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

Keiji Ota[1,2], Thibault Gajdos[3], Anthony Ciston[4], Patrick Haggard[1], 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] Centre de Recherche en Psychologie et Neurosciences, Centre national de la recherche scientifique

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

Corresponding author

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

School of Biological and Behavioural Sciences

Queen Mary University of London

Abstract

In everyday life, decision-makers need to integrate information from various sources which differ in how reliable the correct information is provided. The present study addressed whether people can optimally use explicit information about the trustworthiness of the information source and whether people know how information sources affect their decisions. In each trial, we presented a sequence of six information sources which indicated a correct colour (red or blue) each with a different level of reliability. Each source was explicitly labelled the reliability percentage, i.e., the likelihood that a source provides correct information. Participants were asked to decide which colour was more likely to be correct. In the first series of three experiments, we found that participants failed to make use of explicit reliability cues optimally. In particular, participants were less able to use reliably wrong information sources (reliability below 50%) than reliably correct information sources (reliability above 50%), even though these two types of information sources were equally informative. In addition, participants failed to ignore the colour displayed by unreliable sources (50% reliability), although these sources gave just random information for a binary decision. In the second series of two experiments, we asked participants, after each choice, to report their subjective feelings of whether they followed or opposed a colour suggested by sources with a specific reliability percentage. We found that the ratings of the participants' influence report tracked the amount of evidence which supported the choice they just made. Further, participants were able to introspect their own choice bias guided by unreliable information sources. These findings suggest that human choice behaviour deviates from Bayesian integration. However, people have a good metacognitive monitoring of how their decisions are driven by external stimuli.

Keywords

Evidence accumulation, Bayesian decision theory, metacognition, introspection

Introduction

In most of our everyday decisions, one challenge is the need to integrate information from a range of sources of varying reliability to make the most informed choice. When a parent for instance decide whether to vaccinate their child against the flu, they might consult their doctor, read online articles about the efficiency of the vaccine or ask a friend's opinion. Crucially, their beliefs about the reliability of these sources are key to pondering the different advice: parents may trust their doctor's or friend's advice but doubt information from news sources. While these beliefs might stem from their own assessment, i.e. whether these sources have provided accurate and trustworthy information in the past, they may also have been informed by explicit ratings. For instance, they may have read that doctors are 90% accurate in providing medical advice, been told that their friend is 80% certain that this vaccine is safe or heard that a particular media source spreads false information in 50% of their articles. Crucially, these explicit probabilities about how trustworthy sources of information should be integrated into decision-making, allowing one to ponder the advice they provide. However, can people make use of this information about source reliability optimally? And are they able to provide an account of how source reliability influences their decisions? In the present study, we investigated whether people can use explicit probabilistic knowledge about information reliability to make optimal decisions, assessing both the ability to use reliability and introspect and its influence on choice.

Bayesian decision theory (BDT) provides a mathematical model of how information reliability should be combined with evidence to make decisions (Berger, 1985; Maloney & Zhang, 2010). The BDT has been shown to capture well human behaviour when combining information from different sensory modalities (Ernst & Banks, 2002) or when operating visuomotor control (Kording & Wolpert, 2004; Ota et al., 2020; Tanae et al., 2021; Trommershauser et al., 2005). These studies demonstrate that information is weighted according to the precision of its source, the decision being ultimately influenced more by accurate than noisy signals. Importantly, in most of these studies, experimenters manipulate the quality of the evidence and let participants build their internal beliefs about the reliability of sources through feedback. However, such an implicit inference can be slow and insufficient, particularly in complex tasks, which leads to inaccurate beliefs about uncertainty (Acerbi et al., 2012; Zhang et al., 2015).

Importantly, humans can also form beliefs about the likelihood of information to be true by being told explicitly how likely it is to tell the truth. Indeed, previous research indicates that people make decisions using explicit probabilistic knowledge, for instance when provided information about the likelihood of receiving rewards (Kahneman & Tversky, 1984). Interestingly,

it has been shown that people often struggle to make optimal choices when reasoning with 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). Results suggest that these distortions are not necessarily stable between tasks (Wu et al., 2009) however, showing, for instance, that estimation of probabilities based on experience can lead to the underestimation of rare events (Hertwig et al., 2004). One explanation for these discrepancies is that distortion could be adapted according to the range of probabilities presented (Zhang & Maloney, 2012; Zhang et al., 2020) leading to specific biases and suboptimalities depending on the context. However, it is less clear whether suboptimal choices arise when probabilities are not about the likelihood of an event occurring but about the reliability of providing correct information. In other words, is explicit probabilistic knowledge about the trustworthiness of the information source subject to specific decision biases?

Crucially, the human cognitive system forms probabilistic beliefs about both the external world and its own functions (Koriat, 2006; Son & Schwartz, 2002). Indeed, recent research in the field of metacognition has shown that people introspect the accuracy of their own cognitive processes (Fleming & Dolan, 2012; Katyal & Fleming, 2024) and detect their errors (Charles et al., 2014; Charles et al., 2013; Charles & Yeung, 2019; Yeung & Summerfield, 2012). Beyond simply tracking whether their choice is correct or not, people actually form probabilistic beliefs about their decisions, computing on the level of evidence supporting the decisions they made and the likelihood of that 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). Indeed, people are able to use other people’s confidence when incorporating advice (Carlebach & Yeung, 2023; Pescetelli et al., 2021; Pescetelli & Yeung, 2021) and when making decisions with another person (Bahrami et al., 2010; Hertz et al., 2017). This research starts to shed light on how people interpret the reliability of the decisions of others when making joint decisions. However, it remains unclear how people use reliability estimates that come from external sources.

More importantly beyond understanding decisions, the question of whether people can accurately introspect 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). People 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). However, it remains unclear whether people are aware of how information reliability affects their choices.

In the present study, we have developed a novel paradigm that mimics combining information from sources of variable trustworthiness. 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 the percentage. First, reliability scores that are above 50% mean that the source is more likely to give correct information and is therefore *reliably correct*. For these sources, Bayesian decision theory allows us to compute the weight it should be given according to its reliability score. Second, a reliability score equal to 50% means that the source is as likely to give correct information as incorrect information. Such a source was categorised as *unreliable*. Importantly, Bayesian decision theory predicts that unreliable sources should be ignored as they are not predictive of the correct response, being as informative as a coin flip. Filtering out unreliable information is likely to be cognitively demanding however (Jiwa et al., 2023; Vidal-Perez et al., 2024) and we hypothesized that it might still bias choice. Indeed, preliminary results in a similar task indicate that the mere presence of unreliable sources decreases choice accuracy and also reduces choice confidence. 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 requires participants to reinterpret its prediction as evidence against the content it

displayed. BDT also predicts that reliably wrong sources should be treated as equally informative as reliably correct sources, as the information they provide allows us to make accurate predictions about the correct response. Yet, counterfactual thinking is known to be cognitively demanding (Van Hoeck et al., 2015). Indeed, previous results suggest that people struggle to use reliably wrong information, its mere presence being associated with a decrease in accuracy and confidence (Jiwa et al., 2023; Vidal-Perez et al., 2024).

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 test this, we developed a computational model of participant choices, hypothesising that participants encode the reliability probabilities of the sources in a distorted way and update their belief with some degree of recency, giving different weights to evidence according to its position in the sequence. In a second series of experiments, we also asked participants to rate, using a VAS 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. We show that the weight participants assigned to information is not optimal, with 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 strength of influence. In particular, participants were able to report being biased by unreliable sources. These findings suggest that human behaviour deviates from optimality but metacognitive monitoring allows to introspect to those biases.

Experiment 1

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). 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. 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

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 (**Error! Reference source not found.**A). 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., trustworthiness 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 (**Error! Reference source not**

found.A). 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 (**Error! Reference source not found.**A). 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 correct* 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 (**Error! Reference source not found.**

B). 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.

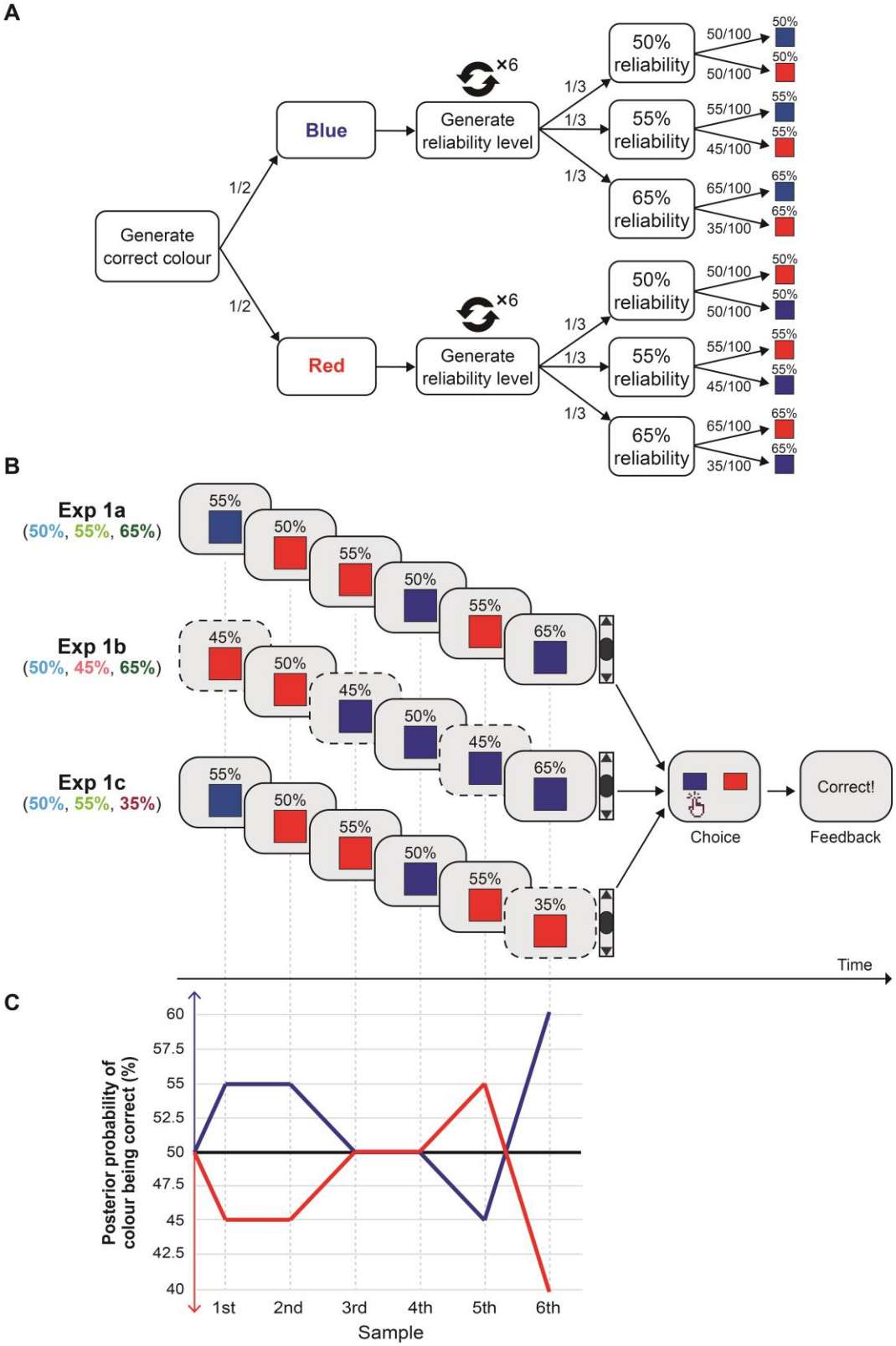

A
Generate correct colour
1/2
1/2
Blue
Red
×6
Generate reliability level
1/3
50% reliability
55% reliability
65% reliability
50/100
55/100
45/100
65/100
35/100
B
Exp 1a
(50%, 55%, 65%)
Exp 1b
(50%, 45%, 65%)
Exp 1c
(50%, 55%, 35%)
Choice
Correct!
Feedback
Time
C
Posterior probability of colour being correct (%)
60
57.5
55
52.5
50
47.5
45
42.5
40
1st
2nd
3rd
4th
5th
6th
Sample

Figure 1. Decision based on information reliability.

(A) Generative process of creating the information sequence. Each information source is generated by choosing a level of information reliability. The colour it displayed is chosen according to the correct colour, with a probability corresponding to its reliability. **(B)** Participants scrolled a webpage to view a sequence of six information 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, the reliability percentage varied between 50%, 55% and 65%. In experiment 1b, sources of 55% reliability were inverted to 45% reliability. In experiment 1c, 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. **(C)** The predictions of the colour were accumulated. This example suggests that the posterior probability of blue is 10% higher than red.

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:

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

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

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

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

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:

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

where parameter $\alpha \geq 0$ reflects the participant's sensitivity to the probabilistic information, and β 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. The Bayesian optimal corresponds to $\alpha = 1$ and $\beta = 0$. If, for instance, $\alpha = 0$ and $\beta > 0$, 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 $\widetilde{\mathrm{Lp}}$ can equivalently be written as:

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

where:

$$\varphi(\pi(s), \alpha, \beta) = \frac{\gamma \pi(s)^{\alpha}}{\gamma \pi(s)^{\alpha} + \left(1-\pi(s)\right)^{\alpha}}$$

and $\gamma = e^{\beta}$. The function $\varphi(\pi(s), \alpha, \beta)$ 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 memory leak, implying that each piece of information is weighted by its order in the presentation sequence. 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, that assigns a non-negative weight to each piece of information, and we normalize it by requiring $w(6) = 1$. We then obtain:

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

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

$$p(blue) = \frac{1}{1 + e^{-\tilde{L}(s_1, \dots, s_6, \alpha, \beta, w)}}$$

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 α or β 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.

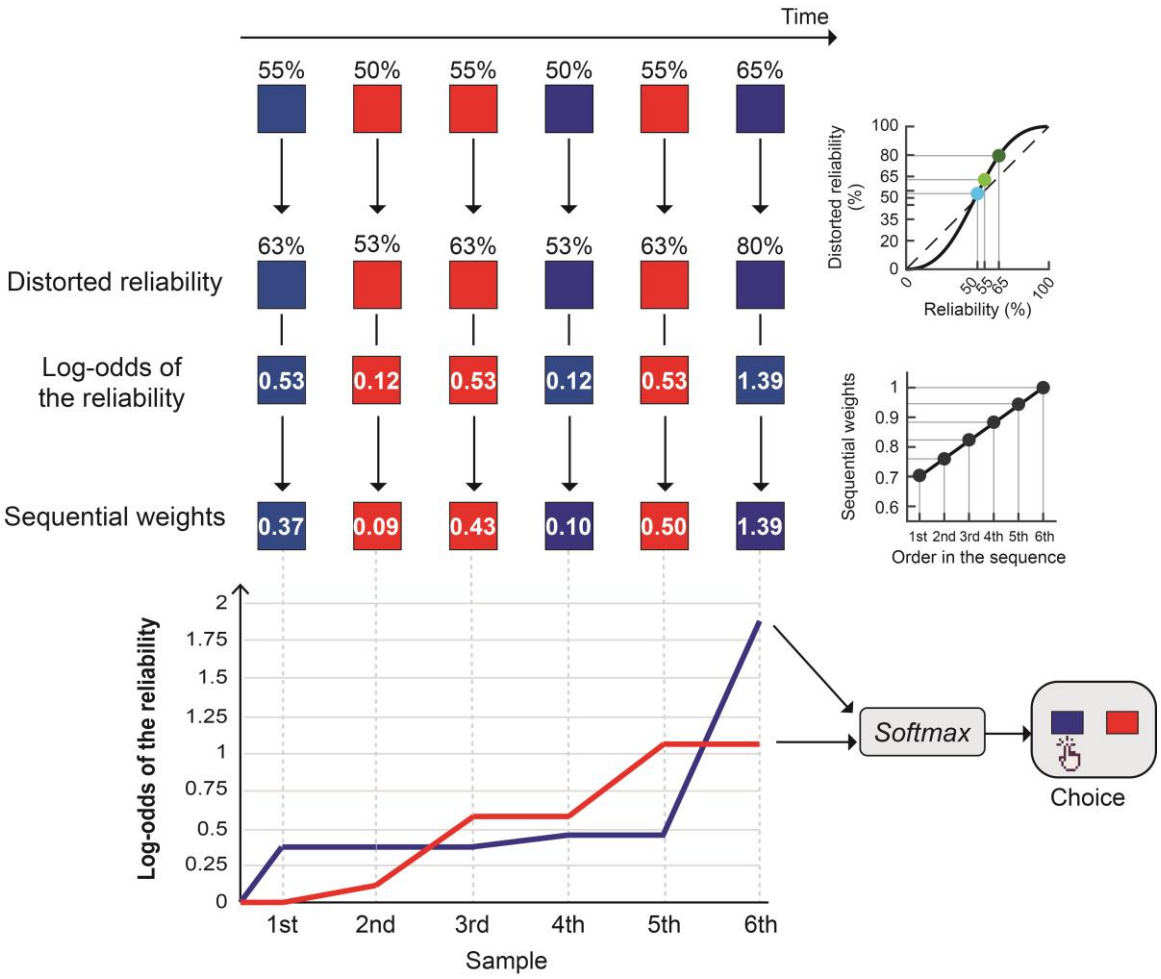


Figure 2. Computational model.

(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 next to it. (2) The computational model of the choice proposed that the reliability percentage displayed for each source is first encoded in a distorted manner and then transform in log-odds. (3) The log-odds of the distorted reliability is weighted depending on the order in which the information source was presented. (4) Evidence for each colour is accumulated by adding the log-odds of the reliability for each sample in the sequence. A choice is modelled based on softmax decision rule applied to the difference in the sum of the log-odds of reliability between the two colours.

We tested the assumption of reliability distortion and the assumption of memory leak 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 $\alpha = 1$, $\beta = 0$, and $w(p(s)) = 1$. In the second model, we introduced memory leak solely, keeping α = 1 and β = 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 $\Theta = \{\alpha, \beta\}$. In the last model, we introduced both memory leak and distortion in reliability, using a full set of free parameters $\Theta = \{\alpha, \beta, w(1), w(2), w(3), w(4), w(5)\}$.

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 individual parameters for each participant. The parameter α was constrained between 0 and 20, while memory leak parameters were constrained between 0 and 2. No constraint was imposed on β. 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).

Measuring the influence of information reliability on choice. 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 given a sequence. However, we could estimate how each of the three information reliability 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 and 50%, 55% and 35% in experiment 1c. For each of the three experiments, we used these three explanatory variables to predict the participant's choice c. Using "lme4" package (Bates et al., 2015), we performed mixed-effects logistic regression with the intercept varied between participants as a random effect.

Results

Choice behaviour is modulated by the presence of reliably wrong source

First, we examined whether choice accuracy was affected by the presence of reliably wrong information sources (Figure 3A). Percentage of correct choices was above chance in all three 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]). In contrast to our previous results (Jiwa et al., 2023), the percentage of correct choices was not significantly different between the 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).

We next considered the percentage of trials where participants made the same choice as the optimal choice predicted by Bayes rule (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]). We found a significant tendency between experiments (exp 1a vs. exp 1b: t [45] = 2.28, p = 0.054; exp 1a vs. exp 1c: t [50] = 2.15, p = 0.07). These results suggest that choice accuracy tends to be reduced when reliably wrong information sources are present.

To further quantify participants' performance, we computed how efficient they were at making correct choices (Juni et al., 2012), because participants were instructed to choose the correct colour as accurately as possible. Any deviation from optimal performance in integrating explicit cues of reliability could reduce one's performance. The percentage of correct choices by the ideal observer was on average 68.0% (67.4% in exp 1a; 68.5% in exp 1b; 68.1% in exp 1c). On the other hand, the mean of participants' chances was 61.9% (64.0% in exp 1a; 61.6% in

exp 1b; 59.9% in exp 1c), which means that their efficiency was 0.909. Participants thus lost 9.1% of the maximum possible accuracy relative to the ideal observer.

In the next analysis, we examined participants' trial-wise choice behaviour. We used a logistic regression approach to determine how the net number of samples of a given colour for a given reliability level would predict the choice. To do so, we computed, in each sequence, the difference between the number of red and blue squares for each of the three levels of information reliability, using this measure as a proxy for the strength of the signal at that reliability level. We then used these explanatory variables to predict trial-wise choice generated by Bayes rule as well as the choice made by the participants. 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).

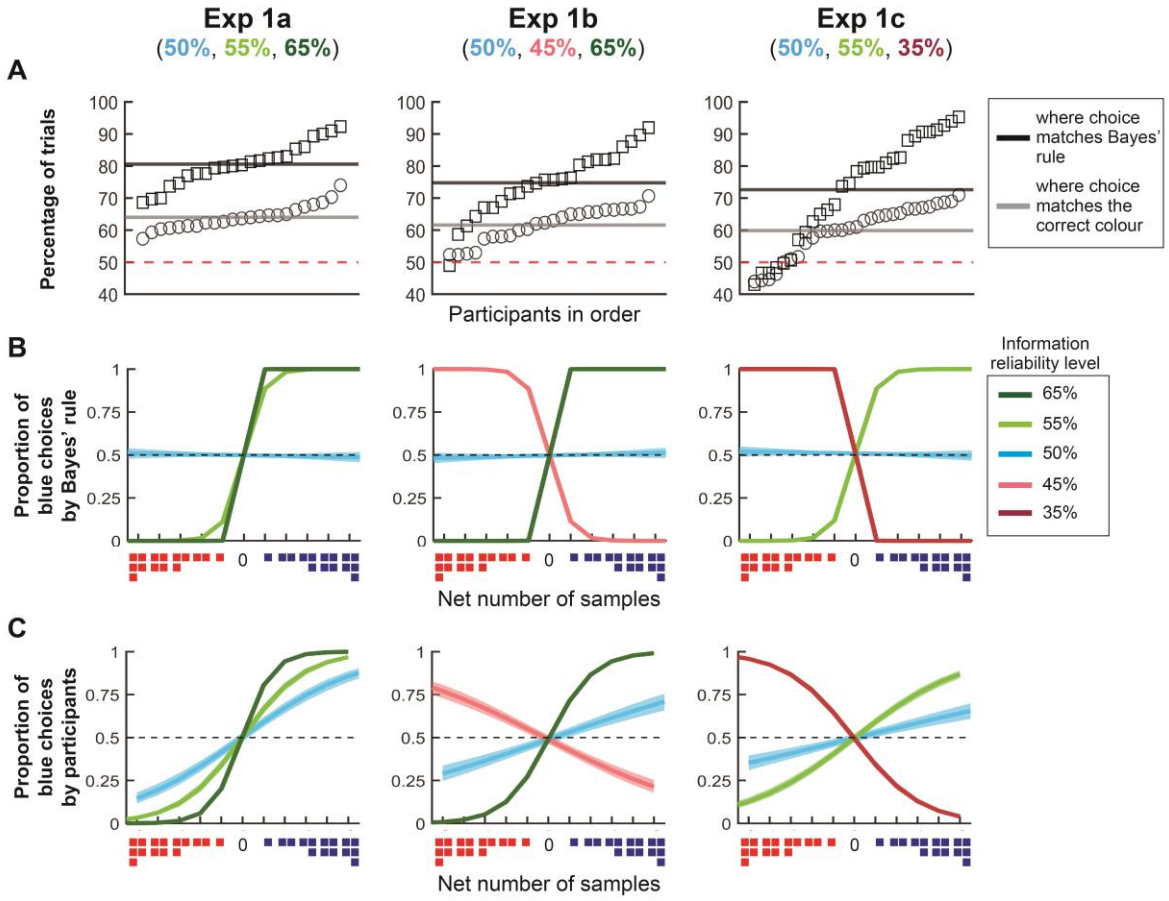


Figure 3. Measuring the influence of information reliability on choice.

(A) Percentage of choosing the correct colour (grey circles) and of following the optimal choices predicted by Bayes rule (black squares) 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.

Error! Reference source not found.B illustrates the choice behaviour of the Bayesian optimal observer predicted by the regression model. As can be guessed intuitively, the proportion of choosing the blue option increased when a given sequence displayed more blue squares of 55% and 65% reliability. A 65% reliability source had the most impact on the optimal choice and led to a step-like increase in the proportion of blue choices. A net difference of only one source displaying either one of the colours produced a 99.9% chance of choosing that colour. As predicted, the proportion of blue choices was reversed for the reliably wrong information source. The optimal choice was now to select the opposite colour to the prediction of the 45% or 35% reliability sources (pink curve or red curve respectively in **Error! Reference source not found.**B). Importantly, for the optimal observer, presenting reliably wrong sources did not change the slope of logistic regression but just reversed the sign of the slope, capturing the fact that according to Bayes rule, reliably wrong information is as informative as reliably correct information.

Turning to the behaviour of participants, however, different results were observed. Comparing the use of reliably wrong and reliably correct information, the slope of logistic regression was shallower for reliably wrong sources than for reliably correct. We observed that the positive regression coefficients of the 55% reliability source on choice behaviour in experiment 1a (β = 0.68, 99%CI = [0.60, 0.76]) or experiment 1c (exp 1c: β = 0.38, 99%CI = [0.33, 0.43]) were significantly greater in absolute value than negative coefficients of the 45% reliability source in experiment 1b (β = -0.25, 99%CI = [-0.32, -0.19]) (exp 1a vs. exp 1b: z = 12.9, $p < 0.001$; exp 1a vs. exp 1c; z = 9.3, $p < 0.001$). A similar pattern of regression coefficients was also confirmed between the 65% reliability source in experiment 1a (β = 1.40, 99%CI = [1.29, 1.50]) or experiment 1b (β = 0.95, 99%CI = [0.88, 1.03]) and the 35% reliability source in experiment 1c (β = -0.63, 99%CI = [-0.69, -0.57]) (exp 1a vs. exp 1c: z = 19.3, $p < 0.001$; exp 1b vs. exp 1c: z = 10.3, $p < 0.001$). Therefore, the presence of reliably wrong information reduced the sensitivity to information reliability. In other words, although reliably correct information is as informative as reliably wrong information, human participants failed to optimally use information that was reliably saying the opposite of the truth.

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 optimal observer did not increase with the net number of samples of a given colour displayed by 50% reliability (**Error! Reference source not found.**B, blue line). In contrast to this prediction, the net number of 50% reliable samples impacted the choice of participants in all three experiments (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]). That is, participants chose the blue option more frequently when one or more unreliable sources (50% reliability) were blue. These results demonstrate that, even when unreliable information is explicitly labelled as such, people fail to ignore unreliable cues by demonstrating a positive bias by the colour that these cues display.

Levels of information reliability

Lastly, we compared the effect across three levels of information reliability within each experiment. In experiment 1a, the slope of regression with the 65% reliability source on choice was greater than those with the 55% and 50% reliability (65% vs. 55%: $z = 16.3$, $p < 0.001$; 65% vs. 50%: $z = 24.6$, $p < 0.001$). The 55% reliability source was more strongly correlated with choice than the 50% reliability (55% vs. 50%: $z = 9.1$, $p < 0.001$). In experiment 1b, the slope of regression with the 65% reliability source was greater, in absolute value, than those with the 45% and 50% reliability (65% vs. 45%: $z = 21.8$, $p < 0.001$; 65% vs. 50%: $z = 24.5$, $p < 0.001$). The absolute regression coefficient of the 45% reliability was greater than the 50% reliability (45% vs. 50%: $z = 2.9$, $p < 0.05$). Finally, in experiment 1c, the absolute regression coefficient of the 35% reliability was greater than the 55% and 50% reliability (35% vs. 55%: $z = 9.6$, $p < 0.001$; 35% vs. 50%: $z = 20.5$, $p < 0.001$). Again, the effect of the 55% reliability on choice was stronger than the 50% reliability (55% vs. 50%: $z = 10.7$, $p < 0.001$). These results confirm that three levels of information reliability gradually increased the likelihood of choosing the colour that these cues display.

A computational mechanism underlying choice behaviour

We demonstrated a clear deviation from the optimal behaviour when assigning weights to information sources according to their reliability. To shed light on this suboptimality in the information integration process, we considered two potential factors: 1) the distortion in how the

reliability score is encoded (i.e. a 50% unreliable square is actually treated as slightly reliable 52%) and 2) how the information is lost through decay during sequential sampling (sequential weighting). To probe the existence of the two sources of errors, we built four models including them gradually to predict the choice made by participants. The first model was the optimal Bayesian observer model (BDT) with no distortion in the encoding of information reliability and no memory loss. The second model corresponded to the optimal observer with no distortion in the encoding of the reliability but with sequential weights mimicking the effect of memory decay. The third model corresponded to an observer that would encode the reliability scores with a certain degree of distortion but without memory loss. 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. We confirmed that the recovery of group-level parameters and individual-level parameters is robust (Supplementary Figure 1 & Supplementary Figure 2).

The bars in Figure 4A show the expected log pointwise predictive density (i.e., the out-of-sample predictive accuracy of the model) relative to the best model. In all the first three experiments, the BDT model performed the worst to fit the data. The full model including both sequential weight and distortion of reliability outperformed the other three models. The reliability distortion model was significantly better than the sequential weights model, suggesting that participants' choices were better explained by a distorted perception of information reliability than by sequential weights.

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 an above chance reliability in all three experiments (61.4% in exp 1a; 53.7% in exp 1b; 54.0% in exp 1c) 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 used the reliably correct information in their decisions as if they overestimated both the sources of 55% reliability (72.4% in exp 1a; 62.3% in exp 1b) and those of 65% reliability (88.0% in exp 1a; 82.0% in exp 1b). In contrast, the participants used reliably wrong information as if they underestimated the sources of 45% reliability (42.7% in exp 1b) and those of 35% reliability (29.2% in exp 1c).

We also illustrate sequential weights on information sources predicted by the full model (Figure 4C). In general, the model predicted that participants reduced their weights on the

information source as it was presented earlier in a sequence, resembling memory loss. In particular, on average across the three experiments, weights placed on the first source were 26% smaller than the last source.

A Bayesian hierarchical model estimates a full posterior sample of the model parameter and thus allows us to compare a difference in parameter values between experiments. The full model included two group-level parameters which define a shape of reliability distortion (i.e., the parameter for the sensitivity to information and the parameter for the presentation bias) as well as five group-level parameters for memory leak (see Methods). We retrieved posterior distributions of these seven parameters for each experiment from the fit of the full model (Supplementary Figure 3). By comparing a pairwise difference of the same parameter between experiments, we found that the parameter capturing the bias by unreliable sources was significantly greater in experiment 1a than in experiments 1b and 1c (exp 1a vs. exp 1b: $p = 0.012$; exp 1a vs. exp 1c: $p < 0.01$). This result indicates that the positive bias by unreliable sources was mitigated by the presence of reliably wrong information sources. No significant difference between experiments was found for the other six parameters.

To summarise, a computational analysis suggests that people distort the objective reliability of sources, even when it is given explicitly. Participants tend to consider unreliable sources to be reliably correct but this effect is mitigated when reliably wrong sources of information are present. Moreover, participants overweight the last information they received compared to the first piece of information. In experiment 1, we quantified the objective influence of explicit reliability cues on decisions. We then turned to the question of whether people were aware of the influence of reliability cues on their decisions.

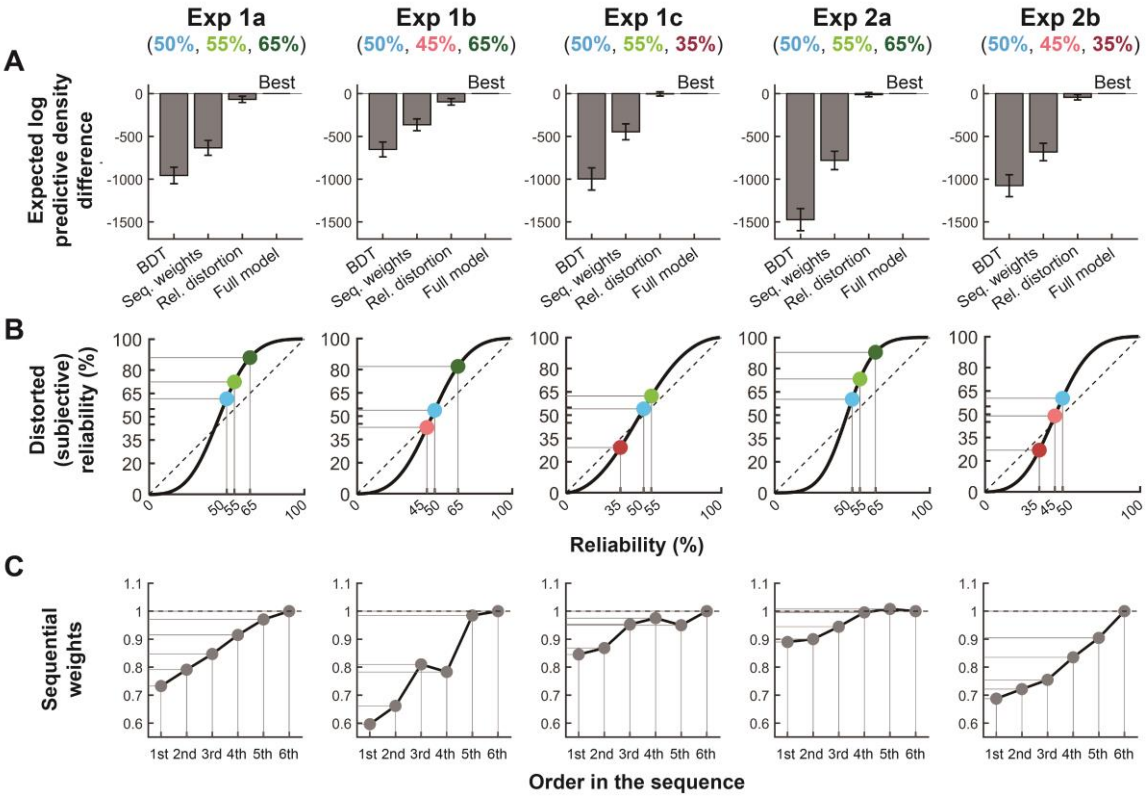


Figure 4. Results of computational analysis.

(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.

Experiment 2

Methods

Participants

Using Prolific, we recruited 64 participants for our two introspection 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 were fluent English speakers and had no history of neurological disorders, and normal vision. Participants received a basic payment of £8 for their participation in a 60-minute experiment and earned a bonus based on their performance on the task. No participants were excluded from the analysis. Participants gave informed consent and all procedures were approved by the Research Ethics Committee of University College London (ID ICN-PH-PWB-22-11-18A).

Experimental design

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 reliably wrong or unreliable information only 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 5). 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.

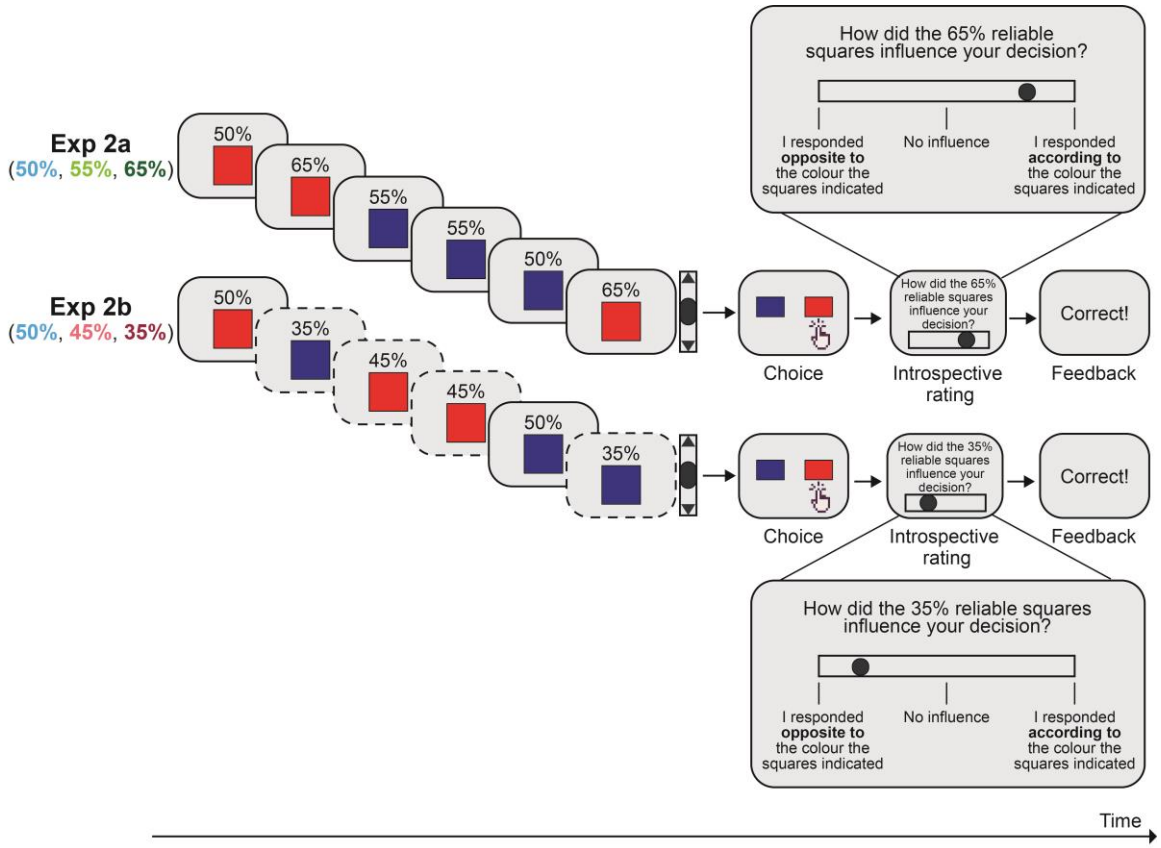


Figure 5. Introspection experiment.

In experiment 2a (i), the reliability reliability varied between 50%, 55% and 65%. In experiment 2b (ii), sources of 55% reliability were inverted to 45% and sources of 65% reliability were inverted to 35% so that they predicted the opposite colour. Two sequences showed the equivalent posterior probability of the true webpage colour. After choosing the colour, participants were 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.

Data analysis

The optimal choice based on Bayes rule. We modelled the optimal choice based on Bayes rule as in the first three experiments.

Computational models. Four computational models from the first three experiments were independently fitted to the participants' choice data in experiment 2a and experiment 2b, respectively.

Measuring the influence of information reliability on choice. The same logistic regression that was performed in the first three experiments was conducted to estimate the influence of information reliability on the participants' choices. A net number of sources between blue and red was counted separately for 50%, 55% and 65% in experiment 2a and 50%, 45% and 35% in experiment 2b.

Measuring the subjective influence of information reliability. 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 introspect 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 introspect 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, 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 assigned 0 at 65% reliability, 2 for congruent sources at 55% reliability and 2 for incongruent sources at 50% reliability. 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. The level of reliability that was chosen for the introspective report interacted with the net number of samples at each reliability level, creating nine explanatory variables in each experiment. Further, the interaction term was added for each level of introspected reliability. For each of the two experiments, we used these twelve explanatory variables to predict the participant's rating of introspective reports. Using "lme4" package (Bates et al., 2015), we performed mixed-effects linear regression with the intercept varied between participants as a random effect.

Results

The presence of unreliable and reliably wrong information impairs choice

In experiments 2a and 2b, we found that the percentage of correct choices was significantly lower when reliably wrong information was presented (Figure 6A; exp 2b: mean = 58.2%, 99%CI = [54.6, 61.9]) compared to when information was reliably right (exp 2a: mean = 64.4%, 99%CI = [62.4, 66.4]) (t [62] = 4.43, $p < 0.001$). Similarly, the percentage of optimal choices was also lower in experiment 2b (mean = 69.4.2%, 99%CI = [61.7, 77.2]) than in experiment 2a (mean = 84.4%, 99%CI = [80.0, 88.8]) (t [62] = 5.04, $p < 0.001$). In terms of how efficient participants were at making correct choices, the percentage of correct choices by the ideal observer was on average 67.6% (67.5% in exp 2a; 67.8% in exp 2b). On the other hand, the percentage of correct choices by participants was 61.3% (64.4% in exp 2a; 58.3% in exp 2b),

which means that their efficiency was 0.906. Participants thus lost 9.4% of the maximum possible accuracy relative to the ideal observer.

To understand how participants treated individual reliability scores, we repeated the logistic regression analysis of the first experiment. For the bayesian optimal observer, as predicted, the amount of squares of a given colour positively correlated with the likelihood of choosing that colour for the reliably correct sources and negatively correlated with the likelihood of choosing that colour for the reliably wrong information (Figure 6B). The participant's choice likelihood increased more steeply with the amount of reliably correct information (exp 2a) than that of the reliably wrong (exp 2b) information (Figure 6C). The slope of the 55% reliability source in experiment 2a (β = 0.72, 99% CI = [0.65, 0.79]) was significantly greater in absolute value than the slope of the 45% reliability source in experiment 2b (β = -0.08, 99% CI = [-0.13, -0.03]; exp 2a vs. exp 2b: z = 21.7, p < 0.001). The same pattern of results was also observed for the 65% reliability source in experiment 2a (β = 1.73, 99%CI = [1.62, 1.84]) and the 35% reliability source in experiment 2b (β = -0.66, 99%CI = [-0.71, -0.60]; exp 2a vs. exp 2b: z = 25.9, p < 0.001). Therefore, the presence of the reliably wrong information source reduced sensitivity to information reliability.

Furthermore, we found that the evidence provided by the 50% reliability source positively impacted participant's choice in both experiments (exp 2a: β = 0.34, 99% CI = [0.28, 0.41]; exp 2b: β = 0.28, 99% CI = [0.23, 0.33]). This again confirmed that unreliable information biased decision-making, even when labelled as such, and even in a context where the only informative evidence came from reliably wrong sources. Indeed, the regression coefficients were not significantly different in both experiments, suggesting that the bias by unreliable information is present regardless of whether other information sources are reliably correct or reliably wrong.

In experiment 2a, the slope of the 65% reliability source on choice was greater than those of the 55% and 50% reliability (65% vs. 55%: z = 22.7, p < 0.001; 65% vs. 50%: z = 32.2, p < 0.001). The 55% reliability source was more strongly correlated with choice than the 50% reliability (55% vs. 50%: z = 11.5, p < 0.001). In experiment 2b, the slope of the 35% reliability source was greater, in absolute value, than those of the 45% and 50% reliability (35% vs. 45%: z = 23.5, p < 0.001; 35% vs. 50%: z = 15.4, p < 0.001). The absolute regression coefficient of the 45% reliability was greater than the 50% reliability (45% vs. 50%: z = 8.3, p < 0.001). Again, these results confirm a gradual increase in the likelihood of choice by three levels of information reliability.

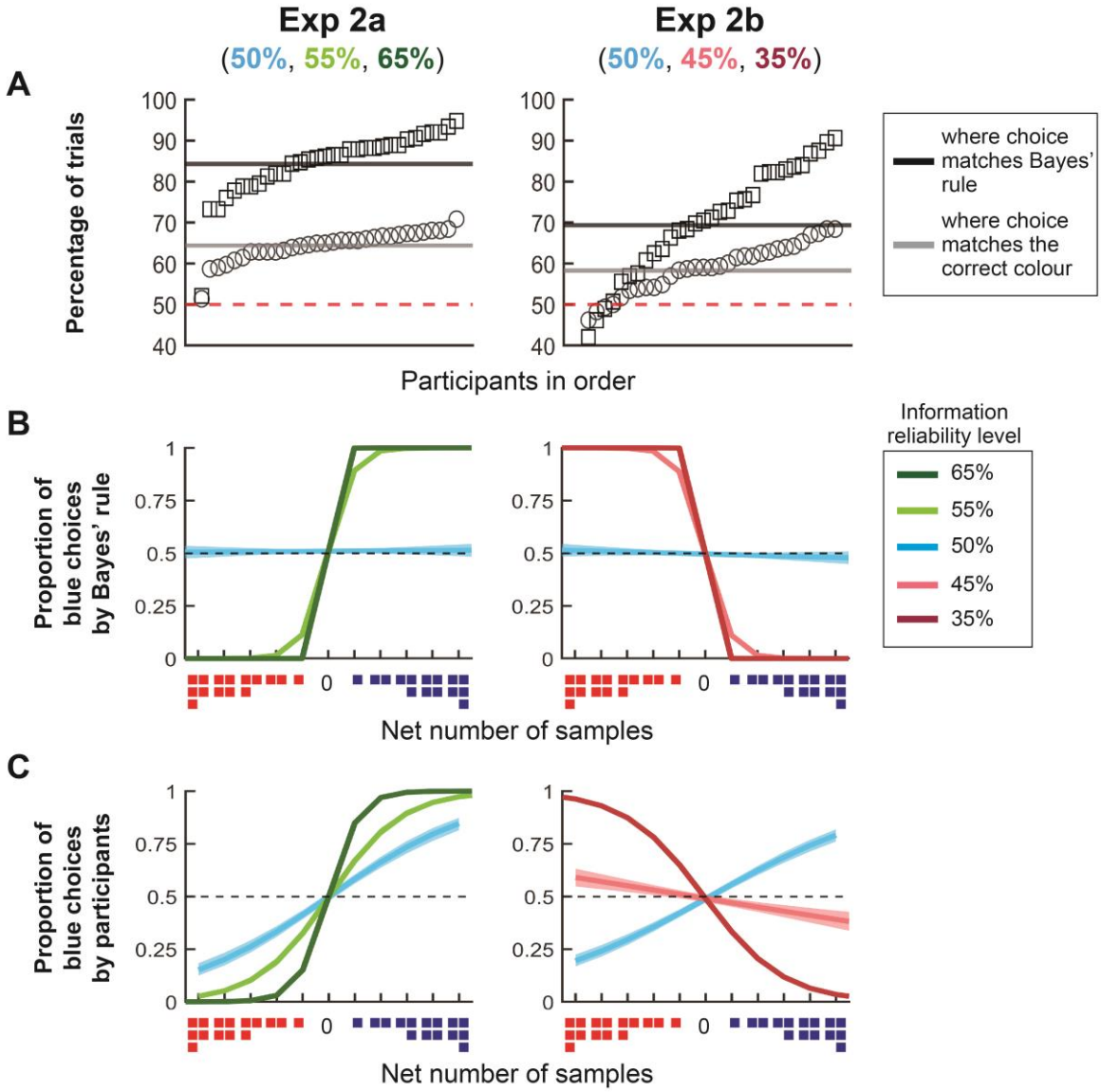


Figure 6. Replicating the results of experiment 1.

(A) Percentage of choosing the correct colour (grey) and following the optimal choices predicted by Bayes rule (black) ordered from the lowest to the highest performance. **(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. **(C)** Proportion of blue choices predicted from participants' choices.

Computational analysis confirmed that, as for experiment 1, the full model including reliability distortion and sequential weights outperformed other models (Figure 4A) in predicting choice behaviour. The patterns of reliability distortion and sequential weights were almost identical to the patterns observed in the first three experiments: The reliability of 50% sources

was overestimated (60.1% in exp 2a; 60.2% in exp 2b). The participants behaved as if they overestimated the reliability of the sources of 55% (73.2%) and 65% (90.4%) in experiment 2a and underestimated the reliability of 35% source (26.8%) in experiment 2b. The estimate of the 45% source was close to the true reliability (48.9%). Further, the participants gave less weights on the first few samples in a sequence.

A comparison of the posterior distribution of the group-level parameter in the full model shows no significant difference in the information sensitivity and the presentation bias between the two experiments (Supplementary Figure 3), suggesting that the shape of reliability distortion was similar. However, we found a significant reduce in the values of sequential weights in experiment 2b compared with experiment 2a (*ps* < 0.05). Therefore, the presence of reliably wrong information increased the memory leak.

How do people feel the influence of information reliability on their choices?

In the next analysis, we addressed people's metacognitive knowledge of how much they knew their decisions were influenced by the evidence provided by each reliability source. We reasoned that if participants were able to introspect the influence of information reliability on their choice, their subjective report would increase with the reliability of the source introspected, 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 introspected and 2) the net number of samples of that reliability level congruent with the choice (see Methods).

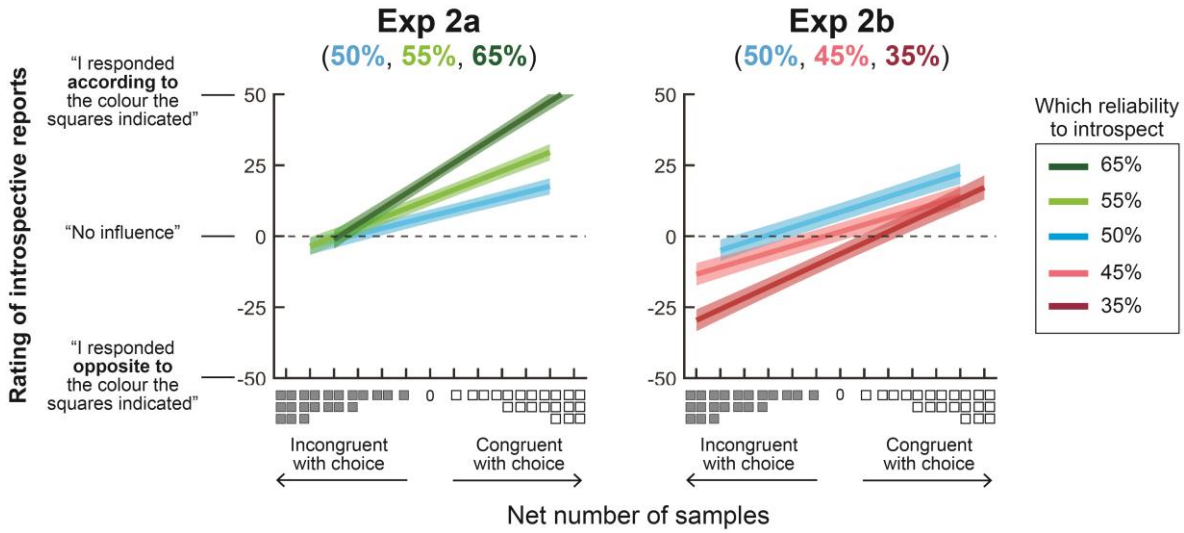


Figure 7. Measuring the subjective feeling of the influence of information reliability on choice.
Rating of introspective reports as a function of the net number of samples for congruent or incongruent with the colour the participant chose and of the reliability level that was chosen for the introspective 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 were plotted against the net number of congruent samples, separately for each level of introspected reliability. The ratings increased with introspected reliability: in experiment 2a, the intercept of regression was 25.6 (99% CI = [22.0, 29.2]), 14.4 (99% CI = [10.8, 18.0]) and 7.4 (99% CI = [3.8, 11.0]) when the introspected reliability was 65%, 55% and 50%, respectively. In experiment 2b, the intercept of regression was -8.6 (99% CI = [-13.5, -3.8]), 1.3 (99% CI = [-3.5, 6.2]) and 9.1 (99% CI = [4.3, 14.0]) when the introspected reliability was 35%, 45% and 50%, respectively.

In addition, the ratings increased with the net number of samples of the colour congruency with the choice. In experiment 2a, the slope of regression against the net number of congruent samples was significantly positive when the reliability chosen for the introspective report was 65% (β = 5.40, 99% CI = [4.76, 6.02]), 55% (β = 3.30, 99% CI = [2.73, 3.88]) and 50% (β = 2.11, 99% CI = [1.58, 2.64]), respectively. Similarly, in experiment 2b, the slope was significantly positive when the introspected reliability was 35% (β = 3.90, 99% CI = [3.26, 4.54]), 45% (β = 2.50, 99% CI = [1.87, 3.12] and 50% (β = 2.70, 99% CI = [2.09, 3.31]), respectively. 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.

We reasoned that if participants remained unaware of the influence of unreliable information on choice (Figure 6C), 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. We found a significant positive slope for the 50% reliability in both experiment 2a (β = 2.11, 99% CI = [1.58, 2.64]) and experiment 2b (β = 2.70, 99% CI = [2.09, 3.31]). This result suggests that participants could detect their own choice bias by unreliable evidence and were aware of their own suboptimal behaviour.

Next, we compared the slopes of regression across the information reliability within each experiment. In experiment 2a, the slope with the 65% reliability source was significantly greater than the slope with the 55% and 50% reliability (65% vs. 55%: z = 7.15, p < 0.001; 65% vs. 50%:

z = 11.6, p < 0.001). The 55% reliability source was more strongly correlated with the introspective report than the 50% reliability (55% vs. 50%: z = 4.6, p < 0.001). In experiment 2b, the slope with the 35% reliability source was significantly greater than the slope with the 45% and 50% reliability (35% vs. 45%: z = 4.7, p < 0.001; 35% vs. 50%: z = 4.1, p < 0.001). We did not find a significant difference in the slopes between the 45% reliability and the 50% reliability (45% vs. 50%: z = 0.7, p = 0.99). Because three levels of information reliability increasingly influenced the likelihood of choice (Figure 6C), these results suggest that participants adjusted their subjective feelings of influence reflecting on the degree to which their choice was actually influenced by the source of information reliability. 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).

Our interpretation was based on the average across participants. To understand whether individuals could introspect their own weight they themselves assigned to each reliability level when deciding a colour, we correlated across participants the individual objective weights of the reliability sources with the subjective influence of each source. To do so, we recovered from the best-fitting computational model the distorted reliability score that individuals assigned to each of the three levels of information reliability (Figure 4). These distorted reliabilities are a proxy of the extent to which the participant gave weight to the information provided by that source, thereby quantifying the *objective measure of the influence* of that source on behaviour. On the other hand, we also retrieved the slope of the regression in Figure 7 which provides a measure of how much the participants were sensitive to the quantity of information displayed by that source when reporting its influence, thereby quantifying the *subjective sense of the influence* of each given source. For each participant, we correlated the distorted reliability score with the slope of subjective influence.

If the distorted reliability of a given source was equal to 50%, it would mean that sources were considered unreliable by the participant, thereby having no influence on choice. If the individual were able to report correctly they considered that source as unreliable, the slope of regression should be close to zero, providing a minimal for the relation between perceived reliability and subjective report. On the other hand, If the distorted reliability of a given source was above 50%, it would correspond to a positive influence on the choice (i.e., following the colour of the information source). If the individual were aware of such a positive influence, the slope of regression should be positive. Lastly, If the distorted reliability of a given source was

below 50%, it would correspond to a negative influence on the choice (i.e., opposing the colour of the information source). If the individual were aware of such a negative influence, the slope of regression should be positive again. For those reasons, we expected a quadratic relationship between the objective measure of influence and the subjective measure.

Figure 8 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 decision weight to a source of information are also able to report a stronger sense of being influenced by that source of information, than those who assign a weak decision weight. Our findings (Figure 7 & Figure 8) provide substantial evidence that people can introspect the degree to which particular information sources influence their behaviour.

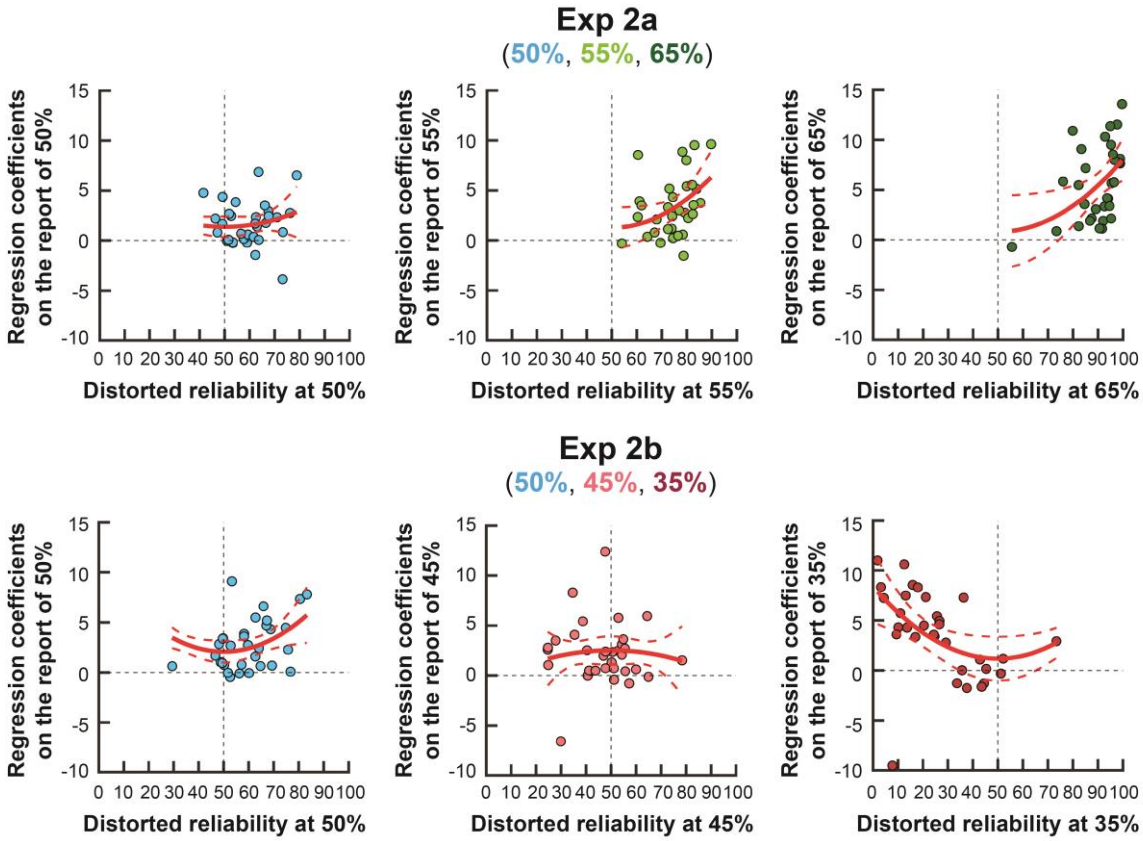


Figure 8. Correlation between the subjective feeling of the influence and the actual strength of the influence.

Across participants, the individual's sensitivity (i.e., the slope of regression) to introspect the influence of a particular level of information reliability on their choices is plotted against the same individual's distortion from that reliability. The distortion in reliability indicates the amount of choice bias induced by a particular level of information reliability and therefore tells the actual strength of the 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 of a source of information when making decisions. Participants saw successive pieces of evidence coming from different sources supporting one of two possible responses (red or blue). Each sample was associated with a percentage corresponding to the reliability of the source of the information i.e. the probability that the source would show information congruent with the correct color for that trial. 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 correct sources of information that were predictive of the correct colour, or reliably wrong sources of information that would show reliably the incorrect colour (experiment 1b-c, experiment 2b). Crucially, they were also presented with unreliable sources of information 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 a distorted representation of the reliability of the sources and whether they gave different weights to each sample in the sequence according to their position. Experiment 1 revealed that participants did not use explicit reliability optimally, overestimating the reliability of the unreliable and reliably correct sources. Participants also struggled to use reliably wrong sources of information, despite these being in theory equally informative for the Bayesian optimal observer than reliably correct 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 these sources supported. In a second series of experiments, we probed whether participants were aware of these influences on their choice, by asking them to report how a given source influenced their choice. Overall, the participants reported a stronger sense of influence when the reliability of the source increased and when the evidence shown was congruent with the choice. Crucially, participants were also aware that they were positively biased by the unreliable information, speaking in favour of participants' ability to introspect their use of explicit reliability.

The question of how humans and other animals estimate and use information about environmental uncertainty has been the subject of intense research (Ma & Jazayeri, 2014; Maloney & Mamassian, 2009). Previous studies on this question have explored how people learn to use information from noisy sensory cues, forming implicit estimations of the reliability of

different sources of information to maximize reward (Ma, 2023; Maloney & Zhang, 2010). Crucially, this research has revealed how people form beliefs about source precision through perceptual experience. Importantly, only very few studies have explored how people use explicit probabilistic information about the reliability of sources (Carlebach & Yeung, 2023; Pescetelli et al., 2021; Pescetelli & Yeung, 2021). Research in social and applied psychology has also started investigating how people form explicit beliefs about source trustworthiness based on their prior beliefs and novel information, looking for instance at the effect of debunking (Radkani et al., 2024; Swire-Thompson et al., 2021). Similarly, research in applied psychology has started to look at how people use reviews and ratings to form opinions on products or companies (Hoffart et al., 2019; Oktar & Lombrozo, 2025). Nevertheless, we lack both empirical evidence and theoretical frameworks for understanding how people use explicit probabilistic knowledge about the reliability of the information they sample to make decisions.

Research in behavioural economics provides the beginning of an answer to this question, revealing clear limitations in making optimal inferences when reasoning with probabilities. It has been established that the perception of the likelihood of events is distorted (Gonzalez & Wu, 1999; Tversky & Kahneman, 1992). The probability of rare events is typically overestimated while frequent events are underestimated. This was demonstrated for decisions under risk where 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). In that situation, it is possible to infer, using the choice patterns of participants, the odds of winning they seem to be basing their choice on. To account for these results, it has been proposed that probabilities undergo a non-linear weighting function resembling an “inverted S-shaped” 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 explicit reliability percentages, based on their patterns of choices. We indeed found that participants behaved as if they underestimated small reliability percentages and overestimated medium-to-large reliability percentages, 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. Indeed we observed that participants particularly overestimated the reliability of the most reliable sources presented. This is in broad agreement with the idea that these distortions happen due to the normalization of the probabilities according to the range of probabilities presented (Zhang et al., 2020). Zhang et al (2020) proposed that when reasoning with probabilistic information,

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 can explain how the distortion function would vary across contexts, depending on the range of reliability presented. Further research will be needed to confirm whether this process explains the difference in distortion we observed, varying systematically the range of reliabilities presented.

Beyond distorting reliabilities, our findings revealed three other types of suboptimal behaviour in participants' decision processes. First, participants were less able to use reliably wrong sources than reliably correct sources. This may be due to an increase in cognitive loads when processing information provided by reliably wrong information. Indeed, we observed that memory decay in the evidence accumulation process increased when reliably wrong information was present in experiment 2b: participants lost the information presented by the first sample by about 30%. Indeed, participants were required to reinterpret the suggestions by reliably wrong sources as evidence in favour of the opposite colour. Counterfactual thinking is known to engage working memory and affect decision-making (Van Hoeck et al., 2015). Second, participants' choices were biased by the colour suggested by the unreliable information sources. Furthermore, as the number of unreliable samples displaying a colour increased, participants became even more likely to choose the colour suggested by the unreliable sources. This 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%. This was true even though participants were explicitly informed that 50% reliability squares were uninformative, as reliable as a coin flip, and instructed to ignore them. However, the presence of one reliably wrong source in experiments 1b and 1c reduced such a bias. 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 uncertain, it is then possible participants relied on unreliable information to break the tie and choose a response based on the random evidence it displayed.

The results of experiment 2 showed that participants were aware that their response was 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 their colour. 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 participants 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 that 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). However, people use their metacognitive knowledge to regulate their behaviour when the use of metacognitive control is pronounced (Boldt & Gilbert, 2022; Risko & Gilbert, 2016). Another possibility 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. 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, three levels of source reliability) had an increased correlation with both the objective performance of participant's choice (Figure 6C) and the subjective report of influence (Figure 7). Across participants, we correlated the individual's objective influence of information source on choice with the same individual's subjective sense of influence (Figure 8). 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 introspect the influence of choice reflecting on how their decisions are made.

Importantly, however, subjective reports of influence were not predicted by a more comprehensive measure of influence encompassing the relative weight of a piece of information on choice. Formally, the influence of a piece of evidence on choice can be retrieved by imagining what would have been the behaviour if one had not observed that piece of evidence. In other words, it consists of computing the likelihood of having made the same choice if that piece of evidence was absent and comparing it to the likelihood of making the choice when that piece of information was present using counterfactual thinking to infer influence. 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 are 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 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.

In everyday life, explicit, descriptive information about the trustworthiness of information is often available. For instance, online shopping services such as Amazon provide a star rating review indicating customers' global agreement of how good products are (Hoffart et al., 2019; Oktar & Lombrozo, 2025). In social networking services, misleading posts and fake facts are easily spread out (Vosoughi et al., 2018), which leads people to form false beliefs (Fazio et al., 2013; Rapp, 2008; Rapp & Salovich, 2018). Social networking services add fact-checking flags in the hope that flags could alert people to fake news (Andersen & Søe, 2019; Gaozhao, 2020). Our experimental work implies that people still overly trust or may distrust information, even when such explicit information about trustworthiness is available to them. Indeed, drivers of forming false beliefs include neglecting source cues (Ecker et al., 2022). For example, a logo banner specifying the publisher for online news items has been found not to decrease belief in fake news or increase belief in factual news (Dias et al., 2020). Therefore, a more sophisticated approach such as prebunking or debunking would be important to prevent such cognitive failures (Ecker et al., 2022).

Competing interests

The authors declare no competing interests.

Author's contributions

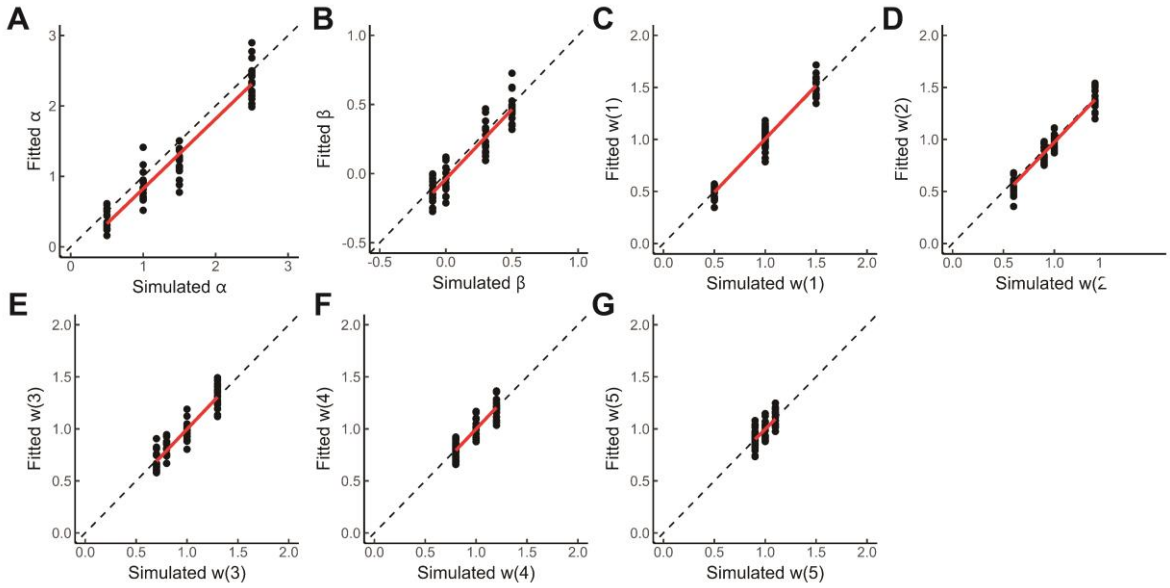
Acknowledgments

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

Data and code availability

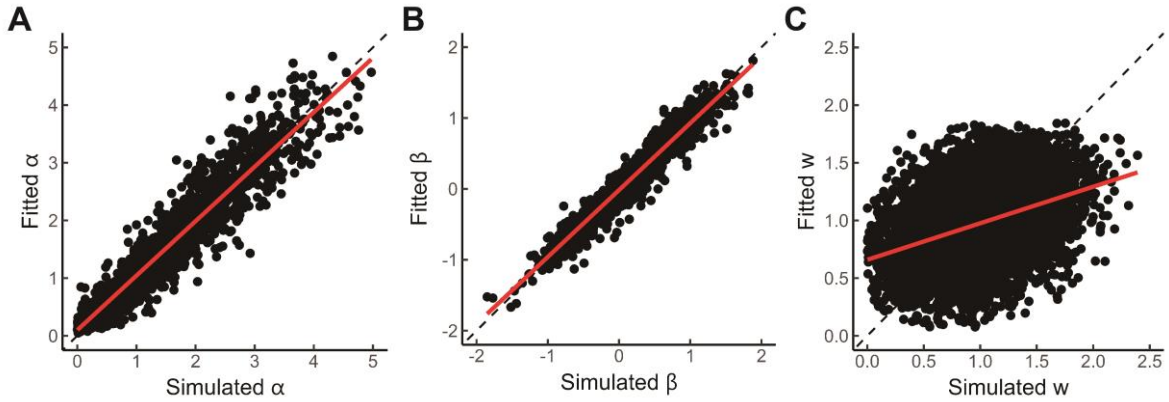
The data frame for a representative participant will be posted. The data frames for all the participants and the codes used to generate the figures will be posted.

Supplementary Information



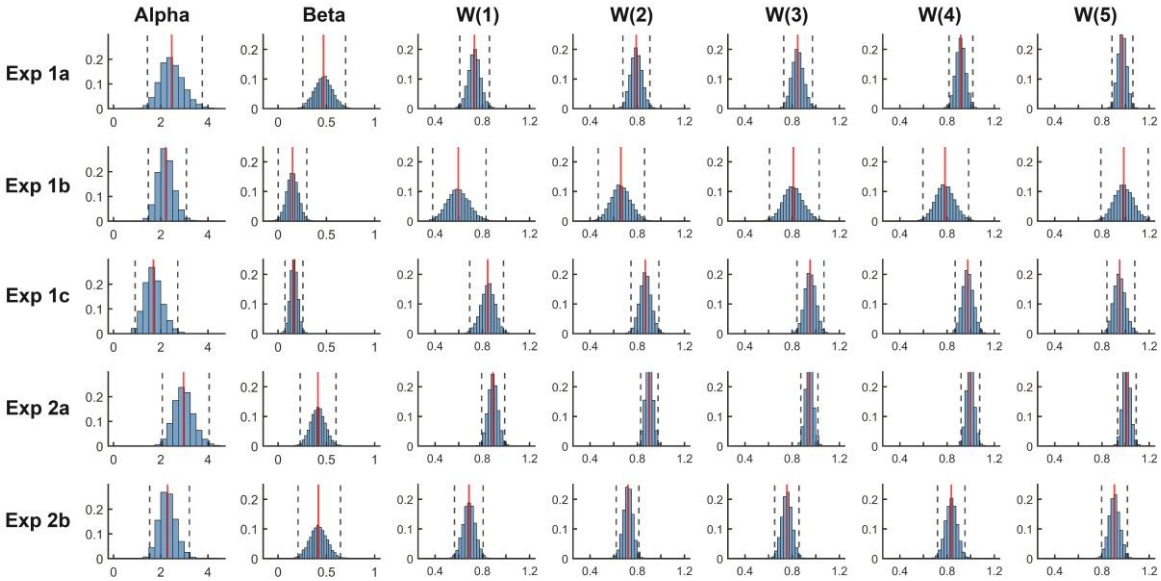
Supplementary Figure 1. Recovery of group-level parameters.

A. Parameter recovery of alpha (i.e., a parameter of sensitivity to the probabilistic information). **B.** Parameter recovery of beta (i.e., a parameter of presentation bias). **C-G.** Parameter recovery of sequential weights (i.e., a parameter of memory leak). 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.

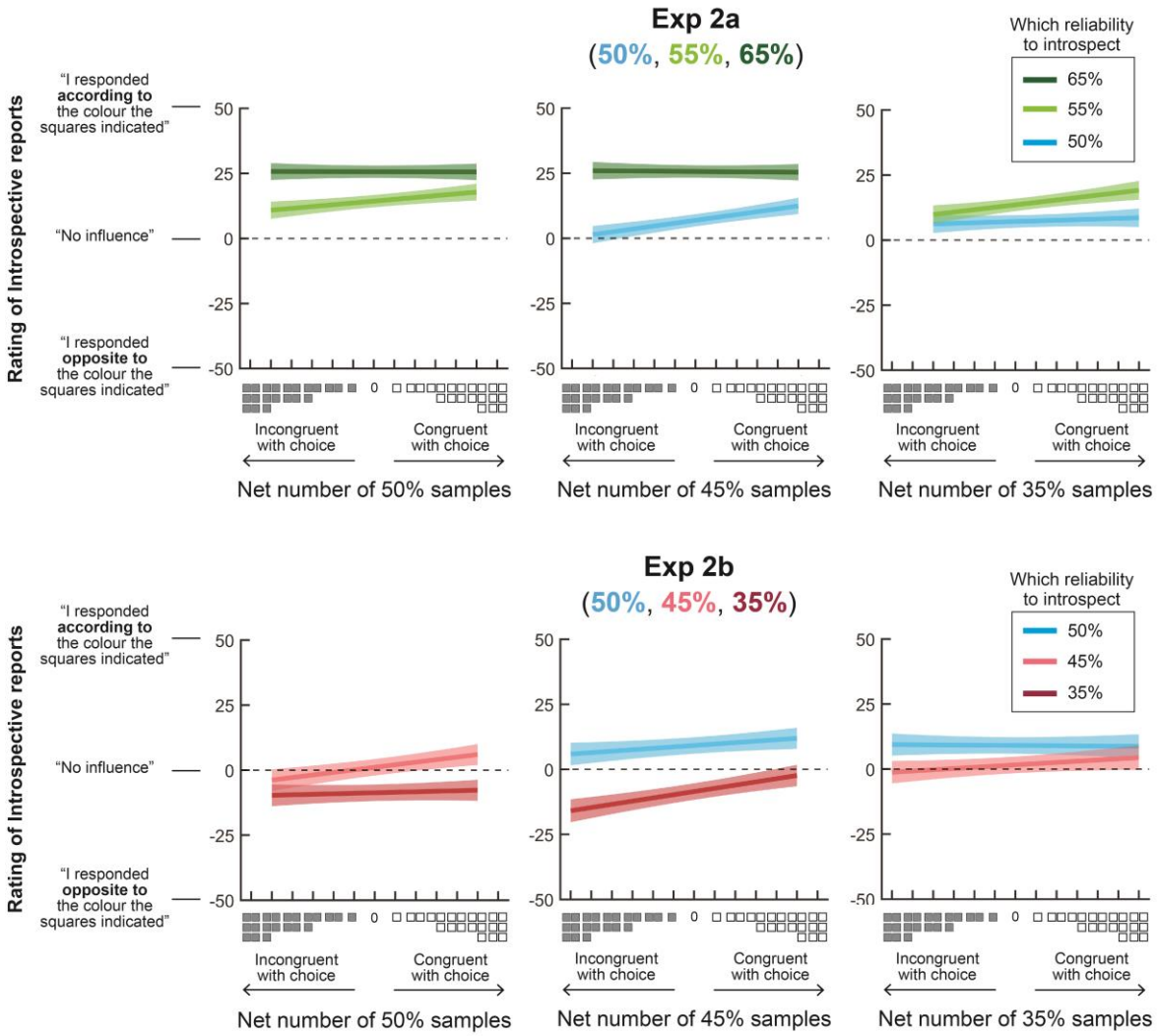


Supplementary Figure 2. Recovery of individual-level parameters.

A. Parameter recovery of alpha. **B.** Parameter recovery of beta. **C.** Parameter 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. Full posterior samples of group-level parameters. We retrieved the estimates of seven group-level parameters from a Bayesian hierarchical model fit of the full model. Histograms show the estimates of parameter values. Red lines denote the mean while dashed lines denote the 99% confidence intervals.



Supplementary Figure 4. Marginal effects on the introspective reports.

See the caption in Figure 7. A flat slope of the regression indicates that the information sources whose reliability levels were not chosen for the introspective report have little effect on the subjective feeling when introspecting the influence of the chosen level of reliability.

References

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Comput Biol, 8*(11), e1002771. https://doi.org/10.1371/journal.pcbi.1002771

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, Feb). 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).

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010, Aug 27). 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, May). 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

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer.

Boldt, A., & Gilbert, S. J. (2022, Apr 27). 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, Jan). 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, Apr). 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, Jan 22). 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, Jun). 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, Jan). 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, Nov 1). 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, Mar). 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

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, May). 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, Aug 18). 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, Jan 24). 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, Feb). 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, Jan 18). 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, May 19). 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, Feb). 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, Jun). 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, Jun). 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, Aug). 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, Dec 19). 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

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2012, Aug 21). Effective integration of serially presented stochastic cues. *J Vis, 12*(8). https://doi.org/10.1167/12.8.12

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, Feb). 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

Kording, K. P., & Wolpert, D. M. (2004, Jan 15). 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, Apr 12). 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

Ma, W. J., Kording, K. P., & Goldreich, D. . (2023). *Bayesian models of perception and action: An introduction*. MIT press.

Maloney, L. T., & Mamassian, P. (2009, Jan-Feb). 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, Nov 23). 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, Jun). 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, May 9). 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, Feb). 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, Mar). 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, May). 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., Tanae, M., Ishii, K., & Takiyama, K. (2020, Jan 22). 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, Jul 28). 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, Jun). 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, Oct). 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, Mar). 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, Jan). 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, Oct). 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, Apr). 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, Sep). 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, Sep 1). 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, Dec). 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, Dec 29). 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, Aug 3). 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, Apr). 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, Jan). Factorial comparison of working memory models. *Psychol Rev, 121*(1), 124-149. https://doi.org/10.1037/a0035234

Van Hoeck, N., Watson, P. D., & Barbey, A. K. (2015). Cognitive neuroscience of human counterfactual reasoning. *Front Hum Neurosci, 9*, 420. https://doi.org/10.3389/fnhum.2015.00420

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

Vidal-Perez, J., Dolan, R., & Moran, R. (2024). Disinformation elicits learning biases. *Preprint at https://www.researchsquare.com/article/rs-4468218/v1*. https://doi.org/https://doi.org/10.21203/rs.3.rs-4468218/v1

Vosoughi, S., Roy, D., & Aral, S. (2018, Mar 9). The spread of true and false news online. *Science, 359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Wenke, D., Fleming, S. M., & Haggard, P. (2010, Apr). 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, Apr 14). 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, May 19). 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., Daw, N. D., & Maloney, L. T. (2015, Aug). Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions. *Nat Neurosci, 18*(8), 1152-1158. https://doi.org/10.1038/nn.4055

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, Sep 8). The bounded rationality of probability distortion. *Proc Natl Acad Sci U S A, 117*(36), 22024-22034. https://doi.org/10.1073/pnas.1922401117