European Journal for Philosophy of Science (2021) 11:88
https://doi.org/10.1007/s13194-021-00404-2

PAPER IN THE PHILOSOPHY OF THE SCIENCES OF MIND AND BRAIN

Self-reflexive cognitive bias

Joshua Mugg[1] · **Muhammad Ali Khalidi**[2]

Received: 5 January 2021 / Accepted: 15 July 2021

Abstract
Cognitive scientists claim to have discovered a large number of cognitive biases, which have a tendency to mislead reasoners. Might cognitive scientists themselves be subject to the very biases they purport to discover? And how should this alter the way they evaluate their research as evidence for the existence of these biases? In this paper, we posit a new paradox (the 'Self-Reflexive Bias Paradox'), which bears a striking resemblance to some classical logical paradoxes. Suppose that research *R* appears to be good evidence for the existence of bias *B*, but if *B* exists, then *R* would have been subject to *B*. Thus, it seems sensible for the researcher to reject *R* as good evidence for the existence of *B*. However, rejecting *R* for this reason admits the existence of *B*. We examine four putative cognitive biases and criticisms of them, each of which seem to be subject to self-reflexivity. In two cases, we argue, paradox is avoidable. In the remaining two, we cannot find a way to avoid the paradox, which poses a practical obstacle to scientific inquiry and results in an intriguing theoretical quandary.

Keywords Cognitive bias · Heuristics · Paradox · Cognitive Science

1 Cognitive bias

Work on cognitive bias has been a lively research program in the past few decades, and some of the ideas have spread beyond academia to capture the popular imagination (e.g. Evans, 2010; Gladwell, 2005; Kahneman, 2011; Stanovich, 2004).[1] In the 'rationality wars' of the 1980s and 1990s, some philosophers (e.g. Cohen, 1981) argued that it was not possible to use human reasoning to conclude that human

[1] There has been such an explosion in the literature on biases that the number of biases have multiplied in popular culture (e.g. Wikipedia lists well over 100 distinct biases).

✉ Muhammad Ali Khalidi
makhalidi@gc.cuny.edu

Joshua Mugg
joshuamugg@gmail.com

[1] Park University, 8700 NW River Park Dr., Parkville, MO 64152, USA

[2] City University of New York Graduate Center, 365 Fifth Avenue, New York, NY 10016, USA

Published online: 14 August 2021

reasoning is itself subject to systematic biases, at least not in *competence* (only in *performance*). It would be like using a scale to weigh itself: the self-reflexivity makes it impossible. But this seems to go too far: it is not as though there is *one thing* 'human reasoning' that we use to show that *that same thing*, 'human reasoning,' is irrational. Rather, we typically use one aspect of human reasoning to demonstrate that some other aspect is subject to certain biases. Nevertheless, we will argue, some biases do exhibit a special kind of self-reflexivity, such that *the very process* used to discover a specific bias may itself be subject to that same bias. It turns out that, sometimes, researchers are trying to weigh a scale using that very scale, and that generates a paradox in at least some cases, which may be an impediment to inquiry.

A bias is usually understood as a systematic pattern or habit of reasoning that constitutes a departure from norms of rationality.[2] However, in some contexts, and given certain constraints (of time, cognitive effort, and so on), some non-ideally rational habits of reasoning may perform better than ideal rules of inference and argument. Though these heuristics may be regarded as biases in some contexts, they may be seen as rational in others. This gives rise to the notion of 'bounded rationality' (Simon, 1969), whereby sub-optimal heuristics are often to be preferred to optimal rules under certain conditions or in specific contexts. Given naturalized approaches to epistemology and the fact that we are cognitively limited, we must rely on biases to some extent: we don't have time or computing power not to. As Antony (2016, 161) comments on implicit bias: 'without bias, we would know less, not more.' In what follows, we shall focus primarily on contexts in which these habits of reasoning distort and mislead, since it is those contexts that generate potentially paradoxical self-reflexivity. We will refer to cognitive shortcuts in general, whether they lead to correct results or errors, as 'heuristics,' and when they mislead or result in errors, we will refer to them as 'biases.'[3]

Here is how the paper will proceed: in Sect. 2, we stave off an objection that cognitive bias does not apply to scientific reasoning. In Sect. 3, we outline the general form of the Self-Reflexive Bias Paradox for cognitive science and compare it to the Liar Paradox. In Sect. 4, we examine four putative cognitive biases and show how the paradox applies to each. We conclude that, although two of the biases can avoid a vicious paradox, in the latter two cases, we are unable to see a way out.

2 Is scientific reasoning insulated?

One might object to our project at the outset on the grounds that the heuristics and biases research program studies *everyday reasoning*, which is subject to errors that *scientific reasoning* is able to avoid for two reasons. First, contemporary

[2] This characterization of cognitive bias is consistent with a number of definitions offered by researchers working in the field. 'Cognitive bias refers to a systematic (that is, nonrandom and, thus, predictable) deviation from rationality in judgment or decision-making' (Blanco 2017). 'A bias is a systematic discrepancy between the (average) judgment of a person or a group and a true value or norm' (Gigerenzer 2018). 'The term biases refers to the systematic errors that people make in choosing actions and in estimating probabilities…' (Stanovich, Toplak & West 2020).

[3] This terminology agrees broadly with standard usage in cognitive science; for example, in an introduction to a seminal collection of papers, Gilovich & Griffin (2002, 3) define biases as 'departures from the normative rational theory that served as markers or signatures of the underlying heuristics.'.

science relies on instrumentation and computation, which are designed to reduce the reliance on the frailty of human cognition and perception (see e.g. Bogen & Woodward, 1992). Second, science has evolved various mechanisms to correct for many cognitive biases, especially in a collective context (e.g. peer review, double-blind experiments, and replication). Third, while a single researcher may be biased, the addition of team members and other research teams will lead to a correction of their mistakes.

While these considerations mitigate the worry about cognitive bias, they do not overcome it completely for both conceptual and empirical reasons. Regarding the first reason, instrumentation and computation are designed and built by humans, and so the worry of the fallibility of human reasoning returns. As for the second reason, while peer-review and replication should catch errors in reasoning, it is entirely possible that the biases at work in the initial author's reasoning will be at work in reviewers' and editors' reasoning as well. Finally, while ideally the social nature of science reduces individual bias, it also creates new biases, such as publication bias due to the 'file-drawer effect' (see e.g. Franco et al., 2014). In other words, although the social nature of science can help us avoid error when scientists have different blind spots (or blind spots that we know how to check), it is of less help if each subject has the same blind spots, as may be the case if certain biases exist within scientific reasoning. As such, we think that we do not have sufficient reason for thinking that scientific reasoning entirely escapes biases. Thus, it remains an empirical question whether and to what extent science is susceptible to biases.

Some cognitive scientists, philosophers, and others have directly broached the question as to whether cognitive biases influence the way in which scientists conduct their inquiries and have found reasons to think that scientific reasoning is indeed subject to some of the same biases as ordinary reasoning. After all, scientists have the same basic perceptual and cognitive makeup as the rest of us: scientists are human too. Thus, many of these biases should apply to them, though maybe not all, since some can be partly overcome by training. One of the earliest expressions of this worry can be found in Francis Bacon's *Novum Organum* (1620/1902), where he cautions his readers against four biases or 'Idols' that distort scientific reasoning and lead to spurious results. The first of these biases, the 'Idols of the Tribe,' so called because they pertain to the entire 'tribe' of human beings, involve various kinds of perceptual and cognitive biases. Bacon (1620/1902, XLI) points out that 'the human mind resembles those uneven mirrors which impart their own properties to different objects, from which rays are emitted and distort and disfigure them.' He thinks that scientists can succumb to these distortions no less than other human beings and that they need to guard against them when conducting research.

More recently, Bacon's concern has been bolstered by empirical evidence. In particular, some empirical work shows that experts are as prone as novices to cognitive biases, notably studies on the base-rate fallacy with physicians (Shanks, 1991; Bergus et al., 1995). Other studies show that scientists often succumb to pre-scientific ways of thinking, even when it comes to their own fields of research. For example, Knobe and Samuels (2013, 72) have provided evidence to show that scientists and non-scientists alike 'are drawn to a conception of innateness that differs from the one at work in contemporary scientific research,' though they also show

that both groups are 'capable of "filtering out" their initial intuitions and using a more scientific approach.'

We do not take these considerations to demonstrate that science is biased. Rather, we mean to show that it is not known that science is unbiased. We think it is an open empirical question whether, and to what extent, cognitive biases are operative in science. As such, we return to our question whether researchers on cognitive biases might succumb to the very biases they purport to discover.

3 Application to cognitive science

Since science may be subject to bias and cognitive science is (presumably) a branch of science, we can raise the following questions. 1) Are there instances in which cognitive science is itself affected by cognitive biases? 2) Could human biases affect research on those biases themselves? That is, can research on a *particular* cognitive bias itself be prone to the very bias it is investigating? 3) If so, does that generate a paradoxical type of self-reflexivity?

In some cases, it doesn't seem as though self-reflexivity is an issue. For example, if we want to know how likely it is that doctors commit the Base Rate Fallacy, we would need to take the base rates of doctors' estimations into account. This is a kind of self-reflexivity, but the researchers need not be prone to that fallacy. Indeed, the Base Rate Fallacy is easy to correct once researchers are mindful of it. However, in other cases, self-reflexivity would seem to be a serious possibility. For example, experiments on the prevalence of Confirmation Bias might give more weight to confirming than disconfirming evidence. In that case, the existence and pervasiveness of the bias seems to cast doubt on the research itself.

If there are cases of self-reflexivity in which the research on a particular bias is subject to that very bias, this generates a puzzle that is reminiscent of the classical logical paradoxes. Suppose a group of scientists *S* have carried out research *R* that purports to show the existence of some bias *B* that tends to manifest in some specific context, and the research is methodologically sound, given our knowledge of how to avoid reasoning errors at the time the research was conducted. This gives *S* reason to believe that bias *B* exists.

Now, in evaluating their own research after purportedly discovering this new bias, rationality requires members of research group *S* to take into account the possibility that they too are susceptible to bias *B.* They must reason about the reasoning process they used to discover *B,* since they now know of one additional way in which human reasoning can be biased (namely, *B*). If they are susceptible to *B*, and *R* was conducted in a way in which one would expect reasoning to be subject to *B*, then rationality requires *S* to withdraw assent to the existence of *B*.[4] This is so because the existence of *B* was supported only by *R,* but since *S* is susceptible to *B, S* may have succumbed to *B,* implying that *R* is not reliable. So it is reasonable to withhold assent to the existence of the putative new bias *B*.

[4] Later, we will explore how the paradox manifests itself depending on whether the existence or prevalence of the bias is at issue.

In fact, if the contexts are relevantly similar, then if the researchers did *not* find the effect in their own reasoning, it seems this would be evidence against the existence of the bias. If the bias really does exist, we should *expect* the cognitive scientists to be subject to it.[5] For example, in a paper arguing for the pervasiveness of confirmation bias, Nickerson (1998, 211) writes:

> …I have argued that the confirmation bias is pervasive and strong and have reviewed evidence that I believe supports this claim. The possibility will surely occur to the thoughtful reader that what I have done is itself an illustration of the confirmation bias at work. I can hardly rule the possibility out; to do so would be to deny the validity of what I am claiming to be a general rule.

Although Nickerson acknowledges the thoughtful reader's worry, he does not seem particularly worried by it. This, we think, is a mistake.

Once *S* realizes that their research *R* was subject to bias *B*, they can say: 'Precisely! Our research *R* was subject to *B*, which gives us reason to believe in the existence of bias *B*, just as we claimed!' However, there is something paradoxical about this response: if *S* does not have reason to believe that bias *B* exists (because *R* is undercut in virtue of being subject to *B*) then *S* has reason to believe that bias *B* exists (since *S* has been subject to *B*), and if *S* has reason to believe that bias *B* exists (because of *R*) then *S* does not have reason to believe that bias *B* exists (because *R* is undercut). We can call this the 'Self-Reflexive Bias Paradox,' which we can put more formally as follows:

> **Self-Reflexive Bias Paradox:** If a subject *S* conducts research *R*, which is evidence for the existence of bias *B* that arises in context *C*, and *R* was conducted in context *C,* then *R* was likely subject to *B* and *S* should not accept research *R* as good evidence for the existence of *B*. If *S* had accepted the existence of *B* purely on the basis of *R*, then *S* should not accept the existence of *B;* but, if *S* does not accept *R* as good evidence for the existence of *B* on the basis that *R* was likely subject to *B,* then *S* should accept the existence of *B*.

While we have framed the paradox from the point of view of the scientists considering their own research, the bias likewise applies to outside observers considering whether *R* supports *B*, and in the following section we will consider cases where the researchers positing the bias and those criticizing the bias may equally succumb to the paradox. Who it is that raises the possibility that the research was subject to the bias does not matter when it comes to the rationality of accepting the bias. The structure of this paradox is similar to that of the Liar's Paradox: if liar *L* makes an utterance *U* ('I am a liar') to *S,* then *S* should believe *L* if and only if *S* should not believe *L*. In the Self-Reflexive Bias Paradox, the research *R,* which is the evidence for the existence of the bias, is analogous to the liar's utterance *U*. The existence of the bias *B* is analogous to *L*'s being a liar. Finally, in each paradox the subjects *S* are determining what attitude they ought to take toward the research *R* and bias *B* (in the Self-Reflexive Bias Paradox) or liar *L* and utterance *U* (in the Liar's Paradox).

[5] Thanks to [redacted] for pointing this out to us.

However, there are two clear differences between the paradoxes, and thus the Self-Reflexive Bias Paradox is not simply another version of the Liar Paradox. First, epistemic support is unlike truth in that the former admits of gradations whereas the later does not. Thus, in the Liar's Paradox, we swing from being compelled to regard the liar's statement as true to regarding it as false, while in the Self-Reflexive Bias Paradox we swing from being compelled to regard the research as well supported to less well supported. Second, there is an additional logical step in the Self-Reflexive Bias Paradox. The reasonableness of accepting the existence of the bias depends on how well supported the bias is, and so the bias being well supported (or not well supported) implies that it is (or is not) rational to accept the existence of the bias.

These differences do not make the Self-Reflexive Bias Paradox any less paradoxical. To see why, imagine that *L* says 'everything I say is unreliable' rather than saying 'everything I say is false.' Would it be rational to accept *L*'s statement as true? If so, then we should regard *L* as unreliable, implying that we should not accept *L*'s statement. However, then we have admitted that *L* is unreliable, implying that we accept *L*'s statement, in which case, we should not accept *L*'s statement.

While people do not tend to claim that they are liars or that everything they say is unreliable, at least not in scientific papers, researchers do claim that they themselves are subject to biases, opening the possibility that the Self-Reflexive Bias Paradox is actual. As we mentioned in Sect. 1, in the 'rationality wars' some critics of the research on cognitive biases occasionally raised concerns of self-reflexivity, but they did so in an unconvincing blanket fashion. In a paper responding to such criticisms, Stein (1997) outlines an argument that is sometimes brought up against the scientific research that purports to show that humans are irrational. According to that argument, if we are justified in believing that humans are irrational, we should not trust the results of our reasoning processes, and since this conclusion is itself based on our reasoning processes, we are not justified in believing it (Stein, 1997, 546). Stein rightly rejects this very general version of the argument on the grounds that if humans employ faulty reasoning principles there is no reason to think that they are invoked in *every* reasoning context.Rather, as we have already pointed out, it is safe to assume that they are only deployed in certain contexts and these contexts need not be the ones that govern the scientific research in question. But Stein (1997, 559–560) also admits that there is a problem in determining which principles of reasoning are being deployed in each context and hence in ruling out the possibility that a cognitive scientist is relying, albeit implicitly, on some faulty reasoning principle. Our claim in this paper is that there are actual contexts in which it can be plausibly maintained that cognitive scientists are indeed falling prey to the same biases in reasoning that they purport to be investigating, and that this leads to a self-reflexive paradox.

4 Case studies

To unravel this puzzle and determine how we should respond to it, we consider four cases in which self-reflexivity is at issue when it comes to research on cognitive biases. We chose these four biases because they demonstrate that the puzzle plays out differently depending on the case involved. In the first two cases, we suggest

two ways one may escape the paradox, but the third and fourth cases present special challenges, which are not so easily overcome.

4.1 Theory-ladenness of observation

Theory-Ladenness of Observation (TLO) refers to the process whereby higher order cognition has a top-down effect on perception or observation, whether positively or negatively (Brewer, 2012). This is at issue wherever science depends on human observers, as in astronomy, geology, psychology, and many other branches of science. Consider an example of a positive influence of theory-ladenness from Darwin's autobiography. In his description of a trip to Wales in 1831, he collected geological samples with one of his former professors, but he did not observe evidence of glacial formations since he did not have the glacial theory in geology. When he returned later, in 1842, having learned about the glacial theory (though he was still not entirely convinced by it), he found the evidence for it overwhelming. In this way, the theories Darwin held (and did not hold) impacted what it was that he observed. Brewer (2012, 325–326) cites this example in his research on the psychology of science as an instance in which TLO had a positive impact on scientific research. But he also notes that some of his own previous work on the subject tended to emphasize the negative instances of TLO, in which top-down cognitive processes threaten the objectivity of observation. Two decades later, he reflected on his earlier research, recognizing that he himself was subject to negative TLO. He writes:

> In retrospect, in our initial papers on theory ladenness… we were eager to show the existence of top-down processes, so we pointed out many examples from cognitive psychology and the history of science where there was evidence for theory distorting perception, biasing attention, shifting interpretations, distorting memory, and so forth. (2012, 290)

Here, Brewer applies TLO to the research that led him and colleagues to posit the existence of TLO in experimental work in cognitive psychology and various episodes from the history of science. Brewer and his colleagues were keen to show the existence of theory-ladenness, so they focused on cases in cognitive psychology and the history of science that they thought demonstrated it and thereby exaggerated its importance. Is this a version of the Bias Paradox? *Prima facie,* it seems so. It appears that Brewer's research is not good evidence for the existence of TLO, and so he and his colleagues should not accept the existence of TLO based on their work, since their research was problematically biased by TLO. However, if even their own observations were themselves influenced by their background theory, then their research *is* good evidence for the existence of TLO, and they can accept the existence of TLO.

One might think that Brewer was guilty of confirmation bias, rather than TLO, in that he and his colleagues went in search of evidence to confirm, rather than disconfirm, TLO. We have two responses. First, biases are not discrete, and the relation between the dozens of biases are only beginning to be explored. We suspect that TLO

and confirmation bias are intimately related. Second, and more importantly, it seems that confirmation bias is also a likely candidate for the paradox, as researchers on the subject sometimes acknowledge in passing. For example, in his article arguing that confirmation bias is better understood as myside bias, Mercier (2017, 99) asks:

> Why wouldn't the researchers who have claimed to prove the existence of the confirmation bias also be biased? Couldn't their experiments be designed in such ways that they are more likely to make participants look biased? Couldn't these scholars have mistaken rational behavior for proof of a confirmation bias? (see also Nickerson, 1998, 211, quoted above)

However, Mercier asks these questions just to set them to the side. Our point here is that there is a *prima facie* worry about self-reflexivity for confirmation bias just as much as TLO, and in the interest of space, we will focus on the latter.

It seems possible that we can avoid a vicious paradox both for Brewer considering his own research, and for those of us looking on. First, it may be that some observation is theory-laden (as in Brewer's own work), but that Brewer's evidence from the history of science and elsewhere doesn't provide good evidence for it. So we cannot conclude from undercutting Brewer's research that TLO does not exist at all. But second, if Brewer and colleagues were just trying to establish the *existence* of the effect, then their flawed methodology would effectively give reason to believe in the existence of TLO. So if we dismiss their evidence on the grounds that it is itself theory-laden, we are at least acknowledging that the effect exists, albeit not where they say it is. What is at issue in the quoted text above is the *scope* and *perniciousness* of TLO, rather than evidence for its existence. Indeed, criticism of TLO has not tended to dispute the claim that cognition alters reported observations. Rather, epistemologists and psychologists debate whether TLO is (or to what extent it is) pernicious and whether the effect happens within perception or the interpretation of perception. Neither side in this debate disputes the mere existence of TLO. Hence, one can consistently hold that TLO is manifested in the research of Brewer and colleagues, but has not been demonstrated in their subjects, and this can also be maintained by these researchers themselves.

4.2 Dunning-Kruger effect

The Dunning-Kruger Effect (D-K Effect) occurs when those who are deficient in some skill tend not to recognize their deficiency and overestimate their own performance or ability (Dunning, 2011). For example, in one study, 95 first-year medical students completed a CPR exercise, then were asked how well they had done. Only three said they 'failed' the exercise (i.e. missed steps, put them in the wrong order, executed them incorrectly, or moved too slowly), but an expert examiner judged that 36 had failed (Vnuk et al., 2006; cited in Dunning, 2011, 255).[6] Thus, at least 33

[6] Studies on the D-K Effect usually involve novice subjects. Vnuk et al. (2006) is a rare instance of a D-K Effect experiment on a type of scientific expert.

participants who had failed thought they had passed; they not only failed the test, they also misjudged how well they had done.

The D-K Effect has been examined under a variety of experimental conditions for a number of different cognitive tasks or abilities, e.g. logic, grammar, social skills, even humor. Part of the explanation given by Dunning and Kruger, who first drew attention to the effect, is that those who are deficient in some skill are also likely to be deficient in the ability to identify who is good at that skill and who is not, including themselves (Kruger & Dunning, 2002). In many tests, participants in the bottom quartile, on average, estimate themselves to be in the second quartile. Also, top performers *under*-estimate their own performance—because they over-estimate the performance of others (as opposed to because they do not think they are good enough), as Dunning and Kruger argue.

At first glance, the D-K Effect might seem an odd candidate for the self-reflexivity required by the Bias Paradox, since the investigators are experts at identifying the relevant skills. However, Dunning and Kruger open themselves up to the paradox when they write:

> Although we feel we have done a competent job in making a strong case for [the existence of the D-K Effect]… our thesis leaves us with one haunting worry that we cannot vanquish. That worry is that this article may contain faulty logic, methodological errors, or poor communication. (Kruger & Dunning, 1999, 1132)

In fact, some critics of the D-K Effect have claimed that the researchers are in fact ignorant, since the results are just a statistical artifact, namely *regression to the mean*, whereby an extreme value on one variable (here, performance on the test) exhibits regression when compared to another variable (here, perception of performance) (see, e.g., Burson et al., 2006), or confusion on how to measure self-assessment and compare it to competence (Nuhfer et al., 2016, 2017). Applying the first criticism to the example above, regression to the mean implies that if we let those same 33 failing students retake the test, even without any feedback, their average score would increase. However, *ability* would remain constant. Thus, at least *some* of those 33 students had not *really* misjudged their ability (though they had misjudged their performance), but were simply unlucky in that first trial. To measure *ability,* researchers would need to observe each individual over several trials.

Dunning offers two replies. First, he replies by saying that regression to the mean does not explain why performance and perception of performance are so badly correlated. However, Nuhfer et al. (2016) have argued that the two are not badly correlated. Self-assessment measures are a blend of meaningful self-assessment and random noise. Because analyzing randomly generated data in the way Dunning and Kruger do results in the same pattern they find, it appears that the alleged correlation is the result of random noise. Additionally, Nuhfer et al. (2017) argue that Kruger and Dunning (1999) and the subsequent literature on the D-K Effect overlook six aspects of numeracy leading them to see patterns where there are none. The details of the errors do not much matter

for our purposes.[7] What is crucial is that the claim is that researchers purporting to show the existence of the D-K Effect need to be competent in numeracy, and critics claim that they are not. Suppose that these criticisms are correct. Dunning's second reply to Burson et al. (2006) can be equally leveled against Nuhfer and his colleagues, and it puts him squarely in the Bias Paradox:

> But, perhaps unknown to our critics, these responses to our work have also furnished us moments of delicious irony, in that each critique makes the basic claim that our account of the data displays an incompetence that we somehow were ignorant of. Thus, at the very least, we can take the presence of so many critiques to be prima facie evidence for both the phenomenon and theoretical account we made of it, whoever turns out to be right. (Dunning, 2011, 265)

Dunning applies the bias to himself and his colleague. On his interpretation of the criticism, the self-reflexivity of the D-K Effect is prevalent because Dunning and Kruger have overestimated their own ability to judge the abilities of other individuals. We seem to have another instance of the Self-Reflexive Bias Paradox. If Dunning and Kruger's own research is subject to the D-K effect, then their research does not give them reason to believe in the existence of the D-K effect; but if the research was unsound because the researchers were subject to the D-K effect, then the D-K effect exists, albeit in the researchers rather than the participants.

As with the TLO case, if Dunning and Kruger succumbed to the D-K Effect in their research, it is only one data point as opposed to many, so it would be a weak vindication of the effect. At the same time, though, if all they are trying to do is establish the *existence* of the effect, not its prevalence, then it provides some evidence of the cognitive bias. So the escape available in the TLO case is not available here.

While Dunning seems to have opened a paradox for himself, no paradox need remain for Dunning and Kruger's critics. They can respond with an alternative explanation by saying that their criticism is not a matter of the D-K Effect at work in Dunning and Kruger's research, but rather is just an instance of being wrong twice. Dunning and Kruger were wrong about their explanation (a better explanation has to do with regression toward the mean and other aspects of numeracy) and Dunning and Kruger were even more wrong about their confidence in the truth of their theory. Dunning's interpretation of the criticism was just wrong. So, it seems, Dunning and Kruger's critics *need not* raise their criticism in a way that (paradoxically) vindicates the D-K Effect. This leads us to

[7] The six numeracy errors are: 1) 'Random noise can generate X-shaped patterns in Kruger-Dunning-type graphs, and researchers can easily misinterpret these patters as meaningful measures of self-assessment.' 2) Data sets are too small to offer reliability. 3) There are strong floor and ceiling effects in the type of graph Kruger and Dunning employ. 4) 'Sorting data pairs by one member of the pair invariably produces the "X-shaped" pattern of Kruger-Dunning graphs and, sorting data by percentile rank renders all expressions of performance as norm-referenced rather than criterion-based.' 5) 'Kruger-Dunning graphs cannot show the distributions of varied self-assessment skills in a populace.' 6) 'Kruger-Dunning graphs fail to reveal the degree of correlation that exists between self-assessed competence and demonstrated competence on a participant-by-participant basis.' (Nuhfer et al., 2017, 9).

a second way to avoid the paradox: if one has an alternative explanation of the phenomenon under study, then that explanation can be proffered for the existence of the effect in both the participants and the researchers. In some cases, one can escape from the paradox by giving an alternative explanation of the effect, which does not invoke a cognitive bias, and one can give the same explanation of the effect as it is manifested in the practice of the researchers themselves. Thus, the paradox is avoidable for Dunning and Kruger's critics.

4.3 Inherence heuristic

The 'inherence heuristic,' introduced by Cimpian and Salomon (2014), is purportedly a basic cognitive tendency that leads people to explain patterns or correlations with reference to inherent (or intrinsic) features rather than extrinsic (i.e. relational or historical) features. For example, when asked why orange juice is consumed at breakfast, subjects tend to explain it by appeal to intrinsic properties (e.g. refreshing citrus odor) rather than historical factors (e.g. marketing campaign by California Fruit Growers Exchange in 1920s). Likewise, when asked why a router broke, subjects tend to use intrinsic properties (e.g. it is cheap) rather than historical properties (e.g. it was stepped on). Thus, Cimpian and Salomon caution us to be critical of our tendency to explain by way of intrinsic properties.[8] Although this cognitive tendency can be considered a heuristic, which leads to correct answers in some contexts, it is often manifested as a bias giving rise to misconceptions, as in the faulty explanation of the widespread practice of drinking orange juice at breakfast. As Cimpian and Salomon (2014, 468) make clear, in some cases 'the heuristic leads to reasonably accurate (perhaps even normatively correct) explanations.' However, 'many of the patterns that currently structure our world are the products of complex chains of historical causes rather than being simply a function of the inherent features of the entities involved. The human mind, however, may be prone to ignore this possibility' (Cimpian & Salomon, 2014, 462). In what follows, we will concentrate on those cases in which the inherence heuristic can be characterized as a bias.

At least some of Cimpian and Salomon's critics have raised the possibility of self-reflexivity in this instance. In a response to a target article on the subject of the inherence heuristic, we ourselves accused Cimpian and Salomon of having the same cognitive tendency that they claim to have uncovered: '…we cannot help entertaining the possibility that Cimpian and Salomon fall prey to the inherence heuristic in positing an innate heuristic to explain certain human cognitive tendencies, rather than explaining them in terms of relations of human beings to the

[8] The Fundamental Attribution Error, the common human tendency to explain aberrant behavior of others in terms of their flawed character while tending to explain one's own aberrant behavior in terms of rationalizations, can be thought of as an analogue of the inherence heuristic in reasoning about the social world (Jones & Harris 1967; Ross 1977).

world' (Khalidi & Mugg, 2014, 493).[9] Cimpian and Salomon note that humans tend to explain and make inferences using intrinsic features. To explain this fact, Cimpian and Salomon appeal to *intrinsic* properties of humans, which itself may be an instance of the inherence heuristic. Moreover, in this case, the claim is that this cognitive tendency is a bias rather than a heuristic since those who succumb to the bias are providing spurious explanations for the phenomenon being discussed.

In our commentary on Cimpian and Salomon, we argued that we need not posit a heuristic or bias in our cognitive makeup that privileges inherent (or intrinsic) properties, since: (a) these properties tend to be perceptually salient (for example, people don't usually have access to the history of orange juice when drinking it), and (b) they tend to be more explanatory (given the causal structure of the world, since causal powers are intrinsic properties). The fact that humans tend to rely on intrinsic properties has more to do with the world and our relation to it than it does with the intrinsic properties of human cognition.

But the alternative explanation that we offered suggests that the researchers (Cimpian and Salomon) may have fallen prey to the very bias that they claim to have uncovered. Our critique opens the possibility that the inherence heuristic is responsible for leading Cimpian and Salomon to explain behavior in terms of an inherent tendency of the human mind. Since this may well be an instance of the inherence heuristic, this criticism of the inherence heuristic seems to vindicate the existence of a cognitive bias in favor of inherent or intrinsic properties. We have another instance of the Self-Reflexive Bias Paradox: if the researchers are subject to the cognitive bias, then we should reject their reason for positing the existence of the inherence heuristic and have no reason to believe in the existence of this new bias; but if the researchers were subject to the cognitive bias, then we have a reason to accept their conclusion that the bias exists, albeit in the researchers rather than the participants. Either the researchers are right and there is an inherence heuristic in their experimental subjects based on their evidence, or they themselves are victims of the inherence heuristic manifested in their own scientific research, in which case we have no reason to believe that the inherence heuristic is present in their subjects.

In the first two cases discussed above, the theory-ladenness of observation and the Dunning-Kruger effect, there seemed to be ways to avoid the paradox that do not seem available in this case. The criticism of TLO can be interpreted as focusing more on the prevalence or extent of the bias rather than its very existence, and so the fact that the criticism of TLO admitted its existence was not problematic. However, our critique of Cimpian and Salomon took issue with the *very existence* (rather than the extent or scope) of the inherence heuristic. Thus, the escape from the paradox in the TLO case is not open here.

[9] Though in our earlier work we wrote 'innate heuristic', it would have been more correct to say 'inherent heuristic.' Our main criticism was not that Cimpian and Salomon were offering an innate inherence heuristic (as opposed to a non-innate one); rather, our criticism relied on the distinction between inherent (or intrinsic) and extrinsic (or relational) properties.

The critics of the D-K Effect proposed an alternative explanation that applies not just to the (putative) bias being investigated, but to the researchers themselves: the researchers are guilty of an error that invalidates their findings. The bias only manifested for Dunning who wanted to maintain that he had, paradoxically, fallen victim to his own bias. But in the case of the inherence heuristic, it is the *critics* who raise the paradox. To be sure, the critique denies Cimpian and Salomon's theoretical explanation and proposes an alternative explanation, but the alternative involves ascribing *inherent properties*, and in this case, the inherent properties pertain to the *human mind* (namely, the minds of the researchers), thus apparently vindicating the cognitive bias. If we were right in thinking that Cimpian and Salomon are wrong, then Cimpian and Salomon were right in thinking that the inherence heuristic exists. In other words, in accusing researchers of favoring inherent properties over other features of the world, the critique of the inherence heuristic seems to be affirming inherent properties in the minds of the researchers. This criticism of the research appears to give credence to the results of that very research, thus generating a paradox.

4.4 Bias bias

Gigerenzer and colleagues (e.g. Gigerenzer, 2018; Gigerenzer & Brighton, 2009; Todd & Gigerenzer, 2003) have criticized much of the research on cognitive biases over the past several decades, claiming that in many cases the purported biases have not been demonstrated. In some cases, cognitive scientists have misinterpreted their own experimental data, while in other cases different experimental results are obtained when the tasks in question are reframed. To cite one case of the first type, some purported instances of the gambler's fallacy are not actually fallacious, since, Gigerenzer claims, participants are 'evaluated against the wrong normative standard' (2018, 315). As for a case of the second type, Gigerenzer shows that participants are less prone to commit the base-rate fallacy when problems are expressed in terms of frequencies rather than probabilities (2018, 308–309). Based on this thoroughgoing critique of the research on cognitive bias, Gigerenzer argues that many researchers in cognitive science[10] are guilty of a 'bias bias,' which he defines as follows:

> Bias bias: The tendency to see systematic biases in behavior even when there is only unsystematic error or no verifiable error at all. (2018, 307)

Many cognitive scientists, he claims, are guilty of a bias bias when devising and interpreting their experiments, which makes them erroneously attribute biases to experimental participants.

Although Gigerenzer does not consider the possibility, we can ask whether he might be guilty of the same bias. Gigerenzer claims that cognitive scientists have

[10] Sometimes Gigerenzer focuses on the sub-discipline of behavioral economics in particular, but it is clear from the wide range of research that he criticizes that he is targeting most of the work on cognitive bias in the cognitive sciences.

a tendency to over-ascribe biases when other explanations of error might be more apt, but he is also ascribing a bias to other cognitive scientists, who may reasonably claim that the accusation of bias is unwarranted.[11] What reason might we have to think that he is prone to the same bias that he accuses other cognitive scientists of? Gigerenzer (2018, 307) provides three motives for the bias bias in other cognitive scientists: 'an academic agenda to question the reasonableness of people,' 'a commercial agenda to discredit the judgment of jurors,' and an inclination 'to promote trust in abstract logic, algorithms, and predictive analytics and distrust in human intuition and expertise.' We might add that there may be *social* pressures to explain human behavior in terms of a new cognitive bias: researchers are more likely to publish and enhance their reputation by demonstrating the existence of a novel mental phenomenon, such as a new human cognitive bias. Gigerenzer might contend that these motives are present in some of his colleagues but not himself. However, it would seem that at least the first motive may obtain for Gigerenzer, who is questioning the reasonableness of his fellow cognitive scientists (by attributing ulterior motives to them) and positing the existence of a novel mental phenomenon. Hence, in the absence of an assurance that he is exempt, it is reasonable to think that Gigerenzer might be guilty of the very same bias that he ascribes to others, and this generates a paradox. If he is right that cognitive scientists are prone to a bias bias, then that gives us some reason to think that he is too, but if he is a victim of the bias bias then we have no reason to think that other cognitive scientists fall prey to a bias as he contends.

This paradox can also be generated without referring specifically to Gigerenzer's critique. In Sect. 1, we mentioned that researchers have currently identified a very large number of alleged cognitive biases. The sheer number of these biases (by some counts, over 100 cognitive biases and a few dozen social biases and memory biases[12]) may naturally lead some researchers to speculate that there has been some inflation in enumerating cognitive biases, and they might explain this by positing the existence of a bias bias.[13] Positing the existence of a bias bias seems to be a cogent claim at first sight, but cognitive scientists who would attempt to establish the prevalence of a bias bias among their fellow researchers need to demonstrate somehow that their research is not itself subject to that bias. If they posit the existence of a bias bias and think that it is pervasive among cognitive scientists, then it would be natural to think that their research is also subject to the same bias, which raises the specter of paradox.

[11] This differs from Brighton and Gigerenzer's 'Bias Bias,' according to which researchers are prone to attribute more importance to the existence of bias than to variance and noise (see Brighton and Gigerenzer 2015).

[12] These figures are based on the Wikipedia entry, 'List of cognitive biases.'.

[13] Several decades ago, Christensen-Szalinski & Beach (1984) analyzed citation patterns and concluded that research papers that claimed to find evidence for cognitive biases were cited significantly more often than ones that did not. They referred to this as a "citation bias" rather than a bias bias. They also surveyed 80 psychologists working in the field and found that they tended to remember the negative results better than the positive ones.

5 Objections and replies

An objector to our argument might say that in our initial presentation of the paradox, we ignored the social nature of scientific practice. In fact, no single experiment, paper, or argument is ever the final word in actual scientific practice. Consider Cimpian and Salomon's inherence heuristic. These researchers posited the existence of this bias in the journal *Behavioral and Brain Sciences,* which publishes long target articles and solicits many short replies (such as our commentary), to which the original authors then reply. While our account of the Self-Reflexive Bias Paradox is presented as though the evidence is being considered at the end of inquiry, scientific inquiry is not at an end. Thus, it may be claimed that the actual social nature of science, as manifested over the course of many interactions, solves the alleged paradox.

We are skeptical that the paradox admits of such an easy solution. Suppose that further research on the inherence heuristic directly replicates the findings from Cimpian and Salomon. We are left with the question of how to interpret these findings. One option is to posit a bias that is an inherent feature of cognition to explain it; another is to explain the finding by way of subjects' interaction with the world. If one research team claims that an inherent feature of cognition is the best explanation, it seems reasonable for a second research team to argue that if the first research team is right then that research team should be leery of their own reasoning process from the empirical findings to the existence of the bias. Here, we still have the paradox. The crucial point is that each reasoning move one research team might make in our initial presentation of the paradox is also a reasonable move for other members of the scientific community to make.

Someone might also object to our argument by distinguishing between two possibilities, namely exhibiting a reasoning error and being subject to a cognitive bias.[14] For example, in the case of the inherence heuristic, one might say that even if Cimpian and Salomon wrongly explain human behavior using inherent properties rather than relational properties, this need not demonstrate the existence of a systematic bias. The explanatory error could just be a one-off reasoning error, or an idiosyncrasy of those researchers. A bias is a *systematic* reasoning error exhibited by humans generally, or some significant subset of them, on a regular basis. Thus, it is possible to reject research supporting the existence of a bias on the grounds that the researchers made an incorrect inference without asserting the existence of a systematic bias.

In response, note that researchers who attempt to identify heuristics and biases are engaged in the effort to explain errors of human reasoning. The fact that the researchers exhibited a certain kind of reasoning error stands in need of explanation, and one has already been provided by the researchers in question: the newly discovered bias. Indeed, the researchers may point out that this is exactly what one should expect if the bias really exists and arises in that context. It is open to critics of the researchers to say that this was merely a one-off error and does not

[14] A variation on this objection would distinguish a reasoning error from the cognitive mechanism that is responsible for the error.

rise to the status of a bias, but the possibility of a bias cannot be easily dismissed and needs to be ruled out for some principled reason. If the researchers have posited the existence of such a bias, then the researchers should give us some reason to think that it does not affect their own reasoning process, and if their critics doubt these findings they need to explain them without citing the very same bias (at the risk of conceding its existence). Furthermore, in many of these cases, other members of the academic community seem to accept the biases in question. Theory-ladenness of observation is widely accepted, as was Dunning-Kruger effect until recently. The inherence heuristic is newer, but most of the replies to Cimpian and Salomon's target article on the subject accept the existence of the bias. Thus, if the objection to the existence of the bias holds, there would still be a mistaken inference pattern among a community of cognitive psychologists. It would be odd if it turned out that these cognitive psychologists were the only ones who exhibited such a pattern of mistaken inference for the reasons outlined in Sect. 2. Indeed, it seems ad hoc to claim that the mistaken inference pattern only holds for the researchers in question. One might reply to the charge of being ad hoc by offering an explanation that partitions the group of researchers off from the rest of humanity. Perhaps psychologists acquire the wrong inference pattern, which isn't a pattern within humanity at large, during their training as psychologists. Our rejoinder is that if this is a plausible hypothesis, it is nevertheless appropriate to think of such a pattern as a cognitive bias, albeit one that is limited to a subgroup of individuals.

When we consider the alleged bias bias, it is even clearer that the distinction between a mere reasoning error and a full-fledged bias cannot guarantee an escape from the paradox. Consider a researcher who is in the position of criticizing the field of cognitive science for an inflated tendency to posit the existence of biases to explain certain behavioral effects. Suppose that such a researcher presents empirical evidence to suggest that such a bias exists among cognitive scientists. Given that biases are dispositions for reasoning in a certain way or tendencies for making inappropriate inferences, this researcher would be pointing out a bias within the field of cognitive science. But in that case, it is difficult to see what grounds the researcher would have for absolving herself of the very same bias, and without such absolution, the claim of a bias would seem to generate a paradox. If we have reason to believe in a bias bias then we ought to doubt that this is a feature of our cognitive makeup, so we have reason not to believe that such a bias exists; and if we have reason not to believe in the bias bias, then we cannot discount the evidence for a genuine bias, and we have some reason to posit the existence of a bias bias.

Another objection concerns this paradox's practical import. Researchers who are aware of the self-reflexivity involved in investigating cognitive bias may say that in practice the paradox can be avoided in actual cases. Although the self-reflexivity may be shown to generate a paradox in certain unusual circumstances, since there is no danger of succumbing to it in the vast majority of real research contexts it can be safely ignored. After all, there are well-known techniques for debiasing (Lopes, 1987) and there are experimental manipulations designed explicitly to remove certain cognitive biases and prevent them from influencing our judgments and reasoning processes (Larrick, 2004). Thus, practicing cognitive

scientists have ways of ensuring that no paradox ensues when they are performing research on cognitive bias.

We would respond to this objection by pointing out that we have conceded that there are certainly cases in which the bias does not arise for the researchers themselves since the context of their inquiry is sufficiently different from that of their experimental subjects. Even when the contexts are similar, there are no doubt instances in which the researchers can demonstrate that they have employed methods designed to prevent them from succumbing to cognitive bias, including the bias under investigation. Nevertheless, we would maintain that there are cases in which there is some reason to think that the bias obtains and where the onus is on the researchers to show that they are not victims of the very bias that they claim to have uncovered. In these cases, it would seem as though the paradox persists: the very same reasons for thinking that the bias exists are ones for doubting that it does. This is not just a theoretical possibility but can also generate a paradox in inquiry for practicing cognitive scientists. For all we know, there may be untold other cases in which certain cognitive biases afflict our investigation of them in ways that are more covert.[15] Stein (1997, 560) points out that it is possible that cognitive scientists are relying on faulty reasoning principles that have not yet been discovered or deploying ones that have not yet been determined to be faulty. Hence, we cannot afford to be complacent. Moreover, the interest of these cases lies partly in the perplexing epistemic situation that they give rise to. The cases that we have been considering are not self-refuting, since the scientists making the claims in question are not strictly contradicting themselves, but they are 'self-undermining' (Stein, 1997). For example, the claim that people (including cognitive scientists) have a bias bias gives rise to a tenuous epistemic position, since anyone supporting the claim is thereby providing good reasons for doubting it.

6 Conclusion

Self-reflexivity of a cognitive bias alone does *not* generate an outright contradiction but it does lead to an apparent paradox, and the way in which the paradox is resolved (or fails to be resolved) in different cases sheds some interesting light on the relationship between theory and evidence in cognitive science. If a research group *S* has reason to believe that bias *B* exists (based on their own research findings) then *S* does not have reason to believe that *B* exists because their research is thereby undercut by

[15] Mugg (2020, 256) makes a similar point within the context of implicit racial bias: 'given that we cannot rely on introspection to assess whether implicit biases are manifesting and that we continue to find new areas in which they manifest, we have good reason to think that implicit biases manifest in ways yet unknown. We have no strategy for blocking access in such cases.' In her influential discussion of the epistemic implications of implicit social bias, Saul (2013) argues that the scope of cognitive bias is more restricted than that of social bias, taking probabilistic judgments as her primary illustration. However, we have argued here that other cognitive biases are not so circumscribed. Alfano (2014) and Carter & Pritchard (2017) both discuss the broader epistemic implications of cognitive bias, but neither discuss the possibility of a self-reflexive bias paradox.

the possible existence of the bias in their own research. But if *S* does not have reason to believe that bias *B* exists then *S* has reason to believe that *B* exists, since the researchers have no reason to doubt that their research establishes the existence of *B*. Thus, the self-reflexivity involved in investigating cognitive biases using human cognitive abilities can lead to a paradoxical result.

In examining four potential instances of the paradox, we found two ways that it can be blocked, and these are instructive in determining how pervasive this paradox is. First, the paradox is blocked if one can show that bias *B* exists only in some context *C*, which applies either only to the subjects or to the researchers, but not both. Implicitly, this is how Brewer avoids falling into the paradox when it comes to the theory-ladenness of observation. He suggests that the bias applies to his former research but not to the subjects drawn from the history of science that he was investigating (contrary to his previous claims). The paradox can also turn into a win–win situation for the researchers if one can distinguish the contexts, and furthermore, all that is at issue is the very existence of a bias rather than its prevalence. Second, if it is safe to assume that the bias is not pervasive and only arises in one of the two contexts, then it may be possible to claim that it is present either in the subjects or the researchers (but not both). This is how Dunning is able to claim that the D-K effect may be vindicated even by those who are critical of his research: either his research establishes the existence of the D-K effect in his subjects (but not in the researchers), or his research manifests the bias in the researchers themselves (but not in the experimental subjects). However, this line of reasoning can be rejected if one can come up with an alternative explanation for invalidating the research, which does not cite the very bias at issue but shows that it is due to some other error in reasoning (whether a one-off error or some other cognitive bias). If such an explanation is available to the critics, as in the D-K effect, researchers are not able to claim vindication either way, since the original research is allegedly flawed for other reasons. But if critics attempt to criticize research purporting to establish the existence of a cognitive bias by invoking an explanation that refers to a pattern of reasoning that resembles the one implicated in the proposed bias, as in the inherence heuristic or the bias bias, then the paradox may return. In such cases, critics can insist that the error on the part of the researchers arises not from some systematic bias but merely from a singular mistake or reasoning error, but if the error seems on the face of it to be an instance of the very bias in question, then the paradox resurfaces. At least, the onus is on the critics to explain the error in some other way than by citing the alleged bias. In the absence of such an alternative explanation, self-reflexivity and paradox threaten to return.

So how wide is the scope of the paradox? To generate the paradox, the bias in question must: 1) be applicable to scientific reasoning, 2) be difficult to avoid, and 3) arise in contexts that are similar in relevant respects to the context of the scientific research. We suggest that the more central a cognitive bias is to human reasoning, the more it will have these three properties. Some cognitive biases seem to arise in very specific contexts, and are therefore unlikely to result in a paradox. Many biases in probability judgments, such as the anchoring effect and base rate neglect, seem to fall into this category. On the other hand, those that concern explanations more generally, such as the availability heuristic and confirmation bias, may lead to the

paradox. Indeed, both the bias-bias and the inherence heuristic concern the generation and acceptance of explanations. It is also possible that research on some bias B_1 is subject to another bias B_2, which is in turn subject to B_1.[16] If such a circle of biases were to exist, which could comprise a larger number of biases B_1, B_2…, B_n, then that might broaden the scope of the paradox even further. We do not take this paradox to imply that the field as a whole is somehow undermined. Rather, our point is that it is important for those of us thinking about the existence of cognitive biases to ask to what extent those biases are affecting our inquiry into cognitive biases, and we should be open to the possibility that there is something deeply paradoxical in our study of this domain. We take such an admission to be an exercise in epistemic humility.

Acknowledgements We are grateful to audiences at the Society for Philosophy and Psychology annual conference (Ann Arbor, Michigan, 2018) and the Biases in Science conference (Munich, 2019) for helpful feedback, especially Momme van Sydow and Bennett Holman. We would also like to thank Brian Huss, Kevin Clark, and anonymous referees for very useful comments on earlier drafts.

Declarations

Ethical approval Not applicable.

Informed consent. Not applicable.

Conflict of interest None.

References

Alfano, M. (2014). Expanding the situationist challenge to reliabilism about inference. In A. Fairweather (Ed.), *Virtue epistemology naturalized* (pp. 103–122). Springer.

Antony, L. (2016). Bias: Friend or foe? Reflections on Saulish skepticism. In Brownstein & Saul (Eds.), *Implicit bias and philosophy.* Oxford University Press.

Bacon, F. (1620/1902). *Novum Organum.* P. F. Collier & Son.

Bergus, G. R., Chapman, G. B., Gjerde, C., & Elstein, A. S. (1995). Clinical reasoning about new symptoms despite preexisting disease: Sources of error and order effects. *Family Medicine, 27*, 314–320.

Blanco, F. (2017). Cognitive bias. In J. Vonk & T. Shackelford (Eds.), *Encyclopedia of animal cognition and behavior.* Springer.

Bogen, J., & Woodward, J. (1992). Observations, theories and the evolution of the human spirit. *Philosophy of Science, 59*(4), 590–611.

Brewer, W. F. (2012). The theory ladenness of the mental processes used in the scientific enterprise: Evidence from cognitive psychology and the history of science. In R. W. Proctor & E. J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes* (pp. 289–334). Oxford University Press.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research, 68*, 1772–1784.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*, 60–77.

Carter, J. A., & Pritchard, D. (2017). Cognitive bias, scepticism and understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 272–292). Routledge.

Christensen-Szalinski, J. J., & Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision literature. *American Psychologist, 39*, 75–78.

[16] We are grateful to an anonymous referee for raising this possibility.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences, 37*(5), 461–480.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4*(3), 317–331.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology, 44*, 247–290.

Evans, J. S. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford University Press.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics, 5*(3–4), 303–336.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*(1), 107–143.

Gilovich, T., & Griffin, D. (2002). Heuristics and biases: Then and now. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* Cambridge University Press.

Gladwell, M. (2005). *Blink*. Little, Brown & Company.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1–24.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Khalidi, M. A., & Mugg, J. (2014). The inherent bias in positing an inherence heuristic: Commentary on Cimpian & Salomon. *Behavioral and Brain Sciences, 37*, 493–494.

Knobe, J., & Samuels, R. (2013). Thinking like a scientist: Innateness as a case study. *Cognition, 126*(1), 72–86.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121.

Kruger, J., & Dunning, D. (2002). Unskilled and unaware—but why? A reply to Krueger & Muller (2002). *Journal of Personality and Social Psychology, 82*(2), 189–192.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Blackwell Publishing.

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica, 64*(2), 167–185.

Mercier, H. (2017). Confirmation Bias—Myside Bias. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (pp. 99–114). Routledge/Taylor & Francis Group.

Mugg, J. (2020). How not to deal with the tragic dilemma. *Social Epistemology, 34*(3), 253–264.

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

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., & Wirth, K. (2016). Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy: Advancing Education in Quantitative Literacy*. https://doi.org/10.5038/1936-4660.9.1.4

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy: Advancing Education in Quantitative Literacy*. https://doi.org/10.5038/1936-4660.10.1.4

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173–220). Academic Press.

Saul, J. (2013). Scepticism and implicit bias. *Disputatio, 5*, 243–263.

Shanks, D. (1991). A connectionist account of base-rate biases in categorization. *Connection Sciences, 3*(2), 143–162.

Simon, H. (1969). *Sciences of the artificial*. MIT Press.

Stanovich, K. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. University of Chicago Press.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2020). Intelligence and rationality. In R. J. Sternberg (Ed.), *Cambridge handbook of intelligence* (2nd ed., pp. 1106–1139). Cambridge University Press.

Stein, E. (1997). Can we be justified in believing that humans are irrational? *Philosophy & Phenomenological Research, 57*, 545–565.

Todd, P. M., & Gigerenzer, G. (2003). Bounding rationality to the world. *Journal of Economic Psychology, 24*(2), 143–165.

Vnuk, A., Owen, H., & Plummer, J. (2006). Assessing proficiency in adult basic life support: Student and expert assessment and the impact of video recording. *Medical Teacher, 28*, 429–434.