8

Evolutionary Explanations of Our Reliability

Sinan Dogramaci

PART I Setup

I. Introduction

It makes sense to think that evolution selects for reliable perceivers. Our ancestors wouldn't have lasted long if they couldn't see those hungry crocodiles floating in the river, or if they couldn't tell the delicious and nutritious blueberries from the nasty poisonous redberries.

Did evolution also select for reliable moral thinking? It seems much harder to feel confident that it did. What seems true is that human morality is, overall, evolutionarily adaptive. We think life and health are good, death and pain are bad, and even our sense of fairness and justice can, very arguably at least, be explained evolutionarily.[1] But can we say—and can we say it without viciously begging the question—that these *adaptive* views are *reliably true* views? That's what seems less clear.

It's easy to have the feeling that it's much *more* plausible that evolution could adequately explain the reliability of perceptual belief than that it could do so for moral belief. Many philosophers have said that they're comparatively confident that evolution explains our perceptual reliability while they doubt, or even argue it's impossible, that it could just as adequately explain our moral reliability.[2]

[1] See e.g., Joyce (2006), Kitcher (2011), and Churchland (2019). See FitzPatrick (2014, sec. 2) and Machery and Mallon (2010) for useful critical overviews.

[2] See e.g., Gibbard (2003, ch. 13), Schechter (2010), Joyce (2013), Street (2006, 2016), and Nagel (2012, ch. 5.3).

Sinan Dogramaci, *Evolutionary Explanations of Our Reliability* In: *Oxford Studies in Metaethics.* Volume 17.
Edited by: Russ Shafer-Landau, Oxford University Press.
DOI: 10.1093/oso/9780192865601.003.0008

My main aim in this chapter is to look closer at the apparent asymmetry in the prospects for an evolutionary explanation of our perceptual and our moral reliability. Is there really any good reason to think that evolution can explain our perceptual reliability while it can't equally well explain our moral reliability? I will try to show that the apparent asymmetry here is an illusion. I hope to show that, upon closer examination, the prospects for these two evolutionary projects are the same, at least as far as any philosophical considerations can show. One notable feature that I'll highlight in the two evolutionary explanations is the role that metasemantic assumptions must play in each of them. I'll try to show that the two explanations make equally plausible appeals to metasemantics.

My claim that the apparent asymmetry is an illusion raises a follow-up question: why is there this illusion then? So, my secondary aim in this chapter is to offer an explanation for what made it seem that evolution is better able to explain our perceptual reliability (given that it's not true). My suggestion will be that the illusion is due to the subtle way the truth predicate needs to be used as a logical device in the one explanation but not in the other (see Section X).

II. Why the Issue Matters Philosophically

The question of the reliability of our own beliefs isn't just an extremely interesting question taken on its own. How we handle the question has repercussions for the rest of our beliefs. Once the question of our reliability has been raised, we are forced to answer it affirmatively. If we doubt that we're reliable, or even suspend judgment or take a middling credence toward our reliability, then we are no longer justified in holding on to those beliefs. You can't justifiably think, "I believe lots of stuff (about this or that) but I also admit that my brain doesn't form beliefs especially reliably on this topic". At that point you have to give up the beliefs.

(Terminology: When I call a set of beliefs reliable I mean they are, by and large, true, and when I call a method of belief formation reliable I mean it tends to produce true output beliefs given true input beliefs, if any.)

So, we need to affirm that we're reliable. And that puts pressure on us to think there is some *explanation* out there of how we've succeeded in being reliable believers.

If you think there could be no explanation of your method's reliability, this admittedly may not *force* you to doubt its reliability (and thereby lose justification), but it certainly creates a serious amount of tension. There's tension because we're strongly inclined to suspect that if it's reliable, then there has to be some explanation of that. Our reliability is what Schechter (2010) calls a "striking" fact that "calls out" for explanation (p. 447), that is, it's a fact that we are strongly inclined to think must have some explanation.

So, there is an important explanatory challenge we face. If evolution can address that challenge, that would be a huge achievement.

III. Two Explanatory Challenges: The Operational Challenge vs. the Etiological Challenge

Schechter (2010) draws an important conceptual distinction between two senses in which we might explain the reliability of our belief-forming method.

First there is what he calls the operational challenge: how it is that this or that given method operates reliably? Explain how the method operates, or what *rules* it follows, or what *program* it implements, to turn true inputs into true outputs.

Then there is what Schechter calls the etiological challenge: *how did it come about* that we employ a reliable method? Explain how it happened that we've ended up using a reliable method.

I think Schechter gave us a very important disambiguation of the philosophical question, "What explains the striking fact that we use a reliable belief-forming method?" He's pointed out that there are two different striking things to explain. One thing that's striking is that the method we use is a reliable one—and the operational challenge demands an explanation. A separate thing that's striking is that we use some method that operates reliably—and the etiological challenge demands an explanation.

Schechter's distinction helps reveal what's harder from what's easier to answer in the challenge of explaining our reliability. What's easier is the operational challenge. Consider first the operational challenge for perception. How does perception operate reliably? The outlines of the answer are clear. There is a causal relation between the fact that *p* and our perceptual belief that *p*. The causal relation makes us reliable: a perceptual belief comes about *because* and (by and large, at least) *only when* it is caused by the fact. So, the operational challenge here is one we take ourselves to know how to answer for perception, even if only roughly and in outline.

It's important to see, now, that we can also answer the operational challenge for moral belief. This is a big pay-off of Schechter's conceptual clarification: we see that the operational challenge is *not* especially hard or perplexing for our moral beliefs. How does our moral belief-forming method operate reliably? The answer is that, since our moral beliefs (like mathematical and modal beliefs) can all be ultimately based on a set of *necessary* fundamental principles—the *basic* moral truths—then we can answer the operational challenge by saying that we were pre-programmed to believe those principles. Any creature that is innately disposed to believe a set of necessary truths will thereby be using a method that operates reliably. Some examples can illustrate the idea. How do you know torturing just for pleasure is wrong? Answer: since that fact is necessary, our innate disposition to believe that fact makes our method reliable here. And how do you know contingent moral truths, for example that Johnny's treatment of that pig is wrong? Answer: such contingent truths are contingent only because of their non-moral contingent component; in this example, what makes us reliable moral thinkers is that we combine a known necessary moral truth (about the wrongness of torture for pleasure) with our empirical knowledge of the non-moral and contingent fact that Johnny's treatment amounts to torturing the pig just for pleasure.

This sketch of how we answer the operational challenge shows something about the important issue of *circularity* as it arises in this area. A threat of circularity is, I believe, the main reason philosophers have to be suspicious of any attempt to give an evolutionary explanation of our moral reliability. But we can already easily see, from our discussion of the

operational challenge, that epistemic circularity can be both unavoidable and unthreatening. With the operational challenge, there is an epistemic circularity in our answers *both* for perception and for moral belief. We must *use* our method of perceptual belief formation to discover that we ourselves indeed are causally hooked up to the world in the right way. If we couldn't already trust our perceptual belief-forming method, we'd be completely unable to answer the operational challenge. Likewise, we *use* our method of moral belief formation to arrive at the knowledge that, say, taking care of your kids is something you ought to do. Knowing that that's the moral truth, we can then answer the operational challenge for our method of moral belief formation (by observing that we're all pre-programmed to believe that you ought to take care of your kids). Again, we'd be unable to answer the moral operational challenge if we couldn't already trust our moral belief-forming method. So, presuming none of us are skeptics, we can agree that in both cases the epistemic circularity is not a vicious circularity. I admit that the reasonable fear of a vicious epistemic circularity is a powerful force in many people's thinking about this topic, and it isn't easy to put it completely to rest. I will return to the issue of circularity again at the end of this chapter, after I've presented my arguments on how evolutionary explanations can address the etiological challenge.

But for now, we can at least see that, as Schechter already rightly said (2010, p. 445): "The cases of mathematics, modality, and morality are similar. For each of these domains there is a straightforward way to answer the operational question: Our cognitive mechanisms are reliable because they involve the employment of necessarily truth-preserving rules. The pressing challenge is that of answering the etiological question."

So, let's henceforth keep our focus on the etiological challenge. How did it come about that we use a reliable perceptual belief-forming method? How did it come about that we use a reliable moral belief-forming method? What explains the striking facts that we find ourselves using methods that, let's now grant that we know, operate reliably?[3] In particular, when does evolution provide the explanation?

[3] Yamada (2010) examines a similar property that he argues is a condition on knowing that *p*: "it is not an accident that one is using a truth-conducive method". Setiya (2012, p. 96), sympathetically develops that idea, and Schafer (2013) has a more critical discussion.

IV. Addressing an Initial Worry: Evolution Made Us, and We're Awful

It can easily look like it's obviously false that evolution can help explain the etiology of our moral reliability because, after all, hasn't evolution produced countless moral monsters? Even setting the behavior of wild animals aside, our own history is full of violence and oppression. If evolution bequeathed humans a natural innate inclination for immoral behaviors, for things like wanton violence and war, racism and misogyny, then how can we even indulge the question of whether evolution explains how we came to use a reliable moral method?

This is an understandable worry about my project. In response, the first thing I'll say is, look, I get it, I'm as horrified as you are. It's grim out there. But, you and I can see that things are so morally bad because we can, reliably enough, tell right from wrong. And *we* are the product of evolution. So, are we just lucky that evolution happened to send us in the right direction while it sent lots of other people in the wrong direction? Or, did you and I somehow overcome the instincts evolution tried to instill in us?

No, of course we're not just lucky, and of course we didn't overcome evolution's influence. You and I operate with the same set of basic inclinations as all other humans. As other philosophers and scientists have argued, there is a substantial moral code that is innate and universal among human beings.[4] Scientists are even finding that basic instincts similar to our own are present in other animals, especially the most intelligent ones.[5] You and I are operating from a shared moral starting point, one we share even with people who hold morally awful views, but somehow you and I manage to arrive at a more compassionate, cooperative, and egalitarian set of moral conclusions. This admittedly shows that evolution *alone* cannot explain why someone is morally reliable, but it leaves open the possibility that evolution *together with some set of enabling conditions* could explain why you and I are so morally reliable.

[4] See Mikhail (2011) and Hamlin (2013).

[5] See de Waal (2006), Tomasello (2018); see also FitzPatrick (2014) and Machery and Mallon (2010) for an overview.

And it's not too mysterious what these enabling conditions might be. We are in very favorable conditions that allow us to widely indulge certain moral instincts we share with others who are in unfavorable conditions. Everyone has deep-seated, innate tendencies to favor health and life, to favor reciprocal altruism and fair cooperation in social arrangements, and sometimes even to just be purely altruistic when it's no gain (and not much cost) to ourselves—we are all ready to move in the right direction. But conditions of scarcity will make any one of us more selfish about our own well-being and more willing to cruelly sacrifice or exploit the well-being of others; such unfavorable conditions force us to silence our more altruistic moral instincts, instincts that we'd indulge if only we could afford to. Favorable conditions are required for any human to craft a morally decent life starting from our primitive instincts. You and I, given our fortunate conditions, can be activists for that vegan lifestyle that would have won over zero converts anytime much earlier in human history. But, just as our evolved faculties for acquiring scientific and mathematical knowledge will lead to success only if we're not deprived of nutrition and security (as well as daylight, free time, data, and other favorable aspects of our circumstance), likewise our faculties for acquiring moral knowledge will lead to success only if we're not deprived of a sufficient degree of nutrition and security. And even after food, security, and free time are provided, we'd still need to bother to engage in moral thinking: even after we're in favorable conditions, we still need to pursue some reflective equilibrium before we'll get the reliable moral views that you and I share.

And even in conditions of extreme scarcity, when we are most willing to treat others badly, that resulting behavior, though we eagerly and correctly denounce it as immoral, still is really only behavior that is comparatively *more* selfish and *less* cooperative and *less* altruistic. Even such "selfish" behavior is among the things we should admit are comparatively *moral* within the huge range of conceptual possibilities. (The severity of human irrationality is similarly overblown. Yeah, it's common to flunk the Wason task on the first try, but it's a rare illness to believe you're Jesus.) Sharon Street (2006, p. 133) doubts that our genealogy led us to the moral truth because she thinks it could have led us *anywhere* in conceptual space: she imagines the possibilities in which we believe it's

morally paramount to firmly clasp your hands all the time, or to scream at the color purple, or to worship plants above animals. But it seems clear that evolution will not lead us to such strange moral views, so it is certainly still an open possibility that evolution could explain why we are morally reliable, or at least why our instincts are as reliably pointed in the right direction as they are. That's a striking enough fact to hope to explain.

The project is ambitious, but not too ambitious. I'm *not* trying to explain why, say, utilitarians are reliable (as opposed to Kantians or whoever). Philosophers who pursue reflective equilibrium reach slightly different views—and I won't say who's reliable or make any attempt to explain why. What I want to explain is only the striking reliability of the major stuff that we all agree on. This is stuff that I take it all evolved creatures would agree on, even creatures from another species who evolved on another planet—at least if they are similar to us in intelligence and sociality, and they are lucky enough to find themselves in those favorable conditions of non-scarcity. I assume (and will not argue here) that evolution would give any of us moral beliefs that favor life and health, and disfavor death, harm, and suffering (of course, with exceptions when they promote survival in the longer run). My aim is to see if evolution can explain our reliability about that much, and if it can do so in a way that's no worse (including no more viciously circular) than how well evolution can explain our reliability on perceptual matters.

PART II My Argument

V. A Simple Thought Experiment to Model the Issue

I now want to argue for my main thesis, which I'll state as a conditional:

Thesis: if evolution selects for the trait of having a reliable perceptual method, then it also selects for the trait of having a reliable moral method.

How can this be argued for? It doesn't seem easy to do. Evolution is complicated. Morality, belief, the status of our reliability and how we can

know it—these are all complicated and philosophically difficult topics. And as we'll see shortly, in both cases metasemantics plays a central role in the evolutionary explanations of both our perceptual and moral reliability, and of course we don't know the correct complete metasemantic theory.

So, my argumentative strategy here will be to start from as simple a model of the issue as I can come up with, and then I'll work, step by step, back to the original issue.

To begin, now, here is a thought experiment about a simple pair of cases. I'll then proceed to analogize the two cases to the evolution of our reliable moral and perceptual methods.

Case 1, Open Sesame:
Imagine an environment where we have some creatures, we have a protective cave whose only entrance is blocked by a boulder, and each creature will survive, let's suppose, iff it manages to enter the cave. The boulder will move aside iff the creature utters "open sesame".

Case 2, Traffic Stop:
This next case is just like case 1, except now the boulder has a traffic light attached to it, and the traffic light, at any time, displays one of its colors—red, green, or yellow. A creature will manage to enter the cave, and thus survive, iff it utters "red" if the light is then red, "yellow" if it is yellow, and "green" if it is green.

I intend for it to be equally intuitive, for each of the two cases, that evolution selects for the trait of using, as we can call it, a "signaling method" that successfully opens the cave. Evolution does not select for a successful signaling method in case 2 in some way that it fails to select for a successful signaling method in case 1.

Granting that intuition, what I aim to argue for now is that the two cases are suitably analogous to the evolution of our moral and perceptual methods so as to support my conclusion: if evolution selected for a reliable perceptual method, then it selected for a reliable moral method.

First I'll spell out the intended analogy a bit more directly (almost clumsily stating what must be obvious, I worry). Then I'll proceed to take

up individual points of intended analogy that I'll further support with arguments or clarifications.

VI. Spelling Out the Analogy a Bit

I'm imagining that, in case 1, Open Sesame, the creatures who signal "open sesame" (as opposed to, say, creatures who signal "abracadabra" or whatever) are analogous to creatures who believe they ought to care for their kids (as opposed to, say, creatures who believe they ought to eat all their kids). The cave in case 1 just requires the one signal, "open sesame". Successful creatures can be born pre-programmed to give this signal. There's no need for the creatures to *causally* respond to any further feature of the cave itself, in particular to any feature that varies across times or possibilities. This is analogous to morality and its foundation of timeless and necessary basic principles—the basic moral truths. To believe a basic moral truth, there's no need to causally respond to any particular momentary or contingent feature of the world. (I will assume the falsity of particularism, the view that morality is not based on a set of principles, certainly not a tractable set. I think this assumption might be ultimately dispensable to my project but it simplifies the presentation so tremendously much that I'm just going to make it.) Because the basic moral truths are timeless and necessary, creatures (like us) can be pre-programmed to innately believe the true basic moral principles (e.g., that you ought to take care of your kids) and thereby survive. The only causal sensitivity to varying local conditions will be sensitivity to the empirical non-moral factor in any situation (e.g., that these here are my kids, or that eating these berries will help my kids survive).

In contrast, the cave in case 2, Traffic Stop, requires casual sensitivity to the varying color of the light. What will be selected for is the trait of having a causally sensitive signaling mechanism. This is analogous to having a causally sensitive belief-forming mechanism, like our perceptual belief-forming mechanism.

Having agreed to move past the operational challenge (as adequately answered for both perceptual and moral beliefs, answered without mentioning evolution), we can now see how, as illustrated in the cases,

evolution could select for our having ended up with either of two different kinds of mechanism: evolution can select for a causally sensitive mechanism or for an insensitive, fully pre-programmed mechanism. In either case, evolution selects for a successful signaling method, and I intend to analogize from these successful signaling methods to our reliable belief-forming methods.

Now I'll consider a number of ways in which you might resist my analogy between the simple signaling cases and the actual evolution of our belief-forming methods, and I'll argue in defense of the analogy. Along the way, I'll try to explain why we are tempted to wrongly think evolution is better able to explain our perceptual reliability.

VII. The Analogy Between Evolution Selecting *Signaling-Behavior* and Its Selecting *Belief-Behavior*

The first point of analogy I'll comment on is that the given cases concern methods of *signaling*, but our real concern is with methods of *belief formation*.

A signaling method is clearly a trait that involves a creature's behavior, and it's clear how behavior can make a difference to survival and thus be selected for. Is there a worry that a belief-forming method is not like this?

No. Beliefs cause (fit or unfit) behavior. Perhaps there are subtle cases of different beliefs with extremely similar causal roles, but our main interest is in very plain and obvious sorts of differences in beliefs and behaviors, like the difference between believing you ought to take care of your kids and believing you ought to eat all your kids, and the behavioral difference between caring for them and eating them. I assume that those different beliefs make a big difference to your performing one or the other of those different behaviors, and thus a belief (or belief-forming method) can be selected for as straightforwardly as a signaling method can be. I likewise assume different perceptual beliefs (about colors, shapes, pitches, odors, etc., as well as all the possible relations among such things) make similar and obvious differences to behavior.

Another worry you might have here is that creatures who issue signals, though they might be a good analogy for our practice of issuing imperatives, aren't a good analogy for our practice of forming truth-apt beliefs. (The image of the creatures declaring "Open Sesame!" or "Red!", as if they're commands to the boulder to move aside, may suggest this.) But that's not a good reason to think the cases can't serve as analogies for the evolution of our truth-apt beliefs. The only feature of the signals that's relevant to evolution is their causal role. And that causal role can be shared by our declarations of imperatives, or our mental states of demanding or desiring something to be true, or (my intended analogy) our mental states of holding truth-apt beliefs in moral propositions. And, as I've said, our moral beliefs certainly have this kind of important causal role, since which moral beliefs you hold can make a big difference to your behavior and thereby to your evolutionary fitness.

VIII. The Analogy Between the Evolution of *Successful Signaling* and the Evolution of *Reliable Belief*

I claimed it's intuitive that in the two cases evolution equally well selects for a *successful* mechanism (whether it's a causally sensitive mechanism or not). But our real interest is not in successful mechanisms for signaling, but *reliable* mechanisms for belief formation. (A reliable mechanism is, again, one that tends to produce a correct belief or accurate representation, given the accuracy of any input it acts on.) So, does the analogy hold up here? Does evolution equally well explain how we end up using a reliable method in each case?

To see how it does, let's start with case 2 and the colors. I said that to get into the cave, the creature must give (that is, must utter) the signal "green" when the light is green, and so on. I called this the successful signal because intuitively getting into the cave is what counts as success here. In the description of the case, these signals have clear causal roles: the color shown on the cave's traffic light causes the creature to give its signal, and the creature's signal causes the cave to open. This is analogous to perceptual belief. A nutritious, or poisonous, berry may have a color that causes a perceptual belief in a creature, and that belief causes the

creature to eat, or avoid, the berry. Evolution selects for the trait that plays the role of a causal intermediary that promotes survival, whether by entering the cave or by appropriately handling the berry. So far, this shows that there is an analogy between a successful signaling method and a *successful* perceptual belief-forming method.

But, our question now: is such a *successful* perceptual method also a *reliable* method? You might worry that reliability is importantly different from such forms of success. This worry might come from the thought that, while I stipulated that "green" is the signal that succeeds when the traffic light is green, we could just as easily imagine things went differently where the successful signals were switched around, for example where uttering "red" opens the cave when the light is green. Does this suggest that the successful signal isn't necessarily the signal that accurately (i.e., reliably correctly) represents the displayed color?

No, I don't think there can be such a disconnect here between success and accuracy or reliability. That's because I think it's a constraint on any plausible proposal for the correct theory of metasemantics that success and accuracy are closely connected in such a way that, for example, the signal "red" would represent the color green if the signal "red" played the functional role of opening the cave when the light is green. The most obvious example of a metasemantic theory that closely connects evolutionary success and accuracy is teleosemantics, for example in a form such as Millikan's (1984) or Neander's (2017) theories, but I don't think I need to be committed to any of the controversial elements of the teleosemantics program. *Any* plausible metasemantic theory should agree with the minimal connections I'm taking for granted here. I take it to be very intuitive that, for example, "red" means green if it opens the cave when the light is green, and so on. (I hope you agree.)

Now let's turn back to case 1, Open Sesame. Can we say that evolution explains not only how it came about that we use a *successful* signaling method but explains how we came to use an *accurate* or *reliable* method (of signaling or, ultimately, belief formation)?

The intuitiveness of saying yes here seems to me just as strong as we saw that it was for case 2, Traffic Stop. Giving the signal "open sesame" opens the cave. Does this signal have a representational content? When

the signal is used to open the cave and allow the creature to survive, is the signal being used accurately? I say the answer is an intuitive yes.

We should ignore the fact that the English phrase "open sesame" is an imperative. Our creatures in both the cases issue *signals*. I picked the phrase "open sesame" for the case just to make it vivid and memorable, but it is no more (or less) essentially an imperative than the signal "red" is in the other case. If we can analogize from signals to beliefs (already argued for above), then the analogy should be equally strong in either case, and "open sesame" is a signal with a representational content, and its accuracy condition is just the condition in which that signal is successful. And that condition is the trivial condition that the creatures are always in.

There is, then, this difference between the signals in the two cases: the successful use of the "open sesame" signal will not be an *effect* that is causally sensitive to different conditions, the way the signal "red" was. But the "open sesame" signal is causally *efficacious*: it causes the cave to open. Does metasemantics require that a representationally significant signal, or belief, be triggered in a way that's causally sensitive to distinct causes in the environment? No, and part of my aim with these thought experiments is to bring out the intuitiveness of saying no here. Facts about representation are wholly grounded in facts about usage and causal role (what we call functional role), but we have enough usage and causal role facts in "open sesame" for that signal to be just as intuitively a representational signal as "red" is in the other case.

(And if it really is this one-sided causal role of moral beliefs that makes you doubt that we can explain our moral reliability, then the real source of your doubt seems to be an old-fashioned source of moral skepticism. It seems you're bothered because you want moral facts to cause moral knowledge like a traffic light causes perceptions. And that's not a good reason to be bothered anyway. It seems to rely on an implausible causal theory of knowledge.)

So, metasemantics, we now see, plays a major role in the explanation of why our creatures' signals, or beliefs, will have contents that represent their success conditions, and normally do so accurately. But weren't we asking whether *evolution* can answer the etiological challenge, whether *evolution* can explain how we came to use a reliable method? Yes. This is

an evolutionary explanation, one that appeals to metasemantics. And it has to. You can't tell any story about how the struggle for survival leads to the use of this or that kind of belief-forming method unless you rely on some claims about how the intentional arises from the non-intentional, some metasemantic claims. (We don't start our evolutionary story with any assumptions about the *intentional* states of the competitors in the struggle for survival.) What I've tried to argue here is that the most plausible such metasemantic claims are ones that will secure the reliability of our evolved creatures' signals, and ultimately their beliefs.

In this section, I've been arguing that the evolution of the use of a successful signaling method is a fair analogy for the evolution of the use of a reliable belief-forming method, and I've argued the analogy is as good for the Open Sesame case as for the Traffic Stop case. The fact that the successful signals in one case are not causally sensitive to different environmental conditions is not, I said, good reason to doubt how such signaling could evolve in the creatures, or good reason to doubt how we could evolve a method for forming reliable basic moral beliefs (which are also not causally sensitive to different environmental conditions).

Even if *all* that is granted, however, you might still have this worry: our basic moral beliefs are distinctive not only for being causally insensitive to the environment, but also for their distinctive roles, such as their special role in deliberation and planning and their special way of manipulating others and enabling social coordination with others. How can I explain these special features of our moral beliefs? My response to this worry is that it is not my aim to explain how we could have evolved concepts and beliefs that play these distinctive roles. I agree that these are certainly important features of our moral attitudes. One reason they're philosophically important is that they are among the factors that plausibly help to give our moral beliefs their distinctive moral contents, distinguishing them from each other and from our beliefs in other necessities, like our mathematical beliefs. But it's not my aim to explain any of the puzzling aspects of these other special roles our moral beliefs play, and it's not my aim to explain how, or in virtue of what, we come to have moral concepts and beliefs rather than, say, mathematical ones. My project is about how to explain the evolution of our moral *reliability*. Certainly, our moral beliefs have further philosophically intriguing

features beyond their reliability, but I leave it as a separate task to explain them. I trust they are compatible with the datum I want to explain, namely that we've come to use a reliable moral belief-forming method.

IX. The Analogy Between *Three Traffic Light Colors* and *Indefinitely Many Perceptible Properties*

I mean to analogize the Traffic Stop case to ordinary perceptual belief formation, but of course there is this huge disanalogy: we perceive more than just three or even a few colors.

There are many continua of color, shape, hardness, pitch, odor, and so on that we're able to perceive. Not only are there so many properties in the world that we can perceive, there are so many mental states we form in response to these environmental prompts. Maybe there are *infinitely* many different properties we can perceive (e.g., shades of colors, lengths and orientations of shape), and maybe there are even *infinitely* many different responsive mental states we could possibly enter into (though of course any human in any particular possible world will only enter into finitely many mental states). Or, slightly more cautiously, we can say that the range of properties we can perceive and the range of responses we can give is a huge open-ended range, a range of properties and perceptual responses that it's impossible for us to tractably enumerate in language. I'll work with that more cautious statement of the situation. I'll say, to put it briefly, that there are *indefinitely many* properties we can perceive and respond to.

What this means is that to have a fair analogy between Traffic Stop and perception, we should re-imagine the traffic light in Traffic Stop. Let's imagine it can display any (visible) shade of color, or whatever perceivable property you like, and imagine that surviving requires giving the right distinct signal corresponding to the displayed property, where the range of signals the creature might need to give is a huge open-ended range of responses. Now things are more analogous to our human abilities to perceive indefinitely many different shades, and to our perceptual and our linguistic/cognitive abilities to form indefinitely many different perceptual experiences and sentences/thoughts.

Now, when we re-imagine our model in this more realistic way, the analogy case for morality and the analogy case for perception exhibit a big difference between them—the first case involves only one easy-to-spot state of the world that the creatures must be responsive to, while the second case involves indefinitely many (maybe even infinitely many) different properties any one of which might be instantiated here now. Why don't we need to also adjust our analogy case for morality, Open Sesame? Why can our analogy case for morality only involve one signal? I say we don't need to adjust the analogy case for morality because, I'm assuming, there is a fixed, finite, and very tractable range of basic moral beliefs we actually hold, and so our analogous creatures can be imagined to only issue a small number of signals (e.g., one) to survive in the Open Sesame case. (Again, moral particularists may, in particular at this point, reject my line of argument. A more complicated presentation of my project would work with the assumption that particularism is less intuitive than moral generalism, but again I just assume particularism is false for simplicity.)

This difference in the size of the range of signals/beliefs involved should—I hope you'll agree—*not* make it seem problematic how evolution can answer equally well the etiological challenge for either case. If anything, the manageably small range of responses involved in the moral case puts it on *firmer* ground.

However, this is now also the place where the need to use the language of *truth* arises, and *that*, I will now suggest, is what causes much of the feeling that evolution can only answer the etiological challenge for perception, not for morality. Let me now explain this. (This point about the role of truth here is what inspired me to write this whole chapter, so I hope it's illuminating to some readers!)

X. The Confounding Role of "Truth" in Evolutionary Explanations

When there are indefinitely many possibilities that our creatures must be responsive to, we can no longer summarize what creatures need to do to survive—*unless*, that is, we use the generalizing device of the truth

predicate or some similar generalizing logical device. We can no longer just list off the things the creatures must do; we can't just say that they must signal "red" when the light is red, and say "100" when the light (or screen?) shows 100 dots (or whatever), on and on. We now have to say something general like: creatures must give whatever is the *correct* signal. There, in that last sentence, I just used the word "correct" in the same way that we use "true" as a device for generalizing an affirmation. This is the distinctive role of the truth predicate that has been highlighted by deflationists about truth.[6]

When we want to affirm, say, the law of excluded middle, we don't just want to affirm that it will rain or not rain, or that birds fly or don't fly, but the infinity of propositions of that form. The only way we can do this in natural language is to say something like, "Every instance of the law of excluded middle *is true*". (Although logicians have other technical devices to serve this purpose as well, these technical tools must be defined in natural language using the truth predicate, or some closely related similarly functioning predicate like "correct".)

The law of excluded middle has infinitely many instances, but we still require the truth predicate even when we want to affirm certain finite ranges of propositions, namely when we can't say just which, or how many, propositions go into the range. For example, to express total deference to the Pope's word, we must say something like, "Everything the Pope says is true", even though the Pope will only say finitely many things.

So, likewise, with our creatures who learn to use an indefinitely large range of representations and behaviors, we can intuitively perfectly well explain how their practice could evolve, but that explanation now must have us say the creatures came to have a mechanism for giving the *true*, the *correct*, or the *right*, or the *successful* signal for getting into the cave and surviving.

This need (the need to use a generalizing device of affirmation) is what gives rise, I conjecture, to the tendency to think that evolution intuitively explains how we came to have a perceptual mechanism for representing *the truth*, that is, a *reliable* perceptual mechanism. Look at how this

[6] See Stoljar and Damnjanovic (2014).

language features in philosophers' reports of the intuitiveness of how evolution can explain our perceptual reliability (my italics):

> We have a reliable visual mechanism because, very roughly, it conferred a heritable survival or reproductive advantage on our ancestors to *correctly* represent their environment using vision.
>
> (Schechter 2010, p. 444) (But Schechter himself *doubts* this is so for morality, p. 456.)

> [A] faculty [for detecting lions] would have had the function of tracking *the truth*. (Joyce 2013, p. 528)

> [Gibbard says that he seeks] a deep vindication of the capacities one exercises—an account of why beings like us would tend to get that sort of thing *right*.
>
> (Gibbard 2003, p. 256) (And, in his view, this is exactly what we have for perception but *lack* for moral thinking.)

> And since things with causal powers in one's immediate surroundings are the kinds of things that can affect one's survival and reproduction, we can see why selective pressures might well have led us to form *accurate* beliefs about them. (Street 2016, p. 322)

But, I urge, we should not see here any special connection between evolution's answer to the etiological challenge for perception and *truth* itself—except for our need to use a generalizing device to describe the mechanism that evolved. The need to use the generalizing device is just due to a fact about our cognitive and linguistic limitations: we can't tractably enumerate all the things a reliable and evolutionarily fit perceiver needs to perceive, or all the psychological responses they need to give. But there's nothing in the nature of evolution, or our perceptual mechanism, or the way facts about one can explain facts about the other, that involves any special role for truth or reliability itself.

With morality, we plausibly have just a small stock of basic principles (assuming particularism is false), and they may even reduce to something as tractable as, very crudely put, the claim that we ought to promote survival, especially of our own group (and that group, we enlightened people will hopefully agree, can encompass all sentient

beings). So we don't have any similar *need* to use the truth predicate. We don't need to say evolution explains how we came to have a mechanism for believing the *correct* things. We can instead say evolution explains how we came to have a mechanism for believing, crudely again, that we ought to promote survival, or whatever. But, my point here is that this difference amounts to no difference in evolution's ability to answer the etiological challenge in the two cases. It is only a difference in what expressive resources are called for when we describe the situations.

(If we assume our creatures would evolve desires for their own survival, we'd also want to say that they will evolve true beliefs, because, as functionalism tells us, actions based on *true* beliefs tend to be successful, i.e., they tend to fulfill desires—but this use of "true" is again just the generalizing use, and does not reveal a special connection that evolved perceptual beliefs have to the truth and which evolved moral beliefs lack. I've left desires out of my very simplified story of signaling creatures, but the story would go pretty much the same if we'd supposed the creatures acted on beliefs and desires, and the desires are for their own survival.)

The confounding role of "truth" also shows up, in an illustrative way I think, if we go back and think again about the operational challenge. One aspect of the perceptual case that I think some people get particularly misled by is that the perceptual method's operation involves causal and thus counterfactual sensitivity, and given the indefinitely large range of things the method must be sensitive to, we describe that as causal and counterfactual sensitivity *to the truth*. (See the talk of "tracking the truth" in the quote from Joyce above.) But, again, we put it that way, that is, we use the truth predicate, only because of the indefinitely large range of signals that the method involves. It's not because evolution is *selecting for truth* or *for reliability* in some special way, some way it isn't also doing so in the moral case.

I'm sure some readers will still insist: evolution is *selecting for truth* or *for reliability* in the perceptual case in a way that it isn't doing so in the moral case. But I hope I've shown that we need to at least be very careful that we're not being misled by our need to use the truth predicate. Look at the perceptual and moral cases again. In either case, evolution is, of course, selecting for having a method that gets you into the cave. In the moral case, this amounts to selecting for giving the signal "open sesame".

In the perceptual case, this amounts to selecting for giving . . . yes, the true signal, but, that is just shorthand for saying evolution selects for giving the signal "green" if the light is green, "red" if it is red, and so on, indefinitely. Evolution is not selecting *for truth* or *for reliability* here in a way that it is not also selecting for it in the moral case.

We could, if you want to, maintain that evolution is selecting *for the reliability* of the signaling method in the perceptual case—that's fine with me. What's important to me is that if we do say that, then we should also agree that evolution is selecting for the reliability of our signaling in the moral case too. I already argued above that evolution (with help from metasemantics) *can* explain why we use a reliable method, but it explains it *equally well* in either case, perceptual or moral, whether or not we want to call that "selecting *for* reliability". We've still seen no reason to think the etiological challenge is harder to answer for moral belief than for perceptual belief.

XI. The Counterfactual Robustness of the Etiology of Our Reliability

At this point, we're now in a good position to critically evaluate another common thought that moral debunkers are often tempted by. Some moral debunkers, it seems to me, think that evolution better explains the etiology of our perceptual reliability because that explanation has some counterfactual robustness. These debunkers emphasize the fact that evolution *would still* make us reliable perceivers even in other counterfactual possible worlds. And this, they'll point out, is a non-trivially true counterfactual, one with a possibly true antecedent:

(i) If the perceivable facts were very different from how they actually are, evolution would still explain our perceptual reliability.

They then contrast this with the moral case where we find the corresponding counterfactual to be either false or at best degenerately true, one with an impossible antecedent:

(ii) If the moral facts were very different from how they actually are, evolution would still explain our moral reliability.

That counterfactual certainly sounds wrong, and even if it's true that's only because it's degenerate.

The debunker's thought here is mistaken, it seems to me. We've by now granted that we use a moral method that operates reliably, and we use a perceptual method that operates reliably, though the means for successful operation differs in each case. One method is pre-programmed to believe the necessities, and the other one is built to be causally (and thus counterfactually) sensitive to contingent and varying conditions. But when we keep our focus on the etiological challenge, rather than the operational challenge, I see no epistemically important difference in the contingency or the counterfactual robustness of how evolution explains our coming to use such a well-operating method. Evolution will lead to our using our same well-operating moral method even in other worlds where we evolve, *and* evolution will lead to our using our same well-operating perceptual method even in other worlds where we evolve. We don't use a different perceptual method in other worlds we evolve in. We would use this same method we actually use, though perhaps applying it to different perceivable conditions. Why, though, should we especially care about those counterfactuals like (i) whose antecedents take us to possible worlds where all that differs are the perceivable facts (the color of a traffic light, and so on). And why should we care, at all or any differently, about those other counterfactuals like (ii) whose antecedents take us to impossible worlds where the moral facts differ? Across the (reasonably nearby) possible worlds where we evolve, evolution, together with metasemantics, is what explains why we evolved to use a reliable method, whether perceptual or moral—that's the only kind of range of modal space that seems relevant to answering the etiological challenge, the space of (reasonably nearby) worlds where we evolve. It's especially strange to think we learn something epistemically important when we look beyond the possible worlds, out in impossible territory, where everything becomes degenerate, including whether we would use a reliable perceptual method. So, I don't see any good reason why either of the counterfactuals, (i) and (ii) above, is more relevant than

the other, or really relevant *at all*, to the explanation of our adoption of one of these reliable methods than the other. The fact that one of the methods is for forming beliefs in contingencies, and the other method is for forming beliefs in necessities, does *not*, it seems to me, justify assigning any special epistemic significance to the one or the other counterfactual. Thinking those counterfactuals reveal something important seems to me a result of confusing the operational and etiological challenges, and not understanding how our reliability gets explained in response to either challenge.

Perhaps the way I've said things seem to me, in the previous paragraph, will not be how things seem to others. At the very least, if this kind of consideration (about counterfactuals like these) is the debunker's reason for doubting the explainability of the reliability of our moral method, at least quite a bit more would need to be said on their behalf, and I myself don't see what it could be.

XII. The Analogy Between Reliable Belief *About a Protective Cave* and Reliable Belief *About Moral Matters*: The Threat of Circularity Revisited

Finally, I now return one more time to the question of whether an evolutionary explanation of our moral reliability must be viciously question-begging.

I've used this simple Open Sesame model in order to argue that we can evolutionarily explain why we'd have a successful signal, and thus a reliable belief, *about something that it would be* ***adaptive*** *to have signals or beliefs about*. But in order for this model to provide any kind of helpful analogy to our actual *moral* beliefs, it may seem I must beg the main question at issue: it may seem I must be simply assuming that *the moral truth* is something *that it's* ***adaptive*** *to signal or to believe*. Does my model then viciously beg the question? Did I show how to evolutionarily explain our moral reliability only on the viciously circular assumption that the moral truth is something it's adaptive to believe?

My reply to this worry is this. I have not aimed to argue that evolution can explain our moral reliability in a completely non-question-begging

way. (That was already conceded in the discussion of the operational challenge.) What I have aimed to argue is that evolution is in no worse a position to explain our moral reliability than it is to explain our perceptual reliability. And, I now argue, the present worry about circularity *is equally applicable* when posed in the perceptual case. (And, in both cases, I find the circularity to be benign, not vicious, but arguing for that isn't part of my aim here.)

Consider, then, the corresponding worry for the perceptual case (presented here as a parody):

What does these creatures' having a reliable belief *about these properties of the traffic signal* have to do with *our actual beliefs about the colors of things*? You are simply *assuming*, in a question-begging way, that our actual color beliefs can be usefully modeled on analogy with these hypothetical properties of the traffic signal that, when you represent those properties, it gets you into the cave. You cavalierly *stipulated* that the properties of the traffic signal in your case are *color properties* (i.e., the properties we actually represent in vision), but it was question-begging to do this. For all we can show, our actual color beliefs *have nothing to do with any real properties in the world* like the traffic signal's real properties.

(This is, in fact, more or less what Hoffman (2019) absurdly claims.)

I don't think this is a reasonable worry about the perceptual case, but all I want to argue now is that our ability to answer such a worry in the perceptual case is no better than our ability to answer it in the moral case.

In our thought experiment about the colors, we could have *tried* to avoid begging any such questions by only saying the traffic light exhibits some unspecified properties F1, F2, and F3, and the creature must issue certain signals S1, S2, and S3, respectively, in order to get into the cave. My skeptical opponent may think it was tendentious when I called F1, F2, and F3 *color* properties and I then went on to argue that evolution will select for using signals (S1, S2, and S3) that thus represent *colors* accurately. Now, *I* don't think this was a tendentious way to describe the model, because in real life, our minds represent the world as

containing colors, and we take it (and reasonably so, I'd say) that these representations are largely accurate—the world really does contain colors. But, whether or not you see a vicious circularity in my claim that evolution will lead us to accurately represent the real colors I take there to be in the world, what's important here is that we're in no worse an epistemic position to say the same things about moral properties and the selection of a mechanism that reliably represents them.

I assume (tendentiously or not, though I'll point out that Street 2006 also begins with this assumption) that there *are* moral properties out there in the world, and there are moral truths. I then intended to sketch how evolution could select for our accurate representation of these. Does my sketch of an evolutionary model really apply to our actual moral beliefs and moral method? Well, just think about our *actual* moral beliefs. I've tried to bring out how *they do* resemble the "open sesame" beliefs of the story. I take it to be plausible, and I've tried to invite you to agree, that our actual moral beliefs play a metasemantic role much like the "open sesame" cave beliefs in the thought experiment. (These beliefs have an important causal effect, while they have no contingent external causal triggers—just like the "open sesame" beliefs. And again I bracket from my project, which is focused on our reliability, any other special roles played by our moral beliefs.) Granting me that we have moral beliefs, granting me that there are moral properties and truths in the world, and granting me that our moral beliefs play a role much like the "open sesame" beliefs, then my skeptical opponent could only be raising the skeptical possibility that what we morally believe and what's morally true are out of sync. But *that* skeptical possibility is just the corresponding case to what we've already considered and dismissed. In the color case, it's the possibility that our color beliefs and the real colors of things are out of sync, and in the moral case, it's the possibility that our moral beliefs and the real moral properties things have (properties I simply assume some things out there have) are out of sync. I already addressed this out-of-sync possibility when we looked earlier at how the metasemantics plausibly has to go (when we imagined the thing we call "red" is green and so on, and I said the metasemantics won't plausibly allow that to happen). So, I don't think we beg the question any worse in the moral case than we did, and had to, in the color case.

Let me directly address one last way that I often hear the accusation of question-begging put against the defenders of moral realism. I often hear moral debunkers make the accusation that, in order to give an evolutionary explanation of the reliability of our moral beliefs, we'd need to *state the moral truths* (that life is good, death is bad, and so on) in the course of giving that explanation, whereas we don't need to state any facts about the colors of things to evolutionarily explain our perceptual reliability. That complaint, whether or not it would have had any merit if it were true, simply is not true with regard to the evolutionary explanation of our moral reliability that I've sketched here. I did not state the moral truths in presenting how moral beliefs evolve. I only said that certain kinds of beliefs (ones like "open sesame" signals) would be moral beliefs, and metasemantics will see to it that they come out reliable. We can go on to ask what content these reliable moral beliefs have (they will have contents like what we evolved humans morally believe, i.e., that life is good, death is bad, and so on), but we have already explained these beliefs' reliability before even getting to the question of their particular content. So, once again, there is no epistemically significant asymmetry between the evolutionary explanation of our moral reliability and our perceptual reliability.

XIII. Conclusion

Of course, the whole story about how metasemantics works with our actual moral vocabulary is complicated. Likewise for our use of color terms. My aim here is only to present the two cases in a simple way where we find no forthcoming reasons to doubt, on philosophical grounds, that evolution is equally well able to answer the etiological challenge for our perceptual and moral belief-forming methods. If there is some asymmetry between morality and perception, or between Open Sesame and Traffic Stop, some asymmetry that suggests evolution can only explain things in some importantly different ways between the two cases, what is the difference? I've found none.

One difference I have found, however, is a difference in the role of the truth predicate in the two cases. And this difference may be what

misleads us into wrongly thinking evolution selects for reliability in some special way in the perceptual case.[7]

Bibliography

de Waal, Frans, 2006, *Primates and Philosophers: How Morality Evolved*, Princeton University Press.

Churchland, Patricia, 2019, *Conscience: The Origins of Moral Intuition*, W. W. Norton.

FitzPatrick, William, 2014, "Morality and Evolutionary Biology", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/morality-biology/.

Gibbard, Allan, 2003, *Thinking How to Live*, Harvard University Press.

Hamlin, Kiley, 2013, "Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core", *Current Directions in Psychological Science* 22(3): 186–93.

Hoffman, Donald, 2019, *The Case Against Reality*, W. W. Norton.

Joyce, Richard, 2006, *The Evolution of Morality*, Cambridge University Press.

Joyce, Richard, 2013, "The Evolutionary Debunking of Morality", in J. Feinberg and R. Shafer-Landau (eds.), *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, 15th edition, Cengage.

Kitcher, Philip, 2011, *The Ethical Project*, Harvard University Press.

Machery, Edouard and Ron Mallon, 2010, "Evolution of Morality", in J. M. Doris (ed.), *The Moral Psychology Handbook*, pp. 3–46, Oxford University Press.

Mikhail, John, 2011, *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge University Press.

Millikan, Ruth Garrett, 1984, *Language, Thought, and Other Biological Categories*, Cambridge University Press.

[7] For help preparing this chapter, I would like to thank Matt Vermaire and audiences at MadMeta, Boulder, Toronto, Harvard, and Texas.

Nagel, Thomas, 2012, *Mind and Cosmos*, Oxford University Press.

Neander, Karen, 2017, *A Mark of the Mental*, MIT Press.

Schafer, Karl, 2013, "Knowledge and Two Forms of Non-Accidental Truth", *Philosophy and Phenomenological Research* 89(2): 373–93.

Schechter, Joshua, 2010, "The Reliability Challenge and the Epistemology of Logic", *Philosophical Perspectives* 24: 437–64.

Setiya, Kieran, 2012, *Knowing Right from Wrong*, Oxford University Press.

Stoljar, Daniel and Nic Damnjanovic, 2014, "The Deflationary Theory of Truth", *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2014/entries/truth-deflationary/.

Street, Sharon, 2006, "A Darwinian Dilemma for Realist Theories of Value", *Philosophical Studies* 127(1): 109–66.

Street, Sharon, 2016, "Objectivity and Truth: You'd Better Rethink It", in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics Volume 11*, pp. 293–333, Oxford University Press.

Tomasello, Michael, 2018, *A Natural History of Human Morality*, Harvard University Press.

Yamada, Masahiro, 2010, "Getting It Right by Accident", *Philosophy and Phenomenological Research* 83(1): 72–105.