Confidence Drives a Neural Confirmation Bias

Max Rollwage (max.rollwage.16@ucl.ac.uk)
Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom
Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

Tobias Hauser (t.hauser@ucl.ac.uk)
Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom
Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

Alisa Loosen (a.loosen.17@ucl.ac.uk)
Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom
Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

Rani Moran (r.moran@ucl.ac.uk)
Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom
Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

Raymond J. Dolan (r.dolan@ucl.ac.uk)
Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom
Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

Stephen M. Fleming (stephen.fleming@ucl.ac.uk)
Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom
Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

Abstract

Contested issues (e.g. climate change) can generate polarised and entrenched opinions. A prominent source of polarisation is confirmation bias, where evidence against one's position tends to be selectively disregarded. Despite the ubiquity of confirmation bias, its computational and neural underpinnings are unknown. Across three studies, we combined psychophysical modelling and magnetoencephalography (MEG) with a perceptual discrimination task to address its neural underpinnings. Convergent evidence from these studies show that, at a neural level, accumulation of confirming evidence is facilitated in comparison to disconfirming evidence. This effect is amplified when people are highly confident in an initial decision, reducing the likelihood of behavioural changes of mind. We conclude that confidence shapes a selective neural gating for choice-consistent information, revealing a neuronal mechanism underlying a confirmation bias.

Keywords: confirmation bias; evidence accumulation; magnetoencephalography; drift-diffusion model; confidence

Background

The tendency to accept information that confirms our beliefs while disregarding disconfirming evidence is known as confirmation bias (Nickerson, 1998). Although an extensive literature has documented this bias in behavior (Nickerson, 1998), the underlying cognitive and neuronal processes are not yet understood.

Here we combine theoretical models and neural metrics of evidence accumulation to identify alterations in information processing that underpin the phenomenon of confirmation bias. Across three experiments, human participants viewed a cloud of dots briefly moving on a computer screen, and tasked to decide whether the dots were moving to the left or the right (Figure 1A). In every trial, participants were presented with a sample of moving dots (pre-decision evidence) before indicating their initial decision and confidence in their choice. They were then presented with a new sample of moving dots (post-decision evidence) before making a final choice and providing a confidence estimate. Importantly, pre- and post-decision evidence indicated the

same direction of motion, such that the post-decision evidence was always helpful.

A causal role of confidence on changes of mind (Study 1, N=28)

In such task, an ideal Bayesian observer changes its mind in light of new evidence that an original choice is wrong, whereas a confirmation bias blunts such belief flexibility (Fleming, van der Putten, & Daw, 2018; Rollwage, Dolan, & Fleming, 2018). In a first experiment we investigated the role of confidence on post-decision evidence processing, as we hypothesized that a confirmation bias would be strongest when participants are highly confident in their decision. In order to dissociate confidence from objective performance we used a psychophysical manipulation (a "positive evidence" manipulation increases the motion coherence in the correct direction but also adds some coherence in the opposite direction; Zylberberg, Barttfeld, & Sigman, 2012) to selectively increase participants confidence ($t(27)=3.2$ $p=.002$, Figure 1B) while leaving performance (Bayesian t-test indicating equality: $BF_{01}=4.61$; Figure 1C) and reaction times (Bayesian t-test indicating equality: $BF_{01}=4.51$) unaffected. As predicted, this experimental manipulation led to a reduction in changes of mind ($t(20)=3.51$, $p=.002$, Figure 1D), an effect fully mediated by a boost in confidence ($a \times b$; $\beta= -4.84$, $p<10^{-6}$; Figure 1E). This initial mediation analysis indicated that confidence reduces changes of mind, plausibly by promoting a bias towards confirmatory evidence.

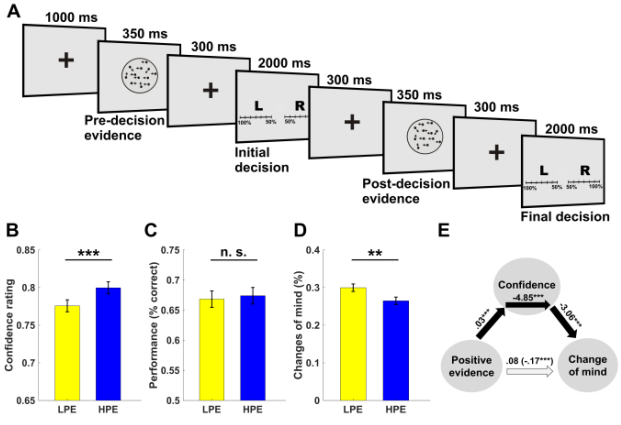

Figure 1: Task design for all experiments and results of study 1. **A** Trial timeline. **B&C** A psychophysical manipulation of positive evidence selectively increased confidence in the first decision (**B**) while keeping accuracy constant (**C**). Group means ± S.E.M **D** A positive evidence manipulation reduced changes of mind. Group means ± S.E.M. **E** Multilevel mediation analysis indicates that the effects of the positive evidence manipulation on changes of mind was fully mediated by a shift in confidence. The direct effect is shown before (path c presented in brackets) and after controlling for confidence (path c' presented outside of the brackets). **p<.01, ***p<.001; LPE= low positive evidence condition; HPE= high positive evidence condition

Confidence induces a selective gain for choice-consistent information (Study 2, N=24)

We next investigated the influence of confidence on post-decision evidence accumulation in more detail by applying drift-diffusion modelling (Gold & Shadlen, 2007) to the final decision. Within the drift-diffusion framework, confidence might reduce changes of mind through two potential mechanisms. First, confidence might shift the starting point of the post-decision accumulation process closer to the bound associated with the initial decision (Figure 2A upper-panel). Second, confidence may induce selective accumulation of evidence that is in line with an initial decision (influence on drift-rate; Figure 2A lower-panel). Out of 11 competing drift-diffusion models the observed data were best accounted for by a model that incorporated an influence of confidence and choice consistency (i.e. post-decision evidence that either confirms or disconfirms the initial decision) on both starting point and drift-rate (Figure 2B-D). The interaction effect of confidence × choice-consistency on the starting point ($p < 10^{-20}$; Figure 2D right-hand panel), indicated that participants started the accumulation process closer to the bound of the initial decision when they were highly confident in their choice. There was a similar interaction effect on the drift-rate ($p < 10^{-20}$; Figure 2D right-hand panel) indicating that participants selectively accumulated evidence supporting their initial decision, and more so when they were highly confident.

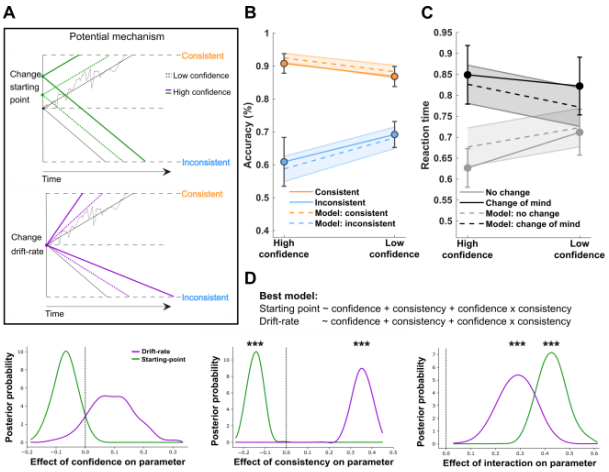

Figure 2: Drift-diffusion model fits to the final decision. **A** Illustration of how confidence reduces changes of mind through either shifting the starting point towards the decision bound of the initial decision (upper-panel) and/or a selective increase of drift-rate for evidence supporting the initial decision (lower-panel). **B & C** Model simulations (of the best fitting model) reproduce behavioural patterns of accuracy and reaction times of the final decision. Model simulations shown in dotted lines and behavioral data

shown in solid lines. Error bars indicate ± 95% confidence intervals. **D** Posterior distribution of model parameters of the best fitting model. The right panel shows dependencies of the drift-rate (purple lines) and starting point (green lines) on initial confidence (left panel), choice-consistency (middle panel) and the interaction between initial decision × choice consistency. Dotted vertical lines represent an effect of zero/no effect. Note that these dependencies are simultaneously fitted, controlling for mutual influences.***p<.0001

MEG recordings reveal neural the mechanisms underlying confirmation bias (Study 3, N=25)

While our drift-diffusion model fits support a distinct influence of initial choice and confidence on post-decisional processing, they allow only indirect inference on how confidence affects evidence accumulation. To directly quantify such changes we used MEG to obtain a time-resolved neural metric of post-decisional accumulation. Specifically, we trained a support-vector machine classifier on brain activity (normalized amplitude of all MEG channels) at each time point in the pre-decision time window to predict which choice (left or right) a person will make on a particular trial. We then applied the trained classifier to brain activity at the corresponding time point in the post-decision time window to derive a probabilistic prediction of internal evidence favouring a leftward versus rightward decision. To summarize this evidence accumulation process, we fitted a linear regression to the time series of neural predictions within each trial (see Figure 3A right panel), giving us a trial-by-trial neural measure of the starting point (intercept) and drift rate (slope). These measures of evidence accumulation (slope) should be responsive to the presented motion direction during the post-decision period, and we show this was indeed the case (β=. 08, $p<10^{-14}$).

Having validated a neural metric of evidence accumulation, we turned next to our central question of whether confidence induces a selective accumulation for choice-consistent information. As hypothesized, we found that after high confidence (vs. low confidence) decisions, accumulation of neural evidence was facilitated if it was consistent with an initial decision, but largely abolished if it was inconsistent (Figure 3B). In other words, our MEG analysis is consistent with high confidence leading to neural evidence accumulation being "blind" to disconfirmatory evidence. To formally quantify this effect, we entered the slope and starting point of neural evidence accumulation into hierarchical regression models with evidence type (confirmatory, disconfirmatory) and initial confidence (high, low) as predictors. We obtained a significant interaction between confidence and evidence type on slope (β=. 039, $p<10^{-5}$, see Figure 3D), but no effect on starting point (β=. 009, p>.3).

Finally, we asked whether high confidence in a decision leads to qualitatively distinct signatures of post-decisional processing. To address this, we evaluated the extent to which the entire time course of classifier predictions trained in the pre-decision phase generalised to the post-decision phase (King & Dehaene, 2014). Strikingly, we found a cluster of time points in which a representation of the initial decision was activated earlier in the post- compared to the pre-decision phase when confidence was high (p=.012, corrected for multiple comparisons; Figure 3E). Such early reinstatement of a later processing stage is consistent with confidence inducing a preparedness or expectation for evidence supporting the initial decision, indicating that confidence not only increases the magnitude of choice-consistent evidence processing, but also changes the manner in which post-decision evidence is processed.

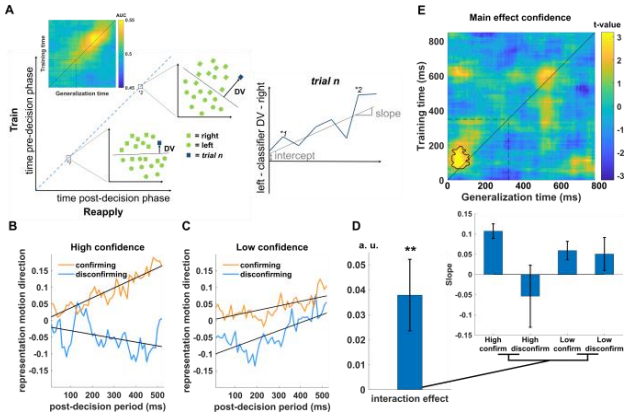

Figure 3: MEG analysis investigating the influence of confidence on processing of post-decision evidence. **A** We trained a machine-learning classification algorithm to predict left/right choices in the pre-decision phase and reapplied the trained classifier to the corresponding time point during the post-decision phase. The distance of each trial to the separating hyperplane provides a graded measure of neural evidence for a left or right decision. Changes in this neural representation within one trial provides a neural metric of evidence accumulation (see right panel). The inset shows the temporal generalization of decoding accuracy from pre- to post-decision phase, indicating that the pre-decision classifier generalizes to the post-decision phase along the major diagonal (i.e. corresponding timepoints). DV = decision variable. **B&C** Change in neural representation in response to post-decision evidence separated into trials in which post-decision evidence confirmed or disconfirmed the initial choice, and as a function of low (**B**) and high (**C**) initial confidence. More positive values on the y-axis indicate better (more veridical) integration of post-decision evidence. Weighted group averages are presented and the regression lines are fits to this averaged data. **D** Interaction effect of initial confidence × choice-consistency predicting the changes in neural representation in response to post-decision evidence (slope). Fixed-effect from hierarchical regression ± S.E.M is presented. Slopes across the four different conditions (low confidence & confirming evidence, low confidence &

disconfirming evidence, high confidence & confirming evidence, high confidence & disconfirming evidence) are shown in the right upward panel. Weighted group averages ± S.E.M are presented. **E** Temporal generalization of decoding accuracy from the pre- to the post-decision phase. Here a modulation of confidence on decoding accuracy is shown, with yellow colors indicating higher decodability of the initial decision (i.e. stronger representation) as a function of confidence. A significant cluster of timepoints with a significant main effect of confidence was found above the main diagonal early in the post-decision phase (enclosed by the solid contour; $p<.05$ corrected for multiple comparisons). Dotted lines indicate the off-set of the stimulus (pre or post-decision stimulus respectively). The time window starts with stimulus presentation (0ms) and ends when the response options are presented (850ms). ***$p<.001$

Conclusion

By combining behavioural and neural modelling, we provide experimental evidence that holding high confidence in a decision leads to a striking modulation of post-decisional processing and the emergence of a behavioural confirmation bias. These findings are consistent with a neural representation of confidence acting as a top-down control mechanism to selectively amplify processing of choice-consistent information.

Acknowledgments

The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). S. Fleming is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (206648/Z/17/Z). TUH is supported by a Wellcome Sir Henry Dale Fellowship (211155/Z/18/Z), a grant from the Jacobs Foundation (2017-1261-04), the Medical Research Foundation, and a 2018 NARSAD Young Investigator grant (27023) from the Brain & Behavior Research Foundation.

References

Fleming, S. M., van der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, 1–8. https://doi.org/10.1038/s41593-018-0104-6

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, *30*, 535–574.

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Current Biology*, *28*(24), 4014–4021.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*, 79.