that transformed probabilities remain in the unit interval (e.g., Berkson, 1944; Agresti, 2013; McCullagh, 2019). Second, the distortion is independent of prior beliefs: conditional on a given Bayesian posterior ρ_s, the adjustment to log-odds depends only on whether the signal is favorable or not.

Importantly, the distortion operates on posterior beliefs about source reliability rather than beliefs about the state of the world, departing from standard models of motivated reasoning (Brunnermeier and Parker, 2005; Mayraz, 2011). This choice reflects the idea that individuals may rationalize information that conflicts with their desired outcome by questioning the credibility of its source rather than the piece of information itself: for instance, a patient facing a harsh diagnosis may doubt the competence of the doctor (and decide to ask for a second opinion). By allowing perceived source reliability to depend on whether the signal is favorable, the model captures this form of motivated skepticism, providing a direct channel through which motivated reasoning can shape incentives for source verification.[10]

[10]With this choice, I do not claim that motivated reasoning might not also directly affect posterior beliefs

Apart from this distortion, the decision problem remains unchanged; so, the verification condition takes the same form as Equation 3, with the Bayesian posterior belief about the source type ρ_s replaced by its motivated counterpart, $\tilde{\rho}_s$. As a result, verification regions retain their interval structure, allowing for a direct comparison of verification choices with and without the bias.

To build intuition, let me first discuss the effect of motivated reasoning at a given prior: one might expect the bias to reduce source verification of favorable signals while increasing it for unfavorable ones. Surprisingly, this need not be the case, as increasing the perceived posterior probability that the source is good when a favorable signal realizes has two opposite effects: on the one hand, if the source was already seen as reliable enough, the distortion reduces uncertainty about the source type, making verification less valuable. On the other, however, if the source was initially regarded as unlikely to be reliable, the bias increases uncertainty: then, when the source reports a favorable signal, motivated reasoning leads the DM to give it the benefit of the doubt, increasing its perceived posterior reliability and, consequently, its value of verification. From a technical point of view, this happens because uncertainty on source reliability is measured by $\tilde{\rho}_s(1-\tilde{\rho}_s)$, which is maximized at $\tilde{\rho}_s = \frac{1}{2}$. A symmetric argument can be made for unfavorable signals. To formalize this mechanism, let:

$$\tilde{\Delta}_s(\mu) = \tilde{\rho}_s(1-\tilde{\rho}_s)(\mu'_{G,s} - \mu'_{B,s})^2 \quad \text{for } s \in \{0,1\}$$

This is the value of verification of the source of signal s at prior belief μ computed at the distorted posterior about the source type $\tilde{\rho}_s$. Then, the following result holds:

Proposition 2 *Assume motivated reasoning follows Equation 8, with $\eta > 0$. Then, $\forall\, \mu \in (0,1)$:*

$$\tilde{\Delta}_1(\mu) \geq \Delta_1(\mu) \quad \Leftrightarrow \quad \rho_1(\mu) \leq \frac{1}{1+e^{\eta/2}}$$

$$\tilde{\Delta}_0(\mu) \geq \Delta_0(\mu) \quad \Leftrightarrow \quad \rho_0(\mu) \geq \frac{1}{1+e^{-\eta/2}}$$

about the underlying state; however, I do not see it as the most natural way to model the phenomenon.

In words, motivated reasoning increases the value of verifying the source of a favorable signal if and only if the Bayesian posterior probability that the source is reliable is sufficiently low: since the value of verification depends on the uncertainty about the source type, increasing this posterior for a favorable signal raises the value of verification only when the resulting belief is more uncertain (i.e. closer to $\frac{1}{2}$) than its Bayesian counterpart. The intuition reverses for unfavorable signals.

This reasoning also suggests when motivated reasoning can reduce the value of source verification *uniformly* across priors. Since the Bayesian posterior probability that the source is reliable following a favorable signal is increasing in the prior μ, it is minimized when the DM is most skeptical about the state being the preferred one (i.e. at $\mu = 0$); thus, if motivated reasoning is sufficiently strong to reduce the value of verification even at this lowest posterior reliability, it must necessarily reduce it for all other priors as well. A similar argument can be applied to unfavorable signals when the DM is most confident about the source being reliable (i.e. at $\mu = 1$). Taken together, these observations allow to determine conditions under which motivated reasoning uniformly reduces the value of verification and, consequently, shrinks the set of priors for which verification is optimal. Formally, let $\tilde{\mathcal{V}}_s^K$ denote the verification region for signal s under motivated reasoning, given a verification cost $K > 0$. Then:

Proposition 3 *Assume motivated reasoning follows Equation 8, with $\eta > 0$. Then, $\forall\, K > 0$:*

$$\eta \geq 2 \log \left[\frac{1 - \rho_1(0)}{\rho_1(0)} \right] \implies \tilde{\mathcal{V}}_1^K \subseteq \mathcal{V}_1^K$$

$$\eta \geq 2 \log \left[\frac{\rho_0(0)}{1 - \rho_0(0)} \right] \implies \tilde{\mathcal{V}}_0^K \subseteq \mathcal{V}_0^K$$

For either signal realization, sufficiently strong motivated reasoning reduces the set of priors for which source verification is optimal. Importantly, both thresholds depend on the corresponding Bayesian posterior about the source type evaluated at the degenerate prior $\mu = 0$, although for opposite reasons: for favorable signals, motivated reasoning increases perceived source reliability; so,

the relevant threshold is the *lowest* attainable Bayesian posterior, $\rho_1(0)$. Conversely, for unfavorable signals, motivated reasoning decreases perceived source reliability; hence, the relevant benchmark is the *highest* Bayesian posterior, $\rho_0(0)$. In both cases, these correspond to the posterior beliefs of the DM who is most skeptical about the preferred state being true.

4.2 Confirmation Bias

Confirmation bias is the tendency to seek, interpret, or recall information in ways that support prior beliefs (Nickerson, 1998). This phenomenon has been widely documented in the psychological and economic literature and is considered a pervasive feature of human cognition across several domains, including political judgment, (Lord et al., 1979; Taber and Lodge, 2006), financial decision-making (Pouget et al., 2017) and evaluation processes (Mahoney, 1977; Tomkins et al., 2017).

In particular, confirmation bias has been associated with selective information processing in the form of selective exposure (i.e. individuals tend to seek information consistent with their existing beliefs while avoiding sources of contradictory evidence, as in Hart et al., 2009), selective perception (i.e. individuals tend to interpret evidence in ways that favor those beliefs, as in Fryer Jr et al., 2019); and selective recall (i.e. individuals are more likely to remember information consistent with prior beliefs, as in Levine and Murphy, 1943 and Eagly et al., 1999). This naturally raises the question of how confirmation bias relates to selective source verification.

To explore this issue, I extend my framework to incorporate confirmation bias as modelled by Rabin and Schrag (1999). To illustrate their approach, suppose that $\mu > \frac{1}{2}$, so that the DM initially believes that $\theta = 1$ is more likely than $\theta = 0$: in this case, a signal is confirmatory if positive ($s = 1$), disconfirmatory otherwise. Confirmation bias then implies that the DM may mistakenly interpret a signal $s = 0$ as the opposite realization $s = 1$, while positive signals $s = 1$ are always correctly processed. Formally, I assume that a disconfirmatory signal is misperceived as confirmatory with

probability $\alpha \in [0,1]$; thus, letting $\hat{s}$ denote the perceived signal realization:

$$\Pr(\hat{s}=1 \mid s=1)=1; \qquad \Pr(\hat{s}=1 \mid s=0)=\alpha; \qquad \Pr(\hat{s}=0 \mid s=0)=1-\alpha \tag{9}$$

The case of $\alpha = 0$ corresponds to the rational benchmark described in Section 2. Importantly, I assume that the DM is *naive*, meaning that she is unaware of her own bias and takes $\hat{s}$ at face value, treating it as the true signal realization. This belief is not even undermined by verification, as verifying the source does not reveal whether the signal has been misperceived, but only the source's type. The assumption of naivety is motivated by empirical findings that show how individuals often fail to recognize the influence of cognitive biases on their own judgments: in particular, evidence found that people tend to perceive themselves as less vulnerable to cognitive biases than others (Pronin et al., 2002) and appear more willing to acknowledge the possibility to be biased in the abstract rather than in specific tasks (Ehrlinger et al., 2005). Together, these findings suggest that personal perceptions may be treated as accurate even when they are systematically distorted.

The wedge between actual and perceived signal realizations is the only modification introduced by confirmation bias; so, conditional on the perceived signal, the DM is observationally Bayesian. In particular, confirmation bias does not affect the perceived value of source verification: the value $\Delta_{\hat{s}}(\mu)$ assigned to verifying the source of a signal $\hat{s}$ taken at face value — and the corresponding verification region — are the same of those of a rational DM who observed the corresponding signal realization $s = \hat{s}$; the key difference is that, because the perceived signal may differ from its actual realization, the value of source verification $\Delta_{\hat{s}}(\mu)$ can be misspecified.

This generates a sharp prediction when both source types are reliable ($p \geq \frac{1}{2}$): in this case, Theorem 1 shows that — all else equal — a rational DM finds it more valuable to verify sources of disconfirmatory signals. With confirmatory bias, however, some disconfirmatory signals may be misperceived as confirmatory, reducing their perceived verification value: then, signals that would

have been worth verifying may go unverified, thereby attenuating the asymmetry in verification choices. To formalize this result, I introduce the notion of a *verification gap*:

Definition 4 *For any $\alpha \in [0,1]$ and $K > 0$, the* verification gap $g(\alpha)$ *is the difference in verification probabilities of disconfirmatory and confirmatory signals, conditional on true signal realizations:*

$$g(\alpha) \equiv \Pr(\text{verify} \mid s \text{ is disconfirmatory}, \alpha) - \Pr(\text{verify} \mid s \text{ is confirmatory}, \alpha)$$

Importantly, for the sake of exposition, I define the verification gap as a *signed* difference: the difference in verification probabilities between the source of a disconfirmatory and confirmatory signal. Then, the gap equals $\Pr(\text{verify} \mid s = 0, \alpha) - \Pr(\text{verify} \mid s = 1, \alpha)$ whenever $\mu \in (\frac{1}{2}, 1)$, while it reverses sign otherwise. Then, the following holds:

Proposition 4 *If $p \geq \frac{1}{2}$, the verification gap $g(\alpha)$ is linearly decreasing in the bias α:*

$$g(\alpha) = (1 - \alpha)\, g(0)$$

In other words, because the DM fails to recognize a fraction of disconfirmatory signals as such in the first place, a biased DM appears more impartial in her verification choices; consequently, the asymmetry in verification choices is greatest under Bayesian updating ($\alpha = 0$), shrinks monotonically as the bias intensifies, and disappears entirely when $\alpha = 1$. Thus, Proposition 4 challenges the common intuition that asymmetric verification choices are evidence of confirmation bias, showing that, in fact, the opposite holds.

5 Discussion

This paper studies when a rational decision maker chooses to verify the reliability of an information source at a cost. I show that a small verification cost is not sufficient for the DM to be willing to verify a source: source verification is valuable only for an intermediate range of prior beliefs,

where uncertainty about source reliability is sufficiently high, and learning the source type can still meaningfully affect beliefs. More importantly, I show that optimal source verification is selective: when all sources are reliable, sources of disconfirmatory signals are systematically more valuable to verify than confirmatory ones; so, asymmetric scrutiny can arise even under Bayesian updating. Instead, when one source is unreliable, this ranking can reverse depending on the trade-off between competence and deceptiveness, and can be reversed between intermediate and extreme priors for some information environments. I further show that behavioral departures from Bayesian updating need not amplify these patterns: in particular, I find that neither confirmation bias nor motivated reasoning is necessary to explain the selective verification choices they are commonly associated with; indeed, they may even attenuate or reverse such asymmetries. Taken together, these results warn against interpreting asymmetric treatment of disconfirmatory information as evidence of biased reasoning; rather, selective source verification may reflect an optimal response to uncertainty about the underlying sources of information. By challenging popular explanations of asymmetric information processing, this paper raises many new questions. Below, I discuss some limitations of the present analysis and outline open tasks left for future research.

Source verification and belief dynamics. The current model considers a single period: the DM observes a signal, decides whether to verify its source, and updates her beliefs about the state accordingly. In practice, however, this process can repeat over time: hence, an open question is how endogenous source verification shapes belief dynamics. The fact that the value of verification is naturally higher for intermediate beliefs suggests that DMs may be more likely to verify information at early stages of their belief formation, but might eventually stop doing so once their beliefs became sufficiently extreme. If this is the case, early stage of information provision may have a priming effect on agnostic DMs, being crucial in shaping the direction of subsequent

beliefs and potentially giving rise to polarization.

Generalizing the information structure. The information structure in this paper is deliberately simple: there are only two states, two signals, and two possible sources, source precisions and the ex-ante probability of a source being bad are known, the verification cost is known and exogenous, and verification is perfect. These restrictive assumptions ensure the model is tractable, but limit its relevance for many real-world settings. Then, one theoretical question is to what extent these assumptions can be relaxed while preserving the model's key findings. Additionally, I assume the DM has perfect knowledge of the information environment, which is also unlikely to hold in reality: if, instead, the DM only holds beliefs about the distribution of source types and their precisions, how do these beliefs evolve over time? How do they interact with beliefs about the state? How does this correlation affect verification decisions? Addressing these questions is crucial for understanding the potential of this framework to describe real-world settings.

Reputational incentives and strategic disclosure. The current model is agnostic about the origin of signals: are they randomly realized, or strategically selected by information providers? In particular, the model abstracts away from the reputational incentives that information providers may face when they have an interest in shaping either receivers' beliefs, as in persuasion or manipulation models, or to match them, as when competing for receivers' attention and trying to build a loyal audience by providing information aligned with their worldview. Then, an open question is how a strategic information provider would optimally disclose information when receivers process it according to the mechanisms described in this paper. Exploring this question could establish interesting connections with the literature on persuasion, media slant, and reputational concerns.

Bringing theory to the field. How well does the model predict behavior? To bridge the gap between decision theory and real-world findings, I plan to gather experimental evidence. A first question is how well the benchmark developed in this paper describes verification choices in a controlled setting: to address this point, I designed a laboratory experiment in which participants have full knowledge of the information environment, while framing effects and other behavioral motives are minimized (Appendix B). However, an equally important question is whether the mechanisms identified by the model can explain verification choices in more realistic settings: when people gather information, for instance through LLM-based tools, what drives their source verification choices? Addressing this topic requires capturing the dynamics of real-world information acquisition situations, which can be naturally done with a field experiment. This approach would help assess the external validity of the theoretical framework and reconnect the analysis with concrete phenomena, including the growing use of AI assistants as information sources.

References

Agresti, Alan (2013). *Categorical data analysis*. John Wiley & Sons (cited on page 29).

Argenziano, Rossella, Sergei Severinov, and Francesco Squintani (2016). "Strategic information acquisition and transmission". In: *American Economic Journal: Microeconomics* 8.3, pages 119–155 (cited on page 5).

Arrow, Kenneth J, David Blackwell, and Meyer A Girshick (1949). "Bayes and minimax solutions of sequential decision problems". In: *Econometrica, Journal of the Econometric Society*, pages 213–244 (cited on page 4).

Barreda, Ines Moreno de (2024). "Cheap talk with two-sided private information". In: *Games and Economic Behavior* 148, pages 97–118 (cited on page 5).

Ben-Porath, Elchanan, Eddie Dekel, and Barton L Lipman (2014). "Optimal allocation with costly verification". In: *American Economic Review* 104.12, pages 3779–3813 (cited on page 5).

Berkson, Joseph (1944). "Application of the logistic function to bio-assay". In: *Journal of the American statistical association* 39.227, pages 357–365 (cited on page 29).

Beus, Johannes (February 2026). *AI Overviews in Deutschland: So stark sinken die Klickraten wirklich*. SISTRIX. URL: https://www.sistrix.de/news/ai-overviews-in-deutschland-so-stark-sinken-die-klickraten-wirklich/ (visited on 09/30/2026) (cited on page 3).

Blackwell, David (1951). "Comparison of Experiments". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Edited by Jerzy Neyman. Berkeley and Los Angeles: University of California Press, pages 93–102 (cited on pages 13, 14, 45).

Brunnermeier, Markus K and Jonathan A Parker (2005). "Optimal expectations". In: *American Economic Review* 95.4, pages 1092–1118 (cited on page 29).

Buja, Andreas, Werner Stuetzle, and Yi Shen (2005). "Loss functions for binary class probability estimation and classification: Structure and applications". In: *Working draft, November* 3, page 13 (cited on page 13).

Carpiano, Richard M, Timothy Callaghan, Renee DiResta, Noel T Brewer, Chelsea Clinton, Alison P Galvani, Rekha Lakshmanan, Wendy E Parmet, Saad B Omer, Alison M Buttenheim, et al. (2023). "Confronting the evolution and expansion of anti-vaccine activism in the USA in the COVID-19 era". In: *The Lancet* 401.10380, pages 967–970 (cited on page 8).

Chapekis, Athena, Anna Lieb, Sono Shah, and Aaron Smith (2026). "Investigating Click Behaviors On Google Search Result Pages That Produce an AI Overview". In: *arXiv preprint arXiv:2608.04831* (cited on page 2).

Charness, Gary and Chetan Dave (2017). "Confirmation bias with motivated beliefs". In: *Games and Economic Behavior* 104, pages 1–23 (cited on pages 6, 7).

Chen, Daniel L, Martin Schonger, and Chris Wickens (2016). "oTree—An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pages 88–97 (cited on page 56).

Cisternas, Gonzalo and Jorge Vásquez (2022). "Misinformation in social media: The role of verification incentives". In: *FRB of New York Staff Report* 1028 (cited on page 6).

Coutts, Alexander (2019). "Good news and bad news are still news: Experimental evidence on belief updating". In: *Experimental Economics* 22.2, pages 369–395 (cited on page 7).

Danz, David, Lise Vesterlund, and Alistair J Wilson (2022). "Belief elicitation and behavioral incentive compatibility". In: *American Economic Review* 112.9, pages 2851–2883 (cited on page 55).

Di Pei, Harry (2015). "Communication with endogenous information acquisition". In: *Journal of Economic Theory* 160, pages 132–149 (cited on page 5).

Druckman, James N and Mary C McGrath (2019). "The evidence for motivated reasoning in climate change preference formation". In: *Nature Climate Change* 9.2, pages 111–119 (cited on page 28).

Eagly, Alice H, Serena Chen, Shelly Chaiken, and Kelly Shaw-Barnes (1999). "The impact of attitudes on memory: An affair to remember." In: *Psychological bulletin* 125.1, page 64 (cited on page 32).

Ehrlinger, Joyce, Thomas Gilovich, and Lee Ross (2005). "Peering into the bias blind spot: People's assessments of bias in themselves and others". In: *Personality and Social Psychology Bulletin* 31.5, pages 680–692 (cited on page 33).

Eil, David and Justin M Rao (2011). "The good news-bad news effect: Asymmetric processing of objective information about yourself". In: *American Economic Journal: Microeconomics* 3.2, pages 114–138 (cited on pages 6, 7, 28).

Epley, Nicholas and Thomas Gilovich (2016). "The mechanics of motivated reasoning". In: *Journal of Economic perspectives* 30.3, pages 133–140 (cited on page 27).

Fryer Jr, Roland G, Philipp Harms, and Matthew O Jackson (2019). "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization". In: *Journal of the European Economic Association* 17.5, pages 1470–1501 (cited on page 32).

Gale, Douglas and Martin Hellwig (1985). "Incentive-compatible debt contracts: The one-period problem". In: *The Review of Economic Studies* 52.4, pages 647–663 (cited on page 5).

Garanko, Jana (December 2025). *Semrush AI Overviews Study: What 2025 SEO Data Tells Us About Google's Search Shift*. Semrush. URL: https://www.semrush.com/blog/semrush-ai-overviews-study/ (visited on 09/30/2026) (cited on page 3).

Garrett, R Kelly (2009). "Echo chambers online?: Politically motivated selective exposure among Internet news users". In: *Journal of computer-mediated communication* 14.2, pages 265–285 (cited on page 7).

Gentzkow, Matthew and Emir Kamenica (2014). "Costly persuasion". In: *American Economic Review* 104.5, pages 457–462 (cited on page 5).

Gentzkow, Matthew and Jesse M. Shapiro (2006). "Media Bias and Reputation". In: *Journal of Political Economy* 114.2, pages 280–316. DOI: 10.1086/499414 (cited on page 6).

Godes, David (2026). "Will the Truth Free Us from Misinformation?" In: *Management Science* (cited on page 6).

Gonçalves, Duarte, Jonathan Libgober, and Jack Willis (2026). "Retractions: Updating from complex information". In: *Review of Economic Studies* 93.1, pages 476–516 (cited on page 53).

Grossman, Sanford J (1981). "The informational role of warranties and private disclosure about product quality". In: *The Journal of law and Economics* 24.3, pages 461–483 (cited on page 6).

Grossman, Sanford J and Joseph E Stiglitz (1980). "On the impossibility of informationally efficient markets". In: *The American economic review* 70.3, pages 393–408 (cited on page 5).

Hart, William, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill (2009). "Feeling validated versus being correct: a meta-analysis of selective exposure to information." In: *Psychological bulletin* 135.4, page 555 (cited on page 32).

Hastall, Matthias R and Silvia Knobloch-Westerwick (2013). "Severity, efficacy, and evidence type as determinants of health message exposure". In: *Health Communication* 28.4, pages 378–388 (cited on page 28).

Healy, PJ and J Kagel (2023). *Testing elicitation mechanisms via team chat*. Technical report. Working paper (cited on page 55).

Hossain, Tanjim and Ryo Okui (2013). "The binarized scoring rule". In: *Review of Economic Studies* 80.3, pages 984–1001 (cited on page 55).

Hu, Desheng, Joachim Baumann, Aleksandra Urman, Elsa Lichtenegger, Robin Forsberg, Aniko Hannak, and Christo Wilson (2026). "Auditing Google's AI Overviews and Featured Snippets: A Case Study on Baby Care and Pregnancy". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Volume 20. 1, pages 1044–1062 (cited on page 3).

Iyengar, Shanto and Kyu S Hahn (2009). "Red media, blue media: Evidence of ideological selectivity in media use". In: *Journal of communication* 59.1, pages 19–39 (cited on page 7).

Keppo, Jussi, Giuseppe Moscarini, and Lones Smith (2008). "The demand for information: More heat than light". In: *Journal of Economic Theory* 138.1, pages 21–50 (cited on page 4).

Knobloch-Westerwick, Silvia, Benjamin K Johnson, and Axel Westerwick (2013). "To your health: Self-regulation of health behavior through selective exposure to online health messages". In: *Journal of Communication* 63.5, pages 807–829 (cited on page 28).

Koehler, Jonathan J (1993). "The influence of prior beliefs on scientific judgments of evidence quality". In: *Organizational behavior and human decision processes* 56.1, pages 28–55 (cited on page 7).

Kunda, Ziva (1990). "The case for motivated reasoning." In: *Psychological bulletin* 108.3, page 480 (cited on pages 27, 28).

Law, Ryan (February 2026). *Update: AI Overviews Reduce Clicks by 58%*. Ahrefs. URL: https://ahrefs.com/blog/ai-overviews-reduce-clicks-update/ (visited on 09/30/2026) (cited on page 3).

Levine, Jerome M and Gardner Murphy (1943). "The learning and forgetting of controversial material." In: *The Journal of Abnormal and Social Psychology* 38.4, page 507 (cited on page 32).

Lord, Charles G, Lee Ross, and Mark R Lepper (1979). "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." In: *Journal of personality and social psychology* 37.11, page 2098 (cited on pages 7, 32).

Luo, Yu and Jiaying Zhao (2019). "Motivated attention in climate change perception and action". In: *Frontiers in psychology* 10, page 1541 (cited on page 28).

Mahoney, Michael J (1977). "Publication prejudices: An experimental study of confirmatory bias in the peer review system". In: *Cognitive therapy and research* 1.2, pages 161–175 (cited on page 32).

Mayraz, Guy (2011). "Priors and Desires: A Model of Payoff-Dependent Beliefs". In: *CEP Discussion Paper* 1047 (cited on pages 28, 29).

McCullagh, Peter (2019). *Generalized linear models*. Routledge (cited on page 29).

Meriam Library, California State University, Chico (2010). *Evaluating Information: Applying the CRAAP Test*. URL: https://library.csuchico.edu/sites/default/files/craap-test.pdf (visited on 10/26/2026) (cited on page 12).

Milgrom, Paul R. (1981). "Good News and Bad News: Representation Theorems and Applications". In: *The Bell Journal of Economics* 12.2, pages 380–391. DOI: 10.2307/3003562 (cited on page 6).

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat (2022). "Managing self-confidence: Theory and experimental evidence". In: *Management Science* 68.11, pages 7793–7817 (cited on pages 7, 28).

Munro, Geoffrey D and Peter H Ditto (1997). "Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information". In: *Personality and Social Psychology Bulletin* 23.6, pages 636–653 (cited on page 7).

Nickerson, Raymond S (1998). "Confirmation bias: A ubiquitous phenomenon in many guises". In: *Review of general psychology* 2.2, pages 175–220 (cited on page 32).

Persico, Nicola (2000). "Information acquisition in auctions". In: *Econometrica* 68.1, pages 135–148 (cited on page 5).

Piccardi, Tiziano, Miriam Redi, Giovanni Colavizza, and Robert West (2020). "Quantifying engagement with citations on Wikipedia". In: *Proceedings of the web conference 2020*, pages 2365–2376 (cited on page 2).

Pouget, Sébastien, Julien Sauvagnat, and Stephane Villeneuve (2017). "A Mind is a Terrible Thing to Change: Confirmatory Bias in Financial Markets". In: *Review of Financial Studies* 30.6, pages 2066–2109. DOI: 10.1093/rfs/hhw100 (cited on page 32).

Pronin, Emily, Daniel Y Lin, and Lee Ross (2002). "The bias blind spot: Perceptions of bias in self versus others". In: *Personality and Social Psychology Bulletin* 28.3, pages 369–381 (cited on page 33).

Rabin, Matthew and Joel L Schrag (1999). "First impressions matter: A model of confirmatory bias". In: *The quarterly journal of economics* 114.1, pages 37–82 (cited on pages 27, 32).

Seybert, Nicholas and Robert Bloomfield (2009). "Contagion of wishful thinking in markets". In: *Management Science* 55.5, pages 738–751 (cited on page 28).

Shi, Xianwen (2012). "Optimal auctions with information acquisition". In: *Games and Economic Behavior* 74.2, pages 666–686 (cited on page 5).

Sims, Christopher A (2003). "Implications of rational inattention". In: *Journal of monetary Economics* 50.3, pages 665–690 (cited on page 11).

Sims, Christopher A (2006). "Rational inattention: Beyond the linear-quadratic case". In: *American Economic Review* 96.2, pages 158–163 (cited on page 11).

Stigler, George J (1961). "The economics of information". In: *Journal of political economy* 69.3, pages 213–225 (cited on page 5).

Taber, Charles S and Milton Lodge (2006). "Motivated skepticism in the evaluation of political beliefs". In: *American journal of political science* 50.3, pages 755–769 (cited on pages 7, 28, 32).

Thaler, Michael (2024). "The Fake News Effect: Experimentally Identifying Motivated Reasoning Using Trust in News". In: *American Economic Journal: Microeconomics* 16.2, pages 1–38. DOI: 10.1257/mic.20220146 (cited on pages 6, 7).

Tomkins, Andrew, Min Zhang, and William D Heavlin (2017). "Reviewer bias in single-versus double-blind peer review". In: *Proceedings of the National Academy of Sciences* 114.48, pages 12708–12713 (cited on page 32).

Townsend, Robert M (1979). "Optimal contracts and competitive markets with costly state verification". In: *Journal of Economic theory* 21.2, pages 265–293 (cited on page 5).

Wason, Peter C (1960). "On the failure to eliminate hypotheses in a conceptual task". In: *Quarterly journal of experimental psychology* 12.3, pages 129–140 (cited on page 27).

A Proofs

A.1 Proof of Lemma 1.

By Bayes rule:

$$\rho_1 \coloneqq \Pr(r = G \mid s = 1, \mu) = \frac{(1-\varepsilon)[\mu(2q-1)+(1-q)]}{\varepsilon[\mu(2p-1)+(1-p)]+(1-\varepsilon)[\mu(2q-1)+(1-q)]} \tag{10}$$

$$\rho_0 \coloneqq \Pr(r = G \mid s = 0, \mu) = \frac{(1-\varepsilon)[\mu(1-2q)+q]}{\varepsilon[\mu(1-2p)+p]+(1-\varepsilon)[\mu(1-2q)+q]} \tag{11}$$

It is easily showed that $\rho_1 \geq \rho_0$ implies $\mu \geq \frac{1}{2}$ and viceversa, with equality exactly at $\mu = \frac{1}{2}$.

A.2 Proof of Lemma 2.

To start, note that if $\mu = 0$, then Bayes rule implies $\mu'_{G,1} = \mu'_{B,1} = \mu'_{G,0} = \mu'_{B,0} = 0$, and, consequently, $\mu'_{G,1} - \mu'_{B,1} = \mu'_{G,0} - \mu'_{B,0} = 0$. By Equation 3, this means $\Delta_1(0) = \Delta_0(0) = 0$. An analogous argument can be made for $\mu = 1$. Next, note that if $\mu = 0$, then $\rho_1 = \rho_0$ (see Equation 10 and Equation 11), and $\mu'_{G,1} - \mu'_{B,1} = \mu'_{G,0} = \mu'_{B,0}$. By Equation 3, this implies $\Delta_1\left(\frac{1}{2}\right) = \Delta_0\left(\frac{1}{2}\right)$.

A.3 Proof of Proposition 1.

I start by introducing an intermediate result:

Lemma 3 *If $u(\mu, \theta) = -(\mu - \theta)^2$, the verification value $\Delta_s(\mu)$ is strictly quasi-concave on $(0, 1)$.*

Proof. Fix $s = 1$ and let $\phi(\mu) \coloneqq \frac{d \ln[\Delta_1(\mu)]}{d\mu}$. If $u(\mu, \theta)$ is a strictly proper scoring rule, the value function $V(\mu)$ is strictly convex (Blackwell, 1951); so, $\Delta_1(\mu) > 0$: then, $\phi(\mu)$ is well defined, and the stationary points of $\Delta_1(\mu)$ coincide with the solutions of $\phi(\mu) = 0$. To see why, note that $\phi(\mu) = \frac{\Delta'_1(\mu)}{\Delta(\mu)}$ for every $\mu \in (0, 1)$: as $\Delta(\mu) > 0$, this means $\phi(\mu) = 0 \Leftrightarrow \Delta'_1(\mu) = 0$. Also, $\phi(\mu)$ and

$\Delta'(\mu)$ always have the same sign. Additionally, if $u(\mu, \theta) = -(\mu - \theta)^2$:

$$\phi(\mu) = \frac{2}{\mu} - \frac{2}{1-\mu} - (2q-1)\left[\frac{1}{P_{G,1}(\mu)} + \frac{2(1-\varepsilon)}{P_1(\mu)}\right]$$

Since $\phi(\mu)$ is continuous, $\phi(0^+) = +\infty$, and $\phi(1^-) = -\infty$, $\phi(\mu)$ admits at least one zero in $(0,1)$ by the intermediate value theorem.[11] Indeed, $\phi(\mu) = 0$ admits exactly one solution: to prove this statement, first rearrange $\phi(\mu)$ as a polynomial $A_0 + A_1\mu + A_2\mu^2 + A_3\mu^3$, where:

$$A_0 \coloneqq (1-q)\left[2(1-\varepsilon)(1-q) + \varepsilon\right] > 0$$

$$A_1 \coloneqq (1-q)\left[(1-\varepsilon)(2q-1) - 4(1-\varepsilon)(1-q) - 2\varepsilon\right] + \frac{\varepsilon(2q-1)}{2}$$

$$A_2 \coloneqq -(2q-1)\left[(1-\varepsilon)(2q-1) + 5(1-\varepsilon)(1-q) + \frac{3\varepsilon}{2}\right] < 0$$

$$A_3 \coloneqq -(1-\varepsilon)(2q-1)^2 < 0$$

Then, note that, since A_1 has the only ambiguous sign, the sequence of coefficients (A_0, A_1, A_2, A_3) changes sign exactly once; so, $\phi(\mu)$ has exactly one positive real root by Descartes rule of signs. Let this be denoted by $\mu^* \in (0,1)$. Since $\phi(\mu)$ is continuous, $\phi(0) = A_0 > 0$, and $\phi(1) = -q[2(1-\varepsilon)q + \varepsilon] < 0$, $\phi(\mu) > 0$ on $(0, \mu^*)$, and $\phi(\mu) < 0$ on $(\mu^*, 1)$; so, Δ_1 is strictly increasing on $(0, \mu^*)$ and strictly decreasing otherwise: in other words, it is strictly quasi-concave. Strict quasi-concavity of Δ_0 can be showed analogously. ■

Now, for each signal realization, $s \in \{0,1\}$, let the critical cost κ be defined as:

$$\kappa \coloneqq \max_{\mu \in (0,1)} \Delta_s(\mu) = \Delta_s(\mu_s^*)$$

Importantly, existence and uniqueness of κ are guaranteed by Lemma 3. If $K > \kappa$, then $\Delta_s(\mu) < K$ for all $\mu \in (0,1)$; so, the verification region $\mathcal{V}_s^K$ is empty. Otherwise, $\mathcal{V}_s^K$ is non-empty as $\mu_s^* \in \mathcal{V}_s^K$;

[11] The intermediate value theorem can be easily generalized to open intervals with a reduction argument: assume $\lim_{x \to a^+} f(x)$ exists and is not equal to zero; then, there exists $c \in (a,b)$ s.t. $f(c)$ has its same sign. The same argument can be applied to $\lim_{x \to b^-} f(x)$ for some $d \in (a,b)$. Without loss of generality, c and d can be chosen so that $c < d$. Then, if the two limits have opposite signs, one can apply the IVT to the interval $[c,d]$ to obtain $x^* \in [c,d] \subset (a,b)$ s.t. $f(x^*) = 0$.

so, $\mathcal{V}_s^K$. Note that $\mathcal{V}_s^K$ must be an interval because Δ_s is strictly quasi-concave by Lemma 3: by definition, the upper level set of a quasi-concave function on $\mathbb{R}$ is convex, and a convex subset of $\mathbb{R}$ is an interval. Also, $\mathcal{V}_s^K$ is closed, since it is the pre-image of the closed set $[K, +\infty)$ under the continuous function Δ_s; therefore, $\mathcal{V}_s^K$ is a compact interval (Heine-Borel theorem).[12]

A.4 Proof of Theorem 1.

I start by introducing some preliminary notation. Let $P_{G,1} \coloneqq \mu q + (1-q)(1-\mu)$, $P_{B,1} \coloneqq p\mu + (1-p)(1-\mu)$, $P_{G,0} \coloneqq (1-q)\mu + q(1-\mu)$, $P_{B,0} \coloneqq (1-p)\mu + p(1-\mu)$, $P_{M,s} \coloneqq (1-\varepsilon)P_{G,s} + \varepsilon P_{B,s}$ and $D_s \coloneqq P_{G,s}P_{B,s}(P_{M,s})^2$ for $s \in \{0,1\}$. Then, the following result holds:

Lemma 4 *If $u(\mu, \theta) = -(\mu - \theta)^2$, then $\forall\, \mu \in [0,1]$:*

$$\Delta_1(\mu) \geq \Delta_0(\mu) \;\Leftrightarrow\; D_0(\mu) \geq D_1(\mu)$$

Proof. By Bayes rule, I can rewrite $\Delta_s(\mu)$ as reported in Equation 3 as:[13]

$$\Delta_s(\mu) :== \frac{\varepsilon(1-\varepsilon)\mu^2(1-\mu)^2(q-p)^2}{D_s} \qquad \text{for each } s \in \{0,1\} \tag{12}$$

Then, is easy to see that $\Delta_1(\mu) \geq \Delta_0(\mu) \Leftrightarrow D_0(\mu) \geq D_1(\mu)$. ■

Therefore, Theorem 1 can be showed by comparing D_1 and D_0 only. To do so, let me consider the following pairwise differences:

$$P_{G,0} - P_{G,1} = (2q-1)(1-2\mu)$$

$$P_{B,0} - P_{B,1} = (2p-1)(1-2\mu)$$

$$P_{M,0} - P_{M,1} = (1-2\mu)[(1-\varepsilon)(2q-1) + \varepsilon(2p-1)]$$

[12] Δ_s is continuous on $(0,1)$ because it is a composition of continuous functions on $(0,1)$.

[13] In particular:

$$\rho_s = \tfrac{(1-\varepsilon)P_{G,s}}{P_{M,s}}; \quad 1-\rho_s = \tfrac{\varepsilon P_{B,s}}{P_{M,s}}; \quad \mu'_{G,1} - \mu'_{B,1} = \tfrac{\mu(1-\mu)(q-p)}{P_{G,1}P_{B,1}}; \quad \mu'_{G,0} - \mu'_{B,0} = -\tfrac{\mu(1-\mu)(q-p)}{P_{G,0}P_{B,0}}$$

Since $q, p \geq \frac{1}{2}$ and $\varepsilon \in (0,1)$, the sign of these expressions is equal to the sign of $1-2\mu$: in particular, $\mu < \frac{1}{2} \Leftrightarrow P_{G,0} > P_{G,1}$, $P_{B,0} \geq P_{B,1}$, and $P_{M,0} > P_{M,1}$. It follows that:

$$\mu > \tfrac{1}{2} \quad \Rightarrow \quad D_0 = (P_{M,0})^2 P_{G,0} P_{B,0} > (P_{M,1})^2 P_{G,1} P_{B,1} = D_1 \quad \Leftrightarrow \quad \Delta_1 > \Delta_0$$

Conversely, $\mu > \frac{1}{2} \Rightarrow \Delta_1 < \Delta_0$. Finally, if $\mu = \frac{1}{2}$, all differences vanish; hence, $\Delta_0 = \Delta_1$.

A.5 Proof of Corollary 1.1.

To prove statement (i), take any $\hat{\mu} > \frac{1}{2}$ s.t. $\hat{\mu} \in \mathcal{V}_1^K$: by definition, this means $\Delta_1(\hat{\mu}) \geq K$. By Theorem 1, since $\hat{\mu} > \frac{1}{2}$, it must also be that $\Delta_0(\hat{\mu}) > \Delta_1(\hat{\mu})$; hence, $\Delta_0(\hat{\mu}) \geq K \Rightarrow \hat{\mu} \in \mathcal{V}_0^K$. The proof for $\hat{\mu} < \frac{1}{2}$ follows symmetrically. To prove statement (ii), I introduce two preliminary results:

Lemma 5 *For every $\mu \in [0,1]$, $\Delta_0(\mu) = \Delta_1(1-\mu)$.*

Proof. Take any $\mu \in [0,1]$. By Equation 12:

$$\Delta_1(1-\mu) = \frac{\varepsilon(1-\varepsilon)(1-\mu)^2\mu^2(q-p)^2}{P_{G,1}(1-\mu)P_{B,1}(1-\mu)P_1(1-\mu)^2} \equiv \frac{\varepsilon(1-\varepsilon)(1-\mu)^2\mu^2(q-p)^2}{P_{G,0}(\mu)P_{B,0}(\mu)P_0(\mu)^2} = \Delta_0(\mu)$$

where the equivalence follows from $P_{G,1}(1-\mu) = P_{G,0}(\mu)$ and $P_1(1-\mu) = P_0(\mu)\ \forall\, \mu \in [0,1]$.

■

Lemma 6 *For any $\mu \in \left(0, \frac{1}{2}\right)$, $\Delta_1(\mu) > \Delta_1(1-\mu)$.*

Proof. $\mu < \frac{1}{2} \Rightarrow \Delta_1(\mu) > \Delta_0(\mu)$ by Theorem 1. But $\Delta_0(\mu) = \Delta_1(1-\mu)$ by Lemma 5.

Hence, $\mu < \frac{1}{2} \Rightarrow \Delta_1(\mu) > \Delta_1(1-\mu)$. ■

Now, fix $s = 1$ and assume $\mathcal{V}_1^K$ is non-empty. Let the endpoints of $\mathcal{V}_1^K$ be denoted by $\{\underline{\mu}_1^K, \bar{\mu}_1^K\}$, with $\underline{\mu}_1^K < \bar{\mu}_1^K$. If $\bar{\mu}_1^K \leq \frac{1}{2}$, then $\underline{\mu}_1^K + \bar{\mu}_1^K \leq 1$ holds trivially. Otherwise, by Lemma 6:

$$\Delta_1(1-\bar{\mu}_1^K) > \Delta_1\big[1-(1-\bar{\mu}_1^K)\big] = \Delta_1(\bar{\mu}_1^K)$$

But since $\Delta_1(\bar{\mu}_1^K) = K$, this means $\Delta_1(1-\bar{\mu}_1^K) > K \;\Rightarrow\; (1-\bar{\mu}_1^K) \in \text{int}(\mathcal{V}_1^K) \;\Rightarrow\; \underline{\mu}_1^K < 1-\bar{\mu}_1^K$. Rearranging, I obtain $\underline{\mu}_1^K + \bar{\mu}_1^K < 1$. The result for $s=0$ follows by Lemma 5.

A.6 Proof of Theorem 2.

When $p > \frac{1}{2}$, Lemma 4 still holds; so, $\text{sign}\big(\Delta_1 - \Delta_0\big) = -\,\text{sign}\big(D_1 - D_0\big)$; therefore, it is enough to study under what conditions $D_1 > D_0$. To reach this goal, I first show a preliminary result. In what follows, let $\tau_N^{\mathcal{I}} = \big(3-2\varepsilon\big)\big(q-\frac{1}{2}\big) + \big(1+2\varepsilon\big)\big(p-\frac{1}{2}\big)$ and $\tau_E^{\mathcal{I}} = qp[(1-\varepsilon)q + \varepsilon p]^2 - (1-q)(1-p)[1-(1-\varepsilon)q - \varepsilon p]^2$, as defined in Section 3:

Lemma 7 *For any $\mu \in (0,1)$, $D_1 - D_0 = (2\mu-1)[\tau_N^{\mathcal{I}}\,\mu(1-\mu) + \tau_E^{\mathcal{I}}\,(2\mu-1)^2]$.*

Proof. First, let $F(\mu) \coloneqq D_1(\mu) - D_0(\mu)$, and note that — since D_1 and D_0 are both polynomials of at most degree 4 — $F(\mu)$ is also a polynomial of at most degree 4. Next, observe that $F(\mu) = D_1(\mu) - D_1(1-\mu)$ by Lemma 5; so, $F\big(\frac{1}{2}\big) = 0$. It follows that $F(\mu) = (2\mu-1)H(\mu)$ for some polynomial $H(\mu)$ of at most degree 3 (Factor theorem).

Next, observe that Lemma 5 requires $H(\mu) = H(1-\mu)$. This symmetry implies that $H(\mu)$ is a linear combination of 1 and $\mu(1-\mu)$.[14] Since $1 = 4\mu(1-\mu) + (2\mu-1)^2$, there must exist constants A, B such that $F(\mu) = (2\mu-1)\left[A\mu(1-\mu) + B(2\mu-1)^2\right]$.

Finally, I prove that $A = \tau_N^{\mathcal{I}}$ and $B = \tau_E^{\mathcal{I}}$. To find B, simply evaluate $F(\mu)$ at $\mu = 1$ to see that $B = F(1) = D_1(1) - D_0(1)$. Using the definitions of D_1 and D_0, this equality becomes $\tau_E^{\mathcal{I}}$. To find A, first observe that $A = 2F'\big(\frac{1}{2}\big)$. By directly computing the derivative, I obtain $A = \tau_N^{\mathcal{I}}$. ■

Next, let $B(\mu) \equiv \tau_N^{\mathcal{I}}\,\mu(1-\mu) + \tau_E^{\mathcal{I}}\,(2\mu-1)^2$. For the sake of exposition, I first assume $\mu \in \big(\frac{1}{2}, 1\big)$, so that $(2\mu-1) > 0$. Then, Lemma 7 implies that $\text{sign}(D_1 - D_0) = \text{sign}\, B(\mu)$. It follows that, if $\tau_N^{\mathcal{I}}\tau_E^{\mathcal{I}} \geq 0$, three cases are possible: if $\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}} \geq 0 \;\wedge\; (\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}}) \neq (0,0)$, then $B(\mu) >$

[14] To see why, simply write H as a polynomial of degree 3, and use the symmetry to derive restrictions on the coefficients.

$0 \,\forall\, \mu \in (0,1) \Rightarrow \Delta_1(\mu) < \Delta_0(\mu) \,\forall\, \mu \in (0,1)$; so, sources of disconfirmatory signals are strictly more valuable to verify regardless of the DM's prior. If $\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}} \leq 0 \wedge (\tau_N^{\mathcal{I}}, \tau_E^{\mathcal{I}}) \neq (0,0)$, then $B(\mu) < 0 \;\forall\; \mu \in (0,1) \Rightarrow \Delta_1(\mu) > \Delta_0(\mu) \;\forall\; \mu \in (0,1)$; so, sources of confirmatory signals are strictly more valuable to verify regardless of the DM's prior. Otherwise, if $\tau_N^{\mathcal{I}} = \tau_E^{\mathcal{I}} = 0$, then $B(\mu) = 0 \;\forall\; \mu \in (0,1) \Rightarrow \Delta_1(\mu) = \Delta_0(\mu) \,\forall\, \mu \in (0,1)$; so, sources of either signals are equally valuable to verify regardless of the DM's prior. Finally, to extend the reasoning to $\mu \in \left(0, \frac{1}{2}\right)$, note that by Lemma 5:

$$\Delta_1(\mu) - \Delta_0(\mu) = \Delta_0(1-\mu) - \Delta_1(1-\mu) = -\big[\Delta_1(1-\mu) - \Delta_0(1-\mu)\big]$$

So, witching from μ to $(1-\mu)$ inverts both the sign of $\Delta_1 - \Delta_0$ and the confirmatory labels, which implies that the ranking at μ coincides with that at $(1-\mu)$.

A.7 Proof of Theorem 3.

Let $p < \frac{1}{2}$ and $\tau_N^{\mathcal{I}} \tau_E^{\mathcal{I}} < 0$. Under these assumptions, Lemma 7 still holds, so:

$$\forall\, \mu \in (0,1), \quad D_1 - D_0 = (2\mu - 1)[\tau_N^{\mathcal{I}}\, \mu(1-\mu) + \tau_E^{\mathcal{I}}\, (2\mu-1)^2]$$

To start, let $B(\mu) \equiv \tau_N^{\mathcal{I}}\, \mu(1-\mu) + \tau_E^{\mathcal{I}}\, (2\mu-1)^2$, and assume $\mu \in \left(\frac{1}{2}, 1\right)$, so that $(2\mu - 1) > 0$. Then:

$$\text{sign}(D_1 - D_0) = \text{sign}[B(\mu)]$$

Note that if $\tau_N^{\mathcal{I}} \tau_E^{\mathcal{I}} < 0$ (i.e. either $\tau_N^{\mathcal{I}} < 0 < \tau_E^{\mathcal{I}}$ or $\tau_E^{\mathcal{I}} < 0 < \tau_N^{\mathcal{I}}$), $\text{sign}(D_1 - D_0) = \text{sign}\, \tau_N^{\mathcal{I}}$ when $\mu \to \frac{1}{2}$, while $\text{sign}(D_1 - D_0) = \text{sign}\, \tau_E^{\mathcal{I}}$ when $\mu \to 1$: in words, $(D_1 - D_0)$ has opposite signs as prior beliefs approach the endpoints. This, together with the fact that $B(\mu)$ is an affine function — hence, continuous and strictly monotone — imply that $(D_1 - D_0)$ has a unique zero on $\mu \in \left(\frac{1}{2}, 1\right)$, denoted by $\hat{\mu}$ (Intermediate Value Theorem adjusted for open intervals — see Footnote

11 — plus a monotonicity argument to show uniqueness).[15] Since $\text{sign}(B) = \text{sign}(\tau_N^{\mathcal{I}})$ near $\mu = \frac{1}{2}$ and $\text{sign}(B) = \text{sign}(\tau_E^{\mathcal{I}})$ near $\mu = 1$, the ranking of sources imposed by $\text{sign}(\tau_N^{\mathcal{I}})$ holds on $\left(\frac{1}{2}, \hat{\mu}\right)$, and reverses on $\left(\hat{\mu}, 1\right)$. At $\mu = \hat{\mu}$, both sources are equally valuable to verify. The result can be extended to $\mu \in \left(0, \frac{1}{2}\right)$ via Lemma 5 using a symmetry argument.

A.8 Proof of Proposition 2.

First, note that Equation 8 implies $\tilde{\rho}_s = \frac{\rho_s e^{\eta\, m(s)}}{1-\rho_s+\rho_s e^{\eta\, m(s)}}$. Then, fix $s = 1$. To solve $\tilde{\Delta}_1(\mu) \geq \Delta_1(\mu)$, observe that, if $\eta > 0$ and $\rho_1 \in (0, 1)$,substituting the distorted posterior $\tilde{\rho}_1$ and rearranging yields:

$$\tilde{\rho}_1(1-\tilde{\rho}_1) \geq \rho_1(1-\rho_1) \;\Leftrightarrow\; e^{\eta} \geq (1-\rho_1+\rho_1 e^{\eta})^2 \;\Leftrightarrow\; e^{\eta/2} \geq 1-\rho_1+\rho_1 e^{\eta} \;\Leftrightarrow\; \rho_1 \leq \frac{1}{1+e^{\eta/2}}$$

Thus, the condition holds if and only if ρ_1 lies below $\frac{1}{1+e^{\eta/2}}$. The proof for $s = 0$ is analogous.

A.9 Proof of Proposition 3.

I start by introducing an intermediate result:

Lemma 8 *The posterior probability that the source is reliable $\rho_s(\mu)$ is strictly increasing in the prior μ given a positive signal ($s = 1$), but strictly decreasing in μ given a negative one ($s = 0$).*

Proof. It is easily showed that $\text{sign}\left(\frac{d\rho_1}{d\mu}\right) = \text{sign}(q-p) > 0$ and $\text{sign}\left(\frac{d\rho_0}{d\mu}\right) = \text{sign}(p-q) < 0$; so, $\rho_1(\mu)$ is strictly increasing in μ and $\rho_0(\mu)$ is strictly decreasing in μ. ■

Next, fix a $K > 0$, and let $s = 1$. By definition, $\tilde{\mathcal{V}}_1^K \subseteq \mathcal{V}_1^K$ is equivalent to $\tilde{\Delta}_1(\mu) \leq \Delta_1(\mu) \;\; \forall\, \mu \in [0, 1]$. By Proposition 2, the inequality above is satisfied if and only if $\rho_1(\mu) \geq \frac{1}{1+e^{\eta/2}}$. By Lemma 8,

[15] Solving $B(\mu) = 0$ under the assumption of $\mu \in \left(\frac{1}{2}, 1\right)$, gives $\hat{\mu} = \frac{1}{2}\left(1 + \sqrt{\frac{\tau_N^{\mathcal{I}}}{\tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}}}}\right)$. Importantly, the delta of the equation always lies if $(0, 1)$: if $\tau_N^{\mathcal{I}} > 0 > \tau_E^{\mathcal{I}}$, then $\tau_N^{\mathcal{I}} > 0 \wedge \tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}} > 0 \Rightarrow \frac{\tau_N^{\mathcal{I}}}{\tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}}} > 0$ and $\tau_E^{\mathcal{I}} < 0 \Rightarrow \frac{\tau_N^{\mathcal{I}}}{\tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}}} < 1$; similarly, if $\tau_N^{\mathcal{I}} < 0 < \tau_E^{\mathcal{I}}$ then $\tau_N^{\mathcal{I}} < 0 \wedge \tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}} < 0 \Rightarrow \frac{\tau_N^{\mathcal{I}}}{\tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}}} > 0$, and $\tau_E^{\mathcal{I}} > 0 \Rightarrow \frac{\tau_N^{\mathcal{I}}}{\tau_N^{\mathcal{I}} - 4\tau_E^{\mathcal{I}}} < 1$.

a sufficient condition for this inequality to be satisfied is requiring it to hold at the lowest prior, i.e. $\mu = 0$. Solving for η leads to:

$$\eta \geq 2 \log \left[\frac{1 - \rho_1(0)}{\rho_1(0)} \right] \Rightarrow \tilde{\mathcal{V}}_1^K \subseteq \mathcal{V}_1^K$$

The proof for $s = 0$ is analogous.

A.10 Proof of Proposition 4.

Fix $\alpha \in [0, 1]$, and let $\mu \in \left(0, \frac{1}{2}\right)$, so that $g(\alpha) = \Pr(\text{verify} \mid s = 0, \alpha) - \Pr(\text{verify} \mid s = 1, \alpha)$. By Equation 9, this implies:

$$g(\alpha) = (1 - \alpha) \Pr(\text{verify} \mid \hat{s} = 0, \alpha = 0) + \alpha \Pr(\text{verify} \mid \hat{s} = 1, \alpha = 0) - \Pr(\text{verify} \mid \hat{s} = 1, \alpha = 0)$$

where $\Pr(\text{verify} \mid s, \alpha) = \Pr(\text{verify} \mid \hat{s}, 0)$ holds because the DM is observationally Bayesian conditional on perceived signals. By Definition 4, this is equivalent to $g(\alpha) = (1 - \alpha)g(0)$. The proof for $\mu \in \left(0, \frac{1}{2}\right)$ follows analogously.

B Lab Experiment

This experimental set-up allows to closely reproduce the information structure proposed in the theoretical framework: in particular, it is designed so that prior beliefs, source precisions, and the share of bad sources can be set exogenously and disclosed to DMs, and verification perfectly reveals the source type. Such features are admittedly restrictive: I see this as a first test of the model's mechanisms, and relaxing such assumptions represents a natural next step for further research.

B.1 Experimental Design

The experiment consists in a belief updating problem with a verification choice. In each *round* t, participants form beliefs over a binary state θ_t, which represents the color of a ball — called *truth ball* — drawn from a box, i.e. $\theta_t \in \{red, blue\}$.[16] Importantly, the box is changed at every period, and the ball is re-drawn. Throughout the paper, balls of the same color of the truth ball are labeled as *good balls*, while I refer to the others as *bad* ones.[17] After the truth ball has been selected, two urns are filled with 10 red and blue balls, each in different proportions, so that one, the *good urn*, contains more good balls than the other, the *bad urn*. Then, an urn is randomly selected, a *signal* ball is drawn from it, and its color is reported to the participant, who can opt to learn which urn the draw came from (i.e. to *verify* the source) at a cost. My main interests lie in this verification choice, and the updated beliefs on the state and the source type. The detailed flow of the experiment is reported in Figure 6.

Environment. To exogenously vary prior beliefs and source precisions, I modify the compositions of the box and urns at each round.[18] To set prior beliefs, I change the number of red balls in the

[16] The terminology "truth ball" is borrowed from Gonçalves et al. (2026).

[17] These labels are only used in the paper for explanatory purposes. Participants never encounter them.

[18] I refer to them as "box" and "urns" to help participants distinguish between the sets of balls used to determine prior beliefs and those used to set source precisions, respectively.

box while keeping the total number equal to 10: in this way, I obtain $\mu = \Pr_t(\theta_t = red) = \frac{\#\, red\, balls}{10}$. Similarly, to vary source precisions, I keep the total number of balls in each urn fixed at 10 while changing the proportion of good balls, so that the precision of each urn (source) is equal to $\frac{\#\, good\, balls}{10}$. Consequently, a good urn with precision $q \in (\frac{1}{2}, 1)$ contains $10q$ good balls and $10(1-q)$ bad ones, while a bad urn with precision $p \in [\frac{1}{2}, q)$ contains $10p$ and $10(1-p)$ good and bad balls, respectively. Participants are informed about all compositions, so that $\{\mu, q, p\}$ are known.

Share of bad sources. To vary the share of bad sources, I use a simple randomization procedure. At each round t, participants are informed of a threshold $N_t \in [1, 9]$; then, an integer from 1 to 10 is drawn from an uniform distribution: if the realized number is at or below the threshold, the signal ball is drawn from the bad urn; otherwise, from the good one. Thus, participants know that the probability of receiving a signal from the bad urn in the corresponding round is $\varepsilon_t = \frac{N_t}{10}$. I vary ε_t across periods by changing the threshold N_t.

Source Verification. At each round t, participants receive a message that looks as follows:

"*The signal ball is red [blue].*"

This message, or signal, is always truthful, meaning that it always correctly reports the color of the signal ball; however, it is not perfectly informative on the state because of the mixed compositions of both urns. After the message has been received, participants are asked their maximum willingness to accept (WTA), expressed in pounds (£), to forego source verification: in other words, they are asked the minimal amount of money they are willing to accept not to learn which urn the signal ball was drawn from. To incentivize this question, I apply the Becker-DeGroot-Marschak (BDM) mechanism: a random number $K \sim U[0, X]$ is drawn and compared with the reported WTA; if

it is weakly above this value, the participant receives £K and does not learn anything; otherwise, they do not receive any premia, but not discover the signal source.

Posterior beliefs. At each round t, I elicit participants' posterior beliefs about both the urn type r and the state θ_t. To minimize the scope for ex-post rationalization, beliefs about the urn type are elicited before the realization of K, while beliefs about the state are elicited after r is revealed when verification occurs, and after the realization of K otherwise. Truthful reporting in both elicitations is incentivized using a binarized scoring rule (BSR) (Hossain and Okui, 2013): specifically, let $\hat{p}$ denote the participant's reported probability of the relevant event (i.e. $\theta_t = red$ or $r = G$), and let I be an indicator equal to one if that event occurs, and zero otherwise; then, participants receive £0 with probability $(I - \hat{p})^2$, and £5 with complementary probability.[19] The instructions explicitly inform participants that, under this payment rule, reporting their true beliefs is in their best interest, and provide a numerical example illustrating this property.[20] Finally, immediately after eliciting posterior beliefs about the source type, I also collect an unincentivized measure of the potential scope for belief updating under verification.

[19]This means that, for posterior beliefs about the state, the probability to receive £L is $(\mathbb{1}\{\theta_t = red\} - \hat{\mu}_{t+1})^2$, while is $(\mathbb{1}\{r = G\} - \hat{\rho}_t^s)^2$ for posteriors about the urn type.

[20]In a recent paper, Danz et al. (2022) show that providing information about the incentive properties of the BSR can increase deviations from truthful reporting. To mitigate this concern, I keep the instructions as simple as possible, and provide a numerical example to illustrate the recommended reporting strategy is indeed optimal. They also show that the BSR tends to distort reported beliefs toward the center of the distribution, although misreporting appears to be less pronounced on online platforms (Healy and Kagel, 2023). This potential distortion does not affect most of the hypotheses tested in the experiment.

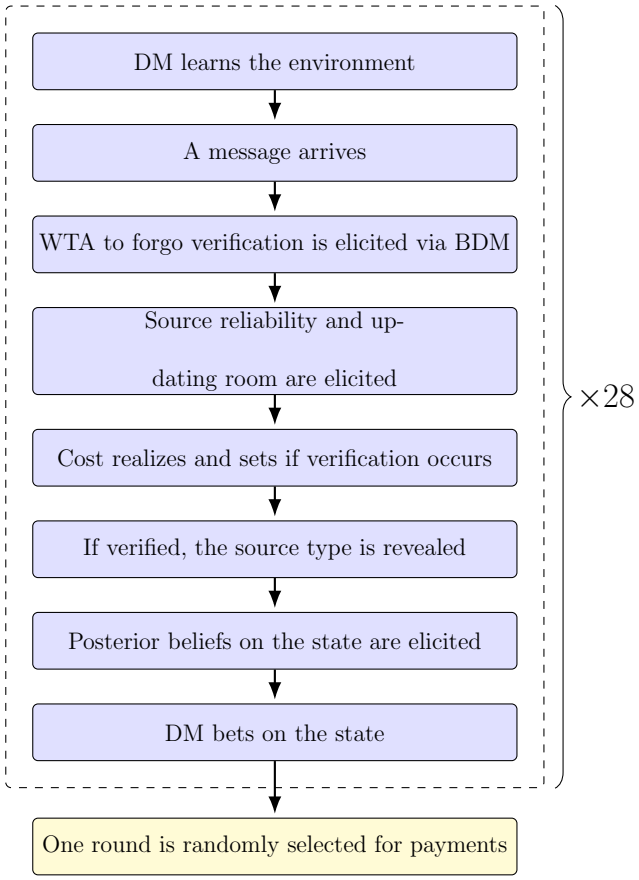


Figure 6: Flow of the experiment.

B.2 Implementation Details

Interface. I will implement the study in oTree (Chen et al., 2016). The experiment will count two slots of 14 incentivized rounds each, for a total of 28 rounds (25 for identification purposes, plus 3 replications to test for internal consistency). The full list of parameters used in the experiment is reported in Table 2. To minimize fatigue, the slots will be divided by a mandatory one-minute interval to rest; also, I will simplify the belief elicitation process by asking participants to report their posteriors using a slider on a grid ranging from 0 to 100. The experiment will open with instructions, followed by two unincentivized practice rounds, and will end with a set of socio-demographic questions. I estimate the experiment to last approximately 1 hour.

Block	Vector $\{\mu, \mathbf{q}, \mathbf{p}, \varepsilon_{\mathbf{t}}\}$	Signal s
1 (Prior)	$\{0.1, 0.80, 0.60, 0.3\}$	Both 0 and 1
	$\{0.2, 0.80, 0.60, 0.3\}$	
	$\{0.3, 0.80, 0.60, 0.3\}$	
	$\{0.4, 0.80, 0.60, 0.3\}$	
	$\{0.5, 0.80, 0.60, 0.3\}$	
	$\{0.6, 0.80, 0.60, 0.3\}$	
	$\{0.7, 0.80, 0.60, 0.3\}$	
	$\{0.8, 0.80, 0.60, 0.3\}$	
	$\{0.9, 0.80, 0.60, 0.3\}$	
2 (Precision Gap)	$\{0.5, 0.90, 0.50, 0.3\}$	Either 0 or 1
	$\{0.5, 0.85, 0.55, 0.3\}$	
	$\{0.5, 0.75, 0.65, 0.3\}$	
3 (Bad sources share)	$\{0.5, 0.80, 0.60, 0.1\}$	
	$\{0.5, 0.80, 0.60, 0.5\}$	
	$\{0.5, 0.80, 0.60, 0.7\}$	
	$\{0.5, 0.80, 0.60, 0.9\}$	

Table 2: List of vectors used in the experiment. The order is randomized. The value of s for blocks 2 and 3 is randomly chosen for each individual. Note that the result corresponding to the highlighted vector at the corresponding signal realization can be used in the other two treatment blocks as well.

Subjects Pool. TBD.

Other Incentives. At the end of the experiment, one round is selected at random for payment. For the selected round, participants receive the payments determined by the BDM and the two BSRs, together with a bonus of £5 if they correctly guessed the color of the true ball. In addition, all participants receive a £10 show-up fee.