Belief updating with misinformation: The role of prior beliefs*

Lars Wittrock† Martin Strobel‡ Elias Tsakas§

Department of Microeconomics and Public Economics, Maastricht University

April, 2026

Click here for most recent version

Abstract

In this paper we study experimentally how people process verifications of previously received information. We propose a theoretical model that builds on the standard framework of Grether (1980) to provide a unified mechanism that describes how people react to retractions and confirmations. Our data corroborates the predictions of this model, showing that reactions to verifications are explained by the prior beliefs, i.e., for low priors subjects underreact to retractions and overreact to confirmations, whereas for high priors they overreact to retractions and underreact to confirmations. Our explanation is consistent with the idea that people overreact to unexpected information, which is a direct consequence of base rate neglect within our model. Our findings are qualitatively robust in various dimensions that we tested, and indicate that the way people react to retractions is more nuanced than it has been thought so far.

Keywords: Belief updating, retractions, confirmations, prior beliefs.
JEL Classification: D83, D91, C91

*We are greatly indebted to the Editor Nageeb Ali, and two anonymous referees for their instrumental input that really helped us improve the paper. We would also like to thank Collin Raymond, Yucheng Liang, Joshua Miller, Ted O'Donoghue, Peter Schwardmann, Ori Heffetz, Alex Rees-Jones, Frauke Stehr, and seminar audiences at the ESA World in Lyon, MBEES in Maastricht, ESA North America in Santa Barbara, BERG (Cornell, Wharton and HUJI), Maastricht University for very useful comments and suggestions.

†Homepage: https://www.larswittrock.eu; E-mail: lars.wittrock@proton.me

‡Homepage: https://martinstrobel.net; E-mail: m.strobel@maastrichtuniversity.nl

§Homepage: www.elias-tsakas.com; E-mail: e.tsakas@maastrichtuniversity.nl

1 Introduction

In many situations it is unclear at first if information is fully reliable. Such situations include reading a news article, hearing a claim from a politician, watching a reel on social media, or reading an academic working paper before publication. In such situations initially uncertain information is frequently checked for the purpose of being verified. And as a result it is sometimes retracted or confirmed.

In practice, despite such fact-checking efforts, many people continue to believe false information, even after it has been publicly retracted, e.g., many people still believe that vaccinations lead to autism, or that Obama was born outside the USA, or that Iraq possessed weapons of mass destruction (Lewandowsky *et al.*, 2012). On the flip side, people often remain skeptical about vaccine safety even after it has been scientifically confirmed that those are safe (Browne, 2018). The bottom line in all such cases is that people differ in how they respond to retractions or confirmations of previous information. This calls for understanding the pattern of how people respond to information checking.

This question is not new. Particularly in psychology it has been studied for years, with the general conclusion being that people in general continue to be influenced by retracted information, in the sense that there are spillovers in the direction of the initial signal even after information has been retracted (Ecker *et al.*, 2022, and references therein). This bias is known as the continued influence effect. Similarly, in political science, it is found that correcting false information is often ineffective (Nieminen and Rapeli, 2018). However, in most of these studies the effect can also be attributed to context-specific characteristics, e.g., the nature of the narrative, political motivation, the presence of motivated reasoning, etc.

As a result, a natural question arises of whether failure to correctly process information retractions and/or verifications is different at a fundamental level from incorporating standard signals in the first place. And in order to address this question, one would need to focus on a neutral experimental setting where all the aforementioned context-specific channels have been exogenously shut down. This approach has been recently taken in the seminal paper of Goncalves *et al.* (2026), who find that the continued influence effect still holds in a variant of the standard balls-and-urns experiment. Their findings suggest that the failure to process information retractions correctly can be classified as a novel updating bias, which they attribute to the increased complexity of retractions compared to regular signals. Nonetheless, given the foundational nature of the question more work is needed towards a general theory on how people update when processing information checks. In this spirit, our paper extends the work of Goncalves *et al.* (2026), by asking two sets of questions that can eventually contribute towards such a theory.

First, do people react to confirmations of previously uncertain information in a different way compared to regular signals? And if yes, is the underlying mechanism similar to the one that drives reaction to retractions? Second, what do the dynamics of processing verifications look like? In particular, do people maintain the same mechanism(s) of how they process verifications over time? And moreover, do past information checks affect future belief updating?

To address these questions we introduce a variant of the design used in Goncalves *et al.*

(2026), who have already modified the standard balls-and-urns experiment (Benjamin, 2019, and references therein) in a way that allows for non-trivial retractions/confirmations of previously received signals. The two key differences in our benchmark experiment are that (a) confirmations do not lead to direct revelation of the true state,[1] and (b) histories are longer, with new signals being received after retractions/confirmations have taken place.

Let us explain in more detail how our benchmark experiment is structured. We begin with two urns that differ in the distribution of balls, e.g., the so-called blue urn that contains 3 blue and 1 red ball each, and the red urn that contains 1 blue and 3 red balls each. An urn is randomly picked, and 6 more balls are added to it, 3 of which are blue and 3 red. This means that the selected urn now contains two types of balls, viz., informative (the 4 balls that were initially contained in the urn) and uninformative (the 6 balls that were added ex post). Then, we proceed by consecutively drawing 9 random balls (with replacement) from the urn, and after each draw we tell the subjects the color of the ball. After receiving each such a signal the subjects report their belief about the red urn having been picked in the first place. In addition, 3 randomly selected draws (between signal 3 and signal 8) are verified, in the sense that we also tell the subjects whether they are informative or uninformative, and we elicit again their beliefs about the color of the urn. Announcement of an informative ball is interpreted as a confirmation, and that of an uninformative ball as a retraction.

Since our first question seeks for analogies between the way in which people process confirmations and the way in which they process retractions, it is important to study both types of verifications carefully in our setting. Surprisingly, we find that the reaction to both types of signals is explained by the prior beliefs within the standard model of belief updating (Grether, 1980). In particular, when subjects have a low prior probability, they will underreact to retractions and overreact to confirmations. On the other hand, when they have a high prior probability, they will overreact to retractions and underreact to confirmations. This finding is illustrated in Figure 1.[2]

The proposed intuition is quite simple in fact. Subjects seem to overreact to unexpected information and underreact to expected information.[3] This is simply a consequence of base-rate neglect, i.e., base-rate neglect implies that the role of the signal increases fast as the prior decreases, and therefore the agent reacts more and more to unexpected information.[4] Specifically, low prior belief implies that the signal is unexpected, and therefore subjects overreact to both the initial signal and the subsequent confirmation, while they underreact to the retraction.[5] This corresponds to the first part of the figure. Analogous logic applies

[1] In one of their robustness treatments towards the end of their paper, Goncalves *et al.* (2026) also garble confirmed information. However, they only use this treatment to study the robustness of their main finding —regarding the continued influence effect— to a setting where confirmations do not fully reveal the state, and they do not discuss how confirmations are treated.

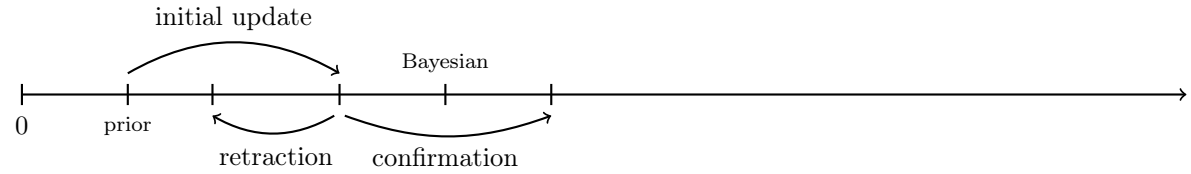
[2] This result was not preregistered. However, given its importance and its robustness to all the tests we ran, we felt compelled to present it as the main result of the paper.

[3] The role of surprise on belief updating has been recently studied within a specific context (Bronnikov *et al.*, 2026).

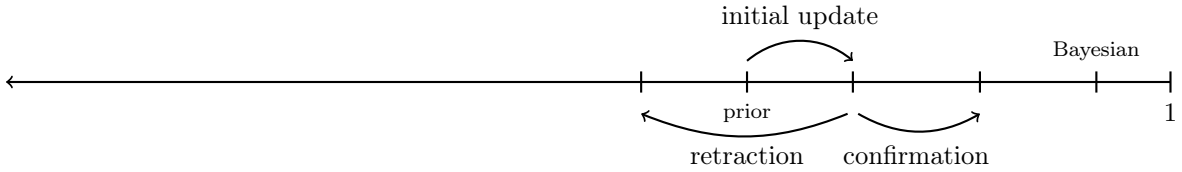
[4] Note that the converse would hold if the agent exhibited base-rate overuse, i.e., he would react less to unexpected information and more to expected information.

[5] This also explains why the reaction to the initial signal is correlated with the reaction to retractions.

to the second case where the prior belief is high, and it corresponds to the second part of the figure.



(a) **Low prior belief**: underreaction to retraction and overreaction to confirmation.



(b) **High prior belief**: overreaction to retraction and underreaction to confirmation.

Figure 1: Illustration of our main findings.

This finding reveals a more nuanced picture than the one that has been documented so far in the literature on retractions (Goncalves *et al.*, 2026, and references therein). In particular, we also find that the continued influence effect holds for a large range of prior beliefs. However, it does not seem to be a universal effect, as previously thought. But most importantly, going back to our original research question, it reveals that the basic mechanisms that drive reactions to retractions and confirmations are analogous to each other.

So, let us now look into whether the aforementioned pattern —both for retractions and confirmations— is consistent over time. To do so, we split our experiment into three blocks of periods (viz., early, middle, late) and we repeated our analysis for each of them separately. What we find is that qualitatively the exact same effects persist. Let us elaborate.

Starting with retractions, for each block of periods there is a corresponding range of low prior beliefs where the subjects underreact to retractions, while they overreact to retractions for the remaining prior beliefs which lie outside this range. But we also find that (1) the magnitude of biases —both overreactions and underreactions— decreases, and (2) the range of high priors where subjects overreact to retractions gradually shrinks. Starting from the former, this observation is consistent with the subjects learning how to process retractions. In turn, this fits with the story of Goncalves *et al.* (2026), according to which retractions are inherently more complex than regular signals. Then, switching to our second conclusion, we see that in late periods our results look much closer to the ones of Goncalves *et al.* (2026), in that we see underreaction to retractions for a rather large range of prior beliefs.

Then, switching to confirmations, we also find that for each block of periods there is a corresponding range of low prior beliefs where the subjects overreact to confirmations, while they underreact to confirmations for the remaining prior beliefs which lie outside this range. Then, similarly to the case of retractions, the range of high priors where the subjects underreact to confirmations shrinks. However, it is not the case that biases decrease uniformly over time.

Putting all this together, suggests that the mechanism is qualitatively stable over time, although the magnitudes of the biases might change, and so does the relative range of underreaction vs overreaction.

Finally, we asked whether verifications have a downstream effect in the way subjects update regular signals. To address this, we split our observations with respect to the profile of verifications the subjects had experienced previously. In particular, we looked at updating of regular signals conditional on (1) no previous verification, (2) a single retraction, (3) a single confirmation, (4) two retractions, (5) two confirmations, and (6) one retraction and one confirmation.

In general the pattern of updating remains the same across the six aforementioned conditions. Yet the sizes of the bias change. In particular, after subjects have observed two confirmations, their base-rate use increases towards the rational level, and their overall updating moves closer towards the Bayesian benchmark, especially for intermediate and high priors.

We also tested whether retractions are processed differently from a signal of the opposite color. Of course, informationally the two are equivalent. However, what our model predicts and we also see in our data is that subjects react more strongly to the signal of the opposite color compared to the retraction, for any prior belief. This is consistent with the finding of Goncalves *et al.* (2026), who also see that retractions are processed in a way which is inherently different to opposite signals. In fact, what we find is also consistent with their explanation, i.e., that retractions are more complex than opposite signals.

The paper is organized as follows. Section 2 contains a literature review. In Section 3 we present the theoretical model that we will use to formalize our hypotheses. Section 4 contains a description of our experimental design and our empirical strategy. In Section 5 we present our main results. Section 6 revisits the difference between retractions and opposite signals. Section 7 presents our additional treatments. Section 8 concludes. Some proofs, the experiments instructions, figures and tables are all relegated to the appendices.

2 Literature

Our paper relates to three separate literatures. Firstly, to the literature on retractions of information, secondly, to the literature on belief updating biases and thirdly, to the political science literature on fact-checking of misinformation.

Psychology: Continued Influence Effect

Starting with the first stream, psychologists have studied belief retractions extensively, albeit always in specific contexts which often pose identification issues, due to the fact that other mechanisms that interact with retractions are in play. We will talk about this literature in the next couple of paragraphs, but first we will refer to the paper of Goncalves *et al.* (2026) which is closest to ours and which first introduced this problem in an abstract setting.

In their work, Goncalves *et al.* (2026) focus on retractions exclusively. Their design consists of shorter tasks with 4 periods and a single verification. As we have already

described in this introduction, their main finding is that subjects underreact to retractions, thus exhibiting the continued influence effect. Then using additional measurements they took during their experiment, they attribute this phenomenon to the fact that retractions are inherently more complex than regular signals. As we have already discussed, a lot of what they find is also corroborated in our experiment. However, we also show that the picture is more nuanced, in that there is also a region of high prior beliefs where subjects systematically overreact to retractions.

Now, going back to the psychology literature, we refer to Ecker *et al.* (2022) for a recent review of the continued influence effect. One of the questions we study is closely related to this effect. However, our experimental design differs to those used in the psychology literature in several ways. We aim to remove any doubt about the credibility of the retraction as well as context dependencies such as different causal structures. We also focus on the statistical inference from retractions rather than memory and recall of past retracted information. Let us explicitly review some of the main papers.

Johnson and Seifert (1994) introduced the "continued influence effect" which states that people continue to be influenced by retracted information. In their setting people continue to rely on retracted information regarding the cause of a warehouse fire. The authors suggest that the reason for this bias is the causal structure of the displayed information.

Roets *et al.* (2017) provide an additional reason: cognitive ability. They study a setting in which participants are asked to evaluate a person on several dimensions based on a variety of information provided to them and one salient piece of information is later retracted. They find that people with high cognitive ability end up with a belief no different to a control group that never saw the retracted information. However, participants with lower cognitive ability continued to rely on the retracted information, leading to the continued influence effect on average. We find some supporting evidence in our setting. People who answered more CRT questions correctly also hold slightly more accurate beliefs after retractions of past information.[6] Nonetheless, categorizing subjects into types is far less predictive than people's initial response.

Pennycook and Rand (2021) provide an overview of the psychology of fake news. They discuss extensively why people believe and share fake information. Also they name the lack of careful reasoning as a key driver for the belief in fake information.

Ecker *et al.* (2010) study the influence of ex-ante warning subjects about misinformation. They consider a setting with a causal narrative and retract a key piece of information at a later point. They find that warnings about misinformation reduce the continued influence effect but fail to fully eliminate it. In our setting we provide subjects clear instructions about the informativeness of the information they see. This could also be understood as a warning about misinformation.

Finally, O'Rear and Radvansky (2020) consider a further reason for the continued influence effect. They use the setting from Ecker *et al.* (2010) and additionally ask participants if they believe the retraction of information. They find that most people do not believe this retraction. They further suggest that some of the wider evidence for the continued influence

[6]We do not claim that CRT questions are a good measure of cognitive ability, however, they are frequently used in this way (see e.g. Hoppe and Kusterer, 2011) and likely positively correlated.

effect may be due to people not accepting the retraction.

Economics: Belief updating biases

Now, let us turn to the literature on belief updating biases, which has recently surged in economics. Studying biases of information processing was pioneered by Phillips and Edwards (1966) as well as Tversky and Kahneman (1971, 1974). Since then, a large number of others have studied how people update their beliefs in a variety of settings. Benjamin (2019) provides a comprehensive overview of the literature. He pools the data from all existing studies into a meta analysis and formalizes a clear framework for analysing belief updating. Benjamin finds evidence that in general people under-infer from new information and that people exhibit base-rate neglect, although there is substantial variation across different studies. Moreover, he finds that in situations with sequential information signals people indeed update sequentially rather than pooling all previous information.

A number of people have explored further questions related to information processing and choice. Charness *et al.* (2021) study the choice between differently biased information sources in a laboratory experiment. They find that people make fundamental mistakes in reasoning about the relative informativeness of biased information sources. The majority of people seeks out confirmatory information. Enke *et al.* (2020) study the formation of beliefs based on memories. To do so they link information to memorable contexts. They find that beliefs are formed based on associative memories from the past. None of these papers mention situations with information uncertainty. Liang (2020) studies a settings with information uncertainty where subjects receive an uncertain signal that is either from a high or low accuracy source. He finds that people under-infer from the uncertain information. A theory of aversion to information uncertainty explains the results. Epstein and Halevy (2021) and Shishkin and Ortoleva (2021) consider a setting with information ambiguity. However, none of these papers considers a setting with additional information about previously uncertain/ambiguous information.

Political science: Fact-checking of Misinformation

Finally, our work relates to the literature on fact-checking (of potential misinformation) in political science. As in the psychology literature, the main focus lies on the correction of false information. Nieminen and Rapeli (2018) provide an overview of the literature on fact-checking of misinformation. While it has become much more popular to check factual information in the last years, results regarding the effectiveness of correcting information are mixed. Some studies find that fact-checking reduces misperceptions while others find that fact-checking is often ineffective.

Walter *et al.* (2020) pool the data from 30 existing studies on fact-checking to evaluate what methods are more or less effective at correcting people's beliefs. They find that in general people do respond to the correction of information by adjusting their beliefs. However, the ability to correct political misinformation depends to a large extent on people's prior belief. Moreover, trust in fact-checking organisations seems to be a further important

factor for correcting beliefs.

Thorson (2015) studies the persistent effects of corrected misinformation. She finds that even when a correction of previous information is fully believed, misinformation continues to affect attitudes of people. This is the case even when the correction takes place right after people observe misinformation.

Also Uscinski *et al.* (2020) study the question of why people continue to believe disproven information. They focus on Covid-19 theories and find that denialism and conspiracy thinking are among the main drivers.

Finally, Clayton *et al.* (2020) study a method to potentially reduce the influence of misinformation: ex-ante labelling misinformation as false. In the social media environment this method does reduce the influence of misinformation, however, only by a small extent.

Overall, much of the political science literature focuses on political settings with motivated reasoning. The credibility of a retraction seems to play a large role in how people respond to a retraction. We choose a neutral framework to remove doubts about the credibility of retractions and to exclude motivated reasoning as a mechanism for biased beliefs. In addition our framework allows us to study confirmations in the same way as retractions.

3 Theoretical framework

In this section, we present a theoretical model that we will use to describe a mechanism behind our findings later in the paper. The model is a simple application of the standard belief updating model of Grether (1980), and in this sense it is already well-established within the literature.

3.1 Different Types of Signals

There is a (binary) state space $\Theta = \{\text{Blue } (B),\ \text{Red } (R)\}$. There are two stochastic signals — an informative and a uninformative— both yielding data from the set $S = \{\text{blue } (b),\ \text{red } (r)\}$. Without loss of generality, because of symmetry, we restrict attention to cases where the red signal has been drawn.

- The INFORMATIVE SIGNAL π_I is symmetric and reveals the true state with probability $1 - \varepsilon > 1/2$. Thus, the likelihood ratio of r is given by

$$\frac{\pi_I(r|R)}{\pi_I(r|B)}.$$

  Throughout the rest of the paper we set $\varepsilon = 0.25$, which is the parameter we also use in our experiment, meaning that the likelihood ratio is equal to 3.

- The UNINFORMATIVE SIGNAL π_U provides no information about true state, i.e., the likelihood ratio of r is given by

$$\frac{\pi_U(r|R)}{\pi_U(r|B)} = 1.$$

In most cases, a subject observes a signal without knowing whether it is informative or uninformative. Instead, he assigns probability α to the signal being informative, thus obtaining the following initial signal:

- The INITIAL SIGNAL $\pi_A := \alpha\pi_I + (1-\alpha)\pi_U$ aggregates the informative and uninformative signal using the reliability parameter α. This means that the likelihood ratio of r is given by
$$\frac{\pi_A(r|R)}{\pi_A(r|B)} = \frac{\alpha\pi_I(r|R) + (1-\alpha)\pi_I(r|B)}{\pi_I(r|B)}.$$
Thus, given the parameter $\varepsilon = 0.25$ that we have already set, the likelihood ratio of the red signal is $(2+\alpha)/(2-\alpha)$. In our experiment we exogenously fix $\alpha = 0.4$, meaning that the likelihood ratio becomes $3/2$. Throughout the paper, we often refer to it as the regular signal.

In some cases, the agent receives information in the form of a verification/check of a previous signal realization. The verification signal is denoted by π_V, and yields one of two possible outcomes:

- A RETRACTION reveals to the subject that the previously observed signal realization was in fact uninformative. A retraction is denoted by ✗. That is, the likelihood ratio of a retraction is given by the inverse of the likelihood ratio of the previously observed initial signal
$$\frac{\pi_V(\text{✗}|R)}{\pi_V(\text{✗}|B)} = \frac{\pi_A(r|B)}{\pi_A(r|R)}. \tag{1}$$
In other words, informationally, a retraction is equivalent to observing a blue initial signal. So, in our experiment the likelihood ratio of retracting a red signal is $2/3$.

- A CONFIRMATION reveals to the subject that the previously observed signal realization was in fact informative. A confirmation is denoted by ✓. That is, the likelihood ratio of a confirmation is given by the inverse of the likelihood ratio of the previously observed red signal multiplied by the likelihood ratio of the informative signal
$$\frac{\pi_V(\text{✓}|R)}{\pi_V(\text{✓}|B)} = \frac{\pi_A(r|B)}{\pi_A(r|R)}\frac{\pi_I(r|R)}{\pi_I(r|B)}. \tag{2}$$
This means that informationally, a confirmation is equivalent to the difference between an informative and an initial signal. So, in our experiment the likelihood ratio of confirming a red signal is equal to 2.

3.2 Belief updating

The agent starts with a prior belief p_0, and receives information in the form of a sequence of signals $\{\pi_1, \pi_2, \ldots, \pi_T\}$. Some of these signals are regular signals, while others are verifications, i.e., for each t we have $\pi_t \in \{\pi_A, \pi_V\}$. By definition, a verification can only follow an initial signal, i.e., if $\pi_t = \pi_V$, then $\pi_{t-1} = \pi_A$.

At each period $t \in \{1, \ldots, T\}$, the agent treats the belief p_{t-1} that he has carried from the previous period as his working prior, and updates his belief to p_t using the realization of signal π_t. Following the standard model of Grether (1980), upon observing a signal realization $s \in \{r, b\}$ (in case of an initial signal) or $s \in \{\text{✗}, \text{✓}\}$ (in case of a verification), the agent is assumed to update according to the formula

$$p_t(R|s) = \frac{p_{t-1}(R)^{c_t} \pi_t(s|R)^{d_t}}{p_{t-1}(R)^{c_t} \pi_t(s|R)^{d_t} + p_{t-1}(B)^{c_t} \pi_t(s|B)^{d_t}}, \tag{3}$$

where $c_t, d_t > 0$ are individual parameters of belief updating biases. In particular, c_t measures the bias in using the priors (i.e. base-rate use), while d_t measures the bias in using the likelihoods (i.e. inference). The rational benchmark of Bayesian updating is a special case where $c_t = d_t = 1$. For an extensive overview of this literature, see Benjamin (2019).

Remark 1 While our model is flexible enough to allow biases to vary with respect to the full history, throughout most of the paper we will only let the updating parameters depend on the type of the signal, i.e., for initial signals we will have c and d, for retractions we will have $c_{\text{✗}}$ and $d_{\text{✗}}$, and for verifications we will have $c_{\text{✓}}$ and $d_{\text{✓}}$. Later on, in Section 5.4, we condition the updating parameters for each signal type on history characteristics.

3.2.1 Belief updating given initial signals

As commonly done in the literature, we reformulate Grether's updating formula in terms of ratios, i.e., for a regular signal $\pi_t = \pi_A$, we obtain

$$\frac{p_t(R|r)}{p_t(B|r)} = \left[\frac{p_{t-1}(R)}{p_{t-1}(B)}\right]^c \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^d. \tag{4}$$

This means that the posterior odds are given by multiplying the prior odds (raised to the power of c) with the likelihood ratio (raised to the power of d). Given that the likelihood ratio is larger than 1, the agent should always update in the direction of the observed signal, regardless of the values of the updating parameters.

3.2.2 Belief updating given retractions

As we have already mentioned, retractions are intrinsically linked to the initial signal realization. This means that the belief p_t that has been already formed using Equation (4) is now treated as the new prior, while the retraction signal of Equation (1) is treated as a new signal. Thus, the retraction leads to updating according to the formula

$$\begin{aligned} \frac{p_{t+1}(R|r,\text{✗})}{p_{t+1}(B|r,\text{✗})} &= \left[\frac{p_t(R)}{p_t(B)}\right]^{c_{\text{✗}}} \left[\frac{\pi_V(\text{✗}|R)}{\pi_V(\text{✗}|B)}\right]^{d_{\text{✗}}} \\ &= \left[\frac{p_{t-1}(R)}{p_{t-1}(B)}\right]^{c \cdot c_{\text{✗}}} \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{d \cdot c_{\text{✗}} - d_{\text{✗}}}. \end{aligned} \tag{5}$$

The key observation here is that the updating biases of the two signals (viz., the initial signal and the retraction) interact and in someway amplify each other.

Then, we classify the different ways in which retractions can be treated:

(R_0) The agent *processes the retraction rationally* whenever the following holds:

$$\frac{p_{t-1}(R)}{p_{t-1}(B)} = \frac{p_{t+1}(R|r,\text{✗})}{p_{t+1}(B|r,\text{✗})}. \tag{6}$$

That is, she processes retractions as if she fully discarded the initial signal, similarly to what a Bayesian agent would have done.

(R_1) The agent *underreacts to the retraction* whenever the following holds:

$$\frac{p_{t-1}(R)}{p_{t-1}(B)} < \frac{p_{t+1}(R|r,\text{✗})}{p_{t+1}(B|r,\text{✗})} < \frac{p_t(R|r)}{p_t(B|r)}. \tag{7}$$

In the literature, underreaction to retractions is typically called the continued influence effect (Goncalves *et al.*, 2026, and references therein). This is what we typically expect to see, and the idea is that the initial updating partially persists even after the signal has been retracted.

(R_2) The agent *overreacts to the retraction* whenever the following holds:

$$\frac{p_{t+1}(R|r,\text{✗})}{p_{t+1}(B|r,\text{✗})} < \frac{p_{t-1}(R)}{p_{t-1}(B)} < \frac{p_t(R|r)}{p_t(B|r)}. \tag{8}$$

This is not commonly observed in the literature, although one can think of reasonable situations where this might happen.

Let us characterize how the agent processes the retraction. First, define the threshold

$$t_{\text{✗}} := \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{\frac{d\cdot c_{\text{✗}} - d_{\text{✗}}}{1-c\cdot c_{\text{✗}}}}.$$

noticing that $t_{\text{✗}} > 0$. Subsequently, define the probability threshold

$$p_{\text{✗}} := \frac{t_{\text{✗}}}{1+t_{\text{✗}}}, \tag{9}$$

noticing that the agent processes the retraction rationally if and only if $p_{t-1}(R) = p_{\text{✗}}$. This is because $t_{\text{✗}}$ is the unique solution to Equation (6).

For the remainder of our characterization, we distinguish two main cases: First, if the agent underuses his prior beliefs, we expect to see underreaction to retractions for small priors, and overreaction to retractions for large priors. If on the other hand, the agent overuses his prior on aggregate, we expect to see overreaction to retractions for small priors, and underreaction to retractions for large priors. These results are the basis for our first hypothesis. They are formalized in Proposition 1 and depicted in Figure 2.

Proposition 1 *1) If $c \cdot c_{\text{✗}} < 1$, the following hold:*

(1.i) The agent processes the retraction of r at $t+1$ correctly if $p_{t-1}(R) = p_{\text{✗}}$.

(1.ii) The agent underreacts to the retraction of r at $t+1$ if $p_{t-1}(R) < p_{\text{✗}}$.

(1.iii) The agent overreacts to the retraction of r at $t+1$ if $p_{t-1}(R) > p_{✗}$.

2) If $c \cdot c_{✗} > 1$, the following hold:

(2.i) The agent processes the retraction of r at $t+1$ correctly if $p_{t-1}(R) = p_{✗}$.

(2.ii) The agent underreacts to the retraction of r at $t+1$ if $p_{t-1}(R) > p_{✗}$.

(2.iii) The agent overreacts to the retraction of r at $t+1$ if $p_{t-1}(R) < p_{✗}$.

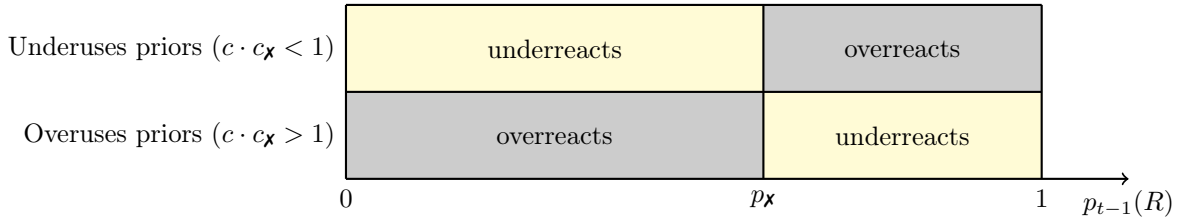


Figure 2: Theoretical predictions of how the agent reacts to retractions.

Let us explain the intuition. First of all, the general insight is that irrational reactions to retractions are caused by the fact that two different updating biases (a lá Grether) interact with one another in a "non-linear way" thus not cancelling each other.

Let us go a bit deeper into how the retraction bias depends on the prior belief. Whenever the prior is small, the initial signal is unexpected, whereas the subsequent retraction is expected (both in relation to the prior). Thus, because of base-rate neglect, the agent puts much weight on the signal compared to the prior, and therefore reacts strongly to the surprise. The reverse takes place when the agent exhibits base-rate overuse. The reasoning is analogous for large prior beliefs.

These theoretical predictions present a more nuanced picture compared to most of the existing literature (Goncalves *et al.*, 2026, and references therein), i.e., Proposition 1 suggests that reactions to retractions are largely driven by the prior beliefs. This means that it may very well be the case that —depending on the updating parameters— overreaction to retractions are only observed if we have relatively high priors. In most of the existing literature, and particularly in Goncalves *et al.* (2026), beliefs remain concentrated relatively close to the uniform belief, which could potentially explain why overreactions have not been extensively documented.

3.2.3 Belief updating given confirmations

Analogously to retractions, confirmation leads to updating according to the formula

$$\begin{aligned} \frac{p_{t+1}(R|r,✓)}{p_{t+1}(B|r,✓)} &= \left[\frac{p_t(R)}{p_t(B)}\right]^{c_✓} \left[\frac{\pi_V(✓|R)}{\pi_V(✓|B)}\right]^{d_✓} \\ &= \left[\frac{p_{t-1}(R)}{p_{t-1}(B)}\right]^{c\cdot c_✓} \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{d\cdot c_✓ - d_✓} \left[\frac{\pi_I(r|R)}{\pi_I(r|B)}\right]^{d_✓}. \end{aligned} \tag{10}$$

Then, we can classify the different ways in which confirmations are treated:

(C_0) An agent *processes the confirmation rationally* whenever

$$\frac{p_{t+1}(R|r,\checkmark)}{p_{t+1}(B|r,\checkmark)} = \frac{p_{t-1}(R)}{p_{t-1}(B)}\frac{\pi_I(r|R)}{\pi_I(r|B)}, \tag{11}$$

i.e., she updates as if she were matching the belief of a Bayesian agent who incorporates the informative signal to p_{t-1}.

(C_1) An agent *underreacts to the confirmation* whenever the following holds:

$$\frac{p_{t-1}(R)}{p_{t-1}(B)} < \frac{p_t(R|r)}{p_t(B|r)} < \frac{p_{t+1}(R|r,\checkmark)}{p_{t+1}(B|r,\checkmark)} < \frac{p_{t-1}(R)}{p_{t-1}(B)}\frac{\pi_I(r|R)}{\pi_I(r|B)}. \tag{12}$$

Intuitively, the updating given the initial signal and then given the confirmation will lead to a belief which is short of the Bayesian update.

(C_2) An agent *overreacts to the confirmation* whenever the following holds:

$$\frac{p_{t-1}(R)}{p_{t-1}(B)} < \frac{p_t(R|r)}{p_t(B|r)} < \frac{p_{t-1}(R)}{p_{t-1}(B)}\frac{\pi_I(r|R)}{\pi_I(r|B)} < \frac{p_{t+1}(R|r,\checkmark)}{p_{t+1}(B|r,\checkmark)}. \tag{13}$$

Intuitively, the updating given the initial signal and then given the confirmation will lead to a belief which goes even beyond the Bayesian update.

Similarly to our analysis of the retraction, we characterize how the agent processes the confirmation. First, define the threshold

$$t_\checkmark := \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{\frac{d\cdot c_\checkmark - d_\checkmark}{1-c\cdot c_\checkmark}} \left[\frac{\pi_I(r|R)}{\pi_I(r|B)}\right]^{\frac{d_\checkmark - 1}{1-c\cdot c_\checkmark}}.$$

noticing that $t_\checkmark > 0$. Then again, we define the probability threshold

$$p_\checkmark := \frac{t_\checkmark}{1+t_\checkmark}. \tag{14}$$

Once again, we distinguish two main cases: First, if the agent underuses his prior beliefs, we expect to see overreaction to confirmations for small priors, and underreaction to confirmations for large priors. If on the other hand, the agent overuses his prior on aggregate, we expect to see underreaction to confirmations for small priors, and overreaction to confirmations for large priors. These results are formalized in Proposition 2 and depicted in Figure 3.

Proposition 2 *1) If $c \cdot c_\checkmark < 1$, the following hold:*

(1.i) The agent processes the confirmation of r at $t+1$ correctly if $p_{t-1}(R) = p_\checkmark$.

(1.ii) The agent overreacts to the confirmation of r at $t+1$ if $p_{t-1}(R) < p_\checkmark$.

(1.iii) The agent underreacts to the confirmation of r at $t+1$ if $p_{t-1}(R) > p_\checkmark$.

2) If $c \cdot c_\checkmark > 1$, the following hold:

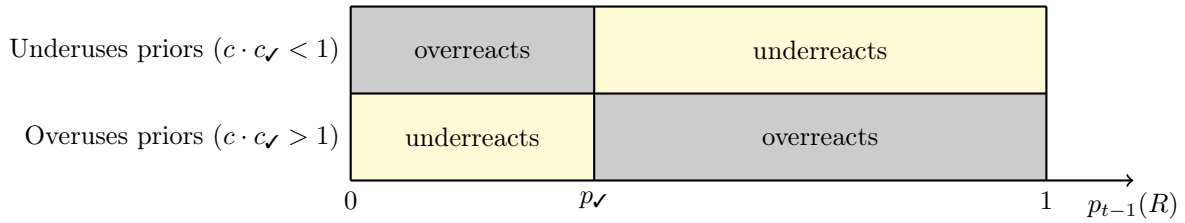


Figure 3: Theoretical predictions of how the agent reacts to confirmations.

(2.i) The agent processes the confirmation of r at $t+1$ correctly if $p_{t-1}(R) = p_{\checkmark}$.

(2.ii) The agent overreacts to the confirmation of r at $t+1$ if $p_{t-1}(R) > p_{\checkmark}$.

(2.iii) The agent underreacts to the confirmation of r at $t+1$ if $p_{t-1}(R) < p_{\checkmark}$.

The intuition is similar to the one of retractions, i.e., the fact that updating biases distort the posterior beliefs twice —once for the initial signal and once for the confirmation— amplifies the overall effect in a way that leads to underreaction or overreaction depending on the prior beliefs. In this sense, once again, the biases due to confirmations can be seen as a special case of Grether's standard model, albeit with different updating parameters.

Let us explain the intuition in a bit more detail. Whenever the prior is small, the initial signal is unexpected, and so is the subsequent confirmation. Thus, because of base-rate neglect, the agent puts much weight on the signal relative to the prior, which induces a strong reaction to surprise. And again, the converse occurs when there is base-rate overuse. The reasoning is analogous for large prior beliefs.

4 Experiment

4.1 Baseline experimental design

In our baseline experiment we modify the novel design of Goncalves *et al.* (2026) in two ways that allow us to address the two main questions of this paper, while remaining in the standard neutral setting of balls and urns. First, we allow for confirmation in a non-trivial way, i.e., a confirmation does not reveal the state. Second, we extend the horizon from 4 to 12 periods, out of which 9 are regular signals and 3 are verifications. Let us describe our design in detail.

In the beginning of the experiment there are two urns: a Red urn containing one blue and three red balls, and a Blue urn containing one red and three blue balls. One of the two urns (Red or Blue) is randomly drawn with equal probability. The 4 balls from the chosen urn are labelled "informative", and are placed inside a black box together with 6 more balls which are labelled "uninformative". The uninformative balls are distributed equally between the two colors, i.e., 3 are blue and 3 are red. This means that only four out of the ten balls in the black box can provide useful information about the color of the urn, i.e., we have $\alpha = 0.4$. Figure 4 illustrates how the black box was composed: a similar figure was also shown to the participants of the study.

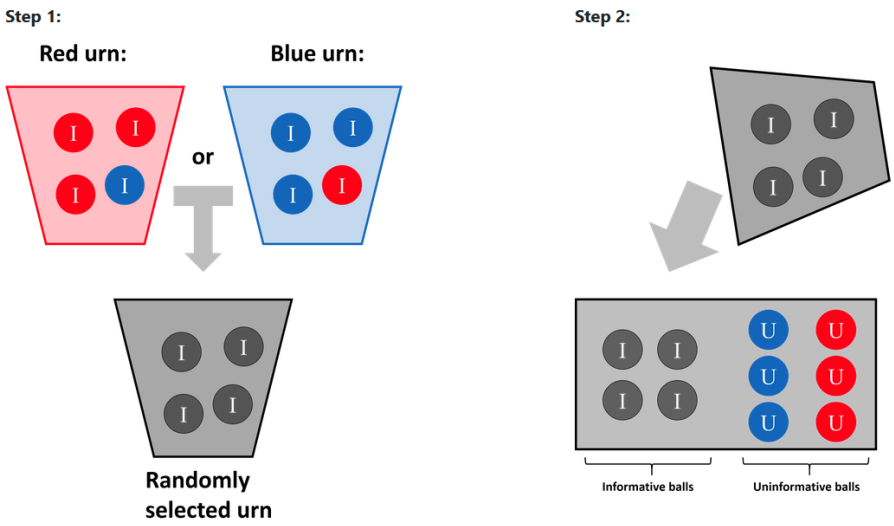


Figure 4: Experimental design with informative balls and uninformative balls

Each subject receives sequentially 12 hints about the color of the selected urn in the form of information about the balls that are drawn (with replacement) from the black box. After each hint, the subject is asked to report her subjective probability of the red urn having been drawn. Out of the 12 hints, 9 correspond to regular signals (i.e., the color of a newly drawn ball is revealed to the subject), and 3 correspond to verifications (i.e., it is revealed to the subject whether the previously seen ball was informative or uninformative). The periods in which verifications occur vary across subjects, albeit in a specific way. In particular, for each subject we randomly pick a number between 2 and 6, and we verify three consecutive regular signals, starting from this number, i.e., formally, we pick some $\tau \in \{2, \ldots, 6\}$, then verifications take place in periods $\{\tau+1, \tau+3, \tau+5\}$. This purposefully excludes verifications in the first and last round.

We conducted three different treatments which were identical in every aspect, except for the way in which we presented information about previous periods to the subjects, in order to control for the effect of memory and anchoring. Specifically, in the first treatment we kept showing to the subjects the entire history of signals that had been previously realized and their last reported belief; in the second treatment we removed the history of signals and maintained the last belief; in the third treatment we kept the history of signals and removed the last belief. The subjects numbers were quite balanced across treatments.

4.2 Empirical strategy

Our main hypotheses with respect to how retractions/confirmations are treated follow directly from the two propositions above. In order to formulate them, we first need to estimate the updating parameters in the Grether regression for all three types of signals. The structural

equations that we estimate look as follows:

$$\text{Initial signal:} \qquad \log \frac{p(R|r)}{p(B|r)} = c \cdot \log \frac{p(R)}{p(B)} + d \cdot \log \frac{3}{2} + \eta \tag{15}$$

$$\text{Retraction:} \qquad \log \frac{p(R|r,\text{✗})}{p(B|r,\text{✗})} = c_{\text{✗}} \cdot \log \frac{p(R|r)}{p(B|r)} + d_{\text{✗}} \cdot \log \frac{2}{3} + \eta_{\text{✗}} \tag{16}$$

$$\text{Confirmation:} \qquad \log \frac{p(R|r,\text{✓})}{p(B|r,\text{✓})} = c_{\text{✓}} \cdot \log \frac{p(R|r)}{p(B|r)} + d_{\text{✓}} \cdot \log 2 + \eta_{\text{✓}} \tag{17}$$

Before proceeding let us make a few comments. Each of these equations has been obtained by rewriting Equation (3) using ratios and then taking logarithms, as usual. The reason we have not included constants is that within each of these three regressions the signal remains fixed, and therefore the variation that would normally be explained by the constant is now picked by the log-likelihood ratio.

4.3 Further design details

The experiment was programmed using oTree (Chen *et al.*, 2016).[7] The subjects were recruited from Academic Prolific. We pre-registered outlier criteria and analysis.[8] The full set of instructions can be found in Appendix C. All belief reports were incentivized using a quadratic scoring rule. Following Danz *et al.* (2020), we hide the quantitative details of the scoring rules behind an additional button and restrict the main text to a qualitative explanation of the mechanism with an emphasis on its incentive compatibility.[9] Subjects received a completion fee of €2.50 as well as a bonus payment up to €3.00 from the scoring rule mechanism. To enter the main part of the experiment subjects had to answer 5 out of 6 instruction comprehension test questions correctly. In the end of the experiment we attached a post-experiment survey with questions about their strategy and regular demographic questions.

4.4 Implementation

The baseline experiment was conducted in January of 2022. In total 606 subjects completed the experiment, from which 66 were removed based on pre-defined criteria, e.g., subjects not reacting to any of the information on screen, completing the survey faster than possible (when seriously reading), or giving primarily dominated responses (i.e., updating in the wrong direction in the majority of cases). The median time to complete the experiment was 17 minutes and subjects received on average €4.75. Each subjects reported 12 different

[7] The oTree code can be found here: https://github.com/lwittrock/belief_updating

[8] The pre-registration can be found here https://aspredicted.org/Z7N_P7W.

[9] This approach is extensively used in recent years in the literature, albeit typically with a binarized version of the quadratic scoring rule. The fact that the payment details are not shown to the subjects by default makes us confident that the actual payment mechanism does not play a major role.

beliefs, leading to a total of 6480 observations used in our analysis. All data files are available online.[10]

5 Main results

5.1 Preliminary analysis and main hypotheses

Subjects reacted to red and blue signals in a symmetric way implying that the color of the ball had no influence on the subjects' behavior. This suggests that in general subjects understood the setup well, and allows us to analyze together the observations that involve blue and red balls. This also allows us to transform our data, so that all beliefs are expressed in the same interval $[0, 1]$. In particular, whenever a blue signal is observed and the subject decreases the probability of R from 0.3 to 0.2, we will encode this as a red signal having been observed and the probability of R having increased from 0.7 to 0.8. Thus, in whatever follows, we will always refer to probabilities attached to the Red urn, and we will always consider the effect of a red ball having been drawn.

There were no significant differences across the three treatments, and therefore we pool the data together for our analysis. For a detailed analysis of the treatment effects, see Section 7.1.

We begin with the Grether regressions. The estimates that we obtain are summarized in Table 1. As seen, for all three types of signals there is significant base-rate neglect, i.e., $\hat{c}, \hat{c}_{\text{✗}}, \hat{c}_{\text{✓}} < 1$, similarly to what most of the literature finds (Benjamin, 2019). At the same time, we also find overinference for all three types of signals, i.e., $\hat{d}, \hat{d}_{\text{✗}}, \hat{d}_{\text{✓}} > 1$. This is in contrast to many papers in the literature. However, it is consistent with the documented evidence, according to which people tend to overinfer from weak signals Augenblick *et al.* (2025).

Based on these estimates, we obtain the estimated thresholds $\hat{p}_{\text{✗}} = 0.33$ and $\hat{p}_{\text{✓}} = 0.63$ which, together with the fact that $\hat{c} \cdot \hat{c}_{\text{✗}} < 1$ and $\hat{c} \cdot \hat{c}_{\text{✓}} < 1$, allows us to form the following two main hypotheses:

($H_1^{\text{✗}}$) Subjects overreact (resp., underreact) to retractions of a signal at period $t+1$ whenever the belief of the same color was lower (resp., higher) than 0.33 at $t-1$.

($H_1^{\text{✓}}$) Subjects underreact (resp., overreact) to confirmations of a signal at period $t+1$ whenever the belief of the same color was lower (resp., higher) than 0.63 at $t-1$.

It is important to note that we have estimated $\hat{p}_{\text{✗}}$ and $\hat{p}_{\text{✓}}$ on population level. It would have been of course more appropriate to obtain these thresholds for each individual. However, we do not have enough power to do so, and therefore we can only rely on population estimates.

[10] See https://github.com/lwittrock/UpdatingMisinformation-Analysis.

	Dependent variable:		
	Observed Log-Posterior-Ratio		
	Initial	Retractions	Confirmations
	(15)	(16)	(17)
Prior	0.697***	0.722***	0.575***
	(0.051)	(0.064)	(0.098)
Signal	1.530***	0.806***	1.453***
	(0.027)	(0.187)	(0.151)
Observations	6777	985	635
Adjusted R^2	0.391	0.438	0.406
Note:	*p<0.1; **p<0.05; ***p<0.01		
	SEs clustered by subject.		

Table 1: Estimates of the Grether regression for the three types of signals.

5.2 Retractions

As we have already discussed, using the estimated parameters, we identify the threshold of prior beliefs where we expect the subjects to switch from underreacting to overreacting to retractions. In fact, we can even predict the magnitude of the retraction bias

$$b_{t+1}(R|r,\text{✗}) := p_{t+1}(R|r,\text{✗}) - p_{t-1}(R), \tag{18}$$

as a function of the prior belief. The predicted bias is denoted by $\hat{b}_{t+1}(R|r,\text{✗})$, and graphically corresponds to the red curve in Figure 5. On the other hand, the grey bars show the actual retraction bias that we see in the data.

Obviously, we can conclude that our directional hypothesis is corroborated, i.e., subjects underreact to retractions for small priors and overreact to retractions for large priors. This means that the continued influence effect does not appear to be a universal phenomenon, as previously suggested in the literature, but rather it holds only for small priors. In simple terms —as we have already mentioned before— this is because base-rate neglect is much more influential for small priors. Hence, there is large reaction towards R when $p_{t-1}(R)$ is small.

But we can actually say something more than just corroborate the directional hypothesis. The model seems to predict quite accurately the magnitude of the retraction bias, despite being so stylized. This is again seen in Figure 5.

One interesting observation that we make is that the retraction bias can also be predicted by the reaction to the initial signal. In particular, denote the initial bias by

$$b_t(R|r) := p_t(R|r) - p_t^{\text{Bayes}}(R|r), \tag{19}$$

where $p_t^{\text{Bayes}}(R|r)$ is the Bayesian at period t update given the red initial signal. Then, it is the case that overreaction (resp., underreaction) to the initial signal is correlated

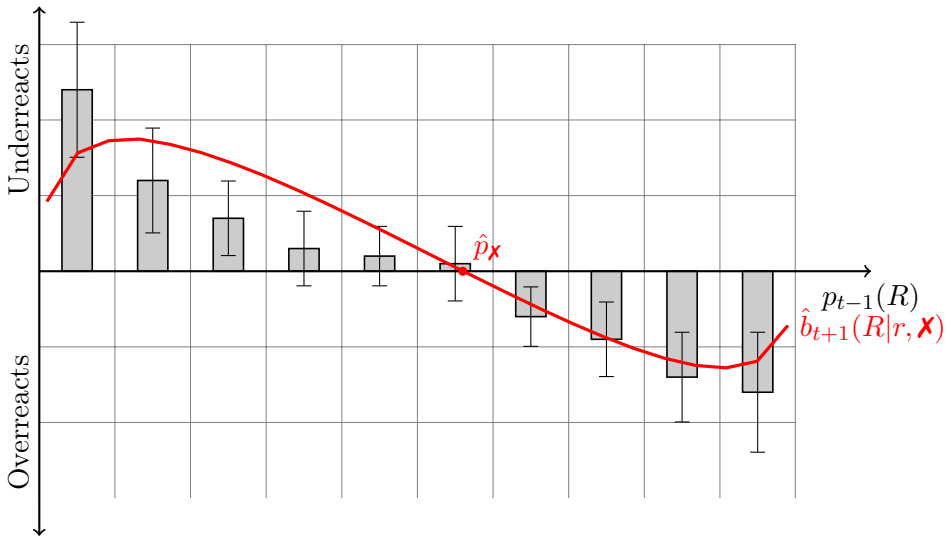


Figure 5: Bias when processing retractions: The red line corresponds to the predicted retraction bias $\hat{b}_{t+1}(R|r,\text{✗})$ according to our model, whereas the grey bars correspond to the actual retraction bias $b_{t+1}(R|r,\text{✗})$ that we obtain from the data. Each step in the grid is equal to 0.1 (both vertically and horizontally).

with overreaction (resp., underreaction) to the subsequent retraction. This correlation is illustrated in Figure B1.

However, this does not mean that there is a causal effect, as overreactions/underreactions in both of them seem to be driven by the prior. In fact, our model predicts that the reaction to the initial signal as a function of the prior follows a similar pattern as the reaction to the retraction that we have already established. This is illustrated in Figure B2. In this sense, the correlation on Figure B1 should probably be attributed to both biases being driven by the prior belief.

5.3 Confirmations

Our analysis of confirmations is analogous to the one of reactions, viz., using the estimated parameters, we start by pinning down the threshold of prior beliefs where we expect the subjects to switch from overreacting to underreacting to confirmations.

The bias in the reaction to confirmation is measured by

$$b_{t+1}(R|r,\text{✓}) := p_{t+1}(R|r,\text{✓}) - p^{\text{Bayes}}_{t+1}(R|r,\text{✓}), \tag{20}$$

as a function of the prior belief. Once again, the predicted bias is denoted by $\hat{b}_{t+1}(R|r,\text{✓})$, and graphically corresponds to the red curve in Figure 6. On the other hand, the grey bars show the observed bias that we see in the data.

Note that the directional effect that we hypothesize in $(H_1^{\text{✓}})$ is corroborated, i.e., subjects overreact to confirmations for small priors and underreact to confirmations for larger priors. This also suggests that the mechanism that drives the updating in response to retractions is similar to the corresponding mechanism for confirmations.

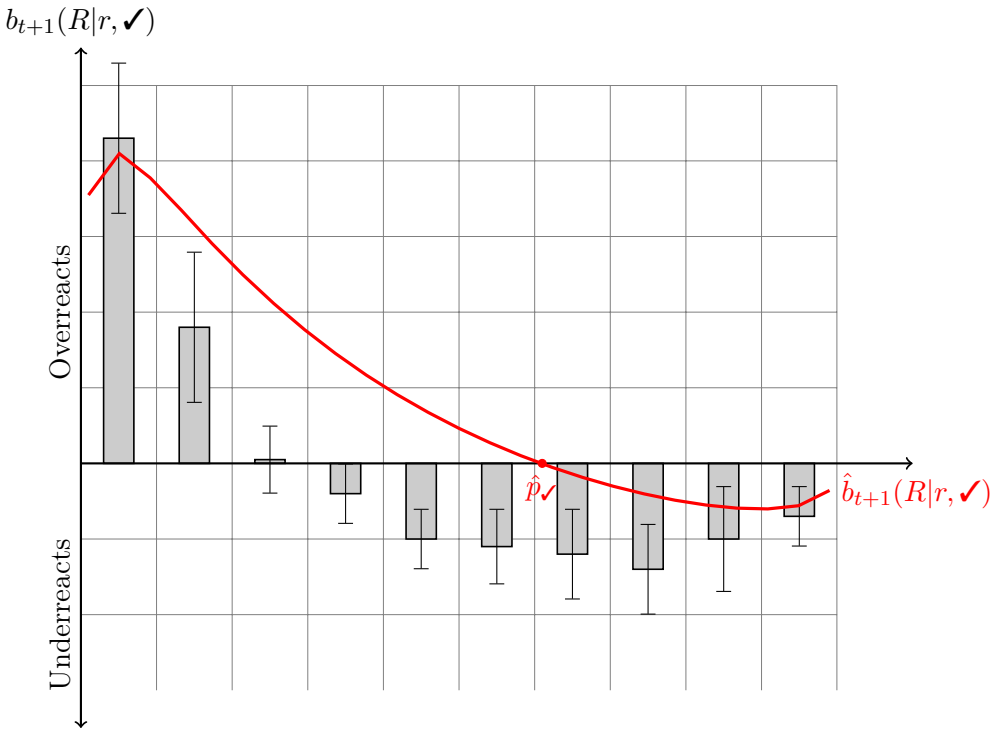


Figure 6: Bias when processing confirmations: The red line corresponds to the predicted bias $\hat{b}_{t+1}(R|r,\checkmark)$ according to our model, whereas the grey bars correspond to the actual bias $b_{t+1}(R|r,\checkmark)$ that we obtain from the data. Each step in the grid is equal to 0.1 (both vertically and horizontally).

In particular, once again, our explanation relies on how subjects react to surprises. Whenever the prior belief is low, both the initial signal and the subsequent confirmation are unexpected (in relation to the prior), and the subjects overreact to this type of unexpected news. Analogously, whenever the prior belief is high, both the initial signal and the subsequent confirmation are expected, and subjects underreact to these pieces of information.

Of course, our prediction of the magnitude of the bias is not accurate for every block of priors. However, this is neither very surprising nor worrisome. Recall that ours is a very stylized model with only a handful of parameters. If we had estimated the updating parameters individually, we would have in all likelihood obtained a better fit of the data, but this is not possible (due to limited power). Thus, we are satisfied with corroborating our directional hypothesis, which at the end of the day is what our hypothesis states.

5.4 Dynamics

The second set of questions in our paper focuses on the dynamics of information verifications. Specifically, first we naturally ask whether the effects that we established in the previous section persist over time, i.e., is it the case that *at all points*, subjects underreact to retractions and overreact to confirmations for small priors, and and vice versa for large priors? Second, we ask whether verifications have a downstream effect on regular updating,

i.e., does the experience of retractions and/or confirmations affect the way subjects process information? As we will see, the two subquestions are related to each other.

In order to formulate the two hypotheses that correspond to the first subquestion, we first allow the updating parameters in Grether's model to depend on the period, as postulated in Equation (3). In order to maintain enough power, we split the periods into 3 blocks: early periods $T_1 := \{3, 4, 5\}$, middle periods $T_2 := \{6, 7, 8\}$, late periods $T_3 := \{9, 10, 11\}$. Then, the estimated updating parameters are the one shown in Table B1. Notice that the priors are measured at period $t - 1$, meaning that initial signals are estimated using the blocks $T_1 = \{2, 3, 4\}$, $T_2 = \{5, 6, 7\}$, $T_3 = \{8, 9, 10\}$.

One crucial observation is that our overall measures of base-rate use improve over time, although they persists significantly below 1, i.e., $c^{T_1} \cdot c_{✗}^{T_1} < c^{T_2} \cdot c_{✗}^{T_2} < c^{T_3} \cdot c_{✗}^{T_3} < 1$ and $c^{T_1} \cdot c_{✓}^{T_1} < c^{T_2} \cdot c_{✓}^{T_2} < c^{T_3} \cdot c_{✓}^{T_3} < 1$. This means that we predict that biases remain qualitatively the same over time, albeit smaller in magnitude. The respective thresholds will be $p_{✗}^{T_1} = 0.48$ and $p_{✓}^{T_1} = 0.43$ for the early periods; $p_{✗}^{T_2} = 0.53$ and $p_{✓}^{T_2} = 0.56$ for the middle periods; $p_{✗}^{T_3} = 0.71$ and $p_{✓}^{T_3} = 0.80$ for the early periods. So, we can formalize the following hypotheses:

($H_2^{✗}$) In each of the three blocks of periods $T \in \{T_1, T_2, T_3\}$ subjects overreact (resp., underreact) to retractions of a signal at period $t + 1$ whenever the belief of the same color was lower (resp., higher) than $p_{✗}^{T}$ at $t - 1$.

($H_2^{✓}$) In each of the three blocks of periods $T \in \{T_1, T_2, T_3\}$ subjects underreact (resp., overreact) to confirmations of a signal at period $t + 1$ whenever the belief of the same color was lower (resp., higher) than $p_{✓}^{T}$ at $t - 1$.

Before going to the results, notice that our theoretical predictions suggest that in later periods, we will have underreaction to retractions more often, i.e., there is a larger range of priors that lead to underreaction. At the same time, we predict that biases —both underreactions and overreactions— will become smaller in magnitude over time (at least for the priors where the bias goes into the same direction).

Now, looking at the data, which are all summarized in Figure 4(a), we observe that our main hypothesis regarding the directional effect is corroborated, i.e., for all blocks small prior beliefs lead to underreaction to retractions, whereas large prior beliefs lead to overreaction. The magnitude seems to get smaller over time, although the differences are not significant. That is, overreactions to retractions seem to gradually vanish. It is plausible that with sufficiently long horizon, underreaction persists (for small priors) and overreaction is eliminated (also for larger priors). This would be consistent with the findings of Goncalves *et al.* (2026). However, with our experiment, we cannot conclusively make this claim.

Now, let us switch to hypothesis ($H_2^{✓}$), which is the mirror image of ($H_2^{✗}$) for confirmations. Once again the main conclusion is that the directional effect that we previously found persists regardless of the period, i..e, subjects overreact to confirmations when the priors are small and underreact to confirmations when the priors are large. In terms of magnitude, our theoretical predictions seem to overestimate the actual reactions, especially when the priors are large. However, similarly to our overall analysis, we are less concerned with the magnitudes, especially given how stylized the model is.

Of course, both for retractions and confirmations, by design it is the case that have relatively few observations in the early periods, more observations in the late periods, and most observations in the middle periods. This is because the first regular signal to be verified is randomly drawn between the second and the sixth (regular signals) and then the two consecutive regular signals are also verified. This also means that in the early periods at most one signal is verified; in middle periods there is quite some variation about how many verifications they have previously seen; while in late periods it is usually the second or theird verification that they see. As a result, the effect of time is confounded by the number of previously observed verifications. We come back to this issue later in this section.

For now, let us switch to our second subquestion, and look at the effect that different types of verifications have in downstream processing of information. To do this, we distinguish the different types of verifications they may have observed. In particular, we look at how subjects update in each of the following situations:

($H_{0,0}$) They have not experienced any retraction or confirmation so far.

($H_{1,0}$) They have experienced 1 retraction and 0 confirmations.

($H_{0,1}$) They have experienced 0 retractions and 1 confirmation.

($H_{2,0}$) They have experienced 2 retractions and 0 confirmations.

($H_{0,2}$) They have experienced 0 retractions and 2 confirmations.

($H_{1,1}$) They have experienced 1 retraction and 1 confirmation.

We estimate the updating parameters for regular signals in each of these situations. The results are shown in Table B2. There are some small differences in the estimates, but overall the picture looks quite similar across the board. One thing for instance that we notice is that base-rate use increases significantly after having observed multiple confirmations. On the other hand retractions seem to increase base-rate neglect, but it also improves overinference.

The overall effects on predicted biases are then shown in Figure B.2, suggesting that we do not expect to see major changes in our data. Indeed, as shown in Figure B6, subjects overreact to information when the priors are small and underreact when the priors are large. The only difference is when 2 confirmations are observed updating is statistically indistinguishable from Bayesian for a larger range of priors. However, we should be careful at interpreting this result, given that the confidence intervals are large due to small sample size.

Finally, going back to our earlier discussion regarding the effect of past observations of verifications on processing retractions and confirmations, we also look into the respective biases in each on the aforementioned six situations. These are depicted in Figures B7 and B8. What we see is that the overall results look nearly identical to the ones in ($H_{0,0}$), both for retractions and confirmations. Furthermore, the way retractions are treated in ($H_{1,0}$) and ($H_{0,1}$) are qualitatively the same as in ($H_{0,0}$), although the bias is smaller in magnitude conditional on a verification having been already observed. Of course, there are also larger errors, which is justified by the fact that there are fewer observations in each of

$(H_{1,0})$ and $(H_{0,1})$, compared to $(H_{0,0})$. Similar conclusions can be drawn, when we compare $(H_{1,0})$ with $(H_{2,0})$ or $(H_{1,1})$, or when we compare $(H_{0,1})$ with $(H_{0,2})$ or $(H_{1,1})$, in terms of how retractions or confirmations are processed. In each of these comparisons, we see similar directional effects, but also increasing errors as we add verifications. Nevertheless, the sample size decrease as we condition with more verifications, and therefore we refrain from making conclusive statements about these last comparisons.

6 Retraction vs opposite regular signal

One of the main messages in Goncalves *et al.* (2026) is that retractions are inherently different to regular signals. They test this hypothesis by comparing reactions to retraction with reactions to observing a new signal of the opposite color.

In our model the latter would correspond to updating according to the Equation (6) with $c_{\text{✗}} = c$ and $d_{\text{✗}} = d$, i.e., the updating parameter of the opposite color are set equal to the updating parameters of the initial signal, as the opposite color is a regular signal itself. Using the parameters that we have estimated, our theoretical model predicts stronger updating in the opposite direction compared to the retractions. Notice that we expect this to be the case regardless of the prior, and a fortiori both when there is underreaction and overreaction. This also suggests that the range where there will be underreaction to the opposite colored signal is smaller than the range where there will be underreaction to retractions. And analogously, the range where there will be overreaction to the opposite colored signal is larger than the range where there will be overreaction to retractions.

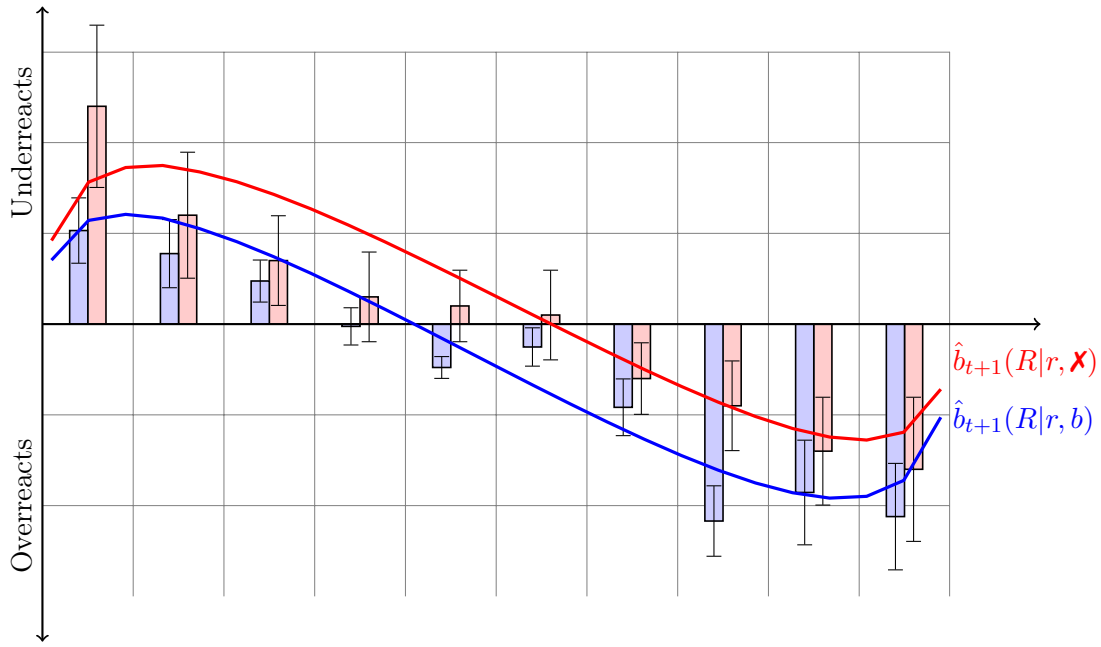


Figure 7: Bias when processing signal of opposite color.

Our results corroborate all these predictions. This is seen in Figure 7. This finding is consistent with the idea that subjects react less to retractions than to opposite signals. In

this sense, we confirm the finding of Goncalves *et al.* (2026) that retractions are inherently different than regular signals, and more specifically with their interpretation that retractions are more difficult to process because they are more complex.

Importantly, this does not mean that the bias is smaller. In fact, as seen in our figure, the bias is smaller only when the subjects underreact to retractions and opposite signals, but not when they overreact. However, as we have already discussed, the latter effect will be diminished as subjects learn, at which point it will be the case that retraction bias will probably be overall smaller than the corresponding bias to observing signals of the opposite color.

7 Different treatments

7.1 Varying information display

As we have already explained, in our baseline experiment, we ran 3 different treatments that differ only in the way information about past signals and previous reports was presented to the subjects, in order to control for the effect of memory and anchoring. Accordingly, in the first treatment the subjects saw the entire history of signals that had been previously realized and their last reported belief (treatment 1); in the second treatment we kept the history of signals and removed the last belief (treatment 2); in the third treatment we removed the history of signals and maintained the last belief (treatment 3). The sample sizes are balanced, i.e., the total sample sizes in the three treatments are 188, 221, 197 respectively, while the corresponding sample sizes after the removal of outliers are 164, 196, 180.

As illustrated in Figures 9(a) and 9(b), there is no significant effect across the three treatments in how retractions or confirmations are processed. Note that this is the case not only on aggregate, but also for each block of priors separately.

7.2 Ex-ante Verifications of Information

In this section we turn to the analysis of the third research question: What is the effect of checking information ex-ante? We analyze 1) whether people react differently to ex-ante verifications in general, as compared to ex-post checks, and 2) whether beliefs are more dispersed after retracted signals than equivalent uninformative signals.

To answer our third research question, what is the effect of verifying information ex-ante, we run an additional experiment (with 243 subjects, out of which 30 were removed based on predefined criteria). The median time to complete the experiment was 17 minutes and subjects received on average €4.75.

This experiment is based on the same design as explained above but varies one crucial detail. Instead of ex-post confirming or retracting the previous signal, subjects are immediately shown if a ball is informative or uninformative. Everything else, remains the same. Subjects thus see 6 uncertain signals and 3 signals which are immediately verified as informative or uninformative. Appendix C.3 provides further details on the differences between the three treatments along with a screenshot of how information was presented to subjects.

Our initial hypothesis was that ex-ante verification will reduce the bias, both in the case of retractions and verifications. What we find for starters is that qualitatively the same pattern applies as previously, i.e., for small priors subjects underreact to retractions and overreact to confirmations, and vice versa for large priors.

However, in terms of the magnitudes, the bias increases in the case of retractions for almost all priors (Figure B.4), and decreases for confirmations again for almost all priors (Figure B.4). This suggests that from a policy perspective, ex ante verifications are not necessarily beneficial.

8 Conclusion

This paper contributes to the literature on how people process information verifications asking whether confirmations are also processed differently from regular signals, and if yes whether the mechanism is the same that governs the processing of retractions. Second we set out to understand the dynamics of processing verifications, focusing on two different questions, i.e., does the same mechanism persist over time, and moreover do early observations of verifications affect updating biases at a later stage.

To answer these questions within a single experiment we modified the novel experimental design of Goncalves *et al.* (2026), by making confirmation not fully revealing of the state, as well as by making the horizon much longer.

The main result that we find is that both retractions and confirmations are driven by the same (previously-unidentified) mechanism, i.e., biases in both processing retractions and confirmations are explained by the prior beliefs. In particular, small prior beliefs lead to underreaction of retractions and overreaction of confirmations. On the other hand, large prior beliefs lead to overreaction of retractions and underreaction of confirmations. This finding is robust across all robustness checks and different treatments that we considered. Our explanation is that people react differently to information verification, depending on how surprised they are both by the initial information and by the verification. While we did not preregister this finding as a hypothesis, upon discovering it, we decided to present it in detail, given how fundamental and how robust it was in our data.

Taken together, our findings have important implications for the way uncertain information is communicated in practice. People might end up with substantially biased beliefs after a retraction or they may infer too little from a confirmation. Moreover, it is likely that the introduction of motivated reasoning and other context dependent features will amplify the bias in information processing we documented. Finally, it implies that the presence of continued misinformation is likely one driver for increasing belief dispersion, and in the extreme, belief polarization.

A Proofs

A.1 Proof of Proposition 1

The agent underreact to a retraction if and only if

$$\frac{p_{t-1}(R)}{p_{t-1}(B)} < \frac{p_{t+1}(R|r,✗)}{p_{t+1}(B|r,✗)}.$$

Using Equation (5), the last inequality can be rewritten as

$$\frac{p_{t-1}(R)}{p_{t-1}(B)} < \left[\frac{p_{t-1}(R)}{p_{t-1}(B)}\right]^{c\cdot c_✗}\left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{d\cdot c_✗ - d_✗},$$

which is in turn equivalent to

$$\left[\frac{p_{t-1}(R)}{p_{t-1}(B)}\right]^{1-c\cdot c_✗} < \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{d\cdot c_✗ - d_✗}, \tag{A.1}$$

Then, suppose that $c\cdot c_✗ < 1$, meaning that Inequality (A.1) holds if and only if

$$\frac{p_{t-1}(R)}{p_{t-1}(B)} < \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{\frac{d\cdot c_✗ - d_✗}{1-c\cdot c_✗}} = t_✗.$$

Hence, if we solve the last inequality, we obtain directly

$$p_{t-1}(R) < \frac{t_✗}{1+t_✗} = p_✗,$$

which completes part 1.ii of the result. The remaining parts follow analogously.

A.2 Proof of Proposition 2

The agent underreacts to a confirmation if and only if

$$\frac{p_{t+1}(R|r,✓)}{p_{t+1}(B|r,✓)} < \frac{p_{t-1}(R)}{p_{t-1}(B)}\frac{\pi_I(r|R)}{\pi_I(r|B)}.$$

Using Equation (10), the last inequality can be rewritten as

$$\left[\frac{p_{t-1}(R)}{p_{t-1}(B)}\right]^{c\cdot c_✓}\left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{d\cdot c_✓ - d_✓}\left[\frac{\pi_I(r|R)}{\pi_I(r|B)}\right]^{d_✓} < \frac{p_{t-1}(R)}{p_{t-1}(B)}\frac{\pi_I(r|R)}{\pi_I(r|B)},$$

This inequality can be directly rewritten as

$$\left[\frac{p_{t-1}(R)}{p_{t-1}(B)}\right]^{1-c\cdot c_✗} > \left[\frac{\pi_A(r|R)}{\pi_A(r|B)}\right]^{d\cdot c_✓ - d_✓}\left[\frac{\pi_I(r|R)}{\pi_I(r|B)}\right]^{d_✓ - 1},$$

and therest of the steps are identical to the ones in the proof of Proposition 1.

B Tables and figures

B.1 Reactions to initial signal vs reaction to retraction

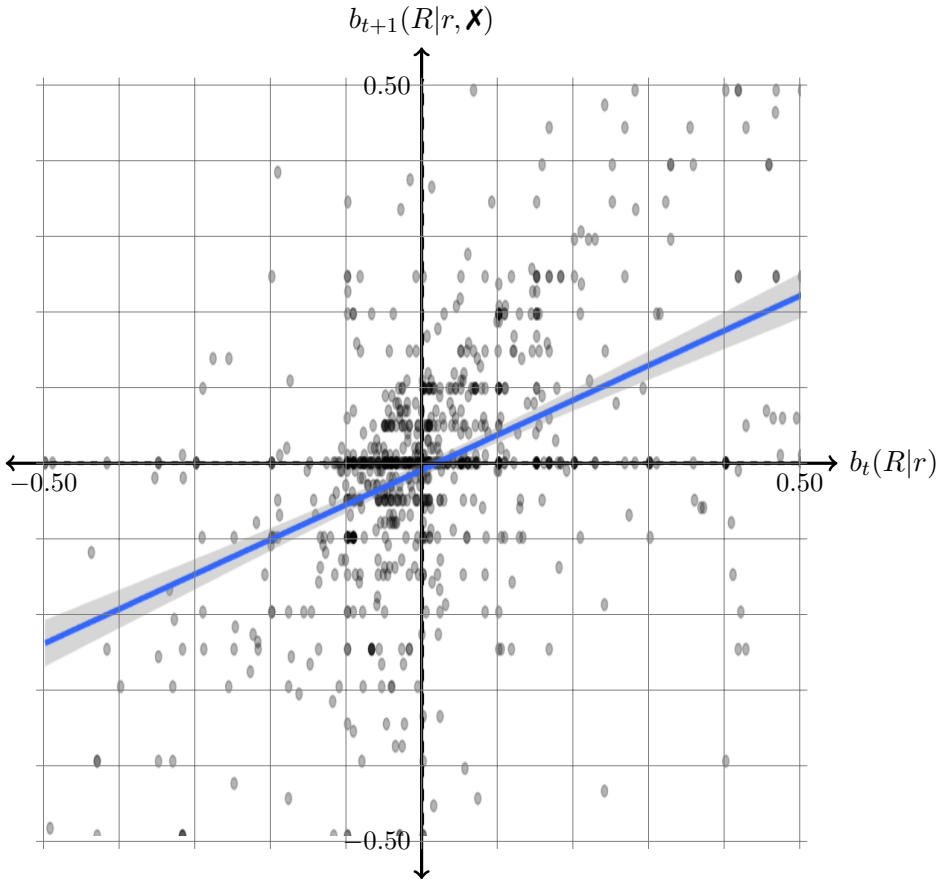


Figure B1: Correlation of initial bias and retraction bias.

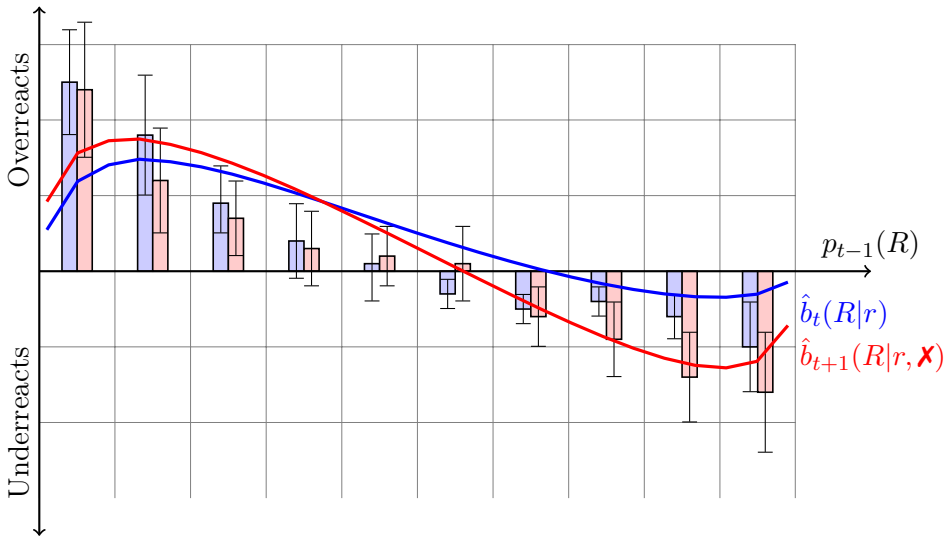
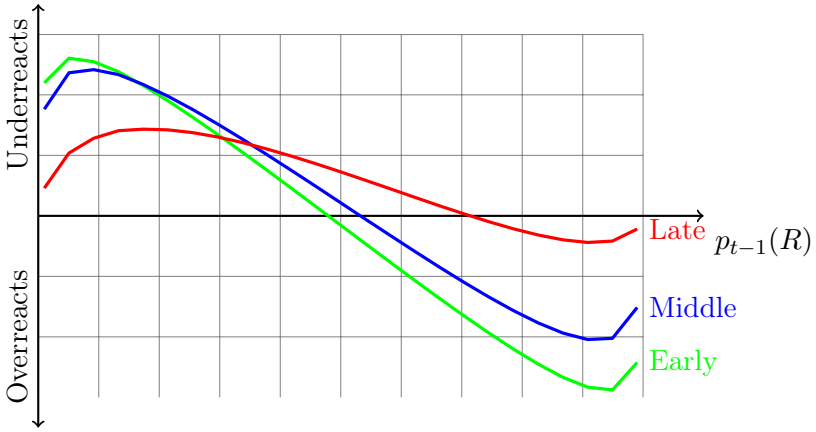


Figure B2: Bias when processing initial signals: The red color corresponds to retractions, and the blue color to initial signals.

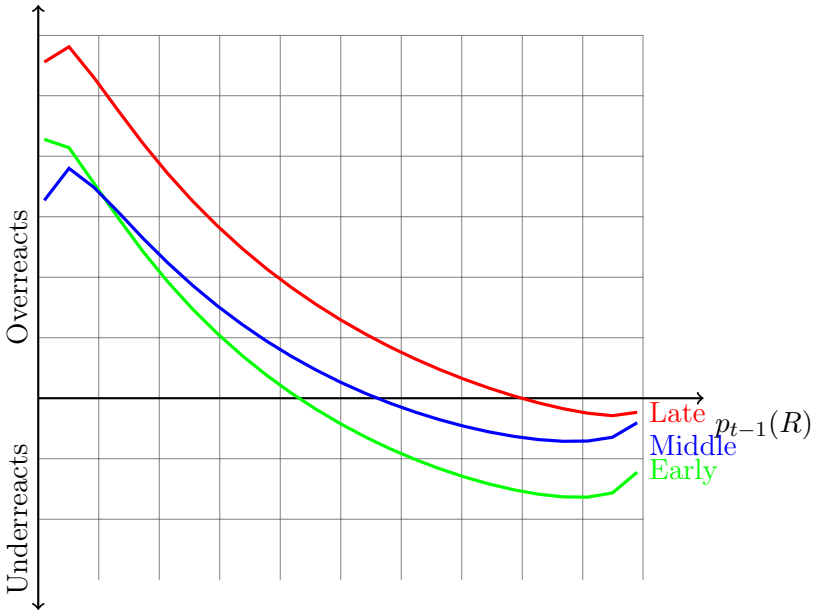
B.2 Reaction to verifications over time

Panel A: Base-Rate Use			
Period Block	Regular (c^T)	Retraction ($c^T_{\text{✗}}$)	Confirmation ($c^T_{\text{✓}}$)
Early (T_1)	0.393^{***}	0.635^{***}	0.589^{***}
	(0.021)	(0.060)	(0.076)
Middle (T_2)	0.570^{***}	0.594^{***}	0.700^{***}
	(0.020)	(0.044)	(0.061)
Late (T_3)	0.787^{***}	0.867^{***}	0.455^{***}
	(0.017)	(0.049)	(0.082)
Panel B: Inference			
Period Block	Regular (d^T)	Retraction ($d^T_{\text{✗}}$)	Confirmation ($d^T_{\text{✓}}$)
Early (T_1)	1.544^{***}	1.136^{***}	0.744^{***}
	(0.079)	(0.248)	(0.184)
Middle (T_2)	1.422^{***}	0.639^{***}	1.215^{***}
	(0.095)	(0.213)	(0.159)
Late (T_3)	1.541^{***}	0.618^{***}	2.474^{***}
	(0.120)	(0.292)	(0.298)

Table B1: Updating parameters estimated by period block: One OLS regression per signal type for each period block, $T \in \{T_1, T_2, T_3\}$. So there are 9 regressions in total.

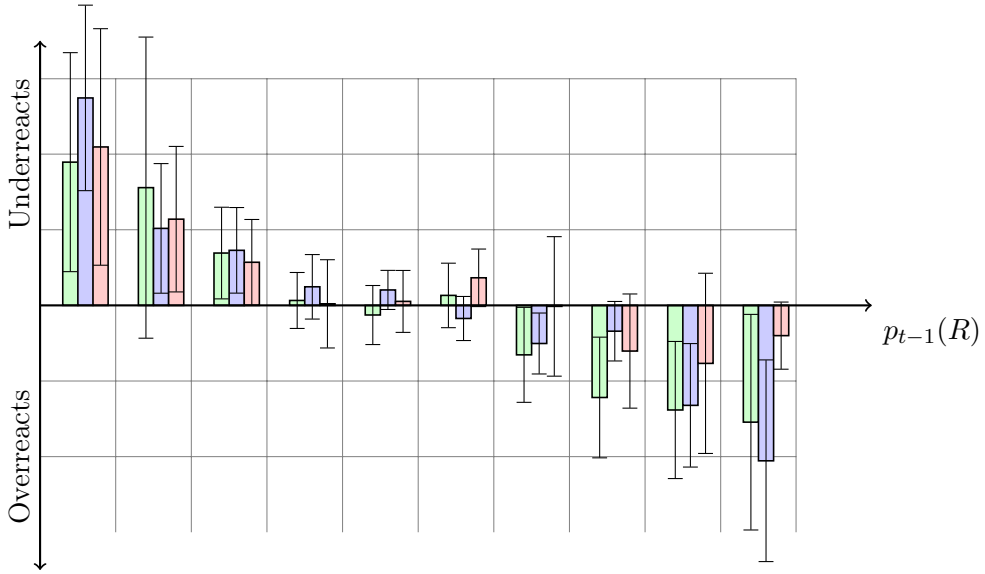


(a) Retractions

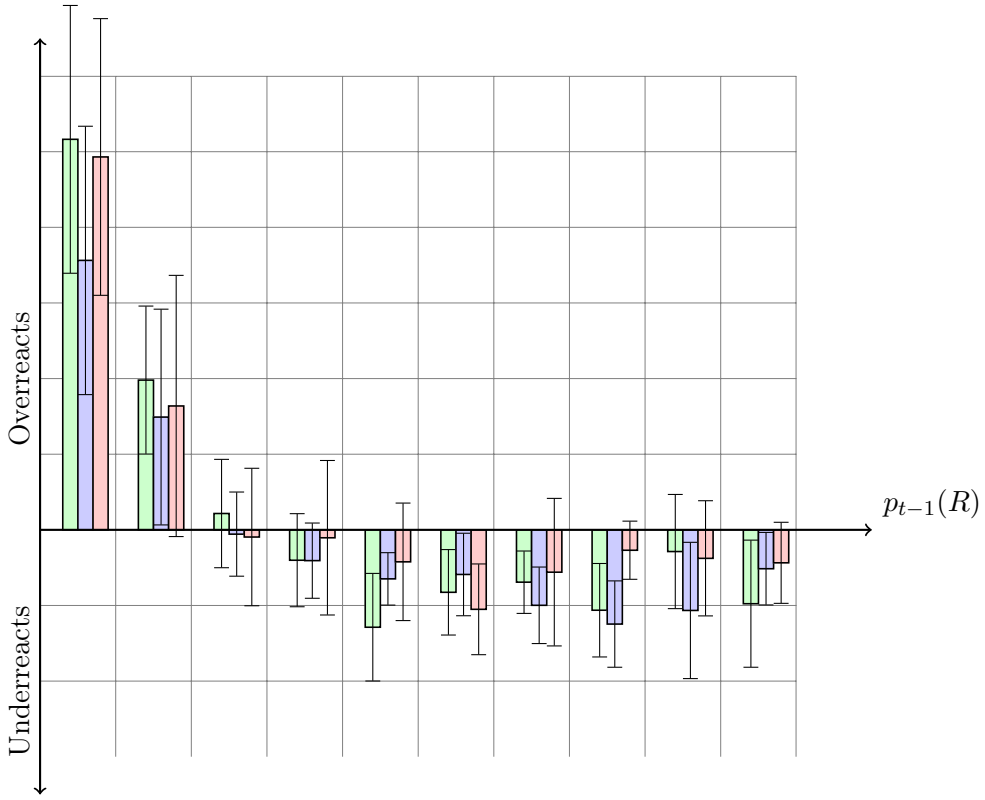


(b) Confirmations

Figure B3: Theoretical predictions per block of periods: The green color corresponds to the early periods, the blue color to the middle periods and the red color to the late periods.



(a) Retractions



(b) Confirmations

Figure B4: Observed bias when processing verifications per block of periods. The green color corresponds to the early periods, the blue color to the middle periods and the red color to the late periods.

	Base-rate use	Inference
	(c)	(d)
$(H_{0,0})$	0.601^{***}	1.540^{***}
	(0.015)	(0.218)
$(H_{1,0})$	0.409^{***}	1.603^{***}
	(0.042)	(0.249)
$(H_{0,1})$	0.483^{***}	1.213^{***}
	(0.046)	(0.137)
$(H_{2,0})$	0.403^{***}	1.239^{***}
	(0.052)	(0.315)
$(H_{0,2})$	0.734^{***}	1.560^{***}
	(0.067)	(0.239)
$(H_{1,1})$	0.653^{***}	1.210^{***}
	(0.035)	(0.152)

Table B2: Updating parameters for regular signals conditional on each profile of previously realized verifications: One OLS regression per signal type for profile of verifications, $H \in \{H_{0,0}, H_{1,0}, H_{0,1}, H_{2,0}, H_{0,2}, H_{1,1}\}$. So there are 6 regressions in total.

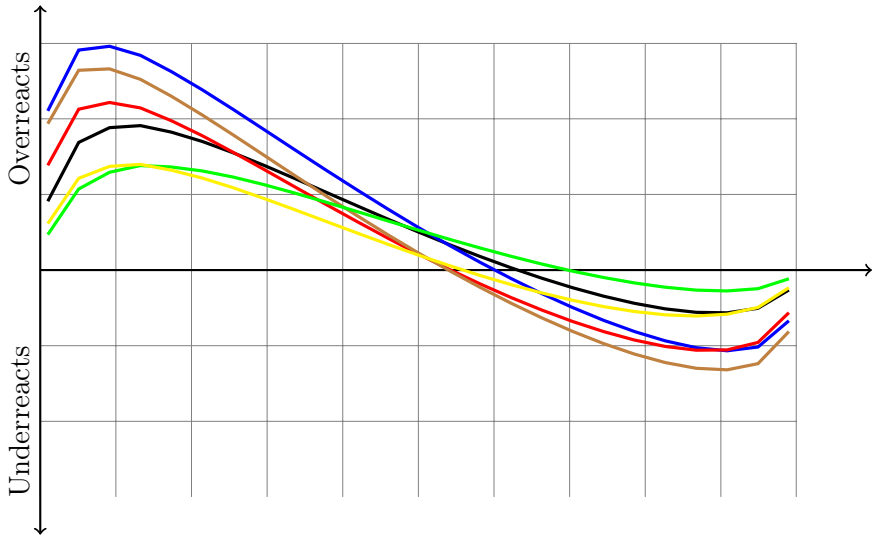
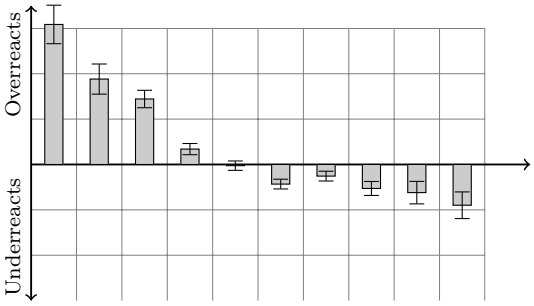
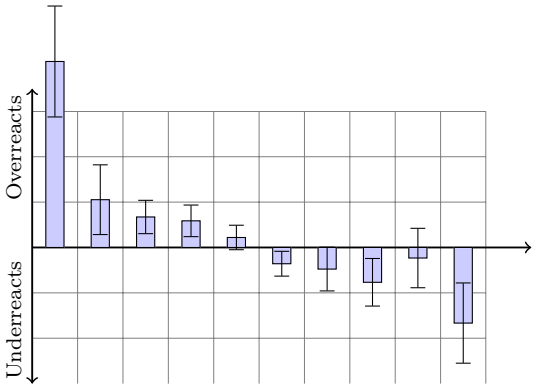


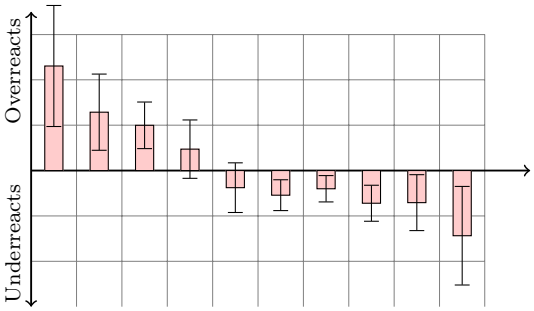
Figure B5: Theoretical predictions for regular signals per verification profile. The black color corresponds to $(H_{0,0})$, the blue color corresponds to $(H_{1,0})$, the red color corresponds to $(H_{0,1})$, the brown color corresponds to $(H_{2,0})$, the green color corresponds to $(H_{0,2})$, and the yellow color corresponds to $(H_{1,1})$.



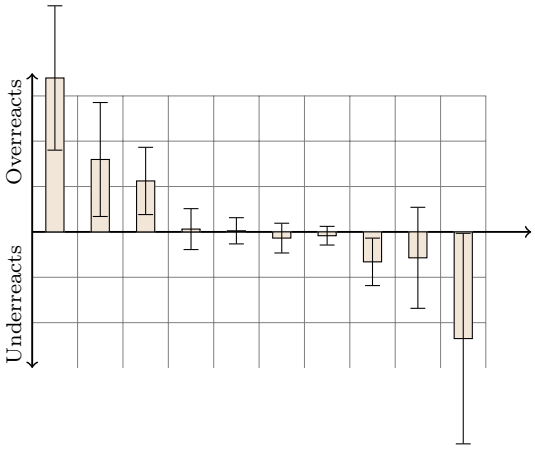
(a) $(H_{0,0})$: 0 retractions / 0 confirmations



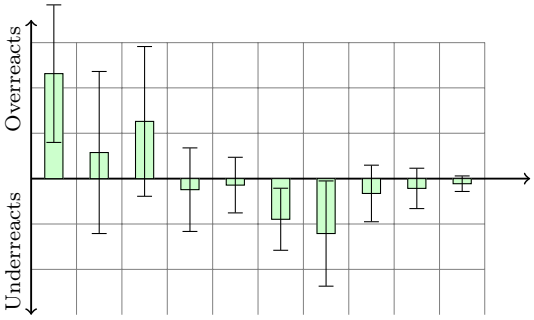
(b) $(H_{1,0})$: 1 retraction / 0 confirmations



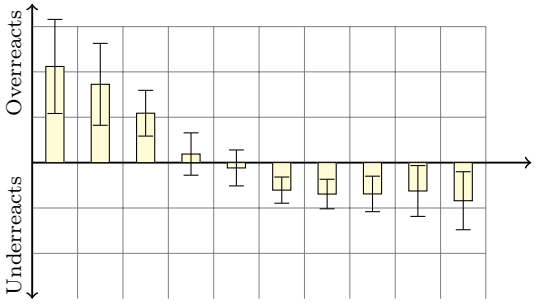
(c) $(H_{0,1})$: 0 retraction / 1 confirmations



(d) $(H_{2,0})$: 2 retractions / 0 confirmations

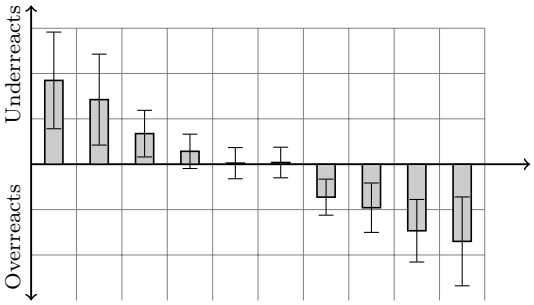


(e) $(H_{0,2})$: 0 retractions / 2 confirmations

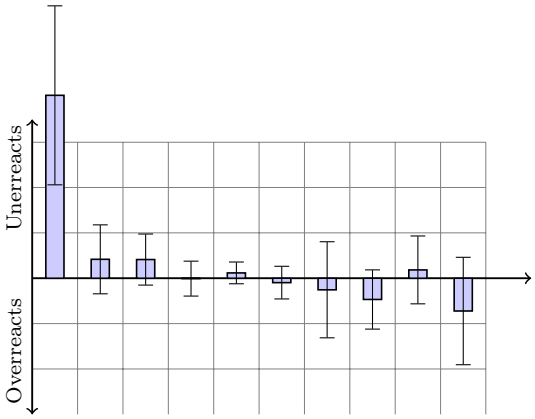


(f) $(H_{1,1})$: 1 retraction / 1 confirmation

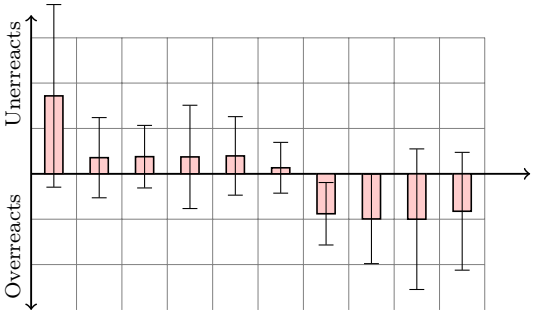
Figure B6: Observed bias for regular signals. The different subfigures correspond to the different verification profiles.



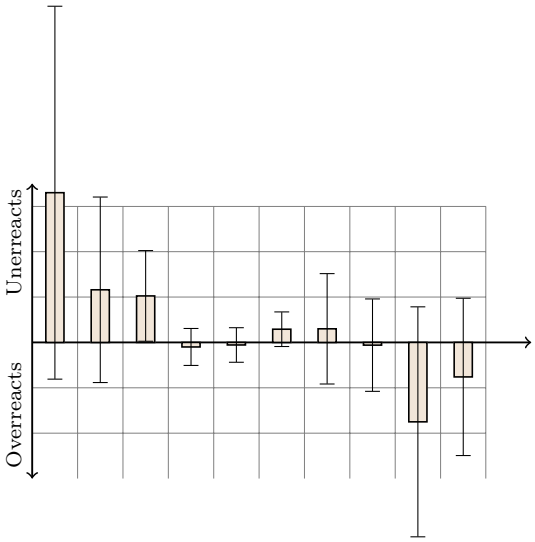
(a) $(H_{0,0})$: 0 retractions / 0 confirmations



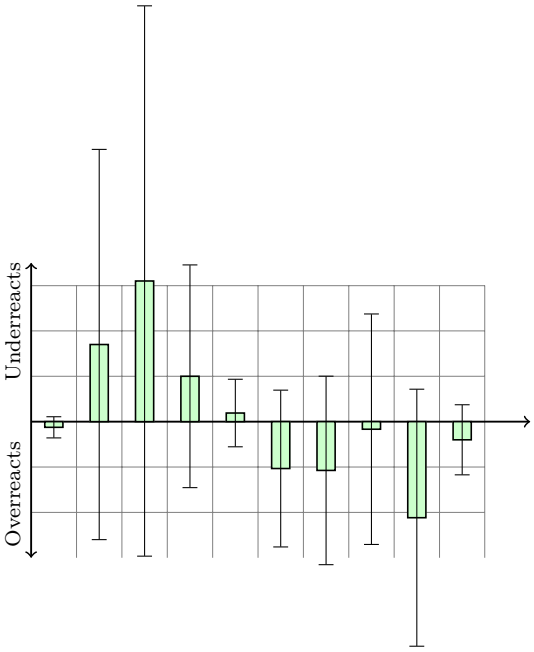
(b) $(H_{1,0})$: 1 retraction / 0 confirmations



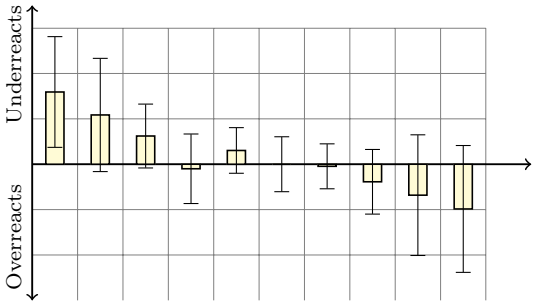
(c) $(H_{0,1})$: 0 retraction / 1 confirmations



(d) $(H_{2,0})$: 2 retractions / 0 confirmations

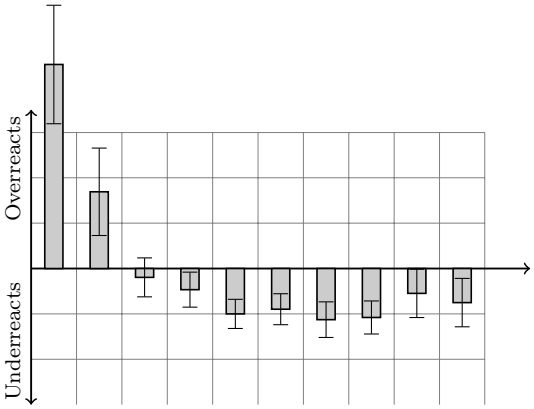


(e) $(H_{0,2})$: 0 retractions / 2 confirmations

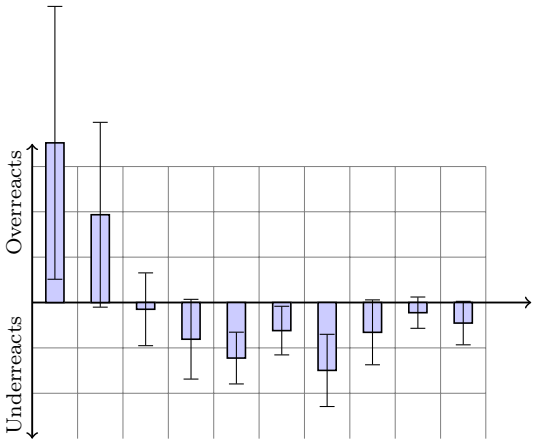


(f) $(H_{1,1})$: 1 retraction / 1 confirmation

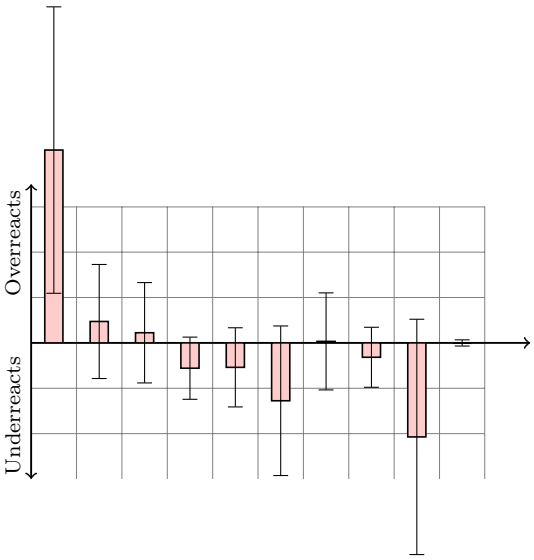
Figure B7: Observed bias for retractions. The different subfigures correspond to the different verification profiles.



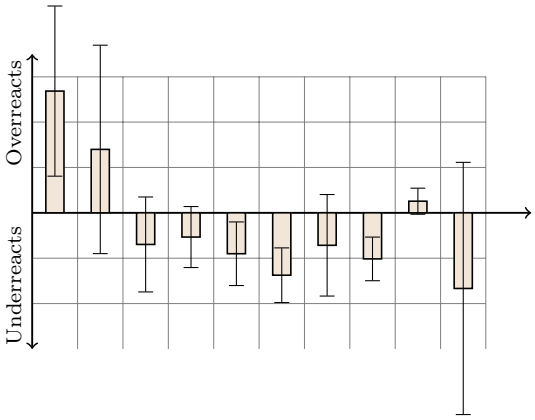
(a) $(H_{0,0})$: 0 retractions / 0 confirmations



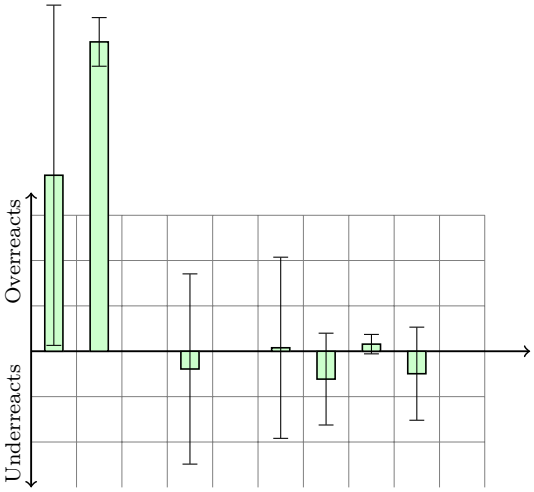
(b) $(H_{1,0})$: 1 retraction / 0 confirmations



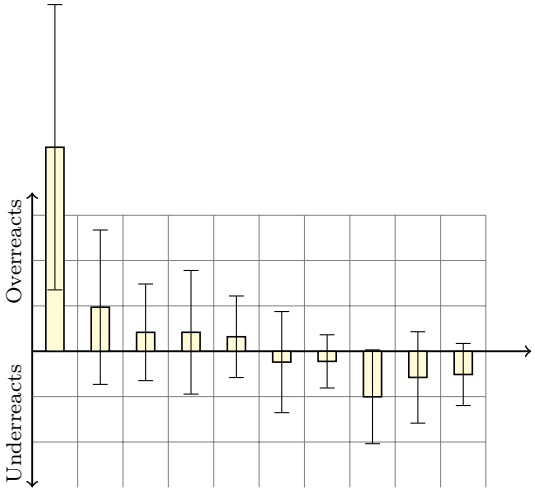
(c) $(H_{0,1})$: 0 retraction / 1 confirmations



(d) $(H_{2,0})$: 2 retractions / 0 confirmations



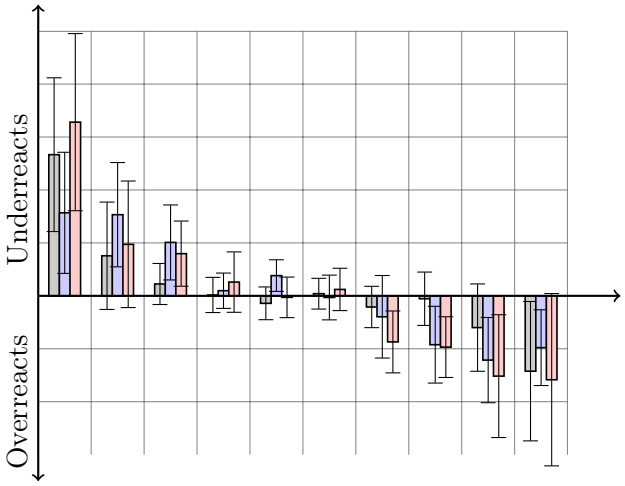
(e) $(H_{0,2})$: 0 retractions / 2 confirmations



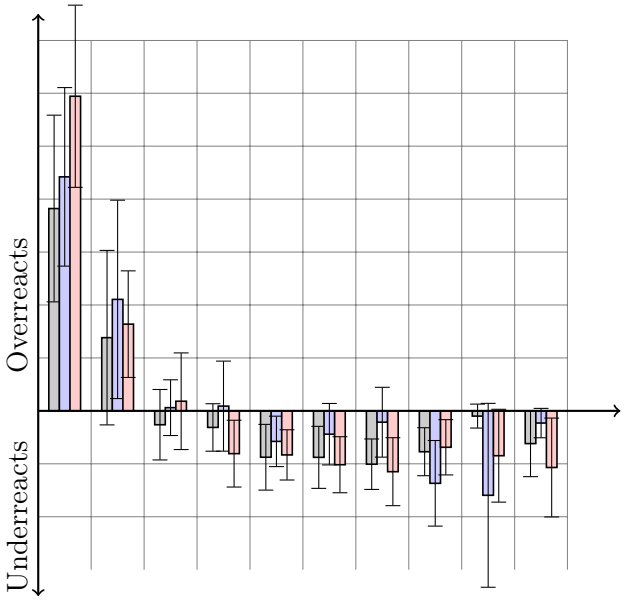
(f) $(H_{1,1})$: 1 retraction / 1 confirmation

Figure B8: Observed bias for confirmations. The different subfigures correspond to the different verification profiles.

B.3 Treatment analysis



(a) Retractions



(b) Confirmations

Figure B9: Bias per treatment. The black color corresponds to Treatment 1, the blue color to Treatment 2, and the red color to Treatment 3.

B.4 Ex ante vs ex post verification

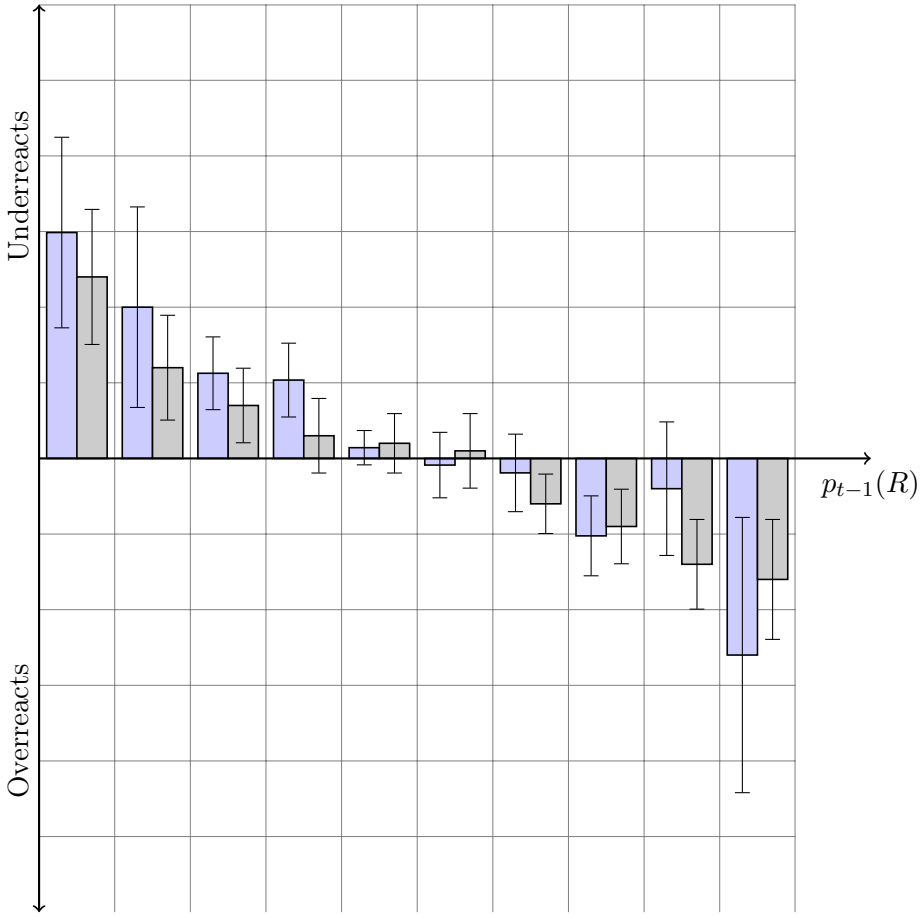


Figure B10: Bias when signal is retracted ex ante: The blue color is the ex ante retraction, and the black color the ex post.

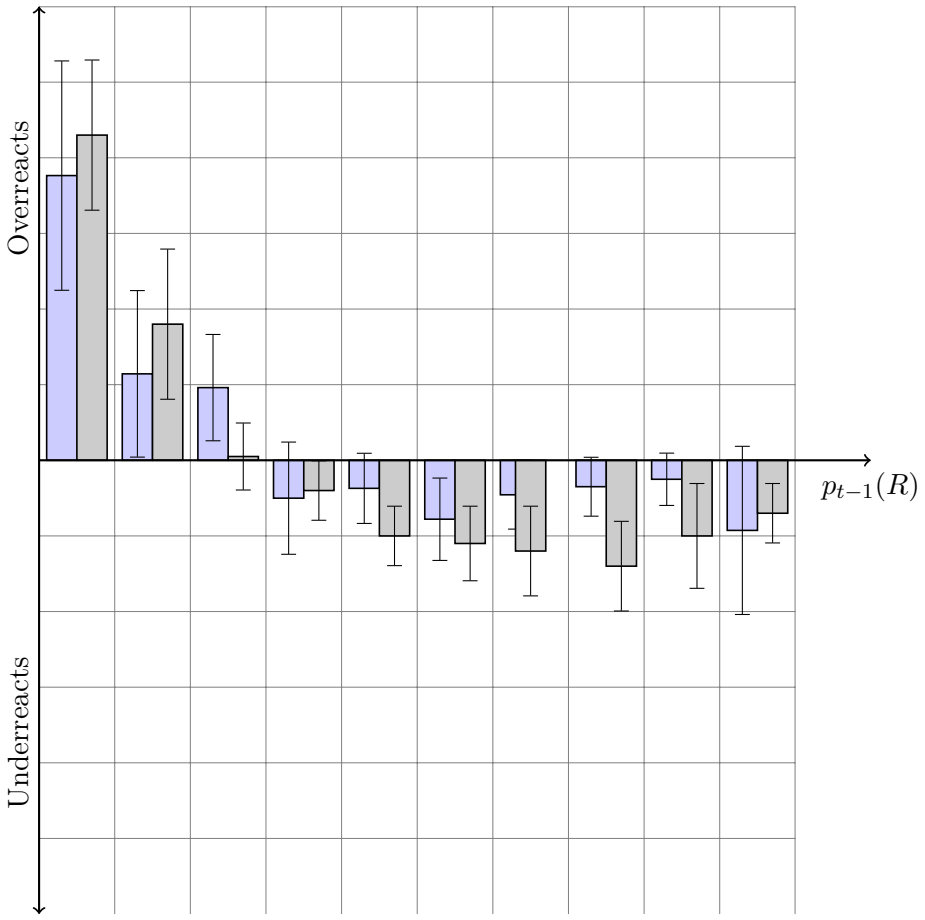
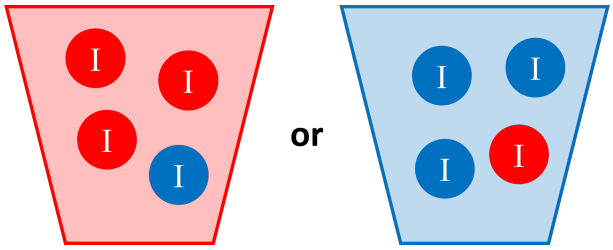


Figure B11: Bias when signal is confirmed ex ante: The blue color is the ex ante confirmation, and the black color the ex post.

C Instructions and Screenshots

C.1 Instructions - Ex-post Information Checks

There are two urns, one red and one blue, each containing 4 balls as displayed below. All balls are labeled with a letter 'I', the meaning of which will be explained on the next page. One of the two urns will be randomly selected in the beginning. You do not know which urn will be selected. It will remain the same throughout this study.

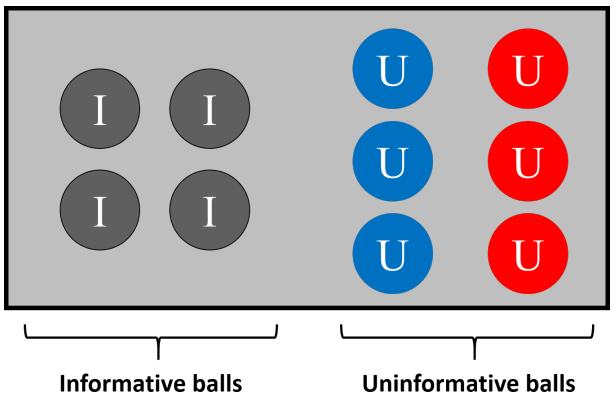


Your **task will be to guess which urn you think was selected.** To do so you will receive hints about the selected urn. This will be described on the next page.

[Page break]

The selected urn contains 4 balls, 3 of the same color and 1 of the opposite color, as shown on the previous page. All balls from the selected urn are put into a black box. They are labeled with the letter 'I', which stands for 'informative'. If you knew the colour of all 4 balls, you would be able to identify the selected urn. For the moment you do not know the color of the selected urn and therefore the 4 informative balls are displayed in grey (although they have a color, either red or blue).

The black box also contains 6 other balls that do not come from the urn. These 6 balls are labeled with the letter 'U', which stands for 'uninformative'. Knowing the color of the uninformative balls does not help to identify the selected urn. **The black box and the 10 balls inside it remain the same throughout the entire study.**



This study has 12 rounds in total. In each round you get one of two possible hints:

- **A ball is drawn** from the black box. You are told the color of the ball. The ball is put back into the box together with the other 9 balls. You do not know if the balls is informative or uninformative. *Example:* A red ball is drawn from the box. ?
- No new ball is drawn, but you receive **information about the ball you saw previously**. You will be told if the ball you saw before was one of the 4 informative balls or one of the 6 uninformative balls. *Example:* Previously you saw a red ball. The ball that was drawn last round was one of the 6 uninformative balls. U

In every round you will be asked to make a guess about the urn that has been selected in the beginning. At the end of the study you will be told the color of this urn. You will receive a bonus which depends on the accuracy of your answer to one of the 12 guesses (you do not know which one). The procedure for calculating your bonus is described in detail below. You may skip these details. The important thing is that the procedure guarantees that you should expect to maximize your bonus by reporting what you truly think the chances are in each question.

[Button: 'Details about the bonus']

We apply the following procedure:
First, we randomly pick one of the questions. For this question, we calculate the error you made. This is how many percentage points your report was away from 100% (if the RED URN was selected) or from 0% (if the BLUE URN was selected). Then, we plug in the error into the following formula:

$$3 - 3 \cdot \text{error}^2$$

This will be your bonus (in Euros).
EXAMPLE: Suppose that you report 60% chance that the RED URN was selected in Step 1. Then, your bonus is calculated as follows:

- If the urn was RED:
  - your error is $(100\% - 60\%) = 40\%$
  - your bonus is $3 - 3 \cdot (40\%)^2 = 2.52$Euros
- If the urn was BLUE:
  - your error is $(60\% - 0\%) = 60\%$
  - your bonus is $3 - 3 \cdot (60\%)^2 = 1.92$Euros

As we have already mentioned, you should expect to maximize your earnings by reporting what you actually think are the chances of a RED URN.
Example: If you actually think that the chances are 60% that the RED URN was selected, then:

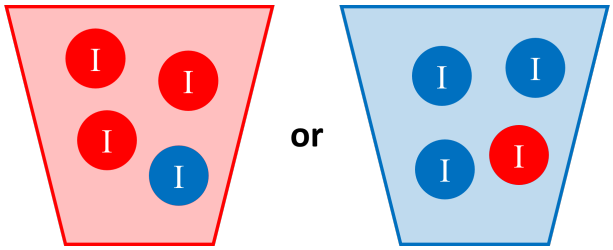
- By reporting 60%, you will make on average 2.28 Euros.

- By reporting 10%, you will make on average 1.53 Euros.
- By reporting 100%, you will make on average 1.80 Euros.

As you see you maximize your earning by reporting exactly 60%. The further away you report from what you actually think, the less money you should expect to make.

C.2 Instructions - Ex-ante Information Verification

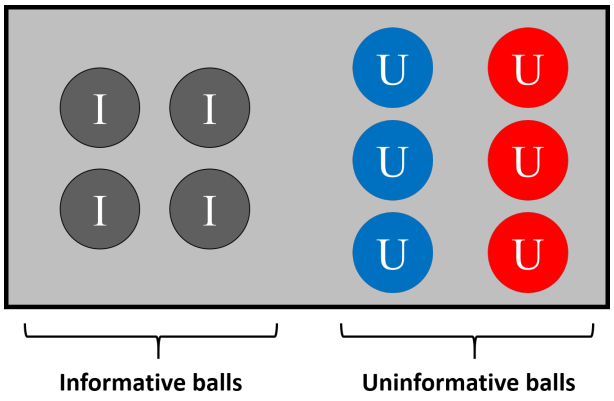
There are two urns, one red and one blue, each containing 4 balls as displayed below. All balls are labeled with a letter 'I', the meaning of which will be explained on the next page. One of the two urns will be randomly selected in the beginning. You do not know which urn will be selected. It will remain the same throughout this study.



Your **task will be to guess which urn you think was selected.** To do so you will receive hints about the selected urn. This will be described on the next page.

[Page break]

The selected urn contains 4 balls, 3 of the same color and 1 of the opposite color, as shown on the previous page. All balls from the selected urn are put into a black box. They are labeled with the letter 'I', which stands for 'informative'. If you knew the colour of all 4 balls, you would be able to identify the selected urn. For the moment you do not know the color of the selected urn and therefore the 4 informative balls are displayed in grey (although they have a color, either red or blue).
The black box also contains 6 other balls that do not come from the urn. These 6 balls are labeled with the letter 'U', which stands for 'uninformative'. Knowing the color of the uninformative balls does not help to identify the selected urn. **The black box and the 10 balls inside it remain the same throughout the entire study.**



This study has 9 rounds in total. In each round a ball is drawn from the black box and you get one of two possible hints:

- You are **told only the color** of the ball. The ball is put back into the box together with the other 9 balls. You do not know if the ball is informative or uninformative. *Example:* A red ball was drawn from the box. ?
- You are **told the color and the letter** on the ball. The ball is put back into the box together with the other 9 balls. You know if the ball is informative or entirely uninformative. *Example:* A red ball was drawn from the box. It is one of the 4 informative balls: U

In every round you will be asked to make a guess about the urn that has been selected in the beginning. At the end of the study you will be told the color of this urn. You will receive a bonus which depends on the accuracy of your answer to one of the 12 guesses (you do not know which one). The procedure for calculating your bonus is described in detail below. You may skip these details. The important thing is that the procedure guarantees that you should expect to maximize your bonus by reporting what you truly think the chances are in each question.

[Button: 'Details about the bonus']

We apply the following procedure:
First, we randomly pick one of the questions. For this question, we calculate the error you made. This is how many percentage points your report was away from 100% (if the RED URN was selected) or from 0% (if the BLUE URN was selected). Then, we plug in the error into the following formula:

$$3 - 3 \cdot \text{error}^2$$

This will be your bonus (in Euros).
EXAMPLE: Suppose that you report 60% chance that the RED URN was selected in Step 1. Then, your bonus is calculated as follows:

- If the urn was RED:

- your error is $(100\% - 60\%) = 40\%$
- your bonus is $3 - 3 \cdot (40\%)^2 = 2.52$Euros

- If the urn was BLUE:
  - your error is $(60\% - 0\%) = 60\%$
  - your bonus is $3 - 3 \cdot (60\%)^2 = 1.92$Euros

As we have already mentioned, you should expect to maximize your earnings by reporting what you actually think are the chances of a RED URN.
Example: If you actually think that the chances are 60% that the RED URN was selected, then:

- By reporting 60%, you will make on average 2.28 Euros.
- By reporting 10%, you will make on average 1.53 Euros.
- By reporting 100%, you will make on average 1.80 Euros.

As you see you maximize your earning by reporting exactly 60%. The further away you report from what you actually think, the less money you should expect to make.

C.3 Example screen with ball draw

We have three treatments to vary information display. One version is displayed below. In the other two we either do not display the table containing the history of previous signals or we do not display the sentence reminding people of their previously reported belief. All else stays the same. It should also be noted that currently no default belief is selected on the slider. Once the subject clicks on the slider an icon appears. Also, the belief is displayed below it in front of the percentage symbol.

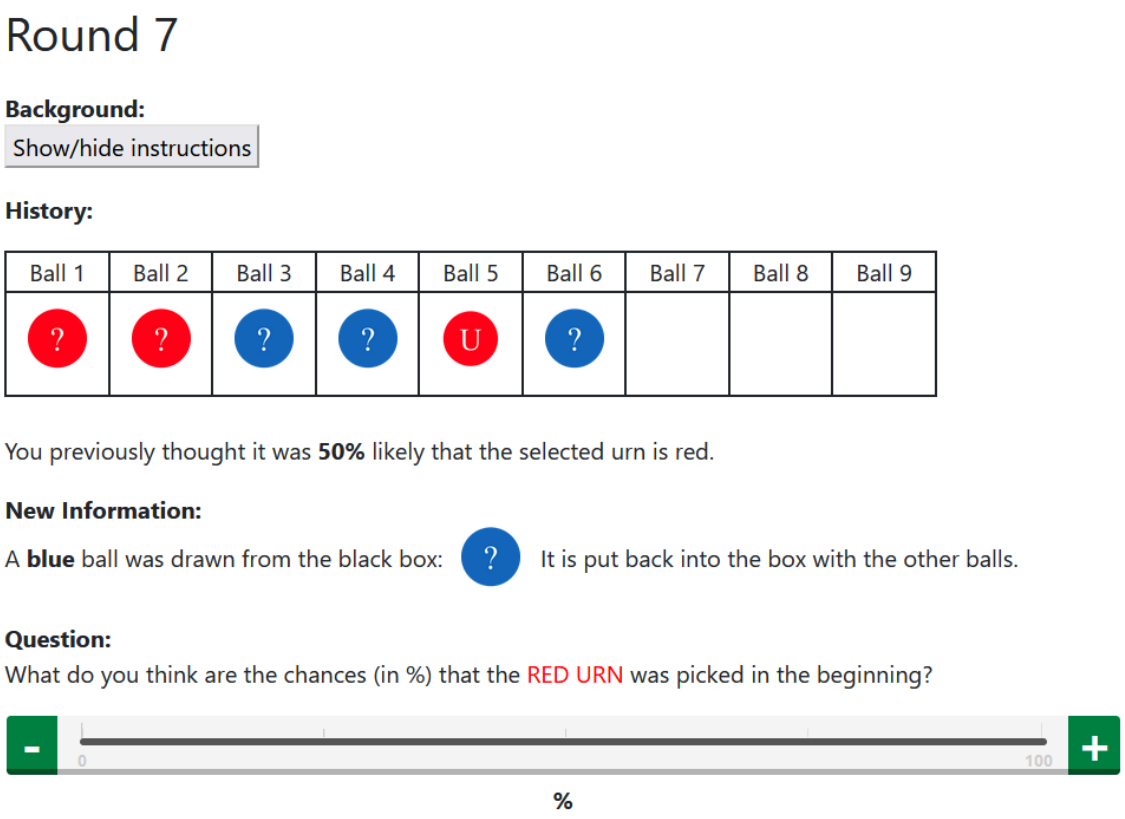


Figure C12: Example Screen

References

Augenblick, N., Lazarus, E. and Thaler, M. (2025). Overinference from weak signals and underinference from strong signals. *Quarterly Journal of Economics*, **140**, 335–401.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics - Foundations and Applications 2*, Elsevier, pp. 69–186.

Bronnikov, E., Tsakas, E., Vostroknutov, A. and Thomsson, K. (2026). Breaking through the information bubble: How surprise shapes belief updating across media sources. *Working Paper*.

Browne, M. (2018). Epistemic divides and ontological confusions: The psychology of vaccine scepticism. *Human Vaccines and Immunotherapeutics*, **14**, 2540–2542.

Charness, G., Oprea, R. and Yuksel, S. (2021). How do people choose between biased information sources? evidence from a laboratory experiment. *Journal of the European Economic Association*, **19** (3), 1656–1691.

Chen, D. L., Schonger, M. and Wickens, C. (2016). oTree - an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E. *et al.* (2020). Real solutions

for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, **42** (4), 1073–1095.

Danz, D., Vesterlund, L. and Wilson, A. (2020). Belief elicitation: Limiting truth telling with information on incentives. *NBER Working Paper*.

Ecker, U., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L., Brashier, N., Kendeou, P., Vraga, E. and Amazeen, M. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, **1** (1), 13–29.

Ecker, U. K., Lewandowsky, S. and Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, **38** (8), 1087–1100.

Enke, B., Schwerter, F. and Zimmermann, F. (2020). Associative memory and belief formation. *NBER Working Paper*.

Epstein, L. G. and Halevy, Y. (2021). Hard-to-interpret signals.

Goncalves, D., Libgober, J. and Willis, J. (2026). Retractions: Updating from complex information. *Review of Economic Studies*, **93**, 476–516.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, **95** (3), 537.

Hoppe, E. I. and Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters*, **110** (2), 97–100.

Johnson, H. M. and Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20** (6), 1420.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N. and Cook, J. (2012). Misinformation and its correction. *Psychological Science in the Public Interest*, **13** (3), 106–131.

Liang, Y. (2020). Learning from unknown information sources. *SSRN Electronic Journal*.

Nieminen, S. and Rapeli, L. (2018). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, **17** (3), 296–309.

O'Rear, A. E. and Radvansky, G. A. (2020). Failure to accept retractions: A contribution to the continued influence effect. *Memory & Cognition*, **48** (1), 127–144.

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

Phillips, L. D. and Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, **72** (3), 346–354.

Roets, A. *et al.* (2017). 'fake news': Incorrect, but hard to correct. the role of cognitive ability on the impact of false information on social impressions. *Intelligence*, **65**, 107–110.

Shishkin, D. and Ortoleva, P. (2021). Ambiguous information and dilation: An experiment.

Thorson, E. (2015). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, **33** (3), 460–480.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76** (2), 105–110.

— and — (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185** (4157), 1124–1131.

Uscinski, J. E., Enders, A. M., Klofstad, C., Seelig, M., Funchion, J., Everett, C., Wuchty, S., Premaratne, K. and Murthi, M. (2020). Why do people believe covid-19 conspiracy theories? *Harvard Kennedy School Misinformation Review*, **1** (3).

Walter, N., Cohen, J., Holbert, R. L. and Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, **37** (3), 350–375.