BEHAVIORAL AND BRAIN SCIENCES (2018), Page 1 of 49
doi:10.1017/S0140525X17001911, e95

The difference between ice cream and Nazis: Moral externalization and the evolution of human cooperation

P. Kyle Stanford
Department of Logic and Philosophy of Science, University of California, Irvine, Irvine, CA 92697
stanford@uci.edu
http://www.lps.uci.edu/lps_bios/stanford

Abstract: A range of empirical findings is first used to more precisely characterize our distinctive tendency to objectify or externalize moral demands and obligations, and it is then argued that this salient feature of our moral cognition represents a profound puzzle for evolutionary approaches to human moral psychology that existing proposals do not help resolve. It is then proposed that such externalization facilitated a broader shift to a vastly more cooperative form of social life by establishing and maintaining a connection between the extent to which an agent is herself motivated by a given moral norm and the extent to which she uses conformity to that same norm as a criterion in evaluating candidate partners in social interaction generally. This connection ensures the correlated interaction necessary to protect those prepared to adopt increasingly cooperative, altruistic, and other prosocial norms of interaction from exploitation, especially as such norms were applied in novel ways and/or to novel circumstances and as the rapid establishment of new norms allowed us to reap still greater rewards from hypercooperation. A wide range of empirical findings is then used to support this hypothesis, showing why the status we ascribe to moral demands and considerations exhibits the otherwise puzzling combination of objective and subjective elements that it does, as well as showing how the need to effectively advertise our externalization of particular moral commitments generates features of our social interaction so familiar that they rarely strike us as standing in need of any explanation in the first place.

Keywords: altruism; cooperation; correlated interaction; ethics; evolution; gossip; hypercooperation; hypocrisy; morality; moral psychology; prosocial behavior

1. Prosocial behavior and moral motivation: A mystery hiding in plain sight

Recent decades have witnessed an explosion of fascinating scientific work concerning prosocial, cooperative, and altruistic behavior in both human and nonhuman organisms. Researchers have now documented an impressive range of instances in which other animals, especially nonhuman primates, exhibit sympathetic and empathetic concern for others, aggression in response to failures of reciprocation, inequity aversion, and other behaviors that would in humans be regarded as characteristic responses to the demands of morality.

Indeed, the pace of recent progress in this area has sometimes obscured the depth of the explanatory challenges that remain. Primatologist Frans de Waal, for example, though careful to concede that nonhuman primates fall short of "full-blown" human morality, nonetheless repeatedly insists that the existing evidence *already* demonstrates that the "foundations," "pillars," or "building blocks" of human moral psychology also exist in nonhuman primates and are explicable in fairly straightforward evolutionary terms (e.g., De Waal 1996; De Waal et al. 2006). But many moral philosophers have resisted this suggestion, arguing that these admittedly fascinating discoveries fail to establish that other organisms share the most central and distinctive features of human moral psychology. Perhaps most importantly, humans experience the demands of morality as somehow *imposed* on us externally: We do not simply enjoy or prefer to act in ways that satisfy the demands of morality; we see ourselves as *obligated* to do so *regardless* of our subjective preferences and desires, and we regard such demands as imposing unconditional obligations not only on ourselves, but also on any and all agents whatsoever, regardless of *their* preferences and desires. As philosopher Richard Joyce notes, even if we

P. Kyle Stanford is a Professor of Logic and Philosophy of Science at the University of California, Irvine. He is the author of many publications in the philosophy of biology and the philosophy of science, including his book, *Exceeding Our Grasp: Science, History and the Problem of Unconceived Alternatives* (2006, Oxford University Press), as well as recent articles concerning moral psychology and its evolution, including "Bending Towards Justice" (2017), and (with Ashley Thomas and Barbara Sarnecka) "No Child Left Alone: Moral Judgments About Parents Affect Estimates of Risk to Children" (2016).

concede that chimpanzees have abundant "inhibitions, aversions, and inclinations" motivating prosocial behavior, and even that these are internalized into what De Waal calls "a sense of social regularity," we must still ask,

> where's the morality? None of the above attributions, nor the sum total of them, amount to a chimpanzee thinking of a negative response as *deserved*, or supposing an act to be a *transgression*, or judging a behavior to be *appropriate*, or considering a trait to be *virtuous*, or assessing a division to be *fair*.... How does one move from having inhibitions to making judgments about prohibitions: from disliking to disapproving, from desiring to regarding-as-desirable? (Joyce 2006, pp. 92–93)

Even if one chimpanzee who punishes another for violating her social expectations is motivated (as De Waal suggests) by concern for the welfare of the larger community, she need not experience this as a matter of respecting or enforcing any kind of *objective* or *externally imposed* moral order, rather than simply discouraging a fellow group member from repeating a violation of what her own subjective preferences (regarding community welfare or anything else) happen to be. De Waal's slide between these distinct possibilities is evident even in his explicit definition of nonhuman primates' "sense of social regularity" as a set of "expectations about the way in which oneself (or others) *ought* to be treated and how resources *ought* to be divided" whose violation provokes protest or punishment (De Waal 1996, p. 95, my emphasis).

Of course, other organisms may well experience demands for prosocial emotions or behaviors as somehow externally imposed upon them, but moral philosophers are right to insist that simply demonstrating the presence of such emotions and behaviors fails to establish that they are motivated in other organisms by any sense of obligation and/or prohibition rather than inclination and aversion. Although De Waal sometimes claims (e.g., 1996, p. 210; see also Suchak et al. 2016) that it is simply parsimonious to assume that similar behavior in similar circumstances arises from similar motivations in nonhuman and human organisms, we would surely want a convincing explanation of *why* humans (or any organisms) would externalize moral motivations, demands, and obligations in this way before deciding whether we think that any particular nonhuman organisms do so.

After all, from an evolutionary point of view, such externalization or objectification is deeply puzzling even in our own case: It seems that mere *subjective preferences* for social interactions with those who are kind, generous, fair, loyal, and so forth, and for avoiding those who are cruel, selfish, deceitful, exploitative, and so forth, would equally effectively guide us towards things that are good for us and away from things that are bad for us, just as it does in the case of our subjective preferences for keeping our heads dry, our tummies full, our orgasms frequent, and for the vast majority of other fitness-enhancing behaviors in which we routinely engage. The real challenge from an evolutionary point of view, then, is not to explain why we *prefer* interacting with kind and generous conspecifics to cruel and selfish ones, but instead why we treat moral demands and obligations as *anything more than such preferences* – as anything more than how we ourselves happen to want to behave and/or what we find appealing in others. Why, that is, do we experience our attitude towards Nazis or slavers any differently than an aversion to kale or a preference for chocolate ice cream over vanilla?

In what follows, I will first appeal to recent findings in cognitive science to more precisely characterize those features of our moral experience and cognition that require such further explanation, and I will argue that existing evolutionary proposals concerning human moral psychology simply do not help meet the distinctive explanatory demand that emerges. I will then go on to argue that externalizing moral demands in this way enabled us to safely discover and take advantage of novel opportunities for productive cooperative interaction by establishing a crucial *connection* between our own motivation to conform to any given distinctively moral norm of behavior and the extent to which we demand that others conform to that same norm if we are to view them as attractive or desirable partners in social interaction more generally. Externalization thus protected our expanding prosocial motivations and commitments from exploitation by ensuring *correlated interaction* between ourselves and those who share those same motivations and commitments, even as new norms governing such behavior were introduced and existing norms were extended in novel ways and applied to novel circumstances. I will then seek to show how this proposal both explains and unifies a number of important further findings from recent experimental cognitive science and how the need to recruit others and advertise ourselves as attractive partners in such forms of interaction structures much more of our resulting social behavior than we realize, including features of our social lives so familiar that we have difficulty recognizing that they require any explanation at all. But before asking whether this proposal offers a convincing explanation for the distinctive character of human moral motivation, let us first seek a more precise characterization of just what phenomena we might be seeking to explain by invoking it.

2. Acquiring the explanatory target

Recent work in cognitive science allows us to characterize our experience of moral demands, considerations, and judgments with unprecedented precision. We should start by noting that our distinctive experience of moral motivation and cognition is not simply a function of either its rule-governed or normative character, for human societies also characteristically recognize normative rules of the sort we regard as mere social conventions (such as norms of etiquette or fashion) to which the further distinctive characteristics of moral demands or norms are not attributed. Social psychologists have documented that, across an impressively heterogeneous variety of societies and cultures, children between the ages of 2.5 and 3 years begin to reliably and systematically distinguish norms and transgressions recognized as genuinely moral in character, such as pulling hair or stealing, from those regarded instead as merely conventional, such as talking out of turn or drinking soup from a bowl (Smetana 2006; Turiel 1983; Turiel et al. 1987). Like moral demands, social conventions or rules are understood normatively, but moral violations are nonetheless regarded as more serious and more deserving of punishment than violations of merely conventional norms, and their wrongness is typically explained by appeal to considerations of harm, fairness,

justice, rights, or the welfare of victims, whereas comparable explanations for conventional norms appeal instead to considerations of social utility or acceptability. Moral norms are systematically viewed both as more "generalizable" (i.e., applicable to people in other places and historical periods, whether or not they care about them) than mere social conventions and as "authority-independent," meaning that their force cannot be suspended by an appropriate individual or institution: for example, children report that it is not wrong to chew gum in class if a teacher has no rule against it, but that hitting another student is wrong no matter what the teacher says. Nucci (1986) showed that this distinction is salient even to Amish and Mennonite teenagers with respect to God's authority: 100% of these subjects agreed that it would not be wrong to leave their heads uncovered or to ignore the prescribed day of worship if God had made no rule against this, but more than 80% insisted that stealing or hitting would *still* be wrong even if God had made no rule forbidding it. Some fascinating evidence suggests that psychopaths may constitute the exception that proves this rule (treating all social norms and rules as purely conventional in character; see Blair 1995; Nichols 2004), but this claim remains controversial (cf. Aharoni, et al. 2012).

More recent work, however, has challenged whether moral norms *must* concern or be justified by appeal to considerations of harm, fairness, justice, rights, or welfare. Haidt et al. (1993), for instance, showed that groups of low socioeconomic status (SES) subjects in both Brazil and the United States judged harmless transgressions such as privately washing the toilet bowl with the national flag and privately masturbating with a dead chicken to be both serious and distinctively moral in character. In a study of children in traditionally Arab villages in Israel, Nisan (1987) found that a moralized response was provoked by norm violations of all kinds, including mixed-sex bathing and addressing a teacher by his first name. Nichols (2004) showed that rules of etiquette prohibiting disgusting activities in particular were judged by American children to be serious, generalizable, and authority-independent, while American college students judged the same prohibitions to be serious and authority-independent, though not generalizable. Such findings illustrate that norms *need not* concern harm, justice, or rights in order to exhibit further characteristics associated with distinctively moral normativity. Moreover, Kelly et al. (2007, pp. 118–21) show in addition that scenarios that *do* involve harm, justice, and rights-based violations (beyond those likely familiar to young children from a school or playground) can be presented in ways that undermine subjects' commitments to features like their generalizability or authority-independence.[1] These results suggest that although moral norms may be frequently or even prototypically concerned with harm, fairness, justice, rights, or welfare, it would be a mistake to treat such concerns as a defining feature of moral norms themselves. Indeed, we might hope for a satisfying account of the reliable and robustly cross-cultural ontogenetic emergence of a distinction between merely conventional and genuinely moral norms to explain *why* the latter are frequently but not invariably concerned with harm, fairness, justice, rights, or welfare.

More recently, Nichols and Folds-Bennett (2003) have shown that in addition to distinguishing moral from merely conventional norms and violations, even very young children also distinguish moral properties from "response-dependent" properties (like "icky," "yummy," or "boring") that are seen as *constituted* by our own reactions to features or aspects of the world. In these experiments, 4- to 6-year-old children did not treat the existence of such response-dependent properties as depending on the presence of actual responders, holding, for example, that grapes were yummy even before anyone was around to taste them. But faced with disagreement, children judged that grapes were yummy and cleaning house boring only "for some people," whereas children facing precisely parallel disagreement did not judge that one monkey helping another who is hurt is good only "for some people," but instead that it is good "for real."

Perhaps most revealing of all, Goodwin and Darley (2008; 2012) show that subjects reliably locate moral judgments at a particular point along a *scale* of increasing objectivity ranging from judgments of taste or preference ("Frank Sinatra was a better singer than is Michael Bolton") to judgments of social convention ("Wearing pajamas and bath robe to a seminar meeting is wrong behavior") to moral judgments ("Robbing a bank in order to pay for an expensive holiday is a morally bad action") to judgments of straightforward empirical or scientific fact ("Boston [MA] is farther north than Los Angeles [CA]"). Specifying such objectivity as a question of (i) whether disagreements require that at least one party be mistaken or (ii) whether there can be a right answer regarding whether the statement in question is true, these authors found that:

> ethical beliefs were treated almost as objectively as scientific or factual beliefs, and decidedly more objectively than social conventions or tastes. Individuals seem to identify a strong objective component to their core ethical beliefs, and thus treat them as categorically different from social conventions. (Goodwin & Darley 2008, p. 1359)

These categorical differences in subjects' willingness to tolerate the possibility of disagreement without error concerning various kinds of judgments provide perhaps the clearest way to characterize the sense in which we treat moral norms and judgments as systematically more objective than judgments of taste or preference or judgments of social convention. Note, however, that along with such categorical differences, Goodwin and Darley (2012) also found considerable variation in the degree of objectivity assigned to particular ethical statements by different experimental subjects, as well as in the relative degree of objectivity assigned to moral judgments with different sorts of content – the statement that it is wrong to rob a bank, for example, was reliably judged far more objective than the statement that anonymous giving is good. Such variation has been particularly emphasized by Wright et al. (2013; 2014) in work that confirms Goodwin and Darley's central findings.

Having specified with somewhat greater precision the features of distinctively moral motivation for which any satisfying explanation of moral externalization or objectification will have to account, I will now go on to argue that existing efforts to understand aspects or features of human moral psychology in evolutionary terms do not help satisfy these explanatory demands, whether or not they are explicitly addressed to them. That is, extant attempts to solve the puzzle explicitly are unpersuasive, while other influential views concerning the function(s) of human moral psychology (even if correct) do not actually

help explain why our experience of moral motivation and moral norms systematically differs from that of mere preferences, desires, and merely conventional norms in the salient ways just described.

3. Illuminating failures

Following his own characterization of our puzzle (see previous sect. 1), Joyce (2006) himself offers two independent lines of explanation for the distinctively externalized character of our moral experience. He begins with the familiar suggestion that moral considerations will motivate us *more effectively* if we take them to be perceptions of the world's own qualities rather than our subjective reactions to it. But this claim would seem to turn on a pair of confusions.

First, truly adaptive moral motivation would not be *as strong as possible* but would instead provide the *right* degree of motivation as balanced against a wide array of other evolutionarily important motivational impulses: Even if acting in morally praiseworthy ways turns out to be important from an evolutionary point of view, so are alleviating hunger, avoiding predation, having sex, and much else besides. Thus, simply increasing the motivational force of any and all moral considerations across the board is not advantageous unless we implausibly suppose that it always favors our evolutionary interests to behave in what we see as the most morally admirable way possible rather than balancing moral considerations against other evolutionarily significant motivations like hunger, fear, and sexual desire. Note that much the same problem afflicts Daniel Dennett's influential proposal (see Dennett 1995, Ch. 17) that moral considerations serve as "conversation stoppers" terminating what would otherwise be endless chains of internal deliberation about what to do. Even more clearly than Joyce's, Dennett's proposal suggests that it is somehow advantageous for moral considerations to trump all other motivational impulses, and this seems to describe neither a plausible strategy for evolutionary survival nor our actual internal deliberations: Rare indeed are those for whom the recognition that an act would less than fully satisfy the demands of morality precludes any further deliberative consideration of whether or not to perform it.

Even if this challenge can be overcome, however, another looms for both Joyce and Dennett, because it seems clear in any case that mere subjective states can be as strongly motivating as any others. After all, pain is a paradigmatically subjective response to the world, but there are few more powerful motivations for humans than avoiding serious pain, so it seems that even purely subjective states can generate motivation of arbitrarily high magnitude. Hence, even if creatures who externalize moral judgments are indeed more motivated by those judgments than they would be by subjective preferences or desires of equal strength (cf. Young & Durwin 2013), the real question is how and why we came to externalize or objectify moral demands in this way in the first place, and the need for more effective or powerful motivation simply does not help explain *that*.

At the heart of this second difficulty lies the fact that sufficiently strong desires and subjective preferences (of just the sort we have for one ice cream flavor over another) seem capable of doing the proposed job of moral motivation just as well as moral considerations that we externalize or objectify. And once this general problem is recognized, we can see how it prevents a wide range of familiar proposals concerning the evolutionary function(s) of human moral psychology from helping us resolve our puzzle. Robert Frank (1988) has influentially argued, for example, that our moral emotions serve to commit us to courses of action that would otherwise not serve our immediate self-interests (ensuring that we will not need to actually carry out such courses of action very often so long as others are aware of these commitments). But this does not help address our puzzle, for there is no reason that the motivation for such commitments could not be (and be advertised as) sufficiently compelling preferences, desires, or affective states experienced as entirely subjective in character. Indeed, at least some paradigmatic instances of such commitment devices, such as jealousy, are ones that we *do* characteristically think of as subjective responses to the world (much like ice cream preferences) rather than responses to any externalized or objectified demand, as Frank perhaps acknowledges in not raising or addressing the question of why moral demands are characteristically externalized or objectified in the first place. Even if Frank's influential proposal is entirely correct, then, it does not help explain the distinctively externalized phenomenology of our moral experience.

Likewise, advocates of so-called strong reciprocity (such as Boyd, Richerson, Bowles, Gintis, and Fehr) have sometimes suggested that the central function of our moral psychology is to motivate costly punishment (especially by third parties) of those who violate prosocial norms, whereas proponents of "indirect reciprocity" (such as Alexander, Nowak, and Sigmund) have instead argued that our moral psychology motivates us to keep track of others' reputations (and manage our own) for engaging in appropriately altruistic, cooperative, and prosocial behavior. But once again, even if one or both of these influential suggestions are entirely correct, such punishment and/or reputation-tracking could be motivated equally effectively by the sorts of desires and subjective preferences that motivate the vast majority of our fitness-enhancing behaviors: We would simply need to be disposed to punish others when they violated *those subjective preferences* (concerning our interactions with them or even with third parties) and/or to keep track of (and manage) reputations for satisfying or violating them. Influential extant efforts to explain human moral psychology in evolutionary terms thus seem to fail quite spectacularly at explaining why moral demands can better do the job for which they are or were supposedly needed if they are externalized or objectified in the ways noted previously (sect. 2).[2]

A more promising recent approach by DeScioli and Kurzban (2009; 2013) suggests that the central problem to which our moral psychology offers a solution is the need for bystanders to a conflict to coordinate in choosing sides: These authors suggest that moral condemnation is a form of signaling that allows bystanders to avoid the costs of escalating conflicts by all choosing the same side, while making such side-taking responsive to actions themselves rather than the identity of actors avoids despotism by preventing the power of such allegiance from becoming permanently concentrated in the hands of particular individuals. This elaborate proposal at least promises to explain our characteristic generalization of moral judgments (i.e., why they must apply to everyone in the same

way), but unfortunately it does so in a way that ignores the most notable characteristic of those judgments, namely, their *pejorative* character: We do not *merely* affiliate with those who share our own moral commitments against those who do not – in addition, we think that we are *right* to do so and would be wrong to do otherwise. But if mere act-guided coordination of side-taking in bystanders to a conflict is the point of our moral psychology, there is no need for violators to be *wrong* rather than merely *different from us*.

To see this point more clearly, note that advocates of strong reciprocity might also sensibly claim that the generalizability of moral judgments is explained by the need for punishment of norm violators *no matter the identity of those violators*, but we would not thereby explain why norm violators could not be (generalizably) punished simply for being different from us rather than for failing to respect externalized moral obligations imposed on us all. Indeed, DeScioli and Kurzban suggest that our "dynamic coordination" strategy uses a common cultural store of proscribed actions to allow bystanders to coordinate their affiliation in disputes *without* forming persistent allegiances that allow power to concentrate in the hands of individuals. But the pejorative character of moral judgment would seem to make it *harder* rather than easier for such coordination to remain dynamic: A person whose conduct is immoral has behaved *wrongly* rather than just differently than we ourselves would have and is therefore someone with whom we will be less inclined to interact or ally ourselves in the future (see sect. 5 below), no matter the rightness of her cause the next time around. Thus, the pejorative character of genuine moral *condemnation* actually leads it to fit rather less well than simple morally grounded *local affiliation* would with the function of dynamic coordination DeScioli and Kurzban propose for our moral psychology more generally.

These challenges might simply magnify our interest in Joyce's second proposal, which begins instead from the suggestion that our *general* tendency to externalize (or "project") features of our experience is "the predictable result of natural selection's tight-fisted efficiency" (Joyce 2006, p. 128), as he seeks to illustrate explicitly with respect to sensory qualities such as color and pain. We do not project pain, he suggests, because the *point* of pain is to orient us towards a problem with the body rather than features of the world that cause pain in us. But we project the redness, warmth, and crackling sound of the fire into or onto the world as we experience it because there is simply nothing to be gained by representing these qualities instead as our subjective responses to the fire's sensation-inducing properties: "A perfectly adequate and simpler solution is if our experience presents itself as being *of the world*: of the fire being red, hot, and crackling" (Joyce 2006, p. 128). Similarly, Joyce suggests, we project moral qualities into the world, regarding our reaction to a particular behavior as *appropriate* and the behavior as *deserving* our praise or blame, for instance, because this is just phenomenologically simpler than representing such judgments to ourselves *as* subjective responses and because it motivates the relevant fitness-enhancing behaviors equally effectively. If so, the "projection" of moral demands and considerations into or onto the world itself might well be an ancestral condition from which no shift to an externalized or objectified moral psychology was ever required.

The fundamental problem with this proposed line of explanation is both subtle and illuminating, in a way that we can best begin to appreciate by recalling why the ubiquity of altruistic, prosocial, and cooperative behavior in nature once seemed to present a striking challenge to evolutionary theory more generally (one famously posed by Darwin [1871] himself in Ch. 5 of *The Descent of Man*). It seems that creatures who systematically contribute to the fitness of others at some cost to their own would be reliably outcompeted or invaded by those who welcome the assistance or sacrifices of others but do not themselves engage in any such costly prosocial or altruistic behavior. The story of how this challenge was ultimately met by the discovery of evolutionary mechanisms like kin selection, reciprocal altruism, and group (or multilevel) selection is by now familiar, but far less widely appreciated is the subsequent recognition that, in a deep and important sense, all of these mechanisms turn out to be variations on a common theme – namely, correlated interaction. As Brian Skyrms (1996) and others have emphasized, what is most *fundamentally* required to protect cooperators from exploitation in a game like the Prisoner's Dilemma or establish cooperation in a game like the Stag Hunt is that there be *some* mechanism in place to ensure that cooperators or altruists are more likely to interact with one another than with members of the population at large, and the various distinct mechanisms that ensure the persistence of altruistic or exploitable cooperative behavior in nature are just different ways of achieving *that*. Moreover, further inquiry has revealed a fascinating variety of ways in which correlated interaction can come about in naturally occurring populations ranging from bacteria to chimpanzees: everything from simple population viscosity or structure to the most cognitively complex forms of signaling, individual recognition, recollection of past behavior, and conditional strategies of interaction.

But the need to *actively maintain* such correlated interaction among cooperators and/or altruists in order to protect their behavioral dispositions from exploitation renders Joyce's (2006) parallel between the "projection" of moral demands or motivations and the color, heat, and crackling sound of the fire profoundly suspect. As the occasional struggles of those who are color-blind remind us, we have every evolutionary incentive to perceive colors and the like just as fellow group members do, but the same is just not true of prosocial, cooperative, and altruistic motivations, where the possibility of increasing our own payoffs by defecting and/or exploiting our interactive partners is omnipresent. To experience moral demands and considerations as aspects of the external world itself in the way we experience the redness, warmth, and crackling sound of the fire is effectively to *presume* that others' experiences of those demands and considerations are identical to our own, thus putting ourselves in danger of exploitation by those who experience the moral demands of a given situation quite differently than we ourselves do (or who would quickly evolve to do so). We therefore cannot remain sensitive to the need to restrict our cooperative and/or altruistic interactions only to those whose perceptions of the moral demands and obligations of a given situation are sufficiently similar to our own while simultaneously "projecting" the source of moral motivation into or onto the world itself as we do the color, warmth, and sound of the fire. Nor, therefore, can we rest content with the suggestion

that natural selection's "tight-fisted efficiency" made the "projection," externalization, or objectification of moral motivation the evolutionary path of least resistance or a likely ancestral condition from which no shift to an externalized moral phenomenology would have been required.

In fact, I will ultimately suggest that externalizing moral motivation was itself favored in ancestral environments precisely because it allowed prosocial, altruistic, and/or cooperative agents to more quickly, efficiently, and effectively correlate their interactions and thereby take better advantage of the adaptive possibilities constituting our transition to a vastly more cooperative form of social life. To evaluate this suggestion, however, we must first consider some characteristic features of human cooperative interaction itself.

4. *Homo sapiens*: Spontaneous, plastic, domain-general cooperators

Biologically altruistic, prosocial, and cooperative behavior is by no means limited to human beings, but in nonhuman organisms such behavior is characteristically restricted to a relatively narrow and well-defined set of recurring contexts (such as alarm-calling or guarding, cooperative breeding, collective defense, or food-sharing) in which the costs and benefits of quite specific forms of behavior in just these contexts have been sufficiently stable to selectively favor equally specific dispositions to engage in them. By contrast, human cooperative, altruistic, and prosocial behavior is not only *domain-general*, but also remarkably *spontaneous* and (like human behavior more generally) extremely *facultative*, *plastic*, and *flexible*: We frequently extend or adapt existing patterns of prosocial or altruistic behavior in new ways and into circumstances that are unfamiliar and/or infrequently occurring, and we routinely innovate entirely novel forms of cooperative interaction and problem-solving in both familiar and unfamiliar contexts. Indeed, among researchers who study behavior comparatively, humans are famous for the extraordinarily wide range of circumstances in which they actively and flexibly try to find ways to assist and cooperate with one another (e.g., Boyd & Richerson 2005; Cheney 2011; Fehr & Fischbacher 2003; Melis & Semmann 2010; Moll & Tomasello 2007; Silk & House 2011).

In addition, recent experimental studies (e.g., Warneken & Tomasello 2006; 2007) have shown that even prelinguistic or barely linguistic (14- to 18-month-old) children spontaneously seek out and take advantage of novel opportunities to cooperate with and assist even nearly complete strangers in achieving their goals to a much greater extent than chimpanzees do. Moreover, even very young children reliably take advantage of low-cost or cost-free opportunities to benefit strangers (Thompson et al. 1997), and although chimpanzees are certainly *capable* of similarly low-cost or cost-free altruistic behavior (e.g., Horner et al. 2011; Melis et al. 2011b), they seem to take advantage of such opportunities far less readily or reliably, and in a far more restricted set of contexts, even for group members with whom they are already familiar (Jensen et al. 2006; Silk & House 2011; Silk et al. 2005). Similarly, using an experimental apparatus and procedure specifically designed to facilitate cross-species comparisons, Burkhardt et al. (2014) found human children to be markedly more spontaneously prosocial than even our nearest primate relatives.

Indeed, Michael Tomasello has recently argued in considerable detail that chimpanzees lack at least some of the motivational and cognitive capacities that make humans' extraordinarily robust cooperative dispositions possible (Tomasello 2009; 2016). More specifically, at about 9 months of age, human infants begin to engage in "joint attentional activities" in which they not only pursue a simple shared goal, but also monitor the attention of a partner to the activity and the goal, whereas chimpanzees seem neither to establish joint attention with others nor to participate in activities with genuinely shared goals (Tomasello 2009, pp. 63–75): Unlike children, for example, chimpanzees do not seek to re-engage a partner who suddenly breaks off her participation in a coordinated problem-solving activity, even when they are highly motivated to complete the task. In further sharp contrast to even very young children, chimpanzees cannot reverse roles with a partner in a task they have learned to complete successfully (Fletcher et al. 2012), they prefer individual over collaborative problem-solving strategies even when the rewards are equal (Bullinger et al. 2011), and they have great difficulty cooperating in circumstances in which obtained resources can easily be monopolized by one party, whereas even very young children (3 years old) spontaneously divide resources obtained collaboratively into equitable shares (Warneken et al. 2011) and continue to cooperate to complete tasks and secure rewards for all participants in a task even if their own reward has already been received (Tomasello 2009, pp. 65–66). Obtaining resources collaboratively makes children but not chimpanzees more likely to share those resources with those who collaborated to obtain them (Hamann et al. 2011; Warneken et al. 2011), and children but not chimps exclude free riders from the spoils of a collective enterprise (Melis et al. 2011a; 2013; though cf. Suchak et al. 2016). Moreover, chimps show no interest in engaging in cooperative social games with no instrumental objective, whereas human children not only do so, but also will often *turn* a task with an instrumental objective into a cooperative game by replacing the reward in the experimental apparatus to restart the activity (Tomasello 2009, pp. 63–65). Thus, although chimpanzees are undoubtedly impressive individual problem-solvers who can effectively learn to use a partner's behavior as an instrumental resource, they seem to lack many of the distinctive cognitive abilities and motivations that enable humans to work *with* one another to solve such problems *together* so frequently, flexibly, and spontaneously.[3]

This remarkable affinity for spontaneous cooperation and prosociality appears even in infants of 6 months (Hamlin et al. 2007) and does not appear to be a function of learning or enculturation. The age at which prelinguistic or barely linguistic children begin to spontaneously cooperate with or assist others and the range of situations in which they do so appear not to be sensitive to the culture in which they are raised: These results have now been replicated, for example, in children from "traditional, small-scale cultural settings in Peru and India" (Callaghan et al. 2011, p. vii). And remarkably, not only does providing verbal encouragement by parents or the prospect of an external reward make no difference to whether or not children help, but also providing such a reward actually *decreases* the extent of helping on future occasions, a well-known "overjustification" effect thought to occur when an intrinsically

rewarding activity or behavior has had its intrinsically rewarding character partially displaced by the substitution of such an external reward (Tomasello 2009, pp. 7–10). Further evidence of this phylogenetic legacy of spontaneous and flexible human hypercooperation can be found in everything from the fact that humans alone seem to engage in direct, active teaching (Kline 2015) to the fact that we alone among mammals have eyes with enlarged white sclera, the better to allow a potential cooperative partner to see where our attention is directed (Tomasello et al. 2007).

Revealingly, influential earlier research suggesting that nonhuman primates had surprisingly limited theory of mind abilities (e.g., Povinelli 2000, Ch. 2) was conducted using experimental paradigms in which subjects were expected to distinguish among human experimenters or conspecifics who could or could not help them find hidden food. When the same abilities were probed in experimental paradigms in which subjects *competed* for food, the performance of nonhuman primates improved dramatically. Most nonhuman primates apparently struggle to understand even the possibility that an interactive partner might be spontaneously offering them voluntary assistance in obtaining food. In one striking example, chimpanzees (unlike 2-year-old human children) could *not* use the communicative gestures (e.g., pointing) of a human or trained conspecific partner as cues to the location of hidden food, but could find the same food if the partner was instead reaching for it unsuccessfully (Hare & Tomasello 2004). These experiments also reveal something about the humans who designed them, however, to whom it did not occur that such spontaneously prosocial or cooperative scenarios might be dramatically more difficult than competitive ones for nonhuman primates to recognize, interpret, or understand. Although chimps are certainly *capable* of impressive feats of altruism and cooperation (Melis et al. 2011b; Suchak et al. 2016), the sort of domain-general, spontaneous, frequent, and flexible cooperative and altruistic interaction so characteristic of human sociality does not seem to be a similarly ubiquitous feature of ordinary chimpanzee social life.

5. Moral externalization: A cooperation-building machine

For creatures whose behavior is as *generally* facultative and plastic as our own, the benefits of such spontaneous, domain-general, and flexible cooperative tendencies seem evident: They enable us to more quickly, easily, and (therefore) frequently identify and take advantage of novel cooperative opportunities in both familiar circumstances as well as those that arise infrequently or are even completely unprecedented. There are many unfamiliar as well as familiar things a group of hominins working together can do far more effectively, quickly, efficiently, safely, or otherwise better than if each worked alone, and many others that hominins working separately could not reliably accomplish at all. And such dispositions towards spontaneous and flexible hypercooperation would have become increasingly useful and important as they enabled ancestral hominins (unlike other primates) to radiate into a wide variety of novel ecological habitats uninhabitable by their phylogenetic ancestors and as humans evolved to become *obligate* cooperators in virtually every environment they occupy.

But of course, such hypercooperative dispositions also expose humans to enormously elevated risks of exploitation (cf. Enquist & Leimar 1993). Accordingly, even very young children do not cooperate *indiscriminately*. For one thing, cooperative and altruistic dispositions appear to be dramatically influenced by indicators of shared membership in a culturally defined in-group, and sensitivity to such indicators is found even among infants (Hamlin et al. 2013; Kinzler et al. 2007; Mahajan & Wynn 2012). Even *within* a well-defined in-group, however, spontaneous hypercooperators remain persistently at risk of exploitation by those who might happily enjoy the fruits of their cooperative and/or altruistic inclinations but not themselves share rewards, forego opportunities for exploitation, maintain costly alliances, or pay the other costs involved in the nearly limitless forms of cooperation possible for human beings in both familiar and unfamiliar or infrequently occurring contexts. A disposition to cooperate with any willing partner in each new and unfamiliar context of interaction is effectively a ticket to sustained, serial exploitation.

The transition to a hypercooperative form of social life therefore required not only that early hominins be motivated to initiate and sustain cooperative and collaborative efforts far more readily than even their nearest primate relatives in response to both familiar and novel environmental challenges, but also some mechanism for ensuring that those efforts be selectively directed (even within a well-defined in-group) towards those who share those same motivations. Faced with a rapidly proliferating array of cooperative and collaborative opportunities, increasingly frequent and consequential decisions concerning the desirability of partners would need to be sensitive not only to characteristics of those partners likely to influence chances of success (like competence, knowledge, or physical prowess), but also to the critical importance of avoiding exploitation. And recent evidence suggests that our moral psychology plays a central and distinctive role in mediating such preferences.

Perhaps most importantly, Skitka et al. (2005) showed that the desirability of partners in social interaction generally is mediated by our awareness of the extent of our distinctively *moral* disagreement with them. These experimenters found that subjects preferred greater social (i.e., closeness in relationship) and even physical distance from those with whom they had identified moral disagreements *over and above* the social and physical distances they preferred from those with whom they differed in equally strongly held non-moral attitudes (also controlling for importance, certainty, and centrality of these attitudes). A stronger moral conviction on the issue in question was associated with greater intolerance of attitudinally dissimilar others in both intimate (e.g., friend) and distant (e.g., owner of a store one frequents) relationships, and these effects were robust when experimenters controlled for age, gender, individual differences in political orientation, and even tendency to see issues overall in a moral light. Measures of attitude strength alone, however, had an inconsistent relationship with preferred social distance, tending to vary across both relationship types and the specific issue, but more often than not were simply unassociated with preferred social distance from those who held conflicting attitudes. It therefore seems to be whether a

particular attitude or conviction is seen as *moral* in character that most reliably determines whether it inclines us to include or exclude those who do or do not share it from social interaction with us.

Skitka et al. (2005) also found lower levels of goodwill and cooperativeness in groups that were heterogeneous with respect to a moral as opposed to an equally strongly held non-moral attitude and a greater inability to generate procedural solutions for resolving disagreements. Importantly, however, groups seeking to resolve non-moral disagreements were constructed by the experimenters so as to be in fact heterogeneous with respect to a moral attitude as well (regarding capital punishment or abortion), but were charged only with finding procedures for resolving their non-moral disagreement instead. Thus, it is not the mere existence but the *awareness* of distinctively moral disagreement that depresses goodwill and cooperativeness in the members of such groups substantially below that of groups with equally strong non-moral disagreements.

Once again, however, moral convictions would not need to be *externalized* in order to mediate our preferred social distance or enthusiasm for social interaction with others: It would be simpler and more continuous with the rest of our motivational psychology for us to regard those who do not share a particular subset of our own attitudes as merely *different* from ourselves in ways that we dislike or find unappealing, rather than as violating demands that we regard all agents as unconditionally responsible for satisfying. Hence, even if we allow that a special conceptual category of motivations and attitudes emerged to mediate our enthusiasm for potential partners in various forms of social interaction, we must still ask how and why such motivating attitudes became externalized into distinctively moral *obligations*.

The key to resolving this further puzzle is to recognize that although we might more simply and easily use subjective preferences or desires to motivate *either* our own conformity to a given set of norms or standards of behavior, *or* our social exclusion of those who do not similarly conform, experiencing moral motivation as externally imposed on both ourselves and others simultaneously is nonetheless an extremely effective and efficient way to ensure that these otherwise distinct motivations arise together and remain systematically connected to one another even as *particular* norms themselves come to be modified, applied, and extended in new ways and into new circumstances, and even as entirely new norms come to be adopted (or come to be moralized) in the first place. That is, experiencing moral demands and obligations as externally imposed simultaneously on both ourselves and others ensures that if I myself come to be motivated to conform to a particular norm or standard of behavior that I experience as distinctively moral in character, I *automatically* demand that others conform to it as well, judging them to be less attractive potential partners in social interaction generally if they do not. If instead my motivation to engage in some particular exploitable prosocial behavior or conform to a given norm were experienced as a purely subjective preference or desire, I would not thereby automatically require that desirable social partners must also engage in that behavior or conform to that norm, leaving me open to exploitation by those who are not similarly motivated. (Similarly, if my motivation to exclude those who do not adhere to or accept a given norm were a merely subjective preference or desire, it would not automatically ensure that I myself am also motivated to adopt or adhere to the norm, ultimately leading to my own exclusion from valuable opportunities for cooperative social interaction by those who externalize it.) Thus, I regard my affinity for ice cream or for chocolate over vanilla as a subjective preference or desire that I neither expect nor require you to share. But I experience my own opposition to Nazis or slavery (or selfish hunting partners, or cowardly defenders of the tribe, or untrustworthy coalition partners) as a response to an externally imposed demand by which I insist that you be similarly motivated.[4] And if I become convinced that you are not motivated by the same distinctively externalized moral demands that I myself am, I will find you a correspondingly less desirable candidate partner for social interaction generally.[5]

The creation of a novel conceptual category of norms or standards of behavior to which I hold both others and myself responsible simultaneously thus established a mechanism for safely *extending* prosocial, altruistic, and cooperative behavior in new ways and into new contexts. That is, it ensured that each such extension was automatically protected from exploitation *from its inception* rather than exposing those motivated by it to exploitation unless and until they evolved a distinct and independent subjective desire or preference to exclude those who do not share that disposition from social interaction with them. From an evolutionary point of view, our characteristic externalization of moral motivation thus represents a mechanism for establishing and maintaining *correlated interaction under plasticity* – that is, for establishing and maintaining correlated interaction between those prepared to accept any given norm, even as the content of the particular norms we embrace remains free to change over time as we apply them in novel ways, extend them into novel or unfamiliar circumstances, and innovate entirely new norms governing our social interactions. As noted previously, such correlated interaction is the crucial condition which must be satisfied in order for cooperative, altruistic, and otherwise exploitable prosocial dispositions to emerge and/or remain evolutionary stable.

This proposal is strongly supported by Goodwin and Darley's (2012) further finding of a direct relationship between the *degree* or *extent* to which we externalize or objectify a given norm that we ourselves accept and the extent to which we are prepared to exclude those who do not accept it from social interaction with us. These authors found that even when they controlled for the strength of agreement with a given moral belief, the more *objective* subjects held the belief in question to be the less comfortable they were with social engagement with (e.g., being a roommate of) someone who disagreed and the more immoral they regarded such a person as being (see also Wright et al. 2014). These results nicely complement Skitka et al.'s finding (see above) that *within the category of externalized norms* (but *not* more generally) we find a direct relationship between the *strength* of our own commitment to a given norm and the extent to which we exclude those who do not follow it from social interaction with us. That is, our enthusiasm for social interaction with those who violate a given moral norm decreases not only in proportion to the strength with which we ourselves are committed to and motivated by that norm (Skitka et al. 2005), but also in proportion to the extent to which we ourselves have *externalized* or *objectified* it into a distinctively moral obligation (Goodwin & Darley 2012).

Note that this proposal obviates the need for any costly punishment of norm-violators in order to get the benefits of externalization off the ground: Each individual simply *protects herself* from exploitation by excluding those who engage in either transgressions of commission (e.g., theft) or omission (e.g., failing to share) from her own circle of voluntary social interaction. As the pool of available partners comes to consist of an ever-higher proportion of exploitative or otherwise undesirable partners rejected by other community members, the selective pressure to identify and avoid undesirable partners becomes even stronger, and such discrimination in the desirability of partners will serve to carve out *networks* of cooperative interaction within a single community. Baumard et al. (2013) describe ethnographic evidence suggesting the ubiquity of such partner choice among contemporary hunter-gatherer groups, noting that "punishment as normally understood … is uncommon in societies of foragers … in these societies, most disputes are resolved by self-segregation" (p. 66; see also Guala 2012). More costly forms of punishment can await the explicit recognition and/or agreement that persistently exploitative or norm-violating partners represent a public nuisance or a threat to the public welfare that can be addressed by coordinated action at the community level with relatively low costs or risks to each individual member.

Moreover, this account explains many of the distinctive central features of our moral psychology noted previously (sect. 2), such as the moral/conventional distinction and the distinction we draw between moral properties and response-dependent properties (like being icky or boring) seen as constituted by our heterogeneous subjective reactions to objects or events in the world. Perhaps most importantly, it offers a natural explanation of the systematic differences in objectivity that Goodwin and Darley (2008) found between judgments of taste/preference, convention, morality, and empirical fact. While judgments of taste or preference carry no demand for or expectation of intersubjective agreement, conventional norms are imposed (explicitly or implicitly) on the members of a group by that group itself (or a recognized authority within it) and thus apply to its members regardless of the subjective preferences and desires of any particular individual but not those of the group as a whole. But we experience prototypical moral norms, demands, and obligations as imposed *unconditionally*, irrespective of not only our own preferences and desires, but also those of any and all agents whatsoever, including those of any social group or arrangement to which we belong: We apply them not only to ourselves and our fellow community members, but also to absolutely any candidate social partner with whom we might interact, no authority is capable of suspending the demands they impose, and we regard violations of them as more serious and more deserving of punishment than violations of the merely conventional rules that we collectively self-impose. There is correspondingly less room for intersubjective disagreement without error concerning them, indeed nearly as little as in the case of empirical or scientific fact.

The characteristic plasticity of the process by which particular norms become moralized also helps to explain why there is considerable variation in the degree of objectivity (see Goodwin & Darley 2012; Wright et al. 2013; 2014) ascribed to different norms by different individuals: It seems natural to see this as a consequence of differences in the *extent* to which particular norms have become moralized by different individuals or groups in the process of acquiring them. Whether and to what extent particular norms become moralized in an individual will presumably be quite strongly influenced by the cultural environment in which she is raised, but a wide range of further factors might also influence whether or to what extent any particular norm becomes moralized (such as the presence of a strong affective response; see Nichols 2004). Such factors might also help explain the various curiously partial or incomplete forms of moralization discussed by Haidt et al. (1993) and Nichols (2004), while the comparatively high importance of avoiding exploitation might help explain Goodwin and Darley's (2012) finding that subjects typically judge the wrongness of immoral acts as significantly more *objective* than the goodness of positive moral acts even when the *strength* of their convictions in the wrongness or rightness of each act is precisely the same.

This remarkable plasticity in the acquisition, externalization, and application of distinctively moral norms does not, of course, show that there are *no* constraints on the content of the norms that can become (or cease to be) moralized, and it is certainly noteworthy that norms regarding harm and fairness appear to be either universal or found across a much wider range of human societies than others (see Nichols 2004). Indeed, we are now finally well-positioned to understand why human moral norms are so *often* concerned with harm, justice, fairness, and protecting the rights of victims, even though they *need not* be so concerned in order to trigger some or all aspects of the moralized response: Although it may be possible for nearly any norm or behavior to become moralized, norms protecting spontaneous prosociality and cooperation from exploitation are those which must be preserved to maintain the most important evolutionary benefits of moralization itself, and norms concerning harm, fairness, justice, and the rights of victims seem like natural candidates for this role.

This suggestion leaves open a wide variety of mechanisms, however, by which such cooperation-enhancing norms could become and remain ubiquitous among human cultures. This might be a straightforward consequence of the fact that existing sets of norms have come from pre-existing sets of norms by descent with modification, or it might be a product of convergent cultural evolution favoring groups better able to foster fitness-enhancing cooperation among their members. It may be that we are biologically predisposed to regard norms with particular kinds of content as moral in character, or their moralization might be the product of the sort of strongly biased learning mechanism that threatens to make the very distinction between learning and predisposition seem unhelpful or inapposite. But we need not here prejudge the answer to the daunting question of which of these mechanisms or what combination of them is most likely to have produced the distinctive patterns we find in the moralization of particular kinds of norms among human cultures and societies, nor is any single explanation likely to be correct in all cases. The ubiquity of norms concerning harm, fairness, justice, and the rights of victims can be explained by recognizing the relative *importance* of such norms in fostering fitness-enhancing cooperative, altruistic, and prosocial interaction among humans, even if we remain agnostic about the specific evolutionary mechanism(s) at work.

The account also leaves open a wide range of important further empirical questions. For example, it may be that the further distinction between conventional and moral norms emerged phylogenetically following the constitution of a more general category of normative beliefs, or it may be that some norms originally seen as merely conventional became externalized or objectified in the ways we have noted to create the distinctive conceptual category of moral norms. It leaves open a further and distinct question concerning how moral norms acquire their characteristic status in the course of individual ontogeny. It could be that at a particular point in human development a subset of existing norms become externalized or objectified and thereby acquire a distinctively moral character, but there is at least some evidence that children instead start out "reifying" (externalizing or objectifying) *all* social norms and rules and simply learn over time to treat some (but only some) particular norms and rules as merely conventional instead (see Gabennesch 1990). Once again, the fundamental attractions of the idea that moral externalization establishes or preserves correlated interaction between those prepared to adopt any particular norm (or to extend an existing norm in a particular way) while leaving the content of those norms free to evolve on a cultural (rather than phylogenetic) timescale do not depend on the answers to these admittedly important further questions.

Finally, this proposal explains an intriguing pattern of variation in our characteristic generalization of moral norms identified by Sarkissian et al. (2011). These authors found that subjects are much more likely to insist that at least one party to a moral disagreement must be in error if that disagreement arises among the subject's peers than if it arises instead between a peer and a member of an isolated Amazonian tribe or an extraterrestrial civilization. Although the authors interpret these results as showing that "the folk" become increasingly relativist as they are forced to confront alternative moral frameworks or perspectives, the account presented here would suggest instead that the demand that others must share our own moral convictions becomes progressively more relaxed as salient contextual cues render the realistic possibility of significant interaction with the agent who makes a contrary moral judgment seem ever more remote. This is perfectly intelligible, of course, if human moral psychology evolved to mediate the evaluation and selection of partners in social interaction: An act can be wrong for agents in all times, places, and cultures even if the demand for intersubjective agreement among *judges* of that act is dramatically weakened when salient further considerations substantially reduce or preclude the possibility of the dissenting judge serving as a partner in significant forms of social interaction in any case. This analysis suggests that if salient cues instead emphasize the realistic prospect of ongoing social interactions with those who dissent from the subject's own moral judgments, we should see a corresponding increase in the extent to which those subjects insist that at least one party to the disagreement must be in error.

6. Becoming human

I have suggested that the emergence of a distinctive conceptual category of motivations or attitudes which others must share (on pain of being judged less desirable partners in social interaction generally) served to protect the extension of increasingly spontaneous cooperation and other forms of adaptively advantageous prosociality from exploitation and thereby facilitated a transition in the hominin line to a vastly more cooperative form of social life. But the degree of protection against exploitation afforded by such moral externalization in the course of extending prosociality and cooperation is strictly limited by the accuracy and detail of the information agents are able to acquire about one another's distinctively moral commitments. Perhaps some of this information might be gleaned merely by observing one another's behavior, but recognizing how important the extent and detail of such information are to the degree of protection against exploitation that it provides invites us to notice the otherwise quite remarkable proportion of our ordinary conversational interaction devoted to soliciting and providing just this information to one another.

As the anthropologist Robin Dunbar has documented, an astounding 60–70% or more of our ordinary conversation is taken up with discussion of "who-was-doing-what-with-whom and personal social experiences" (Dunbar 1996, p. 174; see also Emler 1990; 1994; 2001) – an activity colloquially called "gossip" that is extremely widespread in traditional hunter-gatherer societies (Haviland 1977) just as it is in our own. What these researchers have not sufficiently emphasized, however, is that such gossip does not consist simply of *reports* of our own and others' behavior: We are also extraordinarily interested in expressing our own and hearing others' attitudes towards, rationales for, and justifications of that behavior. I suggest that we judge the desirability of potential partners in cooperative endeavors and other social entanglements neither exclusively nor even primarily by gathering intelligence on cheats, liars, free-riders, and other undesirables from our interlocutors (cf. Dunbar 1996; 2004; Enquist & Leimar 1993), but instead by evaluating the extent to which *those very interlocutors share our own attitudes and reactions* towards these and other specific examples of behavior, including observed, reported, rumored, and even merely imagined or counterfactual instances.

After all, we eagerly engage in such practices of moral evaluation, instruction, and advertisement *with our peers* (cf. Sarkissian et al. 2011; and see sect. 5 above) regarding conduct in rumors, stories, films, television programs, legends, fables, and fairy-tales regarding characters who are fictional, long dead, or otherwise present no possibility themselves of serving as actual prospective partners in social interaction. Of course, we engage in such discussion because we find it "interesting," but the extraordinary interest we take in these particular subjects of conversation can no more be simply taken for granted than facts about what we find painful, pleasurable, or delicious. It is obviously useful for creatures who live in stable social groups and depend upon one another to take an interest in who does what with whom, but our otherwise remarkable interest in sharing, hearing, validating, ratifying, and calibrating our opinions *about* such doings with one another arises only because we thereby learn or establish as much or more about the moral convictions of the prospective cooperative partners who *relate* and *respond to* such gossip as we do about those whom they discuss, and because this affords us the opportunity to similarly advertise our own

moral convictions, negotiate points of substantive agreement and disagreement, and try to establish the *widest* possible range of genuinely shared moral commitments with our interlocutors. In fact, such moral advertisement, evaluation, navigation, and negotiation seem to make up a substantial proportion of ordinary conversational interaction among humans.

Indeed, I suggest that such practices of moral advertisement and evaluation have played a crucial role in facilitating a range of further distinctive features of human cognition and communication. Because we so often spontaneously extend cooperative or exploitable interaction in novel ways and into novel or infrequently occurring contexts, we do well to learn as much as we can about the dispositions and judgments of potential cooperative partners (as well as to advertise our own) across as wide a range of circumstances as possible, including those that are at present merely imagined, hypothetical, or counterfactual. And of course, the ability to advertise, assess, and negotiate agreements concerning such dispositions or judgments effectively would be dramatically expanded not only by the sophistication of our linguistic abilities, but also by other cognitive abilities either unique to humans or present in them to an extraordinary degree, such as theory of mind abilities, perspective-taking, counterfactual reasoning or simulation, chronesthesia (mental time travel), and abstract categorical thought.

I do not mean to suggest that the demands of cooperation and moral advertisement were ultimately *the* key driver of higher cognitive abilities in humans. Indeed, Sterelny (2012) has argued convincingly that the extremely rapid cognitive evolution that produced behaviorally modern humans at least 40,000–50,000 years ago was most likely the product of one or more feedback loops, in which the acquisition or sophistication of some cognitive abilities made the acquisition or sophistication of others even more valuable in an ongoing cycle. I do want to suggest, however, that a feedback loop of precisely this sort connects the various sophisticated linguistic and cognitive capacities described in the previous paragraph, with improvements in each rendering further improvements in the others even more useful *specifically for* the purposes of moral advertisement, evaluation, and the navigation or negotiation of shared moral judgments concerning particular instances of behavior (and their possible variations), character traits, and (much more rarely) general and abstract moral commitments.[6] Increasing the sophistication of any or all of the distinctive cognitive and linguistic abilities noted above would certainly have allowed us to increase the range, complexity, and precision of the information we provide one another in the course of such moral advertisements, evaluations, and negotiation. Therefore, *among other important advantages*, such cognitive and linguistic sophistications would have allowed us to better correlate our interactions with desirable cooperative partners, reduce the risk of exploitation or failed cooperation in new and unfamiliar contexts, and thereby dramatically expand the range of circumstances into which our cooperative dispositions could be safely and effectively extended.[7]

Of course, agents have incentives to represent themselves as more desirable interactive partners than they really are, and in fact people do (intentionally or not) overestimate how ethically they themselves will behave in a given set of circumstances while predicting others' behavior more accurately (Epley & Dunning 2000). But such misrepresentations must typically remain fairly modest in character because they are by no means cost-free. Other members of the community have a wide variety of more-or-less continuous opportunities to compare the *actual* behavior of each community member with moralized attitudes she has *previously* expressed regarding behavior in circumstances that were at the time merely reported, rumored, imagined, or counterfactual, and damaging the credibility of one's moral advertisements is another way to become marked as an undesirable cooperative partner by others in the community. This also helps to explain why the extension of moral concern and spontaneous prosociality is particularly sensitive to indicators of in-group membership: Especially in ancestral environments, it would generally be all and only the members of an in-group with whom such moral advertisements and evaluations could be regularly exchanged.

This view of moral advertisement also explains our otherwise somewhat puzzling attitude towards hypocrisy. If I declare that it was wrong for you to abandon the camp instead of defend it or to keep all of the best nuts for yourself and then I proceed to do just the same in your position, my *hypocrisy* is itself a moral wrong *over and above* my cowardice or my greed. But of course, these behaviors may indeed deserve my condemnation even if I myself am unable to resist engaging in them! This puzzlement dissolves, however, when we note that *public criticism* of others' moral failings expresses the recognition of an unconditional, generalizable, authority-independent externalized demand to which the author of such criticism is (therefore) automatically subject as well *and thus serves as at least a prima facie commitment that she herself would not respond to the same circumstances in the same fashion.* In this way, our advancing linguistic and cognitive abilities would have enabled simple moralizing gossip among humans to convey genuine (albeit implicit) commitments regarding our own future behavior, and undertaking such public commitments would have become yet another way to increase one's attractiveness as a potential partner in exploitable social interactions.[8] Indeed, such moral criticism seems a natural and plausible precursor to fully explicit practices of promise-making and commitment more generally. This also illuminates the distinctive role played in our social lives by practices of apology and repair, by which we seek to reassure others that we do indeed recognize a norm or obligation that we take ourselves to have violated in some particular case and that we can still be trusted to see it as placing demands or constraints on us in the future (despite recent contrary evidence). This is part of why such apologies are not cost free and why we are often reluctant to make them, as well as why they strike us as empty or insincere unless accompanied by modified behavior.

I have suggested not only that plasticity in the *content* of the norms or behaviors we moralize is what allows humans to so spontaneously, frequently, and flexibly take advantage of novel opportunities for cooperative and altruistic interactions, but also that the price of protecting that plasticity from exploitation is the continuous monitoring of others' and advertisement of our own existing moral commitments in the course of ordinary social interaction. If this analysis is correct in even its fundamentals, it seems possible but

unlikely that nonhuman organisms externalize distinctively moral motivation in the same ways that humans do. Although in principle nonhuman primates might better correlate their cooperative and other prosocial behavior if their own motivations for those behaviors arose only in conjunction with comparable demands on the behavior of others[9] (and vice versa), it would nonetheless seem that the most salient advantages of such externalization arise for humans precisely because we so frequently and reliably face the need to correlate such interactions when (1) existing norms are applied in novel ways to familiar circumstances, (2) existing norms are adapted or extended into unfamiliar and/or infrequently occurring contexts in particular ways, and (3) entirely new norms come to be adopted (or come to be moralized) in the first place. After all, if exploitable prosocial behavior arose only in consistent ways across a narrow and/or fixed range of recurring contexts, the extent of an agent's subjective motivation to pursue a given course of action in those contexts and the extent to which she demands that desirable social partners do so as well could independently converge on stable values and then remain there indefinitely. But externalization ensures that this crucial relationship is immediately established and maintained even as we *extend* our exploitable prosocial behavior in novel ways, into novel contexts, or by means of novel norms of behavior. Thus, it is the *combination* of the extraordinary plasticity, flexibility, and facultative character of our behavior with the extraordinary spontaneity and frequency of our prosociality that renders moral externalization so enormously advantageous for human beings. More generally, the selective advantage of externalizing a distinctive category of motivations so as to protect them from exploitation increases directly with the spontaneity, frequency, and plasticity with which a creature engages in cooperative, altruistic, and other exploitable forms of prosociality, and present evidence (tentatively) suggests that humans' enormously spontaneous, frequent, and plastic prosociality is simply not a pervasive feature of chimpanzee social life. Moreover, the advantages of such externalization seem to depend on the extent, precision, and accuracy of the information we can exchange with one another concerning our distinctively moral commitments, and here again it seems that humans are far better equipped to exchange such information than are chimps or other nonhuman primates. Indeed, perhaps this makes it less surprising that nonhuman primates seem to lack some of the fundamental cognitive capacities (e.g., joint attention) that facilitate such spontaneous, flexible, and domain-general cooperative interaction in humans (see sect. 4).

Of course, the very same plasticity in the content of our moralized norms that makes externalization so important for human cooperation allowed (and continues to allow) norms of behavior having nothing whatsoever to do with protecting cooperation, altruism, or prosociality from exploitation to *become* moralized: attitudes concerning homosexuality, say, or gambling, or masturbation, or female genital mutilation, or food taboos, or whatever else a particular person or social group might ultimately seek to convince us (for whatever reasons) are unconditionally prescribed or proscribed in the distinctively externalized manner of moral demands and obligations. And of course, human cultures famously demonstrate an incredible range of variation in the particular norms that have come to be moralized by their members. That is, although human beings reliably begin moralizing *some norms or others* at a particular point in their cognitive development, which *particular* norms become moralized is highly variable and depends sensitively on a complex surrounding social environment.

On the other hand, this sensitive dependence of the content of any particular person's norms on her social and cultural environment also ensures considerable homogeneity in the norms embraced by the members of any given cultural or social group at any given time. It might be tempting to view this group-level variation through the lens of cultural group selection, supposing (as Darwin famously did) that differences in the moral norms to which the members of distinct groups are committed generate selective advantages and disadvantages for each group in competition with others. But many, if not most, of the differences in the content of the particular moral convictions embraced by different social groups would seem to have little or no intrinsic adaptive significance. Indeed, the truly extraordinary range of variation in the content of the moral convictions embraced by even closely related social groups (e.g., Henrich et al. 2010, p. 61) strongly suggests that the content of a group's moral convictions ultimately came to play a further role simply as an effective and reliable marker of in-group membership itself, just as it has in the case of linguistic dialects (cf. Boyd & Richerson 2005, Ch. 6; Dunbar 1996, p. 168f); styles of dress; rituals; tattooing, scarification, other bodily markings or adornment; and virtually every other sociocultural practice with easily recognizable variants whose differences have little or no intrinsic adaptive significance. Cultural evolution routinely takes advantage of nonadaptive or very weakly adaptive variation in such practices to serve as in-group markers helping to identify the members of one's own communities (at a variety of scales or sizes) and distinguish them from members of other such communities. Thus, once the general conceptual category of moralized norms was in place, selective pressures would have favored variation *for its own sake* in the particular norms moralized by different groups, just as they have in the case of many other similarly plastic social and cultural practices, generating in turn the extraordinary range of variation in the moral norms we find among even closely related human cultural groups. Following the inception of large-scale agriculture, it seems that this same plasticity was easily co-opted to establish, extend, and reinforce power asymmetries by means of new norms (perhaps especially new role-dependent and/or asymmetric norms) as human groups became larger, more complex, and more hierarchical.

On this view, our moral psychology represents something of a kludge, whose rough edges, scars, and imperfect fit with the rest of our motivational psychology is revealed in our endless philosophical puzzlement (at least since Plato) concerning how *anything* could have the distinctive combination of characteristics that we seem to unreflectively attribute to moral obligations, facts, and properties. The phenomenology of moral demands combines elements from each side of the fundamental division between the subjective and the objective in ways found nowhere else in nature: Perhaps most saliently, we experience such demands as *imposed* on us without regard for our preferences in something like the way objective empirical facts are, but as nonetheless intimately connected to our

motivational states in ways that such empirical facts never are. Indeed, moral judgment and motivation are so deeply entangled that philosophers have sometimes argued that simply recognizing the existence of a moral demand or obligation automatically entails being motivated by it, unlike the existence of any ordinary matter of empirical fact. It is by now a familiar observation that evolution proceeds by making incremental modifications to whatever materials are already close at hand. And it seems eminently plausible to suppose that among creatures who go in for cognitively complex forms of representation at all, the most fundamental division embodied in their experience will be that between representations of how things stand in the world itself (e.g., "the cat is on the mat") – from which others cannot dissent without *somebody* being wrong about *something* – and our subjective reactions to those states of the world, like pain or the desire for ice cream, that are intrinsically motivating but carry no such demand for intersubjective agreement. This fundamental division was surely the background phenomenological and conceptual framework into which moral norms, demands, considerations, and commitments had to be shoehorned by the conservative, tinkering process of evolution; and this in turn explains why we find their curiously hybrid character so endlessly puzzling. This ongoing puzzlement illustrates how and why our moral cognition is perhaps an imperfect adaptation, but of course, it need not be perfect to confer a selective advantage upon those who possess it. An old Darwinian joke tells of two hikers being chased by a bear, one of whom stops to put on snowshoes. "Are you crazy?" says the other, "You can't outrun a bear in snowshoes." To which the first hiker replies: "I don't have to outrun a bear – I only have to outrun *you*."

NOTES

1. These results are somewhat difficult to interpret, however, as the harmful behaviors in question are presented as elements of widely accepted punishments, military training, and in other ways that might undermine the extent to which subjects see them as norm violations at all.

2. Stanford (2017) argues that a version of this same problem undermines Michael Tomasello's (2016) recent attempt to explain the characteristic externalization or objectification of our moral motivations by appealing to the detailed evolutionary genealogy he proposes for human moral psychology more generally.

3. Although Christophe Boesch (2005) has famously argued that chimps engage in cooperative group hunting of red colobus monkeys in the Tai Forest, Tomasello and others argue that this behavior is best interpreted as mere coordination, with each chimpanzee chasing prey from the most advantageous position possible, given the positions already occupied by others, as in the group hunting of other social mammals like lions and wolves (Moll & Tomasello 2007, p. 642; Tomasello 2009, pp. 61–63, see also 79–80).

4. Again we might ask, "Why not just have a special category of subjective preferences, desires, and motivations that we *do* require others to share if we are to regard them as desirable partners in social interaction?" But this is simply a *re-description* of the sort of moralized motivations I have just characterized, including the salient features (like intolerance of intersubjective disagreement) that lead us to *resist* thinking of such motivations as mere subjective preferences or desires.

5. Baumard et al.'s (2013) "partner choice" models for the evolution of fairness clearly recognize the importance of both establishing one's own motivation for prosocial behavior and for using the corresponding behavior or motivation in others as a criterion in partner choice, but not how externalization serves to ensure that these distinct motivations remain tightly connected even as particular norms or standards of behavior come to be modified, applied, and extended in new ways and into new circumstances, and even as entirely novel norms come to be adopted (or come to be moralized) in the first place.

6. There is an obvious affinity between my proposal and the sort of sophisticated expressivist meta-ethics ably defended in recent decades by thinkers like Alan Gibbard and Simon Blackburn. However, these authors are primarily concerned to provide an account of the *meaning* of our moral discourse, whereas I am instead concerned with offering an empirical explanation of a central aspect of our moral cognition and phenomenology. Indeed, the explanation I have offered fits nearly any account of the meaning of our moral language, and its attractions do not depend on the appeal of this or any other meta-ethical view.

7. This also seems a plausible candidate for gene-culture coevolution, as one of the most remarkable features of the environments in which human beings are raised is the extent to which they are given instruction, encouragement, and opportunities to *practice* the component skills figuring in this feedback loop.

8. This account also gains striking support from the evidence presented by Jordan et al. (2017) in support of what they call a "false-signaling" theory of hypocrisy.

9. Intriguingly, chimpanzees appear much better able to cooperate spontaneously if they are able to select their own partners (see Suchak et al. 2014).

ACKNOWLEDGMENTS

This paper has benefited from discussions with an extremely large number of people over a long period of time, particularly Penelope Maddy, Kim Sterelny, Brian Skyrms, and members of my graduate seminars and the Baboon Club reading group at University of California, Irvine. Others I can recall at the moment include Kevin Zollman, Elliott Wagner, Tucker Lentz, Cailin O'Connor, Nathan Fulton, Bennett Holman, Jonathan Birch, Pat Forber, Marta Halina, Luke McGowan, Greg McWhirter, Peter Kirwan, Joshua Knobe, John Doris, Carl Craver, Ron Mallon, Edouard Machery, Simon Blackburn, Patricia Marino, Jessica Isserow, Michael Poulin, Peter Ditto, Collin Rice, Joan Silk, Alex Rosenberg, Walter Sinott-Armstrong, Jim Woodward, Yoichi Ishida, Jim Bogen, Sarah Brosnan, Alyssa Stanford, Gillian Barker, Elizabeth O'Neil, Joe McCaffrey, Casey Stanford, Rebecca MacIntosh, Gerald Dworkin, Robert Audi, Amy Berg, Rick Grush, Michael Hardimon, Gila Sher, Dick Arneson, Sean Greenberg, and Deena Weisberg, as well as audiences at the University of Pittsburgh; University of Western Ontario; University of Washington; University of California, Irvine's Center for the Scientific Study of Morality; University of California, San Diego; Washington University in St. Louis; the UClub Forum at University of California, Irvine; and Duke University. A large number of especially helpful and constructive *Behavioral and Brain Sciences* reviewers, including editor Paul Bloom, also richly deserve acknowledgement. Special thanks are owed to Richard Joyce, whose provocative book got me wondering about the question, and to Felix Warneken, who generously permitted me to use video clips of his experiments with chimps and children in talks on this material. I am also extremely grateful to have been able to work extensively on this paper during a sabbatical year granted by the University of California, Irvine, and while serving as Senior Fellow at the Center for the Philosophy of Science at the University of Pittsburgh. I would also like to thank the Australian National University, where I completed substantial work on this paper while serving as a Visiting Fellow, as well as some extraordinarily kind and generous interlocutors at the eighth Philosophy of Biology at Dolphin Beach conference. Finally, heartfelt thanks to my moon and stars and to my beamish boy, without whose hard work, patience, love and support nothing else I do would be possible.

Open Peer Commentary

Moral cues from ordinary behaviour

doi:10.1017/S0140525X18000018, e96

Suraiya Allidina and William A. Cunningham

Department of Psychology, University of Toronto, Toronto, Ontario, M5S 3G3, Canada.

suraiya.allidina@mail.utoronto.ca cunningham@psych.utoronto.ca
http://socialneuro.psych.utoronto.ca

Abstract: People want to form impressions of others based on their moral behaviours, but the most diagnostic behaviours are rarely seen. Therefore, societies develop symbolic forms of moral behaviour such as conventional rituals and games, which are used to predict how others are likely to act in more serious moral situations. This framework helps explain why everyday behaviours are often moralized.

People moralize many curious things: standing quietly for the national anthem, dressing modestly, standing to the right on the escalator, and separating recycling from garbage. While some of these actions have clear utility for society, others seem to have a relatively arbitrary nature with a confusing moral status. Indeed, if asked, people can even moralize whether someone wears a sweater-vest or not (Van Bavel et al. 2012). The target article by Stanford, with extension, provides a framework to help understand why everyday behaviours and rituals can take on moral significance. In the target article, Stanford proposes that the externalization of moral demands, which shifts experience from one of internal preference to one of obligation, evolved as a way for people to identify potential partners for productive interaction. Observing behaviour in moral situations provides a great deal of information about a person's character and role in the social collective; someone who donates a kidney to a complete stranger (Marsh et al. 2014) can likely be trusted in times of crisis, whereas someone who loots the apartments of a burning building has already shown themselves to be untrustworthy in social exchange. Comparing people to moral scripts tells us a lot about not only their preferences, but also their moral dispositions (Uhlmann et al. 2015) and the degree to which they have internalized an objective set of society-building rules.

Yet, although behaviour in moral situations is extremely diagnostic, the chances of observing someone make decisions in extreme situations is rare. This presents a challenge – we need to develop models of others that allow for an understanding of their moral character, but we are not often given the experiences that are most diagnostic in forming such representations. To help people make predictions about the moral character of others – predictions that are necessary for productive interaction – we propose that societies develop norms, games, and conventional rituals that allow for moral behaviour to play out in a relatively more symbolic form. By observing symbolic forms of moral behaviour, people can infer the extent to which others understand and are willing to conform to the moral demands imposed by society. On this view, everyday rituals, such as standing for the national anthem or shaking hands at a job interview, transmit the social norms of society (Rossano 2012) and indicate whether others will follow rules (Watanabe & Smuts 1999). Games provide a similar function, allowing individuals across species to learn the standards they will be held to in society (Allen & Bekoff 2005; Bekoff 2001; Rakoczy 2007; Rakoczy & Tomasello 2007) and to predict whether others are likely to follow these standards. Applying this to the moral domain, these low-stakes games allow people to form impressions and predict how others will act in more serious moral situations. In this way, we can collect social information through games and ordinary rituals, using others' adherence to such rituals as indicators of potential adherence to moral norms.

A prominent example is the ritual of gift exchanges on birthdays: the continued yearly passing of a $20 gift card between two people signals that they continue to value reciprocal exchange. A friend who neglects their end of the exchange signals low trustworthiness and therefore might not be a good choice as a business partner. The unspoken rules within these rituals become moralized in themselves – the friend who disregards conventional gift-giving may be viewed as less invested in the group.

The rituals through which we signal moral character are externally imposed by society, providing an easily observable objective standard to which others can be held. Those who abide by the standards are seen as moral rule-followers, whereas failing to meet these standards results in condemnation of one's character and reduced opportunities for productive cooperation. Although such moral signaling may be costly, it persists because of the tendency to focus on others' actions rather than words in determining moral character (Henrich 2009). Therefore, despite any potential inconvenience, we dogmatically abide by the rules within these rituals in order to signal to others that we can be trusted.

Although the standards implied by these rituals are externally imposed, the specific behaviours that are moralized will depend on the moral principles that a particular person values. One important distinction is between individual-based moral foundations, centering around harm and fairness, and group-based moral foundations, focusing on authority, in-group loyalty, and purity (Haidt & Joseph 2004; 2007). The rituals that people use as cues for moral norm-following will likely differ based on the principles they are primarily concerned with – someone who values individual-based moral foundations may judge others who cheat at board games or take the last piece of cake, while someone who values group-based foundations may be more concerned with whether others dress appropriately or stand for the national anthem. In deciding which rituals are most diagnostic of moral behaviour, the social information that is collected can be tailored to each individual's moral concerns.

Critically, this conceptualization provides a framework for understanding why people sometimes moralize everyday behaviours. Because truly exceptional behaviours are rare, we must rely on other ways of forming impressions to adequately predict people's behaviours when extraordinary situations arise. Creating rules by externalizing everyday behaviours, such as dressing in appropriate ways or recycling, allows us to do this. Morality is readily inferred from these everyday behaviours, with people reporting moral or immoral events in almost a third of their reports throughout the day (Hofmann et al. 2014). Because these behaviours are common, adherence to the norms governing them may be used as cues for whether people will abide by more serious moral norms. For example, although support for environmental policies is not readily observable, people can easily see whether others make the effort to recycle. Policy support may therefore be inferred from recycling behaviour; failing to abide by this norm then results in moral condemnation. Through this process, even norms that are not inherently moral may become externalized.

The difference between the scope of a norm and its apparent source

doi:10.1017/S0140525X1800002X, e97

Jonathan Birch

Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London WC2A 2AE, United Kingdom.

j.birch2@lse.ac.uk personal.lse.ac.uk/birchj1

Abstract: We should distinguish between the *apparent source* of a norm and the *scope* of the norm's satisfaction conditions. Wide-scope social norms need not be externalised, and externalised social norms need not

be wide in scope. Attending to this distinction leads to a problem for Stanford: The adaptive advantages he attributes to externalised norms are actually advantages of wide-scope norms.

Stanford has identified an important feature of our moral psychology: our tendency to regard moral norms as externally imposed demands rather than as shared subjective commitments. The motivational force of a moral norm appears to come "from outside" – from an external source – and this appearance of externality calls for explanation. He also argues, plausibly, that existing accounts of the evolution of morality do not explain this phenomenon of "moral externalisation." His proposed explanation is that externalisation generates a link between the strength of an agent's own motivation to comply with a norm and the strength of his or her normative expectation that others likewise comply with – a link that helps maintain correlated interaction between prosocial individuals. Here is a problem for this proposal.

Stanford does not distinguish between the *apparent source* of a norm (i.e., do I regard myself as making a subjective commitment to a norm, or do I regard it as something that is externally imposed on me?) and the *scope* of the norm's satisfaction conditions (i.e., do the satisfaction conditions of the norm pertain to my behaviour only, or to the whole community's behaviour?). The apparent source and scope of a norm are separable: Wide-scope social norms need not be externalised, and externalised social norms need not be wide in scope.

To see the difference between the apparent source of a norm and its scope, consider the following two cases. On the one hand, an agent may have a subjective commitment to a norm that pertains to the behaviour of an entire community. Suppose, for example, I have a subjective commitment to the norm that everyone in my community should pick up litter. The norm is wide in scope (it applies to the behaviour of the whole community), yet I do not regard it as externally imposed: I regard it as a subjective commitment, the motivational force of which derives from my personal desire for clean streets. I like it when people adhere to the norm and dislike it when they violate the norm, but I do not regard these norm violations as transgressions of an externally imposed demand. On the other hand, an agent may externalise a norm that pertains solely to his own behaviour. Suppose, for example, the Pope regards certain norms as applying to himself alone *qua* Pope. These norms of Papal conduct have very narrow scope, and yet they may well be externalised: The Pope regards these norms not as subjective commitments, deriving their motivational force from his own personal desires, but as externally imposed Divine commands.

The problem for Stanford's argument is that the adaptive advantages he attributes to externalised social norms are actually advantages of wide-scope social norms. It is wide-scope social norms, not externalised social norms per se, that maintain correlated interaction between cooperators. For example, my subjective commitment to the litter-picking norm will motivate me to pick up litter myself, to monitor my neighbours' litter-picking, to get upset when neighbours fail to pick up litter, to encourage my neighbours to pick up litter, and to prefer interacting with neighbours who pick up litter to neighbours who don't. If others share my subjective commitment, we will profitably cooperate; if they don't, I will shun them. This adaptive package of correlated interaction and profitable cooperation can arise without any externalisation of the norm, provided it is sufficiently wide in scope. Conversely, an externalised norm may fail to yield any significant correlated interaction if it is excessively narrow in scope. Norms that apply to a single individual, such as norms of Papal conduct, are a limiting case in which there is no correlated interaction at all. Once we distinguish between wide-scope norms and externalised norms, allowing the two properties to come apart, we see that it is the former property, not the latter, that leads to correlated interaction.

The ability of wide-scope social norms to maintain cooperation across extended social networks suggests an important role for these norms in human social evolution. One can imagine a gradual expansion of the scope of social norms from the scale of the band to the scale of the wider kin-group, and from the scale of the kin-group to the scale of even larger ethnolinguistic groups. However, there would have been no need for these wide-scope norms to be perceived as externally imposed: shared subjective commitments would have yielded the same adaptive advantages. Externalisation is a separate phenomenon in need of a separate explanation. Although this is not the place to develop such an explanation, it is worth pointing out that Stanford's article, surprisingly, makes no mention of religion. As the example of the Pope suggests, it may be that our tendency to externalise moral norms is a culturally evolved way of thinking entangled with the concept of a Divine enforcer.

The brighter the light, the deeper the shadow: Morality also fuels aggression, conflict, and violence

doi:10.1017/S0140525X18000031, e98

Robert Böhm,[a] Isabel Thielmann,[b] and Benjamin E. Hilbig[b,c]
[a]*School of Business and Economics, RWTH Aachen University, 52062 Aachen, Germany;* [b]*Cognitive Psychology Lab, University of Koblenz-Landau, 76829 Landau, Germany;* [c]*Max Planck Institute for Research on Collective Goods, 53113 Bonn, Germany.*
robert.boehm@rwth-aachen.de **thielmann@uni-landau.de**
hilbig@uni-landau.de **http://www.robertboehm.info**
http://www.cognition.uni-landau.de/people/isabel-thielmann-msc
http://www.cognition.uni-landau.de/hilbig

Abstract: We argue that, in addition to the positive effects and functionality of morality for interactions among in-group members as outlined in the target article, morality may also fuel aggression and conflict in interactions between morality-based out-groups. We summarize empirical evidence showing that negative cognitions, emotions, and behaviors are particularly likely to appear between out-groups with opposing moral convictions.

In the target article, Stanford argues that moral norms serve as fundamental pillars of human interaction, which regulate cooperation, fairness, and altruism within social groups. In turn, severe punishment and social exclusion of in-group wrongdoers might serve the particular purpose to promote stable in-group cooperation. As such, maintaining a shared in-group morality reduces the necessity of constantly tracking the potential risk of exploitation in in-group interactions, thus providing the basis for spontaneous in-group cooperation. Although we concur with the author's reasoning and conclusion in principal, it is unfortunately only a selective excerpt of the full picture. Specifically, the target article overlooks the dramatic negative consequences of (opposing) moral convictions for human interactions. While morality may indeed foster cooperation and harmony within groups, it may also fuel aggression and conflict between groups. We therefore argue that morality represents an important cue through which both in-group cooperation and intergroup conflict are channeled.

Indeed, the presence of a morally opposing out-group is a central criterion of in-group identification (Parker & Janoff-Bulman 2013). A prominent example is the widespread political separation and radicalization based on opposing moral ideologies regarding issues such as refugees, abortion, or socialized healthcare. In such group constellations, moral disagreement becomes the defining aspect of social identity, uniting a joint moral position in opposition to others. Because immorality is associated with harmful acts (Gray et al. 2012), morality-based out-groups are perceived as a severe threat to the in-group. Specifically, morality-based out-groups (i.e., pro-life vs. pro-choice) have been shown to be perceived with more negative emotions and seen as a greater threat to the in-group than non–morality-based

out-groups (i.e., fans of Boston Red Sox vs. New York Yankees; Parker & Janoff-Bulman 2013). Likewise, actions of morality-based out-groups are likely perceived in terms of offensive aggression toward the in-group (Waytz et al. 2014), which may, in turn, spark intentions to defend and protect the in-group, even by means of actively harming the out-group (Böhm et al. 2016).

This perspective is supported by research comparing intergroup conflict between morality-based and non–morality-based opponent groups: In morality-based intergroup conflict, the motivation to absolutely benefit the in-group (i.e., "in-group love") is closely tied to the motivation to aggressively harm or competitively outperform the out-group (i.e., "out-group hate"; Brewer 1999). Interactions between non-morality-based groups have been shown to be primarily guided by in-group love (e.g., Halevy et al. 2008; Thielmann & Böhm 2016), whereas out-group hate increases substantially in interactions between morality-based out-groups (Parker & Janoff-Bulman 2013; Weisel & Böhm 2015). This suggests that the nature of interactions between morality-based out-groups is crucially different from interactions between individuals with opposing moral demands within the same social group. For example, Weisel and Böhm (2015) measured out-group hate as the willingness to actively diminish out-group members' resources at personal cost in an intergroup social dilemma game. Despite the availability of an outside option that had the same benefit for the in-group without necessarily harming the out-group, findings revealed a clear motivation to harm the out-group. Importantly, out-group hate increased substantially only in interaction with members of a morality-based out-group but not in interaction with members of a non–morality-based, yet high-enmity out-group. This shows that the incongruence between groups' moral values may not only hinder intergroup cooperation, but also will even foster harmful intergroup conflict. By implication, introducing morality into intergroup conflict has the destructive potential to override the human aversion of doing harm to others (e.g., Buhl 1999).

Moreover, there are several psychological factors that may further promote and justify aggression toward morality-based out-groups. Most importantly, members of morality-based out-groups are easily dehumanized, denying them the essential rights of humans and therefore providing the psychological basis of harsh and even immoral treatment (Haslam 2006; Struch & Schwartz 1989). Serving the same purpose, the immoral behaviors of the in-group toward morality-based out-groups (e.g., torture) are reframed (i.e., morality shifting; Leidner & Castano 2012), such that the mistreatment of (alleged) moral opponents becomes psychologically more acceptable.

In addition to the rise of out-group hate, moral opposition in intergroup conflict may also hinder peacemaking and reconciliation in ongoing conflicts (Halperin et al. 2011; Waytz et al. 2014). For instance, group-based anger has been shown to increase the support for compromises in intergroup negotiations when there is low hatred between groups (i.e., low levels of out-group hate), whereas it leads to increased aggression when there is high hatred between groups (Halperin et al. 2011).

Overall, the available evidence clearly demonstrates that there are two sides to how morality shapes human interaction. Whereas morality may indeed foster harmony within groups as reasoned by Stanford, it also fuels conflict between groups: As soon as several individuals with shared moral convictions form opposing groups, the moral foundation is also a basis for long-term intergroup conflict and violence, which aims at actively harming out-group members rather than at restoring cooperative social interactions. Understanding the role of morality for human interactions on a more global level therefore also requires awareness of its downside: Whereas it has the potential to unite, it will concurrently divide. As a consequence, the positive effects of morality for the functioning of interactions within groups are intrinsically tied to the negative and destructive effects for interactions between groups. This inherent link might even have co-evolved as suggested by the theory of parochial altruism (e.g., Choi & Bowles 2007). We thus conclude that morality – though it may well drive cooperation and harmony within groups – also fuels aggression and conflict on a larger scale.

Coordination, conflict, and externalization

doi:10.1017/S0140525X18000043, e99

Justin P. Bruner

School of Politics and International Relations, The Australian National University, Acton, Australian Capital Territory 2602, Australia.
Justin.bruner@anu.edu.au
https://sites.google.com/site/justinbrunerphil/home

Abstract: I argue that the set of moralized norms and beliefs is more expansive than Stanford appears to suggest. In particular, I contend that norms governing behavior in conflictual coordination problems are likely to be moralized.

Stanford's proposal is extraordinarily compelling and cleverly weaves together existing empirical and theoretical findings to construct a plausible story regarding the emergence of the distinctive kind of cooperation in which humans routinely engage. Stanford's evolutionary account centers on protecting prosociality and cooperation from exploitation. Stanford's proposed solution – moral externalization – is an attractive option as it both (1) ensures the focal individual is (reasonably) shielded from exploitation and (2) allows similarly minded individuals to reap the benefits of cooperation in novel environments. Both (1) and (2) enhance fitness and, as a result, ensure that those who externalize moral beliefs often outperform their peers.

I focus on just one facet of Stanford's overall picture and attempt to better delineate the kinds of beliefs and norms which would, on Stanford's proposed story, become moralized. I claim that the set of moralized norms and beliefs is more expansive than Stanford appears to suggest and contend that norms governing behavior in conflictual coordination problems – where the threat of exploitation is minimal or non-existent – are likely to be externalized.

To begin, note that although Stanford acknowledges that "nearly any norm or behavior [can] become moralized" (sect. 5, para. 13), it is clear Stanford nonetheless thinks only certain norms are *likely* to be externalized (namely, those norms protecting prosociality from exploitation). Norms of this kind are particularly crucial because they "maintain the most important evolutionary benefits of moralization itself" (sect. 5, para. 13). Moral externalization is adaptive in large part because it involves and sustains norms regarding "harm, fairness, justice, and the rights of victims" (sect. 5, para. 13). In contrast, rules of etiquette or "mere convention" are unlikely to be externalized as they pertain to cases where the benefits of externalization are minimal. In these circumstances, one does not typically stand to gain or lose much from the behavior of others, and the threat of exploitation is non-existent.

Many strategic scenarios do not involve the threat of exploitation but, unlike the case of etiquette, have real fitness consequences for those involved. Take, for instance, conflictual coordination problems. In these strategic scenarios, individuals must coordinate their behavior to achieve some desired end (perhaps they endeavor to jointly manage a home, or divvy up resources, or establish a property convention). A variety of outcomes allow for successful coordination, but tensions arise as there is disagreement as to which coordinative arrangement is most desirable. These strategic scenarios do not correspond to the Prisoner's Dilemma but are instead best conceived of as either a hawk-dove game or the "battle of the sexes." On this representation, it is clear that free-riding and exploitation are no longer of central concern. Instead, what matters is avoiding

miscoordination and ensuring that disagreement does not escalate into violence or jeopardize cooperative endeavors in other domains.

Not surprisingly, norms can easily guide behavior and thus minimize the probability of conflict and miscoordination. John Maynard-Smith, for instance, famously suggested that individuals condition their behavior on some features of the strategic scenario of interest (Maynard Smith 1982). In the case of resource competition, for example, individuals employing the "bourgeois strategy" (i.e., behave aggressively if in possession of the contested resource, acquiesce if not) navigate these scenarios without incident. Individuals could similarly condition their strategic behavior on some features of themselves (such as gender in the case of joint-home management). However the details are spelled out, the results are the same: If all adhere to the same norm, coordination can easily be attained. As a result, individuals have incentive to selectively interact with those adhering to the same norm as themselves.

Furthermore, these norms can guide behaviors in similar, but novel, conflictual coordination problems. Norms of division can be used to allocate meat from the weekly hunt, as well as divide the fruits of novel cooperative endeavors. A gendered division of labor in the home could help individuals assign responsibilities and tasks in other environments. Not only do these norms allow individuals to avoid miscoordination, but they also allow for relatively seamless collaboration in *unfamiliar scenarios*.

Taken together, these norms secure many benefits and perhaps even rival the benefits associated with the cooperative norms considered by Stanford. Exploitation has severe fitness consequences; the miscoordination in conflictual coordination problems has real and significant costs. Miscoordination can result in the breakdown of collaboration, as well as wasteful and sometimes violent conflict. Demanding that others adhere to similar norms ensures cost-free coordination in familiar and new environments alike. Thus, there appears to be significant pressure for these norms to become moralized.

Is there any evidence to support the claim that norms governing behavior in conflictual coordination problems will become moralized? The experimental literature may not currently be able to provide an answer, as most cases considered in the experiments cited by Stanford involve exploitation or harm. Yet, continuing with the examples given above, conventions of property as well as gendered division-of-labor norms are likely to appear to many as more "objective" and universal than table etiquette or subjective preference for vanilla over chocolate ice cream. At the same time, however, it is far from clear that these norms will be taken to be *as* objective and binding as norms that prohibit murder or various other harms and injustices. The norms governing behavior in conflictual coordination problems matter. However, it may be sensible to position such norms somewhere between conventions of etiquette and the fully moralized norms which shield cooperation from exploitation.

Moral externalisation fails to scale

doi:10.1017/S0140525X18000055, e100

Carl Joseph Brusse and Kim Sterelny

School of Philosophy, RSSS, CASS, The Australian National University, Acton, Australian Capital Territory 2601, Australia.

Carl.brusse@anu.edu.au Kim.sterelny@anu.edu.au

https://researchers.anu.edu.au/researchers/sterelny-k

Abstract: We argue that Stanford's picture of the evolution of externalised norms is plausible mostly because of the idealisations implicit in his defence of it. Once we take into account plausible amounts of normative disagreement, plausible amounts of error and misunderstanding, and the knock-on consequences of shunning, it is plausible that Stanford undercounts the costs of externalisation.

Stanford's theory of moral externalisation supposes that agents (1) have *self-interest-independent* motivations to adhere to enculturated norms, and (2) that agents are motivated to shun those who fail to adhere to these norms, and especially if they take advantage of others' conformity to free-ride. These motivations operate on (culturally?) selected normative content, and in the context of coordination and cooperation problems this motivational combination supposedly opens up a new space of flexibility and adaptive plasticity, through selection of normative content that automatically (via the secondary motivation) sequesters the agent from conspecifics who would otherwise exploit them.

We suggest that the plausibility of Stanford's picture depends on its idealisation. Stanford largely makes the case for moral externalisation through an idealised model of dyadic interaction, with individuals assessing one another as social partners according to compatibility with respect to some specific norm. We think that this idealisation does not scale to realistic social ecologies, to realistic levels of normative complexity, and to realistic degrees of error and error-correction. Moreover, we think Stanford understates the adaptive plasticity made possible by more limited forms of normative cognition.

Our first scaling problem is heterogeneity: What of normative variation? Our lineage has been obligately social for much longer than any plausible inception of norm-following (certainly complex norm-following and reputation-tracking, mediated by language). And in a small group environment, moral externalisation is potentially dangerous: Start shunning without pre-existing positive assortment, and you will shun yourself out of your group. Likewise, novel coordination problems invite normative disagreements, as agents differ on the appropriate response. In Stanford's analysis, the stakes of disagreement rise: They are not just failures to negotiate a contract, a lost opportunity after which everyone then returns to the status quo. They threaten fracture. Unless disagreement was rare or transient, the cost of disagreement would select against norm externalisation.

This suggests that the normative content to be externalised must diffuse rapidly and be *synchronised* with high fidelity; conflict costs will be high if a new consensus is established slowly, piecemeal. Perhaps a not-implausibly strong conformity bias might deliver this, though it would have to be fortuitously present. Moreover, any mechanism that generated normative uniformity is in tension with the promised benefits of the model with respect to adaptive plasticity. A moral mutant should be expected to be shunned or out-grouped fairly swiftly, and for a normative innovation to spread within a group (without either fracturing it or being snuffed out), it seems it must do so before anyone really notices; to notice is to object. So, any shunning mechanism efficient enough to weed out free-riders (and mitigate disagreement risk) is also liable to weed out valuable innovations.

The second scaling problem is normative complexity. Externalisation and shunning is a blunt instrument; it motivates agents to want to have *nothing* to do with norm violators, and this imposes significant opportunity costs in any plausible human behavioural ecology. You don't share the kill with me in the way that I think is right, so externalisation of that norm means I am less likely to hunt with you again. But I am also less likely to coordinate with respect to collective defence or co-parenting, or any number of other fitness-critical activities with a high coordination dividend. Shunning has costly knock-on consequences. So, again, externalisation only has minimal opportunity costs if the agent and their conspecifics are already highly normatively synchronised.

Moreover, normative synchronisation limits costs only if the environment is normatively transparent. Any inaccuracy in identifying normative commitments (such as failure to appreciate mitigating reasons for the violation) risks conflict costs. Perhaps gossip can limit these costs, but it would need to be both honest and accurate.

In short, when we focus on early, ancestral, band-level societies prior to vertical segmentation, we think Stanford has undercounted the costs of externalisation. We also think he has overcounted the relative benefits, because he under-rates the utility of correlated subjective preferences and how they can give rise to quasi-contractual norms. In most circumstances, your preference for ice-cream flavours is of no moment to me because eating ice-cream is an individual activity. But subjective preferences matter a lot if and as they are relevant to collective action. For example, before moving into a joint house, it is important to identify the subjective preferences of the inhabitants with respect to whether they wash up after each meal, once a day, or once a week; if meals are joint, and if so, at what time and who cooks. These are all subjective preferences, but because they need to be coordinated in joint activity, they give rise to quasi-contractual norms; norms neither universal nor externalised, but providing a route through which novel cooperation opportunities can be profitably managed. Many stable and important social forms are based on the alignment of subjective preferences: in complex social worlds, most voluntary clubs and associations. Of course, in the house-share case and many other cases involving these quasi-contractual norms, the agents need a meta-norm or meta-policy: What to do if other agents violate the quasi-contractual norms? In many cases, that policy must be signalled. But the point here is that you do not need anything new as the meta-policy. You just need a standing policy of when and how you cut your losses, and on retaliation against those who fail in their quasi-contractual obligations. Frank's (1988) work on commitment emotions suggests both how the policy works and how it is signalled, without, as Stanford himself notes, requiring externalised norms. On balance, it remains unclear why norms need to be externalised to serve as profitable coordination or signalling devices.

These considerations suggest that normative externalisation is a late-breaking cultural innovation whose adaptive social role (if indeed it is adaptive) post-dates the foundational steps in the evolution of distinctively human cooperation. Rather than being tied to the evolution of gossip and indirect reciprocity in the early or mid-Pleistocene, we see it as more likely linked to the later African and out-of-Africa radiations in the late Pleistocene and early Holocene.

Norms, not moral norms: The boundaries of morality do not matter

doi:10.1017/S0140525X18000067, e101

Taylor Davis and Daniel Kelly

Department of Philosophy, Purdue University, West Lafayette, IN 47906-2098.
taylordavis@purdue.edu drkelly@purdue.edu
http://www.taylordavisphilosophy.com
http://web.ics.purdue.edu/~drkelly/

Abstract: We endorse Stanford's project, which calls attention to features of human psychology that exhibit a "puzzling combination of objective and subjective elements," and that are central to cooperation. However, we disagree with his delineation of the explanatory target. What he calls "externalization or objectification" conflates two separate properties, neither of which can serve as the mark of the moral.

Stanford rightly emphasizes a crucial distinction between a category of judgments that merely express subjective preferences and a different category of judgments that express norms or normative evaluations. We agree that the latter category marks the core of the interesting, perhaps uniquely human, phenomenon at issue. However, we think Stanford mischaracterizes this category.

Consider the kinds of epistemic norms that govern inferences: the norm that says for stronger inductive inferences you *should* base your extrapolations on larger samples rather than smaller ones, or the principle of the disjunctive syllogism that tells you if you know that p or q is true, and you know that p is false, then you *should* accept that q is true. These claims do not express subjective desires or preferences about reasoning, but rather norms: requirements or obligations of good reasoning. Phenomenologically, these can be experienced as being externally imposed upon us, perhaps evincing what Wittgenstein (1983 p. 352) famously described as the feeling of the "hardness of the logical must." Yet they are epistemic norms rather than putatively moral ones. The "oughts" apply to inferences, and certainly do not primarily regulate cooperative behavior.

Even in aesthetics, where "beauty is in the eye of the beholder," there is an intuitive difference between subjective preference and normative evaluation. Take the common idea of a "guilty pleasure": A person might grant that some catchy ditty on the radio is really not very *good*, as a work of art, but may nevertheless enjoy listening to it, and even prefer listening to it rather than other types of music (e.g., Coltrane's more abstract explorations) that she herself would rate as *better* by some normative standard external to, or less subjective than, her own personal preferences. Conversely, a person might appreciate the complex artistry of a songwriter who is drawn to minor keys and sad topics, and feel the force of the claim that by some less ego-centric standard she *should* prefer it to fizzy pop music. Yet she might simply not like its morose vibe. Subjective desires and normative evaluations are both logically and psychologically distinct, and unlike the former, the latter have this property we will call, as a nod to Wittgenstein, *hardness*. We hold that this *hardness* is a property of normative judgments in general, epistemic, aesthetic, prosocial, antisocial, and so forth.

A second distinction, importantly different from the first, marks a difference within the general category of normative judgments. This distinction can capture the idea that (of course?) there is another sense of "objective" according to which aesthetic judgments *are* widely thought to be more subjective than moral judgments. As Stanford notes, a number of empirical studies have followed Goodwin and Darley (2008) in measuring participants' judgments about different cases in which people disagree about a normative issue. In this approach, responses are deemed "objectivist" when participants judge that, in cases of disagreement, at least one party to the disagreement must be wrong; conflicting claims cannot both be correct. These studies consistently show that very few people give objectivist responses on aesthetic matters, and people are much more likely to be objectivist about putatively moral issues (Beebe & Sackris 2016; Goodwin & Darley 2008; 2012; Wright et al. 2013). (Unfortunately, none of these studies has examined objectivism for epistemic matters.) Call normative judgments that have this property *objectivist*.

We take a different lesson from these studies than Stanford. We see them as showing that some putatively moral judgments are objectivist, but *far* from all of them are. Stanford claims that "the statement that it is wrong to rob a bank, for example, was reliably judged far more objective than the statement that anonymous giving is good. Such variation has been particularly emphasized by Wright et al. (2013; 2014) in work that confirms Goodwin and Darley's central findings" (target article, sect. 2, para. 4). First, we are puzzled by the claim of "far more objective," because the methodology does not allow respondents to judge *degrees* of objectivity. Rather, each claim is judged categorically, as either objective or not, in two different ways: In addition to the disagreement question just described, participants were asked whether claims were "true," "false," or "just an opinion or attitude." Second, and more importantly, are the patterns in the data. Wright et al. (2013) report that on the putatively moral issue of anonymous giving to charity, only 11% of participants gave the objectivist response on both measures, with only 35%

giving the objectivist response on even a single measure. In fact, out of seven claims deemed moral by at least half of the participants, only four were judged objective by more than half on both measures. We see the trend continued in Beebe and Sackris (2016), who measured objectivism only using the disagreement measure, and found that, out of 10 putatively moral issues, only 4 were judged objective by more than half of participants. Three of these 10 issues were judged objective by only 25% of participants or less. What we take the data to show, then, is that a substantial proportion of judgments about putatively moral matters are *not* judged objective by *most* people.

What does all this have to do with morality and distinctively moral norms? We agree that understanding human normative cognition will be crucial to understanding human cooperation (cf. Boyd & Richerson 2005; Chudek & Henrich 2011; Sripada & Stich 2007). However, we hold that Stanford conflates two different ways normative judgments might be "objectified or externalized," which we have separated out as the properties of hardness and objectivism. Each is interesting and important on its own terms, but we hold that they do not run in tandem. We also hold that these properties, taken either together or independently, are not unique to what is often pre-theoretically considered *morality*. So, we continue to be skeptical of attempts like this (cf. Kelly & Stich 2007; Nado et al. 2009) to show there is any empirically important – let alone well-delineated – phenomenon deserving of being partitioned off as morality. No subcategory of norms makes up a psychologically distinctive or cooperatively indispensable set of *moral* ones.

How does moral objectification lead to correlated interactions?

doi:10.1017/S0140525X18000079, e102

Geoffrey P. Goodwin

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104.

ggoodwin@psych.upenn.edu

https://psychology.sas.upenn.edu/people/geoffrey-goodwin

Abstract: The objectification of moral norms is purported to occur because it enables correlated interactions between individuals who share the same cooperative norms. But how does this process take place? I suggest two mechanisms beyond those Stanford identifies. I also ask whether there is predictable variation in which moral norms engender the strongest coupling between objectification and discomfort with disagreement.

Stanford astutely identifies a problem that has not been adequately addressed: Why is it that people objectify (or "externalize") moral demands? Although ample evidence exists documenting this tendency, a satisfying explanation for it has not yet been produced. Indeed, the question itself has not been seriously considered by experimentalists working on metaethical belief. Stanford's theory is that this tendency emerged because it facilitates human social cooperation. Specifically, it promotes "correlated interaction" between social partners who share the same moral norms. In this commentary, I raise two questions regarding this original and thought-provoking thesis – one regarding the precise mechanics of how moral objectification gives rise to correlated interactions, and the other regarding the link between objectification and social interaction.

How is it that the objectification of moral norms gives rise to correlated interactions between individuals who share them? Stanford emphasizes one mechanism primarily – namely, that the objectification of moral norms automates the demand for others to conform to such norms, thereby protecting potential cooperators from exploitation by individuals who do not share these norms. A mere subjective preference, he argues, could underpin a similar desire for others' conformity, but it is unlikely to do so *automatically*. In essence, the objectification of moral norms produces a reliable, internal mechanism by which cooperative individuals can select among potential interaction partners. Stanford also intimates at the possibility of a more social psychological mechanism – namely, that by objectifying certain moral norms, individuals *advertise* themselves as desirable social partners (i.e., their objectification indicates their commitment to moral norms). However, it is not clear that objectification is critical here, for the same reason that Stanford regards it as unnecessary for producing individual commitment to moral norms – communicating one's own commitment to a moral norm could equally well be achieved by indicating a strong subjective preference to abide by it. This is perhaps why Stanford does not stress the importance of this advertising mechanism.

There are, however, two further social psychological mechanisms stemming from moral objectification that might make correlated interactions more likely. Both extend a point Stanford emphasizes in describing his primary mechanism. The objectification of a moral norm creates a strong expectation that others abide by this norm, but it also *conveys this social expectation*. Hence, apart from the effect that objectification has on the holder of such an expectation, it should also have a powerful effect on its targets, by exerting social pressure on them to abide by the norm. This may, in turn, shape and constrain the norms endorsed and followed by others in society ("lifting all boats"). Here, objectification does seem critical. Broadcasting a mere subjective preference would not imply the same strength of expectation that others should conform.

A second, related mechanism, is that accordant targets of such expectations should subsequently direct more social attention to the person doing the objectification, in the hopes of cooperating with them. The proclamation of an objectified moral norm thereby enables *assortative matching* among like-minded social partners. However, unlike Stanford's primary assortative mechanism, which relies on processes internal to the person doing the objectification, this effect is mediated by a change in the minds of those who observe the objectification of a moral norm. Once again, though, objectification itself seems critical. In contrast to a mere preference, objectification conveys a particularly strong expectation that others abide by the norm. It therefore informs others who share this norm that they themselves are especially socially desirable in the eyes of the original norm holder. Therefore, I suggest that these two social psychological mechanisms complement Stanford's theory, in addition to the primary mechanism he identifies.

A second question concerns the documented tendency for moral objectification to predict relevant social attitudes. In existing analyses *across* moral norms, the more strongly a norm is objectified, the more people feel uncomfortable with another person who disagrees with it (Goodwin & Darley 2012). This result is highly pertinent to Stanford's theory, as he notes, because it demonstrates a clear link between objectification and social exclusion. But does (and should) this link vary as a function of the content of particular norms? Are there some moral norms for which there is an especially strong coupling between objectification and discomfort with disagreement?

To provide some preliminary data, I re-analyzed these correlations at the level of each individual norm (rather than across norms) in three relevant experiments reported in Goodwin and Darley (2012; Study 1 and its two follow-ups). The correlations between objectification and discomfort with a disagreeing party were significant for each norm, but there was also considerable heterogeneity in their size – they ranged from 0.12 and 0.20 in the case of norms pertaining to the wrongness of robbing a convenience store, and the wrongness of flag-burning; and up to 0.70 and 0.72, in the case of norms pertaining to the goodness of donating significant income to charity, and the wrongness of punching someone in the face at a bar. What explains this variation? One

possibility is valence – given that, overall, negative norms are typically objectified to a greater degree than positive norms, even when matched on how strongly people agree with them (Goodwin & Darley 2012). However, there was no overall difference in the average strength of these correlations for positive ($r = 0.45$) and negative ($r = 0.43$) norms, $t(30) = 0.26$, $p = 0.80$. This null result in fact seems consistent with Stanford's theory. After all, when gauging another person's cooperative potential, it is instructive to know not only how strongly they objectify norms pertaining to the wrongness of harm and exploitation, but also how strongly they objectify norms that encourage kindness and generosity.

But it leaves a lingering question: Is it possible to predict or explain the heterogeneity in these observed correlations? Why should objectification be tied closely to social exclusion for some moral norms and not for others? We do not yet have a full characterization of this variation, but when we do, it will be instructive to learn whether Stanford's theory might elucidate it.

Green beards and signaling: Why morality is not indispensable

doi:10.1017/S0140525X18000080, e103

Toby Handfield,[a] John Thrasher,[a] and Julian García[b]

[a]*School of Philosophical, Historical and International Studies (SoPHIS), Monash University, Melbourne, Victoria 3800, Australia;* [b]*Faculty of Information Technology, Monash University, Melbourne, Victoria 3800, Australia.*

toby.handfield@monash.edu Johnthrasher23@gmail.com
julian.garcia@monash.edu www.tobyhandfield.com
www.JohnJThrasher.com http://garciajulian.com

Abstract: We argue that although objectivist moral attitudes may facilitate cooperation, they are not necessary for the high levels of cooperation in humans. This is implied by evolutionary models that articulate a mechanism underlying Stanford's account, and is also suggested by the ability of merely conventional social norms to explain extreme human behaviors.

The target article argues that the distinctive psychology that regards moral properties as objective, external features of the world is adaptive because it allows us to engage in beneficial cooperation through correlated interactions. Implicit in the account are two theses: (1) the *correlation thesis* that moral commitment serves as a correlating device, allowing fellow norm-followers to associate with each other and to ostracize defectors; and (2) the *indispensability thesis* that externalized moral norms of this sort are necessary to achieve pro-social cooperation, at least at the high rate seen in humans.

We argue that the first thesis is plausible, but that it undermines the second thesis. If the correlation thesis is true, there is good reason to think that externalized moral attitudes are not indispensable for achieving cooperation, but are merely one possible solution among many.

There is already a well-understood set of mechanisms by which agents may ensure correlated interaction with fellow biological altruists. It is not clear how Stanford's account is supposed to relate to this menu of options. Is it a type of reciprocity, costly signaling, group selection, or something else? We suggest there are three possibilities (not mutually exclusive) that are especially promising.

Costly signaling models require a diversity of types in the population (e.g., cooperators and defectors) who – at least in simple cases – face differential signaling costs, or who stand to make differential gains from being believed. Some types can afford to send signals that are uneconomic for other types, and, hence, any sufficiently costly signal is credible. Costly signaling mechanisms have been proposed to explain phenomena like unconditional sharing, costly punishment, apology, and guilt (Gintis et al. 2001; Jordan et al. 2016; Ohtsubo & Watanabe 2009). Could externalizing moral psychology be a form of costly signal also?

It is conceivable. By rigid commitment to moral attitudes, an agent might incur a cost that a defector would not be willing to pay. But for this hypothesis to be plausible, we need an explanation for why externalized moral attitudes are especially burdensome or costly. Prima facie, it is hard to see that they are. Compared to better established examples of costly signals, including rituals such as fasting, bodily mutilation, and animal sacrifice, it does not seem as if having the belief that moral requirements are objective features of the world is especially costly at all.

The *green beard hypothesis* is that altruistic traits are genetically linked to a distinctive phenotypic trait. If only altruists can develop green beards, then a strategy of cooperating with fellow green beards is a plausible evolutionary outcome. A genetic barrier to green-bearded defectors keeps the world safe for cooperation. Recent work (cf. Gardner & West 2010) has shown that green beard mechanisms can sustain cooperation even if the link between beard and altruism is not strict but permits some plasticity. The result is an unstable dynamic, in which the population cycles through different beard colors, with bursts of cooperation in the beginning of each beard cycle, followed by invasion by defectors, followed by development of a new beard color (Jansen & van Baalen 2006; Traulsen & Nowak 2007). Perhaps the best elaborated account of how these sorts of dynamics might explain actual human cooperation is the case of accents, which are certainly hard to fake and are also not tightly linked to any particular genes (Cohen 2012).

Both costly signaling and green beard accounts undermine the indispensability thesis, however, because both imply that the cooperative correlating mechanism is arbitrary. Stanford emphasizes that the moral psychology, which he seeks to explain, appears to be cross-culturally robust. This makes it quite unlike rituals or accents, which vary dramatically across time and space.

Finally, the most promising mechanism to underlie Stanford's account is *social selection*. Suppose that a competitive mating environment exists in which fitness is enhanced by finding reliable long-term cooperative partners. In such a market, it is adaptive to have a reputation for being reliable. So, we predict adaptations that make one sensitive to reputation (cf. Haley & Fessler 2005). Cooperative, altruistic behaviors may then be adaptive because they enhance one's reputation – even if those behaviors are done cynically, for reputation-enhancing reasons. The competitiveness of the mating market may then drive this process so that ever better demonstrations of reliability are required in order to obtain a mate. In this setting, it may be more cost-effective to *be reliable* than to merely *appear reliable* (Sperber & Baumard 2012). One particular way to be reliable is to take moral facts as external, objective demands. Notably, this process explains the development of "high-quality" types as emerging from a competition among lower-quality types who are more self-serving in their pursuit of reputation.

This better explains the adaptive function of externalizing psychology in particular, but without more explicit modeling, it remains open that a population in equilibrium may have only a small minority who display this phenotype. The main mechanism – reputation-sensitive coordination with the prevailing norms – may explain most observed cooperation. Indeed, this is plausible. Consider the evidence of social norms prevailing over moral commitments, revealed both in history and in experimental settings (e.g., Haney et al. 1972). Externalizing psychology is apt to fascinate philosophers, who are in the grip of related metaethical puzzles dating back to Plato's *Euthyphro*, but this is not yet evidence that it plays a significant role in achieving actual cooperation.

Debates about the genesis of human cooperation are unlikely to make significant advances without comparison of models "in the field" – using disciplines such as archaeology, ethnography,

genetics, and experimental economics. This work, however, requires models that can deliver testable predictions. We wait with interest to see whether Stanford's proposal constitutes a novel mechanism, with novel predictions, or if it can be assimilated to existing mechanisms. We have suggested here that if it is assimilated to existing mechanisms, there is little hope for the indispensability thesis. The "categorical imperative" nature of moral commitments may be a contingent artefact of our idiosyncratic evolutionary history.

Externalization is common to all value judgments, and norms are motivating because of their intersubjective grounding

doi:10.1017/S0140525X18000092, e104

Carme Isern-Mas and Antoni Gomila

Human Evolution and Cognition Group (EvoCog), University of the Balearic Islands (UIB), Institute for Cross-Disciplinary Physics and Complex Systems (IFISC), Associated Unit to Consejo Superior de Investigaciones Científicas (CSIC), Campus Carretera Valldemossa, 07122 Palma de Mallorca, Spain.
lsernmas.carme@gmail.com toni.gomila@uib.cat
http://evocog.org/carme-isern/ https://antonigomila.wordpress.com/

Abstract: We show that externalization is a feature not only of moral judgment, but also of value judgment in general. It follows that the evolution of externalization was not specific to moral judgment. Second, we argue that value judgments cannot be decoupled from the level of motivations and preferences, which, in the moral case, rely on intersubjective bonds and claims.

We have two qualms with Stanford's target article: one regarding the way he characterizes the question he raises – the objectification of moral norms – and one regarding the way he answers it – that the objectification of moral norms evolved as a strategy to promote hyper-cooperation and to avoid exploitation. On the first count, we show that objectification is a feature of value judgment in general, whatever the domain. On the second count, we argue that this level of moral norms and judgment cannot be decoupled from the level of motivation and preferences, without which it would lack its motivating strength. We conclude by pointing out the implications of these two points for an evolutionary approach to morality.

First, Stanford seems to take for granted his starting point, that externalization is distinctive of moral judgment. In fact, he does not even consider the possibility that other kinds of value judgments also exhibit this feature. But it is pretty clear that they do as well. Take, for instance, aesthetic judgment: saying that something is beautiful entails that everybody should find it so. Since Kant's third *Kritique* (i.e., *Critique of Judgment*, Kant 1790/1987; see also Ginsborg 2014; Zangwill 2014), it is undisputed that value judgments aim at universal validity and cannot be reduced to subjective preferences. Other kinds of judgment are not so objectified. Neither the judgment of the agreeable, which simply claims that one likes something but not that everyone else ought to like it, nor cognitive judgments, by which we ascribe a property to an object, exhibit this objectivity.

It follows from this that the explanandum of the target article – the externalization of moral judgment – is too narrowly set. An evolutionary account of externalization has to deal with the externalization of value judgments in general, not just of moral judgments. Therefore, the explanans proposed misses the real nature of the phenomenon in question, the fact that human values, not just moral ones, do not reduce to subjective preferences. An evolutionary account of externalization in terms of the benefits of hyper-cooperation misses the point of externalization as such. The fact that externalization contributes to morality does not imply that it was actually selected for this effect (Gould & Lewontin 1979). Sophisticated language, theory of mind, and counterfactual reasoning also contribute indirectly to cooperation in the way that Stanford claims for externalization (transmission of information about others' moral commitments), but they are not considered features of morality. Why should externalization be different?

Second, if for the sake of the argument we accept that Stanford's proposal – that moral norms and values externalization promote group conformity – can be extended to all sorts of value judgments, the question still remains as to how it is that norm objectification manages to do so. In other words, how is it that people's behavior is sensitive to such normative judgments? It is at this point that a link between the level of preferences and the level of norms is unavoidable. While it is clear that moral judgments do not reduce to prosocial preferences – a point already made by Darwin in *The Descent of Man* – it is also important to realize that norms and values are not decoupled from motivations either. Again, we miss an explicit recognition of this fact in the target article, although it is implicitly assumed near the end, when Stanford notices that moral commitments are intrinsically motivating (and carry a demand for intersubjective agreement). Without this link, recognition of a universal, objective duty might be behaviorally inert – and fear of group exclusion is not the psychological way through which we experience the call of duty. In other words, the real evolutionary quiz is not the externalization of moral norms, but the fact that we feel the pull of the norm.

From this point of view, Stanford's evolutionary scenario is unsatisfactory, because it is concerned just with the level of norms and values, as detached from the level of motivations. It is as if humans come to formulate judgments, to recognize them as moral in character, and automatically become prone to conform to them and to demand from others a similar conformity. What is missing is an answer as to why we feel obliged to comply, why we feel intrinsically motivated to behave according to the judgment, and why we expect others to feel the same.

A promising answer to this question, we submit, can be found in Darwall's second-person view of morality (Darwall 2006). In outline, it would go like this: The objective character of moral norms is geared to subjective motivations because such norms are grounded in the patterns of claims and demands that emerge in intersubjective interaction and that a community comes to sanction. From this point of view, the process that explains the link between norms and preferences would be as follows: (a) I interact with another agent with coordination and reciprocity, and out of evolved prosocial preferences; (b) We explicitly or implicitly make demands on each other; (c) We hold each other accountable in case of failure to comply without excuse; (d) We come up with expectations and norms about how others will or ought to behave; (e) We feel motivated to conform to those norms because we know that we can be held accountable by others; (f) The group comes to sanction those norms and expects everybody to conform; (g) We end up experiencing these norms as moral (i.e., externalized) and at the same time are motivated by them.

Taken together, both points – that externalization is common to all value judgments, and that norms are motivating because of their intersubjective grounding – suggest that an evolutionary account should deal, on the one hand, with norm and value externalization as the real cognitive novelty in the human lineage – a novelty that appears to be general rather than domain-specific – and on the other hand, with the social dynamics through which intersubjective claims and demands come to be externally sanctioned, and psychologically internalized. From this point of view, externalization is not a selected feature to ensure conformity to moral norms because of their efficiency in promoting cooperation; rather, externalization is an outcome of the process through which prosocial preferences become normative because of the demands of mutual accountability that mediate the interactions in a species like ours.

From objectivized morality to objective morality

doi:10.1017/S0140525X18000109, e105

Joseph Jebari and Bryce Huebner

Department of Philosophy, Georgetown University, Washington, D.C. 20057.
jdj48@georgetown.edu **Bryce.Huebner@georgetown.edu**
http://brycehuebner.weebly.com

Abstract: Stanford holds that the externalization and objectivization of moral judgments are what sustain human cooperative lifeways. We reply that the central function of human moral psychology is to track and respond to the structural features of our social environment, and we argue that moral obligations are grounded in the relationship between individual agents and the stability of their social groups.

Humans are more helpful, generous, and informative than any other ape (Tomasello 2009). We share information with people who are neither kith nor kin, and we cooperate with people we have never met and will never meet again. Stanford argues that human ultrasociality is possible because we treat moral obligations as part of an *externally imposed* moral order, which applies equally to all; because the experience of moral motivation feels objective, it automatically generates the demand that others be similarly motivated, and in populations of like-minded individuals, this yields correlated interactions that are less likely to be exploited (target article, sect. 5, para. 7). We are skeptical of Stanford's psychologically oriented approach to human morality. People do take others to be governed by the same moral norms as they are, and to be motivated to do so by something more than desires, preferences, or conventional assumptions (target article, sect 2). But if morality is understood as a kind of complex cooperative disposition, objectification becomes unnecessary, as correlated patterns of interactions that require conformity with open-ended and cooperative norms can be sustained by a sophisticated form of homophily (Ohtsuki et al. 2006; Rand et al. 2011). Therefore, we maintain that a plausible understanding of the evolution of objective morality must look beyond human psychology, to the objective features of the world that govern cooperative human ways of life (henceforth "lifeways").

Stanford appeals to an objectification mechanism to explain why we disaffiliate from those who act differently. But this downplays the role of affiliative tendencies in producing automatically coupled values and preferences. We adjust our normative expectations to track prevalent forms of social behavior; we treat conformity as rewarding, and deviation from social norms as aversive (Bicchieri 2016; Klucharev et al. 2009; 2011; Milgram & Sabini 1978; Montague 2006); and within a shared normative framework, feelings of social fluency will promote affiliation among those with similar normative views (Reber & Norenzayan, in press). If such forces are sufficient to drive complex, open-ended, and cooperative forms of behavior – as we believe they are – then objectification will be unnecessary to explain why people who engage in contra-normative behavior are judged to be "less attractive potential partners in social interaction" (target article, sect. 5, para. 7).

On their own, such facts will not explain why we treat moral *obligations* as objective. But as Stanford rightly notes, humans are obligate cooperators, who must rely upon one another to survive (target article, sect 5). Although forms of social tolerance and differential affiliation are present across the ape clade, these tendencies are greatly enhanced in humans, and this opens up lifeways that place social learning and cooperation at center stage (Hare 2017; Henrich 2016). Like other domesticates, we have become more docile and less reactive (Wilkins et al. 2014); development is also slowed, allowing more time for juveniles to engage in social learning and to learn how to trust others (Hrdy 2009). Just as importantly, we have lost the adaptations that would allow us to survive and flourish on our own: We lack sharp teeth for defense and hunting (Henrich 2016), we cannot extract nutrients from uncooked food (Isler & Van Schaick 2012; Wrangham 2009), and without cultural scaffolding, we rapidly lose access to tools and technologies that allow us to overcome these limitations (Henrich 2004; 2016). In acknowledging the critical changes that have emerged over the course of human evolution, it becomes clearer that our reasons for treating moral obligations as external has little to do with feelings of objectivity. Cooperation, coordination, and trust are objective features of our social lifeway. If successful attempts to contravene the cooperative structure of society were widespread, they would trigger population collapse. No less dramatically, defectors would be unable to survive without networks of cultural scaffolding. Consequently, the felt objectivity of the norms governing correlated interactions is not just a fact about motivation; obligations emerge as a consequence of our relationship to the social order, and our moral motivations are determined by our (often tacit) recognition of this relationship.

Stanford's focus on psychological factors obscures the significance of one of the most distinctive features of human evolution: the emergence of complex and adaptive social networks, which are structured by rich patterns of social interaction (Apicella et al. 2012; Fehl et al. 2011; Hill et al. 2011). We do not just rely on culture-sustaining social processes to survive and flourish, we also constitute these processes. For culture-sustaining social processes to persist, people must be motivated to preserve the network structures they constitute. This requires a robust sensitivity to the structural constraints that govern the organization and dynamics of our social lifeway, as significant deviations from these constraints will prevent these networks from playing their constitutive role in sustaining culturally acquired knowledge (Barkoczi & Galesic 2016; Derex & Boyd 2016; Muthukrishna et al. 2014; Reia et al. 2017; Sterelny 2012; cf. Hooker 2013). This yields a motivation to preserve the non-optional features of the social order we depend upon and constitute. But this is not just a fact about human motivation, human preferences, or some other feature of human psychology. The structural and dynamic constraints that characterize culturally robust social phenomena are determined by the emergent patterns of compatibility and incompatibility that obtain between various social practices and routines.

We contend that our moral psychology is grounded in our ability to track and respond to the structural features of our social environment. Likewise, our moral obligations are grounded in the relationship between individuals and the stability of their social groups – a relationship that is largely independent of individual attitudes (Jebari, in preparation). Of course, we can converge on locally stable, but otherwise optional norms, and where this occurs, such norms will feel objective. Here too, the experience of objectivity and the motivation to preserve such norms will be grounded in the relationship between individuals and the adaptive networks they constitute. Because moving away from stability is dangerous, different groups may converge on different norms, and more generally, this will make threats to local stability feel dangerous. Consequently, moral progress will typically require finding points of higher-order stability. But that is another story for another day.

Moral externalization is an implausible mechanism for cooperation, let alone "hypercooperation"

doi:10.1017/S0140525X18000110, e106

Tim Johnson

Center for Governance and Public Policy Research & Atkinson Graduate School of Management, Willamette University, Salem, OR 97301.
tjohnson@willamette.edu **www.tim-j.com**

Abstract: To facilitate cooperation, moral externalization requires truthful and meticulous information about others' moral commitments (Stanford target article, sect. 6). By definition, this information does not exist in the low-information environments where humans display their "hypercooperativeness." Furthermore, collecting that information – if possible – entails costs that other mechanisms for correlated interaction avoid. Hence, moral externalization is an unlikely mechanism for cooperation, let alone "hypercooperation."

I don't feel like I did something spectacular; I just saw someone who needed help. I did what I felt was right.

— Wesley Autrey (the "Subway Hero"), quoted in Buckley (2007, p.A1)

When a stranger collapsed onto the subway tracks at 137th Street and Broadway in New York City's Upper Manhattan, fellow subway-rider Wesley Autrey did not have time to assess the stranger's moral commitments, nor could he access the information needed to do so (see Buckley 2007). Instead, as the No. 1 train barreled toward the stranger, Autrey evaluated the scene and – if we believe the sentiments quoted in the epigram – acted solely on his *own* beliefs. He sprung from the platform and risked his life to save the stranger. This action, albeit exceptional in its potential costs and realized benefits, is but a single example of the anonymous altruism (Engel 2011) and rapid-fire cooperation (Rand et al. 2014) that has distinguished humans as "hypercooperators." Also, it exemplifies a form of prosocial behavior that cannot be explained by moral externalization.

As Stanford explains in the target article, the effect of moral externalization on prosociality "is strictly limited by the accuracy and detail of the information agents are able to acquire about one another's distinctively moral commitments" (sect. 6, para. 1). According to this line of reasoning, Wesley Autrey will act altruistically only after gathering the information needed to compare his moral commitments with those of the guy lying on the tracks. By definition, this accurate and detailed information does not exist in the relatively anonymous environments in which humans display their unique form of prosociality (see, e.g., target article, sect. 4, paras. 1–3, for discussion of such environments). Accordingly, moral externalization seems to be an unlikely mechanism for the evolution of hypercooperation.

Furthermore, these information demands impose costs that other mechanisms for cooperation avoid, thus making moral externalization an unlikely mechanism for prosociality even in more pedestrian social encounters. To see this point, note that Stanford argues that moral externalization is a mechanism for correlated interaction (a.k.a. positive assortment). Correlated interaction ensures that the beneficiaries of prosocial behavior are those who share the same heritable attribute driving that behavior (e.g., genotype or strategy), not merely those with the same inclination to act prosocially in a given encounter (see, e.g., Hamilton 1964a; 1964b; Van Cleve & Akçay 2014). After all, if only the inclination to cooperate is used to discriminate among social partners, then the population could drift to an easy-to-exploit cooperative strategy that opens the door to exploitation by defectors (see, e.g., Johnson & Smirnov 2013). Hence, for moral externalization to promote correlated interaction, agents must rely not only on truthful and meticulous information about the moral commitments that make others likely to cooperate in an encounter (target article, sect. 6, para. 1), but also on information about whether others' moral commitments make them likely to cooperate *because* they share the same heritable attribute triggering cooperation.

Moral externalization thus requires extensive information to facilitate positive assortment. Other mechanisms need less information to do so. For instance, conditioning cooperation on material parity generates positive assortment via a comparison of two pieces of information: the value of one's wealth holdings and the approximate value of a social partner's holdings (Johnson & Smirnov 2012). This information is likely to be available without search, due to the benefits of signaling it, and it is likely to be accurate, due to the costs of generating the signal (see, e.g., Nelissen & Meijers 2011). On the contrary, cooperation via moral externalization requires the comparison of multi-dimensional information (do all of the attributes of your moral code match mine?), and that information can be readily faked, as Stanford's discussion about detecting moral commitments implies (sect. 6, para. 1). As a result, one must wonder: Why would natural selection favor a mechanism for prosociality that relies on intensive information search and complex cognition, when low-information and cognitively simple mechanisms for correlated interaction exist? I do not think one can answer this question in a way favorable to Stanford's thesis.

However, with these criticisms issued, Stanford deserves credit for identifying a mechanism that could enable prosociality in certain environments, though – per the discussion above – it likely would have had to evolve for another purpose. That is, were it a "spandrel" resulting from cognitive mechanisms that proved to be adaptive for other reasons, then moral externalization could be co-opted to facilitate prosocial behavior in information-rich environments. Arguing that it evolved *because* it produces correlated interaction that facilitates humans' unique form of prosociality seems less plausible. The environments in which humans display their distinctive form of prosociality lack the information that moral externalization requires, and, in information-dense environments, less costly mechanisms can facilitate the correlated interaction needed to support cooperation.

Moral externalization may precede, not follow, subjective preferences

doi:10.1017/S0140525X18000122, e107

Artem Kaznatcheev[a,b] and Thomas R. Shultz[c,d]

[a]*Department of Computer Science, University of Oxford, Oxford, OX1 3QD, United Kingdom;* [b]*Department of Translational Hematology & Oncology Research, Cleveland Clinic, Cleveland, OH 44195;* [c]*Department of Psychology, McGill University, Montreal, Québec, H3A 1G1, Canada;* [d]*School of Computer Science, McGill University, Montreal, Québec, H3A 0E9, Canada.*

kaznatcheev.artem@gmail.com thomas.shultz@mcgill.ca

https://egtheory.wordpress.com/ www.tomshultz.net

Abstract: We offer four counterarguments against Stanford's dismissal of moral externalization as an ancestral condition, based on requirements for ancestral states, mismatch between theoretical and empirical games, passively correlated interactions, and social interfaces that prevent agents' knowing game payoffs. The fact that children's externalized phenomenology precedes their discovery of subjectivized phenomenology also suggests that externalized phenomenology is an ancestral condition.

As the naturalization of capitalism places selfishness as the default, and cooperation as the anomaly to be explained, so Stanford places subjectivized phenomenology as the default and seeks to explain what evolutionary pressures lead to moral judgments feeling external and objective. We agree with Stanford that the difference between externalized and subjectivized phenomenology of motivations needs an explanation and that current evolutionary theories have not resolved this question. However, we disagree with him on etiology – that is, which side of this distinction is to be taken as default and which emerges later and needs further explanation. We side with Joyce's alternate proposal (mentioned by Stanford in sect. 3, para. 8; see sect. 4.4 of Joyce 2006) that externalization is the norm. Thus, subjectivization (e.g., our feelings towards ice cream) is the exceptional ability in need of an adaptationist account. For us, the general question becomes: How did humans evolve a phenomenology of the subjective world as separate from the default phenomenology of an external world?

In the second half of section 3, Stanford dismisses the idea that an externalized phenomenology "might well be an ancestral condition from which no shift to an externalized or objectified

moral psychology was ever required" (sect. 3, para. 8) because he believes that cooperation requires correlated interactions and correlated interactions require active maintenance. We think that his dismissal is unjustified on at least four points.

First, if we are looking for ancestral states from which our current moral psychology developed, then, even if moral capacities are required for cooperation, that doesn't necessarily mean that cooperation is required for their ancestral states. For example, we don't appeal to considerations of cooperation if we want to understand why the phenomenology of colors is external.

Second, although correlated interactions are needed for the maintenance (or emergence) of cooperation in games like Prisoner's Dilemma and Stag Hunt, those are not the only games relevant to our distant ancestors. In games like Harmony, for example, it is rational to cooperate, but active mechanisms like ethnocentrism can actually undermine this cooperation and replace it with defection (Kaznatcheev 2010). Note that the example of Harmony and ethnocentrism also casts doubt on Stanford's claim that active mechanisms and morality primarily promote hypercooperation. This makes empirically determining the games played by our distant ancestors an important unresolved question. Many researchers mistakenly conclude that, because games like Prisoner's Dilemma and Stag Hunt are the most studied games (due to their undoubted theoretical interest), this implies that these are the games that were important in early human evolution. There is no solid justification to believe this, as it is a case of the (interestingness) tail wagging the (pervasiveness) dog. Instead, game types should be diagnosed from more direct evidence of a game's ability to explain important phenomena, regardless of whether the game seems theoretically interesting.

Third, for evolutionary games, we believe that activity is required to break correlation rather than to create or maintain it. The classic example would be spatial structure, which is known to facilitate cooperation. Although uncorrelated inviscid models are easier to analyze (and were thus analyzed first), they are not more natural because humans (like all other organisms) are embedded in space. To undo this inherent correlation, we need the quintessential active process of locomotion for decorrelation. The other big example requiring active mechanisms to de-correlate would be the genetic proximity in inclusive fitness, and in mammals such as ourselves, parental care.

Finally, in his argument for active mechanisms for building correlations, Stanford speaks of payoffs of evolutionary games as if they are known by the agents making decisions to cooperate or not. But there is no reason to believe that a person's *perception* of evolutionary payoffs should align with those payoffs' actual *effect on individual fitness*. In particular, Kaznatcheev et al. (2014) show examples in which individuals act rationally on their perceptions of payoffs, but due to evolution those perceptions track inclusive fitness instead of the game's effect on individual fitness. In such cases, agents would not know that defection increases their individual payoff. These useful delusions create a social interface, making it easier to cooperate. And just like the individual interfaces (Hoffman 2009) (e.g., color making it easier to choose ripe fruit) which they extend, such social interfaces can feel external and objective.

Having made a negative case against Stanford's dismissal of Joyce, we also propose a positive argument for externalized phenomenology being the default. Just as biologists turn to embryology for inspiration on the etiology of evolution, we turn to developmental psychology for an initial sketch of the order of evolution of morality. Given the limits on paleontological data for morality, we believe that this is the best available option. And current understanding of child development largely supports objectivity as the default. Emotional arousal and emotional contagion are observed even in newborns (Dondi et al. 1999), but toddlers don't develop a sense of self and the ability to differentiate between their own and others' internal states until around 2 years of age (Decety 2010; Roth-Hanania et al. 2011). Children acquire basic theory of mind capacities for inferring subjective states like desires, intentions, and beliefs between 3 and 5 years (Wellman et al. 2001). This suggests that an understanding of experience as subjective both in oneself and others develops from an objectivized phenomenological precursor.

This pattern also occurs with more explicit tests of moral objectivism. Preschool children exhibit higher levels of moral objectivism than 9-year-olds and adults (Schmidt et al. 2017). It has also been observed that the subjectivization of experience occurs at different times for different classes of experience, with 4- to 6-year-olds treating properties like *fun* and *icky* as more response-dependent than moral properties like *good* and *bad* (Nichols & Folds-Bennett 2003). Overall, this suggests that it is reasonable to view externalized phenomenology as an ancestral condition. Consequently, we should also take this direction as the default for evolutionary etiology unless future empirical evidence on the evolution of morality is found to suggest otherwise.

Generalization and the experience of obligations as externally imposed: Distinct contributors to the evolution of human cooperation

doi:10.1017/S0140525X18000134, e108

Elizabeth O'Neill

Department of Philosophy and Ethics, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands.

erh.oneill@gmail.com www.erhoneill.com

Abstract: It is worth distinguishing two phenomena involved in moral externalization: the experience of moral obligations as externally imposed and the tendency to generalize moral obligations. I propose that each played a distinct role in creating the conditions under which characteristically human cooperation could evolve.

Stanford seeks to explain why humans tend to externalize or objectivize moral obligations and other aspects of morality. His proposal is that externalization helped humans solve an adaptive challenge – namely, the challenge posed by a context in which cooperating flexibly, in a variety of circumstances, with potentially unknown partners, would be adaptive if one could avoid partnering with likely exploiters.

However, there are two distinct phenomena involved in what Stanford terms externalization, and it is worth distinguishing more explicitly what role each of these could have played in creating the conditions under which characteristically human cooperation evolved.

One phenomenon is the experience of moral demands as *externally imposed*. A weak version of this presents moral demands as independent of one's own desires or preferences. A stronger version presents moral demands as independent of the desires and preferences of any other agent, such as an authority or members of a social group. Sometimes this idea is discussed under the label of the conditionality or categoricity of moral principles (Southwood 2011); it is also related to the idea of response-independence (Nichols & Folds-Bennett 2003). A second phenomenon is the tendency to *generalize* – to believe that any moral demand that applies to oneself also would apply to any other sufficiently similar agent in similar circumstances. Southwood (2011) characterizes this idea as a matter of the scope of moral principles.

Conceptually, these two phenomena are separable. Agent-specific (non-generalized) moral demands can be experienced as externally imposed; generalized obligations (applicable to any agent in the relevant circumstances) can be experienced as authority-imposed. The distinction between these two phenomena appears in Tomasello's (2016) characterization of objectivity: He presents objectivity as involving a "three-way

generality – agent, target, and standards" (pp. 102–103). Tomasello's *standards* generality corresponds to the idea of moral demands as externally imposed, and *agent* generality corresponds to the belief that an obligation applies to any agent in the relevant circumstances. (His third form of objectivity, *target* generality, involves a tendency to treat one's partners and members of one's community as equals; Stanford's account of the evolution of cooperation does not involve this form of objectivity, nor do I think it should.)

More needs to be said about the relationship between these two phenomena and the specific role that each could have played in helping produce correlated interactions among humans inclined to cooperate. Here I will suggest that neither tendency would be adequate on its own to sustain the necessary correlated interactions, but combined they could be adequate.

At first glance, it might appear as if the tendency to experience moral demands as externally imposed contributes to the evolution of cooperation in the following way: Having a reputation for experiencing moral demands as desire/preference independent makes one a more attractive partner, by making it less likely that one's perception of one's own obligations – on which a partnership relies – will shift. By itself, however, the tendency to experience one's own moral obligations as desire/preference independent is not enough to resolve the adaptive challenge, because it supplies no mechanism for ensuring selective cooperation: one remains vulnerable to exploitation. This part of the problem, then, must be addressed via the disposition to generalize moral demands: For example, if one believes one is obliged to share excess resources with others in need, then one believes others in similar circumstances are obliged to do the same. Presumably, the tendency to generalize obligations produces or comes with a preference for cooperation partners who believe themselves to be subject to the same obligations that one believes oneself to be subject to (and perhaps, more specifically, for cooperation partners who generalize the same obligations that one generalizes oneself). This sort of preference, it seems, is what would produce the correlated interactions that are necessary to explain the origins of characteristic human cooperation.

So, is the tendency to experience moral demands as desire/preference independent actually required for explaining cooperation – could cooperation have come about just via generalization and a resulting preference for partners who generalize the same obligations? I suggest that the tendency to experience obligations as externally imposed supplements the generalization tendency by stabilizing it.

Experiencing generalized obligations as externally imposed can *preserve* one's perception of obligations as generalized, and thus preserve both a preference for others who generalize their obligations and one's attractiveness to others who prefer partners that generalize their obligations. If one treats a generalized obligation as desire/preference independent, one experiences the obligation to be externally imposed both on oneself and on others in similar circumstances; one may also take the obligation's generality itself as an externally imposed feature. By contrast, imagine that one does not treat the generalized obligation as desire/preference independent – one experiences the generalized obligation to share excess resources with others in need as contingent on, for example, the changeable command of an authority. In such a case, one remains a risky partner for others: in the future, one may cease to believe that anyone has such an obligation, or one may come to de-generalize the obligation and believe that it does not apply to oneself. In addition, for different reasons, the instability of the generalization in such a case also poses a risk to oneself. Suppose the relevant authority de-generalizes the obligation to share excess resources – specifically, by suspending it for others but not for oneself. One would be made vulnerable to exploitation in the sense that one would lose one's preference for partnering with others who take themselves to have this obligation; one would lose one's mechanism for avoid exploitation by partner selection.

So, I propose that the tendency to experience obligations as desire/preference independent and the tendency to generalize obligations each play a distinct role in the explanation for characteristically human cooperation. The generalization tendency can bring about correlated interaction, via a preference to partner with those who generalize their obligations. However, the generalization tendency is not inherently stable. Experiencing generalized obligations as externally imposed can stabilize one's tendency to generalize those obligations, thereby reducing the risks of exploitation that arise if one's own or one's partner's experience of the obligation shifts away from generalization.

Do the folk need a meta-ethics?

doi:10.1017/S0140525X18000146, e109

Shivam Patel[a] and Edouard Machery[b]

[a]*Department of Philosophy, University of Pittsburgh, Pittsburgh, PA 15260;* [b]*Department of History and Philosophy of Science, and Center for Philosophy of Science, University of Pittsburgh, Pittsburgh, PA 15260.*

smp119@pitt.edu **machery@pitt.edu**

http://www.philosophy.pitt.edu/person/shivam-patel-0

http://www.edouardmachery.com

Abstract: Stanford argues that cooperators achieve and maintain correlated interaction through the objectification of moral norms. We first challenge the moral/non-moral distinction that frames Stanford's discussion. We then argue that to the extent that norms are objectified (and we hold that they are at most objectified in a very thin sense), it is not for the sake of achieving correlated interaction.

Stanford proposes that cooperators achieve and maintain correlated interaction through the objectification of moral norms (target article, sect. 5, para. 7). We argue that the notion of objectification is idle in explaining how cooperators come to interact with each other to the exclusion of non-cooperators.

Stanford frames his argument in terms of a distinction between moral and non-moral norms, and takes Turiel (1983) and Skitka et al. (2005), among others, as evidence that this distinction is central to human psychology (sect. 2). We are skeptical (Machery 2018; Machery & Mallon 2010; O'Neill & Machery, forthcoming). While there is no doubt that Westerners distinguish moral and non-moral norms, this distinction appears to be culturally specific. As Stanford himself notes, Turiel's work has come under serious criticism (Kelly et al. 2007). Furthermore, many languages do not lexicalize the distinction between moral and non-moral norms (Wierzbicka 2007, p. 68), a surprising fact if the moral domain were a fundamental feature of human cognition. Furthermore, unpublished results suggest that Indian as well as Muslim participants (of various national origins) do not draw any distinction between moral and non-moral norms (see Machery [2012] for a description of the research project).

Given this evidence, it is a virtue of Stanford's argument that it does not need to be framed in terms of moral norms; it only requires that norms related to cooperation be objectified and that their objectification explains correlated interaction without exploitation. One problem in assessing this argument is that Stanford employs at least two notions of objectification. He begins by endorsing a thin notion of objectification: "We see ourselves as *obligated* to [satisfy the demands of morality] *regardless* of our subjective preferences and desires, and we regard such demands as imposing unconditional obligations not only on ourselves, but also on any and all agents whatsoever, regardless of *their* preferences and desires" (sect. 1, para. 2, emphasis in the original). This Kantian notion combines the "categorical nature" of norms (norms apply to us independently or our desires) and their universality (they apply to everybody). However, when discussing Nichols and Folds-Bennett (2003), Stanford endorses a

stronger notion: Moral facts are not response-dependent in contrast to facts about what is boring or yucky. He also comments approvingly on Goodwin and Darley's (2008) results, according to which moral norms are judged to be almost as objective as empirical facts. Here objectification is closer to what is known as "projectivism," the tendency to treat subjective impressions as if they were mind-independent properties of objects (e.g., Hume 1738/1975).

We do not doubt that in *some* sense objectification is a universal phenomenon (O'Neill & Machery, forthcoming) – in all cultures, *some* norms (probably not the same across cultures) are objectified in *some* sense – but we doubt that objectification is the key to interaction among cooperators without exploitation, contrary to what Stanford proposes (sect. 5, para. 7).

First, projectivism is stronger than needed to explain correlated interaction among cooperators. Correlated interaction is routinely achieved by means of norms cooperators understand as conventional. People routinely view tipping, giving up one's seat in the bus to an elderly person, holding the door for someone else, and so forth, as conventional norms that vary from one culture to another. In all such cases, those within a given community demand norm-conforming behavior from each other, thereby generating correlated interaction within the community, but do not demand such behavior from foreign communities. We are also skeptical that people really *project* their norms in the first place. For example, people in the United States and in Singapore view moral disagreements between extremely different groups as faultless, a fact that is difficult to account for if people treat norms as response-independent (Sarkissian et al. 2011).

Stanford may respond that even conventional norms are objectified in *some* sense: People view conventional norms as holding independently of their own subjective preferences (they are categorical) and expect others in their own community to comply with them independently of their subjective preferences. Whether I like it or not, I must drive on the right in the United States, and I expect other Americans to do the same, whether or not they like it. Although such norms do not have universal application, and so fail to live up to Stanford's thin objectification, they nevertheless resemble thinly objectified norms insofar as they have a *general* application (they apply to all group members).

But we doubt that even this form of thin objectification has much to do with correlated interaction among cooperators. First, norms *in general* (be those about dress code, politeness, taboos, etc.) are objectified in the sense of being categorical and generalized to others, whether or not they bear on cooperation (Foot 1972). Second, the best explanation of cooperative behavior – in terms of punishment and reputation-tracking (e.g., Balliet et al. 2011) – requires no appeal to even this form of objectification. Indeed, Stanford criticizes such accounts on the ground that they do not require that agents objectify norms (sect. 3, para. 5). We think Stanford's criticism only serves as grist for our mill: Since punishment and reputation-tracking provide for a robust explanation of cooperative behavior, and since such explanations can be parsed in terms of subjective preferences, we should expect from the outset that objectified norms will be idle in explaining correlated interaction among cooperators.

Our discussion cleaves apart the objectification of norms from correlated interaction among cooperators. We are left with the question that motivates Stanford's research: What evolutionary reason do we have to treat some norms as objective, at least in the thin sense we have identified? Although we have no space to develop this point here, we believe that thin objectification (the categorical nature of norms plus their generalization to a given group) has more to do with the evolutionary importance of capacities to make promises and engage in contracts. When people make a promise, they commit themselves to fulfill it independently of their current and future desires; similarly, when two parties engage in a contract, they commit themselves to respect it, independently of their current and future desires.

Is all morality or just prosociality externalized?

doi:10.1017/S0140525X18000158, e110

Michael J. Poulin

Department of Psychology, University at Buffalo, Buffalo, NY 14260.
mjpoulin@buffalo.edu **https://ubwp.buffalo.edu/scope/**

Abstract: It is more likely that externalized morality that facilitated cooperation (externalized prosociality) was selected for versus other types of moral impulses. Recent research suggests that those other moral impulses may actually be at root prosocial, in that judgments about them are indirectly about avoidance of harm. Externalized prosociality may help explain why prosocial behavior benefits individuals.

Stanford sets out to answer just one big question: Why do people externalize moral precepts? But in the process he offers an intriguing answer to another, potentially bigger question: Why do people cooperate with each other so much?

One long-standing possible answer to this question has been that people evolved to benefit from the action of indirect reciprocity, whereby people (a) prefer to cooperate with people who have a reputation of being cooperators, and (b) prefer to cooperate when doing so will bolster their own reputations (Nowak 2006). However, it is unclear that eliminating reputational concerns eliminates people's bias towards cooperation (Cooper et al. 1996). Moreover, this framework depends on the likelihood of two independent behavioral preferences being selected for: a preference for cooperators and a preference for behaving in reputation-bolstering ways. Stanford elegantly resolves these issues by noting that a belief that moral demands are real should motivate both a preference for moral action and a preference for moral actors.

This argument is at its most compelling if one imagines that the moral demand that is externalized is directly tied to the fitness of networks of individuals with externalized moral beliefs, such as a moral demand to help or at least not harm others – that is, if externalized morality is in fact externalized prosociality. By contrast, this argument is weakest when it is forced to confront the fact that there appear to be many types of externalized moral demands that do not seem inherently cooperation- or fitness-enhancing. As Stanford acknowledges, work by Haidt and colleagues (e.g., Haidt et al. 1993) has revealed that people judge as objectively wrong activities "such as privately washing the toilet bowl with the national flag and privately masturbating with a dead chicken" (target article, sect. 2, para. 2). Stanford subsequently describes these judgments as "partial or incomplete forms of moralization" (sect. 5, para. 12), but it is unclear on what grounds they do not count as fully moral. Although Stanford contextualizes this by noting that moral norms more directly focused on cooperation (e.g., harm and fairness) are more ubiquitous if not universal, it would nonetheless undermine his argument as to the purpose of moral externalization if other forms of morality could readily and equivalently be externalized.

Fortunately for Stanford's argument, there is actually independent evidence that the diverse kinds of moral judgment described by Haidt and others actually may be explainable in terms of the moral imperative to avoid harming others. That is, those who condemn seemingly victimless immoral acts do so in large part because they perceive that there are in fact victims of those acts (Gray et al. 2014), and perceptions of harm explain nearly all of the intensity of these moral judgments (Gray & Schein 2016; Schein et al. 2016). In other words, even moral precepts that do not directly affect others are nonetheless experienced as moral – and perhaps externalized – to the extent that they can be tied to harm. These findings would seem to bolster Stanford's claim that externalized morality evolved in order to facilitate cooperation, but they also suggest a possible refinement to Stanford's proposal: It may not be the case that people externalize moral norms as a whole, but, more specifically, that people externalize the moral imperative to help or at least not harm others.

This idea is intriguing for many reasons, but one reason is that it suggests something genuinely new about the psychology of altruism and cooperation. Psychologists have recognized for quite a while that people engage broadly in altruistic (or, less contentiously, "prosocial") behavior in part because "it feels like the right thing to do." However, until now, the best explanation for this sense has been that prosocial behavior is incentivized (and antisocial behavior disincentivized) by social norms that are accompanied by rewards and sanctions (e.g., Cialdini et al. 1990). Although these norms and the implied rewards or sanctions can be internalized, it is nonetheless the case that this account of prosocial behavior requires reference to social norms. By contrast, if externalized moral demands have themselves been selected for, per Stanford's account, it is possible that prosocial behavior simply feels right, full stop. This might help explain the finding that prosocial behavior is often automatic and intuitive, and undermined by deliberation (for a review, see Rand 2016).

Moreover, perhaps the fact that cooperating or acting prosocially feels right could help explain why it feels good. Acting prosocially predicts greater levels of emotional well-being (Caprara & Steca 2005; Cialdini & Kenrick 1976; Dunn et al. 2014) and even predicts increased health and longevity (Brown et al. 2003; Poulin & Holman 2013; Poulin et al. 2013). It could be that acting in accordance with externalized moral principles confers some of these benefits. The fact that prosocial behavior appears to be most beneficial when offered to others who are believed to be good and trustworthy (Poulin 2014) fits well with the role of externalized morality in coordinating both prosocial behavior and preferences for prosocial others.

Of course, if the benefits of prosocial behavior accrue because of a belief that one is acting on externalized moral principles, this would also suggest that these benefits would not occur for those who failed to externalize the morality of being prosocial. This is at once a testable prediction and a cause for reflection on how humanity's moral nature can fit into modernity. On the one hand, it seems possible we may have evolved to act morally, and enjoy the benefits of doing so, by believing that moral principles are absolute. On the other hand, we have rational reasons to believe that they are not. This tension may be a burden on individuals, and on society.

Moralization of preferences and conventions and the dynamics of tribal formation

doi:10.1017/S0140525X1800016X, e111

Don Ross[a,b,c]

[a]*School of Sociology, Philosophy, Criminology, Government, and Politics, University College Cork, Cork T12 AW89, Ireland;* [b]*School of Economics, University of Cape Town, Private bag, Rondebosch 7701, South Africa; and* [c]*Center for Economic Analysis of Risk, J. Mack Robinson College of Business, Georgia State University, Atlanta, GA 30303.*

don.ross931@gmail.com http://uct.academia.edu/DonRoss

Abstract: Stanford casts original light on the question of why humans moralize some preferences. However, his account leaves some ambiguity around the relationship between the evolutionary function of moralization and the dynamics of tribal formation. Does the model govern these dynamics, or only explain why there are moralizing dispositions that more conventional modeling of the dynamics can exploit?

Stanford's problem can be succinctly expressed: Why do humans moralize some preferences? This means: Most humans treat some of their preferences as expressing objectively grounded, universally binding norms. Further content consistent with Stanford's evidence and reflections can be suggested. A moralized preference is one that the moralizer cannot propose to trade off against other preferences without expecting to experience shame, and without expecting others to legitimately regard her as shamed, except in cases where two or more moralized preferences unavoidably and clearly enjoin opposed actions.

Stanford's solution to the problem is innovative and broadly convincing. It allows him to explain various psychological and social features of moralization, and yields a neat explanation of why moral judgments have been so philosophically perplexing. How well supported by available evidence is this solution? It depends on evidence about moralization *and* on evidence about the pressures on cultural adaptation that Stanford invokes to explain moralization. I will focus on the latter.

The crucial driver of Stanford's model is what he characterizes as humans' unique plasticity. He is not as explicit about this as he might be. Plasticity typically refers most directly to learning capacity. But what mainly does the work in Stanford's model is a *consequence* of learning capacity – namely, the observed fact that people have colonized a remarkable range of niches. This has in turn given rise to, and required cultural adaptation to, a diversity of lifeways. On Stanford's account, this continuous dynamism disrupts stabilization of coordination and control of free-riding by mere conventions entrenched in motivational drives. Moralization has allowed people to repeatedly segregate themselves into tribes which are, according to Stanford, endogenously equilibrated but still potentially unstable because of their continuing dispositions to construct or find new niches. The potential instability preserves the functional value of moralization.

Stanford says little about the dynamics of tribal formation. One might naturally think of our ancestors radiating from warm grasslands and scrubland into climates with cold winters or dense forests. But human tribes manifestly bifurcate *within* shared physical environments. On the face of it, this seems to be just what Stanford's hypothesis predicts: A subset of a founder population in a niche moralizes some of its new conventions in order to achieve and maintain correlated equilibrium and successfully exclude those most disposed to free-riding. Then, presumably – Stanford is not explicit on this point – the excluded villains interact with one another for lack of an alternative, and form and then moralize different conventions.

Does the tribe of cast-offs moralize for the same reason as the original moralizers? If so, they should be expected to spin off yet another tribe, and we predict a recursive pattern that perhaps terminates in the creation of marginal "sick societies" (Edgerton 1992), where free-riding is impossible because benefits from cooperation have shrunken to their biologically minimal core. One might speculatively imagine a sick society that, forced into geographic isolation, endures for long enough that natural section catches up to cultural selection and we end up with normative psychology resembling that of chimpanzees.

An alternative account, also consistent with Stanford's dynamics but making it less general as a model, might go as follows. We begin with a first stage of social evolution in which the "Stanford process" gives rise to the natural disposition to moralize suggested by the developmental evidence that Stanford cites. Once this biological adaptation has occurred, we enter stage two, and tribal formation with rival moral codes is supported by more familiar strategic dynamics of cultural group selection: To compete successfully, groups need effective solidarity; therefore, they need costly entry barriers and membership fees that reliably signal commitment; moralization that limits individual freedom of action serves this function, along with the function of making members inadmissible by rival groups with different moral codes and aligning their self-interest with militant patriotism. Tribal formation itself might then be mainly governed by exogenously varying resource constraints that constitute parameters on stable tribe sizes, with new tribes forming whenever, on the margin, some people are better off forming a new tribe than receiving a diminishing share of the pie generated by the existing tribe. Generation of new niches (i.e., "plasticity") on this second interpretation might mainly, in stage two, help make new moral codes relatively economically adaptive, thereby counterbalancing the initial disadvantages typical of a smaller start-up.

On the first interpretation given above, Stanford's model governs the dynamics of tribal formation. On the second interpretation, it explains why these dynamics find moralizing dispositions to exploit, but the dynamics themselves are modeled by a fusion of anthropological group selection and the economics of dynamic industrial organization.

One would be forced to disambiguate these interpretations if the model were formalized. This leads to a methodological observation about evolutionary psychology. Economists prefer formal models partly because these generate relatively precise discriminating empirical predictions that might not be evident to the theorist in advance of the formal specification. Such predictions are important not because prediction is the primary goal of science (it is not), but because specification of predictions is a crucial tool for identifying a model's empirical scope.

We see scope ambiguity in Stanford's informal model if we compare humans not only with chimpanzees, but also with more genetically distant animals that are more similar to humans along some social dimensions. Whereas chimpanzees form rival, warring groups but *not* morally differentiated tribes, orcas resemble humans in forming geographically overlapping communities with strikingly different core behaviors, communication codes, and social organization. Individuals drawn from different groups housed together in captivity don't seem to get along well. Orcas inhabit all oceans, and we have no independent metric for determining how profoundly or shallowly their inhabited functional niches vary *with respect to what matters to them*. Is Stanford's model intended to be sufficiently general to be used in deciding whether we should predict morality in, for example, killer whales? Or is it intended mainly to explain a dimension of divergence in the ape/hominid evolutionary line? Formalization of the model might usefully force the distinction.

Externalization of moral demands does not motivate exclusion of non-cooperators: A defense of a subjectivist moral psychology

doi:10.1017/S0140525X18000171, e112

Armin W. Schulz

Department of Philosophy, University of Kansas, Lawrence, KS 66045.
awschulz@ku.edu http://people.ku.edu/~a382s825/

Abstract: It is not clear how a moral demand alone can motivate an agent to exclude those who fail to act as the demand states. A more plausible hypothesis for the evolution of human moral cognition is based on seeing moral demands as subjective, but inherently conjunctive. This subjectivist-conjunctive proposal can still account for the apparent externalization of moral demands.

Stanford suggests that an explanation for the apparent "externalization" of moral demands can be found in the fact that this externalization ensures agents are simultaneously motivated to (a) engage in (frequently adaptive) cooperative interactions, and (b) exclude from these interactions (or even punish) those that are not motivated to cooperate. In this way, the externalization of moral demands is said to be an efficient way of leading to the twin behaviors – cooperation and exclusion/punishment of non-cooperators – needed to maintain adaptive cooperative arrangements (also taking into account the ancestral state of the cognitive and conative aspects of human psychology).

However, a key element of this explanation of the (apparent) externalization of moral demands is unconvincing. In particular, while it is relatively clear how an externalized moral demand – such as "injured in-group members need to be helped" – can motivate an agent to act as the demand states, it is not clear how an externalized moral demand alone can motivate an agent to exclude (or even punish) those who *fail* to act as the demand states. After all, moral demands do not (typically) state what should be done when they are *violated*: they prescribe what agents are to do, not what agents are to do when others have *not* done what they are to do.

For this reason, Stanford's proposal implicitly presumes that the externalization of moral demands co-evolved with other motivational structures. In particular, it must be assumed that agents do not just externalize moral demands, but are also motivated to exclude/punish those that violate them – rather than, say, try to educate them or learn from them (both of which are also a priori coherent responses to this situation). In this way, though, the major attraction of the author's explanation of the externalization of moral demands is lost: We are back at requiring separate motivational states for cooperation and the exclusion/punishment of non-cooperators.

Given this, I suggest that a more plausible hypothesis of the evolution of (the relevant aspects of) human moral cognition is based on seeing moral demands as subjective, but inherently *conjunctive*. More specifically, I suggest that there are evolutionary reasons to expect agents to be motivated by moral demands of the following sort: "Injured in-group members need to be helped *and* non-helpers of injured in-group members are not to be helped." Note that this is not just a re-description of Stanford's proposal. Stanford's suggestion is that moral demands are somewhat akin to objective facts, not personal opinions. By contrast, my suggestion is that moral demands are akin to personal opinions, not objective facts – it is just that they are personal opinions that speak to, and thus motivate, a wider range of actions than other personal opinions. There are two reasons for why this purely subjective proposal is more plausible than Stanford's suggestion.

First, the subjectivist-conjunctive proposal still leads to the adaptive connection between cooperation and the exclusion/punishment of non-cooperators that Stanford has rightly emphasized. In turn, this implies that the same selective pressures appealed to in the target article – namely, the fact that the existence and maintenance of adaptive (hyper)cooperative arrangements depend on cooperators being consistently likely to interact with other cooperators (Skyrms 1996; 2004; Sober & Wilson 1998) – also operate here.

Second, the present subjectivist-conjunctive proposal is more in line with the ancestral condition of human (moral) psychology. As Stanford notes, there is good reason to think that humans started out with a decision-making machinery that included a battery of stored reflexes as well as representational (content-based) mental states, some of which are imperative (conative) in form, and some of which are indicative (cognitive) in form (see also Millikan 2002; Schulz 2011; 2013; 2018; Sterelny 2003). In Stanford's proposal, moral cognition evolved by taking a subjective motivational (conative) state, and shifting it closer to a cognitive state. However, a far simpler change – with, as just noted, the same dual-motivational outcome – would be to just expand on the *content* of some of the agent's motivational states: Instead of motivating just one type of behavior, it changed to motivating several types of behavior simultaneously. Since smaller changes to an organism's existing traits are more likely to evolve than larger ones, this thus favors my proposal over Stanford's.

Finally, it is important to emphasize that the subjectivist hypothesis suggested here still provides an explanation of the seeming externalization of moral demands. In the present proposal, the objectivity of moral demands stems from the fact that they are expressions of subjective preferences *both* for acting in certain ways *and* for not interacting with (or even punishing) those that do not share these preferences. That is, when people are asked to rank the "objectivity" of a moral demand, they rank it as closer to a fact than to a personal preference (Goodwin & Darley 2008) because moral demands motivate more behaviors than (most) personal preferences do – a feature they share with cognitive states, which are also relevant to a wide variety of different actions. In other words, the difference in how "objective" a

norm is taken to be just rests on the content of the relevant motivational state, not the (second-order) attitude the agent takes towards that motivational state. Or, to put it succinctly: In my proposal, the difference between ice cream and Nazis is that, while I simply want to eat ice cream, I want to not be a Nazi *and* to not interact with those that want to be Nazis.

Do we really externalize or objectivize moral demands?

doi:10.1017/S0140525X18000183, e113

Stephen Stich

Department of Philosophy, Rutgers University, New Brunswick, NJ 08901.
stich@philosophy.rutgers.edu **http://www.rci.rutgers.edu/~stich/**

Abstract: Stanford's goal is to explain the uniquely human tendency to externalize or objectify "distinctively moral" demands, norms, and obligations. I maintain that there is no clear phenomenon to explain. Stanford's account of which norms are distinctively moral relies on Turiel's problematic work. Stanford's justification of the claim that we "objectify" moral demands ignores recent studies indicating that often we do not.

Stanford has offered an intriguing explanation of the uniquely human cross-cultural tendency to externalize or objectify "distinctively moral" demands, norms, and obligations. I am impressed by the ingenuity and sophistication of the explanation. But I am less impressed by the explanandum. To put my concern rather bluntly, I am not convinced there is anything to explain.

Let me start with "distinctively moral" – an expression that occurs 15 times in Stanford's article, modifying (among other things) "norm(s)," "transgressions," "obligation(s)," and "commitments." Which norms, transgressions, obligations, and commitments are distinctively moral? For the last half of the twentieth century, this was a *hot* topic in philosophy. Indeed, according to Alistair MacIntyre, writing in 1957, "[t]he central task to which moral philosophers have addressed themselves is that of listing the distinctive characteristics of moral utterances" (MacIntyre 1957, p. 325). But that project made very little headway. Though lots of accounts were offered, none gained wide acceptance. During the last two decades, discussions of how to distinguish moral norms, judgments, and transgressions from their non-moral counterparts have largely disappeared from the philosophical literature. Despite trying very hard for half a century, philosophers have been unable to tell us which norms (etc.) are distinctively moral. (For details and references, see Stich, forthcoming.)

So why does Stanford think that there *is* a distinctively moral class of norms, demands, and obligations, and which norms (etc.) are they? The answer, it appears, relies heavily on the work of Elliot Turiel and his associates. While philosophers were still actively debating the appropriate definition of morality, Turiel dipped into the philosophical literature, borrowed a few of the items that had been proposed as distinctive features of moral judgments and norms, added a few ideas of his own, and used these to construct a test – the moral/conventional task – that, he maintained, would tell us whether a person's normative judgment was a moral judgment or a conventional judgment. The moral judgments are the ones that a person takes to be authority independent, will generalize in time and space (if it is wrong here and now, it is also wrong at all other locations and at all other times), and will justify by appeal to harm, justice, or rights. There is, however, something rather puzzling about this. Why are *these* features the ones that characterize judgments that are distinctively moral? The puzzle is underscored when we note that in Turiel's account many norms that that have traditionally been taken to be prototypically moral (norms prohibiting masturbation, for example, or norms prohibiting blasphemy) are not moral norms at all.

One way of responding to this puzzle is to suggest that the features that Turiel specifies pick out a psychological natural kind of norms, and that it is appropriate to consider these to be the *moral* norms because the members of this class include many norms that would intuitively be classified as moral. This idea was first proposed by Kelly et al. (2007) and is elaborated in Kumar (2015) and Stich (forthcoming). But for this response to be workable, the features have to be a "nomological cluster" with a strong tendency to all be present or all be absent, and there is now a long list of studies finding that the features don't cluster in this way. In response, one might add and drop features to the set that putatively characterizes the moral natural kind. There is some suggestion that Stanford is inclined to explore this option since he suggests that we should not consider being concerned with "harm, fairness, justice rights, or welfare" to be a "defining feature of moral norms" (sect. 2, para. 2). And, in contrast with Turiel and his followers, he apparently takes "seriousness" to be a defining feature. But on my reading of the evidence, there is little reason to believe that this set of features is a nomological cluster either. So I am left wondering what, exactly, Stanford has in mind when he talks of "distinctively moral" norms and why he thinks those norms really are distinctively moral.

Let us turn, now, to the tendency to externalize or objectify distinctively moral norms. What does this come to? Sometimes Stanford relies on the language of phenomenology: "[we] experience the demands of morality as somehow *imposed* on us externally" and "we regard such demands as imposing unconditional obligations not only on ourselves, but also on any and all agents whatsoever" (sect. 1, para. 2, emphasis in target article). Well, perhaps Stanford experiences the demands of morality in this way. But I don't recognize this as part of *my* moral phenomenology. Which one of us is an outlier? Stanford maintains that the work of Goodwin and Darley indicates that most people share something like his moral phenomenology. These researchers found that when participants were asked questions designed to determine whether they thought that ethical beliefs are objectively true or false, "ethical beliefs were treated almost as objectively as scientific or factual beliefs" (Goodwin & Darley 2008, p. 1359, quoted in Stanford target article, sect. 2, para. 4). Stanford notes, and attempts to accommodate, the work of Sarkissian et al. (2011), which seems to indicate that ordinary folk are much less objectivist than Goodwin and Darley suggest. But the Sarkissian et al. paper is just the first of a recent cascade of papers, all of which cast doubt on the conclusion that people are consistently objectivist about moral judgments (Beebe 2014; 2015; Beebe & Sackris 2016; Quintelier et al. 2014; for a review, see Sarkissian 2016). Moreover, even if we put these recent studies aside, there is a disconnect between the Goodwin and Darley findings and the account of "distinctively moral" norms proposed by Turiel. Neither the Goodwin and Darley studies nor other studies exploring moral objectivism make any effort to show that the moral judgments they focus on would pass Turiel's test, or anything like it. The assumption that the moral judgments that are the focus of Goodwin and Darley–style studies are "distinctively moral" (as Stanford apparently uses this term) is purely speculative. (For more on this issue, see section 3 in Stich, forthcoming.)

Not as distinct as you think: Reasons to doubt that morality comprises a unified and objective conceptual category

doi:10.1017/S0140525X18000195, e114

Jordan Theriault[a] and Liane Young[b]

[a]*Department of Psychology, Northeastern University, Boston, MA 02115;*
[b]*Department of Psychology, Boston College, Chestnut Hill, MA 02467.*
jordan_theriault@northeastern.edu **liane.young@bc.edu**
http://www.jordan-theriault.com/ **http://moralitylab.bc.edu/**

Abstract: That morality comprises a distinct and objective conceptual category is a critical claim for Stanford's target article. We dispute this claim. Statistical conclusions about a distinct moral domain were not justified in prior work, on account of the "stimuli-as-fixed-effects" fallacy. Furthermore, we have found that, behaviorally and neurally, morals share more in common with preferences than facts.

In the target article, Stanford argues that moral demands inhabit a distinct conceptual category, where they are experienced as externally imposed obligations; and that evolutionarily, this externalization protected prosocial individuals from exploitation, ensuring that any felt obligation to conform with a social norm was paired with a conviction that others should conform as well. Thus, externalizing moral demands (i.e., experiencing moral demands as objective) allowed individuals to reap the benefits of prosociality while also policing defectors.

One critical claim for this argument is that morality comprises "a distinctive conceptual category" (target article, sect. 5, para. 15, sect 6, para. 1). Work was reviewed, showing that children categorically distinguish morals from social conventions (Smetana 2006; Turiel 1983); moral properties are distinguished from response-dependent properties (e.g., "yucky"; Nichols & Folds-Bennett 2003); and, critically, morals are rated as categorically more objective than preferences and social conventions (Goodwin & Darley 2008; 2012), licensing the conclusion that "[moral beliefs are] treated almost as objectively as scientific or factual beliefs … [and] as categorically different from social conventions (Goodwin & Darley 2008, p. 1359). However, we doubt that morals comprise a distinct category – at least on the basis of their objectivity, universality, and authority-independence, as has traditionally been argued. First, prior work has not licensed statistical generalizations about a moral domain, as its authors had assumed. Second, our recent work suggests that objectivity is not an essential feature of morality; rather, behaviorally and neurally, moral claims are more akin to preferences.

To preface our statistical criticism: in most cases, morality must be studied using specific stimuli. For instance, stimuli might include asking children whether hitting (e.g., Wainryb et al. 2004) or stealing (Tisak & Turiel 1988) is acceptable. Goodwin and Darley (2008) asked about discrimination, robbery, and firing into a crowd, among others. The statistical problem is that one must move, by inference, from the specific stimuli to a sampled population (e.g., a moral domain). To make this inference, stimuli must be treated as a *random effect* (i.e., as a random sample from a population). If stimuli are averaged, then statistical conclusions (e.g., a *t*-test across subjects) apply only to those stimuli; in this case, stimuli are a *fixed effect* (and this leaves aside issues with randomly sampling moral stimuli, a problem beyond the scope of this commentary). This "stimuli-as-fixed-effects" fallacy has been identified in other fields (Clark 1973) and is easily solved using mixed effects analyses (Baayen et al. 2008), but the problem has been largely ignored within psychology (Judd et al. 2012; Westfall et al. 2014). The moral/conventional distinction has been criticized on the basis of stimulus content (e.g., stimuli typically describe "schoolyard" violations; Kelly et al. 2007, p. 121), and these criticisms may be justified; but ultimately, such content-based criticisms are unnecessary, as the original findings never licensed conclusions about a moral domain at all. They licensed conclusions about the *exact* stimuli that were used.

Prior work has argued that morality is essentially objective, but how to measure meta-ethical judgment has also been a long-standing concern (for an excellent discussion, see Goodwin & Darley 2010). Stanford rightly calls attention to the "hybrid character" of morality, where moral claims fall somewhere between "[objective] representations of how things stand in the world itself … and our subjective reactions to those states of the world" (sect. 6, para. 11); however, prior work has rarely allowed participants to express this hybrid nature. For example, if participants are forced to classify moral claims as true, false, or an opinion/attitude (Goodwin & Darley 2008), then distinctions between morals, facts, and preferences may appear to be more discrete than they actually are. We attempted to address this issue in a recent study (Theriault et al. 2017), where participants read moral claims (presented alongside facts and preferences) and simultaneously rated the extent that each was "about facts," "about preferences," and "about morality" (1–7; "not at all" to "completely"). Moral claims should be more moral-like than fact-like or preference-like; however, the question of interest was which secondary feature would dominate: Are morals largely fact-like? Or are they largely preference-like?

Although prior work has emphasized that moral claims are essentially objective, our work suggested the opposite: that moral claims were perceived as largely preference-like. Among a set of 24 moral claims that we had generated, and also among 22 claims adapted from the moral foundations questionnaire (Graham et al. 2011; Iyer et al. 2012), participants rated moral claims as significantly more preference-like than fact-like. Furthermore, we scanned subjects as they read the same claims, and found that moral claims elicited widespread activity in brain regions for social cognition and theory of mind (Schurz et al. 2014; Van Overwalle 2009), overlapping with activity for preferences, but not facts. Stanford argues that humans have "[gone] in for cognitively complex forms of representation," and that moral norms have likely been shoehorned into an evolved framework where "the most fundamental division … [is] between how things stand in the world … and our subjective reactions to those states" (sect. 6, para. 11). If this fundamental representational division exists, and if moral demands were (in part) externalized by co-opting cognitive processes that evolved to represent the world, as Stanford seems to argue, then we should see at least some significant overlap between processing for morals and facts. Instead, morals were behaviorally perceived, and neurally represented, as akin to preferences.

Nevertheless, we agree that moral demands are often experienced as external, even if the moral domain is not as unified as prior work has suggested. But this moral externalization may exist along a spectrum: Some moral claims may be experienced as more objective than others. Indeed, we characterized this variability in a recent study, where by-stimuli moral objectivity tracked with activity in social brain regions (Theriault et al., under review). Understanding this variability will be critical for an account of why some moral demands are experienced as obligatory and enforced (e.g., "murder is wrong") whereas others are not (e.g., "eating meat is wrong").

Moral demands truly *are* externally imposed

doi:10.1017/S0140525X18000201, e115

Jan-Willem van Prooijen

Department of Experimental and Applied Psychology, Vrije Universiteit (VU) Amsterdam, 1081 BT Amsterdam, The Netherlands; and The Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), 1008 BH Amsterdam, The Netherlands.

j.w.van.prooijen@vu.nl http://www.janwillemvanprooijen.com

Abstract: Most moral demands indeed *are* externally imposed, as violations are subject to social condemnation. While in modern society objectified moral demands may serve as a cue for desirable interaction partners, human morality evolved in small tribes that offered little choice regarding with whom to cooperate. Instead, it was adaptive to objectify moral demands to avoid the costs of social exclusion.

Why do people experience moral demands as externally imposed? People feel obliged to adhere to central moral demands independent of one's own subjective preferences, and also impose these demands on others regardless of their preferences. To solve this puzzle, Stanford points at the functionality of moral externalization to select social interaction partners, to jointly solve common

challenges in both familiar and unfamiliar circumstances. By experiencing morality as externalized, people require others to conform to those norms, allowing one to build cooperative networks that are prosocial yet protected from exploitation. In the present commentary, I propose that although there needs to be little dispute over the importance of human cooperation in explaining the evolution of morality, Stanford ignores a more parsimonious and plausible explanation for the process of moral externalization: People experience moral demands as externally imposed because, frequently, these demands actually *are* imposed by their immediate social environment, and they *are* obliged to follow them. Such conformity pressures stimulate intrinsic agreement with these moral demands. Human morality evolved in small tribes of ancient hunter-gatherers that offered little choice as to whom to cooperate with, making it adaptive for individuals to maximize their own adherence to the moral demands of their group by objectifying them.

Unlike ice-cream preferences or norms of convention, moral norms have a special status in protecting the common interests of social groups. Violations of moral norms typically are intentional, place the self-interest above the collective interest, and lead to outcomes that are detrimental to the group. These considerations converge with the five foundations of human morality (Graham et al. 2009). Moral violations reduce the evolutionary fitness of fellow group members through harm, unfairness, disloyalty, disobedience, and possible contamination. To sustain high levels of cooperation, group members therefore have good reason to act in a condemning, punitive manner towards moral offenders. Such condemnation does not need to depend on the altruistic acts of a single punisher and can in fact be relatively cost-free for the community. Common responses to moral offenders include public ridicule, gossip, or coalitional punishment. But while the costs of such sanctions for the community are low, they often are high for the offender who might face social exclusion, reputation damage, and decreased access to resources or reproductive opportunities. It has been noted that ancestral humans evolved a moral conscience, which includes moral externalization, to protect them against these condemning responses by promoting intrinsic motivations to not only follow, but also actively enforce themselves, the moral demands of their group (Boehm 2012; DeScioli & Kurzban 2009).

This explanation more parsimoniously captures the origins of moral externalization, given that it does not need to make one central yet problematic assumption in Stanford's line of reasoning, which is that group members are free to choose whomever to cooperate with. Such freedom might apply to contemporary large states, which offer virtually infinite interaction opportunities among citizens and an unprecedented flexibility in the possible cooperative networks that one does, or does not, wish to be part of. It is questionable, however, whether the ancient hunter-gatherer societies in which human morality evolved offered equal flexibility in selecting interaction partners or cooperation opportunities. These small-scale societies often faced threats that needed to be dealt with collectively, including extreme climate variations, droughts, famines, and wars. Such challenging circumstances require high levels of cooperation among all group members, and hence are likely to promote strong norms against exploitation and free-riding that apply unequivocally across the group. Individual members usually could not switch to a different tribe if they disagreed with the moral demands of their group. Instead, a more adaptive strategy for individual group members was to "go along to get along," implying selection pressure to evolve a sense of morality that includes experiencing central moral demands as objective truths that need to be enforced.

This dynamic interplay where externally enforced norms shape moral judgments and actions is still visible in contemporary research within the social sciences. For instance, research on moral hypocrisy suggests that people often are more strongly motivated by *appearing* moral than by *being* moral. When given the opportunity, many research participants chose to flip a coin in order to fairly divide tasks between themselves and another person, yet rigged the coin flip to acquire the most positive outcome for themselves. Intriguingly, participants who rigged the coin flip considered themselves as more moral than participants who chose selfishly without flipping a coin, suggesting that only appearing to satisfy moral demands – to both oneself and others – is sufficient to positively shape moral self-perception (Batson et al. 1999). Relatedly, the actual or implied presence of others increases prosocial behavior, which underscores the role of reputation in human morality (Haley & Fessler 2005). These research examples illuminate that moral judgments and actions are highly susceptible to social evaluations, which is consistent with the idea that human morality is largely shaped by conformity pressures.

An important element of Stanford's theory is the notion that people build and maintain cooperative communities in which they are protected from exploitation. Establishing such cooperative communities indeed is a motivation of human beings where the difference between moral versus amoral preferences is particularly salient. Information that a person prefers vanilla over chocolate ice-cream implies nothing for the likelihood that this person will exploit other group members, but information that a person has a history of moral violations is highly diagnostic for the type of contributions this person will make to a group. But whereas in modern times the implications of these issues are highly flexible, and in line with Stanford's theory – we can avoid people with Nazi sympathies, but people with Nazi sympathies can form a cooperative network with other Nazis – in ancestral times it is questionable whether moral deviants could easily choose to form their own network. Given the costs of social exclusion, more likely is that they largely adjusted to the norms of the majority, and internalized crucial moral demands as their own. Moral demands truly are externally imposed.

The objectivity of moral norms is a top-down cultural construct

doi:10.1017/S0140525X18000213, e116

Burton Voorhees,[a] Dwight Read,[b] and Liane Gabora[c]

[a]*Center for Science, Athabasca University, Athabasca, Alberta, T9S 3A3, Canada;* [b]*Department of Anthropology, University of California, Los Angeles, Los Angeles, CA 90095;* [c]*Department of Psychology, University of British Columbia, Kelowna, British Columbia, V1V 1V7, Canada.*

burt@athabascau.ca **dread@anthro.ucla.edu**
liane.gabora@ubc.ca
http://science.athabascau.ca/staff-pages/burtv
http://www.anthro.ucla,edu/faculty/dwight-read
http://people.ok.ubc.ca/lgabora

Abstract: Encultured individuals see the behavioral rules of cultural systems of moral norms as objective. In addition to prescriptive regulation of behavior, moral norms provide templates, scripts, and scenarios regulating the expression of feelings and triggered emotions arising from perceptions of norm violation. These allow regulated defensive responses that may arise as moral idea systems co-opt emotionally associated biological survival instincts.

Regarding the evolutionary advantage of objectifying of systems of moral norms, Stanford says: "The creation of a novel conceptual category of norms or standards of behavior to which I hold both others and myself responsible simultaneously thus established a mechanism for safely *extending* prosocial, altruistic, and cooperative behavior in new ways and into new contexts" (sect. 5, para. 8, emphasis in target article). But he does not say *how* the creation of this novel conceptual category comes about and admits ignorance as to how exteriorization arises in individuals. Is it an individual trait, or is it something that is culturally induced? If seen as an

individual trait, a number of problems arise. In particular, how could it have ever arisen in a group of non-externalizers, and how could a group come to all externalize the same norms? In arguing the advantage of norm exteriorization, Stanford begs the question of why different individuals exteriorize the same norms.

It is important to note that a fundamental cognitive shift has taken place in humans from evolution at the individual level to evolution at the organizational level (Lane et al. 2009; Read et al. 2009). Conceiving a category of norms or standards of behavior required crossing a cognitive threshold – *individuals must be able to consciously conceptualize themselves as members of a reified group*. With this capacity, cultural idea systems (Leaf & Read 2012) become possible as complexes of beliefs and/or organizational rules that operate in a top-down manner so that individuals gain functionality only by adherence to these rules and/or constraints.

Cultural idea systems are internalized by individuals through enculturation and are taken by culture bearers as having objective reality (Spradley & Mann 1975), thereby providing shared meaning for the events of social life. In this way, human culture creates a "virtual" world, including moral norms that are experienced as universals, applicable to anybody who is considered as "one of us" (Bar-Tal 2000; Hardin & Higgins 1996). In the Upper Paleolithic, we see the beginning of cultural idea systems in the form of kinship systems (Bergendorff 2016; Read 2012), and moral norms are incorporated as part of kin expectation and obligation (Fortes 1969). These patterns of expectations and obligations provide the structures for coordinated cooperation within a group when all members share a kinship relation. Acting in accordance with the behavior expected of kin is important because survival depends on being integrated with one's kin.

If our ancestors' moral norms are part of the cultural idea system acting in a top-down manner within social systems organized through kinship relations, then kinship itself provides the objectivity and coherence of norm exteriorization. In hunter-gatherer bands, where kinship is the basis of all social relations, the obligation to cooperate with others sharing a kinship relation becomes part of the identity of group members. Those who act improperly as kinsmen are sanctioned by the group.

In much work on the evolution of cooperation, punishment is seen as an important factor for maintaining cooperative groups against free-riders. Stanford claims that norm exteriorization removes the need for punishment because individuals will protect themselves from exploitation by simply shunning those recognized as norm violators. This is not sufficient to establish the stability of a system of norms, especially in small hunter-gatherer groups where it may not be possible to avoid contact with or reliance on untrustworthy partners. Punishment-based arguments, however, must deal with the second-order free-rider problem – punishment requires that group members agree to bear the cost of punishing a transgression at some undefined future time; yet, if punishing becomes necessary, some group members may renege on their commitment.

We argue that the second-order free-rider problem, and also Stanford's question of "how moral norms acquire their characteristic status in the course of individual ontogeny" (sect. 5, para. 15), is solved through the linkage of culturally laden feelings and biological emotions within a cultural setting. Emotions are physiological responses to stimuli related to biological survival and are controlled by genetically established neural circuits. The feeling of an emotion is the mental experience accompanying the physiological sensations of the emotion (e.g., Damasio 2012; LeDoux 2012). Through association of feelings and behavior, culture provides functional vehicles for the social expression of emotional responses. Likewise, feelings triggered by culturally salient cues can evoke associated emotions (Damasio 2012; De Leersnyder et al. 2013; Kim & Sasaki 2012).

A cultural system of moral norms is not just a set of rules for behavior; it directs feelings associated with moral behavior or misbehavior that have been interjected by group members and arise automatically when cued. Misbehavior by a group member may lead to feelings of guilt or shame, while perception of a norm violation by another may evoke feelings of anger and indignation (Dubreuil 2010). These feelings may trigger emotional responses, and because the emotions are grounded in biological survival instincts, the perceived norm violation may be responded to defensively as if it threatened biological survival (Ellemers 2012; Ellemers et al. 2002; Voorhees et al. 2018). Culturally determined defensive responses can range from shunning (as posited by Stanford and others), to an impulse to punish, eliminate, or otherwise correct a violation of what is seen as objectively "right and proper."

In sum, cultural ideas, acquired through enculturation, are internalized by culture bearers and seen by them as objective reality. Among hunter-gatherers, behavioral norms are coded as patterns of expectations and obligations that are part of a kinship system. These provide the structure that facilitates coordination and cooperation of group activities. Rather than simply being collections of objectified behavioral rules, moral norm systems provide templates, scripts, or scenarios regulating the expression of feelings and emotions arising through the experience of violating a norm, or seeing another violate a norm. Only humans appear to have the psychological and neurological basis for both norm-following and sanctioning of violators (Dubreuil 2010; Read 2012), and we attribute this to the fact that only humans have the cognitive capacity to grasp the abstract concepts involved in cultural idea systems.

Disgust as a mechanism for externalization: Coordination and disassociation

doi:10.1017/S0140525X18000225, e117

Isaac Wiegman

Department of Philosophy, Texas State University, San Marcos, TX 78666.
isaac.wiegman@txstate.edu **https://isaacwiegman.wordpress.com/**

Abstract: I extend Stanford's proposal in two ways by focusing on a possible mechanism of externalization: disgust. First, I argue that externalization also has value for solving *coordination* problems where interests of different groups *coincide*. Second, Stanford's proposal also holds promise for explaining why people "over-comply" with norms through *disassociation*, or the avoidance of actions that merely appear to violate norms.

I want to suggest a few ways that Stanford's proposal can be extended by focusing on an emotion that may play a role in externalization – namely, disgust (cf. Nichols 2014, pp. 736–40). The psychological profile of disgust includes a sensitivity to contamination via contact, similarity, or association, and that results in avoidance of contaminated objects (Rozin & Fallon 1987). This suggests that disgust evolved for the avoidance of poisons, parasites, and pathogens (cf. Kelly 2011, pp. 52–59). Moreover, disgust appears to play an important role in human moral psychology (e.g., Rozin & Haidt 2013). For example, involuntary effects on facial muscles associated with distaste and disgust at contaminants have also been observed in reaction to perceptions of unfair treatment (Chapman et al. 2009).

Given the functional role of disgust and its penetration into the social and moral domain, it is plausible that – as Kelly has suggested – the process of gene-culture co-evolution co-opted disgust to motivate norm compliance and enforcement (Kelly 2011, pp. 116–22). On this view, if one comes to feel that certain acts are disgusting (e.g., acts that violate certain kinds of norms), one will avoid committing such acts oneself and one will avoid those who commit such acts (because they are contaminated thereby). Importantly for my purposes, one may also think

that others have reason to avoid such acts (since they would be contaminated thereby). If this is an accurate picture of disgust, then disgust appears to offer one way in which a person can come to experience "moral motivation as externally imposed on both ourselves and others simultaneously" (Stanford target article, sect. 5, para. 7). For instance, even if there is salient disagreement about what is disgusting, disgust naturally lends itself to the thought that others should not contaminate themselves via contact with what I deem disgusting. Thus, insofar as a moral norm becomes linked to disgust, disgust will provide an immediate route to externalization. (Though it is unlikely that this will be a good explanation for the externalization of all norms, since it is unlikely that disgust is linked in these ways to the entire range of moral norms.)

This hypothesis suggests two extensions of Stanford's proposal. First, although Stanford highlights the value of externalization for stabilizing cooperation norms against the *competing* interests of cooperators and free riders, externalization also has value for solving coordination problems where interests of different groups *coincide*. To see this, consider part of Kelly's (2011, pp. 123–25) co-opt hypothesis: that disgust was co-opted to implement tribal instincts that function to preserve boundaries between ethnic groups. Disgust could accomplish this by motivating one not to interact with members of other groups, in part because they violate the disgust-linked norms of one's group. If so, then disgust may be a mechanism for what Stanford calls "correlated interaction under plasticity" (sect. 5, para. 8). Howsoever the norms of two groups might diverge over time, disgust will tend to motivate or cause correlated interactions within groups but not between them. However, this suggestion actually applies to Stanford's proposal in a way quite unlike the one he discusses. For example, if two cultural groups have different moral norms (as is likely to be the case even for closely related cultural groups), intergroup interactions will not be as profitable for members of either group, because expectations regarding the interaction are more likely to diverge (McElreath et al. 2003). Thus, members of each group have a shared interest in avoiding interactions with members of the opposite group. In this case, externalization of moral norms – realized in part by mutual disgust – can result in a tendency to minimize interactions with different groups, and this tendency may benefit *both* groups.

Second, Stanford points out that externalization can explain why hypocrisy is considered a moral violation above and beyond the wrongness of acts committed (though also condemned) by the hypocrite. However, in light of the above hypothesis, his proposal may also be able to explain why people tend to over-comply with norms to the point of disassociating from actions that merely appear to violate the norm. As Stanford suggests, externalization likely led to a feedback loop in which cognitive and linguistic abilities became more useful specifically for purposes of moral advertisement. The increased importance of moral advertisement makes it all the more crucial to prevent misleading advertisements like innocently giving the appearance of evil. Thus, one would predict that if Stanford's externalization story is correct, then being motivated to comply with moral norms and avoid those who do not will also be coupled with a tendency to avoid the appearance of violating a moral norm. For instance, moral vegetarians tend to avoid eating meat that would otherwise be discarded (e.g., their roommate's leftovers), even when eating it would not spare any animals from harm. Disgust is one sort of mechanism that might explain this behavior. Because the contamination sensitivity of disgust operates via contact, similarity, and association, acts that merely resemble or are associated with disgust-linked moral violations (e.g., eating meat that has been "tainted" by the cruel practices of factory farms) will also tend to be avoided. Moreover, because disgust is linked to evaluations of bad character (Giner-Sorolla & Chapman 2017), it may also motivate avoidance of actions that are merely associated with bad character.

Although avoiding the appearance of evil may seem to require metarepresentation, disgust provides a mechanism for implementing such avoidance without having to think about one's actions from another perspective. For example, moral vegetarians clearly have an interest in avoiding the appearance of contributing to animal suffering (e.g., by eating the leftovers from a catered event). Nevertheless, if they are disgusted by meat consumption, they will be motivated to avoid eating the meat without considering what others might think of them. Thus, disgust provides a simple and economical way of implementing externalization of norms in some moral domains and hence may have been an early and influential driver of externalization.

A cognitive, non-selectionist account of moral externalism

doi:10.1017/S0140525X18000237, e118

Jason Zinser

Department of Philosophy, University of Wisconsin–Stevens Point, Stevens Point, WI 54481-3897.

jzinser@uwsp.edu

https://www.uwsp.edu/philosophy/Pages/AboutUs/FacultyStaff/jZinser.aspx

Abstract: A general feature of our moral psychology is that we feel that some moral demands are motivated externally. Stanford explains this feature with an evolutionary account, such that moral externalism was selected for its ability to facilitate prosocial interactions. Alternatively, I argue that a cognitive, non-selectionist account of moral externalism is a more parsimonious explanation.

Stanford is providing an evolutionary account for a peculiar feature of morality – namely, that we view the motivations for moral actions, both in ourselves and in others, as originating externally. Being externally motivated means that an action is right or wrong not because it exclusively aligns with one's preferences, but *because it is right*. It seems puzzling to explain moral externalism from an evolutionary perspective, since it appears that subjective states alone would be strong enough to motivate action (Stanford raises the example of pain responses, which are subjective and can be strongly motivational). Therefore, why would moral externalism arise if existent mechanisms, like subjective preferences, can provide the same function? The answer Stanford gives is that moral externalism was selected for as a guide to identify conspecifics with whom it would be good to collaborate. Group members whose actions align with justice, rights, and other externally sanctioned norms would likely be good partners for cooperative actions and also, and perhaps more importantly, less apt to exploit others.

I will argue that a more parsimonious explanation of moral externalism is that it is merely a psychological by-product of underlying affective responses; it is a story we tell ourselves to make sense of a particular class of pre-existing subjective states. This view is predicated on the dual-process model of the mind (for an overview, see Kahneman 2011). In rough outline, the dual-process model suggests that our mind consists of both an automatic, non-conscious processing system and an executive, rational control center, which we experience as our mental selves. The revolutionary nature of this view is that a surprising amount of our decision-making is conducted by our automatic, non-conscious system, which is, only after the initial decision is made, endorsed, ignored, or revised by our conscious selves. Relating the dual-process model to moral psychology, Haidt (2007) nicely summarizes: "[The] basic point is that brains are always and automatically evaluating everything they perceive, and that higher-level thinking is preceded, permeated, and influenced by affective reactions (simply feelings of like and dislike)

which push us gently (or not so gently) toward approach or avoidance" (p. 998).

This view aligns with a popular evolutionary approach to grounding our moral intuitions in our affective responses, in terms of emotions, disgust, or intuitions (Greene 2013; Haidt 2001; Joyce 2006; Nichols 2004; Prinz 2007). The problem with this approach, as Stanford emphasizes, is that if our affective responses motivate moral behavior, why do we then have this "extra" feeling that some of our moral intuitions are motivated externally? A dual-system response could be that our executive, conscious system has to make sense of what our automatic, nonconscious system has decided. We have an inclination, perhaps as a result of natural selection, to have certain affective responses, and then, after we experience the affect, our rational selves have to go about explaining why it is that we have this moral intuition. The conscious explaining of our innate intuitions is called confabulation, which has also been experimentally identified in other non-moral situations (Haidt 2001; Nisbett & Wilson 1977).

An important point from Stanford remains, which is: Why have this extra step for a certain class of affective response? An answer might be that the moral intuitions that invoke an externalist motivation differ in a salient way from those that do not. My favorable response to peanut butter ice-cream is justified by my subjective approval upon consuming it. My moral rebuke of the Ku Klux Klan, as Stanford suggests, is more than me disliking it, but that I judge that it is wrong apart from what I or anyone else might feel. The difference in the affective responses between the two cases might be that the latter response is other-centered whereas the prior response is wholly subjective. My conscious, executive self struggles to find an explanation for this strong affective response that is seemingly concerned with the well-being of others (or concerns with justice, fairness, etc.). The answer we tell ourselves, given that the motivation seems to explicitly rule out merely self-centered motivation, is that justification for such preferences must be external. What else, my conscious self contends, could ground these other-centered affective responses?

The significance of this approach suggests that Stanford has the causal relationship backwards. For Stanford, we derive motivation apart from our affective responses. On the dual-process account, the cause of externalized motivation is the underlying affective responses combined with a post-hoc explanation from our conscious, rational self. This approach has a two-fold advantage on Stanford's account. First, natural selection could simply work on a pre-existing system of affective responses identified even in chimpanzees (De Waal 1996). Second, it is more challenging, it would seem, for natural selection to select a higher-order cognitive feature like "external motivation" for a set of moral judgments. Third, the motivations for behaviors are often opaque, so selection may be blind to this distinction (in trying to identify others who adhere to externalized moral principles).

A final reason for seeing affect as driving moral externalism is that it seems strange for "motivation" to be external. Hume famously claimed that reason is and must be a "slave of the passions" (Hume 1738/1975, p. 415). It seems like Stanford is collapsing *justification* of moral intuitions with *motivation* for moral intuitions. Even if you acknowledge externally sanctioned moral principles, you still have to want to follow them. Understanding that moral externalization is a confabulation would avoid this problem.

So Stanford may be right in arguing that the distinctive features of human morality arose as a selective function to facilitate prosocial behavior, promote cooperation, and to avoid exploitation. This alone is an interesting and important contribution to our understanding of the evolutionary roots of morality. There is no need, however, to attribute this to moral externalism, but to the affective responses that underlie such behavior. Why give an evolutionary explanation of our sense of moral externalism when it can be explained away?

Author's Response

Moral externalization and normativity: The errors of our ways

doi:10.1017/S0140525X17002254, e119

P. Kyle Stanford

Department of Logic and Philosophy of Science, University of California, Irvine, Irvine, CA 92697.

stanford@uci.edu http://www.lps.uci.edu/lps_bios/stanford

Abstract: I respond to the many thoughtful suggestions and concerns of my commentators on a wide variety of questions. These include whether moral norms form a unified category, whether they have a distinctive phenomenology, and/or whether moral normativity is a cultural construct; whether moral externalization is necessary for correlated interaction or human prosociality; precisely how such externalization generates correlated interactions among prosocial agents; and whether there are any convincing alternative explanations for it.

R1. Introduction

Let me begin by thanking the commentators for the evident care, thoughtfulness, and generosity with which they have engaged my work. I have learned a great deal from reflecting on their concerns and suggestions regarding the account I offer in the target article, and I am sincerely grateful to them. I also very much appreciate the opportunities they offer me to expand on, revise, and clarify the case made in the target article. I have tried to organize my efforts to do so as answers to a range of broad and foundational questions, each raised or addressed in some way by multiple commentators.

In Section 2, I agree with the suggestion of several commentators that the externalized phenomenology characteristic of many prototypically moral norms does not pick out all and only those norms identified as moral in some other way (e.g. by their content) or constitute a unique and distinctive "moral domain" of our experience. What nonetheless requires explanation, I suggest, is the complex *pattern* in which humans externalize various judgments of various kinds to various degrees in various circumstances. In Section 3, I reject the idea that externalization is *necessary* for achieving correlated interaction sufficient to generate and/or sustain human hypercooperation, but I suggest that we nonetheless have compelling empirical evidence that externalization is in fact how humans came to achieve the extraordinary degrees and precision of correlated interaction that we do. In Section 4, I consider the suggestion that externalization is itself a cultural construct, and I explain how the externalization of norms creates stable networks of preferential interaction even within well-defined in-groups. In Section 5, I consider whether the sort of externalization I've described is in fact also characteristic of conventional and/or other sorts of non-moral (e.g. aesthetic) norms, and I consider whether externalizing norms and obligations might have been an ancestral condition from which no shift to an externalized moral psychology would ever have been required. In Section 6, I address a variety of alternative explanations proposed by

various commentators for the emergence or generation of moral externalization, and I suggest that where plausible, these alternatives amplify or complement the explanation offered in the target article, rather than competing with it. In Section 7, I consider and evaluate a range of welcome suggestions by various commentators that help refine, supplement, and/or elaborate the account given in the target article itself. And in Section 8, I briefly consider whether or not the evolutionary explanation I've proposed for the externalization of moral norms and obligations should be regarded as a so-called "error" theory of morality or as "debunking" our ordinary practices of moral judgment and evaluation.

R2. Are moral norms and obligations truly externalized? Do moral norms form a unified category?

Perhaps most fundamentally, a substantial number of the commentators argue that the sort of externalized or objectified phenomenology that I repeatedly characterize as "distinctively moral" does not actually exist and therefore needs no explanation of the sort I have proposed. **Stich** frames his argument for this claim around several criticisms of the work of Turiel and others, rightly pointing out that the particular characteristics they use to try to classify norms and violations as moral or conventional have been scrutinized and challenged by other scholars. **Patel & Machery** are similarly "skeptical" that "this distinction is central to human psychology." And **Davis & Kelly** argue that neither "hardness" nor "objectivity" (two properties they suggest I have conflated in characterizing externalization; see sect. R5 below), nor both together, pick out a distinctively moral domain and indeed that "[n]o subcategory of norms makes up a psychologically distinctive or cooperatively indispensible set of *moral* ones." Likewise, **Theriault & Young** are right to be concerned about the "stimuli-as-fixed-effects" fallacy and to suggest that we are far from having convincing evidence that "morals comprise a distinct category – at least on the basis of their objectivity, universality, and authority independence, as has traditionally been argued."

But I am afraid that these authors misunderstand the use I intend to make of the work of Turiel and others on the moral/conventional distinction. Rather than identifying necessary and sufficient features shared by all and only moral judgments or even (much more plausibly) a "nomological cluster" of such features, my goal is instead to use this line of research to help us pick out a phenomenological difference between our experience of many prototypically moral norms (like those prohibiting lying or murder) and many prototypically conventional norms (like that specifying which spoon to use for soup). That phenomenological difference itself is what concerns us most directly, though of course I have suggested that a wide range of further empirical findings are rather neatly elucidated and unified by the evolutionary explanation I offer for it. And we already know that the features to which Turiel and others appeal offer *at best* a rough and ready guide to that phenomenological difference, for there are domains (like moralized disgust) in which the phenomenology of norms seems to exhibit some but not all of those features. In fact, I think the clearest indicator of objectification or externalization (moral or otherwise) is one that Goodwin and Darley (2008) use as an experimental probe for this phenomenological difference: the judgment that when agents disagree at least one party to that disagreement must be mistaken. But even here Goodwin and Darley's own results suggest that there is a *continuum* of such externalization or objectification along which various sorts of judgments tend to cluster (pace **Theriault & Young**) with considerable individual and contextual variation between them, rather than a set of qualitatively discrete categories with hard edges.

What all of this suggests is not that there is no phenomenon to explain, but instead that the phenomenon demanding explanation is considerably more complex than Turiel and others initially supposed: What requires explanation is the complex *pattern* in which humans tend to externalize or objectify various sorts of judgments to varying degrees under various conditions. The explanation I have offered does much more to illuminate some features of that pattern (e.g., why many prototypically moral judgments are strongly externalized) than others (e.g., why moralized disgust seems to exhibit some but not all of the features that Turiel and others suggested were characteristic of moral norms in general; though see my discussion of **Wiegman** in sect. R6). *But it is the complex pattern itself we are seeking to understand.* Recognizing this complex pattern as our explanatory target (especially in conjunction with the clear context-sensitivity of externalization, elaborated below) also leaves me largely untroubled by **Davis & Kelly**'s further suggestion that only a minority of even prototypically moral judgments are in fact objectified or externalized.

This same recognition also reveals, however, that these authors are all quite right to criticize my repeated references to a special conceptual *category* of externalized norms (to which particular norms can simply be added or removed) or to a distinctive phenomenology supposedly distinguishing an independently specifiable "moral domain" from all others. Although this seemed a harmless simplification of both the relevant explanatory demand and my own proposed solution to it, I now see the error of my ways: This shorthand characterization indeed suggests that the attractions of the account I offer themselves *depend* on the existence of such a distinctive moral category. But this suggestion is mistaken: My account of the role played by externalization in protecting prosocial and cooperative tendencies from exploitation in humans is threatened neither by the recognition that such externalization comes in degrees that vary between individuals, contexts, and types of judgment, nor by the fact that the phenomenology of externalization does not itself pick out all and only norms that are judged to be moral on some independent ground (e.g., their content). What matters to the explanation is that we selectively externalize some norms (to varying degrees) and that shifting sets of prosocial or cooperative norms can be effectively and efficiently protected from exploitation by externalizing them in this way, even if the distinctive phenomenology of externalization does not characterize all and only norms independently identified (in some other way) as moral or pick out a categorically distinct moral domain.

Stich also rightly emphasizes that Sarkissian et al. (2011) is only the first in a "recent cascade of papers, all of which cast doubt on the conclusion that people are consistently

objectivist about moral judgments." What much of this cascade reveals is that the extent to which we externalize or objectify judgments (moral or otherwise) is considerably influenced by a wide variety of different cues that include not only the content of the judgment, but also the degree of perceived consensus regarding that judgment, its valence (i.e., prescribing good conduct or proscribing bad conduct; though cf. **Goodwin**), and much else besides. But I resist any suggestion that I have *merely* accommodated Sarkissian et al.'s results or tried to explain them away. These authors suggest that their findings show the folk to become increasingly relativist as they are forced to confront or consider moral perspectives increasingly remote from their own. I have proposed a competing hypothesis: that externalization increases as contextual cues make the need for actual social interaction with a morally deviant agent a more salient, concrete, or realistic possibility. I have also suggested that there is a natural experiment to perform to evaluate these competing hypotheses: replicating Sarkissian et al.'s (2011) study with a further condition emphasizing the realistic possibility of ongoing social interaction with those who disagree with the subject's own moral judgment (e.g., in which the morally deviant Amazonian tribesmen or extraterrestrials described in the original experiment will soon arrive in our town and we will need to decide how to interact with them). This experiment is now underway.

Moreover, other findings in this "recent cascade" actually increase my confidence in the alternative hypothesis I have proposed, in particular the experimental work of Beebe (2014) to which **Stich** directs our attention. Most importantly, Beebe shows that when the deviant agent with whom the subject disagrees in a moral judgment is described more *concretely* (by providing a name and brief description of their academic courses and major, or even just a name and photograph), we find a corresponding increase in the extent to which we externalize or objectify the judgment or norm with which that deviant agent disagrees. This difference is perfectly intelligible if subjects are more concerned to objectify or externalize norms when faced with a more concrete, realistic, or salient possibility of actual social interaction with a deviant agent, but I do not see how they can be easily reconciled with Sarkissian et al.'s (2011) competing hypothesis that it is instead exposure to alternative moral frameworks that moderates our objectivist or externalizing tendencies.

I find it difficult, however, to reconcile **Stich**'s enthusiasm for this recent experimental work highlighting the variability and context-sensitivity of our externalizing tendencies with his (apparent) further claim that there simply *is no* such externalized phenomenology in the first place. He says,

> Sometimes Stanford relies on the language of phenomenology: "[we] experience the demands of morality as somehow *imposed* upon us externally" and "we regard such demands as imposing unconditional obligations not only on ourselves, but also on any and all agents whatsoever" (sect. 1, para. 2, emphasis in target article). Well, perhaps Stanford experiences the demands of morality in this way. But I don't recognize this as part of *my* moral phenomenology. Which one of us is an outlier? (Stich commentary, para. 5, second emphasis his)

Here Stich seems to report that he himself doesn't experience even the most prototypical moral judgments as having the character I describe in these phenomenological terms. It might seem tempting to suppose that Stich instead means only to deny that such externalization is a consistent part of his *moral* phenomenology (i.e., is not reliably exhibited by all and only judgments he regards [on some independent ground] as moral in character). But this cannot be right: For one thing, Stich emphatically denies that there *is* any independently unified category of moral judgments whose degrees of externalization we might then go on to consider. It might therefore seem natural to respond simply by noting that the target article itself already recognizes the existence of reliable variation among human beings in whether they externalize or objectify even prototypically moralized judgments, and by suggesting that perhaps we should not be shocked to discover further intersubjective variation in this respect (I should emphasize that I do *not* mean to be suggesting that Professor Stich is a psychopath or sociopath). Perhaps Stich does lack the sort of externalized moral phenomenology I have described, but my remaining commentators seem to recognize such selective externalization of some but not all norms as part of their own moral phenomenology, even if they also think I have gone on to mischaracterize or misunderstand that phenomenology in a truly daunting number of further ways.

R3. Is externalization *necessary* for correlated interaction and/or human prosociality?

Patel & Machery suggest that even when externalization does occur it is probably not what protects prosociality from exploitation, because correlated interaction can be achieved by other means, such as merely conventional norms (see also **Birch**; **Brusse & Sterelny**; **Handfield, Thrasher, & García** [**Handfield et al.**]; **Jebari & Huebner**; **O'Neill**). Patel & Machery go on to note that what they suggest is "the best explanation of cooperative behavior" we have (Balliet et al. 2011) makes no appeal to externalization, using instead punishment and reputation-tracking as mechanisms for achieving correlated interaction. It is certainly true that externalization is not *necessary* to produce correlated interaction, but this is also not all that externalization does. For one thing, externalization selects out only a subset of norms and judgments (rather than all or none of those we accept) as those by which the comparative desirability of social partners will be judged. In addition, externalization ensures that *one and the same set* of privileged norms is used simultaneously to motivate ourselves and to evaluate candidate social partners even as the membership of this set of privileged norms changes over time. It might even be *possible* to achieve correlated interaction sufficiently precise and robust to protect humans' spontaneous prosociality from exploitation using only subjective preferences and conventional norms, but a wide range of evidence, including perhaps most importantly the systematic sensitivity of preferred social distance to specifically moral (but not other kinds of) disagreement (Skitka et al. 2005), strongly suggests that externalization is *in fact* the means by which such sufficiently precisely correlated interaction was achieved among early humans.

Birch and **O'Neill** similarly emphasize that objectification of the sort I have described is not *necessary* to achieve correlated interaction between cooperators. Birch, for example, suggests that once we distinguish the

apparent source of a norm from its scope, it becomes clear that the latter rather than the former is what really matters for protecting cooperative dispositions from exploitation. He rightly points out that social norms need not be externalized to be of wide scope (e.g., a subjective commitment to removing litter that "applies to the behaviour of the whole community") and need not be of wide scope to be externalized (e.g., externalized or objectified norms that apply to and concern only the conduct of the Pope). Although this latter example is an extreme case of the sort of "role-dependent or asymmetric norms" that I have suggested arose later in our phylogenetic history, it nevertheless illustrates that even norms of narrow scope can be externalized. The first illustration seems more problematic: Externalization leads us to apply a norm with unrestricted scope, to absolutely any potential social partner, so applying "to the behaviour of the whole community" does not exhibit the sort of unrestricted scope that externalization or objectification confers. It seems instead a textbook example of a conventional norm, applied only to the members of the relevant cultural group or community.

But let us concede the possibility of both externalized norms of narrow scope and non-externalized norms applied with at least very wide scope (e.g., to all members of a particularly expansive community). Does this show that it is wide scope *rather than* externalization that does the work of ensuring correlated interaction? Surely not, for my proposal is that the selective externalization of particular norms is precisely *how* such norms came to be applied by early humans with unlimited scope in the first place. Wide scope and externalization are not *competing* explanations for the stability of exploitable norms; instead it is *by means of externalization* that humans actually came to apply particular sets of norms with unrestricted scope. **Birch** is right to think that any alternative mechanism leading us to apply a subset of our norms with arbitrarily wide scope (and to exclude violators from interaction with us) could generate correlated interaction equally effectively, but the suggestion here is that externalizing norms was in fact the way that hominins came to apply them with unrestricted scope rather than only to the members of a particular (even very large) social group.

Similarly, **O'Neill** argues that generalization and desire/preference-independence are distinct features that I have conflated in characterizing the externalization or objectification of moral norms. She is quite right to emphasize that these features are different and play distinct roles in protecting prosociality from exploitation. She goes on to suggest, however, that generalization alone generates correlated interaction while desire/preference-independence serves merely to "stabilize" the resulting prosocial interaction by protecting it from particular kinds of exploitation and/or failure. Here I am less convinced, but I am also unsure of how much is at stake in any residual disagreement, as we seem to agree that both of these characteristic features of moral normativity were important for the evolution of human prosociality.

Moreover, I suspect that **O'Neill**'s analysis here is flawed in an illuminating way. When she argues that generalization alone suffices to produce correlated interaction, she suggests that: "Presumably, the tendency to generalize obligations produces or comes with a preference for cooperation partners who believe themselves to be subject to the same obligations that one believes oneself to be subject to." (**Birch** does not make this claim explicitly but may presuppose it; see also **Wiegman**.) But this is again simply to hide a crucial part of the problem for which externalization itself constitutes such an elegant solution – namely, the challenge of ensuring that at any given time I use one and the same set of some, but not all, of the norms I accept *both* to motivate my own behavior *and* to evaluate the desirability of candidate social partners, even as the members of that set remain open to modification, extension, and replacement on a cultural or historical (rather than biological) timescale. Nothing about evaluating candidate social partners by means of their adherence to a particular set of norms requires these to be the very same norms by which I myself am motivated – it is externalizing the source of moral normativity that establishes and maintains this crucial connection. My suggestion is that externalization is how humans *actually* managed to combine normative plasticity with the tendency to experience (at any given time) *one and the same set* of privileged norms *both* as motivating our own behavior in a distinctive, preference-independent way *and* as the standards by which we evaluate the desirability of candidate social partners.

I hope it is now clear why I decline to defend the "indispensability thesis" that **Handfield et al.** attribute to me, which they describe as the view "that externalized moral norms of this sort are necessary to achieve pro-social cooperation, at least at the high rate seen in humans." Once again, moral externalization is certainly *not* necessary for human hypercooperation, because any source of (sufficiently) correlated interaction will do. I fully agree with Handfield et al. that the mechanism of correlated interaction I describe cannot be usefully assimilated to costly signaling, a green beard hypothesis, or social selection – indeed, I think it is not usefully assimilated to *any* mechanism for generating correlated interaction that we have found elsewhere in nature. Moreover, although correlated interaction can indeed be generated by **Patel & Machery**–style punishment and reputation-tracking (though costly punishment is *itself* a second-order form of altruism whose stability would require some further explanation), **Birch**-style application of norms with wide scope, **O'Neill**-style generalization, and in many other ways besides, my claim is that we have convincing evidence that externalization is the mechanism *actually* responsible for the distinctive *forms* of generalization, reputation-tracking, preference-independence, and so forth, which constitute salient features of our own moral psychology.

R4. Is externalization a cultural construct? How does externalization generate social networks of precisely correlated prosocial interaction?

Patel & Machery offer a further reason for worrying about the centrality or importance of the moral/conventional distinction for human psychology: They claim that the distinction itself is culturally parochial, found only in some human cultural traditions and not others, and they cite unpublished data suggesting that Indian and Muslim subjects with a variety of national origins simply do not recognize any distinction between moral and non-moral norms. I cannot evaluate the unpublished data that Patel & Machery mention, but I remain skeptical. The case seems

at least superficially similar to many others in which we have concluded that concepts or distinctions found in a wide range of human cultures are simply absent from some particular culture or population, only to find that in fact we failed to probe for those concepts or distinctions in a way that was sufficiently sensitive to the subtleties of the particular culture in question. This sort of danger is highlighted by **Poulin**'s suggestion (following Gray) that those who condemn apparently harmless moral violations often do so because they see those violations as in fact having victims who are harmed by them.

Nonetheless, suppose that **Patel & Machery** are right about the cultural parochiality of the moral/conventional distinction or even externalization itself. What follows? Simply that the selective advantages of externalization will have to be recast using only the machinery of cultural evolution, rather than the complex combination of cultural and biological co-evolution that I have proposed. In that case, we would still appeal to the role of externalization in generating correlated interaction and (thereby) protecting prosociality from exploitation and exclusion. But now this would be part of a larger explanation of how a particular *cultural* innovation (viz., the selective externalization of norms) came to play an important role in scaffolding human prosociality in whatever cultures do externalize some, but not all, norms.

This would also be the appropriate response if we were convinced by the very different reasons offered by **Brusse & Sterelny** for thinking that the externalization of some but not all of the norms we embrace "is a late-breaking cultural innovation." But I do not think we should be convinced. The problems raised by Brusse & Sterelny for the idea that externalization evolved in humans to establish and maintain correlated interaction between prosocial or cooperative agents all depend on a subtle but important misunderstanding of how externalization itself works. Brusse & Sterelny are quite right to think (see also **Voorhees, Read, & Gabora** [**Voorhees et al.**]) that excluding or shunning any potential partner with whom we have identified any point of substantive moral disagreement would be far too costly to be adaptive in realistic ancestral environments. But we did not and do not simply shun those with whom we have moral disagreements, whether concerning specific moralized norms or (far more frequently) simply our moralized evaluations of particular cases. Instead, identified points of moral disagreement simply count against the desirability of a given potential social partner *relative to others* and increase the social distance we prefer to maintain from that potential partner (see Skitka et al. 2005). The extent or degree to which any particular disagreement influences our evaluation of a potential social partner is mediated by factors like the extent to which the norm in question is externalized (Goodwin & Darley 2012), how strongly the norm itself is held (Skitka et al. 2005), and presumably much else besides. But what such disagreement generates is not a list of tribemates to shun because I have identified some point of moral disagreement with them, but instead simply a preference ordering (or perhaps a set of preference orderings relativized to different activities) over such potential partners that is a complex function of our identified moral agreements and disagreements with them (as well as non-moral considerations like their competence or physical prowess).

Mistakes and accidents are also, therefore, not as consequential as **Brusse & Sterelny** imagine, though it is nonetheless remarkable that human cultures so reliably include elaborate procedures of apology and repair by which members seek to advertise their commitment to precisely those moral judgments and convictions that recent mistaken or accidental conduct might lead others to suspect they do not share (see also **Allidina & Cunningham** on the moralization of everyday behavior and **Wiegman** on over-compliance with norms). Externalizing a norm leads me to devalue, rather than shun, interaction with those who violate or fail to externalize it, ensuring in turn that identifying points of moral disagreement with my tribemates neither precludes me from cooperative or other forms of prosocial interaction with them, nor leaves me (as Brusse & Sterelny suppose) with very few or no candidate partners with whom I am both willing and able to cooperate. Of course, there may well be some threshold amount or variety of moral disagreement beyond which I will indeed shun a fellow group member, but this is not a consequence of identified moral disagreement in general. This is also why the demands of adaptive plasticity and conformism do not conflict: Two agents can agree on how to extend or adapt an externalized norm in new ways or into new circumstances (or even to externalize an entirely new norm), thus protecting their further prosocial interaction from exploitation, without being shunned for doing so by other group members.

The challenge posed by **Johnson** suffers from a related misunderstanding of the role played by moralized agreement and disagreement in mediating our prosocial dispositions. He argues that my proposal implies that I would never act prosocially or cooperatively towards another agent unless and until I had abundant and detailed positive evidence of many specific points or respects of moral agreement (or, even more challenging, extensive evidence that the agent and I "share the same heritable attribute triggering cooperation"). He therefore suggests that anonymous prosociality and prosociality in environments where we have little or no information about others represent "a form of prosocial behavior than cannot be explained by moral externalization." But nothing about externalization requires positive evidence of moral agreement (much less of heritable properties in common) to be a *condition* of prosociality. Humans are much more *spontaneously* prosocial than other primates, and what is needed to protect that spontaneous prosociality from exploitation is simply that an agent's enthusiasm for social interaction with a given partner should *decline* when evidence of moral *dis*agreement with that partner arises. Nearly any particular degree or respect of spontaneous prosociality with ingroup members (or even strangers) could serve as a (culturally and contextually variable) default starting point for humans, so long as acquiring evidence of moral disagreement with any particular agent reduces the degree (and/or forms) of exploitable prosociality we spontaneously extend to her. What matters is that we hold others *responsible* for living up to our own externalized norms and increase our preferred social distance from them if they fail to do so, and this does not require that we must have any evidence (much less abundant evidence) that they will in fact do so before we are willing to risk interacting cooperatively or prosocially with them. Indeed, it was precisely to protect and facilitate dramatic increases in the

range and degree of our *spontaneous* prosociality that our moral psychology evolved mechanisms ensuring increasingly precise, specific, and powerful correlated interaction in the first place.

For related reasons, I must decline **Ross**'s intriguing invitation to see the account offered in the target article as advancing a comprehensive explanation for the dynamics of tribe formation, according to which:

> A subset of a founder population in a niche moralizes some of its new conventions in order to achieve and maintain correlated equilibrium and successfully exclude those most disposed to free-riding. Then, presumably – Stanford is not explicit on this point – the excluded villains interact with one another for lack of an alternative, and form and then moralize different conventions. (para. 4 in Ross's commentary)

I have several reservations about seeing this as a general model of the dynamics of tribe formation. Perhaps most importantly, it obscures the fact that externalization is so powerful precisely because it can generate correlated interaction between agents prepared to moralize a given norm (or a particular extension or application of a norm) even while they remain accepted members of a larger social group in which that particular extension, application, or norm is not generally moralized (cf. **Brusse & Sterelny**, and see **Böhm, Thielmann, & Hilbig** [**Böhm et al.**] on moral homogeneity within groups). That is, moral externalization ensures correlated interaction between agents prepared to engage in (and demand) a particular form of prosocial interaction *without* the need to form a new (physically or spatially distinct) tribal group in which this behavior is homogeneously moralized. It does seem perfectly plausible to suggest that such moralized differences might emerge, accumulate, and grow into points of persistent substantive moral conflict or disagreement with others, ultimately leading to the formation of a new tribe in something like the process Ross suggests. But even here I suspect this would most often be a matter of members of two groups within a tribe finding themselves with a sufficient number of sufficiently important moral disagreements that members of *each* group are motivated to exclude members of the other from social interaction with them, with neither group consisting of "excluded villains [who] interact with one another for lack of an alternative" (see Baumard et al. [2013] on self-segregation in contemporary hunter-gatherer societies, and Böhm et al. on out-group hatred). This last difference matters, because it suggests that members of the two groups begin to moralize different normative demands long before either group's members are simply shunned or excluded from social interaction generally by members of the other, rather than one tribe being composed of "villains" or "cast-offs" who simply moralize fewer behaviors (and therefore might ultimately threaten to collapse into what Ross, following Edgerton [1992], calls "sick societies"). This proposal would, of course, still help explain why "human tribes manifestly bifurcate *within* shared physical environments" (Ross, para. 4), although it also seems likely that moralized disagreement and/or moralized group identity is just one of many different factors at work in the dynamics of tribe formation.

R5. Do we also externalize conventional norms, aesthetic judgments, and/or other kinds of value judgments? Might externalization have been an ancestral condition?

I hope it is already clear why the very thin sort of objectification that **Patel & Machery** rightly point out is attributable to conventional as well as moral norms was never our explanatory concern. Likewise, **Van Prooijen** is right to suggest that prototypically moral norms are "externalized" in the sense that they are imposed on us by others, but *this* sort of externalization is equally attributable to prototypically conventional norms and is therefore again simply beside the point. And the same is true for the sense of objectification ("hardness") that **Davis & Kelly** argue characterizes any and all normative judgments whatsoever. From the beginning, our explanatory target has been the fact that many prototypically moral norms exhibit a further and much stronger form of phenomenological externalization or objectification, one revealed most clearly in the unwillingness of subjects to tolerate disagreement without error concerning those norms and/or their implications (the "objectivism" Davis & Kelly suggest I have conflated with "hardness"), and it is this further, more robust form of externalization or objectification for which I have sought to provide a convincing evolutionary explanation.

Similarly, perhaps **Isern-Mas & Gomila** are right to think that it has been undisputed at least since Kant that we ascribe *some* distinctive form of objectivity ("aim[ing] at universal validity") not only to moral judgments, but also aesthetic judgments and indeed any value judgments whatsoever. However, this cannot be the same form of objectivity for which Goodwin and Darley (2008) probe by asking subjects about the possibility of disagreement without error. Subjects ascribe *that* form of objectivity in the highest degree to judgments of empirical fact, more modestly to judgments of moral norm violation, followed by judgments of conventional norm violation, and least of all by judgments of taste and preference, including putatively aesthetic judgments like "Frank Sinatra is a better singer than Michael Bolton" (Goodwin and Darley's example). It is unsurprising, then, that the process of objectification Isern-Mas & Gomila (following Darwall 2006) go on to describe is one that applies to conventional norms in just the same way that it does to moral norms and therefore cannot explain the emergence of the sort of objectivity that subjects ascribe to prototypically moral (but not prototypically conventional) norms. Fortunately, I do not think the problem they seek to solve by invoking this process (supplying a supposedly missing connection between our externalized norms/values and our motivations) actually exists: Moral judgments can be intrinsically motivating, I suggest, just like desires and preferences, despite the fact that externalizing the former but not the latter certainly generates *other* salient phenomenological differences between our experiences of intrinsic motivation in the two cases. Indeed, this and other characteristics shared between intrinsically motivating moral judgments and intrinsically motivating preferences or desires leaves me unsurprised by **Theriault & Young**'s evidence that the

neural and behavioral signatures of moral judgments are more like those of preferences than those of objective facts (see also **Schulz**).

Schulz has a distinct worry about the motivational adequacy of externalized moral norms: Although externalizing such a norm can certainly motivate us to follow that norm itself, he suggests, he does not see how or why it would motivate us to disfavor social interaction with those who do not comply with (or do not externalize) the norm. Here I think he is misled by the idea that norms cannot motivate us to do anything that is not explicitly specified as part of the content of that norm itself (hence the need for a further, conjunctive component of that content). But this does not seem to be how our motivational psychology actually works. Notice, for example, that the content of a *conventional* norm does not include any explicit description of the consequences of noncompliance – it simply articulates the norm itself, and it is simply a matter of empirical fact about us that we respond to any particular violation of such norms with criticism, exclusion, forgiveness, shock, glee, disappointment, or in any other particular way. On the account I offer, the same is true of moral norms and obligations: The norm itself does not specify the consequences of noncompliance; instead it is simply an empirical fact about human beings that they devalue social interaction with those who do not comply with (or externalize) the norms that they themselves externalize. Indeed, it is because the consequences of noncompliance are *not* specified as a conjunctive part of the content of the norm itself (as, surprisingly, Schulz himself seems to recognize in his second paragraph) that we must *discover* that increasing preferred social distance is *in fact* the consequence of failing to comply with (and/or failing to externalize) a norm that we ourselves externalize (Skitka et al. 2005). Moreover, Schulz's description of how his conjunctivist/subjectivist proposal explains the *apparent* externalization of moral demands is unconvincing, because the fact that moral demands systematically motivate not only compliance, but also the consequences of others' noncompliance (whatever they may be) also applies straightforwardly to the case of merely conventional norms – again, whatever kind of apparent externalization Schulz can explain in this way is simply not what we set out to understand.

We can now also usefully approach **Kaznatcheev & Shultz**'s suggestion (following Joyce 2006) that the objectification or externalization of norms was a likely ancestral condition from which no shift to an externalized moral phenomenology would have been required. Sometimes it seems that these authors slip into answering the wrong question, as when they seek to establish that "an understanding of experience as subjective both in oneself and others develops from an objectivized phenomenological precursor." The issue is not, of course, whether the *general* capacity to represent experiences as reflecting objective states of the world itself did or did not precede the capacity to represent experiences as reflections of my own subjective states, but rather whether the capacity to experience *norms* as reflecting objective states of the world itself preceded the capacity to represent them as merely reflecting our own and others' subjective preferences and/or social conventions. The more general capacity to discriminate representations of states of the world from my own subjective states is far more fundamental, and indeed it seems we could hardly hope to formulate or even understand norms regarding our own or others' conduct without having this more fundamental capacity already in place. Moreover, the considerations to which Kaznatcheev and Shultz appeal that *do* bear on the externalization of norms in particular seem intended simply to blunt the suggestion that an objectified or externalized normative phenomenology *must have* emerged from an ancestral subjectivist normative phenomenology rather than vice versa. I agree with this modal judgment, but given the *exploitability* of human cooperative and other prosocial dispositions, the fact that insensitivity to intersubjective variation in prosocial norms *would indeed* invite and generate such exploitation, and the fact that we strongly externalize only *some* of the norms we embrace (including disproportionately many that *do in fact* serve to protect prosociality from exploitation), I remain skeptical that an externalized or objectified *normative* phenomenology represents the ancestral condition of human beings.

R6. Are there lacunae in our understanding of how and why externalization occurs and/or any convincing alternative explanations for it?

Poulin suggests that the wide array of apparently heterogeneous externalized moral judgments we find in different human cultures may all simply be consequences of externalizing a more generalized "moral imperative to help or at least not harm others," often in culture-specific ways that may appear harmless to those outside the culture in question (cf. **Patel & Machery**). This is certainly a possibility, though I am more impressed than Poulin with evidence that humans regularly externalize and/or moralize norms that *they themselves* see as unrelated to harm. Poulin's suggestion is, of course, an extreme version of one of the potential explanations proposed in the target article itself for the patterns of similarity in externalized norms we find across human cultures: strongly biased learning in the externalization of particular norms or types of norms (whether categorized by their content or in some other way). But I continue to think that the other (non-exclusive) explanations suggested in the target article (convergent cultural evolution, common descent with modification, and enhancing in-group identification) and perhaps others besides are also quite likely to be part of the story of how different human cultures came to externalize the remarkably diverse and heterogeneous, but nonetheless systematically related, collections of different moral norms that they do. But if (or to the extent that) Poulin is right, the adaptive advantage of the remarkable plasticity with which norms are selectively externalized by members of different cultures at different times is simply that it allows us to quickly update, extend, and modify our (culturally specific) judgments about whether and/or how others can be helped or harmed.

For similar reasons, I doubt that the account proposed by **Voorhees et al.** represents anything like a complete answer to their question, "how could a group come to externalize all of the same norms?" Again, I expect the other mechanisms noted previously to play an important part in this story, though I am happy to regard the intricate machinery these authors propose of "cultural idea systems," "functional vehicles for the social expression of

emotional responses," and "complexes of beliefs and/or organizational rules that operate in a top-down manner so that individuals gain functionality only by adherence to these rules and/or constraints" as an attempt to provide a more detailed and systematic description of many of the processes of vertical and horizontal cultural transmission (and reinforcement) through which the members of any particular cultural group come to share many of the same particular externalized norms and judgments. In that case, however, these authors provide a sympathetic extension or refinement of the account I have offered rather than (as they seem to think) a competitor to it.

Sometimes, however, **Voorhees et al.** seem to suggest that it is the existence (or emergence in ontogeny) of externalization *itself* that they seek to explain by invoking this machinery of cultural idea systems and the like, and here I am unconvinced, in large part for reasons like the ubiquity (if not universality, cf. **Patel & Machery**) of the moral/conventional distinction across human cultures and the fact that human children in very different cultures seem to reliably and spontaneously start externalizing norms at the same point in ontogeny. Voorhees et al. are also wrong to think that my own proposal faces daunting challenges concerning the adaptive value of externalization for a single initiating agent and/or the second-order free-rider problem for punishment. With respect to the first, if I am motivated by a given exploitable prosocial norm, it is beneficial for me to externalize that norm (treating adherence to it as a relevant consideration in my own selection of candidate social partners) to avoid exploitation whether or not others embrace and/or externalize either that same norm or any norms at all. And the target article's account of "punishment" faces no serious second-order free-rider problem because the only cost incurred by a punishing agent is the loss of further opportunities for social interaction with a potential partner she already sees as prone to exploitation.

Most importantly, however, **Voorhees et al.** simply do not offer a plausible account of how the distinctive phenomenology of externalization itself arises or emerges in the course of ontogeny. They claim, for example, that "[i]f our ancestors' moral norms are part of the cultural idea system acting in a top-down manner within social systems organized through kinship relations, then kinship itself provides the objectivity and coherence of norm exteriorization." Here they seem to suggest that the distinctively externalized phenomenology of many prototypically moral judgments derives from the fact that they are embedded in kinship structures, but this is very implausible. For one thing, the fact that a norm is acquired from a member of one's kin gives us no reason to expect that the norm *itself* will be experienced as objective, even if the set of kinship relations structuring this acquisition is nonetheless itself viewed or experienced as an objective fact about the world. But in any case, moral normativity and/or externalization do not seem sensitive to whether or not norms are acquired from (or concerned with) kin. If they were, why would we not also externalize the merely *conventional* norms that we acquire from the same kin at the same time? A similar problem applies, of course, to Voorhees et al.'s more general suggestion that "cultural ideas, acquired through enculturation, are internalized by culture bearers and seen by them as objective reality." If this were indeed the source of the distinctive phenomenology of externalization, then again it seems we should expect to find conventional norms externalized in just the same way.

Jebari & Huebner propose yet another supposedly alternative explanation of moral externalization that I think does not conflict or compete with that offered in the target article. These authors suggest that "a plausible understanding of the evolution of objective morality must look beyond human psychology, to the objective features of the world that govern cooperative human ways of life (henceforth 'lifeways')" – that is, to the sorts of objective facts about the world and ourselves that render humans *obligate* cooperators, such as our affiliative tendencies, conformism, greater social tolerance and docility, and the loss of adaptations like sharp teeth for hunting and defense and the ability to extract nutrients from uncooked food. Jebari & Huebner go on to say:

> In acknowledging the critical changes that have emerged over the course of human evolution, it becomes clearer that our reasons for treating moral obligations as external have little to do with feelings of objectivity. Cooperation, coordination, and trust are objective features of our social lifeway." (Jebari & Huebner, para. 3)

Indeed, they suggest that "the felt objectivity of … obligations emerge as a consequence of our relationship to the social order, and our moral motivations are determined by our (often tacit) recognition of this relationship."

There is very little in this proposal with which I am inclined to disagree, besides the explicit claim that it obviates the need for any appeal to externalization itself. A full explanation of the evolution of human prosociality will indeed appeal to objective facts about the world including the sorts of evolutionary changes that **Jebari & Huebner** rightly suggest have helped render humans obligate cooperators. But these sorts of facts about humans are simply not in competition with "feelings of objectivity" as a *proximate mechanism* for motivating and protecting human prosociality. Far from "obscur[ing] the significance of … the emergence of complex and adaptive social networks … structured by rich patterns of social interaction" (as Jebari & Huebner allege), I appeal to externalization itself to explain how those very social networks become established, maintained, and modified over time, and to explain the emergence and persistence of many of the very adaptations facilitating human prosociality that Jebari & Huebner describe. The only alternative *mechanism* to which Jebari & Huebner appeal is "the role of affiliative tendencies in producing automatically coupled values and preferences" – they suggest that "such forces are sufficient to drive complex, open-ended, and cooperative forms of behavior" and therefore that objectification is unnecessary (cf. sect. R3) to explain our preference for partners who resist "contra-normative" behavior. This appeal, of course, faces precisely the same problem as **Van Prooijen**'s much simpler claim that moral norms are experienced as objective because they are imposed on us by the expectations and demands of others: It offers no explanation whatsoever for the fact that we externalize prototypically moral norms but not prototypically conventional norms. Conventional norms and norm-violations (not to mention subjective preferences) *also* mediate our affiliative tendencies and are *also* part of the social reality of obligate cooperation that Jebari & Huebner describe, so their proposal offers no explanation of the fact that we externalize some, but not all, of the norms and obligations we embrace at any given time.

Wiegman seeks to extend the account offered in the target article with the intriguing suggestion that "disgust provides a simple and economical way of implementing externalization of norms in some moral domains and hence may have been an early and influential driver of externalization." As he explains: "if one comes to feel that certain acts are disgusting (e.g., acts that violate certain kinds of norms), one will avoid committing such acts oneself and one will also avoid those who commit such acts (because they are contaminated thereby)" and, most importantly of all, "one may also think that others have reason to avoid such acts (since they would be contaminated thereby)." Consider, first, the suggestion that those who commit disgusting acts thereby become contaminated, and that we avoid them in order to avoid being contaminated ourselves. Although this does provide a means by which we might acquire simultaneous motivations to avoid both particular acts and potential social partners who commit them, notice that the proposed motivation itself consists in nothing more than a strong subjective preference for avoiding contamination *wherever we find it*. *Externalizing* such motivations and/or the norms they motivate is instead a matter of coming to regard myself and others as *responsible* for or *obligated* to refrain from the disgusting acts in question, rather than simply seeking to avoid potentially contaminating social partners as well as contaminating actions. Thus, if disgust is to provide "a simple and economical way of implementing externalization," this will have to be a consequence of the fact that when an agent finds an act disgusting she "may also think that others have reason to avoid such acts (since they would be contaminated thereby)." But again, thinking that others have good *practical* reasons for refraining from disgusting acts provides no ground at all for thinking that they are in any way responsible for or obligated to refrain from such acts. Externalization requires that we devalue social interaction with others *because* they have violated such responsibilities or obligations, not simply because they have failed to act on the good practical reasons they themselves have for avoiding particular actions, may therefore be or become contaminated themselves, and therefore threaten to contaminate us if we interact with them. Wiegman's claim that "disgust naturally lends itself to the thought that others *should* not contaminate themselves via contact with what I deem disgusting" (para. 2, my emphasis) trades on an ambiguity between such prudential and moralized senses of "should." It remains possible that paired aversions to both particular acts and those who commit such acts (simultaneously generated by my more general desire to avoid contamination) played some important role in the emergence of genuinely externalized norms and obligations, but disgust itself does not seem to provide the "immediate route to externalization" that Wiegman suggests.

This complexity also helps to highlight further reasons we might be well-advised to regard moralized disgust as an unusual or perhaps even unique form of externalization in any case. Following Kelly (2011), **Wiegman** suggests that moralized disgust is ultimately generated by the co-optation of a pre-existing system evolved to avoid poisons, pathogens, and parasites by a further and distinct system responsible for "motivat[ing] norm compliance and enforcement." If so, it seems more likely that moralized disgust emerged against the backdrop of an existing normative psychology, rather than providing the bridge by which such a normative psychology might have emerged in the first place. Note, however, that this complicated and unusual (perhaps even unique) phylogenetic history might well make it less surprising that moralized disgust often exhibits only a subset of the features suggested by Turiel and others to be characteristic of moral norms in general (sect. R2).

Zinser also seeks to propose an alternative explanation for moral externalization, suggesting that it is "merely a psychological by-product of underlying affective responses; it is a story we tell ourselves to make sense of a particular class of pre-existing subjective states" in which "our executive, conscious system has to make sense of what our automatic, non-conscious system has decided." As he notes, such "conscious explaining of our innate intuitions is called confabulation." This proposal, he suggests, offers not only a more parsimonious explanation of moral externalization itself, but also one that better coheres with both dual-process models of the mind and the sort of social intuitionism ably defended in recent years by thinkers like Jonathan Haidt.

There is little question that humans engage in a great deal of this sort of post-hoc confabulation, *especially* to defend or rationalize strongly held moral intuitions, and I have considerable sympathy for both dual-process models of the sort **Zinser** references and Haidt's social intuitionism. But Zinser's proposal nonetheless offers an unconvincing explanation of human moral externalization. For one thing, it ignores the fact that we *feel* (rather than merely believe) moral norms and obligations to be somehow imposed upon us externally. While there is little question that holding particular beliefs can influence both the phenomenological character and the content our experiences, I am aware of no other case in which our conscious minds are proposed to have engineered a phenomenologically novel form of experience (viz., that of feeling morally obligated) and introduced it *back* into the stream of our conscious experiences in order to rationalize or scaffold the confabulation we have adopted in response to that very experience. A further class of reasons includes facts like the ubiquity (if not universality, cf. **Patel & Machery**) of the moral/conventional distinction across human cultures and the invariance of the age at which children in different cultures begin to spontaneously externalize some, but not all, of the norms they accept. If, as Zinser goes on to suggest (see next paragraph), moral externalization were simply a conscious, post-hoc rationalization of the fact that some of the norms we embrace are other-regarding, we should expect to find considerable cultural variation in both the particular confabulations we arrive at and the point in ontogeny when they first appear.

But most importantly, there is simply no genuine phenomenon demanding explanation that **Zinser**'s proposal actually serves to explain. He suggests that the moral intuitions that invite or provoke us to externalize do so because they are other-regarding:

> My conscious, executive self struggles to find an explanation for this strong affective response that is seemingly concerned with the well-being of others (or concerns with justice, fairness, etc.). The answer we tell ourselves, given that the motivation seems to explicitly rule out merely self-centered motivation, is that justification for such preferences must be external. What else, my

> conscious self contends, could ground these other-centered affective responses? (para. 4 in Zinser commentary)

But the fact that a preference is other-*regarding* in this way does not mean it requires a special form of motivational rationalization: such other-regarding preferences are still just *my* subjective preferences about those others. It is not as if the fact that *others* are external to me somehow implies or even suggests that my own motivations to behave prosocially or altruistically towards them must also themselves have an external source. I have lots of other-regarding preferences that I do not externalize: I might, for example, want very much for my children (or yours) to have fulfilling and happy lives without thinking that either they or anyone is morally obligated to live such lives – these are simply the subjective preferences I hold *regarding* others. So there is no special puzzle about the source of motivation for our other-regarding preferences, and no need for us to confabulate moral externalization to make sense of them. Zinser's suggestion cannot actually explain why some (but not all) other-regarding attitudes are externalized, and of course, it also makes no sense of the fact that we do sometimes seem to externalize attitudes or demands that are purely self-regarding, in cases like moral duties to the self (of the sort familiar from the work of Immanuel Kant) and moral obligations of the sort that **Birch** suggests might both apply to and concern only the Pope and his own conduct. No special motivational rationalization is required for our other-regarding norms, preferences, and obligations, and even if there were such a demand, Zinser is wrong to think that confabulating externalization as such a rationalization would help satisfy it.

R7. Extensions, elucidations, and friendly amendments

A number of commentators seek to supplement, extend, or refine the account given in the target article in a variety of further ways that also repay careful consideration. **Bruner**, for example, argues that the set of norms and beliefs we should regard as likely to be or become externalized is wider than I have suggested (see also **Wiegman**) and includes norms governing "conflictual coordination" problems in which agents must coordinate to achieve a desired end but disagree about which of several different possible coordinative arrangements is most desirable. With such norms, free-riding and exploitation are not the outcomes that must be avoided, but instead miscoordination as well as any obstacles to resolving disagreements quickly, easily, and peacefully, which provides agents an "incentive to selectively interact with those adhering to the same norm as themselves" (para. 5 in Bruner's commentary). Bruner acknowledges the target article's claim that it may well be *possible* to moralize nearly any norm or behavior, but he suggests that norms governing conflictual coordination problems are particularly *likely* to be moralized for just the same reasons I suggest that norms protecting prosociality from exploitation are – namely, externalizing these particular norms is what generates the positive fitness consequences of externalization itself. I think Bruner is entirely correct and in fact points the way towards an even broader moral: It is not just norms enhancing cooperation and/or preventing exploitation of prosociality we should expect to be widespread among human societies, but instead any and all norms whose externalization would have significant positive consequences for our fitness. This will include norms governing conflictual coordination, just as Bruner suggests, but it will likely include a wide array of other norms and/or types of norms, too. In the target article, I focused on the fact that we can explain the ubiquity of norms enhancing cooperation or protecting prosociality from exploitation in this way, but Bruner's larger point is that the same explanation will apply to any norm whose externalization reliably ensures substantial positive consequences for our fitness. I should emphasize, however, that this recognition leaves entirely open the question of what process or combination of processes (descent with modification, convergent cultural evolution, biased learning, etc.) are those by which any particular norm or type of norm became and/or remains widespread across human cultures, just as it did in the more specific case of norms enhancing cooperation and protecting prosociality from exploitation.

Böhm et al. seek to extend the account offered in the target article in a different way. These authors suggest that the target article emphasizes the benefits of moralizing for facilitating cooperation and prosociality (especially within in-groups) to the exclusion of recognizing that these same moralizing tendencies play a central role in "fuel[ing] aggression and conflict between groups." My only reservation about their description of the target article's central morals is that the function of moral externalization is not simply (nor even primarily) to enhance and protect cooperation among all and only the members of a particular in-group; it does serve that end, but it does so by forming much more fine-grained networks of correlated interaction *even within a well-defined in-group* between agents inclined and/or willing to externalize particular norms and judgments. In fact, this is part of what makes externalization under normative plasticity so powerful – existing norms can be extended and applied to new situations and new norms moralized for the first time by just one or a few members of a community (generating correlated interaction between them) even when those agents constitute only a tiny fraction of the larger social group to which they belong (cf. **Brusse & Sterelny**; **Jebari & Huebner**; **Ross**).

Setting aside this quibble, I certainly concur with **Böhm et al.**'s suggestion that: "While morality may indeed foster cooperation and harmony within groups, it may also fuel aggression and conflict between groups." Moreover, I suggest that the further experimental evidence which Böhm et al. go on to report (Weisel & Böhm 2015) elegantly coheres with the account offered in the target article itself. These experiments measure out-group hatred as the "willingness to actively diminish out-group members' resources at personal cost in an intergroup social dilemma game." As Böhm et al. describe their results:

> Despite the availability of an outside option that had the same benefit for the in-group without necessarily harming the out-group, findings revealed a clear motivation to harm the out-group. Importantly, out-group hate increased substantially only in interaction with members of a morality-based out-group but not in interaction with members of a non–morality-based, yet high-enmity out-group. (para. 3 in Böhm et al. commentary)

This fascinating finding seems to offer further evidence of the distinctive role that specifically moral (but not other kinds of) disagreement plays in regulating human prosociality. Notice that mere *enmity* was insufficient to motivate agents to *pay* to punish members of another group – actual moral disagreement (for which membership in groups with moralized identities serves as a proxy) was required. So it is not just that externalization and moral advertisement often interact with in-group/out-group dynamics to leave us without sufficient information about members of out-groups to be willing to risk spontaneous prosociality with them: Weisel and Böhm show in addition that when group identities are themselves moralized, our resulting awareness of our moral disagreements with members of another group can motivate even costly efforts to actively harm the members of that group. This result nicely complements Skitka et al.'s (2005) finding that goodwill and cooperativeness among the members of a group trying to solve a problem are not lowered by the *existence* of moral disagreement between them but instead by their *awareness* of that disagreement.

Goodwin offers two similarly sympathetic elaborations of the account provided in the target article. The first are a pair of proposed mechanisms by which moral objectification might give rise to correlated interaction. One such possibility is that we may direct increased social attention towards those who are discovered to objectify a given norm that we ourselves also objectify, which seems plausible and is simply overlooked in the target article. I am somewhat puzzled by Goodwin's description of the other mechanism he proposes, which seems to assume that objectification is itself a public and intersubjectively available process, such that "[t]he objectification of a moral norm creates a strong expectation that others abide by this norm, but it also *conveys this social expectation*" (para. 3 in Goodwin commentary, original emphasis). What I have called externalization or objectification is a matter of private experience, in which I become motivated by a particular norm or obligation in a way that I experience as externally imposed on me and (therefore) upon others as well. It is instead the *advertisement* of one's externalized commitment to a given norm or obligation that represents a public and intersubjectively available act *conveying* "the strong expectation that others abide by this norm." It therefore seems to me that the additional mechanism Goodwin proposes just *is* the mechanism of moral advertisement. Accordingly, I think we must resist Goodwin's further suggestion that moral advertisement "could equally well be achieved by indicating a strong subjective preference to abide by" the norm in question. Not only does advertising our objectification of that norm (rather than simply a subjective preference for compliance with it) convey the expectation that others should also comply (as Goodwin clearly recognizes in formulating his own version of this mechanism), but it also assures others that we are motivated by it *in the right way*, such that we do not feel free to simply change our minds (as we do with merely subjective preferences) or faultlessly trade-off our subjective preference for compliance with the norm against other subjective preferences we might have. Those who externalize the demand to oppose Nazis prefer to interact with those who *externalize* the demand to oppose Nazis rather than those who regard their opposition to Nazis as a mere subjective preference or desire. This also illuminates the process by which moral demands become *iterated*: We devalue not only Nazis as social partners, but also those who tolerate Nazis as social partners, those who tolerate those who tolerate Nazis as social partners, and so on, further correlating interactions between those who do in fact externalize this norm (although, unsurprisingly, the relevant moral demand appears to become progressively weaker at each stage of such iteration).

But **Goodwin** also wants to know whether further features of norms (like their content) play a role in determining the *extent* to which disagreement concerning them motivates social exclusion. Re-analyzing the data reported in Goodwin and Darley (2012), he finds a significant correlation *for each particular norm* between its degree of objectification by an agent and social avoidance or discomfort with a disagreeing party, though also quite a wide and heterogeneous range in the magnitude of those correlations (suggesting that some forms of moralized disagreement are much more important than others in determining our preferred social distance from those with whom we disagree). Given earlier results, it is somewhat surprising that he finds no effect of valence, but he is quite right to suggest that this null result is perfectly consistent with the account offered in the target article, as the valence of a norm does not seem a particularly reliable indicator of the extent to which adhering to it will enhance cooperation and/or protect prosociality from exploitation. In any case, Goodwin is certainly right to suggest that important open questions remain concerning why objectification is "tied closely to social exclusion for some moral norms and not others."

A further sympathetic extension of the account proposed in the target article is offered by **Allidina & Cunningham**, who suggest that because we have relatively few opportunities to witness one another's genuinely consequential moralized behavior, "societies develop norms, games, and conventional rituals that allow for moral behaviour to play out in a relatively more symbolic form" that "allow[s] people to form impressions and predict how others will act in more serious moral situations" (para. 2 in their commentary). Although I suspect this is only one of a number of different ways in which norms seemingly unconcerned with protecting prosociality and cooperation from exploitation can become moralized, it does seem to offer a natural explanation (or partial explanation) for the fact that we moralize *so much* of our ordinary or everyday behavior, such as standing quietly for the national anthem, dressing modestly, recycling, and our conduct in sports and games (Allidina & Cunningham's examples). This explanation would also seem to have the endorsement of legendary men's basketball coach John Wooden, widely regarded as the original source of the adage "sports do not build character; they reveal it."

R8. Conclusion: In defense of the folk

Let me close by thanking my commentators once again for their insightful contributions and by returning to a broad theme that runs throughout many of the commentaries addressed specifically to the phenomenology of moral experience and to the beliefs of "the folk" concerning moral objectivity. Moral philosophers will certainly want to know whether the proposal I have offered constitutes a

so-called "error theory" of morality (or a "debunking" view of morality), which is to say one that implies that moral claims are generally false because the folk are gravely and systematically in error concerning the character of the objectivity they attribute to moral norms, obligations, motivations, and the like. There are a number of reasons to resist this characterization of the account I have offered, not least of which is the fact that I am certainly prepared to accept the reinterpretation it offers for my own moralizing as a broadly accurate description of what I myself have been up to all along. But I find it hard to even make sense of this question as applied to the beliefs and/or experiences of the folk. The folk tend not to have considered views on subjects like the nature of moral objectivity. They don't mistakenly think that moral norms and obligations are external in the same way that rocks and trees are, or in any other particular way – what the folk know is that such norms and obligations are *somehow* external to us and motivate us in a way that is *somehow* importantly different from that of mere preferences and conventional norms (which is true). Perhaps the collection of claims to which many of the folk would assent regarding the character of moral objectivity and/or motivation even includes demonstrable falsehoods or inconsistencies, but we have known since Plato wrote the *Euthyphro* that *nothing* could have the entire collection of properties the folk are thought by philosophers to attribute to moral norms and obligations. And it is philosophers who attribute such sharp and determinate beliefs about the nature of moral objectivity and motivation to the folk, not the folk themselves. Or to put things another way, it seems to me a slander against the folk to attribute to them beliefs about the nature of moral objectivity sufficiently clear and determinate to be falsified by the account I have offered and defended here.

References

[The BBS paper source for each reference is indicated by the author's name initials given in square brackets after the reference. The letters "a" and "r" before author's initials in square brackets stand for target article and response references, respectively]

Aharoni, E., Sinott-Armstrong, W. & Kiehl, K. A. (2012) Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology* 121:484–97. [aPKS]

Allen, C. & Bekoff, M. (2005) Animal play and the evolution of morality: An ethological approach. *Topoi* 24(2):125–35. Available at: https://doi.org/10.1007/s11245-005-5050-8. [SA]

Apicella, C. L., Marlowe, F. W., Fowler, J. H. & Christakis, N. A. (2012) Social networks and cooperation in hunter-gatherers. *Nature* 481(7382):497–501. [JJ]

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59:390–412. Available at: http://dx.doi.org/10.1016/j.jml.2007.12.005. [JT]

Balliet, D., Mulder, L. B. & Van Lange, P. A. (2011) Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin* 137:594–615. [SP, rPKS]

Barkoczi, D. & Galesic, M. (2016) Social learning strategies modify the effect of network structure on group performance. *Nature Communications* 7: article no. 13109. doi:10.1038/ncomms13109. [JJ]

Bar-Tal, D. (2000) *Shared beliefs in a society*. Sage. [BV]

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H. & Strongman, J. A. (1999) Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology* 77:525–37. [J-WvP]

Baumard, N., André, J.-B. & Sperber, D. (2013) A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences* 36:59–122. [arPKS]

Beebe, J. R. (2014) How different kinds of disagreement impact folk metaethical judgments. In: *Advances in experimental moral psychology*, ed. H. Sarkissian & J. Wright, pp. 167–87. Bloomsbury. [SS, rPKS]

Beebe, J. R. (2015) The empirical study of folk metaethics. *Etyka* 50/2015. (Online article). Available at: http://etyka.uw.edu.pl/en/archive/empirical-study-folk-metaethics/ [SS]

Beebe, J. R. & Sackris, D. (2016) Moral objectivism across the lifespan. *Philosophical Psychology* 29(6):912–29. [TD, SS]

Bekoff, M. (2001) Social play behaviour: Cooperation, fairness, trust, and the evolution of morality. *Journal of Consciousness Studies* 8(2):81–90. Available at: https://doi.org/10.2307/1309460. [SA]

Bergendorff, S. (2016) *Kinship and human evolution: Making culture, becoming human*. Lexington Books. [BV]

Bicchieri, C. (2016) *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press. [JJ]

Blair, R. (1995) A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57:1–29. [aPKS]

Boehm, C. (2012) *Moral origins: The evolution of virtue, altruism, and shame*. Basic Books. [J-WvP]

Boesch, C. (2005) Joint cooperative hunting among wild chimpanzees: Taking natural observations seriously. *Behavioral and Brain Sciences* 28(5):692–93. [aPKS]

Böhm, R., Rusch, H. & Gürerk, Ö. (2016) What makes people go to war? Defensive intentions motivate retaliatory and preemptive intergroup aggression. *Evolution and Human Behavior* 37:29–34. [RB]

Boyd, R. & Richerson, P. J. (2005) *The origin and evolution of cultures*. Oxford University Press. [aPKS, TD]

Brewer, M. B. (1999) The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues* 55:429–44. [RB]

Brown, S. L., Nesse, R. M., Vinokur, A. D. & Smith, D. M. (2003) Providing social support may be more beneficial than receiving it: Results from a prospective study of mortality. *Psychological Science* 14:320–27. [MJP]

Buckley, C. (2007) Man is rescued by stranger on subway tracks. *The New York Times*, January 3, 2007, p. A1. Available at: http://www.nytimes.com/2007/01/03/nyregion/03life.html. [TJ]

Buhl, T. (1999) Positive-negative asymmetry in social discrimination: Meta-analytical evidence. *Group Processes & Intergroup Relations* 2:51–58. [RB]

Bullinger, A. F., Melis, A. P. & Tomasello, M. (2011) Chimpanzees, *Pan troglodytes*, prefer individual over collaborative strategies towards goals. *Animal Behaviour* 82:1135–41. [aPKS]

Burkhardt, J. M., Allon, O., Amici, F., Fichtel, C., Finkenwirth, A., Heschl, A., Huber, J., Isler, K., Kosonen, Z. K., Martins, E., Meulman, E. J., Richiger, R., Rueth, K., Spilmann, B., Wiesendanger, S. & van Schaik, C. P. (2014) The evolutionary origin of human hyper-cooperation. *Nature Communications* 5: article no. 4747. doi: 10.1038/ncomms5747. [aPKS]

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liskowski, U., Behne, T. & Tomasello, M. (2011) Early social cognition in three cultural contexts. *Monographs of the Society for Research in Child Development* 76:vii–viii, 1–142. [aPKS]

Caprara, G. V. & Steca, P. (2005) Self-efficacy beliefs as determinants of prosocial behavior conducive to life satisfaction across ages. *Journal of Social and Clinical Psychology* 24:191–217. [MJP]

Chapman, H. A., Kim, D. A., Susskind, J. M. & Anderson, A. K. (2009) In bad taste: Evidence for the oral origins of moral disgust. *Science* 323(5918):1222–26. [IW]

Cheney, D. L. (2011) Extent and limits of cooperation in animals. *Proceedings of the National Academy of Sciences USA* 108(Suppl. 2):10902–909. [aPKS]

Choi, J.-K. & Bowles, S. (2007) The coevolution of parochial altruism and war. *Science* 318:636–40. [RB]

Chudek, M. & Henrich, J. (2011) Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences* 15(5):218–26. [TD]

Cialdini, R. B. & Kenrick, D. T. (1976) Altruism as hedonism: A social development perspective on the relationship of negative mood state and helping. *Journal of Personality and Social Psychology* 34:907–14. [MJP]

Cialdini, R. B., Reno, R. R. & Kallgren, C. A. (1990) A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58:1015–26. [MJP]

Clark, H. (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12:335–59. Available at: http://dx.doi.org/10.1016/s0022-5371(73)80014-3. [JT]

Cohen, E. (2012) The evolution of tag-based cooperation in humans: The case for accent. *Current Anthropology* 53(5):588–616. Available at: https://doi.org/10.1086/667654. [TH]

Cooper, R., DeJong, D. V., Forsythe, R. & Ross, T. W. (1996) Cooperation without reputation: Experimental evidence from Prisoner's Dilemma games. *Games and Economic Behavior* 12(2):187–218. [MJP]

Damasio, A. (2012) *Self comes to mind: Constructing the conscious brain*. Vintage Books. [BV]

Darwall, S. (2006) *The second-person standpoint: Morality, respect, and accountability*. Harvard University Press. [CI-M, rPKS]
Darwin, C. (1871) *The descent of man, and selection in relation to sex*. John Murray. [aPKS]
Decety, J. (2010) The neurodevelopment of empathy in humans. *Developmental Neuroscience* 32(4):257–67. [AK]
De Leersnyder, J., Boiger, M. & Mesquita, B. (2013) Cultural regulation of emotion: Individual, relational, and structural sources. *Frontiers in Psychology* 4, Article 55:1–11. (Online article). [BV]
Dennett, D. (1995) *Darwin's dangerous idea*. Simon and Schuster. [aPKS]
Derex, M. & Boyd, R. (2016) Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences USA* 113:2982–87. [JJ]
DeScioli, P. & Kurzban, R. (2009) Mysteries of morality. *Cognition* 112:281–99. [aPKS, J-WvP]
DeScioli, P. & Kurzban, R. (2013) A solution to the mysteries of morality. *Psychological Bulletin* 139:477–96. [aPKS]
De Waal, F. (1996) *Good natured: The origins of right and wrong in humans and other animals*. Harvard University Press. [aPKS, JZ]
De Waal, F. (2006) *Primates and philosophers: How morality evolved*, ed. J. Ober & S. Macedo. (Includes commentaries by R. Wright, C. Korsgaard, P. Kitcher, & P. Singer.) Princeton University Press. [aPKS]
Dondi, M., Simion, F. & Caltran, G. (1999) Can newborns discriminate between their own cry and the cry of another newborn infant? *Developmental Psychology* 35(2):418–26. [AK]
Dubreuil, B. (2010) Punitive emotions and norm violations. *Philosophical Explorations* 13(1):35–50. [BV]
Dunbar, R. I. M. (1996) *Grooming, gossip and the evolution of language*. Harvard University Press. [aPKS]
Dunbar, R. I. M. (2004) Gossip in evolutionary perspective. *Journal of General Psychology* 8:100–10. [aPKS]
Dunn, E., Aknin, L. & Norton, M. (2014) Prosocial spending and happiness: Using money to benefit others pays off. *Current Directions in Psychological Science* 23:41–47. [MJP]
Edgerton, R. (1992). *Sick societies*. Free Press. [DR, rPKS]
Ellemers, N. (2012) The group self. *Science* 336:848–52. [BV]
Ellemers, N., Spears, R. & Doosje, B. (2002) Self and social identity. *Annual Review of Psychology* 53:161–86. [BV]
Emler, N. (1990) A social psychology of reputation. *European Journal of Social Psychology* 1:171–93. [aPKS]
Emler, N. (1994) Gossip, reputation, and social adaptation. In: *Good gossip*, ed. R. F. Goodman & A. Ben-Ze'ev, pp. 117–38. Kansas University Press. [aPKS]
Emler, N. (2001) Gossiping. In: *The new handbook of language and social psychology*, ed. W. P. Robinson & H. Giles, pp. 317–38. John Wiley. [aPKS]
Engel, C. (2011) Dictator games: A meta study. *Experimental Economics* 14:583–610. [TJ]
Enquist, M. & Leimar, O. (1993) The evolution of cooperation in mobile organisms. *Animal Behaviour* 45:747–57. [aPKS]
Epley, N. & Dunning, D. (2000) Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction. *Journal of Personality and Social Psychology* 79:861–75. [aPKS]
Fehl, K., van der Post, D. J. & Semmann, D. (2011) Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters* 14 (6):546–51. [JJ]
Fehr, E. & Fischbacher, U. (2003) The nature of human altruism. *Nature* 425:785–91. [aPKS]
Fletcher, G., Warneken, F. & Tomasello, M. (2012) Differences in cognitive processes underlying the collaborative activities of children and chimpanzees. *Cognitive Development* 27:136–53. [aPKS]
Foot, P. (1972) Morality as a system of hypothetical imperatives. *The Philosophical Review* 81:305–16. [SP]
Fortes, M. (1969) *Kinship and the social order: The legacy of Lewis Henry Morgan*. Aldine. [BV]
Frank, R. H. (1988) *Passions within reason: The strategic role of the emotions*. W. W. Norton. [aPKS, CJB]
Gabennesch, H. (1990) The perception of social conventionality by children and adults. *Child Development* 61:2047–59. [aPKS]
Gardner, A. & West, S. A. (2010) Greenbeards. *Evolution* 64(1):25–38. Available at: https://doi.org/10.1111/j.1558-5646.2009.00842.x. [TH]
Giner-Sorolla, R. & Chapman, H. A. (2017) Beyond purity: Moral disgust toward bad character. *Psychological Science* 28(1):80–91. Available at: https://doi.org/10.1177/0956797616673193. [IW]
Ginsborg, H. (2014) Kant's aesthetics and teleology. In: *The Stanford encyclopedia of philosophy*, Fall 2014 Online edition, ed. E. N. Zalta. Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. Available at: https://plato.stanford.edu/entries/kant-aesthetics/. [CI-M]
Gintis, H., Smith, E. & Bowles, S. (2001) Costly signaling and cooperation. *Journal of Theoretical Biology* 213(1):103–19. Available at: https://doi.org/10.1006/jtbi.2001.2406. [TH]
Goodwin, G. P. & Darley, J. M. (2008) The psychology of meta-ethics: Exploring objectivism. *Cognition* 106:1339–66. Available at: http://dx.doi.org/10.1016/j.cognition.2007.06.007. [arPKS, TD, SP, AWS, SS, JT]
Goodwin, G. P. & Darley, J. M. (2010) The perceived objectivity of ethical beliefs: Psychological findings and implications for public policy. *Review of Philosophy and Psychology* 1:161–88. Available at: http://dx.doi.org/10.1007/s13164-009-0013-4. [JT]
Goodwin, G. P. & Darley, J. M. (2012) Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology* 48 (1):250–56. Available at: http://dx.doi.org/10.1016/j.jesp.2011.08.006. [arPKS, TD, GPG, JT]
Gould, S. J. & Lewontin, R. C. (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London B: Biological Sciences* 205(1161):581–98. Available at: http://doi.org/10.1098/rspb.1979.0086. [CI-M]
Graham, J., Haidt, J. & Nosek, B. A. (2009) Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96:1029–46. [J-WvP]
Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S. & Ditto, P. H. (2011) Mapping the moral domain. *Journal of Personality and Social Psychology* 101 (2):366–85. Available at: http://dx.doi.org/10.1037/a0021847. [JT]
Gray, K. & Schein, C. (2016) No absolutism here: Harm predicts moral judgment 30× better than disgust – Commentary on Scott, Inbar, & Rozin (2016). *Perspectives on Psychological Science* 11:325–29. [MJP]
Gray, K., Schein, C. & Ward, A. F. (2014) The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General* 143:1600–15. [MJP]
Gray, K., Young, L. & Waytz, A. (2012) Mind perception is the essence of morality. *Psychological Inquiry* 23:101–24. [RB]
Greene, J. (2013) *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin. [JZ]
Guala, F. (2012) Reciprocity: Strong or weak? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences* 35(1):1–15. [aPKS]
Haidt, J. (2001) The emotional dog and its rational tail. *Psychological Review* 108:814–34. [JZ]
Haidt, J. (2007) The new synthesis in moral psychology. *Science* 316:998–1002. [JZ]
Haidt, J. & Joseph, C. (2004) Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133(4):55–66. [SA]
Haidt, J. & Joseph, C. (2007) The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The Innate Mind* 3:367–91. Available at: https://doi.org/10.1093/acprof:oso/9780195332834.003.0019. [SA]
Haidt, J., Koller, S. H. & Dias, M. G. (1993) Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65:613–28. [aPKS, MJP]
Halevy, N., Bornstein, G. & Sagiv, L. (2008) "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science* 19:405–11. [RB]
Haley, K. J. & Fessler, D. M. T. (2005) Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior* 26(3):245–56. Available at: https://doi.org/10.1016/j.evolhumbehav.2005.01.002. [TH, J-WvP]
Halperin, E., Russell, A. G., Dweck, C. S. & Gross, J. J. (2011) Anger, hatred, and the quest for peace: Anger can be constructive in the absence of hatred. *Journal of Conflict Resolution* 55:274–91. [RB]
Hamann, K., Warneken, F., Greenberg, J. R. & Tomasello, M. (2011) Collaboration encourages equal sharing in children but not in chimpanzees. *Nature* 476:328–31. [aPKS]
Hamilton, W. D. (1964a) The genetical evolution of social behaviour. I. *Journal of Theoretical Biology* 7:1–16. [TJ]
Hamilton, W. D. (1964b) The genetical evolution of social behaviour. II. *Journal of Theoretical Biology* 7:17–52. [TJ]
Hamlin, J. K., Mahajan, N., Liberman, Z. & Wynn, K. (2013) Not like me=bad: Infants prefer those who harm dissimilar others. *Psychological Science* 24:589–94. [aPKS]
Hamlin, J. K., Wynn, K. & Bloom, P. (2007) Social evaluation by preverbal infants. *Nature* 450:557–59. [aPKS]
Haney, C., Banks, C. & Zimbardo, P. (1972) Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology* 1:69–97. [TH]
Hardin, C. D. & Higgins, E. T. (1996) Shared reality: How social verification manes the subjective objective. In: *Handbook of motivation and cognition, vol. 3*, ed. R. M. Sorrentino & E. T. Higgins, pp. 28–84. Guilford Press. [BV]
Hare, B. (2017) Survival of the friendliest: *Homo sapiens* evolved via selection for prosociality. *Annual Review of Psychology* 68:155–86. [JJ]

Hare, B. & Tomasello, M. (2004) Chimpanzees are more skillful in competitive than in cooperative cognitive tasks. *Animal Behaviour* 68:571–81. [aPKS]
Haslam, N. (2006) Dehumanization: An integrative review. *Personality and Social Psychology Review* 10:252–64. [RB]
Haviland, J. B. (1977) Gossip as competition in Zinacantan. *Journal of Communication* 27:186–91. [aPKS]
Henrich, J. (2004) Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses-the Tasmanian case. *American Antiquity* 69:197–214. [JJ]
Henrich, J. (2009) The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior* 30(4):244–60. Available at: https://doi.org/10.1016/j.evolhumbehav.2009.03.005. [SA]
Henrich, J. (2016) *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press. [JJ]
Henrich, J., Heine, S. J. & Norenzayan, A. (2010) The weirdest people in the world? *Behavioral and Brain Sciences* 33(2):61–135. [aPKS]
Hill, K. R., Walker, R. S., Božičević, M., Eder, J., Headland, T., Hewlett, B., Hurtado, A. M., Marlowe, F., Wiessner, P. & Wood, B. (2011) Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science* 331(6022):1286–89. [JJ]
Hoffman, D. D. (2009) The interface theory of perception: Natural selection drives true perception to swift extinction. In: *Object categorization: Computer and human vision perspectives*, ed. S. Dickinson, M. Tarr, A. Leonardis & B. Schiele, pp. 148–65. Cambridge University Press. [AK]
Hofmann, W., Wisneski, D. C., Brandt, M. J. & Skitka, L. J. (2014) Morality in everyday life. *Science* 345(6202):1340–43. Available at: https://doi.org/10.1126/science.1251560. [SA]
Hooker, C. (2013) On the import of constraints in complex dynamical systems. *Foundations of Science* 18(4):757–80. [JJ]
Horner, V., Carter, J. D., Suchak, F. & de Waal, F. B. M. (2011) Spontaneous prosocial choice in chimpanzees. *Proceedings of the National Academy of Sciences USA* 108:13847–51. [aPKS]
Hrdy, S. (2009) *Mothers and others: The evolutionary origins of mutual understanding*. Harvard University Press. [JJ]
Hume, D. (1738/1975) *A treatise of human nature*, ed. L. A. Selby-Bigge, rev. P. H. Nidditch, Clarendon Press. [SP, JZ]
Isler, K. & Van Schaik, C. P. (2012) How our ancestors broke through the gray ceiling: Comparative evidence for cooperative breeding in early *Homo*. *Current Anthropology* 53(Suppl. 6):S453–65. [JJ]
Iyer, R., Koleva, S., Graham, J., Ditto, P. & Haidt, J. (2012) Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE* 7(8):e42366. Available at: http://dx.doi.org/10.1371/journal.pone.0042366. [JT]
Jansen, V. A. A. & van Baalen, M. (2006) Altruism through beard chromodynamics. *Nature* 440(7084):663–66. Available at: https://doi.org/10.1038/nature04387. [TH]
Jebari, J. (in preparation) Empirical moral rationalism and the social constitution of normativity. [JJ]
Jensen, K., Hare, B., Call, J. & Tomasello, T. (2006) What's in it for me? Self-regard precludes altruism and spite in chimpanzees. *Proceedings of the Royal Society of London, Series B* 273:1013–21. [aPKS]
Johnson, T. & Smirnov, O. (2012) An alternative mechanism through which economic inequality facilitates collective action: Wealth disparities as a sign of cooperativeness. *Journal of Theoretical Politics* 24:461–84. [TJ]
Johnson, T. & Smirnov, O. (2013) Cooperate with equals: A simple heuristic for social exchange. In: *Simple heuristics in a social world*, ed. R. Herwig, U. Hoffrage & the ABC Research Group, pp. 135–70. Oxford University Press. [TJ]
Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. (2016) Third-party punishment as a costly signal of trustworthiness. *Nature* 530(7591):473–76. Available at: https://doi.org/10.1038/nature16981. [TH]
Jordan, J. J., Sommers, R., Bloom, P. & Rand, D. G. (2017) Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Sciences* 28:356–68. [aPKS]
Joyce, R. (2006) *The evolution of morality*. MIT Press. [arPKS, AK, JZ]
Judd, C. M., Westfall, J. & Kenny, D. A. (2012) Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology* 103:54–69. Available at: http://dx.doi.org/10.1037/a0028347. [JT]
Kahneman, D. (2011) *Thinking, fast and slow*. Farrar, Straus and Giroux. [JZ]
Kant, I. (1790/1987) *Kritique der Urtailskraft* [*Critique of judgment*], ed. W. Pluhar. Hackett. (Original work published in 1790; Hackett English edition in 1987). [CI-M]
Kaznatcheev, A. (2010) Robustness of ethnocentrism to changes in interpersonal interactions. In: *Papers from the AAAI Fall Symposium (FS-10-03): Complex Adaptive Systems – Resilience, Robustness, and Evolvability*, pp. 71–75. Association for the Advancement of Artificial Intelligence. Available as pdf file at: https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/download/2314/2630 [AK]
Kaznatcheev, A., Montrey, M. & Shultz, T. R. (2014) Evolving useful delusions: Subjectively rational selfishness leads to objectively irrational cooperation. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* 36:731–36. (Online article published in April 2014; available at: arXiv:1405.0041v1 [q-bio.PE]). [AK]
Kelly, D. (2011) *Yuck! The nature and moral significance of disgust*. MIT Press. [IW, rPKS]
Kelly, D. & Stich, S. (2007) Two theories of the cognitive architecture underlying morality. In: *The innate mind, vol. 3: Foundations and future horizons*, ed. P. Carruthers, S. Laurence & S. Stich, pp. 348–66. Oxford University Press. [TD]
Kelly, D., Stich, S., Haley, K. J., Eng, S. J. & Fessler, D. M. T. (2007) Harm, affect, and the moral/conventional distinction. *Mind and Language* 22:117–31. Available at: http://dx.doi.org/10.1093/acprof:oso/9780199733477.003.0013. [aPKS, SP, SS, JT]
Kim, H. S. & Sasaki, J. Y. (2012) Emotional regulation: The interplay of culture and genes. *Social and Personality Psychology Compass* 6(12):865–77. [BV]
Kinzler, K. D., Dupoux, E. & Spelke, E. S. (2007) The native language of social cognition. *Proceedings of the National Academy of Sciences USA* 104:12577–80. [aPKS]
Kline, M. A. (2015) How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain Sciences* 38(1):1–71. [aPKS]
Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A. & Fernández, G. (2009) Reinforcement learning signal predicts social conformity. *Neuron* 61:140–51. [JJ]
Klucharev, V., Munneke, M., Smidts, A. & Fernández, G. (2011) Downregulation of the posterior medial frontal cortex prevents social conformity. *Journal of Neuroscience* 31:11934–40. [JJ]
Kumar, V. (2015) Moral judgment as a natural kind. *Philosophical Studies* 172:2887–910. doi: 10.1007/s11098-015-0448-7. [SS]
Lane, D. R. M, Maxfield, D., Read, D. & van der Leeuw, S. (2009) From population to organization thinking. In: *Complexity perspectives in innovation and social change*, ed. D. Lane, D. Pumain, S. E. van der Leeuw & G. West, pp. 11–42. Springer. [BV]
Leaf, M. & Read, D. (2012) *Human thought and social organization: Anthropology on a new plane*. Lexington Press. [BV]
LeDoux, J. (2012) Rethinking the emotional brain. *Neuron* 73(4):653–76. [BV]
Leidner, B. & Castano, E. (2012) Morality shifting in the context of intergroup violence. *European Journal of Social Psychology* 42:82–91. [RB]
Machery, E. (2012) Delineating the moral domain. *Baltic International Yearbook of Cognition, Logic and Communication* 7:1–14. Available at: http://dx.doi.org/10.4148/biyclc.v7i0.1777. [SP]
Machery, E. (2018) Morality: A historical invention. In: *Atlas of moral psychology*, ed. K. Gray & J. Graham, pp. 259–65. Guilford Press. [SP]
Machery, E. & Mallon, R. (2010) Evolution of morality. In: *The moral psychology handbook*, ed. J. M. Doris & the Moral Psychology Research Group, pp. 3–46. Oxford University Press. [SP]
MacIntyre, A. (1957) What morality is not. *Philosophy* 32:325–35. [SS]
Mahajan, N. & Wynn, K. (2012) Origins of "us" versus "them": Prelinguistic infants prefer similar others. *Cognition* 124:227–33. [aPKS]
Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W. & Cardinale, E. M. (2014) Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences USA* 111(42):15036–41. Available at: https://doi.org/10.1073/pnas.1408440111. [SA]
Maynard Smith, J. (1982) *Evolution and the theory of games*. Cambridge University Press. [JPB]
McElreath, R., Boyd, R. & Richerson, P. J. (2003) Shared norms and the evolution of ethnic markers. *Current Anthropology* 44(1):122–30. Available at: https://doi.org/10.1086/345689. [IW]
Melis, A. P., Altricher, K. & Tomasello, M. (2013) Allocation of resources to collaborators and free-riders in 3-year-olds. *Journal of Experimental Child Psychology* 114:364–70. [aPKS]
Melis, A. P., Schneider, A.-C. & Tomasello, M. (2011a) Chimpanzees, *Pan troglodytes*, share food in the same way after collaborative and individual food acquisition. *Animal Behaviour* 82:485–93. [aPKS]
Melis, A. P. & Semmann, D. (2010) How is human cooperation different? *Philosophical Transactions of the Royal Society of London, B: Biological Sciences* 365:2663–74. [aPKS]
Melis, A. P., Warneken, F., Jensen, K., Schneider, A., Call, J. & Tomasello, M. (2011b) Chimpanzees help conspecifics to obtain food and non-food items. *Philosophical Transactions of the Royal Society of London, Series B* 278:1405–13. [aPKS]
Milgram, S. & Sabini, J. (1978) On maintaining urban norms: A field experiment in the subway. *Advances in Environmental Psychology* 1:31–40. [JJ]
Millikan, R. (2002) *Varieties of meaning*. MIT Press. [AWS]

Moll, H. & Tomasello, M. (2007) Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B* 362:639–48. [aPKS]

Montague, R. (2006) *Why choose this book?* Dutton Press. [JJ]

Muthukrishna, M., Shulman, B. W., Vasilescu, V. & Henrich, J. (2014) Sociality influences cultural complexity. *Proceedings of the Royal Society of London B: Biological Sciences* 281(1774):e20132511. doi: 10.1098/rspb.2013.2511. [JJ]

Nado, J., Kelly, D. & Stich, S. (2009) Moral judgment. In: *The Routledge companion to the philosophy of psychology*, ed. J. Symons & P. Calvo, pp. 621–33. Routledge. [TD]

Nelissen, R. M. A. & Meijers, M. H. C. (2011) Social benefits of luxury brands as costly signals of wealth and status. *Evolution and Human Behavior* 32:343–55. [TJ]

Nichols, S. (2004) *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press. [aPKS, JZ]

Nichols, S. (2014) Process debunking and ethics. *Ethics* 124:727–49. [IW]

Nichols, S. & Folds-Bennett, T. (2003) Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition* 90(2): B23–32. Available at: http://dx.doi.org/10.1016/s0010-0277(03)00160-4. [aPKS, AK, EO, SP, JT]

Nisan, M. (1987) Moral norms and social conventions: A cross-cultural comparison. *Developmental Psychology* 23:719–25. [aPKS]

Nisbett, R. E. & Wilson, T. D. (1977) Telling more than we can know. *Psychological Review* 84:231–59. [JZ]

Nowak, M. A. (2006) Five rules for the evolution of cooperation. *Science* 314:1560–63. [MJP]

Nucci, L. P. (1986) Children's conceptions of morality, social convention, and religious prescription. In: *Moral dilemmas: Philosophical and psychological reconsiderations of the development of moral reasoning*, ed. C. Harding, pp. 137–74. Precedent Press. [aPKS]

Ohtsubo, Y. & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior* 30(2):114–23. Available at: https://doi.org/10.1016/j.evolhumbehav.2008.09.004. [TH]

Ohtsuki, H., Hauert, C., Lieberman, E. & Nowak, M. A. (2006) A simple rule for the evolution of cooperation on graphs. *Nature* 441(7092):502–505. [JJ]

O'Neill, E. & Machery, E. (forthcoming) What is the normative sense? Cross cultural evidence. In: *Routledge handbook of moral epistemology*, ed. A. Zimmerman, K. Jones & M. Timmons. Routledge. [SP]

Parker, M. T. & Janoff-Bulman, R. (2013) Lessons from morality-based social identity: The power of outgroup "hate," not just ingroup "love." *Social Justice Research* 26:81–96. [RB]

Poulin, M. J. (2014) Volunteering predicts health among those who value others: Two national studies. *Health Psychology* 33:120–29. [MJP]

Poulin, M. J., Brown, S. L., Dillard, A. J. & Smith, D. M. (2013) Giving to others and the association between stress and mortality. *American Journal of Public Health* 103:1649–55. [MJP]

Poulin, M. J. & Holman, E. A. (2013) Helping hands, healthy body? Oxytocin receptor gene and prosocial behavior interact to buffer the association between stress and physical health. *Hormones and Behavior* 63:510–17. [MJP]

Povinelli, D. J. (2000) *Folk physics for apes: The chimpanzee's theory of how the world works*. Oxford University Press. [aPKS]

Prinz, J. (2007) *The emotional construction of morals*. Oxford University Press. [JZ]

Quintelier, K., De Smet, D. & Fessler, D. (2014) Agent versus appraiser moral relativism: An exploratory study. In: *Advances in experimental moral psychology*, ed. H. Sarkissian & J. Wright, pp. 209–30. Bloomsbury. [SS]

Rakoczy, H. (2007) Play, games, and the development of collective intentionality. *New Directions for Child and Adolescent Development* 2007(115):53–67. Available at: https://doi.org/10.1002/cad. [SA]

Rakoczy, H. & Tomasello, M. (2007) The ontogeny of social ontology: Steps to shared intentionality and status functions. In: *Intentional acts and institutional facts*, ed. S. L. Tsohatzidis, pp. 113–37. Springer. [SA]

Rand, D. G. (2016) Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science* 27:1192–206. [MJP]

Rand, D. G., Arbesman, S. & Christakis, N. A. (2011) Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences USA* 108(48):19193–98. [JJ]

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A. & Greene, J. D. (2014) Social heuristics shape intuitive cooperation. *Nature Communications* 5: article 3677. doi:10.1038/ncomms4677. [TJ]

Read, D. (2012) *How culture makes us human*. Left Coast Press. [BV]

Read, D., Lane, D. & van der Leeuw, S. (2009) The innovation innovation. In: *Complexity perspectives in innovation and social change*, ed. D. Lane, D. Pumain, S. E. van der Leeuw & G. West, pp. 43–84. Springer. [BV]

Reber, R. & Norenzayan, A. (in press) Shared fluency theory of social cohesiveness: How the metacognitive feeling of processing fluency contributes to group processes. In: *Metacognitive diversity*, ed. J. Proust & M. Fortier. Oxford University Press. [JJ]

Reia, S. M., Herrmann, S. & Fontanari, J. F. (2017) Impact of centrality on cooperative processes. *Physical Review E* 95(2):022305. (Online article). Available at: https://doi.org/10.1103/PhysRevE.95.022305 [JJ]

Rossano, M. J. (2012) The essential role of ritual in the transmission and reinforcement of social norms. *Psychological Bulletin* 138(3):529–49. Available at: https://doi.org/10.1037/a0027038. [SA]

Roth-Hanania, R., Davidov, M. & Zahn-Waxler, C. (2011) Empathy development from 8 to 16 months: Early signs of concern for others. *Infant Behavior and Development* 34(3):447–58. [AK]

Rozin, P. & Fallon, A. E. (1987) A perspective on disgust. *Psychological Review* 94 (1):23–41. [IW]

Rozin, P. & Haidt, J. (2013) The domains of disgust and their origins: Contrasting biological and cultural evolutionary accounts. *Trends in Cognitive Sciences* 17(8):367–68. Available at: https://doi.org/10.1016/j.tics.2013.06.001. [IW]

Sarkissian, H. (2016) Aspects of folk morality: Objectivism and relativism. In: *A companion to experimental philosophy*, ed. J. Sytsma & W. Buckwalter, pp. 212–24. Wiley Blackwell. [SS]

Sarkissian, H., Park, J., Tien, D., Wright, J. C. & Knobe, J. (2011) Folk moral relativism. *Mind and Language* 26:482–505. [arPKS, SP, SS]

Schein, C., Ritter, R. S. & Gray, K. (2016) Harm mediates the disgust-immorality link. *Emotion* 16:862–76. [MJP]

Schmidt, M. F., Gonzalez-Cabrera, I. & Tomasello, M. (2017) Children's developing metaethical judgments. *Journal of Experimental Child Psychology* 164:163–77. [AK]

Schulz, A. (2011) The adaptive importance of cognitive efficiency: An alternative theory of why we have beliefs and desires. *Biology and Philosophy* 26(1):31–50. [AWS]

Schulz, A. (2013) The benefits of rule following: A new account of the evolution of desires. *Studies in History and Philosophy of Science, Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44(4, Part A):595–603. [AWS]

Schulz, A. (2018) *Efficient cognition: The evolution of representational decision making*. MIT Press. [AWS]

Schurz, M., Radua, J., Aichhorn, M., Richlan, F. & Perner, J. (2014) Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews* 42:9–34. Available at: http://dx.doi.org/10.1016/j.neubiorev.2014.01.009. [JT]

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J. & Schapiro, S. J. (2005) Chimpanzees are indifferent to the welfare of unrelated group members. *Nature* 437:1357–59. [aPKS]

Silk, J. B. & House, B. R. (2011) Evolutionary foundations of prosocial sentiments. *Proceedings of the National Academy of Sciences USA* 108(Suppl. 2):10910–17. [aPKS]

Skitka, L. J., Bauman, C. W. & Sargis, E. G. (2005) Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology* 8:895–917. [arPKS, SP]

Skyrms, B. (1996) *Evolution of the social contract*. Cambridge University Press. [aPKS, AWS]

Skyrms, B. (2004) *The Stag Hunt and the evolution of social structure*. Cambridge University Press. [AWS]

Smetana, J. (2006) Social-cognitive domain theory: Consistencies and variations in children's moral and social judgments. In: *Handbook of moral development*, ed. M. Killen & J. Smetana, pp. 119–53. Erlbaum. [aPKS, JT]

Sober, E. & Wilson, D. S. (1998) *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press. [AWS]

Southwood, N. (2011) The moral/conventional distinction. *Mind* 120(479):761–802. [EO]

Sperber, D. & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind and Language* 27(5):495–518. Available at: https://doi.org/10.1111/mila.12000. [TH]

Spradley, J. P. & Mann, B. J. (1975) *The cocktail waitress: The woman's work in a man's world*. Wiley. [BV]

Sripada, C. & Stich, S. (2007) A framework for the psychology of norms. In: *The innate mind: Culture and cognition*, ed. P. Carruthers, S. Laurence & S. Stich, pp. 280–301. Oxford University Press. [TD]

Stanford, P. K. (2017) Bending towards justice. *Philosophy of Science* 84:369–78. [aPKS]

Sterelny, K. (2003) *Thought in a hostile world: The evolution of human cognition*. Wiley-Blackwell. [AWS]

Sterelny, K. (2012) *The evolved apprentice: How evolution made humans unique*. MIT Press. [aPKS, JJ]

Stich, S. (forthcoming) The quest for the boundaries of morality. In: *The Routledge handbook of moral epistemology*, ed. K. Jones, M. Timmons & A. Zimmerman. Routledge. [SS]

Struch, N. & Schwartz, S. H. (1989) Intergroup aggression: Its predictors and distinctness from in-group bias. *Journal of Personality and Social Psychology* 56:364–73. [RB]

Suchak, M., Eppley, T. M., Campbell, M. W. & de Waal, F. B. M. (2014) Ape duos and trios: Spontaneous cooperation with free partner choice in chimpanzees. *PeerJ* 2:e417. (Online journal). Available at: http://dx.doi.org/10.7717/peerj.417. [aPKS]

Suchak, M., Eppley, T. M., Campbell, M. W., Feldman, R. A., Quarles, L. F. & de Waal, F. B. M. (2016) How chimpanzees cooperate in a competitive world. *Proceedings of the National Academy of Sciences USA* 113:10215–20. [aPKS]

Theriault, J., Waytz, A., Heiphetz, L. & Young, L. (2017) Examining overlap in behavioral and neural representations of morals, facts, and preferences. *Journal of Experimental Psychology: General* 146(3):305–17. Available at: http://dx.doi.org/10.1037/xge0000350. [JT]

Theriault, J., Waytz, A., Heiphetz, L. & Young, L. (under review) Theory of mind network activity is associated with metaethical judgment: An item analysis. *PsyArXiv*, Available at: http://dx.doi.org/10.17605/OSF.IO/GB5AM. [JT]

Thielmann, I. & Böhm, R. (2016) Who does (not) participate in intergroup conflict? *Social Psychological and Personality Science* 7:778–87. [RB]

Thompson, C., Barresi, J. & Moore, C. (1997) The development of future-oriented prudence and altruism in preschoolers. *Cognitive Development* 12:199–212. [aPKS]

Tisak, M. S. & Turiel, E. (1988) Variation in seriousness of transgressions and children's moral and conventional concepts. *Developmental Psychology* 24:352–57. Available at: http://dx.doi.org/10.1037/0012-1649.24.3.352I. [JT]

Tomasello, M. (2009) *Why we cooperate.* MIT Press. [aPKS, JJ]

Tomasello, M. (2016) *A natural history of human morality*. Harvard University Press. [aPKS, EO]

Tomasello, M., Hare, B., Lehman, H. & Call, J. (2007) Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution* 52:314–20. [aPKS]

Traulsen, A. & Nowak, M. A. (2007) Chromodynamics of cooperation in finite populations. *PLoS ONE* 2(3):e270. Available at: https://doi.org/10.1371/journal.pone.0000270. [TH]

Turiel, E. (1983) *The development of social knowledge: Morality and convention*. Cambridge University Press. [aPKS, SP, JT]

Turiel, E., Killen, M. & Helwig, C. (1987) Morality: Its structure, functions, and vagaries. In: *The emergence of morality in young children*, ed. J. Kagan & S. Lamb, pp. 155–243. University of Chicago Press. [aPKS]

Uhlmann, E. L., Pizarro, D. A. & Diermeier, D. (2015) A person-centered approach to moral judgment. *Perspectives on Psychological Science* 10(1):72–81. Available at: https://doi.org/10.1177/1745691614556679. [SA]

Van Bavel, J. J., Packer, D. J., Haas, I. J. & Cunningham, W. A. (2012) The importance of moral construal: Moral versus non-moral construal elicits faster, more extreme, universal evaluations of the same actions. *PLoS ONE* 7(11):e48693. Available at: https://doi.org/10.1371/journal.pone.0048693. [SA]

Van Cleve, J. & Akçay, E. (2014) Pathways to social evolution: Reciprocity, relatedness, and synergy. *Evolution* 68:2245–58. [TJ]

Van Overwalle, F. (2009) Social cognition and the brain: A meta-analysis. *Human Brain Mapping* 30:829–58. Available at: http://dx.doi.org/10.1002/hbm.20547. [JT]

Voorhees, B., Read, D. & Gabora, L. (2018) Identity, kinship, and the evolution or cooperation. (Preprint, research project paper). Available at: https://www.researchgate.net/project/Identity-Kinship-and-Evolution-of-Cooperation [BV]

Wainryb, C., Shaw, L. S., Langley, M., Cottam, K. & Lewis, R. (2004) Children's thinking about diversity of belief in the early school years: Judgments of relativism, tolerance, and disagreeing persons. *Child Development* 75:687–703. Available at: http://dx.doi.org/10.1111/j.1467-8624.2004.00701.x. [JT]

Warneken, F., Lohse, K., Melis, A. & Tomasello, M. (2011) Young children share the spoils after collaboration. *Psychological Science* 22:267–73. [aPKS]

Warneken, F. & Tomasello, M. (2006) Altruistic helping in human infants and young chimpanzees. *Science* 311:1301–303. [aPKS]

Warneken, F. & Tomasello, M. (2007) Helping and cooperation at 14 months of age. *Infancy* 11:271–94. [aPKS]

Watanabe, J. M. & Smuts, B. B. (1999) Explaining religion without explaining it away: Trust, truth, and the evolution of cooperation in Roy A. Rappaport's "The Obvious Aspects of Ritual." *American Anthropologist* 101(1):98–112. Available at: https://doi.org/10.1525/aa.1999.101.1.98. [SA]

Waytz, A., Young, L. L. & Ginges, J. (2014) Motive attribution asymmetry for love vs. hate drives intractable conflict. *Proceedings of the National Academy of Sciences USA* 111:15687–92. [RB]

Weisel, O. & Böhm, R. (2015) "Ingroup love" and "outgroup hate" in intergroup conflict between natural groups. *Journal of Experimental Social Psychology* 60:110–20. [RB, rPKS]

Wellman, H., Cross, D. & Watson, J. (2001) Meta-analysis of theory-of-mind development: The truth about false-belief. *Child Development* 72:655–84. [AK]

Westfall, J., Kenny, D. A. & Judd, C. M. (2014) Statistical power and optimal design in experiments in which samples of participant respond to samples of stimuli. *Journal of Experimental Psychology: General* 143:2020–45. Available at: http://dx.doi.org/10.1037/xge0000014. [JT]

Wierzbicka, A. (2007) Moral sense. *Journal of Social, Evolutionary, and Cultural Psychology* 1:66–85. [SP]

Wilkins, A. S., Wrangham, R. W. & Fitch, W. T. (2014) The "domestication syndrome" in mammals: A unified explanation based on neural crest cell behavior and genetics. *Genetics* 197(3):795–808. [JJ]

Wittgenstein, L. (1983) *Remarks on the foundations of mathematics*. MIT Press. [TD]

Wrangham, R. (2009) *Catching fire: How cooking made us human*. Basic Books. [JJ]

Wright, J. C., Grandjean, P. T. & McWhite, C. B. (2013) The meta-ethical grounding of our moral beliefs: Evidence for meta-ethical pluralism. *Philosophical Psychology* 26:336–61. [aPKS, TD]

Wright, J. C., McWhite, C. B. & Grandjean, P. T. (2014) The cognitive mechanisms of intolerance: Do our meta-ethical commitments matter? In: *Oxford studies in experimental philosophy, vol. 1*, ed. T. Lombrozo, S. Nichols & J. Knobe, pp. 28–61. Oxford University Press. [aPKS, TD]

Young, L. & Durwin, A. J. (2013) Moral realism as moral motivation: The impact of meta-ethics on everyday decision-making. *Journal of Experimental Social Psychology* 49:302–306. [aPKS]

Zangwill, N. (2014) Aesthetic judgment. In: *The Stanford encyclopedia of philosophy*, Fall 2014 Online edition, ed. E. N. Zalta. Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. Available at: https://plato.stanford.edu/entries/aesthetic-judgment/. [CI-M]