Synthese
https://doi.org/10.1007/s11229-020-02910-x

Confirmation bias without rhyme or reason

Matthias Michel[1,2] · **Megan A. K. Peters**[3]

Received: 3 January 2020 / Accepted: 10 October 2020

Abstract
Having a confirmation bias sometimes leads us to hold inaccurate beliefs. So, the puzzle goes: why do we have it? According to the influential argumentative theory of reasoning, confirmation bias emerges because the primary function of reason is not to form accurate beliefs, but to convince others that we're right. A crucial prediction of the theory, then, is that confirmation bias should be found only in the reasoning domain. In this article, we argue that there is evidence that confirmation bias does exist outside the reasoning domain. This undermines the main evidential basis for the argumentative theory of reasoning. In presenting the relevant evidence, we explore why having such confirmation bias may not be maladaptive.

Keywords Confirmation bias · Myside bias · Cognition · Perception · Confidence

Across a wide variety of experiments, subjects exhibit a *confirmation bias*, "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson 1998, p. 175). This bias is widespread, even among those who are (supposed to be) searching for objective truths such as judges, scientists, and physicians (Nickerson 1998). At a social level, confirmation bias could lead to the aggregation of like-minded individuals and group polarization, thus altering public debates (Sunstein 2002; Myers 1982).

As ubiquitous as it is, that we have a confirmation bias is somewhat bewildering. To survive (and, all things considered, why not?), it is a good thing to acquire accurate beliefs and avoid holding inaccurate ones. But having a confirmation bias often leads

✉ Matthias Michel
matthias.michel.curtil@gmail.com

1 Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, UK

2 Consciousness, Cognition and Computation Group, Université Libre de Bruxelles, Brussels, Belgium

3 Department of Bioengineering, Interdepartmental Graduate Program in Neuroscience, Department of Psychology, University of California Riverside, Riverside, USA

Published online: 22 October 2020

us to hold inaccurate beliefs, with potentially disastrous consequences down the road. So, here's a puzzle: why do we have a confirmation bias?

One of the most influential responses to this challenge has been put forward by proponents of the so-called "argumentative theory of reasoning" (Mercier and Sperber 2011, 2017; Mercier 2016; for other accounts, see e.g., Evans 1989; Kunda 1990; Stanovich 2004). In this article, our aim is to critically evaluate a central prediction of this theory.

The argumentative theory of reasoning makes two central claims (Mercier and Sperber 2011; Mercier 2016). The first is about the *nature* of reasoning. As Mercier (2016) notes:

> The vast bulk of cognition performs perceptual, motoric, and inferential functions without any attention being paid to reasons. By contrast, reasoning *evaluates* reasons … and reasoning *produces* reasons, whether in solitary ratiocination or in argumentation. (p. 690, our emphasis)

Reasoning consists in *evaluating* and *producing* reasons. For the purpose of this article, we grant that this characterization of the nature of reasoning is correct. It follows that a cognitive process that is not engaged in either the evaluation or production of reasons is not a *reasoning* process.

The second, central claim of the theory is about the *function* of reasoning. Here, the main tenet of the argumentative theory is that the function of reasoning is not primarily to form accurate beliefs, but to produce arguments to convince others (Mercier and Sperber 2017). Reasoning is for argumentation. And argumentation, in turn, is essential for efficient communication and cooperation, as it allows to transmit messages without having to rely entirely on trust (Mercier 2016, p. 690; Sperber et al. 2010).

Mercier and Sperber (2017) hold that if argumentation is the function of reasoning, we should expect it to be biased. In particular, we should expect reasoning to systematically work to "find reasons for our ideas and against ideas we oppose" (Mercier and Sperber 2017; p. 218). They write:

> When we produce reasons, we should be heavily biased towards our point of view. We're not going to appear more rational by providing reasons why what we did was stupid; we're not going to convince someone by giving them arguments for their point of view or against ours. This explains an otherwise puzzling feature of reason: the myside bias. (p. 73)

Being biased leads one to find more arguments in favor of one's view, as well as to counter-argue more effectively. This obviously has some benefits when it comes to convincing others that one is right.

The tendency to look for reasons in favor of our own ideas, and against ideas we oppose, is what Mercier and Sperber call *myside bias*. They avoid the term "confirmation bias", as it could convey the idea that subjects have a tendency to confirm *any* view they happen to entertain. As Sperber and Mercier (2017) write: "What they find difficult is not looking for counterevidence or counterarguments in general, but only when what is being challenged is their own opinion" (p. 218). With this caveat in mind, in this article, we use 'confirmation bias' and 'myside bias' interchangeably. Accordingly, confirmation bias—or myside bias—is the tendency to look for arguments and

evidence in favor of one's own beliefs, and to neglect looking for arguments and evidence against one's own beliefs. As defined by Nickerson (1998), it is the "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (p. 175).

A reasoning-based confirmation bias of this kind has been observed in multiple studies (Nickerson 1998; Mercier 2016), and seems independent from general factors, like intelligence (Stanovich and West 2007). For instance, in the Wason selection task (Wason 1968), once participants make an intuitive—and often wrong—judgment, they spend most of their time thinking about the cards they have selected, and do not attempt to disconfirm their intuitive decision (Ball et al. 2003; Lucas and Ball 2005). As another example, in a study by Hall et al. (2012), participants had to answer a questionnaire about how strongly they agreed with a set of moral principles. Following a sleight of hand performed by the experimenter, some of the participants' answers were inverted, such that their answers now indicated disagreement with statements they previously agreed with, and vice versa. Not only did 69% of the participants fail to detect at least one of the changes, but they also subsequently had a tendency to justify views they thought they expressed—when in fact they reported agreeing with the opposite views. In this case, merely making participants believe they held certain views led them to look for arguments supporting views they previously reported disagreeing with (For other cases, See Nickerson 1998; Mercier 2016).

The argumentative theory of reasoning accounts for cases like these by holding that the evolutionary *function* of reasoning is to convince others. If that's the case, it is no surprise that subjects mainly look for arguments in favor of their own views: confirmation bias is a feature of reasoning, not a bug.

One of the most central predictions of the theory naturally follows, in Mercier and Sperber's words: "If it were shown that the myside bias [or confirmation bias] is widespread in our cognitive system, instead of being restricted to reason as we claim, our hypotheses would be weakened." (2019, p. 151). According to this theory, confirmation bias emerges as a straightforward consequence of the function of reasoning, namely, to evaluate and produce arguments effectively. For this reason, Mercier (2017) writes that "from the perspective of the argumentative theory of reasoning, the myside bias should be specific to reasoning—as it seems to be" (p. 111), and that

> there is no evidence of a confirmation or myside bias in cognitive mechanisms besides reasoning. This should not be surprising given that such a bias would be widely maladaptive. An animal eager to confirm mistaken beliefs – that there are no predators around, for instance – would not survive very long. (p. 109)

If, instead, confirmation bias were found outside the reasoning domain, the theory could not claim to have the advantage of providing a parsimonious explanation of confirmation bias: the argumentative theory does not explain how confirmation bias could have evolved outside the reasoning domain. Parsimony would dictate that we should prefer a theory with a single evolutionary story for confirmation bias both within and outside the reasoning domain.

Second, Mercier and Sperber (2011) take the very existence of confirmation bias in the reasoning domain as evidence that reasoning evolved primarily for social, argumentative purposes. As such, finding confirmation bias outside the reasoning domain

would additionally suggest that the evolution of confirmation bias occurred relatively independently from the evolution of reasoning. This independence in turn would mean that one cannot interpret the existence of a confirmation bias as evidence that argumentation is the function of reasoning. That is, the existence of a confirmation bias, in and of itself, would not tell us much about the evolutionary function of reasoning.

In this article, we argue that there is ample evidence of confirmation bias outside the reasoning domain. We present the relevant evidence, and then explain why having such confirmation bias may not be maladaptive. Finally, we answer potential objections. To end on a positive note, we emphasize that the presence of confirmation bias outside the reasoning domain could pave the way for significant experimental progress for research on confirmation bias in general.

1 Confirmation bias outside the reasoning domain

The best evidence against the argumentative theory would come from evidence that confirmation bias exists (a) in at least one non-reasoning domain, and (b) in animals other than humans. We argue that both of these possibilities are well established in the literature.

2 Confirmation bias in perception

You are driving down a foggy road at night. There is a vague shape up ahead in the distance. Having decided that the object is a deer, and not a car or tree, your visual system is then tasked with judging other properties of the object: its trajectory, speed, and so on. These types of subsequent perceptual decisions have been shown to exhibit confirmation bias. The perceived speed of the object is *biased*, or conditioned on the decision that the object is a deer, and not a car.

One early study by Stocker and Simoncelli (2008) demonstrated this effect using simple dot-motion perception in humans, following an experiment by Jazayeri and Movshon (2007; see also Zamboni et al. 2016). Observers viewed random dot motion kinematograms (RDKs) on a computer screen. Some percentage of the dots moved coherently in one direction, either clockwise (up and to the right) of a marker, or counterclockwise (up and to the left) of the marker (Fig. 1a), while the other dots moved randomly. On 30% of the trials, after indicating their decision about whether the net motion of the dots was clockwise or counterclockwise of the marker, observers estimated the actual angle of the perceived motion flow (i.e., how much clockwise or counterclockwise it appeared, relative to the marker; Fig. 1b). The results indicate that motion estimates are biased *away* from the reference point: the angle of the motion that participants report perceiving had a larger magnitude than the actual motion direction angle.

One explanation of this effect is that the subjects' estimates of motion direction depend on their prior perceptual decision, namely, whether they decided that the dots are moving clockwise or counterclockwise of the marker. Having decided that the dots are moving clockwise of the marker, the perceptual system evaluates the *magnitude*

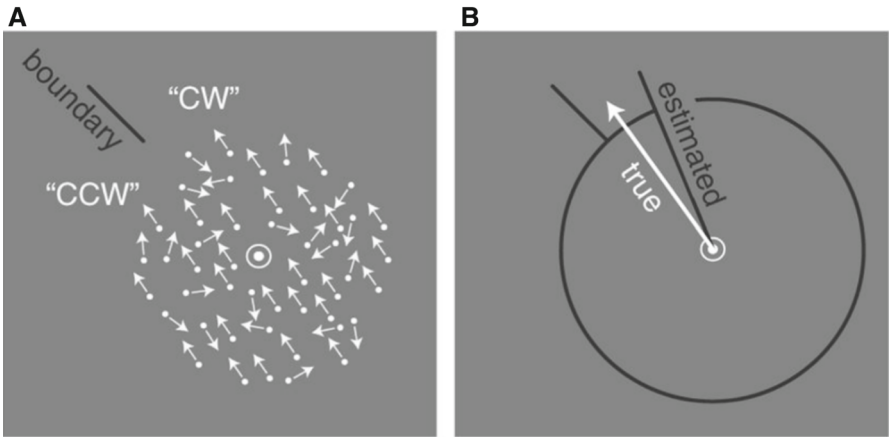

Fig. 1 Dot-motion discrimination and motion estimation task used by Jazayeri and Movshon (2007) and referenced by Stocker and Simoncelli (2008). **a** Subjects see random dot kinematograms of semi-coherent motion moving up and to the right (clockwise; CW) or up and to the left (counterclockwise; CCW) of a boundary marker, and judge whether the primary motion direction is CW or CCW. **b** On some trials, after the discrimination decision, subjects estimate the direction of motion flow in a continuous fashion. Figure reproduced with permission from Jazayeri and Movshon (2007)

of this motion by evaluating the motion of the dots that conform with the decision, while ignoring any motion inconsistent with the decision (but see Fritsche and de Lange 2019, for an alternative view). Consistent with this interpretation, the results were most successfully modeled with a Bayesian observer which essentially resets the probability of the *un*chosen motion category to zero after the choice.[1] That is, observers begin by making a motion decision (clockwise or counterclockwise), and then infer the exact direction of this motion *conditioned* on their initial decision. As a result, the estimates are pushed away from the decision boundary and observers display a confirmation bias: they overweight motion information supporting their categorical choice.

This result was subsequently followed by more recent work on conditional perceptual confirmation bias, and has been replicated in several perceptual tasks, such as line orientation estimation (Luu and Stocker 2018), estimation of the orientation of gabor patches (Fritsche and de Lange 2019) or estimation of stimulus brightness (Brezis et al. 2015). In these studies, participants systematically overweight evidence in favor of their perceptual decisions when estimating the magnitude of some perceptual features. That is, the perceptual systems in charge of evaluating the magnitude of perceptual features seem to look primarily for information that conforms with the initial categorical decisions.

It has been recently suggested that this confirmation bias in perception may have an attentional source. Using a similar task to that used by Jazayeri and Movshon (2007), Talluri et al. (2018) developed an accumulating evidence neural network computational

[1] While observers are optimal on the discrimination task, they are suboptimal on the subsequent motion estimation task. In that sense, it is justified to talk of a *bias* for this latter task, because the participants' behaviors systematically derive from optimality.

model positing that additional signal gain is applied to stimulus features that are consistent with a given perceptual choice; this was recently followed by work from Urai et al. (2019). In essence, feedback from later-stage cognitive areas directs attention to, and thus boosts signal gain on, elements of the physical stimulus that support or bolster the interpretation the system has selected (see also Bronfman et al. 2015). Adaptation-like confirmation biases in sequential perceptual decisions have also been reported (Abrahamyan et al. 2016).

Although the neural mechanisms of this confirmation bias in perceptual decision making remain an area of active investigation, mounting evidence suggests that these low-level, automatic perceptual decisions are susceptible to confirmation bias just as observed in reasoning (Tversky and Kahneman 1974; Festinger 1957; Nickerson 1998). The perceptual system seems to primarily look for sensory information that conforms with its prior categorical decisions. The presence of this bias outside the reasoning domain strongly suggests that the bias may stem from a core information processing source in the brain rather than being unique to the argumentative aspects of rational high-level cognition.

3 Confirmation bias in perceptual confidence

Perceptual confidence refers to the subjective sense of certainty we feel when we are making perceptual decisions, and it also exhibits a well-studied confirmation bias. For example, when asked to report how *sure* you feel in your decision that the object on that foggy, nighttime road is a deer, car, or tree, you overly rely on evidence supporting your perceptual decision, while down-weighting evidence for other possible interpretations. Critically, confirmation bias in perceptual confidence has been observed not only in humans, but in other non-human animals as well.

Over the past decade and a half, numerous studies have reported evidence for confirmation bias in perceptual confidence. Zylberberg et al. (2012) first described that in rating their confidence in dot-motion direction discrimination decisions, observers tend to over-weight evidence supporting the decision that was just made. Subsequently, this observation has been exploited in a number of studies to experimentally dissociate decision accuracy from decision confidence by changing the contrast or other elements of the visual perceptual stimuli.

In these studies, experimenters typically create two experimental conditions designed to induce the same level of accuracy, but where one condition ought to lead to higher confidence than the other. In one condition, subjects have to discriminate the identity of "high energy" stimuli, for example, high contrast or high motion coherence stimuli. In the other condition, the same task is performed with "low energy" stimuli, for instance, low contrast or low motion coherence stimuli. Critically, in both conditions, the stimuli are matched in the signal-to-noise ratio (SNR) of the internal response. That is, experimenters adjust the stimuli such that the discrimination task is just as difficult for the subjects in both conditions. As a result, observers have the same discrimination performance in the "high energy" and "low energy" stimuli conditions (measured as task accuracy or via the signal detection theoretic metric d') (Macmillan and Creelman 2005).

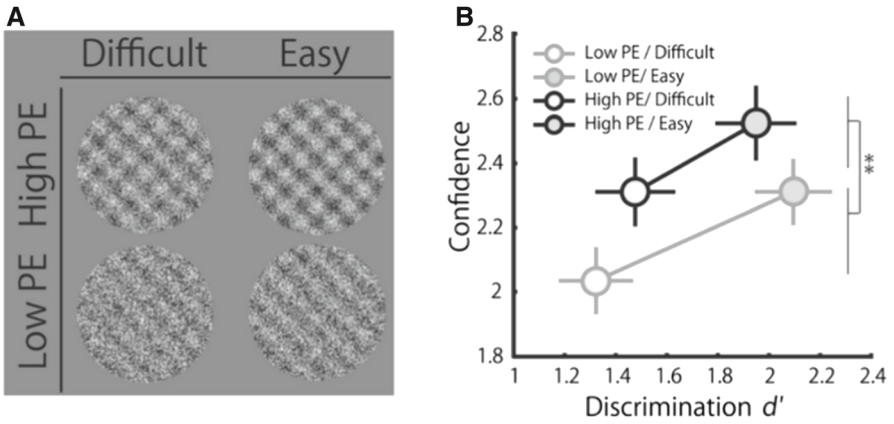

Fig. 2 Stimuli and results used by Koizumi et al. (2015). **a** On each trial, subjects viewed a single stimulus composed of two superimposed left- or right-tilted gratings, with one grating having higher contrast ("energy" or "evidence") than the other. Four stimulus variants were used: high positive evidence ("high energy"/High PE/high contrast) or low positive evidence ("low energy"/Low PE/high contrast), with difficult (the two tilted gratings had very different contrasts) and easy (the two tilted gratings had similar contrasts) versions of each. The subjects' task was to indicate on each trial which of the two superimposed tilted gratings (i.e., the one tilted left or right of vertical) had higher contrast. **b** Subjects performed better in the easy condition than the difficult condition, but there was no effect of energy level (contrast level) on performance. Despite no significant differences in task performance as a function of energy level, subjects rated significantly higher confidence in the "high energy" (High PE) condition. Figure reproduced with permission from Koizumi et al. (2015)

For instance, in an experiment by Koizumi et al. (2015), subjects saw stimuli consisting of two superimposed sinusoidal gratings, one of which had higher contrast, tilted 45° left and right of vertical (Fig. 2a). On each trial, subjects had to determine the orientation of the grating with higher contrast, and then rate their confidence in their decision. The stimuli could take on one of four possible conditions in a 2x2 design. In the easy conditions, stimuli had high SNR as they were constituted of two superimposed gratings with very dissimilar contrasts. In the difficult conditions, stimuli had low SNR because they were constituted by two superimposed gratings with similar contrasts. Within each difficulty level, the stimuli could have either high overall contrast, or low overall contrast. For instance, in the easy, low contrast condition, one grating was at 25% contrast and the other at 15% contrast, but the easy, high contrast condition had a ratio of 50:30%. The corresponding hard conditions could be 25:20% (low contrast) or 50:40% (high contrast). Since the *relative* contrasts of the two stimuli were always preserved across low contrast and high contrast conditions, the authors predicted that participants would have the same performance in both conditions.

As expected, subjects performed better in the easy compared to the hard conditions, and had the same performance in the high contrast and low contrast conditions (as confirmation that the SNR was matched within each difficulty level). Crucially, however, observers rated higher confidence in the high contrast conditions *even though their performance was matched across low and high contrast conditions* (Fig. 2b). That is, despite nearly identical levels of performance between the high and low contrast

conditions within each level of difficulty, subjects felt *more* confident in their decisions in the high contrast compared to the low contrast conditions.

This effect is somewhat surprising given that one could expect subjects to rate their confidence depending on the strength of the sensory evidence in favor of their decision *relative to* the strength of the evidence *against* their decision (as determined by the SNR, i.e. the difficulty of the task or the subject's accuracy on the task). This is not what happens here. Instead, observers seem to rely on the absolute strength of perceptual evidence in favor of their decision ("positive evidence" as termed by Koizumi and colleagues), while tending to ignore perceptual evidence against their decision, in order to compute their level of confidence.

This effect is now well confirmed in the empirical literature. Several additional experiments have shown that similar experimental manipulations systematically result in matched discrimination performance (e.g., the same ability to tell "left tilted grating" from "right tilted grating") but mismatched confidence (e.g., in the high contrast condition, observers report feeling more confident) (Maniscalco et al. 2016; Peters et al. 2017; Samaha et al. 2016, 2017, 2019).

Peters et al. (2017) even demonstrated that it is possible to directly read out this confirmation bias in perceptual confidence by using machine learning techniques to decode electrophysiological recordings (electrocorticography; ECoG) in awake, behaving human observers as they perform perceptual decisions and rate their confidence. Participants' perceptual decisions depended on a balance of evidence for both the chosen and unchosen alternatives as decoded from the ECoG recordings, but confidence was predicted better by computational rules that relied on decision-congruent (i.e., confirmatory) evidence alone. That is, when computing confidence, subjects seem to primarily look for "positive evidence" that confirms their categorical perceptual decisions.

4 Confirmation bias in non-human animals

A final strong piece of evidence that confirmation biases can occur outside the reasoning domain is that the perceptual confidence confirmation bias is also observed in non-human animals—which, presumably, do not reason as humans do.

Odegaard et al. (2018) presented Rhesus macaques with a random-dot motion discrimination task, in which the animals had to indicate whether coherent dot motion flowed in a leftward or rightward direction. The monkeys were rewarded for correct, but not incorrect decisions, and after some decisions, were also given the opportunity to "opt-out", which provided a small but guaranteed reward. Choosing to opt-out was interpreted as indicating lower confidence: the monkey could have earned a higher reward if it made a decision that ended up being correct, but instead decided to opt for the small but guaranteed reward.

In this task, Odegaard et al. manipulated confidence by varying the amount of "positive evidence" (dot motion strength toward the correct choice) and "negative evidence" (dot motion strength toward the incorrect choice). As anticipated, monkeys' motion direction discrimination decisions seemed to depend on the ratio of positive to negative evidence on a given trial. However, whether monkeys decided to opt-out or

not depended mainly on the amount of *positive evidence*. As such, by manipulating the amount of evidence in favor of the decision without changing the *ratio* between positive and negative evidence, Odegaard et al. (2018) could induce suboptimal confidence judgments—essentially, making the monkeys act as if they were more confident or less confident irrespective of their ability to actually tell rightward versus leftward motion. As with other experiments presented above, the monkeys behaved as if they were more confident when the amount of positive evidence was high, and as if they were less confident when the amount of positive evidence was low, despite having the same objective performance in both conditions. This result is a direct replication of the confirmation bias described previously in humans as shown by Koizumi et al. (2015): the magnitude of evidence supporting the decision seems to overly influence confidence judgments over and above the probability that the decision was correct.

Stolyarova et al. (2019) even showed this confirmation bias in rats using grating stimuli very much like those used by Koizumi and colleagues (2015). Rats were trained to discriminate the orientation (vertical or horizontal) of Gabor patches with added Gaussian noise. Following their decisions, they could either wait for a delayed sugar pellet reward if confident that their decision was correct, or directly initiate a new trial to get another chance. The measure of confidence is how long the rats waited for their sugar pellet: the longer they waited, the more confident it was assumed they felt, following previous literature (Lak et al. 2014; Kepecs et al. 2008). In this task, 'high positive evidence' and 'low positive evidence' stimuli were created by increasing or decreasing the contrast of *both* the Gabor patch and the added noise. As before, since the overall ratio between signal and noise remained the same in both conditions, the rats' performances did not significantly differ between the two conditions. Confidence measures, on the other hand, changed significantly: despite having the same performance in both conditions, the rats were willing to wait significantly longer in the high positive evidence condition, compared to the low positive evidence condition. As such, it seems that even non-human animals have a confirmation bias when computing levels of confidence in their perceptual decisions: they disproportionately rely on *positive* sensory evidence (i.e. evidence supporting their decision), and tend to ignore sensory evidence contradicting their decision.

5 Summary

Taken together, these results strongly suggest that the computations underlying both perception and perceptual confidence are influenced by a *positive evidence bias*, i.e. an over-reliance on evidence supporting one's decision, analogous to a form of confirmation bias in these domains. Perceptual decisions lead to subsequent downweighting of evidence against the decision, and an increase in the absolute strength of sensory evidence *in favor* of a decision leads subjects to feel more confident, even if this increase is accompanied by a proportional increase in the evidence *against* their decision. Following Nickerson's influential definition of confirmation bias (Nickerson 1998), these biases should be viewed as a classic form of *confirmation bias*: subjects are interpreting *perceptual* evidence "in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" in order to compute their levels of confidence in their perceptual

decisions. Moreover, the fact that this confirmation bias is present in non-human animals explicitly goes against Mercier and Sperber's contention that "no confirmation bias emerges from studies of animal behavior." (Mercier and Sperber 2017, p. 217).

6 Objections and replies

We now outline three potential objections to our view that confirmation bias exists outside the reasoning domain. The first objection is that reasoning could be involved in the tasks we described. If so, the studies we reviewed do not demonstrate confirmation bias outside the reasoning domain. The second objection is that such bias would be maladaptive, which should make us suspicious that biases outside the reasoning domain could have evolved. The third objection is that the kind of biases we describe are not *genuine* instances of confirmation bias. We answer these three objections in turn.

7 Is reasoning involved in confidence confirmation bias?

One counter-argument could be that the tasks we described do not in fact fall outside the reasoning domain. According to this view, asking participants to provide confidence judgments could trigger an argumentative strategy, leading participants to think—perhaps non-consciously—about the reasons they had for making a particular decision. Since reasoning is biased, the confirmation bias observed in confidence judgments tasks would simply derive from post-perceptual reasoning, and not from the properties of the system in charge of computing confidence judgments.

This explanation is unlikely, for four main reasons. First, it does not account for the same confirmation bias in non-human animals (Odegaard et al. 2018; Stolyarova et al. 2019). Even if they had the required cognitive capacities, it is doubtful that non-human animals would feel the need to justify their confidence judgments to the experimenters, or to themselves, by engaging in reasoning. Of note, the presence of reasoning (in the sense attributed to this term by Mercier and Sperber) in non-human animals would also undermine the argumentative theory.

Second, neurophysiological evidence suggests that the confirmation bias observed in confidence judgments operates by directly biasing the rate of accumulation of evidence in early sensory areas (Talluri et al. 2018; Rollwage et al. 2020). If this bias were entirely driven by reasoning, we should expect it to work instead through the implementation of abstract decision rules outside of the sensory cortices. This is not what the empirical evidence suggests.

Third, one prediction of the 'reasoning' account of confirmation bias in perceptual confidence judgments is that the magnitude of bias should depend on the time that participants have to provide their confidence judgments. Additional time would mean that participants can generate additional reasons to convince themselves that their decisions are right, thus leading them to be more biased. However, Samaha and Denison (2020) recently found the same bias in a task where participants provided confidence judgments at the same time as their perceptual decisions, with a single

key press, immediately after stimulus presentation. Samaha and Denison (2020) interpret this finding as evidence against a post-perceptual account of confirmation bias in confidence judgments; this finding using simultaneous button presses has also been replicated by Maniscalco et al. (2020).

Finally, this hypothesis would involve subjects performing a reasoning task on top of the perceptual and confidence tasks. Given the number of trials involved in those tasks, we consider it quite unlikely that participants would systematically engage in costly reasoning strategies. Why would participants do this, if they can instead simply answer based on a read-out of their own confidence states? Without an independent justification, the hypothesis that participants systematically engage in (possibly unconscious) reasoning, in tasks that do not require them to do so, remains ad hoc.

For these reasons, we consider it unlikely that asking participants to provide confidence judgments triggers argumentative strategies, which in turn account for the observed biases. Instead, the confirmation bias in confidence judgments seems to stem directly from biases imposed on sensory evidence accumulation during perceptual decisions.

It is also worth noting that while we have provided numerous examples here of confirmation bias in confidence judgments, there are also a number of examples of confirmation bias in perception alone even when metacognition is not involved, as discussed above. The presence of these suggests that even if confirmation bias in confidence could be explained away as reasoning-based, it would not be possible to do so for perceptual effects (e.g., orientation judgments in Jazayeri and Movshon's (2007) and Stocker and Simoncelli's (2008) works; see also Jin et al. 2019).

8 Why confirmation bias in the perceptual domain won't kill you

Proponents of the argumentative theory of reasoning have systematically argued that the existence of confirmation bias outside the reasoning domain would be maladaptive. For instance, Mercier and Sperber write:

> Our survival and reproduction very much depend on the quality of the information provided by intuitions. (…) some specific biases may on the whole be advantageous when they lower the costs of cognition or make less likely particularly costly kinds of mistake. The confirmation bias carries none of these advantages. Unsurprisingly, then, no confirmation bias emerges from studies of animal behavior. A mouse bent on confirming its belief that there are no cats around and a rat focusing its attention on breadcrumbs and ignoring other foods to confirm its belief that breadcrumbs are the best food would not pass on their genes to many descendants. (Mercier and Sperber 2017, p. 217).

On this view, the kind of confirmation bias outside the reasoning domain we have reviewed would be maladaptive, which gives rise to a puzzle: why would a wide variety of cognitive and perceptual capacities be subject to confirmation bias?

One strong hypothesis rests on the observation that our environment is typically likely to be stable across time rather than abruptly changing at any given moment (Urai et al. 2019; Talluri et al. 2018; Gold and Shadlen 2007). That is, from one moment

to the next, the state of the environment is unlikely to radically change: if the water in this river is currently flowing one way, it is unlikely to suddenly start flowing in the opposite direction; if the color of the leaves is currently green, it is unlikely to suddenly become purple; and so on. In essence, it is highly likely that the state of the environment at any time point can be strongly predicted by whatever its state was only a moment prior. One hypothesis is that the brain exploits this temporal stability (or temporal dependence) in its continuous evaluation of evidence. Having reached the inference that X is occurring at time t, a reasonable guess for what is occurring at time $t + 1$ is also X.

Additionally, in deciding that X is currently occurring, the brain rules out a large (potentially infinite) number of alternative hypotheses. It would be highly inefficient to maintain the joint distribution of all possible interpretations and reevaluate incoming evidence at every time point as if all previous timepoints had never occurred. Indeed, in the vast space of possible hypotheses, computing the probability of all the unchosen alternatives at every moment (not to mention confidence in those alternatives) could occupy essentially all available neural resources[2]—especially if one assumes that the brain computes exact probabilistic inferences and not variational (e.g. Beck et al. 2012) or sampling-based approximations (e.g. Fiser et al. 2010).

Thus, given temporal stability of the environment, and the vast number of possible states of the environment, it would be computationally efficient for a system to select one categorical inference about the state of the environment and then stick to it (Stocker and Simoncelli 2008; Luu and Stocker 2018). In the face of limited computational resources, these computational properties strongly support the evolutionary utility of committing to a single high-level interpretation by iteratively *updating* a stable belief about the most likely state of the world, rather than continuously monitoring the probability of a very large number of unchosen alternatives. In fact, Qiu et al. (2019) have recently described how such a strategy appears mathematically optimal rather than suboptimal, suggesting that confirmation bias—even outside the reasoning domain—is not only not maladaptive but actually evolutionarily desirable (Qiu et al. 2019).

9 If *this* is not confirmation bias outside of the reasoning domain, what could be?

Another objection to our claim that confirmation bias exists outside the reasoning domain could be that what we have described above is not *really* a kind of confirmation bias. That is, scientists studying "confirmation bias" in the domains of perception, or confidence, are not studying a genuine kind of confirmation bias, but an altogether different phenomenon. We have two responses to this objection.

First, these phenomena do demonstrate a *bias* in the sense that the participants' behaviors are suboptimal (Rahnev and Denison 2018). Optimal bayesian models cast confidence as a readout of the probability of being correct (e.g. Kepecs et al. 2008).

[2] This is true especially if one assumes the brain computes confidence after a decision has been reached (Pouget et al. 2016).

On the contrary, in the studies we presented, the participants' confidence judgments are *suboptimal*. Participants rely more heavily on evidence supporting their decisions than evidence against their decisions, which leads them to systematically derive from optimality (Rahnev and Denison 2018; see also Li and Ma 2020).

Following others (Talluri et al. 2018; Rollwage et al. 2020), we hold that this bias is a *confirmation* bias, because it fits standard definitions of this phenomenon: participants are suboptimal because they—or rather their perceptual systems—"look for" evidence supporting their perceptual decisions, and neglect evidence against those decisions. The effects we reviewed are also consistent with Nickerson's (1998) definition of confirmation bias: "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson 1998), which is used by Mercier and Sperber themselves (2011, p. 63).

Indeed, we have argued that, when making perceptual judgments such as estimating the motion of dots relative to a reference point (Stocker and Simoncelli 2008), subjects have a bias towards making judgments that are "partial" to their existing perceptual beliefs (Brezis et al. 2015; Jazayeri and Movshon 2007; Fritsche and de Lange 2019; Luu and Stocker 2018; Zamboni et al. 2016). In the same way, extensive data suggests that, when computing confidence judgments, subjects overly rely on sensory evidence in favor of their perceptual decisions, and neglect contradictory sensory evidence (Peters et al. 2017; Zylberberg et al. 2012; Samaha et al. 2016, 2017, 2019; Maniscalco et al. 2016).

Again, these effects seem consistent with the standard definition of confirmation bias, in that they demonstrate a systematic overweighting of evidence *in favor* of choices, beliefs, or hypotheses held by the subjects, relative to the weight given to evidence against those choices, beliefs, or hypotheses. That is, once a decision is made, the cognitive systems computing confidence "look for" evidence confirming the decision, while ignoring evidence that contradicts it. Given that these effects fit standard definitions of confirmation bias—which are accepted by proponents of the argumentative theory of reasoning too—we hold that it is natural to consider them as instances of confirmation bias, as others have done (Talluri et al. 2018; Rollwage et al. 2020).

Second, if the various effects we have presented *do not* count as confirmation bias outside the reasoning domain, one might wonder what could ever count as demonstrating such bias. In other words, it is not clear what "confirmation bias outside the reasoning domain" could mean, if not to denote precisely the type of effects we have just presented, namely, the systematic overweighting of information in favor of the participants' (perceptual) decisions, and relative neglect of contradictory information.

On the other hand, if Mercier and Sperber (2011, 2017) implicitly define 'confirmation bias'—or 'myside bias'—such that it applies only to the reasoning domain, their prediction would reveal itself as a moot point. If proponents of the argumentative theory of reasoning have a different definition of 'myside bias' in mind, the burden of proof is on them to provide it and justify it. They would also need to show that their definition does not *directly* rule out the possibility of myside bias outside the reasoning domain. For otherwise, they would simply beg the question. To be meaningful, their definition of 'confirmation bias' should leave the possibility of discovering this bias outside the reasoning domain open. To the extent that this possibility is not ruled out

as a matter of definition, we hold that the effects reviewed in this article demonstrate that confirmation bias *does* exist outside the reasoning domain, thus contradicting the argumentative theory of reasoning.

10 Conclusion

The recognition of confirmation bias outside the reasoning domain could spark a revolution in the way in which confirmation bias is studied. Indeed, the simple psychophysical paradigms and stimuli used to study confirmation bias in the perceptual domain are ideally suited for realizing a wide variety of experimental manipulations. Using methods borrowed from psychophysics to study confirmation bias allows us to exactly manipulate the amount of evidence that subjects are exposed to, and measure the subjects' ability to integrate contradictory evidence. Several preliminary studies have shown promise in linking perceptual and cognitive confirmation biases, suggesting that studying confirmation in the perceptual domain may provide a fruitful avenue for eventually characterizing the source and nature of confirmation bias (e.g., Rollwage et al. 2018). Such **a** level of precision is currently impossible to reach in the study of confirmation bias in the reasoning domain, thus preventing a careful study of the neural and computational sources of such biases. Future research seeking to reveal correlations between confirmation bias in the perceptual and reasoning domains may reveal whether unique, parallel mechanisms give rise to each individually, or if a single source can drive and explain both together.

References

Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M., & Gardner, J. L. (2016). Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences of the United States of America, 113,* E3548–E3557.

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology, 56,* 1053–1077.

Beck, J., Heller, K., & Pouget, A. (2012). Complex inference in neural circuits with probabilistic population codes and topic models. *Neural Information Processing Systems*, *2*, 3059–3067.

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

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,* 20150228.

Evans, J. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Lawrence Erlbaum.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford: Stanford University Press.
Fiser, J., Berkes, P., Orban, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences, 14*(3), 119–130.
Fritsche, M., & de Lange, F. P. (2019). Reference repulsion is not a perceptual illusion. *Cognition, 184*(June 2018), 107–118.
Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30,* 535–561.
Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS ONE, 7*(9), 1–8.
Jazayeri, M., & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature, 446*(7138), 912–915.
Jin, M., Beck, J. M., & Glickfeld, L. L. (2019). Neuronal adaptation reveals a suboptimal decoding of orientation tuned populations in the mouse visual cortex. *Journal of Neuroscience, 39*(20), 3867–3881.
Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227–231.
Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics, 77*(4), 1295–1306.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498.
Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., & Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron, 84*(1), 190–201.
Li, H. H., & Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nature Communications, 11*(1), 1–11.
Lucas, E. J., & Ball, L. J. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking & Reasoning, 11,* 35–66.
Luu, L., & Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *ELife, 7,* 1–24.
Macmillan, N., & Creelman, C. (2005). *Detection theory: A user's guide*. New York, NY: Erlbaum.
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.
Maniscalco, B., Graham Castaneda, O., Morales, J., Odegaard, B., Rajananda, S., & Peters, M. A. K. (2020). The metaperceptual function: Exploring dissociations between confidence and task performance with type 2 psychometric curves. *PsyArxiv.* https://doi.org/10.31234/osf.io/5qrjn.
Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences, 20*(9), 689–700.
Mercier, H. (2017). *Confirmation bias—Myside bias. Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (2nd ed.). New York, NY: Routledge/Taylor & Francis Group.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences, 34*(2), 57–74. **(discussion 74–111)**.
Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge: Harvard University Press.
Mercier, H., & Sperber, D. (2019). Replies to critics. *Teorema, 38*(1), 139–156.
Myers, D. G. (1982). Polarizing effects of social interaction. In H. Barndstatter et al. (Eds.), *Group decision making*, *125*, 137–138.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.
Odegaard, B., Grimaldi, P., Hah Cho, S., Peters, M. A. K., Lau, H., & Basso, M. (2018). Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. In *Proceedings of the* .
Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., et al. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour, 1*(7), 0139.
Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience, 19*(3), 366–374.
Qiu, C., Luu, L., & Stocker, A. A. (2019). Benefits of commitment in hierarchical inference. *BioRxiv*.
Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences, 41,* 1–107.
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,* 1–11.
Samaha, J., & Denison, R. (2020). The positive evidence bias in perceptual confidence is not post-decisional. *BioRxiv*. https://www.biorxiv.org/content/10.1101/2020.03.15.991513v1.full.
Samaha, J., Barrett, J. J., Sheldon, A. D., LaRocque, J. J., & Postle, B. R. (2016). Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Frontiers in Psychology*, *7*, 1–8.
Samaha, J., Lemi, L., & Postle, B. R. (2017). Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Consciousness and Cognition, 54,* 47–55.
Samaha, J., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation estimation. *Journal of Vision, 19*(4), 25.
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind and Language, 25*(4), 359–393.
Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: Chicago University Press.
Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13*(3), 225–247.
Stocker, A., & Simoncelli, E. P. (2008). A Bayesian model of conditioned perception. In*: Advances in neural information processing systems* (pp. 1409–1416).
Stolyarova, A., Rakhshan, M., Peters, M. A. K., Lau, H., Soltani, A., & Izquierdo, A. (2019). Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nature Communications, 10*(4704), 1–14.
Sunstein, C. (2002). The law of group polarization. *Journal of Political Philosophy, 10*(2), 175–195.
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.e8.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.
Urai, A. E., de Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *ELife, 8,* 1–34.
Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20,* 273–281.
Zamboni, E., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2016). Do perceptual biases emerge early or late in visual processing? Decision-biases in motion perception. *Proceedings of the Royal Society B: Biological Sciences, 283*(1833), 20160263.
Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience, 6*(September), 1–10.