



Optimistic Update Bias in Response to False Information

Hamid Razi^{1,2} · Thomas Sambrook¹ · Neil Garrett¹

Received: 11 November 2025 / Accepted: 21 September 2026
© The Author(s) 2026

Abstract

We investigate the degree to which labelling information as true or false after exposure to it can influence the degree to which it alters beliefs and test whether valence (whether the information is better or worse than expected) plays a key role in modulating this change. To investigate this, we adapt a classic belief updating paradigm by having participants estimate the likelihood of different life events occurring in the future and then varying information presented to participants along two dimensions: valence (whether the information is better or worse than thought) and accuracy (whether information is labelled true or false). By examining how much participants' beliefs change post exposure to the information, we find a significant effect along both dimensions. Specifically, participants change beliefs more following information labelled true compared to false and following information that is better compared to worse than expected. This latter effect is consistent with past studies that have revealed a valence effect in belief updating (in response to unlabelled information). The novel finding here is that this valence effect exists in response to false information, and we show evidence—using a Bayes Factor analysis—that it is just as strong in response to false as it is in response to true information. Computational modelling analysis suggests that these patterns of updating arise from the differential use of error signals during learning. Together, the results suggest that labelling information post exposure to it as false has the potential to curb the degree to which it can change beliefs. But this is likely to be less effective in cases where that information presents a rosier view of one's future.

Belief Updating, False Information, Optimistic Update Bias

Keywords Belief updating · False information · Optimistic update bias

Introduction

Unrealistic optimism (Weinstein, 1980) is the tendency to underestimate the likelihood of negative events (such as illness, accident, and natural disasters) and overestimate the likelihood of experiencing positive events such as professional success (Wiswall & Zafar, 2015). It impacts human decision making in domains ranging from politics to medical care (Paling, 2003; Staats et al., 2018; Tappin et al., 2017). Understanding the mechanism that generates unrealistic optimism is crucial for developing methods to mitigate

the bias, for example in relation to false beliefs that persist about the risk of climate change (Sunstein et al., 2017) or COVID-19 (Globig et al., 2022).

A core mechanism that has been shown to generate and maintain positive beliefs is asymmetric belief updating (Benjamin, 2019; Eil & Rao, 2011; Kappes et al., 2018; Korn et al., 2014; Kube & Rozenkrantz, 2021; Kuzmanovic et al., 2015, 2018; Mobius et al., 2011) whereby individuals change their beliefs to a greater extent when they receive information that is better than expected (e.g., learning that the likelihood of being a victim of card fraud is lower than expected) compared to worse than expected (learning it is higher). This asymmetry in belief updating can lead to optimistic beliefs, biased self-perceptions, and flawed financial predictions (Barber & Odean, 1999; Dunning et al., 2004; Eil & Rao, 2011; Greenwald, 1980; Kube et al., 2022; Kube & Rozenkrantz, 2021; Kuhnen, 2015; Peysakhovich & Karmarkar, 2016; Russo et al., 1996; Shefrin, 2009).

However, to date studies (Kappes et al., 2018; Kube et al., 2022; Kuzmanovic et al., 2015, 2016, 2018; Kuzmanovic &

✉ Hamid Razi
H.Razi@uea.ac.uk

¹ School of Psychology, University of East Anglia, Norwich Research Park, Norwich, UK

² Institut des Sciences Cognitives Marc Jeannerod, Centre National de la Recherche Scientifique, UMR 5229, Bron 69675, France

Rigoux, 2017; Ma et al., 2016; Oganian et al., 2019; Sharot et al., 2011) have examined optimistically biased updating in response to information that is accurate. The statistics presented to participants in the belief updating task – the main task used to examine optimistic belief updating (Kuzmanovic & Rigoux, 2017; Sharot et al., 2011; Sharot & Garrett, 2016, 2022) – have to date reflected genuine population base rates about the likelihood of different life events occurring in the future. Belief updating – measured as the change in participants’ estimates about the events occurring before and after the statistics are presented – occurs under the assumption (and made explicit in the task instructions) that the base rates can be trusted. This leaves an important open question: does the asymmetry exist in the face of information that is false? Here, by *false information*, we mean specifically a statistic presented to a participant as a base rate that does not in fact correspond to that event (a number that has been fabricated and explicitly flagged as inaccurate immediately after being shown). This is a narrow, content-free construct relative to the terms *misinformation* and *disinformation* (as they are typically studied elsewhere).

Answering this question is important in an age in which we can be exposed to a large amount of false information on a regular basis. This is partly the result of algorithms that power social media platforms that optimise on engagement over accuracy (Pennycook & Rand, 2021; Vosoughi et al., 2018) alongside a decreased cost to generating and disseminating false information (Ceylan et al., 2023; Pennycook & Rand, 2021). One proposal (Christner et al., 2024; Van Der Linden, 2024) to help safeguard against exposure to false information is to attach labels to information calling out cases where information is unreliable. However, this may be limited in its efficacy if biased belief updating remains in the face of information that is labelled false.

Here, we investigate whether this is the case by adapting the belief updating task. We adapt it by presenting explicit “true” and “false” labels alongside population base rates. We then test whether the characteristic optimistic belief updating pattern that has been found in the original version of the task is present under each of these labels. To do this, we formalise belief updating as a computational process driven by estimation errors: the discrepancy between prior beliefs and the information provided. The size of this error, scaled by a learning rate, determines how much beliefs shift. Modelling belief updating computationally in this way enables belief updating asymmetries that exist under true, false or both labels, to be understood as arising from an asymmetry in learning rates. Specifically, a larger learning rate for better-than-expected information than for worse-than-expected information (Garrett et al., 2014; Garrett & Sharot, 2017; Kuzmanovic & Rigoux, 2017; Sharot et al., 2011). We use the term *belief updating* throughout

to refer to the observed behavioural change between a participant’s first and second estimate, and reserve *learning rate* for the model parameter that governs how efficiently a given estimation error is converted into that change; the latter term follows standard usage in the reinforcement-learning literature (Sutton & Barto, 2018) and is not intended to imply that participants acquire durable general knowledge over the course of the task.

A considerable body of evidence from psychology and neuroscience has shown that learning rates can adapt in response to a range of factors, including: arousal (Filipowicz et al., 2020; Nassar et al., 2012), how volatile an environment we are in (Behrens et al., 2007; Pearce & Hall, 1980; Pulcu & Browning, 2019), perceived threat (Garrett et al., 2018), working memory (Cheng et al., 2024), and mood (Kao et al., 2023). Hence at the outset it stands to reason that learning ought to be attenuated in response to the presence of false information labels. However, learning is often not completely curtailed when people are explicitly told information they have received is either irrelevant or false (Ecker et al., 2010; Porter & Wood, 2021; Ross et al., 1975; Tversky & Kahneman, 1974). A well-known case is the continued influence of misinformation (Ecker et al., 2010) whereby individuals continue to recall narratives they encountered that were subsequently retracted or discredited.

We hypothesised that failures to discount false information might not in fact be uniform but could depend on valence (whether the false information was better or worse than initially expected). This follows from the theory that people maintain positive beliefs (via asymmetric learning) in part because these beliefs elicit positive feelings (and thus have positive utility). This is in contrast to negative beliefs that elicit negative feelings (and negative utility) (Brunnermeier & Parker, 2005; Loewenstein, 2006). If positive beliefs do indeed have value in and of themselves, one ought to be less motivated to attenuate learning upon discovery that information is false in cases where information was better compared to worse than expected, just as one is motivated to ignore accurate bad news in the original belief updating task.

How to interpret differences in belief updating between information classified as “good” versus “bad” (as well as how exactly to classify information as good and bad) in the belief updating task has been contested. Whilst some (Garrett & Sharot, 2014, 2017, 2023; Sharot et al., 2011; Sharot & Garrett, 2016, 2022) argue that this pattern is indicative of a bias, others have argued that this pattern can in fact arise from (unbiased) perfectly rational Bayesian updating (Burton et al., 2022; Harris et al., 2013; Shah et al., 2016). To acknowledge and speak to this ongoing debate, we analyse the data using both approaches: one advocated by Sharot & Garrett and another proposed by Shah et al. Finally,

we included measures of depression, obsessive-compulsive symptoms and schizotypal traits to be able to explore the extent to which these might explain heterogeneity in belief updating. Previous studies have reported that optimistic belief updating varies with psychiatric symptom severity in mood and anxiety disorders (Garrett et al., 2014, 2018; Khalid et al., 2024; Korn et al., 2014; Ossola et al., 2020) whilst obsessive-compulsive disorder and schizophrenia are characterised, to differing degrees, by alterations in belief evaluation (Woodward et al., 2006), uncertainty processing and the maintenance of implausible beliefs (Fradkin et al., 2020; Knowles & Olatunji, 2023), making them plausible candidates for exploratory analyses in the context of learning from false information.

Methods

Participants

One hundred and twenty-seven students were recruited from the University of East Anglia to participate in the study. Nineteen were excluded from all analysis (one participant did not pass attention checks during the task, three did not complete the task, 13 failed catch questions in the questionnaires, and three did not have trials we could assign to each of our four experimental conditions – one of whom also failed the catch questions). The final sample size was

108 (mean (SD) age: 19.48 (1.51)). Power analysis based on effect sizes from a previous study ($d=0.40$) using this paradigm (Garrett & Sharot, 2017) indicated that a sample size of 80 was sufficient to detect an effect size with 95% power and an alpha level of 0.05 (paired sample t-test comparing good versus bad news). We increased the sample size further to enable us to use Bayes Factors where a sample size of at least 100 participants is recommended (Fu et al., 2021, 2022) to infer evidence for/against the null in comparing optimistic update bias for false versus true information. The study was approved by the University of East Anglia's ethics committee. Participants received study credits as compensation for participating in the study.

Behavioural Task and Stimuli

The task (see Fig. 1) was adapted from a belief updating task previously used to investigate optimistic updating biases (Khalid et al., 2024; Kuzmanovic et al., 2019; Kuzmanovic & Rigoux, 2017; Ma et al., 2016; Sharot et al., 2011; Sharot & Garrett, 2016) and followed best practices for using the task (Sharot & Garrett, 2022).

Stimuli

Stimuli consisted of two lists of 25 different negative life events (e.g., domestic burglary — see Supplementary Table 2 for events used). Events and their base rates were obtained

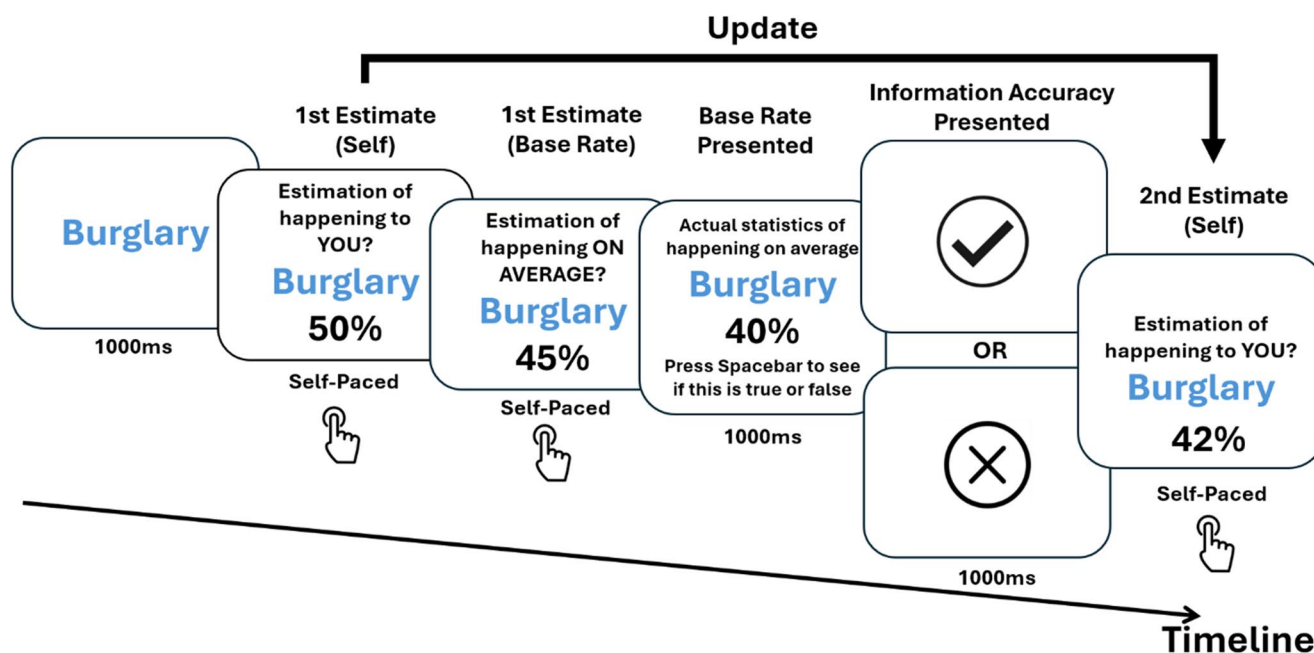


Fig. 1 Behavioural task. On each trial, participants were presented with a short description of an adverse event and asked to imagine the event happening to them in the future. They were then asked to estimate how likely this event was to occur to them personally in the future and how likely the event was to happen on average in a demographically similar

population, i.e., the base rate. The order of these two estimates was randomised. They were then shown the base rate which was followed by a cue indicating whether this statistic was true (tick) or false (cross). Finally, participants were asked to provide a second estimate of how likely this event was to occur to them in the future

from stimuli used in the original study (Sharot et al., 2011), which compiled the base rate of each event occurring at least once to someone from the United Kingdom from reputable online resources (including the Office for National Statistics and PubMed). Very rare or very common events were not included; all base rates lay between 10 and 70% and were normally distributed around the midpoint of the range (Sharot & Garrett, 2022), see Supplementary Fig. 2. To ensure that the range of possible overestimation was equal to the range of possible underestimation, participants were told that the range of base rates lay between 3 and 77%, and they were only permitted to enter estimates (for both base rates and self-estimates) within this range. For each participant one of the two lists was randomly chosen to carry true base rates and one to carry false base rates. For the latter, the base rates were randomly shuffled between events, e.g., the likelihood for domestic burglary was shown for bicycle theft. This procedure meant that the statistical properties of the base rates (the median, range, distribution, etc.) remained unchanged. The two lists were then merged to create a final list of 50 events with 25 accurate base rates and 25 false base rates.

Behavioural Task

On each trial, one of the 50 events was shown. Participants were then prompted to estimate how likely the event was to happen to them personally in the future (E_1) and how likely it was to occur on average in a demographically similar population (eBR; their estimate of the base rate). In half of the trials, the order of the two estimations (E_1 and eBR) was reversed (i.e. eBR followed by E_1). After this, the base rate (BR) was provided. Participants were then prompted to press the spacebar to see whether this base rate was true or false. They then saw a stimulus indicating whether the statistic they had been shown was true (tick) or false (cross). Finally, they were asked to estimate how likely the event was to happen to them again (E_2). There were no time constraints for submitting responses (E_1 , eBR, E_2). The task was created using Qualtrics.

Questionnaires

At the end of the experiment, participants were asked to complete three psychiatric questionnaires: the Beck Depression Inventory (Beck et al., 1961), the Obsessive Compulsive Inventory (Foa et al., 2002), and the Schizotypy short scale (Mason et al., 2005). These measures were included as exploratory analyses because prior work has shown that symptoms of depression are associated with belief updating in the original paradigm (Garrett et al., 2014; Korn et al., 2014), while obsessive-compulsive and schizotypal traits

have been linked to altered belief updating more broadly (Reed et al., 2020). In each of the questionnaires, a catch question was included, which was used to exclude inattentive participants (Zorowitz et al., 2023). Participants ($N=13$) were excluded if they got 1 or more of the catch questions wrong.

Controls

At the end of the experiment, participants rated stimuli on six-point scales for Negativity (“How negative you found this event?”, 1=Not at all to 6=Very), Past Experience (“Has this event happened to you before?”, 1=never to 6=very often), Vividness (“How vividly could you imagine this event?”, 1=not vivid to 6=very vivid), Familiarity (“Regardless of if this event has happened to you before, how familiar do you feel it is to you from TV, friends, movies and so on?”, 1=not at all familiar to 6=very familiar); and Arousal (“When you imagine this event happening to you how emotionally arousing is the image in your mind?”, 1=not arousing at all to 6=very arousing). These ratings were included as covariates in control analysis. Descriptive statistics for updates, estimation errors, trial counts and these subjective ratings, broken down by trial type, are reported in Supplementary Table 3.

Behavioural Analysis

Trials were categorised into four types (Good News True, Bad News True, Good News False and Bad News False; see Supplementary Fig. 1) and analysed using Linear Mixed Models (LMMs) testing for effects of valence (good/bad), accuracy (true/false). To acknowledge and speak to the debate in the literature about whether belief updating patterns in the belief updating task we use here can be read as evidence for a bias, we analyse the data using two approaches: one advocated by Sharot & Garrett and another proposed by Shah et al. Accuracy (True/False) was determined in both approaches according to whether the trial presented a true or false base rate. But how trials were classified into good/bad and the dependent variable used in the LMM differed between the two.

All models were fitted in R using the lme4 package (Bates et al., 2015) with significance tests implemented by lmerTest (Kuznetsova et al., 2017). To interpret and visualize the results of the LMMs, we extracted the Estimated Marginal Means (EMMs) using the emmeans package in R (Lenth et al., 2026). Unlike raw descriptive averages, EMMs represent the predicted means for each condition marginalized over all other predictors and covariates in the model. These adjusted means and their corresponding 95% confidence intervals were extracted for plotting. Further, we used the emmeans

package to conduct post-hoc pairwise comparisons – paired t-tests - between Good and Bad news for True and False information. The package uses the Kenward-Roger approximation to calculate the p-values (Kenward & Roger, 1997).

All models were fitted with frequentist LMMs (lme4/lmerTest), which we use for all tests of main effects, interactions and simple effects. Because a non-significant frequentist interaction cannot distinguish evidence for the null from insensitive data, we additionally refitted the same models in a Bayesian framework solely to compute Bayes Factors (BFs) quantifying evidence for or against the Valence $\times$ Accuracy interaction (i.e. whether the strength of the valence effect was the same for true and false trials). The two approaches address complementary questions, and the conclusions do not depend on the choice of framework. Specifically, we used the brms package (Bürkner, 2017) with weakly informative priors, 16,000 iterations (3,000 for warmup), and four chains to generate posterior distributions for all model parameters. These were then used to generate BFs in bayestestR (Makowski et al., 2019). We used Jeffreys' scale to interpret BFs (Jeffreys, 1961; Wagenmakers et al., 2011), according to which:

- $\text{BFs} < 1/30$ indicate extreme evidence supporting the null hypothesis
- $1/30 < \text{BF} < 1/10$ indicate strong evidence in favour of the null hypothesis
- $1/10 < \text{BF} < 1/3$ indicate moderate evidence in favour of the null hypothesis
- $1/3 < \text{BF} < 1$ indicate anecdotal evidence in favour of the null hypothesis

Approach One

Based on the methodology advocated by Sharot and Garrett (Sharot & Garrett, 2022), we classified good/bad information based on whether participants underestimated (bad news) or overestimated (good news) the likelihood of an event happening to them personally (i.e. E_1) relative to the provided base rate (BR). Trials where the initial self-estimate (E_1) was equal to the base rate ($E_1 = \text{BR}$), $\sim 2\%$ of total trials, were excluded as they could not be categorised. An update score was calculated as the difference between first and second estimates. This difference was calculated such that a positively signed value indicated a change in initial self-estimates towards the provided base rate and a negatively signed value indicated a change away from it. In practice this meant that for Good News trials, update was calculated as first estimate minus second estimate and for Bad News trials second estimate minus first estimate.

Update scores were entered as the dependent variable in a LMM, with Valence (coded as 1=good news, -1=bad

news) and Accuracy (1=true, -1=false) as fixed effects, along with their interaction. Random intercepts and random slopes for Valence, Accuracy, and their interaction were included by participant. The model was implemented in the syntax of R as follows:

LMM1: $\text{Update} \sim \text{Valence} * \text{Accuracy} + (1 + \text{Valence} * \text{Accuracy} \mid \text{Participant})$

Next, we ran a second LMM and included potential confounds (Sharot & Garrett, 2022). Since the difference between the initial self estimate (E_1) and the provided base rate (BR) imposes a ceiling on update scores, for each trial, an estimation error (EE) term was calculated as the absolute difference between the presented base rate (BR) and the participant's initial estimate (E_1) for that trial:

$$EE = |E_1 - \text{BR}|$$

EE on each trial and the five subjective ratings for each event shown on each trial (Vividness, Past Experience, Arousal, Familiarity and Negativity) were entered as additional fixed effects in the model, fitted with random intercepts and slopes by participant. The model was implemented in the syntax of R as follows:

LMM2: $\text{Update} \sim \text{Valence} * \text{Accuracy} + \text{EE} + \text{Vividness} + \text{Past_Experience} + \text{Familiarity} + \text{Arousal} + \text{Negativity} + (0 + \text{Valence} + \text{Accuracy} + \text{Valence} * \text{Accuracy} \mid \text{Participant})$

Note that, following best practices (Matuschek et al., 2017), we opted for a reduced version of the maximal random effects structure due to the singularity issue. Full results of this model, separately for Approach One and Approach Two, are reported in Supplementary Table 1.

Approach Two

In the second approach, we followed the methodology proposed by Shah et al. (2016) and Burton et al. (2022), see also Harris et al. (2013). Here we classified good/bad information based on whether participants underestimated (bad news) or overestimated (good news) the base rate (i.e. eBR) relative to the provided base rate (BR). Trials where the base rate estimate (eBR) was equal to the base rate ($\text{eBR} = \text{BR}$), 1.51% of total trials, were excluded as they could not be categorised.

Next, we calculated a rational Bayesian update for each trial to serve as a benchmark for evaluating participants' actual updates. Following the framework established by Shah et al. (2016) and Burton et al. (2022), the Bayesian

benchmark assumes that participants have specific personal knowledge about their own risk, which must be combined with the objective base rate.

First, we quantified the participant's personal knowledge as a Likelihood Ratio (LR_{personal}). This is calculated by taking the ratio of the odds of their initial self-estimate (E_1) to the odds of their estimated base rate (eBR):

$$LR_{\text{personal}} = \text{Odds}(E_1) / \text{Odds}(eBR) \\ = [E_1 / (1 - E_1)] / [eBR / (1 - eBR)]$$

LR_{personal} captures how much more or less likely the participant believes they are to experience the event compared to the average person.

When the base rate (BR) is presented, a rational Bayesian agent treats the odds of the base rate as the new prior and applies their personal knowledge (LR_{personal}) to compute the predicted posterior odds of their own risk:

$$\text{Odds}(E_{2,\text{predicted}}) = \text{Odds}(BR) \times LR_{\text{personal}} \\ = [BR / (1 - BR)] \times [E_1 / (1 - E_1)] / [eBR / (1 - eBR)]$$

We then convert the predicted posterior odds back into a probability to obtain the rational Bayesian posterior estimate ($E_{2,\text{predicted}}$):

$$E_{2,\text{predicted}} = \text{Odds}(E_{2,\text{predicted}}) / (1 + \text{Odds}(E_{2,\text{predicted}}))$$

Finally, we calculate the magnitude of the Bayesian Update by taking the difference between the initial estimate (E_1) and the predicted rational estimate ($E_{2,\text{predicted}}$). To ensure the update is consistent with the valence of the information, it is calculated as:

$$\text{Update}_{\text{Bayesian}} = E_1 - E_{2,\text{predicted}}$$

if $BR < eBR$ (Good News)

$$\text{Update}_{\text{Bayesian}} = E_{2,\text{predicted}} - E_1$$

if $BR > eBR$ (Bad News)

The participant's Observed Update on each trial is calculated as the change between their first (E_1) and second (E_2) estimates, calculated to match information valence:

$$\text{Update}_{\text{Observed}} = E_1 - E_2$$

$$\text{Update}_{\text{Observed}} = E_2 - E_1$$

Participants' actual updating was then evaluated against this benchmark by calculating a Bayesian deviation score

on each trial, the difference between the Bayesian Update and the participant's Observed Update magnitude (i.e. $\text{Update}_{\text{Bayesian}} - \text{Update}_{\text{Observed}}$). The sign of the deviation score indexes how a participant weights evidence relative to a rational agent. Positive score ($\text{Update}_{\text{Bayesian}} > \text{Update}_{\text{Observed}}$) means less updating than the Bayesian benchmark whilst a negative score ($\text{Update}_{\text{Bayesian}} < \text{Update}_{\text{Observed}}$) means *more* updating. A score of zero indicates perfect alignment with the Bayesian benchmark.

Examining these deviation scores as a function of information valence (Good vs. Bad News) allows us to quantify valence-dependent updating biases - such as the optimism bias - while controlling for structural differences in the probability scale and the inherent statistical asymmetry of prior beliefs (Shah et al., 2016).

Bayesian deviation scores were entered as the dependent variable in a LMM, with Valence (coded as 1=good news, -1=bad news) and Accuracy (1=true, -1=false) as fixed effects, along with their interaction. Random intercepts and random slopes for Valence, Accuracy, and their interaction were included by participant. The model was implemented in the syntax of R as follows:

LMM3: Bayesian Deviation Score ~ Valence * Accuracy + (1 + Valence * Accuracy | Participant)

Finally, we ran a fourth LMM and now included potential confounds: BR_{error} (calculated as: $BR_{\text{error}} = |eBR - BR|$), Vividness, Past Experience, Arousal, Familiarity and Negativity. The model was implemented in the syntax of R as follows:

LMM4: Bayesian Deviation Score ~ Valence * Accuracy + BR_{error} + Vividness + Past Experience + Familiarity + Arousal + Negativity + (0 + Valence + Accuracy + Valence * Accuracy | Participant)

Whilst calculating the Bayesian benchmark for true trials follows logically from Shah et al. (2016) (assuming that Bayesian agents fully trust information explicitly labelled true in this version of the belief updating task), for false trials there is no obvious normative standard as to how a rational Bayesian agent ought to update their beliefs. Rather than assuming agents systematically adjust the (false) base rate upwards or downwards by some fixed amount, our primary analyses assume that the Bayesian agent combines the presented base rate with their prior beliefs in the same manner as for true information. This is because participants are only told that the presented base rate is false (without any information about the direction or magnitude of its deviation from the true value). Nonetheless, the deviation scores for false trials should not as readily be interpreted as indexing how

close/far participants are from a rational Bayesian agent in the same way as they can be for true trials because the normative benchmark itself depends on assumptions one makes about how false information ‘ought’ to be incorporated. To assess the sensitivity of our results to this assumption, we repeated this analysis but included a scaling constant (C) which was applied to the BR odds on false trials, such that $\text{Odds}(\text{BR})_{\text{adjusted}} = \text{Odds}(\text{BR}) \times C$, before combining with $\text{LR}_{\text{personal}}$ to calculate the normative Bayesian benchmarks and re-examined the pattern of results (for $C=0.5$, $C=0.8$, $C=1.2$ and $C=1.5$).

Computational Modelling

We adapted the computational model used by Kuzmanovic and Rigoux (2017, see also Sharot & Garrett, 2022) to test if the relationship between trial-by-trial estimation errors and subsequent updates was modulated by accuracy and valence. We note that the models here do not attempt to make claims as to what time points in the task belief change occurs; we apply these models primarily as an analytical tool to understand the degree to which error signals in the task correspond to subsequent belief change and test whether this relationship is modulated by accuracy, valence and their interaction.

In each model, update on each trial (t), i.e. the change between participants’ first and second self estimates [calculated as: update (good news) = first estimate - second estimate; update (bad news) = second estimate - first estimate] was predicted via the following equation:

$$\text{Update}_t = \alpha * \delta_t * (1 - rP_t * w)$$

rP (relative personal knowledge) for each item was calculated by comparing the estimated base rate (eBR) and the first estimate (E_1) as:

$$rP_t = (eBR_t - E_{1t}) / (eBR_t - 3)$$

if $E_1 < eBR$

$$rP_t = (E_{1t} - eBR_t) / (77 - eBR_t)$$

if $E_1 > eBR$

$$rP_t = 0$$

if $E_1 = eBR$

rP ranges from 0 to 1. $rP=0$ indicates the person does not see themselves as different from the average. $rP=1$ indicates the greatest disparity possible between the individual’s

perception of the likelihood of the event occurring to themselves and the average (which might occur if say the individual has a range of reasons – such as specific lifestyle choices, family history, etc.).

w is a free parameter which indexes the degree to which rP impacts belief updating. When $w=0$, rP has no influence on belief updating. When $w=1$, rP has maximum influence on belief updating.

α is a learning rate, which governs the degree to which participants updated their beliefs in response to the size of the estimation error (Sharot et al., 2011; Garrett et al., 2014, 2018). We tested 4 different models with between two and four learning rates. By varying the number of learning rates we were able to test if estimation errors were integrated to differing degrees depending on valence, accuracy or both factors.

δ is an error term. When running the models following approach one we followed past attempts by Sharot and Garrett to relate errors to updating in the belief updating task by setting this equal to the estimation error on each trial (i.e. $\delta = EE = |E_t - BR|$). When running the models under approach two we followed the approach advocated by Shah et al., 2016 and used by Kuzmanovic and Rigoux (2017) by setting this equal to the base rate estimation error on each trial (i.e. $\delta = BR_{\text{error}} = |eBR - BR|$).

The four models were formulated as follows:

Model 1 (M1)

$$\text{Update}_t = \alpha_{\text{true}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue = True

$$\text{Update}_t = \alpha_{\text{false}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue = False

Free parameters¹ ($n=4$): α_{true} , α_{false} , w , σ

Model 2 (M2)

$$\text{Update}_t = \alpha_{\text{true}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue = True

$$\text{Update}_t = \alpha_{\text{false, goodnews}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue = False and valence = Good

$$\text{Update}_t = \alpha_{\text{false, badnews}} * \delta_t * (1 - rP_t * w)$$

¹ σ is not part of the learning models tested but is used in the model fitting procedure, outlined in the next section.

if information accuracy cue=False and valence=Bad

Free parameters* ($n=5$): α_{true} , $\alpha_{\text{false, goodnews}}$, $\alpha_{\text{false, badnews}}$, w , σ

Model 3 (M3)

$$Update_t = \alpha_{\text{false}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue=False

$$Update_t = \alpha_{\text{true, goodnews}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue=True and valence=Good

$$Update_t = \alpha_{\text{true, badnews}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue=True and valence=Bad

Free parameters* ($n=5$): α_{false} , $\alpha_{\text{true, goodnews}}$, $\alpha_{\text{true, badnews}}$, w , σ

Model 4 (M4)

$$Update_t = \alpha_{\text{true, goodnews}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue=True and valence=Good

$$Update_t = \alpha_{\text{true, badnews}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue=True and valence=Bad

$$Update_t = \alpha_{\text{false, goodnews}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue=False and valence=Good

$$Update_t = \alpha_{\text{false, badnews}} * \delta_t * (1 - rP_t * w)$$

if information accuracy cue=False and valence=Bad

Free parameters* ($n=6$): $\alpha_{\text{false, goodnews}}$, $\alpha_{\text{false, badnews}}$, $\alpha_{\text{true, goodnews}}$, $\alpha_{\text{true, badnews}}$, w , σ

In each model, we converted the predicted Update on each trial into a predicted 2nd Estimate ($\widehat{E}_2$) by adding or subtracting the predicted Update from participants' 1st Self Estimate (E_1). This depended on whether the base rate (BR) presented on that trial was lower than E_1 (first approach) or eBR (second approach) - in which case beliefs shift down - or above E_1 (first approach) or eBR (second approach) - in which case beliefs shift up.

Therefore, in the first approach (Sharot & Garrett, 2022) we have:

$$\widehat{E}_2 = E_{1t} + Update_t$$

if $BR_t > E_{1t}$

$$\widehat{E}_2 = E_{1t} - Update_t$$

if $BR_t < E_{1t}$

And in the second approach (Shah et al., 2016):

$$\widehat{E}_2 = E_{1t} + Update_t$$

if $BR_t > eBR_t$

$$\widehat{E}_2 = E_{1t} - Update_t$$

if $BR_t < eBR_t$

When running the models under approach one, valence (whether information was classified as “good” or “bad”) was based on whether participants underestimated (bad news) or overestimated (good news) their own likelihood (i.e. E_1) relative to the provided base rate (BR). When running the models under approach two, we used the classification procedure advocated by Shah et al. (2016) where we classify good/bad information based on whether participants underestimated (bad news) or overestimated (good news) the base rate (i.e. eBR) relative to the provided base rate (BR).

Model Fitting Procedure

2nd estimates predicted on each trial ($\widehat{E}_2(t)$) were compared to participants' actual 2nd estimates ($E_2(t)$) to find the best fitting parameters in the model fitting process.

We fitted the models hierarchically using Expectation Maximization (EM) algorithm (Huys et al., 2011) in the Julia language (Bezanson et al., 2012) version 1.9.4. Hierarchical parameter estimation has been shown to provide superior cross-validation performance on unobserved data (Scheibehenne & Pachur, 2015).

To find the best fitting set of parameters for each model (given the data we observed) using the EM algorithm, we used a log-likelihood function. For each observation, we calculated the probability density of the observed second estimate (E_2) given the model's predicted second estimate and other parameters. Specifically, we modelled E_2 as following a normal distribution (similar to Nassar et al., 2021) but truncated the distribution such that it was bounded between 3 and 77 (the range of possible values in our task):

$$\log L = \sum_{t=1}^N \log \left[\frac{\Phi \left(\frac{E_2(t) - \widehat{E}_2(t)}{\sigma} \right)}{\sigma \left[\Phi \left(\frac{77 - \widehat{E}_2(t)}{\sigma} \right) - \Phi \left(\frac{3 - \widehat{E}_2(t)}{\sigma} \right) \right]} \right]$$

Where:

- $E_2(t)$ is the observed second estimate for trial t
- $\widehat{E}_2(t)$ is the model's predicted estimate for trial t
- σ is the standard deviation as a free parameter
- Φ is the standard normal cumulative distribution function (CDF)
- φ is the standard normal probability density function (PDF)

To check the robustness of the results, we separately fit the four models in a Bayesian framework using Stan (Carpenter et al., 2017). This allowed us to estimate the complete joint posterior distribution of all parameters, better accounting for the uncertainty around the parameter estimation, and compute expected log predictive density (ELPD) scores which we used as a complementary approach to model comparison. ELPD provides a measure of a model's out-of-sample predictive accuracy (Vehtari et al., 2017). For each model, we ran 4 parallel chains with 4,000 iterations per chain. The first 2,000 iterations of each chain were discarded as warmup, resulting in a total of 8,000 posterior samples. To ensure stable convergence and prevent divergent transitions, we increased the target average proposal acceptance probability (`adapt_delta`) to 0.99 and set the maximum tree depth to 15. Model convergence was confirmed for all parameters, with all Gelman-Rubin convergence diagnostics ($R^{\wedge}$) falling below 1.01 and effective sample sizes indicating reliable estimation in accordance with the best practices (Vehtari et al., 2017).

Model Comparison

To compare the different models we followed two approaches. First we calculated the difference in ELPD scores from the Stan pipeline (denoted ΔELPD) between the top-performing model and each competing model, alongside the standard error of that difference (`se_diff`). Positive ΔELPD values indicate that the top-performing model provided a better fit to the data than the competing model, with larger values reflecting a bigger difference in fit between the two models. We also calculated unbiased subject-level leave-one-out cross-validation (LOOcv) scores from the em pipeline for each participant for each model. LOOcv scores (from the em fitting pipeline) and ELPD scores (from the Stan pipeline) were entered into the mbb-vb-toolbox in MATLAB (Daunizeau et al., 2014). This toolbox uses the Variational Bayesian Approach (VBA), where models are handled as random effects that may vary across subjects, without assuming a known population distribution (instead, the population distribution is characterised in terms of model frequencies). Model frequency represents the posterior probability that a particular

model generated the data for a randomly chosen subject. This value should be compared to chance level (1 divided by the total number of models, so in our study $1/4=25\%$). The toolbox also provides a measure of exceedance probability that measures the probability that a specific model outperforms all other models in the set - essentially quantifying belief that a model has the highest frequency among all competing models (Daunizeau et al., 2009).

Statistical Tests on the Learning Rates

Once the winning model was identified, we reparametrized it such that one parameter indexed the optimistic update bias for true information, one for false information, and a third captured their interaction. We then estimated these parameters again using hierarchical EM. This regularisation shrinks individual estimates toward the group mean, which reduces variance and can artificially inflate statistical significance if standard tests are applied. Moreover, these estimates are not independent across participants, violating a key assumption of traditional frequentist inference. As a result, applying standard t-tests to hierarchical estimates can lead to biased p-values and inflated false positives (Piray et al., 2019). Instead, therefore, we used hierarchical t-tests, which account for this regularisation. This test uses the posterior distribution over each group-level parameter to compute a t-value and corresponding credible interval, based on the estimated group mean and its hierarchical standard error (Piray et al., 2019). We used these hierarchical t-tests to assess whether the difference in good news vs. bad news differed significantly from zero, providing statistical evidence for optimistic update bias in response to true and/or false information (see the [Supplementary Material](#) for the reparameterization).

Model Recovery

We ran model recovery to validate the degree to which we could robustly identify each of our 4 models (M1-M4). The basic logic of this approach is that if the underlying data has verifiably been generated by one of the four specific models, this model should outperform the other 3 in model comparison.

We performed model recovery separately under both Approach One and Approach Two, since the two approaches classify trials as good or bad news differently and use different error terms (EE and BR_{error}). Therefore, they could differ in how distinguishable the four models are from one another. For each approach, we first simulated 120 synthetic participants completing 50 trials under each of the four models, using the parameter ranges observed when that model was fitted to the real data. This generated four synthetic datasets per approach. We then fitted each of these

four synthetic datasets to each of the four models and calculated the resulting LOOcv scores. We fed these LOOcv scores into the Variational Bayesian Analysis toolbox (VBA) in MATLAB (Daunizeau et al., 2014) to obtain, for each simulated dataset, the exceedance probability (XP) and estimated frequency of each of the four candidate models. The best-fitting model for a given dataset was determined based on the highest exceedance probability and model frequency (this is the same approach we followed when fitting the actual data to each of the four models and comparing them). This process was repeated 50 times per approach, and for each iteration we recorded whether the model used to simulate the data was correctly identified as having the highest exceedance probability, creating a confusion matrix for each approach (e.g., if data were simulated using model M1, M1 should have the highest XP). We also recorded the estimated model frequency for each candidate model at each iteration, to assess how often, out of 50 iterations, the model that generated the data turned out as the most frequent. Confusion matrices and model frequency distributions for both approaches are shown in Supplementary Fig. 3.

Parameter Recovery

To evaluate parameter identifiability, we conducted a parameter recovery analysis on the winning model (M4), separately for Approach One and Approach Two, since the group-level parameters of the winning model were estimated independently under each approach. For each approach, we simulated behaviour for 200 synthetic participants, each completing 200 trials, using parameter values randomly drawn from uniform distributions spanning the widest ranges possible. The same model-fitting procedure used for real participants was then applied to each simulated dataset. Recovery success was assessed in two ways. First, by computing the Pearson correlation between each parameter's true (simulated) value and its recovered (estimated) value, to establish whether that parameter could be accurately recovered from the fitting procedure. Second, by computing the Pearson correlation between different parameters within the recovered estimates, to establish whether parameters were traded off against one another during estimation. For each approach, these correlations are displayed together in a single correlation matrix, with diagonal values reflecting recovery accuracy and off-diagonal values reflecting cross-parameter confounds, shown separately for Approach One and Approach Two in Supplementary Fig. 4.

Simulations

To evaluate each model's ability to reproduce the patterns of belief updating, we simulated data for each of the four models

(M1–M4) using the mean group-level parameters obtained when that model was fitted to the real data. For each model, we generated data for 120 synthetic participants, each completing 50 trials, with trial characteristics (base rates, the number of true/false trials, and the range of estimations) matched to the structure and range of the actual task. We then analysed the simulated data in the same way as the real behavioural data, separately under both Approach One and Approach Two. Under Approach One, belief updating was quantified for each simulated participant as the raw difference between their first (pre-information) and second (post-information) estimates, and trials were classified as good or bad news according to whether the simulated participant's first estimate fell above or below the base rate provided (following Sharot & Garrett, 2022). Further, estimation error (EE) was defined as the absolute difference between the first estimate and the provided base rate. Under Approach Two, belief updating was quantified as the difference between a normative Bayesian agent's update and the simulated participant's update, and trials were classified as good or bad news according to whether the simulated participant's estimate of the base rate fell above or below the actual base rate provided (following Shah et al., 2016) while the absolute difference between estimated base rate and the provided base rate was defined as the base rate error (BR_{error}). For each model and each approach, we fitted the same linear mixed-effects model used for the real data to the simulated dataset, with belief updating as the dependent variable, accuracy (true/false) and valence (good/bad news) as fixed effects, and participant as a random effect. Within each model's simulated data, we tested for a significant main effect of accuracy from the mixed-effects model and, using two-tailed paired t-tests, for optimistic update bias (comparing simulated updating following good versus bad news, separately for true and false information), mirroring the statistical approach applied to the real behavioural data. This allowed us to formally establish, for each model and under both analytical approaches, whether it reproduced not only the qualitative direction but also the statistically reliable pattern of accuracy and valence effects observed in the real data.

Results

Positively biased belief updating We examined whether valence (information being positive or negative) as well as the information labels impacted the degree to which beliefs changed using two separate approaches. One consistent with Sharot and Garrett (2022) and another in line with a proposal by Shah et al. (2016).

The first approach (Fig. 2a) used the methodology proposed by Sharot and Garrett (2022). This quantifies belief

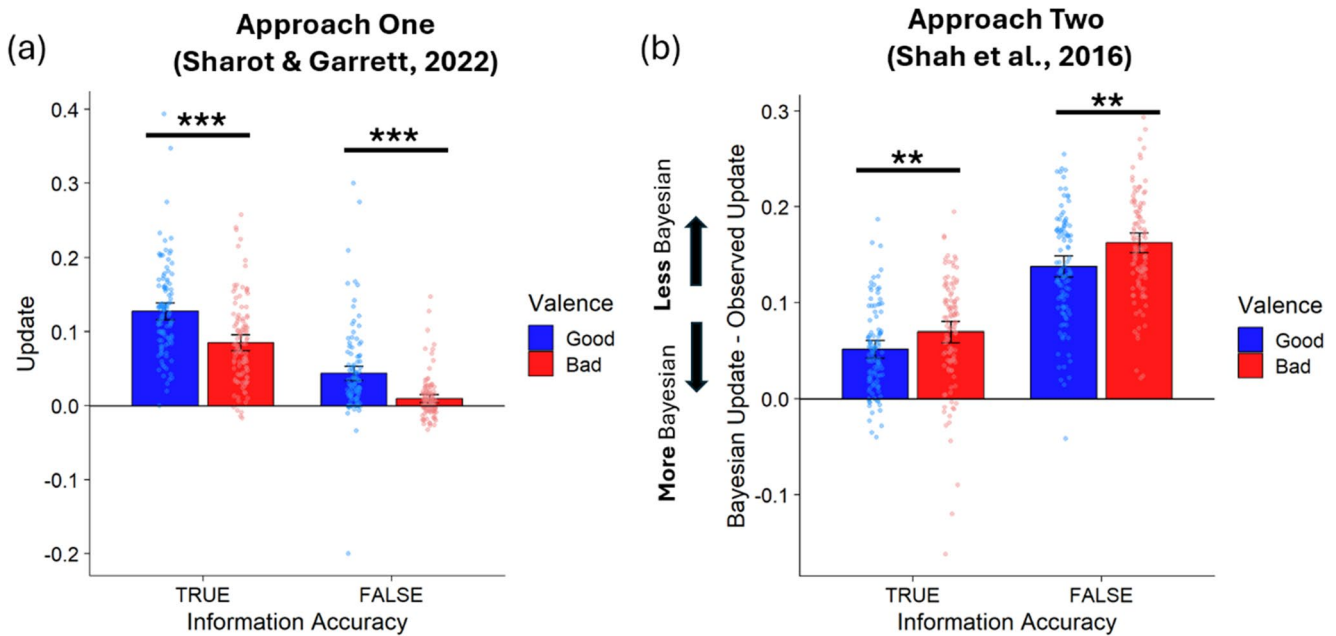


Fig. 2 Belief updating patterns observed under two analytical approaches. **(a)** Approach One (Sharot & Garrett, 2022) compares the raw difference between participants' first (pre information) and second (post information) estimates. Learning was attenuated under false information labels but participants updated to a greater degree following good compared to bad information under both true and false labels. **(b)** Approach Two (Shah et al., 2016) compares how participant updates to a normative Bayesian Agent (the closer to 0 the scores, the

closer to the Bayesian benchmark). Participant updates were closer to a Bayesian Agent following true compared to false information but participants were closer to the Bayesian benchmark following good compared to bad information under both true and false labels. Bars represent the estimated marginal means from linear mixed-effects models. Error bars represent 95% confidence intervals. ** $p < 0.01$, *** $p < 0.001$ two-tailed paired t-test

change as the raw difference between participants first (pre information) and second (post information) estimates and categorises trials as good or bad based on a comparison between participants first estimate of their own likelihood and the base rate provided. This analysis revealed a main effect of valence ($t(101.32) = 7.47$, $p < 0.001$), with greater updating for good news compared to bad news, and a main effect of accuracy ($t(107.84) = 18.18$, $p < 0.001$), with greater updating for true compared to false trials. There was no interaction between valence and accuracy ($t(128.63) = 1.13$, $p = 0.25$). Follow-up paired t-tests confirmed that participants updated more following good compared to bad news for both true ($t(107) = 5.75$, $p < 0.001$) and false ($t(107) = 6.83$, $p < 0.001$) information. To interrogate whether the lack of an interaction provided evidence in favour of the null (i.e., that the valence effect was similarly strong for true and false trials), we conducted a Bayes Factor analysis (see Methods). This revealed extreme evidence (Jeffreys, 1961; Wagenmakers et al., 2011) in favour of the null hypothesis ($BF_{10} = 0.030$) suggesting that the positive valence effect for false trials was of a similar magnitude to the positive valence effect for true trials. Repeating this analysis controlling for potential confounds revealed the same pattern of results (see Supplementary Table 1). This revealed a main effect of valence ($t(98.86) = 10.17$, $p < 0.001$), a main effect

of accuracy ($t(107.15) = 19.18$, $p < 0.001$) and no interaction ($t(105.85) = 1.26$, $p = 0.20$).

The second approach (Fig. 2b) followed the methodology proposed by Shah et al. (2016, see also Burton et al., 2022 and Garrett & Sharot, 2023). Since participants provide an estimate of the base rate as well as a self-estimate on each trial a benchmark of how a normative Bayesian agent would update can be computed (see Methods). We then compare the difference between this Bayesian benchmark and participants actual updates for the different trial types. In this analysis, trials are categorised as good or bad based on a comparison between participants estimate of the base rate and the base rate provided (in line with Shah et al., 2016).

The results under this second approach were consistent with the results under approach one. Specifically, we again observed a main effect of accuracy ($t(106.82) = -17.31$, $p < 0.001$) suggesting participants updated closer to a Bayesian agent for true compared to false information; a main effect of valence ($t(107.78) = -3.78$, $p < 0.001$) suggesting participants' updating was closer to a Bayesian agent for good compared to bad news; but no interaction between accuracy and valence ($t(104.83) = 0.75$, $p = 0.46$). Follow-up paired t-tests confirmed that participants' updates were closer to the Bayesian benchmark following good compared to bad news for both true ($t(107) = -2.89$, $p =$

0.004) and false ($t(107) = -3.89, p = 0.0002$) information. A Bayes Factor analysis on the interaction revealed extreme evidence in favour of the null hypothesis in the main analysis ($BF_{10}=0.021$). Repeating this analysis controlling for potential confounds also revealed a main effect of accuracy ($t(108.10) = -16.44, p < 0.001$) and a main effect of valence ($t(111.50) = -4.32, p < 0.001$), but no significant interaction between accuracy and valence ($t(105.50) = 1.89, p = 0.061$).

An important caveat to this analysis however is that unlike true trials (where calculating the Bayesian benchmark is straightforward if one assumes Bayesian agents fully trust and integrate base rates labelled as true), for false trials it is not obvious how a Bayesian agent ought to treat the information provided. In the analysis here, we assumed that a normative Bayesian Agent fully integrates the base rate for false trials. We adopted this approach because the false label provides information about *reliability* but not *direction*: participants are told only that the presented base rate is false, without any information about the direction or magnitude of its deviation from the true value. But for completeness, we also tried adjusting the base rate used to calculate the Bayesian agent update for false trials by a constant (see Methods). Across the range of values used (0.5–1.5), participants deviated from Bayesian updating more for false compared to true information labels ($C=1.5$: $t(107.27) = -18.22, p < 0.001$; $C=1.2$: $t(107.01) = -17.61, p < 0.001$; $C=0.8$: $t(109.3) = -16.14, p < 0.001$; $C=0.5$: $t(107.10) = -14.04, p < 0.001$). However, the magnitude and direction of the valence asymmetry relative to the Bayesian benchmark depended on the constant selected. Specifically, when $C > 1$ the positive asymmetry remained ($C=1.5$: $t(107.87) = -17.30, p < 0.001$; $C=1.2$: $t(107.03) = -10.27, p < 0.001$). But when $C < 1$ the asymmetry was in the opposite direction ($C=0.8$: $t(108.29) = 2.82, p = 0.005$; $C=0.5$: $t(108.29) = 15.32, p < 0.001$), i.e. greater deviation from the benchmark for good compared to bad news (see Supplementary Fig. 5).²

Computational models of belief updating Formal models suggest that beliefs are updated by a prediction error signal (that quantifies the difference between prior expectations and outcomes) and a learning rate which governs the rate at which errors drive belief change (Sutton & Barto, 2018). Next, we turned to computational modelling of beliefs in this task to understand the relationship between the belief updating pattern observed in the data under the two approaches (Fig. 2) and the use of an analogous error signal, *estimation errors* (Sharot et al., 2011), which express the difference between prior beliefs (E_1) and the information provided (BR). To do this, we adapted and extended

the computational modelling approach used by Kuzmanovic and Rigoux (2017) which model 2nd estimates as a function of estimation errors.

We began by fitting trial by trial 2nd estimates to four different computational models, which were identical except for the number of learning rates (learning rates dictate how sensitive belief changes are to the estimation errors generated on each trial). Briefly (see Methods for full details), Model 1 (M1) had two learning rates: one for true and one for false information; Model 2 (M2) had three learning rates: one for true information and two for false information – one for false good news and another for false bad news; Model 3 (M3) also had 3 learning rates: one for false information and two for true information – one for true good news and another for true bad news. Model 4 (M4) had 4 learning rates (one for true good news, true bad news, false good news and false bad news). Each model also had two additional parameters: w and σ . w enabled the model to dampen the effect that estimation errors had in changing beliefs according to participants' personal knowledge about each event (e.g., personal risk factors for developing an illness, such as family history). σ was used to estimate the Gaussian that the estimates were drawn from in the model fitting process (see Methods).

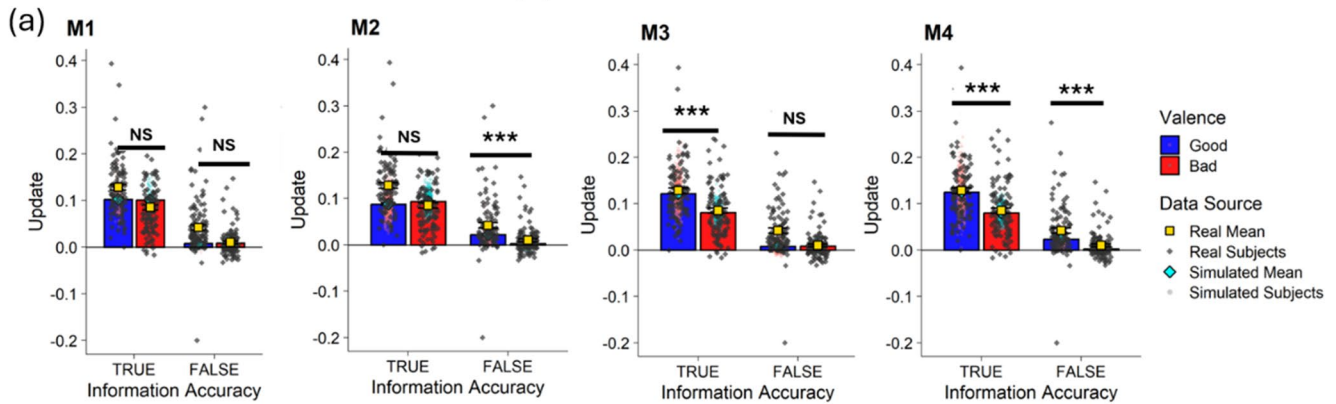
We fit the data to the four models twice. Once where we classified good/bad news by comparing participants first estimates and the base rate provided (as advocated by Sharot & Garrett) with estimation error as the error term, and again where we classified good/bad news by comparing participants first estimate of the base rate and the base rate provided (as advocated by Shah et al., 2016) with base rate error as the error term.

Four learning rate model recovers observed pattern of belief updating Next, using the best fit parameters from the model fitting process (see Methods), we simulated the pattern of updating that each model produced (Fig. 3). M4 was best able to reproduce a pattern of belief updating that most closely resembled the pattern we observed among participants both when we examined the raw difference between participants first (pre information) and second (post information) estimates (i.e. following approach one) and when we examined how close updates were to an ideal normative Bayesian Agent (note as per the behavioural analysis we report above, we assume that a normative Bayesian Agent fully integrates the base rate provided for false trials).

M1 generated updating that was significantly greater for true than false information under the first approach (main effect of Accuracy: $t(164.2) = 52.11, p < 0.001$) and under the second approach (main effect of Accuracy: $t(166.5) = -29.78, p < 0.001$), but there was no optimistic update bias from true information (Approach One: $t(118) = 0.57$,

² We thank an anonymous reviewer for suggesting this approach.

Approach One



Approach Two

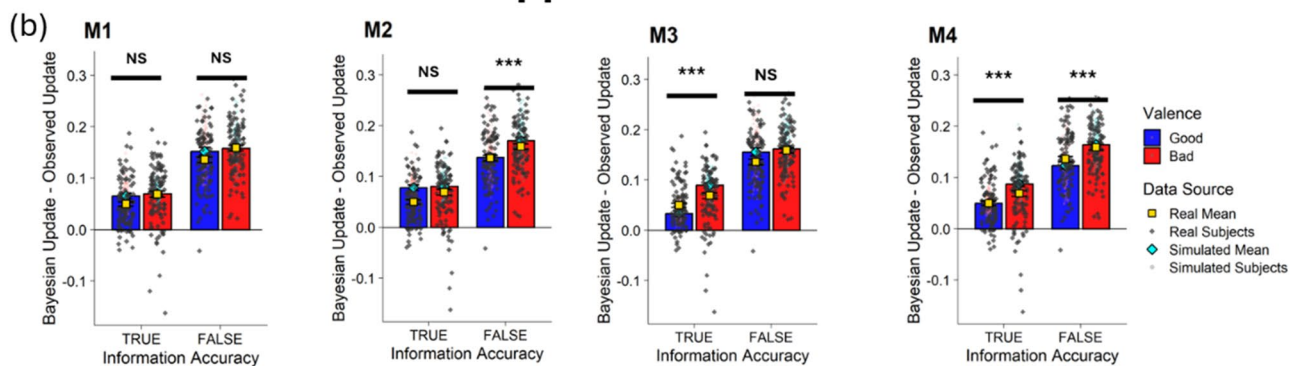


Fig. 3 Simulations from each model producing (a) raw differences in first and second estimates, classifying information as good/bad according to whether first estimates are above/below the base rates provided (approach one following Sharot & Garrett), (b) differences between a Bayesian agent update and participant's update, classifying information as good/bad according to whether estimates of the base rate are above/below the actual base rate provided (approach two following Shah et al., 2016 and Kuzmanovic & Rigoux, 2017). Real data from individual participants overlaid as grey diamonds, with the true empirical means plotted as solid gold squares. Simulated data plotted as solid bars and cyan diamonds for the model means, with individual simulated agents mapped as transparent coloured dots. M1 (which has 2 learning rates: one for true and one for false information) recovers the main effect of accuracy but cannot generate an effect of valence.

$p=0.572$; Approach Two: $t(117) = -0.57$, $p=0.570$) or from false information (Approach One: $t(117) = -0.53$, $p=0.601$; Approach Two: $t(117) = -0.41$, $p=0.682$). M2 also showed greater updating for true than false information and exhibited the bias for false information (Approach One: $t(117)=11.33$, $p<0.001$; Approach Two: $t(117) = -7.53$, $p<0.001$), but not for true information (Approach One: $t(118)=0.69$, $p=0.492$, Approach Two: $t(116) = -0.63$, $p=0.532$). Conversely, M3 showed the bias for true information (Approach One: $t(118)=19.47$, $p<0.001$; Approach Two: $t(116) = -14.84$, $p<0.001$), but not for false information (Approach One: $t(117) = -0.11$, $p=0.910$; Approach Two: $t(118) = -1.14$, $p=0.259$). Finally, by allowing learning

M2 (which has 3 learning rates: a single learning rate for true information and two for false information – one for false good news and another for false bad news) recovers a main effect of accuracy but a valence effect for false information only. M3 (which also has 3 learning rates: a single learning rate for false information and two for true information – one for true good news and another for true bad news) recovers a main effect of accuracy and a valence effect for true information only. M4 – the winning model – which has 4 learning rates captures the updating patterns we observe in the real data under both approaches – a main effect of information accuracy and a valence effect both in response to true information and false information. Error bars represent 95% confidence intervals. *** $p < 0.001$, n.s.: not significant, two-tailed paired t-test

to vary for each combination of accuracy and valence, M4, the winning model, produced both greater updating for true than false information and a significant optimistic update bias for both true (Approach One: $t(118)=18.10$, $p<0.001$; Approach Two: $t(117) = -11.10$, $p<0.001$) and false information (Approach One: $t(116)=10.14$, $p<0.001$; Approach Two: $t(118) = -7.37$, $p<0.001$), a pattern that closely resembled the actual updating we observed in the experimental data (Fig. 3).

Four learning rate model provides the best quantitative fit to the data Next, we tested which model provided the best quantitative fit to the data by comparing how well each of the

four models fit participants' responses against one another. First, we compared leave-one-out cross-validation (LOOcv) scores using a Variational Bayesian Approach (VBA) (Daunizeau et al., 2014; Rigoux et al., 2014). Under both approach one and two, M4 had the highest model frequency ($\sim 75\%$ in approach one and $\sim 93\%$ in approach two), well above the chance level (25%), and a protected exceedance probability of 0.99 under both approaches. Model recovery analysis confirmed that models were discriminable using this approach (see [Supplementary Materials](#)). Next, we computed trial-level expected log predictive densities (ELPD), summed these for each participant, and used the summed values as model evidence (Vehtari et al., 2017). Model 4 outperformed Model 3 in approach one with an ELPD improvement of 33.6 (se_diff=14.5), Model 2 ($\Delta\text{ELPD}=242.7$, se_diff=40.0) and Model 1 ($\Delta\text{ELPD}=256.2$, se_diff=41.1) and similarly in approach two (M3: $\Delta\text{ELPD}=17.35$, se_diff=10.05; M2: $\Delta\text{ELPD}=164.5$, se_diff=35.6; M1: $\Delta\text{ELPD}=188.3$, se_diff=36.4). We then compared the ELPD scores again using VBA. Consistent with the VBA

on the LOOcv scores, Model 4 again had the highest model frequency (Approach one: 78%; Approach two: 66%) and a protected exceedance probability of 1 under approach one and 81% under approach two (Fig. 4). Finally, we compared models directly against one another using pairwise comparisons (instead of all 4 together) again using VBA. Using LOOcv scores, M4 had the highest model frequency when compared against M1 (Approach 1: 86.6%; Approach 2: 92.1%), M2 (Approach 1: 90.1%; Approach 2: 87.7%) and M3 (Approach 1: 72.2%; Approach 2: 98%). Similarly, using ELPD scores, M4 had the highest model frequency when compared against M1 (Approach 1: 81.8%; Approach 2: 87.7%), M2 (Approach 1: 86.3%; Approach 2: 81.3%) and M3 (Approach 1: 79.9%; Approach 2: 72.2%).

Higher learning rate for good news vs. bad news for both true and false information Next, we examined the pattern of learning rates from the winning model (M4) and tested for differences between them. This showed an optimistic update bias present both from true and false

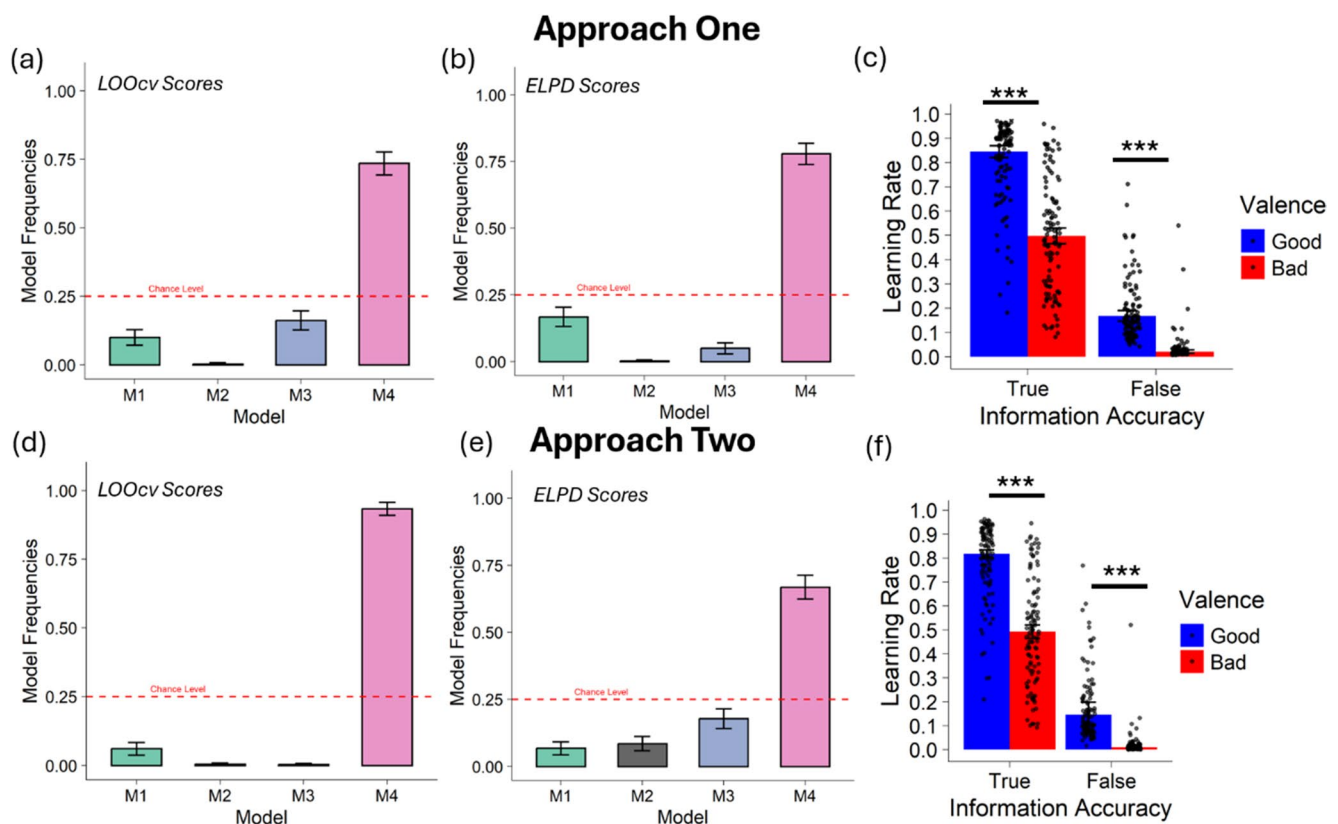


Fig. 4 (a–b) The estimated model frequencies using two metrics of model evidence – LOOcv & ELPD – and the estimates of the winning model under approach one. M4 had the highest frequency no matter the model evidence metric and this was substantially higher than chance level. (c) The four learning rates from the winning model (M4) showed a positive learning bias both for true ($t(107)=8.56$, $p<0.001$, hierarchical t-test comparing $\alpha_{\text{true_goodnews}}$ with $\alpha_{\text{true_badnews}}$) and false ($t(107)=6.74$, $p<0.001$, hierarchical t-test comparing $\alpha_{\text{false_goodnews}}$

with $\alpha_{\text{false_badnews}}$) information. (d–e) The estimated model frequencies and estimates of the winning model under the second approach. Again, M4 emerged as the winner – albeit with weaker ELPD scores – and (f) its estimates showed similar patterns with a positive learning bias both for true ($t(107)=6.77$, $p<0.001$, hierarchical t-test comparing $\alpha_{\text{true_goodnews}}$ with $\alpha_{\text{true_badnews}}$) and false ($t(107)=4.00$, $p<0.001$, hierarchical t-test comparing $\alpha_{\text{false_goodnews}}$ with $\alpha_{\text{false_badnews}}$) information. *** $p<0.001$, hierarchical t-test

information (see Fig. 4c and f & Tables 1 and 2). Specifically, participants learnt more from good than bad news for true (Approach One: $t(107)=8.56$, $p<0.001$; Approach Two: $t(107)=6.77$, $p<0.001$, hierarchical t-test comparing $\alpha_{\text{true_goodnews}}$ with $\alpha_{\text{true_badnews}}$) and false (Approach One: $t(107)=6.74$, $p<0.001$; Approach Two: $t(107)=4.00$, $p<0.001$, hierarchical t-test comparing $\alpha_{\text{false_goodnews}}$ with $\alpha_{\text{false_badnews}}$) information (note we do not enter the four learning rates into 2*2 ANOVA to test for main effects and interaction due to the non-independence of these parameters; it is possible to conduct hierarchical t-tests to compare learning from good and bad news separately for true and false information by reparametrizing the model, see Methods and [Supplementary Materials](#)). There was no interaction in the magnitude of the learning rates between valence (good versus bad) and information accuracy (true vs. false) in approach one ($t(107)=0.59$, $p=0.55$) and two ($t(107)=1.21$, $p=0.22$).

Updating from false information and questionnaire scores Finally, as previous studies have reported that optimistic belief updating varies with psychiatric symptom severity (Garrett et al., 2014, 2018; Khalid et al., 2024; Korn et al., 2014; Ossola et al., 2020), we ran an exploratory analysis to examine whether heterogeneity in updating from false information from good or bad news was associated with OCD, depression or schizotypal traits

(each measured via questionnaires), controlling for age, gender and updating from true information of the same valence. None of the demographic or psychiatric measures significantly predicted updating from false good or false bad news. In both models, the strongest predictor was the corresponding updating following true information (false good: $\beta=0.294$, $p=0.004$; false bad: $\beta=0.204$, $p<0.001$). See [Supplementary Materials](#) for full details of the regression models and regression results.

Discussion

Understanding the circumstances under which individuals are prone to integrating false information is important for understanding how erroneous beliefs such as conspiracy theories can persist in the face of flawed evidence (Douglas et al., 2017) and developing effective strategies to counter the prominent rise of misinformation (Hanley & Durumeric, 2024; Lewandowsky et al., 2012; Nyhan & Reifler, 2010). Here, by adapting a classic belief updating task which presents individuals with information that can vary on two dimensions – valence (whether the information is better or worse than expected) and accuracy (whether the information is true or false) – we show that both of these factors exert important roles in governing the degree to which beliefs are influenced by the information and change as a consequence.

Table 1 Mean parameter estimates under *Approach One* for each model (95% confidence intervals are shown in square brackets) and the comparison of M4 with the rest of the models using differences in expected log predictive density (ΔELPD) with their corresponding uncertainty (se_diff). The mean parameter estimates were used for simulating data under each model shown in Fig. 3

Model	α_{true}	α_{false}	$\alpha_{\text{true_goodnews}}$	$\alpha_{\text{true_badnews}}$	$\alpha_{\text{false_goodnews}}$	$\alpha_{\text{false_badnews}}$	w	σ	ΔELPD (se_diff)
M1	0.63 [0.58–0.68]	0.042 [0.02–0.06]	-	-	-	-	0.67 [0.57–0.76]	8.02 [7.43–8.67]	256.24 (41.08)
M2	0.62 [0.57–0.68]	-	-	-	0.16 [0.12–0.21]	0.02 [0.009–0.04]	0.80 [0.72–0.87]	7.65 [7.14–8.20]	242.66 (40.14)
M3	-	0.051 [0.03–0.072]	0.84 [0.79–0.88]	0.50 [0.43–0.56]	-	-	0.72 [0.63–0.80]	7.24 [6.73–7.79]	33.65 (14.46)
M4	-	-	0.84 [0.79–0.88]	0.49 [0.43–0.56]	0.16 [0.12–0.21]	0.02 [0.01–0.039]	0.70 [0.61–0.79]	7.13 [6.63–7.65]	0

Table 2 Mean parameter estimates under *Approach Two* for each model (95% confidence intervals are shown in square brackets) and the comparison of M4 with the rest of the models using differences in expected log predictive density (ΔELPD) with their corresponding uncertainty (se_diff). The mean parameter estimates were used for simulating data under each model shown in Fig. 3

Model	α_{true}	α_{false}	$\alpha_{\text{true_goodnews}}$	$\alpha_{\text{true_badnews}}$	$\alpha_{\text{false_goodnews}}$	$\alpha_{\text{false_badnews}}$	w	σ	ΔELPD (se_diff)
M1	0.63 [0.58–0.68]	0.04 [0.03–0.06]	-	-	-	-	0.61 [0.51–0.71]	8.16 [7.55–8.81]	188.32 (36.39)
M2	0.63 [0.58–0.68]	-	-	-	0.14 [0.11–0.18]	0.001 [0.0000–0.46]	0.60 [0.50–0.69]	8.11 [7.52–8.76]	164.47 (35.57)
M3	-	0.04 [0.03–0.06]	0.77 [0.72–0.82]	0.53 [0.46–0.59]	-	-	0.55 [0.44–0.66]	7.82 [7.21–8.48]	17.35 (10.05)
M4	-	-	0.77 [0.71–0.82]	0.52 [0.45–0.58]	0.13 [0.10–0.18]	0.01 [0.0005–0.06]	0.48 [0.28–0.68]	7.79 [7.19–8.44]	0

First, we show that information is used to update beliefs to a greater degree when that information is revealed to be true compared to false. This suggests that warnings about the reliability of a piece of information may contribute towards determining whether that information serves to have an impact on altering beliefs or not over the long term. Second, we show that information is integrated to a greater degree when it presents a shift in beliefs towards a “good” (i.e. better than first thought) compared to a “bad” (i.e. worse than first thought) direction. This is consistent with past findings (Garrett et al., 2014; Garrett & Daw, 2020; Kappes et al., 2018; Korn et al., 2014; Kube & Rozenkrantz, 2021; Kuzmanovic et al., 2018, 2019; Kuzmanovic & Rigoux, 2017; Ma et al., 2016; Oganian et al., 2019; Sharot et al., 2011; Sharot & Garrett, 2016), suggesting that learning is often biased in a positive direction. What is new in the findings we present here is that this bias is equally strong in response to false information (in fact, we find evidence against any modulation of the bias by accuracy). This suggests that whilst debunking information after it has been encountered can act to mitigate its influence, such measures are likely to be less successful in cases where false information presents news that is better than expected.

Computational modelling of participants’ responses suggested that both of these patterns of belief updating arose out of the differential use of error signals, generated when new information is encountered and deviates from prior expectations. In general (Garrett & Daw, 2020; Sharot et al., 2011; Sutton & Barto, 2018), the more surprising a piece of information is, the more beliefs can be expected to shift up or down. Here we show that this relationship between degree of surprise (i.e. the size of the error signal) and subsequent change in beliefs (parametrised in computational models as learning rates) is stronger when information is revealed to be true compared to when it is revealed to be false and when information is better compared to worse than expected. We show this by comparing models which parse information according to its accuracy only, according to valence for only true or only false information and according to *both* accuracy and valence with this latter model providing the best fit to the data both quantitatively - using Bayesian model comparison - and qualitatively - by comparing how well belief updating generated by each model (using simulations) was able to capture the updating pattern we observe in the human data (Palminteri et al., 2017).

An important question is whether differences in updating between better than expected (good news) and worse than expected (bad news) information reflect a genuine psychological bias or can instead arise from rational Bayesian updating (Harris et al., 2013; Shah et al., 2016; Burton et al., 2022). To address this issue, we analysed the data using both the conventional analysis employed in previous belief

updating studies (Sharot & Garrett, 2022) and a Bayesian comparison approach following Shah et al. (2016). The results from both approaches revealed that participants updated to a greater degree (approach one) and closer to a normative Bayesian benchmark (approach two) following information labelled true compared to false, suggesting that explicit true/false labels can moderate belief updating. However, calculating a Bayesian benchmark requires making an assumption about how a rational Bayesian agent ought to treat information explicitly labelled as false, for which there is no obvious analytical solution. Our primary analyses assumed that the Bayesian agent combines the presented base rate with their prior beliefs in the same manner as for true information. We adopted this approach because the false label provides information about reliability but not direction: participants are told only that the presented base rate is false, without any information about the direction or magnitude of its deviation from the true value. When we applied different adjustments to the base-rate odds used to calculate the Bayesian benchmark for false information trials, participants deviated from this benchmark more for false compared to true across the range of adjustments we made. However, the magnitude and (under some assumptions) the direction of the valence asymmetry relative to the Bayesian benchmark depended on the (assumed) treatment of false information, suggesting that normative interpretations of optimistic updating are sensitive to assumptions about how rational agents should respond to explicitly false information labels (Harris et al., 2013; Shah et al., 2016).

In the set of computational models we tested, updates were implemented in a single step, which (effectively) collapses over two distinct stages of a trial in the task: (1) when information is first received and an estimation error generated; (2) when cues about the reliability of the information are received. One possibility is that the process of belief change occurs in a single step. Under this scenario, information presented at (1) would need to be maintained in working memory before the true/false cue is shown at (2), which could act much akin to a go/no-go signal (Logan et al., 1984) in determining whether to then integrate the information (i.e. implement an update), possibly via recruiting brain circuits previously suggested to be involved in belief updating in this task to differing degrees dependent both on whether the information was true or false and whether the information was better or worse than expected. These brain regions include the left inferior frontal gyrus and medial frontal cortex (Garrett et al., 2014; Sharot et al., 2011) for good news and the right inferior frontal gyrus (Sharot et al., 2011) and right inferior parietal lobule (Garrett et al., 2014) for bad news. But an alternative possibility is that the process we observe is a two-stage one in which participants first make an initial update to beliefs at (1) before the accuracy

of the information is known and then make a revision to this updated belief at (2), depending on the identity of the true/false cue. At this second stage, learning from the first stage could either be undone (in the presence of a false cue) and/or further boosted (in the presence of a true cue).

This second possibility is consistent with the influential ‘Spinozan’ account of belief formation (Gilbert et al., 1990) whereby participants provisionally accept information as true and move to dismiss it if they subsequently learn the information was false. Under this framework, false information cues would trigger a corrective process that attenuates (or reverses) the impact of the initial update. However, more recent work (Vorms et al., 2022) has challenged the robustness of this account by manipulating how plausible the initial information is to begin with, such that highly implausible information may instead be provisionally treated as false until evidence to the contrary is provided (Vorms et al., 2022). The present data cannot distinguish between these competing temporal accounts because both predict the same final belief estimates. But future work could look to build a more complete temporal picture of the updating process as it unfolds by incorporating other neuroscience methods such as EEG and functional Magnetic Resonance Imaging (Glover, 2011) alongside experimental paradigms which manipulate the plausibility of the presented base rates, interrupting when the information accuracy cue is presented and/or obtaining intermediate estimates of beliefs (i.e. in-between the base rate presentation and information accuracy cue).

We adapted a widely used belief updating task (Garrett et al., 2014; Garrett & Daw, 2020; Garrett & Sharot, 2023; Kappes et al., 2018; Korn et al., 2014; Kube & Rozenkrantz, 2021; Kuzmanovic et al., 2018, 2019; Kuzmanovic & Rigoux, 2017; Ma et al., 2016; Oganian et al., 2019; Sharot et al., 2011; Sharot & Garrett, 2016) shown to test for the presence of biases in belief updating by adding explicit cues on each trial which signalled to participants whether the information they had just observed had been true or false. This allowed us to test whether the bias exists and is similar in strength when true and false information is received. A natural question that arises from these findings is how biases in belief updating might manifest in the absence of any such cues or in the presence of a third “unknown accuracy” cue? The former is how original versions of the belief updating task have been run in the past. Interestingly, the magnitude of the optimistic belief updating observed when using this original version (without any information accuracy cues, Sharot et al., 2011) is of a similar strength to the optimistic belief updating we observe in the true condition here. This might lead one to hypothesise that participants could treat information explicitly signalled to be true similarly to cases where information is presented without any cues provided

about accuracy. However, caution is warranted. Even though cues about information accuracy were not provided in the Sharot et al. study, the default position of participants undertaking original instantiations of the belief updating task may be to assume that the statistics presented were true (rather than unknown). Indeed, the statistics presented in the original design were factually correct and task instructions primed participants to believe they were accurate. It remains to be tested therefore how belief updating from true and false information differs from cases where information accuracy is withheld.

Nonetheless, the results here show that sensitivity to new information is able to adapt in response to signals about how accurate the information is. This extends past findings showing that sensitivity to new information is often not fixed but can flexibly adjust to a range of factors including: volatility of the environment (Behrens et al., 2007), surprise (Pearce & Hall, 1980), age (Moutsiana et al., 2013), psychiatric symptoms (Garrett et al., 2014; Ossola et al., 2020), reward uncertainty (Nassar et al., 2012), working memory (Collins & Frank, 2012), social context (Diaconescu et al., 2014), arousal levels (Eldar et al., 2016; Garrett et al., 2018) and confidence (Desender et al., 2019; Meyniel et al., 2015; Rollwage et al., 2020; Yeung & Summerfield, 2012). Potentially some or all of these factors may also exert roles in governing the degree to which information known to be false can impact beliefs, making certain groups of individuals more susceptible than others to the effects of malicious misinformation attempts which in turn could warrant greater measures being put in place to protect those likely to be more vulnerable.

An important question concerns the extent to which the present findings generalise beyond personally relevant beliefs. The present task focuses on estimates of participants’ own future life events. Several studies have argued that valence-dependent asymmetries are weaker, absent or even reversed for impersonal beliefs (Barron, 2021; Coutts, 2019; Gotthard-Real, 2017). Yet other work suggests that motivational influences on learning are not necessarily restricted to self-relevant beliefs. For example, asymmetric learning has been observed in reinforcement learning paradigms involving abstract rewards (Lefebvre et al., 2017), political belief revision (Tappin et al., 2017), and more recently in learning from false information that confirms prior choices (Vidal-Perez et al., 2025; Razi et al., 2026). Hence motivational significance (including personal relevance, prior commitments and identity-relevant beliefs) rather than self-relevance per se may be a key determinant of where one might expect to see valence-based learning biases emerge. Future work should look to examine the conditions under which one might expect a valence bias to be attenuated or absent when information cues are presented.

A further limitation concerns participants' motivations for revising their beliefs. As participants were not incentivised to maximise the accuracy of their estimates, the present design cannot distinguish whether belief revision reflects a desire to improve accuracy, compliance with perceived task demands, or other interpretations of the experimental goals. Recent work suggests that participants' assumptions about the purpose of belief updating tasks can meaningfully influence observed behaviour (Spektor & Wulff, 2024). However, whilst such factors (and potentially others) may contribute to the overall propensity to revise beliefs, they do not readily explain the differential effects of information accuracy and valence we observe. Future work could explicitly manipulate task framing, participant expectations, and incentives for accurate responding to determine how these factors interact with the processing of true and false information as well as learning asymmetries.

Our findings suggest that warnings and labels that call out cases of potential false information could be a means to help prevent false beliefs being generated and sustained in the face of accurate information to the contrary. However, explicit labels and the like are unlikely to be a panacea. This is not least because - as we show here - individuals are likely to be more susceptible to false information (even when this is made explicit) in cases where it provides a better-than-expected view of the future. Indeed, recent complementary findings (Vidal-Perez et al., 2025) suggest that a related ('positivity') bias also exists whereby individuals learn to a greater degree from false information that confirms past choices and decisions. An important concern is that alongside these biases in how individuals learn when they receive false information, individuals also have increasing agency over where they choose to source their information from in the first place, particularly in digital environments. This can result in selective exposure to certain types of information and skewed informational environments (Flaxman et al., 2016). This skew can be further compounded if individuals are also selective about what information they choose to share with peers in their network, choosing to predominantly share false information that perpetuates a specific view of the world that one finds desirable (Pennycook & Rand, 2022). There is then a potential dual challenge to be met in order to counter disinformation - how biased the information is that we receive and the role that biases play in how we then choose to learn from that information. Other promising avenues exist to meet this challenge such as using Large Language Models to deploy reasoning to move individuals away from deep seated false beliefs (Costello et al., 2024), providing rewards (such as 'likes') in return for sharing accurate information (Vellani et al., 2024) and 'pre-bunking' (Van Der Linden, 2024). Awareness of the powerful role that biases can play and how these can arise from

core learning principles are important factors to consider in evaluating the effectiveness of these and others as the scientific community looks to develop and test different ways to counteract inaccurate information from successfully perpetuating and sustaining false beliefs.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42113-026-00341-4>.

Acknowledgments We thank Nathaniel Daw for the EM code we adapted to fit the models.

Author Contributions H.R., T.S., and N.G. conceived and designed the study and wrote the manuscript. H.R. collected and analysed the data.

Funding None.

Data Availability Data available at: [https://github.com/HamidRazi7/BeliefUpdate_FalseInformation](https://github.com/HamidRazi7/BeliefUpdate_FalseInformation).

Code Availability Code available at: https://github.com/HamidRazi7/BeliefUpdate_FalseInformation.

Declarations

Ethics Approval The study was approved by the University of East Anglia's ethics committee.

Consent to Participate Informed consent was obtained from all individual participants.

Consent for Publication Consent for publication was obtained from all authors.

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Barber, B. M., & Odean, T. (1999). The courage of misguided convictions. *Financial Analysts Journal*, 55(6), 41–55.
- Barron, K. (2021). Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics*, 24(1), 31–58. <https://doi.org/10.1007/s10683-020-9653-z>

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, 4, 561–571. <https://doi.org/10.1001/archpsyc.1961.01710120031004>
- Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9), 1214–1221. <https://doi.org/10.1038/nn1954>
- Benjamin, D. J. (2019). Chapter 2—Errors in probabilistic reasoning and judgment biases. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics: applications and foundations 1* (Vol. 2, pp. 69–186). North-Holland. <https://doi.org/10.1016/bs.hesbe.2018.11.002>
- Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). *Julia: A fast dynamic language for technical computing*.
- Brunnermeier, M. K., & Parker, J. A. (2005). Optimal expectations. *American Economic Review*, 95(4), 1092–1118. <https://doi.org/10.1257/0002828054825493>
- Bürkner, P. C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, 80, 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Burton, J. W., Harris, A. J. L., Shah, P., & Hahn, U. (2022). Optimism where there is none: Asymmetric belief updating observed with valence-neutral life events. *Cognition*, 218, 104939. <https://doi.org/10.1016/j.cognition.2021.104939>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: a probabilistic programming language. *Journal of Statistical Software*, 76, 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Ceylan, G., Kim, D. R., Zhang, B., & Goldring, E. A. (2023). Extended exposure to false information corrupts beliefs. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001464>
- Cheng, Y., Luo, Y., & Gu, R. (2024). Working memory capacity moderates the optimistic update bias. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001674>
- Christner, N., Plötner, M., & Dorrrough, A. R. (2024). Warning labels reduce susceptibility to misinformation. *Psychological Science*. <https://doi.org/10.1177/09567976231222574>
- Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7), 1024–1035. <https://doi.org/10.1111/j.1460-9568.2011.07980.x>
- Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science (New York NY)*, 385(6714), eadq1814. <https://doi.org/10.1126/science.adq1814>
- Coutts, A. (2019). Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics*, 22(2), 369–395. <https://doi.org/10.1007/s10683-018-9572-5>
- Daunizeau, J., Friston, K. J., & Kiebel, S. J. (2009). Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D Nonlinear Phenomena*, 238(21), 2089–2118. <https://doi.org/10.1016/j.physd.2009.08.002>
- Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLOS Computational Biology*, 10(1), e1003441. <https://doi.org/10.1371/journal.pcbi.1003441>
- Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed-accuracy tradeoff for subsequent decisions. *eLife*, 8, e43499. <https://doi.org/10.7554/eLife.43499>
- Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., & Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical bayesian learning. *PLOS Computational Biology*, 10(9), e1003810. <https://doi.org/10.1371/journal.pcbi.1003810>
- Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current Directions in Psychological Science*, 26(6), 538–542. <https://doi.org/10.1177/0963721417718261>
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3), 69–106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>
- Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, 38(8), 1087–1100. <https://doi.org/10.3758/MC.38.8.1087>
- Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2), 114–138. <https://doi.org/10.1257/mic.3.2.114>
- Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, 20(1), 15–24. <https://doi.org/10.1016/j.tics.2015.07.010>
- Filipowicz, A. L., Glaze, C. M., Kable, J. W., & Gold, J. I. (2020). Pupil diameter encodes the idiosyncratic, cognitive complexity of belief updating. *eLife*, 9, e57872. <https://doi.org/10.7554/eLife.57872>
- Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption (SSRN Scholarly Paper 2363701). *Social Science Research Network*. <https://doi.org/10.2139/ssrn.2363701>
- Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment*, 14(4), 485–496. <https://doi.org/10.1037/1040-3590.14.4.485>
- Fradkin, I., Ludwig, C., Eldar, E., & Huppert, J. D. (2020). Doubting what you already know: Uncertainty regarding state transitions is associated with obsessive compulsive symptoms. *PLoS Computational Biology*, 16(2), e1007634. <https://doi.org/10.1371/journal.pcbi.1007634>
- Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods*, 53(1), 139–152. <https://doi.org/10.3758/s13428-020-01408-1>
- Fu, Q., Moerbeek, M., & Hoijtink, H. (2022). Sample size determination for Bayesian ANOVAs with informative hypotheses. *Frontiers in Psychology*, 13, 947768. <https://doi.org/10.3389/fpsyg.2022.947768>
- Garrett, N., & Daw, N. D. (2020). Biased belief updating and suboptimal choice in foraging decisions. *Nature Communications*, 11(1). Article 3417. <https://doi.org/10.1038/s41467-020-16964-5>
- Garrett, N., & Sharot, T. (2014). How robust is the optimistic update bias for estimating self-risk and population base rates? *PLoS One*, 9(6), e98848. <https://doi.org/10.1371/journal.pone.0098848>
- Garrett, N., & Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah. *Consciousness and Cognition*, 50, 12–22. <https://doi.org/10.1016/j.concog.2016.10.013>
- Garrett, N., & Sharot, T. (2023). There is no belief update bias for neutral events: Failure to replicate Burton et al. (2022). *Journal of Cognitive Psychology (Hove England)*, 35(8), 876–886. <https://doi.org/10.1080/20445911.2023.2245112>
- Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Frontiers in Human*

- Neuroscience*, 8 Article 639. <https://doi.org/10.3389/fnhum.2014.00639>
- Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating beliefs under perceived threat. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 38(36), 7901–7911. <https://doi.org/10.1523/JNEUROSCI.0716-18.2018>
- Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59(4), 601–613. <https://doi.org/10.1037/0022-3514.59.4.601>
- Globig, L. K., Blain, B., & Sharot, T. (2022). Private optimism, public despair: Differential effects of the COVID-19 pandemic on well-being. *PLOS ONE*, 17(3), e0263726. <https://doi.org/10.1371/journal.pone.0263726>
- Glover, G. H. (2011). Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America*, 22(2), 133–139. <https://doi.org/10.1016/j.nec.2010.11.001>
- Gotthard-Real, A. (2017). Desirability and information processing: An experimental study. *Economics Letters*, 152, 96–99.
- Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist*, 35(7), 603–618. <https://doi.org/10.1037/0003-066X.35.7.603>
- Hanley, H. W. A., & Durumeric, Z. (2024). Machine-made media: monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites (arXiv:2305.09820). *arXiv*. <https://doi.org/10.48550/arXiv.2305.09820>
- Harris, A., Shah, P., Catmur, C., Bird, G., & Hahn, U. (2013). Autism, optimism and positive events: Evidence against a general optimistic bias. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35(35), 555–560. <https://escholarship.org/uc/item/76c4g9g9>
- Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Computational Biology*, 7(4), e1002028. <https://doi.org/10.1371/journal.pcbi.1002028>
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
- Kao, Y. C., Gruber, M. J., & Lin, C. Y. (2023). Affective state modulates the optimism bias in belief updating. *Cognition*, 233, 105360. <https://doi.org/10.1016/j.cognition.2022.105360>
- Kappes, A., Faber, N. S., Kahane, G., Savulescu, J., & Crockett, M. J. (2018). Concern for others leads to vicarious optimism. *Psychological Science*, 29(3), 379–389. <https://doi.org/10.1177/0956797617737129>
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3), 983–997.
- Khalid, I., Morlaas, O., Bottemanne, H., Thonon, L., Costa, T. D., Fos-sati, P., & Schmidt, L. (2024). Effects of experiencing the COVID-19 pandemic on optimistically biased belief updating. *eLife*, 13. Article RP101157. <https://doi.org/10.7554/eLife.101157>
- Knowles, K. A., & Olatunji, B. O. (2023). Intolerance of uncertainty as a cognitive vulnerability for obsessive-compulsive disorder: a qualitative review. *Clinical Psychology*, 30(3), 317–330. <https://doi.org/10.1037/cps0000150>
- Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, 44(3), 579–592. <https://doi.org/10.1017/S0033291713001074>
- Kube, T., & Rozenkrantz, L. (2021). When beliefs face reality: an integrative review of belief updating in mental health and illness. *Perspectives on Psychological Science*, 16(2), 247–274. <https://doi.org/10.1177/1745691620931496>
- Kube, T., Rozenkrantz, L., Rauber, R., Glombiewski, J. A., & Rief, W. (2022). When open-minded thinking and optimism are not enough: Self-referential processing in belief updating. *Journal of Psychopathology and Clinical Science*, 131(4), 434–447. <https://doi.org/10.1037/abn0000751>
- Kuhnen, C. M. (2015). Asymmetric learning from financial information. *Journal of Finance*, 70(5), 2029–2062. <https://doi.org/10.1111/jofi.12223>
- Kuzmanovic, B., & Rigoux, L. (2017). Valence-dependent belief updating: computational validation. *Frontiers in Psychology*, 8, 1087. <https://doi.org/10.3389/fpsyg.2017.01087>
- Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2015). Self-specific optimism bias in belief updating is associated with high trait optimism. *Journal of Behavioral Decision Making*, 28(3), 281–293. <https://doi.org/10.1002/bdm.1840>
- Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural correlates of optimism in memory updating. *Frontiers in Human Neuroscience*, 10, 223. <https://doi.org/10.3389/fnhum.2016.00223>
- Kuzmanovic, B., Rigoux, L., & Tittgemeyer, M. (2018). Influence of vmPFC on dmPFC predicts valence-guided belief formation. *Journal of Neuroscience*, 38(37), 7996–8010. <https://doi.org/10.1523/JNEUROSCI.0266-18.2018>
- Kuzmanovic, B., Rigoux, L., & Vogeley, K. (2019). Brief report: reduced optimism bias in self-referential belief updating in high-functioning autism. *Journal of Autism and Developmental Disorders*, 49(7), 2990–2998. <https://doi.org/10.1007/s10803-016-2940-0>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82, 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1, 0067. <https://doi.org/10.1038/s41562-017-0067>
- Lenth, R. V., Piaskowski, J., Banfai, B., Bolker, B., Buerkner, P., Giné-Vázquez, I., Hervé, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2026). *emmeans: estimated marginal means, aka least-squares means* (Version 2.0.3) [Computer software]. <https://cran.r-project.org/web/packages/emmeans/index.html>
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131. <https://doi.org/10.1177/152910612451018>
- Loewenstein, G. (2006). Social science. The pleasures and pains of information. *Science*, 312(5774), 704–706. <https://doi.org/10.1126/science.1128388>
- Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice reaction time responses: A model and a method. *Journal of Experimental Psychology Human Perception and Performance*, 10(2), 276–291. <https://doi.org/10.1037/0096-1523.10.2.276>
- Ma, Y., Li, S., Wang, C., Liu, Y., Li, W., Yan, X., Chen, Q., & Han, S. (2016). Distinct oxytocin effects on belief updating in response to desirable and undesirable feedback. *Proceedings of the National Academy of Sciences*, 113(33), 9256–9261. <https://doi.org/10.1073/pnas.1604285113>
- Makowski, D., Ben-Shachar, M. S., & Lüdtke, D. (2019). bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 4(40), 1541. <https://doi.org/10.21105/joss.01541>
- Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, 78(2), 293–296. <https://doi.org/10.1016/j.schres.2005.06.020>

- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315. <https://doi.org/10.1016/j.jml.2017.01.001>
- Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: from neural origins to behavior. *Neuron*, 88(1), 78–92. <https://doi.org/10.1016/j.neuron.2015.09.039>
- Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). Managing self-confidence: theory and experimental evidence (Working Paper No. 17014). National Bureau of Economic Research. <https://doi.org/10.3386/w17014>
- Moutsiana, C., Garrett, N., Clarke, R. C., Lotto, R. B., Blakemore, S. J., & Sharot, T. (2013). Human development of the ability to learn from bad news. *Proceedings of the National Academy of Sciences*, 110(41), 16396–16401. <https://doi.org/10.1073/pnas.1305631110>
- Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasley, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, 15(7), 1040–1046. <https://doi.org/10.1038/nn.3130>
- Nassar, M. R., Waltz, J. A., Albrecht, M. A., Gold, J. M., & Frank, M. J. (2021). All or nothing belief updating in patients with schizophrenia reduces precision and flexibility of beliefs. *Brain*, 144(3), 1013–1029. <https://doi.org/10.1093/brain/awaa453>
- Nyhan, B., & Reifler, J. (2010). When corrections fail: the persistence of political misperceptions. *Political Behavior*, 32(2), 303–330. <https://doi.org/10.1007/s11109-010-9112-2>
- Oganian, Y., Heekeren, H. R., & Korn, C. W. (2019). Low foreign language proficiency reduces optimism about the personal future. *Quarterly Journal of Experimental Psychology*, 72(1), 60–75. <https://doi.org/10.1177/1747021818774789>
- Ossola, P., Garrett, N., Sharot, T., & Marchesi, C. (2020). Belief updating in bipolar disorder predicts time of recurrence. *eLife*, 9, e58891. <https://doi.org/10.7554/eLife.58891>
- Paling, J. (2003). Strategies to help patients understand risks. *Bmj*, 327(7417), 745–748. <https://doi.org/10.1136/bmj.327.7417.745>
- Palminteri, S., Wyart, V., & Koehlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6), 425–433. <https://doi.org/10.1016/j.tics.2017.03.011>
- Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87(6), 532–552.
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1). Article 2333. <https://doi.org/10.1038/s41467-022-30073-5>
- Peysakhovich, A., & Karmarkar, U. R. (2016). Asymmetric effects of favorable and unfavorable information on decision making under ambiguity. *Management Science*, 63(4), 1178–1192. <https://doi.org/10.1287/mnsc.2015.2233>
- Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Computational Biology*, 15(6), e1007043. <https://doi.org/10.1371/journal.pcbi.1007043>
- Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, 118(37), e2104235118. <https://doi.org/10.1073/pnas.2104235118>
- Pulcu, E., & Browning, M. (2019). Affective bias as a rational response to the statistics of rewards and punishments. *eLife*, 8, e46138. <https://doi.org/10.7554/eLife.46138>
- Razi, H., Sambrook, T., & Garrett, N. (2026). Confirmation bias exists in the face of false information. *bioRxiv*. <https://doi.org/10.6489/82026.05.07.723487>
- Reed, E. J., Uddenberg, S., Suthaharan, P., Mathys, C. D., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2020). Paranoia as a deficit in non-social belief updating. *eLife*, 9, e56345. <https://doi.org/10.7554/eLife.56345>
- Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—Revisited. *Neuroimage*, 84, 971–985. <https://doi.org/10.1016/j.neuroimage.2013.08.065>
- Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, 11(1), 2634. <https://doi.org/10.1038/s41467-020-16278-6>
- Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, 32(5), 880–892. <https://doi.org/10.1037/0022-3514.32.5.880>
- Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes*, 66(1), 102–110. <https://doi.org/10.1006/obhd.1996.0041>
- Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, 22(2), 391–407. <https://doi.org/10.3758/s13423-014-0684-4>
- Shah, P., Harris, A. J. L., Bird, G., Catmur, C., & Hahn, U. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology*, 90, 71–127. <https://doi.org/10.1016/j.cogpsych.2016.05.004>
- Sharot, T., & Garrett, N. (2016). Forming beliefs: why valence matters. *Trends in Cognitive Sciences*, 20(1), 25–33. <https://doi.org/10.1016/j.tics.2015.11.002>
- Sharot, T., & Garrett, N. (2022). A guideline and cautionary Note: How to use the belief update task correctly. *Methods in Psychology*, 6, 100091. <https://doi.org/10.1016/j.metip.2022.100091>
- Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11), 1475–1479. <https://doi.org/10.1038/nn.2949>
- Shefrin, H. (2009). How psychological pitfalls generated the global financial crisis. In L. B. Siegel (Ed.), *Voices of wisdom: understanding the positive and negative lessons from financial history* (pp. 224–256). Research Foundation of CFA Institute.
- Spekter, M. S., & Wulff, D. U. (2024). Predecisional information search adaptively reduces three types of uncertainty. *Proceedings of the National Academy of Sciences*, 121(47), e2311714121. <https://doi.org/10.1073/pnas.2311714121>
- Staats, B. R., Kc, D. S., & Gino, F. (2018). Maintaining beliefs in the face of negative news: the moderating role of experience. *Management Science*, 64(2), 804–824. <https://doi.org/10.1287/mnsc.2016.2640>
- Sunstein, C., Bobadilla-Suarez, S., Lazzaro, S., & Sharot, T. (2017). How people update beliefs about climate change: good news and bad news. *Cornell Law Review*, 102(6), 1431–1443.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*, 2nd ed. The MIT Press.
- Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General*, 146(8), 1143–1149. <https://doi.org/10.1037/xge0000298>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Van Der Linden, S. (2024). Countering misinformation through psychological inoculation. In *Advances in experimental social*

- psychology (Vol. 69, pp. 1–58). Elsevier. <https://doi.org/10.1016/bs.aesp.2023.11.001>
- 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>
- Vellani, V., Glickman, M., & Sharot, T. (2024). Three diverse motives for information sharing. *Communications Psychology*, 2(1), 107. <https://doi.org/10.1038/s44271-024-00144-y>
- Vidal-Perez, J., Dolan, R. J., & Moran, R. (2025). Disinformation elicits learning biases. *eLife*, 14. Article RP106073. <https://doi.org/10.7554/eLife.106073>
- Vorms, M., Harris, A. J. L., Topf, S., & Hahn, U. (2022). Plausibility matters: A challenge to Gilbert's Spinozan account of belief formation. *Cognition*, 220, 104990. <https://doi.org/10.1016/j.cognition.2021.104990>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 100(3), 426–432. <https://doi.org/10.1037/a0022790>
- Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39(5), 806–820. <https://doi.org/10.1037/0022-3514.39.5.806>
- Wiswall, M., & Zafar, B. (2015). How do college students respond to public information about earnings? *Journal of Human Capital*, 9(2), 117–169. <https://doi.org/10.1086/681542>
- Woodward, T. S., Moritz, S., Cuttler, C., & Whitman, J. C. (2006). The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, 28(4), 605–617. <https://doi.org/10.1080/13803390590949511>
- Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1310–1321. <https://doi.org/10.1098/rstb.2011.0416>
- Zorowitz, S., Solis, J., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nature Human Behaviour*, 7(10), 1667–1681. <https://doi.org/10.1038/s41562-023-01640-7>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.