Are there really so many moral emotions? Carving morality at its functional joints

Léo Fitouchi, Jean-Baptiste André, Nicolas Baumard

To appear in L. Al-Shawaf & T. K. Shackelford (Eds.), *The Oxford Handbook of Evolution and the Emotions*. New York: Oxford University Press.

Are there really so many moral emotions? Carving morality at its functional joints

Léo Fitouchi[1*], Jean-Baptiste André[1], Nicolas Baumard[1]

[1] Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL University, CNRS, Paris, France
*Corresponding author
leo.fitouchi@ens.fr, nbaumard@gmail.com, jeanbaptisteandre@gmail.com

Abstract

In recent decades, a large body of work has highlighted the importance of emotional processes in moral cognition. Since then, a heterogeneous bundle of emotions as varied as anger, guilt, shame, contempt, empathy, gratitude, and disgust have been proposed to play an essential role in moral psychology. However, the inclusion of these emotions in the moral domain often lacks a clear functional rationale, generating conflations between merely social and properly moral emotions. Here, we build on (i) evolutionary theories of morality as an adaptation for attracting others' cooperative investments, and on (ii) specifications of the distinctive form and content of moral cognitive representations. On this basis, we argue that only indignation ("moral anger") and guilt can be rigorously characterized as moral emotions, operating on distinctively moral representations. Indignation functions to reclaim benefits to which one is morally entitled, without exceeding the limits of justice. Guilt functions to motivate individuals to compensate their violations of moral contracts. By contrast, other proposed moral emotions (e.g. empathy, shame, disgust) appear only superficially associated with moral cognitive contents and adaptive challenges. Shame doesn't track, by design, the respect of moral obligations, but rather social valuation, the two being not necessarily aligned. Empathy functions to motivate prosocial behavior between interdependent individuals, independently of, and sometimes even in contradiction with the prescriptions of moral intuitions. While disgust is often hypothesized to have acquired a moral role beyond its pathogen-avoidance function, we argue that both evolutionary rationales and psychological evidence for this claim remain inconclusive for now.

Keywords: morality, emotions, evolutionary psychology, anger, outrage, guilt, shame, empathy, moral disgust, purity, fairness, cooperation

1. Introduction: The messy landscape of "moral" emotions

Evolutionary approaches to psychology have proven fruitful in considering emotions as cognitive adaptations, evolved to coordinate the activity of multiple (e.g. physiological, attentional, motivational) systems in the solution of specific adaptive problems (Al-Shawaf et al., 2016; Al-Shawaf & Lewis, 2017). Fear functions to protect organisms from fitness-costly dangers (Cosmides & Tooby, 2000). Sexual jealousy decreases the probability of infidelity (Buss, 2003), and romantic love facilitates pair-bonding (Fletcher et al., 2015). What about morality?

Morality, too, is undoubtedly emotional: we feel outraged in the face of others' immoral acts, and experience genuine guilt about our own moral shortcomings. Accordingly, moral psychology has departed from its "rationalist" origins (Kohlberg & Kramer, 1969; Piaget, 1997/1932) to embrace an "affective" perspective stressing the importance of emotional processes in moral cognition (Greene, 2013; Haidt, 2012, 2001, 2007). In this general movement, a large number of emotions have been proposed to play an essential role in moral cognition, as varied as guilt, anger, empathy, contempt, shame and disgust (Fessler & Haley, 2003; Haidt, 2003; Hutcherson & Gross, 2011; Rozin et al., 1999; Tangney et al., 2007)

This leaves us with an heterogenous bundle of emotions, loosely labelled "moral" because of their co-occurrence with social phenomena we often vaguely refer to as "moral". For example, because it motivates prosocial acts, empathy (or compassion), is often deemed a key component of moral cognition (Flack & De Waal, 2000; Haidt, 2003; Price Tangney et al., 2007). But as many have noted, empathy-driven behavior can be at odds with moral intuitions, e.g. when we unfairly favor people with whom we empathize more (e.g. kins, friends) in situations where they do not morally deserve more than other individuals (Batson & Ahmad, 2009; Baumard, 2016; Bloom, 2017). This may be because empathy tracks something different than our moral obligations toward others, that simply happen to often, but not systematically, lead to behaviors approved by our moral sense (e.g. generosity). Similarly, while shame is often triggered by a moral violation made public, it sometimes motivates people to further violate moral obligations (e.g. by lying, hiding our crime) rather than repairing the original violation that triggered it (see Baumard et al., 2013a). Why? Again, this would easily be explained if shame didn't track, by design, the respect of moral obligations, but something else that simply happened to covary with moral behavior. What if, then, many emotions commonly deemed "moral" in fact serve non-moral functions? And, if so, how are these functions distinct from morality?

These brief examples illustrate the two intertwined problems faced by any attempt to systematize moral emotions:

1. A problem of classification: How to distinguish, in a non-arbitrary and fruitful way, properly moral emotions from merely social ones?
2. A problem of functional specification: What is the specific functional role each emotion plays in moral (or social) cognition?

As researchers have argued, an evolutionary approach to emotions offers an avenue for jointly solving these two problems, as it provides non-arbitrary criteria for classifying emotions, that are precisely based on the evolved functional role each plays in the general cognitive architecture (Al-Shawaf et al., 2016; Al-Shawaf & Lewis, 2017).

The corollary is that a functional view of moral emotions requires an evolutionary and cognitive picture of morality in general. Here, we build on characterization of (i) the specific

cognitive *content* of moral representations, based on the notion of moral *obligation* (Stanford, 2018; Tomasello, 2020), and (ii) the distinctive ultimate *function* of moral representations, rooted in the evolution of human cooperation. We suggest that such specifications allow to more clearly distinguish, among the emotions seemingly involved in human moral life (e.g. empathy, shame, disgust, guilt, anger), those that serve a properly moral function from those that are merely social.

In particular, we make the somewhat deflationary argument that only guilt and indignation ("moral anger") can be rigorously characterized as moral emotion. They indeed, by design, manipulate distinctively moral representations of individuals' *duties* and *rights* in the context of cooperative interactions, and adjust behavior in accordance to what cooperative partners "owe each other". By contrast, shame functions to manipulate representations of one's *social status*, and empathy to motivate prosocial behavior toward individuals with whom one's fitness is *interdependent*, which is different from respecting moral obligations. Finally, while disgust has often been proposed as a generator of, or emotional reaction to moral representations, we argue that both evolutionary rationales and psychological evidence for this claim remain inconclusive for now.

2. Situating emotions in human morality

2.1. What is morality *about*? And what is it *for*?

Obligation, duty, right and wrong: the cognitive specificity of moral representations

In psychological research, morality is often conflated with prosocial, other-regarding preferences and the disapproval of harm or selfishness (e.g. Haidt, 2012; Schein & Gray, 2018). When it comes to moral emotions, this sometimes leads to a definition of moral emotions as what simply motivates prosocial behavior or condemns selfishness. For example, Haidt (2003, p. 853) defines moral emotions as "those emotions that are linked to the interests or welfare either of society as a whole or at least of persons other than the judge or agent."

Similarly, in a more computational framework, moral emotions seem construed, along with other social emotions, as devices that recalibrate people's "Welfare Trade-Off Ratio" (WTR), defined as the value the self places on the welfare of another individual relative to his own, thus determining his disposition for prosocial behavior toward the target (Tooby et al., 2008). Guilt, for example, functions to avoid persistently under-valuing the welfare of another individual who, in fact, makes positive contributions to one's fitness, so that up-regulating one's valuation of her welfare is adaptive (Sznycer, 2019).

While morality is manifestly related to providing others with benefits, and avoiding imposing costs on them, it seems characterized, at the cognitive level, by a distinctive type of mental representation that is not well captured in terms of other-regarding preferences, prosocial desires, or welfare trade-off ratios.

Moral representations are, strictly speaking, not about *desires* or *preferences* – however prosocial they may be – but about *obligations* or *duties*. Specifically, moral representations are *prescriptive* mental states, representing certain behaviors as what people *ought* to do, even if they don't *desire* to do so. In other words, the content of moral representations seems to be a kind of desire-independent, self-imposed *obligation* that one has toward someone else (Tomasello, 2020), that exists in virtue of a higher, preference-independent moral demand (Stanford, 2018). This intuitive notion of obligation represents a precise quantity of benefits that we *ought* to provide others because they *deserve* it, or, equivalently, because they are *rightfully* entitled to it. A common set of terms in natural moral language points towards this precise quantity of benefits constituting

the content of moral obligations: people *deserve* X, i.e. I *owe* them X, i.e. they have a *right* to X, i.e. I have the *duty* to provide them X.

And this kind of mental representation has the puzzling specificity of depicting these benefit-providing (or cost-imposition-avoiding) obligations independently of my immediate incentives or my current bargaining position: even if I don't want to give them X, because, e.g., I am not immediately incentivized to do so, the obligation to do so remains. Even if I actually *refuse* to provide these benefits, the obligation remains: I simply *violated* it, and my behavior is accordingly tagged as morally *wrong* (see Darwall, 2010). In other words, as philosophers have long noted, behaving in conformity to moral obligations entails behaving "as if" we respect the terms of an implicit contract previously agreed upon with others (Gauthier, 1986; Rawls, 2001; Scanlon, 2000)

Importantly, morality so understood – i.e. as the cognitive calculation of, and behavioral conformity to, "what we owe to each other" (Scanlon, 2000) – is different from the mere motivation for being prosocial, the latter of which can emerge from psychological mechanisms whose function does not involve moral obligations. We can *desire* to provide benefits to some individuals (e.g. kin, friends), without this being aligned with moral obligations, for example in the case of morally condemned nepotism (see Baumard, 2016; Boehm, 2012; Vollan et al., 2020).

Reputation and the evolution of moral contracts

Where, then, do these specific representations of moral obligations ultimately come from? What fitness-relevant regularities do they encode?

Researchers have long argued that morality likely evolved as a cognitive adaptation to the challenges of cooperation recurrent in human social life (Alexander, 1987; Baumard et al., 2013b; Boehm, 2012; Curry et al., 2019; Haidt, 2012; Tomasello, 2016; Trivers, 1971). Many have proposed that the ultimate function of moral behavior is to secure a good reputation as a cooperator, in order to attract cooperative partners (Alexander, 1987; Baumard et al., 2013b; André & Baumard, 2011; Debove et al., 2015a, 2015b), or to avoid the costs of the punishment and social control imposed on uncooperative individuals (Boehm, 2012; Wrangham, 2019).

Integrating the above logics, a more general way of phrasing it is the following (see André et al., 2021 for a longer argument and formalization). In a positive-sum world offering the opportunity of many beneficial cooperative interactions (e.g. in resource production, collective defense, parental investment, limitations of interpersonal conflict, communication of reliable information, etc.), individuals face two important selective pressures. The first is to ensure that I invest only in cooperative interactions that *pay off*, i.e. that provide me more benefits than they cost (including opportunity costs) – otherwise, I would have been better off doing something else (e.g. forage alone, choose other partners, extort benefits by force).

The second selective pressure results from the first: because others only invest in cooperative interactions that pay off, in order to *attract* their cooperation, I have to ensure that cooperation pays more than it costs for them as well. As with any investment, the costliness of a cooperative investment depends critically on its opportunity costs. This means that I have to make sure that, for my partners, investing in a cooperative interaction with me pays more than *the best alternative option* they could have adopted instead of cooperating with me. If it does not, people are simply better off doing something else. And this includes *lots* of alternative options they could have adopted instead of cooperating with me, e.g. not cooperating at all (e.g. forage alone), defecting in the same interaction (Axelrod & Hamilton, 1981; Trivers, 1971), choosing other partners (Baumard et al., 2013b), ostracizing me (Boehm, 1999) or extorting benefits by force at a cost to me (Boehm, 1999, 2012; Clutton-Brock & Parker, 1995; Wrangham, 2019).

As a result, these two symmetrical selective pressures constrain cooperative interactions to be *mutually beneficial* — as if regulated by a *contract* agreed upon with others: in order to enjoy the benefits generated by cooperative surpluses, I should both ensure than my cooperative investments are more beneficial than they are costly for me, *and simultaneously* ensure that my partners, too, do not afterward "regret" having cooperated with me, i.e. I must ensure that cooperation offers a net benefit for them, too.

A likely possibility is thus that, in the human mind, a dedicated information-processing mechanism flexibly calculates, based on the costs (including opportunity costs) invested by each partner in each particular cooperative interaction, the *rights* and *duties* of each individual (i.e. the "terms of the contract"). The resulting representations should specify what I *deserve*, but also what I *owe* to others. In other words, they should specify, as moral representations actually do, "what we owe to each other" (Scanlon, 2000) for cooperation to be mutually beneficial. Moreover, even if people don't have an immediate incentive to honor these duties, e.g. because they enjoy a temporary strategic advantage on a resource, they should, in order to further attract others' cooperative investments, feel *obligated* to do so by a strange demanding force. In other words, seems to lead to the specific and distinctive phenomenology of moral obligation (see Tomasello, 2020).

This enables to give a precise functional content to moral obligations, beyond vague formulations such as "behaving cooperatively", "being fair", "not being selfish" or "not harming others". The precise amount of benefits morally owed to others should at least compensate the total cost they invested in the cooperative interaction, including the opportunity cost of having renounced their best alternative behavioral option.

And this seems to fit the actual way humans compute moral obligations. For example, children and adults intuit, across cultures, that the amount of benefits owed to others is proportional to the effort they invested in the cooperative interaction (i.e. direct costs), and their talent or skills (i.e. opportunity costs) (Baumard et al., 2011, 2013b; Liénard et al., 2013; Starmans et al., 2017; see André et al., 2021; Baumard & Sheskin, 2015, for a broader review).

Relatedly, contrary to what researchers have sometimes suggested (Curry, Jones Chesters, et al., 2019; Haidt, 2012) this fairness-based logic of moral obligation seems not just a particular "foundation" of morality, applying only to a restricted domain of social life (see André et al., 2021; Baumard et al., 2013, for a longer discussion). Rather, it underlies the logic of intuitions about moral obligations across the so-called "domains" of morality: most of the moral wrongs condemned across cultures appear as unfair advancements of one's interest at the expense of mutual net benefit. For example, adultery, in the context of the cooperative interaction constituting pair-bonding (Gurven et al., 2009), amounts to enjoy the benefits of one's partner reproductive resources and parental investment, while not repaying, by oneself remaining faithful and parentally investing, the opportunity costs they paid through their fidelity. Betrayal of one's in-group amounts to failing to repay to others, through the benefits brought by my own loyalty, the costs they have paid by not betraying me, i.e. by forgoing the opportunity of cooperation with rival coalitions. Regarding duties of deference to authority, recent research suggest that leadership is fundamentally about a mutually beneficial division of labor in which leaders provide computational, decision-making services to followers (Hagen & Garfield, 2019). Accordingly, authority appears as morally *legitimate* only when it works for the common good: if leaders take advantage of followers' obedience to selfishly advance their interests at others' expense, people do not feel any duty to obey them (Boehm, 1999; Vollan et al., 2020), and withdraw their cooperative investment from the relationship (e.g. not obeying, alternative leader choice, reputational sanctions or punitive aggression; see Vollan & al., 2020; Hagen & Garfield, 2019; Boehm, 1999).

2.2. Moral emotions in this context

This rapid detour through the evolution of morality allows us to more clearly articulate the role of emotional systems in the general economy of moral cognition.

What do moral emotions do?

A first question pertains to the role of emotions in general in moral cognition. It has often be argued that emotions have a *constitutive* role in morality, i.e. that some behaviors are intuited as morally right or wrong *because* of the emotional processes accompanying their perception (Frank, 1988; Greene, 2013; Haidt, 2001; Nichols, 2002).

The very existence of moral emotions, however, seems dependent on a capacity for moral *judgement* in the first place. What makes an emotion moral is the moral character of the representational content it manipulates: Indignation is anger *at injustice*; guilt is a regret *of one's immoral behavior*. In other words, something must tag a behavior as unjust or immoral for "moral" emotions to be triggered in the first place. In line with this idea, the overall empirical evidence for an essential role of emotions in moral judgement is weak (McAuliffe, 2019; Avramova & Inbar, 2013; Landy & Goodwin, 2015).

Accordingly, we posit that there is a cognitive system that first generates moral representations, which moral emotions then *process* rather than secrete in the first place (see Figure 1). In the evolutionary perspective sketched above, this system of moral judgement (the "moral sense") evolved as a barometer of what each partner *owes* and *deserves* in the context of cooperative interactions. It calculates, based on the costs (including opportunity costs) invested by each partner in entering the interaction, what each should receive for cooperation to provide a mutual net benefit.

This is where moral emotions enter the picture: these representations of moral rights and duties, generated by the moral sense, must then be processed by systems orchestrating physiology, cognition and behavior adaptively in accordance with the specific information conveyed by moral representations — i.e., by moral emotions (see Figure 1).

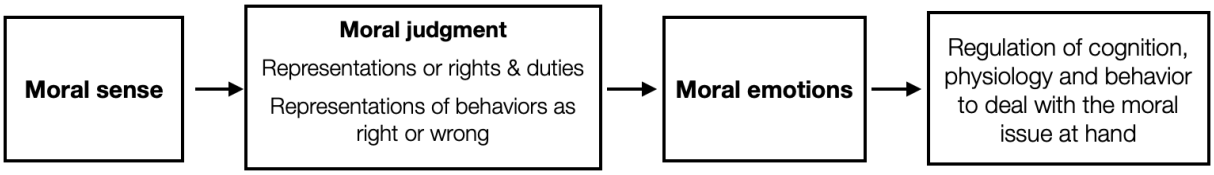

Figure 1. Hypothesized general structure of moral cognition. In this perspective, the moral emotions are not the cause of moral representations. Moral representations are instead generated by a mechanism of moral judgment (the "moral sense"), evolved to calculate what cooperative partners "owe each other" in the context of cooperative interactions. The function of moral emotions is then to orchestrate cognition, physiology and behavior in accordance with the content of these moral representations.

Which emotions serve a moral function?

This functional definition of moral emotions provides a principled way to distinguish between properly moral and merely social emotions. In the remainder of this chapter, we argue that only indignation and guilt satisfy the above conditions. While shame has often been considered, alongside guilt, as a key self-conscious moral emotion, only guilt appears to manipulate distinctively moral representations of rights and duties, to adjust behavior to fit moral contracts (Sect. 3). We then consider the case of empathy, arguing that its interdependence-based prosocial motivation operates independently of moral representations, excluding it from a strictly defined set of moral emotions (Sect. 4.). We finally turn to other-condemning emotions, arguing that only indignation ("moral anger"), and not disgust, function to orchestrate cognition and behavior in the face of others' moral violations (Sect 5).

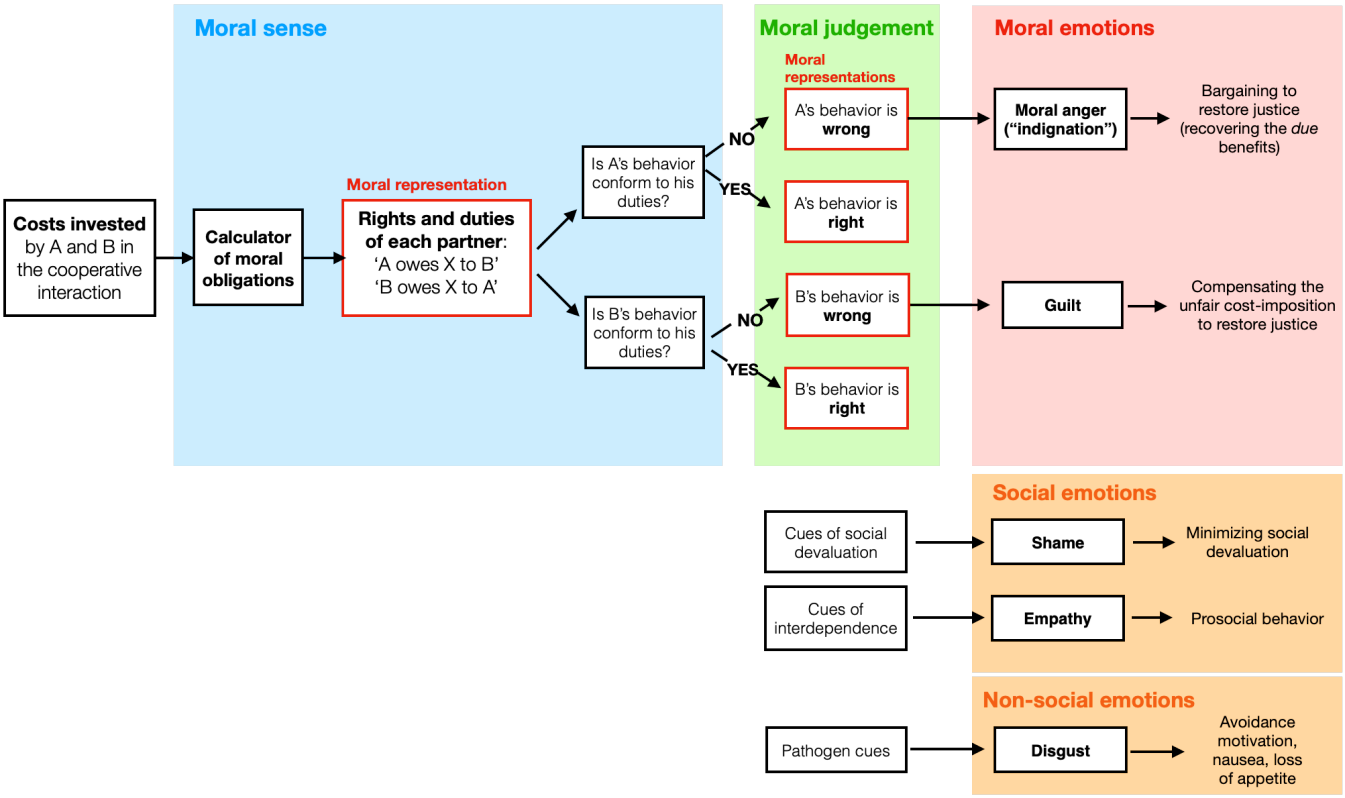

Figure 2. Distinction between moral and non-moral emotions based on the different types of representation they process. Moral emotions process moral representations, produced by the moral sense. The moral sense calculates the rights and duties of each cooperative partner, on the basis of the costs each has invested in the cooperative interaction. These costs must be reimbursed for cooperation to be mutually net beneficial. If A provides B with fewer benefits than he owes her, his behavior is tagged as morally wrong, and triggers in B the emotion of moral anger. Moral anger orchestrates physiology, cognition and behavior to bargain with A, in order to regain the benefits to which she is rightfully entitled. If B provides A with less benefits than she owes him, she feels guilt. Guilt functions to motivate B to provide A with more benefits, in order to compensate the gap between her initial behavior and what she morally owed to A. By contrast, non-moral emotions such as shame, empathy or disgust do not compute distinctively moral representations, but rather, respectively, cues of social devaluation, interdependence, or pathogenic content.

3. Protecting one's social status vs. doing one's duty: Guilt (not shame) as the main self-conscious moral emotion

An often-mentioned class of moral emotions is the *self-conscious* emotions, i.e. emotions reacting to moral violations that one has committed. Shame and guilt have been widely considered as the key self-conscious moral emotions (e.g. Haidt, 2003; Prinz & Nichols, 2010; Sheikh & Janoff-Bulman, 2010; Tangney et al., 2007). However, while guilt indeed seems to compute distinctively moral representations of obligations, by disincentivizing (*ex ante*), and motivating to compensate (*ex post*), the violation of a moral obligation, shame serves a different function. Shame works prevent losses of social status (or "social devaluation") (Sznycer, 2019), a function that does not necessarily overlap with, and sometimes even directly contradicts, moral contracts. As a result, we suggest that only guilt, and not shame, should be considered a properly, functionally defined moral emotion.

In many species, including humans, status hierarchies define individuals' relative access to contested resources (e.g. mates, food), so that gaining and avoiding status losses is a key adaptive problem (von Rueden et al., 2011; von Rueden & Jaeggi, 2016). Traits contributing to social status mainly pertain to abilities to inflict costs (e.g. physical or coalition-derived formidability) and to confer benefits (e.g. competence, intelligence, knowledge, attractiveness, generosity; Durkee et al., 2019; von Rueden & Jaeggi, 2016). In this context, adaptations for avoiding losses of social status appear particularly beneficial to secure the benefits contingent on one's social value. Recent work strongly suggests that shame precisely functions to limit social devaluation by others, by disincentivizing behaviors leading to social devaluation, limiting the spread of socially devaluing information, and mitigating the costs of status loss (e.g. by motivating hiding from others, denial of the socially devaluing action, or displays of remorse or submission) (Durkee et al., 2019; Sznycer et al., 2016, 2018)

This social function is different from the properly moral function of guilt, as suggested by their respective cognitive features.

First, avoiding status devaluation requires being sensitive to others' perceptions of many more traits of the self than merely moral ones. The disposition to repay others' cooperative investments, as component of the willingness to confer them benefits, is indeed only one trait, *among many non-moral others*, that brings social status. In line with this idea, guilt is mainly triggered by moral transgressions, but not by non-moral threats to one's social status (e.g. incompetence, unattractiveness) (Smith et al., 2002). By contrast, in line with a status-management, non-moral function of shame, shame tracks *all types* of threats of social devaluation across cultures: it is typically activated by behaviors that are morally irrelevant yet indicate a low potential to impose costs on or benefit others, e.g. being physically weak, unattractive, incompetent, dumb, not socially influential (Durkee et al., 2019; Sznycer et al., 2016; Tracy & Matsumoto, 2008).

Second, an important reason for the confusion over the moral status of shame is that shame and guilt sometimes co-occur. Acting immorally (which elicits guilt) is one path, among others, to social devaluation (which elicits shame). As a result, immoral actions often generate not only guilt, but also shame, leading to the understandable inference that shame is a moral emotion. But a crucial point here is that *even in these cases, shame is sensitive to what is socially devaluing in immoral actions, whereas guilt is sensitive to what is properly immoral in immoral actions*, namely the gap between what one owed to others and the benefits one actually provided them (or the costs one imposed on them).

In other words, shame fundamentally processes representations of one's *social value* in the eyes of others, rather than properly moral representations of one's *obligations* toward others. In line

with this idea, experimental evidence shows that wrongdoing in itself (i.e. a gap between one's moral obligations and one's actual behavior) is neither necessary nor sufficient to elicit shame: the simple detection that others *falsely believe* that we behaved immorally, when we know we are innocent, is sufficient to elicit shame (Robertson et al., 2018). And conversely, low contributions to a public good on one's part do not independently predict the intensity of shame: only cues that others have devalued us (e.g. through social exclusion) do (Robertson et al., 2018).

Third, regarding their respective outputs, consistent with a properly moral function of guilt, guilt-proneness predicts a greater likelihood of honoring cooperative obligations (Cohen et al., 2013) and a lower likelihood of criminal recidivism (Tangney et al., 2014). After a moral failure, guilt motivates reparative actions, such as apology, confession, acceptance of responsibility and compensation for the harm done (de Hooge et al., 2007, 2011; Ketelaar & Au, 2003; June Price Tangney et al., 2007). These outcomes are all well-suited to reduce the gap between one's actual behavior and what other deserve. By contrast, shame is only contingently associated to cooperative behavior (de Hooge et al., 2007) and can even lead to immoral behaviors, such as aggression, lying or hiding our crime (e.g. Elison et al., 2014; Gausel et al., 2016). These behaviors are geared toward minimizing status loss, not toward reducing the gap between what we owe others and what we actually offered them.

And this, importantly, seems to be a feature rather than a bug, due to the following logic. There are indeed several paths to social status. Sometimes, when we depend on others' cooperation, rehabilitating one's social value can be done through cooperative behaviors (e.g., compensation, apologies, submissive displays), up-regulating others' valuation of ourselves *as a cooperative partner*. Accordingly, in such cases, shame seems to motivate compensation and cooperation (de Hooge et al., 2008; Hooge et al., 2010; Tangney et al., 2014). In other cases, however, the most effective way to regain social status will be the competitive bargaining, cost-imposing route. Accordingly, in these cases, shame adaptively switches motivations toward competitive tendencies (e.g. anger; Sell, 2017), incentivizing others to value one's welfare through aggressive means signaling one's bargaining power (see e.g. Elison et al., 2014).

In other words, even when shame leads us to conform to moral obligations, it does this not because it tracks moral obligations per se, but because it tracks one's social value, which *sometimes happens to be conditioned on one's disposition to act morally*. Stated differently, when it motivates prosocial behavior, shame seems to do so not out of a genuinely moral sense of obligation, as guilt does ("It was my duty to do X, and I failed"), but out of an extrinsic, Machiavellian motivation for the social valuation that the moral behavior will bring ("I should do X because otherwise others won't respect me") (Sperber & Baumard, 2012).

In sum, while guilt and shame are widely considered the two key moral self-conscious emotions, it seems that only guilt, and not shame, operates on distinctively moral representations of obligations (see Figure 2). Guilt, but not shame, orchestrates cognition and behavior in the solution of a distinctively moral challenge – compensating one's violations of others' rights.

4. Disentangling prosocial motivation from moral obligation: Empathy as a prosocial yet non-moral emotion

By empathy, we are referring to the putatively moral emotion sometimes called "empathic concern" (Batson & Ahmad, 2009), "compassion" (Haidt, 2003) or "sympathy" (Smith, 1759), referring to a sensibility to others' suffering, associated with an urge to care for them.

As it vicariously motivates people to benefit others (Batson, 2017), empathy so conceived has often been considered as a main emotional foundation of moral cognition. Psychologists often

consider it a psychological cornerstone of human morality (Graham et al., 2013; Haidt, 2003; Tangney et al., 2007), and ethologists in search of precursors of morality in non-human primates often point to their ability to empathize (Flack & de Waal, 2000).

We suggest, however, that a careful examination of both the cognitive features of empathy and its likely evolutionary function seem to exclude it from a rigorously defined set of moral emotions.

At the proximate level, empathy operates independently of moral representations

At the proximate level, the apparently moral character of empathy comes from the frequent yet superficial co-occurrence of empathy and moral duties to benefit others. For example, in the face of undeserved suffering, empathy will motivate prosocial actions aimed at limiting these unfair costs imposed on an individual.

However, as many scholars have noted, empathy-driven altruism often clashes with human moral intuitions (Batson & Ahmad, 2009; Baumard, 2016; Bloom, 2017; A. Smith, 1822). In particular, empathy-induced altruism can lead to actions judged as immoral by introducing unfair *partiality*. For example, participants who were induced to feel empathy for a terminally ill child were more likely to give him priority in the allocation of end-of-life care over children who needed such care more urgently. And importantly, they themselves judged that this decision was less fair and less moral than allocation decisions not biased by their empathy-induced altruism (Batson et al., 1995; for similar results in other settings, see Batson et al., 1999; Batson & Ahmad, 2009). Such dissociations suggest that empathy-driven prosocial motivation is independent, at the cognitive level, from the calculations produced by people's moral sense.

Relatedly, in economic games where participants are asked to help others by compensating for their bad outcomes, empathic concern is a better predictor than moral outrage of helping behaviors directed toward individuals who did *not* suffer an injustice. By contrast, in conditions where the target's bad outcome is due to an injustice (e.g. a partners' refusal to reciprocate), and is thus *undeserved*, it is moral outrage that is a better predictor than empathic concern of directing help toward the cheated person (Thulin & Bicchieri, 2016). In other words, empathy appears here again as sensitive to individuals' suffering regardless of the *deservingness* of that suffering, i.e. regardless of representations of rights and duties generated by the moral sense. By contrast, the above example suggests that moral outrage is sensitive to individuals' suffering only when that suffering is undeserved.

Empathy thus appears functionally different from properly moral emotions, in the sense that it does not operate on moral representational content: it motivates prosocial behavior independently of, and sometimes even *contrary* to the representation of moral rights and obligations generated by the moral sense. The perception of signs of suffering and distress in other individuals seems to directly generate an urge to care for their welfare, without resorting to the representation of a moral duty to do so, underpinned by the intuition that they *deserve* this help (see Figure 2).

But if empathy is not a moral emotion, what is it for?

Empathy and fitness interdependence

Many scholars have argued that the evolutionary origins of empathy lie in the parental care of offspring based in kin selection (Decety et al., 2016; Tomasello, 2016). In some species, especially humans, empathic concern extends beyond the circle of genetic relatives to friends and collaborative partners (de Waal, 2008). This suggests that empathy evolved not only for kin-

altruism, but for the more general adaptive challenge of fitness interdependence, of which genetic relatedness is just a particular instance (Tomasello, 2016).

Two organisms are (positively) interdependent if an increase in the fitness of one generates an increase in the others' fitness (Aktipis et al., 2018; Roberts, 2005). In such a context, individuals have a "stake" in the fitness of their partners, which makes cooperative behaviors adaptive when their costs are outweighed by the cooperator's stake in the recipient's benefits (Roberts, 2005). A particular way of generating fitness interdependence is genetic relatedness. But on top of that, fruitful collective actions between non-kin generates another type of interdependence: if A and B are, say, hunting partners, and B breaks her leg, A has an interest in helping B to recover quickly, as his fitness depends on B's efficiency in their future cooperative interactions. As a result, it is advantageous for emotions to promote caring for the welfare of individuals with whom one is interdependent, as caring for their welfare leads to increases in one's own fitness.

Several lines of evidence suggest that empathy is one such mechanism. The intensity of empathic feelings is indeed typically modulated by cues of fitness interdependence. In humans and other primates, empathy is amplified by familiarity and social closeness (Preston & de Waal, 2002). For example, children display more empathy-related behaviors toward their mother than toward an unfamiliar individual, and feel more empathy toward in-group rather than out-group members (Davidov et al., 2013; Masten et al., 2010). Activity in the pain neural network is enhanced when individuals view or imagine their loved ones in pain compared with strangers (Cikara et al., 2011). And studies report a modulation of empathic response as a function of racial group membership (Xu et al., 2009), which the mind may consider a proxy of coalitional affiliation (Kurzban et al., 2001).

Disentangling morality from fitness interdependence

How is interdependence-based empathy different from the adaptive challenges linked to moral rights and duties, to *obligations* to benefit others and ideas of *deservingness* of receiving benefits?

Crucially, morality functions to ensure the mutually beneficial character of cooperation when the latter is not guaranteed, i.e. when individuals still have a short-term incentive to cheat by benefiting from others' cooperative investment without repaying it in the future – hence the need for mental representations encoding the terms of an implicit "contract". In other words, representations of moral rights and obligations function to regulate interactions in which I benefit from cooperation only *conditionally on my partners' response* to my cooperative behavior. They regulate interactions in which reaping the cooperative surpluses requires the individual to willingly temporarily weaken his strategic position, by putting himself in a situation of vulnerability to exploitation, and trust that his partner will not succumb to the short-term temptation to take advantage of this vulnerability – i.e. trust that the partner will respect the implicit "contract".

By contrast, the fitness benefits of interdependence-based prosocial behavior are *not* conditioned on the recipient's responses to my cooperative behavior: When individuals protect their mates, children or friends — at least if they do so in proportion to the fitness-stake they have in them— they *automatically* benefit from this prosocial behavior. As there is no short-term incentive to cheat, there is no need for representations encoding a morally legitimate quantity of benefits that each individual should receive, i.e. no need for representations about what each individual "deserves" or "owes". Relatedly, empathy does not feel like a "constraint" imposed on us, demanding to go against our selfish will: instead, it feels like a spontaneous urge to help (see Tomasello, 2020)

As an illustration, consider two individuals, A (a female) and B (a male), who are in a committed long-term pair-bond. Generally, their level of fitness interdependence is high: If A is hurt, B does not benefit by letting his partner incur damage to her embodied capital on which his own reproductive success depends. Accordingly, high levels of empathy-driven spontaneous prosociality will motivate him to care for her welfare. Generally speaking, the fitness payoff of B's helping is to a large extent *not conditioned* on A's response to the benefits he provides her – fitness interdependence does the job.

By contrast, in other respects, A and B's interests are not totally aligned. In particular, they both have a short-term interest to cheat their partner by engaging in extra-pair mating. By mutually guaranteeing sexual fidelity to each other, they both pay the short-term opportunity costs of forgoing alternative, extra-pair reproductive encounters (see Gangestad & Simpson, 2000). They do this in order to reap the larger, future benefits of sustained cooperation in committed pair-bonds (see Gurven et al., 2009). Yet, in order to reap this long-term cooperative benefit by remaining faithful, they place themselves in a situation of vulnerability to exploitation: the benefits of B's sexual fidelity are largely conditioned on A repaying this investment by also not cheating. We suggest that it is precisely this kind of "cooperation dilemma" (Rand & Nowak, 2013) that the moral sense is designed to deal with (see André et al., 2021). It does so by representing individuals' moral *rights* (here, to fidelity), that individuals *deserve* because of the costs they invested for the *common good* (here, the long-term mutual benefits of committed pair-bonds), that other individuals thus have a *duty* to honor, and that it would be morally *wrong* to violate (here, through adultery).

This allows us to grasp why emotions promoting interdependence-based prosociality (e.g. empathy) and properly moral emotions are functionally distinct mechanisms. The function of respecting moral obligations is not to benefit others to the extent that their welfare immediately makes positive contributions to *my* fitness, but rather to the extent necessary to make *them* better off cooperating with me rather than doing something else (e.g. defecting, choosing another partner, using a power struggle to their benefit). By respecting my partners' rights in such a way, I am both incentivizing them to continue investing in our cooperative relationship, and securing a good moral reputation attracting cooperative investments from other partners.

Accordingly, interdependence-based prosociality and cooperative behavior out of moral obligation do not necessarily overlap: In the same way as my personal selfish interest can conflict with moral obligations, my interdependence-mediated "selfish" desires to see my kin and friends favored over equally deserving individuals can conflict with moral obligations (see Batson & Ahmad, 2009).

5. Moral indignation (not disgust) as the main other-condemning moral emotion

5.1. Enacting justice: Moral indignation is for enforcing moral contracts

Whereas morality is about the mutually beneficial management of cooperative interactions, anger probably initially evolved in the context of competitive, zero-sum interactions, as a bargaining mechanism for deterring future aggression through retaliation (Fessler, 2010; McCullough et al., 2013) or resolving conflicts of interests in favor of the angry individual, by coercing others to give the angry person more benefits (Sell et al., 2009; Aaron Sell, 2017).

Anger is indeed typically triggered by an insufficient delivery of benefits to the self *relative to its bargaining power*: Across cultures, anger's intensity is predicted by the perception that the target does not value the angry individual's welfare enough relative to his own (Sell et al., 2017)

and individuals with greater bargaining power (e.g. physical strength, coalition support, attractiveness), are more prone to anger (Sell et al., 2009). Its outputs typically instantiate two bargaining tactics: threatening to inflict costs (e.g., through physical or coalition-derived formidability) or withhold benefits (Fessler, 2010; Sell et al., 2009; Sell et al., 2017)

In the human species, a peculiar form of anger appears in moral life, variously called "moral outrage", "indignation" or "righteous anger". Accordingly, anger has widely been considered one of the main moral emotions (Fessler & Haley, 2003; Haidt, 2003; Hutcherson & Gross, 2011; Rozin et al., 1999; Tangney et al., 2007). A likely possibility is that moral indignation evolved by recycling useful design features of "competitive" anger for use in moral situations in which considerations of people's rights and duties in cooperative interactions are paramount.

Evolving moral indignation from competitive anger

In cooperative interactions, there is always the possibility that cooperative partners will not respect cooperative obligations, i.e. my partner might provide me with fewer benefits than the costs I invested (including opportunity costs) in the collaborative interaction. This would violate the requirement of mutual beneficence. In this situation, an emotion motivating the recruitment of bargaining mechanisms to regain the deserved-but-missing benefits would be useful, and there is manifest fit between this requirement and the bargaining-oriented design of anger.

An emotion solving this moral problem should be slightly distinct from "raw" competitive anger with respect to some of its computational features. Regarding its triggers, it should be sensitive to the *wrong* or *injustice* that I have suffered, i.e. to the gap between benefits actually provided by others and the benefits they morally *owe* me. This should correspond, at the ultimate level, to the quantity of benefits I should receive to repay the costs I invested in the cooperative interaction. This "fairness gap" is distinct from the gap to which "competitive anger" is sensitive, namely the gap between the benefits actually received and the benefits that my *bargaining power* (e.g. formidability) could allow me to *extort* from others by brute power struggle or cost infliction. The two emotions should also have distinct outputs: competitive anger motivates me to obtain as many benefits as I can within the constraints of my ability to impose costs on others. By contrast, moral indignation should only motivate me to take back the *limited* quantity of benefits that I morally deserve and have been denied. Demanding more than these due benefits would lead *me* to appear as a cheater, with negative consequences for my reputation.

However, in existing empirical studies, it is difficult to precisely distinguish indignation as a moral emotion from competitive anger functioning for raw power struggle (i.e. without concern for partners' rights). Economic games experiments, for example, consistently report "anger" as a key emotional response to unfair distributions of benefits, motivating the punishment of free-riders (Dawes et al., 2007; Fehr & Gächter, 2002; Molleman et al., 2019; Nelissen & Zeelenberg, 2010). But it is unclear if such negative emotional reactions to unfairness emerge from a moral motivation to restore justice, or instead from non-moral, competitive retaliation simply aimed at deterring future cost infliction (McCullough et al., 2013; O'Mara Kunz et al., 2011).

Still, some evidence suggests that anger operates on moral representations of rights and duties in the context of cooperative interactions, and at least partly seeks the satisfaction of the moral rights of the harmed individual, rather than a raw retaliation without concern for what each party deserves.

First, in developmental studies where a child takes more than his fair share, or inefficiently plays his role in a cooperative interaction, other children's resentful protest is expressed through the *normative* language typically deriving from moral representations (e.g. "One must do X", "It's not fair"), rather than in terms of personal preferences or desires (e.g. "I don't like when you do

X") (Kachel et al., 2017; Rakoczy et al., 2016). In other words, children's protest seem aimed at "mak[ing] (the partner) sensible, that the person whom he injured did not deserve to be treated in that manner" (Engelmann & Tomasello, 2019, p. 458, quoting Adam Smith, 1759, pp. 95-96). Moreover, this does not seem to trigger a competitive dynamic: the wronged child then trusts her partner to decide to do the right thing, which he often does by re-equalizing the payoffs (Engelmann & Tomasello, 2019).

Second, people's intuitions about punishment, emotionally underpinned by indignation, include a strong requirement of retributive *justice,* that fits the design expected from a system functioning to ensure a fair distribution of the costs and benefits between cooperative partners. When assigning punishment, people want it to conform to "*just desert*": the costs imposed on the culprit should be *proportionate* to the harm done to the victim, in order to restore a fair balance of the interests between individuals (Baumard, 2010; Baumard et al., 2013b; Carlsmith & Darley, 2008; Darley & Pittman, 2003; Osgood, 2017). Tellingly, the developmental trajectory of childrens' retributive justice intuitions parallels the trajectory of their symmetrical *distributive* justice intuitions, prescribing the fair way to share collectively produced goods (Smith & Warneken, 2016), which are clearly impregnated with moral intuitions about what each partner deserves (Corbit et al., 2017; Engelmann & Tomasello, 2019).

Is there such a thing as third-party moral outrage?

A third key element put forward for the moral character of anger is its capacity to be triggered by a moral violation of which one is not the direct victim – often called "third party moral outrage" (e.g. Haidt, 2003; Haley & Fessler, 2003; Tangney et al., 2007). Various functions have been ascribed to this emotional reaction and the "third-party" punishment that flows from it, such as an evolutionarily altruistic enforcement of cooperation favored by group selection (Henrich & Boyd, 2001) or individual-level reputational fitness benefits from a credible signal of one's cooperative quality (Barclay, 2005; Jordan et al., 2016).

While the existence of such a moral emotion is often taken for granted, we note that it is unclear if there really is such a thing as a truly *third*-party moral outrage, i.e. an anger-like emotional reaction to an immoral action whose cost to oneself is *really totally zero.* Surely, there exists such a thing as a moral *judgment* of actions that do not affect us at all, i.e., a representation of that action as violating a moral obligation. The function of such a third-party moral judgement would likely be to encode that future cooperation with the moral violator should be avoided. But are the bargaining-oriented physiological, cognitive and behavioral mechanisms of anger really triggered by third-party moral violations?

Ethnographers have long noted that in small-scale societies, in the context of which human cognition is often assumed to have evolved, third-parties often appear indifferent to moral violations toward an unrelated individual: when outrage and punishment are directed toward moral violators, it is typically administered by the aggrieved parties themselves (Baumard, 2010; Berndt, 1988; Black, 2000; Evans-Pritchard, 1940; Wiessner, 2005).

A first possibility is that what sometimes appears as third-party outrage is in fact only anger at costs imposed on an individual with whom one is interdependent (e.g. kin, friends), so that the fitness cost toward oneself is real, albeit indirect. In line with this idea, psychological evidence from modern populations suggests that people do not punish and feel limited outrage toward violations harming individuals with whom they are not interdependent (Pedersen et al., 2018, 2019), or with whom they have not been induced to empathize (Batson et al., 2007).

Another, compatible possibility is that outraged individuals often misleadingly appear as third parties when in fact they are actually second parties, involved in a larger cooperative

interaction in which the moral violation took place. For example, a moral violation such as knocking down a pedestrian in the street can misleadingly appear as wronging only one person (this particular pedestrian), while it probably also constitutes a cheating behavior in a more general contract involving all members of the society. There is indeed manifestly, in modern societies, a collective action in which people cooperate by paying attention to pedestrians *in general* while driving, so that everyone mutually benefits from safely walking in the street. In this context, injuring a pedestrian because of unsafe driving amounts to violating a more general moral obligation toward all other members of the society, potentially explaining the moral outrage of apparently "third"- but in fact second-parties. In line with this idea, when people are collectively outraged and punish deviants in small-scale and tribal societies, they do so because the target's behavior, even if apparently harming only some individuals, is perceived as dangerous or harmful for themselves too (Baumard & Liénard, 2011; Boehm, 1999, 2012).

5.2. Disgust is (probably) not a moral emotion

In recent decades, disgust has received enormous attention in moral psychology. The idea that disgust could be a moral emotion emerged from seminal studies in which participants were found to morally condemn harmless yet disgusting actions, such as masturbating in a dead chicken before eating it (e.g., Haidt, 2001; Haidt et al., 1993; Haidt & Hersh, 2001). This generated investigations of two distinct ways by which disgust could be a moral emotion.

The first hypothesis proposes that disgust is a *moralizing* emotion, i.e. that the emotional experience of disgust *itself generates* moral representations of right and wrong. In other words, some actions would be intuited as morally wrong *because* they are disgusting. This hypothesis is often intertwined with another one, according to which disgust is responsible for moral judgements of a specific part of the moral domain, often related to sexuality, purity and piety (Graham et al., 2013; Haidt, 2012; Rozin et al., 1999).

The second, weaker hypothesis is that disgust does not cause moral condemnation, but is *elicited* by moral violations in general – just as, for example, moral indignation is (Hanah A. Chapman & Anderson, 2013; Hutcherson & Gross, 2011). This weaker and more probable claim is generally associated to another one, according to which this role of disgust in moral cognition is not restricted to a particular domain of morality (e.g. purity), but is observed for violations across the entire moral domain (Hanah A. Chapman & Anderson, 2014; Molho et al., 2017).

In the following, we consider these two families of hypotheses.

Disgust is not a *moralizing* emotion

From an evolutionary-functional perspective, it is unclear why merely perceiving an action as disgusting should generate a representation of this action as morally wrong. Indeed, pathogen-related behaviors have no reason to be represented as morally bad if they don't unfairly harm, in some way, the interests of cooperative partners. And indeed, in many cases, disgusting behaviors (e.g. picking one's nose in private), when harmless, are simply disgusting, and not immoral (Pizarro et al., 2011).

Rather than generating moral representations, disgust might however play the less important role of an *input* to the moral sense. In other words, the disgust experienced could function as indicating some of the costs imposed of some individuals, thereby influencing the moral evaluation of the interaction at hand (see Tybur et al., 2013; Baumard et al., 2013). For example, farting during a meal can in some contexts be perceived as both disgusting and immoral.

Does this mean that the mere disgusting character of farting *moralizes* it? Probably not: What may be intuited as immoral is unfairly *causing* disgust, as a negative psychological experience, in other people who don't deserve to feel this way while eating. To take a less trivial example, acting in a way that endangers others' lives by exposing them to pathogens (e.g., by spreading an infected substance on them) is likely to be both perceived as immoral and disgusting. But again, this will be probably intuited as immoral only insofar at the pathogens at hand constitute an undeserved cost imposition on others. In other words, to be immoral, disgusting actions should have to be somehow unfair, in the sense of imposing illegitimate costs on other people in the context of cooperative interactions.

In line with these suspicions, experimental evidence for a moralizing role of disgust is overall weak (Piazza et al., 2018, for an extensive review; Landy & Goodwin, 2015, for a meta-analysis). Here, we focus on the following important empirical points.

First, consistent with the above ideas, when participants are asked to judge disgusting actions (e.g., spitting into a napkin while at a dinner party), it is the perception that the action negatively affected the welfare of other people, and not the disgust elicited by the action *per se*, that significantly predicts people's moralization (Royzman et al., 2009). Relatedly, the recurrently reported correlations between disgust-sensitivity and moralizations of purity and sexual behaviors (e.g. Horberg et al., 2009; Inbar, Pizarro, & Bloom, 2009; Inbar, Pizarro, Knobe, et al., 2009) appear largely mediated by perceptions of harm and feelings of anger, and disappear when the latter are controlled for (Schein et al., 2016).

Second, in many (most?) cases, the association between disgust and moral judgements seems merely coincidental: Some unfairly harmful actions, naturally judged immoral, also *happen* to have pathogen-related properties, so that they also trigger disgust (e.g., forbidden sex can include sexual fluids, violence can include blood) (Kayyal et al., 2015). Conversely, actions that have *positive* cooperation-related properties, yet also contain disgusting pathogen cues (e.g. a nurse changing an elderly patient feces-covered sheets), are not morally condemned, but morally *praised* (Kayyal et al., 2015). This suggests that disgust (tracking pathogen cues) and moral judgment (tracking unfair cost imposition or benefit-providing) mostly operate independently, and simply sometimes co-occur in the case of immoral actions that also happen to be disgusting.

Third, and relatedly, even sex- and religion-related violations of so-called "Purity/Sanctity" moral concerns, the apparently harmless character of which initially justified the idea of a moralizing disgust, increasingly appear as tied to perceptions of unfair harm (Gray et al., 2014; Royzman et al., 2015; Schein et al., 2016). For example, the famous, intrinsically harmless scenario of "Julie's and Mark's" sibling incest (Haidt, 2001) often taken as evidence that harmless actions are moralized because they are disgusting, in fact fails to convince participants that the action they are judging is really harmless (Royzman et al., 2015). This suggests that moralizations of "purity", too, may be underpinned by computations unrelated to disgust, and have been mistakenly causally associated with disgust because of the coincidentally disgusting character of some of these behaviors (e.g. sibling incest). In line with this idea, studies addressing the confounding effect of pathogen cues find that pathogen-free violations of the morality of Purity/Sanctity (e.g., stepping on the Quran) are not associated with disgust-related phenomenology or action tendencies, but rather with moral anger, an emotion commonly associated with cooperation-related moral judgements (Royzman et al., 2014).

Fourth, while a range of studies find that experimentally induced disgust amplifies moral judgement (e.g. Horberg et al., 2009; Schnall et al., 2008; Wheatley & Haidt, 2005), a recent meta-analysis of 50 published and unpublished studies found no overall effect of incidental disgust after accounting for a probable publication bias (Landy & Goodwin, 2015; see also Johnson et al., 2016).

Putting these pieces of evidence together suggests the following picture:

1. In many cases, the co-occurrence of disgust and moral representations is merely coincidental (e.g. adulterous sexual pleasure involves sexual fluids, but is not immoral *because* of that) (Kayyal et al., 2015).
2. When this is not the case, disgusting actions are judged immoral only insofar as they are perceived to unjustly impose costs on other people (Royzman et al., 2009; Schein et al., 2016). In other words, it is not disgust *per se* that produces moralization, but perceptions of unfair cost imposition, to which disgusting behaviors can contribute. In this case, disgust would only play the role of an input informing the moral sense of the costs imposed on other people — just as, say, perceptions of people's pain influence our moral judgements.
3. Indignation (moral anger) is the predominant emotional response to moral violations *across moral domains* (even when it comes to "purity" violations) (Kayyal et al., 2015; Piazza et al., 2018; Schein et al., 2016; Royzman et al., 2016), consistent with the idea that moral indignation is the main other-condemning emotion processing moral representations.

Overall, consistent with the idea that human moral cognition is mostly designed to ensure mutually beneficial cooperation, there is little conclusive evidence for pathogen-avoidance mechanisms playing a strong role in moral representations or moral condemnation.

Is immorality disgusting?

A weaker hypothesis has been put forward, according to which disgust does not *generate* moral condemnation, but the reverse: immoral behavior generates a reaction of disgust (Hanah A. Chapman & Anderson, 2013). In support for this view, researchers have noted that people report feeling "disgusted" or "morally disgusted" in response to immoral behaviors (e.g., lying, cheating, stealing) (Hutcherson & Gross, 2011), choose facial expressions of disgust as corresponding to their reaction to immoral behaviors (Molho et al., 2017; Rozin et al., 1999), and express disgust-related facial expressions in response to unfair offers in economic games and moral violations (Cannon et al., 2011; Chapman et al., 2009)

From a functional perspective, the theoretical grounding of such "moral disgust" is however unclear. A possibility would be that the typical design features of disgust, mainly its avoidance motivation, have been co-opted over evolutionary time to serve the secondary function of avoiding immoral individuals with whom cooperation results in net costs (Curtis & Biran, 2001; Hutcherson & Gross, 2011). Relatedly, it has been suggested that, as opposed to the directly punitive function of anger, moral disgust could function to motivate less costly "indirect" punishments of immoral behavior (e.g. through ostracism or reputational sanctioning) (Mohlo et al., 2017). In a similar vein, scholars have also proposed that moral disgust allows people to facilitate, through the communicative function of disgust's typical facial expression, the coordination of moral condemnation with surrounding individuals, by signaling one's disapproval of a moral violation (Tybur et al., 2013).

In solving each of these adaptive problems, however, it is not clear where the added value of disgust's design-features lies. For one thing, why would humans need 'moral disgust', to avoid or ostracize immoral individuals, when they already have *contempt*? Contempt, indeed, seem to unambiguously function for social valuation and partner choice: it is triggered by cues of low relational value (e.g. incompetence, norm transgression), and generates action tendencies of avoidance, exclusion and relationship dissolution (Gervais & Fessler, 2017). Moreover, both in the lab (Molleman et al., 2019) and in the wild (Boehm, 1999, 2012), moral anger appears to readily motivate the coordination of punishment, condemnation, and ostracism of moral violators – consistent with the bargaining-oriented design of anger. Relatedly, a specific communicative payoff of

disgust's facial expression is not obvious when linguistic communication (in the form of gossip) is an efficient and widely used mean for indirect reputational punishment and social valuation coordination (Boehm, 2012; Wiessner, 2005). In other words, is there a place, and a need, for such a thing as a 'moral disgust' when we already have moral anger, contempt, and gossip serving these key functions?

These theoretical issues are consistent with widely noted methodological questions regarding the measurement of "moral disgust" (Armstrong et al., 2020; Piazza et al., 2018). The most common operationalizations of moral disgust are difficult to distinguish from moral anger, contempt, or moral disapproval more generally. First, as long noted, it is not clear that studies in which participants self-report being "disgusted" by moral violations imply that the cognitive system of disgust is really triggered. Indeed, the lay meaning of "disgust", when applied to moral transgressions, has been found to largely overlap with reports of "anger", "moral anger", "contempt" and "moral contempt" (Nabi, 2002; Russell & Giner-Sorolla, 2013).

Second, more implicit measures of facial expressions of disgust, initially used to overcome these limitations (e.g. Chapman et al., 2009), appear to face similar problems. Indeed, the "standard" disgust face has been found to potentially express, and be associated with, more emotions than disgust: in particular, again, with anger and contempt (Widen & Russell, 2008; Widen et al., 2013; Gervais & Fessler, 2017). Moreover, even if participants select or express a truly disgust related-face in response to a moral violation, this can be a metaphorical way of communicating their disapproval, not associated to an actual *experience* of disgust – just as we sometimes say we are "hungry" for knowledge without the cognitive and physiological mechanisms of hunger really being triggered (Royzman & Sabini, 2001; Royzman & Kurzban, 2011).

Future research using more discriminant measures of disgust's typical physiological, cognitive, and phenomenological signatures (e.g. nausea, gagging, loss of appetite), as done by Royzman et al. (2014), could probably help settle these issues with more clarity.

5. Conclusion

In this chapter, we have suggested that a specification of the form and function of moral representations leads to a clearer picture of moral emotions. In particular, it enables a principled distinction between moral and non-moral emotions, based on the particular types of cognitive representations they process. Moral representations have a specific content: they represent a precise quantity of benefits that cooperative partners owe each other, a legitimate allocation of costs and benefits that *ought to be*, irrespective of whether it is achieved by people's actual behaviors. Humans intuit that they have a *duty* not to betray their coalition, that innocent people do not *deserve* to be harmed, that their partner has a *right* not to be cheated on. Moral emotions can thus be defined as superordinate programs orchestrating cognition, physiology and behavior in accordance with the specific information encoded in these moral representations.

On this basis, indignation and guilt appear as prototypical moral emotions. Indignation ("moral anger") is activated when one receives fewer benefits than one deserves, and recruits bargaining mechanisms to enforce the violated moral contract. Guilt, symmetrically, is sensitive to one's failure to honor one's obligations toward others, and motivates compensation to provide them the missing benefits they deserve. By contrast, often-proposed "moral" emotions – shame, empathy, disgust – seem not to function to compute distinctively moral representations of cooperative obligations, but serve other, non-moral functions – social status management, interdependence, and pathogen avoidance (Figure 2).

Acknowledgments. We thank Laith Al-Shawaf for his thoughtful feedback and corrections. This work was supported by the EUR FrontCog grant ANR-17-EURE-0017.

6. References

Aktipis, A., Cronk, L., Alcock, J., Ayers, J. D., Baciu, C., Balliet, D., Boddy, A. M., Curry, O. S., Krems, J. A., Muñoz, A., Sullivan, D., Sznycer, D., Wilkinson, G. S., & Winfrey, P. (2018). Understanding cooperation through fitness interdependence. *Nature Human Behaviour*, *2*(7), 429–431. https://doi.org/10.1038/s41562-018-0378-4

Alexander, R. D. (1987). *The biology of moral systems*. A. de Gruyter.

Al-Shawaf, L., Conroy-Beam, D., Asao, K., & Buss, D. M. (2016). Human Emotions: An Evolutionary Psychological Perspective. *Emotion Review*, *8*(2), 173–186. https://doi.org/10.1177/1754073914565518

Al-Shawaf, L., & Lewis, D. (2017). *Evolutionary Psychology and the Emotions*. https://doi.org/10.1007/978-3-319-28099-8_516-1

André, J.-B., & Baumard, N. (2011). The evolution of fairness in a biological market. *Evolution: International Journal of Organic Evolution*, *65*(5), 1447–1456.

Armstrong, T., Wilbanks, D., Leong, D., & Hsu, K. J. (2020). *Is There a Measurement Crisis in Disgust Research?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/a8u5m

Avramova, Y. R., & Inbar, Y. (2013). Emotion and moral judgment: Emotion and moral judgment. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 169–178. https://doi.org/10.1002/wcs.1216

Axelrod, R., & Hamilton, W. D. (1981). *The Evolution of cooperation*.

Barclay, P. (2005). *Reputational benefits for altruistic punishment*.

Batson, C., & Ahmad, N. (2009). Empathy-induced altruism: A threat to the collective good. *Advances in Group Processes*, *26*, 1–23. https://doi.org/10.1108/S0882-6145(2009)0000026004

Batson, C. D. (2017). *The Empathy-Altruism Hypothesis* (E. M. Seppälä, E. Simon-Thomas, S. L.

Brown, M. C. Worline, C. D. Cameron, & J. R. Doty, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190464684.013.3

Batson, C. D., Ahmad, N., Yin, J., Bedell, S. J., Johnson, J. W., & Templin, C. M. (1999). Two Threats to the Common Good: Self-Interested Egoism and Empathy-Induced Altruism. *Personality and Social Psychology Bulletin*, *25*(1), 3–16. https://doi.org/10.1177/0146167299025001001

Batson, C. D., Kennedy, C. L., Nord, L.-A., Stocks, E. L., Fleming, D. A., Marzette, C. M.,

Lishner, D. A., Hayes, R. E., Kolchinsky, L. M., & Zerger, T. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology*, *37*(6), 1272–1285. https://doi.org/10.1002/ejsp.434

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology*, *68*(6), 1042.

Baumard, N. (2016). *The origins of fairness: How evolution explains our moral nature*. Oxford University Press.

Baumard, Nicolas. (2010). Punishment is not a group adaptation: Humans punish to restore fairness rather than to support group cooperation. *Mind & Society*, *10*(1), 1–26. https://doi.org/10.1007/s11299-010-0080-3

Baumard, N., & Sheskin, M. (2015). Partner choice and the evolution of a contractualist morality. *The moral brain: a multidisciplinary perspective*, *20*, 35-48.

Baumard, Nicolas, André, J.-B., & Sperber, D. (2013a). Partner choice, fairness, and the extension of morality. *Behavioral and Brain Sciences*, *36*(1), 102.

Baumard, Nicolas, André, J.-B., & Sperber, D. (2013b). A mutualistic approach to morality: The

evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(01), 59–78. https://doi.org/10.1017/S0140525X11002202
Baumard, Nicolas, & Liénard, P. (2011). Second-or third-party punishment? When self-interest hides behind apparent functional interventions. *Proceedings of the National Academy of Sciences*, *108*(39), E753–E753.
Baumard, Nicolas, Mascaro, O., & Chevallier, C. P. (2011). Preschoolers are able to take merit into account when distributing goods. *Developmental Psychology*, *48*(2), 492–498. https://doi.org/10.1037/a0026598
Berndt, R. M., Berndt, C., Berndt, R., & Berndt, C. H. (1988). *The world of the first Australians: Aboriginal traditional life: Past and present*. Aboriginal Studies Press.
Black, D. (2000). On the origin of morality. *Journal of Consciousness Studies*, *7*, 107–119.
Bloom, P. (2017). *Against Empathy: The Case for Rational Compassion*. Random House.
Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Harvard University Press.
Boehm, C. (2012). *Moral origins: The evolution of virtue, altruism, and shame*. Basic Books.
Buss, D. M. (2003). *The evolution of desire: Strategies of human mating* (Rev. ed). BasicBooks.
Cannon, P. R., Schnall, S., & White, M. (2011). Transgressions and Expressions: Affective Facial Muscle Activity Predicts Moral Judgments. *Social Psychological and Personality Science*, *2*(3), 325–331. https://doi.org/10.1177/1948550610390525
Carlsmith, K. M., & Darley, J. M. (2008). Psychological Aspects of Retributive Justice. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 193–236). Elsevier. https://doi.org/10.1016/S0065-2601(07)00004-4
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), 1222–1226. https://doi.org/10.1126/science.1165565
Chapman, Hanah A., & Anderson, A. K. (2013). Things rank and gross in nature: A review and synthesis of moral disgust. *Psychological Bulletin*, *139*(2), 300–327. https://doi.org/10.1037/a0030964
Chapman, Hanah A., & Anderson, A. K. (2014). Trait physical disgust is related to moral judgments outside of the purity domain. *Emotion*, *14*(2), 341–348. https://doi.org/10.1037/a0035120
Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us Versus Them: Social Identity Shapes Neural Responses to Intergroup Competition and Harm. *Psychological Science*, *22*(3), 306–313. https://doi.org/10.1177/0956797610397667
Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(6511), 209–216. https://doi.org/10.1038/373209a0
Cohen, T. R., Panter, A. T., & Turan, N. (2013). Predicting Counterproductive Work Behavior from Guilt Proneness. *Journal of Business Ethics*, *114*(1), 45–53. https://doi.org/10.1007/s10551-012-1326-2
Corbit, J., McAuliffe, K., Callaghan, T. C., Blake, P. R., & Warneken, F. (2017). Children's collaboration induces fairness rather than generosity. *Cognition*, *168*, 344–356. https://doi.org/10.1016/j.cognition.2017.07.006
Cosmides, L., & Tooby, J. (2000). *Evolutionary Psychology and the Emotions*.
Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality*, *78*, 106–124. https://doi.org/10.1016/j.jrp.2018.10.008
Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies. *Current Anthropology*, *60*(1), 47–69. https://doi.org/10.1086/701478
Curtis, V., & Biran, A. (2001). Dirt, Disgust, and Disease: Is Hygiene in Our Genes? *Perspectives in Biology and Medicine*, *44*, 17–31. https://doi.org/10.1353/pbm.2001.0001
Darley, J. M., & Pittman, T. S. (2003). The Psychology of Compensatory and Retributive Justice. *Personality and Social Psychology Review*, *7*(4), 324–336. https://doi.org/10.1207/S15327957PSPR0704_05

Darwall, S. (2010). III-Moral Obligation: Form and Substance. *Proceedings of the Aristotelian Society (Hardback)*, *110*(1pt1), 31–46. https://doi.org/10.1111/j.1467-9264.2010.00278.x
Davidov, M., Zahn-Waxler, C., Roth-Hanania, R., & Knafo, A. (2013). Concern for Others in the First Year of Life: Theory, Evidence, and Avenues for Research. *Child Development Perspectives*, *7*(2), 6.
Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–796. https://doi.org/10.1038/nature05651
de Hooge, I. E., Breugelmans, S. M., & Zeelenberg, M. (2008). Not so ugly after all: When shame acts as a commitment device. *Journal of Personality and Social Psychology*, *95*(4), 933–943. https://doi.org/10.1037/a0011991
de Hooge, I. E., Nelissen, R. M. A., Breugelmans, S. M., & Zeelenberg, M. (2011). What is moral about guilt? Acting "prosocially" at the disadvantage of others. *Journal of Personality and Social Psychology*, *100*(3), 462–473. https://doi.org/10.1037/a0021459
de Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion*, *21*(5), 1025–1042. https://doi.org/10.1080/02699930600980874
de Waal, F. B. M. (2008). Putting the Altruism Back into Altruism: The Evolution of Empathy. *Annual Review of Psychology*, *59*(1), 279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625
Debove, S., André, J.-B., & Baumard, N. (2015). Partner choice creates fairness in humans. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1808), 20150392. https://doi.org/10.1098/rspb.2015.0392
Debove, S., Baumard, N., & André, J.-B. (2015). Evolution of equal division among unequal partners: BRIEF COMMUNICATION. *Evolution*, *69*(2), 561–569. https://doi.org/10.1111/evo.12583
Decety, J., Bartal, I., Uzefovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behaviour: Highly conserved neurobehavioural mechanisms across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150077. https://doi.org/10.1098/rstb.2015.0077
Durkee, P. K., Lukaszewski, A. W., & Buss, D. M. (2019). Pride and shame: Key components of a culturally universal status management system. *Evolution and Human Behavior*, *40*(5), 470–478. https://doi.org/10.1016/j.evolhumbehav.2019.06.004
Elison, J., Garofalo, C., & Velotti, P. (2014). Shame and aggression: Theoretical considerations. *Aggression and Violent Behavior*, *19*(4), 447–453. https://doi.org/10.1016/j.avb.2014.05.002
Engelmann, J. M., & Tomasello, M. (2019). Children's Sense of Fairness as Equal Respect. *Trends in Cognitive Sciences*, *23*(6), 454–463. https://doi.org/10.1016/j.tics.2019.03.001
Evans-Pritchard, E. E. (1940). *The Nuer: A description of the modes of livelihood and political institutions of a Nilotic people*.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 5.
Fessler, D. M., & Haley, K. J. (2003). The strategy of affect: Emotions in human cooperation 12. *The Genetic and Cultural Evolution of Cooperation, P. Hammerstein, Ed*, 7–36.
Fessler, D. M. T. (2010). *Madmen: An Evolutionary Perspective on Anger and Men's Violent Responses to Transgression*. 21.
Flack, J. C., & De Waal, F. B. (2000). 'Any animal whatever'. Darwinian building blocks of morality in monkeys and apes. *Journal of Consciousness Studies*, *7*(1–2), 1–29.
Fletcher, G., Simpson, J., Campbell, L., & Overall, N. C. (2015). Pair-Bonding, Romantic Love, and Evolution. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*. https://doi.org/10.1177/1745691614561683
Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. W W Norton & Co.
Gangestad, S. W., & Simpson, J. A. (2000). The evolution of human mating: Trade-offs and strategic pluralism. *Behavioral and Brain Sciences*, *23*(4), 573–587. https://doi.org/10.1017/S0140525X0000337X
Gausel, N., Vignoles, V. L., & Leach, C. W. (2016). Resolving the paradox of shame: Differentiating

among specific appraisal-feeling combinations explains pro-social and self-defensive motivation. *Motivation and Emotion*, *40*(1), 118–139. https://doi.org/10.1007/s11031-015-9513-y
Gauthier, D. (1986). *Morals by agreement*. Oxford University Press on Demand.
Gervais, M. M., & Fessler, D. M. T. (2017). On the deep structure of social affect: Attitudes, emotions, sentiments, and the case of "contempt." *Behavioral and Brain Sciences*, *40*, e225. https://doi.org/10.1017/S0140525X16000352
Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00002-4
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*(4), 1600–1615. https://doi.org/10.1037/a0036149
Greene, J. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin.
Gurven, M., Winking, J., Kaplan, H., von Rueden, C., & McAllister, L. (2009). A Bioeconomic Approach to Marriage and the Sexual Division of Labor. *Human Nature*, *20*(2), 151–183. https://doi.org/10.1007/s12110-009-9062-8
Hagen, E. H., & Garfield, Z. (2019). *Leadership and prestige, mothering, sexual selection, and encephalization: The computational services model* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/9bcdk
Haidt. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Knopf Doubleday Publishing Group.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814.
Haidt, J. (2003). The moral emotions. *Handbook of Affective Sciences*, *11*(2003), 852–870.
Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651
Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals 1. *Journal of Applied Social Psychology*, *31*(1), 191-221.
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*(4), 613.
Henrich, J., & Boyd, R. (2001). Why People Punish Defectors. *Journal of Theoretical Biology*, *208*(1), 79–89. https://doi.org/10.1006/jtbi.2000.2202
Hooge, I. E. de, Zeelenberg, M., & Breugelmans, S. M. (2010). Restore and protect motivations following shame. *Cognition and Emotion*, *24*(1), 111–127. https://doi.org/10.1080/02699930802584466
Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*(6), 963–976. https://doi.org/10.1037/a0017423
Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology*, *100*(4), 719–737. https://doi.org/10.1037/a0022408
Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition & Emotion*, *23*(4), 714–725. https://doi.org/10.1080/02699930802110007
Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, *9*(3), 435–439. https://doi.org/10.1037/a0015960
Johnson, D. J., Wortman, J., Cheung, F., Hein, M., Lucas, R. E., Donnellan, M. B., Ebersole, C. R., & Narr, R. K. (2016). The Effects of Disgust on Moral Judgments: Testing Moderators. *Social Psychological and Personality Science*, *7*(7), 640–647. https://doi.org/10.1177/1948550616654211
Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981
Kachel, U., Svetlova, M., & Tomasello, M. (2017). Three-Year-Olds' Reactions to a Partner's Failure

to Perform Her Role in a Joint Commitment. *Child Development*, *89*. https://doi.org/10.1111/cdev.12816
Kayyal, M. H., Pochedly, J., McCarthy, A., & Russell, J. A. (2015). On the limits of the relation of disgust to judgments of immorality. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00951
Ketelaar, T., & Au, W. T. (2003). *The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction*. 26.
Kohlberg, L., & Kramer, R. (1969). Continuities and discontinuities in childhood and adult moral development. *Human Development*, *12*(2), 93–120.
Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*(26), 15387–15392.
Landy, J. F., & Goodwin, G. P. (2015). Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence. *Perspectives on Psychological Science*, *10*(4), 518–536. https://doi.org/10.1177/1745691615583128
Liénard, P. (2013). Early Understanding of Merit in Turkana Children. *Journal of Cognition and Culture*, 10.
Masten, C. L., Gillen-O'Neel, C., & Brown, C. S. (2010). Children's intergroup empathic processing: The roles of novel ingroup identification, situational distress, and social anxiety. *Journal of Experimental Child Psychology*, *106*(2–3), 115–128. https://doi.org/10.1016/j.jecp.2010.01.002
McAuliffe, W. H. B. (2019). *Do Emotions Play an Essential Role in Moral Judgments?* https://doi.org/10.31234/osf.io/ajbc9
McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160
Molho, C., Tybur, J. M., Güler, E., Balliet, D., & Hofmann, W. (2017). Disgust and Anger Relate to Different Aggressive Responses to Moral Violations. *Psychological Science*, *28*(5), 609–619. https://doi.org/10.1177/0956797617692000
Molleman, L., Kölle, F., Starmer, C., & Gächter, S. (2019). People prefer coordinated punishment in cooperative interactions. *Nature Human Behaviour*, *3*(11), 1145–1153. https://doi.org/10.1038/s41562-019-0707-2
Nabi, R. L. (2002). The theoretical versus the lay meaning of disgust: Implications for emotion research. *Cognition & Emotion*, *16*(5), 695–703. https://doi.org/10.1080/02699930143000437
Nelissen, R. M. A., & Zeelenberg, M. (2010). *Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions*. 11.
Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*(2), 221–236. https://doi.org/10.1016/S0010-0277(02)00048-3
O'Mara Kunz, E., Eckstein, L., Batson, C., & Gaertner, L. (2011). Will moral outrage stand up?: Distinguishing among emotional reactions to a moral violation. *European Journal of Social Psychology*, *41*, 173–179. https://doi.org/10.1002/ejsp.754
Osgood, J. M. (2017). Is revenge about retributive justice, deterring harm, or both?: JUSTICE, DETERRENCE, OR BOTH? *Social and Personality Psychology Compass*, *11*(1), e12296. https://doi.org/10.1111/spc3.12296
Pedersen, E. J., McAuliffe, W. H. B., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514–544. https://doi.org/10.1037/xge0000410
Pedersen, E. J., McAuliffe, W. H. B., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M. E. (2019). When and Why Do Third Parties Punish Outside of the Lab? A Cross-Cultural Recall Study. *Social Psychological and Personality Science*, 1948550619884565. https://doi.org/10.1177/1948550619884565
Piaget, J. (1997). *The Moral Judgement of the Child*. Simon and Schuster.
Piazza, J., Landy, J. F., Chakroff, A., Young, L., & Wasserman, E. (2018). What disgust does

and does not do for moral cognition. *The Moral Psychology of Disgust*, 53–81.
Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, *25*(1), 1–20. https://doi.org/10.1017/S0140525X02000018
Prinz, J. J., & Nichols, S. B. (2010). Moral emotions. *The Moral Psychology Handbook*.
Rakoczy, H., Kaufmann, M., & Lohse, K. (2016). Young children understand the normative force of standards of equal resource distribution. *Journal of Experimental Child Psychology*, *150*, 396–403. https://doi.org/10.1016/j.jecp.2016.05.015
Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003
Rawls, J. (2001). A theory of justice: Original edition. *Cambridge: Harvard University*.
Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, *70*(4), 901–908. https://doi.org/10.1016/j.anbehav.2005.02.006
Robertson, T. E., Sznycer, D., Delton, A. W., Tooby, J., & Cosmides, L. (2018). The true trigger of shame: Social devaluation is sufficient, wrongdoing is unnecessary. *Evolution and Human Behavior*, *39*(5), 566–573. https://doi.org/10.1016/j.evolhumbehav.2018.05.010
Royzman, E., Atanasov, P., Landy, J. F., Parks, A., & Gepty, A. (2014). CAD or MAD? Anger (not disgust) as the predominant response to pathogen-free violations of the divinity code. *Emotion*, *14*(5), 892–907. https://doi.org/10.1037/a0036829
Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, *10*(4), 296.
Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral–conventional distinction. *Cognition*, *112*(1), 159–174. https://doi.org/10.1016/j.cognition.2009.04.004
Royzman, E. B., & Sabini, J. (2001). Something it Takes to be an Emotion: The Interesting Case of Disgust. *Journal for the Theory of Social Behaviour*, *31*(1), 29–59. https://doi.org/10.1111/1468-5914.00145
Royzman, E., & Kurzban, R. (2011). Minding the Metaphor: The Elusive Character of Moral Disgust. *Emotion Review*, *3*(3), 269–271. https://doi.org/10.1177/1754073911402371
Rozin, P., Lowery, L., Haidt, J., & Imada, S. (1999). *The CAD Triad Hypothesis: A Mapping Between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity). 76*(4), 574–586.
Russell, P. S., & Giner-Sorolla, R. (2013). Bodily moral disgust: What it is, how it is different from anger, and why it is an unreasoned emotion. *Psychological Bulletin*, *139*(2), 328–351. https://doi.org/10.1037/a0029319
Scanlon, T. (2000). *What we owe to each other*. Belknap Press.
Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288
Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. *Emotion*, *16*(6), 862–876. https://doi.org/10.1037/emo0000167
Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as Embodied Moral Judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1096–1109. https://doi.org/10.1177/0146167208317771
Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*(35), 15073–15078. https://doi.org/10.1073/pnas.0904312106
Sell, Aaron. (2017). Recalibration Theory of Anger. In T. K. Shackelford & V. A. Weekes-
Shackelford (Eds.), *Encyclopedia of Evolutionary Psychological Science* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-16999-6_1687-1
Sell, Aaron, Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Aneta, F., Rascanu, R., Sugiyama, L.,
Cosmides, L., & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110–128. https://doi.org/10.1016/j.cognition.2017.06.002

Sheikh, S., & Janoff-Bulman, R. (2010). The "'Shoulds'" and " 'Should Nots'" of Moral Emotions: A Self-Regulatory Perspective on Shame and Guilt. *Personality and Social Psychology Bulletin*, 13.

Smith, A. (1822). *The theory of moral sentiments* (Vol. 1). J. Richardson.

Smith, C. E., & Warneken, F. (2016). Children's reasoning about distributive and retributive justice across development. *Developmental Psychology*, *52*(4), 613–628. https://doi.org/10.1037/a0040069

Smith, R. H., Webster, J. M., Parrott, W. G., & Eyre, H. L. (2002). The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology*, *83*(1), 138–159. https://doi.org/10.1037/0022-3514.83.1.138

Sperber, D., & Baumard, N. (2012). Moral Reputation: An Evolutionary and Cognitive Perspective: Moral Reputation. *Mind & Language*, *27*(5), 495–518. https://doi.org/10.1111/mila.12000

Stanford, P. K. (2018). The difference between ice cream and Nazis: Moral externalization and the evolution of human cooperation. *Behavioral and Brain Sciences*, *41*, e95. https://doi.org/10.1017/S0140525X17001911

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, *1*(4), 0082. https://doi.org/10.1038/s41562-017-0082

Sznycer, D. (2019). Forms and functions of the self-conscious emotions. *Trends in Cognitive Sciences*, *23*(2), 143–157.

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences*, *113*(10), 2625–2630. https://doi.org/10.1073/pnas.1514699113

Sznycer, D., Xygalatas, D., Agey, E., Alami, S., An, X.-F., Ananyeva, K. I., Atkinson, Q. D., Broitman, B. R., Conte, T. J., Flores, C., Fukushima, S., Hitokoto, H., Kharitonov, A. N., Onyishi, C. N., Onyishi, I. E., Romero, P. P., Schrock, J. M., Snodgrass, J. J., Sugiyama, L. S., … Tooby, J. (2018). Cross-cultural invariances in the architecture of shame. *Proceedings of the National Academy of Sciences*, 201805016. https://doi.org/10.1073/pnas.1805016115

Tangney, June P., Stuewig, J., & Martinez, A. G. (2014). Two Faces of Shame: The Roles of Shame and Guilt in Predicting Recidivism. *Psychological Science*, *25*(3), 799–805. https://doi.org/10.1177/0956797613508790

Tangney, June Price, Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annu. Rev. Psychol.*, *58*, 345–372.

Thulin, E. W., & Bicchieri, C. (2016). I'm so angry I could help you: Moral outrage as a driver of victim compensation. *Social Philosophy and Policy*, *32*(2), 146–160. https://doi.org/10.1017/S0265052516000145

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

Tomasello, M. (2020). The moral psychology of obligation. *Behavioral and Brain Sciences*, *43*, e56. https://doi.org/10.1017/S0140525X19001742

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. *Handbook of Approach and Avoidance Motivation*, *15*, 251.

Tracy, J. L., & Matsumoto, D. (2008). The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences*, *105*(33), 11655–11660.

Trivers, R. (1971). *The Evolution of Reciprocal Altruism*. *46*(1), 35–57.

Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review*, *120*(1), 65–84. https://doi.org/10.1037/a0030778

Vollan, B., Blanco, E., Steimanis, I., Petutschnig, F., & Prediger, S. (2020). Procedural fairness and nepotism among local traditional and democratic leaders in rural Namibia. *Science Advances*, *6*(15), eaay7651. https://doi.org/10.1126/sciadv.aay7651

von Rueden, C., Gurven, M., & Kaplan, H. (2011). Why do men seek status? Fitness payoffs to dominance and prestige. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1715), 2223–2232. https://doi.org/10.1098/rspb.2010.2145

von Rueden, C., & Jaeggi, A. (2016). Men's status and reproductive success in 33 nonindustrial

societies: Effects of subsistence, marriage system, and reproductive strategy. *Proceedings of the National Academy of Sciences*, *113*. https://doi.org/10.1073/pnas.1606800113

Wheatley, T., & Haidt, J. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science*, *16*(10), 780–784. https://doi.org/10.1111/j.1467-9280.2005.01614.x

Widen, S. C., Pochedly, J. T., Pieloch, K., & Russell, J. A. (2013). Introducing the sick face. *Motivation and Emotion*, *37*(3), 550–557. https://doi.org/10.1007/s11031-013-9353-6

Widen, S., & Russell, J. (2008). Children's and adults' understanding of the "disgust face.". *Cognition and Emotion*, *22*. https://doi.org/10.1080/02699930801906744

Wiessner, P. (2005). Norm enforcement among the Ju/'hoansi Bushmen: A case of strong reciprocity? *Human Nature*, *16*(2), 115–145. https://doi.org/10.1007/s12110-005-1000-9

Wrangham, R. (2019). *The goodness paradox: The strange relationship between virtue and violence in human evolution*. Vintage.

Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses. *Journal of Neuroscience*, *29*(26), 8525–8529. https://doi.org/10.1523/JNEUROSCI.2418-09.2009