Erkenntnis (2022) 87:1351–1376
https://doi.org/10.1007/s10670-020-00252-1

ORIGINAL RESEARCH

# What Is the Function of Confirmation Bias?

**Uwe Peters**[1,2]

Received: 7 May 2019 / Accepted: 27 March 2020 / Published online: 20 April 2020


## Abstract

Confirmation bias is one of the most widely discussed epistemically problematic cognitions, challenging reliable belief formation and the correction of inaccurate views. Given its problematic nature, it remains unclear why the bias evolved and is still with us today. To offer an explanation, several philosophers and scientists have argued that the bias is in fact adaptive. I critically discuss three recent proposals of this kind before developing a novel alternative, what I call the 'reality-matching account'. According to the account, confirmation bias evolved because it helps us influence people and social structures so that they come to match our beliefs about them. This can result in significant developmental and epistemic benefits for us and other people, ensuring that over time we don't become epistemically disconnected from social reality but can navigate it more easily. While that might not be the only evolved function of confirmation bias, it is an important one that has so far been neglected in the theorizing on the bias.

In recent years, confirmation bias (or 'myside bias'),[1] that is, people's tendency to search for information that supports their beliefs and ignore or distort data contradicting them (Nickerson 1998; Myers and DeWall 2015: 357), has frequently been discussed in the media, the sciences, and philosophy. The bias has, for example, been mentioned in debates on the spread of "fake news" (Stibel 2018), on the "replication crisis" in the sciences (Ball 2017; Lilienfeld 2017), the impact of cognitive diversity in philosophy (Peters 2019a; Peters et al. forthcoming; Draper and Nichols 2013; De Cruz and De Smedt 2016), the role of values in inquiry (Steel 2018; Peters

[1] Mercier and Sperber (2017) and others prefer the term 'myside bias' to 'confirmation bias' because people don't have a general tendency to confirm any hypothesis that comes to their mind but only ones that are on 'their side' of a debate. I shall here use the term 'confirmation bias' because it is more common and in any case typically understood in the way just mentioned.

✉ Uwe Peters
u.peters@ucl.ac.uk

[1] Department of Philosophy, University of Southern Denmark, Odense, Denmark

[2] Department of Psychology, King's College London, De Crespigny Park, Camberwell, London SE5 8AB, UK

2018), and the evolution of human reasoning (Norman 2016; Mercier and Sperber 2017; Sterelny 2018; Dutilh Novaes 2018).

Confirmation bias is typically viewed as an epistemically pernicious tendency. For instance, Mercier and Sperber (2017: 215) maintain that the bias impedes the formation of well-founded beliefs, reduces people's ability to correct their mistaken views, and makes them, when they reason on their own, "become overconfident" (Mercier 2016: 110). In the same vein, Steel (2018) holds that the bias involves an "epistemic distortion [that] consists of unjustifiably favoring supporting evidence for [one's] belief, which can result in the belief becoming unreasonably confident or extreme" (897). Similarly, Peters (2018) writes that confirmation bias "leads to partial, and therewith for the individual less reliable, information processing" (15).

The bias is not only taken to be epistemically problematic, but also thought to be a "ubiquitous" (Nickerson 1998: 208), "built-in feature of the mind" (Haidt 2012: 105), found in both everyday and abstract reasoning tasks (Evans 1996), independently of subjects' intelligence, cognitive ability, or motivation to avoid it (Stanovich et al. 2013; Lord et al. 1984). Given its seemingly dysfunctional character, the apparent pervasiveness of confirmation bias raises a puzzle: If the bias is indeed epistemically problematic, why is it still with us today? By definition, dysfunctional traits should be more prone to extinction than functional ones (Nickerson 1998). Might confirmation bias be or have been *adaptive*?

Some philosophers are optimistic, arguing that the bias has in fact significant advantages for the individual, groups, or both (Mercier and Sperber 2017; Norman 2016; Smart 2018; Peters 2018). Others are pessimistic. For instance, Dutilh Novaes (2018) maintains that confirmation bias makes subjects less able to anticipate other people's viewpoints, and so, "given the importance of being able to appreciate one's interlocutor's perspective for social interaction", is "best not seen as an adaptation" (520).

In the following, I discuss three recent proposals of the adaptationist kind, mention reservations about them, and develop a novel account of the evolution of confirmation bias that challenges a key assumption underlying current research on the bias, namely that the bias thwarts reliable belief formation and truth tracking. The account holds that while searching for information supporting one's pre-existing beliefs and ignoring contradictory data is disadvantageous when that what one takes to be reality is and stays different from what one believes it to be, it is beneficial when, as the result of one's processing information in that way, that reality is changed so that it matches one's beliefs. I call this process *reality matching* and contend that it frequently occurs when the beliefs at issue are about people and social structures (i.e., relationships between individuals, groups, and socio-political institutions). In these situations, confirmation bias is highly effective for us to be confident about our beliefs even when there is insufficient evidence or subjective motivation available to us to support them. This helps us influence and 'mould' people and

social structures so that they fit our beliefs,[2] which is an adaptive property of confirmation bias. It can result in significant developmental and epistemic benefits for us and other people, ensuring that over time we don't become epistemically disconnected from social reality but can navigate it more easily.

I shall not argue that the adaptive function of confirmation bias that this reality-matching account highlights is the only evolved function of the bias. Rather, I propose that it is one important function that has so far been neglected in the theorizing on the bias.

In Sects. 1 and 2, I distinguish confirmation bias from related cognitions before briefly introducing some recent empirical evidence supporting the existence of the bias. In Sect. 3, I motivate the search for an evolutionary explanation of confirmation bias and critically discuss three recent proposals. In Sects. 4 and 5, I then develop and support the reality-matching account as an alternative.

## 1 Confirmation Bias and Friends

The term 'confirmation bias' has been used to refer to various distinct ways in which beliefs and expectations can influence the selection, retention, and evaluation of evidence (Klayman 1995; Nickerson 1998). Hahn and Harris (2014) offer a list of them including four types of cognitions: (1) hypothesis-determined information seeking and interpretation, (2) failures to pursue a falsificationist strategy in contexts of conditional reasoning, (3) a resistance to change a belief or opinion once formed, and (4) overconfidence or an illusion of validity of one's own view.

Hahn and Harries note that while all of these cognitions have been labeled 'confirmation bias', (1)–(4) are also sometimes viewed as components of 'motivated reasoning' (or 'wishful thinking') (ibid: 45), i.e., information processing that leads people to arrive at the conclusions they favor (Kunda 1990). In fact, as Nickerson (1998: 176) notes, confirmation bias comes in two different flavors: "motivated" and "unmotivated" confirmation bias. And the operation of the former can be understood as motivated reasoning itself, because it too involves partial information processing to buttress a view that one wants to be true (ibid). Unmotivated confirmation bias, however, operates when people process data in one-sided, partial ways that support their predetermined views no matter whether they favor them. So confirmation bias is also importantly different from motivated reasoning, as it can take effect in the absence of a preferred view and might lead one to support even beliefs that one wants to be *false* (e.g., when one believes the catastrophic effects of climate change are unavoidable; Steel 2018).

Despite overlapping with motivated reasoning, confirmation bias can thus plausibly be (and typically is) construed as a distinctive cognition. It is thought to be a subject's largely automatic and unconscious tendency to (i) seek support for her

[2] Researchers working on folk psychology might be reminded of the 'mindshaping' view of folk psychology (Mameli 2001; Zawidzki 2013). I will come back to this view and demarcate it from my account of confirmation bias here in Sect. 5.

pre-existing, favored or not favored beliefs and (ii) ignore or distort information compromising them (Klayman 1995: 406; Nickerson 1998: 175; Myers and DeWall 2015: 357; Palminteri et al. 2017: 14). I here endorse this standard, functional concept of confirmation bias.

## 2 Is Confirmation Bias Real?

Many psychologists hold that the bias is a "pervasive" (Nickerson 1998: 175; Palminteri et al. 2017: 14), "ineradicable" feature of human reasoning (Haidt 2012: 105). Such strong claims are problematic, however. For there is evidence that, for instance, disrupting the fluency in information processing (Hernandez and Preston 2013) or priming subjects for distrust (Mayo et al. 2014) reduces the bias. Moreover, some researchers have recently re-examined the relevant studies and found that confirmation bias is in fact less common and the evidence of it less robust than often assumed (Mercier 2016; Whittlestone 2017). These researchers grant, however, the weaker claim that the bias is real and often, in some domains more than in others, operative in human cognition (Mercier 2016: 100, 108; Whittlestone 2017: 199, 207). I shall only rely on this modest view here. To motivate it a bit more, consider the following two studies.

Hall et al. (2012) gave their participants (N = 160) a questionnaire, asking them about their opinion on moral principles such as 'Even if an action might harm the innocent, it can still be morally permissible to perform it'. After the subjects had indicated their view using a scale ranging from 'completely disagree' to 'completely agree', the experimenter performed a sleight of hand, inverting the meaning of some of the statements so that the question then read, for instance, 'If an action might harm the innocent, then it is not morally permissible to perform it'. The answer scales, however, were not altered. So if a subject had agreed with the first claim, she then agreed with the opposite one. Surprisingly, 69% of the study participants failed to detect at least one of the changes. Moreover, they subsequently tended to justify positions they thought they held despite just having chosen the *opposite*. Presumably, subjects accepted that they favored a particular position, didn't know the reasons, and so were now looking for support that would justify their position. They displayed a confirmation bias.[3]

Using a similar experimental set-up, Trouche et al. (2016) found that subjects also tend to exhibit a selective 'laziness' in their critical thinking: they are more likely to

[3] It might be proposed that when participants in the experiment seek reasons for their judgments, perhaps they take themselves already to have formed the judgements for *good reasons* and then wonder what these reasons might have been. Why would they seek reasons *against* a view that they have formed (by their own lights) for good reasons? However, we might equally well ask why they would take themselves to have formed a judgment for good reasons in the first place even though they *don't know* any of them? If it is a general default tendency to assume that any view that one holds rests on good reasons, then that would again suggest the presence of a confirmation bias. For a general tendency to think that one's views rest on good reasons *even when one doesn't know them* is a tendency to favor and confirm these views while resisting balanced scrutiny of their basis.

avoid raising objections to their own positions than to other people's. Trouche et al. first asked their test participants to produce arguments in response to a set of simple reasoning problems. Directly afterwards, they had them assess other subjects' arguments concerning the same problems. About half of the participants didn't notice that by the experimenter's intervention, in some trials, they were in fact presented with their *own* arguments again; the arguments appeared to these participants as if they were someone else's. Furthermore, more than half of the subjects who believed they were assessing someone else's arguments now rejected those that were in fact their own, and were more likely to do so for invalid than for valid ones. This suggests that subjects are less critical of their own arguments than of other people's, indicating that confirmation bias is real and perhaps often operative when we are considering our own claims and arguments.

## 3 Evolutionary Accounts of the Bias

Confirmation bias is typically taken to be epistemically problematic, as it leads to partial and therewith for the individual less reliable information processing and contributes to failures in, for instance, perspective-taking with clear costs for social and other types of cognition (Mercier and Sperber 2017: 215; Steel 2018; Peters 2018; Dutilh Novaes 2018). *Prima facie*, the bias thus seems maladaptive.

But then why does it still exist? Granted, even if the bias isn't an adaptation, we might still be able to explain why it is with us today. We might, for instance, argue that it is a "spandrel", a by-product of the evolution of another trait that is an adaptation (Gould and Lewontin 1979). Or we may abandon the evolutionary approach to the bias altogether and hold that it emerged by chance.

However, evolutionary explanations of psychological traits are often fruitful. They can create new perspectives on these traits that may allow developing means to reduce the traits' potential negative effects (Roberts et al. 2012; Johnson et al. 2013). Evolutionary explanations might also stimulate novel, testable predictions that researchers who aren't evolutionarily minded would overlook (Ketelaar and Ellis 2000; De Bruine 2009). Moreover, they typically involve integrating diverse data from different disciplines (e.g., psychology, biology, anthropology etc.), and thereby contribute to the development of a more complete understanding of the traits at play and human cognition, in general (Tooby and Cosmides 2015). These points equally apply when it comes to considering the origin of confirmation bias. They provide good reasons for searching for an evolutionary account of the bias.

Different proposals can be discerned in the literature. I will discuss three recent ones, what I shall call (1) the *argumentative-function* account, (2) the *group-cognition* account, and the (3) *intention–alignment* account. I won't offer conclusive arguments against them here. The aim is just to introduce some reservations about these proposals to motivate the exploration of an alternative.

### 3.1 The Argumentative-Function Account

Mercier and Sperber (2011, 2017) hold that human reasoning didn't evolve for truth tracking but for making us better at convincing other people and evaluating their arguments so as to be convinced only when their points are compelling. In this context, when persuasion is paramount, the tendency to look for material supporting our preconceptions and to discount contradictory data allows us to accumulate argumentative ammunition, which strengthens our argumentative skill, Mercier and Sperber maintain. They suggest that confirmation bias thus evolved to "serve the goal of convincing others" (2011: 63).

Mercier and Sperber acknowledge that the bias also hinders us in anticipating objections, which should make it more difficult for us to develop strong, objection–resistant arguments (2017: 225f). But they add that it is much less cognitively demanding to *react* to objections than to anticipate them, because objections might depend on particular features of one's opponents' preferences or on information that only they have access to. It is thus more efficient to be 'lazy' in anticipating criticism and let the audience make the moves, Mercier and Sperber claim.

There is reason to be sceptical about their proposal, however. For instance, an anticipated objection is likely to be answered more convincingly than an immediate response from one's audience. After all, "forewarned is forearmed"; it gives a tactical advantage (e.g., more time to develop a reply) (Sterelny 2018: 4). And even if it is granted that objections depend on private information, they also often derive from obvious interests and public knowledge, making an anticipation of them easy (ibid). Moreover, as Dutilh Novaes (2018: 519) notes, there is a risk of "looking daft" when producing poor arguments, say, due to laziness in scrutinizing one's thoughts. Since individuals within their social groups depend on their reputation so as to find collaborators, anticipating one's audience's responses should be and have been more adaptive than having a confirmation bias (ibid). If human reasoning emerged for argumentative purposes, the existence of the bias remains puzzling.

### 3.2 The Group-Cognition Account

Even if confirmation bias is maladaptive for *individuals*, it might still be adaptive for *groups*. For instance, Smart (2018) and Peters (2018) hold that in groups with a sufficient degree of cognitive diversity at the outset of solving a particular problem, each individual's confirmation bias might help the group as a whole conduct a more in-depth analysis of the problem space than otherwise. When each subject is biased towards a different particular proposal on how to solve the problem, the bias will push them to invest greater effort in defending their favored proposals and might, in the light of counterevidence, motivate them to consider rejecting auxiliary assumptions rather than the proposals themselves. This contributes to a thorough exploration of them that is less likely with less committed thinkers. Additionally, since individuals appear to have a particular strength in detecting flaws in *others'* arguments (Trouche et al. 2016), open social criticism within the group should ensure that the

*group's* conclusions remain reliable even if some, or at times most, of its members are led astray by their confirmation bias (Smart 2018: 4190; Peters 2018: 20).

Mercier and Sperber (2011: 65) themselves already float the idea of such a social "division of cognitive labor". They don't yet take its group-level benefits to explain why confirmation bias evolved, however (Dutilh Novaes 2018: 518f). Smart (2018) and Peters (2018) also don't introduce their views as accounts of the *evolved* function of the bias. But Dutilh Novaes (2018: 519) and Levy (2019: 317) gesture toward, and Smith and Wald (2019) make the case for, an evolutionary proposal along these lines, arguing that the bias was selected for making a group's inquiry more thorough, effective, and reliable.

While I have sympathies with this proposal, several researchers have noted that the concept of 'group selection' is problematic (West et al. 2007; Pinker 2012). One of the issues is that since individuals reproduce faster than groups, a trait *T* that is an adaptation that is good for groups but bad for an individual's fitness won't spread, because the rate of proliferation of groups is undermined by the evolutionary disadvantage of *T* within groups (Pinker 2012). The point equally applies to the proposal that confirmation bias was selected for its group-level benefits.

Moreover, a group arguably only benefits from each individual's confirmation bias if there is a diversity of viewpoints in the group and members express their views, as otherwise "group polarization" is likely to arise (Myers and Lamm 1976): arguments for shared positions will accumulate without being criticized, making the group's average opinion more extreme and less reliable, which is maladaptive. Crucially, ancestral 'hunter-gather' groups are perhaps unlikely to have displayed a diversity of viewpoints. After all, their members traveled less, interacted less with strangers, and were less economically dependent on other groups (Simpson and Beckes 2010: 37). This should have homogenized them with respect to race, culture, and background (Schuck 2001: 1915). Even today groups often display such homogeneity, as calls for diversity in academia, companies etc. indicate. These points provide reasons to doubt that ancestral groups provided the kind of conditions in which confirmation bias could have produced the benefits that the group-cognition account highlights rather than maladaptive effects tied to group polarization.

### 3.3 The Intention–Alignment Account

Turning to a third and here final extant proposal on the evolution of confirmation bias, Norman (2016) argues that human reasoning evolved for facilitating an "intention alignment" between individuals: in social interactions, reasons typically 'overwrite' nonaligned mental states (e.g., people's divergent intentions or beliefs) with aligned ones by showing the need for changing them. Norman holds that human reasoning was selected for this purpose because it makes cooperation easier. He adds that, in this context, "confirmation bias would have facilitated intention alignment, for a tribe of hunter-gatherers prone to [the bias] would more easily form and maintain the kind of shared outlook needed for mutualistic collaboration. The mythologies and ideologies taught to the young would accrue confirming evidence and tend to stick, thereby cementing group solidarity" (2016: 700). Norman takes his view

to be supported by the "fact that confirmation bias is especially pronounced when a group's ideological preconceptions are at stake" (ibid).

However, the proposal seems at odds with the finding that the bias inclines subjects to ignore or misconstrue their opponents' objections. In fueling one-sided information processing to support one's own view, the bias makes people less able to anticipate and adequately respond to their interlocutor's point of view (Dutilh Novaes 2018: 520). Due to that effect, the bias arguably makes an intention alignment with others (especially with one's opponents) harder, not easier. Moreover, since our ancesteral groups are (as noted above) likely to have been largely viewpoint homogenous, in supporting intention-alignment in these social environments, confirmation bias would have again facilitated group polarization, which is *prima facie* evolutionarily disadvantageous.

All three proposals of the adaptive role of confirmation bias considered so far thus raise questions. While the points mentioned aren't meant to be fatal for the proposals and might be answerable within their frameworks, they do provide a motivation to explore an alternative.

## 4 Towards an Alternative

The key idea that I want to develop is the following. Confirmation bias is typically taken to work against an individual's truth tracking (Mercier and Sperber 2017: 215; Peters 2018: 15), and indeed searching for information supporting one's beliefs and ignoring contradictory data is epistemically disadvantageous when what one takes to be reality is and stays different from what one believes it to be. However, reality doesn't always remain unchanged when we form beliefs about it. Consider social beliefs, that is, beliefs about people (oneself, others, and groups) and social structures (i.e., relationships between individuals, groups, and socio-political institutions). I shall contend that a confirmation bias pertaining to social beliefs reinforces our confidence in these beliefs, therewith strengthening our tendency to behave in ways that cause changes in reality so that it corresponds to the beliefs, turning them (when they are initially inaccurate) into *self-fulfilling prophecies* (SFPs) (Merton 1948; Biggs 2009). Due to its role in helping us make social reality match our beliefs, confirmation bias is adaptive, or so I will argue. I first introduce examples of SFPs of social beliefs. Then I explore the relevance of these beliefs in our species, before making explicit the adaptive role of confirmation bias in facilitating SFPs.

### 4.1 Social Beliefs and SFPs

Social beliefs often lead to SFPs with beneficial outcomes. Here are four examples.

1. *S* (false) believes he is highly intelligent. His self-view motivates him to engage with intellectuals, read books, attend academic talks, etc. This makes him increas-

ingly more intelligent, gradually confirming his initially inaccurate self-concept (for relevant empirical data, see Swann 2012).

2. Without a communicative intention, a baby boy looking at a kitten produces a certain noise: 'ma-ma'. His mother is thrilled, believing (falsely) that he is beginning to talk and wants to call her. She responds accordingly, rushing to him, attending to him, and indicating excitement. This leads the boy to associate 'ma-ma' with the arrival and attention of his mother. And so he gradually begins using the sounds to call her, confirming her initially false belief about his communicative intention (for relevant empirical data, see Mameli 2001).
3. A father believes his adolescent daughter doesn't regularly drink alcohol, but she does. He acts in line with his beliefs, and expresses it in communication with other people. His daughter notices and likes his positive view of her, which motivates her to increasingly resist drinks, gradually fulfilling her father's optimistic belief about her (for relevant empirical data; see Willard et al. 2008).
4. A teacher (falsely) believes that a student's current academic performance is above average. She thus gives him challenging material, encourages him, and communicates high expectations. This leads the student to increase his efforts, which gradually results in above-average academic performance (for relevant evidence, see Madon et al. 1997).

SFPs of initially false positive trait ascriptions emerge in many other situations too. They also occurred, for instance, when adults ascribed to children traits such as being tidy (Miller et al. 1975), charitable (Jensen and Moore 1977), or cooperative (Grusec et al. 1978). Similarly, in adults, attributions of, for example, kindness (Murray et al. 1996), eco-friendliness (Cornelissen et al. 2007), military competence (Davidson and Eden 2000), athletic ability (Solomon 2016), and even physiological changes (Turnwald et al. 2018) have all had self-fulfilling effects. Moreover, these effects don't necessarily take much time to unfold but can happen swiftly in a single interaction (e.g., in interview settings; Word et al. 1974) right after the ascription (Turnwald et al. 2018: 49).

SFPs are, however, neither pervasive nor all-powerful (Jussim 2012), and there are various conditions for them to occur (Snyder and Klein 2007). For instance, they tend to occur only when targets are able to change in accordance with the trait ascriptions, when the latter are believable rather than unrealistic (Alfano 2013: 91f), and when the ascriber holds more power than the ascribee (Copeland 1994: 264f). But comprehensive literature reviews confirm that SFPs are "real, reliable, and occasionally quite powerful" (Jussim 2017: 8; Willard and Madon 2016).

### 4.2 The Distribution of Social Beliefs and Role of Prosociality in Humans

Importantly, SFPs can be pernicious when the beliefs at the center of them capture negative social conceptions, for instance, stereotypes, anxious expectations, fear, or hostility (Darley and Gross 1983; Downey et al. 1998; Madon et al. 2018). In these cases, SFPs would be maladaptive. Given this, what do we know about the

distribution of social beliefs, in general, and positive ones, in particular, in ancestral human groups?

Many researchers hold that our evolutionary success as a species relies on our being "ultra-social" and "ultra-cooperative" animals (e.g., Tomasello 2014: 187; Henrich 2016). Human sociality is "spectacularly elaborate, and of profound biological importance" because "our social groups are characterized by extensive cooperation and division of labour" (Sterelny 2007: 720). Since we live in an almost continuous flow of interactions with conspecifics, "solving problems of coordination with our fellows is [one of] our most pressing ecological tasks" (Zawidzki 2008: 198). A significant amount of our beliefs are thus likely to be social ones (Tomasello 2014: 190f).

Moreover, when it comes to oneself, to group or "tribe" members, and to collaborators, these beliefs often capture positive to overly optimistic ascriptions of traits (e.g., communicativeness, skills, etc.; Simpson and Beckes 2010). This is well established when it comes to one's beliefs about oneself (about 70% of the general population has a positive self-conception; Talaifar and Swann 2017: 4) and one's family members (Wenger and Fowers 2008). The assumption that the point also holds for 'tribe' members and collaborators, more generally, receives support from the "tribal-instincts hypothesis" (Richerson and Boyd 2001), which holds that humans tend to harbor "ethnocentric attitudes in favor of [their] own tribe along with its members, customs, values and norms", as this facilitates social predictability and cooperation (Kelly 2013: 507). For instance, in the past as much as today, humans "talk differently about their in-groups than their out-groups, such that they describe the in-group and its members [but not out-groups] as having broadly positive traits" (Stangor 2011: 568). In subjects with such 'tribal instincts', judgments about out-group members might easily be negative. But within the groups of these subjects, among in-group members, overly optimistic, cooperation-enhancing conceptions of others should be and have been more dominant particularly in "intergroup conflict, [which] is undeniably pervasive across human societies" (McDonald et al. 2012: 670). Indeed, such conflicts are known to fuel in-group "glorification" (Leidner et al. 2010; Golec De Zavala 2011).

Given these points, in 'ultra-cooperative' social environments in which 'tribe' members held predominantly positive social conceptions and expectations about in-group subjects, positive SFPs should have been overall more frequent and stronger than negative ones. Indeed, there is evidence that even today, positive SFPs in individual, dyadic interactions are more likely and pronounced than negative ones.[4] For instance, focusing on mothers' beliefs about their sons' alcohol consumption, Willard et al. (2008) found that children "were more susceptible to their mothers'

[4] SFPs can also accumulate when they occur across different interactions, and in contemporary societies, overall accumulative SFP effects of negative social beliefs capturing, e.g., stereotypes might be stronger than those of positive social beliefs in individual dyadic interactions (Madon et al. 2018). However, in ancestral, 'tribal' groups of highly interdependent subjects, even accumulative SFPs of, e.g., stereotypes would perhaps still have contributed to conformity and social stability. I shall return to the possible SFP-related benefits of nowadays highly negative social conceptions, i.e., stereotypes, ethnocentrism etc. below.

positive than negative self-fulfilling effects" (499): "mothers' false beliefs buffered their adolescents against increased alcohol use rather than putting them at greater risk" (Willard and Madon 2016: 133). Similarly, studies found that "teachers' false beliefs raised students' achievement more than they lowered it" (Willard and Madon 2016: 118): teacher overestimates "increase[d] achievement more than teacher underestimates tended to decrease achievement among students" (Madon et al. 1997: 806). Experiments with stigmatized subjects corroborate these results further (ibid), leading Jussim (2017) in his literature review to conclude that high teacher expectations help students "more than low expectations harm achievement" (8).

One common explanation of this asymmetry is that SFPs typically depend on whether the targets of the trait ascriptions involved accept the expectations imposed on them via the ascriptions (Snyder and Klein 2007). And since subjects tend to strive to think well of themselves (Talaifar and Swann 2017), they respond more to positive than negative expectations (Madon et al. 1997: 792). If we combine these considerations with the assumption that in ancestral groups of heavily interdependent subjects, positive social beliefs about in-group members (in-group favoritism) are likely to have been more prevalent than negative ones, then there is reason to hold that the SFPs of the social conceptions in the groups at issue were more often than not adaptive. With these points in mind, it is time to return to confirmation bias.

### 4.3 From SFPs to Confirmation Bias

Notice that SFPs depend on trait or mental-state ascriptions that are 'ahead' of their own truth: they are formed when an objective assessment of the available evidence doesn't yet support their truth. Assuming direct doxastic voluntarism is false (Matheson and Vitz 2014), how can they nonetheless be formed and confidently maintained?

I suggest that confirmation bias plays an important role: it allows subjects to become and remain convinced about their social beliefs (e.g., trait ascriptions) when the available evidence doesn't yet support their truth. This makes SFPs of these beliefs more likely than if the ascriber merely verbally attributed the traits without committing to the truth of the ascriptions, or believed in them but readily revised the beliefs. I shall argue that this is in fact adaptive not only when it comes to positive trait ascriptions, but also to negative ones. I will illustrate the point first with respect to positive trait ascriptions.

#### 4.3.1 Motivated Confirmation Bias and Positive Trait Ascriptions

Suppose that you ascribe a positive property *T* to a subject *A*, who is your ward, but (unbeknownst to you) the available evidence doesn't yet fully support that ascription. The more convinced you are about your view of *A* even in the light of counterevidence, the better you are at conveying your conviction to *A* because, generally, "people are more influenced [by others] when [these] others express judgments with high confidence than low confidence" (Kappes et al. 2020: 1; von Hippel and

Trivers 2011). Additionally, the better you are at conveying to *A* your conviction that he has *T*, the more confident he himself will be that he has that trait (assuming he trusts you) (Sniezek and Van Swol 2001). Crucially, if *A* too is confident that he has *T*, he will be more likely to conform to the corresponding expectations than if he doesn't believe the ascription, say, because he notices that you only *say* but don't believe that he has *T*. Relatedly, the more convinced you are about your trait ascription to *A*, the clearer your signaling of the corresponding expectations to *A* in your behavior (Tormala 2016) and the higher the normative impetus on him, as a cooperative subject, to conform so as to avoid disrupting interactions with you.

Returning to confirmation bias, given what we know about the cognitive effect of the bias, the more affected you are by the bias, the stronger your belief in your trait ascriptions to *A* (Rabin and Schrag 1999), and so the lower the likelihood that you will reveal in your behavior a lack of conviction about them that could undermine SFPs. Thus, the more affected you are by the bias, the higher the likelihood of SFPs of the ascriptions because conviction about the ascriptions plays a key facilitative role for SFPs. This is also experimentally supported. For several studies found that SFPs of trait ascriptions occurred only when ascribers were *certain* of the ascriptions, not when they were less confident (Swann and Ely 1984; Pelham and Swann 1994; Swann 2012: 30). If we add to these points that SFPs of trait ascriptions were in developmental and educational contexts in ancestral tribal groups more often beneficial for the targets than not, then there is a basis for holding that confirmation bias might in fact have been selected for sustaining SFPs.

Notice that the argument so far equally applies to motivated reasoning. This is to be expected because, as mentioned above, *motivated* confirmation bias is an instance of motivated reasoning (Nickerson 1998). To pertain specifically to confirmation bias, however, the evolutionary proposal that the bias was selected for facilitating SFPs of social conceptions also has to hold for *unmotivated* confirmation bias. Is this the case?

#### 4.3.2 Unmotivated Confirmation Bias and Negative Trait Ascriptions

Notice that when we automatically reinforce any of our views no matter whether we favor them, then our preferences won't be required for and undermine the reinforcement process and the SFPs promoted by it. This means that such a general tendency, i.e., a confirmation bias, can fulfil the function of facilitating SFPs more frequently than motivated cognitions, namely whenever the subject has acquired a social conception (e.g., as the result of upbringing, learning, or testimony). This is adaptive for at least three reasons.

First, suppose that as a parent, caretaker, or teacher you (unknowingly) wishfully believe that *A*, who is your ward, has a positive trait *T*. You tell another subject (*B*) that *A* has *T*, and, on your testimony, *B* subsequently believes this too. But suppose that unlike you, *B* has no preference as to whether *A* has *T*. Yet, as it happens, she still has a confirmation bias toward her beliefs. Just like you, *B* will now process information so that it strengthens her view about *A*. This increases her conviction in, and so the probability of an SFP of, the trait ascription to *A*, because now both you and *B* are more likely to act toward *A* in ways indicating ascription-related

expectations. As a general tendency to support any of one's beliefs rather than only *favored* ones, the bias thus enables a social 'ripple' effect in the process of making trait ascriptions match reality. Since this process is in ultra-social and ultra-cooperative groups more often than not adaptive (e.g., boosting the development of a positive trait in *A*), in facilitating a social extension of it, confirmation bias is adaptive too.

Secondly, in ancestral groups, many of the social conceptions (e.g., beliefs about social roles, gender norms, stereotypes etc.) that subjects unreflectively acquired during their upbringing and socialization will have been geared toward preserving the group's function and *status quo* and aligning individuals with them (Sterelny 2006: 148). Since it can operate independently of a subject's preferences, a confirmation bias in each member of the group would have helped the group enlist each of its members for re-producing social identities, social structures, traits, and roles in the image of the group's conceptions even when these individuals disfavored them. In sustaining SFPs of these conceptions, which might have included various stereotypes or ethnocentric, prejudicial attitudes that we today consider offensive negative trait ascriptions (e.g., gender or racist stereotypes) (Whitaker et al. 2018), confirmation bias would have been adaptive in the past. For, as Richerson and Boyd (2005: 121f) note too, in ancestral groups, selection pressure favored social conformity, predictability, and stability. That confirmation bias might have evolved for facilitating SFPs that serve the 'tribal' *collective*, possibly even against the preference, autonomy, and better judgment of the individual, is in line with recent research suggesting that many uniquely human features of cognition evolved through pressures selecting for the ability to conform to other people and to facilitate social projects (Henrich 2016). It is thought that these features may work against common ideals associated with self-reliance or "achieving basic personal autonomy, because the main purpose of [them] is to allow us to fluidly mesh with others, making us effective nodes in larger networks" (Kelly and Hoburg 2017: 10). I suggest that confirmation bias too was selected for making us effective 'nodes' in social networks by inclining us to create social reality that corresponds to these networks' conceptions even when we dislike them or they are harmful to others (e.g., out-group members).

Thirdly, in helping us make social affairs match our beliefs about them even when we don't favor them, confirmation bias also provides us with significant epistemic benefits in social cognition. Consider Jack and Jill. Both have just seen an agent *A* act ambiguously, and both have formed a first impression of *A* according to which *A* is acting the way he is because he has trait *T*. Suppose neither Jack nor Jill has any preference as to whether *A* has that trait but subsequently process information in the following two different ways. Jack does *not* have a confirmation bias but impartially assesses the evidence and swiftly revises his beliefs when encountering contradictory data. As it happens, *A*'s behavior soon does provide him with just such evidence, leading him to abandon his first impression of *A* and reopen the search for an explanation of *A*'s action. In contrast, Jill *does* have a confirmation bias with respect to her beliefs and interprets the available evidence so that it supports her beliefs. Jill too sees *A* act in a way that contradicts her first impression of him. But unlike Jack, she doesn't abandon her view. Rather, she reinterprets *A*'s action so that it bolsters her view. Whose information processing might be more adaptive?

For Jack, encountering data challenging his view removes certainty and initiates a new cycle of computations about *A*, which requires him to postpone a possible collaboration with *A*. For Jill, however, the new evidence strengthens her view, leading her to keep the issue of explaining *A*'s action settled and be ready to collaborate with him. Jack's approach might still seem better for attaining an accurate view of *A* and predicting what he'll do next. But suppose Jill confidently signals to *A* her view of him in her behavior. Since people have a general inclination to fulfil others' expectations (especially positive ones) out of an interest in coordinating and getting along with them (Dardenne and Leyens 1995; Bacharach et al. 2007), when *A* notices Jill's conviction that he displays *T*, he too is likely to conform, which provides Jill with a correct view of what he will do next. Jill's biased processing is thus more adaptive than Jack's approach: a confirmation bias provides her with certainty and simpler information processing that simultaneously facilitates accurate predictions (via contributing to SFPs). Generalizing from Jill, in everyday social interactions we all form swift first impressions of others without having any particular preference with respect to these impressions either way. Assuming that confirmation bias operates on them nonetheless, the bias will frequently be adaptive in the ways just mentioned.

### 4.3.3 Summing Up: The Reality-Matching Account

By helping subjects make social reality match their beliefs about it no matter whether they favor these beliefs or the latter are sufficiently evidentially supported, confirmation bias is adaptive: when the bias targets positive social beliefs and trait ascriptions, it serves both the subject and the group by producing effects that (1) assist them in their development (to become, e.g., more communicative, cooperative, or knowledgeable) and (2) make social cognition more tractable (by increasing social conformity and predictability). To be sure, when it targets negative trait ascriptions (pernicious stereotypes, etc.), the bias can have ethically problematic SFP effects. But, as noted, especially in ancestral 'tribal' groups, it would perhaps still have contributed to social conformity, predictability, and sustaining the *status quo*, which would have been adaptive in these groups (Richerson and Boyd 2005) *inter alia* by facilitating social cognition. Taken together, these considerations provide a basis for holding that confirmation bias was selected for promoting SFPs. I shall call the proposal introduced in this section, the *reality-matching* (RM) account of the function of confirmation bias.

## 5 Supporting the RM Account

Before offering empirical support for the RM account and highlighting its explanatory benefits, it is useful to disarm an objection: if confirmation bias was selected for its SFP-related effects, then people should not also display the bias with respect to beliefs that *can't* produce SFPs (e.g., beliefs about physics, climate change, religion, etc.). But they do (Nickerson 1998).

### 5.1 From Social to Non-social Beliefs

In response to the objection just mentioned, two points should be noted. First, the RM account is compatible with the view that confirmation bias was *also* selected for adaptive effects related to *non*-social beliefs. It only claims that facilitating the alignment of social reality with social beliefs (i.e., reality matching) is one of the important adaptive features for which the bias was selected that has so far been neglected.

Second, it doesn't follow that because confirmation bias also affects beliefs that can't initiate SFPs that it could not have been selected for affecting beliefs that can and do initiate SFPs. The literature offers many examples of biological features or cognitive traits that were selected for fulfilling a certain function despite rarely doing so or even having maladaptive effects (Millikan 1984; Haselton and Nettle 2006). Consider the "baby-face overgeneralization" bias (Zebrowitz and Montepare 2008). Studies suggest that people have a strong readiness to favorably respond to babies' distinctive facial features. And this tendency is overgeneralized such that even adults are more readily viewed more favorably, treated as likeable (but also physically weak, and naïve) when they display babyface features. While this overgeneralization tendency often leads to errors, it is thought to have evolved because *failures* to respond favorably to babies (i.e., false negatives) are evolutionarily more costly than overgeneralizing (i.e., false positives) (ibid).

Might our domain-general tendency to confirm our own beliefs be similarly less evolutionarily costly than not having such a general tendency? It is not implausible to assume so because, as noted, we are ultra-social and ultra-cooperative, and our beliefs about people's social standing, knowledge, intentions, abilities, etc. are critical for our flourishing (Sterelny 2007: 720; Tomasello 2014: 190f; Henrich 2016). Importantly, these beliefs, unlike beliefs about the non-social world, are able to and frequently do initiate SFPs contributing to the outlined evolutionary benefits. This matters because if social beliefs are pervasive and SFPs of them significant for our flourishing, then a domain-general tendency to confirm *any* of our beliefs ensures that we don't miss opportunities to align social reality with our conceptions and to reap the related developmental and epistemic benefits. Granted, this tendency overgeneralizes, which creates clear costs. But given the special role of social beliefs in our species and our dependence on social learning and social cognition, which are facilitated by SFPs, it is worth taking seriously the possibility that these costs can often outweigh the benefits.

While this thought doesn't yet show that the RM account is correct, it does help disarm the above objection. For it explains why the fact that confirmation bias *also* affects beliefs that cannot initiate SFPs doesn't disprove the view that the bias was selected for reality matching: the special role of social beliefs in our species (compared to others species) lends plausibility to the assumption that the costs of the bias' overgeneralizing might be lower than the costs of its failing to generalize. I now turn to the positive support for the RM account.

### 5.2 Empirical Data

If, as the RM account proposes, confirmation bias was selected for facilitating the process of making reality match our beliefs, then the bias should be common and pronounced when (1) it comes to social beliefs, that is, beliefs (a) about oneself, (b) about other people, and (c) about social structures that the subject can determine, and when (2) social conditions are conducive to reality matching. While there are no systematic comparative studies on whether the bias is more frequent or stronger with respect to some beliefs but not others (e.g., social vs. non-social beliefs), there is related empirical research that does provide some support for these predictions.

(a) *Self-related Beliefs*

In a number of studies, Swann and colleagues (Swann 1983; Swann et al. 1992; for an overview, see Swann 2012) found that selective information processing characteristic of confirmation bias is "especially pronounced with regards to self-concepts" and so self-related beliefs (Müller-Pinzler et al. 2019: 9).[5] Interestingly, and counter-intuitively, the data show that "just as people with positive self-views preferentially seek positive evaluations, those with *negative* self-views preferentially seek *negative* evaluations" (Talaifar and Swann 2017: 3). For instance, those "who see themselves as likable seek out and embrace others who evaluate them positively, whereas those who see themselves as dislikeable seek out and embrace others who evaluate them negatively" (ibid). Much in line with the RM account, Swann (2012) notes that this confirmatory tendency "would have been advantageous" in "hunter-gatherer groups": once "people used input from the social environment to form self-views, self-verification strivings would have stabilized their identities and behavior, which in turn would make each individual more predictable to other group members" (26).

Similarly, in a study in which subjects received feedback about aspects of their self that can be relatively easily changed (e.g., their ability to estimate the weights of animals), Müller-Pinzler et al. (2019) found that "prior beliefs about the self modulate self-related belief-formation" in that subjects updated their performance estimates "in line with a confirmation bias": individuals with prior negative self-related beliefs (e.g., low self-esteem) showed increased biases towards factoring in negative (vs. positive) feedback, and, interestingly, this tendency was "modulated by the social context and only present when participants were exposed to a potentially judging audience" (ibid: 9–10). This coheres with the view that confirmation bias might serve the 'collective' to bring subjects into accordance with its social conceptions (positive or negative).

[5] Relatedly, neuroscientific data show that a positive view of one's own traits tends to correlate with a reduced activation of the right inferior prefrontal gyrus, which is the area of the brain processing self-related content, when the subject receives negative self-related information (Sharot et al. 2011). That is, optimists about themselves display a diminished sensitivity for negative information that is in tension with self-related trait optimism (ibid).

(b) *Other-Related Beliefs*

If confirmation bias was selected for sustaining social beliefs for the sake of reality matching then the bias should also be particularly pronounced when it comes to beliefs about *other* people especially in situations conducive to reality matching. For instance, powerful individuals have been found to be more likely to prompt subordinates to behaviorally confirm their social conceptions than relatively powerless subjects (Copeland 1994; Leyens et al. 1999). That is, interactions between powerful and powerless individuals are conducive to reality matching of the powerful individuals' social beliefs. According to the RM account, powerful individuals should display a stronger confirmation bias with respect to the relevant social beliefs. Goodwin et al. (2000) found just that: powerful people, in particular, tend to fail to take into account data that may contradict their social beliefs (capturing, e.g., stereotypes) about subordinates and attend more closely to information that supports their expectations. Relative to the powerless, powerful people displayed a stronger confirmation bias in their thinking about subordinates (ibid: 239f).

Similarly, if confirmation bias serves to facilitate social interaction by contributing to a match between beliefs and social reality then the bias should be increased with respect to trait attributions to other people in subjects who care about social interactions compared to other subjects. Dardenne and Leyens (1995) reasoned that when testing a hypothesis about the personality of another individual (e.g., their being introverted or extroverted), a preference for questions that match the hypothesis (e.g., that the subject is introverted) indicates social skill, conveying a feeling of being understood to the individual and contributing to a smooth conversation. Socially skilled people ('high self-monitors') should thus request 'matching questions', say, in an interview setting, for instance, when testing the introvert hypothesis, an interviewer could ask questions that are answered 'yes' by a typical introvert (e.g., 'Do you like to stay alone?'), confirming the presence of the hypothesized trait (ibid). Dardenne and Leyens did find that matching questions pertaining to an introvert or an extrovert hypothesis were selected most by high self-monitors: socially skilled subjects displayed a stronger confirmatory tendency than less socially skilled subjects (ibid).

Finally, there is also evidence that confirmation bias is more pronounced with respect to social beliefs compared to non-social beliefs. For instance, Marsh and Hanlon (2007) gave one group of behavioral ecologists a specific set of expectations with respect to sex differences in salamander behavior, while a second group was given the opposite set of expectations. In one experiment, subjects collected data on variable sets of live salamanders, while in the other experiment, observers collected data from identical videotaped trials. Across experiments and observed behaviors, the expectations of the observers biased their observations "only to a small or moderate degree", Marsh and Hanlon note, concluding that these "results are largely optimistic with respect to confirmation bias in behavioral ecology" (2007: 1089). This insignificant confirmation bias with respect to beliefs about non-social matters contrasts with findings of a significant confirmation bias with respect to beliefs about people (Talaifar and Swann 2017; Goodwin et al. 2000; Marks and Fraley 2006; Darley and Gross

1983), and, as I shall argue now, social affairs whose reality the subject can determine.

(c) *Non-personal, Social Beliefs*

One important kind of social beliefs are *political* beliefs, which concern social states of affairs pertaining to politics. Political beliefs are especially interesting in the context of the RM account because they are very closely related to reality matching. This is not only because subjects can often directly influence political affairs via voting, running as a candidate, campaigning, etc. It is also because subjects who are highly confident about their political beliefs are more likely to be able to convince other people of them too (Kappes et al. 2020). And the more widespread a political conviction in a population, the higher the probability that the population will adopt political structures that shape reality in line with it (Jost et al. 2003; Ordabayeva and Fernandes 2018).

If, as the RM account proposes, confirmation bias was selected for sustaining social beliefs for the sake of reality matching then the bias should be particularly strong when it comes to beliefs about political states of affairs. And indeed Taber and Lodge (2006) did find that "motivated [confirmation] biases come to the fore in the processing of political arguments", in particular, and, crucially, subjects "with weak […] [political] attitudes show less [confirmation] bias in processing political arguments" (767). In fact, in psychology, attitude strength, especially, in politically relevant domains of thinking has long been and still is widely accepted to increase the kind of selective exposure constitutive of confirmation bias (Knobloch-Westerwick et al. 2015: 173). For instance, Brannon et al. (2007) found that stronger, more extreme political attitudes are correlated with higher ratings of interest in attitude-consistent versus attitude-discrepant political articles. Similarly, Knobloch-Westerwick et al. (2015) found that people online who attach high importance to particular political topics spent more time on attitude-consistent messages than users who attached low importance to the topics, and "[a]ttitude-consistent messages […] were preferred", reinforcing the attitudes further (171). While this can contribute to political group polarization, such a polarization also boosts the group-wide reality-matching endeavour and can so be adaptive itself (Johnson and Fowler 2011: 317).

In short, then, while there are currently no systematic comparative studies on whether confirmation bias is more frequent or stronger with respect to social beliefs, related empirical studies do suggest that when it comes to (positive or negative) social beliefs about oneself, other people, and social states of affairs that the subject can determine (e.g., political beliefs), confirmation bias is both particularly common and pronounced. Empirical data thus corroborate some of the predictions of the RM account.

### 5.3 Explanatory Benefits

The theoretical and empirical considerations from the preceding sections offer support for the RM account. Before concluding, it is worth mentioning three further reasons for taking the account seriously. First, it has greater explanatory power than the three alternative views outlined above. Second, it is consistent with, and provides new contributions to, different areas of evolutionary theorizing on human cognition. And it casts new light on the epistemic character of confirmation bias. I'll now support these three points.

For instance, the argumentative-function account holds that confirmation bias is adaptive in making us better arguers. This was problematic because the bias hinders us in anticipating people's objections, which weakens our argumentative skill and increases the risk of us appearing incompetent in argumentative exchanges. The RM account avoids these problems: if confirmation bias was selected for reinforcing our preconceptions about people to promote SFPs then, since in one's own reasoning one only needs to justify one's beliefs to oneself, the first point one finds acceptable will suffice. To convince *others*, one would perhaps need to anticipate objections. But if the bias functions to boost primarily only one's own conviction about particular beliefs so as to facilitate SFPs then 'laziness' in critical thinking about one's own positions (Trouche et al. 2016) shouldn't be surprising.

Turning to the group-cognition account, the proposal was that confirmation bias is adaptive in and was selected for making group-level inquires more thorough, reliable, and efficient. In response, I noted that the concept of 'group selection' is problematic when it comes to traits threatening an individual's fitness (West et al. 2007; Pinker 2012), and that confirmation bias would arguably only lead to the group-level benefits at issue in groups with viewpoint diversity. Yet, it is doubtful that ancestral groups met this condition. The RM account is preferable to the group-cognition view because it doesn't rely on a notion of group selection but concerns primarily individual-level benefits, and it doesn't tie the adaptive effects of the bias to conditions of viewpoint diversity. It proposes instead that the adaptive SFP-related effects of the bias increase individuals' fitness (e.g., by facilitating their navigation of the social world, aligning them/others with their group's conceptions etc.) and can emerge whenever people hold beliefs about each other, interact, and fulfill social expectations. This condition is satisfied even in groups with viewpoint homogeneity.

The RM account also differs from the intention–alignment view, which holds that confirmation bias evolved for allowing us to synchronize intentions with others. One problem with this view was that the bias seems to hinder an intention alignment of individuals by weakening their perspective-taking capacity, and inclining them to ignore or distort people's objections. The RM account avoids this problem because it suggests that by disregarding objections or counterevidence to one's beliefs, one can remain convinced about them, which helps align social reality (not only, e.g., people's intentions) with them, producing the adaptive outcomes outlined above. The account can also explain why confirmation bias is particularly strong in groups in which shared ideologies are at stake (Taber and Lodge 2006; Gerken 2019). For subjects have a keen interest in reality corresponding to their ideological conceptions. Since the latter are shaping social reality via their impact on behavior

and are more effective in doing so the more convinced people are about them (Kappes et al. 2020), it is to be expected that when it comes to ideological propositions in like-minded groups, confirmation bias is more pronounced. And, as noted, the resulting group polarization itself can then be adaptive in strengthening the reality-matching process.

Moving beyond extant work on the evolution of confirmation bias, the RM account also contributes to and raises new questions for other areas of research in different disciplines. It, for instance, yields predictions that psychologists can experimentally explore in comparative studies such as the prediction that confirmation bias is more common and stronger when targeting social versus non-social beliefs, or when conditions are conducive to reality matching as opposed to when they are not. The account also adds a new perspective to research on SFPs and on how social conceptions interact with their targets (Hacking 1995; Snyder and Klein 2007; Jussim 2017). Relatedly, the RM account also contributes to recent philosophical work on, *folk-psychology*, i.e., our ability to ascribe mental states to agents to make sense of their behavior. In that work, some philosophers argue that folk-psychology serves "mindshaping", that is, the moulding of people's behavior and minds so that they fit our conceptions, making people more predictable and cooperation with them easier (Mameli 2001; Zawidzki 2013; Peters 2019b). There are clear connections between the mindshaping view of folk psychology and the RM account, but also important differences. For instance, the RM account pertains to the function of confirmation bias, not folk psychology. Moreover, advocates of the mindshaping view have so far left the conditions for effective mindshaping via folk-psychological ascriptions and the possible role of confirmation bias in it unexplored. The RM account begins to fill this gap in the research and in doing so adds to work on the question of how epistemic (or 'mindreading') and non-epistemic (or 'mindshaping', e.g., motivational) processes are related in folk-psychology (Peters 2019b: 545f; Westra 2020; Fernández-Castro and Martínez-Manrique 2020).

In addition to offering contributions to a range of different areas of research, the RM account also casts new light on the epistemic character of confirmation bias. Capturing the currently common view on the matter, Mercier (2016) writes that "piling up reasons that support our preconceived views is not the best way to correct them. […] [It] stop[s] people from fixing mistaken beliefs" (110). The RM account offers a different perspective, suggesting that when it is directed at beliefs about social affairs, confirmation bias does often help subjects correct their mistaken conceptions to the extent that it contributes to SFPs of them. Similarly, Dutilh Novaes (2018) holds that the bias involves or contributes to a failure of perspective taking, and so, "given the importance of being able to appreciate one's interlocutor's perspective for social interaction", is "best not seen as an adaptation" (520). The RM account, on the other hand, proposes that the bias often facilitates social understanding: in making us less sensitive to our interlocutor's opposing perspective, it helps us remain confident about our social beliefs, which increases the probability of SFPs that in turn make people more predictable and mindreadable.

## 6 Conclusion

After outlining limitations of three recent proposals on the evolution of confirmation bias, I developed and supported a novel alternative, the reality-matching (RM) account, which holds that one of the adaptive features for which the bias evolved is that it helps us bring social reality into alignment with our beliefs. When the bias targets positive social beliefs, this serves both the subject and the group, assisting them in their development (to become, e.g., more communicative or knowledgeable) while also making their social cognition more effective and tractable. When it targets negative social beliefs, in promoting reality matching, the bias might contribute to ethically problematic outcomes, but it can then still support social conformity and predictability, which were perhaps especially in ancestral tribal groups adaptive. While the socially constructive aspect of confirmation bias highlighted here may not be the main or only feature of the bias that led to its evolution, it is one that has so far been overlooked in the evolutionary theorizing on confirmation bias. If we attend to it, an account of the function of confirmation bias becomes available that coheres with data from across the psychological sciences, manages to avoid many of the shortcomings of competitor views, and has explanatory benefits that help advance the research on the function, nature, and epistemic character of the bias.

**Acknowledgements** Many thanks to Andreas De Block, Mikkel Gerken, and Alex Krauss for comments on earlier drafts. The research for this paper was partly funded by the Danmarks Frie Forskningsfond Grant no: 8018-00053B allocated to Mikkel Gerken.



## References

Alfano, M. (2013). *Character as moral fiction*. Cambridge: CUP.

Bacharach, M., Guerra, G., & Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision, 63,* 349–388.

Ball, P. (2017). The trouble with scientists. How one psychologist is tackling human biases in science. *Nautilus*. Retrieved May 2, 2019 from http://nautil.us/issue/54/the-unspoken/the-trouble-with-scientists-rp.

Biggs, M. (2009). Self-fulfilling prophecies. In P. Bearman & P. Hedstrom (Eds.), *The Oxford handbook of analytical sociology* (pp. 294–314). Oxford: OUP.

Brannon, L. A., Tagler, M. J., & Eagly, A. H. (2007). The moderating role of attitude strength in selective exposure to information. *Journal of Experimental Social Psychology, 43,* 611–617.

Copeland, J. (1994). Prophecies of power: Motivational implications of social power for behavioral confirmation. *Journal of Personality and Social Psychology, 67,* 264–277.

Cornelissen, G., Dewitte, S., & Warlop, L. (2007). Whatever people say I am that's what I am: Social labeling as a social marketing tool. *International Journal of Research in Marketing, 24*(4), 278–288.
Dardenne, B., & Leyens, J. (1995). Confirmation bias as a social skill. *Personality and Social Psychology Bulletin, 21*(11), 1229–1239.
Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20–33.
Davidson, O. B., & Eden, D. (2000). Remedial self-fulfilling prophecy: Two field experiments to prevent Golem effects among disadvantaged women. *Journal of Applied Psychology, 85*(3), 386–398.
De Bruine, L. M. (2009). Beyond 'just-so stories': How evolutionary theories led to predictions that non-evolution-minded researchers would never dream of. *Psychologist, 22*(11), 930–933.
De Cruz, H., & De Smedt, J. (2016). How do philosophers evaluate natural theological arguments? An experimental philosophical investigation. In H. De Cruz & R. Nichols (Eds.), *Advances in religion, cognitive science, and experimental philosophy* (pp. 119–142). New York: Bloomsbury.
Downey, G., Freitas, A. L., Michaelis, B., & Khouri, H. (1998). The self-fulfilling prophecy in close relationships: Rejection sensitivity and rejection by romantic partners. *Journal of Personality and Social Psychology, 75,* 545–560.
Draper, P., & Nichols, R. (2013). Diagnosing bias in philosophy of religion. *The Monist, 96,* 420–446.
Dutilh Novaes, C. (2018). The enduring enigma of reason. *Mind and Language, 33,* 513–524.
Evans, J. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87,* 223–240.
Fernández-Castro, V., & Martínez-Manrique, F. (2020). Shaping your own mind: The self-mindshaping view on metacognition. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-020-09658-2.
Gerken, M. (2019). Public scientific testimony in the scientific image. *Studies in History and Philosophy of Science Part A*. https://doi.org/10.1016/j.shpsa.2019.05.006.
Golec de Zavala, A. (2011). Collective narcissism and intergroup hostility: The dark side of 'in-group love'. *Social and Personality Psychology Compass, 5,* 309–320.
Goodwin, S., Gubin, A., Fiske, S., & Yzerbyt, V. (2000). Power can bias impression formation: Stereotyping subordinates by default and by design. *Group Processes and Intergroup Relations, 3,* 227–256.
Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B, 205*(1161), 581–598.
Grusec, J., Kuczynski, L., Rushton, J., & Simutis, Z. (1978). Modeling, direct instruction, and attributions: Effects on altruism. *Developmental Psychology, 14,* 51–57.
Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, et al. (Eds.), *Causal cognition* (pp. 351–383). New York: Clarendon Press.
Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In H. R. Brian (Ed.), *Psychology of learning and motivation* (pp. 41–102). New York: Academic Press.
Haidt, J. (2012). *The righteous mind*. New York: Pantheon.
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), e45457.
Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10,* 47–66.
Henrich, J. (2016). *The secret of our success*. Princeton, NJ: Princeton University Press.
Hernandez, I., & Preston, J. L. (2013). Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology, 49*(1), 178–182.
Jensen, R. E., & Moore, S. G. (1977). The effect of attribute statements on cooperativeness and competitiveness in school-age boys. *Child Development, 48*(1), 305–307.
Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution, 28,* 474–481.
Johnson, D. D. P., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature, 477,* 317–320.
Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129*(3), 339–375.
Jussim, L. (2012). *Social perception and social reality*. Oxford: OUP.

Jussim, L. (2017). Précis of social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy. *Behavioral and Brain Sciences, 40,* 1–20.
Kappes, A., Harvey, A. H., Lohrenz, T., et al. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience, 23,* 130–137.
Kelly, D. (2013). Moral disgust and the tribal instincts hypothesis. In K. Sterelny, R. Joyce, B. Calcott, & B. Fraser (Eds.), *Cooperation and its evolution* (pp. 503–524). Cambridge, MA: The MIT Press.
Kelly, D., & Hoburg, P. (2017). A tale of two processes: On Joseph Henrich's the secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter. *Philosophical Psychology, 30*(6), 832–848.
Ketelaar, T., & Ellis, B. J. (2000). Are evolutionary explanations unfalsifiable? Evolutionary psychology and the Lakatosian philosophy of science. *Psychological Inquiry, 11*(1), 1–21.
Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation, 32,* 385–418.
Knobloch-Westerwick, S., Johnson, B. K., & Westerwick, A. (2015). Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. *Journal of Computer-Mediated Communication, 20,* 171–187.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498.
Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin, 36*(8), 1115–1129.
Levy, N. (2019). Due deference to denialism: Explaining ordinary people's rejection of established scientific findings. *Synthese, 196*(1), 313–327.
Leyens, J., Dardenne, B., Yzerbyt, V., Scaillet, N., & Snyder, M. (1999). Confirmation and disconfirmation: Their social advantages. *European Review of Social Psychology, 10*(1), 199–230.
Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science, 12*(4), 660–664.
Lord, C., Lepper, M., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231–1243.
Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology, 72,* 791–809.
Madon, S., Jussim, L., Guyll, M., Nofziger, H., Salib, E. R., Willard, J., et al. (2018). The accumulation of stereotype-based self-fulfilling prophecies. *Journal of Personality and Social Psychology, 115*(5), 825–844.
Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy, 16,* 597–628.
Marks, M. J., & Fraley, R. C. (2006). Confirmation bias and the sexual double standard. *Sex Roles: A Journal of Research, 54*(1–2), 19–26.
Marsh, D. M., & Hanlon, T. J. (2007). Seeing what we want to see: Confirmation biasin animal behavior research. *Ethology, 113,* 1089–1098.
Matheson, J., & Vitz, R. (Eds.). (2014). *The ethics of belief: Individual and social*. Oxford: OUP.
Mayo, R., Alfasi, D., & Schwarz, N. (2014). Distrust and the positive test heuristic: Dispositional and situated social distrust improves performance on the Wason Rule Discovery Task. *Journal of Experimental Psychology: General, 143*(3), 985–990.
McDonald, M. M., Navarrete, C. D., & van Vugt, M. (2012). Evolution and the psychology of intergroup conflict: The male warrior hypothesis. *Philosophical Transactions of the Royal Society, B, 367,* 670–679.
Mercier, H. (2016). Confirmation (or myside) bias. In R. Pohl (Ed.), *Cognitive illusions* (pp. 99–114). London: Psychology Press.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–111.
Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, MA: Harvard University Press.
Merton, R. (1948). The self-fulfilling prophecy. *The Antioch Review, 8*(2), 193–210.
Miller, R., Brickman, P., & Bolen, D. (1975). Attribution versus persuasion as a means for modifying behavior. *Journal of Personality and Social Psychology, 31*(3), 430–441.
Millikan, R. G. (1984). *Language thought and other biological categories*. Cambridge, MA: MIT Press.
Müller-Pinzler, L., Czekalla, N., Mayer, A. V., et al. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports, 9,* 14416. https://doi.org/10.1038/s41598-019-50821-w.
Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996). The self-fulfilling nature of positive illusions in romantic relationships: Love is not blind, but prescient. *Journal of Personality and Social Psychology*, *71*, 1155–1180.

Myers, D., & DeWall, N. (2015). *Psychology*. New York: Worth Publishers.
Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83,* 602–627.
Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220.
Norman, A. (2016). Why we reason: Intention–alignment and the genesis of human rationality. *Biology and Philosophy, 31,* 685–704.
Ordabayeva, N., & Fernandes, D. (2018). Better or different? How political ideology shapes preferences for differentiation in the social hierarchy. *Journal of Consumer Research, 45*(2), 227–250.
Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology, 13*(8), e1005684.
Pelham, B. W., & Swann, W. B. (1994). The juncture of intrapersonal and interpersonal knowledge: Self-certainty and interpersonal congruence. *Personality and Social Psychology Bulletin, 20*(4), 349–357.
Peters, U. (2018). Illegitimate values, confirmation bias, and mandevillian cognition in science. *British Journal for Philosophy of Science*. https://doi.org/10.1093/bjps/axy079.
Peters, U. (2019a). Implicit bias, ideological bias, and epistemic risks in philosophy. *Mind & Language*, *34*, 393–419. https://doi.org/10.1111/mila.12194.
Peters, U. (2019b). The complementarity of mindshaping and mindreading. *Phenomenology and the Cognitive Sciences*, *18*(3), 533–549.
Peters, U., Honeycutt, N., De Block, A., & Jussim, L. (forthcoming). Ideological diversity, hostility, and discrimination in philosophy. *Philosophical Psychology*. Available online: https://philpapers.org/archive/PETIDH-2.pdf.
Pinker, S. (2012). The false allure of group selection. Retrieved July 20, 2012 from http://edge.org/conversation/the-false-allure-of-group-selection.
Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics, 114*(1), 37–82.
Richerson, P., & Boyd, R. (2001). The evolution of subjective commitment to groups: A tribal instincts hypothesis. In R. M. Nesse (Ed.), *Evolution and the capacity for commitment* (pp. 186–202). New York: Russell Sage Found.
Richerson, P., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.
Roberts, S. C., van Vugt, M., & Dunbar, R. I. M. (2012). Evolutionary psychology in the modern world: Applications, perspectives, and strategies. *Evolutionary Psychology, 10,* 762–769.
Schuck, P. H. (2001). The perceived values of diversity, then and now. *Cardozo Law Review, 22,* 1915–1960.
Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience, 14,* 1475–1479.
Simpson, J. A., & Beckes, L. (2010). Evolutionary perspectives on prosocial behavior. In M. Mikulincer & P. Shaver (Eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature* (pp. 35–53). Washington, DC: American Psychological Association.
Smart, P. (2018). Mandevillian intellingence. *Synthese, 195,* 4169–4200.
Smith, J. J., & Wald, B. (2019). Collectivized intellectualism. *Res Philosophica, 96*(2), 199–227.
Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge–advisor system. *Organizational Behavior and Human Decision Processes, 84,* 288–307.
Snyder, M., & Klein, O. (2007). Construing and constructing others: On the reality and the generality of the behavioral confirmation scenario. In P. Hauf & F. Forsterling (Eds.), *Making minds* (pp. 47–60). John Benjamins: Amsterdam/Philadelphia.
Solomon, G. B. (2016). Improving performance by means of action–cognition coupling in athletes and coaches. In M. Raab, B. Lobinger, S. Hoffman, A. Pizzera, & S. Laborde (Eds.), *Performance psychology: Perception, action, cognition, and emotion* (pp. 88–101). London, England: Elsevier Academic Press.
Stangor, C. (2011). *Principles of social psychology*. Victoria, BC: BCcampus.
Stanovich, K., West, R., & Toplak, M. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science, 22,* 259–264.
Steel, D. (2018). Wishful thinking and values in science: Bias and beliefs about injustice. *Philosophy of Science*. https://doi.org/10.1086/699714.

Sterelny, K. (2006). Memes revisited. *British Journal for the Philosophy of Science, 57,* 145–165.
Sterelny, K. (2007). Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society B, 362,* 719–730.
Sterelny, K. (2018). Why reason? Hugo Mercier's and Dan Sperber's the enigma of reason: A new theory of human understanding. *Mind and Language, 33*(5), 502–512.
Stibel, J. (2018). Fake news: How our brains lead us into echo chambers that promote racism and sexism. *USA Today*. Retrieved October 8, 2018 from https://eu.usatoday.com/story/money/columnist/2018/05/15/fake-news-social-media-confirmation-bias-echo-chambers/533857002/.
Swann, W. B. (1983). Self-verification: Bringing social reality into harmony with the self. In J. Suls & A. G. Greenwald (Eds.), *Social psychological perspectives on the self* (Vol. 2, pp. 33–66). London: Erlbaum.
Swann, W. B., Jr. (2012). Self-verification theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 23–42). Beverley Hills, CA: Sage Publications Ltd.
Swann, W., & Ely, R. (1984). A battle of wills: Self-verification versus behavioral confirmation. *Journal of Personality and Social Psychology, 46,* 1287–1302.
Swann, W. B., Jr., Stein-Seroussi, A., & Giesler, B. (1992). Why people self-verify. *Journal of Personality and Social Psychology, 62,* 392–406.
Taber, C., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50,* 755–769.
Talaifar, S., & Swann, W. B. (2017). Self-verification theory. In L. Goossens, M. Maes, S. Danneel, J. Vanhalst, & S. Nelemans (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–9). Berlin: Springer.
Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology, 44,* 187–194.
Tooby, J., & Cosmides, L. (2015). The theoretical foundations of evolutionary psychology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 3–87). Hoboken, NJ: Wiley.
Tormala, Z. L. (2016). The role of certainty (and uncertainty) in attitudes and persuasion. *Current Opinion in Psychology, 10,* 6–11.
Trouche, E., et al. (2016). The selective laziness of reasoning. *Cognitive Science, 40,* 2122–2136.
Turnwald, B., et al. (2018). Learning one's genetic risk changes physiology independent of actual genetic risk. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-018-0483-4.
von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34*(1), 1–16.
Wenger, A., & Fowers, B. J. (2008). Positive illusions in parenting: Every child is above average. *Journal of Applied Social Psychology, 38*(3), 611–634.
West, S. A., Griffin, A. S., & Gardiner, A. (2007). Social semantics: How useful has group selection been? *Journal of Evolutionary Biology*, *21*, 374–385.
Westra, E. (2020). Folk personality psychology: Mindreading and mindshaping in trait attribution. *Synthese*. https://doi.org/10.1007/s11229-020-02566-7.
Whitaker, R. M., Colombo, G. B., & Rand, D. G. (2018). Indirect reciprocity and the evolution of prejudicial groups. *Scientific Reports*, *8*(1), 13247. https://doi.org/10.1038/s41598-018-31363-z.
Whittlestone, J. (2017). *The importance of making assumptions: Why confirmation is not necessarily a bias*. Ph.D. Thesis. Coventry: University of Warwick.
Willard, J., & Madon, S. (2016). Understanding the connections between self-fulfilling prophecies and social problems. In S. Trusz & P. Przemysław Bąbel (Eds.), *Interpersonal and intrapersonal expectancies* (pp. 117–125). London: Routledge.
Willard, J., Madon, S., Guyll, M., Spoth, R., & Jussim, L. (2008). Self-efficacy as a moderator of negative and positive self-fulfilling prophecy effects: Mothers' beliefs and children's alcohol use. *European Journal of Social Psychology, 38,* 499–520.
Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology, 10,* 109–120.
Zawidzki, T. (2008). The function of folk psychology: Mind reading or mind shaping? *Philosophical Explorations, 11*(3), 193–210.
Zawidzki, T. (2013). Mindshaping: A new framework for understanding human social cognition. Cambridge: MIT Press.
Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass, 2,* 1497–1517.