Title: Confirmation bias is adaptive when coupled with self-awareness

Authors: Max Rollwage[1, 2*], Stephen M. Fleming[1,2,3]

Affiliations:

[1] Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom

[2] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

[3] Department of Experimental Psychology, University College London, London WC1H 0AP, United Kingdom

*Correspondence to: Max Rollwage (max.rollwage.16@ucl.ac.uk)

Keywords

confirmation bias; metacognition; belief updating; adaptive rationality; selective information processing

Abstract

Selective consideration of information is a prominent feature of human behaviour, and recent studies have identified proneness to confirmation bias as a cognitive feature underlying dogmatic beliefs. While such altered information processing typically leads to detrimental performance in laboratory tasks, the ubiquitous nature of confirmation bias makes it unlikely that selective information processing is universally detrimental. Here we suggest that confirmation bias is adaptive to the extent that agents have good metacognition, and are able to seek information when they realise they are wrong. Using simulation-based modelling, we explore how the adaptiveness of holding a confirmation bias depends on such metacognitive insight. We find that the behavioural consequences of selective information processing are highly dependent on agents' introspective abilities. Strikingly, selective information processing can even improve decision-making when compared to unbiased evidence accumulation, as long as it is accompanied by good metacognition. These results further suggest that interventions which boost people's metacognition might be efficient in alleviating the negative effects of selective information processing.

Introduction

Humans often engage in a one-sided consideration of the evidence when making decisions. Independently of whether we have a heated fight with our spouse or a debate with a co-worker, we often only consider our side of the argument. While this might be a useful strategy to win a debate, it can also lead to an inaccurate picture of the existing facts. This selective incorporation of evidence is known as confirmation bias (Nickerson, 1998) and includes both active search for confirming evidence (Klayman & Ha, 1987; Wason, 1960) as well as a reluctance to incorporate disconfirming information (Klayman, 1995; Lord et al., 1979). Confirmation bias has been reported in a variety of settings (Klayman, 1995), including the formation of clinical diagnosis (Groopman, 2008), inference about people's character (Snyder & Swann, 1978), investment decisions (Park et al., 2010), views about societal issues such as capital punishment (Lord et al., 1979) and climate change (Sunstein et al., 2016), and even scientific investigations (Koehler, 1993; Mynatt et al., 1977). Recent cognitive neuroscience studies have identified the selective integration of choice-consistent information as a key mechanism underpinning confirmation bias (Palminteri et al., 2016; Rollwage et al., 2020; Talluri et al., 2018).

Perhaps most prominently, confirmation bias has been commonly reported in relation to politically charged beliefs, such that people are generally prone to process information in line with their political convictions (Kaplan et al., 2016; Nyhan & Reifler, 2010; Redlawsk, 2002; Taber et al., 2009; Taber & Lodge, 2012). On a societal level, such skewed information intake might lead to entrenched beliefs, and, in turn, the prevalence of dogmatic groupings and societal polarization (Zmigrod, 2020; Rollwage et al., 2019). In line with this hypothesis, people who show a resistance to belief updating are also more likely to show extreme political beliefs (Zmigrod, Rentfrow, et al., 2019b), aggression towards opposing political views

(Zmigrod, Rentfrow, et al., 2019a), and authoritarian (Sinclair et al., 2019) or dogmatic traits (Rollwage et al., 2018; Zmigrod, Zmigrod, et al., 2019).

Investigating the computational mechanisms that underpin these associations, we recently showed that dogmatic participants were characterized by two cognitive alterations in the context of a perceptual decision-making task (Rollwage et al., 2018). First, dogmatic participants showed a reduction in metacognitive ability, manifesting as a selective overconfidence after making errors. Second, metacognitive ability (i.e. the accuracy of confidence judgments) was predictive of post-decision evidence integration, where people with poorer metacognition showed less sensitivity for corrective information.

These results indicate that confidence may act as an internal control signal that guides future information processing (Atiya et al., 2019; Desender et al., 2018, 2019; Meyniel, Schlunegger, et al., 2015; Meyniel & Dehaene, 2017; Murphy et al., 2015). In studies using magnetoencephalography (MEG), we found evidence for this hypothesis, with confidence strongly modulating the extent of neural post-decision processing. Evidence accumulation was largely unbiased after low confidence decisions but displayed a confirmation bias after high confidence decisions (Rollwage et al., 2020). In other words, people appear especially resistant to corrective information when they hold high confidence in a wrong decision. However, when confidence is well aligned with performance – when metacognitive ability is high – such selective post-decision integration is likely to be less problematic, as people will tend to be open to new information when they need it (as they will be less confident after making errors). This line of reasoning indicates that people's metacognitive ability might be a crucial driver of the degree to which selective information processing leads to negative behavioural outcomes.

Here we test a hypothesis that selective information integration might be adaptive when coupled with high metacognitive ability. Our proposal is in line with the broader hypothesis

that the ubiquitous nature of selective evidence integration makes it unlikely that this cognitive characteristic is always maladaptive (Gigerenzer, 2008; Klayman & Ha, 1987). For instance, others have interpreted confirmation bias as a heuristic that reduces computational complexity (Evans, 1989) or allows for robustness against noise (Lefebvre et al., 2020; Tsetsos et al., 2016). Here we offer an alternative perspective: that confirmation bias is adaptive to the extent it is tempered by a metacognitive ability to effectively monitor and recognise when we might be wrong (Fleming & Dolan, 2012). We use simulation-based modelling to compare different evidence integration strategies and test their respective performance. Specifically, we compared unbiased evidence integration with a simple confirmation bias as well as with a confidence-driven confirmation bias (as observed in Rollwage et al., 2020). Since we expected the performance of a confidence-weighted confirmation bias to depend on the reliability with which confidence judgments indicate choice accuracy, we also investigated the influence of metacognitive ability on the adaptiveness of confirmation bias.

Modelling behavioural impact of selective evidence integration strategies

We model a simple situation in which agents make a binary decision between two choice options based on noisy information drawn from a world state that is unknown to the agent. This situation is easily modelled with existing frameworks for characterising belief updating (Bronfman et al., 2015; Fleming et al., 2018; Talluri et al., 2018), and closely resembles common perceptual decision-making paradigms. This setting also acts as a minimal framework within which more complex decision problems can be modelled, as exemplified by debates about the existence of climate change. For instance, to reach an opinion about whether human activity causes global warming (the ground truth), we have to form beliefs based on multiple noisy evidence samples (e.g. scientific publications and newspaper articles). Importantly, this process requires the updating of pre-existing beliefs whenever more evidence becomes

available. In such a situation, Bayesian belief updating is often used as benchmark model (Mathys et al., 2011; O'Reilly et al., 2013). Thus, we incorporated a Bayesian updater as our "unbiased agent" against which other evidence integration strategies can be compared. A Bayesian agent keeps track of its graded belief that one or other choice option is correct by calculating the probability of the chosen option, given the evidence, compared to the alternative: $\frac{P(choice|evidence)}{P(alternative|evidence)}$

For simplicity, we simulate a situation in which participants only receive two samples of information: X_{pre} represents the initial information and X_{post} represents the additional (or post-decision) evidence. Both X_{pre} and X_{post} are sampled from normal distributions:

$$X_{pre} \sim N(\mu, \sigma_{pre}^2)$$

$$X_{post} \sim N(\mu, \sigma_{post}^2)$$

$$\mu = [-1, 1]$$

where the mean (μ) of these distributions corresponds to the actual world state that needs to be inferred. While pre- and post-decision evidence distributions have the same mean (i.e. indicate the same underlying world state), they might differ in their variance. The variance represents the reliability of the information, with higher variances indicating less reliable information. We simulate different reliabilities of pre- and post-decision evidence (see Figure 1A) as the consequences of a confirmation bias are likely to depend on this balance.

After receiving initial information, agents make an initial decision ($decision_{initial}$) which depends solely on X_{pre}. This $decision_{initial}$=1, if X_{pre} has a positive value ($X_{pre}>0$), whereas $decision_{initial}$=-1 if X_{pre} has a negative value ($X_{pre}<0$). An estimate of confidence in this initial decision is derived by calculating the log-odds in favour of the chosen world state:

$$LogDir_{pre} = \frac{2 * \mu * X_{pre}}{\sigma_{pre}^2}$$

These log-odds can be transformed into a confidence rating between 0 and 1 as follows:

$$confidence_{initial} = \frac{e^{LogDir_{pre}}}{(1 + e^{LogDir_{pre}})}$$

After the initial decision the agent receives additional information X_{post}. In order to reach a final decision, both evidence samples can be integrated in an unbiased Bayesian fashion by simply summing the log-odds:

$$LogDir_{final} = LogDir_{pre} + LogDir_{post}$$

Note that by adding the log-odds of pre- and post-decision evidence, the certainty/reliability of these two evidence samples is implicitly considered, i.e. evidence is combined in line with Bayesian principles. The final decision depends on the sign of the posterior log-odds ($LogDir_{final}$). If the final decision corresponds to the actual state of the world, the agent can be said to have formed an accurate belief and performs the task correctly. The average decision accuracy over many simulated trials forms our measure of the agent's performance under different decision strategies (we simulated 20,000 trials for each agent and each level of evidence reliability). Depending on the different reliabilities of X_{pre} and X_{post}, an unbiased Bayesian observer will achieve different final performances (see Figure 1A), with more reliable information yielding better performance.

A confirmation bias can be modelled as an altered incorporation of post-decision evidence dependent on whether this new information confirms or disconfirms the initial decision:

If sign(X_{post}) = sign ($decision_{initial}$):

$$LogDir_{final} = LogDir_{pre} + \omega_{confirm} * LogDir_{post}$$

Else if sign(X_{post}) ≠ sign ($decision_{initial}$):

$$LogDir_{final} = LogDir_{pre} + \omega_{disconfirm} * \ LogDir_{post}$$

As observed empirically (Rollwage et al., 2020; Talluri et al., 2018), confirmation bias manifests as an amplification of confirmatory evidence integration ($\omega_{confirm} > 1$) and a reduction in disconfirmatory evidence integration ($\omega_{disconfirm} < 1$). This form of confirmation bias can range from .5 (no bias) to 1 (where maximal confirmation bias is obtained when processing of disconfirmatory information is abolished and $\omega_{disconfirm}$ is 0):

$$\omega_{confirm} = confirmation\ bias * 2$$

$$\omega_{disconfirm} = (1 - confirmation\ bias) * 2$$

ωs vary between 0 and 2, whereby a value of 0 represents an absence of processing and a value of 2 represents maximal processing of post-decision evidence. Values of 1 represent unbiased information processing.

Results

We modelled agents with gradually increasing levels of confirmation bias and investigated their performance relative to an unbiased agent (see Figure 1C). As hypothesized, a selective accumulation of information results in detrimental performance, with higher levels of confirmation bias leading to a more pronounced detriment. This effect was most starkly present when the agent received relatively weak initial information but reliable post-decision evidence (see Figure 1B). In other words, a confirmation bias has its most damaging effects when initial beliefs are already based on weak information, as the bias prevents the incorporation of new and potentially more reliable corrective information.

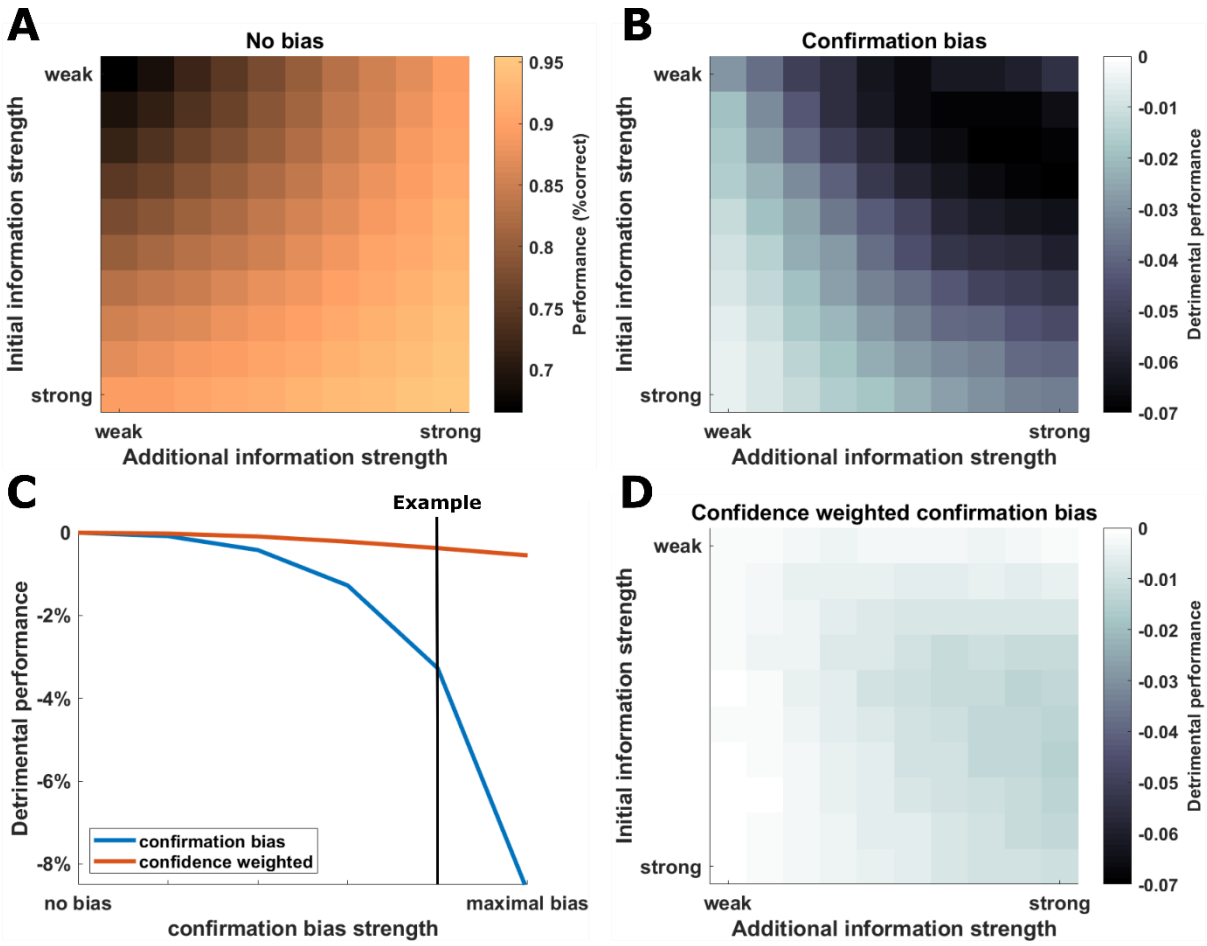

Figure 1 Comparison of agents' performance with different biases in information processing. ***A*** *Performance of an agent that integrates initial and additional information in an unbiased Bayesian manner. Depending on the evidence strength, this agent shows different levels of accuracy, with better performance when both evidence samples are strong/reliable.* ***B*** *Difference in performance between an unbiased agent and a confirmation bias agent as a function of the reliability of the initial and additional evidence. A confirmation bias has especially detrimental effects when initial evidence is relatively weak.* ***C*** *Comparison of agents with a simple confirmation bias and a metacognitive agent against an unbiased agent (detrimental performance of zero would indicate the same performance as an unbiased agent), as a function of confirmation bias strength. Here, performance is averaged over all combinations of initial and additional evidence strengths. The vertical line indicates the strength of confirmation bias used in panels B and D.* ***D*** *Difference in performance between an unbiased agent and a metacognitive agent as a function of the reliability of initial and additional evidence. Overall, the metacognitive agent shows only a relatively small disadvantage in comparison to an unbiased agent. In comparison to a simple confirmation bias, the metacognitive agent suffers less detriment in situations with weak initial evidence.* ***B&D*** *Dark colours indicate more detrimental performance of confirmation bias strategies when compared to unbiased evidence integration.*

In a previous empirical study, we observed a modulation of confirmation bias by confidence (Rollwage et al., 2020) such that participants were relatively unbiased in their use of new evidence when less confident, but showed an enhanced confirmation bias after high confidence decisions. Here we simulate an agent that shows a similar modulation of confirmation bias by

initial confidence which we term a “metacognitive” agent. Such an agent shows a canonical confirmation bias when it is confident in an initial choice (confidence=1), but is unbiased when unsure (confidence=.5):

$$\omega Conf_{confirm} = 1 + (\omega_{confirm} - 1) * \big((confidence - .5) * 2\big)$$

$$\omega Conf_{disconfirm} = 1 + (\omega_{disconfirm} - 1) * \big((confidence - .5) * 2\big)$$

We note that a metacognitive agent differs from a Bayesian agent in that its initial confidence directly modulates the extent to which post-decision evidence is incorporated. For a Bayesian agent there is of course also a sense in which confidence “weights” the subsequent incorporation of evidence, in that a highly confident decision will require more disconfirming evidence to be overturned. Such updates, however, are in keeping with the linear accumulation of the log-odds of one or other hypothesis. In contrast, our metacognitive agent shuts down the processing of disconfirmatory evidence when it is confident, representing a non-linear effect of confidence on the incorporation of post-decisional log-odds. Thus, our metacognitive agent shows similarities with a circular inference model (Bouttier et al., n.d.; Jardri et al., 2017), in which prior beliefs directly alter sensitivity to new evidence.

Interestingly, such a confidence-weighted confirmation bias outperforms a simple confirmation bias in all settings and only shows slight impairments in relation to an unbiased agent (see Figure 1C). When we compared a confidence-weighted confirmation bias to a simple confirmation bias, we found that this metacognitive agent clearly outperforms a simple confirmation bias in situations when the initial evidence is weak (see Figure 1D). While a simple confirmation bias has the strongest decrement in performance in these situations, a metacognitive agent “realizes” it is dealing with weak initial evidence (by assigning lower confidence to these decisions) and thus shows a more equal sensitivity to confirming and disconfirming information.

Up until now we have assumed that agents calculate confidence in an initial choice by directly evaluating the reliability of evidence that informed the decision. In this case, the agent uses exactly the same information for both their decision and confidence rating. However, empirical studies have shown that confidence judgments and decisions may rely on (partially) separate information (Cortese et al., 2016; Maniscalco et al., 2016; Miyoshi & Lau, 2020; Odegaard et al., 2018; Peters et al., 2017; Zylberberg et al., 2012). The type (and quality) of evidence that informs confidence judgments in turn determines the degree to which confidence is a good indicator of belief accuracy (Maniscalco & Lau, 2012). This correspondence between confidence and performance can be formally quantified as the ratio of *meta-d'/d'* (known as metacognitive efficiency), where *meta-d'* reflects metacognitive sensitivity and *d'* reflects primary task performance within a signal detection theoretic framework. Several reasons for a dissociation between *meta-d'* and *d'* have been suggested. For instance, confidence may reflect a noisy read-out of the decision evidence or a decline of decision evidence in working memory prior to a confidence judgment (Maniscalco & Lau, 2012), leading to a *meta-d'/d'* ratio of less than 1. On the other hand, confidence might be informed by evidence that was not available at the time of decision (Moran et al., 2015; Moreira et al., 2018; Pleskac & Busemeyer, 2010), or on correlated evidence that is accumulated in parallel (Fleming & Daw, 2017), both of which may lead *meta-d'/d'* ratios to surpass 1.

Here, we assessed the degree to which the level of metacognitive efficiency influences the performance of our metacognitive agent. To do so, we relaxed our assumption that confidence is directly derived from the evidence informing the initial decision. Instead, the evidence informing decisions (X_{pre}) and confidence estimates (X_{conf}) were distinct but correlated, and we allowed the reliability of the confidence (σ^2_{conf}) and decision (σ^2_{pre}) samples to differ. Hence, X_{pre} and X_{conf} were sampled from a bivariate normal distribution with the same mean (μ) and the following covariance:

$$\begin{bmatrix} X_{pre} \\ X_{conf} \end{bmatrix} \sim N(\mu, \Sigma)$$

$$\Sigma = \begin{bmatrix} \sigma_{pre}^2 & \rho * \sigma_{pre} * \sigma_{conf} \\ \rho * \sigma_{pre} * \sigma_{conf} & \sigma_{conf}^2 \end{bmatrix}$$

Σ describes the covariance between X_{pre} and X_{conf}, whereby ρ represents the correlation between these variables. As suggested by Fleming & Daw (2017), in this situation confidence can be inferred based on a combination of X_{conf}, the initial decision, and the covariance of X_{conf} and X_{pre} (see the Appendix of Fleming & Daw for further details on this calculation):

$$Confidence = P(decision_{initial} = worldstate | X_{conf}, decision_{initial}, \Sigma)$$

Importantly, modelling separate samples of X_{conf} and X_{pre} allows for dissociations between *meta-d*' and *d'*, thus making it possible to simulate varying degrees of metacognitive efficiency.

A decrease in the reliability of the evidence informing the confidence rating ($\sigma_{conf}^2 > \sigma_{pre}^2$) naturally reduces metacognitive efficiency and leads confidence judgments to be less reliable predictors of choice accuracy, when keeping ρ constant (for this condition we fixed ρ=.8; however, we note that our findings are not dependent on the specific value of ρ, see Figure S1). Our simulation shows that reduced metacognitive efficiency leads a metacognitive agent to show impaired performance (see left side of Figure 2A), as the weighting of new information by confidence is less effective. We also modelled a situation in which confidence and decision information have the same reliability ($\sigma_{pre}^2 = \sigma_{conf}^2$), but the evidence samples have a variable degree of correlation (ρ=[.1-.8]). Such a setting can result in increased metacognitive efficiency, as the "confidence rater" has additional information on which to base an evaluation of the decider (Fleming & Daw, 2017). Surprisingly, when metacognitive efficiency is high (*meta-d'/d'*>1) we found that a metacognitive agent can even outperform an unbiased agent (see Figure 2A right side). This result is striking as it suggests that a selective integration of

information is not necessarily a “bias” but represents an advantageous strategy for achieving optimal performance in the context of a realistic cognitive architecture (i.e. one in which metacognition is particularly efficient).

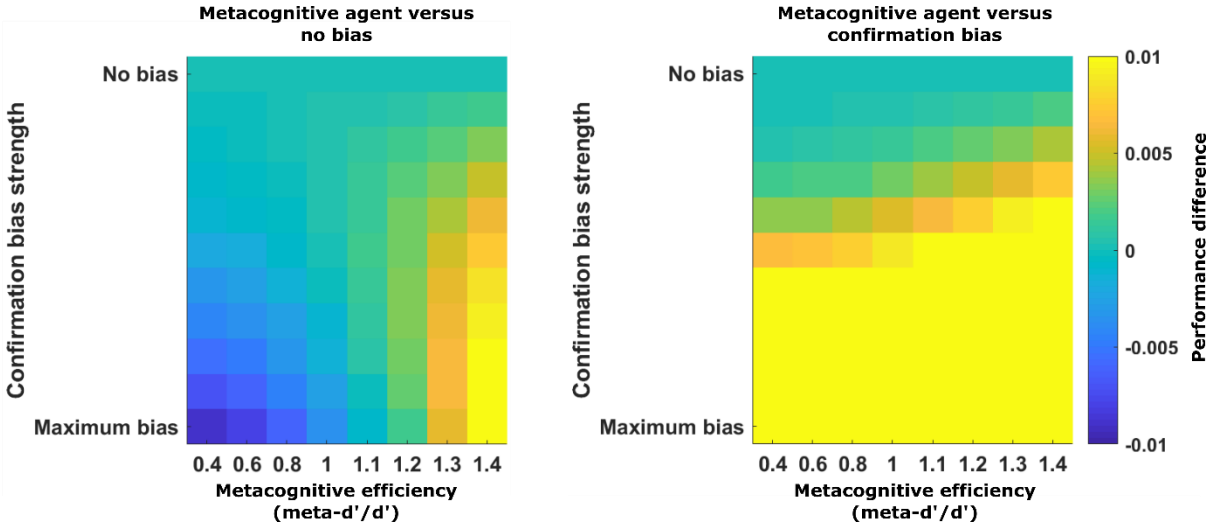

Figure 2 Performance of a metacognitive agent as function of metacognitive ability. ***A*** *Performance difference of a metacognitive agent compared to an unbiased agent. The performance of a confidence weighted confirmation bias is sensitive to metacognitive efficiency (i.e. the accuracy of confidence ratings), with the greatest benefits obtained when metacognitive ability is high. When the ratio of meta-d’/d’ is above 1, a metacognitive agent outperforms an unbiased agent.* ***B*** *Performance difference of a metacognitive agent compared to a simple confirmation bias. The metacognitive agent shows superior performance in all settings. Even in situations with low metacognitive ability, the metacognitive agent outperforms a simple confirmation bias.* ***A&B*** *Hotter colours indicate better performance of the metacognitive agent compared to the other information processing strategies.*

Discussion

Here we investigated the effects of confirmation bias on the accuracy of belief formation. Our central proposal is that when confirmation bias is a feature of a self-aware (metacognitive) agent, it ceases to be detrimental, and may even become adaptive. We used simulation-based modelling to compare the performance of agents with different forms of confirmation bias against an unbiased agent. In general, a simple (non-metacognitive) confirmation bias showed detrimental effects compared to an unbiased agent in all settings. In comparison, a metacognitive agent which modulates the extent of a confirmation bias by its confidence (as

shown empirically in human observers; Rollwage et al., 2020) outperformed a simple confirmation bias agent, and was in many cases not substantially worse than an unbiased agent. In turn, by simulating varying degrees of metacognitive efficiency, we found that the performance of our metacognitive agent was highly sensitive to its level of self-awareness. Strikingly, a metacognitive agent with high self-awareness could in some cases even outperform an unbiased agent, indicating that selective information processing might be particularly adaptive when coupled with good metacognitive abilities.

Why should a metacognitive agent have an advantage in this case? The core mechanism appears to be the capacity of a confidence estimate to provide a "second look" on a decision, similarly to how an external adviser might give us a separate view on a topic. The adaptivity of this information processing mechanism crucially depends on the agent's metacognitive ability – as agents with good metacognition provide the most effective "internal" advisory signals. Interestingly, however, a confidence-weighted confirmation bias outperformed a simple confirmation bias in all settings, even when metacognitive ability was relatively low (see Figure 2B). Conversely, when metacognition is high, this evidence-weighting strategy can even outperform an unbiased agent. These results are in accordance with a view that cognitive biases may have originally evolved for good evolutionary reasons and are often adaptive when considered in the context of the agent's environment, including its broader mental toolkit (Gigerenzer, 2008).

By incorporating confirmation bias as part of a broader cognitive architecture in which different mental processes can interact with each other (e.g. decisional and metacognitive processes), selective information processing may become adaptive when compared to the same "bias" considered in isolation. In the same spirit, it has been argued that heuristics that may appear as biases in simple and constrained laboratory tasks become beneficial in more complex environments (Pleskac & Hertwig, 2014). More broadly, our study indicates that considering

cognitive biases in isolation from other mental processes might lead to the wrong conclusions about the impact of a particular cognitive feature on behaviour.

We note that a clear behavioural benefit for confirmation bias was only present when simulating agents with "hyper" metacognitive efficiency (i.e. *meta-d'/d*>1), such that metacognition becomes more acute than first-order task performance. This might seem odd at first glance, as it implies that the system is not using all the information available to it at the time of making an initial choice, and only afterwards becomes more sensitive to whether it was right or wrong. However, this kind of pattern is commonly observed in empirical data (Fleming & Daw, 2017; Moreira et al., 2018), and is thought to be driven either by additional post-decisional processing, differences in the variance of signal compared to noise (Miyoshi & Lau, 2020) or (as simulated here) a parallel processing of information that allows the system to detect and correct its errors (Moreira et al., 2018). The capacity for rapid error detection is well established in human studies (Murphy et al., 2015; Rabbitt, 1966; Resulaj et al., 2009; Van Den Berg et al., 2016; Yeung & Summerfield, 2012) and thus it is reasonable to assume that hyper metacognitive efficiency may be common in the healthy population.

Importantly, the confidence system does not need to be more reliable than the decision system to achieve high metacognitive ability: it needs only to rely on partially separate information (as used here; see Figure S1 for simulations using a wider set of parameters). In this respect, our results also contribute to a long-running debate over why it might be useful for the brain to encode a separate confidence signal that is distinguishable from decision evidence (Maniscalco et al., 2016; Meyniel, Sigman, et al., 2015). A metacognitive agent that realizes its own mistake (and assigns low confidence to this decision) will then tend to become more open to new information, due to the confidence weighting applied to selective information processing (and in fact, if confidence falls below 0.5 in a two-choice scenario we would predict that

metacognitive agents would show a “disconfirmation” bias). Our results suggest that this benefit is only seen when metacognition is partly independent of first-order cognition.

Selective information processing has been assumed to lead to skewed, entrenched and potentially inaccurate beliefs about a range of societal and political issues (Rollwage et al., 2019; Zmigrod, 2020). However, the current results suggest that the detrimental effects of selective information processing depend on people’s broader self-awareness. In turn, metacognitive deficits might represent core drivers of polarised or radical beliefs, due to their consequence for maladaptive confirmation bias. Interestingly, this hypothesis is in line with empirical observations (Rollwage et al., 2018), showing that more dogmatic participants show reduced metacognitive sensitivity which in turn is predictive of reduced post-decision evidence processing. Recognising metacognition as a central driver of belief polarisation may make it possible to develop new strategies for debiasing decision-making (Rollwage et al., 2019). Excitingly, there are existing interventions that have been shown to boost people’s metacognitive ability (Baird et al., 2014; Carpenter et al., 2019). Our results indicate that cognitive training which improves domain-general self-awareness and metacognitive efficiency may help to alleviate the negative behavioural outcomes of selective information processing, and foster resilience against misinformation and belief polarisation.

Funding

M.R. is a predoctoral Fellow of the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research. The participating institutions are the Max Planck Institute for Human Development and University College London (UCL). The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). S.M.F. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (206648/Z/17/Z).

Author Contributions

M.R. and S.M.F. conceived the study and wrote the paper. M.R. developed the methodology and ran the simulations with input from S.M.F.

References

Atiya, N. A. A., Rañó, I., Prasad, G., & Wong-Lin, K. (2019). A neural circuit model of decision uncertainty and change-of-mind. *Nature Communications*, *10*(1), 1–12.

Baird, B., Mrazek, M. D., Phillips, D. T., & Schooler, J. W. (2014). Domain-specific enhancement of metacognitive ability following meditation training. *Journal of Experimental Psychology: General*, *143*(5), 1972.

Bouttier, V., Deneve, S., & Jardri, R. (2019). Confirmation Bias is explained by Descending Loops in the Cortical Hierarchy. *Conference on Cognitive Computational Neuroscience, 13-16 September 2019, Berlin, Germany* https://doi.org/10.32470/CCN.2019.1433-0

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1810), 20150228. https://doi.org/10.1098/rspb.2015.0228

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51.

Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications*, *7*(1), 1–18.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*. https://doi.org/10.1177/0956797617744771

Desender, K., Murphy, P., Boldt, A., Verguts, T., & Yeung, N. (2019). A postdecisional neural marker of confidence predicts Information-Seeking in Decision-Making. *Journal*

of Neuroscience, *39*(17), 3309–3319.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Phil. Trans. R. Soc. B*, *367*(1594), 1338–1349.

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

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29.

Groopman, J. (2008). *How doctors think*. Houghton Mifflin Harcourt.

Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, *8*, 14218.

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*(December), 1–11. https://doi.org/10.1038/srep39589

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in

hypothesis testing. *Psychological Review*, *94*(2), 211.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*(1), 28–55.

Lefebvre, G., Summerfield, C., & Bogacz, R. (2020). A normative account of confirmatory biases during reinforcement learning. *BioRxiv*.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*(3), 923–937.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39.

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, *114*(19), E3859–E3868.

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Computational Biology*, *11*(6), 1–25. https://doi.org/10.1371/journal.pcbi.1004305

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

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review*.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002

Moreira, C. M., Rollwage, M., Kaduk, K., Wilke, M., & Kagan, I. (2018). Post-decision wagering after perceptual judgments reveals bi-directional certainty readouts. *Cognition*, *176*, 40–52.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *ELife*, *4*, 1–23. https://doi.org/10.7554/eLife.11946

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, *29*(1), 85–95.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

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

O'Reilly, J. X., Schuffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M.

F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38), E3660–E3669. https://doi.org/10.1073/pnas.1305373110

Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A. K., Lau, H., & Basso, M. A. (2018). Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences*, *115*(7), E1588–E1597.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. (2016). Confirmation bias in human reinforcement learning. *BioRXiv*, 1–21. https://doi.org/http://dx.doi.org/10.1101/090654

Park, J., Konana, P., Gu, B., Kumar, A., & Raghunathan, R. (2010). Confirmation bias, overconfidence, and investment performance: Evidence from stock message boards. *McCombs Research Paper Series No. IROM-07-10*.

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., & Halgren, E. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 139.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*(5), 2000.

Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264.

Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated

reasoning on political decision making. In *Journal of Politics* (Vol. 64, Issue 4, pp. 1021–1044). https://doi.org/10.1111/1468-2508.00161

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266. https://doi.org/10.1038/nature08275

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.

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), 1–11.

Rollwage, M., Zmigrod, L., de-Wit, L., Dolan, R. J., & Fleming, S. M. (2019). What underlies political polarization? A manifesto for computational political psychology. *Trends in Cognitive Sciences*, *23*(10), 820–822.

Sinclair, A. H., Stanley, M., & Seli, P. (2019). Closed-Minded Cognition: Right-Wing Authoritarianism is Negatively Related to Belief Updating. *psyArXiv*

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*(11), 1202.

Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., & Sharot, T. (2016). How people update beliefs about climate change: Good news and bad news. *Cornell L. Rev.*, *102*, 1431.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*(2), 137–155. https://doi.org/10.1007/s11109-008-9075-8

Taber, C. S., & Lodge, M. (2012). Motivated skepticism in the evaluation of political beliefs (2006). *Critical Review*, *24*(2), 157–184. https://doi.org/10.1080/08913811.2012.711019

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, *28*(19), 3128–3135.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 3102–3107.

Van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *ELife*, *5*, 1–21. https://doi.org/10.7554/eLife.12192

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Phil. Trans. R. Soc. B*, *367*(1594), 1310–1321.

Zmigrod, L. (2020). The role of cognitive rigidity in political ideologies: theory, evidence, and future directions. *Current Opinion in Behavioral Sciences*, *34*, 34–39.

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2019a). Cognitive inflexibility predicts extremist attitudes. *Frontiers in Psychology*, *10*, 989.

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2019b). The partisan mind: Is extreme political partisanship related to cognitive inflexibility? *Journal of Experimental Psychology: General*.

Zmigrod, L., Zmigrod, S., Rentfrow, P. J., & Robbins, T. W. (2019). The psychological roots of intellectual humility: the role of intelligence and cognitive flexibility. *Personality and Individual Differences*, *141*, 200–208.

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

Supplementary material

In the main text we focussed on the influence of metacognitive efficiency on the performance of a confidence-weighted confirmation bias, showing that this information processing strategy can outperform an unbiased agent when metacognitive efficiency is high. This indicates that the performance of a metacognitive agent scales according to its self-awareness.

However, when sampling X_{pre} and X_{conf} from separate but correlated distributions, metacognitive efficiency can be influenced by either ρ, σ_{conf} or both. Metacognitive abilities will increase when the correlation (ρ) between X_{pre} and X_{conf} is reduced, as well as when X_{conf} is more reliable, i.e. σ_{conf} is small (Fleming & Daw, 2017; see Figure S1A). Importantly, independently of how these boosts in metacognitive efficiency are achieved, the results we obtain on the benefits for post-decision processing are maintained (Figure S1B). In other words, the performance of a metacognitive agent compared to an unbiased agent is determined by its metacognitive efficiency, but not particular parameters of the confidence model used here. This inference is supported by a very high correlation between variation in metacognitive efficiency and post-decision integration performance as a function of ρ and σ_{conf} (r=.97, Figure S1C).

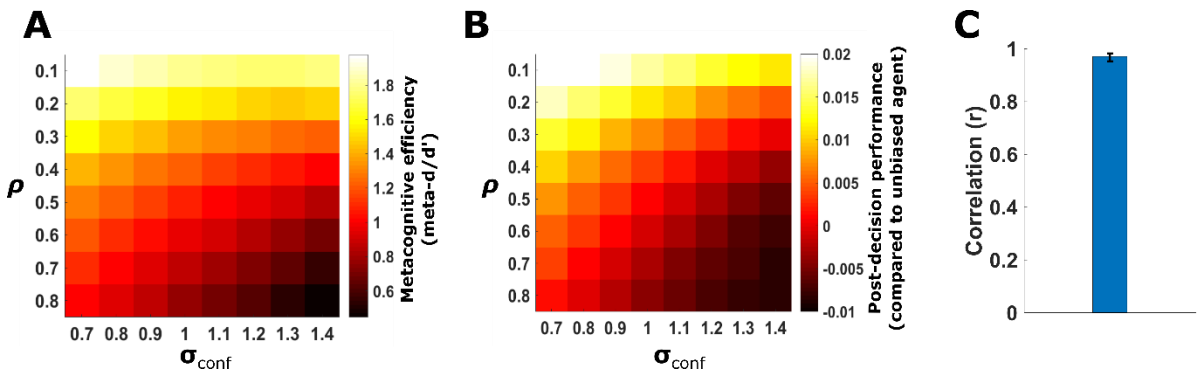

Figure S1. Relationship between metacognitive efficiency and post-decision integration performance. ***A*** *The agent's metacognitive efficiency (meta-d'/d') is determined by the settings of σ_{conf} and ρ. Lower values of σ_{conf} indicate that confidence relies on more reliable evidence samples, whereby lower levels of ρ mean that the evidence for the decision and confidence are only mildly correlated, i.e. both provide independent information.* ***B*** *Performance difference of a metacognitive agent compared to an unbiased agent, as a function of σ_{conf} and ρ. Interestingly, these parameters have the same influence on post-decision evidence integration as they do on metacognitive efficiency (compared panels A and B). This indicates that, independently of how a specific metacognitive ability is achieved, it has the same influence on post-decision evidence integration.* ***C*** *Support for a link between metacognition and post-*

decision processing benefit is obtained by computing the correlation between these statistics across all values of σ_{conf} and ρ, i.e. the correlation between panels A and B (± 95% confidence interval).