Evolutionary Explanations and the Debunking of Moral Beliefs

Nathan Cofnas

Balliol College

Thesis submitted for the degree of DPhil in Philosophy

Michaelmas 2020

University of Oxford

Contents

Acknowledgments

The luckiest thing to ever happen in my professional life was being assigned Guy Kahane and Andreas Mogensen as advisors. I would be happy to remain a DPhil student forever, but am told this is not an option.

In the course of writing this thesis, some of my extracurricular work generated intense political controversy. I am grateful to the many people at Oxford and elsewhere who supported me during this time, particularly Jeff McMahan.

Thanks to Jonathan Anomaly and Neven Sesardić for their support in both personal and philosophical matters over the past several years.

Viewed from the outside, the life of a graduate student might seem rather monotonous: reading, typing, reading, typing, and so on for hours a day every day for years. But as those who have experienced it know, there is plenty of drama—and that was especially true for me. I am grateful to my wife, Youngsan Goo, who was there for everything.

Abstract

1. Are Moral Norms Rooted in Instincts? The Sibling Incest Taboo as a Case Study

According to Westermarck's widely accepted explanation of the incest taboo, cultural prohibitions on sibling sex are rooted in an evolved biological disposition to feel sexual aversion toward our childhood coresidents. Bernard Williams posed the "representation problem" for Westermarck's theory: the content of the hypothesized instinct (avoid sex with *childhood coresidents*) is different from the content of the incest taboo (avoid sex with *siblings*)—thus the former cannot be causally responsible for the latter. Arthur Wolf posed the related "moralization problem": the instinct concerns personal behavior whereas the prohibition concerns everyone. This paper reviews possible ways of defending Westermarck's theory from the representation and moralization problems, and concludes that the theory is untenable. A recent study purports to support Westermarck's account by showing that unrelated children raised in the same peer groups on kibbutzim feel sexual aversion toward each other and morally oppose third-party intra-peer-group sex, but this study has been misinterpreted. I argue that the representation and moralization problems are general problems that could potentially undermine many popular evolutionary explanations of social/moral norms. The cultural evolution of morality is not tightly constrained by our biological endowment in the way some philosophers and evolutionary psychologists believe.

2. Power in Cultural Evolution and the Spread of Prosocial Norms

According to cultural evolutionary theory in the tradition of Boyd and Richerson, cultural evolution is driven by individuals' learning biases, natural selection, and random forces. Learning biases lead people to preferentially acquire cultural variants with certain contents or in certain contexts. Natural selection favors individuals or groups with fitness-promoting variants. Durham (1991) argued that Boyd and Richerson's approach is based on a "radical individualism" that fails to recognize that cultural variants are often "imposed" on people regardless of their individual decisions. Fracchia and Lewontin (2005) raised a similar challenge, suggesting that the success of a variant is often determined by the degree of *power* backing it. With power, a ruler can impose beliefs or practices on a whole population by diktat, rendering all of the forces represented in cultural evolutionary models irrelevant. It is argued here, based on work by Boehm (1999, 2012), that, from at least the time of the early Middle Paleolithic, human bands were controlled by powerful *coalitions of the majority* that deliberately guided the development of moral norms to promote the common good. Cultural evolutionary models of the evolution of morality have been based on false premises. However, Durham (1991) and Fracchia and Lewontin's (2005) challenge does not undermine cultural evolutionary modeling in *nonmoral* domains.

3. A Debunking Explanation for Moral Progress

According to "debunking arguments," our moral beliefs are explained by evolutionary and cultural processes that do not track objective, mind-independent moral truth. Therefore (the debunkers say) we ought to be skeptics about moral realism. Huemer

counters that "moral progress"—the cross-cultural convergence on liberalism—cannot be explained by debunking arguments. According to him, the best explanation for this phenomenon is that people have come to recognize the objective correctness of liberalism. Although Huemer may be the first philosopher to make this explicit empirical argument for moral realism, the idea that societies will eventually converge on the same moral beliefs is a notable theme in realist thinking. Antirealists, on the other hand, often point to seemingly intractable cross-cultural moral disagreement as evidence against realism (the "argument from disagreement"). This paper argues that the trend toward liberalism is susceptible to a debunking explanation, being driven by two related non-truth-tracking processes. First, large numbers of people gravitate to liberal values for reasons of self-interest. Second, as societies become more prosperous and advanced, they become more effective at suppressing violence, and they create conditions where people are more likely to empathize with others, which encourages liberalism. The latter process is not truth tracking (or so this paper argues) because empathy-based moral beliefs are themselves susceptible to an evolutionary debunking argument. Cross-cultural convergence on liberalism per se does not support either realism or antirealism.

4. Realist Social Selection: How Gene–Culture Coevolution Can (but Probably Did Not) Track Mind-Independent Moral Truth

Standard evolutionary debunking arguments (EDAs) in metaethics target moral beliefs by attributing them to natural selection. According to the debunkers, natural selection does not track mind-independent moral truth, so the discovery that our moral beliefs (realistically construed) were caused by natural selection renders them unjustified. I argue that our innate moral faculty is likely *not* the product of natural selection, but rather

social selection. Social selection is a kind of gene–culture coevolution driven by the enforcement of collectively agreed-upon rules. Unlike natural selection, social selection is teleological and could potentially track mind-independent moral truth by a process that I term *realist social selection*: early humans could have acquired moral knowledge via *reason* and enforced rules based on that knowledge, thereby creating selection pressures that drove the evolution of our innate moral faculty. Given anthropological evidence that early humans designed rules with the conscious aim of preserving individual autonomy and advancing their collective interests, realist social selection appears to be an attractive theory for moral realists. However, I propose a new EDA to show that realist social selection is unlikely to have occurred.

5. A Debunking How-Possibly Explanation for the Principle of Universal Benevolence

According to Street's evolutionary debunking argument (EDA), evolutionary biology provides "powerful" explanations of our "basic evaluative judgements." The discovery that our moral beliefs (realistically construed) are "saturated with evolutionary influence" renders them unjustified, since natural selection does not track mind-independent moral truth. De Lazari-Radek and Singer agree that most of our commonsense moral beliefs are debunked in the way Street claims, but they argue that belief in Sidgwick's *principle of universal benevolence* cannot be explained by natural selection and is therefore immune from EDAs. I argue that Street oversold the power of her evolutionary explanations, thus leaving an opening for realists to claim that moral beliefs with less powerful evolutionary explanations can escape debunking. In fact, all naturalistic theories of morality—including those invoked by Street and de Lazari-Radek and Singer—are speculative

"how-possibly" explanations. If how-possibly explanations are *not* debunking, then both Street's (global) and de Lazari-Radek and Singer's (selective) debunking arguments fail. If how-possibly explanations *are* debunking, then selective debunkers must show that there is no plausible way that naturalistic forces could have produced the beliefs they want to defend. I argue that naturalistic how-possibly explanations can debunk moral beliefs by appealing to ontological parsimony, and provide a debunking how-possibly explanation for belief in the principle of universal benevolence.

Introduction

Many evolutionary psychologists and philosophers believe that natural selection shaped our core moral beliefs. In their view, the evaluative tendencies that underlie moral judgment were selected to promote adaptive behavior in the ancestral environment. Thus, for example, we think incest is wrong because we evolved to avoid maladaptive inbreeding. We think people have an obligation to take care of their family members because we evolved to favor those who share our genes.

Moral realists believe that there are "stance-independent" moral truths, which "*are not made true by virtue of their ratification from within any given actual or hypothetical perspective*" (Shafer-Landau 2003: 15). Advocates of evolutionary debunking arguments (EDAs) point to the alleged influence of natural selection on our moral beliefs to undermine moral realism. According to evolutionary debunkers, natural selection aims at fitness, not (stance-independent) moral truth, so it would be a huge coincidence if it endowed us with moral intuitions that align with truth. The discovery of the evolutionary origins of our moral beliefs (realistically construed) renders them unjustified. We should reject (or at least be skeptics about) moral realism.

This dissertation is about the evolutionary origins of our moral beliefs and the implications for EDAs. Each chapter defends its own hypothesis and is self-standing. However, taken together they aim to present a general approach to studying the evolution of morality that incorporates ideas from several different research programs.

In my view, there should be no sharp separation between moral psychology/anthropology and metaethics. Questions about the reality and/or nature of moral facts are seen as the subject matter of metaethics, but the scientific study of morality cannot be metaethically neutral. If real moral facts exist and we have epistemic

access to them, the best explanation for at least some of our moral beliefs—and the resulting behavior—may be that people acquired knowledge of the stance-independent moral truth. And if our moral intuitions are the product of gene–culture coevolution, hundreds of thousands of years of human societies enforcing rules informed by moral knowledge may have led us to evolve innate intuitions corresponding to those truths. My conclusion is that real moral facts are explanatorily superfluous and should be rejected on grounds of ontological parsimony. But I contend that moral realism should be regarded as a theory with material implications for our scientific understanding of morality, even if we ultimately reject it.

Chapter 1 argues that Westermarck's widely accepted explanation of the incest taboo fails on conceptual and empirical grounds. My analysis brings out some conceptual issues that (I argue) undermine many popular evolutionary explanations of moral beliefs, including those invoked by evolutionary debunkers such as Street (2006). For a variety of reasons, we cannot explain the existence of a moral norm that demands or prohibits Xing by showing that there was natural selection for Xing or for not Xing, respectively.

Chapter 2 considers another influential approach to explaining the evolution of moral norms—Boyd-and-Richerson style gene–culture coevolutionary theory (e.g., Richerson and Boyd 2005). According to cultural evolutionary theorists in this tradition, moral norms (like other cultural variants) spread largely as a consequence of individuals with certain learning biases copying each other, as well as natural selection acting on individuals and groups. Building on Boehm's (2012) theory of social selection, I argue that moral norms were originally imposed on people by force. People were not *choosing* which norms to adopt in the way assumed by cultural evolutionary models.

Chapter 3 proposes a naturalistic explanation for cross-cultural convergence on liberalism, which appeals to a combination of (non-moral-truth-tracking) evolutionary and cultural forces.

Chapter 4 proposes the theory of *realist social selection*. According to Boehm (2012), moral evolution was driven by social selection: Coalitions of the majority in early human groups collectively imposed rules on people in order to bring about social outcomes that they saw as desirable, which favored the evolution of a conscience. I argue that social selection is teleological and therefore different from natural selection. If our moral intuitions evolved due to social selection, the empirical premise of standard EDAs—i.e., our moral intuitions evolved due to *natural selection*—is false. It would have been possible for us to have teleologically evolved moral intuitions in line with stance-independent truth if early rule makers were informed by knowledge of moral truth acquired via reason. However, I propose a new debunking argument, which purports to show that realist social selection did not occur.

Chapter 5 considers the use of so-called selective EDAs, focusing on de Lazari-Radek and Singer's (2012) argument for the *principle of universal benevolence*, which says we have a reason to impartially maximize the utility of all sentient beings. De Lazari-Radek and Singer agree with (global) evolutionary debunkers that most of our commonsense moral beliefs are debunked in virtue of being explained by natural selection, but they argue that belief in the principle of universal benevolence has no evolutionary explanation and is therefore immune from EDAs. I argue that all evolutionary (and other naturalistic) explanations of moral beliefs are at best speculative *how-possibly* explanations. If such how-possibly explanations are not debunking, then both global and selective EDAs fail. If they *are* debunking, then one only needs to come up with a how-possibly explanation for belief in the principle of universal benevolence

in order to rebut de Lazari-Radek and Singer. I argue that naturalistic how-possibly explanations are debunking, since they make real moral facts explanatorily superfluous. I propose a naturalistic how-possibly explanation for belief in the principle of universal benevolence.

In broad outline, the theory that emerges from this analysis is as follows. A few general norms (e.g., prohibitions on indiscriminate violence directed at ingroup members) have been enforced consistently in all human societies since the beginning of our species. These norms have to some extent been encoded in the genome via gene–culture coevolution, and these impose very broad constraints on morality. We also evolved a host of amoral dispositions to perform/avoid certain behaviors and favor/disfavor certain states of affairs, which do not directly generate moral norms. But people tend to enforce norms that serve their perceived interests, which in turn are partly determined by their innate preferences. Which norms prevail in a particular society depends largely on the balance of power among competing factions. Theoretically, there are few constraints on what systems of norms can be imposed on people, but those that frustrate many people's evolved impulses are less stable because they are more likely to trigger resistance.

I would consider my view to be anti-nativist, but with some qualifications. I certainly deny that there is anything like a "universal moral grammar" (Mikhail 2007), and would agree with Prinz (2007: 246) that none of our "biological predispositions...qualify as *moral* rules without cultural elaboration." However, I think we evolved (due to social selection) a disposition to acquire some general moral values that have been consistently promoted in human societies for hundreds of thousands of years. In the absence of cultural inputs, these genetic adaptations are not enough to generate full-fledged morality, but some elements of morality may be *partly* encoded in

the genome. Biology influences the specific content of our moral beliefs primarily by indirect means, namely, by shaping the (nonmoral) desires of people who design and enforce moral norms.

References

Boehm, Christopher. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* New York: Basic Books.

de Lazari-Radek, Katarzyna, and Peter Singer. 2012. "The Objectivity of Ethics and the Unity of Practical Reason." *Ethics* 123 (1): 9–31.

Mikhail, John. 2007. "Universal Moral Grammar: Theory, Evidence and the Future." *Trends in Cognitive Sciences* 11 (4): 143–152.

Prinz, Jesse. 2007. *The Emotional Construction of Morals.* Oxford: Oxford University Press.

Richerson, Peter J., and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: University of Chicago Press.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence.* Oxford: Oxford University Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127 (1): 109–166.

Are Moral Norms Rooted in Instincts?

The Sibling Incest Taboo as a Case Study[*]

Abstract: According to Westermarck's widely accepted explanation of the incest taboo, cultural prohibitions on sibling sex are rooted in an evolved biological disposition to feel sexual aversion toward our childhood coresidents. Bernard Williams posed the "representation problem" for Westermarck's theory: the content of the hypothesized instinct (avoid sex with *childhood coresidents*) is different from the content of the incest taboo (avoid sex with *siblings*)—thus, the former cannot be causally responsible for the latter. Arthur Wolf posed the related "moralization problem": the instinct concerns personal behavior whereas the prohibition concerns everyone. This paper reviews possible ways of defending Westermarck's theory from the representation and moralization problems, and concludes that the theory is untenable. A recent study purports to support Westermarck's account by showing that unrelated children raised in the same peer groups on kibbutzim feel sexual aversion toward each other and morally oppose third-party intra-peer-group sex, but this study has been misinterpreted. I argue that the representation and moralization problems are *general* problems that could potentially undermine many popular evolutionary explanations of social/moral norms. The cultural evolution of morality is not tightly constrained by our biological endowment in the way some philosophers and evolutionary psychologists believe.

Keywords: representation problem; moralization problem; incest taboo; Westermarck effect; evolutionary psychology

---

[*] Published in *Biology & Philosophy* 35, no. 5 (2020): 47.

1. Introduction

There is a widespread prohibition on brother–sister marriage across human cultures, reflecting a widespread taboo on brother–sister sex (Wolf and Durham 2004). Some scholars believe the sibling incest taboo has a clear biological basis, and they see it as a paradigm case of biology influencing (or determining) a social norm. Following Westermarck (1891), they argue that, in order to avoid close inbreeding, we evolved to feel sexual aversion toward people with whom we are raised as young children—people who are almost always our biological siblings. This innate impulse, which came to be known as the "Westermarck effect," expresses itself as a norm prohibiting sibling sex. Ruse and Wilson (1986: 184) wrote that the following explanation of the incest taboo is "widely accepted":

> Lowered genetic fitness due to inbreeding led to the evolution of the juvenile sensitive period [to develop sexual aversion] by means of natural selection; the inhibition experienced at sexual maturity led to prohibitions and cautionary myths against incest or (in many societies) merely a shared feeling that the practice is inappropriate. Formal incest taboos are the cultural reinforcement of the automatic inhibition, an example of the way culture is shaped by biology.

Wilson (1998: 196) later acknowledged that some cultures might have recognized the negative consequences of close inbreeding for offspring, and instituted the taboo for this reason. But he reiterated that, in the majority of cases, "[t]he taboos seem likely to have arisen from the Westermarck effect," and he referred to the process of "translating the Westermarck effect into incest taboos." Lieberman and Lobel (2012: 27) argue that new evidence bolsters Westermarck's theory that (in their words) "personal sexual aversions that target childhood associates manifest themselves on the cultural stage as proscriptions against incest." Others stop short of endorsing this hypothesis, but treat it as a serious

contender. De Waal (1999: 99) says that an "intriguing question is whether the incest taboo merely serves to formalize and strengthen the Westermarck effect or whether it adds a substantially new dimension."

The causal story Westermarck gives for how we get from a biological disposition to a social norm faces some often-overlooked conceptual challenges. Wilson flouts these challenges when he asserts that Westermarckian inhibitions get "translat[ed]" into the incest taboo. Even de Waal, who does not necessarily jump on the Westermarck bandwagon, apparently sees no problem in principle with the idea that the inhibition could simply be "formalize[d]" in the taboo. But private inhibitions and social taboos are very different things. When we discover a biological disposition that seems to correspond to a norm or institution, we cannot automatically conclude that the latter is rooted in the former. What it would even mean for a norm to be "rooted" in a biological disposition is itself unclear.

Another set of examples illustrates the conceptual problems we face in seeking the biological basis of norms. Street (2006: 115) asks us to consider some "judgements about reasons" including the following: "The fact that something would promote the interests of a family member is a reason to do it." "We have greater obligations to help our own children than we do to help complete strangers." What accounts for the widespread acceptance of these judgements? According to Street: "Evolutionary biology offers powerful answers to [this question], very roughly of the form that *these* sorts of judgements about reasons tended to promote survival and reproduction much more effectively than the alternative judgements." The theory of kin selection, she says, explains why we accept the propositions listed above.

But there is a huge gap between the instincts predicted by the theory of kin selection and the *evaluative judgements* formulated by Street. For one thing, the parental

instinct that leads us to care more about the welfare of our own children than the welfare of strangers has a different *content* than the widely accepted moral/legal principle that "We have greater obligations to help our own children than we do to help complete strangers." If I have children, my instinct should lead me to want to help *these* children. Insofar as I care about other people's preferences, I should want them to be more favorably disposed toward *my* children than toward their own. Furthermore, Street's propositions have a moral dimension that the instincts predicted by kin-selection theory do not have.

Street (2006: 120) does acknowledge that there is often a gap of some kind between instinctual dispositions and the (supposedly) corresponding evaluative judgements when she writes:

> I do not mean that we automatically or inevitably accept the full-fledged evaluative judgements that line up in content with our basic evaluative tendencies....My point here is instead the simple and plausible one that had the general content of our basic evaluative tendencies been very different, then the general content of our full-fledged evaluative judgements would also have been very different, and in loosely corresponding ways.

But, as noted above, the content of our parental instinct (give preferential treatment to *my* children) does not "line up" with the content of the corresponding evaluative judgement (everyone *ought* to give preferential treatment to *their* children). It is not clear how or if the instinct is related to the judgement at all. So it is premature to conclude that, had the instinct been different, our evaluative judgement would be different in a "loosely corresponding" way.

This paper examines the sibling incest taboo as a case study to understand the relationship between biological dispositions and social norms, particularly moral norms. The analysis will bring out a number of conceptual problems that arise when we study

the biological underpinnings of social/moral norms generally, and incest taboos specifically. Ultimately, I argue that Westermarck's hypothesis cannot explain the sibling incest taboo.

The following section outlines the representation and moralization problems (Williams 1983; Wolf 2004b), which concern two different ways in which the Westermarck effect and the incest taboo differ in their content. Given that the Westermarck effect targets *childhood coresidents* and *personal behavior*, and the incest taboo targets *siblings* and *everyone's behavior*, it seems that the former cannot be responsible for the latter. The third section clarifies what "content" means in this context, and considers and rejects a proposed solution to the representation problem. The fourth section considers a possible solution to the moralization problem—what is termed the vicarious disgust hypothesis—and concludes that it is empirically unsupported. The fifth section considers alternative theories of the sibling incest taboo, focusing on the idea that it was established to promote social cohesion within families, or in response to recognition of the dangers of inbreeding. I argue that versions of the representation and moralization problem apply to the former but not the latter theory. The final section argues that the representation and moralization problems are *general* problems that could potentially undermine many popular evolutionary explanations of social/moral norms.

2. The Representation and Moralization Problems

Williams (1983) posed the "representation problem" for Westermarck's hypothesis: the content of the Westermarck effect—*feel sexual aversion toward childhood coresidents*—is different from the content of the incest taboo—*do not have sex with biological siblings*. Even if the Westermarck effect is real, he argued, it could have nothing to do with the

incest taboo, since the aversion and the taboo concern different behaviors. If the Westermarck effect gave rise to a moral norm it would be *do not have sex with childhood coresidents*.

Although Williams and subsequent commentators have treated the representation problem as a challenge for Westermarck's hypothesis specifically, the problem applies in virtually the same form to other theories about how instincts give rise to moral norms. Consider Street's (2006) suggestion that our innate kin-directed altruistic tendencies (implanted in us by kin selection) gave rise to the norm saying that we have a reason to promote the interests of family members. This theory is subject to the representation problem because natural selection did not give us an instinct to prefer our *family members*. Rather, it gave us a disposition to be more altruistic toward people who exhibit cues that were *historically correlated* with being our family members. Such cues presumably include spending a lot of time with and/or growing up in close proximity to oneself (see Lieberman et al. 2007; Silk 2009). But according to the moral norm mentioned by Street we might have greater obligations toward a cousin we barely know than we have toward a classmate who grew up with us, even if the latter elicits our innate altruistic response much more strongly.

Wolf (2004b: 11–12) posed the "moralization problem" for Westermarck's hypothesis: the content of the Westermarck effect concerns *one's own behavior*, whereas the content of the incest taboo concerns *everyone's behavior*. That is, there is an amoral–moral gap in content. As will be discussed in more detail in the following section, the content of the Westermarck effect (and many other adaptive dispositions) concerns the behavior of a specific person toward another specific person(s). As a general rule, the content of moral norms is not relativized to a particular person or concrete relationship. The content of the Westermarck effect, *avoid sex with my childhood coresident(s)*, is

different in *two* ways from the content of the incest taboo, *everyone ought to avoid sex with their sibling(s)*.

One might be tempted to think that the moralization problem has an easy solution. Isn't it obvious that a strong personal desire should lead us to endorse a moral norm that gives free expression to the desire? Since moral norms have to be universalizable, the content of the norm would have to take on a universalizable form, and this explains why the content of the disposition must change (so one could argue). But this a mistake. First, there is no obvious reason why we should think that, in general, strong desires will give rise to moral norms supporting the desired behavior. If most people have a desire to *X*, why create a norm saying that *everyone ought to X*? Why would a desire to *X* per se make us care whether or not other people *X*? Second, if we have a strong personal desire to *X*, this can be the basis for establishing a moral norm that *prohibits* or *curtails*, not *demands*, *X*ing. Street (2006) says that our impulse to help our kin leads us to endorse the norm that family members should help each other. But, as a matter of fact, many if not all societies—including *all* nomadic foragers (Boehm 1999)—have anti-nepotism norms (sometimes codified in law) that prohibit certain forms of family-directed support. In short, there is no reason why we should expect any particular personal desire (that does not directly concern third-party behavior) to be moralized.

We often face the following situation: we believe we have identified a naturally selected biological disposition, *D*, that prompts behavior that is similar to that prescribed by a social norm, *N*. We assume that *D* must have given rise to *N*—that the former is somehow causally responsible for the latter. But on closer inspection we find that the behaviors and attitudes produced by *D* are actually quite different from those demanded by *N*. Often *D* is an amoral instinct while *N* is moralized. If *D* really is causally responsible for *N*, it was not by a simple process of "formaliz[ing]" widely shared

individual dispositions. It may not even be necessary to refer to *D* at all in order to explain *N*.

3. The Content of Biological Dispositions and Social Norms

Street (2006) and Williams (1983) refer to the "content" of biological dispositions, which they compare with the "content" of corresponding social/moral norms. Street does this to call attention to supposed similarities—Williams, to differences—in the content of dispositions and norms. But what *is* the content of a disposition or a norm?

3.1. Social Norms

Sripada and Stich (2006: 281) define "norms" as "rule[s] or principle[s] that specif[y] actions that are required, permissible, or forbidden independently of any legal or social institution." Norms can be formally enforced, but they do not have to be. Sripada and Stich describe norms as having "*independent normativity*." That is to say, conforming to a norm is seen by the community that accepts it as being an end in itself. (In practice, of course, people often have ulterior motives.) Machery and Mallon (2010: 12) similarly note that the "content [of norms] essentially involves deontic concepts, such as SHOULD or OUGHT." Although Sripada and Stich say that norms regulate "actions," which suggests *physical behavior*, norms can also require, permit, or forbid emotional and attitudinal responses (see Machery and Mallon 2010: 12).

There is more to say about the concept of *norms*, but the definition given above is sufficient for the present discussion. Westermarck, Street, et al. are essentially

interested in the origin of the collectively accepted rules and principles regulating our behavior and attitudes.

Social norms may or may not be explicit. Some are never—or rarely—explicitly stated. For example, there is a rarely stated norm that you do not sit directly next to a stranger in a public place as long as there are other seats available.[1] Most people probably follow this rule without having been taught it explicitly and without thinking about it consciously. We might become conscious of such norms only when they are violated. Nevertheless, unstated norms are always stateable *in principle*. It is always possible to state the rules or principles to which people in a community expect each other to conform.

The *content* of a social norm is the semantic content conveyed when the norm is stated. We obtain the content of a norm simply by disquotation. The content of "do not steal" is do not steal. What exactly constitutes "stealing" may not be easy to articulate. But the content of the norm is embodied in its more or less vague linguistic expression as a proposition/imperative. The contents of Street's "evaluative judgements" are exactly what is expressed when those judgements are stated.

3.2. Biological Dispositions

Instincts are behavioral adaptations to respond in (historically) adaptive ways to certain environmental conditions. They have two component parts, namely, a part that tracks states of the environment and a part that executes a response. In classical ethology these are called "innate releasing mechanisms" and "fixed action patterns," respectively (Eibl-Eibesfeldt 1989: chapter 2). The way in which instincts operate as proximate causes of

[1] Incidentally, Eibl-Eibesfeldt (1989: 336–339) argues that this norm has a biological basis.

our behavior can be described by verbal propositions. A description of an instinct has the form "I exhibit response *R* when faced with stimulus *S*," where *S* is the cue(s) that the world-tracking part of the instinct uses to determine when to trigger *R*. The content of an instinct corresponds to the content of the proposition describing the proximate mechanisms that constitute the instinct. The content of the Westermarck effect is clearly different from that of the incest taboo.

Instincts are by definition adaptations, but they do not derive their content (in the sense described above) in virtue of being adaptations. Biological dispositions that are not adaptations—for example, accidental byproducts of instincts—have content in virtue of being describable with propositions of the form "I exhibit response *R* when faced with stimulus *S*."[2] If the Westermarck effect were real but, contra Westermarck, not an adaptation, it would still have the same content—"I avoid sex with childhood coresidents"—and this content could be compared with that of the incest taboo.

3.3. Another Way to Determine the Content of Dispositions: A Solution to the Representation Problem?

The most promising solution to the representation problem was proposed by Sesardic (1998). He argued that, according to Westermarck's hypothesis, it does not matter what property in our siblings *objectively* causes our sexual aversion toward them. It only matters what property we *subjectively* experience as the cause of the inhibition. When

[2] Such dispositions would lack *teleosemantic* content, which requires the mechanism that "consumes" a representation of the world to have a proper function based on its evolutionary history, which in turn determines the content of the representation (Millikan 1989). Because the evolutionary history of a biological disposition is not *as such* relevant to how we normally experience it, the notion of "content" upon which the representation problem is based cannot be grounded in evolutionary history.

we find ourselves unattracted to our siblings, we might "make sense of [our] own aversion...in terms of the most 'natural' category available (close kin relationship), and 'feel' that what is repellent...is sex with *close relatives*" (Sesardic 1998: 424). To explain our aversion, we "choose the line of least resistance and simply reach for the more meaningful social category like *brother* or *sister*, rather than a seemingly irrelevant and queer characteristic like *someone with whom I spent the first years of my childhood*" (Sesardic 2004: 114–115).

In light of Sesardic's observation, we need to distinguish between the *objective* content and the *subjectively experienced* content of biological dispositions. Williams (1983) implicitly assumed that these two kinds of content are always the same, but there is no reason to think that they are. Propositions expressing the objective (O) and the subjectively experienced (SE) content of a disposition have the same form, namely, "I exhibit response *R* when faced with stimulus *S*," but they may designate different *S*'s. When faced with an object exhibiting *R*-eliciting stimulus S_O, we may *feel* or *believe* that *R* is being elicited by some other property associated with the object, S_{SE}. Sometimes S_O and S_{SE} may be the same, sometimes they may be different.

However, there is a sense in which Sesardic's solution to the representation problem assumes what needs to be explained. He notes that *brother* and *sister* are "natural" and "meaningful" social categories whereas *childhood coresident* is not. That's true. But why are the former categories socially meaningful in the first place? These are salient, meaningful categories *because* we have social norms regulating the relationship between brothers and sisters—norms saying, for example, that they have certain obligations toward each other and prohibiting sex and marriage between them. To explain how the categories *brother* and *sister* became significant, we have to explain why

our ancestors decided to create social norms referring to these categories, thereby *causing* them to be significant.

Silk (2009: 3243) claims that "other primates distinguish kin from non-kin, form enduring relationships with their offspring, selectively groom, support and reconcile conflicts with their relatives and are aware of the kinship relationships between other group members." But the evidence she reviews does not support the idea that nonhuman primates are "aware" of relationships between kin *qua kin*. Rather, they instinctually bond with childhood associates, and are aware of similar bonds between third parties. Silk notes that the "primary basis for kin recognition...in primates" is "[c]lose association early in life....The importance of early familiarity is supported by the evidence that captive 'foster' mothers routinely accept strange infants, even when they are not the same sex, exact ages or species as their own infants" (p. 3244). There is no basis for thinking that nonhuman primates experience this bonding as "kin recognition," or have any such concept. They simply develop special feelings for other individuals—who do not necessarily even need to be the same species, let alone close kin—with whom they have specific kinds of close interaction.

We have to imagine a time before social norms regulating familial relationships were established, when our ancestors were equipped with dispositions to be altruistic and feel sexual aversion toward those with whom they had close associations when growing up (who were often but not always close kin). Would *brother* and *sister* have been more "natural" categories than *childhood coresident* for these early humans? The answer is not obvious. Childbirth is a conspicuous event, and under normal circumstances the bond between mothers and their children is strong in all primates. It is plausible that the categories *mother* and *offspring* would have been easy to recognize. Once *mother* and *offspring* became significant categories, perhaps it would have been natural to recognize

the category of *offspring born to a single woman*, i.e., maternal siblings. On the other hand, early human children probably often grew up and had intimate relationships with people who were not born to the same mother. Often enough these relationships might have produced feelings of altruism and sexual aversion greater than or comparable to what people felt toward their biological siblings, particularly siblings separated in age by several years.

One of the main sources of evidence for the Westermarck effect comes from studies of arranged marriages in Taiwan. In so-called "minor marriages," a bride was adopted by the husband's family as an infant and the future couple was raised as if they were siblings. Wolf (2004a) found evidence that these marriages were marked by sexual dissatisfaction reflected in lower fertility and higher divorce rates. However, if the future husband was at least nine years older than bride, there was no evidence of sexual dissatisfaction. In the same vein, Lieberman et al. (2007) found that the strength of sibling altruism depends on the amount of time spent together before age 18. It is thus reasonable to suppose that, historically, biological siblings separated in age by several years would have frequently failed to develop strong Westermarckian inhibitions or feelings of altruism, while unrelated children sometimes formed strong bonds. *Childhood coresident*—someone with whom one grew up in intimate contact—could conceivably have been recognized as a significant category, but for some reason it was not.

Even if Sesardic's solution solved the representation problem—and I have suggested reasons to doubt that it does—it would not solve the moralization problem. If I have the subjective experience of feeling sexual aversion toward my sibling (qua sibling), why should I support a rule that you cannot have sex with *your* sibling?

4. The Vicarious Disgust Hypothesis: A Solution to the Moralization Problem?

Suppose for the sake of argument that the Westermarck effect is real and that its subjectively experienced content is *avoid sex with my siblings*. Even still, the content of the Westermarck effect does not concern third-party behavior. There is no logical link between the biological disposition to *avoid sex with my siblings* and the moral norm *everyone should avoid sex with their siblings*. The process described by Westermarck requires a psychological mechanism that transforms personal sexual aversions into moral opposition to third parties engaging in sexual behavior analogous to that which we find aversive.

Westermarck (1921) explains the process by which sexual aversion toward our own close kin (due to the Westermarck effect) gives rise to a moral/legal prohibition on incest as follows: "sexual indifference is very generally combined with sexual aversion when the act is thought of....And...aversions which are generally felt readily lead to moral disapproval and prohibitory customs or laws" (pp. 197–198). Because "[p]ersons who have been living closely together from childhood are as a rule near relatives[,]...their aversion to sexual relations with one another displays itself in custom and law as a prohibition of intercourse between near kin" (p. 193). We can call this the *vicarious disgust hypothesis*.

The vicarious disgust hypothesis makes two claims. First, third-party incest triggers empathic disgust (which is rooted in our Westermarckian aversion toward our childhood coresidents). Second, people judge disgusting actions to be morally wrong—that is, disgust leads us to condemn acts that would *otherwise be considered morally neutral*. But the second claim is not well supported, and there is direct empirical evidence

against the idea that Westermarckian aversions lead to vicarious disgust and moral condemnation.

The idea that we judge disgusting actions to be morally wrong is intuitively plausible, and is often assumed to be correct by psychologists. Is it true? There is evidence that inducing disgust (especially via olfaction) can *amplify* the moral condemnation of acts that *would be judged immoral in the absence of the disgust prime*. But even defenders of the amplification theory acknowledge that the amplification effect is small and usually only found in individuals with certain personality traits such as low mindfulness or high disgust sensitivity (Schnall et al. 2015). In any case, the vicarious disgust hypothesis assumes not that disgust *amplifies* moral condemnation but that it leads us to condemn acts that would otherwise be judged morally neutral. The evidence for this is weak.

Wheatley and Haidt (2005) gave highly hypnotizable subjects a posthypnotic suggestion to feel disgust when they encountered one of two words ("often" or "take"). Subjects then read a vignette describing innocuous behavior: a student council representative "[tries to take/often picks] topics that appeal to both professors and students in order to stimulate discussion" (p. 782). Wheatley and Haidt report that "[p]articipants judged the actions to be more morally wrong when their hypnotic word was present" (p. 782). This study has been widely cited to support the claim that disgust causes moral condemnation. However, it is questionable whether Wheatley and Haidt's subjects thought that the behavior described in the vignette was actually *immoral*. When moral evaluations were converted to a 100-point scale (from *not at all morally wrong* to *extremely morally wrong*), the average ratings for the control and experimental groups were 2.7 and 14.0, respectively. As May (2014: 129) observes, a mean rating of 14 on a 100-point scale "still seems to count the action as *not* morally wrong."

Several other studies have found that disgust can cause people to condemn (what would normally be regarded as) morally neutral acts, but the effect is small. Landy and Goodwin (2015: 530) point out that the effect is smaller in unpublished than published studies, indicating possible publication bias in the literature. Both liberals and conservatives say that "[w]hether or not someone did something disgusting" is one of the factors they consider when deciding if their action was wrong, with conservatives assigning slightly more evaluative importance to disgustingness (Haidt and Graham 2007). However, as Pizarro et al. (2011: 268) note, "there is a plethora of behaviors that are judged by most people as disgusting but not immoral, such as eating pig brains or picking one's nose in private." So disgust does not lead inevitably to moral condemnation. If our disgust toward incest is what motivates us to establish the incest taboo, there must be something special about Westermarckian-based sexual disgust that makes us morally condemn the disgusting act in others.

Lieberman and Lobel (2012) claim to provide evidence for a direct link between Westermarck effect-induced sexual disgust and moral opposition to the disgusting sexual behavior. But a closer look at their data suggests that this conclusion is highly questionable.

Children in Israeli kibbutzim were traditionally raised from birth in same-age peer groups composed of around 15 nonrelatives. Growing up they would eat, sleep, and (at least as children) shower together. Each day, they would spend at most three hours with their parents and biological siblings. A frequently cited piece of evidence for the Westermarck effect is the (alleged) finding that when these children grew up, they did not feel sexual attraction toward each other despite the fact that there was no formal prohibition on sex between them (Shepher 1971, 1983; Wolf 2004b).

Lieberman and Lobel (2012) conducted a survey of communally raised kibbutz children to measure their sexual attraction to their peers vs. siblings, and also their moral attitudes toward peer vs. sibling sex. They claim that their results support Westermarck's theory that childhood coresidence leads to sexual aversion and moral condemnation of the aversive act. Sexual aversion toward peer-group members was positively correlated with time spent residing with them, and moral opposition to third-party intra-peer-group sex was positively correlated with subjects' own aversion toward their peers. Thus, they conclude, "personal sexual aversions shape attitudes relating to third-party sexual behavior" (p. 26). Let's look more closely at the evidence for this.

Because children joined the kibbutzim at different ages, peer-group members resided with each other for varying lengths of time. Lieberman and Lobel predicted that the kibbutzniks' sexual aversion toward each of their peers would be correlated with the amount of time they spent together before age 18. Following Westermarck's logic, they also predicted that the total amount of time people spent in communal peer groups—and the degree of aversion they felt toward their peers—would be associated with greater moral opposition to *third-party* intra-peer-group sex.

Lieberman and Lobel (2012) conducted two surveys. In both surveys, respondents reported their feelings about kissing as well as having sex with each of their opposite-sex peers on a scale ranging from "−5 (*very disgusting*) to 5 (*very appealing*); midpoint 0 (*neutral, not disgusting* or *appealing*)" (p. 28). For their analyses, Lieberman and Lobel combined the two responses to form a "sexual attraction measure" ranging from −10 to 10. They calculated each subject's "total sexual attraction score" by adding the sexual attraction scores for all of the subject's opposite-sex peers.

They measured moral attitudes toward third-party sexual practices in different ways in the two studies. In Study 1, subjects read two vignettes (based on the Mark-and-

Julie vignette used by Haidt 2001) describing consensual and secret sex between siblings or peer-group members that was enjoyed by both participants and had no chance of leading to pregnancy. "For each vignette, subjects indicated how wrong it would be for the individuals described to start a sexual relationship and, separately, to get married on a scale ranging from −5 (*very morally wrong*) to 5 (*morally ok*); midpoint 0 (*neutral, not morally wrong* or *morally ok*)" (Lieberman and Lobel 2012: 29). Scores for each vignette were combined to calculate a "morality dependent measure toward third-party peer sex and...third-party sibling sex" (p. 29).

In Study 2, subjects ranked 10 behaviors, including sibling and intra-peer-group sex, in order of how morally wrong they perceived them to be.

Was the Westermarck-inspired hypothesis—that time spent coresiding with peers would be associated with increased moral opposition to third-party intra-peer-group sex—supported? In Study 1, Lieberman and Lobel found *no statistically significant effect* of time spent coresiding with peers on moral attitudes toward third-party peer sex. That is, people who spent more time coresiding with opposite-sex peers were *not* more likely to say it would be morally wrong for two peer-group members to start a sexual relationship or to get married.

In Study 2, however, they did find a statistically significant effect. When rank ordering 10 behaviors from most to least morally wrong, people who spent more time coresiding with opposite-sex peers ranked "Two Kibbutz classmates having sex" slightly lower (i.e., more morally wrong). The average rank of "Two Kibbutz classmates having sex" was 8.86 (SD = 1.55), which was just above "Speeding on the highway" (average rank = 6.66) and below the highest ranked behavior, "Converting to another religion" (average rank = 9.00). The average rank of "A brother and sister having sex" was much lower at 3.76 (SD = 2.15). Subjects' total coresidence duration with opposite-sex peers

had a correlation of −.34 with their ranking of intra-peer-group sex. Lieberman and Lobel found that the association between coresidence duration and rank ordering for intra-peer-group sex was entirely mediated by personal sexual aversion—i.e., total coresidence duration did not predict rank ordering when controlling for subjects' sexual attraction/aversion toward their opposite-sex peers.

Lieberman and Lobel (2012: 32) conclude that their findings

> provide the missing psychological link for past studies documenting the virtual absence of marriage between coreared Kibbutz peers (Shepher 1983)—marriage rates were likely reduced not only because of mating opportunities outside the Kibbutz...but also inherent sexual aversions that developed to those within....These results support Westermarck's claim that "aversions which are generally felt readily lead to moral disapproval"....

But Lieberman and Lobel (2012) seem to be significantly overstating their case. The relationship between coresidence duration with (and sexual aversion toward) opposite-sex peers and "moral views relating to sex between peers" (p. 30) was extremely weak. Again, Lieberman and Lobel's Study 1 did not find *any* effect of total coresidence duration on reported moral attitudes toward third-party sex. Study 2 found that coresidence duration (mediated by sexual aversion) had a very small effect on where people placed "Two Kibbutz classmates having sex" when rank ordering 10 behaviors from most to least immoral. But subjects ranked "A brother and sister having sex" on average 5.1 places lower than they ranked "Two Kibbutz classmates having sex," despite the fact that they had much closer contact with their non-sibling classmates during the supposedly critical period for developing sexual aversion. There is no evidence that they

regarded kibbutz peer-group members having sex as genuinely immoral. On the other hand, they clearly regarded sibling sex as immoral.

The vicarious disgust hypothesis was a possible solution to the moralization problem, but not to the representation problem. Even if it were empirically supported, it would not explain why our ancestors came to believe that sex between *siblings* but not *childhood coresidents* is morally problematic. (The representation problem did not arise in kibbutzim, because we know that *peer-group member* was made into a socially significant category via social engineering.) But, as the evidence reviewed above suggests, there is no compelling evidence for a psychological mechanism that converts disgust in general—or Westermarckian sexual aversion specifically—directly into moral condemnation.

5. Alternative Theories of Incest Taboos

Incest taboos vary enormously across time and place, and there is no specific prohibition found in all societies. Even the taboo on sibling sex is far from being universal. Many traditional societies simply lacked norms concerning sibling incest (Thornhill 1991). For at least two-and-a-half centuries in Roman Egypt there seems to have been a positive preference for sibling marriage among commoners (Hopkins 1980). For one-and-a-half millennia Zoroastrians practiced sibling as well as parent–child marriage with strong encouragement from their religious authorities (Scheidel 1996). Some cultures that *did* have sibling incest taboos nevertheless made exceptions for the ruling families (Middleton 1962).

Any society's norms governing sex will be shaped by a panoply of historical, sociological, and evolutionary forces, all of which interact with human biological

dispositions. There is almost certainly no single explanation of the sibling incest taboo that fully explains every instance of it. But it may be possible to identify factors that contributed to the development or persistence of the taboo in many cultures.

Theories of incest taboos can be divided into two general categories: *intentional* and *selectionist*. According to intentional theories—which can also be described as *sociological*—taboos were (consciously or unconsciously) designed by people to serve some (perceived) function. According to selectionist theories, taboos are the direct or indirect product of natural selection. That is, natural selection could (directly) favor individuals or groups with adaptive taboos, or it could (indirectly) endow people with adaptions from which taboos emerge as a byproduct. (Westermarck's hypothesis is selectionist, with natural selection producing the taboo indirectly via the Westermarck effect.) Again, different combinations of forces may shape the incest taboos of different societies. Natural selection always plays some role in cultural evolution by default because all cultural innovations are subject to natural selection. Regardless of how cultural practices originate, individuals or groups with more adaptive variants tend to proliferate.

Sex and marriage taboos may have profound implications for group functioning and adaptedness. Many social scientists in the twentieth century believed that, without the nuclear family incest taboo forcing people to make alliances outside the family unit, distinctively human society would be impossible. Lévi-Strauss (1956: 278), for example, declared:

> it will never be sufficiently emphasized that, if social organization had a beginning, this could only have consisted in the [nuclear family] incest prohibition....It is there, and only there, that we find a passage from nature to culture, from animal to human life, and that we are in a position to understand the very essence of their articulation.

The fact that sibling and parent–child incest taboos are not universal casts doubt on the claim that they are the distinguishing features of human social life, but it is true that forced exogamy can foster interfamilial cooperation. Sex and marriage taboos can have other effects that influence group adaptedness in ways that could not be foreseen by any human designer. Schulz et al. (2019) argue that the Catholic Church's ban on cousin marriage triggered a cascade of cultural changes that made Westerners more individualistic, independent, and impersonally prosocial and less conformist and loyal to their ingroup. The taboo on sibling sex and marriage may have conferred advantages to groups that adopted it, which help account for its spread.

This section considers two promising sociological theories of incest taboos: they were established (a) to maintain social cohesion within family units or (b) in response to recognition of the dangers of inbreeding. The social cohesion theory turns out to be subject to versions of the representation and moralization problems, and probably does not explain the origin of the nuclear family incest taboo, though it may help explain its persistence. The recognition hypothesis is not subject to the representation and moralization problems, and it is supported by some suggestive evidence.

5.1. The Social Cohesion Hypothesis

According to Shor and Simchai (2009), people in small and highly cohesive "nonvoluntary groups" will suppress their sexual attraction to each other and/or establish norms against ingroup sexual relations. People do this because they are aware that ingroup sexual relationships will threaten the group's unity and "may even lead to its dissolution" (p. 1814). People also fear the potential awkwardness that would ensue if

they are rebuffed by, or have a failed romantic relationship with, someone with whom they must continue living in close contact. In Shor and Simchai's words:

> when the institutional frame (be it family or any other small nonvoluntary group) is greatly cohesive and associations are dense, individuals identify the social and personal price of intimate dyadic relationships in terms of both group cohesiveness and potential embarrassment. Under such conditions, any expression of sexual emotions or drives may be consciously or unconsciously suppressed. (p. 1814)

They marshal evidence for this theory from an unlikely source, namely, the same kibbutzim that are usually held up as supporting Westermarck. Proponents of Westermarck claim that coreared kibbutzniks felt sexual aversion toward each other in the absence of a taboo on intra-peer-group sex (Shepher 1971, 1983). Shor and Simchai found evidence that this is doubly wrong. Careful interviews with former kibbutzniks revealed that they did not typically feel sexual aversion toward childhood coresidents, and in many kibbutzim there *was* a taboo on intra-peer-group sex. Most of Shor and Simchai's interviewees felt attraction to at least one of their childhood coresidents: 33.3% reported "strong" and 20% reported "moderate" sexual attraction toward at least one peer-group member (p. 1822). Slightly less than half reported sexual *indifference*, and almost no one reported sexual *aversion*. Many interviewees said that their peer group had norms against forming intragroup romantic relationships, and some said that they avoided making romantic overtures for fear of embarrassment. Kibbutzniks who described their peer group as being relatively cohesive were considerably more likely to report feeling sexual indifference toward their peers.

Shor and Simchai's claim that being part of a cohesive, nonvoluntary group leads at most to sexual *indifference* raises the question of why people typically feel sexual *aversion* toward their nuclear family members. They offer two possible answers. First,

the sexual indifference we feel toward close family members in virtue of belonging to a cohesive, nonvoluntary group is reinforced by a strong cultural taboo, which turns the indifference into an aversion. Second, intrafamilial attraction may be more common than we usually think: "despite this powerful taboo, studies have found considerable rates of incest inside nuclear families, in America and elsewhere, including between siblings" (p. 1833). The second suggestion is questionable in light of the evidence that a large proportion of people—including Shor and Simchai's own interviewees—report genuine aversion toward first- and third-party incest.

Shor and Simchai seem to have identified a real phenomenon, namely, people sometimes supress sexual attraction and/or establish norms against intragroup sex for the (more or less explicit) purpose of preserving group cohesion. But as an explanation for the incest taboo, this theory faces its own version of the representation and moralization problems. If I suppress my incestuous attractions to avoid damaging my close relationships with my family members, why should I support a rule prohibiting specifically *incestuous* attractions? That is to say, the content of my inhibition is *avoid sex with fellow members of a cohesive, nonvoluntary group*, but the content of the taboo is *avoid sex with close family members (whether or not they are part of a cohesive, nonvoluntary group)*. And why should I care what *other* people do? Maybe I have a reason to want my own family members to suppress their attraction to *me*, but there is no reason to be concerned with the desires of third parties.

Even if Shor and Simchai's social cohesion hypothesis does not account for the origin of the sibling incest taboo, it may help explain the power and resilience of the taboo. Many nuclear families are relatively cohesive, nonvoluntary groups, and our tendency to suppress sexual attraction within such groups may reinforce incest taboos.

5.2. The Recognition Hypothesis

According to the recognition hypothesis, people recognized the link between inbreeding and birth defects. In hunter–gatherer and other small-scale societies, the birth of physically or mentally disabled people could burden the entire group. People had clear reasons to oppose third-party sexual behavior likely to produce offspring who would never become effective workers or warriors. After incest prohibitions were established, many societies—including Western societies—forgot their original purpose.

Burton (1973) was the first person to argue that incest prohibitions were established in response to conscious recognition of the dangers of inbreeding. He reported his impression, based on a "cross-cultural study of resistance to temptation," that "the most common reason given in both primitive and modern societies for the incest taboo is that it produces bad stock" (pp. 504–505).

In a more systematic study, Durham (1991: 347–348) found evidence for fairly widespread recognition of the dangers of inbreeding. He investigated the "consequences of [incest] taboo violations, real or hypothetical," according to people in 60 cultures. In 40 of the cultures, people reported that some negative socially or divinely imposed consequence would follow from incest. In 29, people reported that incest was punished by a social sanction, "most commonly death." In 16, "there was mention of a *supernatural* sanction, most commonly disease or death" (p. 347). (In five cultures people who engaged in incest were subject to social sanctions *and* were considered liable for supernatural punishment.) Most important, in 20 of these societies—i.e., one-half of those that reported some sort of negative consequence for incest—Durham found references to a "'bad stock' argument." For several more of these societies there was ambiguous evidence, or evidence from other ethnographic databases, that they

recognized the negative consequences of inbreeding for offspring. Durham also notes that there was a correlation between the richness of the ethnographic data for a given society and the chance that he could find evidence of a folk theory about incest producing disabled offspring (pp. 348–349, n. 56). Thus, 20 out of 60 is likely an *underestimate* of the proportion of societies that achieved and preserved awareness of the negative effects of inbreeding.

Durham (1991) made another discovery that he sees as supporting the recognition hypothesis: incest rules in different societies reflect the relative dangers of inbreeding. The risk that close inbreeding will produce disabled offspring varies among populations. Large, exogamous populations tend to accumulate more deleterious recessive alleles. Because reproductive partners are, on average, distantly related to each other, the partners are less likely to have inherited the same deleterious mutations. Consequently, the mutations are less likely to be weeded out through the production of nonviable homozygotes. This makes inbreeding relatively dangerous. In contrast, reproductive partners in small, endogamous populations must be relatively closely related. In these populations deleterious recessive alleles will be more quickly weeded out, making inbreeding less dangerous. The data analyzed by Durham suggest that "small endogamous populations do seem to have less extensive incest taboos than do their exogamous counterparts" (p. 356). He suggests that taboos were tailored to the relative risks of inbreeding in different societies by human design. However, the data might also be explained as a consequence of natural selection favoring groups with locally adaptive taboos.

What about the sibling incest taboo in Western societies? The feeling of visceral disgust that incest arouses (Royzman et al. 2008) is not typically bound up—at least not consciously—with awareness of any negative consequences for offspring. In fact, until

the 1960s there was a near scientific consensus in the West that inbreeding does not pose any dangers (Wolf 2004b: 1–3).[3] Now that the risks of inbreeding are widely known, people do often cite this as a reason to oppose incest, though they frequently cite psychological harm to the participants, and many say that incest is wrong in principle regardless of the consequences (Royzman et al. 2015). It is possible that our predecessors established the sibling incest taboo in response to recognition of the dangers of close inbreeding, and we retained the taboo after (temporarily) forgetting its original purpose.

There is evidence that this is exactly what happened with the taboo on cousin sex. Marriage between first cousins is perfectly accepted—if not encouraged—in many cultures, and used to be common in the West. In the late sixth century, Pope Gregory I claimed that it was prohibited because of the consequences for offspring:

> A certain secular law in the Roman State allows that the son or daughter of a brother and sister, or of two brothers or two sisters may be married. But we have learned from experience that the offspring of such marriages cannot thrive. Sacred law forbids a man to uncover the nakedness of his kindred. Hence it is necessary that the faithful should only marry relations [at least] three or four times removed. (quoted in Durham 1991: 331)

Theoretically the same thing could have happened with the sibling incest taboo. Alternatively, self-interest may have been the true reason for the Catholic Church's ban

[3] Most of Westermarck's contemporaries rejected his theory of inbreeding depression. Malinowski (1927: 243) asserted that "biologists are in agreement on the point that there is no detrimental effect produced upon the species by incestuous unions" (quoted in Wolf 2004b: 2). This (near) consensus lasted into the 1950s, and went largely unquestioned at a conference devoted to the incest taboo held at Stanford in 1956 (Aberle et al. 1963; Wolf 2004b: 3). Mainstream scientific opinion did not reverse until the 1960s in response to advances in genetics (Aberle et al. 1963: 256–257).

on cousin marriage. In the fourth century the Church established a number of laws—including prohibitions on adoption, polygyny, and cousin marriage—that increased the likelihood of people dying without heirs, in which case their property was inherited by the Church (Goody 1983; Prinz 2007: 231–232). In any case, the fact that Westerners continue to taboo sex and marriage between first cousins suggests that taboos can persist and continue to elicit emotional commitment long after their original purpose has been forgotten. Cultural evolution is often conservative, and for good reason. It is usually adaptive for people to accept the beliefs, values, and practices that they inherit unless they have a compelling reason to reject them (Henrich 2016). And, as Buchanan and Powell (2018: 251–252) observe, the elements of a culture are interconnected. Norms—particularly those that "implicate group identity or moral identity"—may become "culturally entrenched," since they "occupy a central, highly connected position in the cultural web." It may be possible for culturally entrenched taboos to be transmitted down the generations even if the original justification is forgotten.

According to the recognition hypothesis, people in the distant past noticed that sexual relations between those with certain biological relationships—including siblings—tended to lead to negative consequences for offspring. There is no representation problem for this theory. If sex between childhood coresidents who were *not* siblings led to negative consequences, then people would likely have noticed this and established a taboo to keep childhood coresidents apart. But our biology does not work that way, so *childhood coresident* never became an especially important social category.

6. Discussion

The widespread prohibition on sibling incest has been held up by many scholars as the

best or one of the best examples of a moral norm that has an evolutionary explanation. According to Westermarck's (1891, 1921) influential theory, we evolved to feel sexual aversion toward childhood coresidents (who are usually our siblings) as an adaptation for inbreeding avoidance. This aversion leads us to condemn incest as immoral, which explains the incest taboo. However, the representation and the moralization problems show that there is no logical connection between feeling sexual disgust toward one's own childhood coresidents and judging sibling sex to be immoral. I have argued that Westermarck's account of the sibling incest taboo is probably wrong.

Other moral norms for which evolutionary explanations have been proposed will have to be analyzed separately. However, there is reason to believe that representation and moralization problems are pervasive. Instincts to *have empathy for people we grew up with* or to *retaliate when someone harms us* do not lead directly to moral norms such as *everyone ought to care for their children* or *anyone who forcibly takes the property of another person should be punished*. The gap between the content of our instincts and moral norms would help explain the fact that there is virtually no substantive universal moral norm (Prinz 2008), despite the fact that basic instincts *are* universal. As noted, even the prohibition on nuclear family incest is far from being a human universal (Thornhill 1991). However, moral variation is not infinite. Sripada and Stich (2006: 284) observe that "norms tend to cluster under certain general *themes*," and Curry et al. (2019) found that categories of behavior related to cooperation are especially likely to be moralized across cultures. The appearance of general themes in morality does not necessarily mean that moral norms are determined by instincts. Across cultures people with a capacity for normative reasoning are endowed with similar impulses, face some of the same challenges, and are subject to some of the same selection pressures. It is not surprising that general themes would emerge independently in different moral systems

even if a wider range of variation is theoretically possible given the constraints imposed by our biology.

No one believes that every one of our moral beliefs can be explained as the direct product of natural selection. On all accounts, sociological factors also exert at least some influence. But a number of scholars believe that certain core, widespread moral norms are the more or less direct expression of biologically based adaptations. Street (2006: 119–120) assumes that "natural selection has had a tremendous *direct* influence on...our 'more basic evaluative tendencies', and...these basic evaluative tendencies, in their turn, have had a major influence on the evaluative judgements we affirm." Even this qualified position may be too strong. Our evaluative tendency to view incest as immoral is among the most "basic" that we have. Yet there is reason to believe that our opposition to incest not only *was not* but also *could not* have been directly shaped by natural selection via an instinct with different content.

Acknowledgements

I am grateful to Roger Crisp, Guy Kahane, Andreas Mogensen, Neven Sesardić, and two anonymous reviewers for helpful comments on earlier drafts of this paper.

References

Aberle, David F., Urie Bronfenbrenner, Eckhard H. Hess, Daniel R. Miller, David M. Schneider, and James N. Spuhler. 1963. "The Incest Taboo and the Mating Patterns of Animals." *American Anthropologist* 65 (2): 253–265.

Boehm, Christopher. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

Buchanan, Allen, and Russell Powell. 2018. *The Evolution of Moral Progress: A Biocultural Theory*. New York: Oxford University Press.

Burton, Roger V. 1973. "Folk Theory and the Incest Taboo." *Ethos* 1 (4): 504–516.

Curry, Oliver Scott, Daniel Austin Mullins, and Harvey Whitehouse. 2019. "Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies." *Current Anthropology* 60 (1): 47–69.

de Waal, Frans B. M. 1999. "The End of Nature Versus Nurture." *Scientific American* 281 (6): 94–99.

Durham, William H. 1991. *Coevolution: Genes, Culture, and Human Diversity*. Stanford, CA: Stanford University Press.

Eibl-Eibesfeldt, Irenäus. 1989. *Human Ethology*. New York: Aldine de Gruyter.

Goody, Jack. 1983. *The Development of the Family and Marriage in Europe*. Cambridge: Cambridge University Press.

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–834.

Haidt, Jonathan, and Jesse Graham. 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize." *Social Justice Research* 20 (1): 98–116.

Henrich, Joseph. 2016. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Hopkins, Keith. 1980. "Brother–Sister Marriage in Roman Egypt." *Comparative Studies in Society and History* 22 (3): 303–354.

Landy, Justin F., and Geoffrey P. Goodwin. 2015. “Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence.” *Perspectives on Psychological Science* 10 (4): 518–536.

Lévi-Strauss, Claude. 1956. “The Family.” In *Man, Culture, and Society*, edited by Harry L. Shapiro, 261–285. New York: Oxford University Press.

Lieberman, Debra, and Thalma Lobel. 2012. “Kinship on the Kibbutz: Coresidence Duration Predicts Altruism, Personal Sexual Aversions and Moral Attitudes among Communally Reared Peers.” *Evolution and Human Behavior* 33 (1): 26–36.

Lieberman, Debra, John Tooby, and Leda Cosmides. 2007. “The Architecture of Human Kin Detection.” *Nature* 445 (7129): 727–731.

Machery, Edouard, and Ron Mallon. 2010. “Evolution of Morality.” In *The Moral Psychology Handbook*, edited by John M. Doris, 3–46. Oxford: Oxford University Press.

Malinowski, Bronislaw. 1927. *Sex and Repression in Savage Society*. London: Kegan Paul, Trench, Trubner.

May, Joshua. 2014. “Does Disgust Influence Moral Judgment?” *Australasian Journal of Philosophy* 92 (1): 125–141.

Middleton, Russell. 1962. “Brother–Sister and Father–Daughter Marriage in Ancient Egypt.” *American Sociological Review* 27 (5): 603–611.

Millikan, Ruth Garrett. 1989. “Biosemantics.” *The Journal of Philosophy* 86 (6): 281–297.

Pizarro, David, Yoel Inbar, and Chelsea Helion. 2011. “On Disgust and Moral Judgment.” *Emotion Review* 3 (3): 267–268.

Prinz, Jesse. 2007. *The Emotional Construction of Morals.* Oxford: Oxford University Press.

———. 2008. “Is Morality Innate?” In *Moral Psychology, Vol. 1: The Evolution of Morality: Adaptations and Innateness,* edited by Walter Sinnott-Armstrong, 367–406. Cambridge, MA: MIT Press.

Royzman, Edward B., Kwanwoo Kim, and Robert F. Leeman. 2015. “The Curious Tale of Julie and Mark: Unraveling the Moral Dumbfounding Effect.” *Judgment and Decision Making* 10 (4): 296–313.

Royzman, Edward B., Robert F. Leeman, and John Sabini. 2008. “‘You Make Me Sick’: Moral Dyspepsia as a Reaction to Third-Party Sibling Incest.” *Motivation and Emotion* 32 (2): 100–108.

Ruse, Michael, and Edward O. Wilson. 1986. “Moral Philosophy as Applied Science.” *Philosophy* 61 (236): 173–192.

Scheidel, Walter. 1996. “Brother–Sister and Parent–Child Marriage Outside Royal Families in Ancient Egypt and Iran: A Challenge to the Sociobiological View of Incest Avoidance?” *Ethology and Sociobiology* 17 (5): 319–340.

Schnall, Simone, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan. 2015. “Landy and Goodwin (2015) Confirmed Most of Our Findings Then Drew the Wrong Conclusions.” *Perspectives on Psychological Science* 10 (4): 537–538.

Schulz, Jonathan F., Duman Bahrami-Rad, Jonathan P. Beauchamp, and Joseph Henrich. 2019. “The Church, Intensive Kinship, and Global Psychological Variation.” *Science* 366 (6466): eaau5141.

Sesardic, Neven. 1998. “From Biological Inhibitions to Cultural Prohibitions, or How Not to Refute Edward Westermarck.” *Biology & Philosophy* 13 (3): 413–426.

———. 2004. “From Genes to Incest Taboos: The Crucial Step.” In *Inbreeding, Incest, and the Incest Taboo: The State of Knowledge at the Turn of the Century,* edited by Arthur P. Wolf and William H. Durham, 109–120. Stanford, CA: Stanford University Press.

Shepher, Joseph. 1971. “Mate Selection among Second Generation Kibbutz Adolescents and Adults: Incest Avoidance and Negative Imprinting.” *Archives of Sexual Behavior* 1 (4): 293–307.

———. 1983. *Incest: A Biosocial View.* New York: Academic Press.

Shor, Eran, and Dalit Simchai. 2009. “Incest Avoidance, the Incest Taboo, and Social Cohesion: Revisiting Westermarck and the Case of the Israeli Kibbutzim.” *American Journal of Sociology* 114 (6): 1803–1842.

Silk, Joan B. 2009. “Nepotistic Cooperation in Non-Human Primate Groups.” *Philosophical Transactions of the Royal Society B* 364 (1533): 3243–3254.

Sripada, Chandra Sekhar, and Stephen Stich. 2006. “A Framework for the Psychology of Norms.” In *The Innate Mind, Vol. 2: Culture and Cognition,* edited by Peter Carruthers, Stephen Laurence, and Stephen Stich, 280–301. Oxford: Oxford University Press.

Street, Sharon. 2006. “A Darwinian Dilemma for Realist Theories of Value.” *Philosophical Studies* 127 (1): 109–166.

Thornhill, Nancy Wilmsen. 1991. “An Evolutionary Analysis of Rules Regulating Human Inbreeding and Marriage.” *Behavioral and Brain Sciences* 14 (2): 247–261.

Westermarck, Edward. 1891. *The History of Human Marriage.* London: Macmillan.

———. 1921. *The History of Human Marriage.* Vol. 2. 5th edition. London: Macmillan.

Wheatley, Thalia, and Jonathan Haidt. 2005. “Hypnotic Disgust Makes Moral Judgments More Severe.” *Psychological Science* 16 (10): 780–784.

Williams, Bernard. 1983. “Evolution, Ethics, and the Representation Problem.” In *Evolution from Molecules to Men*, edited by D. S. Bendall, 555–566. Cambridge, UK: Cambridge University Press.

Wilson, Edward O. 1998. *Consilience: The Unity of Knowledge.* New York: Vintage Books.

Wolf, Arthur P. 2004a. “Explaining the Westermarck Effect: Or, What Did Natural Selection Select For?” In *Inbreeding, Incest, and the Incest Taboo: The State of Knowledge at the Turn of the Century,* edited by Arthur P. Wolf and William H. Durham, 76–92. Stanford, CA: Stanford University Press.

———. 2004b. “Introduction.” In *Inbreeding, Incest, and the Incest Taboo: The State of Knowledge at the Turn of the Century,* edited by Arthur P. Wolf and William H. Durham, 1–23. Stanford, CA: Stanford University Press.

Wolf, Arthur P., and William H. Durham, eds. 2004. *Inbreeding, Incest, and the Incest Taboo: The State of Knowledge at the Turn of the Century*. Stanford, CA: Stanford University Press.

Power in Cultural Evolution and the Spread of Prosocial Norms[*]

Abstract: According to cultural evolutionary theory in the tradition of Boyd and Richerson, cultural evolution is driven by individuals' learning biases, natural selection, and random forces. Learning biases lead people to preferentially acquire cultural variants with certain contents or in certain contexts. Natural selection favors individuals or groups with fitness-promoting variants. Durham argued that Boyd and Richerson's approach is based on a "radical individualism" that fails to recognize that cultural variants are often "imposed" on people regardless of their individual decisions. Fracchia and Lewontin raised a similar challenge, suggesting that the success of a variant is often determined by the degree of power backing it. With power, a ruler can impose beliefs or practices on a whole population by diktat, rendering all of the forces represented in cultural evolutionary models irrelevant. It is argued here, based on work by Boehm, that, from at least the time of the early Middle Paleolithic, human bands were controlled by powerful coalitions of the majority that deliberately guided the development of moral norms to promote the common good. Cultural evolutionary models of the evolution of morality have been based on false premises. However, Durham and Fracchia and Lewontin's challenge does not undermine cultural evolutionary modeling in nonmoral domains.

Keywords: cultural evolution; gene-culture coevolution; power; prosocial norms; evolution of morality

[*] Published in *The Quarterly Review of Biology* 93, no. 4 (2018): 297–318.

1. Introduction

According to mainstream cultural evolutionary theory, individuals possess "*cultural variants*"—"*idea[s], skill[s], belief[s], attitude[s],* and *value[s]*" (Richerson and Boyd 2005: 63) mostly "stored in human brains" (p. 61). Through interacting with each other, people may adopt and transmit these variants. Cultural evolutionists seek to identify the dispositions that underlie the transmission process, and explain why those dispositions were favored by natural selection under ancestral conditions (Henrich and McElreath 2003; Lewens 2015: 17). Having (purportedly) identified how people acquire and transmit culture, mathematical modelers can, given the fitness values of cultural variants and their distribution in a population, predict the future evolution of a culture, or explain retroactively why some variants proliferated and others disappeared.

Fracchia and Lewontin (2005: 21–22) raise a fundamental objection to this research program. They note that the success or failure of a cultural variant can depend "crucially on the amount of power behind it" (Fracchia and Lewontin 2005: 22). Cultural evolutionary models assume that people more readily acquire beliefs and values with certain intrinsic properties, or preferentially learn from certain kinds of people in certain contexts. The models do not seem to account for the possibility that cultural variants can be imposed on people coercively, regardless of their "learning rules" or "transmission biases" (Richerson and Boyd 2005). Indeed, in a "list of cultural evolutionary forces discussed" in *Not by Genes Alone*, Richerson and Boyd (2005: Table 3.1) say nothing about power or coercion. All of the forces mentioned involve voluntary imitation (see also Henrich and McElreath 2003: Box 3). This potentially leaves out a big factor that operates in real life. Whether "punishment," which Richerson and Boyd do not include on their list of cultural evolutionary forces but which they do discuss in other parts of the

book and in other works, or "prestige bias" captures the phenomenon of power as it is exercised in real life will be discussed presently.

Some years before Fracchia and Lewontin (2005), Durham (1991) criticized Boyd and Richerson for what he described as their "radical individualism" that ignores "structured asymmetries or power relations, and coercion" (Durham 1991: 182). When Boyd and Richerson do mention power or coercion at all (e.g., Boyd and Richerson 1985: 229–230), they treat it as "individually delivered 'punishment of noncooperators'" (Durham 1991: 182).

Mainstream cultural evolutionists have not changed their approach in response to Durham or Fracchia and Lewontin. To illustrate, consider a classic empirical study in the Boyd and Richerson tradition in which Gürerk et al. sought to investigate "institutional selection" (Gürerk et al. 2006: 108). In their experiment, *anonymous* subjects played a 30-round public goods game, with each round consisting of three stages: an "institution choice," a "voluntary contribution," and a "sanctioning" stage. In the first stage, subjects "simultaneously and independently" chose to adopt either a "sanctioning institution" (SI) or a "sanction-free institution" (SFI; Gürerk et al. 2006: 108). In the second stage, subjects (anonymously) played a public goods game with all of the others who had chosen the same institution, and they were informed about how much each of the other players in *both* groups had contributed to the public good. (Thus the SI group members knew how well they were doing compared with SFI group members and vice versa.) In the third stage, subjects in the SI group were allowed to pay to punish or reward each other.

In the first couple rounds, most subjects chose the SFI. However, freeriding in the SFI group led contributions to fall precipitously. Punishment in the SI group quickly led contributions to rise to the near-maximum level. The vast majority of subjects who

initially chose the SFI ended up switching to the SI and adopting a prosocial strategy. Those who participated in the SI received a significantly higher payoff than those who stuck with the SFI. Henrich describes this study as an "experimental demonstration of cultural group selection in action." A key lesson, he says, is that "the players' degree of rationality did not permit them to foresee the final outcome and select the higher payoff institution on the first interaction" (Henrich 2006: 61).

Can we really draw conclusions about "cultural group selection" from Gürerk et al.'s (2006) experiment? Or were there "power relations" and "coercion" that were decisive in the real-life cultural evolution of cooperation and punishment but which were not captured in the experimental setup? It is true that subjects in the SI group were able to "punish" each other, and in that sense they wielded a kind of power. But there are some essential differences between power, punishment, and coercion in real life and that among Gürerk et al.'s (2006) subjects in the SI group. First, subjects were able to choose beforehand whether they wished to be subject to punishment or not (i.e., they chose whether to participate in the SI or the SFI). Henrich (2006) imagines that this is analogous to hunter-gatherers in the ancestral environment choosing which band to migrate to. However, the option to avoid being subjected to punishment by preemptively migrating was probably only an option for our hunter-gatherer ancestors in very limited circumstances. (Because migration opportunities are limited, hunter-gatherers often receive the threat of expulsion as a death threat. See the section below titled The Reverse Dominance Hierarchy and Deliberate Guidance of Cultural Evolution for a real-life illustration. When hunter-gatherers do have the chance to migrate, it is typically to culturally similar bands.) Second, power in the experiment was symmetrical not, as it typically is in real life, asymmetrical (see Singh et al. 2017: 461–462)—in the experiment anyone could punish anyone for any reason. (That is to say, in real life, those with less

power cannot punish individuals—or groups—with more power. And a noncooperative individual who violates group rules cannot go around punishing cooperative group members.) Third, and perhaps most important, the experimental setup totally eliminated the possibility of collective decision-making. Again, *anonymous* subjects "*simultaneously and independently*" chose whether to participate in the SI or the SFI. Henrich points out that most subjects failed to "foresee" that higher payoffs would ultimately be reaped by participants in the SI. He argues that this shows that successful institutions are not generally designed with rational foresight—people just cannot predict the consequences of different institutional arrangements. But if instead of simultaneously and independently choosing their "institution" subjects had been allowed to meet and coordinate a strategy, it is quite likely that most everyone would have been able to "foresee" that the best strategy would be to adopt the SI and to collectively punish freeloaders. The experimental setup *artificially* prevented this from happening because it was predicated on radical individualism that sees cultural evolution as driven by the decisions of uncoordinated, independently acting individuals.

The results of one of Ostrom et al.'s (1992: 412–413) experiments seem to support the assumption that, if people are allowed to communicate and make a collective decision, they *can* anticipate the benefits of punishment. Groups of eight subjects played a common pool resource game for 20 rounds. After the 10th round, subjects met face-to-face for 10 minutes and were allowed to decide (by a majority vote) whether to institute a "sanctioning mechanism" with a fee-to-fine ratio of 1/2 for their future interactions. Out of four groups in this experimental condition, two instituted a sanctioning mechanism (with fee-to-fine ratios of 10¢/20¢ and 20¢/40¢, respectively). However, Ostrom et al. (1992) note that the subjects in this experiment had previously participated in another experiment where they played a common pool resource game *with a*

sanctioning mechanism but no chance for communication. Out of the 14 subjects who voted against the sanctioning mechanism, 11 had previously participated in a design with a relatively high fee-to-fine ratio of 20¢/80¢. Out of the 18 who voted in favor of the mechanism, just *three* had been in the 20¢/80¢ design. They "infer from this result that the high level of sanctioning activity in the 20¢/80¢ design, the lack of overall efficiency gains and the presence of blind revenge combined to impede the willingness of participants to choose a sanctioning mechanism" (Ostrom et al. 1992: 413). They also suggest that "the experience of the first 10 rounds of the...game had an effect on mechanism choice" (Ostrom et al. 1992: 413). To test this possibility, they conducted two more experiments where subjects could communicate and adopt a sanctioning mechanism *before playing any rounds*. The result: "In both of these experiments, the subjects quickly agreed to an investment strategy and a sanctioning mechanism to punish defectors. Across the two experiments, net yields averaged 95%–94% with fees and fines included" (Ostrom et al. 1992: 413).

2. What Is Power?

In order to properly articulate Durham and Fracchia and Lewontin's (DFL) challenge, we should have a clear idea of what power is. Fracchia and Lewontin (2005; Lewontin 2005) refer simply to "power," which they illustrate with examples of rulers imposing cultural variants by decree. They do not explicitly describe the mechanisms that make such decrees effective. Durham usually refers just to "power," and at one point to "structured asymmetries or power relations, and coercion" (Durham 1991: 182)—phraseology that seems to imply that "power relations" and "coercion" are different

things. “Power” is clearly a multifarious concept. Which type(s) of power might pose a challenge to traditional cultural evolutionary modeling?

In his influential analysis, Dahl (1957) suggests that the “intuitive idea of power...is something like” the following: “*A* has power over *B* to the extent that he can get *B* to do something that *B* would not otherwise do” (Dahl 1957: 202–203; cf. Lukes 2005: 30). Whether or not this definition describes our intuitive idea of power, it is clearly too broad for the present purposes. On Dahl’s definition, virtually all cultural transmission would have to be conceived as a manifestation of power. Any time *B* copies *A*, *A* exercises power over *B*, even if *B* is guided by one of the learning biases represented in standard cultural evolutionary models.

Bachrach and Baratz (1970) offer a useful taxonomy of power (see discussion in Lukes 2005: 21–22), which may allow us to pick out those forms that present a challenge for cultural evolutionary modeling. Their taxonomy includes influence, authority, force, and manipulation. *A* influences *B* when *A*, “without resorting to either a tacit or an overt threat of severe deprivations, causes [*B*]...to change his course of action” (Bachrach and Baratz 1970: 30). *A* exercises authority when *B* “complies because he recognizes that the command is reasonable in terms of his own values” (Bachrach and Baratz 1970: 34). *A* uses force when *A* obtains compliance by depriving *B* of choice. *A* manipulates *B* when *B*’s “compliance is forthcoming in the absence of recognition on the complier’s part either of the source or the exact nature of the demand upon him” (Bachrach and Baratz 1970: 28).

According to Bachrach and Baratz’s (1970) jargon, “power” should refer only to cases where *B* complies with *A* because *B* fears that *A* “will deprive him of a value or values which he regards more highly than those which would have been achieved by noncompliance” (Bachrach and Baratz 1970: 24). “Force,” they say, should refer only to

cases of physical manipulation. However, following Lukes (2005)—and common usage—we can regard influence, authority, force, and manipulation as species of power. When *A* deprives *B* of the option of following a course of action, this is an exercise of force.

Influence, as Bachrach and Baratz define it, is a form of power that can, in general, be easily accommodated by cultural evolutionary models. When people freely copy each other, this is a way of being influenced. Standard cultural evolutionary models simply reflect the rules, or the *learning biases*, that determine how people freely copy each other.

Force is the form of power that poses the most obvious (apparent) challenge to the models. If cultural variants are imposed on people because the option to reject them is taken away by force, then it seems that standard cultural evolutionary models cannot explain what happened.

Authority is established when individuals freely recognize the right of a leader to make certain demands on them. If cultural variants spread on a large scale because of the exercise of authority, this would seem to undermine cultural evolutionary modeling, at least to some extent. If people accept the authority of a leader to tell them what cultural variants to adopt, to the extent that the leader exercises their authority, the learning biases of cultural evolutionary models will cease to play a decisive role in (further) cultural evolution. Learning biases might explain why people choose to accept the authority of a particular leader in the first place, but they would not determine which variants subsequently spread. When religious leaders promulgate opinions about values or doctrine, their followers may adopt those opinions because they accept the authority of the leader. To reiterate, cultural evolutionary models might help explain how authority itself comes to be *established* (e.g., how people become Catholic and thereby accept the Pope's authority). But, just as when cultural variants are spread through force, when they

are spread through the exercise of authority the learning biases of cultural evolutionary models play no role. When Durham (1991) distinguished between "power relations" and "coercion," perhaps he had in mind the distinction between "authority" and "force."

Authority and force can interact in interesting ways. For example, one reason that we accept authority might be that it is backed by force. (Would Prime Minister Theresa May retain her authority if the military and the police disbanded? Probably not for long.) Conversely, the power to employ force is often made possible by widespread acceptance of the legitimacy of the authority. (People tolerate Theresa May giving orders to the military and the police because they accept her right to be Prime Minister.) Sorting out how this works in detail is not important for the present discussion. The point is that when cultural variants are spread through the exercise of either force or authority, the learning biases featuring in cultural evolutionary models are *to that extent* irrelevant.

Manipulation—surreptitiously generating desires in people to make certain choices—generally works by taking advantage of learning biases. For example, advertisers use celebrity endorsements to take advantage of prestige bias. The Sara Lee Corporation paid former basketball star Michael Jordan a great deal of money to endorse Hanes underwear because many people are motivated to copy successful athletes. Richerson and Boyd (2005: 124) note that some men also started shaving their heads because this was Michael Jordan's practice. They cite imitating Jordan's choice of underwear and imitating his hairstyle as illustrations of "prestige bias." But these two cases of imitation illustrate different kinds of *power*. When men shave their heads because of Jordan, they are being *influenced*. When they buy Hanes because of him, they are (most likely) being *manipulated*.

In principle, cultural evolution driven by manipulation can be modeled in the same way as that driven by *influence*. Of course, if manipulation is a decisive factor, an

explanation of cultural evolution that leaves this fact out and appeals only to the learning biases of the people who were manipulated would be impoverished. But there is no reason why standard cultural evolutionary models cannot be used to track how cultural variants spread in such circumstances. *Prestige*, however, has a superficial resemblance to *authority*. The following section considers whether the latter can be reduced to the former.

The fact that which cultural variants people adopt may be determined by their position in networks of power—specifically networks of force and authority—cannot be captured, at least in any obvious way, by models that are predicated on a strong form of methodological individualism. Models that are committed to a highly individualistic picture of decision-making and action will also miss the fact that, when it comes to force and authority, the *locus of power* is very often a *group* of collectively acting individuals.

3. Can Power Be Reduced to "Coordinated Punishment" or "Prestige Bias"?

Coordinated punishment and prestige bias are two phenomena that already feature in mainstream cultural evolutionary models. If either one captures the phenomenon of power—namely, the exercise of *force* or *authority*—in a sufficiently realistic way, then cultural evolutionary modeling would not be threatened by DFL's challenge. This section considers coordinated punishment and prestige bias in turn.

Boyd et al. acknowledge that previous models of the evolution of punishment have been unrealistic in their assumption that "punishment is an unconditional and uncoordinated individual action automatically triggered by defection" (Boyd et al. 2010: 617). To rectify this shortcoming, they propose a model (also defended in Bowles and Gintis 2011: chapter 9) of what they call "coordinated punishment of defectors." This

sounds like it may be a step toward incorporating collective social phenomena (including collectively exercised *force*) into cultural evolutionary models. However, closer inspection suggests that the kind of coordination captured by Boyd et al.'s (2010) model is still unrealistically individualistic. In their model, "coordinated punishment" means that, when individuals decide whether to punish defectors, they take into account whether other individuals have signaled *their* intention to punish defectors. But, as shall be argued, the decision to punish—as it is represented in this model—is made without the sort of collective plan that supports cooperative behavior in real life, and that, according to anthropological evidence, played an essential role in the evolution of prosocial norms.

Boyd et al. say that, in their model, "punishment is coordinated among group members so that it is contingent on the number of others predisposed to participate in the punishment" (Boyd et al. 2010: 617). But there is much more to the coordinated punishment described by anthropologists, or even by Boyd et al. in their verbal description of anthropological observations. As Boyd et al. themselves say:

> ethnographic evidence indicates that punishment is coordinated by means of gossip and other communication among punishers, is contingent on the expected effectiveness of punishment in inducing cooperation, and it is not undertaken unless it is judged as legitimate by most group members. (Boyd et al. 2010: 617)

The collective agreement and action that emerges from this kind of cooperation cannot be reduced to punishing in response to signs that others will punish.

In Boyd et al.'s (2010) model, individuals are drawn from a population of "punishers" and "nonpunishers" to form groups. Members of each group engage in an initial three-stage interaction. In the first stage, punishers exhibit a "signal." In the second stage, if a threshold number of other individuals in the group gave the signal, punishers cooperate (with probability $1 - e$)—otherwise they defect. Nonpunishers always defect

at this stage. In the third stage, if the threshold number of punishers revealed themselves in the first stage, the punishers impose a cost on those who did not cooperate. Group members then engage in further interactions consisting only of cooperation/defection and punishment stages.

In administering a punishment, each punisher incurs a cost. The more punishers there are in a group, the less the *total* cost of inflicting punishment on a defector, since the defector's ability to retaliate at all can be assumed to diminish exponentially with increasing numbers of punishers.

Because punishment imposes a cost on defectors that is, on average, greater than the cost of cooperating, in general, cooperation maximizes expected payoff if defectors face punishment. In the model, defectors who are punished in the first round subsequently behave like punishers, cooperating with probability $1 - e$. They defect with probability e either because of error or because, due to random differences in circumstances, cooperation may be costlier than being punished.

In the model, after engaging in a (randomly varying) number of interactions, individuals produce offspring according to the payoff they received, and new groups are formed from a mix of the offspring. Offspring inherit the strategy of "punishing" or "nonpunishing" from their parent.

Boyd et al.'s basic conclusion is that "the initial proliferation of punishment occurs underplausible levels of group genetic differences and results in persistent and high levels of cooperation" (Boyd et al. 2010: 620). Punishers are able to proliferate in their model because—unlike in other models—they refrain from inflicting costly punishment unless there are enough other punishers in their group to reduce the cost, and they are able to do this because punishers "coordinate" by signaling their intentions. This paper will argue that prosocial-norm-enforcing punishment did not evolve in this way,

with individual contingent, signaling punishers gradually infiltrating groups of nonpunishers in the ancestral environment. Rather, individuals within groups of nonpunishers formed coalitions to collectively impose prosocial norms on all group members. And, through a process of gene-culture coevolution, we became genetically adapted to be receptive to such norms, including norms to punish noncooperators. The evidence will show that Boyd et al.'s (2010) model gives an incorrect historical account of how punishment arose in human evolution, and of how we evolved a disposition to punish norm violators.

The discussion above suggests that power cannot be reduced to coordinated punishment. Can it be reduced to "prestige bias"? Prestige bias refers to our (well-established) tendency to preferentially copy individuals who are successful, or to whom *others* are paying attention or deferring (see, e.g., Chudek et al. 2012). This is, of course, one of the central learning biases taken into account by cultural evolutionary models.

It is true that prestigious and prominent people have (ceteris paribus) more influence over culture than the less prestigious, and this influence is a form of "power." When celebrity Kylie Jenner tweeted that she did not use the app Snapchat anymore, the company's stock lost 1.3 billion dollars in value. That is because when Jenner says she uses or does not use something, many people will be inclined to imitate her.

But the "power" of Kylie Jenner is fundamentally different from the power of a king, a hunter-gatherer coalition enforcing rules, or even the Pope. Jenner's influence derives from the fact that many people preferentially imitate her. When she exhibits behavior, or expresses an opinion, we can explain why it spreads by appealing to the learning biases (namely, the prestige bias) of her audience. The kind of power that threatens cultural evolutionary modeling is that which is backed by *force* or *authority*, and which thereby neutralizes all of the learning biases featuring in mathematical models.

Neither force nor authority can be reduced to prestige bias, since force and authority neutralize all learning biases. When, in 1492, Isabella I and Ferdinand II issued the Alhambra Decree ordering the Jews of Spain to convert to Catholicism, leave, or be executed (as discussed in more detail below), it is theoretically possible that a small number of Jews were impressed by the religious passion of the prestigious queen and king and converted for that reason. But most of the Jews who converted probably did so because their freedom to choose had been coercively restricted. Living in Spain as a practicing Jew was literally no longer a *possible choice*. In hunter-gatherer groups, individuals did not have the *choice* to refrain from sharing meat or contributing to group defense—if they did not do these things, and were recalcitrant enough in the face of reproach, they would be subject to execution or (what was often effectively the same thing) expulsion.

4. Approaches to Accommodating Power

Lewontin (2005) illustrates the role of power in cultural evolution with an example. Bavarians are mostly Catholic, Westphalians Protestant. This state of affairs was not brought about (according to Lewontin) by any mechanism recognized by cultural evolutionists—not by biased transmission, the relative appeal of the content of the two religions, drift, Darwinian selection, or any other (see the list in Richerson and Boyd 2005: Table 3.1). Rather, Lewontin says, in 1555 the German princes and the Holy Roman Emperor adopted "the rule of *cuius regio, eius religio*, which allowed rulers to enforce their own religion in their own dominions and to expel those who were recalcitrant" (Lewontin 2005). If cultural change is usually driven in this way—by sweeping diktats issued by powerful individuals or small groups—then Boyd and

Richerson-type modeling, which tracks "the aggregated effects of small-scale events" (Lewens 2015: 17), will have a very limited range of application. Most of history—and the Paleolithic culture to which our minds are adapted—would better be explained by a study of how individuals take power and impose their favored cultural variants on the masses, and cultural evolutionary models would not shed much light on the selective forces that shaped our social learning dispositions.

As Lewens notes, it would surely be a huge overestimation of the power of the typical ruler to suppose that they can change the religious beliefs of a population by mere decree. In regard to the specific example of *cuius regio, eius religio,* it is not clear that the German princes even tried to change people's religious beliefs on a large scale. They tended to "enforce" Catholicism or Lutheranism on their subjects according to what was already the prevailing religion, and even subjects in the minority did not necessarily convert in response to the official order (Lewens 2015: 133).

But that is just one example. It seems clear that sometimes cultural variants, even religious beliefs, can spread as a result of being backed by power. As discussed, the forms of power that seem to undermine cultural evolutionary modeling are *force* and *authority*. Sometimes power makes all of the difference in which cultural variants prevail in a group. The Alhambra Decree, which expelled practicing Jews from Spain in 1492, caused more than two-thirds of the not already-converted Jewish population to become Catholic, at least nominally. The decree, initiated by one person (Isabella I) and signed by two (the aforementioned and Ferdinand II) radically changed the distribution (and attractiveness) of certain cultural variants in Spain, and its consequences reverberated for centuries. In the 4th century, Christianity surely got a boost from being supported by Constantine, although whether this was responsible for its ultimate triumph is difficult to say.

Lewens (2015: 138–139) proposes three ways in which cultural evolutionary research programs could accommodate the phenomenon of power. First, restriction: we could simply restrict the application of models to times and places where cultural evolution is not driven by Lewontin scenarios—"where human societies have been free of interest groups, governments, unions, and perhaps gross asymmetries of individual power" (Lewens 2015: 138). Second, presumption: we could take systems of power as "fixed background conditions against which individuals interact" (Lewens 2015: 138), using cultural evolutionary theory to track cultural change resulting from those interactions. And, finally, reduction: we could try to reduce systems of power—"governmental structures, judicial procedures, and so forth" (Lewens 2015: 139)—to small-scale interactions dealt with by standard cultural evolutionary theory.

By restricting cultural evolutionary modeling to times and places "free of interest groups, governments, unions, and...gross asymmetries of individual power" (Lewens 2015: 138), *restriction* undermines any modeling applied to modern societies. Lewens suggests that we could still use cultural evolutionary models to investigate early hominin history, since "hunter-gatherers were largely egalitarian with respect to individuals' access to economic resources" (Lewens 2015: 138, citing Boehm 1999 and Knauft 1991). But even this is not certain. Despite their lack of formal leaders, great differences in individual power among adult males, or great differences in individual access to economic resources, hunter-gatherers do have systems of rules, coercion, and punishment. In fact, as shall be argued below, the egalitarianism of hunter-gatherers is itself maintained by the exercise of power wielded *by* the rank and file over their potential dominators (or alphas), creating what Boehm (1993, 1997, 1999) calls a "reverse dominance hierarchy." This means that even hunter-gatherers possess a power apparatus that could theoretically be used to impose cultural variants on individuals willy-nilly and

force deviants into line. Whether hunter-gatherers actually do use this power to influence the spread of cultural variants in a way that undermines Boyd and Richerson-type modeling is the central question addressed in the present paper.

The problem with *presumption*—taking power relations as "fixed background conditions against which individuals interact" (Lewens 2015: 138)—is that it seems to ignore DFL's challenge. Suppose we want to explain the spread of different sects of Christianity in mid-16th-century Germany. An explanation emphasizing the forces listed in Richerson and Boyd (2005: Table 3.1) and relegating power relations to "fixed background conditions" would clearly be impoverished and misleading, in that it would be taking for granted the factor that ought to be a primary target of a scientific explanation (even if, as Lewens argues, it was not the whole story). How could we construct an explanation of the spread of Catholicism and Lutheranism taking the power of the princes as "fixed background conditions"? Perhaps we could treat the positive and negative incentives offered for adopting different beliefs in different places as being features of the environment like the weather. The fact that people prefer, ceteris paribus, denominations that do not cause them to be exiled from their home could be considered a "content bias." Although it may be possible to model the decision-making process of German peasants along these lines, this approach would ignore the fact that a major driving force of cultural evolution is the asymmetrical power relations of the princes and the peasants. Fracchia and Lewontin's (2005) point is that explaining historical trends—trends about how cultural variants spread—requires us to recognize power as a major factor. In Durham's words, "to ignore [power]...is to ignore what may often be the leading cause of transformation" (Durham 1991: 182). The fact that we can treat the influence of power as a given and focus on individual decision-making does nothing to counter the claim that power often plays a decisive role in the fate of cultural variants.

Lewens (2015: 138) raises a similar problem for presumption, saying that as a strategy it "simply omits to explain many features that are admitted as important in determining cultural change," although he says presumption can be useful for "understanding...individual interactions and their long-term effects."

This leaves *reduction*—reducing the influence of power to individual-level interactions that can be handled by standard cultural evolutionary models. In Lewens's view, the justification for this approach is that power relations are all grounded in individual interactions, which should in principle be amenable to populational modeling. This does seem to be the approach advocated by cultural evolutionary modelers themselves (e.g., Boyd and Richerson 2002; Richerson and Henrich 2012), although in practice their models do not generally reflect power relations at all.

But redescribing the exercise of power in terms of individual interactions does not negate the fact that power is a factor, and it does not neutralize DFL's challenge. Consider again the Alhambra Decree. Isabella and Ferdinand issued an order (Jews had to convert, leave, or be executed). Since those with the ability to enforce the order—the soldiers—accepted the authority of the monarchy, they conveyed the threat to the Jews, who chose whether to convert or leave (no one volunteered for execution). If we want to explain why Spanish soldiers accepted the monarchy, we would probably need to refer to all of the learning biases that figure in cultural evolutionary models. If we want to explain how Jews responded to the threat, we could say that the cost of being overtly committed to Judaism had increased, and since people have a content bias to acquire cultural variants with lower costs, Judaism became less attractive. So rather than saying that the Jews converted in response to the exercise of power by Isabella and Ferdinand, we can say that they decided which practices to adopt in light of the incentives. But we still cannot explain what happened without referring to the fact that, because the social

structure gave power to the monarchs, two individuals drove cultural evolution in a specific direction. To explain why the powerful individuals acted as they did, we would need to engage in traditional historical, sociological, or psychological analysis.

5. Can Cultural Evolutionary Modeling Be Restricted to the Ancestral Environment?

Modern societies are founded on extensive, complex systems of power. The information we are able to obtain, the opinions we are exposed to, and the options that are presented to us are constrained in all sorts of ways by the (often hidden) exercise of force and authority. DFL's challenge to cultural evolutionary modeling appears to have great force when it comes to understanding how cultural variants spread in extremely hierarchical societies where powerful individuals and groups exercise such far-reaching control. This is not to deny that the models can be used to explain some aspects of cultural evolution even in modern societies. When people's choices in a particular domain are largely unconstrained by power, the models can sometimes explain how and why culture develops as it does. For example, cultural evolutionary models can explain the demographic transition in the West (Richerson and Boyd 2005: chapter 5), since people in Western countries in recent history have been largely free to choose how many children to have. Power is rarely the *only* factor influencing what people believe and how they act, so models might be used in conjunction with traditional historical/anthropological/sociological methods. But to explain phenomena such as the spread of communism and the concomitant values in the 20th century, a study of the influence of *power* will often be more fruitful than a study of the aggregate effect of the decisions of many individuals.

Cataloging all of the ways in which power can or does influence the spread of cultural variants, and in which it might undermine cultural evolutionary modeling, is not the aim of the present paper. This paper has a narrower focus, which is to consider whether or to what extent power undermines cultural evolutionary modeling applied to nomadic, foraging societies. It seeks to test the possibility raised by Lewens (2015) that we can *restrict* the application of Boyd and Richerson-style models to conditions that prevailed during the long period before the advent of sedentarism, food storage, and agriculture, when human society was reputedly egalitarian.

There are two reasons why it is worthwhile to investigate whether Boyd and Richerson-style models can be applied unproblematically to nomadic foragers. First, the models have been used to give accounts of the historical origins of specific cultural variants, such as certain ethical norms. If the models cannot be applied to nomadic foraging societies (where the variants in question originated), then we will have to reconsider these accounts. Second, the models—predicated on the idea that cultural variants spread in the ancestral environment due largely to content and context learning biases—have been used to make predictions about what psychological adaptations we might have. Given that variants spread due to content and context learning biases, cultural evolutionary theorists ask what psychological dispositions we might have evolved in order to extract adaptive information from the cultural environment. But if *power* was a major factor influencing how variants spread in the ancestral environment, then we might make different predictions about how our psychology evolved. We might predict, for example, that we evolved to have different attitudes toward variants imposed on us by power and those that we adopted due to learning biases.

Chimpanzees (*Pan troglodytes*) have a strong dominance hierarchy based on physical force. The alpha individual rules because he (sometimes with the help of his

coalition partner) can beat up every other chimp. Beta can beat up gamma, and so on down the line. Under natural conditions, all adult males, being physically stronger, are dominant to all females (Goodall 1986).

Dominant chimps use their social position primarily to secure access to food and mating opportunities. Dominance relations have little direct influence on cultural transmission. Chimps have different, culturally transmitted ways of fishing for termites, building nests, and performing a few other tasks. Which cultural variant an individual acquires depends on the practice of the most accessible models and features of the local environment (McGrew 2004: chapter 7). Under experimental conditions, chimps preferentially attend to dominant individuals when learning how to solve foraging problems, and are more likely to copy their methods (Horner et al. 2010). But dominant individuals have no interest in—and may even lack awareness of—the cultural practices of their subordinates, hence they make no attempt to forcibly spread their favored practices. The common ancestor of chimps and humans probably had a similar hierarchical social structure (Boehm 1999).

At some point, hominins developed much greater awareness of cultural possibilities, and a much greater capacity for deliberately transmitting learned practices. It became feasible for dominant individuals to force subordinates to adopt their preferred cultural variants. Richerson and Henrich note that "[p]redatory elites and other self-interested subgroups with some form of coercive power" could establish norms and institutions that "disproportionately benefit them. Ideologically motivated groups with coercive power may sustain equilibria at mad extremes, at least for brief periods of time" (Richerson and Henrich 2012: 50). (They do not comment on the implications this might have for cultural evolutionary modeling.) But elites who impose cultural variants on others do not necessarily have to be motivated by narrow self-interest—they could be

prompted by altruism, concern for the commonweal, or the belief that a well-functioning group will ultimately benefit themselves (i.e., enlightened self-interest). In any case, if dominant individuals or coalitions in the Paleolithic did exercise power to impose cultural variants on others, it could create a problem for cultural evolutionary modeling for the reasons discussed above.

There are two crucial empirical questions to answer. First, was there any locus of power in Paleolithic communities *capable* of imposing cultural variants on subordinates? Second, if the answer to the first question is yes, was this power routinely exercised in a way that supports DFL's critique of cultural evolutionary modeling? These questions are dealt with in turn in the following two sections. Based on work by Boehm, the first will be answered in the affirmative. The second will also be answered in the affirmative, but with a qualification. It will be argued that, in the ancestral environment, power was used primarily as a means of enforcing moral norms. Therefore, DFL's critique undermines standard cultural evolutionary modeling applied to moral norms that arose in the past 250,000 years, but it does not necessarily undermine modeling applied to nonmoral norms in the ancestral environment. It also does not necessarily undermine modeling applied to the evolution of moral norms and intuitions that might have occurred before 250,000 bp. (The specific example of Baumard et al.'s 2013 model will be discussed.)

6. The Existence of Power in the Paleolithic

Knauft notes that "male status differentiation in human evolution is U-shaped" (Knauft 1991: 397). The social organization of the chimp-human common ancestor is presumed to have been as hierarchical as that of modern chimps. After hunter-gatherers settled down and developed the means to stockpile food, human societies became even more

hierarchical than those of chimps. For some millions of years after the chimp-human lineages split and before sedentary, hierarchical societies were established, humans lived as nomadic hunter-gatherers. All nomadic hunter-gatherers that have ever been described, without exception, live by an ideology of extreme egalitarianism (among men). No nomadic hunter-gatherer society tolerates adult males issuing direct orders to other adult males. Some groups have a titular leader, but his role never goes beyond leading by example and helping to organize group discussions aimed at reaching a consensus on group actions (whether to go to war, where to migrate, and the like; Boehm 1999).

As noted earlier, Boehm (1993, 1997, 1999) argues that nomadic foraging societies are characterized by a "reverse dominance hierarchy": men who would otherwise be subordinate to an alpha band together to keep would-be alphas from assuming positions of dominance. In contrast, Erdal and Whiten (1994: 177) argue that there is an *absence*, not a *reversal*, of hierarchy. According to them, would-be subordinates simply refuse to obey would-be leaders. A hunter-gatherer group does not (in their view) "start with a hierarchy and [then] reverse it" (Erdal and Whitten 1994: 177). Rather, "counterdominant" behavior prevents leaders from arising in the first place.

The debate between Boehm and Erdal and Whiten is not just scholastic. If Erdal and Whiten are right that forager egalitarianism is based on a widespread refusal to obey orders, then there would have been essentially no locus of power among Paleolithic adult male foragers that could have potentially directed cultural evolution. If no one ever obeys authority, no individual or coalition can exercise power over anyone else. But if Boehm is right that egalitarianism is maintained by the asserted efforts of the rank and file to keep down ambitious, aggressive, and accomplished individuals, Paleolithic bands

would have had groups—coalitions of the majority—with the power to potentially impose cultural variants on those whom they dominated.

There are compelling reasons to side with Boehm in this debate. Perhaps most significant is the fact that nomadic hunter-gatherers deliberately and actively *enforce* egalitarianism, employing a variety of *coercive* methods ranging from gossip to ridicule to ostracism to expulsion (often a de facto death sentence) to outright execution. They do not simply refuse to obey commands and rebuff would-be dominators. They espouse an explicit egalitarian ideology (Cashdan 1980) and take active measures to head off possible power grabs by individuals. As Boehm (1999) argues, our primate heritage makes us readily susceptible to developing orthodox hierarchies (i.e., with alpha-types dominating everyone else). To prevent orthodox hierarchies from emerging requires the group to continually keep down those who would take control if given the opportunity.

What is more, the group actively controls the behavior of potential alphas, compelling them to procure meat for others and allowing them to act as *informal* leaders in hunting and warfare (so that the group benefits from their expertise without submitting to their domination; Boehm 1997: S104). Again, this goes beyond refusal to obey orders and reflects bona fide domination and control of potential alphas.

Everyone, not only potentially dominant individuals, is subject to group rules. People who are not suspected of political ambitions but who horde food or engage in deception can be targeted by some of the same sanctions as those who try to gain *individual* power. Boehm (1999) quotes Service's (1975: 48–49) observation that in foraging societies social life is intensely regulated by "codes, rules, expectations, habits, and customs that are related to etiquette, ethic, and role. And because these are not [normally] explicit, nor revealed by frequent breaches, the society might give the impression of freedom and lack of conflict" (Boehm 1999: 84). But the threat of

collectively imposed sanctions of varying degrees of severity hangs over everyone's heads.

The foregoing suggests that, in the sort of hunter-gatherer bands that existed since at least the early Middle Paleolithic and to which our social psychology is presumed to be adapted, there did exist a locus of power that was *theoretically* capable of driving cultural evolution in the way DFL suggested (i.e., by fiat issued by a consensus of groups). Whether this threatens cultural evolutionary modeling turns on the answer to an empirical question: Did hunter-gatherer coalitions of the majority, as a matter of *fact*, exercise the power described above to promote their favored cultural variants, or did they use it only to suppress power grabs by would-be alphas? It will be argued that the empirical evidence supplied by Boehm (1999, 2012) strongly suggests that hunter-gatherers *did* use their collective power to impose cultural variants that they thought would promote group success and well-being—variants broadly related to morality. (Note that a great variety of behavior can be moralized, from food preferences to factual beliefs.) They did *not* use power to impose variants related to practices that were not thought to directly concern the group, like individuals' specific spear-making techniques. Consequently, cultural evolutionary models can be applied more or less unproblematically to the *nonmoral* domain of culture. It is cultural evolutionary modeling of the transmission of group-concerning moral practices such as reciprocity and punishment (e.g., Henrich and Boyd 2001) that may be problematic.

7. The Reverse Dominance Hierarchy and Deliberate Guidance of Cultural Evolution

This section briefly outlines Boehm's (1999, 2012) account of how subordinate humans

circa 250,000 bp overthrew their alphas and, around the same time, used their newly acquired collective power to impose a deliberately engineered moral code. In the light of Boehm's theory, the following section will critically analyze some well-known modeling work on the coevolution of prosocial norms and norm psychology.

From our primate ancestors we inherited three key dispositions: an enjoyment of dominating, a capacity for submission, and a dislike of being dominated. Among chimpanzees, these same dispositions lead to the development of a dominance hierarchy. Subordinates—particularly males—wait for the opportunity to challenge higher-ups and ascend the hierarchy, if possible. Subordinate chimpanzees (*P. troglodytes*) often form coalitions to challenge the power of alphas (Boehm 2012: 95–96) but, because of limits on their coordination abilities and the physical supremacy of the dominant individuals, they never succeed in doing away with the practice of alpha rule itself.

Chimps, bonobos, and gorillas have the ability to make generalizations about what sort of positive or negative behavior will elicit the wrath of their superior, and purposefully comply with the superior's demands (Boehm 2012: 106). (The behavior demanded by dominant apes mostly concerns feeding priority and mating; Boehm 2012: 107–108.) This was an important preadaptation for our ancestors to develop the ability to appreciate group-imposed *rules*. Taking advantage of our ability to understand and adopt regular patterns of behavior, coalitions of the majority in human groups instituted rules to enforce egalitarianism and promote the good of the collective. For hunter-gatherers, the coalition of the majority is the object of fear, rather than, as for chimps or gorillas, powerful dominant individuals. Boehm expounds: "People like Bushmen or Pygmies gossip incessantly and are highly judgmental, and group opinion is something to be feared because moral outrage can lead to ostracism, expulsion from the group, or even execution. This is true of all hunter-gatherers" (Boehm 2012: 107).

Darwin speculated that, when language developed, "the wishes of the members of the same community could be distinctly expressed, the common opinion how each member ought to act for the public good, would naturally become to a large extent the guide to action" (Darwin 1871: 72). The "power" of public opinion derived, he said, not from coercion but from our natural love of approbation and "horror of scorn and infamy," which itself is rooted in our "[i]nstinctive sympathy" inherited from our distant ancestors (Darwin 1871: 86; see Lewens 2007: 162–167 for an overview of Darwin's account; Darwin 1871 does not explain precisely how "[i]instinctual sympathy" gives rise to a desire for moral approval). Although Darwin was right about the new possibilities that language created for social organization, the ethnographic evidence discussed below suggests that hunter-gatherers employ coercion to enforce behavior approved by public opinion. There is good reason to think that coercive methods were employed by ancient humans as well. We do value the moral approval our fellows, and this is also an important source of motivation to conform to group rules (Mameli 2013: 911). But our desire for moral approval is most likely an evolved response to the widespread social practice of collectively punishing norm violators, and is not simply an outgrowth of "sympathy."

Language allows members of coalitions in a hunter-gatherer band to collectively agree on what Boehm terms a "blueprint" upon which to model their society. Echoing Darwin, he says that because "humans are able to communicate in great detail,...groups can develop precise notions about the kind of society in which they wish to live" (Boehm 1999: 193). Whether, as a matter of historical fact, humans began collectively suppressing alphas *and then* started enforcing other moral behavior or vice versa (see Boehm 1999: 194), ultimately hunter-gatherers came to implement blueprints of egalitarian, moral societies. Boehm (2012: Table 1) reports the number of cases recorded by ethnologists where capital punishment was meted out for various offenses in 50

mobile foraging bands. There are records of capital punishment in 24 of the 50 bands, for a total of 45 instances. Boehm notes that this is likely a significant underestimate of the prevalence of capital punishment, because foragers have learned to conceal such practices from outside authorities. (The count here also does not include expulsion, which, as noted, is often a de facto death sentence.) The most common crimes eliciting a lethal response were various forms of intimidating the group with aggression, violence, or "malicious sorcery." But people were also executed for such moral violations as theft, failing to share meat, incest, adultery, and failing to respect taboos. To reiterate, execution is just the most extreme form of punishment. The vast majority of norm violators in hunter-gatherer bands reform in response to less serious sanctions such as ridicule. People use the collective power of the group to make society conform to a desired blueprint.

Humans were hunting with wooden weapons at least 400,000 bp, and were taking down big game as a regular source of food 250,000 bp (Boehm 2012: 146). Boehm speculates that the key impetus for reversing the dominance hierarchy came when meat from large animals became a crucial part of our diet. Hunting bands can only be effective if all hunters are reasonably well nourished and motivated, and this requires that the meat be shared rather than hoarded by an alpha and his coalition partners (Boehm 2012: 151–152) or taken by "greedy thieves and cheaters" (Boehm 2012: 155).

When coalitions of the majority started to collectively enforce group rules, this created a strong selection pressure for the development of a *conscience*, and concomitant emotions such as shame and moral pride. To have a conscience is to "*personally [identify] with community values*, which means internalizing your group's rules" so that you are emotionally invested in following them (Boehm 2012: 113). The human capacity for language allows for an extreme degree of effective surveillance. No matter who

witnesses a rule violation, word can spread throughout the whole group. People who "failed to control their predatory tendencies" (Boehm 2012: 67)—who bullied or cheated or otherwise violated group norms—would have been at a great disadvantage. Those who were able to internalize norms as goals in themselves would have avoided trouble much better than those who, like chimps, observed the rules only when they thought an authority was watching (Mameli 2013). Given the prospect of group punishment, developing a good reputation became an important way to increase individual fitness.

Consider an illustrative example of group enforcement of norms, reported by Turnbull (1961: 94–108) and cited by Boehm (2012: 37–43). The Mbuti have a hunting practice where the men set up long nets in the shape of a semicircle while the women and children make noise to drive animals (small- and medium-sized game) toward them. According to their rules, a man can kill any animal that falls into his own net and take the meat for his own family. Although there may be chance variation in how well a man fares on a given hunting expedition relative to other men, on average this is a fair way of distributing the spoils.

In the course of one such hunt, a man named Cephu felt he was getting shortchanged, so he surreptitiously moved his net ahead of the others. The strategy worked as far as increasing his catch but, unfortunately for him, he was spotted by another hunter. Word quickly spread of his misdeed. Most of the other families arrived back at the camp before him, and a man named Kenge announced: "Cephu is an impotent old fool. No, he isn't, he is an impotent old animal—we have treated him like a man for long enough, now we should treat him like an animal. Animal!" (Boehm 2012: 38–39). This triggered an outpouring of gossip and condemnation of Cephu as, in Boehm's words, "a group consensus materialized" (Boehm 2012: 39).

Cephu soon returned, and Kenge shouted at him that he was an animal. Everyone else was silent and for a short while just ignored him. Then, after the whole crowd confronted Cephu with vague accusations that he was selfish, someone finally made a direct accusation against him that he had stolen meat. One man said that "he hoped Cephu would fall on his spear and kill himself like the animal he was. Who but an animal would steal meat from others? There were cries of rage from everyone, and Cephu burst into tears" (Turnbull 1961; quoted in Boehm 2012: 39– 40).

At first Cephu argued that he had made an honest mistake. When no one accepted this excuse, he claimed that he deserved to place his net in a better position, considering that he was "an important man, a chief, in fact, of his own band" (referring to his extended family; Boehm 2012: 40). Someone responded that the Mbuti do not have chiefs. If Cephu was a chief of his own band, "let him go with it and hunt elsewhere and be a chief elsewhere" (Boehm 2012: 40)—a suggestion that, were it followed, would have meant starvation in the forest for Cephu and his family. Faced with the threat of expulsion, Cephu began to apologize profusely. He agreed to turn over all of his meat from the day's catch to his accusers, and the other band members then took everything from him and his wife. Despite all the drama, a few hours later he participated with everyone else in the evening singing and all was set right again—and presumably he never repeated his misdeed.

Notice how this real-life example of enforcing a cooperative norm is fundamentally different from what was possible in the experimental setup of Gürerk et al. (2006; discussed in the introduction to this paper). All of the members of the group explicitly coordinated in advance, agreeing on what they *foresaw* would be a set of practices that would lead the group to collectively acquire as much meat as possible and to distribute it fairly. Members of the society did not have a realistic option to opt out of

the system—they could not choose to participate in a "sanction free institution." The decision to adopt a cooperative institution and punish defectors/deviants was made collectively. In Cephu's case, we see that punishment was preceded by collective coordination. Kenge led the group by asserting that they should change the way they treat him: for his misdeed they should treat him as an animal. Before Cephu arrived, all of the members of the group (besides his immediate family members) had shared their sense of outrage with each other, and established that they were on the same page. When everyone confronted Cephu, they were not doing so as individuals expressing their personal anger, but as members of a group that had already come to a collective understanding.

The collective rebuke and punishment of Cephu illustrates how our moral sense plays an important role in undermining the radical individualistic approach of cultural evolutionary models. Chimp society is despotic and without morality. In humans, collectively imposed and accepted moral rules allow coordinated group action among all those who accept the rules. Hunter-gatherers use morality to enforce egalitarianism, whereas sedentary, resource-hording people use morality to enforce despotism. Whether morality is used to promote egalitarianism or despotism, it binds people into a collective decision-making body that cannot be legitimately atomized, as in cultural evolutionary models.

8. Cultural Evolutionary Models of the Development of Prosocial Norms

In cultural evolutionary models, cultural variants are distributed in a population, and individuals with content and context learning biases choose whom to imitate. Modeling work suggests that, in populations composed of social learners interacting with each other, stable behavioral patterns will emerge as a "by-product" of our learning biases

(Chudek et al. 2013: 442). These behavioral equilibria are more likely to be group *detrimental* than beneficial (Boyd and Richerson 1992; Henrich and Boyd 2001: 86; Chudek and Henrich 2011: 222). But, according to other models, cultural group selection favors the spread of those equilibria that happen to be beneficial (e.g., Boyd and Richerson 2002). Then, in a process of gene-culture coevolution, we become genetically disposed to be receptive to prosocial behavioral patterns, or "norms" (Henrich and Boyd 2001; Chudek and Henrich 2011; Chudek et al. 2013).

The aforementioned models assume that there is no organized enforcement of norms—such as the collective enforcement described in the previous section. Boyd and Richerson say that "[f]or most of human history, states were weak or non-existent, and norms"—such as "rules against murder"—"were not enforced by external sanctions" (Boyd and Richerson 2002: 287). Each individual makes a decision whether or not to imitate the behavioral patterns represented in the group. Among these behavioral patterns are tendencies to *punish* those who fail to exhibit certain other behaviors. *Punishing norm violators* is just another cultural variant that can spread among individuals. A norm (whether it is beneficial or detrimental to the group) can be stabilized (in part) by punishment. Punishment is presumed to involve a cost, but it can be reinforced by the punishment of nonpunishers. Punishment of nonpunishers does not have to go on forever (punishing those who fail to punish nonpunishers) because a combination of conformism and fear of punishment will drive cooperation at the first level to near fixation, which in turn lowers the cost of first-order punishment (Boyd and Richerson 1992).

As noted, Boyd and Richerson's models are based on the assumption that during the period of gene-culture coevolution in which norms first evolved and we became adapted to live in a norm-governed environment, "norms were not enforced by external sanctions" (Boyd and Richerson 2002: 287). The ethnographic evidence discussed in

connection with Boehm's theory suggests that this is false. Although there were no "states" or governments in the modern sense, there was a decision-making body—a coalition of the majority—that exercised coercive power, which amounts to what is effectively the same thing. The cultural variants to refrain from murder, to share meat, and so on were collectively adopted and formally *enforced*, not just punished by isolated individuals who possessed a variant to punish those who did not exhibit the relevant behaviors (or to punish in response to signals that others will join in their punishment, as in Boyd et al.'s 2010 model, discussed earlier).

Chudek et al. (2013) list three possible ways that different behavioral equilibria could be selected. The first is that "rational, forward-looking individuals" perceive the ultimate benefits of being in a cooperative equilibrium, "assume others are similarly sensible, and choose the prosocial state" (Chudek et al. 2013: 439). They give three reasons why they think this was not a significant factor in real life. First, Chudek et al. say, people are not actually good at making the calculations required to determine what practices would be beneficial. Second, "group decisions are often heavily influenced by leaders and coalitions whose interests diverge from the overall group" (Chudek et al. 2013: 439). Third, we see many examples of patently *non*prosocial institutions in societies throughout the world. Chudek et al. seem to recognize the possibility of "group decisions" (in the second reason), but they do not explore the implications of this phenomenon for cultural evolutionary theory. Insofar as group decisions do play a role in cultural evolution, these authors think that they tend to reflect the narrow self-interest of powerful political factions. They do not consider the possibility that group decisions can be made by a coalition of the majority that is more or less interested in the success of the group as a whole—but if Boehm was right then this is in fact a big part of how

groups in the ancestral environment selected among different possible behavioral equilibria to collectively enforce (cf. Singh et al. 2017: 470–471).

The second possible way Chudek et al. (2013) mention to select among behavioral equilibria is "stochasticity." By chance, groups move from one equilibrium to another, and groups are likely to spend more time in equilibria with larger basins of attraction.

The third way is cultural group selection, which acts on the variation provided (in part) by stochasticity. Cultural evolutionists identify cultural group selection as by far the most important factor in the spread of group-beneficial norms. Groups randomly adopted more or less group-beneficial or detrimental norms, and those with beneficial norms survived, enjoyed more immigration, and their individuals were preferentially copied by individuals in other groups.

Boyd and Richerson (2002) provide a model showing that, in a population consisting of small groups with behavioral norms at different equilibria, *group-beneficial* norms can spread if people tend to copy more successful individuals in either their own or in neighboring groups. Individuals in groups with more group-beneficial norms will tend to be more successful and, therefore, their norms will be copied by the members of other groups. But this model fails to account for the empirical fact that moral norms are formally enforced (by collective action) within foraging bands. Individuals from one band cannot copy the moral norms of individuals from another. (Imagine if Cephu responded to the accusations against him by saying that he subscribed to the different meat-sharing practices of a neighboring people.) However, a group can *collectively* decide to copy another, more successful group, and begin enforcing a new suite of norms.

Henrich describes "prestige-biased group transmission" (Henrich 2016: 168) as a matter of "individuals" in one group copying "individuals" in more successful groups.

But he also describes some striking cases where the *leadership* of one group decides to adopt the practices of another group. In the case of the Irakia Awa of New Guinea, "*senior men*" (Henrich 2016: 173; emphasis added) decided to copy the pig-rearing practices of their more economically successful neighbors, the Fore. The process of transmission did not involve individuals copying individuals. In Henrich's words: "[T]his transmission between groups occurred rapidly because the Irakia already had a political institution in the village, which involved a council of the senior members of each clan, who were empowered by tradition (social norms) to make community-level decisions" (Henrich 2016: 174).

Power played a complicated role in the adoption of new communal practices among the Irakians. A few years after they copied the Fore, some young men did not want to continue raising pigs anymore. When they communicated their preference to the village elders, the elders "would not even discuss it [and]...disparaged the idea and criticized the younger people for being lazy and unwilling to lead proper lives....The young men...admitted that it would be impossible to make such a change with the elders firmly against it" (D. J. Boyd 2001: 270). However, another group of young adults converted to Seventh-Day Adventism—a religion that prohibits the consumption of pork. They outright refused to raise pigs, and no serious sanctions were imposed on them. Apparently the power (that probably took the form of *authority*) of the village elders was substantial but not absolute.

Although the Irakians were not foragers, this story illustrates how the assumptions of cultural evolutionary models (that transmission is from individual to individual and that variants are not imposed on groups of people all at once by means of power) do not apply to many real-life scenarios that have been studied by cultural

evolutionists. Henrich's *verbal* description of how the Irakians adopted the Fore's practices refers to forces that cultural evolutionary *models* do not accommodate.

9. Gene-Culture Coevolution: The Development of "Norm Psychology"

As discussed, cultural evolutionary theorists argue that the learning biases that lead us to acquire adaptive cultural information tend to lead, when many individuals interact with each other, to the development of norms—"stable group-wide patterns of behavior" (Chudek et al. 2013: 442). Cultural evolutionary models of the development of transmitted behavioral patterns show that a variety of norms can be stable (Chudek et al. 2013: 438–439). Some equilibria are group beneficial, such as those involving widespread cooperation and punishment of defection. Most equilibria are group detrimental, such as those involving widespread noncooperation, or widespread, enforced performance of costly rituals. (Chudek and Henrich 2011: 222 give several examples, including disease-spreading endocannibalism.) Although group-detrimental norms are *more* likely than beneficial norms to develop out of the interaction among individuals within a group (according to the modelers), cultural group selection will nevertheless tend to favor the proliferation of the latter.

Groups that happen to have cooperative norms tend to outsurvive other groups, receive more immigrants, and (as discussed above) their individuals tend to be selected as models to imitate by individuals in other groups. On Chudek et al.'s (2013) account, because norms inevitably arise as a "by-product" of cultural learning, natural selection would have favored innate dispositions for handling norms—a "norm psychology." They suggest that our innate norm psychology prepares us to recognize, and motivates us to observe, the social norms in our environment. Since cultural group selection caused

cooperative norms (including norms to punish defectors and punish nonpunishers) to prevail in the ancestral environment, our norm psychology should make us especially disposed to acquire prosocial norms and punish violators (see also Henrich and Boyd 2001: 87). Indeed, in accordance with Chudek et al.'s predictions, evidence suggests that young children automatically infer behavioral rules by observing people (particularly adults), and that they internalize rule adherence as a personal goal and even enforce observance in other children (Chudek and Henrich 2011; Chudek et al. 2013).

The theories of both Boehm and Chudek et al. predict the existence of a "norm psychology." Both claim that Paleolithic hunter-gatherers lived in norm-governed societies and that individual success was tied to the ability to recognize and follow local norms, which tended to be prosocial. Laboratory experiments revealing our disposition to infer and follow norms will not adjudicate between the theories.

The theories, although making virtually the same predictions about how recent hominin evolution shaped our psychology, give different explanatory accounts of how prosocial norms initially spread, and they paint radically different pictures of the human capacity to *deliberately* guide our own evolution (both cultural and genetic). The empirical evidence reviewed in this paper seems to favor Boehm. Among nomadic hunter-gatherers around the world—all that have been studied by anthropologists—powerful coalitions enforce prosocial norms with a great deal of explicit awareness of the social benefits that observance of these norms will produce. These coalitions enforce the norms because individuals have an explicit blueprint in their minds for what kind of society they want to live in, and norm enforcement is a deliberate strategy for bringing that society about. In many cases humans are blind to the consequences of their socially transmitted practices—the evidence for that is undeniable (Henrich 2016). But our hunter-gatherer ancestors were not blind to the consequences of all of their practices.

When coalitions of the majority seized power from alphas, they gained the ability to impose practices that had consequences that they favored.

9.1. An Alternative Explanation for the Development of Prosocial Behavior: The "Mutualistic Theory" of Morality

Baumard et al. (2013) defend an alternative explanation for the evolution of moral judgment and prosocial behavior. They suggest that, in the ancestral environment, people could benefit by cooperating with each other in a variety of ways: they could hunt together, share food with the expectation of future reciprocation, and so on. Each person would bring a certain amount of resources, effort, and talent to cooperative ventures. This led people to *compete* to be chosen as partners for cooperation. If *A* will get a lower return by cooperating with *B* than *A* would get on average from cooperating with someone else, *A* will be better off not choosing *B* as a partner, and *B* will lose out. Our ancestors faced the adaptive challenge of seeking out good exchange partners while making themselves attractive exchange partners to others. Baumard et al. argue that, as a consequence, we evolved a moral disposition to value "fairness"—the notion that people are entitled to share in the product of cooperation in proportion to their contribution to creating that product. In hunter-gatherer tribes the moral order was not sustained by collective agreement about rules and collectively administered punishment, but by individuals exercising "partner choice."

Baumard et al.'s (2013) theory seems to explain a range of findings in experimental games. For example, in dictator games, where one person (the dictator) decides how to divide a pot of money with an anonymous partner, dictators tend to keep most of the pot for themselves. But suppose the dictator is asked to distribute money that

was *earned* by the anonymous partner by performing well in a quiz contest or on an exam. In that case, dictators tend to become very generous, and give their partners more or less what they earned (Ruffle 1998; Oxoby and Spraggon 2008). We seem to intuitively feel that people are entitled to the product of their efforts and, according to Baumard et al. (2013), we object to inequality only when it is the consequence of *unfairness*.

The theories of Baumard et al. and Boehm can, to some extent, be reconciled. It could be that groups began collectively enforcing moral codes around 250,000 bp (as Boehm argues), but voluntary mutualistic cooperation contributed to the evolution of our moral sense (in the way described by Baumard et al.) *before* that time. Even if early human groups were dominated by alphas who bullied group members and sometimes appropriated resources for themselves by force, not all interactions had to be based on bullying: group members could still have cooperated with each other in some contexts, and competed to be chosen as cooperation partners. Furthermore, even after hunter-gatherers established a reverse dominance hierarchy and egalitarian political norms, people were still allowed to obtain unequal rewards due to greater ability in some contexts. Although the rules among nomadic foragers for sharing meat from big game generally demand a more or less equal distribution, the rules for sharing other kinds of food are often much looser, and people who contribute more to obtaining such food may have a degree of freedom to distribute it as they wish (see Gurven 2004). Even if, since 250,000 bp, hunter-gatherer bands were politically egalitarian (among males) and big game hunting/distribution was a *largely* socialist enterprise, people still formed cooperative relationships with each other on a smaller scale, and it would have been advantageous for them to follow norms of fairness even if these norms were not always enforced by the collective effort of the group.

Although some of our behavior and moral intuitions can be explained by Baumard et al.'s (2013) account, we cannot discount the evidence that antisocial behavior is (and presumably was) often collectively punished in egalitarian bands, and that our moral psychology is to some extent tailored to living in such egalitarian bands. As noted, Baumard et al. suggest that antisocial hunter-gatherers are, in general, not formally "punished," but rather they suffer when other members of their group refuse to cooperate with them. But the anthropological evidence reviewed above suggests that hunter-gatherers do not merely withdraw cooperation from antisocial individuals. Groups collectively mete out positive punishments—the death penalty is widely administered for a range of offenses. It is true that, as Baumard et al. say, "[p]unishment...is uncommon in societies of foragers" (Baumard et al. 2013: 66). But this can be explained by the fact that the threat of punishment preemptively stops most serious offenses (cf. Service 1975: 48–49). The death penalty rarely needs to be administered because it is preceded by many escalating warnings of what is to come should offenders fail to reform.

The hypothesis that our moral sense is calibrated for living under conditions of somewhat enforced egalitarianism also explains some results from experimental games, which Baumard et al. (2013) cannot so easily explain. They discuss the following experiment conducted by Dawes et al. (2007). Subjects were divided into groups containing four anonymous members for a one-shot interaction. Each subject was given a random amount of money by a computer, and was informed how much money had been granted to the other three members. The players then had the chance to give "positive" or "negative" tokens to each other. Giving a positive or a negative token cost the giver one monetary unit, and increased or decreased (respectively) the recipient's payoff by three monetary units. Groups were broken apart and formed with new anonymous members to play again for a total of five rounds.

The results of Dawes et al.'s study suggest that people are willing to pay a penalty to equalize outcomes. In the course of five rounds, 68% of subjects "reduced another player's income at least once, 28% did so five times or more, and 6% did so ten times or more....74%...increased another player's income at least once" (Dawes et al. 2007: 794). Those who received a very high initial endowment tended to receive many negative tokens, while those who received a very low initial endowment tended to receive many positive ones. Subjects' token-distributing behavior cannot be seen as a rational strategy because, again, the interactions were one shot. It cannot be interpreted as retaliation or punishment, since subjects knew that all of the payoffs had been determined randomly by a computer. Rather, subjects seemed to be bothered by inequality per se. Baumard et al. offer the following explanation: "Overall, the distribution of [tokens]...displays the logic of fairness: The more a participant received money, the more others would 'tax' her. Conversely, the less she received, the more she would get 'compensated'" (Baumard et al. 2013: 75). However, the way subjects reacted to inequality does *not* make them attractive partners for cooperation. Paying a personal cost in order to reduce the payoff of luckier individuals simply reduces the expected payoff from engaging in cooperation. Unlike a "tax," which *redistributes* payoffs and thereby benefits the recipient, the negative tokens administered by 68% of subjects reduced both their own and the recipients' payoffs *without benefiting anyone*. Thus, the disposition that motivates negative-token giving does *not* make us attractive partners in cooperation. But the behavior is easily explained if we assume that we are adapted to living in hunter-gatherer bands where a certain degree of egalitarianism was enforced, and people were prohibited from accumulating significantly more resources than their fellows.

10. Discussion

Mainstream cultural evolutionary theory in the Boyd and Richerson tradition assumes that cultural variants spread as a consequence of individuals' content and context learning biases, Darwinian selection, and random forces in a way that is amenable to mathematical modeling. Durham (1991), Fracchia and Lewontin (2005), and Lewontin (2005) raise the challenge that often cultural variants are spread through the exercise of power, implying that the target of explanation for cultural evolution should be the behavior of powerful individuals and groups. This paper argued that Durham and Fracchia and Lewontin's critique may be valid as far as the evolution of morality and prosociality goes (at least since 250,000 bp), but less so for the spread of nonmoral cultural variants (perhaps until the advent of sedentarism, food storage, and agriculture).

Cultural evolutionists emphasize that their models are not intended to capture everything that happens in real life (Richerson and Boyd 1987; McElreath and Boyd 2007: 4–6). Like all models, they are meant to be simplifications that isolate some of the main forces at play. But if something like Boehm's (1999, 2012) account of the evolution of morality is correct, cultural evolutionary models of the evolution of morality and the coevolution of norms and norm psychology are *distortions*, not simplifications. A legitimate model may isolate one or a few forces among the multitude that exist in the real world. It cannot postulate forces that are not operative at all, or ignore those that have a decisive influence on the phenomena under investigation.

In light of the anthropological evidence reviewed in this paper, we can see how Boyd et al.'s (2010) model of the evolution of "coordinated punishment" distorts history. According to Boyd et al., there was a genetic mutation(s) associated with the behavior to *punish cheaters if τ other people in my group (honestly) signal that they are prepared to*

punish cheaters. (Again, this mutation can survive better than the mutation associated with *always punish cheaters*, since always punishers are disadvantaged when they are a small minority.) In a population containing some conditional punishers, by chance a few groups will have the threshold number (i.e., $\tau + 1$), and they will prosper and replace other groups, thereby increasing the number of conditional punishers.

The story that Boehm (1999, 2012) tells is very different. He says that coalitions within human groups formed and explicitly agreed to enforce certain kinds of prosocial behavior through punishment. The people involved did not have any preexisting disposition to punish. The practice of punishing people to enforce prosocial norms arose all at once. In contrast, on Boyd et al.'s (2010) account the genetically based disposition to punish arose *first*, then people developed a social norm to punish. On Boehm's (1999, 2012) account, because groups started practicing collective punishment, this created selection pressures for people with a genetic disposition to punish norm transgressors.

To be clear, the point of this paper is *not* to say that there are real-life complexities that cultural evolutionary models fail to capture. The point is that some of the fundamental forces that drive cultural evolution, and drove the coevolution of norms and norm psychology, cannot be accommodated by the sorts of models used in cultural evolutionary theory. Most of the mathematical models employed in cultural evolutionary theory are best adapted to cases where forces act in regular, iterative ways. Perhaps evolution in some domains of culture does work like that—cultural evolutionary modeling in these domains is safe from Durham (1991) and Fracchia and Lewontin's (2005) challenge. In other domains—including the moral—a more traditional historical/anthropological/sociological approach may be more fruitful (and legitimate).

Acknowledgments

I am grateful to Andreas Mogensen for valuable conversations on this topic and feedback on multiple drafts of this paper. I received helpful comments on previous drafts from Christopher Boehm, Neven Sesardić, and two anonymous reviewers.

References

Bachrach, Peter, and Morton S. Baratz. 1970. *Power and Poverty: Theory and Practice*. New York: Oxford University Press.

Boehm, Christopher. 1993. "Egalitarian Behavior and Reverse Dominance Hierarchy." *Current Anthropology* 34 (3): 227–240.

———. 1997. "Impact of the Human Egalitarian Syndrome on Darwinian Selection Mechanics." *The American Naturalist* 150 (S1): S100–S121.

———. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

———. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution.* Princeton, NJ: Princeton University Press.

Boyd, David J. 2001. "Life without Pigs: Recent Subsistence Changes among the Irakia Awa, Papua New Guinea." *Human Ecology* 29 (3): 259–282.

Boyd, Robert, Herbert Gintis, and Samuel Bowles. 2010. "Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare." *Science* 328 (5978): 617–620.

Boyd, Robert, and Peter J. Richerson. 1985. *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

———. 2002. "Group Beneficial Norms Can Spread Rapidly in a Structured Population." *Journal of Theoretical Biology* 215 (3): 287–296.

Cashdan, Elizabeth A. 1980. "Egalitarianism among Hunters and Gatherers." *American Anthropologist* 82 (1): 116–120.

Chudek, Maciej, Sarah Heller, Susan Birch, and Joseph Henrich. 2012. "Prestige-Biased Cultural Learning: Bystander's Differential Attention to Potential Models Influences Children's Learning." *Evolution and Human Behavior* 33 (1): 46–56.

Chudek, Maciej, and Joseph Henrich. 2011. "Culture–Gene Coevolution, Norm-Psychology and the Emergence of Human Prosociality." *Trends in Cognitive Sciences* 15 (5): 218–226.

Chudek, Maciej, Wanying Zhao, and Joseph Henrich. 2013. "Culture–Gene Coevolution, Large-Scale Cooperation and the Shaping of Human Social Psychology." In *Cooperation and Its Evolution,* edited by Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser, 425–457. Cambridge, MA: MIT Press.

Dahl, Robert A. 1957. "The Concept of Power." *Behavioural Science* 2 (3): 201–215.

Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex.* Vol. 1. London: John Murray.

Dawes, Christopher T., James H. Fowler, Tim Johnson, Richard McElreath, and Oleg Smirnov. 2007. "Egalitarian Motives in Humans." *Nature* 446 (7137): 794–796.

Durham, William H. 1991. *Coevolution: Genes, Culture, and Human Diversity.* Stanford, CA: Stanford University Press.

Erdal, David, and Andrew Whiten. 1994. “On Human Egalitarianism: An Evolutionary Product of Machiavellian Status Escalation?” *Current Anthropology* 35 (2): 175–178.

Fracchia, Joseph, and Richard C. Lewontin. 2005. “The Price of Metaphor.” *History and Theory* 44 (1): 14–29.

Goodall, Jane. 1986. *The Chimpanzees of Gombe: Patterns of Behavior*. Cambridge, MA: Harvard University Press.

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach. 2006. “The Competitive Advantage of Sanctioning Institutions.” *Science* 312 (5770): 108–111.

Gurven, Michael. 2004. “To Give and to Give Not: The Behavioral Ecology of Human Food Transfers.” *Behavioral and Brain Sciences* 27 (4): 543–560.

Henrich, Joseph. 2006. “Cooperation, Punishment, and the Evolution of Human Institutions.” *Science* 312 (5770): 60–61.

———. 2016. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich, Joseph, and Robert Boyd. 2001. “Why People Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas.” *Journal of Theoretical Biology* 208 (1): 79–89.

Henrich, Joseph, and Richard McElreath. 2003. “The Evolution of Cultural Evolution.” *Evolutionary Anthropology* 12 (3): 123–135.

Horner, Victoria, Darby Proctor, Kristin E. Bonnie, Andrew Whiten, and Frans B. M. de Waal. 2010. “Prestige Affects Cultural Learning in Chimpanzees.” *PLoS ONE* 5 (5): e10625.

Knauft, Bruce M. 1991. “Violence and Sociality in Human Evolution.” *Current Anthropology* 32 (4): 391–409.

Lewens, Tim. 2007. *Darwin.* London: Routledge.

———. 2015. *Cultural Evolution: Conceptual Challenges.* Oxford: Oxford University Press.

Lewontin, Richard C. 2005, October 20. “The Wars over Evolution.” *The New York Review of Books*. Retrieved from http://www.nybooks.com/articles/2005/10/20/the-wars-over-evolution/

Mameli, Matteo. 2013. “Meat Made Us Moral: A Hypothesis on the Nature and Evolution of Moral Judgment.” *Biology & Philosophy* 28 (6): 903–931.

McElreath, Richard, and Robert Boyd. 2007. *Mathematical Models of Social Evolution: A Guide for the Perplexed.* Chicago: University of Chicago Press.

McGrew, W. C. 2004. *The Cultured Chimpanzee: Reflections on Cultural Primatology.* Cambridge, UK: Cambridge University Press.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. “Covenants with and without a Sword: Self-Governance Is Possible.” *American Political Science Review* 86 (2): 404–417.

Oxoby, Robert J., and John Spraggon. 2008. “Mine and Yours: Property Rights in Dictator Games.” *Journal of Economic Behavior & Organization* 65 (3–4): 703–713.

Richerson, Peter J., and Robert Boyd. 1987. “Simple Models of Complex Phenomena: The Case of Cultural Evolution.” In *The Latest on the Best: Essays on Evolution and Optimality,* edited by John Dupré, 27–52. Cambridge, MA: MIT Press.

———. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: University of Chicago Press.

Richerson, Peter J., and Joseph Henrich. 2012. “Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems.” *Cliodynamics* 3 (1): 38–80.

Ruffle, Bradley J. 1998. “More Is Better, but Fair Is Fair: Tipping in Dictator and Ultimatum Games.” *Games and Economic Behavior* 23 (2): 247–265.

Service, Elman R. 1975. *Origins of the State and Civilization: The Process of Cultural Evolution.* New York: Norton.

Turnbull, Colin M. 1961. *The Forest People: A Study of the Pygmies of the Congo.* New York: Simon & Schuster.

A Debunking Explanation for Moral Progress[*]

Abstract: According to "debunking arguments," our moral beliefs are explained by evolutionary and cultural processes that do not track objective, mind-independent moral truth. Therefore (the debunkers say) we ought to be skeptics about moral realism. Huemer counters that "moral progress"—the cross-cultural convergence on liberalism—cannot be explained by debunking arguments. According to him, the best explanation for this phenomenon is that people have come to recognize the objective correctness of liberalism. Although Huemer may be the first philosopher to make this explicit empirical argument for moral realism, the idea that societies will eventually converge on the same moral beliefs is a notable theme in realist thinking. Antirealists, on the other hand, often point to seemingly intractable cross-cultural moral disagreement as evidence against realism (the "argument from disagreement"). This paper argues that the trend toward liberalism is susceptible to a debunking explanation, being driven by two related non-truth-tracking processes. First, large numbers of people gravitate to liberal values for reasons of self-interest. Second, as societies become more prosperous and advanced, they become more effective at suppressing violence, and they create conditions where people are more likely to empathize with others, which encourages liberalism. The latter process is not truth tracking (or so this paper argues) because empathy-based moral beliefs are themselves susceptible to an evolutionary debunking argument. Cross-cultural convergence on liberalism per se does not support either realism or antirealism.

Keywords: Evolutionary debunking arguments, Moral realism, Argument from disagreement, Moral progress, Liberalism

---

[*] Published in *Philosophical Studies* 177, no. 11 (2020): 3171–3191.

1. Introduction

Moral realists believe that there are facts in virtue of which (at least some of) our moral beliefs are objectively and non-relatively true or false (Tersman 2006). Realists typically also hold that moral knowledge is possible, and that some (or many) of our ethical views are in fact true.

Skeptics about moral realism often appeal to "debunking arguments" to undermine the justification of our moral beliefs. Debunking arguments come in several varieties (Sauer 2018: chapter 1) which roughly conform to the following schema:

> *Causal premise.* S's belief that p is explained by X.
> *Epistemic premise.* X is [a non-truth-tracking] process.
> Therefore
> S's belief that p is unjustified. (Kahane 2011: 106)

There is ongoing debate about exactly what epistemic principles underly debunking arguments, and how such arguments should be formulated (e.g., Bogardus 2016; White 2010). Nichols (2014) favors "process" over "best-explanation" debunking arguments: *S*'s belief that *p* is unjustified when it is the product of an "epistemically defective" process. On Vavova's (2018) account, beliefs are rendered unjustified if we discover that they were (decisively) shaped by an "irrelevant influence": "An **irrelevant influence** for me with respect to my belief that *p* is one that (a) has influenced my belief that *p* and (b) does not bear on the truth of *p*" (p. 136).

According to some popular *evolutionary* debunking arguments (EDAs), our core moral beliefs reflect our basic evaluative tendencies, which in turn were implanted in us by natural selection for the purpose of increasing inclusive fitness (Street 2006; see also Joyce 2006: chapter 6). Natural selection favors evaluative tendencies that increase

inclusive fitness, and cares nothing for whether they lead to judgments that align with objective moral truths (if such truths existed). This is the epistemically defective cause/process—or "irrelevant influence"—that undermines our belief in objective moral truth.

While EDAs have become popular in recent years, the *argument from disagreement* is, as Sauer (2018: 99) notes, "arguably the most common challenge to metaethical moral realism." Cultures—and to some extent individuals within cultures—seem to disagree about fundamental moral principles. Antirealists often claim that we would not expect such disagreement if everyone had the potential ability to perceive objective moral truth (Mackie 1977: 36–37). Realists often counter that moral disagreement is more superficial than it first appears, or that disagreement—at least about "normatively significant core issues" (Sauer 2018: 100)—does not (or would not) persist under ideal conditions. The mere fact that cultures disagree about morality does not mean that there can be no objective truth of the matter, since cultures disagree about many matters of objective fact (Railton 1986: 195).

The argument from disagreement is essentially empirical—in light of an observation (moral disagreement) we should reject realism. Huemer (2016) turns the argument from disagreement on its head, claiming that cultures are *converging* on certain moral beliefs, which supports realism.

Huemer (2016) advances an "empirical case" not just for moral realism, but for the claim that *liberalism* is the correct moral theory. Liberalism is a "broad ethical orientation [that] (1) recognizes the moral equality of persons, (2) promotes respect for the dignity of the individual, and (3) opposes gratuitous coercion and violence" (p. 1987). "[W]hile this broad orientation is mostly uncontroversial today," he says, "human history has been dominated by highly illiberal views" (p. 1987).

He says that the standard debunking arguments, which purport to explain our moral beliefs by appealing to non-truth-tracking forces such as natural selection and culture, cannot explain the historical drift toward liberalism. Debunking arguments "lack credibility because they afford no explanation for the most important fact about the history of moral thought: the spread of liberalism across the world over the course of human history, especially recent history" (p. 2007). The trend occurred far too rapidly to be explained by biological evolution, and

> [p]urely cultural accounts of the source of morals leave us at a loss to explain why the culture itself has moved in a given direction over time. At first glance, it may seem that many explanations are possible—for instance, perhaps changing technologies or changing forms of economic organization have somehow necessitated different values. But the list of potential explanations dwindles as we try to take into account the entire phenomenon: it is not just, for example, that slavery was abolished in the United States. It is that societies around the world have been liberalizing with respect to many different issues—slavery, war, torture, execution, democracy, women's suffrage, segregation, and so on—and this has been going on for centuries. It is very difficult to come up with explanations for this broad phenomenon that don't require us to posit large coincidences. (p. 2007)

Liberal realism, however, "can offer a plausible account of the data," i.e., data "concern[ing] the development of moral values over the course of human history" (p. 1988)—what some philosophers describe as "moral progress."

Huemer's argument is an inference to the best explanation, which can be spelled out as follows:

(1) Across cultures, people have been converging on liberal practices and values on a wide range of disparate issues.

(2) Moral realism predicts convergence on practices and values.

(3) Moral antirealism predicts divergence in practices and values.

(4) There are two candidate hypotheses to explain the historical convergence on liberalism (1).

   a. The debunking hypothesis: For each issue in question, people have become more liberal for "[p]urely cultural" reasons that have nothing to do with recognizing objective moral truth.

   b. The moral realist hypothesis: Over time, people have increasingly adopted liberal practices and values *because* liberalism is true.

(5) In light of (2) and (3), convergence on liberal practices and values is best explained by moral realism (4b) rather than by a non-realist debunking explanation (4a).

(6) Therefore, liberal realism is true.

Huemer may be the first philosopher to explicitly make the inference from moral convergence to the truth of moral realism. However, many realists have said that realism predicts some degree of convergence, and attributed convergence to our perception of objective truth. Parfit (2011: 538) writes: "[T]hough humanity's earliest moral beliefs were in several ways distorted by evolutionary forces, those distortions are being overcome, so that true moral beliefs are becoming more and more widely held." Brink (1989: 208–209) says that, given the difficulty in acquiring moral knowledge and the fact that many moral disputes "depend on complex nonmoral issues," realists should not necessarily expect perfect convergence. Nevertheless, there has been "significant convergence of moral belief [and] moral progress over time....[The] relevant changes in moral consciousness have all been changes in the same direction." Smith (1994: 188–189) says that "there has been considerable moral progress" and that "we should...be quite optimistic about the possibility of an agreement about what is right and wrong being reached under more idealized conditions" (conditions which, in his view, may or may

not be established at some point). Singer (1981) famously argues that the perception of moral truth via reason is driving an expansion of our circle of moral concern.

This paper challenges step (3) of Huemer's argument—the idea that moral antirealism predicts divergence in practices and values. As noted, this is a claim that antirealists themselves generally accept. The lack of an explanation for the convergence on liberal moral beliefs (insofar as it has occurred) is a serious lacuna in standard debunking arguments. It will be argued here that there are non-truth-tracking processes that tend to push cultures toward liberalism. We do not need to posit "large coincidences" to explain why many cultures have drifted toward liberalism on a variety of issues.

2. Structure of the Argument

The question is whether the following principle of inference is epistemically reliable: *If cultures converge on a set of moral beliefs X, conclude that X is objectively, non-relatively true (i.e., true in realist terms).* A debunking account of moral progress should show that this is not epistemically reliable, because the process that drives cross-cultural moral convergence does not track moral truth (conceived in realist terms).

There are various ways of understanding the notion of tracking the truth. For example, a debunker may claim that a targeted belief is *insensitive*, i.e., the believer would still have believed that *p* if *p* had been false (Nozick 1981). But there are serious concerns about this kind of counterfactual when the truth in question is (said to be) metaphysically necessary (Clarke-Doane 2015: 87–92). Furthermore, sensitivity does not seem to be a requirement for justified belief or knowledge (Bogardus 2016; Vogel 1987; White 2010). Alternatively, a debunker can claim that it is not necessary to assume the truth of the targeted belief in order to explain why we hold it (see Harman 1977). In

other words, an "evolutionary [or other naturalistic] explanation makes objective morality redundant" (Ruse and Wilson 1986: 187; cf. Joyce 2006: 220). This second approach is the one taken here.

The aim of this paper is to show that the phenomenon of cross-cultural moral convergence is susceptible to a naturalistic explanation. In regard to cross-cultural convergence on liberalism, antirealism and liberal realism are (largely) predictively equivalent (though they might make different predictions about *how* that outcome will come about). Whether the debunking or the realist account is more convincing will depend to some extent on one's prior metaethical views. The mere fact of convergence per se does not support realism or antirealism. However, some reasons will be given for why the naturalistic, debunking story may give a more plausible account of the facts (sections 7.2–7.3).

In outline, the argument of this paper is as follows:

(1) Early human societies were structured like those of chimpanzees. Foraging bands were dominated by an alpha male who ruled tyrannically in pursuit of his own narrow self-interest. A couple hundred thousand years ago, hunter–gatherers overthrew their alphas and established strong egalitarian norms (at least among adult men) to prevent anyone from attaining a position of dominance. Although they were often violently aggressive toward outgroups, within societies they were quasi-socialist.

(2) With the advent of agriculture, hierarchies reemerged, and highly illiberal social systems were established. Hierarchical, militaristic, agricultural societies proliferated as a result of competition among groups. Egalitarian hunter–gatherers who came into contact with these societies were exterminated or absorbed.

(3) For reasons of self-interest, most people in hierarchical, oppressive societies do not like being abused for the benefit of those in power, or living with the constant threat of violence from within or outside their group. They may acquiesce to mistreatment or violence for a number of reasons, but when the opportunity arises to resist they often take it. For the same reason that hunter–gatherers overthrew their tyrannical alphas and established egalitarian and cooperative norms, people in modern hierarchical societies have, when possible, frequently pushed back against abuse and demanded more liberal treatment. In large hierarchical societies where rulers preside over armies and police forces it is more difficult for the masses to push back against their oppressors than it is for a couple dozen hunter–gatherers to overthrow their alpha. For this reason, the liberalizing process has been slow and tortuous in agricultural societies, but the trend is evident.

(4) As societies become more liberal—less violent within the group, more cooperative with other groups, and so on—this reinforces psychological dispositions, such as empathy and aversion to violence, that in turn drive further liberalism.

(5) The main drivers of the trend toward liberalism, namely, the pursuit of self-interest and empathic concern for others, do not track objective moral truth.

(6) The debunking argument outlined in (1)–(5) does not disprove realism, but it blocks a seemingly strong argument against antirealism.

The claim in (6) has already been addressed above. It is worth making some preliminary comments about (1)–(5).

(1) has been defended (with some variations) by a number of anthropologists (e.g., Cashdan 1980; Knauft 1991; Service 1975), and perhaps most persuasively by Boehm (1999, 2012). The evidence Boehm gives will be briefly reviewed below. (2) has

been defended by anthropologists including Boehm, Diamond (1987), and Turchin (2009, 2010; Turchin and Gavrilets 2009). Of course, almost no anthropological theory about the distant past will be without controversy. Nevertheless, (1) and (2) are, if not definitively proved, highly plausible accounts of our history. If they are correct, then the trend toward liberalism has *in part* been a historical return to liberalism. (It is a return only *in part* because there are important differences between hunter–gatherer and modern liberalism, as shall be discussed.) It is reasonable to suppose that some of the same forces that led hunter–gatherers to converge on liberalism (described in [3]) are responsible for the contemporary convergence on liberalism.

(4) is difficult to test directly, but it is based on two well-established principles of psychology. First: desensitization. Repeated exposure to a stimulus tends to reduce its stimulatory effect. By the same token, *lack* of exposure to a stimulus *increases* its stimulatory effect (Mrug et al. 2016). Second, the presence of external threats promotes tribalism and increases hostility toward outgroups (Buchanan and Powell 2016, 2018).

According to (5), the interrelated processes that drive cross-cultural moral convergence—viz., the pursuit of narrow self-interest and empathizing—do not track moral truth. That is, we do not need to assume the existence of objective moral truth to give a complete explanation of moral beliefs that are generated by these processes. First, the pursuit of narrow self-interest. Moral beliefs are sometimes created merely to advance the self-interests of their creators. For example, throughout history powerful people have often promoted the idea that we have a moral obligation to respect social hierarchies and grant special privileges to ruling classes. Ambitious politicians spread the idea that certain classes of people ought to be despised in order to create conflicts that help them solidify their own power. People whose self-interest is served by opposite moral beliefs have sometimes successfully promoted *those* beliefs. When different people or groups

compete to spread moral beliefs that serve their respective self-interests, which one prevails is determined by sociological facts (who has the most social/political clout). To explain the resulting token moral beliefs we do need to appeal to moral truth. Second, empathizing. Empathy-based moral beliefs are vulnerable to a strong EDA (spelled out in more detail in Sect. 7.1). According to this EDA, the explanation for our empathic tendencies appeals only to natural selection, not to objective moral truth.

3. Naturalistic Explanations for Moral Progress

An important element of (what is called) "moral progress" is the apparent trend toward more "inclusivist" morality (Buchanan and Powell 2016). Most traditional moralities are "exclusivist"—they ascribe much greater moral worth to in- than outgroup members. Over time, many societies have adopted moral outlooks that to some extent dissolve ingroup/outgroup distinctions, and affirm the moral value of everyone.

Buchanan and Powell (2016) offer a "naturalistic theory" to explain the trend toward inclusivist morality. They suggest that "exclusivist moral psychology is...an 'adaptively plastic' trait" (p. 998). In the ancestral environment of the Pleistocene, outgroups could present either a threat or an opportunity. Outgroups could have posed *threats* "relate[d] to the transmission of infectious disease, competition over scarce resources, external physical dangers, or beliefs and practices that are dissonant with in-group values and thus imperil group cohesion" (p. 997). They also could have provided *opportunities* for trade or mate exchange. On Buchanan and Powell's account, our moral psychology evolved to respond adaptively to signs of either threat or opportunity. If outgroups pose a threat (in terms of disease, competition, etc.) we develop an *exclusivist* morality. Otherwise we develop a more *inclusivist* morality.

According to Buchanan and Powell, in "circumstances that mimic the harsh conditions" (p. 997) that prevailed most of the time in the ancestral environment, people regard outgroups as dangerous threats and develop exclusivist moral outlooks. When societies become more prosperous, and conditions less harsh, this triggers our moral psychology to become more inclusivist.

Buchanan and Powell are surely right that inclusivist moralities—and liberalism in general—will not flourish among people who perceive each other as dangerous threats. But their account cannot explain all elements of the trend toward inclusivist morality (or liberalism in general). Consider, for example, the treatment of women. An important aspect of moral progress is the dramatic improvement in the treatment of women. But the reason for female oppression was never that men regarded women as a dangerous outgroup carrying disease or posing a violent threat. Rather, for much of history, men regarded women as people who had less power and whose preferences they did not need to take into consideration.

Moral progress is often driven by people forcefully asserting their own rights. Oppressed people do not always passively wait for the moral circle of their oppressors to expand. To gain recognition as moral equals they protest, elicit sympathy, acquire political influence, and engage in all manner of wheeling and dealing. Part of the explanation for the spread of liberalism is that victimized people have fought for better treatment. Blacks played a major role in the civil rights movement, women in women's liberation, colonized people in overthrowing empires. Railton (1986: 194–195) makes a similar point when he says that "in the long haul, barring certain exogenous effects, one could expect an uneven secular trend toward the inclusion of the interests of (or interests represented by) social groups that are capable of some degree of mobilization." Commenting on Railton, Buchanan and Powell (2018: 82–83) observe that societies

often develop moral concern for groups that are *not* "capable of some [seriously threatening] degree of mobilization." Both Railton and Buchanan and Powell may be right. Oppressed people can and often do employ various means of encouraging their oppressors to expand their moral circle, but this is not always necessary. Animals are incapable of protesting on their own behalf, but throughout the world people have become increasingly concerned with animal welfare (see Buchanan and Powell 2018, p. 83).

According to Hopster (2020: 1263), "[i]n a world with a global traffic of goods and information, in which societies depend on each other for resources and share mutual goals, it should not come as a surprise that they also come to adopt a roughly shared moral outlook." As people participate in an increasingly global culture, "social conformity" leads everyone to adopt common values. A society's values are influenced by its historical experience and its social and material conditions. Since many societies shared important experiences (e.g., world wars) and are converging on common ways of life, this also drives moral development in similar directions. In Hopster's view, it is to some extent an accident that the global culture is (currently) relatively liberal. While there may be something to this theory, it will be argued here that convergence on liberalism is not an accident, and the impetus to move in this direction is coming from within every human society.

This paper argues that the trend toward liberalism can be explained by two related sociological phenomena that do not involve recognition of moral facts. First, basic liberal values—moral equality, respect for the dignity of individuals, opposition to "gratuitous coercion and violence"—have *always* been appealing to the majority of people for reasons of self-interest, at least *within* groups. Throughout history going back to the days of nomadic foragers, when majorities within populations have gained political influence

they frequently established relatively liberal norms in order to protect themselves from exploitation by the strong or powerful. Second, as a byproduct of successfully implementing self-protective liberal policies, people have become conditioned to be less aggressive and more empathic, and this has led to an expansion of the moral circle and the strengthening of liberal moral intuitions.

4. The Ancient Origins of Liberalism

Huemer (2016) claims that "human history has been dominated by highly illiberal views" (p. 1987), and "[t]he vast majority of governments in history have been dictatorial" (p. 1992). This may not be entirely correct.

There is reason to believe that until the beginning of sedentary living and agriculture, certain liberal moral norms were literally universal among humans, and political organization was egalitarian. If hunter–gatherer bands can be described as having a "government," then the majority of governments for the vast majority of history have been quasi socialist, not dictatorial. The main evidence for this comes from studies of hunter–gatherers in the twentieth century. All nomadic foraging societies that have ever been observed—from the Amazon to the Outback—actively prohibit any political hierarchy, at least among men. There is no recorded group of nomadic foragers that will tolerate an adult male issuing direct orders to another adult male (see the extensive documentation of this fact in Boehm 1999, 2012). All nomadic foragers have strict rules requiring that food—particularly meat from large game animals—be distributed among all members of the group.

This means, as one anthropologist put it, that "male status differentiation in human evolution is U-shaped" (Knauft 1991: 397). Chimpanzee (*Pan troglodytes*)

society is strictly hierarchical, with a tyrannical alpha ruling because he (perhaps along with his coalition partner) can physically dominate every other individual (Goodall 1986). The common ancestor of chimps and humans is presumed to have had a similar social structure (Boehm 1999). At some point, however, our hunter–gatherer ancestors overthrew their alphas and instituted what Boehm (1999) calls a "reverse dominance hierarchy": Men who would otherwise have been subordinate to an alpha banded together to prevent anyone from achieving a position of dominance. It was only in relatively recent history—when we moved away from nomadic hunting and gathering—that we reestablished traditional hierarchies, and adopted ideologies that affirmed the moral (and sometimes divine) legitimacy of rulers.

The upshot is that, until fairly recently, human societies had *converged* on what were essentially liberal moral norms—at least *within groups*—for reasons of self-interest among the majority. Certainly, there were important differences in the moral outlook of hunter–gatherers across the world. But some core moral norms were universal. Within groups, people were more or less recognized as moral equals, everyone was granted a fairly high degree of respect, and coercion was used mainly to protect people from being bullied, cheated, or pushed around by others. Women typically had less political influence than men, and their treatment varied among hunter–gatherer societies. But they were afforded many of the same protections as men. Women certainly were not subjugated to nearly the extent that they were in more recent history after the transition to agriculture (Boehm 1999; Hayden et al. 1986).

Hunter–gatherer liberalism typically did not extend beyond the ethnolinguistic group. Between-group relations were characterized by extreme *illiberalism*. Neighbors were often in a permanent state of war governed by few or no rules. Whole communities—particularly the men—were sometimes exterminated. In a survey of

twentieth-century studies on hunter–gatherers, Keeley (1996: Table 6.2) found that the percentage of male deaths due to warfare ranged from 8.3% for the Gebusi (Papua New Guinea) to 59.0% for the Jivaro (South America). However, as shall be argued below, even violent hunter–gatherers regarded peace as a desirable condition for reasons of pure self-interest, namely, they did not like living under the threat of attack. They simply lacked mechanisms to establish peace. The trend toward less war was driven not by recognition of the objective, mind-independent truth that war is morally bad. Rather, it was driven by institutional and technological innovations that allowed people to escape something that, for reasons of self-interest, they never liked in the first place.

5. The Origins of Modern Illiberal Hierarchies

When hunter–gatherers became sedentary, hierarchies eventually reemerged. The exact reason why sedentarism—which is usually accompanied by food storage—can lead to a breakdown in egalitarianism is not entirely clear, but it is probably related to the resulting increase in population size (cf. Boehm 1999: 143–144). When populations become much larger than a mobile foraging band, it is no longer possible for coalitions of the majority to surveil and collectively control everyone's behavior, or for the group to make collective decisions. After reaching a certain size, societies become nonfunctional without hierarchical leadership. The larger the population, the more extensive the hierarchy must be. Turchin and Gavrilets (2009: Fig. 2) find that, among six historical empires, every order of magnitude increase in population size over time was accompanied by the addition of another level of hierarchy. They report: "It appears that an acephalous tribe is the largest social scale a human group can achieve without the benefit of centralized organization" (p. 172).

When agricultural societies first developed around 10,000 years ago, they quickly exterminated or absorbed all mobile foragers with whom they had contact. This happened virtually everywhere that agriculture arose. It was not because agriculturalists were necessarily better off than foragers in terms of health or wellbeing—indeed, the lot of the average farmer 10,000 years ago was probably worse in many ways. Diamond (1987) calls the transition to agriculture "the worst mistake in the history of the human race" partly for that reason. But agriculture can sustain a much larger population than foraging. Because of the surplus it produces, it can support classes of people not directly involved in food production—classes that can specialize in things like bureaucracy, tool/weapon making, and warfare. Mobile foraging bands—even coalitions of bands—are no match for armies fielded by agricultural populations (Diamond 1987; Turchin and Gavrilets 2009).

The development of large-scale hierarchical communities led to intense intergroup competition. Small agricultural societies can defeat mobile foragers. Bigger or more organized agricultural societies can defeat smaller or less well-organized ones. Turchin has developed mathematical models suggesting that, given some realistic assumptions, the trend toward larger societies—including eventually massive empires—was the inevitable outcome (Turchin 2009, 2010; Turchin and Gavrilets 2009). People were not necessarily *choosing* these increasingly illiberal forms of social organization because they preferred them, or believed them to be aligned with moral truth. It's just that, wherever they developed, aggressively militaristic and hierarchical groups exterminated or absorbed their neighbors.

6. The Trend toward Liberalism

To say that there has been a trend toward "liberalism" can be misleading, since commitment to core liberal values takes different forms. It is important to be clear on what kind of liberalism we are talking about. The most basic distinction is between what we can call "within-" and "between-society" liberalism. Within-society liberalism applies the principles of liberalism—viz., moral equality, respect for the dignity of individuals, opposition to gratuitous coercion and violence—within a single society. Between-society liberalism applies those principles to outsiders. Within-society liberalism can be further divided into what we can call "narrow" and "broad" varieties. *Narrow* within-society liberalism applies to some restricted class(es) of people, generally those wielding political power—e.g., men or a dominant ethnic group. *Broad* within-society liberalism (theoretically) applies to everyone within a society. (Between-society liberalism can also be divided into narrow and broad varieties, but that distinction is not relevant for the present purposes.)

Each of these types of liberalism can vary in terms of its *intensity*. People can be deemed to have greater or lesser moral worth. They can be more or less respected and protected from gratuitous coercion and violence.

Nomadic foragers were characterized by intense narrow within-society liberalism (among men), more or less intense broad within-society liberalism (extending to women), and, among most groups, very low levels of between-society liberalism. In the modern era, arguably we seem to have been converging on fairly intense broad within-society liberalism, and only moderately intense between-society liberalism. That is to say, across cultures people generally support—if not in practice then in principle—the idea that *within their society* liberal principles apply to everyone: all citizens are equal

under the law, minorities and less powerful individuals should be protected, resources should be distributed to meet everyone's basic needs, people should resolve disputes nonviolently, and so on. No society lives up to these ideals perfectly, but the gap between performance and ideals is closing steadily. This is why we can say that there has been convergence on *intense* broad within-society liberalism. We can only say there has been convergence on *moderately* intense between-society liberalism because it is not a mainstream idea in any society that outsiders should be treated as genuine equals. No mainstream politician says that their country's resources should be distributed to the people of the world based on need, regardless of who or where they are. The US government spends less than 1% of the federal budget on foreign aid. A 2017 survey of likely voters in the US found that 57% think the government spends *too much* on foreign aid, 27% think it spends the right amount, and just 6% think it spends too little (Rasmussen Reports 2017). The US is hardly an exception. Sweden—famous for its generous social welfare programs that benefit citizens and residents—spends only a slightly higher percentage of its federal budget on foreign aid (Sida 2019).

It is not surprising that, throughout the history of civilization, people low on the hierarchy within groups—plebeians, peasants, and the like, who invariably comprise the majority of the population—have often supported more within-society liberal policies. Oppressed majorities have always wished to have their moral worth and dignity recognized, and to be free from excessive coercion by their rulers, each other, or anyone else. But it is also not surprising that progress toward liberalism has been slow. As the previous section argued, for much of history civilizations have been engaged in continual, vicious conflict with each other, with the strongest and most unified exterminating, brutally subjugating, or absorbing their neighbors. Only when local powers were evenly matched could temporary, peaceful equilibria *sometimes* be reached.

It is easy to see that, under those conditions, it would have been difficult for people to launch revolutions against oppressive rulers—much more difficult than it was for coalitions of a dozen hunter–gatherers to keep down would-be alphas. Nevertheless, there has been a tendency for majorities, when they have gained political power, to move society in a liberal direction, at least when it comes to their own treatment.

Clearly, we do not need to appeal to the existence of objective moral facts to explain why there has been a fair amount of cross-cultural convergence on *narrow within-society* liberalism that is restricted to politically influential classes of people. The potentially difficult questions are (a) why has there been some cross-cultural convergence on *broad* within-society liberalism? and (b) why there has been an *increase* in *between-society* liberalism?

7. Broad within-Society Liberalism

There are at least three forces that push cultures in the direction of greater broad within-society liberalism. First, people tend to empathize with those with whom they have close contact. Our tendency to empathize means that moral norms prescribing better treatment of people within society can easily take root. Second, when a subgroup within a society (e.g., men) successfully pursues *narrow* within-society liberalism, this can create a revolutionary atmosphere that inspires *other* groups (e.g., women) to demand the same benefits for themselves, which can lead to increasingly broad within-society liberalism. Third, *pacification* is an almost inevitable concomitant of a society becoming more prosperous and advanced. Prosperous people have less reason to resort to criminality and violence. Consequently, they have less need to adopt an aggressive stance to protect themselves from attack. As a society advances, policing becomes more effective, which

further reduces violence. As a result of living in a peaceful society, our impulses become more pacific, our sensitivity to violence increases, and we adopt increasingly extreme antiviolence moral norms. These three forces are considered here in turn.

7.1. Empathy

Humans clearly have a capacity to dehumanize perceived enemy outgroups. Under the right conditions, empathy toward an outgroup can effectively be set to zero. Many hunter–gatherers across the world had a practice of butchering and eating members of enemy tribes (Keeley 1996: 103–106). Ancient civilizations in the Americas, Europe, and Asia perpetuated killings—often in brutal ways—of defeated enemies on as large a scale as their technology allowed. Examples of people committing atrocities against outgroup members in recent history are too well known to require discussion. These observations may inspire a very cynical view of human nature. We might assume that treating members of an outgroup as worthy of moral concern requires us to overcome our natural impulses.

However, humans have a natural inclination—shared with our primate relatives (de Waal 1996)—to *empathize* with each other. Like other primates, we tend to empathize most strongly with close ingroup members, particularly our family and friends with whom we have established special bonds. But most people (besides psychopaths) respond empathically to cues of suffering and distress in others, at least to some extent, unless this reaction is suppressed by feelings of intense fear or hatred (cf. Buchanan and Powell 2016). Even under conditions of war people can fail to dehumanize their enemies, and enter into cooperative, friendly relationships. During World War I soldiers on the front lines developed a culture of "live and let live" in which they purposefully avoided

killing each other. During the famous Christmas truce of 1914, large numbers of British, French, and German soldiers left the trenches to fraternize and sing carols with each other in no man's land. (The live-and-let-live culture dissolved in December 1915 when the Germans introduced gas attacks.) The point is that our tendency to regard outgroup members as inhuman monsters worthy of death is hardly automatic or inevitable. Yes, under certain conditions we are capable of turning empathy off and engaging in extreme brutality. But this is only a *possible*, not the *default*, mode of human interaction, even vis-à-vis outgroup members. We do not have records of the inner psychological conflicts of waring hunter–gatherers, but there is some evidence that they were capable of empathizing with their enemies. In rare cases they may even have let adult male captives live, although "the ethnographic accounts indicate that such acts of mercy were at least unusual, if not exceptional" (Keeley 1996: 213, n. 8). Our tendency to empathize does not invariably lead to broad within-society liberalism. But it does make us psychologically prepared to move in that direction.

Nichols (2004) provides evidence that "affect-backed norms"—norms that "prohibit an action that is emotionally upsetting" (p. 128)—tend to be taken especially seriously, and are more likely to be passed on from generation to generation than affectively neutral norms. He finds, for example, that disgust-backed etiquette rules from the Middle Ages—i.e., rules prohibiting actions that people find disgusting—are much more likely to have survived to the present day than non-disgust-backed etiquette rules (Nichols 2002). Since harm to other people can trigger an emotionally upsetting empathic response, harm-prohibiting norms tend to be particularly enduring. People take harm-prohibiting norms very seriously, judging them to be *moral* as opposed to *conventional* rules (Turiel et al. 1987).

According to Nichols's (2004) theory, the fact that an action is emotionally

upsetting (e.g., spitting while eating in public, killing people) does not automatically compel us to prohibit it. Rather, once norms are established, those that do prohibit emotionally upsetting actions (e.g., don't spit while eating in public, don't kill people) have a special advantage in cultural transmission. Following this logic, we would not expect empathy (such as it is) toward members of oppressed classes of people within a society to *automatically* lead those in power to adopt norms to protect them from abuse. But feelings of empathy toward a class of people can potentially lead those in power to adopt norms that take their interests into account. Once broad within-society liberal norms gain a foothold, they may be especially resilient—likely to be passed on and to inspire strong commitment.

Consider the status of women. Collectively, men can dominate women by physical strength. Since the beginning of civilization, men have used their strength to arrange society according to their own preferences, often with little regard for the preferences of women. But most men have also cared about women—at least those with whom they had personal relationships. So it is not surprising that they would, in some places, heed women's demands for more favorable treatment, even equality.

The practice of slavery can be undermined by our tendency to form bonds and empathize with people. In the United States slaves were often abused, and some Whites seemed to have genuinely viewed Africans as a dangerous enemy population. But many Whites who had contact with slaves, including slaveholders, formed bonds with them. White and black children played together. Some slaves were closely integrated in white households. As W. E. B. Du Bois (1903/1986: 382) said, there could be "something of kindliness, fidelity, and happiness" in the relationship between slaves and slaveholders. White slaveholders such as Thomas Jefferson and George Washington even helped lay the groundwork for emancipation. (Washington supported the abolition of slavery in

principle, and his will stipulated that his slaves should be freed upon the death of his wife. Jefferson spearheaded some antislavery legislation, and clearly supported eventual emancipation.) Many white non-slaveholders saw slavery as nothing but abuse of a completely harmless people. Harriet Beecher Stowe's *Uncle Tom's Cabin*—the worldwide best-selling book of the nineteenth century after the Bible—is believed to have played an important role in ending slavery and triggering the US Civil War. The book emphasized the abuse of slaves while portraying them as childishly helpless and nonthreatening. In the preface, Stowe explained her intention to "awaken sympathy and feeling for the African race" (Stowe 1852: vi), and the sympathy she awoke did indeed lead many people to reject slavery. Buchanan and Powell (2016, p. 1002) note that a particularly effective tactic of abolitionists was to distribute drawings of the terrible conditions of slaves being transported on the Middle Passage.

The line of argument so far has assumed that we can debunk a moral belief by showing that it is caused by empathy. The moral realist might object. The realist could argue that our tendency to empathize is one of the ways that we detect moral truth. I.e., the realist could argue that *empathy is a truth-tracking emotion*—in fact, Marshall (2018) makes exactly this argument. When we empathize with slaves, this is a way of perceiving the wrongness of slavery. But our empathy-based moral beliefs are clearly susceptible to an evolutionary debunking argument. There is a large literature on the evolutionary purpose of empathy. The general conclusion of evolutionary biologists is that, as Sapolsky (2017) puts it, empathy "is a state on a continuum with what occurs in a baby or in another species" (pp. 652–653). "Lots of animals display building blocks of empathic states" (p. 655) for clear evolutionary reasons. Each species is endowed with the building blocks of empathy that conferred a fitness advantage in its ancestral environment. De Waal expounds:

> Evolutionary theory postulates that altruistic behavior evolved for the return-benefits it bears the performer....Empathy is an ideal candidate mechanism to underlie so-called directed altruism, i.e., altruism in response to another's pain, need, or distress. Evidence is accumulating that this mechanism is phylogenetically ancient, probably as old as mammals and birds....The dynamics of the empathy mechanism agree with predictions from kin selection and reciprocal altruism theory. (de Waal 2008: 279)

If evolutionary biologists are right that the "dynamics" or our empathic responses are explained by kin selection and reciprocal altruism theory, the onus is on the realist to show why empathy is nevertheless a truth-tracking emotion.

7.2. The Contagiousness of Revolutions

When one group within a society successfully fights for better treatment, this can inspire others to follow suit. The moral realist might argue that this is because when group *A* wins better treatment for itself, this can help group *B* recognize the objective truth that it is entitled to the same benefit. However, the debunker can argue that the pursuit of narrow self-interest is a better explanation of cases like this. If group *A* fights for its own rights, and *B* follows, why, according to the moral realist, did *A* fail to recognize that the same rights should have been extended to *B* in the first place? It seems that *A* was just concerned with its own interests. If *A* was motivated by self-interest, why suppose that *B* was motivated by recognition of moral truth?

Take the American Revolutionary War, which was a fight for narrow within-society liberalism, i.e., liberalism for the white population. The colonists sought to expel the British, who had treated them illiberally—who had exploited and disrespected them. They rallied around slogans expressing the value of freedom and their right to be treated

as equals vis-à-vis other nations. The famous phrase "all men are created equal" appeared in the Declaration of Independence and was included with variations in some state constitutions as well. Although slaves had been left out of the revolution, the colonists' success at securing their own rights inspired many slaves to demand the same rights for themselves. At a public reading of the Massachusetts Constitution in 1780, the slave Bett (later known as Elizabeth Freeman) heard the assertion that "All men are born free and equal." She sued for her freedom in a Massachusetts court, claiming that slavery violated the state constitution, and won. (Presumably she won because there was already a strong abolitionist movement in Massachusetts that made the court receptive to her arguments.) In other states, the struggle for freedom did not, of course, succeed so quickly. But the American Revolution helped trigger fiercer resistance to slavery.

The French Revolution was fought under the slogan "liberty, equality, fraternity." But when the revolutionary men took power, they decided that women were not to be treated as political or social equals, or to have the same liberties as men. Naturally, many women objected. The "Declaration of the Rights of Man and of the Citizen of 1789" was followed two years later by Olympe de Gouges' "Declaration of the Rights of Woman and the Female Citizen." A number of more or less militant feminist organizations were established, most notably the Society of Revolutionary Republican Women. Feminists were unsuccessful in their struggle for female equality, and in 1793 the National Convention banned women's clubs and organizations altogether. Gouges was executed during the Reign of Terror. But the seeds for women's equality had been planted: If men could demand equality, then so could women, although it was many years before woman gained the political clout to be successful.

7.3. Pacification

Huemer (2016) notes that there has been a significant, worldwide decline in violence over time. Although "many factors...may have contributed to this decline," he says, "one is of particular interest here: there has been a dramatic shift in human values over history" (p. 1988). In 1300 CE, the murder rate in Europe was 35 per 100,000 people per year. Today it is 3 per 100,000. "Again, many factors may have contributed to the decline—among them changing attitudes toward murder. Men of the past perceived many more things as reasons for killing" (p. 1990).[4] Huemer notes that, in 1804, former US Treasury Secretary Alexander Hamilton was killed in a duel with Vice President Aaron Burr, fought over some insulting remarks made about Burr by Hamilton. "Such behavior on the part of respected men would be unthinkable today" (p. 1990).

An alternative explanation of the trend toward lower murder rates is that, although people have always strongly opposed murder, we have just become better at preventing it. It is true that our attitudes toward killing in some specific circumstances have changed—dueling, for example, has become counter-normative. But this may be because law enforcement has become more effective and it is no longer necessary (in most parts of the world) for people to settle grievances on their own. There is reason to think that no one ever enjoyed living in a society where killing was common, or where anyone could be pressured into dueling. The trend toward less killing and less dueling was a matter of people becoming better at protecting their own interests, not recognizing objective moral truth.

[4] The word "murder" by definition refers to killing that is counter-normative and is viewed negatively by the moral community in question. Presumably Huemer is referring to the various types of extrajudicial killing, about which people may have a range of positive or negative attitudes.

Evidence suggests that murder rates fell precipitously as soon as people made the transition from hunting and gathering to living in primitive states. For example, an analysis of pre-Columbian Native American skeletons suggests that city dwellers were much less likely to have died violently than hunter–gatherers. Only 2.7% of Incan, Aztec, and Mayan skeletons showed signs of violent trauma compared with 13.4% of hunter–gatherer skeletons (Pinker 2011: 51). We do not know the circumstances of these violent deaths, but it seems reasonable to surmise that the states had lower rates of what we would call "murder." Why would murder be less common in states than in hunter–gatherer societies? Is it because people who make the transition from hunting and gathering to agriculture are suddenly more likely to discover the objective moral truth that murder is wrong?—or that it is wrong to kill people under a wider range of conditions? This seems unlikely. A better explanation is probably that states, which were controlled by powerful central governments, were simply better at keeping the peace.

Over time, governments have become increasingly better at preventing crime. This has sometimes resulted in murder rates falling dramatically without there being any meaningful change in values. Since the 1990s, the murder rate in the US has decreased by about 50%. No one would claim that Americans have adopted stronger antimurder values today than they had twenty-five years ago. The reason for the decline in murder is due primarily to better policing and social controls (cf. Pinker 2011: chapter 3). (To be clear, Huemer would not deny this. The point is that it is not necessary to invoke a change in values to explain even an enormous reduction in murder rates.)

When governments are not effective at keeping the peace and protecting people's rights, so-called "cultures of honor" are liable to develop: People take the law into their own hands, not because they necessarily *want* to do so, but because they *have* to if they are to survive and flourish. Nisbett and Cohen (1996) argue that, when law enforcement

is lax, herdsmen are particularly likely to develop cultures of honor, because cattle can easily be appropriated. A culture of honor developed among the herders who settled in the American South—elements of that culture have persisted to the present day. Cultures of honor lie on a continuum, and as law enforcement becomes increasingly effective, most people readily cede responsibility for defending themselves to the government.

We can conclude that people do not and never did enjoy living in violent societies where things like murder and dueling are common. We do not need to appeal to the recognition of objective, mind-independent moral truth to explain why people support social changes that reduce violence.

When a type of stimulus is rarely encountered we tend to become more sensitive to it (see Mrug et al. 2016). As society becomes less violent, it is inevitable that our sensitivity to will increase and we will respond to violence with more and more abhorrence. In Europe in the Middle Ages people would attend public executions where criminals were tortured to death—drawn and quartered or burned at the stake, sometimes for minor crimes. In later years the methods of execution became less gruesome, and the practice of torture was completely rejected. Instead of being publicly burned, criminals were publicly hanged. As people became more sensitive to violence, they objected to public hangings, and hangings began to be conducted behind closed doors. Eventually the death penalty was completely abolished in many countries.

8 Between-Society Liberalism

As discussed, hunter–gatherers are generally low on between-society liberalism. This is reflected in their intense intergroup warfare. Ethnographic evidence suggests that warfare is at least a biyearly occurrence for 65 to 70% of hunter–gatherer groups. Ninety percent

"engage in war at least once a generation, and virtually all the rest report a cultural memory of war in the past" (Pinker 2011: 52).

Huemer (2016: 2000) argues that the "most simple and natural" explanation for why "human beings become increasingly reluctant to go to war" is "[b]ecause war is horrible"—i.e., it is an *objective moral truth* that war is horrible. An alternative explanation, however, is that war is horrible in the sense of being very unpleasant. Although it is possible to find prominent thinkers of the past celebrating war and martial values (Huemer provides quotes by Nietzsche, Henry Adams, and Emile Zola), most people who experience war do not like it. People go to war for a variety of reasons, and not only from necessity. But war itself is, in general, highly unpleasant. People do not like living in societies where they are constantly under threat of attack.

Even though people may dislike war, it can be difficult to establish peace even if peace is desired by all parties to a conflict. Hunter–gatherers can agree not to attack each other, but what will stop one side from reneging? If each side knows that the other is liable to attack in spite of a truce, they will be tempted to respond preemptively—thus they fall into the "Hobbesian trap." Therefore, high levels of warfare may not tell us much about the actual preferences of the people involved. Yanomamö society, for example, is notoriously violent—around 42% of male deaths are due to warfare (Keeley 1996: Table 6.2). Do the Yanomamö live this way because they fail to recognize that engaging in constant, brutal violence is objectively morally wrong? It seems that their beliefs about moral truth have little to do with it. Rather, they simply fell deep into the Hobbesian trap. A Yanomamö warrior explained to an anthropologist: "We are tired of fighting. We don't want to kill anymore. But the others are treacherous and cannot be trusted" (Wilson 1978/2004: 119–120; see Pinker 2011: 46).

When warring groups are brought under the control of a central authority, it becomes possible for them to lay down their arms without fear of being wiped out by their enemies. Given the assurance that a central authority will punish aggressors, people are no longer tempted to launch preemptive attacks. As Pinker notes:

> The various "paxes" that one reads about in history books—the Pax Romana, Islamica, Mongolica, Hispanica, Ottomana, Sinica, Britannica, Australiana (in New Guinea), Canadiana (in the Pacific Northwest), and Praetoriana (in South Africa)—refer to the reduction in raiding, feuding, and warfare in the territories brought under the control of an effective government. (Pinker 2011: 55)

In regard to Pax Australiana, an Auyana man of New Guinea reported: "life was better since the government had come" because "a man could now eat without looking over his shoulder and could leave his house in the morning to urinate without fear of being shot" (Thayer 2004: 140; see Pinker 2011: 55). When large states are established, fighting among small groups is forcibly suppressed to everyone's benefit. Chiefs, kings, and emperors have every reason to maintain order—to prevent villages or gangs from raiding or exterminating their neighbors. The vast majority of people welcome such law enforcement.

In recent history, multiple forces have helped reduce warfare between states. States that are connected by trading relationships have less to gain—and more to lose—by fighting each other (Pinker 2011: 165). International organizations, particularly the United Nations, have taken the role of world governments, suppressing belligerent or so-called "rogue" states. Perhaps most important, though, was the development of nuclear weapons. Large nation states armed with (or capable of developing) nuclear weapons cannot seriously fight each other without risking their mutual destruction. Serious, direct

military confrontations between major powers ended abruptly with the development of nuclear weapons.

The abandonment of warfare in many parts of the world has, predictably, been accompanied by a rejection of martial values—values that no longer serve any purpose. In order to explain the adoption of more liberal values with respect to war, we do not need to appeal to people's recognition of the moral truth that war is horrible. We need only assume that, in Mackie's words, the values that prevail in different societies "reflect ways of life" (Mackie 1977: 37).

9. Conclusion

This paper has provided a naturalistic explanation for cross-cultural convergence on liberalism. The alternative explanation is the liberal realist one, which says that people converged on liberalism because they recognized its objective correctness. If both realism and antirealism predict convergence then the fact of convergence per se does not support either metaethical view. However, if we determine that the naturalistic account is superior, it would suggest that our liberal beliefs were produced by non-truth-tracking processes and are therefore not likely to correspond to objective truth.

Acknowledgments

I am especially grateful to Guy Kahane and Andreas Mogensen for extensive feedback on multiple drafts of this paper. I received very helpful comments from Maximilian Kiener, Neven Sesardić, audiences at the University of Oxford, and two anonymous reviewers.

References

Boehm, Christopher. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior.* Cambridge, MA: Harvard University Press.

———. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* New York: Basic Books.

Bogardus, Tomas. 2016. "Only All Naturalists Should Worry About Only One Evolutionary Debunking Argument." *Ethics* 126 (3): 636–661.

Brink, David O. 1989. *Moral Realism and the Foundation of Ethics.* Cambridge, UK: Cambridge University Press.

Buchanan, Allen, and Russell Powell. 2016. "Toward a Naturalistic Theory of Moral Progress." *Ethics* 126 (4): 983–1014.

———. 2018. *The Evolution of Moral Progress: A Biocultural Theory*. New York: Oxford University Press.

Cashdan, Elizabeth A. 1980. "Egalitarianism among Hunters and Gatherers." *American Anthropologist* 82 (1): 116–120.

Clarke-Doane, Justin. 2015. "Justification and Explanation in Mathematics and Morality." In *Oxford Studies in Metaethics,* vol. 10, edited by Russ Shafer-Landau, 80–103. Oxford: Oxford University Press.

de Waal, Frans B. M. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals.* Cambridge, MA: Harvard University Press.

———. 2008. "Putting the Altruism Back into Altruism: The Evolution of Empathy." *Annual Review of Psychology* 59: 279–300.

Diamond, Jared. 1987. "The Worst Mistake in the History of the Human Race." *Discover* 8 (5): 64–66.

Du Bois, W. E. B. 1903/1986. “The Souls of Black Folk.” In *Writings,* edited by Nathan Huggins, 357–547. New York: Library of America.

Goodall, Jane. 1986. *The Chimpanzees of Gombe: Patterns of Behavior.* Cambridge, MA: Harvard University Press.

Harman, Gilbert. 1977. *The Nature of Morality: An Introduction to Ethics.* New York: Oxford University Press.

Hayden, B., M. Deal, A. Cannon, and J. Casey. 1986. “Ecological Determinants of Women’s Status among Hunter/Gatherers.” *Human Evolution* 1 (5): 449–473.

Hopster, Jeroen. 2020. “Explaining Historical Moral Convergence: The Empirical Case against Realist Intuitionism.” *Philosophical Studies* 177 (5): 1255–1273.

Huemer, Michael. 2016. “A Liberal Realist Answer to Debunking Skeptics: The Empirical Case for Realism.” *Philosophical Studies* 173 (7): 1983–2010.

Joyce, Richard. 2006. *The Evolution of Morality.* Cambridge, MA: MIT Press.

Kahane, Guy. 2011. “Evolutionary Debunking Arguments.” *Noûs* 45 (1): 103–125.

Keeley, Lawrence H. 1996. *War before Civilization: The Myth of the Peaceful Savage.* New York: Oxford University Press.

Knauft, Bruce M. 1991. “Violence and Sociality in Human Evolution.” *Current Anthropology* 32 (4): 391–409.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong.* New York: Penguin Books.

Marshall, Colin. 2018. *Compassionate Moral Realism.* Oxford: Oxford University Press.

Mrug, Sylvie, Anjana Madan, and Michael Windle. 2016. “Emotional Desensitization to Violence Contributes to Adolescents’ Violent Behavior.” *Journal of Abnormal Child Psychology* 44 (1): 75–86.

Nichols, Shaun. 2002. “On the Genealogy of Norms: A Case for the Role of Emotion in Cultural Evolution.” *Philosophy of Science* 69 (2): 234–255.

———. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment.* Oxford: Oxford University Press.

———. 2014. "Process Debunking and Ethics." *Ethics* 124 (4): 727–749.

Nisbett, Richard E., and Dov Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South.* Boulder, CO: Westview Press.

Nozick, Robert. 1981. *Philosophical Explanations.* Cambridge, MA: Harvard University Press.

Parfit, Derek. 2011. *On What Matters.* Vol. 2. Oxford: Oxford University Press.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined.* New York: Viking.

Railton, Peter. 1986. "Moral Realism." *The Philosophical Review* 95 (2): 163–207.

Rasmussen Reports. 2017. "Most See U.S. Foreign Aid as a Bad Deal for America." Retrieved from http://www.rasmussenreports.com/public_content/politics/general_politics/march_2017/most_see_u_s_foreign_aid_as_a_bad_deal_for_america

Ruse, Michael, and Edward O. Wilson. 1986. "Moral Philosophy as Applied Science." *Philosophy* 61 (236): 173–192.

Sapolsky, Robert M. 2017. *Behave: The Biology of Humans at Our Best and Worst.* New York: Penguin Press.

Sauer, Hanno. 2018. *Debunking Arguments in Ethics.* Cambridge, UK: Cambridge University Press.

Service, Elman R. 1975. *Origins of the State and Civilization: The Process of Cultural Evolution.* New York: Norton.

Sida. 2019. "Development Cooperation Budget." Retrieved from https://www.sida.se/English/About-us/Budget/

Singer, Peter. 1981. *The Expanding Circle: Ethics and Sociobiology.* New York: Farrar, Straus and Giroux.

Smith, Michael. 1994. *The Moral Problem.* Oxford: Blackwell.

Stowe, Harriet Beecher. 1852. *Uncle Tom's Cabin; or, Life among the Lowly.* Vol. 1. Boston: John P. Jewett.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127 (1): 109–166.

Tersman, Folke. 2006. *Moral Disagreement.* Cambridge, UK: Cambridge University Press.

Thayer, Bradley A. 2004. *Darwin and International Relations: On the Evolutionary Origins of War and Ethnic Conflict.* Lexington, KY: University Press of Kentucky.

Turchin, Peter. 2009. "A Theory for Formation of Large Empires." *Journal of Global History* 4 (2): 191–217.

———. 2010. "Warfare and the Evolution of Social Complexity: A Multilevel-Selection Approach." *Structure and Dynamics* 4 (3): 1–37.

Turchin, Peter, and Sergey Gavrilets. 2009. "Evolution of Complex Hierarchical Societies." *Social Evolution & History* 8 (2): 167–198.

Turiel, Elliot, Melanie Killen, and Charles C. Helwig. 1987. "Morality: Its Structure, Functions, and Vagaries." In *The Emergence of Morality in Young Children,* edited by Jerome Kagan and Sharon Lamb, 155–244. Chicago: University of Chicago Press.

Vavova, Katia. 2018. "Irrelevant Influences." *Philosophy and Phenomenological Research* 96 (1): 134–152.

Vogel, Jonathan. 1987. “Tracking, Closure, and Inductive Knowledge.” In *The Possibility of Knowledge: Nozick and His Critics,* edited by Steven Luper-Foy, 197–215. Totowa, NJ: Rowman & Littlefield.

White, Roger. 2010. “You Just Believe That Because...” *Philosophical Perspectives* 24 (1): 573–615.

Wilson, Edward O. 1978/2004. *On Human Nature: With a New Preface.* Cambridge, MA: Harvard University Press.

Realist Social Selection

How Gene–Culture Coevolution Can (but Probably Did Not) Track Mind-Independent Moral Truth

Abstract: Standard evolutionary debunking arguments (EDAs) in ethics target moral beliefs by attributing them to natural selection. According to the debunkers, natural selection does not track mind-independent moral truth, so the discovery that our moral beliefs (realistically construed) were caused by natural selection renders them unjustified. I argue that our innate moral faculty is likely not the product of natural selection, but rather *social selection*. Social selection is a kind of gene–culture coevolution driven by the enforcement of collectively agreed-upon rules. Unlike natural selection, social selection is teleological and could potentially track mind-independent moral truth by a process that I term *realist social selection*: early humans could have acquired moral knowledge via reason and enforced rules based on that knowledge, thereby creating selection pressures that drove the evolution of our innate moral faculty. Given anthropological evidence that early humans designed rules with the conscious aim of preserving individual autonomy and advancing their collective interests, realist social selection appears to be an attractive theory for moral realists. However, I propose a new EDA to show that realist social selection is unlikely to have occurred.

Keywords: evolutionary debunking arguments; moral realism; natural selection; social selection; teleology

1. Introduction

Moral realists believe that there are moral truths that are objective, mind-independent, and—according to most realists—knowable (Tersman 2006). As Shafer-Landau (2003: 15) puts it, moral truths are "stance-independent": "they obtain independently of any preferred perspective, in the sense that *the moral standards that fix the moral facts are not made true by virtue of their ratification from within any given actual or hypothetical perspective*."

Moral antirealists often appeal to "debunking arguments." Debunking arguments target a belief that *p* by purporting to show that the cause of the belief does not track the truth about *p*. There is some controversy about what it means to *not track the truth*. It could mean that we would have believed that *p* whether or not *p* is true, or the truth of *p* played no role in explaining our belief, or our belief-forming process cannot be expected to generate reliable beliefs concerning *p* (see, e.g., various approaches described in Joyce 2006; Kahane 2011; Nichols 2014; Street 2006). Standard *evolutionary* debunking arguments (EDAs) in ethics claim that our moral beliefs were caused by natural selection, and that natural selection does not track mind-independent moral truth (e.g., Joyce 2006; Ruse 1986; Ruse and Wilson 1986; Sauer 2018; Street 2006).

According to Street's (2006) influential EDA, the content of our core moral beliefs is explained by natural selection. Certain behaviors are adaptive and will therefore be favored: protecting one's own life, caring for one's own children, returning favors (when this cultivates beneficial cooperative relationships), punishing those who have inflicted deliberate harm on oneself (when this will discourage them or others from inflicting further harm), and so on. In order to prompt these adaptive behaviors, natural selection endowed us and some other animals with "basic evaluative tendencies" to judge

that such behavior is "called for." In other animals these evaluative tendencies exist only in a "proto" form. Our close relatives such as chimpanzees, who have been subject to similar selection pressures, have proto evaluative tendencies that are analogous—or homologous—to ours. Humans have the capacity to articulate our intuitions about what is "called for" with language in order to generate "full-fledged" evaluative judgments such as *it is good to return favors, criminals should be punished,* or *what George did was wrong*. But these full-fledged judgments are rooted in the evaluative tendencies—which on Street's account seem to be equivalent to our innate *moral intuitions*—that were implanted in us by natural selection simply because they increased inclusive fitness in the ancestral environment. Our moral intuitions concerning family obligations, for example, are explained by the theory of kin selection. Our intuitions about reciprocity/justice are explained by the theory of reciprocal altruism.

It is important to emphasize that, although Street (2006) herself often refers to generic "evolutionary forces," her EDA is based on the claim that our moral beliefs (or the underlying evaluative tendencies/intuitions) are explained, not by just by any evolutionary forces, but specifically by *natural selection*. (The processes described by the theories of kin selection and reciprocal altruism are forms of natural selection.) This is why, in Street's view, they do not track mind-independent moral truth. Natural selection aims only at inclusive fitness. It favors evaluative dispositions that lead their possessors to make evaluative judgments that increase inclusive fitness, regardless of whether they are true. This is not to say that natural selection never tracks *any* kind of mind-independent truth. Having true beliefs about (e.g.) the behavior of medium-sized physical objects can be adaptive for obvious reasons, so a tendency to form true beliefs about such objects would generally be favored. But in the case of *evaluative* judgments, no advantage is conferred by the fact per se that a judgment is mind-independently true

(if any evaluative judgments are true in that sense). It would be, as Street (2006: 125, 143) says, an "incredible" and "implausible" coincidence if the evaluative tendencies that promoted inclusive fitness in our species, and were therefore favored by natural selection, also happened to align with mind-independent truth.

The scientific assumptions behind her argument have not been a major point of contention in the EDA literature. Although there are a few skeptics (e.g., Deem 2016; Isserow 2019; Levy and Levy 2020; Machery and Mallon 2010; Parfit 2011: 535–538), many philosophers—including many realists—accept that natural selection has played a decisive role in shaping our core, commonsense moral beliefs in more or less the way Street describes. In short, it is often taken for granted that, even if Street got some details wrong, her basic thesis is correct, namely: "one enormous factor in shaping the content of human values has been the forces of natural selection, such that our system of evaluative judgements is thoroughly saturated with evolutionary influence" (Street 2006: 114).

This paper argues that the evolutionary explanations of our moral beliefs that feature in standard, Street-style EDAs may be mistaken in ways that undermine the arguments. Section 2 outlines the concept of gene–culture coevolution, and considers whether it poses a threat to standard EDAs. It concludes that, in many cases, gene–culture coevolution can be reduced to natural selection, and gene–culture coevolutionary theories of morality that reduce to natural selection can easily be incorporated into standard EDAs. Section 3 argues that strong evidence supports a gene–culture coevolutionary account of morality that *cannot* be reduced the natural selection: the theory of *social selection*. Social selection occurs when members of a species unconsciously direct their own evolution toward a mentally represented endpoint. Anthropological evidence suggests that early humans drove the evolution of our moral

faculty by collectively establishing rules to preserve individual autonomy and promote group interests, and imposing fitness-reducing punishments on rule violators. Section 4 argues that social selection is teleological and fundamentally different from natural selection. It raises the possibility of *realist social selection*—the theory that social selection caused us to teleologically evolve an innate moral faculty that lines up with mind-independent moral truth (i.e., social selection might be truth tracking). That is, our ancestors established rules to protect autonomy and promote their collective interests because they recognized the mind-independent moral imperative of these rules via reason. Section 5 offers a new EDA, which concludes that realist social selection is unlikely to have occurred.

2. Gene–Culture Coevolution and Tracking Mind-Independent Moral Truth

Gene–culture coevolution refers to the phenomenon whereby cultural evolution affects genetic evolution and vice versa (Durham 1991; Henrich 2016; Lewens 2015; Richerson and Boyd 2005). Some scientists have defended gene–culture coevolutionary accounts of the evolution of morality, which appear strikingly different from the theories that feature in popular EDAs such as Street's (2006). Do these theories pose a challenge for EDAs? Before considering this question, we should be clear on how *natural selection* is traditionally understood and why it is said to not track mind-independent moral truth.

Following Lewontin (1970), Sober (2000: 36) writes that "Natural selection occurs when there is heritable variation in fitness." To distinguish natural from *artificial* selection, we should add that the former is *nonteleological*. For selection to be nonteleological two conditions must be met. First, variations (mutations) are introduced *randomly*. Second, no selecting agent—such as a human breeder or a divine "intelligent

designer"—has a mental representation of an evolutionary *telos* that determines what properties confer *fitness*. There is no more perspicuous summary of the basic idea of natural selection than that given by Darwin himself:

> As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be *naturally selected*. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form. (Darwin 1859: 5)

The *heritability* requirement means that there must be a mechanism of inheritance that makes offspring resemble their parent(s). We now know that—at least in the sorts of cases that concerned Darwin—the mechanism of inheritance is DNA. But natural selection does not require DNA specifically. It requires that offspring resemble their parent(s), regardless of how this resemblance is brought about (Lewontin 1970: 1; Sober and Wilson 1998: 107).

As a matter of fact, DNA is not the only way in which organisms inherit traits from their parents. Humans and (to a very limited extent) some other animals can also inherit traits via *cultural* transmission. Culturally transmitted traits, or "cultural variants"—ideas, skills, beliefs, attitudes, values (Richerson and Boyd 2005: 63)—can be passed from parents to children by *learning* (which involves imitation, emulation, or guided discovery; Sterelny 2012). Since these heritable cultural variants can both vary and affect fitness, adaptive variants can spread due to natural selection acting on individuals or (theoretically) groups. As Richerson and Boyd (2005: 76) say, "To the extent that people acquire beliefs from their parents, natural selection acts on culture in almost exactly the same way as it does on genes."

DNA transmission is (in animals) only vertical (parents to offspring). In contrast, cultural transmission can also occur horizontally (organism to other members of its generation) and obliquely (adults to non-offspring children) (Cavalli-Sforza and Feldman 1981). A consequence of this is that the *variants themselves* can undergo Darwinian evolution in the cultural realm (Richerson and Boyd 2005: 76; cf. Dawkins 1976/2006: 245–260). That is, *variants* can be more or less fit, in the sense that they are more or less likely to be copied by other people and transmitted down the generations.

This brings us to the phenomenon of *gene–culture coevolution*. *Culture*—the panoply of cultural variants—is a fitness-relevant aspect of the enculturated organism's environment. As such it can impose selection pressures that favor certain *genetic* variants. The selection pressures acting on individuals depend in complex ways on the cultural variants of others in their society, and on the variants that they themselves adopt. Genetic adaptations make culture possible, and culture in turn creates conditions that favor new genetic adaptations, which affect culture, and so on (Richerson and Boyd 2005).

Whereas we cannot choose our DNA, we have some liberty to choose among competing cultural variants. This creates selection pressure for *the ability/tendency to identify and adopt the most fitness-promoting cultural variants in our environment*. While the variants themselves are undergoing their own Darwinian selection, the individuals who participate in the culture are subject to Darwinian selection for the tendency to choose those that are best from their own fitness perspective. According to cultural evolutionary theorists, this led us to evolve certain adaptive, innate *learning biases*. The leading cultural evolutionary theorists Richerson and Boyd have, based on a combination of modeling work and empirical evidence, identified three categories of learning biases under the headings "*Content-based (or direct) bias*," "*Frequency-based*

bias," and "*Model-based bias*." These refer, respectively, to our preference for cultural variants that are (a) intrinsically appealing (due to their perceived benefit, or because "the structure of cognition makes some variants easier to learn or remember"), (b) common or rare (we tend to conform to the majority, i.e., adopt common variants), or (c) exhibited by people with certain characteristics (we tend to copy others who are successful, prestigious, or similar to ourselves) (Richerson and Boyd 2005: Table 3.1).

Richerson and Boyd (2005: Table 3.1) observe that natural selection can result in direct "[c]hanges in the cultural composition of a population caused by the effects of holding one cultural variant rather than others. The natural selection of cultural variants can occur at individual or group levels." As noted, insofar as people acquire cultural variants vertically, natural selection drives cultural evolution in essentially the same way it drives genetic evolution. The way natural selection acts on culture can be different insofar as cultural and genetic transmission are disanalogous: culture can be transmitted horizontally and obliquely (Cavalli-Sforza and Feldman 1981), variants do not necessarily arise randomly (Lewens 2015: 33), cultural practices can be tested and abandoned during an individual's lifetime (Eibl-Eibesfeldt 1989: 16), and there could potentially be other differences (see Sterelny 2006). But when individuals or groups survive or perish because of the transmittable cultural variants they hold, this is still natural selection (Eibl-Eibesfeldt 1989: 11–12, 16; Richerson and Boyd 1984; 2005: 4, 13–14, 68–79; Rogers and Ehrlich 2008; Shea 2012: 2240; Sober and Wilson 1998: 149–154). In such cases, individuals or groups vary in fitness due to traits that are heritable (offspring tend to resemble their parents), although the individuals/groups in question need not have inherited the traits vertically.

Now we can consider whether the phenomenon of cultural evolution, or the processes of gene–culture coevolution outlined above, has implications for EDAs in ethics.

Standard EDAs typically claim that natural selection influenced our moral beliefs by acting on genetically transmitted traits. One might suppose that EDAs lose their force if the traits in question are *culturally* transmitted. Parfit appears to advocate this view in the following passage:

> When Street and others claim that our normative beliefs were mostly produced by evolutionary forces, these writers are in part referring to cultural evolution. Some normative beliefs became more widely spread when and because communities of people with these beliefs were more likely to be successful. It is much less clear how we should assess the claim that certain normative beliefs were in this way, not *reproductively*, but *socially* or *culturally* advantageous. It is less clear, for example, whether and how such explanations of our normative beliefs should be assumed to *debunk* or undermine these beliefs. When the acceptance of certain normative beliefs made some community or culture more likely to survive and flourish, this fact does not as such cast doubt on the truth or plausibility of these beliefs. Such explanations of our normative beliefs do not obviously, in Street's phrase, *contaminate* these beliefs. (Parfit 2011: 537)

If, when Parfit refers to moral beliefs that proliferate because they make a community "likely to survive and flourish," he has in mind something like natural selection, then his reasoning seems to be based on a straightforward mistake. *If* explaining a moral belief in terms of natural selection debunks it, then it does not matter what the underlying mechanism of transmission is—whether it is genetic, epigenetic, cultural, Lamarckian, or anything else we could imagine. As discussed above, natural selection is not defined by genes or DNA. The reason that natural selection—whether it acts on individuals or groups—is said to not track mind-independent moral truth is not because the mechanism for transmitting moral beliefs is genetic. Parfit says that "[w]hen the acceptance of certain normative beliefs made some community or culture more likely to survive and flourish,

this fact does not as such cast doubt on the truth or plausibility of these beliefs." This statement is correct in a limited sense, but it misses the point of the EDA. If the causal explanation of our moral belief that *p* is that societies that believed that *p* were more likely to survive and flourish (i.e., were favored by natural selection), then our belief is just as "contaminated" as any other moral belief explained by natural selection.

Parfit's claim that when Street and other debunkers refer to "evolutionary forces" they are "in part referring to cultural evolution" is probably wrong, and is certainly wrong with respect to Street. (Street [2006: 118] makes it clear that she is talking about natural selection acting on traits that "must be genetically heritable.") But that is not important. The possibility that natural selection acts on culturally as well as genetically transmitted moral beliefs is easily accommodated by standard EDAs.

What about the phenomenon of gene–culture coevolution? Does the fact that gene frequencies change in response to cultural change and vice versa pose a challenge for EDAs?

The apparent fact that we evolved innate learning biases to acquire adaptive cultural variants is consistent with the claim, which is the basis of standard EDAs, that our moral beliefs were shaped by natural selection. However, cultural evolutionists in the Boyd-and-Richerson tradition have developed mathematical models of the evolution of morality, which suggest that morality—or at least certain core elements of it—evolved in a somewhat different way than Street and other debunkers have proposed.

Cultural evolutionary models assume that cultural variants are distributed in a population, and individuals with the learning biases mentioned above choose whom to imitate. Modeling work shows that, in populations of social learners, stable behavioral patterns emerge as a "by-product" of the learning biases, particularly when the conformist bias is combined with punishment of deviants (Boyd and Richerson 1992;

Chudek et al. 2013: 442). A key assumption of these models is that stable behavioral patterns, or norms, are not the product of conscious design. Groups vary in their initial distributions of norms, and individuals choose whom to imitate based on their learning biases, not based on an envisioned endpoint. Cultural evolutionary theorists in this tradition emphasize their view that people are simply unable to anticipate the consequences of different norms, so attempts to deliberately design a system of group-beneficial norms is futile (e.g., Henrich 2006). Many norm equilibria are possible, and the vast majority are *detrimental*, not beneficial, to group fitness (Boyd and Richerson 1992; Chudek and Henrich 2011: 222; Henrich and Boyd 2001: 86). So how do group beneficial norms ever evolve? According to the modelers, a small number of groups will randomly stumble on group-beneficial equilibria—namely, prosocial norms enforced by punishment—and these practices will be favored by cultural group selection (Boyd and Richerson 2002). After group-beneficial prosocial norms spread by cultural group selection, in a process of gene–culture coevolution people become genetically adapted to be innately receptive to such prosocial norms (Chudek and Henrich 2011; Chudek et al. 2013; Henrich and Boyd 2001; see Cofnas 2018: 311).

According to the story given above, our innate moral intuitions are the product of three stages of nonteleological Darwinian selection. First, there is selection among competing cultural variants, namely, different *norms*. This is Darwinian—i.e., natural—selection because (according to the story) people do not choose among the variants with the aim of bringing about any particular end result. Second, there is cultural group selection, which mostly involves natural selection acting on groups. Third, natural selection acts on individuals due to selection pressures created by the fact that society enforces prosocial norms.

Street-style EDAs assume that our moral beliefs are the product of natural selection acting on traits that (in Street's words, as quoted above) "must be genetically heritable." The Boyd-and-Richerson account of morality suggests that moral evolution was driven in large part by natural selection acting on culturally transmitted beliefs, which only later become genetically incorporated via gene–culture coevolution. This might seem to pose a problem for EDAs, but in fact it does not. The epistemic premise of standard EDAs is that natural selection does not track mind-independent moral truth. As we have seen, natural selection does not depend on any particular mechanism of transmission. The three processes that feature in Boyd and Richerson's account are all cases of natural selection, and therefore not truth tracking. Their theory about the origin of morality might conflict with Street's, but not in a way that undermines her debunking argument. On the cultural-evolutionary modelers' account, our moral intuitions were produced by natural selection, and therefore it would be an "incredible coincidence" (to use Street's expression) if they happened to align with mind-independent moral truth.

3. The Social-Selection Account of Morality

The previous section argued that gene–culture coevolution does not necessarily pose a threat to standard EDAs, which are predicated on the epistemic principle that natural selection does not track mind-independent moral truth. The processes described in mainstream cultural evolutionary theory—selection for adaptive learning biases, cultural group selection, biased transmission, etc.—are just special kinds of natural selection.

This section describes Boehm's (1999, 2012) account of the evolution of morality, which (it shall be argued in the following section) *does* pose a genuine threat to standard EDAs. According to Boehm's theory, morality evolved by a process of gene–

culture coevolution that he calls "social selection": our ancestors *consciously* designed blueprints of desirable societies, enforced behavior demanded by the blueprint, and in this way *artificially* selected people to have certain moral intuitions. Social selection is teleological in a way that makes it fundamentally different from natural selection. Moral beliefs produced by social selection cannot be undermined—at least not directly—by EDAs predicated on the idea that *natural* selection does not track mind-independent moral truth.

Regarding the evolutionary psychological theories upon which her EDA is based, Street (2006: 113) writes: "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." For the purposes of the present paper we can take the same attitude toward the social-selection theory of morality. The details are uncertain, but, in light of the evidence discussed below, the key claim is highly plausible. Human societies have been organized around deliberately engineered social codes for an evolutionarily significant amount of time. For thousands of generations, individuals' fitness was to some degree tied to their ability to conform to these codes. This created *teleological* selection pressures favoring individuals who could internalize the (explicitly represented) norms of their society—that is, who had a *conscience*. Contra Boyd and Richerson, these norms were not the product of natural selection, but were deliberately created by agents to promote their collective goals. The present paper will not try to adjudicate between the accounts of morality offered by evolutionary psychology, cultural evolutionary theory, and the theory of social selection, but some (nondefinitive) reasons for favoring the last are suggested (cf. Cofnas 2018). Insofar as the social-selection account of morality is convincing, Street's certainty in the evolutionary psychological account would seem to have been misplaced.

Boehm's (1999, 2012) theory goes as follows. Until around 250,000 years ago, hominin society resembled that of chimpanzees (*Pan troglodytes*), with each group being dominated by an alpha male and his coalition partner(s). Like chimps, subordinate hominins sometimes formed coalitions to challenge the alpha, and they sometimes ganged up on him as an entire group to prevent an instance of abuse or even depose him. But although one alpha might be dethroned, a new one would always take his place—the social system itself remained stable. Alphas enforced rules designed almost entirely to benefit themselves, and were often resented, but their subordinates lacked the ability and/or inspiration to change the system. At some point, however, there was a revolution. Our ancestors evolved the requisite levels of intelligence, foresight, and ability to communicate and cooperate, which allowed them to conceive and collectively execute a plan to redesign their political system. Subordinate males banded together to overthrow their alphas and establish a "reverse dominance hierarchy" with themselves occupying the leadership position. Instead of an alpha ruling for his own benefit, the new coalitions of the majority ruled for their *collective* benefit. They designed an explicit "blueprint" (Boehm 1999: 12, 193–194) for a desirable society, which called for extreme egalitarianism (at least among adult men), measures to protect individual autonomy, and various prosocial behaviors. They punished those who failed to live up to the demands of the blueprint, and rewarded those who did, in ways that affected fitness.

"[S]ocial control by groups" was "initially nonmoralistic" (Boehm 2012: 15), but it created unique selection pressures in the human lineage that led to the evolution of a *conscience*, which Boehm (2012: 113) defines as the tendency to "*personally [identify] with community values*, which means internalizing your group's rules."[5] At first, people

[5] Under certain conditions people's consciences can lead them to *reject* their group's rules. But the evolutionary function of the conscience is to prompt conformity. For most

followed group rules for the same reason they previously followed rules imposed by tyrannical alphas, namely, to avoid punishment. When rules are made and enforced only by an alpha (as among chimps), they tend to be followed only when the alpha is watching or otherwise likely to discover deviance. When rules were made to serve the collective interests of the majority of a group, however, almost all members had a stake in enforcing them, and violations witnessed by one person could be made public to everyone by means of language. Many generations of highly effective law enforcement favored individuals who were particularly inclined to follow group rules, and the evolutionary function of the conscience is to compel us to follow group rules—or at least think very carefully before transgressing (Boehm 2012: 114–115). In Boehm's (2012) words: "prehistorically humans began to make use of social control so intensively that individuals who were better at inhibiting their own antisocial tendencies, either through fear of punishment or through absorbing and identifying with their group's rules, gained superior fitness" (p. 17). There was likely a positive feedback loop at work. As people evolved stronger consciences, "group punishment increasingly became driven by moral outrage" (p. 88), and consequently more intense and effective. In addition to negative social selection (punishment), rule conformists could also have been favored as marriage or exchange partners.

The present paper cannot do justice to all the evidence from multiple sources that Boehm provides to support his theory, but the following is a brief summary of some highlights. The primary empirical evidence comes from studies of twentieth-century nomadic foragers who lived more or less as our ancestors in the Pleistocene. These

---

of our evolutionary history (including to some extent today), seriously opposing the values of one's community would have elicited swift, fitness-reducing punishments.

societies are marked by striking commonalities. Without exception,[6] from Australia to South America to Africa, they all enforce strict political egalitarianism among men. Some nomadic foragers have nominal leaders, but neither the leader nor any other adult male is allowed to issue direct orders to another adult male. "These foragers very predictably share a core of moral beliefs with an egalitarian emphasis on every hunter's being a political equal, while the political positions of women as nonhunters are much more subject to diversity" (Boehm 2012: 80). They also have rules requiring that food, particularly meat from large game, be shared among all group members. They are *consciously aware* that their rules have the effect of preserving individual autonomy and protecting people from bad luck (e.g., an unsuccessful hunt). Nomadic foragers all over the world employ the death penalty against recalcitrant rule breakers—almost always men—and especially against men exhibiting alpha-like behavior (Boehm 2012: Table 1). Approximately half of documented death penalty cases across these societies involved males who were

> *intimidating* their groups...by greedily or maliciously using supernatural power to seriously threaten the welfare or lives of others; by being far too ready to kill, repeatedly, out of greed or anger; by otherwise managing to seriously dominate others; or (much more rarely) by being aggressively insane. (p. 85)

[6] Some critics of Boehm have pointed to examples of hunter–gatherer societies that are not egalitarian. However, these hunter–gatherers are sedentary and therefore not *nomadic* foragers. Wengrow and Graeber (2015) do not deny that nomadic foragers were egalitarian, but they "propose a relationship between seasonality and the conscious reversal of political structures" (p. 600). According to their theory, some Upper Pleistocene hunter–gatherers alternated seasonally between nomadic and sedentary social systems. During the nomadic phase they were egalitarian, and during the sedentary phase they were (weakly) hierarchical. But even if some Upper Paleolithic hunter–gatherers sometimes deviated from nomadic foraging and strict egalitarianism, this would not undermine the claim that social selection was a decisive factor in moral evolution, particularly in the beginning. At the time our moral faculty initially evolved—and for a substantial period of time afterwards—our ancestors presumably *were* nomadic foragers.

Other crimes inviting capital punishment included theft, cheating in the context of meat sharing, taboo violations, and proscribed sexual behavior. However, the death penalty is a rare last resort. Most instances of rule violation are dealt with by less serious (but escalating) punishments ranging from teasing to full-blown shaming and ostracism, and to expulsion from the group (which, under certain circumstances, can be a de facto death sentence).

Boehm dates the origin of the reverse dominance hierarchy and egalitarianism to around 250,000 years ago, as this was when our ancestors began relying on big game as a primary source of food. In his view, this way of life would not have been possible unless all hunters were relatively well fed and satisfied, which meant meat could not have been hoarded by dominant individuals as it is in chimps. Boehm's speculation has some support from a study of cut marks on the bones of butchered animals in the Middle East. The cut marks on bones from 400,000 years ago are "relatively chaotic" compared to those from 200,000 years ago (Stiner et al. 2009). The former seem to have been made by people cutting from many different directions, while the latter seem to have been made by one person from one position. This suggests that meat distribution 400,000 years ago may have been carried out as it is in modern chimps, with an alpha hoarding the lion's share for himself while allowing his allies to grab something for themselves. In contrast, 200,000 years ago, when big game was a staple food, one person was distributing meat in an organized way, as is done among contemporary foragers (Boehm 2012: 159–160).

Street (2006) believes that our moral intuitions are built on "'proto' forms of evaluative judgement" shared with animals (p. 119). She says: "We may view many of our evaluative judgements as conscious, reflective endorsements of more basic evaluative tendencies that we share with other animals" (p. 117). In contrast, the social-

selection account implies that there is a much larger—even a qualitative—gap between human morality and any precursors in other animals. Our moral faculty is the product of selection pressures created by generations of people deliberately engineering their societies and systematically enforcing codes of behavior. This process did not occur in any other species. After reviewing the evidence, Boehm (2012: 20–24, 116–129) concludes that no nonhuman animal gives any indication of having a conscience in the sense of being motivated to follow rules for reasons other than reward or punishment, or exhibiting *shame* in response to being caught violating a rule.

Some primatologists, however, claim that nonhuman primates possess intuitions about justice that are homologous to ours. Brosnan and de Waal (2003: 297) claim to show that brown capuchin monkeys (*Cebus apella*) display "inequity aversion," which they link to the "'sense of fairness' [that] is probably a human universal." There is a well-known video showing a capuchin angrily rejecting a cucumber when her partner received a grape (which monkeys prefer) for equal work—supposedly an expression of moral outrage (de Waal 2013). The claim that capuchins have homologous intuitions about justice conflicts with the theory that our moral faculty evolved in response to recent social selection in the hominin line—it suggests that our moral intuitions are the product of *natural* selection. But there are a number of serious problems with the idea that capuchins in Brosnan and de Waal's experiments were acting on the basis of moral intuitions about equality. For example, in the control condition, monkeys were equally likely to refuse the cucumber if they could simply observe a pile of grapes. The natural interpretation is that those in the inequity condition refused cucumbers because they saw that they were being deprived of a better reward, not because they were bothered by inequity per se (Prinz 2008: 399–400). This is a controversial issue and cannot be resolved here. But it is worth noting that the existence of homologous moral intuitions in monkeys or any

other animal is far from being established, and there are good reasons to doubt their existence (see Henrich 2004; Machery and Mallon 2010: 6–11; Prinz 2008; Wynne 2004).

4. Social Selection Is Teleological and Could Track Mind-Independent Moral Truth

4.1. Social Selection Is Teleological

Social selection is teleological and therefore cannot be a type of natural selection (which is by definition *non*teleological). Standard EDAs that target moral beliefs by attributing them to natural selection have no force against moral beliefs that are the product of social selection. Before addressing the question of how social selection could track mind-independent moral truth, we should clarify exactly how it is teleological and thus different from various forms of natural selection, some of which are superficially similar.

The first way for selection to be teleological is for mutations to be introduced nonrandomly (see Section 2). This could be due to (e.g.) a Lamarckian alchemical force driving evolution toward greater complexity, God systematically causing adaptive mutations, or genetic engineering by humans. The second way for selection to be teleological is for the selection pressures to be determined by a *telos* represented in the mind of an agent—this makes selection *artificial*. Social selection is teleological because it is a kind of artificial selection. But the distinction between natural and artificial selection is a bit subtle.

We can begin by making a distinction between *conscious* and *unconscious* selection, which cuts across the distinction between natural and artificial selection.

Conscious selection—what is normally called *breeding*—is a paradigm type of artificial (and therefore teleological) selection. It occurs when an intelligent designer with at least some minimal understanding of heritability—such as a human—chooses individuals with certain traits for propagating the next generation in order to bring about a desired end goal. Breeding is what produced pit bulls and Thoroughbreds.

Darwin (1859: 36) coined the term "unconscious selection" to describe what happens when agents preserve organisms with the most desirable characteristics (the sweetest corn, the friendliest wolf) and discard the others without conscious awareness that this will drive evolution toward a specific endpoint. In his words, it is "that which follows from men naturally preserving the most valued and destroying the less valued individuals, without any thought of altering the breed" (Darwin 1868: 193). He believed that unconscious selection is different from natural selection, listing them as separate forces (pp. 193–194). He did not seem to think there is a sharp line separating unconscious selection and breeding. He wrote:

> Unconscious selection graduates into methodical, and only extreme cases can be distinctly separated; for he who preserves a useful or perfect animal will generally breed from it with the hope of getting offspring of the same character; but as long as he has not a predetermined purpose to improve the breed, he may be said to be selecting unconsciously....[E]xcept that in the one case man acts intentionally, and in the other unintentionally, there is little difference between methodical and unconscious selection....Unconscious selection so blends into methodical that it is scarcely possible to separate them. (Darwin 1868: 193–194, 210–211)

But Darwin overlooks the important distinction between natural and artificial/teleological selection. Breeding—what he called "methodical" selection—is paradigmatically teleological. On the other hand, unconscious selection can be either teleological or nonteleological depending on whether the selecting agents impose selection pressures based on a *representation* of a *telos*. *Nonteleological* unconscious

selection occurs when the selecting agents *choose what they prefer without reference to a mentally represented paradigm or blueprint*, and as a side effect drive evolution in a certain direction. This is how wolves became dogs and various grasses initially evolved into grains. Most cases of sexual selection also fall in this category. An individual sexually selects a mate by choosing the one it prefers among the options available, but generally (and in all nonhuman cases) without having a representation of the ideal mate. Selection does not cease to be natural merely because it depends on the actions or choices of agents—many paradigm cases of natural selection involve selection pressures created by agents (e.g., lions create selection pressures for gazelles by chasing them). And selection does not cease to be natural merely because it is *predictable*—many paradigm cases of natural selection have predictable outcomes (e.g., predation by lions predictably leads to populations of fleeter gazelles). Nonteleological unconscious selection may be a special type of natural selection, but it is natural nonetheless.[7] In contrast, *teleological* unconscious selection occurs when agents *choose what conforms most to a mentally represented paradigm or blueprint*, and so drive evolution to conform to the paradigm or blueprint. It is teleological for the same reason as breeding, namely, the future endpoint is represented at the time of selection.

Social selection is—at least in the vast majority of cases—a kind of teleological unconscious selection. It occurs when agents in a society explicitly represent a desired social arrangement and create selection pressures (rewards and punishments) that lead to the evolution of populations of individuals who are more disposed to follow the necessary

[7] Some evolutionary theorists, including Darwin (1859: 88) and Mayr (2001: 137–138), distinguish sexual from natural selection. As far as EDAs are concerned, the important point is that *under normal circumstances* sexual and natural selection share the essential property of being nonteleological. Because it is possible for humans to choose a mate based on a mentally represented paradigm, sexual selection could be teleological, and could in fact be a (probably minor) component of social selection.

rules. We should not rule out the possibility that some nomadic foragers occasionally had a vague idea that, in killing a recalcitrant rule violator, they were ending his (potentially) evil lineage. Such cases, if they ever occurred, were presumably atypical. But if they did then social selection would have taken the form of (teleological) *breeding*.

4.2. Realist Social Selection

In cases of social selection, a population evolves (teleologically) toward an outcome represented in the minds of the selecting agents. In principle, it could track mind-independent moral truth very simply. The selecting agents merely have to enforce behavior in conformity with a representation of morality formed in response to their recognition of the truth. In other words, if our ancestors acquired moral knowledge via *reason* and imposed fitness-reducing punishments on those who refused to behave morally or espouse moral views, we would have been selected to have a moral faculty in line with mind-independent morality. We can call this the *realist social selection* theory.

The possibility of realist social selection should be distinguished from a different possibility raised by Mogensen (2015). He suggests that our moral beliefs are the product of natural selection, but their *proximate* cause is our recognition of moral truth. Mogensen makes the following point. Some advocates of EDAs claim that (a) our moral beliefs are the product of natural selection, (b) selection aims at fitness, and does not track mind-independent moral truth, (c) therefore mind-independent moral truth cannot figure into the explanation of our moral beliefs. The reasoning from (a) to (c) is based on a mistake, namely, a confusion of ultimate and proximate causes. The fact that we have been naturally selected to have trait T to perform adaptive function F tells us nothing about the ontogeny of T—*how* we acquire T in our lifetime. That is, T's *ultimate* cause

(selection to F) does not determine its *proximate* cause. Evolutionary debunkers like Street say that the belief that *we ought to return favors* is explained (ultimately) by the theory of kin selection. But what is the proximate cause of that belief? It could be that selection equipped us to acquire the belief that *we ought to return favors* by recognizing the mind-independent truth of this proposition. (It would still be a "coincidence" that the same behaviors were both adaptive and mind-independently morally right, although in Mogensen's view this does not undermine realism.)

Suppose Mogensen's idea is right: natural selection favored a disposition to perceive moral truth because holding true moral beliefs happens to be adaptive from a fitness perspective. Even so, natural selection would not *track* moral truth. The belief that (e.g.) *we ought to return favors* is adaptive for reasons described by the theory of reciprocal altruism, not because it is moral per se. We could have been led to acquire this belief by any number of proximate causes—for example, natural selection could have endowed us with innate moral intuitions leading us to acquire the belief without needing to perceive its mind-independent truth. This scenario is different from realist social selection. In realist social selection, selection tracks mind-independent moral truth *qua truth*. There is a direct causal link from objective moral truth → perception of the truth → the imposition of selection pressures. It is not a "coincidence" that realist social selection selects for true moral beliefs. Whereas the process described by Mogensen is (nonteleological) natural selection—again, it is just a coincidence that it is adaptive to hold true moral beliefs—realist social selection is teleological for the reasons described above.

A *coincidence* provides prima facie support for a theory that we regard as having a low prior probability (Hopster 2019: 263). As noted above, Mogensen does not think it is necessarily a problem for realists to explain the correspondence between adaptive

evaluative beliefs and mind-independent moral truth as a mere coincidence. As he says: "There is...no general ban on believing in very surprising coincidences, provided that we have some evidence for their occurrence" (Mogensen 2015: 1811). But the idea that realists can dissolve Street's (2006) "Darwinian dilemma" by appealing to this principle is controversial (Hopster 2019), and would even be resisted by many realists. Without getting into the details of the controversy, it is not clear that the alleged correspondence between adaptive evaluative beliefs and mind-independent moral truth has the features usually associated with mere coincidences (see Hopster 2019). Realists who accept that our moral faculty was (largely) the product of natural selection generally attempt to find ways to show that the appearance of coincidence is illusory (e.g., Enoch 2010; Skarsaune 2011; Wielenberg 2010). It at least seems reasonable to suppose that, ceteris paribus, explanations that do not posit big coincidences are preferable to those that do. The theory of realist social selection may be a more attractive option for realists than the Mogensen scenario, as it allows them to provide a compelling evolutionary explanation of our moral faculty without having to assume a surprising coincidence about how selection pressures favored moral dispositions in line with mind-independent moral truth.

5. A New Evolutionary Debunking Argument

To reiterate, the social selection theory of morality says that group social control was initially nonmoral. Early humans who lacked a conscience but possessed high levels of general intelligence and foresight, and the ability to consider their own and others' well-being from a neutral point of view, designed and enforced rules to preserve individual autonomy and advance their collective interests. In doing this, they acted *exactly as (most) moral realists would expect*. Moral realists have a range of views about mind-

independent morality, but almost all would agree with two basic claims: (1) morality dictates that we should respect people's interests and autonomy impartially and (2) moral truth is epistemically accessible via the cognitive abilities listed above. From the perspective of the realist, it sounds like early humans grasped some key elements of moral truth and enforced rules in accordance with that truth, which caused us to evolve our innate moral faculty. In other words, our moral faculty is the product of realist social selection.

It is true that the moral codes of nomadic foragers did not impartially recognize the moral equality of *all* people (let alone conscious beings). They varied greatly in their conception of women's rights, and frequently deemed outgroup members to have a much lesser moral status (if any at all). But this is not a serious problem for the realist. There is no reason to expect nomadic foragers to get everything right. Realist social selection only requires that early humans grasped some basic principles of morality—however imperfectly—and enforced rules in light of this knowledge. That is all that is needed for realist social selection.

As discussed, if we accept social selection as the best explanation of our innate moral faculty, this would undermine standard EDAs because social selection is not natural selection. If evolutionary debunking is to succeed, a new EDA needs to be formulated to target beliefs generated by a moral faculty that evolved due to social selection. It needs to show that a causal factor(s) that had decisive influence over the social-selection process was not truth tracking, and therefore, in light of this discovery, the beliefs generated by our moral faculty are unjustified.

This can be done by pushing standard EDAs back a step. Instead of targeting the moral faculty by attributing it to natural selection, evolutionary debunkers can target the *motivation* of the social selectors. The source of epistemic contamination was not natural

selection acting directly on our moral faculty (as in standard EDAs). Rather, natural selection determined the desires and dispositions of the social selectors, and thus drove the evolution of our moral faculty *indirectly*.

Consider the process that led to the evolution of our conscience (according to the theory of social selection). It began when people who were being exploited by a tyrannical alpha banded together to depose him. The underlying motivation is readily explained by natural selection. Because the individual fitness of primates is highly correlated with their place in the dominance hierarchy, natural selection endowed them with a dislike of being dominated in order to motivate them to ascend the hierarchy if possible (Boehm 1999: 170, 237–239; Eibl-Eibesfeldt 1989: 297). Some nonhuman primates, such as chimps, can express their resentment against alphas by forming coalitions of a few individuals to attack him, or ganging up on him as a larger group. The difference between human and nonhuman primates is that our ancestors evolved the necessary intellectual and psychological capacities to do something *permanent* about alpha tyranny. Our ancestors were motivated to depose the alpha by their naturally evolved primate resentment of being dominated.

When coalitions of the majority of men overthrew their alphas, each man realized that he was unlikely to become an alpha himself, and even if he did he would likely meet the same fate as the previous one. It was in the narrow self-interest of each individual in the coalition for them to rule as a society of equals—to agree that *no one* would be subordinate to anyone. In this way, they were using their cognitive abilities not to acquire knowledge of mind-independent moral truth, but to satisfy their naturally selected primate desires.

Coalitions of the majority also imposed rules demanding that certain foods (particularly meat) should be shared more or less equally. This served as both an

insurance policy and a means to prevent individual hoarding that could lead to significant social inequalities and alpha-like behavior. Everyone knew that they themselves would sometimes be unlucky, and in the long run everyone would benefit from a system that ensured as much as possible that everyone would be taken care of. Our ancestors who instituted food-sharing rules were acting on self-interest, which of course can be explained by natural selection.

Most of the other rules established by our ancestors can similarly be explained by the pursuit of self-interest. Prohibitions against thievery, lying, and cheating all benefit the group as a whole, and there is reason to think that our ancestors were consciously aware of this fact. Boehm (2012: 52) describes nomadic foragers as "intuitive applied sociologists who are purposefully trying to shape their society in ways that will help themselves because everyone's life is helped by better cooperation."

According to this debunking account, although social selection is not natural selection, the social-selection regime was determined by our ancestors' desires which were *in turn* caused by natural selection. It would be an implausible coincidence if natural selection endowed our lineage with desires to impose rules that just so happened to correspond to mind-independent moral truth. Therefore, social selection driven by naturally selected motivations does not track mind-independent moral truth. Our moral faculty—and the evaluative judgments it generates—are, to use Street's (2006: 114) expression, "thoroughly saturated" with the influence of natural selection, but the influence is *indirect*. Natural selection explains why *social* selection took the course that it did.

A similar strategy of pushing EDAs back a step is employed by Kahane (2014), and his EDA can supplement the one proposed here against realist social selection. He was responding to de Lazari-Radek and Singer's (2012) defense of utilitarianism. De

Lazari-Radek and Singer argue that belief in the utilitarian principle of universal benevolence is uniquely immune to evolutionary debunking, therefore we should accept utilitarianism. They acknowledge, however, that the principle of universal benevolence requires a theory of well-being in order to tell us what benevolence consists it—without such a theory, the principle is "empty of content" (p. 23). Kahane (2014) counters that our ideas about well-being appear to be strongly determined by natural selection. If utilitarians accept the epistemic premise that natural selection does not track mind-independent moral truth, they have to acknowledge that our beliefs about well-being are unjustified, which (for all practical purposes) renders utilitarianism contentless. In a parallel way, the theory of realist social selection assumes that nomadic foragers were acting on a mind-independently correct theory of well-being when they established rules to promote their collective interests. But their views about well-being—namely, that it consists in individual autonomy, protection from attack, freedom from being hungry, and so on—have compelling natural-selection-based explanations, as discussed above.

5.1. Why Favor the Debunking Account?

FitzPatrick (2015: 893) argues that (standard) EDAs are "question-begging." He says that debunkers cannot defeat realism "simply by proposing a story that, *if true*, would cause problems for realism, and then claim that simply because of greater parsimony we should accept it as true." A version of FitzPatrick's objection could be applied to the EDA proposed here. We have two mutually exclusive theories that fit the facts, viz., realist social selection and the debunking account. Both theories predict that, having evolved the necessary intellectual and psychological capacities, hominins would overthrow their alpha tyrants, establish egalitarian political systems to preserve

individual autonomy, and institute rules to promote collective well-being. Why should we prefer the debunking explanation? Why assume that our ancestors were acting on their naturally selected desires rather than knowledge of mind-independent morality?

But the FitzPatrick objection does not have the same force against the social-selection-based EDA as it has against standard EDAs. FitzPatrick's original point was that the evolutionary explanations upon which standard EDAs are predicated are highly speculative. Only prior metaethical commitment to antirealism would justify favoring the speculative debunking theory over equally speculative realist accounts of the etiology of our moral beliefs. The situation is different when we choose between realist social selection and the debunking account of social selection. Realists who reject *standard* EDAs like Street's can deny the entire evolutionary story upon which they are based. But advocates of realist social selection must agree with the debunkers about the basic evolutionary story. They differ only in what motivations they attribute to the social selectors. Unless realists deny that our naturally selected desires would have prompted exactly the same behavior predicted by realist social selection, they cannot avoid postulating a big coincidence. The debunking account is not preferable just because it is more *ontologically* parsimonious (in that it does not postulate mind-independent moral truth). The debunking account explains why social selection unfolded *as if* it was being guided by our naturally selected desires—namely, it *really was* guided by these desires. For advocates of realist social selection, this would be a huge coincidence. They must postulate not only extra ontology (mind-independent moral truth) but *also* an inexplicable coincidence.

6. Discussion

Which evolutionary account of morality we accept has profound implications for EDAs in ethics. Different theories postulate different kinds of selection forces, and each one of these forces must be assessed to determine whether or under what conditions it could track mind-independent moral truth.

This paper has considered what follows metaethically from the premise that Boehm's (1999, 2012) social-selection account of moral evolution is correct. There are several influential competing accounts of moral evolution—some of which have been discussed in this paper—and scientists are far from reaching a consensus on which is best (Bloom 2019). Arguably, however, evidence is accumulating in support of something along the lines of the social-selection account. Although they do not discuss the evolution of morality, Levine et al. (2018) provide evidence that moral reasoning (at least in some contexts) follows a contractualist logic, which fits with the predictions of social-selection theory. Curry et al. (2019) argue persuasively that categories of behavior that are relevant to cooperation are universally moralized, which is what we would expect if morality was designed to advance collective group interests.

If the theory of social selection is correct, metaethics potentially has something substantial to contribute to the *science* of the evolution of morality. The major scientific question for evolutionary anthropology is what motives our ancestors were acting on when they established and enforced social rules. Realist social selection is, in essence, a scientific hypothesis about what was motivating them. Although I have argued in favor of a debunking account, realist social selection must be acknowledged as a legitimate hypothesis. Scientists investigating the biological evolution of morality cannot remain metaethically neutral.

References

Bloom, Paul. 2019. “Comments.” *Current Anthropology* 60 (1): 59–60.

Boehm, Christopher. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior.* Cambridge, MA: Harvard University Press.

———. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* New York: Basic Books.

Boyd, Robert, and Peter J. Richerson. 1992. “Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups.” *Ethology and Sociobiology* 13 (3): 171–195.

———. 2002. “Group Beneficial Norms Can Spread Rapidly in a Structured Population.” *Journal of Theoretical Biology* 215 (3): 287–296.

Brosnan, Sarah F., and Frans B. M. de Waal. 2003. “Monkeys Reject Unequal Pay.” *Nature* 425 (6955): 297–299.

Cavalli-Sforza, L. L., and M. W. Feldman. 1981. *Cultural Transmission and Evolution: A Quantitative Approach.* Princeton, NJ: Princeton University Press.

Chudek, Maciej, and Joseph Henrich. 2011. “Culture–Gene Coevolution, Norm-Psychology and the Emergence of Human Prosociality.” *Trends in Cognitive Sciences* 15 (5): 218–226.

Chudek, Maciej, Wanying Zhao, and Joseph Henrich. 2013. “Culture–Gene Coevolution, Large-Scale Cooperation and the Shaping of Human Social Psychology.” In *Cooperation and Its Evolution,* edited by Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser, 425–457. Cambridge, MA: MIT Press.

Cofnas, Nathan. 2018. “Power in Cultural Evolution and the Spread of Prosocial Norms.” *The Quarterly Review of Biology* 93 (4): 297–318.

Curry, Oliver Scott, Daniel Austin Mullins, and Harvey Whitehouse. 2019. “Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies.” *Current Anthropology* 60 (1): 47–69.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London: John Murray.

———. 1868. *The Variation of Animals and Plants under Domestication.* Vol. 2. London: John Murray.

Dawkins, Richard. 1976/2006. *The Selfish Gene.* 30th anniversary edition. New York: Oxford University Press.

de Lazari-Radek, Katarzyna, and Peter Singer. 2012. “The Objectivity of Ethics and the Unity of Practical Reason.” *Ethics* 123 (1): 9–31.

de Waal, Frans B. M. 2013. *Two Monkeys Were Paid Unequally: Excerpt from Frans De Waal’s Ted Talk*. https://www.youtube.com/watch?v=meiU6TxysCg

Deem, Michael J. 2016. “Dehorning the Darwinian Dilemma for Normative Realism.” *Biology & Philosophy* 31 (5): 727–746.

Durham, William H. 1991. *Coevolution: Genes, Culture, and Human Diversity.* Stanford, CA: Stanford University Press.

Eibl-Eibesfeldt, Irenäus. 1989. *Human Ethology.* New York: Aldine de Gruyter.

Enoch, David. 2010. “The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It.” *Philosophical Studies* 148 (3): 413–438.

FitzPatrick, William J. 2015. “Debunking Evolutionary Debunking of Ethical Realism.” *Philosophical Studies* 172 (4): 883–904.

Henrich, Joseph. 2004. “Inequity Aversion in Capuchins?” *Nature* 428 (6979): 139.

———. 2006. “Cooperation, Punishment, and the Evolution of Human Institutions.” *Science* 312 (5770): 60–61.

———. 2016. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich, Joseph, and Robert Boyd. 2001. “Why People Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas.” *Journal of Theoretical Biology* 208 (1): 79–89.

Hopster, Jeroen. 2019. “Striking Coincidences: How Realists Should Reason About Them.” *Ratio* 32 (4): 260–274.

Isserow, Jessica. 2019. “Evolutionary Hypotheses and Moral Skepticism.” *Erkenntnis* 84 (5): 1025–1045.

Joyce, Richard. 2006. *The Evolution of Morality*. Cambridge, MA: MIT Press.

Kahane, Guy. 2011. “Evolutionary Debunking Arguments.” *Noûs* 45 (1): 103–125.

———. 2014. “Evolution and Impartiality.” *Ethics* 124 (2): 327–341.

Levine, Sydney, Max Kleiman-Weiner, Nichols Chater, Fiery Cushman, and Joshua B. Tenenbaum. 2018. “The Cognitive Mechanisms of Contractualist Moral Decision-Making.” *CogSci*.

Levy, Arnon, and Yair Levy. 2020. “Evolutionary Debunking Arguments Meet Evolutionary Science.” *Philosophy and Phenomenological Research* 100 (3): 491–509.

Lewens, Tim. 2015. *Cultural Evolution: Conceptual Challenges*. Oxford: Oxford University Press.

Lewontin, R. C. 1970. “The Units of Selection.” *Annual Review of Ecology and Systematics* 1: 1–18.

Machery, Edouard, and Ron Mallon. 2010. “Evolution of Morality.” In *The Moral Psychology Handbook,* edited by John M. Doris, 3–46. Oxford: Oxford University Press.

Mayr, Ernst. 2001. *What Evolution Is.* New York: Basic Books.

Mogensen, Andreas L. 2015. “Evolutionary Debunking Arguments and the Proximate/Ultimate Distinction.” *Analysis* 75 (2): 196–203.

Nichols, Shaun. 2014. “Process Debunking and Ethics.” *Ethics* 124 (4): 727–749.

Parfit, Derek. 2011. *On What Matters.* Vol. 2. Oxford: Oxford University Press.

Prinz, Jesse. 2008. “Is Morality Innate?” In *Moral Psychology, Vol. 1: The Evolution of Morality: Adaptations and Innateness,* edited by Walter Sinnott-Armstrong, 367–406. Cambridge, MA: MIT Press.

Richerson, Peter J., and Robert Boyd. 1984. “Natural Selection and Culture.” *BioScience* 34 (7): 430–434.

———. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: University of Chicago Press.

Rogers, Deborah S., and Paul R. Ehrlich. 2008. “Natural Selection and Cultural Rates of Change.” *Proceedings of the National Academy of Sciences, USA* 105 (9): 3416–3420.

Ruse, Michael. 1986. *Taking Darwin Seriously: A Naturalistic Approach to Philosophy.* Oxford: Blackwell.

Ruse, Michael, and Edward O. Wilson. 1986. “Moral Philosophy as Applied Science.” *Philosophy* 61 (236): 173–192.

Sauer, Hanno. 2018. *Debunking Arguments in Ethics.* Cambridge, UK: Cambridge University Press.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence.* Oxford: Oxford University Press.

Shea, Nicholas. 2012. "New Thinking, Innateness and Inherited Representation." *Philosophical Transactions of the Royal Society B* 367 (1599): 2234–2244.

Skarsaune, Knut Olav. 2011. "Darwin and Moral Realism: Survival of the Iffiest." *Philosophical Studies* 152 (2): 229–243.

Sober, Elliott. 2000. *Philosophy of Biology.* 2nd edition. Boulder, CO: Westview Press.

Sober, Elliott, and David Sloan Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge, MA: Harvard University Press.

Sterelny, Kim. 2006. "The Evolution and Evolvability of Culture." *Mind & Language* 21 (2): 137–165.

———. 2012. *The Evolved Apprentice: How Evolution Made Humans Unique.* Cambridge, MA: MIT Press.

Stiner, Mary C., Ran Barkai, and Avi Gopher. 2009. "Cooperative Hunting and Meat Sharing 400–200 Kya at Qesem Cave, Israel." *Proceedings of the National Academy of Sciences, USA* 106 (32): 13207–13212.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127 (1): 109–166.

Tersman, Folke. 2006. *Moral Disagreement.* Cambridge, UK: Cambridge University Press.

Wengrow, David, and David Graeber. 2015. "Farewell to the 'Childhood of Man': Ritual, Seasonality, and the Origins of Inequality." *Journal of the Royal Anthropological Institute* 21 (3): 597–619.

Wielenberg, Erik J. 2010. "On the Evolutionary Debunking of Morality." *Ethics* 120 (3): 441–464.

Wynne, Clive D. L. 2004. “Fair Refusal by Capuchin Monkeys.” *Nature* 428 (6979): 140.

A Debunking How-Possibly Explanation for the Principle of Universal Benevolence

Abstract: According to Street's evolutionary debunking argument (EDA), evolutionary biology provides "powerful" explanations for our "basic evaluative judgements." The discovery that our moral beliefs (realistically construed) are "saturated with evolutionary influence" renders them unjustified, since natural selection does not track mind-independent moral truth. De Lazari-Radek and Singer agree that most of our commonsense moral beliefs are debunked in the way Street claims, but they argue that belief in Sidgwick's *principle of universal benevolence* cannot be explained by natural selection and is therefore immune from EDAs. I argue that Street oversold the power of her evolutionary explanations, thus leaving an opening for realists to claim that moral beliefs with less powerful evolutionary explanations can escape debunking. In fact, all naturalistic theories of morality—including those invoked by Street and de Lazari-Radek and Singer—are speculative "how-possibly" explanations. If how-possibly explanations are *not* debunking, then both Street's (global) and de Lazari-Radek and Singer's (selective) EDAs fail. If how-possibly explanations *are* debunking, then selective debunkers must show that there is no plausible way that naturalistic forces could have produced the beliefs they want to defend. I argue that naturalistic how-possibly explanations of moral beliefs can be debunking by appealing to the value of ontological parsimony. I provide a debunking how-possibly explanation for belief in the principle of universal benevolence.

Keywords: evolutionary debunking arguments; impartiality; utilitarianism; moral realism; nihilism; parsimony

1. Introduction

According to Street's (2006) influential evolutionary debunking argument (EDA), natural selection explains our "basic evaluative judgements." Consequently, "our system of evaluative judgements is thoroughly saturated with evolutionary influence" (p. 114), i.e., the influence of natural selection. Street lists several "judgements about reasons" including the following:

(1) The fact that something would promote one's survival is a reason in favor of it.
(2) The fact that something would promote the interests of a family member is a reason to do it.
(3) We have greater obligations to help our own children than we do to help complete strangers.
(4) The fact that someone has treated one well is a reason to treat that person well in return. (p. 115)

She expounds: "Evolutionary biology offers powerful" explanations for why we make these judgements, "very roughly of the form that *these* sorts of judgements about reasons tended to promote survival and reproduction much more effectively than the alternative judgements" (p. 115). The explanation for judgement (1) she claims is "obvious": our ancestors with this "general evaluative tendency" survived and reproduced more than those with contrary tendencies. Judgements (2) and (3) are explained by the theory of kin selection, (4) by the theory of reciprocal altruism. This knowledge about the etiology of our moral beliefs (realistically construed) removes any justification we had for holding

them, because natural selection is not truth tracking[1] with respect to mind-independent moral truth.

Street (2006) presents her EDA as a challenge, not just for the specific moral beliefs for which she proposes evolutionary explanations, but as a *general* challenge for moral realism. In showing that *some* of our core moral beliefs are caused by natural selection[2] via the influence of our evolved "basic evaluative judgements," she intends to undermine the view that *any* of our moral beliefs are mind-independently true. This leaves realists with an opening. Realists can accept that many or even most of our core moral beliefs are undermined by EDAs. But if they can point to a moral belief that is not explained by natural selection, *this* belief would escape evolutionary debunking (Sauer 2018: 31–32). Since realism only requires there to be a single mind-independent moral truth, realists only need to find one moral belief that is immune from EDAs.

This is the strategy adopted by de Lazari-Radek and Singer (2012). They agree with Street that many of our commonsense moral beliefs are debunked. But they say that at least one moral belief—belief in the *principle of universal benevolence* (PUB)—could not be the product of natural selection, and therefore remains unscathed.

Sidgwick (1907) argued that there is a conflict between two "self-evident" principles of practical reason, namely, the principles of *rational egoism* and *universal benevolence*. The first says that we should act in our own best interests. The second, that

[1] For the purposes of this paper, I understand a process to be non-truth tracking if the truth of the beliefs it generates is explanatorily redundant. I.e., we do not need to assume that *p* in order to explain token beliefs that *p* (see Harman 1977; Joyce 2008: 217; Ruse and Wilson 1986: 187). This follows Street's (2006) practice. As she puts it, Darwinian explanations of our moral beliefs "make no reference to the realist's independent evaluative truths" (p. 155).

[2] Street's (2006) EDA is based on the so-called "positive" view of natural selection according to which we can explain the phenotype of individuals by appealing to past selection pressures. There is some controversy surrounding this, but that is outside the scope of the present paper. For an overview of the debate and a defense of the positive view see Birch (2012).

we should act impartially to maximize the good of all sentient beings. Sidgwick deemed this the "dualism of practical reason." De Lazari-Radek and Singer (2012) argue that the conflict can be resolved in favor of universal benevolence because the principle of rational egoism is susceptible to an EDA whereas universal benevolence is not—the former but not the latter increases fitness.[3] Since it was not natural selection that made us receptive to PUB, it must be our capacity to recognize mind-independent moral truth via reason.

Toward the end of their paper, de Lazari-Radek and Singer propose a three-step procedure for "establishing that an intuition has the highest possible degree of reliability." One must show that the following conditions hold:

1. careful reflection leading to a conviction of self-evidence;
2. independent agreement of other careful thinkers; and
3. the absence of a plausible explanation of the intuition as the outcome of an evolutionary or other non-truth-tracking process. (p. 26)

In regard to PUB, they summarize their argument as follows:

> We form the intuition as a result of a process of careful reflection that leads us to take, as Sidgwick puts it, "the point of view of the universe." This idea is not specific to any particular cultural or religious tradition. On the contrary, the leading thinkers of distinct traditions have independently reached a similar principle and have regarded it as the essence of morality. In addition to the well-known Jewish and Christian versions of the Golden Rule, we find similar ideas in the Confucian, Hindu, and Buddhist traditions. Finally, there is no plausible explanation of this principle as the direct outcome of an evolutionary process, nor is there any other obvious non-truth-tracking explanation. Like our ability to do higher mathematics, it can most plausibly be explained as the outcome of our capacity to reason. (pp. 25–26)

[3] Crisp (2012) notes that natural selection would actually favor "something like kin altruism, which is neither egoistic nor impartially benevolent." But, for the sake of argument, we can grant de Lazari-Radek and Singer their claim that belief in the principle of rational egoism can be debunked.

They do not deny that the conviction of self-evidence can be—and frequently is—produced by non-truth-tracking forces. Careful reflection led Kant (1797/2017: 6:424–425) (and many other thinkers throughout history) to conclude that homosexuality and onanism are self-evidently, objectively immoral behaviors.[4] De Lazari-Radek and Singer would presumably attribute his intuitions to non-truth-tracking cultural factors. But if careful thinkers living under a variety of cultural conditions independently converge on the same belief, and there is no debunking account of how they formed the belief, this could arguably be taken as prima facie evidence for its truth.

When de Lazari-Radek and Singer (2012: 26) explicitly outline their procedure for defending an intuition, they say (as quoted above) that one must establish "the absence of a plausible explanation of the intuition as the outcome of an evolutionary or other non-truth-tracking process." Note that they refer to "evolutionary *or other non-truth-tracking process[es]*." However, despite asserting that belief in PUB does not have any evolutionary or "other obvious non-truth-tracking explanation," they do not devote a single sentence to exploring any possible *non*obvious, non-truth-tracking processes besides evolution that could be responsible for it. In fact, in the sentence immediately following the one mentioning "evolutionary or other non-truth-tracking process[es]," they drop the reference to other non-truth-tracking processes and go back to talking only about evolution ("If the third requirement were not met—if the intuition could be explained as the outcome of an evolutionary process—that would not show the intuition

[4] Kant (1797/2017: 6:425): "That such an unnatural use (and so misuse) of one's sexual attribute is a violation of duty *to oneself*, and indeed one contrary to morality in its highest degree, occurs to everyone immediately, with the thought of it...."

to be false..."). Elsewhere, they present the evolutionary explanation and the realist explanation for belief in PUB as the only two options:

> In the absence of an appeal to our evolved capacity to reason as the basis for our ability to grasp moral truth...it is difficult to see what plausible evolutionary explanation there could be for the idea of equal concern for the interests of complete strangers [including all sentient beings] who do not belong to one's own group. (p. 20)

In practice, de Lazari-Radek and Singer conflate *evolutionary* debunking arguments with debunking arguments generally, which leads them to conclude that if a moral belief lacks a Street-style evolutionary explanation then it cannot be debunked.

I argue that the mistake in de Lazari-Radek and Singer's defense of PUB is rooted in Street's (2006) own formulation of her EDA. Street casts her evolutionary explanations as "powerful" and scientifically well established. This gives a misleading picture of the state of the science. Evolutionary and other naturalistic theories of morality are, at best, plausible but speculative *how-possibly* explanations. This raises two possibilities: such explanations of morality are either debunking or not debunking. If they are debunking, then belief in PUB (or any other moral principle) can be debunked simply by showing how it *could* plausibly be the product of non-truth-tracking forces. If they are not debunking, then both Street's and de Lazari-Radek and Singer's arguments fail, since they both rely on the same how-possibly explanations to debunk commonsense morality. I argue that how-possibly explanations *can* be debunking, and propose a debunking account of belief in PUB and its alleged precursor, the Golden Rule.

Section 2 argues that Street's evolutionary explanations of our moral beliefs are not nearly as powerful as she claims, they leave many of our beliefs unexplained, and there are no alternative naturalistic explanations of morality that meet the high bar she

demands. Section 3 proposes a new way of formulating debunking arguments relying on "how-possibly" explanations. Given the availability of plausible, naturalistic how-possibly explanations of our moral beliefs, realist explanations that appeal to non-natural moral facts should be rejected on grounds of ontological parsimony. Section 4 argues that the appearance of the Golden Rule in different traditions does not mean that "careful thinkers" independently agree about the self-evidence of PUB. Section 5 offers a debunking how-possibly explanation for why belief in PUB arose in a certain intellectual community.

2. The Empirical Premise of Street's Evolutionary Debunking Argument

The empirical premise of Street's (2006) EDA is that "our system of evaluative judgements is thoroughly saturated with evolutionary influence" (p. 114). In asserting this, she is not making the trivial claim that natural selection played a causal role in producing our evaluative judgements. (Without natural selection we—along with our moral beliefs—would not exist at all.) Rather, she is making a claim about *how* natural selection shaped the *content* of our moral beliefs, namely: natural selection favored certain evaluative judgements, which in turn give rise to "loosely corresponding" (p. 120) moral beliefs. In contrast, realists typically claim that we grasp truths in domains such as mathematics and ethics as a byproduct of our *general reasoning ability*, which was favored by natural selection to solve adaptive problems in the ancestral environment (e.g., Cuneo and Shafer-Landau 2014: 427; de Lazari-Radek and Singer 2012: 16–18; Huemer 2005: 215–216; Parfit 2011: 494–497).

This section argues for two points. First, Street's account of how natural selection shaped our moral beliefs is not nearly as "powerful" as she claims. Second, even

supposing for the sake of argument that her account is correct, it is not clear that our system of evaluative judgements is *saturated* with the influence of natural selection in the way she assumes.

2.1. Are Moral Beliefs Rooted in Naturally Selected Basic Evaluative Judgements?

Street (2006) argues that the "influence of Darwinian selection pressures on the content of human evaluative judgements is best understood as indirect" (p. 119), and we do not "automatically or inevitably accept the full-fledged evaluative judgements that line up in content with our basic evaluative tendencies" (p. 120). But, she says, if we had evolved under different conditions, our basic evaluative judgements—and consequently our "full-fledged" judgements—"would also have been very different, and in loosely corresponding ways" (p. 120). When male and female lions pair up, the males kill the cubs that were fathered by their partner's previous mate, and female lions accept this behavior (p. 120). According to Street, if we had followed an evolutionary trajectory like that of lions, we would make very different judgements about the morality of killing children. If we had evolved like social insects, our full-fledged evaluative judgements would reflect our tendency to view survival as "'good' only insofar as it was of some use to the larger community" (pp. 120–121). Other evolutionary debunkers have made similar suggestions. Ruse and Wilson (1986: 186) note that many species engage in behaviors such as cannibalism, incest, and parricide "with gusto and in order to survive." Had our lineage been subject to the same selection pressures, we would consider these things "highly moral" rather than paradigmatically immoral. "In short," they say, "ethical premises are the peculiar products of genetic history, and they can be understood solely

as mechanisms that are adaptive for the species that possess them." Darwin also proposed an argument along these lines:

> If...men were reared under precisely the same conditions as hive-bees, there can hardly be a doubt that our unmarried females would, like the worker-bees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters; and no one would think of interfering. (Darwin 1871: 73)

The theory that different selection pressures would have led us to form correspondingly different moral beliefs is questionable. Certainly, if humans had followed an evolutionary trajectory more like that of lions, males might have evolved an impulse to kill their stepchildren. But would we think it *morally right* to act on this impulse? Mothers, the biological fathers of the children, and the children themselves would be against it, for reasons that are themselves explained by natural selection. Why would we form the belief that newly paired males should act on *their* naturally selected desires at the expense of everyone else's desires? In the real world, many of our moral beliefs concern the *immorality* of acting on natural impulses, including impulses that may (to some extent) be adaptations. People believe it is morally wrong to kill our rivals, rape, steal, or engage in extreme nepotism even when such actions would promote the fitness of the perpetrator. If males had evolved a desire to kill their stepchildren, we might simply have formed the moral belief that men should curb this desire.

In fact, our species may be more like lions than Street assumes. Lowe et al. (2020) provide evidence that male chimpanzees routinely kill infants in order to make their mothers sexually receptive. Lowe et al.'s conclusion should be treated with caution. But if it is correct—and given that similar behavior seems to occur among other primates—humans could well have inherited lion-like infanticidal tendencies from the last common

ancestor of humans and chimps. The fact that no society tolerates this (possibly) naturally selected behavior directly contradicts Street's prediction.

Suppose we had a haplodiploid reproductive system like that of bees. Our ancestors would have been selected for stronger altruistic tendencies toward other members of their (highly genetically related) group. Presumably human societies would be arranged in profoundly different ways, and it seems plausible to suppose that this would have implications for our moral beliefs. But how would our core moral beliefs be different? Street (as quoted above) speculates that we would see survival as "'good' only insofar as it was of some use to the larger community." It is worth noting that this concept of "good" is not too far from our own. Being useful to the larger community is generally thought to be among the highest goods, though we do assign moral value to classes of people who are not and never will be useful in this sense. If we evolved like bees, individuals would presumably have greater personal concern for the common good. But we would also be capable of reflecting on these impulses, and potentially modifying our values. It is not entirely obvious if or how our moral beliefs would be different. (Cf. Sidgwick's comments on Darwin's thought experiment about an intelligent, bee-like species, quoted in Darwin 1874: 99, n. 6.)

Similar objections can be applied (mutatis mutandis) to the examples of cannibalism and parricide.

A more general concern is that the instincts implanted in us by natural selection have a different content than the allegedly corresponding evaluative judgements. Take the judgement, "The fact that something would promote the interests of a family member is a reason to do it." Kin selection would implant in each person a disposition to *promote the interests of my kin* (or, more precisely, *promote the interests of people who exhibit cues that were historically correlated with being kin, such as having grown up in close*

proximity to me). It is not obvious that an impulse to favor *my* kin would generate a moral intuition that *everyone* should favor *their* kin (Cofnas 2020a).

The point is not that Street is necessarily wrong that our naturally selected basic evaluative judgements exert a causal influence over our moral beliefs, resulting in a loose correspondence between the former and the latter. Rather, the point is that this is highly speculative and far from being established science.

2.2. Is Morality Saturated with Evolutionary Influence?

Suppose for the sake of argument that all of Street's evolutionary explanations are correct: there was a direct causal chain from the selection pressures described by reciprocal altruism theory to our belief that reciprocity is morally good, and so on. Even granting this, it is not clear that our whole evaluative system is *saturated* with evolutionary influence (cf. Hopster 2018; Kahane 2011: 118).

Evolutionary debunkers are aware that natural selection is not the only force that has shaped the content of our moral beliefs. Natural selection could not possibly explain why moral beliefs can change rapidly, or why they sometimes vary substantially between groups that have been separated for only a few generations. There must be other factors at work. The question for metaethics is whether these factors are potentially truth tracking.

Street (2006) herself highlights the "crucial and *sui generis* influence of rational reflection" (p. 114), which can make our evaluative judgements "stray, perhaps quite far, from alignment with our more basic evaluative tendencies" (p. 120). She expounds:

> we are reflective creatures, and as such are capable of noticing any given evaluative tendency in ourselves, stepping back from it, and deciding on

> reflection to disavow it and fight against it rather than to endorse the content suggested by it. (p. 120)

On the face of it, this sounds like something a realist would say. But Street denies the realist implications, arguing that rational reflection does not provide a

> means of standing apart from our evaluative judgements, sorting through them, and gradually separating out the true ones from the false...[I]f the fund of evaluative judgements with which human reflection began was thoroughly contaminated with illegitimate [evolutionary] influence...then the tools of rational reflection were equally contaminated.... (p. 124)

In other words, all reasoning can do is elaborate on our "starting fund of evaluative judgements" (p. 124), which was implanted in us by natural selection. But why should we believe that this naturally selected starting fund is the *only* source of moral premises? Street gives six examples of basic evaluative judgements (the four mentioned above and two others), which she claims are the product of natural selection. Do *all* human moral beliefs arise through reflection on these six premises? This is far from being established. Street denies without argument that some other moral premises might be accessible via reason.

Societies across time and place often adopt opposing moral beliefs: slavery is right vs. slavery is wrong, some people have a hereditary right to rule vs. everyone is born politically equal, and so on. Are all of these opposing beliefs rooted in debunked basic evaluative judgements? Even if Street is right that natural selection endowed us with a starting fund of evaluative judgements, she does not show that these judgements contaminate *all* of our moral beliefs.

2.3. Limits of the Science of Morality

Are there evolutionary or other naturalistic theories of morality that are less speculative than Street's (2006), or that could potentially explain a wider range of our moral beliefs? In short: no. For several decades, morality has been a central area of research in theoretical biology, psychology, anthropology, sociology, philosophy, and, more recently, neuroscience. Although there has been plenty of interesting empirical and theoretical work, we are far from having definitive answers. Consider Bloom's remarks about the state of the science:

> Morality is ultimately about fairness and justice. Or it is about maximizing the welfare of sentient beings. Or intertemporal choice, giving up immediate satisfaction for long-term gain. Or maybe it is really all about altruism. Morality has one foundation—harm. Or it has three foundations: autonomy (which includes harm), but also divinity and community. Or five foundations: care, fairness, loyalty, authority, and purity. Or perhaps six—do not forget about liberty. Morality is innate and universal, a modular system of the sort proposed by Chomsky and Fodor for language. Or it is innate and universal, but nothing like a modular system. Or it is partially innate and partially learned. Or entirely learned. At least we know that morality is a distinct domain, a cognitive natural kind. Unless it is not. (Bloom 2019: 59)

Given that scientists have not been able to agree on what morality even *is*, we are clearly far from being able to determine the role natural selection played in shaping the content of our moral beliefs.[5] This is not to deny that the science of morality has made great progress. But scientists are still debating fundamental issues. More important from the

[5] The point is not that scientists cannot agree on a *definition* of morality, but that they cannot agree on its *nature*. Compare the debate about morality with the debate about how to define *gene*. Different scientists prefer different gene concepts, but this is largely a semantic issue about how to describe an underlying reality (i.e., DNA) that is well understood. When it comes to our moral faculty, the underlying reality is not well understood at all.

perspective of EDAs, none of the major theories even purports to offer a complete explanation for all of our moral beliefs.

Take the highly influential moral foundations theory (MFT), which Bloom alludes to in the quote above. According to MFT, there are (at least) five "foundations of intuitive ethics": "harm/care, fairness/reciprocity, in-group/loyalty, authority/respect, and purity/sanctity" (Haidt and Joseph 2007: 381). Haidt and Joseph (2007: 381) describe these foundations as "sets of concerns," each one "linked to an adaptive challenge and to one or more moral emotions." Our innate "concerns" do not *determine* our moral beliefs. Rather, they provide the emotional motivation—or "preparedness" (p. 381)—to moralize certain domains and not others. "MFT assumes that nothing is hardwired or insensitive to influence" (Haidt and Joseph 2011: 2121). "[C]lasses of social concerns are likely to become moralized during development. Social issues that cannot be related to one of the foundations are much harder to teach, or to inspire people to care about" (Haidt and Joseph 2007: 381). MFT does not purport to explain the specific *content* of our moral beliefs. Moral foundations theorists say that "cultures build incommensurable moralities on top of a foundation of shared intuitions" (Haidt and Joseph 2004: 54), and "cultures create moralities that are unique yet constrained in their variations" (Haidt and Joseph 2007: 381). Even if this is true—and in some general sense it is highly plausible (cf. Prinz 2007: chapter 7)—there is a lot left unexplained about the etiology of our moral beliefs. How do we choose among the wide range of possible and incompatible moral beliefs related to vague "sets of concern" like "harm/care"?

There are a number of other interesting scientific theories of morality, and it is impossible to review them all. The upshot is that no theory comes close to providing naturalistic explanations for all of our core moral beliefs, or to showing that "our system of evaluative judgements is thoroughly saturated with evolutionary influence" in the way

Street (2006: 114) claims. While it seems plausible to think that our naturally selected biological endowment puts some sort of constraints on morality (see Prinz 2007), how this works is largely a matter of speculation. Non-evolutionary factors clearly play a significant role in shaping many (if not all) of our beliefs. Any mainstream evolutionary theory of morality that debunkers might appeal to will leave some important moral beliefs unexplained.

3. Debunking Arguments Based on How-Possibly Explanations

Traditional debunking arguments have something like the following form:

> *Causal premise.* S's belief that p is explained by X.
> *Epistemic premise.* X is an off-track process.
> Therefore
> S's belief that p is unjustified. (Kahane 2011: 106)

The causal premise of *evolutionary* debunking arguments is a proposition that a moral belief(s) was shaped (directly or indirectly) by natural selection. But if the arguments of the previous section are right, theories of how natural selection shaped morality are highly speculative and leave many of our beliefs unexplained.

The debunker might hope to draw upon a wider range of scientific theories of morality, not just those that appeal to natural selection. Debunking arguments work just as well if the targeted beliefs can be attributed to off-track psychological, anthropological, or sociological processes. But social-scientific theories of morality are no less sketchy and speculative than evolutionary theories. There is no scientific theory from either evolutionary biology or social science that provides a genuinely powerful (by

Street's standards) explanation for all of our core moral beliefs. The state of the science is unlikely to change dramatically in the foreseeable future.

In light of the fact that "complete or even nearly complete adaptation explanations are going to be rare in evolutionary biology," Brandon (1990: 177) argues that evolutionary theorists must often settle for "an account of how some adaptation *could have* evolved." He calls these "how-possibly explanations": "potential explanations, none of whose explanatory premises contradict or conflict with 'known facts' (i.e., the thing we believe based on good evidence)" (p. 179). According to a standard view, we *explain* an event by showing that it follows from the conjunction of initial conditions and generalizations or laws. How-possibly explanations work differently. They are "based on generalizations or laws we have good reason to believe are true, but whose initial conditions are speculative" (p. 179).

Although Brandon was concerned with evolutionary biology, the social sciences frequently deal in how-possibly explanations, too. Social scientists trying to explain past events may have no choice but to speculate (to some extent) about the relevant initial conditions. Even when studying well-documented events, we might not have access to *all* of the relevant data, or have a reliable way of determining what elements of the initial conditions were causally important. In these cases, explanations will necessarily take a how-possibly form, combining speculations about initial conditions with generalizations about how individuals or groups could be expected to react to such conditions.

The best scientific accounts of our moral beliefs will in most cases be how-possibly explanations drawing on theories and data from evolutionary biology and the social sciences. Can such how-possibly explanations be debunking? How-possibly-explanation debunking arguments would take the following form:

Hypothetical causal premise. Assuming plausible initiation conditions, S's belief that p—the truth of which implies the existence of an ontologically radical phenomenon—could be explained by a known causal mechanism(s) X.

Epistemic premise. X is an off-track process.

Therefore

S's belief that p is unjustified.

Consider an example to illustrate how this kind of debunking argument works. People who have near-death experiences sometimes have visions of an afterlife in response to which they form the belief that God and an afterlife are real. As a matter of fact, scientists do not have a definitive explanation for why these experiences occur, but there are some plausible naturalistic theories (Blanke et al. 2016). A religious belief formed in response to a near-death experience can be targeted with the following how-possibly-explanation debunking argument:

Hypothetical causal premise. S's belief that God and an afterlife exist *could* be explained by hallucination-inducing neurological processes, which may be stimulated under conditions associated with near-death experiences.

Epistemic premise. Hallucination-inducing neurological processes are off-track with respect to the existence of God and an afterlife.

Therefore

S's belief in God and an afterlife is unjustified.

The metaphysical principle underlying how-possibly-explanation debunking arguments is that we should not postulate ontologically radical causal mechanisms whose reality

cannot be independently verified to explain beliefs that *could* be explained by known causal mechanisms. This can be derived from a more general (and widely accepted) principle of commonsense reasoning, namely, we should not postulate extra ontology to explain phenomena that are *consistent* with known facts and laws. Whether we should postulate new ontology to explain some observations is a function of (a) how plausible our ontologically parsimonious how-possibly explanation is in light of (what we regard as) established knowledge and (b) how ontologically radical the alternative explanation would be. If a parsimonious how-possibly explanation is plausible (in the light of established knowledge), and the alternative explanation postulates some radical ontology whose reality cannot be otherwise verified, we are especially justified in rejecting the latter.

Because how-possibly-explanation debunking arguments appeal to ontological parsimony, they can only be used to target *non-naturalist* versions of moral realism (which were also the target of Street's EDA). They provide no reason to reject naturalist versions of moral realism that do not postulate any special ontology.

3.1. The Force of the Argument from Parsimony

FitzPatrick (2015: 883) objects to EDAs on the grounds that they "beg the question against realism from the start." They *assume* that a naturalistic explanation of our moral beliefs is true, but this is the very thing that realists deny. He argues:

> The mere availability of [a naturalistic] story does not by itself actually debunk the realist alternative: it simply provides a rival account of our moral beliefs, which will succeed in its debunking ambitions only if it is actually *correct*. Those attracted to it will of course cite virtues such as greater parsimony (it explains our

> beliefs without having to appeal to real moral properties and facts), and this may contribute to their own justification for believing it, given the rest of their views and commitments. But this can hardly be expected to have offensive force against *realists*. Greater parsimony is a theoretical virtue only where the world is obligingly austere, and that is exactly what is at issue in this debate.

How-possibly-explanation debunking arguments explicitly appeal to the virtue of parsimony. As stated above, the underlying metaphysical principle is that one should not postulate ontologically radical causal mechanisms whose reality cannot be independently verified (e.g., interaction with real moral properties) to explain beliefs that *could* be explained by known causal mechanisms (e.g., natural selection). FitzPatrick would say that these debunking arguments fail because the realist can propose alternative how-possibly explanations according to which real moral properties did play a causal role in producing our moral beliefs. Which explanation one accepts will depend on "the rest of [one's] views and commitments."

But FitzPatrick's argument overlooks the fact that favoring less parsimonious hypotheses requires *justification*—and greater justification the more ontologically radical its postulates are. One cannot justify belief in extra ontology simply by referring to "the rest of [one's] views and commitments." *This* would be begging the question. Either the extra ontology should be independently verifiable, or the less parsimonious hypothesis should have considerably more explanatory power than the parsimonious one. This is where realism seems to come up short. There has been no really compelling argument for the existence of real (non-naturalist) moral properties that does not rely on mere intuition. There is no *independent* justification for the existence of real moral facts beyond the disposition to believe in them.[6] If there are plausible, naturalistic how-

[6] Although non-naturalist moral realists typically appeal to intuition, Huemer (2016) proposed an empirical argument for realism. He suggested that the best explanation for cross-cultural convergence on liberalism is that liberalism is objectively correct. This

possibly explanations for why we hold certain moral beliefs (realistically construed), the onus is on the realist to give a reason to prefer the less parsimonious explanation.

Huemer (2005) attacks parsimony-based arguments against moral realism from another angle. He describes the "argument from weirdness" (p. 199) as follows:

1. Objective values would be weird.
2. If something would be weird, then it (probably) does not exist.
3. Therefore, objective values (probably) do not exist. (p. 200)

According to Huemer, if "weird" means "very dissimilar from all or most other things that exist," then premise (2) could be rendered, "If something is very dissimilar from other things that exist, then it (probably) does not exist." Huemer rejects the argument from weirdness. He asks us to consider a "partial list of what exists," including time, space, numbers, relationships, physical states, aesthetic properties, gravitational fields, the past, and moral properties. Each one of these things is "*very* different" in kind from the others. "So, on the face of it, this version of the argument from weirdness goes nowhere" (p. 200).

But there is an important difference between things like time, space, physical states, gravitational fields, and the past, on the one hand, and aesthetic and moral properties, on the other. The former are bound together in a causally interacting ontological web, so that we have multiple routes to verify each kind of thing. To justify our beliefs about gravitational fields one has to assume the existence of time, space, the past, etc. Denying the existence of any one of these phenomena causes the entire web to collapse. *Numbers* are a more difficult case. According to one view, the utility of

was disputed by Cofnas (2020b) and Hopster (2020), who offered debunking explanations for this sociological phenomenon.

mathematics in our dealings with the physical world suggests that our beliefs about numbers correspond to some sort of mind-independent reality (Colyvan 2001). According to another view, our scientific theories do not actually require us to postulate numbers (Field 2016). If the former is right, numbers are part of the ontological web. If the latter is right, there may be no reason to believe in numbers at all. Moral and aesthetic properties, whose mind-independent existence is denied by nihilists, are different. In these cases, *all we have* are intuitions that cannot be independently corroborated. Nihilists/skeptics do not claim that "If something is very dissimilar from other things that exist, then it (probably) does not exist." Rather, the claim is that strict criteria must be met in order to establish the existence of such phenomena. Perhaps a mere psychological disposition to believe provides some prima facia justification for the belief. But if the disposition itself *could* be explained in terms of phenomena that are known to exist via multiple, independent lines of evidence, this substantially weakens, if not removes, the justification for the belief (assuming no better justification can be found). Consider the fact that many people have a strong intuition that they interact with a personal God who is "very dissimilar from other things that exist." This intuition is not *by itself* a compelling reason to accept God's existence given that we have plausible how-possibly explanations that attribute the intuition to (not fully understood) neuropsychological processes.

4. The Golden Rule: Do Careful Thinkers Independently Agree about the Principle of Universal Benevolence?

Some careful thinkers—Sidgwick, de Lazari-Radek, Singer, and others—have a strong feeling of conviction that the good of all sentient beings should be maximized impartially. But if this conviction is shared mainly by a group of thinkers coming out of

a particular (Western) intellectual tradition, that would undermine step (2) of de Lazari-Radek and Singer's argument and lower the burden on the debunking how-possibly explanation. The debunking explanation would only need to explain why belief in PUB could have arisen under a specific set of cultural conditions, not why it is consistently endorsed by careful thinkers independently of each other.

PUB says that we ought to impartially maximize "the good" of individuals. Without a theory of what that good consists in, PUB is "empty of content" (de Lazari-Radek and Singer 2012: 27). However, PUB is most often combined with a utilitarian theory of the good (i.e., the good consists in pleasure and the absence of pain). It was explicitly formulated for the first time as the "principle of utility" by Jeremy Bentham in England in 1780.[7] (As will be discussed in the following section, some pre-Benthamite thinkers in China and Europe followed different lines of reasoning to formulate principles that share some properties with, but still fall short of, PUB.) If taken as the single foundational principle of morality, PUB combined with a utilitarian theory of the good is utilitarianism. Utilitarianism has been influential among philosophers, though it has always been a minority position. A recent survey of philosophers found that less than a quarter identify as consequentialists of any kind (Bourget and Chalmers 2014: 476). It is technically possible to regard PUB as one among multiple, potentially conflicting ethical principles. That is, one could hold that we have a pro tanto reason to impartially maximize the good of sentient beings, which could be overridden by competing reasons. It is unclear how common this view is. PUB is primarily associated with utilitarians who, again, are a minority.

[7] Bentham's *An Introduction to the Principles of Morals and Legislation*, which introduced the principle of utility, was first printed in 1780 and officially published in 1789.

At first glance, rather than being something about which "careful thinkers" tend to independently agree, PUB is associated with a small number of mostly anglophone philosophers in fairly recent history. But, as noted earlier, de Lazari-Radek and Singer (2012: 25–26) claim that "the leading thinkers of distinct traditions"—they mention Judaism, Christianity, Confucianism, Hinduism, and Buddhism—have taken the "point of view of the universe" and "independently reached a similar principle [in the form of the Golden Rule] and have regarded it as the essence of morality....[T]here is no plausible explanation of this principle as the direct outcome of an evolutionary process, nor is there any other obvious non-truth-tracking explanation." Parfit (2011) similarly argues that the Golden Rule was "independently proclaimed and accepted in several of the world's earliest civilizations" (p. 536). He says that belief in the Golden Rule—like the belief that "everyone's wellbeing matters equally"—is "clearly *not* the product of evolutionary forces" (p. 538). Interestingly, Mill (1863/2015: 130) regarded Jesus's Golden Rule as expressing "the complete spirit of the ethics of utility." He did not address the question of whether belief in this teaching is immune from EDAs. But the Golden Rule is not sufficiently "similar" to PUB to support de Lazari-Radek and Singer's argument. Furthermore, the idea that the Golden Rule is the "essence of morality" in all the aforementioned traditions seems to be based on a tendentious reading of some cherry-picked quotes. Contra de Lazari-Radek and Singer (and Parfit), there is a compelling debunking (how-possibly) explanation for why versions of the Golden Rule appear in different traditions.

Let's begin with Christianity. Though Jesus does not comment explicitly on "morality" (his concern is the religious law), the Gospels quote him stating that the Golden Rule is the essence of the religious commandments: "whatever you want men to do to you, do also to them, for this is the Law and the Prophets" (Matthew 7:12 NKJV).

In another passage, he agrees with a lawyer's suggestion that one inherits eternal life by "'lov[ing] the Lord your God with all your heart, with all your soul, with all your strength, and with all your mind', and 'your neighbor as yourself'" (Luke 10:27). Paul expounds:

> Owe no one anything except to love one another, for he who loves another has fulfilled the law. For the commandments...are *all* summed up in this saying, namely, "You shall love your neighbor as yourself." Love does no harm to a neighbor; therefore love *is* the fulfillment of the law. (Romans 13:8–10)

In the Sermon on the Mount, Jesus offers these instructions, which put the Golden Rule in context:

> Love your enemies, do good to those who hate you, bless those who curse you, and pray for those who spitefully use you. To him who strikes you on the *one* cheek, offer the other also. And from him who takes away your cloak, do not withhold *your* tunic either. Give to everyone who asks of you. And from him who takes away your goods do not ask *them* back. And just as you want men to do to you, you also do to them likewise. (Luke 6:27–31)

There are a few things to notice about the Golden Rule as it was taught by Jesus. First, it is clearly not concerned with "sentient beings" generally, but people specifically. There is no evidence that Jesus (or any other Christian thinkers) thought that people merit love in virtue of being sentient or experiencing pleasure and pain. Second, even restricted to human beings, the Golden Rule does not recommend impartially maximizing the good. Jesus tells his followers to allow themselves to be abused and to love their abusers, whom Jesus promises to later "cast...into the furnace of fire," i.e., Hell, for all eternity (Matthew 13:42). This is not at all "similar" to PUB (again, even restricted to members of our species). Third, Jesus does not purport to take the "point of view of the universe." He explicit takes the point of view of a specific agent, namely, *God*. He interprets God's law

from the perspective of *God's* priorities, which do not appear to be informed by PUB. Although God may "love" people, he will nevertheless "gather out of his kingdom all things that offend, and those who practice lawlessness, and will cast them into the furnace of fire" (Matthew 13:41–42). To suggest, as de Lazari-Radek and Singer do, that Jesus took the "point of view of the universe" to discover a moral principle that is "similar" to PUB is a big stretch.

The version of the Golden Rule in the Hebrew Bible is most naturally read as nonimpartial even with respect to human beings: "Do not take revenge and do not hold a grudge against *bnei amechoh* (= sons of your people/nation) and love *reiachoh* (= your neighbor) like yourself" (Leviticus 19:18). Most of the rabbinic commenters interpret *reia* (neighbor) to refer to Jews, or just religious Jews. Maimonides defines *reia* in this context as a Jew who is your "brother in Torah and mitzvos [i.e., religious commandments]" (*Mishneh Torah, Avel* 14:1; in Maimonides n.d.). Even if *reia* is interpreted universally (as it was by some Jewish commentators), there is no indication in the Bible that this is the "essence of morality."

Proponents of the "modern claim that the central character and essential originality of Judaism lies in its universalist, humanistic ethics" (Reinhard 2005) often cite an exchange recorded in the Talmud between the sage Hillel and a potential convert. The would-be convert demanded that Hillel teach him the entire Torah while he stood on one foot, to which Hillel replied: "What you hate, do not do to others. All the rest is commentary. Go study" (*Shabbos* 31a in Schorr 2001). At first glance this sounds similar to the Christian idea that the essence of the law is the Golden Rule. However, Hillel's statement needs to be understood in context. In a search of the classic Rabbinic literature, Navon (2010) found 80 instances where 15 different mitzvahs or states of affairs are declared to be "equal to all" the Torah or the commandments. Things that are equated

with all the Torah or all the commandments include the sabbath, circumcision, *tzitzis* (putting fringes on four-cornered garments), studying the Torah, peace, charity, and living in Israel. The idea that Hillel's statement to the convert shows that the essence of Judaism or Jewish ethics is the Golden Rule is not supported when considering it in the context of the tradition.

Even if we take Hillel's statement that all of the Torah is a commentary on the principle, "What you hate, do not do to others," it would not mean that PUB is the essence of Jewish ethics. The injunction does not say anything about maximizing the good of all people—let alone of all sentient beings. It is a general exhortation to prosocial behavior. And Hillel does not purport to take the "point of view of the universe."

In a footnote, de Lazari-Radek and Singer (2012: n. 44) cite three non-Western sources to support their claim that the Golden Rule is seen as the essence of morality in different traditions: the *Analects of Confucius*, the Hindu *Mahabharata*, and the Buddhist *Saṃyutta Nikāya*. Let's look at these examples.

De Lazari-Radek and Singer (2012: n. 44) quote the *Mahabharata*: "One should not behave towards others in a way which is disagreeable to oneself. This is the essence of morality. All other activities are due to selfish desire." (The word translated as "morality" is *dharma*.) But presenting this quote in isolation may paint a misleading picture of Hindu ethics. Bakker (2013: 49) observes that

> the Hindu commentaries do not pay special attention to the Golden Rule as a subject in its own right....The Hindu authors focus on themes like *dharma*, self-control, *ahimsa* (non-violence), asceticism, the effects of certain deeds on afterlife and the importance of living in accordance with the caste hierarchy, while the Golden Rule itself is scarcely mentioned separately.

As Davis (2008: 147) says, "Indic formulations of the Golden Rule" (including the one

quoted above) “repeatedly point toward a principle that is much more significant within classical Indic ethical discourse, namely the principle of *ahimsa*, nonharming or nonviolence” (quoted in Bakker 2013: 49). It may be that some Western commentators eager to find commonalities between Christianity and the other great religions have assigned the Golden Rule in Hinduism a meaning and significance that it does not have. A few statements in the Hindu literature are superficially similar to the Golden Rule of the Gospels, but they are rooted in fundamentally different philosophical and ethical perspectives.

De Lazari-Radek and Singer cite but do not quote passages from Confucius and the Buddha.

Confucius says:

> Tzu-kung asked, ‘Is there a single word which can be a guide to conduct throughout one’s life?’ The Master said, ‘It is perhaps the word “*shu*”. Do not impose on others what you yourself do not desire.’ (Analects 15.24; in Confucius 1992)

How this statement should be interpreted is not obvious, but the idea that the essence of Confucian morality is a Christian-style Golden Rule is implausible on its face. It simply does not fit with any of the conspicuous features of Confucianism. Consider the following observations by Csikszentmihalyi (2008: 157):

> The earliest Chinese expressions of Golden Rule-style injunctions existed somewhat uneasily within a system that otherwise emphasized acting out of a set of virtues. While post-Buddhist Confucians were better able to integrate the general principle of reflexivity into their moral system, they still had difficulty reconciling it with classical aspects of their tradition. A close examination of both early and late traditional writing on Golden Rule passages in the Confucian canon reveals that the scope of the application of the rule was often restricted, sometimes even to the point of being used as a metaphor for reflexivity in action rather than as a moral imperative.

Like Hillel, Confucius seems to be making a general exhortation to prosocial behavior, not instructing us to impartially maximize the good of all people or sentient beings, and certainly not purporting to take the "point of view of the universe." Confucians explicitly rejected the idea of impartial moral concern. Mencius called Mozi a "beast" for promoting an ethics of "inclusive care," which Mencius interpreted (perhaps wrongly) to mean that people do not have special moral obligations to their fathers (Fraser 2016: xi). (More on Mozi in the following section.)

The Buddha says: "What is displeasing and disagreeable to me is displeasing and disagreeable to the other too. How can I inflict upon another what is displeasing and disagreeable to me?" (*Saṃyutta Nikāya* 353). Nothing the Buddha says here or elsewhere suggests that this principle is the essence of morality, that the good of all beings should be maximized impartially, or that he purports to take the "point of view of the universe."

It is easy to come up with a plausible, debunking how-possibly explanation for the Golden Rule. Moral educators in different traditions independently realized that they could foster prosocial behavior by harnessing our capacity for empathy—by encouraging people to imagine themselves in others' shoes. It is not surprising that people who were concerned with the commonweal and understood basic human psychology would hit upon some of the same rhetorical and pedagogical strategies.

In summary, step (2) of de Lazari-Radek and Singer's argument is an empirical, anthropological claim, namely, PUB was derived independently by thinkers in different traditions in the form of the Golden Rule. The evidence reviewed above suggests that this is false. First, the Golden Rule is fundamentally different from PUB. De Lazari-Radek and Singer (2012: 21) specifically emphasize that "the principle of universal benevolence bids us to have concern not only for the good of our own species but for all

sentient beings." It is supposed to spring from the recognition that, from the "point of view of the universe," every sentient being's pleasure and pain has equal importance. But the Golden Rule says nothing about nonhuman animals, suggesting that it is motivated by different considerations. The Golden Rule—particularly in its more common negative formulation (*do not* do to others what you yourself dislike)—does not even suggest maximizing the good impartially among members of our own species. Second, the Golden Rule appears to be seen as a fundamental principle only in Christianity. Third, none of the thinkers in the traditions de Lazari-Radek and Singer refer to claim to derive the Golden Rule by taking the "point of view of the universe." Jesus, for example, takes the point of view of a non-PUB-committed *God*. Fourth, in most traditions, the Golden Rule seems to be best understood as a general exhortation to prosocial behavior rather than the foundation of morality.

5. A Debunking How-Possibly Explanation for the Principle of Universal Benevolence

The how-possibly explanation for why a small group of (mostly anglophone) intellectuals have come to believe in PUB will invoke the following four non-truth-tracking forces.

First, normal members of our species have a tendency to *empathize*, which, under certain conditions, leads us to be concerned with the well-being of others. Empathy is an adaptation with a genuinely powerful evolutionary explanation. According to de Waal (2008: 279):

> Evolutionary theory postulates that altruistic behavior evolved for the return-benefits it bears the performer....Empathy is an ideal candidate mechanism to

> underlie so-called directed altruism, i.e., altruism in response to another's pain, need, or distress....The dynamics of the empathy mechanism agree with predictions from kin selection and reciprocal altruism theory.

Given that empathy can be explained by natural selection, any moral beliefs to which it gives rise are subject to the strongest possible EDA (Cofnas 2020b: 14–15).

Second, people pursue their narrow *self-interest*. As de Lazari-Radek and Singer (2012) themselves point out, our tendency to pursue our self-interest would obviously have been favored by natural selection. Any moral belief that can be traced back to the pursuit of self-interest is (in the absence of alternative justification) thereby debunked.

Third, we engage in *moral consistency reasoning* (Campbell and Kumar 2012)—what moral philosophers call the *method of reflective equilibrium*. When we discover that we make opposing judgements about ostensibly morally similar cases, we seek to resolve the inconsistency by changing our judgements or modifying our principles. According to Campbell and Kumar (2012: Fig. 1), this process works as follows. We begin by making intuitive moral evaluations of similar cases. These evaluations are generated by "system 1"—they are emotional and automatic. If our judgements are opposed, we employ the rational "system 2" to identify possible morally relevant differences between the cases, which are in turn fed back into system 1. If the "differentiating features" identified by system 2 fail to activate any intuitions in system 1, there is no affective response and we fail to resolve the inconsistency. Then we either "attempt, in system 2, to revise the less tenable response" (i.e., reverse one of our judgements), or "identify a heretofore unseen relevant difference between the two cases." If we identify a new morally relevant difference, we can re-describe the cases in a way that resolves the inconsistency.

One might think that moral consistency reasoning is truth tracking, since consistency is a mind-independent property that is required for a set of beliefs to be true. But consistency reasoning per se is not truth tracking. It only becomes truth tracking when the procedure employed to bring about consistency is truth tracking. To illustrate the point, consider Kahane's (2011: 119–120) example of a man with a compulsion to count every blade of grass in his backyard. We might criticize him for making an arbitrary distinction between his backyard and everywhere else. There is "nothing special" about his backyard, ergo he ought to count every blade of grass on earth. But if the original motivation (count the grass in his backyard) does not correspond to some mind-independent moral imperative, neither does its "reasoned extension" (count all the grass on earth).[8] If Campbell and Kumar's (2012) model is correct, moral consistency reasoning is driven by moral intuitions encoded in (the nonrational) system 1, so could only be truth tracking if the underlying intuitions were epistemically reliable. According to Campbell and Woodrow (2003: 361), the drive for moral consistency itself has a naturalistic (and possibly evolutionary) explanation. They observe that normative consistency promotes "stability in behavior" and thus predictability. "A creature that is inconsistent in normative expression is unreliable and consequently a poor bet for any cooperative endeavor." For this reason, people will come to value moral consistency in others, and seek to be consistent in their own moral judgements.

Fourth, people have a facultative adaptation to make stronger or weaker ingroup/outgroup distinctions depending on the circumstances. In the presence of threats related to things like disease, scarcity, or physical danger, we tend to become more

[8] Kahane (2011: 120) is commenting on how natural selection endowed us with a disposition for partial altruism (i.e., toward kin and cooperation partners). He notes that "extending this [selectively altruistic] concern through reasoning does nothing to salvage its epistemic status."

ingroup oriented and hostile to outsiders. When conditions are more favorable, this response is attenuated and we are relatively tolerant and open to outsiders. As Buchanan and Powell (2018) put it, the presence or absence of threats conduces to "exclusivist" or "inclusivist" moralities, respectively. The "adaptively plastic moral psychological mechanism" (p. 191) underlying this response was designed by natural selection to track states of affairs relevant to group competition, not moral truth.

Before considering whether there is a plausible, debunking how-possibly explanation for belief in PUB, it is worth noting that there is already a compelling debunking explanation for why we tend to be realists about whatever moral beliefs we happen to hold. Namely, realist intuitions would have been favored by natural selection (Joyce 2006: 108–118; Ruse 1986: 103; Tooby and Cosmides 2010: 221–222). As Joyce (2006: 111) expounds, the benefits from following moral rules—such as an "enhanced reputation"—are often reaped in the far future, if they can be foreseen at all. If we experienced the motivation to follow moral rules as commensurable with other desires, we would be much more likely to succumb to immoral temptations, especially when the benefits of being moral are distant, uncertain, or unforeseen. In Joyce's words: "The distinctive value of imperatives imbued with practical clout is that they silence further calculation, which is a valuable thing when our prudential calculations can so easily be hijacked by interfering forces and rationalizations." Since we evolved to be moral realists, virtually *any* moral belief will, once adopted, appear "self-evident." There is nothing special about the fact that believers in PUB see *this* principle as self-evident, because all people tend to see their moral beliefs the same way. (Remember, for example, how Kant saw the immorality of "unnatural" sex as an obvious truth that "occurs to everyone immediately, with the thought of it"; see note 4 above.) And this can be explained by natural selection.

Now consider the social conditions that prevailed in late eighteenth-century England, where PUB was originally formulated. In the wake of its victory in the Seven Years' War in 1763, Britain rose meteorically as both an economic and military power. Britain gained control of the entire North Atlantic and attained dominance in India. Although it would lose the colonies that became the United States, India turned out to be a bonanza of wealth (Lynn 2000: 202). "The constantly increasing importance of naval power, trade, and colonies had made Britain a hegemon....The fleet rendered Britain invulnerable to direct attack, while its wealth allowed it to intervene on the continent even though Britain did not possess a large army" (Lynn 2000: 204–205). The Industrial Revolution also began around this time, with Britain taking the lead (Allen 2009). This unprecedented security and prosperity created ideal conditions for triggering an inclusivist morality.

In the late eighteenth century, European nations were increasingly entering into mutually beneficial trading relationships. Although each party was motivated by narrow self-interest, the new economic system led to a widening of Europeans' moral circle as an accidental byproduct. Kant (1795/1991: 114) took notice of this phenomenon, writing that nations can be united

> by means of their mutual self-interest. For the *spirit of commerce* sooner or later takes hold of every people, and it cannot exist side by side with war....Thus states find themselves compelled to promote the noble cause of peace, though not exactly from motives of morality.

In the same vein, Pinker (2011: xxvi) observes that people tied to each other through commerce "become more valuable alive than dead, and they are less likely to become targets of demonization and dehumanization." Trade does not inevitably lead to peace and friendship, as some of the more naïve advocates of *doux commerce* (gentle

commerce) theory have suggested. Trade has frequently been accompanied by violence and exploitation. Kant and Bentham were writing at the height of the slave trade, and at a time when trading relationships routinely descended into violence (Hirschman 1997: 62). Nevertheless, in the late eighteenth century, Europeans—and particularly the British—were increasingly coming to understand and experience the benefits of peaceful economic cooperation. This encouraged them to adopt much more inclusivist moralities, which attributed value to a larger swath of humanity.

All of the trends that started in the late eighteenth century accelerated in the nineteenth. Britain was very briefly threatened with attack at the beginning of the Napoleonic Wars (1803–1815) (see Lynn 2000: 211), but ultimately emerged victorious as the "first modern international superpower" (Duffy 2002: 213). The era in which John Stuart Mill (the most influential promoter of Bentham's ideas) and Sidgwick came of age is known as *Pax Britannica* (Gough 2014).

It is not difficult to see how moral philosophers engaged in consistency reasoning in late-eighteen- and nineteenth-century Britain could have stumbled upon PUB. The sociological developments discussed above had cultivated what were (by historical standards) highly inclusivist moral intuitions, which coexisted with widespread acceptance of colonialism and animal cruelty. This led people to form apparently inconsistent moral judgements, which different thinkers resolved in different ways. What justifies treating different races differently—granting freedom to one and colonizing or even enslaving others? Given people's enhanced tendency to empathize with foreigners, the idea that race itself was a morally relevant criterion had become psychologically unpalatable. Moral consistency reasoners needed to find some difference between races to justify unequal treatment, revise their judgements, or both. Slavery became untenable, since it could not possibly be squared with a disposition to feel concern for all people.

The slave trade was outlawed by popular demand in all British-controlled territories in 1807, and by 1843 there was blanket emancipation. Colonialism, however, was often justified by pointing to some alleged difference between Europeans and colonized people, such as different capacities for rationality (Marshall 2000: 244–245). If non-Europeans were less rational, colonialism could be framed as an expression of empathy. Mill (1859/2015: 13), for example, argued that certain "race[s]...may be considered as in [their] nonage," and would therefore benefit from the rule of their superiors.

To get to PUB, thinkers had to expand their circle of moral concern beyond our species—to *all sentient beings*. How might this have happened? We can potentially feel empathy toward animals as a byproduct of empathic dispositions that evolved to promote benevolence toward kin and cooperation partners. In late eighteenth-century Britain, social conditions had raised people's tendency to empathize to extremely high levels. One might expect that this empathy would spill over into (some) people's feelings toward animals. Given system-2 rooted concern for animals, philosophers had to confront the question of why we assign so much more value to human than animal well-being—an apparent moral inconsistency. Most people in late eighteenth-century Britain appealed to religion to justify human specialness: all people are equal before God (Marshall 2000: 244), and are spiritually superior to animals in virtue of their divine gifts. This was obviously unacceptable to philosophers who rejected traditional Christianity. The idea that human moral superiority derives from the faculty of rationality, although appealing to some intellectuals, did not stand up to scrutiny. As Bentham (1789: 309) wrote:

> What...is it that should trace the insuperable line [distinguishing humans and animals]? Is it the faculty of reason, or, perhaps, the faculty of discourse? But a full-grown horse, or dog, is beyond comparison a more rational, as well as a more conversible animal, than an infant of a day, or a week, or even a month, old.

Any theory that drew a sharp distinction between humans and animals would fail to accommodate the system 2-rooted empathic concern that (some) people felt toward animals. But the idea that the ability to experience pain or pleasure is what is morally relevant could survive the test of consistency reasoning. Bentham (1789: 309) concluded: "The question is not, Can they *reason?* nor, Can they *talk?* but, Can they *suffer?*" This principle could be approved by system 2, since cues—or for some people even the thought—of suffering and happiness in other people or animals can elicit empathic pain or pleasure, respectively. It followed logically (for those whose emotion-driven consistency reasoning led them down this path) that more pleasure is good, less pain is bad, which implies (a utilitarian version of) PUB. Some thinkers living under conditions that cultivate the requisite sensitivities toward humans and animals have, and continue to be, attracted to PUB as a way to make sense of their intuitions.

5.1. A Religious Route to Impartial Morality

Some pre-Benthamite thinkers followed a religious line of reasoning, which led them to conclude that the interests of all human beings count equally—not from the perspective of the universe, but of *God*. These thinkers were not concerned with sentience or pleasure and pain per se, so they had no regard for animals and did not advocate PUB. But it is noteworthy that there are multiple routes to the view the morality should be impartial. Assuming the religious premises are false, theological reasoning that leads to impartial morality is presumably not truth tracking.

First, there was Mozi, who lived during the Warring States period (4th century BCE) in China. This was a time of great economic development due to increased job specialization and trade, though, as the name of the era indicates, it was also marked by

violent conflict. Mozi is most famous for his doctrine of "inclusive care," which enjoins us to consider all the people of the world like ourselves. He based his ethics on the perspective of Heaven, which was seen as a quasi-personal god (Fraser 2016: 16). "Heaven's conduct is broad and impartial," he said (p. 36). From its perspective, all people count equally: "Now in the world there are no great or small states—all are Heaven's towns. Among people there are no younger or elder, noble or common—all are Heaven's subjects" (p. 36). According to Mozi, action should be guided by the following principle:

> Does it benefit people? Then do it. Does it not benefit people? Then stop....In all statements and all actions, what is beneficial to Heaven, ghosts, and the people, do it. In all statements and all actions, what is harmful to Heaven, ghosts, and the people, reject it.

In Mozi's philosophy, the "benefit [*lì*] of all," despite its utilitarian ring, does not refer to "the total sum of welfare in society or to individuals' average level of welfare" (p. 141). As Fraser (2016: 138) explains, "benefit" includes three kinds of goods: "material prosperity, an adequate population or family size, and sociopolitical order, including social stability and personal and national security. These three goods are the concrete criteria against which the Mohists evaluate statements, conduct, practices, and institutions." Despite the fact that his theory of the good was very different from that of Bentham, Sidgwick, or any other advocates of PUB in the West, Mozi still endorsed impartially maximizing the good with respect to human beings. In the long run, Mohism "had little direct influence on the Chinese philosophical tradition" (p. 19). As mentioned earlier, his doctrine of impartiality was explicitly rejected by the Confucians who came to dominate Chinese philosophy.

A century before Bentham formulated the principle of utility, some Anglican theologians developed a religious version of what scholars now call utilitarianism, which was restricted to our species. Like Mozi, they reasoned that God has an impartial interest in human well-being. They combined this view with a theory of the good that was much closer to Bentham's, namely, the good consists in maximizing pleasure and minimizing pain in rational beings (Heydt 2014). This version of utilitarianism was squarely rooted in religion. As Heydt (2014: 26) expounds, the theological utilitarians "argue that moral obligation requires God and God's sanctions." The theologians derived their conclusions by making inferences about God's intentions, not by engaging in moral consistency reasoning. It is possible that Bentham's moral intuitions, which drove his moral consistency reasoning, were influenced to some extent by the teachings of theologians, whose conclusions were based on religious premises that Bentham rejected.

6. Conclusion

Advocates of EDAs typically claim that there are "powerful" evolutionary explanations for our core moral beliefs—or at least the basic evaluative tendencies that give rise to those beliefs. Advocates of selective EDAs accept that empirical claim, but say that at least one moral belief lacks an evolutionary explanation and is therefore immune from debunking.

I argued that almost none of our moral beliefs have genuinely powerful evolutionary (or other naturalistic) explanations. For the foreseeable future, the best scientific explanations of our moral beliefs will, in most cases, be mere how-possibly explanations appealing to evolution, history, culture, and psychology. Such how-possibly explanations are either debunking or not debunking. Some philosophers would

presumably welcome the idea that how-possibly explanations are not debunking. This would mean that both global EDAs to defend antirealism (e.g., Street 2006) and selective EDAs to defend *realism* (e.g., de Lazari-Radek and Singer 2012) fail. I argued that, as long as naturalistic how-possibly explanations surpass a certain threshold of plausibility, they are debunking—realist explanations should be rejected on grounds of ontological parsimony. By exercising a little imagination, it will generally be possible to come up with plausible, naturalistic how-possibly explanations for moral beliefs that realists have held up as immune from debunking—such as belief in the principle of universal benevolence.

References

Allen, Robert C. 2009. *The British Industrial Revolution in Global Perspective.* Cambridge, UK: Cambridge University Press.

Bakker, Freek L. 2013. "Comparing the Golden Rule in Hindu and Christian Religious Texts." *Studies in Religion* 42 (1): 38–58.

Bentham, Jeremy. 1789. *An Introduction to the Principles of Morals and Legislation.* London: T. Payne and Sons.

Birch, Jonathan. 2012. "The Negative View of Natural Selection." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (2): 569–573.

Blanke, Olaf, Nathan Faivre, and Sebastian Dieguez. 2016. "Leaving Body and Life Behind: Out-of-Body and near-Death Experience." In *The Neurology of Consciousness: Cognitive Neuroscience and Neuropathology,* 2nd edition, edited by Steven Laureys, Olivia Gosseries, and Giulio Tononi, 323–347. San Diego, CA: Academic Press.

Bloom, Paul. 2019. “Comments.” *Current Anthropology* 60 (1): 59–60.

Bourget, David, and David J. Chalmers. 2014. “What Do Philosophers Believe?” *Philosophical Studies* 170 (3): 465–500.

Brandon, Robert N. 1990. *Adaptation and Environment.* Princeton, NJ: Princeton University Press.

Buchanan, Allen, and Russell Powell. 2018. *The Evolution of Moral Progress: A Biocultural Theory.* New York: Oxford University Press.

Campbell, Richmond, and Victor Kumar. 2012. “Moral Reasoning on the Ground.” *Ethics* 122 (2): 273–312.

Campbell, Richmond, and Jennifer Woodrow. 2003. “Why Moore’s Open Question Is Open: The Evolution of Moral Supervenience.” *The Journal of Value Inquiry* 37 (3): 353–372.

Cofnas, Nathan. 2020a. “Are Moral Norms Rooted in Instincts? The Sibling Incest Taboo as a Case Study.” *Biology & Philosophy* 35 (5): 47.

———. 2020b. “A Debunking Explanation for Moral Progress.” *Philosophical Studies* 177 (11): 3171–3191.

Colyvan, Mark. 2001. *The Indispensability of Mathematics.* New York: Oxford Universty Press.

Confucius. 1992. *The Analects.* 2nd edition. Translated by D. C. Lau. Hong Kong: Chinese University Press.

Crisp, Roger. 2012. “Katarzyna de Lazari-Radek and Peter Singer, ‘The Objectivity of Ethics and the Unity of Practical Reason.’” *PEA Soup.* Retrieved from http://peasoup.us/2012/12/ethics-discussions-at-pea-soup-katarzyna-de-lazari-radek-and-peter-singer-the-objectivity-of-ethics-1/

Csikszentmihalyi, Mark A. 2008. “The Golden Rule in Confucianism.” In *The Golden Rule: The Ethics of Reciprocity in World Religions,* edited by Jacob Neusner and Bruce Chilton, 157–169. London: Continuum.

Cuneo, Terence, and Russ Shafer-Landau. 2014. “The Moral Fixed Points: New Directions for Moral Nonnaturalism.” *Philosophical Studies* 171 (3): 399–443.

Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex.* Vol. 1. London: John Murray.

———. 1874. *The Descent of Man, and Selection in Relation to Sex.* 2nd edition. London: John Murray.

Davis, Richard H. 2008. “A Hindu Golden Rule, in Context.” In *The Golden Rule: The Ethics of Reciprocity in World Religions,* edited by Jacob Neusner and Bruce Chilton, 146–156. London: Continuum.

de Lazari-Radek, Katarzyna, and Peter Singer. 2012. “The Objectivity of Ethics and the Unity of Practical Reason.” *Ethics* 123 (1): 9–31.

de Waal, Frans B. M. 2008. “Putting the Altruism Back into Altruism: The Evolution of Empathy.” *Annual Review of Psychology* 59: 279–300.

Duffy, Michael. 2002. “Contested Empires, 1756–1815.” In *The Eighteenth Century, 1688–1815,* edited by Paul Langford, 213–244. Oxford: Oxford University Press.

Field, Hartry. 2016. *Science without Numbers: A Defense of Nominalism.* 2nd edition. Oxford: Oxford University Press.

FitzPatrick, William J. 2015. “Debunking Evolutionary Debunking of Ethical Realism.” *Philosophical Studies* 172 (4): 883–904.

Fraser, Chris. 2016. *The Philosophy of the Mòzǐ: The First Consequentialists.* New York: Columbia University Press.

Gough, Barry. 2014. *Pax Britannica: Ruling the Waves and Keeping the Peace before Armageddon.* New York: Palgrave Macmillan.

Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues." *Daedalus* 133 (4): 55–66.

———. 2007. "The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules." In *The Innate Mind, Vol. 3: Foundations and the Future,* edited by Peter Carruthers, Stephen Laurence, and Stephen P. Stich, 367–391. Oxford: Oxford University Press.

———. 2011. "How Moral Foundations Theory Succeeded in Building on Sand: A Response to Suhler and Churchland." *Journal of Cognitive Neuroscience* 23 (9): 2117–2122.

Harman, Gilbert. 1977. *The Nature of Morality: An Introduction to Ethics.* New York: Oxford University Press.

Heydt, Colin. 2014. "Utilitarianism before Bentham." In *The Cambridge Companion to Utilitarianism,* edited by Ben Eggleston and Dale E. Miller, 16–37. Cambridge, UK: Cambridge University Press.

Hirschman, Albert O. 1997. *The Passions and the Interests: Political Arguments for Capitalism before Its Triumph.* Twentieth anniversary edition. Princeton, NJ: Princeton University Press.

Hopster, Jeroen. 2018. "Evolutionary Arguments against Moral Realism: Why the Empirical Details Matter (and Which Ones Do)." *Biology & Philosophy* 33 (5–6): 41.

———. 2020. "Explaining Historical Moral Convergence: The Empirical Case against Realist Intuitionism." *Philosophical Studies* 177 (5): 1255–1273.

Huemer, Michael. 2005. *Ethical Intuitionism.* New York: Palgrave Macmillan.

———. 2016. "A Liberal Realist Answer to Debunking Skeptics: The Empirical Case for Realism." *Philosophical Studies* 173 (7): 1983–2010.

Joyce, Richard. 2006. *The Evolution of Morality*. Cambridge, MA: MIT Press.

———. 2008. "Preçis of *the Evolution of Morality*." *Philosophy and Phenomenological Research* 77 (1): 213–218.

Kahane, Guy. 2011. "Evolutionary Debunking Arguments." *Noûs* 45 (1): 103–125.

Kant, Immanuel. 1795/1991. "Perpetual Peace: A Philosophical Sketch." In *Kant: Political Writings,* 2nd edition, translated by H. B. Nisbet, edited by H. S. Reiss, 93–130. Cambridge, UK: Cambridge University Press.

———. 1797/2017. *The Metaphysics of Morals.* Revised edition. Translated by Mary Gregor. Edited by Lara Denis. Cambridge, UK: Cambridge University Press.

Lowe, Adriana E., Catherine Hobaiter, Caroline Asiimwe, Klaus Zuberbühler, and Nicholas E. Newton-Fisher. 2020. "Intra-Community Infanticide in Wild, Eastern Chimpanzees: A 24-Year Review." *Primates* 61 (1): 69–82.

Lynn, John A. 2000. "International Rivalry and Warfare." In *The Eighteenth Century: Europe 1688–1815,* edited by T. C. W. Blanning, 178–217. Oxford: Oxford University Press.

Maimonides, Moses. n.d. "Mishneh Torah: Avel - Chapter 14." Retrieved from https://www.chabad.org/library/article_cdo/aid/1181895/jewish/Avel-Chapter-14.htm#lt=he

Marshall, P. J. 2000. "Europe and the Rest of the World." In *The Eighteenth Century: Europe 1688–1815,* edited by T. C. W. Blanning, 218–246. Oxford: Oxford University Press.

Mill, John Stuart. 1859/2015. “On Liberty.” In *On Liberty, Utilitarianism and Other Essays,* edited by Mark Philp and Frederick Rosen, 5–112. Oxford: Oxford University Press.

———. 1863/2015. “Utilitarianism.” In *On Liberty, Utilitarianism and Other Essays,* edited by Mark Philp and Frederick Rosen, 115–155. Oxford: Oxford University Press.

Navon, Mois. 2010. “Equal to All the Mitzvot in the Torah.” Retrieved from http://www.divreinavon.com/pdf/EqualToAll.pdf

Parfit, Derek. 2011. *On What Matters.* Vol. 2. Oxford: Oxford University Press.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined.* New York: Viking.

Prinz, Jesse. 2007. *The Emotional Construction of Morals.* Oxford: Oxford University Press.

Reinhard, Kenneth. 2005. “The Ethics of the Neighbor: Universalism, Particularism, Exceptionalism.” *Journal of Textual Reasoning* 4 (1).

Ruse, Michael. 1986. “Evolutionary Ethics: A Phoenix Arisen.” *Zygon* 21 (1): 95–112.

Ruse, Michael, and Edward O. Wilson. 1986. “Moral Philosophy as Applied Science.” *Philosophy* 61 (236): 173–192.

[*Saṃyutta Nikāya*] 2000. *The Connected Discourses of the Buddha: A Translation of the Saṃyutta Nikāya.* Translated by Bhikkhu Bodhi. Boston: Wisdom Publications.

Sauer, Hanno. 2018. *Debunking Arguments in Ethics.* Cambridge, UK: Cambridge University Press.

Schorr, Yisroel Simcha, ed. 2001. *Talmud Bavli - Shabbos.* Vol. 1. New York: Mesorah.

Sidgwick, Henry. 1907. *The Methods of Ethics.* 7th edition. London: Macmillan.

Street, Sharon. 2006. “A Darwinian Dilemma for Realist Theories of Value.” *Philosophical Studies* 127 (1): 109–166.

Tooby, John, and Leda Cosmides. 2010. “Groups in Mind: The Coalitional Roots of War and Morality.” In *Human Morality and Sociality: Evolutionary and Comparative Perspectives,* edited by Henrik Høgh-Olesen, 91–234. New York: Palgrave Macmillan.