

Socially adaptive belief

Daniel Williams^{1,2} 

¹Corpus Christi College, University of Cambridge, Cambridge, UK

²Centre for Philosophical Psychology, University of Antwerp, Antwerpen, Belgium

Correspondence

Daniel Williams, Corpus Christi College, Trumpington Street, Cambridge CB2 1RH, UK.
Email: dw473@cam.ac.uk

Funding information

Fonds Wetenschappelijk Onderzoek, Grant/Award Number: 7032; Corpus Christi College, University of Cambridge; Fonds voor Wetenschappelijk Onderzoek (FWO), Grant/Award Numbers: G0C7416N, G.0020.12N

I clarify and defend the hypothesis that human belief formation is sensitive to social rewards and punishments, such that beliefs are sometimes formed based on unconscious expectations of their likely effects on other agents – agents who frequently reward us when we hold ungrounded beliefs and punish us when we hold reasonable ones. After clarifying this phenomenon and distinguishing it from other sources of bias in the psychological literature, I argue that the hypothesis is plausible on theoretical grounds and I show how it illuminates and unifies a range of psychological phenomena, including confabulation and rationalisation, positive illusions, and identity-protective cognition.

KEYWORDS

belief, bias, irrationality, motivated cognition, self-deception, social cognition

1 | INTRODUCTION

Many animals navigate their environments with the use of internal representations. For such animals, it is plausible to think that the utility of these internal representations is dependent on their accuracy. For rats making their way back to their nests, it pays to exploit cognitive maps that accurately represent the spatial layout of their environments. For chimpanzees living complex social lives, it pays to accurately represent the relative positions of other chimpanzees in the local dominance hierarchy. And for humans choosing where to go on holiday, it pays to command a whole body of accurate information: about the relative prices and climates of different destinations, the travel time, which period in the year different destinations are busiest, and so on.

From this perspective, human beliefs fall into place as a highly complex species of a broader genus: internal representations whose job is to provide accurate information that an agent can exploit in guiding his or her inferences and actions (Fodor, 1975; Millikan, 1984; Papineau, 1984). In a popular metaphor, beliefs thus function as maps by which we steer (see Ramsey, 1990, p.

146). Like maps, their practical value is dependent on their accuracy: If your map of London misrepresents its spatial layout, you will get lost; if your beliefs are false, you will not be able to satisfy your desires or achieve your goals. Given this, it is natural to conclude – and many psychologists and philosophers have concluded – that, ‘the proper function of cognition is ... the fixation of true beliefs’ (Fodor, 2001, p. 68; see also Millikan, 1984).

In this article, I argue that this perspective on beliefs and belief formation neglects an important feature of human social life: In our species, our beliefs are objects of intense social scrutiny. Because other agents have reliable access to what we believe and frequently reward us when we hold ungrounded beliefs and punish us when we hold reasonable ones, this creates powerful incentives for otherwise rational individuals to form beliefs in ways that are sensitive to such social rewards and punishments. I contend that we frequently capitulate to such incentives, such that the way in which we form beliefs is highly sensitive to the actual or anticipated effects of candidate beliefs on other agents. Given this, many of our systematic departures from epistemic rationality are driven not by *irrationality* or the use of cost-effective heuristics but rather by well-calibrated rational self-interest: When the demands of truth conflict with social expedience, it can be practically advantageous to jettison the former in favour of the latter.

Importantly, the hypothesis that belief formation is sensitive to social rewards and punishments is not original. In fact, variations on it appear in a multitude of disparate theoretical contexts and projects, many of which I will touch on below.¹ Nevertheless, to the best of my knowledge, the hypothesis has never been clearly explicated or explored in the manner that I will do here. Furthermore, many of its manifestations in the scientific literature either focus on narrow examples of the more general phenomenon or else burden the hypothesis with additional commitments that are more controversial and less plausible. For these reasons, perhaps, it has been largely ignored throughout much of contemporary philosophy, despite its many important philosophical implications.² My central aim in this article is to rectify this state of affairs. The hypothesis that human belief formation is sensitive to social rewards and punishments warrants much greater attention in philosophy than it currently enjoys.

I structure the article as follows. In Section 2, I relate the phenomenon of socially adaptive belief formation to other sources of epistemic irrationality identified in the psychological and philosophical literature, and I argue that it consists of a highly neglected example of motivated cognition. In Section 3, I argue that the hypothesis is plausible on theoretical grounds: In a species with substantial social scrutiny of beliefs, forming beliefs in a way that is sensitive to their effects on other agents leads to practical success. I also clarify the hypothesis and address several likely confusions and objections. In Section 4, I identify plausible manifestations of socially adaptive belief formation, focusing especially on confabulation and rationalisation (Section 4.1), positive illusions (Section 4.2), and identity protective cognition (Section 4.3). I conclude in Section 5 by identifying important areas for future research.

¹This includes Trivers' hypothesis that self-deception is socially strategic (Pinker, 2005; Trivers, 2000), the hypothesis that beliefs function as ‘signals’ or ‘displays’ (Funkhouser, 2017; Simler & Hanson, 2017; Sterelny, 2015), the hypothesis that reasoning evolved to facilitate social functions of persuasion and reputation management (Haidt, 2013; Mercier & Sperber, 2011, 2017), and research in social science on the relationship between beliefs and group identity (especially Kahan, 2013, 2017a). I have been especially influenced by Kurzban (2012), Simler and Hanson (2017), and Trivers (2011).

²Recent counterexamples to this claim include Sterelny's (2015) work on beliefs as ‘displays’ and Funkhouser's (2017) hypothesis that beliefs function as signals (see Section 3.1) (see also Williams, 2018).

2 | SOURCES OF EPISTEMIC IRRATIONALITY

When it comes to rational agency, there is a traditional distinction between epistemic and pragmatic criteria for success (Harman, 2004). As Bortolotti (2015) puts it, '[t]he rational agent is either the one *who gets things right* (e.g., the one with true beliefs), or the one *for whom things go well* (e.g., the one with beliefs that are conducive to happiness or flourishing)' (p. 7). Nevertheless, it is both intuitive and widely held that these two forms of rationality are closely connected, such that an agent's practical success is dependent on epistemic rationality. One form of this argument, for example, stresses the dependence of *instrumental rationality* on epistemic rationality: Because an individual will be better able to satisfy her desires if her beliefs are true than if they are false, it is instrumentally rational to form and evaluate beliefs in a way that is most likely to lead to true beliefs (Cowie, 2014). A similar argument focuses on *evolutionary success* and 'the prevailing assumption ... that beliefs that maximize the survival of the believer will be those that best approximate reality' (McKay & Dennett, 2009, p. 493). Because epistemic rationality describes the norms conducive to forming true beliefs, many cognitive scientists therefore assume that a useful starting point for modelling human cognition is that it conforms to norms of logic (Fodor, 1975) or probability theory (Anderson, 1990).

Notoriously, these perspectives on the relationship between practical success and epistemic rationality generate a puzzle. Recent decades of psychological research have generated an enormous body of experimental results documenting systematic deviations from epistemic rationality in human cognition (Kahneman, 2011; Mercier & Sperber, 2011). As Bortolotti (2015) puts it, '[p]sychologists have repeatedly and convincingly argued that people make basic mistakes in deductive and inductive reasoning and violate basic rules of statistics, probability theory and decision theory' (p. 67). Given the foregoing perspective, these and other systematic deviations from epistemic rationality can seem strange (Mercier & Sperber, 2011; Trivers, 2011). Why would we frequently thwart our practical success by reasoning in irrational ways? And why would evolution have given rise to a systematically irrational organism?

My aim in this article is to address a part of this puzzle. Specifically, I will argue that distinctive characteristics of human social life routinely put practical success into direct conflict with epistemic rationality. This is by no means the only source of epistemic irrationality in our species, however. Given this, it will be useful to briefly relate the phenomenon of socially adaptive belief formation that I intend to describe in later sections to other more well-known explanations of biased belief formation. As will be clear, these explanations are neither exhaustive nor mutually exclusive, but they have been the focus of the most influential research on epistemic irrationality in the psychological and philosophical literature. Although it is something of a simplification, these explanations can be divided into those that appeal to *constraints* of various kinds and those that appeal to *motivational influences* on belief formation.

2.1 | Constraints

The most influential explanation of epistemic irrationality in the psychological literature points to constraints on *time*, *resources*, and *computational power* in human cognition (Gigerenzer & Selten, 2002; Kahneman, 2003). As it is often put, our rationality is not absolute but *bounded*; we *satisfice* rather than optimise (Simon, 1956). Such practical constraints on cognition entail that we must often rely on 'fast and frugal' heuristics and approximations that sacrifice

reliability across all contexts for reliability across most important contexts at a reduced processing cost (Gigerenzer & Selten, 2002).

A related but distinct explanation of epistemic irrationality points to constraints on the evolutionary process. Due to phenomena such as evolutionary inertia and the problem of local maxima in optimization procedures such as natural selection, for example, evolution does not reliably give rise to optimal systems, even once one factors in resource limitations. Just as the injury prone human spinal column is arguably not the optimal solution for supporting the weight of a biped but rather a 'kluge' arrived at via tinkering with previous designs, many aspects of human cognition might be sub-optimal for similar reasons (Marcus, 2009). Relatedly, evolution lacks foresight. Evolved systems adapted to one environment can become suboptimal when environments rapidly change. For example, proponents of evolutionary psychology have argued that the human susceptibility to the 'gambler's fallacy', in which individuals treat independent trials of a random process as dependent, reflects the fact that truly random processes would have been absent in our ancestral environment (Pinker, 1999).

Finally, some constraints are intrinsic to certain tasks (Stich, 1990). For example, there are some tasks in which the only way to reliably reduce one kind of error is to be biased towards another kind of error. If these errors are associated with different costs, this can lead to outcomes in which biased systems are more adaptive than unbiased ones. To take the most famous example, false positives and false negatives in the task of predator detection both constitute errors, but 'the cost of getting killed even once is enormously higher than the cost of responding to a hundred false alarms' (Nesse & Williams, 1995, p. 213). Under such conditions, evolution can therefore favour systems biased towards the less costly error over unbiased systems if the former systems are better at avoiding the costlier error.

2.2 | Motivational influences

A second major source of epistemic irrationality involves *motivational influences* on belief formation (Kunda, 1990; Sharot & Garrett, 2016). Such motivational influences arise when individuals sample or process information to arrive at conclusions that they *want* to arrive at for reasons independent of their truth, and they are encapsulated in expressions of contemporary folk psychology such as 'wishful thinking', 'denial', 'burying your head in the sand', and 'drinking your own Kool-Aid'.

Motivational influences on belief formation are standardly understood as forms of *motivated cognition* in contemporary psychology (Kunda, 1990) and in terms of *belief-based utility* in the social sciences (Loewenstein & Molnar, 2018). These two ideas relate to one another directly, however. 'Belief-based utility' refers to the fact that beliefs do not merely inform our efforts to satisfy our preferences; they are important targets of our preferences. This therefore creates an incentive to sample and process information in ways intended to satisfy these belief-based preferences – to arrive at beliefs that one wants to arrive at and avoid forming beliefs that one wants to avoid. When individuals capitulate to this incentive, they engage in motivated cognition.

Why would individuals assign value to beliefs for reasons independent of their truth? The simplest answer is that beliefs generate what I will call 'non-epistemic effects'. Epistemic effects can be thought of as the purely content-based effects that enable beliefs to inform us concerning the state of the world. As McKay and Dennett (2009) put it, however, '[b]elief states have complex effects *beyond simply informing our deliberations*' (p. 508; my emphasis). For example, beliefs can have a powerful emotional impact, rendering us happy, proud, ashamed, depressed,

and so on. When individuals engage in motivated cognition, they consciously or unconsciously factor in the relative value of these non-epistemic effects into the way in which they seek out and process information.

In most accounts of motivated cognition, the relevant non-epistemic effects associated with beliefs arise from the individual's attitudes towards the states of affairs represented by those beliefs, such that the motive to believe that *P* arises from the desire that *P*. Importantly, however, the concepts of belief-based utility and motivated cognition are more general than this, and apply whenever an individual is motivated to form (or avoid forming) certain beliefs for reasons independent of their truth (or falsity) (Bénabou & Tirole, 2016). In this article, I therefore want to focus on a very different form of motivated cognition – one that arises when the motivation to form, or avoid forming, certain beliefs is driven by their actual or anticipated effects *on other agents*. Specifically, I aim to defend the following hypothesis:

Socially adaptive belief (henceforth SAB): *Belief formation is sensitive to social rewards and punishments.*

It is worth briefly unpacking the various elements of this claim. By 'social rewards and punishments', I mean social outcomes that we strive to bring about or avoid, where this includes not just our conscious social goals but also unconscious social motives and the social outcomes that psychological mechanisms are adapted to bring about. By 'sensitivity', I mean that such social rewards and punishments causally influence the way in which we sample and process information. An important question is how this causal influence works (see Section 5). My principal aim in this article is to substantiate the claim *that* socially adaptive belief formation occurs, however, not to explore *how* it occurs. Finally, what do I mean by 'belief' in this context? Of course, there is nothing like a philosophical consensus concerning the nature of belief. For much of this article, I will therefore assume a maximally general account, consistent with how the term is understood throughout much of contemporary psychology and philosophy: 'A belief is a functional state of an organism that implements or embodies that organism's endorsement of a particular state of affairs as actual' (McKay & Dennett, 2009, p. 493). I return to this issue in more depth in Section 5, however.

Insofar as motivated cognition exists in general, there seems to be no *in-principle* reason for denying the possibility of SAB. Nevertheless, is there any positive reason for endorsing its truth? I will now advance two complementary lines of argument in its defence. First, in Section 3, I will argue that the hypothesis is plausible on theoretical grounds: Given distinctive characteristics of human social life, forming beliefs in a way that is sensitive to social rewards and punishments leads to practical success. I will also clarify the hypothesis and address several likely objections and confusions. In Section 4, I will then focus on several plausible examples of socially adaptive belief formation as identified in the psychological and social sciences.

3 | SOCIALLY ADAPTIVE BELIEF AND PRACTICAL SUCCESS

In this section, I argue that SAB is theoretically plausible on the following grounds: Given distinctive characteristics of human social life, forming beliefs in a way that is sensitive to their effects on other agents leads to practical success. I will not commit to a specific account of what 'practical success' means in this context – for example, whether it should be understood as

desire-satisfaction, physical or psychological wellbeing, or evolutionary success – because I think that the argument will likely go through on any of these interpretations.

There are three characteristics of human social life that undermine the connection between true belief and practical success. First, other agents have reliable access to what we believe. Although there are substantial disagreements concerning the mechanisms that underlie this capacity of mindreading and its presence and sophistication in other species, there is no doubt that we *can* mindread with a scope and flexibility far beyond that of other animals. Of course, this capacity is not infallible. We are often wrong about what others believe. Sometimes they consciously deceive us. At other times, we are simply mistaken. In general, however, we are highly successful at attributing beliefs to other agents.

Second, other agents do not merely attribute beliefs to us. They *care* what we believe, responding to us differently as a consequence of which beliefs they attribute to us in ways that can have dramatic effects on our wellbeing. In general, our practical success is highly dependent on the impression that we make on others, and the beliefs that we hold are highly relevant to this impression, providing information to other individuals about our traits, motivations, trustworthiness, loyalty, and ethical, social, and legal obligations. The judgements that other agents form concerning such phenomena influence their assessments of us in myriad ways.

Finally, not only do the social effects of our beliefs have dramatic consequences for our wellbeing, but the kinds of beliefs that cause desirable social effects in the complex social environments that we inhabit are different from the kinds of beliefs that we would have a practical incentive to form in the absence of such social effects. Specifically, ungrounded beliefs frequently elicit desirable responses from other agents and grounded beliefs frequently elicit undesirable ones. The most obvious example of this phenomenon involves cases in which individuals are ostracised or even murdered for failure to believe in the religious and political myths of their surrounding communities, but I will return in Section 4 to several other examples.

Given these considerations, I propose that the following claim is *prima facie* plausible: Forming beliefs in a way that is sensitive to social rewards and punishments leads to practical success, such that individuals who form beliefs in a way that is sensitive to their effects on other agents will on average achieve greater practical success than individuals who do not. Of course, this claim clearly needs careful handling. It would obviously not deliver practical success to wholly capitulate to social incentives when one forms beliefs, for example. Whether factoring social incentives into the process of belief formation is conducive to practical success is dependent on the relative costs and benefits of doing so, which are themselves highly dependent on the social context and contents of the relevant beliefs.

The *benefits* of socially adaptive belief formation increase in proportion to the degree of social scrutiny of beliefs. That is, it is only under conditions in which other agents care what one believes that one has any incentive to factor them into the way in which one seeks out and processes information. The *costs* of socially adaptive belief formation can be understood in terms of the practical costs associated with deviating from how one would form beliefs in the absence of their social effects. Insofar as other agents reward epistemic irrationality, these costs in effect reduce to the potential costs associated with holding false or ungrounded beliefs. Crucially, however, such costs are themselves highly variable: Although in most cases such beliefs frustrate one's ability to satisfy one's desires, there are some cases in which they do not, either because one is unlikely to ever act on the beliefs over and above asserting one's commitment to them or because they concern phenomena that one has little ability to influence. It has long been noted that individuals are more likely to be swayed by motivated cognition and other

biases in such cases. As Bénabou and Tirole (2016) put it, '[b]eliefs for which the individual cost of being wrong is small are more likely to be distorted by emotions, desires, and goals' (p. 150).

Given the plausible assumption that there are conditions in which the practical benefits from socially adaptive belief formation outweigh the costs, then, an individual could improve her practical success by factoring social incentives into the way in which she samples and processes information. This, I think, should at least make one take SAB seriously as a hypothesis. Of course, it should *only* make one take it seriously. So far, I have not offered any positive evidence for the existence of socially adaptive belief formation. I take up this challenge in Section 4. First, however, it will be useful to clarify the hypothesis and address several likely confusions and objections.

3.1 | Clarifications and objections

First, one might object that *pretending* to form socially adaptive beliefs would be more practically advantageous than actually forming them.³ Such deception, after all, seems to deliver the best of both worlds: One reaps all the benefits of socially adaptive belief without incurring any of the potential costs associated with genuinely forming and thus potentially acting upon those beliefs. People already have the resources to intentionally deceive and there is some evidence suggesting that people are in fact bad at detecting deception (Bond & DePaulo, 2006). Furthermore, evidence from political science suggests that individuals do sometimes express opinions that they do not in fact believe in order to signal their allegiance to their political coalition (Schaffner & Luks, 2018). Taking these considerations together, one might worry that the theoretical case for SAB just outlined comes to seem less compelling.

Of course, we do sometimes consciously deceive others about our beliefs. Nevertheless, conscious deception itself brings its own costs. It typically requires substantial energy and attention and often elicits strong punishment if it is discovered (von Hippel & Trivers, 2011), and these costs increase in proportion to the degree of social scrutiny. This is true even if people are bad at detecting deception. The evidence alleged to demonstrate this fact, however, is highly controversial. For example, most studies designed to test the detection of deception involve socially atypical conditions, such as little or no punishment for the deceiver, the inability of the deceived to question the deceiver, and a lack of familiarity or personal history between the two agents (see von Hippel & Trivers, 2011, p. 3). Furthermore, the relevant question is not the frequency with which deception is detected, but the *expected costs* of detection, which can of course be high even if the rate of detection is low. Finally, it is crucial to stress that there are some cases in which conscious deception itself brings few benefits: If one has little practical incentive to hold true beliefs anyway, one will also have little practical incentive to consciously deceive. For these reasons, the claim that there are cases in which the practical benefits of genuine socially adaptive belief formation outweigh the benefits of conscious deception is plausible. Of course, its *truth* must ultimately be decided based on its ability to illuminate concrete psychological phenomena – a task that I take up in Section 4.

Second, one might worry that SAB is committed to the thesis of doxastic voluntarism, the highly unpopular view that individuals are capable of forming beliefs at will. It is not. Specifically, it does not assert that individuals go through a process of conscious reasoning in forming beliefs of the form: 'If I believe that P, this will have desirable effect E on such and such people;

³I thank an anonymous reviewer for pressing this point.

therefore, I should believe that P'. As I have argued, socially adaptive belief formation is best understood as a form of motivated cognition, and motivated cognition is generally something that we are unconscious of engaging in (Kunda, 1990). Of course, one might reasonably ask what the psychological mechanisms and processes underlying socially adaptive belief formation are. As with motivated cognition in general, there are likely many routes by which social incentives influence belief formation: for example, through the wilful avoidance of information, the adjustment of time spent sampling and processing information, the adjustment of the evidential standards required to accept or reject given propositions, the opportunistic assignment of trust, and more (Kahan, 2017a; Kunda, 1990). Nevertheless, a thorough investigation of this complex question lies beyond the scope of this article, and it constitutes a core task for future work (see Section 5 below).

Finally, it is crucial to reiterate what I stressed in Section 1: Even if SAB is often neglected in philosophy and psychology, variants and manifestations of the core idea have surfaced numerous times in psychology and the social sciences. It is worth briefly relating SAB to two of the most influential of these.

First, the proposal that SAB shares most in common with is Trivers' (2000, 2006, 2011) famous hypothesis that the capacity for self-deception evolved to facilitate interpersonal deception. At the core of Trivers' evolutionary hypothesis is the following simple idea:

'If ... deceit is fundamental in animal communication, then there must be strong selection to spot deception and this ought, in turn, to select for a degree of self-deception, rendering some facts and motives unconscious so as not to betray—by the subtle signs of self-knowledge—the deception being practiced' (Trivers, 2006, p. xx).

According to Trivers (2011), self-deception is therefore *socially strategic* (we deceive ourselves the better to deceive others), such that it 'evolved to facilitate interpersonal deception by allowing people to avoid the cues to conscious deception that might reveal deceptive intent' (von Hippel & Trivers, 2011, p. 1).

There is an obvious kinship between this proposal and SAB. Both seek to illuminate various biases in human information processing by appeal to the way in which the distinctive character of our social environments sometimes incentivises epistemic irrationality. Nevertheless, I noted in Section 1 that variants of SAB in the psychological literature often burden this core insight with additional assumptions or claims that are more controversial and thus less plausible. This applies to Trivers' hypothesis, which augments the basic thesis embodied in SAB with additional claims that are strictly inessential to it.

First, Trivers advances the hypothesis as a specific theory of *self-deception*, understood 'as a variety of different processes that are directly comparable to those involved in interpersonal deception' (von Hippel & Trivers, 2011, p. 2). As several theorists have noted, however, many examples of self-deception do not seem to involve a socially strategic element, and these might also be adaptive (McKay & Dennett, 2009). More importantly, it is not clear that all examples of socially adaptive belief are best understood as forms of self-deception. In Section 4.3, for example, I will turn to the phenomenon of identity protective cognition, which occurs when individuals seek out and process information to arrive at beliefs that signal their membership of and loyalty to their respective coalitions. The function of such (often ungrounded) beliefs is not to enable agents to better persuade other agents of their *truth*, as might be the case when I turn to confabulations and positive illusions below (Section 4.2 and Section 4.3), and it is difficult to

understand them in terms of an intrapersonal analogy to interpersonal deception. Furthermore, SAB does not require any double bookkeeping among an agent's beliefs, which is often held to be necessary for genuine self-deception (see, e.g., Pinker, 2011).⁴ SAB is thus less committal than Trivers' hypothesis: It is a general claim about the tendencies underlying belief formation, not a specific proposal about self-deception.

Second, Trivers sets his hypothesis in the theoretical context of evolutionary psychology with a focus on specific genetic adaptations and 'unconscious modules favoured by selection' (Trivers, 2000, p. 116). Although SAB is consistent with this highly controversial framework, SAB does not entail it and its specific evolutionary account (the 'adaptive' in SAB refers to the fact that such beliefs *adapt* the individual to her social environment, not that the processes by which they are formed are genetic adaptations *for that purpose*). Specifically, it might be that SAB emerges as a by-product or consequence of other adaptations: for example, the intense social motivation exhibited in our species and the general capacity for motivational influences on belief formation. This seems to be Heyes' (2018) view (although she does not elaborate on it beyond this brief passage):

Increased social motivation also makes minds more malleable by the social environment. Highly attentive to the actions of others, and craving social approval, *developing humans are inclined to adopt those actions, beliefs, and ways of thinking that yield social rewards.* (Heyes, 2018, pp. 57–58; my emphasis).

SAB is thus intended to strip away the core insight that the presence of other agents often incentivizes epistemic irrationality without committing to the additional – more controversial – claims that Trivers supplements it with.

A second and related idea that SAB has a strong affinity with is the proposal that beliefs sometimes function as social *signals* (Funkhouser, 2017; Kurzban & Athena Aktipis, 2007; Simler & Hanson, 2017). Although this claim is often made without elaboration in the context of discussing certain beliefs (e.g., Kahan, 2017a), Funkhouser (2017) has recently advanced a more systematic defence of this idea (see also Simler & Hanson, 2017). Drawing on the resources of signalling theory, Funkhouser argues that some beliefs are formed not in order to represent the world accurately but rather to signal specific information about the believer to other agents in order to manipulate their beliefs and behaviour. That is, Funkhouser argues not just that our beliefs *do* sometimes convey information to other agents; he argues that it is at least some of the time their *function* to convey such information – a function that can take precedence over their primary representational function, leading to various kinds of biased belief.

Again, there is an obvious kinship between SAB and this proposal, and the current paper can be understood as building on and complementing Funkhouser's work. Nevertheless, it is important to draw a conceptual distinction between SAB and the signalling hypothesis, even if they are consistent and potentially complementary. This is for three reasons.

First, it is important to distinguish the loose way in which beliefs are often described as signals from the technical sense in which Funkhouser intends this claim. In the former case, there is little explanatory loss in replacing the claim that the belief functions as a signal with the less controversial claim that forming the relevant belief allows the agent to signal certain information to other agents *through her outward behaviour*. This might seem like a trivial distinction, but at the core of Funkhouser's theoretical proposal is the claim that we can draw on the

⁴I thank an anonymous reviewer for pointing this out.

resources of signalling theory to illuminate the nature of beliefs themselves. To many, this will likely be objectionable: Signals are typically thought of as perceptible traits or behaviours, for example, whereas beliefs are typically thought of as paradigmatically unobservable mental states (Glazer, 2018). At the very least, Funkhouser's proposal likely requires controversial assumptions about the nature of beliefs that the hypothesis outlined in this paper does not. The signalling hypothesis is thus more controversial than SAB.

Second, even if some beliefs *are* fruitfully described as signals (see Section 4.3), it is doubtful that the signalling hypothesis can account for all of the influences of social incentives on belief formation. To take only the most obvious example, one way that social rewards and punishments might influence cognition is by *detering* agents from forming certain beliefs – for example, by systematically avoiding certain kinds of information or avoiding drawing conclusions that they would otherwise infer. Under such conditions, there is no belief to function as the relevant signal. Of course, one might argue that the absence of a belief itself signals something important – for example, the relevant agent's ignorance. But this returns us to the first issue: It is no longer *beliefs* that are functioning as the signal, even though the influence of social motivation is still central.

Finally, Funkhouser (2017) seems to treat the signalling hypothesis as in some sense *sui generis*, identifying two core functions of belief, a representational function and a signalling function. By contrast, I have argued that socially adaptive belief formation should be understood in terms of the more general phenomenon of motivated cognition in which an individual's goals and emotions influence the process of belief formation. Given this, it is consistent with my view – and indeed widely supported by the available evidence – that there are other forms of motivated cognition that do not involve socially adaptive belief formation but that also systematically bias belief formation away from truth and that even sometimes promote practical success (see, e.g., Bortolotti, 2015; McKay & Dennett, 2009; Section 4.2 below). SAB thus situates the social functions of beliefs in the broader context of psychological phenomena already widely recognised in psychology and philosophy.

For these reasons, I think that one should resist the identification of SAB with the signalling hypothesis put forward by Funkhouser, even if – as will be clear below – there might be some cases in which socially adaptive beliefs are usefully understood as signals.

4 | THREE EXAMPLES OF SOCIALLY ADAPTIVE BELIEF FORMATION

My aim in this section is to identify several plausible examples of socially adaptive belief formation. As will be clear, these examples are not intended to be exhaustive, and some are more controversial than others. In conjunction with one another, however, they build a compelling case for both the existence and importance of socially adaptive belief formation in human psychology.

4.1 | Confabulation and rationalization

At least since Nietzsche and Freud, it has been common wisdom that we are often mistaken about the causes of our attitudes and choices. This ignorance has been extensively confirmed in recent decades of cognitive neuroscience and experimental psychology. One of the most striking

features of this research has been its demonstration of the human proclivity to *confabulate* (Bortolotti, 2017). Rather than admit our ignorance, we typically advance sincere – but often demonstrably mistaken – explanations of our attitudes and choices. In a common experimental set-up, for example, subjects are asked to choose from a set of products (e.g., pantyhose, detergent, wine, etc.) that are identical except for seemingly superficial differences (e.g., position, packaging, and presentation). When asked to explain their choices, people largely neglect these superficial differences and instead point to non-existent differences between the products. For example, individuals choosing between identical pairs of pantyhose showed a strong bias towards the pantyhose on the right side of the table. When asked to explain their choice, however, they confabulated, pointing to (non-existent) subtle differences in features such as their colour or knitting (Nisbett & Wilson, 1977).

A defining feature of confabulations is that they are advanced without any conscious intention to deceive. Furthermore, it is widely assumed that agents believe the content of their confabulations (Bortolotti, 2017). This raises an obvious question: *Why* do people confabulate? According to the standard view in contemporary philosophy and cognitive science, confabulations constitute attempts to accurately represent the causes of the agent's attitudes. These attempts are *unsuccessful*, however, because the agent lacks relevant information (Strijbos & de Bruin, 2015).

A very different answer argues that confabulations are in large part optimised for *social consumption* (Bergamaschi Ganapini, 2020; Haidt, 2013; Kurzban, 2012; Mercier & Sperber, 2011, 2017; Simler & Hanson, 2017). On this view, although agents harbour no conscious intention to deceive in cases of confabulation, a primary function of confabulation is *social*, enabling agents to present their attitudes and choices as rational and morally justifiable (Bergamaschi Ganapini, 2020; Haidt, 2013). As Mercier and Sperber (2017) put it, '[t]he reasons people attribute to themselves ... are chosen less for their accuracy than for their high ... value as justifications' (p. 186).

There are several reasons for favouring the view that a primary function of confabulation is public relations. First, the standard view that confabulation aims exclusively at accurate representation confronts several problems. Most fundamentally, it fails to explain why confabulation is so selective in the reasons that are identified as causes of the agent's attitudes, and why confabulators are often positively resistant to acknowledging the actual causes of those attitudes when they are presented as possibilities (Mercier & Sperber, 2017). As Bergamaschi Ganapini (2020) puts it, the standard view 'overlooks that confabulations are generally presented as *good* (or proper) grounds' (p. 6). That is, confabulations typically consist of post hoc *rationalisations* designed to show that the relevant attitude or choice was *justified*. In addition, the standard view does not offer a satisfactory explanation of why individuals confidently confabulate rather than admit their ignorance. Bortolotti (2017), for example, argues that, 'people do not acknowledge their ignorance because *they do not know that they do not know* some of the key factors contributing to their attitudes and choices' (p. 237). This is merely to restate the problem, however: *Why* are individuals ignorant of their ignorance? This is puzzling if the exclusive function of confabulation is accurate representation. It is not puzzling if confabulation is designed to cast the agent's attitudes and choices in a socially attractive light.

Second, even those who endorse the standard view acknowledge that confabulation brings social benefits and that the failure to confabulate – to simply acknowledge one's ignorance or identify the real causes of one's attitudes (e.g., the position of a pair of pantyhose) – would often result in social costs. Bortolotti and Cox (2009), for example, note that, 'giving a confident answer is socially rewarded and advantageous as opposed to saying "I don't know"' (p. 961).

Given this, confabulation is a prime *candidate* for socially adaptive belief: Offering ill-grounded explanations of our attitudes and choices that present them in a socially attractive light leads to practical success.

Finally, the idea that confabulation is socially strategic fits with an increasingly influential body of research on the social functions of reasoning more generally. Mercier and Sperber (2011, 2017), for example, have argued that the capacity to produce and evaluate reasons evolved for social purposes of persuasion, justification, and reputation management (see also Haidt, 2013). On their view, confabulations are not atypical but rather emerge from the more general fact that ‘most of the reasons we provide to explain what we think and do are just after-the-fact rationalizations’, such that ‘the main role of reasons is not to motivate or guide us in reaching conclusions but to explain and justify after the fact the conclusions we have reached’ (Mercier & Sperber, 2017, p. 112; my emphasis). Like many others, they ground this approach to understanding reasoning in the importance of reputation and impression management to the survival, reproductive success, and wellbeing of members of our species (see Haidt, 2013; Tetlock, 2002). Such considerations have led several theorists to argue that a good metaphor for understanding the explanations that we provide for our attitudes and choices is a *press secretary* who cares ‘a great deal more about appearance and reputation than about reality’ (Haidt, 2013, p. 86; see Dennett, 1981, p. 152; Kurzban & Athena Aktipis, 2007; Kurzban, 2012; Simler & Hanson, 2017).

For these reasons, the idea that confabulation is a form of socially adaptive belief is plausible. Nevertheless, it also introduces another puzzle: If confabulation is a form of impression management, why do agents genuinely *believe* such confabulations rather than simply pretend to believe them? To understand this, recall from Section 3 that one should expect to see socially adaptive belief under conditions of high social scrutiny of beliefs and low personal risk associated with false beliefs. Confabulations seem to satisfy both conditions. First, they are almost always advanced in interactive contexts in response to a request for explanation from another agent (Bergamaschi Ganapini, 2020). Second, as per the comments from Mercier and Sperber above, the reasons identified in confabulations are typically advanced to *explain* our attitudes and choices, not to *guide* them. As Abelson (1986) remarks, ‘We are very good at finding reasons for what we do, but not so good at doing what we have reasons for’ (p. 228). The very fact that confabulations identify reasons for our attitudes and choices that did not in fact guide them thus explains why their epistemic irrationality is less costly.

Of course, none of these considerations is intended to be decisive. The nature of confabulation and rationalization remain topics of both scientific and philosophical controversy, and much remains to be understood. Nevertheless, the idea that confabulation is a form of socially adaptive belief is increasingly influential in philosophy and the psychological sciences, and there are powerful considerations in its favour.

4.2 | Positive illusions

A second area where socially adaptive belief formation plausibly occurs is the domain of *positive illusions* (see Kurzban, 2012; Kurzban & Athena Aktipis, 2007; von Hippel & Trivers, 2011). Positive illusions include excessively positive self-appraisals and overly optimistic beliefs about one’s future and ability to control the environment, and they are ubiquitous in the general population in Western countries (Taylor & Brown, 1988). A classic example of positive illusions, for example, is the *better-than-average effect*, the fact that the majority of Westerners believe

themselves to be better than average with respect to most socially desirable traits (Alicke, 1985). Indeed, most Westerners believe themselves to be less prone to such self-serving and self-aggrandizing biases than the average individual (Friedrich, 1996).

Evidence suggests that positive illusions are not just ubiquitous but also *beneficial*, correlating with greater psychological wellbeing, physical health, and perhaps even reproductive success, leading McKay and Dennett (2009) to classify them as 'adaptive misbeliefs' (see also Taylor & Brown, 1988). In the psychological and philosophical literature, the standard explanation of both the *cause* and *benefits* of positive illusions focuses mostly or even exclusively on the individual. For example, it is widely held that the function of positive illusions is to protect the 'self', the 'self-concept', the 'self-image', and/or the agent's self-esteem (Taylor & Brown, 1988; see Kurzban, 2012 for a review). A different (and perhaps complementary – see below) explanation of positive illusions, however, contends that they serve *social functions* – that positive illusions constitute a form of impression management by which we manipulate the beliefs of other agents.

The logic underlying a conception of positive illusions as socially adaptive beliefs is straightforward. First, it is strongly in our interests to persuade other people that we possess socially desirable traits, that we have a rosy future, and that we have greater control over the environment than we in fact do, especially for positive outcomes. Second, we are typically better able to persuade other agents of propositions that we ourselves believe. This creates a powerful practical incentive to genuinely *form* such self-serving and self-aggrandizing beliefs. This idea is at the core of Trivers' famous evolutionary account of self-deception outlined above. As Kurzban and Athena Aktipis (2007) summarise the argument in the context of positive illusions:

Positive illusions ... derive from selection pressures associated with persuasion. If others can be made to believe that one is healthy, in control, and has a bright future, then one gains in value as a potential mate, exchange partner, and ally because of one's ability to generate positive reciprocal benefits in the future. (Kurzban & Athena Aktipis, 2007, p. 137).

As noted above, this social perspective on positive illusions is controversial, with the dominant explanation of positive illusions largely ignoring their effects on other people. Indeed, when Taylor and Brown (1988) consider the possibility that the function of positive illusions is 'public posturing' or 'self-presentation' in their classic article on the topic, they reject this explanation on the grounds that individuals genuinely believe positive illusions. Given this, they – and many others – do not even consider the possibility that positive illusions are both genuinely believed *and* socially strategic. That is, they do not consider the possibility of SAB. Again, however, there are several considerations that weigh in favour of an interpretation of positive illusions as socially adaptive beliefs.

First, the dominant explanation of positive illusions is difficult to understand. Holding the social effects of beliefs constant, it is difficult to see how false beliefs would be more useful or adaptive for an individual than true beliefs (Kurzban, 2012). Consider the dominant explanation of why it is useful to hold false beliefs about one's ability to control the environment, for example (see Bandura, 1989, p. 1177). According to this explanation, such illusions of control are useful because they motivate one to undertake risky actions that one would not take if one's beliefs were accurate. This is puzzling, however. In the absence of the social effects of one's beliefs, an individual who acts on accurate information would seem to be at an advantage relative to an individual who acts on inaccurate information. Indeed, there is extensive

experimental evidence confirming that – holding the social effects of one's beliefs constant – overconfidence in fact hinders one's ability to achieve one's goals (Baumeister, Heatherton & Tice, 1993; Fenton-O'Creevy, Nicholson, Soane & Willman, 2003; see Kurzban, 2012, pp. 113–115).

Second, there are characteristics of positive illusions that seem to support an interpretation in terms of SAB. Most obviously, the better-than-average effect concerns 'almost any dimension that is both subjective and *socially desirable*' (McKay & Dennett, 2009, p. 505; my emphasis). Because of this, the content of positive illusions appears to change according to what is socially desirable (see Endo, Heine & Lehman, 2000; Trivers, 2011, p. 17).⁵ Furthermore, people typically only hold positive illusions that they consider *defensible*, such that 'self-presentation is ... the result of a trade-off between favorability and plausibility' (Baumeister, 1999, p. 8), which is difficult to understand if their function is merely the protection of one's self-image or self-esteem (Kurzban, 2012).

Finally, it is important to note that the two explanations are not mutually exclusive: It could be that individuals reap both personal and social benefits from positive illusions. Indeed, if positive illusions are not in fact personally harmful, this could help to explain *how* social incentives come to play a role. As I noted in Section 3, socially adaptive belief is likely to be most pronounced under conditions in which the personal risks associated with false beliefs are minimal. Insofar as positive illusions have personal benefits independent of their effects on other agents, this might provide an explanation of how they can end up being recruited for impression management as well.

Once again, none of this is intended to be decisive. The causes of positive illusions and the causes of their various benefits are still dimly understood, especially given that an interpretation in terms of SAB is rarely even considered in the psychological and philosophical literature. As I have sought to show, however, there are plausible reasons – both theoretical and evidential – for thinking that positive illusions serve social functions, and this view has been influential in the cognitive sciences.

4.3 | Identity protective cognition

The final example of socially adaptive belief formation that I will consider is identity protective cognition (henceforth IPC), the tendency of individuals to sample and process information in ways designed to protect their status as members of desirable groups or subcultures (Kahan, 2013, 2017a). IPC arises whenever certain beliefs become strongly associated with coalitions that individuals want to belong to. In many political and religious coalitions, for example, holding certain beliefs is effectively part of the membership criteria, such that dissent from those beliefs leads to various kinds of exclusion, ostracism, or even – in certain parts of the world, at least, and throughout much of human history – murder. When such beliefs are not those best licensed by the evidence, an individual therefore has a practical incentive to seek out and process information not to arrive at the truth but rather to protect her group membership. As Kahan (2017b) puts it:

⁵Endo et al. (2000) report findings that in Japan, positive illusions involving self-enhancement that are pervasive in Western countries vanish; instead, individuals typically show biases towards both *self-effacement* and unrealistically positive views of their *relationships*, which are plausibly driven by cultural differences in the valuation of humility and interdependence (Trivers, 2011, p. 17).

When individuals apprehend—largely unconsciously—that holding one or another position is critical to conveying who they are and whose side they are on, they engage information in a manner geared to generating identity-consistent rather than factually accurate beliefs. (Kahan, 2017b, p. 6).

A large body of recent work on IPC focuses on its role in how members of the general public form beliefs concerning politically contested matters of societal risk, such as those associated with climate change, genetically modified organisms, fracking, nuclear waste disposal, and gun control measures. In this context, the explanandum is why ‘members of the public disagree—sharply and persistently—about facts on which expert scientists largely agree?’ (Kahan, Jenkins-Smith & Braman, 2011, p. 147). An intuitive and widespread explanation of such divergence between public opinion and scientific consensus points to factors such as ignorance, low numeracy, a lack of scientific literacy, and the exploitation of unreliable cognitive heuristics when assessing risk. As Kahan (2013, 2017a) and colleagues (Kahan et al. 2011) have pointed out, however, a large body of data and experimental results are inconsistent with such explanations. Specifically, those who diverge from expert consensus appear to be no less informed, scientifically literate, or numerate than those who align with it.⁶ In fact, polarization on such issues is greatest among those who score highest on tests of scientific literacy, numeracy, and ‘cognitive reflection’ (an individual’s ability to override intuitive judgements and engage in careful deliberative reasoning) (Kahan, 2017a). Instead, the only factor that significantly correlates with the positions that people take on these issues is their political identity (Kahan, 2013, 2017a).

Given this data, Kahan and others speculate that what drives people’s beliefs in this area is not a dispassionate concern for the truth but rather the desire to protect their respective group identities. Because issues such as climate change and nuclear waste disposal have become highly politicised, the positions that one takes on them become ‘badges of social membership’ (Haidt & Kesebir, 2010, p. 818):

Sometimes ... positions on a disputed societal risk become conspicuously identified with membership in competing groups ... In those circumstances, individuals can be expected to attend to information in a manner that promotes beliefs that signal their commitment to the position associated with their group. (Kahan, 2017a, p. 1).

Researchers have demonstrated several mechanisms by which IPC occurs. An especially important one involves the differential trust assigned to testimony based on its congruency with one’s group’s position.⁷ In one experiment, for example, subjects were asked to assess whether highly credentialed scientists were experts on various issues such as climate change, fracking, and gun control. Their judgements were highly dependent on whether the relevant scientists endorsed the beliefs held within their own political community (Kahan et al., 2011). In addition, IPC can also interfere with reasoning and deliberation. In another experiment, Kahan and colleagues demonstrated that highly numerate individuals capable of judging whether evidence

⁶Although see Ranney and Clark (2016), who show that climate-specific mechanistic and statistical knowledge does seem to increase acceptance of the existence of climate change across the political spectrum, at least temporarily. I thank an anonymous reviewer for pointing this out.

⁷In a forthcoming paper, De Cruz (in press) also argues that there are social (in addition to epistemic) demands when it comes to testimony.

from a controlled experiment supports a given hypothesis effectively lose this ability if shown an experiment that supports a hypothesis inconsistent with their respective group's position on a given issue (Kahan, Peters, Dawson & Slovic, 2017).

As per the discussion in Section 3 above, Kahan points out that engaging in IPC is often perfectly (practically) rational: Given that individuals have a negligible impact on the phenomena that they form beliefs about in this area, they have little incentive to believe what is true; given the high levels of social scrutiny of such beliefs, they have a strong incentive to believe what signals their group membership. As Kahan (2017b) puts it:

Far from evincing irrationality, this pattern of reasoning [i.e., IPC] promotes the interests of individual members of the public, *who have a bigger personal stake in fitting in with important affinity groups than in forming correct perceptions of scientific evidence.* (Kahan, 2017b, p. 1; my emphasis).

Although Kahan's research is on a very specific topic, the basic logic of IPC generalises to any case in which beliefs that are not best licensed by the available evidence become strongly associated with desirable coalitions of various kinds. Under such conditions, an individual's group attachments clash with the aim of truth and thereby undermine the link between practical success and epistemic rationality. This clash between group identity and epistemic rationality has long been recognised. Writing of his experiences in the Spanish civil war, for example, Orwell (1968) famously noted that, 'everyone believes in the atrocities of the enemy and disbelieves in those of his own side, without ever bothering to examine the evidence' (p. 252). This observation was experimentally vindicated in the 1950s when one of the earliest studies on motivated cognition demonstrated that students at Dartmouth and Princeton overwhelmingly reported more infractions by the other side in a penalty filled football match between their universities (Hastorf & Cantril, 1954).

Kahan's research raises the question of how beliefs become strongly associated with certain coalitions to begin with. There are likely many routes by which this occurs, including deliberate efforts by those with vested interests in creating the association. One interesting suggestion in this area, however, is that beliefs that function as badges of group membership are inherently biased towards implausibility and absurdity precisely because out-group members have no incentive to hold such beliefs, thereby ensuring that they function more effectively to differentiate in-group members from outsiders (Tooby, 2017; see also Simler & Hanson, 2017, p. 279).

The relationship between beliefs and loyalty identified by IPC plausibly extends beyond the examples just described. In totalitarian regimes, for example, people are harshly punished if evidence comes to light that they do not subscribe to the regimes' myths, generating a powerful incentive to seek out and process information to jettison the truth in favour of beliefs that signal their loyalty. As Hannah Arendt (1953) remarked, '[t]he ideal subject of totalitarian rule is not the convinced Nazi or the dedicated communist, but people for whom the distinction between fact and fiction, true and false no longer exists' (p. 474). Ordinary life is replete with more prosaic examples of this conflict between loyalty and epistemic rationality. We often expect our friends and family to take our side on factual disputes involving others, for example, even though our side invariably constitutes a self-serving interpretation of those facts (Simler & Hanson, 2017, p. 75).

Importantly, IPC does not provide the only explanation of cases in which group membership leads to ungrounded beliefs. In a recent paper, for example, Levy (2019) argues that the greater trust we assign to in-group testimony is an adaptation for acquiring *knowledge* under

the plausible assumptions that in-group members evince greater benevolence than out-group members and that benevolence is a useful cue for filtering testimony. Given this, misinformation can arise whenever issues become politicised (i.e., aligned with specific groups) without the influence of anything like socially adaptive belief formation.

This purely epistemic explanation of the relationship between group identity and ungrounded beliefs no doubt plays an important role in many cases. Nevertheless, there are several phenomena that are difficult to reconcile with this hypothesis. Most obviously, group identity interferes with information processing even in cases that do not involve trust. As I noted above, for example, otherwise highly numerate individuals have been shown to lose the ability to understand the results of a controlled experiment when it supports a hypothesis inconsistent with their group's position (Kahan et al., 2017). Similarly, Levy's account also fails to explain why polarization on politically contested issues is *greatest* among those who score highest on tests of scientific literacy, numeracy, and 'cognitive reflection'. If political identity is a cue that individuals rely on when they have to resort to testimony, one would expect those most reliant on testimony to be most reliant on such cues. In fact, the opposite is true (Lelkes, Malka & Bakker, 2018). IPC explains this phenomenon by appeal to the greater cognitive resources certain individuals have to rationalize conclusions that they want to believe for reasons of group identity (Kahan, 2017a). It is not clear how Levy's account could explain it. Finally, and most importantly, purely epistemic explanations fail to explain why individuals are so emotionally *invested* in the relevant beliefs, why they go to great lengths to advertise such beliefs to in-group members, and more generally the emotional and motivational influences at play when it comes to such politicised issues (see Kahan, 2017a, 2017b; Kahan et al., 2017).

Focusing on the relationship between beliefs, loyalty, and group identity offers an especially clear-cut example of socially adaptive belief. As these remarks should make clear, however, this is a topic that warrants substantially more research in the future.

5 | CONCLUSION

The core claim of this article has been simple: The way in which we form beliefs is sensitive to their effects on other agents. I have argued that this hypothesis is plausible on theoretical grounds in light of distinctive characteristics of human social life, and I have identified several putative examples of this phenomenon in a range of different areas. These three examples are not supposed to be exhaustive. Collectively, however, they illustrate important features of human cognition that theorists from a range of different fields have sought to illuminate by appeal to the influence of social incentives on belief formation. As I have noted, some of these examples are more controversial than others. My aim in this article has not been to conclusively vindicate SAB but to render it plausible in the hope that this might spur future research on this phenomenon. To that end, I will conclude by noting three important areas for future research.

First, it would be beneficial in future work to have a more formal taxonomy of the various ways in which social motives influence belief formation. These motives are heterogeneous: to be socially and sexually desirable, to build, maintain, and strengthen relationships and alliances, to attain social dominance and prestige, and so on. It would be useful to have a more systematic understanding of how this diverse array of complex social goals guides the way in which we seek out and process information.

Second, I have not addressed in any detail the psychological mechanisms and processes that underlie socially adaptive belief formation. A more rigorous treatment in the future should

rectify this. As I noted in Section 3.1, motivated cognition in general is facilitated by a variety of different strategies and there is no reason to think that socially adaptive belief formation would be different. Nevertheless, the treatment of this topic here has been shallow. Remedying this defect is a crucial task for future work in both philosophy and psychology.

Finally, and most importantly, future research should focus on a more rigorous examination of the evidence for and against SAB. Importantly, there are really two issues here. First, although I have tried to explain why SAB offers a plausible explanation of the phenomena outlined in Section 4, I have also noted that in most cases there are competing explanations of such phenomena that make no reference to social incentives: for example, the second-order ignorance widely thought to drive confabulation, the purely personal hedonic and motivational benefits of positive illusions, and the combination of in-group trust and unfortunate epistemic circumstances alleged to underpin the relationship between group identity and ungrounded beliefs. Future work should search for more effective ways of adjudicating such controversies. To take only one example, SAB makes a straightforward prediction: Manipulating people's expectations about the social consequences of candidate beliefs should influence the way in which they seek out and process information.⁸ Future experimental work should look for ways to test this prediction.

Just as important, however, is a more theoretical question that I have largely ignored throughout this article: Even granting that social incentives influence the way in which we seek out and process information, why treat the cognitive attitudes that result from such incentives as *beliefs* (see Levy, 2018)?⁹ Many philosophers and psychologists have sought to draw a distinction between different kinds of cognitive attitudes that are often subsumed under the general term 'belief'.¹⁰ Although none of the distinctions such theorists have drawn that I am aware of align straightforwardly with the difference between socially adaptive beliefs and ordinary world-modelling beliefs that I have outlined here, one might nevertheless worry that the functional properties of the former are sufficiently different from the latter to warrant status as a different kind of cognitive attitude. To take only the most obvious example, socially adaptive beliefs are typically much less responsive to evidence than ordinary beliefs. If one individuates cognitive attitudes by their functional properties, does this not threaten the idea that they constitute the same kind of attitude?

From my perspective, this line of argument is better thought of as a potential clarification of SAB than a critique. After all, implicit in the theoretical argument of Section 2 is that one should *expect* socially adaptive beliefs to function differently from ordinary world-modelling beliefs. That is, insofar as their function is to elicit desirable responses from other agents, one would expect their functional properties to be adapted to this function. For example, one would expect agents to shield socially adaptive beliefs from counter-evidence, to be emotionally invested in such beliefs, to advertise them to others, to be reluctant to draw implications from such beliefs that are not themselves socially adaptive, and to be reluctant to act on such beliefs outside of social contexts. Indeed, I noted in Section 3 that beliefs that we are less likely to act on are the prime *candidates* for the influence of motivational influences such as social goals. If one concludes from such functional differences that socially adaptive beliefs are not really

⁸I am grateful to an anonymous reviewer for pointing this out.

⁹I thank an anonymous reviewer for pressing this point.

¹⁰Some influential distinctions include: Dennett (1978), who distinguishes beliefs from *opinions*; Bratman (1992), who distinguishes beliefs from *acceptances*; Rey (1988), who distinguishes 'central beliefs' from *avowals*; and Van Leeuwen (2014, 2018), who distinguishes 'factual beliefs' from *religious credences*.

beliefs at all but rather a different kind of cognitive attitude merely *masquerading* as beliefs, that would be an important theoretical clarification of SAB.

Nevertheless, it is a notoriously difficult philosophical question how to functionally individuate beliefs (and psychological kinds more generally), and there are equally persuasive considerations for treating socially adaptive cognitive attitudes as a kind of belief. For example, they guide sincere verbal assertions, which is plausibly the most important cue used by ordinary people for belief ascription (Rose, Buckwalter & Turri, 2014), and the functional differences just outlined are themselves differences of *degree*, not kind. Agents *do* in fact act on the kinds of socially adaptive attitudes outlined in Section 4, they do use them in reasoning, and they are not literally impervious to counterevidence. Furthermore, the appeal to motivational influences is intended to explain the functional differences between motivated beliefs and non-motivated beliefs without appealing to a difference of kind in the relevant cognitive attitudes. A drug addict motivated to deny her drug problem, for example, also harbours beliefs with fundamentally different functional properties to ordinary beliefs (Pickard, 2016). Rather than explaining such functional differences by introducing a distinct cognitive attitude, it is plausibly more illuminating to explain them in terms of the way in which a single kind of cognitive attitude adapts to the influence of the agent's motivations. It may be that something similar should be said about socially adaptive beliefs.

These brief remarks barely scratch the surface of this complex issue. For a fully satisfying understanding of the way in which the contents of our minds are shaped by the structure of our social worlds, however, this is an issue that must be addressed in future work.

ACKNOWLEDGEMENT

For helpful feedback and comments on earlier drafts, I would like to thank two anonymous reviewers and Bence Nanay, Stephen Gadsby, Marcella Montagnese, Ben Tappin, Nick Wiltscher, Thomas Raleigh, Kevin Lande, Gerardo Viera, and Magdalini Koukou.

ORCID

Daniel Williams  <https://orcid.org/0000-0002-9774-2910>

REFERENCES

- Abelson, R. (1986). Beliefs are like possessions. *Journal for the Theory of Social Behaviour*, 16(3), 223–250. <https://doi.org/10.1111/j.1468-5914.1986.tb00078.x>
- Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, 49(6), 1621–1630.
- Anderson, J. (1990). *The adaptive character of thought*. Hoboken, NJ: Taylor and Francis.
- Arendt, H. (1953). Ideology and terror. In *The origins of totalitarianism* (pp. 460–479). New York, NY: Harcourt Brace.
- Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist*, 44(9), 1175–1184.
- Baumeister, R. F., Heatherton, T. F. & Tice, D. M. (1993). When ego threats lead to self-regulation failure: Negative consequences of high self-esteem. *Journal of Personality and Social Psychology*, 64(1), 141–156.
- Baumeister, R. F. (1999). The nature and structure of the self: An overview. In R. F. Baumeister (Ed.), *The self in social psychology* (pp. 1–20). Philadelphia, PA: Psychology Press.
- Bénabou, R. & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3), 141–164. <https://doi.org/10.1257/jep.30.3.141>
- Bergamaschi Ganapini, M. (2020). Confabulating reasons. *Topoi*, 39(1), 189–201.
- Bond Jr., C. F. & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3), 214–234.

- Bortolotti, L. & Cox, R. E. (2009). "Faultless" ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition*, 18(4), 952–965.
- Bortolotti, L. (2015). *Irrationality*. Cambridge: Polity Press.
- Bortolotti, L. (2017). Stranger than fiction: Costs and benefits of everyday confabulation. *Review of Philosophy and Psychology*, 9(2), 227–249.
- Bratman, M. E. (1992). Practical reasoning and acceptance in a context. *Mind*, 101(401), 1–15.
- Cowie, C. (2014). In defence of instrumentalism about epistemic normativity. *Synthese*, 191(16), 4003–4017. <https://doi.org/10.1007/s11229-014-0510-6>
- De Cruz, H. (2020). Believing to belong: Addressing the novice-expert problem in polarized scientific communication. *Social Epistemology*, 1–13.
- Dennett, D. (1978). How to change your mind. In D. Dennett (Ed.), *Brainstorms: Philosophical essays on mind and psychology* (pp. 300–309). Cambridge, MA: MIT Press.
- Dennett, D. (1981). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.
- Endo, Y., Heine, S. & Lehman, D. (2000). Culture and positive illusions in close relationships: How my relationships are better than yours. *Personality and Social Psychology Bulletin*, 26(12), 1571–1586.
- Fenton-O'Creevy, M., Nicholson, N., Soane, E. & Willman, P. (2003). Trading on illusions: Unrealistic perceptions of control and trading performance. *Journal of Occupational and Organizational Psychology*, 76(1), 53–68.
- Fodor, J. (1975). *The language of thought*. New York, NY: Thomas Y. Crowell.
- Fodor, J. (2001). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.
- Friedrich, J. (1996). On seeing oneself as less self-serving than others: The ultimate self-serving bias? *Teaching of Psychology*, 23(2), 107–109.
- Funkhouser, E. (2017). Beliefs as signals: A new function for belief. *Philosophical Psychology*, 30(6), 809–831.
- Gigerenzer, G. & Selten, R. (2002). *Bounded rationality*. Cambridge, MA: MIT Press.
- Glazer, T. (2018). Are beliefs signals? *Philosophical Psychology*, 31(7), 1114–1119.
- Haidt, J. (2013). *The righteous mind*. London: Penguin Books.
- Haidt, J. & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 792–832). Hobekon, NJ: Wiley.
- Harman, G. (2004). Practical aspects of theoretical reasoning. In A. R. Mele & P. Rawlins (Eds.), *The Oxford handbook of rationality* (pp. 45–56). New York, NY: Oxford University Press.
- Hastorf, A. & Cantril, H. (1954). They saw a game: A case study. *The Journal of Abnormal and Social Psychology*, 49(1), 129–134.
- Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. London: Harvard University Press.
- Kahan, D. (2017a). The expressive rationality of inaccurate perceptions. *Behavioral and Brain Sciences*, 40, 26–28. <https://doi.org/10.1017/S0140525x15002332>
- Kahan, D. (2017b). Misconceptions, misinformation, and the logic of identity-protective cognition. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2973067>
- Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8, 407–424.
- Kahan, D. M., Peters, E., Dawson, E. & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, 1(1), 54–86.
- Kahan, D., Jenkins-Smith, H. & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, 14(2), 147–174.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5), 1449–1475.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Random House.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498.
- Kurzban, R. (2012). *Why everyone (else) is a hypocrite*. London: Princeton University Press.
- Kurzban, R. & Athena Aktipis, C. (2007). Modularity and the social mind. *Personality and Social Psychology Review*, 11(2), 131–149.
- Leikes, Y., Malka, A. & Bakker, B. N. (2018). An expressive utility account of partisan cue receptivity: Cognitive resources in the service of identity expression. Unpublished manuscript. Annenberg School for Communication,

- University of Pennsylvania. Retrieved from <https://pdfs.semanticscholar.org/dc57/456399c71dfb8ed13bc52bb65c0857987a0f.pdf>
- Levy, N. (2018). Showing our seams: A reply to Eric Funkhouser. *Philosophical Psychology*, 31(7), 991–1006.
- Levy, N. (2019). Due deference to denialism: Explaining ordinary people's rejection of established scientific findings. *Synthese*, 196(1), 313–327.
- Loewenstein, G. & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, 2(3), 166–167.
- Marcus, G. (2009). How does the mind work? Insights from biology. *Topics in Cognitive Science*, 1(1), 145–172.
- McKay, R. & Dennett, D. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32(6), 493–510.
- Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2), 57–74.
- Mercier, H. & Sperber, D. (2017). *The enigma of reason*. Cambridge, MA: Harvard University Press.
- Millikan, R. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
- Nesse, R. & Williams, G. (1995). *Why we get sick*. New York, NY: Vintage Books.
- Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.
- Orwell, G. (1968). Looking back on the Spanish war. In S. Orwell & I. Angus (Eds.), *The collected essays, journalism and letters of George Orwell* (Vol. 2, pp. 249–267). New York, NY: Harcourt, Brace, and World.
- Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, 51(4), 550–572.
- Pickard, H. (2016). Denial in addiction. *Mind & Language*, 31(3), 277–299.
- Pinker, S. (1999). *How the mind works*. New York, NY: W. W. Norton.
- Pinker, S. (2005). So how does the mind work? *Mind & Language*, 20(1), 1–24.
- Pinker, S. (2011). Representations and decision rules in the theory of self-deception. *Behavioral and Brain Sciences*, 34(1), 35–37.
- Ramsey, F. P. (1990). In D. H. Mellor (Ed.), *Philosophical papers*. New York, NY: Cambridge University Press.
- Ranney, M. A. & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, 8(1), 49–75.
- Rey, G. (1988). Toward a computational account of akrasia and self-deception. In B. P. McLaughlin & A. O. Rorty (Eds.), *Perspectives on self-deception* (Vol. 6), pp. 264–296. Los Angeles, CA: University of California Press.
- Rose, D., Buckwalter, W. & Turri, J. (2014). When words speak louder than actions: Delusion, belief, and the power of assertion. *Australasian Journal of Philosophy*, 92(4), 683–700.
- Schaffner, B. F. & Luks, S. (2018). Misinformation or expressive responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly*, 82(1), 135–147.
- Sharot, T. & Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in Cognitive Sciences*, 20(1), 25–33. <https://doi.org/10.1016/j.tics.2015.11.002>
- Simler, K. & Hanson, R. (2017). *The elephant in the brain*. Oxford: Oxford University Press.
- Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2), 129–138.
- Sterelny, K. (2015). Content, control and display: The natural origins of content. *Philosophia*, 43(3), 549–564.
- Stich, S. (1990). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation*. Cambridge, MA: MIT Press.
- Strijbos, D. & de Bruin, L. (2015). Self-interpretation as first-person mindshaping: Implications for confabulation research. *Ethical Theory Moral Practice*, 18(2), 297–307.
- Taylor, S. E. & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103(2), 193–210.
- Tetlock, P. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, 109(3), 451–471.
- Tooby, J. (2017). *Coalitional instincts*. Edge. Retrieved from <https://www.edge.org/response-detail/27168>
- Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, 907(1), 114–131.
- Trivers, R. (2006). Foreword to Richard Dawkins' "*The selfish gene*". In R. Dawkins (Ed.), *The selfish gene: 30th anniversary edition* (pp. xix–xx). Oxford: Oxford University Press.
- Trivers, R. (2011). *The folly of fools*. London: Basic Books.

- von Hippel, W. & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(1), 1–16.
- Van Leeuwen, N. (2014). Religious credence is not factual belief. *Cognition*, 133(3), 698–715.
- Van Leeuwen, N. (2018). The factual belief fallacy. *Contemporary Pragmatism*, 15(3), 319–343.
- Williams, D. (2018). Hierarchical Bayesian models of delusion. *Consciousness and Cognition*, 61, 129–147.

How to cite this article: Williams D. Socially adaptive belief. *Mind & Language*. 2021; 36:333–354. <https://doi.org/10.1111/mila.12294>