Preface: Rational Social Animals Go Wild

We live in polarized times. We are politically and ideologically divided, and our divisions seem to reflect, or be constituted by, divergent *beliefs.* As I write, the US is split into two camps that seem to have wildly divergent beliefs over the efficacy of facemasks. For one side, they are sensible precautions against the spread of COVID-19; for the other, they represent an outrageous infringement of civil liberties. The value of facemasks is just one topic that divides right and left over the virus and its risks. The two sides diverge on the origin of the virus, on the advisability of lockdowns and on the harmfulness of the disease, among other topics.

COVID-19 is just the latest front in an ongoing conflict that pits beliefs against beliefs. While the science of COVID-19 is young, and there's room to doubt whether *either* side has a right to much confidence in its views, on many of the issues that divide us the evidence is overwhelmingly on one side. Bad beliefs—beliefs that appear to be wildly at variance with the great preponderance of evidence—seem to animate opposition to vaccination and to the teaching of evolution. Much more consequentially—quite probably catastrophically—bad beliefs have played a central role in the world's failure to tackle climate change. Until very recently, US policy on climate change was decided by a president who thinks that global warming is a hoax. Together with Australia, Brazil, and other countries governed by climate change deniers, the US was able to stymie international efforts at tackling the problem. The world is already paying a heavy price.

This is a book about beliefs, good and bad, about how they are generated and how they might best be improved. Epistemology, the subdiscipline of philosophy concerned with beliefs and their justification, is ancient. Up to quite recently, however, modern epistemology was focused on theoretical questions; in particular, on the analysis of knowledge. It wasn't much concerned with the practical questions that will be

my focus. My concern isn't with the analysis of knowledge. Rather, it is with how knowledge is acquired and what factors lead to good and bad beliefs. Correlatively, my exemplars of belief will not be the uncontroversial cases that feature in a great deal of contemporary discussion; cases in which, say, one agent believes that another owns a car of a particular make, or that there's a barn hereabouts. Instead, my exemplars will be cases that are controversial but *shouldn't* be: beliefs about anthropogenic climate change, evolution, and the safety and efficacy of vaccines. These examples are chosen because there's an expert consensus on these issues, but many people reject the expert view. Are they rational in doing so? What explains their dissent? Should we attempt to change their minds, and if so, how should we do so? I'll defend some controversial answers to these questions about controversial beliefs.

Philosophers love definitions, so before I go on let me say a few words about what I mean by "bad beliefs." There are lots of ways in which beliefs might be bad. A belief might be morally bad (racist beliefs are bad in that kind of way). I'm not concerned with moral badness, but with epistemic badness; that is badness in the belief's relationship to evidence and to the world it aims to reflect. Epistemic badness itself comes in a variety of forms. One way in which a belief can be epistemically bad is by being false. My primary examples of bad belief are false: climate change denial, anti-vaxxer beliefs, creationism, and so on. But not all false beliefs are bad beliefs, in the sense I'm concerned with. I'm an atheist: I don't believe that any religion is true. But I don't think theists are bad believers in my sense. They're not bad believers because I think that religious belief can be rational: a thoughtful person who is familiar with the evidence for and against the existence of God can reasonably conclude that God exists.

A bad belief, in my sense, is therefore not (necessarily) a false belief, but an unjustified belief (a bad belief could even be a true belief, if the totality of evidence is misleading). That still isn't enough to pin down the kind of belief I'm concerned with, though: there are different ways in which beliefs can be unjustified, and I'm concerned with only one. A belief might be unjustified because it's not supported by the evidence available to the believer herself or because it's not supported by the totality of the evidence. Obviously, these things can come apart: a detective

might conclude her suspect is guilty of the crime based on nothing but prejudice and therefore believe badly (in one way) even if all the available evidence (including, say, the forensic report that she hasn't yet read) actually supports the suspect's guilt. Bad beliefs, in my sense, are not those that are subjectively unjustified in this kind of way, though.

A bad belief, in my sense, is a belief that (a) conflicts with the beliefs held by the *relevant epistemic authorities* and (b) held despite the widespread public availability either of the evidence that supports more accurate beliefs or of the knowledge that the relevant authorities believe as they do. The "relevant epistemic authorities" are those people and institutions that are widely recognized as being in the best position to answer questions in the domain: scientists are the relevant epistemic authorities when it comes to evolution; historians the relevant epistemic authorities on the Holocaust; and so on. I don't intend this characterization as a *definition* of bad beliefs (philosophers, save your counterexamples). It's meant to pick out a set of beliefs without begging some questions I want to leave open, such as the question whether bad believers can be subjectively justified in holding their beliefs. My aim is to explain why people come to hold beliefs that are bad in *that* kind of way. Why do they reject climate change, in defiance of the scientific authorities? Why do they reject vaccines, in defiance of the medical profession? And so on.

There are already many books and papers which aim to answer this question or questions that overlap very considerably with this one. Many of these books and papers argue that bad beliefs are explained (in important part) by the ways in which we humans are supposed to depart from some ideal of rational deliberation. On these kinds of views, bad beliefs are explained by a range of irrational (or *a*rational) psychological dispositions characteristic of human beings (albeit perhaps more pronounced on one side of politics than the other). According to views like this, people reject science (say) due to a need for stability, or out of an irrational respect for authority figures or (in more scientific language) due to the confirmation bias or a need for cognitive closure. These biases or dispositions are irrational or arational, though *having* them might itself be rational. The world is complex and time is short; often we must make decisions on the basis of limited information and before we can

properly think things through. It is often better to rely on processes that tend to get things right most of the time, even if they are not themselves rational. On a now standard story, we have evolved to think rationally only under certain conditions: when time permits and resources are plentiful and we're motivated to draw on those resources. These influential views explain bad beliefs by our tendency to make decisions using heuristics and shortcuts, reliance on which is inevitable and adaptive.

This book defends a very different view. It argues that bad beliefs are (in central cases) the product of genuinely and wholly *rational* processes. These processes are rational in the sense that they respond appropriately to evidence, *as* evidence. To say that they're rational is not to say that they always get things right (though given the link between evidence and getting things right, rational mechanisms tend to get things right). If the evidence is misleading, rational processes go wrong. If I get lost in a foreign city because the map I got from the tourist office has been tampered with by pranksters, my confusion is not due to any rational failing on my part. I accessed and processed information in a way that's beyond criticism (assuming there are no grounds for me to suspect tampering). The problem arises in the *inputs* into my navigational reasoning, not my reasoning itself. Analogously, I'll suggest that bad beliefs tend to arise—very significantly, at any rate—through the rational reasoning processes of those who end up believing badly. Just as I might find myself lost because someone tampered with the inputs into my navigation, so people end up believing badly because their epistemic environment has been manipulated.

Another way in which this book departs from more familiar views is that it heavily emphasizes social processes in the generation of knowledge. Knowledge, I'll argue, arises from distributed epistemic labor: epistemic labor distributed across space, time and across agents. Moreover, the knowledge thereby generated often is itself not an individual possession: it is parceled out across multiple agents, and even across the environment. It is *normal* and *rationally appropriate* for agents not to fully understand their own epistemic tools or the role they themselves play in the generation of knowledge, nor even the knowledge thereby generated. This is not just a limitation to which people like you and I are subject, because we're not scientists. Rather, it is the expected

upshot of the way cognitive labor is distributed, and scientists are and must be limited in just the same kind of way.

I'll argue that those who come to hold bad beliefs do so for roughly the same sorts of reasons as those who come to hold good beliefs. It isn't because *they're* irrational and *we're* not. It is largely because *we* defer to reliable sources of evidence and *they* defer to unreliable. This deference, which may be explicit or implicit, is itself rational on both sides. Given that we're epistemically social animals, it's largely through deference that we come to know about the world and generate further knowledge. The processes are much the same in our case and in theirs, and for the most part beyond reproach. Accounting for why some of us go astray in belief formation requires us to understand mechanisms of deference, the features of agents and the world that lead us to trust one source rather than another, and how testimony can be implicit as well as explicit. It also opens us onto the world: it requires us to scrutinize the features of the epistemic landscape and how that landscape can come to be epistemically polluted.

The rationality of bad belief formation has escaped recognition by philosophers, social scientists, and the general public, I suggest, because bad beliefs are so at odds with so much of the evidence. Climate change skeptics have beliefs that are at odds with the record of climate change and with well-established theories about the relationship between CO_2 and temperature. Anti-vaxxers have beliefs about the safety of vaccines that are at odds with the medical literature. And so on. I'll argue that nevertheless these beliefs are not at odds with the *higher-order* evidence. Higher-order evidence is evidence that concerns not the issues about which we're trying to make up our minds, but the reliability of the first-order evidence and how other people are responding to that evidence. Higher-order evidence is genuine evidence, and we rely on it all the time. But philosophers and psychologists overlook its pervasiveness and its significance. Once we come to see the ubiquity of higher-order evidence and the extent to which cognition is reliant on it, we'll be forced to rethink the extent of irrationality in human reasoning.

In effect, the argument I offer from high-order evidence parallels Cecilia Heyes' (2018) argument against nativist accounts of cognition. Nativists appeal to the "poverty of the stimulus" to motivate their

accounts: given that infants receive so little instruction and have so few examples to imitate, the acquisition of species-typical behaviors and capacities must be due to genes and not environment. Heyes argues that this is false: the stimulus is rich, not impoverished. She argues that the infant has ample opportunities for learning. I suggest that an analogous appeal to poverty underwrites arguments for pervasive human irrationality: given how impoverished the evidence for bad beliefs is, something other than rational response to evidence must explain their formation. In the experiments designed to demonstrate irrationality, first-order evidence for the beliefs adopted may be thin, but there's plenty of evidence nevertheless, and our responses are typically sensitive to it, I'll argue.

An account of knowledge production that emphasizes how the epistemic environment is saturated with higher-order evidence yields a distinctive account of how we best improve knowledge production. Bad belief has bad political effects, but it also has causes that are themselves political, in a broad sense of "political." We are (I'll suggest) epistemic individualists: we prize individual cognition and take it to be responsible for the great bulk of our cognitive achievements.[1] This individualism causes us to overlook or underestimate the need to attend to the epistemic environment: to the social mechanisms underlying knowledge production and the social cues that modulate deference. I'll suggest that understanding knowledge and belief requires combating epistemic individualism, and being more attentive to our environment and to the pollutants that have been allowed to accumulate in it.

A focus on the epistemic environment leads to different kinds of remedies for bad beliefs than those suggested by more familiar views. Deficit accounts of bad belief formation (a deficit of knowledge, of motivation, of rationality) suggest remedies that turn on correcting the deficit(s). If the deficit is in information, then we might improve beliefs by broadcasting the truth more widely. If it is in rationality, we might address it

[1] Who is the "we" here? The evidence is strongest with regard to WEIRD people; that is, those who live in the Western, Educated, Industrialized, Rich and Developed world. It is WEIRD people who are the participants in most psychological studies (in most research in most fields), and therefore less is known about other groups. There is some, albeit contested, evidence that other cultures are less individualistic than WEIRD cultures, though as we will see there is evidence that East and South Asians are also epistemic individualists.

through the education system (perhaps by teaching critical thinking). Alternatively, perhaps it would best be addressed by presenting information in a way that minimizes its potential threat to identity or to people's values. At least some of the approaches inspired by these explanations are valuable. What's not to like about improving the education system? I'm confident that some of these initiatives would actually improve people's beliefs to some degree. But I'll suggest that they should not be our sole, or even our main, focus. Rather, we should focus on improving the epistemic environment. That doesn't just mean we need to address what messages are circulating (this isn't the information deficit hypothesis in a new guise). The epistemic environment consists in much more than explicit messages. It consists in agents and institutions as well as messages, and the former may often be more significant than the latter.

Consider, for instance, the cues that we use to decide how much weight to give to testimony. Some of these cues are obvious: for instance, we weigh testimony by those we perceive as expert more heavily than testimony from those we perceive as less expert, and we weigh testimony from multiple sources more heavily than testimony from a lone individual. Only a little less obviously, we weigh testimony by those we perceive as sharing our values more heavily than from those we perceive as malevolent or as ideological opponents. Given these facts, one important way of improving people's beliefs is by way of attending to these kinds of cues. We can improve belief formation through what I will call *epistemic engineering*: the management of the epistemic environment. For instance, we might take care to ensure that people who lack expertise can't easily give themselves an unearned *appearance* of expertise.

Epistemic engineering raises significant ethical issues, of course. Aren't we manipulating others when we engineer the environment in this kind of way? To see the force of the objection, contrast such engineering with more traditional ways of changing minds: by giving reasons and presenting evidence. These more traditional ways are (surely) maximally respectful of agents and their rationality. In contrast, changing the ways in which cues for belief are distributed seems disrespectful at best, perhaps even subversive of agents' autonomy. The worry becomes even more pressing if a great deal of testimony is implicit, delivered not

via assertion but implicated by what is not said, and even by subtle features of the context. Manipulating such features, in the way I advocate, seems to be engaging in *nudging*, and nudging is hugely controversial.

I'll argue that in its typical guises, nudging (and therefore epistemic engineering) is unproblematic. It's controversial only because it's misunderstood. Both philosophers and cognitive scientists typically understand nudging as taking advantage of non-rational mechanisms; it's because nudges bypass genuine reasoning that nudging is controversial. I argue instead that nudges should be understood as implicit testimony. Being guided by a nudge is being guided by testimony, and there's nothing irrational about such guidance (here my defense of the rationality of the processes underlying belief formation becomes relevant to the assessment of the policies aimed at improving it). Hence nudging is usually respectful of agency, and questions concerning manipulation or epistemic paternalism can be set aside.

The Book: A Preview

Darwin called *The Origin of Species* "one long argument." This book, too, is an attempt at one long argument: it's designed to be read through (here, I'm afraid, comparisons with Darwin come to an end). Nevertheless, readers might appreciate a sense of the contents to come.

In Chapter 1, I introduce the topic of belief and belief formation, and set out the case for thinking that the quality of our beliefs is crucial to the quality of our social and political lives. I make this case against the belief skeptics: those who think that the beliefs that agents express play a smaller role in explaining their behavior than we might have thought. I then turn to existing explanations of belief acquisition and update (i.e, how they change over time), drawn from the social sciences. I argue that the influential deficit and motivated cognition accounts fall short of explaining how people come to hold entrenched views that conflict with settled science. In Chapter 2, I turn to a very different body of work in the cognitive sciences: work on cultural evolution. Drawing heavily on the so-called Californian school of researchers, I argue that we owe much of our success at colonizing a dizzying variety of environments to

cumulative culture, which embodies valuable knowledge. This knowledge, I suggest, is deeply social: it's the product of cognition distributed across many agents and across time, and it is never fully grasped by any individual. I then argue that contemporary science does not free us from heavy reliance on socially embedded knowledge production. Rather, if anything, it increases it: science, too, is the product of distributed cognition and individual scientists are never in a position fully to understand their own work.

Chapter 3 turns to distributed cognition in everyday contexts. I argue that the outsourcing of knowledge to others is routine for us. We take ourselves to be epistemically autonomous beings, but we form and update our beliefs very heavily through social referencing (looking to others, especially to those with whom we identify) and deference. These kinds of processes are rational, I'll argue: they're ways of responding appropriately to genuine evidence. They're also highly adaptive, though it's not that fact that makes them rational (it's the other way round). Having established that distributed cognition is more powerful than we tend to think, I turn in Chapters 4 and 5 to the converse question: how successful is individual cognition? In Chapter 4 I argue, contrary to what seems to be the consensus in epistemology and contrary to widespread intuition, that unaided individual cognition is highly unreliable. Without deference (to the right people to the right extent), we're epistemically at sea. In Chapter 5 I argue that we live in a polluted epistemic environment, which ensures that individual cognition fares even worse than it might've done. The focus on individual reasoning has led us to neglect this environment, I'll argue, thereby handing its management over to frauds and merchants of doubt.

In advocating attention to the epistemic environment, I'm advocating nudging, and nudging is highly controversial. In Chapter 6, I address this issue. I argue that nudging is not autonomy-subversive, as is often thought. It's not autonomy-subversive because it relies on mechanisms of deference—the same sorts of mechanisms that in earlier chapters I suggested were rational mechanisms. Nudging is in effect arguing, and being guided by nudges is being guided by (higher-order) evidence. Chapter 6 has an additional aim. Not only does it aim to show that nudges provide higher-order evidence: it also aims to show that the

cultural and social cues that in earlier chapters I argued are essential to human flourishing and to knowledge production themselves work through the provision of higher-order evidence. We orient ourselves and make decisions *centrally* by reference to higher-order evidence.

In a brief concluding chapter, I pull these threads together. Higher-order evidence is genuine evidence: in being guided by it, we're acting and thinking rationally. It follows, I argue, that much of the evidence commonly cited in support of the view that we are pervasively ir- or arational does not in fact support it. We've tended to conclude, on the basis of evidence that we're often responsive to cues and manipulations that don't involve the presentation of first-order evidence, that we're responding arationally (albeit adaptively). But we're more rational than we think: we're social and cultural animals, and we respond to the genuine evidence that our fellows and our cultural environment provide to us. Perhaps we're rational animals after all.

That's the agenda this book will pursue. Let me finish this introduction with a few words on methodology, and the sources of evidence that will guide my argument. I am a philosopher, working in that tradition of post-analytic philosophy that takes the sciences as exemplary (though not, to my mind, exhaustive) of knowledge production (this is a branch of what Eric Schliesser (2019) calls *synthetic philosophy*). The kind of philosophy I aim to engage in develops theories that systematize and interpret evidence from a broad range of sources, but especially from the cognitive sciences: cognitive and social psychology, the cognitive science of religion and work in cultural evolution. I engage in this kind of philosophy—call it naturalistic synthetic philosophy—because I believe it's more likely to generate knowledge about the kinds of questions I am interested in (here) than alternatives. That doesn't mean that I think other ways of doing philosophy are worthless. Far from it: Other approaches may be better for pursuing other valuable ends. Further, other ways of doing philosophy are often relevant to my project, and I'll draw on them when they are.

In particular, I'll draw on work in epistemology. Above, I mentioned that the focus of modern epistemology has been on the analysis of knowledge. But recently there's been a flowering of work in analytic epistemology focused on more practical questions. Analytic philosophy

differs from naturalistic synthetic philosophy in that while the latter takes the sciences as its most important source of evidence, the former relies heavily on the tools of conceptual analysis, the construction of thought experiments and the generation of counterexamples. Analytic epistemology has recently spawned social epistemology: epistemology concerned with the epistemic workings and effects of social interaction and institutions. I've already signaled my indebtedness to social epistemology by referring to testimony above. I'll draw on work in social epistemology and in analytic epistemology more broadly: for instance, work on higher-order evidence and on the epistemic significance of disagreement.

In a recent book, Nathan Ballantyne (2019) describes his project as an exercise in "inclusive regulative epistemology." Regulative epistemology is practical: it aims at guiding belief formation. Ballantyne's work is inclusive because unlike some other regulative epistemologists (he cites Bishop and Trout (2004) and Roberts and Wood (2007)) it does not aim to replace other methods, but instead draws on them. Ballantyne's project is an exercise in inclusive analytic regulative epistemology. Mine– in an even bigger mouthful– is an exercise in inclusive naturalistic synthetic regulative epistemology (don't worry—there'll be no call for me to use this phrase again).

Having situated the project on the philosophical field, let me now say something about its relationship to the cognitive sciences. In recent years, psychology has been rocked by a replication crisis: when experiments have been repeated, researchers have often been unable to reproduce the original findings. For instance, one group attempted to replicate 100 experiments previously published in high-profile journals, but succeeded in replicating only 41 (Open Science Collaboration 2015).[2] This crisis has made some philosophers reluctant to utilize evidence from psychology, and has led others to dismiss the entire field and the philosophy that draws on it. Caution is warranted, but dismissal is not.

[2] It's important to note, however, that these data are difficult to interpret. A failure to reach significance—the criterion they used for successful replication—might in many cases be due to the power of the replication attempt. It's also important to note that failures of replication may occur for many reasons, even when the underlying effect is real. See Earp (2016) for discussion of how we should understand this project.

While it's not strictly relevant to my project, let me first say a few words about why many areas of psychology have no problem with replicability. The replication crisis arises, in important part, from the use of sample sizes that were too small to rule out chance as a plausible explanation of the results reported (too often, unscrupulous or surprisingly ignorant psychologists took advantage of this fact to massage data in ways more or less guaranteed to produce statistically significant results; for example, by shelving unsuccessful experiments and simply repeating them until, by chance, they got the results they wanted). But small sample sizes are only a problem in some areas. In cognitive psychology, while sample sizes (in terms of numbers of participants) are sometimes tiny, tasks are often repeated a very large number of times. The high number of trials ensures that studies have an extremely high power, and are able reliably to detect small effects. The *p* value for significance—the threshold used to assess the likelihood that evidence against the null hypothesis is due to chance—is conventionally set at 0.05. Roughly, that is, an effect is taken to be (provisionally) established, or "significant," if the probability that we would see it by chance if there were no genuine relationship between the variables of interest was 5 per cent or less. One sign that some areas of science suffer from a serious problem is that there is a suspicious clustering of published results just below the cut-off for significance (Leggett et al. 2013; Masicampo & Lalande 2012); this is evidence that researchers have engaged in *p*-hacking—manipulation—to massage the data until it reaches significance (this can be done, for instance, by dividing the data in unprincipled ways until a subpopulation is identified for whom the finding is significant). But in cognitive psychology, tiny *p* values (e.g., $p < .001$) are not uncommon (see Scholl, 2017 for discussion).

But while much of psychology is untouched by the replication crisis, I can't take a great deal of comfort in that fact. Much of the work on which I'll draw comes from social and political psychology, which are ground zero for the crisis. In the absence of better evidence, I'll draw on this work freely, albeit carefully and reflectively.

To the extent I can, I'll rely on more recent work. Methodological standards have risen dramatically since the replication crisis first came to widespread attention. Many studies are now preregistered, which

dramatically reduces the risk of *p*-hacking (if I test for a hypothesis that differs from the one I registered, or split my sample in a way that was not motivated by the hypothesis I registered, this is now plain to everyone). Sample sizes have increased dramatically, increasing the power to detect genuine effects and lowering the risk of generating a chance finding. It is now easier than before to publish null results, and even when they are not published formally, such studies are now routinely made available online. This dramatic rise in standards ensures that newer work is more likely to be reliable than older. At the same time, we now have a better sense of which older work is especially unreliable. Researchers have developed statistical tests for detecting *p*-hacking and estimating true effect sizes in the light of the file drawer effect (the shelving of experiments when they failed to cross the magic $p = 0.05$ threshold, thereby ensuring that their results do not become part of the public record), which enables us to identify unreliable work. In fact, even without formal testing researchers often have a good sense of which work is reliable and which is not (Camerer et al. 2018). If a result seems too good to be true, it probably is.

We should be cautious in drawing on cognitive science. We should be attentive to effect sizes, to sample sizes, to replicability and to how well hypotheses cohere with other work. But we shouldn't be skeptical across the board. Naturalistic synthetic philosophy—or, at any rate, my version—is motivated in part by the conviction that evidence from the special sciences is routinely better than evidence from other sources. It's not always good evidence, and I'll approach it with a skeptical eye when warranted. But on many topics, it's a far better source of evidence than philosophers' intuitions, and I'll treat it as such. I'm confident that some of the work I'll cite will prove to be flawed, sometimes seriously. But the account I'll develop should be able to withstand such blows.

For all its use of empirical findings, this is an exercise in philosophy. It depends not on the science (directly), but on interpretations of the science and on philosophical argument. The general drift of the book is increasingly philosophical: the earlier chapters describe and interpret scientific work, and the later chapters engage much more in philosophical argument. The overall conclusions run contrary to widespread views within the cognitive sciences. If the account offered here is correct, we're

much more rational than psychology and naturalistic philosophy usually holds. But our rationality depends very heavily on others and on how we're embedded in epistemic networks.

Of course, individuals are predictably overfond of their own work and predictably limited in their capacity to assess it. It is through the scrutiny of the epistemic community that my account will be tested and its strengths and flaws revealed.