Failure to Replicate Burton, Harris, Shah & Hahn (2021):

There is No Belief Update Bias for Neutral Events

Neil Garrett[1] & **Tali Sharot**[2,3]

[1]School of Psychology, University of East Anglia
[2]Affective Brain Lab, Department of Experimental Psychology, University College London
[3]The Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London

Highlights

- Only one of four experiments the authors conducted shows an update bias in neutral events. Experiments 2, 3 and 4 do not replicate the authors' own results, neither does aggregating data over all 4 experiments.
- The authors alter a well-established task, introducing confounds – including a skewed set of base rates - that are absent in the original task and are well known to produce false findings.
- We attempt to replicate Burton's et al.'s results in a new study, using the established version of the task, and fail to find an update bias in neutral events.

Abstract
We investigate Burton's et al.'s, recent findings of a belief update bias for neutral events. First, we find that Burton et al. fail to replicate their own findings in three out of the four experiments they conduct. When aggregating their data over their four experiments (500 participants) the results do not support a belief update bias for neutral events. In an attempt to replicate their findings, we collect a new data set employing the original belief update task design, but with neutral events. A belief update bias for neutral events is not observed. Finally, we highlight the wide range of statistical errors and confounds in Burton et al.'s design and analysis and the misleading statements they make.

We attempted to replicate Burton et al.'s findings of seemingly biased belief updating with neutral stimuli. We were unable to replicate these results when analysing Burton et al.'s own data (which uses a confounded set of stimuli). Neither were we able to replicate these results when collecting new data using an unconfounded set of stimuli.

Burton et al., loosely base their study on the belief update task (Garrett et al., 2018; Garrett and Sharot, 2014; Ossola et al., 2020; Sharot et al., 2011a; Sharot and Garrett, 2016; Garrett and Sharot, 2017; Kappes et al., 2020; Kuzmanovic et al., 2016; Ma et al., 2016; Moutsiana et al., 2015). The belief update task has previously revealed that healthy individuals tend to update their beliefs to a greater extent in response to unexpected positive information (e.g., learning that the likelihood of being a victim of card fraud is lower than expected) than negative information (e.g., learning it is higher than expected). This phenomenon can lead to optimistic beliefs, is absent in depression (Garrett et al., 2014; Korn et al., 2014) and is reduced when individuals come under threat (Garrett et al., 2018).

Rather than examining biases in response to information about events that are either negative (events one would prefer not to occur, such as robbery) or positive (events one would prefer to occur, such as winning a prize), Burton et al. examine if people update their beliefs to a greater degree after learning that neutral events (events one is indifferent about) are less likely than previously thought compared to more likely. This is the pattern that is often observed for negative events (e.g., Sharot et al., 2011), and the reverse pattern has been observed for positive events (Garrett and Sharot, 2017).

Instead of using the classic task (Garrett et al., 2014, 2018; Garrett and Sharot, 2014, 2017; Kappes et al., 2018; Korn et al., 2014; Ossola et al., 2020; Sharot et al., 2011a; Sharot and Garrett, 2016), to investigate this, Burton et al. alter the task, changing the response scale and the distribution of probability base rates, among other modifications. These modifications - as discussed in detailed below - have been previously documented by us to introduce confounds, not present in the original task, which will lead to false positives (Garrett and Sharot, 2017).

1. Failure to find an update bias in neutral events in Burton's et al., own data.

Burton et al., claim to show a bias in belief update for neutral events. The problem, put simply, is that Burton et al.'s own data does not show a bias in updating beliefs about neutral events in three of their four studies or in the aggregated data.

Burton et al., conduct four experiments, each analysed using four approaches: Linear Mixed Models (LMM) (Marks & Baines, 2017), Bayesian analysis (Shah et al., 2016), Reinforcement Learning (Kuzmanovic and Rigoux, 2017) and the classic approach - Linear Regression (Garrett et al., 2018; Garrett and Sharot, 2014; Kappes et al., 2020; Moutsiana et al., 2013; Sharot et al., 2011a; Ossola et al., 2020). The three approaches which they report in the supplementary material - Bayesian, Reinforcement Learning and Linear Regression - fail to show a bias in updating for neutral events (see **Table 1** below). This includes the Bayesian ratio approach that the authors themselves had advocated for previously (Shah et al., 2016).

Moreover, the authors aggregate their results over three experiments, despite not pre-registering this approach. Why aggregate if you aim to replicate? Regardless, they fail to aggregate over all four experiments, rather they quite oddly select to aggregate over only three experiments. When we use their data and code to aggregate over all four experiments, we find that the aggregated results do not show a bias in neutral events in three out of the four analytic approaches reported in supplementary material, including the classic approach. This means that the claim the authors make repeatedly in the manuscript - that the aggregated supplementary results show a belief update bias in neutral events - is patently false.

In the code the authors used to compile the aggregate data for the Bayesian Ratio measure (available here: https://osf.io/3nteq/?view_only=9ea1dcb105164bda9f35228b3bb3495c, see code lines 562 and 613), the authors set up the data frame to compile the data with 500 rows (one row for each participant) - which is participants for all four studies (there are 100 participants each in studies 1, 2, and 3, 200 participants in study 4). This results in 200 rows of missing data. This suggests that the authors initially set up the code to aggregate data over all four experiments, but then removed the relevant parts of the code to read in the data from Experiment 4, perhaps after observing the null result.

	Exp 1 (N = 100)	**Exp 2 (N = 100)**	**Exp 3 (N = 100)**	**Exp 4 (N = 200)**	**Aggregate (N = 500)**
Bayesian Difference	**Marginal (0.049)**	**Marginal (0.044)**	**Yes (0.013)**	**NO (0.22)**	**Yes (0.001)**
Bayesian Ratio	**Yes (0.001)**	**NO (0.784)**	**NO (0.449)**	**NO (0.93)**	**NO (0.06)**
Reinforcement Learning	**Yes (0.001)**	**NO (0.704)**	**NO (0.324)**	**NO (0.94)**	**NO (0.06)**
Regression Coefficient	**Yes* (0.001)**	**NO (0.662)**	**NO (0.540)**	**NO (0.28)**	**NO (0.07)**

TABLE 1 – Belief Update Bias in Neutral Stimuli? Burton et al.'s data reveals an effect of belief update bias for neutral events in study 1. This effect is highlighted in the title of the paper "Asymmetric Belief Updating Observed with Valence-Neutral Life Events". Yet, they fail to replicate their own effect in studies 2,3, 4 and in the aggregated data. p values are in parentheses
*This specific effect holds only if the authors unique trial exclusion protocol is followed. This is a protocol that is bespoke to them and has not - to our knowledge - ever been followed by researchers using the Belief Update Task. If all trials are included (as would normally be the case), this effect also disappears ($t(96) = 1.47$, $p=0.15$).

In the main text the authors use linear mixed models committing two errors. First, they fail to account for random effects (i.e. they only include a random intercept), thus inflating degrees of freedom by 10-40 fold. Doing this, they find the effect they are looking for. But it is well known that failure to incorporate these random effects would increase Type-1 error rates substantially, theoretically by 100% (Barr et al., 2013; Judd et al., 2012; Murayama et al., 2014). In fact, at least

one paper has recently been retracted for this reason (Fisher et al., 2015). The authors excuse this by saying that the correct model does not converge. This, however, does not change the fact that they are inflating the likelihood of type-1 error, which is why this analytic approach reveals an effect, but none of the other approaches do. We note that if one was to report LLMs that inflate degrees of freedom due to non-convergence, it would then be necessary to show that the same effect is observed using a different, statistically sound approach. In supplementary tables S2-S6 they report additional LLMs that do not converge. It is well-known that models that do not converge give unreliable estimates and should not be reported (Barr et al., 2013). The authors state they use LMMs because they wanted to follow a "precedent in this literature" citing one study - Marks & Baines (2017).

Second, they commit an analytical mistake well-known to produce false results (see Garrett & Sharot 2017 and Sharot & Garrett, 2021 for details). Namely, they do not control for *estimation errors*. An estimation error is the difference between a participants' estimate of the probability of an event occurring and the information provided about the actual probability. Certain task designs, such as Burton et al's, create a situation in which the estimation errors are greater in one condition than the other. In such cases, if participants are paying attention, they would by necessity update their beliefs more in the condition in which the estimation error is greater. This does not reflect a bias, rather it reflects basic learning. To avoid such a confound it is critical to control for estimation errors, as done in all past papers using the update bias task (Garrett et al., 2018; Garrett and Sharot, 2014; Ossola et al., 2020; Sharot et al., 2011a; Sharot and Garrett, 2016; Garrett and Sharot, 2017; Kappes et al., 2020; Kuzmanovic et al., 2016; Ma et al., 2016; Moutsiana et al., 2015). Burton et al., however, select not to do so and then unsurprisingly find a dubious effect.

As we detail in Section 3, when we attempt to replicate Burton et al's results by collecting fresh data, we fail to do so even when using LMMs that inflate degrees of freedom and that do not control for estimation errors.

2. Burton's et al., knowingly insert confounds into the task that have been documented to produce false results.

The authors take a task that has been very carefully designed and then change it. This includes **changing the response scale and skewing event probabilities** (see Supplementary Figure 1), **thus introducing confounds** that are well known (Garrett and Sharot, 2017; Sharot & Garrett, 2021), not least to the authors themselves who were previously criticised for generating spurious results in this way.

In the original task, very rare or very common events are not included - all event probabilities lie between 10% and 70%. Participants are told that the range of probabilities is between 3% and 77% and are only permitted to enter estimates within this range. This is done for two reasons. First, It is known that people's perception of very low probabilities is distorted (Kahneman and Tversky, 1979). **Second, it is important to ensure that the range of possible overestimation is equal to the range of possible underestimation**. That is, if all event probabilities lie between 11%

and 78% and participants are allowed to enter numbers between 0% and 100% then by design, they will not be able to update upwards as much as downwards. As a result, it has been established that if this paradigm is used to make claims about differences between downwards and upwards updating (regardless of whether the events are neutral, positive, negative or anything else), care has to be taken to use a set of base rates that are centred around the midpoint of the scale (Garrett and Sharot, 2017; Sharot & Garrett, 2021).

Burton et al., fail to do this in any of their experiments (see Supplementary Figure 1). The mean event base rate in their experiments are close to 30% on a 0-100 scale. We have been very clear in the past (Garrett and Sharot, 2017; Sharot and Garrett, 2021) that such a large positive skew in the base rate distribution like this, can artificially create greater updating for "downwards trials" (where the base rate presented is lower than participants first estimate) compared to "upwards trials" (where the base rate presented is higher than participants first estimate), which is exactly the pattern observed by Burton et al. **It is baffling why the authors deliberately and consistently chose to test their hypothesis over 4 experiments using a scale and base rate set well-known to them to produce false positives, and select not to control for this confound.**

Moreover, despite all past papers of the update bias including 20-40 trials per condition (that is per "good news" and "bad news") the authors have on average **7 trials per condition** for neutral stimuli. They are thus increasing noise, which increases the likelihood of false findings. In addition, the authors fail to collect ratings of possible confounds which are always collected and controlled for when using the task (Garrett et al., 2018, 2014; Garrett and Sharot, 2017, 2017; Moutsiana et al., 2015, 2013; Ossola et al., 2020; Sharot et al., 2011b). They say this is because it has been shown that controlling for these variables does not change the results. This logic is flawed. The fact that the effect holds after confounds are controlled for when it is a true effect does not mean that it will hold when it is not a true effect (that is a bias for neutral stimuli). Thus, their results are unreliable and uninterruptable.

It is surprising that the authors selected not to follow the known rigorous protocol of the belief update task, choosing instead not to control for well-known confounds, inserting statistical artefacts and distorting the task such that the data becomes non-informative.

3. Failure to replicate Burton, Shah, Harris & Hahn.

We run a study in an attempt to replicate Burrton et al., and failed to find an update bias in neutral events. In particular there was no difference in the amount of updating in response to observing probabilities that are lower than expected relative to probabilities that are higher than expected. This failure was observed regardless of the analytic approach adopted.

Our study follows the loose footed interpretation by Burton et al., of the Belief Update Task, while correcting for confounds they introduced (listed in section 2 above), which are absent in the original task (Sharot and Garrett, 2021). All analysis is restricted to events participants rated as neutral.

First, we ran linear mixed effects models, exactly as implemented by Burton et al. Update was entered as the dependent variable, direction of error (upwards/downwards) as the independent variable. Intercepts and slopes were taken as random effects (i.e., allowed to vary across participants). This revealed no difference in updating beliefs as a function of whether updating is upwards or downwards ($F(1, 81.94) = 0.11$, $p=0.74$, **Figure1c**). Even when re rerunning the model in a manner that inflates degrees of freedom (Barr et al., 2013; Judd et al., 2012; Murayama et al., 2014) by only including intercepts as random effects (as per the main analysis of Burton et al.), we still did not observe a bias in belief updating for neutral stimuli ($F(1, 1918.7)=0.68$, $p=0.41$). Finally, we reran the LMM excluding trials that could potentially be misclassified as upwards or downwards which can occur if the base rate presented sits between participants own estimate of the event occurring and their estimate of the base rate (Garrett and Sharot, 2017, 2014) (see **Methods**). Once again, this revealed no difference in updating beliefs as a function of whether updating is upwards or downwards ($F(1, 83.29) = 1.05$, $p=0.31$).

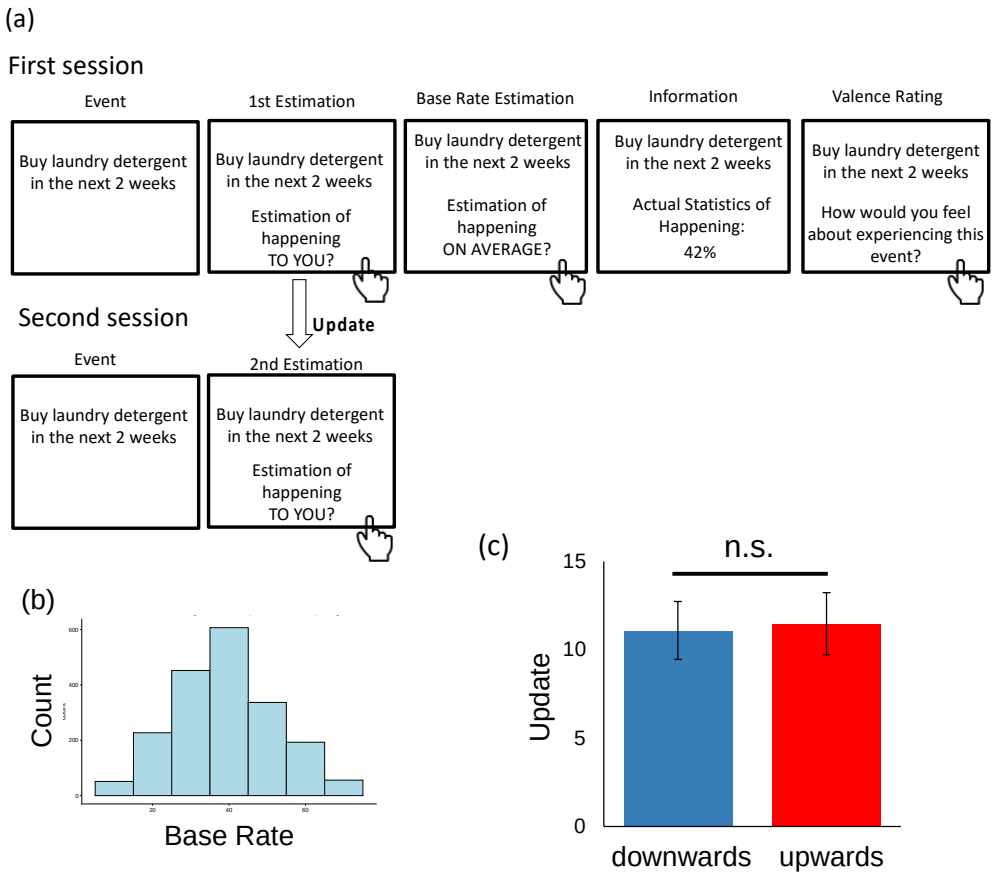


Figure 1. Task design. **(a)** On each trial, participants (N=100) were presented with a short description of 1 to 39 events and asked to estimate how likely this event was to occur to them. Estimates were entered into a text box displayed on the computer screen using a computer keyboard on a scale between 3% and 77%. Participants were then asked to estimate how likely the event was to occur on average in the population on the same scale. They were then presented with the average probability of that event occurring to a person like themselves (derived from factual sources, see Supplementary Materials). In a

second session, participants were asked to re-estimate how likely the event was to occur to them. For each event, an update term was calculated as the difference between the participant's first and second estimations, such that positive numbers indicate a move towards the base rate. **(b)** All events probabilities lay between 10% and 70% with a midpoint of 40. **(c)** Following Burton et al., 2021, we plot the magnitude of belief updating for events rated as neutral by participants, predicted by the linear mixed effects model with bars representing 95% confidence intervals. As can be observed, there is no asymmetry in belief updating for trials in which participants learned the event is more likely than they had originally estimated (upwards) or less likely (downwards). In other words, we were unable to replicate the difference in updating Barton et al report.

Next, we turned to examine whether learning scores differed on trials when participants received numbers that are higher than expected vs. lower than expected. Learning scores are regression coefficients which express the degree to which participants are updating their beliefs in proportion to the error made. This is the analytic method we use in our studies (e.g., Garrett et al., 2018, 2014; Kappes et al., 2018; Moutsiana et al., 2013; Sharot et al., 2011b) and Burton et al., report in their supplementary material. Comparing these for downwards versus upwards once again revealed no difference in learning rates about neutral events, regardless of whether participants learned the event was more likely than anticipated or less likely ($t(88) = 0.43$, $p=0.67$, paired sample ttest).

Burton et al., use three more analytic approaches. Two are Bayesian Analysis methods the authors have tried to popularise (Shah et al., 2016) and a third is a Reinforcement Learning approach developed by Kuzmanovic and Rigoux (Kuzmanovic and Rigoux, 2017). Note, however, that Burton et al.,'s implementation of this Reinforcement Learning approach is flawed and at odds with what was actually proposed by Kuzmanovic and Rigoux (see **Methods** for further details). Analysing the new data using these methods we find that two of these approaches reveal the *opposite* effect to that reported by Burton et al., (Bayesian ratio approach: median downwards = 0.46, median upwards = 0.67, $Z = 3.07$, $p=0.002$, paired sample Wilcoxon test; Burton et al.,'s "Reinforcement Learning"; median downwards = 0.58, median upwards = 0.80, $Z=3.27$, $p=0.001$, paired sample Wilcoxon test). Only one approach - the Bayesian Difference approach revealed an effect in the same direction as in Burton et al., (mean downwards = 0.08, mean upwards = 0.04, $t=-2.99$, $p=0.004$, paired sample t test). In sum, we failed to replicate Burton et al., 's (2021) findings of belief update bias for neutral events.

Conclusion

Burton et al., claim to show an update bias in neutral events. Yet, three out of four of their own experiments fail to show the effect in three of four analytic approaches they use (Bayesian, "Reinforcement Learning", Linear Regressions). Moreover, they claim to show the effect over the aggregated data of the experiments, but only include data of three of the four experiments conducted. Once data is aggregated over the four experiments, the effect is not observed. The effect is only observed using an LMM that inflated degrees of freedom by 10 fold to 40 fold, failing to account for random effects. An attempt to replicate Burton et al.'s findings by running a fresh study also fails to find a belief update bias in neutral events. Finally, there are an alarming number of nonsensical statistics and false reporting throughout Burton et al.'s manuscript (we

outline additional examples in the Supplementary Materials). In sum, we show that the claims made by Burton et al., are clearly not supported by their data, or anyone else's.

Methods

Participants

One hundred participants were recruited via the online platform Prolific. This sample size is the same as the one used by Burton et al., in Experiments 1-3. Completion of the experiment took approximately 1 hour and participants were compensated for their time. The study was approved by the UCL's Ethics Committee.

Task

The study involved two sessions (**Fig. 2**). In a first session, each participant was presented with one of 39 life events (e.g., "Buy laundry detergent in the next two weeks) and asked to imagine the event happening to them. They were then asked to estimate how likely that event was to happen to them (E1); participants were also asked to give a second estimate of the likelihood of the event happening to an average person in the population (eBR; an estimate of the base rate). Participants were instructed to type in each estimate between 3% and 77% and were not able to enter responses outside of this range. There were no restrictions on participants' response time. The order of the two estimates (E1 and eBR) was counterbalanced between subjects by randomly assigning each participant to one of two conditions: E1 followed by eBR (N=51), or eBR followed by E1 (N=49). After these two initial estimates were recorded, participants were shown the base rate statistic (BR) of the event happening to someone from the same socioeconomic environment as them, which ranged from 10% to 70%. Finally, participants were asked to rate how negative or positive they found the event on a five point scale (1 = very negative, 2 = negative, 3 = neural, 4 = positive, 5 = very positive). In a second session, which took place immediately after the first session, participants were asked to re-estimate how likely each event was to happen to them (E2). Again, there were no restrictions on participants' response time.

After completion of the task, we tested participants' memory for the information presented. Participants were asked to recall the information previously presented (BR) of each event. Subsequently, participants were then asked to rate all life events according to their past experience with each event (''Has this event happened to you before?" From 1 = never to 6 = very often), vividness of imagination (''How vividly could you imagine this event?" From 1 = not vivid to 6 = very vivid); familiarity (''Regardless if this event has happened to you before, how familiar do you feel it is to you from TV, friends, movies and so on?" From 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?" From 1 = not arousing at all to 6 = very arousing). The survey was constructed and presented using web based survey service Qualtrics.

Analysis

Life events were categorized as neutral for each participant individually according to their own evaluation. Specifically, events were classified as neutral if the participant rated the event as 3

during the task (mean [sd] number of trials rated neutral per participant: 20 [6.56]). Events that received a rating other than 3 were discarded.

Participants could either receive information in a 'downwards direction" or an 'upwards direction" depending on whether the participant initially overestimated or underestimated the probability of the event relative to the base rate, respectively. Specifically, if their first estimate (E1) was higher than the base rate presented (BR), the information would be categorized as ''downwards" and if their first estimate (E1) was lower than the base rate presented (BR), the information would be categorized as ''upwards". Trials in which the initial estimate was equal to the statistic presented were excluded from subsequent analyses as these could not be categorized into either condition. In addition, we followed the exclusion criterion employed by Burton et al.. Specifically, mean updates in each of the two conditions (upwards/downwards) were calculated and outliers were removed (±3 × the interquartile range).

Linear Mixed Effect Models

Belief update was calculated for each trial and participant as the difference between first and second estimate. As done previously (Garrett & Sharot, 2014; Garrett et al., 2014; Moutsiana et al., 2013, 2015; Sharot, Kanai et al., 2012) and followed by Burton et al., update was calculated such that positive scores indicate a move towards the base rate and negative scores a move away from the base rate:

update (downwards) = E1 – E2

update (upwards) = E2 – E1

Following Burton et al., we used a linear mixed effects (LMM) model with update entered as the dependent variable, direction of error (upwards/downwards) as a fixed factor, and participant as a random factor, including intercepts and slopes as random effects. In the syntax of the lme4 package, the specification for the regression was as follows:

update ~ direction + (1 + direction | Participant)

Following Burton et al., we then used Type III tests and Satterthwaite's approximation for degrees of freedom to calculate the statistical significance of the fixed effects. We also examined whether we could detect an effect if we ran the LMM without random slopes, i.e.

update ~ direction + (1 | Participant)

To be clear, we do **not** think this is a valid model specification but we wanted to test whether even with this very lenient approach that Burton et al. took to the data, a false positive could arise for neutral events when the proper experimental design was used.

Finally, we reran the LMM (with both random intercepts and slopes) excluding trials (25% of trials rated neutral) that would be assigned into a different category under and alternate classification scheme (Garrett and Sharot, 2017, 2014) in which trials were partitioned into downwards/upwards according to whether participants estimate of the base rate (eBR) was higher (downwards) or lower (upwards) than the base rate presented (BR).

Linear Regression

Next we examined the relationship between estimation errors and update. For each trial, an estimation error term was calculated as the unsigned difference between the probability presented and participants' first estimate on that trial (the likelihood the event happens to them, i.e. E1)

estimation error = | probability presented - first estimate |

We estimated the extent to which participants integrated new information into their beliefs by regressing absolute estimation errors against update scores separately for upwards and downwards trials for each participant:

Update (downwards) = b0 + b1*estimation error

Update (upwards) = b0 + b1*estimation error

This resulted in two scores (the unstandardized regression coefficient b1 in the equations above) for each participant: one for upwards trials and one for downwards trials. These were compared with one another using paired sample ttests.

Bayesian Analysis

This analysis directly follows the procedure of Burton et al.

Participants estimate of each event occurring to themselves in the future (E1) and estimate of the base rate (eBR) were used to calculate an Implied likelihood Ratios (LHR) on each trial as:

$$LHR = \frac{E1}{1 - E1} \div \frac{eBR}{1 - eBR}$$

This LHR was then used in conjunction with the base rate presented (BR) to calculate trial by trial predicted posterior odds, calculated as:

$$Posterior\ Odds = \frac{BR}{1 - BR} \times LHR$$

Finally, Posterior Odds were used in conjunction with E1 to calculate the degree to which a rational Bayesian agent would update on each trial, as:

$$Bayesian\ Update = \left| \frac{E1 - Posterior\ Odds}{1 + Posterior\ Odds} \right|$$

From here, two measures were calculated (Bayesian Difference, Bayesian Raito), both of which compare Bayesian Update with participants actual update (defined as above) observed:

$$Bayesian\ Difference = Bayesian\ Update - Update$$

$$Bayesian\ Ratio = \frac{Update}{Bayesian\ Update}$$

Each of these measures were compared for upwards trials versus downwards trials using Wilcoxon paired difference test or paired sample ttests.

Reinforcement Learning

This analysis directly follows the procedure of Burton et al. which claims to follow a modelling approach presented by Kuzmanovic and Rigoux (2017).

Updates (calculated as above) are modelled as:

$$Update = \alpha \times \delta \times (1 - rP \times w)$$

δ is a prediction error, defined as the difference between participants estimate of the base rate (eBR) and the actual base rate presented (BR):

δ = *eBR - BR.*

rP - "relative personal knowledge" - is calculated according to whether estimates of the base rate are higher or lower than estimates of ones own likelihood, as:

rP = (eBR - E1)/eBR if E1 < eBR
rP = (E1 - eBR)/(100 – eBR) if E1 > eBR
rP = 0 if E1 = eBR

α and w are free parameters. α, the learning rate, determines the degree to which beliefs change in proportion to the prediction error. w accounts for participants' individual variability in their sensitivity to rP. Rather than fit this model to participants updates to derive α and w estimates

for each participant – which would be the normal approach for an RL model of this form – Burton et al. instead do the following.

First, they assume that w is 1 for all participants. This enables them to reduce the update equation to:

$$Update = \alpha \times \delta \times \ (1 - rP)$$

Which in turn enables them to rearrange the terms of the Update equations such that α sits as the dependent variable:

$$\alpha = \frac{Update}{\delta \times (1 - rP)}$$

Second, they use the above formulation to calculate a 'trial by trial' learning rate (trials where update = 0, i.e. beliefs stay the same, the authors assume that $\alpha = 0$). We note that the approach to modelling here is rather at odds with a conventional RL approach whereby a single best fit learning rate (and w parameter) would be derived to account for *all* of each participants' updates and model comparison is used to compare models with different combinations of parameters. We do not suggest others try to follow this approach, we are simply following Burton et al.'s flawed recipe.

α is then averaged for each participant for each condition (upwards, downwards) and then the two conditions compared using a Wilcoxon paired difference test.

Acknowledgements

Neil Garrett is funded by a Sir Henry Wellcome Fellowship (209108/Z/17/Z). Tali Sharot is funded by a Wellcome Trust Senior Researcher Fellowship.

References

Burton, J; Harris, A; Shah, P; Hahn, U; (2021) Optimism Where There is None: Asymmetric Belief Updating Observed with Valence-Neutral Life Events. Cognition .

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang* **68**:10.1016/j.jml.2012.11.001. doi:10.1016/j.jml.2012.11.001

Fisher CI, Hahn AC, DeBruine LM, Jones BC. 2015. Women's Preference for Attractive Makeup Tracks Changes in Their Salivary Testosterone. *Psychol Sci* **26**:1958–1964. doi:10.1177/0956797615609900

Garrett N, González-Garzón AM, Foulkes L, Levita L, Sharot T. 2018. Updating beliefs under perceived threat. *Journal of Neuroscience* **38**:7901–7911.

Garrett N, Sharot T. 2017. Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and cognition* **50**:12–22.

Garrett N, Sharot T. 2014. How robust is the optimistic update bias for estimating self-risk and population base rates? *PLoS One* **9**:e98848.

Garrett N, Sharot T, Faulkner P, Korn CW, Roiser JP, Dolan RJ. 2014. Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Frontiers in human neuroscience* **8**:639.

Judd CM, Westfall J, Kenny DA. 2012. Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology* **103**:54–69. doi:10.1037/a0028347

Kahneman D, Tversky A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**:263–291. doi:10.2307/1914185

Kappes A, Faber NS, Kahane G, Savulescu J, Crockett MJ. 2018. Concern for Others Leads to Vicarious Optimism. *Psychol Sci* **29**:379–389. doi:10.1177/0956797617737129

Kappes A, Harvey AH, Lohrenz T, Montague PR, Sharot T. 2020. Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience* **23**:130–137. doi:10.1038/s41593-019-0549-2

Korn CW, Sharot T, Walter H, Heekeren HR, Dolan RJ. 2014. Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine* **44**:579–592. doi:10.1017/S0033291713001074

Kuzmanovic B, Jefferson A, Vogeley K. 2016. The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage* **133**:151–162. doi:10.1016/j.neuroimage.2016.02.014

Kuzmanovic B, Rigoux L. 2017. Valence-Dependent Belief Updating: Computational Validation. *Frontiers in Psychology* **8**:1087. doi:10.3389/fpsyg.2017.01087

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 of the United States of America* **113**:9256–61. doi:10.1073/pnas.1604285113

Moutsiana C, Charpentier CJ, Garrett N, Cohen MX, Sharot T. 2015. Human Frontal-Subcortical Circuit and Asymmetric Belief Updating. *J Neurosci* **35**:14077–14085. doi:10.1523/JNEUROSCI.1120-15.2015

Moutsiana C, Garrett N, Clarke RC, Lotto RB, Blakemore S-J, Sharot T. 2013. Human development of the ability to learn from bad news. *Proceedings of the National Academy of Sciences* **110**:16396–16401.

Murayama K, Sakaki M, Yan VX, Smith GM. 2014. Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **40**:1287–1306. doi:10.1037/a0036914

Ossola P, Garrett N, Sharot T, Marchesi C. 2020. Belief updating in bipolar disorder predicts time of recurrence. *eLife* **9**:e58891. doi:10.7554/eLife.58891

Shah P, Harris AJ, Bird G, Catmur C, Hahn U. 2016. A pessimistic view of optimistic belief updating. *Cognitive Psychology* **90**:71–127.

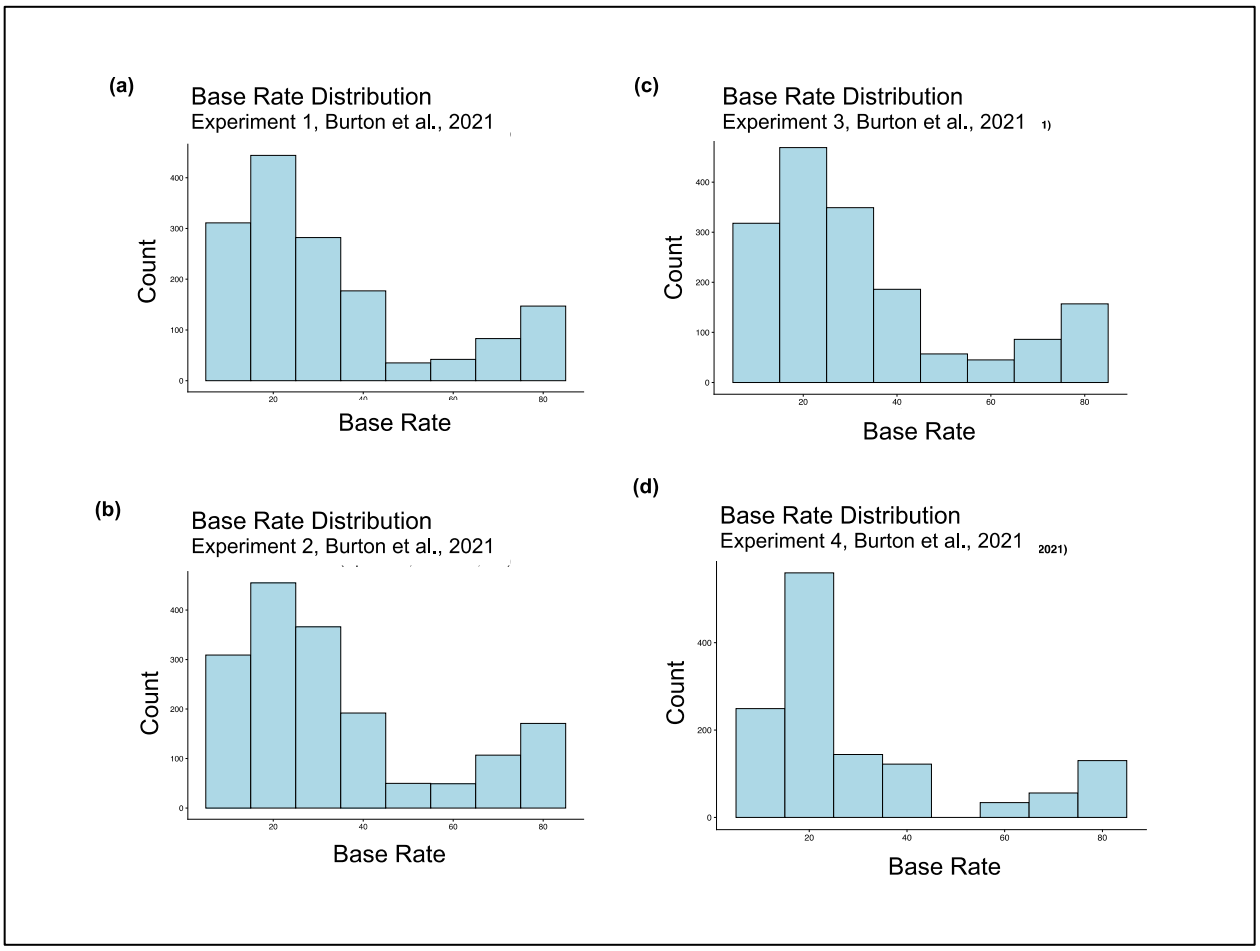
Sharot T, Garrett N. 2021. A Guideline and Cautionary Note: How to Use the Belief Update Task Correctly. doi:10.31234/osf.io/st4vu

Sharot T, Garrett N. 2016. Forming beliefs: Why valence matters. *Trends in cognitive sciences* **20**:25–33.

Sharot T, Korn CW, Dolan RJ. 2011a. How unrealistic optimism is maintained in the face of reality. *Nature neuroscience* **14**:1475–1479. doi:10.1038/nn.2949

Sharot T, Korn CW, Dolan RJ. 2011b. How unrealistic optimism is maintained in the face of reality. *Nature neuroscience* **14**:1475–1479.

Supplementary Materials



Supplementary Figure 1. Base rates used by Burton at al. for neutral events in each experiment. The mean base rate in each experiment sits well below 50, the midpoint of the response scale they opt to use, which was 0 to 100 (Experiment 1: $t(1520)=-31.40$, $p<0.001$; Experiment 2: $t(1698)=-30.22$, $p<0.001$; Experiment 3: $t(1666)=-32.69$, $p<0.001$; Experiment 4: $t(1294)=-33.19$, $p<0.001$, one sample ttests vs 50).

List of Stimuli

Event	Base Rate (BR)	Source
Meet with your supervisor in the next four weeks	56	Garrett & Sharot, 2017
Participate in a game of sport in the next four weeks	29	Burton et al., 2021
The next car that passes you is the colour black	20	Burton et al., 2021
Use more than 3.7GB of mobile data over the next four weeks	17	Burton et al., 2021
Meet your future spouse through an online dating service	38	Burton et al., 2021
Marry someone with a different political affiliation to you	26	Burton et al., 2021
The next person that you talk to has a positive impression of Golf cars	62	Webpage no longer available
The next person that you talk to has a positive impression of Ford Focus car	59	Webpage no longer available
The next new person you meet has a reduced ability to digest luctose	65	https://medlineplus.gov/genetics/condition/lactose-intolerance/#:~:text=Approximately%2065%20percent%20of%20the,people%20affected%20in%20these%20communities.
The next email sent to you will be spam	55	https://www.statista.com/statistics/270899/global-e-mail-spam-rate/
The likelihood that you will receive less than 28 spam call in the next 4 weeks	52	https://techcrunch.com/2020/12/07/spam-calls-grew-18-this-year-despite-the-global-pandemic/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAL-epd69o2ik7B0KqAGYLjcKmZtyDMlkgyEm-v6sgnuE9md9-C8_dGLTYhY1NygVeGl0WXOcDlax4XgKENOo-BHSnDLHSt6iYDs13r4heTOx0AGZzGKt-LbuTRp39U822o7Pfqgz9PASadRpHg17PoA5XUB2AvvmHukjru5OBy35
The next salesperson you see will have brown hair	48	https://beachwaveperm.com/most-common-hair-color-in-uk/#:~:text=Report%20Ad-,2.,South%20and%20amongst%20indigenous%20Brits.
Have an extra artery in the arm	30	https://www.sciencefocus.com/news/humans-are-evolving-an-extra-artery-in-the-arm/

The next car you pass has been cleaned at least once in the past 3 months	55	https://www.intelligentcarleasing.com/blog/how-clean-is-your-car-study/
Buy laundry detergent in the next two weeks	42	http://datacolada.org/22#footnote_0_574
Download 1-3 new apps for your phone in the next month	32	https://techcrunch.com/2017/08/25/majority-of-u-s-consumers-still-download-zero-apps-per-month-says-comscore/
The next woman you walk past has a foot that measures (from heel to toe) 240mm or less	41	https://www.nature.com/articles/s41598-019-55432-z.pdf
Inhale and exhale 11,000 liters of air tomorrow	48	https://www.sharecare.com/health/air-quality/oxygen-person-consume-a-day
Use 90 gallons of water or less on a weekday next week	49	https://www.usgs.gov/special-topic/water-science-school/science/water-qa-how-much-water-do-i-use-home-each-day?qt-science_center_objects=0#
The next stranger you see, walks at an average speed of 1.24 m/sec or less	44	https://core.ac.uk/download/pdf/82088742.pdf
The next stranger you walk past, walked between 6000 and 10000 steps the previous day	33	https://www.researchgate.net/figure/Distribution-of-average-number-of-steps-per-day_tbl2_6747371
The next stranger you walk past that owns a TV will watch 41 or less TV adverts tomorrow	51	https://www.statista.com/statistics/486685/number-of-tv-ads-seen-daily-in-the-uk/
The next stranger you walk past does not drink tea	37	https://www.statista.com/chart/23081/most-consumed-drink-types-uk/
The next student you pass in the street (aged 13-27) either types at a speed of 10-20 or 30-40 words per minute	43	https://onlinetyping.org/blog/average-typing-speed.php#students
The next stranger you walk past aged 65 or more does not have a smartphone	39	https://www.statista.com/statistics/489255/percentage-of-us-smartphone-owners-by-age-group/
The next stranger you walk past lives in a house with two, three or four other persons	35	https://www.statista.com/statistics/281627/households-in-the-united-kingdom-uk-by-size/#statisticContainer
The next adult male you meet has a body mass index between 18.5 and 24.9	30	https://researchbriefings.files.parliament.uk/documents/SN03336/SN03336.pdf
The next woman you meet (aged 45-54) is 5ft 5inches tall or less	70	https://allcountries.org/uscensus/230_cumulative_percent_distribution_of_population_by.html
The next male you meet (aged 35-44) is 6ft 1inch tall or above	10	https://allcountries.org/uscensus/230_cumulative_percent_distribution_of_population_by.html
A prime number or a number less than 10 is drawn first in the UK national lottery this Saturday	23	https://www.national-lottery.co.uk

The first car you see next Monday has driven over 10,000 miles in the past year	23	https://www.statista.com/statistics/513456/annual-mileage-of-motorists-in-the-united-kingdom-uk/
The next adult male you meet drinks 5-7 cups of coffee per day	36	https://www.researchgate.net/figure/Frequency-distribution-of-coffee-consumption-cups-day-at-age-32-years-and-its_tbl1_228615882
The next song you hear is 210 seconds or less in duration	33	http://theinformationdiet.blogspot.com/2011/11/probability-distribution-of-song-length.html
The next stranger you pass in the street is less than 15 or older than 65	36	https://www.statista.com/statistics/270370/age-distribution-in-the-united-kingdom/
Get a haircut in the next 4 weeks	45	Burton et al., 2021
Drink between 56 and 84 cups of coffee over the next four weeks	43	Burton et al., 2021
Yawn 6 times or less tomorrow	38	https://reader.elsevier.com/reader/sd/pii/0031938495020144?token=0F37B34E4D887C55A3583F9FCAA6C508C641E69209FDFA6EA0C834EBBD60B87D13545017AA93044168AB22595708E9AB&originRegion=eu-west-1&originCreation=20211019115718
The next 50 year old man you meet has an arm span of 173cm or less	33	https://allcountries.org/uscensus/230_cumulative_percent_distribution_of_population_by.html
Use 893 KwH of electricity (or more) in a month at least once in the next year	47	https://www.eia.gov/tools/faqs/faq.php?id=97&t=3

Supplementary Table 1. List of events used in Garrett and Sharot, 2021 along with their srouces. These are normally distributed around a mean of 40 (the midpoint of the scale used).

Erroneous reporting by Burton et al.

Burton et al. is riddled with an alarming number of erroneous statistics. We highlight some examples below.

Study	*Event Valence*	*Median ratio measure for downwards trials*	*Median ratio measure for upwards trials*	*Z*	*p-value*
1	Positive	0.00	0.00	1.41	0.921
	Neutral	0.57	0.04	-3.73	< 0.001
	Negative	0.49	0.00	-4.92	< 0.001
2	Positive	0.00	0.17	-1.19	0.116
	Neutral	0.53	0.28	0.79	0.784
	Negative	0.51	0.07	-2.26	0.012
3	Positive	0.00	0.46	-2.14	0.016
	Neutral	0.53	0.28	-0.13	0.449
	Negative	0.51	0.09	-3.71	< 0.001

Aggregate	Positive	0.00	0.10	-2.36	0.009
	Neutral	0.53	0.25	-2.34	0.010
	Negative	0.51	0.00	-6.25	< 0.001

Supplementary Table 2. Incorrect results reported by Burton et al., of Wilcoxon signed rank tests comparing Bayesian ratio measures that compare participants' updating to rational Bayesian predictions. All of the test statistics reported in the 5th column are nonsensical (by means of an example, the first entry reports the same ratio for downwards and upwards – that cannot generate a Z score – i.e. a meaningful difference - of 1.41 with a pvalue of 0.921. Checking each of the z statistics against the pvalues reported (e.g., here: https://www.socscistatistics.com/pvalues/normaldistribution.aspx) verifies that all of these statistics are in fact false. Note also that results from the final experiment – Experiment 4 – are omitted and are not included in the aggregate calculations.

Study	*Event Valence*	*Median learning rate for downwards trials*	*Median learning rate for upwards trials*	*Z*	*p-value*
1	Positive	0.00	0.00	0.28	0.610
	Neutral	0.61	0.04	-3.04	0.001
	Negative	0.60	0.00	-4.83	< 0.001
2	Positive	0.00	0.19	-0.89	0.187
	Neutral	0.68	0.45	0.53	0.704
	Negative	0.56	0.14	-1.86	0.032
3	Positive	0.00	0.55	-2.45	0.007
	Neutral	0.64	0.30	-0.46	0.324
	Negative	0.57	0.12	-4.02	< 0.001
Aggregate	Positive	0.00	0.15	-2.15	0.016
	Neutral	0.66	0.28	-2.14	0.016
	Negative	0.58	0.00	-6.12	< 0.001

Supplementary Table 3. Incorrect results reported by Burton et al., of Wilcoxon signed rank tests comparing the learning rate measure derived from the reinforcement learning model presented by Kuzmanovic and Rigoux (2017). All of the test statistics (Z scores) reported in the 5th column are incorrect. Note also that results from the final experiment – Experiment 4 – are omitted and are not included in the aggregate calculations.

Study	*Event Valence*	*Mean coefficient for downwards trials*	*Mean coefficient for upwards trials*	*t*	*p-value*
1	Positive	-0.02	0.10	-1.90	0.060
	Neutral	0.20	0.00	3.57	< 0.001
	Negative	0.20	0.20	2.23	0.028
2	Positive	0.04	0.30	-3.22	0.002
	Neutral	0.11	0.06	0.44	0.662
	Negative	0.23	0.11	1.40	0.165
3	Positive	-0.14	0.11	-2.06	0.042
	Neutral	0.13	0.08	0.62	0.540
	Negative	0.40	0.13	1.52	0.132

Aggregate	Positive	-0.04	0.17	-4.03	< 0.001
	Neutral	-0.15	0.05	1.75	0.081
	Negative	0.27	0.15	2.76	0.006

Supplementary Table 4. Incorrect results reported by Burton et al. comparing regression coefficients whereby estimation errors (the difference between first estimates and the information provided) are used to predict update values. Note that the coefficient for downwards trials is positive for neutral events in Experiments 1, 2 and 3 yet negative when this is calculated in aggregate (highlighted) – this is clearly impossible. Note also that Neutral Downwards and Negative Upwards Coefficients are identical in Experiments 1, 2 and 3 - this is not a coincidence and occurs because the incorrect value is reported for Upwards Negative in each case (the value for Downwards Neutral is duplicated and the correct value for Upwards Negative is not presented). The same error occurs in Table 5 (below). Note also that results from the final experiment – experiment 4 – are omitted and not included in the aggregate calculations.

Study	*Event Valence*	*Mean coefficient for downwards trials*	*Mean coefficient for upwards trials*	*t*	*p-value*
1	Positive	0.04	0.11	-0.85	0.398
	Neutral	0.21	0.04	3.04	0.003
	Negative	0.30	0.21	3.78	< 0.001
2	Positive	0.08	0.06	0.28	0.777
	Neutral	0.21	0.06	1.83	0.071
	Negative	0.33	0.21	1.07	0.285
3	Positive	-0.04	-0.06	0.08	0.934
	Neutral	0.17	0.01	1.89	0.061
	Negative	0.31	0.17	1.80	0.075
Aggregate	Positive	0.03	0.04	-0.13	0.895
	Neutral	0.19	0.04	3.62	< 0.001
	Negative	0.31	0.19	3.20	0.002

Supplementary Table 5. Incorrect results reported by Burton et al. comparing regression coefficients whereby base rate errors (the difference between estimates of the base rates and the information provided) are used to predict update values. Note that Neutral Downwards and Negative Upwards Coefficients are identical in Experiments 1, 2, 3 and in the aggregate results (this is highlighted in yellow). Just as is the case in the previous analysis they report, occurs because the incorrect value is reported for Upwards Negative in each case (the value for Downwards Neutral is duplicated and the correct value for Upwards Negative is not presented). The same error occurs in Table 4 (above). Note also that results from the final experiment – experiment 4 – are omitted and not included in the aggregate calculations.