HPLS (2022) 44:23
https://doi.org/10.1007/s40656-022-00505-y

ORIGINAL PAPER

The evolution of moral belief: support for the debunker's causal premise

Michael T. Dale[1]

Received: 23 March 2021 / Accepted: 22 April 2022

Abstract The causal premise of the evolutionary debunking argument contends that human moral beliefs are explained by the process of natural selection. While it is universally acknowledged that such a premise is fundamental to the debunker's case, the vast majority of philosophers focus instead on the epistemic premise that natural selection does not track moral truth and the resulting skeptical conclusion(s). Recently, however, some have begun to concentrate on the causal premise. So far, the upshot of this small but growing literature has been that the causal premise is likely false due to the seemingly persuasive evidence that our moral beliefs are in fact *not* the result of natural selection. In this paper, I argue that this view is mistaken. Specifically, I advocate the Innate Biases Model, which contends that there is not only compelling evidence for an evolved cognitive capacity for acquiring norms but also for the existence of an evolutionarily instilled set of cognitive biases that make it either more or less likely that we adopt certain moral beliefs.

Keywords Evolution · Debunking · Causal premise · Innate biases model · Ethics · Metaethics

✉ Michael T. Dale
michaeldale@utexas.edu

[1] The University of Texas at Austin, Austin, USA

Published online: 01 June 2022

1 Introduction

According to the evolutionary debunking argument (EDA), the possibility that our moral beliefs were to a large extent shaped by natural selection[1] spells trouble for the moral realist.[2] Here are the main premises of the argument (Cf. Kahane, 2011):

> *Causal premise*: Our moral beliefs are explained by the process of natural selection.
> *Epistemic premise:* Natural selection does not track moral truth.

Importantly, these premises (and the resulting conclusions) are unpacked in different ways for different versions of the EDA. For example, Joyce (2006, pp. 108–142) claims that our moral beliefs are the result of an innate moral capacity that evolved to foster long term cooperation. If this is true, though, then that means that our moral beliefs can be explained without invoking the existence of moral truths. This makes such truths explanatorily specious, and, as a result, our moral beliefs are no longer justified.

Street (2006), on the other hand, claims that the specific *content* of our moral beliefs was shaped by natural selection. For example, the belief that "one ought to help one's kin before helping a stranger" was shaped by tendencies that came about as a result of natural selection.[3] Of course, that idea in and of itself doesn't necessarily mean that our moral beliefs are false. Indeed, our perceptive beliefs were also shaped by natural selection, but we generally view them as (at least somewhat) truth-tracking. However, according to Street, these accounts are not analogous because we do not have a reason to think that it would have been fitness enhancing for our moral beliefs to be true in the same way that we have a reason to think that it would have been fitness enhancing for our perceptive beliefs to be true (i.e. it would have been fitness enhancing to have true beliefs about where cliffs and predators are). In particular, she writes:

> Exactly why would it promote an organism's reproductive success to grasp the independent evaluative truths posited by the realist? The realist owes us an answer here. It is not enough to say, ''Because they are true.'' We need to know more about why it is advantageous to apprehend such truths before we have been given an adequate explanation (2006, pp. 129–130).

For Street, such an "adequate explanation" does not exist and, as a result, our moral beliefs are probably not true.[4]

[1] Natural selection is the process whereby organisms with phenotypes better adapted to their environment tend to pass on more genes to succeeding generations than organisms with phenotypes that are less well adapted. Fitness is the quantitative representation of this process.

[2] A moral realist can be understood as someone who believes that there are moral truths that hold independently of our moral beliefs and attitudes.

[3] In the case of this particular belief, the tendencies that shaped it probably came about as a result of kin selection.

[4] For a version of the EDA that leans more towards outright error theory, see Ruse (1998).

Yet no matter how the premises (and conclusions) are unpacked, in every EDA there is always some account of how our moral beliefs are explained by the process of natural selection. Which is to say, there is always some version of the causal premise—and this makes sense, as such a premise is crucial. Interestingly, though, most philosophers opt to pass over this premise in order to engage with the epistemic premise and the resulting conclusion(s). They do this by assuming for the sake of argument that the causal premise is true, even though most of them—debunkers and realists alike[5]—agree that there is probably not enough evidence to support such a premise.

Recently, however, philosophers have begun to focus on the causal premise.[6] So far, the upshot of this small but growing literature has been that the causal premise is likely false due to the seemingly compelling evidence that our moral beliefs are in fact *not* the result of natural selection.[7] In this paper, I disagree with both majorities. That is, I disagree with the overall majority view that there is not (yet) enough evidence to indicate whether the causal premise is true or false. Furthermore, I disagree with the majority of the (admittedly very limited) empirically focused literature that contends that once the evidence is looked at, the causal premise is likely false. Which is all to say, I argue that there is compelling evidence to indicate that it is in fact true that our moral beliefs are to a significant extent shaped by natural selection, and I do this by advocating the Innate Biases Model (IBM).[8]

The specific plan of the paper is as follows. In the second section, I explicate the IBM, which consists of explaining what it would mean to have innate biases, as well as a cognitive capacity that enables us to acquire norms. In the third section, I show why previous arguments attempting to show that our moral beliefs are *not* the result of natural selection are unconvincing and in so doing present evidence and argument in support of the IBM. In the final section, I briefly discuss the philosophical implications of using the IBM in support of the EDA.

[5] Some quotes: "[I]t must suffice to emphasize the hypothetical nature of my arguments, and to say that while I am skeptical of the *details* of the evolutionary picture I offer, I think its *outlines* are certain enough to make it well worth exploring the philosophical implications" (Street, 2006, § 3). "[T]he evolutionary explanations in the above examples, as plausible as they may sound, are a long way from even beginning to fill out the empirical details needed to fully secure this premise" (Kahane, 2011, p. 111). "Reconstructing the *actual* history of [our moral past], from its beginning to the present, is plainly beyond the evidence available…[The] data are too sparse to screen out rival hypotheses about the sequence of events" (Kitcher, 2011, p. 11). "No one, not even the debunker, thinks [the empirical claim] is conclusive" (Vavova, 2014, p. 79).

[6] See Deem (2016), Fraser (2014), Isserow (2019), Levy and Levy (2020), and Machery and Mallon (2010).

[7] Deem (2016), Fraser (2014), Isserow (2019), and Levy and Levy (2020) all argue that the empirical evidence does not support the debunker's case.

[8] Of course, I won't be arguing that the evidence is *conclusive*, but I do believe that the evidence is at the very least *suggestive*.

2 The Innate Biases Model[9]

In his paper discussing three possible hypotheses that a moral nativist could adopt, Sripada (2008, pp. 332–340) puts forward what he calls the *Innate Biases Model* (IBM). According to this model,

> [S]ome element of innate structure…serves to make the presence of some moral norms…*more likely* relative to the case in which the bias is absent. For example, the widespread presence of certain kinds of moral norms across groups may be explained in terms of the fact that the innate structure of the mind is predisposed to *favor* the emergence or propagation of these norms or predisposed to *disfavor* the emergence or propagation of other competing norms (p. 332).

To help explain how this model functions on a more concrete level, Sripada argues extensively for why incest avoidance is best understood through the IBM. According to the Westermarck hypothesis, humans have an innate aversion against engaging in incestuous behavior (Westermarck, 1922).[10] In the context of evolution, such an aversion makes adaptive sense because inbreeding often leads to deleterious fitness consequences (Durham, 1991). Indeed, most nonhuman animals do not engage in incestuous behavior and this is likely due to the genetic risks of such behavior (Keller & Waller, 2002; Pusey & Wolf, 1996).

Among human populations, some of the most compelling evidence of such an aversion comes from Israeli kibbutzim. In these communal villages, genetically unrelated children were raised together, and even though they were never told that romantic or sexual relationships between members of the same group were incestuous or to-be-avoided, sexual intercourse and marriage was exceedingly rare among members of the same group. Indeed, many studies could not find a single instance of a non-platonic relationship (Durham, 1991; Shepher, 1983; Spiro, 1958). According to Sripada, the IBM provides a compelling explanation of this intriguing phenomenon. In particular, if evolution instilled an innate bias in humans to avoid sexual relationships with genetically related conspecifics, and the way in which we pick out who is a genetic relation is by automatically labeling those we are raised around as genetic relations,[11] then such a mechanism would often correctly pick up on our true genetic relations but would sometimes "misfire" in the cases in which we are raised alongside genetically unrelated individuals. And it is such "misfiring" that supposedly occurred among children raised in Israeli kibbutzim.[12]

[9] It's important to remember that the primary objective of this section is to ***explain*** the IBM through one particular example. Putting forward a full empirical case in support of the model doesn't come until Sect. 3.

[10] For an updated defense of the Westermarck hypothesis, see Wilson (2019).

[11] Such a labeling process makes sense in the context of natural selection, for it would be extremely difficult (and costly) to instill a mechanism that can in some way *sense* or *pick up on* who is actually a genetic relation.

[12] While this interpretation is fairly well supported, it isn't entirely uncontroversial. For an alternative interpretation, see Cofnas (2020).

Of course, if this were the only evidence about incestuous behavior (or lack there-of), then we would have little reason to suppose that there is an *evolutionarily* instilled bias against incest. Indeed, for such a case to be made, such a bias needs to be (fairly) universal.[13] Fortunately for the IBM, there is overwhelming evidence of such universality, as almost all recorded human societies have some type of moral norm prohibiting incest (Murdock, 1949; Sripada, 2008). Thus, what is supposedly occurring—according to the IBM—is that an innate bias against incestuous behavior is contributing to the formation of certain norms, taboos, and laws against incest.

Some critics, however, are quick to point out that norms about incest are quite variable between cultures (e.g. Levy & Levy, 2020; Prinz, 2007). Indeed, while almost all incest norms extend beyond the nuclear family, they sometimes stop short right outside of that (e.g. the Bedouins of the Arabian Peninsula encourage first cousin marriage) while at other times they extend as far as the entire tribal unit (Murdock, 1949). But, interestingly, such evidence is actually *supportive* of the IBM—at least insofar as it is a rival to other innateness hypotheses. If the hypothesis were that evolution had selected for the *specific content* of our moral beliefs,[14] then this variation in incest norms would be a significant strike against the hypothesis. Indeed, there does not seem to be any specific incest belief that is universal.[15] Importantly, though, the IBM allows for such variation. If the instilled mechanism is an innate *bias*, then there would likely be some type of universal aversion to incest, but the specifics of the aversion would be manifested differently across cultures, depending on the history and etiology of the particular culture. For example, due to more of an emphasis on the importance of royal bloodlines (which would be a culturally—as opposed to an evolutionarily—instilled norm), one society might encourage first and second cousin marriage. However, in another culture, it might be the case that ancient religious stories tell of angry spirits that visit people who engage in sexual relations with *anyone* known to be a genetic relation.[16]

[13] Some are skeptical of using universality as evidence of an evolutionarily instilled mechanism, and such skepticism isn't entirely unfounded. Indeed, all known cultures control and use fire, but this does not mean that we have evolutionarily instilled mechanisms for fire use. Nevertheless, universality is still accepted as a fairly good indicator of the existence of an evolutionarily instilled mechanism, especially when it comes to mechanisms related to moral norms (including biases and beliefs). For a full defense of why this is the case, see Machery and Mallon (2010) and Pölzler (2018).

[14] This view is very close to the one advocated by Street (2006). More on this in the next section.

[15] Except, perhaps, something along the lines of "don't have sexual intercourse with anyone that is *at the very least* a member of your nuclear family." Though even this possibility comes off as more of a bias than a specific belief.

[16] And, just as it did in the Israeli kibbutzim, it would make sense that such a bias sometimes "misfires." That is, it sometimes leads to incest norms that aren't necessarily optimal in the evolutionary sense, such as when it leads to norms that target non-genetically connected relationships. Of course, an important question to ask here is: how many "misfirings" is too many? That is, when exploring the anthropological evidence of incest norms, at what point can it be said that the IBM can no longer account for the data and a more culturally focused explanation is needed? Indeed, if it were the case that *most* or even *a significant number* of incest norms "misfired" insofar as they prohibited sexual relationships between non-genetically related individuals, then I would be willing to admit that the IBM is perhaps less compelling than alternative cultural explanations. However, as far as I know, while there are certainly a handful of examples of such "misfirings" found in the literature (e.g. Henrich, 2015; Cofnas, 2020) the overwhelming consensus supports the idea that *most* incest taboos do indeed target genetically connected relationships.

At this point, the IBM should be coming into focus, but some may still wonder what exactly is meant by a *bias*. Unfortunately, Sripada does not delve into much detail about this, but if we turn to Nichols's (2004) *Affective Resonance Account*, we can further flesh out the view. The details of his account are unnecessary for the purposes of this paper, but it's helpful to understand that Nichols's account is very similar to Sripada's IBM, except that instead of innate biases, Nichols advocates for a collection of innate *emotions* that, in turn, lead us to adopting certain moral beliefs. So, in the case of incest, Nichols would likely argue that the innate bias can be understood specifically as an innate disgust reaction against the thought of having sexual intercourse with a genetic relation. While I am going to resist reducing all biases to emotions in the way that Nichols advocates because I want the concept of a bias to allow room for preferences, aversions, likes, and dislikes (which may or may not be understood as emotional), I do agree that, at the end of the day, most biases will have at the very least a significant emotional component.

There is one final step before the initial explication of the IBM is complete. In order for natural selection to influence our moral beliefs in the way the IBM suggests, it isn't enough for us to only have biases. We also need some sort of a capacity that enables us to acquire norms. Here is Machery and Mallon's (2010) understanding of the concept of a norm[17]:

> [N]orms are attitudes toward types of actions, emotions, thoughts, or other traits. These norms are typically shared by many members of a given group and regulate people's behaviors, thoughts, emotions, characters, and so on. Their content essentially involves deontic concepts, such as SHOULD or OUGHT. Such norms can prescribe or forbid a thought, behavior, or any other characteristic, and may be associated with a disposition to punish those individuals that do not comply with the norms (p. 12).

Thus, what the IBM claims is that we not only have a *normative capacity* which enables us to acquire norms, but we also have a set of cognitive biases that influence our normative capacity in such a way that they significantly determine the types of normative attitudes that we assimilate. These attitudes then shape our moral beliefs. (Importantly, the capacity I am referring to should be understood as a *general* capacity for acquiring *all* types of norms—including anything from "one ought not drink alcohol until one is twenty-one years old" to "one ought not skate counter-clockwise on the ice-skating rink" to "one ought not engage in incestuous behavior"—and thus it should not be understood as a capacity that specializes in any type of *moral* cognition.)

I now want to present some evidence[18] in favor of the idea that a capacity like this was instilled in us by evolution. All known societies have norms and ways of

[17] See also Sripada and Stich (2006).

[18] A more complete case can be found in Sripada and Stich (2006) and Machery and Mallon (2010, pp. 13–16).

policing and enforcing such norms (Brown, 1991). Of course the particular strategies of enforcement vary across cultures, but the existence of enforced norms is an uncontroversially universal phenomenon. Even in historical records, there is no trace of a culture without norms. Moreover, humans are especially talented in their ability to reason about norms and detect norm violations,[19] even though they have fairly poor general logical reasoning abilities.[20][21] Importantly, children show similar differences in their reasoning abilities (Cummins, 1996; Harris & Núñez, 1996). This universality of norms and our early development of the ability to reason about norms and detect who is violating them is good evidence that we have a genetically instilled capacity for acquiring norms. As I now want to move on to defending my more controversial thesis about innate biases, I am not going to extensively explain why such a genetic endowment is likely evolutionary;[22] suffice it to say that it seems fairly straightforward that individuals and groups who were able to acquire norms and punish norm violators (e.g. free loaders and other threats to social cohesion and cooperation) were at an evolutionary advantage over those who did not have such an ability.

2.1 The IBM and the causal premise

At this point, the explication of the IBM is complete. However, now that we see the specifics of the model, there is a question about its relation to the causal premise. In particular, the causal premise asserts that our moral beliefs are explained by the process of natural selection. But the IBM does not go quite that far. It claims, first, that we have an innate capacity that enables us to acquire norms; and, second, that we have an innate collection of biases that make it either more or less likely that we adopt certain moral beliefs. Thus, it doesn't go as far as Street (when she claims that the specific content of our moral beliefs was shaped by natural selection) or Joyce (when he claims our moral beliefs stem from an evolved moral capacity). Specifically, nowhere in the IBM is it stated that *all* of our moral beliefs are shaped by natural selection. As the IBM claims that many of our moral beliefs are shaped by our normative attitudes, and such attitudes are influenced by biases instilled by natural selection, it maintains that *many* or *most* of our moral beliefs are shaped by natural selection,[23] but it can't be said that *all* of them are, and it can't be said of the ones that are shaped by natural selection that they are *completely* shaped by it.[24] Indeed,

[19] For example, people have a fairly easy time figuring out who the norm violators are when they are told to enforce such norms as (the more familiar) "If you are under twenty-one years of age, you cannot drink beer in the bar" or (the less familiar) "If you eat mongogo nut, then you must have a tattoo on your chest" (Cosmides & Tooby, 2005).

[20] Indeed, people have a much more difficult time with conditionals not related to norms, such as determining when an indicative conditional is false under the following rule: "If there is a red bird in the drawing on top, then there is an orange on the drawing below" (Cosmides, 1989; Sugiyama et al., 2002).

[21] Of course, it should be acknowledged that this evidence isn't uncontroversial. See, for example, Sperber and Girotto (2002) for an alternative explanation.

[22] See Machery and Mallon (2010, pp. 16–19) for a such a discussion.

[23] Indeed, we will see evidence in support of this in the next section.

[24] Of course, Joyce and Street would not claim that our moral beliefs are completely shaped by natural selection either, but they would claim that they are significantly shaped by natural selection at their most

there could be other emotions, biases, and cognitive attitudes that develop—in response to, say, cultural practices or early childhood experiences—and such emotions, biases, and cognitive attitudes could also influence our normative attitudes.

Thus, if the IBM is going to support the causal premise, the premise will need to modified slightly to:

*Causal Premise**: Most of our moral beliefs are to a significant extent explained by the process of natural selection.

Of course, the causal premise* is still supportive of the debunker's case, but it may not be *as* supportive as the original causal premise, and this might have interesting implications for the debunker's skeptical conclusions. I briefly discuss these implications in the final section. For now, though, I want to turn to the primary objective of this paper: showing why the IBM is the most empirically supported account of the causal premise*.

3 The anti-debunker's empirical case

As mentioned in the first section, a few recent philosophical papers have addressed the causal premise. While all of them contribute to the overall discussion, the most developed and compelling case put forward so far comes from Levy and Levy (2020). In their paper, the Levys argue for why the empirical evidence is not supportive of either Joyce's or Street's hypothesis about the evolution of moral belief. In this section, I want to look closely at the Levys' arguments against Joyce and Street with the following goal in mind. As will be shown, the IBM is distinctly different from either Joyce's or Street's evolutionary hypothesis, and so the Levys never address it specifically in their paper (though of course there is a considerable amount of overlap between all hypotheses advocating the evolution of moral belief). Thus, as I move through the Levys arguments and explain why some are compelling and others are not, I will show why the IBM is in fact compatible with their more convincing points, even when such points do a significant amount of damage to Joyce's and Street's hypotheses. This will lead to an understanding that the IBM—a hypothesis that, again, neither Joyce nor Street nor the Levys consider—is the most compelling, well-supported empirical hypothesis regarding the evolution of moral belief.

3.1 Joyce's moral sense theory

The Levys first address Joyce's *moral sense* theory. Joyce (2006, pp. 108–142) argues that humans evolved to have an innate *moral capacity* or moral sense, which he understands as a specific, universal, and characteristic tendency to make moral judgments. He believes we evolved such a capacity because of the importance of long term cooperation. Indeed, we needed some type of 'motivational bulwark' that

Footnote 24 (continued)
fundamental level. The IBM can't go quite that far, as other biases might be shaping our beliefs even at the most fundamental level.

would encourage us to look beyond our short term, selfish interests (Joyce, 2006, p. 121). One body of evidence that he uses to support this view comes from developmental moral psychology. In particular, he points to the (near) universality of the moral/conventional distinction, and argues that this is indicative of an evolved moral capacity. According to these studies (Nucci & Turiel, 1978; Smetana, 1981; Tisak & Turiel, 1984), children from a very young age are able to differentiate between a moral norm—which is defined as a norm that prohibits a fairly serious harm to a person or a person's rights; is authority-independent; and can be generalized to other cultures ("don't punch your friend for no reason")—and a conventional norm—which is defined as a less serious transgression that is frequently justified by a specific authority figure or institution and is usually not generalizable to other cultures ("don't chew gum in class"). Joyce argues that this early tendency to think in moral terms and to view the difference between moral and conventional norms as a difference in kind is indicative of a particular moral capacity that characteristically emerges at a young age.

In response, the Levys point out that quite a few follow up studies have cast doubt on the moral/conventional distinction (2020, pp. 6–7). In particular, cross-cultural studies have found that what is seen as moral and what is seen as conventional often depends on the specifics of a culture (Haidt et al., 1993; Shweder & Miller, 1985). For example, in India, food, sex, and clothing are moralized; in Brazil, strange sexual conduct (e.g. between a human and a chicken) is moralized; but in the USA, most of these acts are seen as merely conventional. Moreover, as Gabennesch (1990) points out in his review, some studies have found that the moral/conventional distinction often does not emerge until much later, sometimes not even until late-teens, which suggests that the environment is having a more important role than Turiel and his colleagues originally suggested.

While I don't think the evidence that the Levys present is conclusive[25] (they don't think it is either), it is enough to muddy the waters, and when the evidence for a certain innate capacity is unclear, it becomes difficult to maintain its existence. Indeed, the very bread and butter of innate capacity research is to show universality, and if that isn't possible, this is good reason to start considering other options, including environmental influences. Importantly, though, this research fits well with the IBM. According to the IBM, there is no specific moral capacity; there is only a normative capacity that acquires certain norms, and these norms differ only in degree of seriousness. This aspect of the hypothesis is significant because if it is correct, then there is no innately instilled psychological difference in kind between a moral norm and conventional norm. True, different cultures understand some norms as moral and others as conventional, but these labels depend on the specifics of the culture, and probably only boil down to the strength of the emotional reaction a person feels when considering a specific norm violation. And indeed, as the evidence that the Levys point to suggests, the moral/conventional distinction seems very much to be a culturally influenced phenomenon.

[25] For example, there still could be an innate moral capacity, but perhaps it is more general and open to environmental influence.

But this isn't the only aspect of this evidence to consider. If it were the case that norms and taboos were *very* different from culture to culture, then that would cast doubt on the idea that there is a universal set of innate biases. Indeed, the IBM asserts that there are specific, evolutionarily instilled biases for and against certain behaviors, and these biases almost always manifest across cultures, even if their particular manifestations depend on cultural context. Fortunately for the IBM, this is exactly what the evidence just discussed shows. All cultures have taboos for actions related to, for example, sex, food, and attire (thus indicating that there are probably innate biases regarding such topics), but the type and seriousness of each taboo depends on the history and etiology of the culture that it is part of.

At this point, some might be wondering just how much cultural variation an innate bias allows for. Indeed, if it allows for too much variation, the IBM starts to look less plausible, as it might come off as insulated from empirical data (see footnote 16). Fortunately, though, the IBM can be fairly explicit about how much variation there can be before we can start to doubt the existence of an innate bias. Of course, as mentioned, if there are completely disparate norms and taboos between cultures concerning a certain issue, then that is clear evidence that there is no innate bias at work. For example, if it were the case that in one culture there were taboos for food, sex, and clothing, while in another culture the taboos centered on the color teal, the amount of hair in one's left eyebrow, and the angle of one's feet, and such variation continued across most cultures, then that would be indicative that there are *not* innate biases related to food, sex, or clothing.

But what if there is a norm that is common to a collection of similar cultures? For example, in many Australian Aboriginal cultures, it is taboo to be alone with one's mother-in-law (Hiatt, 1984). While this can be considered *more* universal than the taboos and norms considered in the previous paragraphs, it is still only local to a certain collection of (related) cultures. Thus, there is still no reason to believe that it stems from an evolutionarily instilled bias. For such a bias to exist, there needs to a certain amount of universality across *most* cultures, including cultures that are not recently related to each other.[26]

The Levys also point out that if Joyce's moral capacity really did evolve, it probably did so in the form a specific neural mechanism (2020, pp. 7–8). Unfortunately for Joyce's hypothesis, however, there is little neuroscientific evidence for such a mechanism. Indeed, there have been numerous fMRI studies carried out on subjects while they made moral judgements, and such brain scans have revealed not only that

[26] With that said, there may in fact be evidence for a more general innate bias concerning deference and respect for elders and/or authority figures. To explore such evidence (with the IBM in mind) would be to look for near universality in norms that advocate respecting elders and/or authority figures, but a diversity in the particular way these norms are manifested (e.g. in one culture, there might be a strong norm about respecting grandparents, while in another culture, there might be a more general respect norm, along the lines of, "if the person is older than you, you must respect them"). If it turned out either (a) that there wasn't a near universality of such respect norms across cultures, or (b) that there was a near universality in both the existence of such norms and their particular contents, then we would have reason to doubt that an innate bias is at work.

multiple neural areas are involved (Sinnott-Armstrong & Wheatley, 2012), but also that different areas are associated with different moral judgments (Dale, 2020; Dale & Gawronski, 2022; Greene et al., 2001; Moll et al., 2005; Parkinson et al., 2011). If this is the case, then it is very unlikely that there is some specific cognitive (sub) system that is dedicated to moral cognition.

While we should be careful about drawing conclusions about the functional specialization of certain brain regions based on fairly coarse grained mapping techniques that pick up only on slightly more blood flow[27] to those regions (Jasanoff, 2018, pp. 71–90), I will grant the Levys this point. Indeed, as far as the neuroscience suggests, there is little evidence for the moral mechanism that Joyce hypothesizes. But notice how this evidence is actually compatible with the IBM. As the IBM posits that we have a collection of innate biases, and these biases have evolved for potentially disparate reasons, it makes sense that they would be generated from different systems in the brain (although it is more likely that they are generated from the more emotion based regions). True, the IBM does require a general normative capacity, as discussed in the previous section, but even if one wanted to understand this capacity as a type of specialized mechanism, such a system would be fairly broad and require many different neural subsystems (which would activate in particular ways when certain types of normative beliefs—e.g. moral beliefs—were generated). Thus, the neuroscientific evidence is very much in line with the IBM.

3.2 Street's basic evaluative tendencies

The Levys also address Street's evolutionary account, which is significantly different from Joyce's and, importantly, a bit more in line with the IBM. Street (2006, § 1–4) argues that the specific *content* of our moral beliefs was to a large extent shaped by natural selection. Thus, it isn't that a capacity or tendency to form moral beliefs evolved (à la Joyce), but instead that the moral beliefs themselves (e.g. "one ought to help one's kin before helping a stranger") were shaped by evolutionary forces. Of course Street does not believe that every specific moral belief is the direct result of natural selection; it's more that "natural selection has had a tremendous *direct* influence on…our basic evaluative tendencies, and…these basic evaluative tendencies, in their turn, have had a major influence on the evaluative judgments we affirm" (Sect. 4). So, there is a basic "calling out for" or "counting in favor" of something that is written into our genes, and this then shapes our specific moral beliefs.

Before moving on to the Levys' critique of Street, it's important to understand how the IBM differs from Street's model. Indeed, in some ways they are similar because they both maintain that there are basic, underlying, evolutionarily instilled mechanisms that shape our particular moral beliefs. However, there is a significant difference between a Streetian *basic evaluative tendency* and an *innate bias*. The

[27] And just because there is slightly more blood flow to a certain region doesn't mean that many other regions aren't also experiencing activity, which may or may not be fundamental to the particular function in question (Jasanoff, 2018, pp. 71–90).

former is a "proto" evaluative belief that comes loaded with a specific evaluative pull towards understanding a situation in a particular way (Sect. 4). Street uses the following example: when someone treats me well, that calls out for a specific action; that is, treating them well in return (Sect. 4). The particular way in which I do that (or believe that I ought to do that) depends on the situation and my culture, but the tendency to experience the situation as demanding a particular kind of response is evolutionarily endowed.

The IBM, on the other hand, is much more basic and straightforward. There is no evaluative appraisal or "calling out for" or "proto" belief or even being pulled to respond in a certain way. There is only a basic (primarily emotional) bias concerning certain situations and states-of-affairs. So, in the example of reciprocal altruism mentioned above, if it were the case that there is an innate bias with regard to the situation, it would be more along the lines of a positive feeling towards a person in response to her treating me well.[28] Then, from there, this bias manifests into a moral belief, the particulars of which depend on cultural and environmental factors.

To take another example, consider nepotism. It's no secret that we often help family members over strangers or even friends. According to Street, such nepotistic beliefs are formed because we have an evolutionarily instilled "proto" belief that we ought to do actions that help our kin. According to the IBM, however, there is nothing nearly as sophisticated as a "proto" belief or any type of "calling out" to do certain actions. There is only a basic positive bias towards our family members such that we view them in a more positive light. This, in turn, leads us to form nepotistic beliefs.

All this is to say, there is more specific moral content (in the manner of action or reason guidance) built into Street's basic evaluative tendencies than in the biases of the IBM. And that is why she needs to find fairly compelling evidence of the universality of moral *belief* if she wants her causal premise to be empirically supported. One obvious contender is of course incest.[29] For moral beliefs about incest to be the product of Streetian basic evaluative tendencies, incestuous behavior must call out for a particular type of response, and this would mean that—if Street is right—there would be a fairly universal response to incest. However, as the Levys (2020, p. 13) correctly point out (and as I already discussed in Sect. 2), taboos, laws, and beliefs about incest vary quite widely from culture to culture (Prinz, 2007). Thus, there doesn't seem to be any compelling reason to accept that we evolved to see incestuous behavior as calling out for a particular type of evaluative response.

However, as discussed, there is without question a significant amount of universality in incest response, as almost all recorded human societies have some type of moral norm prohibiting incest (Murdock, 1949). Thus, what seems to fit the bill is not a response loaded with a particular evaluative motivation but instead a basic, fairly straightforward but nonetheless powerful emotional response bias to

[28] Indeed, there is evidence of reciprocal altruism among apes (De Waal, 1996), and because it is less likely that apes understand a situation as "calling out for" a certain response, the more basic positive-feelings-towards-someone-who-helps-me hypothesis is more parsimonious and promising.

[29] Street herself does not discuss incest.

incestuous thoughts and behavior. Such an innate bias can not only allow for the variation that we see between cultures with regard to incest taboos, but it can also account for the universality of the aversion to incestuous behavior. Which is to say, the IBM occupies a nice middle ground between Street's more content loaded account and an (even less plausible) account that incest taboos primarily stem from cultural influence.

Next, the Levys discuss harm. As the Levys themselves admit, "*some* kind of prohibition against harm is present in very many societies, perhaps every society" (2020, p. 13). However, many cultures differ in their specific harm norms. Some prohibit almost all types of harm; some allow harm only to out-group members; some allow harm to certain in-group members, such as women, the disabled, and other marginalized groups; and some allow for ceremonial harms such as cannibalism, scarification, piercing, and circumcision (Chagnon, 1992; Prinz, 2007; Robarchek & Robarchek, 1992; Silverberg & Gray, 1992; Sripada, 2008). Due to this cultural diversity, it is difficult to claim—as Street's version of the causal premise does—that harm situations "call out" for a particular type of normative response. Indeed, if they did, we would see much more uniformity in particular harm norms. For example, we might see most (if not all) cultures with a norm prohibiting harm to disabled clan members, as such a norm would be underwritten by a "proto" belief about how harming such clan members ought to be avoided. Or, there might be a fairly universal norm against cannibalism. However, we just don't see this type of universality when it comes to harm norms.[30] Thus, there is little reason to believe that Street's basic evaluative tendencies are the most compelling explanation.

Yet, again, the data fit nicely with the IBM. With regard to harm, the IBM posits that there is some sort of evolutionarily instilled bias against harming conspecifics, and this bias is then manifested in different ways in particular cultures. I'll admit that harm is a bit trickier than the previously discussed moral norms, as it is likely that there are multiple factors (and perhaps multiple biases) at work, but that doesn't mean that the IBM isn't the best explanation of the data. Indeed, it does seem to be, and there are plenty of specific hypotheses that could cash it out. For example, humans could have an innate bias against harming in-group members (which would be advantageous to fitness as in-group members are and always have been an important asset to an individual's survival and reproductive success), and it could then be up to the culture to decide who is part of the "in-group" and who is part of the "out-group." For some, only other clans are out-groups; for others, women, the disabled, and other marginalized groups within the community are viewed as out-group members; and for still others, there is an attempt to classify every person on earth (as well as some nonhuman animals) as in-group members (e.g. modern, liberal, western morality). Another innate bias that probably affects harm norms is empathy.[31] Perhaps as a result of our evolutionarily instilled empathic abilities, we feel averse

[30] Of course, what counts as a harm probably varies between cultures. For example, cultures with ceremonial cannibalism and scarification probably don't view such actions as harms. But, again, this is supportive of the IBM. If harm situations "called out" for a particular normative response, we probably wouldn't see so much variation between cultures in what counts as a harm. Indeed, such variation implies the existence of a less content-laden mechanism.

[31] For a discussion of this possibility, see Dwyer (2006).

to harming those we identify with, and it is up to the specifics of culture to decide who it is we identify with. For the purposes of this paper, though, the particulars are not important. What's important is to understand that the existence of (some type of) innate biases does the best job at explaining both the across-culture universality and the between-culture diversity of harm norms.[32]

Furthermore, consider again food taboos. Fessler and Navarette (2003) looked at food taboos in 78 different cultures and found that meat is the type of food most likely (by a significant margin) to have a taboo attached to it. Of course, the specifics of the taboo (and the type of meat that it targets) varies by culture, but the universality of this aversion hints at some type of innate disposition. And indeed, the existence of an evolutionarily instilled bias (probably disgust) makes sense in the context of evolution, as meat has "an extremely high potential for food-borne infection and other pathogenic consequences" (Sripada, 2008).[33] This is an especially interesting example to consider because while almost all cultures and religions seem to have some type of taboo against meat, for some societies it has manifested into a strong moral norm but for others it is merely a convention or even a personal choice (e.g. vegetarianism and veganism). This is further evidence of a lack of any (psychologically instilled) difference in kind between moral and conventional norms and instead a continuum of moral norms differing only in the perceived grievousness of the actions they concern (which is most likely dictated by emotional response), with some cultures designating the more grievous actions as "moral" transgressions and the less grievous actions as "conventional" transgressions. And as such a psychological continuum is a necessary aspect of the IBM, the food taboo literature provides further support for the hypothesis.

Some might point out here that other interpretations are possible. For example, with or without the existence of a norm, everyone would probably agree that rotting meat is disgusting—and the thought of *eating* rotting meat is even more disgusting! So why think an innate bias is behind these norms and taboos? Why can't such norms be due simply to the fact that we all find the idea of rotten meat disgusting? This is an interesting possibility, but a bias still remains the most plausible explanation. Indeed, all spoiled and rotting foods are disgusting! So why would it be significantly more common for cultures to have taboos about *meat* in particular? The best explanation of this is that there is some kind of mechanism (i.e. a bias) underlying our aversion towards meat. Of course, if we only saw these norms and taboos in certain collection of cultures, or if we saw a significant amount of uniformity in the particular kind of meat taboos, we would have reason to doubt the existence of a bias. But as the data reveal *universality with variation* with regard to meat taboos, an innate bias remains the best explanation.

[32] The Levys (p. 13) mention that harm norms might best be understood as a culturally explained "good trick" (Dennett, 1995) in response to a common human problem, much like covering one's head in the sun. However, these two examples are not analogous, as harming others can often be in a person's short-term interest. Thus, letting people figure out for themselves that (at least some types of) harm is bad in the long run would likely not be an evolutionarily stable strategy.

[33] Interestingly, many nonhuman animals also develop aversions to meat (Fessler & Navarette, 2003).

3.3 The frugality of natural selection

Finally, it's important to take our theoretical understanding of evolution into consideration. Right now, we have three options on the table: the IBM, Joyce's moral capacity, and Street's basic evaluative tendencies. Which makes the most sense with our current understanding of the process of natural selection? One thing we know for sure about natural selection is that it is frugal, and if there is an efficient and easy way to select for a trait that enhances fitness, then it will often select for that trait even if it is not the overall best design and it leads to other problems for the organism. The poor design of the vertebrate eye;[34] the use of the same pipe for both breathing and eating; the fact that the human birth canal passes through the pelvis;[35] all of these are the result of natural selection opting for efficiency at the expense of the overall *bauplan*[36] of the organism. Now, if natural selection has this tendency to select for *just enough* to get the job done, then it is difficult to see why it would opt for an entire moral capacity or even basic evaluative tendencies (which, again, require a significant amount of specific evaluative content) when more basic innate biases would be sufficient.[37] Of course, for reasons already discussed, there was good reason to select for *some* type of mechanism that influenced our moral beliefs, but it seems overall unlikely for natural selection to have opted for a costly, more restrictive design that specializes in moral content or cognition. Indeed, the more efficient and effective[38] design would be to simply instill some sort of bias when an adaptive problem becomes significant enough to call for such a bias.

4 Conclusion

In this paper, I argued that the IBM is the most compelling account of the causal premise*. However, as discussed in Sect. 2.1, the causal premise* is weaker than the original causal premise that most evolutionary debunkers put forward. Indeed, any debunker using the IBM in support of their case will not be able to claim that all moral beliefs are explained by the processes of natural selection, and this might have interesting implications for discussions surrounding the EDA. For example, it may give the anti-debunker a bit more room to maneuver in response to Street's version of EDA. Specifically, I am thinking of David Copp's *quasi-tracking thesis* (2008, p. 194) and William FitzPatrick's discussion of critical reflection (2015, p. 887),

[34] The nerves and blood vessels of the primate eye are *in front* of the retina, which makes little sense from a design standpoint. See the cephalopod eye for a superior design.

[35] This of course results in a significant portion of women dying during childbirth, a risk which most other female animals do not suffer from.

[36] A *bauplan* is commonly understood as the overall body plan or blue print of an organism. See Gould and Lewontin (1979) for a complete explanation.

[37] True, for the IBM to be correct, a normative capacity also had to evolve, but such a capacity is probably necessary for Joyce's and Street's accounts, as well. Indeed, both accounts would need to explain why we have non-moral, normative beliefs, and the kind of normative capacity posited by the IBM would do that well.

[38] Effective because having a broader, less restrictive mechanism would allow for more variation, which has likely always been an important aspect of human and hominin life.

according to both of which even if natural selection has tainted our moral beliefs to a certain extent, there is still the possibility that we can use our ability to rationally reflect to come to at least some sort of understanding of the moral truths—perhaps through the process of *reflective equilibrium*.

Needless to say, Street (2006, § 5) rejects this possibility because—as she argues—our moral beliefs are likely tainted all the way down. That is, rational reflection on our moral beliefs must always stem from some moral starting point, and because all of our moral attitudes and beliefs are shaped by a non-truth-tracking process, there is simply no finger hold for our moral beliefs to grab onto. However, if it is the case that there are only innately instilled moral *biases*—as opposed to more content specific evaluative tendencies—then perhaps there is more room for rational reflection. Indeed, as explained in the second section, the innate biases account allows for other influences at the most fundamental level, and this might mean that our moral beliefs aren't tainted all the way down in the way Street claims that they are.

Of course, I am only briefly gesturing at this possibility, and a lot more work needs to be done before such a case can be said to be convincingly made. However, the possibility is there; and with it comes further recognition of the importance of exploring the empirical details of the evolutionary debunking argument.

References

Brown, D. (1991). *Human universals*. McGraw-Hill.

Chagnon, N. (1992). *Yanomamö: The last days of Eden*. Harcourt Brace Javanovich.

Cofnas, N. (2020). Are moral norms rooted in instincts? The sibling incest taboo as a case study. *Biology and Philosophy, 35*, 47.

Copp, D. (2008). Darwinian skepticism about moral realism. *Philosophical Issues, 18*, 186–206.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason Studies with the Wason selection task. *Cognition, 31*, 187–276.

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 584–627). Wiley.

Cummins, D. D. (1996). Evidence of deontic reasoning in 3- and 4-year-olds. *Memory and Cognition, 24*, 823–829.

Dale, M. T. (2020). Neurons and normativity: A critique of Greene's notion of unfamiliarity. *Philosophical Psychology, 33*, 1072–1095.

Dale, M. T., & Gawronski, B. (2022). Brains, trains, and ethical claims: Reassessing the normative implications of moral dilemma research. *Philosophical Psychology.* Advance online publication. https://doi.org/10.1080/09515089.2022.2038783.

De Waal, F. (1996). *Good natured*. Harvard University Press.

Deem, M. (2016). Dehorning the Darwinian dilemma for normative realism. *Biology and Philosophy, 31*, 727–746.

Dennett, D. (1995). *Darwin's dangerous idea*. Simon & Schuster.

Durham, W. (1991). *Coevolution*. Stanford University Press.

Dwyer, S. (2006). How good is the linguistic analogy? In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Culture and cognition* (pp. 237–256). Oxford University Press.

Fessler, D., & Navarette, C. (2003). Meat is good to taboo: Dietary prescriptions as a product of the interactions of psychological mechanisms and social processes. *Journal of Cognition and Culture, 3*(1), 1–40.

Fitzpatrick, W. J. (2015). Debunking evolutionary debunking of ethical realism. *Philosophical Studies, 172*, 883–904.
Fraser, B. (2014). Evolutionary debunking arguments and the reliability of moral cognition. *Philosophical Studies, 168*, 457–473.
Gabennesch, H. (1990). The perception of social conventionality by children and adults. *Child Development, 61*, 2047–2059.
Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, B, Biological Sciences, 205*(1161), 581–598.
Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.
Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*, 613–628.
Harris, P., & Núñez, M. (1996). Understanding of permission rules by preschool children. *Child Development, 67*, 1572–1591.
Henrich, J. (2015). *The secret of our success*. Princeton University Press.
Hiatt, L. (1984). You mother-in-law is poison. *Man, 19*(2), 183–198.
Isserow, J. (2019). Evolutionary hypotheses and moral skepticism. *Erkenntnis, 84*, 1025–1045.
Jasanoff, A. (2018). *The biological mind*. Basic Books.
Joyce, R. (2006). *The evolution of morality*. MIT Press.
Kahane, G. (2011). Evolutionary debunking arguments. *Nous, 45*, 103–125.
Keller, L. F., & Waller, D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology and Evolution, 17*, 230–241.
Kitcher, P. (2011). *The ethical project*. Harvard University Press.
Levy, A., & Levy, Y. (2020). Evolutionary debunking arguments meet evolutionary science. *Philosophy and Phenomenological Research, 100*, 491–509.
Machery, E., & Mallon, R. (2010). Evolution of morality. In J. M. Doris and the Moral Psychology Research Group (Eds.), *The moral psychology handbook*. Oxford University Press.
Moll, J., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafhian, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*, 799–809.
Murdock, G. P. (1949). *Social structure*. Free Press.
Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press.
Nucci, L., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development, 49*, 400–407.
Parkinson, C., Sinnott-Armstrong, W., Koralus, P., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience, 23*(10), 3162–3180.
Pölzler, T. (2018). *Moral reality and the empirical sciences*. Routledge.
Prinz, J. J. (2007). Is morality innate? In W. Sinnott-Armstrong (Ed.), *Moral psychology*, Vol. 1 (pp. 367–407). MIT Press.
Pusey, A. E., & Wolf, M. (1996). Inbreeding avoidance in animals. *Trends in Ecology and Evolution, 11*, 201–206.
Robarchek, C. A., & Robarchek, C. J. (1992). Cultures of war and peace: A comparative study of Waorani and Semai. In J. Silverberg & P. Gray (Eds.), *Aggression and peacefulness in humans and other primates* (pp. 189–213). Oxford University Press.
Ruse, M. (1998). *Taking Darwin seriously: A naturalistic approach to philosophy*. Prometheus Books.
Shepher, J. (1983). *Incest: A biosocial view*. Academic Press.
Shweder, R., & Miller, J. (1985). The social construction of a person: How is it possible? In K. Gergen & K. Davis (Eds.), *The social construction of the person.* Springer.
Silverberg, J., & Gray, P. (1992). *Aggression and peacefulness in humans and other primates*. Oxford University Press.
Sinnott-Armstrong, W., & Wheatley, T. (2012). The disunity of morality and why it matters to philosophy. *The Monist, 95*(3), 355–377.
Smetana, J. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52*, 1333–1336.
Sperber, D., & Girotto, V. (2002). Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides and Tooby. *Cognition, 85*(3), 277–290.

Spiro, M. (1958). *Children of the kibbutz*. Harvard University Press.
Sripada, C. (2008). Nativism and moral psychology: Three models of the innate structure that shapes the content of moral norms. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, Vol. 1 (pp. 319–344). MIT Press.
Sripada, C., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Culture and cognition* (pp. 280–301). Oxford University Press.
Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies, 127*, 109–166.
Sugiyama, L., Tooby, J., & Cosmides, L. (2002). Cross-cultural evidence of cognitive adaptations for social exchange among the Shiwiar of Ecuadorian Amazonia. *Proceedings of the National Academy of Sciences, 99*, 11537–11542.
Tisak, M. S., & Turiel, E. (1984). Variation in seriousness of transgressions and children's moral and conventional concepts. *Developmental Psychology, 24*(3), 352–357.
Vavova, K. (2014). Debunking evolutionary debunking. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics*, Vol. 9 (pp. 76–101). Oxford University Press.
Westermarck, E. (1922). *History of human marriage* (Vol. 2). Allerton.
Wilson, R. (2019). Incest, incest avoidance, and attachment: Revisiting the westermarck effect. *Philosophy of Science, 86*(3), 391–411.