Integrating explicit reliability for optimal choices: effect of trustworthiness on decisions and metadecisions

Keiji Ota[1,2], Anthony Ciston[4], Patrick Haggard[1], Thibault Gajdos Preuss[3*] & Lucie Charles[1,2*]

[1] Institute of Cognitive Neuroscience, University College London, London, United Kingdom

[2] School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

[3] Aix Marseille Univ, CNRS, CRPN, Marseille, France

[4] Max Planck Institute for Human Cognitive and Brain Sciences

** these authors contributed equally*

Corresponding author

Keiji Ota Email: keiji.ota@qmul.ac.uk

School of Biological and Behavioural Sciences

Queen Mary University of London

Abstract

A key challenge in today's fast-paced digital world is to integrate information from various sources, which differ in their reliability. Yet, little is known about how explicit probabilistic information on the likelihood of a source to provide correct information is used in decision-making. Here, we investigated how such explicit reliability markers are integrated in decision-making and the extent to which individuals have metacognitive insight into this process. We developed a novel paradigm where participants viewed the opinions from sources of varying reliability to decide between two options. After each decision, they rated how much a given source influenced their choice. Using computational modelling, we found that participants systematically misrepresented source reliability, inflating the likelihood that a source would provide correct information. They were also biased by unreliable sources of information, even when they were explicitly labelled as random. Additionally, sources that were labelled as reliably wrong, suggesting consistently the incorrect answer, impaired performance even though participants knew that they could use them to infer the correct choice. This effect was captured by an increased recency in the evidence accumulation process likely due to increased memory load when those sources were present. Despite these biases, participants showed metacognitive awareness of what influenced their choices: they were generally accurate in reporting how much a source influenced them and were aware of the impact of unreliable sources on their decisions. These results suggest that people show a suboptimal use of explicit information about reliability but have some metacognitive knowledge of their own choice biases.

[250 words]

Keywords

Evidence accumulation, Bayesian decision theory, metacognition, introspection

Introduction

In today's fast-paced digital world, we are inundated with information from diverse sources, including social media, news outlets, governments, and private companies, on complex and critical issues such as health, environment or economy. A key challenge is integrating information, even when different sources contradict each other or vary in reliability. Intuitively, providing explicit indicators of a source's reliability should help people assess and use information more accurately. However, empirical evidence on the effectiveness of displaying information trustworthiness remains mixed. For instance, some studies found that star ratings about news sources' credibility can reduce the tendency to believe and share false information (Celadin et al., 2023; Epstein & Robertson, 2015; Hansen et al., 2014) while others have observed that reliability labels do not affect the consumption of information from unreliable sources or the perception of misinformation (Aslett et al., 2022). On a more fundamental level, the cognitive processes underlying how people use explicit information about source reliability remain poorly understood. Therefore, empirical evidence and theoretical models of how reliability is encoded and integrated in the decision process are essential to design effective information policies and safeguard informed decision-making.

It has been shown that humans implicitly learn to weight information according to the precision of its source, for instance, when combining information from different sensory modalities (Deneve & Pouget, 2004; Ernst & Banks, 2002; Kording et al., 2007) or when operating visuomotor control (Kording & Wolpert, 2004; Ota et al., 2020; Tanae et al., 2021; Trommershauser et al., 2005). These studies demonstrate that behaviour, in this case, is well modelled by a Bayesian inference process where evidence is weighted according to its predictive power and that choices ultimately are more influenced by accurate than noisy signals (but see (Ota & Maloney, 2024; Ota et al., 2015, 2016; Ota et al., 2019)). Importantly, in most of these studies, experimenters manipulate the reliability of the evidence and let participants form beliefs about the likelihood of the information being true through feedback. It is reasonable to assume that in this case, the reliability of a source is somehow (not necessarily consciously, though) represented as a probability that the information provided by this source is accurate. Indeed, several models have proposed how the brain could encode such measures of uncertainty to optimise behaviour (Knill & Pouget, 2004; Walker et al., 2023)

However, humans can also develop internal representations of those probabilities explicitly, by being told how likely a source is to provide correct and truthful information. Indeed, previous research indicates that people make decisions using explicit probabilistic knowledge,

for instance when given information about the likelihood of receiving rewards or the occurrence of an event (Kahneman & Tversky, 1984). Interestingly, it has been shown that people often struggle to make optimal choices when reasoning with explicit probabilities. For instance, it was demonstrated that people make choices as if they overestimated the probability of rare events and underestimated the probability of frequent events (Gonzalez & Wu, 1999; Kahneman & Tversky, 1979). This distortion of probabilistic information through a non-linear probability weighting function resembling an “inverted S-shape” has been confirmed in several studies (Tversky & Kahneman, 1992) and is the subject of intense research (Martins, 2006; Ungemach et al., 2009). However, it is less clear whether suboptimal choices arise when probabilities are not about the likelihood of an event occurring but about the likelihood of providing correct information. In other words, is explicit probabilistic knowledge about the reliability of a source subject to specific distortion?

Crucially, the human cognitive system forms probabilistic beliefs not only about the external world but also its own functions (Koriat, 2006; Son & Schwartz, 2002). Indeed, recent research in the field of metacognition has shown that people accurately monitor their own cognitive processes (Fleming & Dolan, 2012; Katyal & Fleming, 2024) and evaluate the accuracy of their decisions, sometimes in non-conscious ways (Charles et al., 2014; Charles et al., 2013; Charles & Yeung, 2019; Yeung & Summerfield, 2012). Moreover, people form probabilistic beliefs about their decisions, computing the level of evidence supporting their choice and the likelihood of a given decision to be correct (Meyniel, 2020; Meyniel & Dehaene, 2017). These judgments of confidence are essential for people to communicate their beliefs (Bang et al., 2014) and to weigh advice coming from others (Carlebach & Yeung, 2023; Pescetelli et al., 2021; Pescetelli & Yeung, 2021) when making joint decisions (Bahrami et al., 2010; Hertz et al., 2017). While these studies begin to reveal how people interpret a source’s self-reported reliability, there is little evidence on the seemingly simpler case of using objective reliability estimates, computed independently of what the source thinks of itself.

Beyond understanding decisions, the question of whether people can accurately monitor how much weight they, in fact, give to different sources of information remains underexplored. In other words, if people are assigning a certain weight to a source, can they produce a clear account of how they do it? Preliminary research suggests that when people must base their choice on a combination of several attributes, they often use quick heuristics to solve problems (Glöckner & Betsch, 2023), but struggle to know the factors that influenced them the most (Cash & Oppenheimer, 2024a, 2024b). People often misattribute the true reasons for their decisions

(Celadin et al., 2023; Epstein & Robertson, 2015; Hansen et al., 2014), falling victim to "bias blindness" (Greenwald et al., 1998; Nosek et al., 2005; Pronin, 2007; Pronin et al., 2002). They can even experience an illusory sense of control (Sidarus et al., 2013; Wenke et al., 2010) over decisions that are in fact, manipulated (Charles & Haggard, 2020; Kummen et al., 2023). Nonetheless, it remains unclear whether the same metacognitive blind spots exist when introspecting the effect of sources' reliability on choices.

The aim of this study is to tackle these two questions: How is information reliability integrated into decision-making processes, and are people able to recognize the extent to which they are influenced by pieces of information with varying levels of reliability? For that purpose, we have developed a novel paradigm that mimics combining information from sources of variable reliability. In the task, participants see successive samples (red and blue squares) supporting one of two possible responses (red or blue). Each sample is associated with a percentage corresponding to the reliability of the source—the probability that the source is showing correct information about the true underlying colour. Participants must combine the evidence (the colour of the squares) with the reliability score of the source (the percentage) to decide which response it supports. After viewing 6 samples, they decide which response (red or blue) is likely to be correct.

We conceptualised three categories of source reliability based on their percentage. First, reliability scores that are above 50% mean that the source is more likely to give correct information and is therefore *reliably right* and informative for the decision. Second, a reliability score equal to 50% means that the source is as likely to give correct information as incorrect information. Such source can be categorized as *unreliable* and, as it is not predictive of the correct response, should be ignored. Finally, reliability scores below 50% mean that the source is more likely to give incorrect information than correct information, making it *reliably wrong*. Importantly, a reliably wrong source is informative for the choice but its prediction should be reinterpreted as evidence against the content it displayed.

The first aim of our study was therefore to investigate how people used these three categories of sources reliability and whether they could assign appropriate weight to the evidence provided by sources based on explicit percentages about reliability. To account for the decision process, we developed a computational model of participant choices, proposing a framework of how explicit cues about source reliability are used and combined with the evidence to make decisions. We hypothesised that participants encode the reliability of the evidence in a distorted way, misrepresenting the likelihood of the source to give correct information.

Furthermore, we predicted that participants would update their belief about the correct response when new evidence is presented, but with some degree of recency, giving higher weight to newer information. Such approach allows us to quantify precisely for each participant which source and which piece of evidence would influence their choice maximally. Additionally, we also probed participants' awareness about their own decision-making process, asking them to rate, using a visual analogue scale, the extent to which they felt their choice was influenced by a given source of information. Therefore, this design allowed us to compare *objective influence*, the extent to which a choice is influenced by information from a given source, with *subjective influence*, the extent to which participants perceive a given source influences their choice.

Our results confirmed that the weight participants assigned to information is not optimal, giving too much weight to unreliable information and too little weight to reliably wrong information. Nevertheless, we demonstrate that the subjective feeling of influence was correlated with the actual degree of influence of a given source. In particular, participants were able to report being biased by unreliable sources of information. These findings suggest that human behaviour deviates from optimality but metacognitive monitoring allows people to become aware of some of those limitations.

Results

Choice Behaviour Analysis

Task accuracy

In five experiments, participants performed a decision-making task where they had to sample information from different sources of varying reliability to decide between two options. On each trial, the correct colour (red or blue) for that trial was randomly drown from two equally probable possibilities. To decide which colour was the correct one, participants were shown a sequence of six evidence samples consisting in the opinion of six sources about the correct colour, displayed as red and blue square (Figure 1A). The reliability of the source, i.e. its likelihood to give correct information was shown on top of each square as a percentage. For each information source, the reliabilities of the sources were chosen from three levels with equal probability (50%, 55% and 65% in Experiment 1a; 50%, 45% and 65% in Experiment 1b; 50%, 55% and 35% in Experiment 1c, 50%, 55% or 65% in Experiment 2a; 50%, 45% or 35% in Experiment 2b). The participant's task was to choose a colour that they believed was more likely to be correct after sampling six information sources. Feedback was provided after each choice. In Experiments 2a and 2b, participants were asked additionally to report how *they felt* their choice was influenced by a particular source (for instance, 65% or 35% in Figure 1A), chosen at random (50%, 55% or 65% in Experiment 2a; 50%, 45% or 35% in Experiment 2b). Participants rated the influence on a scale ranging from negative influence (i.e., I responded opposite to the colour the squares indicated) to positive influence (i.e., I responded according to the colour the squares indicated), with no influence presented in the middle.

Importantly, the sources of information were overall equally informative in all experiments. The sources with 55% and 65% reliability can be labelled as *reliably right*, because these sources will display a colour that is more likely to be the correct than the incorrect one. The sources with 45% and 35% reliability can be labelled as *reliably wrong*, because these sources will display a colour that was more likely to be the incorrect than the correct one. Crucially however, a source of 45% reliability is formally as informative as a 55% reliable source to predict the correct colour, as was a source of 65% reliability compared to a 35% reliable source. The source with 50% reliability is considered *unreliable* however because the colour it predicts is as equally likely to be correct than the incorrect colour – making the opinion it provides uninformative. Participants

were given careful explanations about these three types of reliability levels to make sure they understood what these reliability levels meant (see Methods).

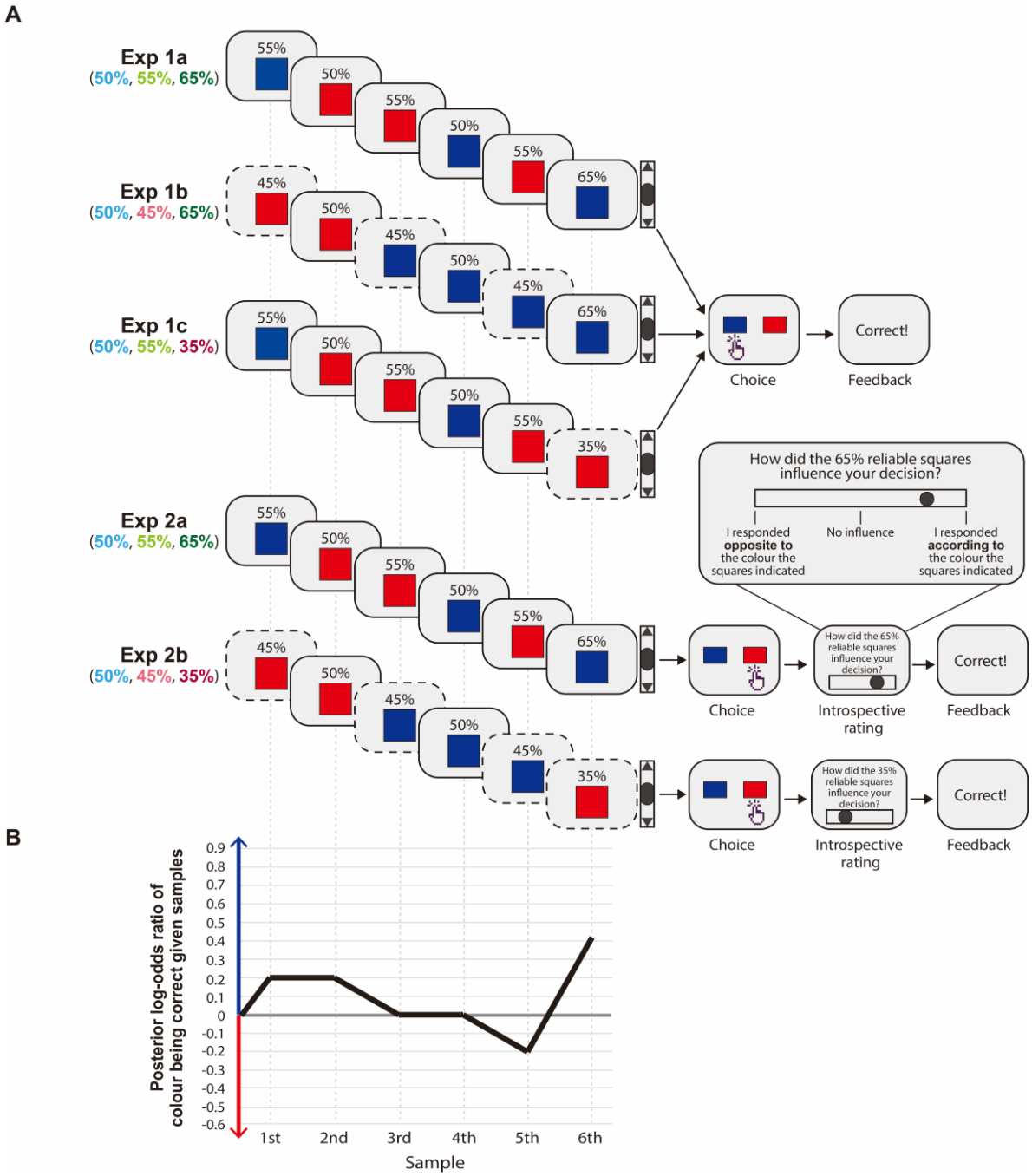


Figure 1. Decision based on information reliability.

(A) Participants scrolled down a webpage to reveal a sequence of six sources, each predicting a colour. The reliability of each piece of information was displayed as a percentage value which corresponded to the likelihood

that the source would show the true webpage colour. In Experiment 1a and 2a, the reliability percentage varied between 50%, 55% and 65%. In Experiment 1b and 2b, sources of 55% reliability were inverted to 45% reliability. In Experiment 1c and 2c, sources of 65% reliability were inverted to 35% reliability. Sources below 50% reliability predicted the opposite colour but with equivalent informativeness. After viewing the predictions, participants were asked to choose which colour they believed would be the correct colour which had generated the sequence. In Experiment 2, after choosing the colour, participants were additionally asked to rate their subjective sense about the degree to which the information sources of a particular reliability level influenced their choice. Participants used a continuous scale to indicate whether they followed the colour the squares indicated or opposed to the colour the squares indicated. In each trial, one of the three levels of information reliability was chosen at random to ask the subjective feeling of influence at that reliability on their choice. **(B)** The predictions of the colour were accumulated from one sample to another. The posterior log-odds ratio between blue being correct and red being correct given samples can be computed. The ideal decision maker chooses blue once the log-odds ratio at the final sample exceeds 0.

We found participants could perform the task with good accuracy overall (grey circles in Figure 2A), being above chance in all experiments (Exp 1a: mean = 64.0%, 99% confidence interval [CI] = [61.7, 66.4]; Exp 1b: mean = 61.6%, 99%CI = [58.3, 64.9]; Exp 1c: mean = 59.9%, 99%CI = [55.4, 64.4]; Exp 2a: mean = 64.4%, 99%CI = [62.4, 66.4]; Exp 2b: mean = 58.2%, 99%CI = [54.6, 61.9]). Accuracy was significant lower however for Experiment 2b, where all the sources of information were reliably wrong, compared to Experiment 2a when information was reliably right (t [62] = 4.43, $p < 0.001$). We observed a trend towards lower accuracy when reliably wrong sources were present (Experiments 1c, 1d and 2b) although it did not reach significance in the first set of experiments (Exp 1a vs. Exp 1b: t [45] = 1.78, p = 0.16; Exp 1a vs. Exp 1c: t [50] = 2.24, p = 0.06). Overall, these results suggest that participants were less able to use information appropriately when it was framed as reliably wrong.

Because the sources presented varied randomly across trials, the decision difficulty fluctuated as well. For each trial sequence, we computed the posterior probability that blue was the correct response given the evidence presented using Bayes rule (see Methods & Figure 1B). An optimal Bayesian observer maximising decision accuracy should choose blue if the final posterior probability is superior to 50% and the corresponding posterior log-odds ratio exceeds 0, and red otherwise. Crucially, a Bayesian optimal observer choice should be influenced by reliably right than reliably wrong sources but should ignore information provided by the unreliable (50%) sources.

We compared the participant's choice with the choice predicted by the Bayesian optimal observer. We found participants made optimal choices in 70-80% of the trials (Exp 1a: mean = 80.3%, 99%CI = [76.0, 84.6]; Exp 1b: mean = 74.6%, 99%CI = [68.1, 81.2]; Exp 1c: mean = 73.0%, 99%CI = [63.7, 82.4], Exp 2a: mean = 84.4%, 99%CI = [80.0, 88.8]; Exp 2b: mean = 69.4.2%, 99%CI = [61.7, 77.2]), showing good performance overall. Crucially, we found

participants made less optimal choices when reliably wrong information was present, both in Experiment 1 (Exp 1a vs. Exp 1b: t [45] = 2.28, p = 0.054; Exp 1a vs. Exp 1c: t [50] = 2.15, p = 0.07) and Experiment 2 (Exp 2a vs. Exp 2b: t [62] = 5.04, p < 0.001), confirming participants struggled to use reliably wrong information optimally.

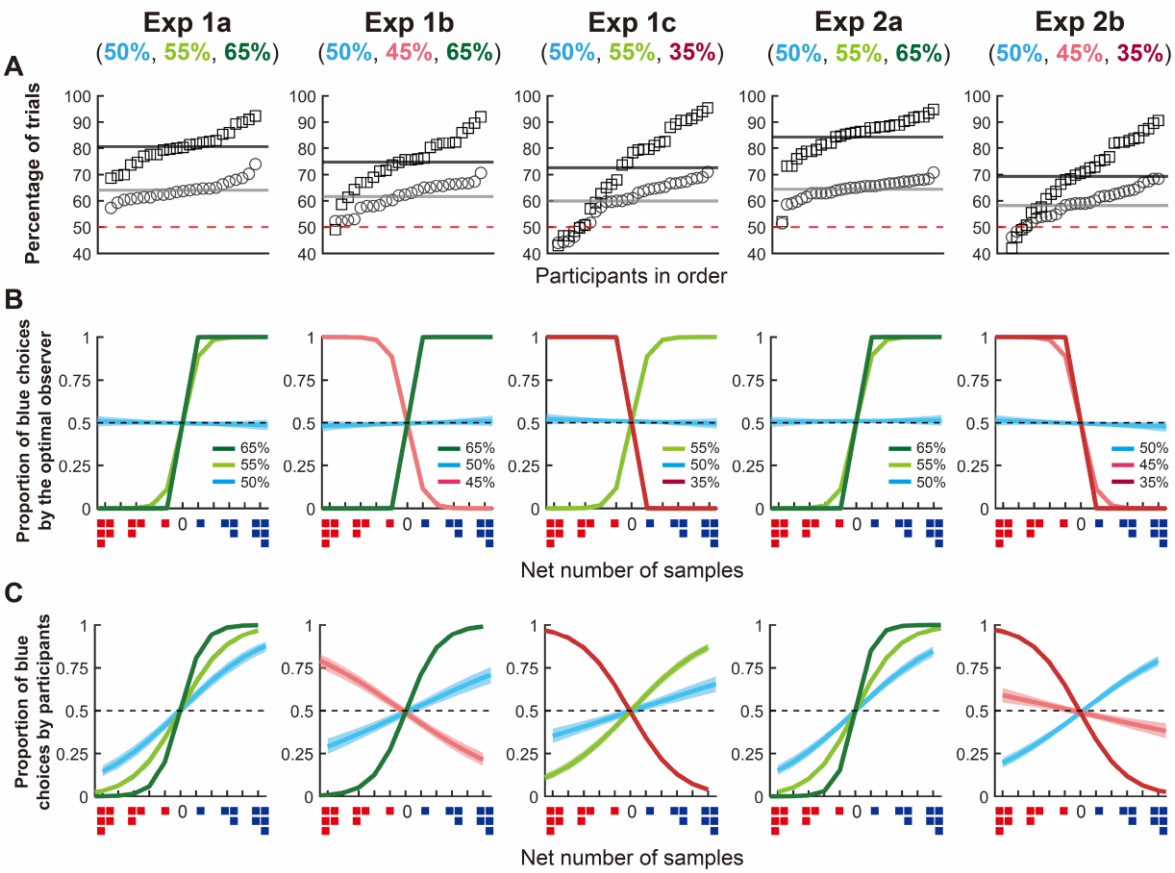


Figure 2. Measuring the influence of information reliability on choice.

(A) Grey circles: Percentage of trials where the participant's choice matched the correct colour. Black squares: Percentage of trials where the participant's choice matched the optimal choice predicted by the optimal Bayesian observer. Percentages are ordered from the lowest to the highest performance. Horizontal lines denote the mean across participants. **(B)** Proportion of choosing the blue option predicted from the optimal choices as a function of the net number of samples for blue or red. Coloured curves denote different levels of information reliability. **(C)** Proportion of choosing the blue option predicted from participants' choice. A positive slope indicates that the proportion of blue choices increases as the net number of blue samples displayed by each level of information reliability increases.

The influence of information on choice is proportional to source reliability

Next, we quantified the extent to which each of the three sources influenced the participant's choices and compared it to the influence predicted by a Bayesian optimal agent. To do so, we computed, for each sequence and for each of source of a given reliability, the net number of red and blue squares it displayed. This can be understood as the amount a source of a given reliability level supports one colour compared to the other (see Methods). Using a logistic regression approach, we probed for each reliability level how much evidence in favour of one colour would influence the choice made by the participant.

Considering first the behaviour or an optimal Bayesian observer (Figure 2B), the proportion of choosing the blue option increased when a given sequence displayed more blue squares of 55% and 65% reliability (Figure 2B: light green curve and dark green curve, respectively). As can be guessed intuitively, a 65% reliability source had the most impact on choice and led to a step-like increase in the proportion of blue choices, with a net difference of only one more coloured sample increasing the chance of choosing that colour to 99.9%.

Participants' behaviour was qualitatively similar, evidence coming from sources of higher reliability leading to a steeper increase in choice likelihood than evidence from less reliable sources. This was the case in Experiment 1a (65% vs. 55%: $z = 16.3$, $p < 0.001$; 55% vs. 50%: $z = 9.1$, $p < 0.001$), in Experiment 1b (65% vs. 45%: $z = 21.8$, $p < 0.001$; 45% vs. 50%: $z = 2.9$, $p < 0.05$) and in Experiment 1c (35% vs. 55%: $z = 9.6$, $p < 0.001$; 55% vs. 50%: $z = 10.7$, $p < 0.001$), as well as in in Experiment 2a (65% vs. 55%: $z = 22.7$, $p < 0.001$; 55% vs. 50%: $z = 11.5$, $p < 0.001$) and in Experiment 2b (35% vs. 45%: $z = 23.5$, $p < 0.001$; 45% vs. 50%: $z = 8.3$, $p < 0.001$). Overall, these results confirmed that participants gave more weight to more reliable sources. However, participants gave less weight than the optimal Bayesian observer ($zs > 40.1$, $ps < 0.001$), suggesting they did not use the reliability indexes optimally.

Choice behaviour is modulated by the presence of reliably wrong source

Considering again the optimal Bayesian observer, we confirmed that reliably wrong sources were treated symmetrically to reliably right ones: the optimal observer now selected the opposite colour to the prediction of the 45% or 35% reliability sources (pink curve or red curve respectively in Figure 2B), the slope of the logistic regression being identical than for reliably right sources but with opposite sign. This captures the fact reliably wrong information is as informative as reliably right information.

Participants, however, differed in that respect compared to an optimal observed. Overall, the slope of logistic regression for reliably wrong sources was less steep than for reliably right

ones, both in Experiment 1 and Experiment 2. Regression coefficients of the 45% reliability source on choice behaviour were significantly smaller in absolute value than positive coefficients of the 55% reliability source (Exp 1a: β = 0.68, 99%CI = [0.60, 0.76] vs Exp 1b: β = -0.25, 99%CI = [-0.32, -0.19], z = 12.9, p < 0.001; Exp 1a vs Exp 1c: β = 0.38, 99%CI = [0.33, 0.43], z = 9.3, p < 0.001; Exp 2a: β = 0.72, 99% CI = [0.65, 0.79] vs Exp 2b: β = -0.08, 99% CI = [-0.13, -0.03], z = 21.7, p < 0.001). Similarly, positive regression coefficients of the 65% reliability source were greater in absolute value than negative regression coefficients of the 35% reliability source (Exp 1a: β = 1.40, 99%CI = [1.29, 1.50] vs Exp 1c: β = -0.63, 99%CI = [-0.69, -0.57], Exp 1a vs. Exp 1c: z = 19.3, p < 0.001; Exp 1b: β = 0.95, 99%CI = [0.88, 1.03] vs Exp 1c, z = 10.3, p < 0.001; Exp 2a: β = 1.73, 99%CI = [1.62, 1.84] vs Exp 2b: β = -0.66, 99%CI = [-0.71, -0.60], z = 25.9, p < 0.001). Therefore, the presence of reliably wrong information reduced influence on choice for human participants, although reliably right information was theoretically as informative as reliably wrong information.

The presence of unreliable information sources biases choice behaviour

According to the Bayesian optimal observer, the colour shown by a 50% reliable source (i.e., unreliable information) should not impact the choice because it is not predictive of the correct response. Indeed, the proportion of blue choices by the simulated optimal observer did not increase with the net number of samples of a given colour displayed by 50% reliability (Figure 2B, blue line). In contrast to this prediction, the net number of 50% reliable samples impacted the choice of participants in both Experiment 1 (Exp 1a: β = 0.36, 99%CI = [0.29, 0.43]; Exp 1b: β = 0.17, 99%CI = [0.11, 0.23]; Exp 1c: β = 0.12, 99%CI = [0.07, 0.17]) and Experiment 2 (Exp 2a: β = 0.34, 99% CI = [0.28, 0.41]; Exp 2b: β = 0.28, 99% CI = [0.23, 0.33]). That is, participants chose the blue option more frequently when one or more unreliable sources (50% reliability) were blue, deviating from the optimal Bayesian observer (zs > 6.4, ps < 0.001). These results demonstrate that, even when unreliable information is explicitly labelled as such, people fail to ignore it and are positively biased towards the content the unreliable source presents.

A computational account for choice behaviour

This first set of analyses suggests a deviation from optimality when using explicit indices of sources' reliability. To shed light on the origin of these suboptimalities, we turned to computational modelling. We considered two potential factors leading to reliability misuse: 1) a distortion in the encoding of the reliability scores and 2) a sequential effect, where decisions are

influenced depending on the order in which information is presented. Figure 3 illustrates how these two sources of errors were modelled and used to predict choice. When presented with a colour square of a given reliability, the computational model first encodes the source's reliability on a log-odds scale and then applies a linear distortion to it (Zhang & Maloney, 2012; Zhang et al., 2020). Note that this is mathematically equivalent to first transforming the probability corresponding to the source's reliability through a non-linear probability distortion function and then mapping this distorted probability to the log-odds scale (Gonzalez & Wu, 1999). Each transformed log-odds is then weighted according to a sequential effect. In Figure 3, we illustrate a recency effect as an example of sequential effects, i.e., more recent information influences decisions more than early information. Note, however, that our model could capture any sequential effects (see Methods). Finally, these weighted log-odds are subsequently summed. The decision is made based on the sign of the sum.

We built four models, including progressively both distorted encoding of the reliability and sequential weighting to determine which of these two sources of errors predicted the choice best (see Methods). The first model included no distortion in the encoding of information reliability and no sequential effect, mimicking the behaviour of an optimal Bayesian observer model (BDT). The second model corresponded to the optimal observer with no distortion in the encoding of the reliability but with sequential weights. The third model corresponded to an observer that would encode the reliability scores with a certain degree of distortion but without sequential effect. The fourth model included the two sources of suboptimality: sequential weights and distortion in the encoding of reliability. We used a hierarchical Bayesian model fit procedure to model trial-wise choice behaviour, confirming the recovery of group-level parameters and individual-level parameters was robust (Supplementary Figure 1 & Supplementary Figure 2).

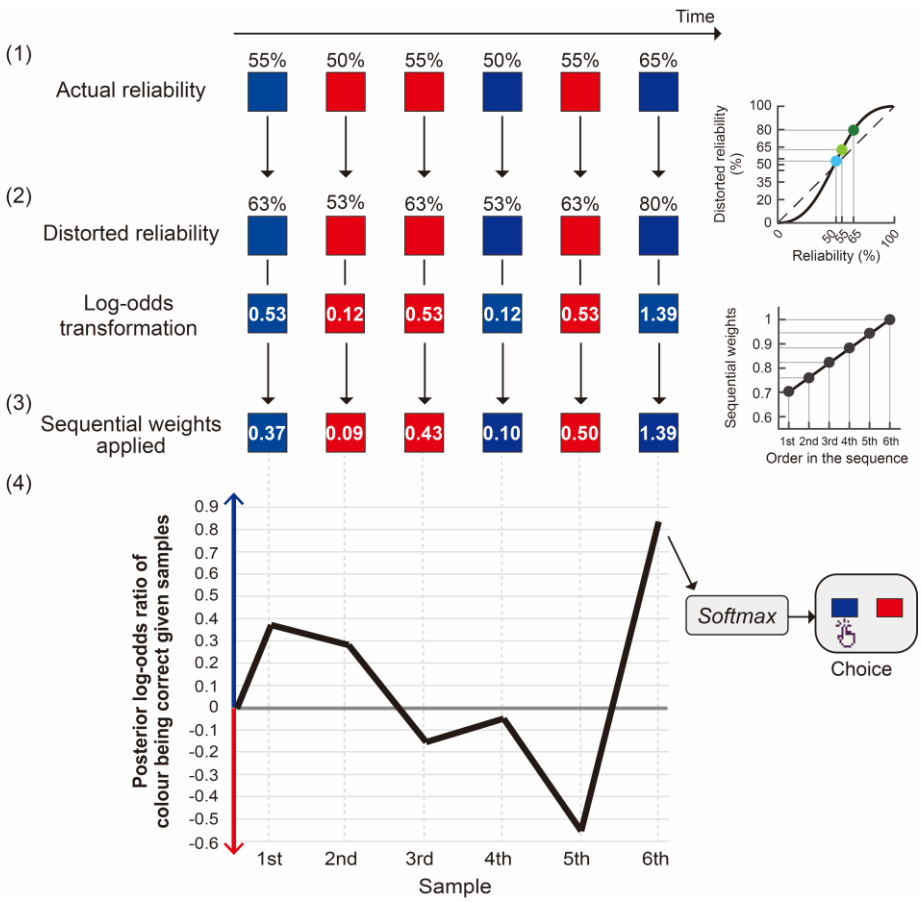


Figure 3. A schematic illustration of computational model.

(1) In each trial, a sequence of six information samples coming from six sources was presented. Each source displayed the prediction of the colour (blue or red) with its reliability score labelled by a percentage above it. (2) The computational model of the choice proposes that the decision maker transforms the log-odds of the objective reliability percentage (Zhang & Maloney, 2012; Zhang et al., 2020), which is equivalent to a biased linear distortion of reliability percentage via non-linear probability distortion function (right panel) (Gonzalez & Wu, 1999). (3) The subjective log-odds is further distorted by sequential effects. Each source of information is weighted depending on the order in which the information source was presented (right panel). (4) Evidence for each colour is accumulated by adding the subjective log-odds of the reliability from one sample to another. After the final sample is presented, a softmax decision rule is applied to the subjective posterior log-odds ratio to determine a choice. The decision maker becomes more likely to choose blue as the log-odds ratio increases.

In all experiments, the BDT model performed the worst to fit the data (Figure 4A). On the opposite, the full model including both sequential weight and distortion of reliability outperformed the other three models. Additionally, the reliability distortion model was significantly better than

the sequential weights model to predict choice, suggesting that participants' choices were better explained by a distorted perception of information reliability than by sequential weights.

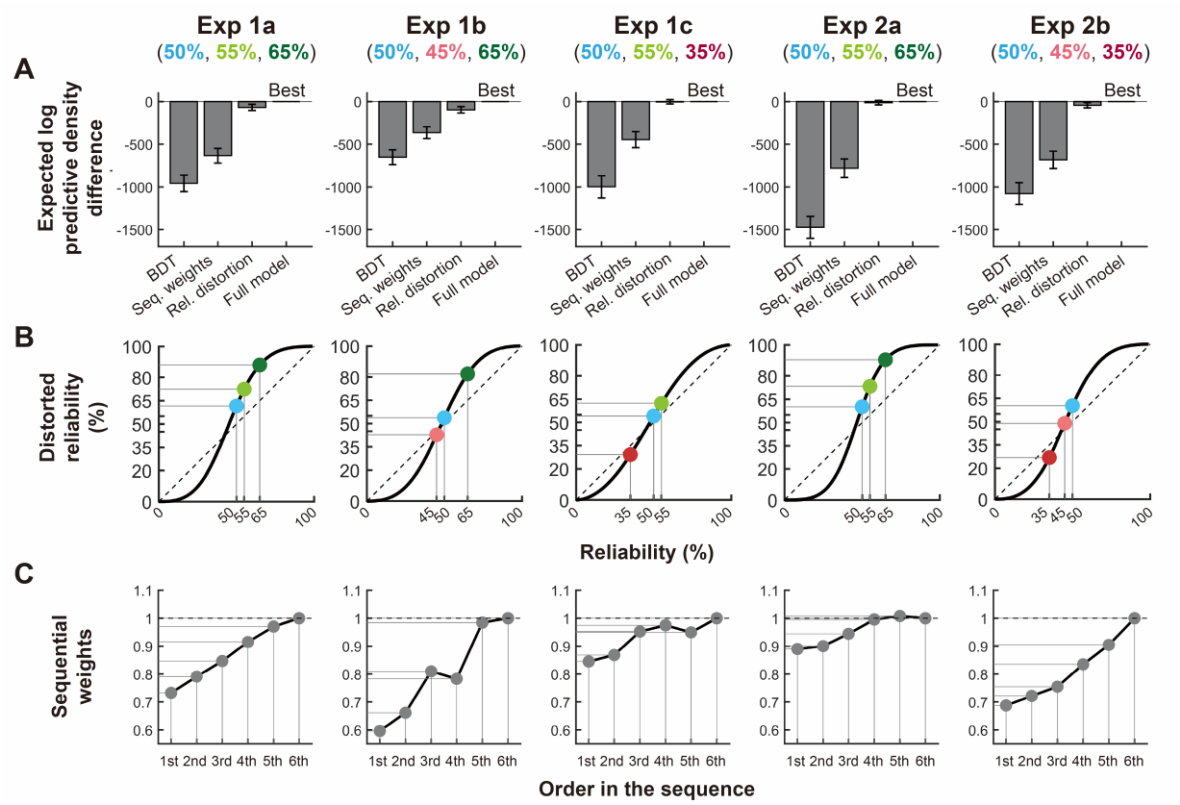


Figure 4. Results of computational analysis.

(A) Expected log pointwise predictive density denotes the out-of-sample predictive accuracy of the model explaining the choice data. The lower the value, the worse the model. The full model which had distortion in reliability and sequential weights on information sources accounted for the data best. **(B)** Distorted reliability percentage recovered from the full model. **(C)** Sequential weights recovered from the full model.

From the full model, we could retrieve, for each information reliability level, the fitted percentages corresponding to the estimates of reliability across participants (Figure 4B, coloured circles). The model predicted that the 50% reliability squares were treated as having above chance reliability in both the first three experiments (61.4% in Exp 1a; 53.7% in Exp 1b; 54.0% in Exp 1c) and the second two experiments (60.1% in Exp 2a; 60.2% in Exp 2b). This pattern of the result is consistent with the previous descriptive result of the positive bias by the colour displayed by the unreliable (50%) information sources.

Further, the participants treated the reliably right and reliably wrong sources as more reliable than they actually were, the sources of 55% reliability (72.4% in Exp 1a; 62.3% in Exp 1b; 73.2% in Exp 2a) and 65% reliability (88.0% in Exp 1a; 82.0% in Exp 1b; 90.4% in Exp 2a) being treated as *more* likely to be correct than they actually were and the 45% reliability (42.7% in Exp 1b) and 35% reliability (29.2% in Exp 1c; 26.8% in Exp 2b) as being more likely to be wrong than they actually were. Regarding the sequential weights, we found that participants gave less weight to information sources presented earlier in a sequence, resembling memory loss or underweighting of prior information when updating a belief with new information. In particular, weights placed on the first source were, on average, 27.5% smaller than the last source in Experiment 1 and 21.1% smaller in Experiment 2.

We found that the parameter capturing the bias by unreliable sources was significantly smaller when a reliably wrong sources was present in Experiment 1 b&c (Exp 1a: Mean parameter value = 0.47, 99%CI = [0.22, 0. 73]; Exp 1b: Mean = 0.15, 99%CI = [-0.01, 0.32]; Exp 1c: Mean = 0.16, 99%CI = [0.06, 0.27]; randomisation tests, Exp 1a vs. Exp 1b: p = 0.008; Exp 1a vs. Exp 1c: p = 0.005; see Supplementary Figure 3) but not in Experiment 2b compared to Experiment 2a (Exp 2a: Mean = 0.41, 99%CI = [0.20, 0.63]; Exp 2b: Mean = 0.42, 99%CI = [0.18, 0.68]; randomisation test, Exp 2a vs. Exp 2b: p = 0.49). This result indicates that the positive bias by unreliable sources was impacted by the presence of reliably wrong sources. When it contributed to the increasing the range of reliability presented (Exp 1b&c), the presence of reliably wrong sources mitigated the effect of unreliable sources on choice but this mitigation was not observed when reliably wrong information sources were present and the range of reliabilities were the same (Exp 2b).

We found that the parameter values for five sequential weights were significantly smaller in Experiment 2b than in Experiment 2a ($ps < 0.037$; see Supplementary Figure 3), suggesting that memory loss was increased when the sources of information were reliably wrong. No significant difference between experiments was found for the other six parameters.

To summarise, a computational analysis suggested that people placed subjective weights on the reliability of sources, even when the reliability scores were given explicitly. Participants tended to consider unreliable sources to be reliably right but this effect was mitigated when a wider range of reliabilities were present. Moreover, participants overweight the last information they received compared to the first piece of information, and this effect seemed to be increased when a greater amount of reliably wrong sources were present.

Influence Report Analysis

In Experiment 2, we asked participants to report how much their decisions were influenced by the evidence provided by each reliability source. We reasoned that if participants were aware of the influence of information reliability on their choice, their subjective report would increase with the reliability of the source considered for the introspective report, as well as with the net number of congruent samples of that source supporting the choice. Therefore, we built a regression model predicting the trial-by-trial ratings of influence using 1) the reliability value of the source chosen for the introspective report and 2) the net number of samples of that reliability level congruent with the choice (see Methods).

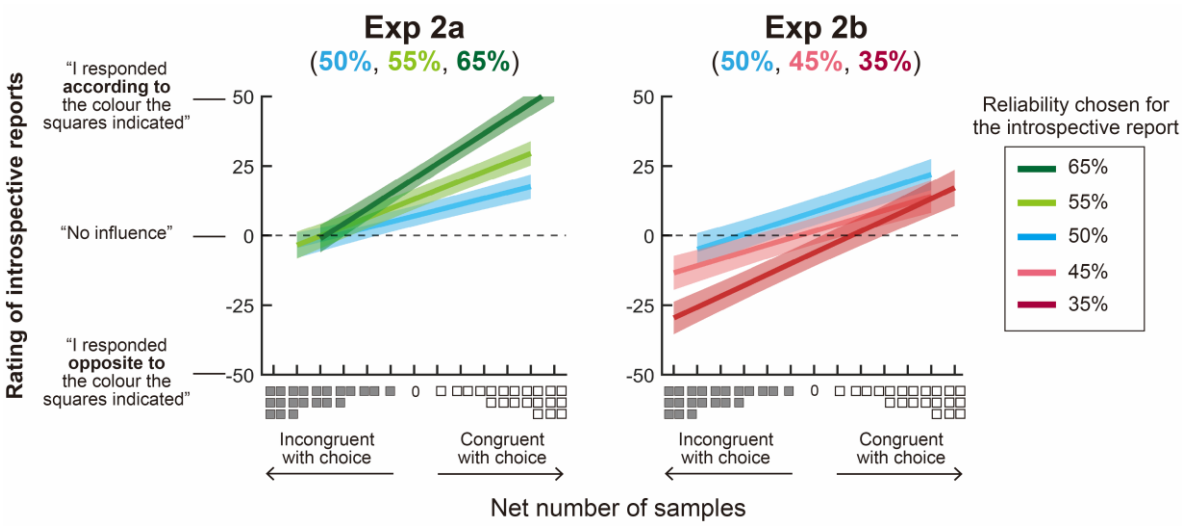


Figure 5. Measuring the subjective feeling of the influence of information reliability on choice. Rating of introspective reports as a function of the net number of samples for congruent or incongruent with the colour the participant chose and of the reliability level that was chosen for the introspective report. A positive slope indicates that the subjective feeling of following the colour increases as the net number of congruent samples displayed by a particular level of information reliability increases.

The ratings of influence increased with the net number of samples congruent with the choice, as shown by the positive regression slope for the 65% (β = 5.40, 99% CI = [4.76, 6.02]) and 55% reliable sources (β = 3.30, 99% CI = [2.73, 3.88]). Similarly, in Experiment 2b, we also observed increased ratings of negative influence when the net number of samples incongruent with the choice increases for the 35% (β = 3.90, 99% CI = [3.26, 4.54]) and 45% reliable sources

(β = 2.50, 99% CI = [1.87, 3.12]). These results suggest that the participant's subjective reports accurately tracked the amount of evidence provided by the source of information chosen for the introspective report. Note that the ratings of influence were not affected by the information sources that were not chosen for the introspective report (e.g., the net number of 55% or 65% sources when introspecting the influence of 50% reliability, see Supplementary Figure 4).

Next, we examined whether increased source reliability was also associated with a stronger sense of being influenced by the source. We tested this idea by comparing between reliability levels the slopes of the correlation of the net number of samples supporting the choice with the subjective influence. We found that a source with stronger reliability had a steeper slope than a source with weaker reliability in Experiment 2a (65% vs. 55%: z = 7.15, p < 0.001; 55% vs. 50%: z = 4.6, p < 0.001) and in Experiment 2b (35% vs. 45%: z = 4.7, p < 0.001; 35% vs. 50%: z = 4.1, p < 0.001), confirming participants were able to report they more influenced by them. No difference in the slopes between the 45% reliability and the 50% reliability was observed however (z = 0.7, p = 0.99).

Considering the unreliable sources, we predicted that if participants remained unaware of the influence of unreliable information on choice (Figure 2C), the ratings for unreliable (50%) information would be independent of the net number of unreliable samples congruent with the choice. In contrast to this prediction, the ratings of influence for the 50% reliable sources increased with the net number of samples congruent with the choice, as reflected by significant positive slope when the 50% reliability was chosen for the introspective report (Exp 2a: β = 2.11, 99% CI = [1.58, 2.64]; Exp 2b: β = 2.70, 99% CI = [2.09, 3.31]). This result suggests that participants could track that unreliable evidence biased their choice. Taken together with the results of choice behaviour, our findings suggest that participants adjusted their subjective feelings of influence to reflect the degree to which their choice was actually influenced by the source of information.

Computational approach to measuring influence

Finally, we explored whether across individuals, the introspective reports reflected how much weight they assigned to information at each reliability level. To do so, we recovered from the best-fitting computational model the reliability that individuals assigned to each of the three levels of information reliability (Figure 4). This distorted reliability is a proxy for the weight the participants gave to that source, thereby quantifying the *objective measure of the influence* of that source on behaviour. On the other hand, we also retrieved for each participant the slope of

the linear regression between the net number of samples and the subjective influence, providing a measure of how influenced participants felt by the amount of evidence displayed by a particular source, thereby quantifying the *subjective sense of the influence* of each given source. For each participant, we correlated the distorted reliability score with the regression slope of subjective influence.

We expected a quadratic relationship between the objective and the subjective measure of influence. This is because, if an individual was aware of the influence of a source on choice, the regression slope of subjective influence should be positive and maximal for sources that are perceived as highly wrong or highly correct. Only for a source with a perceived reliability of 50%, the slope should be null. Figure 6 illustrates this relation for each reliability level across participants. We fitted both a quadratic model ($y = b_1 * x^2 + b_0$) and a constant model ($y = b_0$) to the data points. We found that in Experiment 2a, the quadratic fit was significantly better than the constant model for the 65% reliable source ($F = 8.27$, $p < 0.01$) and the 55% reliability source ($F = 6.00$, $p < 0.05$) but not for the 50% reliability source ($F = 0.95$, $p = 0.34$). In Experiment 2b, we found a better fit by the quadratic model for the 35% reliability source ($F = 8.68$, $p < 0.01$) and the 50% reliability source ($F = 5.06$, $p < 0.05$) but not for the 45% reliability source ($F = 0.27$, $p = 0.61$). Taken together, these results demonstrate the awareness of the influence of information at the individual level: those who assign a strong weight to a source of information are also able to report a stronger sense of being influenced by that source of information. Our findings (Figure 5 & Figure 6) provide substantial evidence that people are aware of the degree to which particular information sources influence their behaviour.

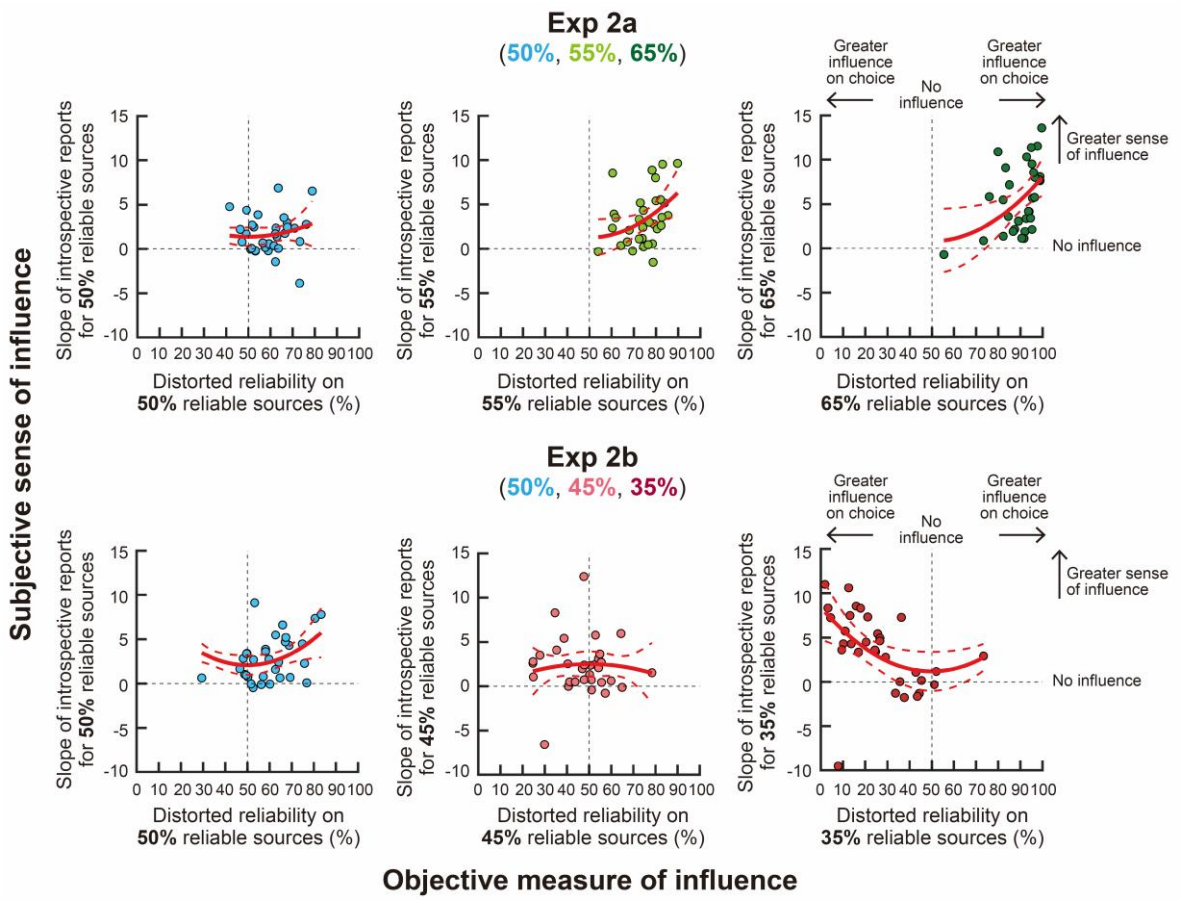


Figure 6. Correlation between the subjective sense of the influence and the objective measure of the influence.

Across participants, the individual's sensitivity (i.e., the slope of regression) to report the influence of a particular level of information reliability on their choice is plotted against the same individual's distorted percentage for that reliability. The distorted reliability percentage provides a proxy of the weight the individual assigns to a particular level of information reliability when making a choice and therefore tells the objective influence on choice. We expected a quadratic relationship between these two measures (see text). In each panel, a solid red curve indicates the quadratic model best fitted to the data points and dashed red curves indicate 95% confidence intervals.

Discussion

In this study, we developed a novel paradigm investigating how people use explicit probabilistic knowledge about the reliability sources of information when making decisions. Participants saw successive pieces of evidence representing the “opinions” of different sources about which one of two possible responses was correct (red or blue). Each sample was associated with a percentage corresponding to the reliability of the source of the information it came from, i.e., the probability that the source would show information congruent with the correct choice. Participants had to combine the evidence (the colour of the squares) with its reliability index (the percentage) to decide which response it supported. After viewing 6 samples, participants had to decide which of the two colours they thought was more likely to be correct. In each experiment, participants were presented with reliably right sources of information that were predictive of the correct colour, or reliably wrong sources that would reliably show the incorrect colour (Experiment 1b-c, Experiment 2b). Crucially, they were also presented with unreliable sources that were as likely to show correct than incorrect information (50% reliability).

We used a computational modelling approach to determine how participants used the reliability information, testing whether their choice reflected an accurate representation of the reliability of the sources and whether they gave similar weights to each sample in the sequence. The first set of experiments revealed that participants did not use explicit reliability optimally, overweighting the reliability of reliably right sources. Participants also struggled to use reliably wrong sources of information, despite these being in theory as informative as reliably right sources. We also found that participants failed to ignore the information provided by the unreliable sources and instead tended to act in accordance with the response suggested by these sources. By asking participants to report how a given source influenced their choice, we could probe their awareness of these influences on their choice. Overall, participants reported a stronger sense of influence when the reliability of the source increased and when the evidence shown was congruent with the choice, suggesting good introspection of the information influencing their decisions. Surprisingly, participants were also aware of being influenced by unreliable sources, suggesting some metacognitive knowledge of their own decision bias.

The question of how humans and other animals use estimates of uncertainty in their environment has been the subject of intense research (Ma & Jazayeri, 2014; Maloney & Mamassian, 2009). Previous studies have shown that people make implicit estimations of uncertainty to maximize reward or information gain (Maloney & Zhang, 2010; Zhang et al., 2020) . Importantly, very few studies (Carlebach & Yeung, 2023; Pescetelli et al., 2021; Pescetelli &

Yeung, 2021) have explored the reverse situation where people have to use explicit probabilistic knowledge about the reliability of sources to make decisions. Research in social and applied psychology provides some insights showing how prior beliefs about the credibility of sources shape judgments, looking for instance at the effect of debunking (Radkani et al., 2024; Swire-Thompson et al., 2021) or at the way people use reviews and ratings to form opinions on products or companies (De Martino et al., 2017; Hoffart et al., 2019; Oktar & Lombrozo, 2025). Importantly, it has been shown that one of the drivers of forming false beliefs include neglecting source cues (Ecker et al., 2022). Nonetheless, even when using explicit cues for reliability, people are known to fall for misinformation. Indeed, banners identifying the source of online news items has been shown to be unsuccessful to decrease belief in fake news (Dias et al., 2020). Even, pre-bunking or de-bunking seem to be insufficient to prevent biasing by misinformation (Ecker et al., 2022). In this context, it is crucial to provide both empirical evidence and theoretical frameworks for understanding why people fail to use correctly explicit knowledge about source trustworthiness and clarify how people use probabilistic information about reliability in the case of simple decisions.

Research in behavioural economics investigated the ability to reason with probabilities. When making decisions under risk for instance, when participants are asked to choose between two lotteries where information about the probability of winning and potential rewards is explicitly described (Hertwig et al., 2018), it has been established that the perception of the likelihood of events is often distorted (Gonzalez & Wu, 1999; Tversky & Kahneman, 1992). In this situation, it is possible to infer the participant's internal odds of winning from their decision patterns. Using this approach, it was revealed that participants overestimate the probability of rare events, while those of frequent events is underestimated. To explain those findings, it has been proposed that probabilities undergo a non-linear weighting function resembling an "inverted S-shaped" when perceived, explaining the overweighting of rare events compared to frequent events (Hertwig et al., 2004; Tversky & Kahneman, 1992; Zhang et al., 2020). In the present paper, we adopted a similar approach to estimate how participants might have distorted the reliability percentages, based on their choices. We found that participants behaved as if they underestimated small reliability percentages (reliably wrong) and overestimated medium-to-large reliability percentages (reliably right), resembling an "S-shaped" function. This suggests that a pattern of probability distortion might extend from distorting the likelihood of an event to distorting the likelihood of information being correct. Interestingly however, the pattern we observed is opposite than the one often reported when making choices on the likelihood of events (Gonzalez & Wu,

1999; Tversky & Kahneman, 1992) where rare events are treated as more likely to occur than they should. Results suggest however that these distortions are not necessarily stable between tasks (Wu et al., 2009) and that in some cases, for instance when probabilities are based on experience, it can lead to the underestimation of rare events (Hertwig et al., 2004). Interestingly, in the context of reliability probabilities, this result suggests that participants tend to overestimate the predictive power of information sources, whether they were reliably right or reliably wrong. Indeed, underestimating the reliability of wrong sources and overestimating reliability of right sources means that participants would overall inflate the likelihood of the sources to give relevant information. Such pattern of distortion might be specific to encoding reliability and reveal a novel type of suboptimality in processing explicit knowledge about source trustworthiness.

Beyond distorting reliabilities, our findings revealed two other types of suboptimal behaviour in participants' decision processes. First, using the regression approach we found that participants were less able to use reliably wrong sources than reliably right sources (Figure 2C), leading to decrease accuracy when those were present (Figure 2A). Our computational model provides two possible interpretations to account for this effect. The first interpretation is that information from reliably wrong sources was encoded less well than those of reliably correct information, affecting the quality of the evidence accumulation process when those were present. Corroborating this idea, we observed an increased recency effect resembling memory decay when reliably wrong information was present in experiment 2b (Figure 4C & Supplementary Figure 3), suggesting participants neglected the information presented by earlier samples compared to later samples. Indeed, reinterpreting the suggestions by reliably wrong sources as evidence in favour of the opposite colour can be thought as cognitively costly and require increased working memory resources affecting the quality of decision-making. A second interpretation is that participants encoded the likelihood of the reliably wrong sources in a more distorted way. This account was unlikely as we did not observe any difference in the distortion parameters when all sources presented were reliably or when reliably wrong ones were present (Figure 4B; see Supplementary Figure 3 for a comparison of model parameter values). The small number of sources reliabilities present in our experiments however makes such effect difficult to detect even if it was present. Further research will be needed to test whether, when a wider range of reliably wrong and reliably present are present, different types of distortions are observed for those types of sources.

Second, participants' choices were biased by the colour suggested by the unreliable information sources. As the number of unreliable samples displaying one colour increased,

participants became more likely to choose that colour. This was true even though participants were explicitly informed that 50% reliability squares were as informative about the correct colour as a coin flip and that they should therefore ignore those. This biasing effect was captured by our computational model, showing that participants assigned a probability slightly higher than 50% to unreliable sources, in the range of 53.7% to 61.5%. How can we explain that participants were biased by unreliable sources? One interpretation of this finding is that participants tended to be biased by the overall colour present in the sequence, independently of the reliability of the squares. Indeed, it has been shown that repeated exposures to a stimulus enhances preference for that stimulus (Zajonc, 1968, 2001), a phenomenon called “the mere exposure effect”. In that context, bias can result from exposure to the colour, similar to a global priming effect by the dominant colour of the sequence. Such interpretation would also explain why an increasing number of unreliable samples biased the choice more. An alternative interpretation is that participants did not actually believe the instruction that unreliable information was truly uninformative for the choice. Indeed, it has been proposed in theoretical work on communication that people assume that communication is cooperative in nature and that the evidence a source provides is by definition informative and as truthful as possible (Grice, 1991). This concept, captured in Grice’s maxims could explain why participants fail to conceive the existence of a random source of information, especially in a context where most sources are reliable. Finally, another possible interpretation of this finding is that participants correctly understood that the information provided by the 50% squares was unreliable and random but used it the same way people flip a coin when they are unsure of how to decide between two options. Indeed, for many sequences presented, the evidence in favour of both options could be ambiguous, one response being favoured only marginally over the other. In this context of uncertainty, it is then possible participants relied on unreliable information to break the tie and choose a response based on the random evidence it displayed.

In any case, we also observed that the biasing effect by unreliable sources of information was less pronounced when both reliably wrong and reliably right sources were intermixed (Experiment 1b&c) compared to when all sources were either reliably right or reliably wrong. It has been proposed that differences in distortion in different contexts could happen due to the normalization of the probabilities according to the range of probabilities presented (Zhang et al., 2020). According to this view, people first transform the probability to log odds so that they can be combined by addition or subtraction rather than multiplication. Then crucially, the log odd values are normalized according to the range of values occurring in the task. This mechanism

could explain how the distortion function is stronger when the range of probabilities presented is narrower. Considering the small range of probabilities considered in the present study, it was not possible to test this hypothesis explicitly and further research will be needed to confirm whether varying systematically the range of reliabilities presented also affects how reliabilities are distorted. If confirmed however, such findings could explain why we observed reduced distortion and less biasing effect by unreliable information when the overall range of probabilities was higher.

Most studies of metacognition have focussed on how participants form beliefs about their confidence in their choice in perceptual, memory or motor tasks (Fleming, 2024). A standard approach to measure the accuracy of confidence is to compare *subjective* confidence judgements with the accuracy of *objective* behavioural performance, assuming that high confidence is coupled with an accurate operation of a given task (Fleming & Lau, 2014). In the present study, we developed a similar approach, asking participants the extent to which they felt influenced by a given source of information and comparing this *subjective* sense of influence to an *objective* measure of the influence of that source on choice. By doing so, it allowed us to provide a quantitative measure of the metacognitive sensitivity in detecting influence in choice. We first confirmed that the strength of sensory cue (in our case, the three levels of source reliability) had an increased correlation with both the objective performance of participant's choice (Figure 2) and the subjective report of influence (Figure 5). Across participants, we correlated the individual's objective influence of information source on choice with the same individual's subjective sense of influence (Figure 6). Indeed, we found a significant relationship between them: those who were strongly influenced by sources of a particular level of reliability also had a stronger sense of being influenced, compared to those who were less influenced. Taken together, these findings suggest that people have a good metacognitive ability to detect influence and reflect on how their decisions are made.

Surprisingly, we also found that participants were aware that their responses were biased by unreliable information. Indeed, when asked to report the influence of the squares of 50% reliability on their choice, participants confirmed that their decision was positively influenced by the colour the unreliable sources supported. Crucially, participants did not simply report a constant influence but accurately reported increased influence when the number of unreliable sources favouring their choice increased. Taken together, these results suggest that participants were aware of being biased by explicit cues of 50% reliability, confirming they had some degree of metacognitive access to the factors influencing their choice. One may wonder why participants

still followed the colour suggestions from unreliable sources, despite they were aware of such a positive bias. One possibility is that this finding reflects a form of metacognitive knowledge rather than metacognitive control. It has been suggested that the metacognitive processes rely on a map between monitoring one's own behavioural performance and control of behaviour (Koriat, 2006; Son & Schwartz, 2002). In theory, people should use their metacognitive knowledge to regulate their behaviour and exert a form of metacognitive control (Boldt & Gilbert, 2022; Risko & Gilbert, 2016) but dissociation between these two processes exists (Jiwa et al., 2023). An alternative explanation is that, as suggested above, participants knowingly used the 50% reliable squares to break the tie between options when reliable information about the correct response was sparse. This would be consistent with the increased report of influence when more unreliable squares were present.

Importantly, these findings seem in contradiction with some social psychological studies showing that people perform poorly when attempting to detect their own biases or reasons for choice (Epstein & Robertson, 2015; Greenwald et al., 1998; Pailhes & Kuhn, 2020; Sidarus et al., 2013). For instance, people often remain unaware that their behaviour is manipulated by search engine rankings (Epstein & Robertson, 2015) or person's speech and gestures (Pailhes & Kuhn, 2020, 2021). People also tend to evaluate themselves as less subject to various biases and more objective than others (Armor, 1999; Ehrlinger et al., 2005; Hansen et al., 2014; Pronin et al., 2002; Schwalbe et al., 2020). Indeed, our other research has revealed that people suffer from metacognitive blind spots when trying to understand the reasons for their choice, mistaking autonomy for acting contrarian (Kummen et al., 2023). Our design is, of course, different from the design of these studies and these differences might explain the distinct pattern of results here. While the previous studies use cues or manipulations which implicitly guide action choices (Epstein & Robertson, 2015; Greenwald et al., 1998; Pailhes & Kuhn, 2020; Sidarus et al., 2013), our task presented explicit and descriptive cues of source reliability. This difference seems important to raise a conscious awareness about drivers of actions.

Overall, the present study is paving the way for a new exploration of how people use explicit indicators of information trustworthiness and their ability to introspect their influence on decisions. In everyday life, explicit, descriptive information about the trustworthiness of information has become more and more available. Most online shopping services now provide a star rating review indicating customers' global agreement on how good products are (Hoffart et al., 2019; Oktar & Lombrozo, 2025). Similarly, social networking services have started to rely heavily on fact-checking indicators in the hope to help reduce the influence of disinformation

(Andersen & Søe, 2019; Gaozhao, 2020) and their dissemination (Fazio et al., 2013; Rapp, 2008; Rapp & Salovich, 2018; Vosoughi et al., 2018). Our experimental work and computational modelling bring new light on the fundamental cognitive mechanisms underlying the use of reliability in decision-making, showing that people still overly trust information even when explicit trustworthiness is available.

Methods

Participants

Ninety-four participants were recruited on the online platform Prolific (https://www.prolific.co/), to participate in three decision-making experiments: experiment 1a (20 female, 8 male, 2 other, Mean age = 27.2, SD = 4.9), experiment 1b (18 female, 12 male, 0 other, Mean age = 24.6, SD = 5.2), experiment 1c (24 female, 9 male, 1 other, Mean age = 30.0, SD = 4.6). We recruited 64 participants for our two metacognitive experiments: experiment 2a (22 female, 9 male, 1 other, Mean age = 26.3, SD = 5.2) and experiment 2b (25 female, 6 male, 1 other, Mean age = 28.6, SD = 4.7). Recruitment was restricted to the United Kingdom. All participants were fluent English speakers and had no history of neurological disorders. Participants received a basic payment of £8 for their participation in a 60-minute experiment. Participants who took less than half the expected time to complete the task (i.e., 30 mins) were excluded from the analysis as they were considered not to have engaged with the instructions properly. This exclusion left final samples of 23 participants for experiment 1a, 24 for experiment 1b and 29 for experiment 1c. No participants were excluded from the analysis in experiments 2. All participants gave informed consent. The procedures were approved by the Research Ethics Committee of University College London (ID ICN-PH-PWB-22-11-18A).

Experimental design

Experiment 1

Apparatus. The online task was programmed using jsPsych (de Leeuw, 2015) and the experiment was hosted on the online research platform Gorilla (https://gorilla.sc/) (Anwyl-Irvine et al., 2020).

Stimuli and task. On each trial, a colour (blue or red) was chosen at random as the correct colour for that trial. The participants' task was to guess which colour was more likely to be correct. To help participants guess, we generated six *information sources* which predicted the correct colour. Each piece of evidence provided by the sources was a red or blue square (200x200 pixels) accompanied with a percentage that indicated the likelihood of the source to give accurate information about the correct colour. That is, each coloured square was associated with an explicit cue indicating the reliability score of the source (i.e., reliability of the evidence). Reliability scores varied between three levels in each experiment: 50%, 55% and 65% in Experiment 1a; 50%, 45% and 65% in Experiment 1b; 50%, 55% and 35% in Experiment 1c. To generate the

stimulus sequence, the reliability score of the source was fist chosen at random among the three levels. Then the colour of the square was drawn from a Bernoulli distribution with a probability of choosing the correct colour equal to the reliability score. For instance, if the correct colour was blue and a 65% reliable source had been chosen to be displayed, the colour of the square was generated at random, with a 65% chance of it being blue and a 35% of it being red. The reliability score (in this case 65%) was labelled above the centre of the square. Therefore, each information source explicitly signalled the probability that the displayed colour is the correct one (i.e., the likelihood of the correct colour). To generate the stimuli sequence, this process was repeated six times, generating six information samples shown to the participant. As a consequence of the stochastic generative process, the strength of evidence supporting a correct colour varied across trials. Crucially, it could be that the information would by chance support the incorrect response, although this would be a rare occurrence.

Participants were provided instructions about information reliability scores as follows. First, participants were informed that information sources above 50% reliability are *reliably right* because the displayed colour predicts the correct colour. In particular, participants were told that, for a blue square with 90% reliability, "*According to this square, the page colour should be blue. The source is 90% reliable, meaning there is a 90% chance that this piece of evidence shows the correct page colour*". Second, participants were explicitly told information sources with 50% reliability were uninformative because they are as likely to give the correct as the incorrect colour and are considered *unreliable information*. Participants were told that "*A blue square with 50% reliability indicates the page colour is blue, but it is only 50% reliable.50% reliability means this square gives you random information, as reliable as if the square's colour had been determined by flipping a coin*". Finally, participants were explicitly told that information sources below 50% reliability are informative but should be interpreted as evidence in favour of the opposite colour than they display and therefore are considered *reliably wrong* information. We provided the following explanation: "*A blue square with a 25% reliability is only 25% reliable. That's even less accurate than flipping a coin. Less than 50% reliability means the evidence provided by this square is reliably wrong. That is to say: if there is a 25% chance the square colour is correct, then there is a 75% chance that it is incorrect. So, while this square indicates the page colour is blue, likely it is actually red*".

On each trial, participants progressively scrolled down the page at their own pace using their mouse to reveal new sources and their predictions of the colour. In total, participants could see six information samples before they made a response (Figure 1A). At the bottom of the page,

two buttons corresponding to blue and red appeared on the left and right of the screen. Participants were asked to click a button to indicate which colour they believed would be more likely to be correct. The position of the blue-choice button and the red-choice button was randomised across participants. No time limit was imposed, and participants were encouraged to make their guesses as accurately as possible. Participants were given immediate feedback on whether their guess was correct or wrong. They proceeded to the next trial at their own pace by clicking the “next” button on the screen.

Each of the three experiments consisted of 6 blocks of 50 trials (300 trials in total). Participants could take a short break in between blocks as long as the duration of the experiment did not exceed 3 hours. A short version of our task will be available to play online.

Experiment 2

The apparatus, stimuli and task in the two introspective experiments were the same as in the first three experiments with the following differences: information reliability scores now varied between 50%, 55% and 65% in Experiment 2a and between 50%, 45% and 35% in Experiment 2b. We included one unreliable source and two reliably wrong sources in Experiment 2b to avoid potential contaminated effects by the presence of reliably right information. In addition to this change, we included a subjective estimate question in each trial. Immediately after guessing the colours, participants were now asked to provide a subjective rating to report the degree to which they felt they were influenced by the squares of a given reliability level (Figure 1A). In each trial, one information reliability level (e.g., 65%) was randomly selected and the following question appeared on the screen: “How did the 65% reliable squares influence your decision?”. Participants rated their subjective sense of being influenced by the squares of that reliability level on a continuous scale ranging from “I responded opposite to the colour the squares indicated” (-50) to “No influence” (0) to “I responded according to the colour the squares indicated” (+50). Participants were instructed to use the full range of the scale and to make a rating uniquely based on the decision they just made. More specifically, we provided the following instructions “*After each response, we will ask you to evaluate how the squares of a given reliability level influenced the choice you just made. You will see a scale with a slider like the one below. You will have to indicate for your most recent choice to what extent you felt you followed the colour shown by those squares or chose the opposite colour of what those squares were suggesting.*” No time limit was imposed. A rating score below 0 indicated a negative influence by the colour of information reliability, a score of zero indicated total independence and a score

above 0 indicated a positive influence. Each of the two experiments consisted of 8 blocks of 36 trials (288 trials in total). Participants could take a short break in between blocks.

Data analysis

Models of optimal choice and participant's choice. Let C be the actual colour of the stimulus in the current trial, which can be either equal to B (blue) or R (red). We assume that $\Pr(C = B) = \Pr(C = R)$. The participants were instructed that both colours are equiprobable. Each information sample the participant receives can be seen as a signal s, providing a colour $c(s)$ and a probability $\pi(s)$. For instance, the sample "Blue with reliability 55%" is defined by $c(s) = B$ and $\pi(s) = 0.55$. We denote by S the set of all pieces of information presented to the participant in a given trial, and by $S_B = \{s \in S | c(s) = B\}$ and $S_R = \{s \in S | c(s) = R\}$ the pieces of information with blue and red colours, respectively. To illustrate, assume that the participant observes three samples: "Blue with reliability 55%", "Red with reliability 50%" and "Blue with reliability 60%". We then have $S = \{s_1, s_2, s_3\}$, with $c(s_1) = B, \pi(s_1) = 0.55$, $c(s_2) = R, \pi(s_2) = 0.5$, $c(s_3) = B, \pi(s_3) = 0.60$. In that example, the set of Blue samples is $S_B = \{s_1, s_3\}$ and the set of Red samples is $S_R = \{s_2\}$. Note that, by definition, for all pieces of information s:

$$\Pr(C = c(s)|s) = \pi(s) \tag{1}$$

The posterior ratio corresponding to the stimulus colour given the information $(s_1, \dots, s_6)$ is then:

$$\begin{aligned} \frac{\Pr(C = B|s_1, \dots, s_6)}{\Pr(C = R|s_1, \dots, s_6)} &= \prod_{s \in S} \frac{\Pr(C = B|s)}{\Pr(C = R|s)} \quad \text{(because the pieces of information are independent)} \\ &= \left(\prod_{s \in S_B} \frac{\pi(s)}{1 - \pi(s)}\right)\left(\prod_{s \in S_R} \frac{1 - \pi(s)}{\pi(s)}\right) \end{aligned} \tag{2}$$

where the last equality follows from equation (1). We thus have the following posterior log-odds

$$L(s_1, \dots, s_6) = \sum_{s \in S_B} \log\left(\frac{\pi(s)}{1 - \pi(s)}\right) - \sum_{s \in S_R} \log\left(\frac{\pi(s)}{1 - \pi(s)}\right) \tag{3}$$

An optimal Bayesian observer would thus choose Blue whenever $L(s_1, \dots, s_6) > 0$, and Red otherwise. See Figure 1B.

The decision maker might be, however, biased. A simple way to model such a bias is to assume that she linearly distorts the log-odds corresponding to each piece of information (Zhang

& Maloney, 2012; Zhang et al., 2020). In other words, she would compute the following subjective log-odd ratio:

$$\tilde{L}(s_1, \dots, s_6, a, b) = \sum_{s \in S_B} \left(a\log\left(\frac{\pi(s)}{1-\pi(s)}\right) + b \right) - \sum_{s \in S_R} \left(a\log\left(\frac{\pi(s)}{1-\pi(s)}\right) + b \right) \quad (4)$$

where parameter $a \geq 0$ reflects the participant's sensitivity to the probabilistic information, and b is interpreted as a measure of her "presentation bias", i.e., to which extent she is influenced by the mere presentation of a given colour (Figure 3). The Bayesian optimal corresponds to $a = 1$ and $b = 0$. If, for instance, $a = 0$ and $b > 0$, the participant does not take at all into account the reliability of the information, and simply counts the number of samples corresponding to each colour. Crucially, it turns that $\tilde{L}$ can equivalently be written as:

$$\tilde{L}(s_1, \dots, s_6, a, b) = \sum_{s \in S_B} \log\left(\frac{\varphi(\pi(s), a, b)}{1-\varphi(\pi(s), a, b)}\right) - \sum_{s \in S_R} \log\left(\frac{\varphi(\pi(s), a, b)}{1-\varphi(\pi(s), a, b)}\right) \quad (5)$$

where:

$$\varphi(\pi(s), a, b) = \frac{\gamma\pi(s)^a}{\gamma\pi(s)^a + \left(1-\pi(s)\right)^a} \quad (6)$$

and $\gamma = e^b$. The function $\varphi(\pi(s), a, b)$ is interpreted as a subjective probability distortion function (Gonzalez & Wu, 1999). In other words, the biased log-odd can equivalently be seen as an unbiased log-odd applied to distorted probabilities or a biased log-odd applied to actual probabilities. In turn, distorted probability can be interpreted as the weight the participant gave to each information reliability level in the decision process.

The subjective log-odd might be further distorted by a sequential weight, implying that each piece of information is weighted by its order in the presentation sequence (Figure 3). To formalize this idea, let $\rho(s)$ denotes the rank order of the piece of information s (in other words, $\rho(s_i) = i$). We further define a subject-dependent weight function $w > 0$, that assigns a non-negative weight to each piece of information, and we normalize it by requiring $w(6) = 1$. We then obtain:

$$\tilde{L}(s_1, \dots, s_6, a, b, w) = \sum_{s \in S_B} w\left(\rho(s)\right) \left(a\log\left(\frac{\pi(s)}{1-\pi(s)}\right) + b \right) - \sum_{s \in S_R} w\left(\rho(s)\right) \left(a\log\left(\frac{\pi(s)}{1-\pi(s)}\right) + b \right) \quad (7)$$

If, for instance, $w(1) < 1$, the participant reduces a weight on the first sample relative to the last sample, resembling a memory leak or a recency effect. If, on the other hand, $w(1) > 1$, the participant increases a weight on the first sample relative to the last sample, resembling a primacy effect.

Finally, we model the fact that the decision process might be noisy by assuming that the actual choice follows a softmax rule (Figure 3). Thus, the decision maker chooses Blue with probability:

$$p(blue) = \frac{1}{1 + e^{-\tilde{L}(s_1,\ldots,s_6,a,b,w)}} \tag{8}$$

In a softmax rule, an inverse temperature parameter tunes decision stochasticity. We fixed this parameter to 1 due to a problem of parameter identifiability. If we found a difference in the values of a or b parameter between the experiments, it may indicate that two groups of participants were different in sensitivity to the probabilistic information or in presentation bias. However, because of this parameter constraint, we cannot rule out the possibility that two groups of participants were different in decision stochasticity.

We tested the assumption of reliability distortion and the assumption of sequential weights by a factorial model comparison (van den Berg et al., 2014). The first model was the BDT model, representing the behaviour of an optimal Bayesian observer who does not distort the reliability probabilities and has a perfect memory capacity (no free parameter). This is equivalent to constraining $a = 1$, $b = 0$, and $w(p(s)) = 1$. In the second model, we introduced sequential weights solely, keeping a = 1 and b = 0, allowing only the following set $\Theta = \{w(1), w(2), w(3), w(4), w(5)\}$ of free parameters to change. Conversely, the third model introduced distortion in reliability solely, using a set of free parameters $\Theta = \{a, b\}$. In the last, full model, we introduced both sequential weights and distortion in reliability, using a full set of free parameters $\Theta = \{a, b, w(1), w(2), w(3), w(4), w(5)\}$.

For each model, we fitted the model probability of choosing blue to the participant's choice data using a Bayesian hierarchical modelling approach. The four models were fitted using Hamiltonian Monte Carlo sampling as a Markov Chain Monte Carlo method (MCMC) in Stan (Carpenter et al., 2017; StanDevelopmentTeam., 2023) and RStan (version 2.32.6). MCMC approximates the posterior distribution of the free parameters of the model. We estimated the means of group parameters, the variances of group parameters and the individual parameters for each participant. The parameter a was constrained between 0 and 20, while sequential

weight parameters were constrained between 0 and 2. No constraint was imposed on b. Models were fit from four parallel chains with 2,000 warm-up samples, followed by 2,000 samples drawn from converged chains. Model comparison was performed using the *loo* package in R, which uses a version of the leave-one-out estimate that was optimized using Pareto smoothed importance sampling (PSIS) (Vehtari et al., 2017). *loo* estimates the expected log pointwise predictive density (elpd) without one data point using posterior simulations. This index is a measure of the out-of-sample predictive accuracy, i.e., how well the entire data set without one data point predicts this excluded point. PSIS-loo is sensitive to overfitting: neither more complex models nor simpler models should be preferred by the inference criterion (Danwitz et al., 2022). For each experiment, the last 2,000 MCMC posterior samples of the means of the seven group parameters in the full model can be seen in Supplementary Figure 3.

Measuring the influence of information reliability on choice (Experiment 1 and 2). We performed a regression analysis to estimate the influence of three levels of explicit information reliability cue on the participant's binary action choices. The participant's choice would be ultimately determined by an overall difference between the evidence that favoured the blue option and the evidence that favoured the red option in a given sequence. However, we could estimate how each of the three information reliabilities partially affected the choice using a regression approach. To compute the difference in the evidence values, we counted, for each sequence of six sources, how many blue sources and red sources were shown respectively by each information reliability level. We then computed the net number of samples in favour of each colour by calculating the difference between blue and red for each information reliability level. For instance, if a participant had the following sequence: 65% blue, 55% blue, 55% red, 55% red, 50% blue and 50% red, 65% reliability had a net value of +1 blue source and 0 red source, 55% reliability had a net value +1 blue source and +2 red sources, 50% reliability had a net value of +1 for each colour. Therefore, the net number displayed at 65% reliability is +1 for blue, +1 for red at 55% reliability and 0 at 50% reliability. A net number of sources between blue and red was counted separately for 50%, 55% and 65% in Experiment 1a, 50%, 45% and 65% in Experiment 1b,50%, 55% and 35% in Experiment 1c, 50%, 55% and 65% in Experiment 2a and 50%, 45% and 35% in Experiment 2b. We performed the logistic regression to predict the participants' binary choices, with running separately for the first three decision-making experiments and the second two metacognition experiments. We used a categorical variable *exp* to represent the experiment (1a, 1b or 1b in the first experiment; 2a or 2b in the second experiment). We then coded *source 1* as the net number of sources with 50% reliability, *source 2* as the net number of sources with either

55% reliability (in Exp 1a, Exp 1c and Exp 2a) or 45% reliability (in Exp 1b and Exp 2b) and *source 3* as the net number of sources with either 65% reliability (in Exp 1a, Exp 1b and Exp 2a) or 35% reliability (in Exp 1c and Exp 2b). Finally, for each set of the experiments (Experiment 1a-1c or Experiment 2a-2b), we used "lme4" package (Bates et al., 2015) to perform mixed-effects logistic regression with the following formula: glmer (*choices* ~ (1 | *ID*) + *exp* × *source 1* + *exp* × *source 2* + *exp* × *source 3*). The intercept varied between participants *ID* as a random effect. See Figure 2B&C.

Measuring the subjective influence of information reliability (Experiment 2 only). A regression approach allows us to characterise discrepancies between human and optimal performance (Dal Martello et al., 2023; Ota et al., 2024; Ota & Maloney, 2024). We performed a linear regression analysis to estimate how the rating of introspective reports was correlated to the level of the reliability of information. We reasoned that the strength of the introspective rating would be determined by the information reliability per se as well as by the net number of squares present in that trial for the squares of the reliability level, and by the direction in which it influenced the decision (following or opposing the colour of sources). Suppose the participant was presented with two blue squares of 65% reliability, one red square of 65% reliability and three red squares of 50% reliability. If the participant chose blue and then was asked to report the influence of 65% reliability squares on their choice, the participant would report that they followed the colour displayed by the 65% squares (i.e., positive influence) because their blue choice was likely supported by two blue squares of 65% reliability. In another example, suppose the following sequence was presented: two blue squares of 35% reliability, one red square of 35% reliability and three red squares of 50% reliability. If the participant chose red and then was asked to report the influence of 35% reliability squares on their choice, the participant would report that they opposed the colour displayed by the 35% squares (i.e., negative influence) because their red choice was likely supported by two blue squares of 35% reliability.

Therefore, for each level of information reliability in a given sequence of six sources, we counted 1) how many sources were *congruent* with the colour chosen by the participant and 2) how many sources were *incongruent* with the chosen colour. We then computed a net difference between the congruent and incongruent sources at each reliability level. For instance, if a participant had the following sequence: 50% red square, 55% blue square, 55% blue square, 65% red square, 65% blue square and 50% red square, then the 65% reliability displayed 1 blue source and 1 red source, 55% reliability displayed 2 blue sources and 50% reliability displayed

2 red sources. If the participant chose blue, this means that two 55% reliable sources were congruent with the colour the participant chose, and two 50% reliable sources were incongruent with the chosen colour. In such a trial, a net number was 0 at 65% reliability, 2 for congruent sources at 55% reliability and 2 for incongruent sources at 50% reliability. Therefore, a net number between congruent and incongruent sources was 0 at 65% reliability, 2 at 55% reliability and -2 at 50% reliability. As such a net number between congruent and incongruent sources was counted separately for 50%, 55% and 65% in Experiment 2a and 50%, 45% and 35% in Experiment 2b. We performed the linear regression to predict the participants' introspective reports, separately for Experiment 2a and Experiment 2b. We used a categorical variable *level* to represent which reliability was chosen for the introspective report for a given trial (50%, 55% or 65% in Exp 2a; 50%, 45%, 35% in Exp 2b). We then coded *source 1* as the net number between congruent and incongruent sources with 50% reliability, *source 2* as the net number with either 55% reliability (in Exp 2) or 45% reliability (in Exp 2b) and *source 3* as the net number with either 65% reliability (in Exp 2a) or 35% reliability (in Exp 2b). Finally, for experiment, we performed mixed-effects linear regression with the following formula: lmer (*reports* ~ (1 | *ID*) + *level* × *source 1* + *level* × *source 2* + *level* × *source 3*). The intercept varied between participants *ID* as a random effect. See Figure 5 and Supplementary Figure 4.

Competing interests

The authors declare no competing interests.

Author's contributions

Conceptualization, PH & LC; data collection, AC & LC; formal analysis, KO, TGP & LC; Methodology, AC, PH & LC; Software, AC & LC; Project administration, PH & LC; supervision PH, TGP & AC; visualization, KO, TGP & LC; writing – original draft, KO, TGP & LC; writing – review & editing, KO, AC, PH, TGP & LC; funding acquisition, LC and PH.

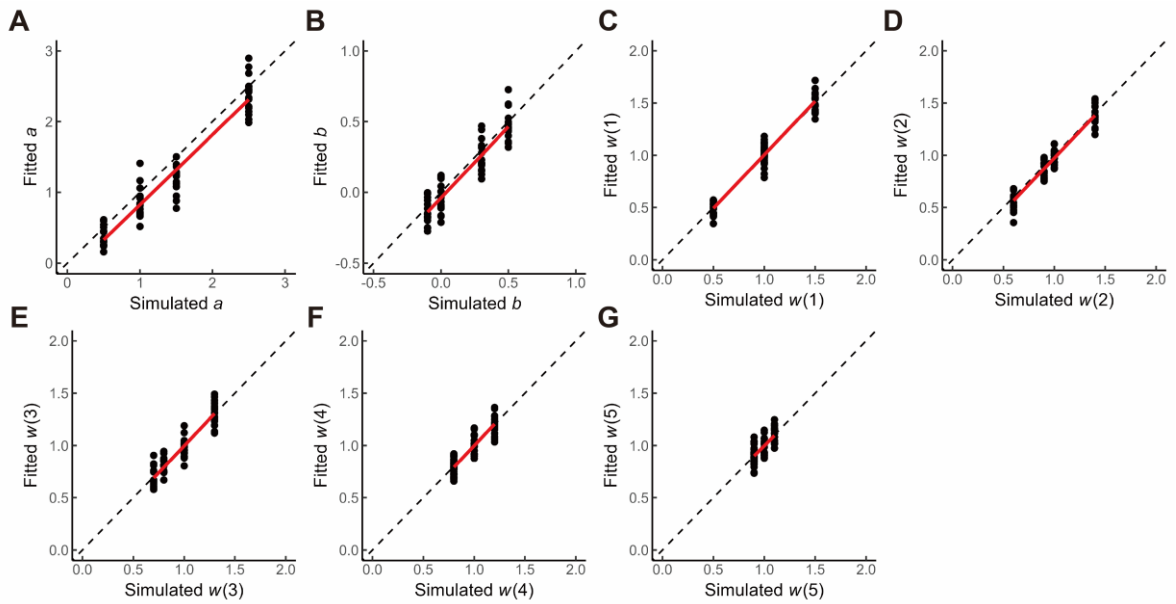
Acknowledgments

This work was supported by an ESRC grant ES/V00378X/1 awarded to LC and PH.

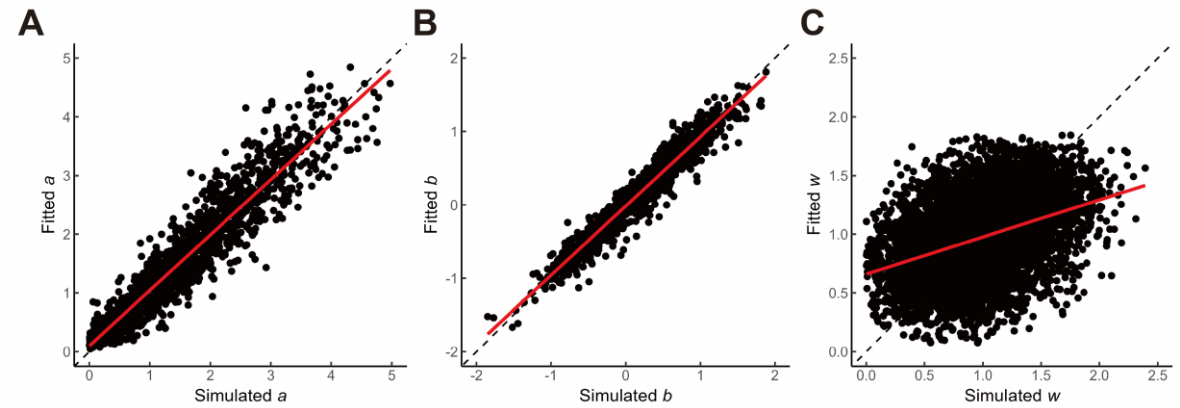
Data and code availability

The data frame for a representative participant will be posted. The data frames for all the participants and the codes used to generate the figures will be posted.

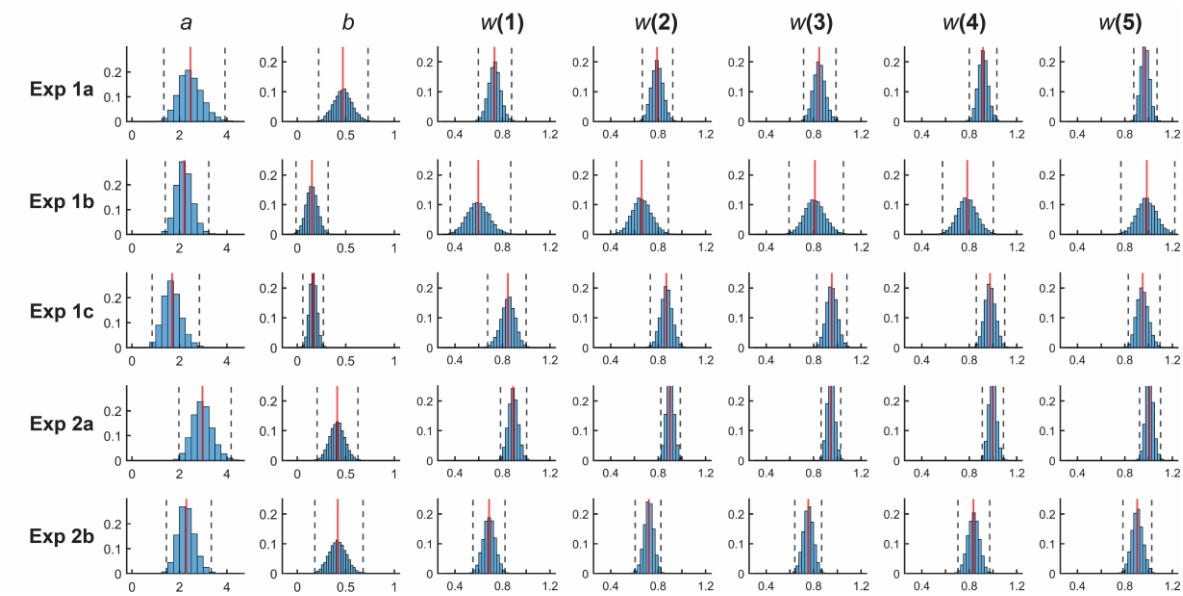
Supplementary Information



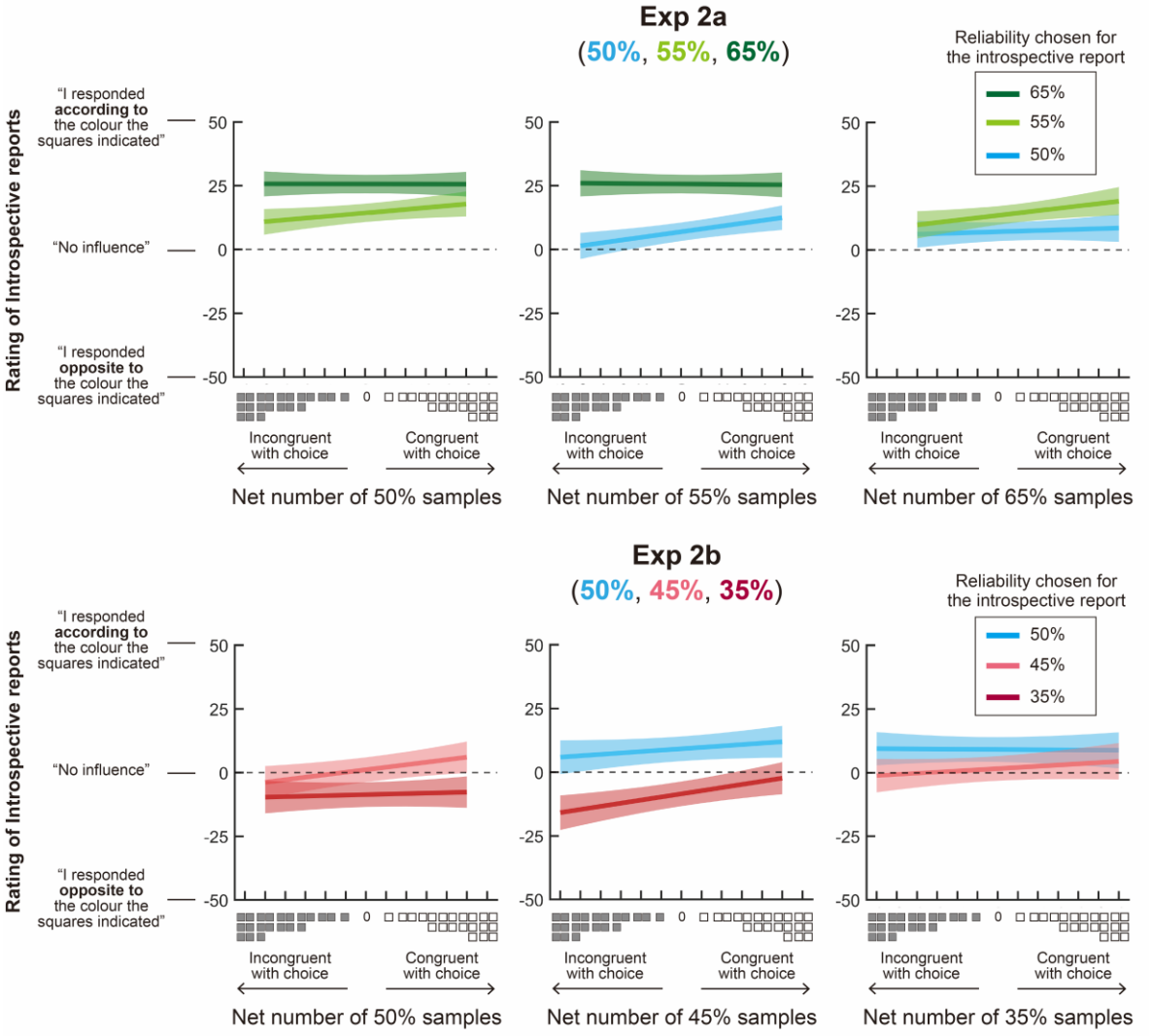
Supplementary Figure 1. Parameter recovery of means of group-level parameters. **A.** Recovery of the parameter for sensitivity to the probabilistic information. **B.** Recovery of the parameter for presentation bias. **C-G.** Recovery of sequential weights. Scatter plots show the correlation of the simulated values (i.e., the input values to the simulation) and the recovered values obtained in the fitting process. Black dashed lines show diagonal lines for a perfect parameter recovery. Red lines show the linear regression lines.



Supplementary Figure 2. Parameter recovery of individual-level parameters. **A.** Recovery of the parameter for sensitivity to the probabilistic information. **B.** Recovery of the parameter for presentation bias. **C.** Recovery of sequential weights. Scatter plots show the correlation of the simulated values (i.e., the input values to the simulation) and the recovered values obtained in the fitting process. Black dashed lines show diagonal lines for a perfect parameter recovery. Red lines show the linear regression lines.



Supplementary Figure 3. MCMC posterior samples of means of group-level parameters. We plotted the MCMC samples of the means of the seven group-level parameters from a Bayesian hierarchical model fit in the full model. Each histogram provides a proxy of the posterior distribution of the parameter value. Red lines denote the average of the posterior distribution while dashed lines denote the 99% confidence intervals.



Supplementary Figure 4. Marginal effects on the introspective reports.

See the caption in Figure 5. A flat slope of the regression indicates that the reliability sources that were not chosen for the introspective reports (x-axis) have little effect on the subjective feeling when reporting the influence of the chosen level of reliability on choice (coloured lines).

References

Andersen, J., & Søe, S. O. (2019). Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news – the case of Facebook. *European Journal of Communication*, *35*(2), 126-139. https://doi.org/10.1177/0267323119894489

Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behav Res Methods*, *52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Armor, D. (1999). The illusion of objectivity: A bias in the perception of freedom from bias. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *59*(5163).

Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Sci Adv*, *8*(18), eabl3844. https://doi.org/10.1126/sciadv.abl3844

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-1085. https://doi.org/10.1126/science.1185718

Bang, D., Fusaroli, R., Tylen, K., Olsen, K., Latham, P. E., Lau, J. Y., Roepstorff, A., Rees, G., Frith, C. D., & Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Conscious Cogn*, *26*(100), 13-23. https://doi.org/10.1016/j.concog.2014.02.002

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Boldt, A., & Gilbert, S. J. (2022). Partially Overlapping Neural Correlates of Metacognitive Monitoring and Metacognitive Control. *J Neurosci*, *42*(17), 3622-3635. https://doi.org/10.1523/JNEUROSCI.1326-21.2022

Carlebach, N., & Yeung, N. (2023). Flexible use of confidence to guide advice requests. *Cognition*, *230*, 105264. https://doi.org/10.1016/j.cognition.2022.105264

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *J Stat Softw*, *76*. https://doi.org/10.18637/jss.v076.i01

Cash, T. N., & Oppenheimer, D. M. (2024a). Assessing metacognitive knowledge in subjective decisions: The knowledge of weights paradigm. *Thinking & Reasoning*, 1-43. https://doi.org/10.1080/13546783.2024.2426543

Cash, T. N., & Oppenheimer, D. M. (2024b). Parental rights or parental wrongs: Parents' metacognitive knowledge of the factors that influence their school choice decisions. *PLoS One*, *19*(4), e0301768. https://doi.org/10.1371/journal.pone.0301768

Celadin, T., Capraro, V., Pennycook, G., & Rand, D. G. (2023). Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts. *Journal of Online Trust and Safety*, *1*(5). https://doi.org/10.54501/jots.v1i5.100

Charles, L., & Haggard, P. (2020). Feeling free: External influences on endogenous behaviour. *Q J Exp Psychol (Hove)*, *73*(4), 568-577. https://doi.org/10.1177/1747021819888066

Charles, L., King, J. R., & Dehaene, S. (2014). Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. *J Neurosci*, *34*(4), 1158-1170. https://doi.org/10.1523/JNEUROSCI.2465-13.2014

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80-94. https://doi.org/10.1016/j.neuroimage.2013.01.054

Charles, L., & Yeung, N. (2019). Dynamic sources of evidence supporting confidence judgments and error detection. *J Exp Psychol Hum Percept Perform*, *45*(1), 39-52. https://doi.org/10.1037/xhp0000583

Dal Martello, M. F., Ota, K., Pietralla, D. E., & Maloney, L. T. (2023). Detecting visual texture patterns in binary sequences through pattern features. *J Vis*, *23*(13), 1. https://doi.org/10.1167/jov.23.13.1

Danwitz, L., Mathar, D., Smith, E., Tuzsus, D., & Peters, J. (2022). Parameter and Model Recovery of Reinforcement Learning Models for Restless Bandit Problems. *Computational Brain & Behavior*, *5*(4), 547-563. https://doi.org/10.1007/s42113-022-00139-0

de Leeuw, J. R. (2015). jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behav Res Methods*, *47*(1), 1-12. https://doi.org/10.3758/s13428-014-0458-y

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. *J Neurosci*, *37*(25), 6066-6074. https://doi.org/10.1523/JNEUROSCI.3880-16.2017

Deneve, S., & Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *J Physiol Paris*, *98*(1-3), 249-258. https://doi.org/10.1016/j.jphysparis.2004.03.011

Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-001

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29. https://doi.org/10.1038/s44159-021-00006-y

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: people's assessments of bias in themselves and others. *Pers Soc Psychol Bull*, *31*(5), 680-692. https://doi.org/10.1177/0146167204271570

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proc Natl Acad Sci U S A*, *112*(33), E4512-4521. https://doi.org/10.1073/pnas.1419828112

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433. https://doi.org/10.1038/415429a

Fazio, L. K., Barber, S. J., Rajaram, S., Ornstein, P. A., & Marsh, E. J. (2013). Creating illusions of knowledge: learning errors that contradict prior knowledge. *J Exp Psychol Gen*, *142*(1), 1-5. https://doi.org/10.1037/a0028649

Fleming, S. M. (2024). Metacognition and Confidence: A Review and Synthesis. *Annu Rev Psychol*, *75*, 241-268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philos Trans R Soc Lond B Biol Sci*, *367*(1594), 1338-1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Front Hum Neurosci*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Gaozhao, D. (2020). Flagging Fake News on Social Media: An Experimental Study of Media Consumers' Identification of Fake News. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3669375

Glöckner, A., & Betsch, T. (2023). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, *3*(3), 215-228. https://doi.org/10.1017/s1930297500002424

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cogn Psychol*, *38*(1), 129-166. https://doi.org/10.1006/cogp.1998.0710

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol*, *74*(6), 1464-1480. https://doi.org/10.1037//0022-3514.74.6.1464

Grice, P. (1991). *Studies in the Way of Words*. Harvard University Press.

Hansen, K., Gerbasi, M., Todorov, A., Kruse, E., & Pronin, E. (2014). People Claim Objectivity After Knowingly Using Biased Strategies. *Pers Soc Psychol Bull*, *40*(6), 691-699. https://doi.org/10.1177/0146167214523476

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychol Sci*, *15*(8), 534-539. https://doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Hogarth, R. M., & Lejarraga, T. (2018). Experience and Description: Exploring Two Paths to Knowledge. *Current Directions in Psychological Science*, *27*(2), 123-128. https://doi.org/10.1177/0963721417740645

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nat Commun*, *8*(1), 2191. https://doi.org/10.1038/s41467-017-02314-5

Hoffart, J. C., Olschewski, S., & Rieskamp, J. (2019). Reaching for the star ratings: A Bayesian-inspired account of how people use consumer ratings. *Journal of Economic Psychology*, *72*, 99-116. https://doi.org/10.1016/j.joep.2019.02.008

Jiwa, M., Yu, C., Boonyaratvej, J., Ciston, A., Haggard, P., Charles, L., & Bode, S. (2023). Exposure to misleading and unreliable information reduces active information-seeking. *Preprint at https://osf.io/preprints/psyarxiv/4zkxw*. https://doi.org/10.31234/osf.io/4zkxw

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2). https://doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341-350. https://doi.org/10.1037/0003-066x.39.4.341

Katyal, S., & Fleming, S. M. (2024). The future of metacognition research: Balancing construct breadth with measurement rigor. *Cortex*, *171*, 223-234. https://doi.org/10.1016/j.cortex.2023.11.002

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*, *27*(12), 712-719. https://doi.org/10.1016/j.tins.2004.10.007

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One*, *2*(9), e943. https://doi.org/10.1371/journal.pone.0000943

Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244-247. https://doi.org/10.1038/nature02169

Koriat, A. (2006). *Metacognition and consciousness*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511816789.012

Kummen, A., Haggard, P., Williams, G., & Charles, L. (2023). Mistaking opposition for autonomy: psychophysical studies on detecting choice bias. *Proc Biol Sci*, *290*(1996), 20221785. https://doi.org/10.1098/rspb.2022.1785

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annu Rev Neurosci*, *37*, 205-220. https://doi.org/10.1146/annurev-neuro-071013-014017

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: testing Bayesian transfer. *Vis Neurosci*, *26*(1), 147-155. https://doi.org/10.1017/S0952523808080905

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Res*, *50*(23), 2362-2374. https://doi.org/10.1016/j.visres.2010.09.031

Martins, A. C. R. (2006). Probability biases as Bayesian inference. *Judgment and Decision Making*, *1*(2), 108-117. https://doi.org/10.1017/s1930297500002321

Meyniel, F. (2020). Brain dynamics for confidence-weighted learning. *PLoS Comput Biol*, *16*(6), e1007935. https://doi.org/10.1371/journal.pcbi.1007935

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc Natl Acad Sci U S A*, *114*(19), E3859-E3868. https://doi.org/10.1073/pnas.1615773114

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Pers Soc Psychol Bull*, *31*(2), 166-180. https://doi.org/10.1177/0146167204271418

Oktar, K., & Lombrozo, T. (2025). How aggregated opinions shape beliefs. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-024-00398-7

Ota, K., Charles, L., & Haggard, P. (2024). Autonomous behaviour and the limits of human volition. *Cognition*, *244*, 105684. https://doi.org/10.1016/j.cognition.2023.105684

Ota, K., & Maloney, L. T. (2024). Dissecting Bayes: Using influence measures to test normative use of probability density information derived from a sample. *PLoS Comput Biol*, *20*(5), e1011999. https://doi.org/10.1371/journal.pcbi.1011999

Ota, K., Shinya, M., & Kudo, K. (2015). Motor planning under temporal uncertainty is suboptimal when the gain function is asymmetric. *Front Comput Neurosci*, *9*, 88. https://doi.org/10.3389/fncom.2015.00088

Ota, K., Shinya, M., & Kudo, K. (2016). Sub-optimality in motor planning is retained throughout 9 days practice of 2250 trials. *Sci Rep*, *6*, 37181. https://doi.org/10.1038/srep37181

Ota, K., Shinya, M., Maloney, L. T., & Kudo, K. (2019). Sub-optimality in motor planning is not improved by explicit observation of motor uncertainty. *Sci Rep*, *9*(1), 14850. https://doi.org/10.1038/s41598-019-50901-x

Ota, K., Tanae, M., Ishii, K., & Takiyama, K. (2020). Optimizing motor decision-making through competition with opponents. *Sci Rep*, *10*(1), 950. https://doi.org/10.1038/s41598-019-56659-6

Pailhes, A., & Kuhn, G. (2020). Influencing choices with conversational primes: How a magic trick unconsciously influences card choices. *Proc Natl Acad Sci U S A*, *117*(30), 17675-17679. https://doi.org/10.1073/pnas.2000682117

Pailhes, A., & Kuhn, G. (2021). Subtly encouraging more deliberate decisions: using a forcing technique and population stereotype to investigate free will. *Psychol Res*, *85*(4), 1380-1390. https://doi.org/10.1007/s00426-020-01350-z

Pescetelli, N., Hauperich, A. K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition*, *215*, 104810. https://doi.org/10.1016/j.cognition.2021.104810

Pescetelli, N., & Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. *J Exp Psychol Gen*, *150*(3), 507-526. https://doi.org/10.1037/xge0000960

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends Cogn Sci*, *11*(1), 37-43. https://doi.org/10.1016/j.tics.2006.11.001

Pronin, E., Lin, D. Y., & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality and Social Psychology Bulletin*, *28*(3), 369-381. https://doi.org/10.1177/0146167202286008

Radkani, S., Landau-Wells, M., & Saxe, R. (2024). How rational inference about authority debunking can curtail, sustain, or spread belief polarization. *PNAS Nexus*, *3*(10), pgae393. https://doi.org/10.1093/pnasnexus/pgae393

Rapp, D. N. (2008). How do readers handle incorrect information during reading? *Mem Cognit*, *36*(3), 688-701. https://doi.org/10.3758/mc.36.3.688

Rapp, D. N., & Salovich, N. A. (2018). Can't We Just Disregard Fake News? The Consequences of Exposure to Inaccurate Information. *Policy Insights from the Behavioral and Brain Sciences*, *5*(2), 232-239. https://doi.org/10.1177/2372732218785193

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends Cogn Sci*, *20*(9), 676-688. https://doi.org/10.1016/j.tics.2016.07.002

Schwalbe, M. C., Cohen, G. L., & Ross, L. D. (2020). The objectivity illusion and voter polarization in the 2016 presidential election. *Proc Natl Acad Sci U S A*, *117*(35), 21218-21229. https://doi.org/10.1073/pnas.1912301117

Sidarus, N., Chambon, V., & Haggard, P. (2013). Priming of actions increases sense of control over unexpected outcomes. *Conscious Cogn*, *22*(4), 1403-1411. https://doi.org/10.1016/j.concog.2013.09.008

Son, L. K., & Schwartz, B. L. (2002). The relation between metacognitive monitoring and control. In *Applied Metacognition* (pp. 15-38). https://doi.org/10.1017/cbo9780511489976.003

StanDevelopmentTeam. (2023). Stan Modeling Language: User's Guide and Reference Manual. http://mc-stan.org/manual.html

Swire-Thompson, B., Cook, J., Butler, L. H., Sanderson, J. A., Lewandowsky, S., & Ecker, U. K. H. (2021). Correction format has a limited role when debunking misinformation. *Cogn Res Princ Implic*, *6*(1), 83. https://doi.org/10.1186/s41235-021-00346-6

Tanae, M., Ota, K., & Takiyama, K. (2021). Competition Rather Than Observation and Cooperation Facilitates Optimal Motor Planning. *Frontiers in Sports and Active Living*, *3*. https://doi.org/10.3389/fspor.2021.637225

Trommershauser, J., Gepshtein, S., Maloney, L. T., Landy, M. S., & Banks, M. S. (2005). Optimal compensation for changes in task-relevant movement variability. *J Neurosci*, *25*(31), 7169-7178. https://doi.org/10.1523/JNEUROSCI.1906-05.2005

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297-323.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychol Sci*, *20*(4), 473-479. https://doi.org/10.1111/j.1467-9280.2009.02319.x

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychol Rev*, *121*(1), 124-149. https://doi.org/10.1037/a0035234

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432. https://doi.org/10.1007/s11222-016-9696-4

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Walker, E. Y., Pohl, S., Denison, R. N., Barack, D. L., Lee, J., Block, N., Ma, W. J., & Meyniel, F. (2023). Studying the neural representations of uncertainty. *Nat Neurosci*, *26*(11), 1857-1867. https://doi.org/10.1038/s41593-023-01444-y

Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control over effects of action. *Cognition*, *115*(1), 26-38. https://doi.org/10.1016/j.cognition.2009.10.016

Wu, S. W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making compared with an equivalent motor task. *Proc Natl Acad Sci U S A*, *106*(15), 6088-6093. https://doi.org/10.1073/pnas.0900102106

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philos Trans R Soc Lond B Biol Sci*, *367*(1594), 1310-1321. https://doi.org/10.1098/rstb.2011.0416

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2), 1-27. https://doi.org/10.1037/h0025848

Zajonc, R. B. (2001). Mere Exposure: A Gateway to the Subliminal. *Current Directions in Psychological Science*, *10*(6), 224-228. https://doi.org/10.1111/1467-8721.00154

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front Neurosci*, *6*, 1. https://doi.org/10.3389/fnins.2012.00001

Zhang, H., Ren, X., & Maloney, L. T. (2020). The bounded rationality of probability distortion. *Proc Natl Acad Sci U S A*, *117*(36), 22024-22034. https://doi.org/10.1073/pnas.1922401117