What Roles Do Emotions Play in Morality?

Antti Kauppinen

Revised draft, January 30, 2022

For Andrea Scarantino (ed.), *The Routledge Handbook of Emotion Theory*.

For better or for worse, our moral lives are significantly shaped by our emotions. Everyone agrees that they influence what we regard as good or bad as well as our behavior towards others: if we are envious of someone, we'll easily find flaws in them, and if we feel guilty about a making a cruel joke, we are apt to make amends and take more care in the future. What is a matter of debate is how far this influence reaches and whether we should welcome it. Some, like the Scottish enlightenment thinker David Hume (1739-40/1978), argue that certain emotions are necessary elements of moral thought or virtue, while others, like the German philosopher Immanuel Kant (1785/1998), see them as merely distorting factors, or at best consequences of unemotional moral judgments.

In this chapter, I examine some of these various roles emotions might play in morality. I'll begin with a look at their significance to the very emergence of unselfish behavior and moral agency. Many scientists and philosophers believe that the evolution of emotions and emotional processes like empathy, fear, resentment, and shame is crucial to explaining how people came to act in ways that benefit others and to follow and enforce social and moral norms. Second, many psychologists and neuroscientists have recently emphasized that affective reactions shape people's moral judgments in their different varieties. There is, however, much debate about what the evidence shows about the relative significance of affect, other kinds of non-conscious cognition, and conscious reasoning.

In the third section, I move to less empirical philosophical questions about the possible constitutive role of emotion in moral thinking and even moral truth. Philosophical

sentimentalists have often argued that when we judge that something is wrong, for example, what we believe is that the object either causes or merits negative emotional responses. On such views, moral truths will be truths about what causes or merits the suitable responses. The main challenge for such approaches is accounting for the seeming objectivity, authority, and normativity of morality. So-called expressivist views sidestep some of these challenges by arguing that moral judgments are not *about* sentiments in any way, but rather *express* them. Finally, in Section 4 I turn to questions about moral knowledge and justification. Some philosophers argue that the direct and indirect causal influence that emotions have on our moral beliefs undermines our justification for them, because our emotions have evolved to enhance our biological fitness rather than track moral truths. Others reject these arguments, and some even argue that emotions constitute our most fundamental mode of access to value.

In a relatively short chapter, I must set aside several important roles that emotions seem to have in our moral practices. To mention just two important aspects, many ethicists from the 4th century BCE philosophers Mencius (1970) and Aristotle (2000) onwards have argued that emotions are essential to being *virtuous*: virtuous people don't just do the right thing, but also feel the right way. And second, reactive attitudes like resentment and indignation are arguably constitutive to our practices of *holding people responsible* for actions that manifest ill will or indifference, as P. F. Strawson (1962) held. I will only touch on these debates in passing below.

Essential Readings

Among the best sources for understanding the role of emotion in the evolution of moral behavior and thought are de Waal (2006) (which contains comments by several leading philosophers) and Tomasello (2016). Haidt (2001) is a classic, if controversial overview of the role of emotion in moral judgment, and Greene (2013) summarizes a lot of neuroscientific

research on the topic, arguing for the importance of reason in transcending tribal partiality. May (2018) offers a critical perspective on some of the more inflated claims made for affective primacy in ethics. When it comes to questions about the nature of moral judgment and moral facts, Prinz (2007) offers a comprehensive defense of relativist sentimentalism, while D'Arms and Jacobson (2000) influentially articulate a less radical neo-sentimentalist account. Schroeder (2010) offers an accessible but sophisticated overview of the non-cognitivist tradition in ethics. Finally, Street (2006) is a classic discussion of the epistemic challenges posed by the influence of evolved affective reactions. Joyce (2006) draws radically skeptical conclusions from such considerations, though he recommends that we should nevertheless pretend that morality is real. Tappolet (2016) is the most thorough defense of the positive epistemic role of emotion.

1. Emotions and the Emergence of Morality

Worker ants routinely sacrifice themselves for their colony, evidently out of evolved instinct (Tofilski et al. 2008). But why do human beings sometimes sacrifice their good and even their lives for others, including total strangers? Is such apparently morally laudable behavior always ultimately selfish? Emotions may offer a part of the answer, as Darwin (1871) himself speculated.

What Is Morality?

When we talk about moral behavior, we may mean two different things. First, we may talk about doing things that we regard as morally good or right. Plausibly, paradigm instances of morally good behavior are *altruistic* or *prosocial*, which is to say they promote the interests of others at some (potential) cost to oneself (e.g. Tomasello & Vaish 2013, 232). For clarity, I will talk about *altruistic behavior* in such cases. Such behavior can result from motives that

have nothing to do with morality, such as hard-wired instincts in ants or self-interested calculation in humans. So it is common to distinguish altruistic behavior from *psychological altruism*, in which altruistic behavior results from *altruistic motives*, desiring the good of another instead of (or in addition to) one's own good, and not merely as a means to future benefits for oneself (Kitcher 2011; for finer distinctions, see Kokkonen 2021). While everyone grants that there is altruistic behavior, the very existence of psychological altruism is sometimes disputed.

But second, by moral behavior we may also mean behavior that results from the agent herself thinking that it is morally good or required. I will call this *moral agency*. It requires the ability to form moral thoughts or judgments, or at least sensitivity to purported moral norms or reasons as such. Many philosophers emphasize that it is a form of self-governance: moral agents can take a step back from their non-moral motives and reflect on what they *ought* to do (Korsgaard 1996). Moral agency often results in altruistic behavior, but perceived moral demands are not limited to benefiting others – they may include vengeance for being wronged, for example. As it turns out, many scientists and philosophers believe that emotions have an important role in the emergence of both altruistic behavior and moral agency.

1.1 Emotions and the Evolution of Altruism

The Darwinian theory of natural selection offers a simple explanation of why organisms engage in behaviors that promote their own survival and reproduction in their environment: roughly, the genes that program for such behavioral dispositions are more likely to be passed on to future generations than those that don't, and consequently gradually become prevalent. This may easily suggest a picture of ruthless competition among individuals, since those who sacrifice for others reduce their own chances of surviving and reproducing. But in fact,

cooperation and even individual sacrifice is common in nature. The most interesting species for understanding the evolution of human morality are our closest relatives, primates like chimpanzees and bonobos, who appear to lack the capacity for moral thought proper, but engage in some altruistic behavior. Most obviously, family members care for the young, share food, and defend each other against outsiders (Silk 2002). But there is also limited cooperation with non-kin individuals. For example, chimpanzees sometimes groom each other, showing a preference for those who have groomed them in the past (Gomes, Mundry, and Boesch 2009). They also to some extent collaborate in hunting and share the resulting meat (Boesch 1994), and form coalitions to make gains in dominance hierarchy (de Waal 1982).

There are two well-known evolutionary explanations for how traits that cause altruistic behavior might result from natural selection. In what is known as *kin selection* (Hamilton 1964), the genes of a parent who is innately disposed to sacrifice for her offspring (and other relatives in proportion to genetic relatedness) will become more prevalent in later generations than those of a parent who lacks such a disposition, since the former's offspring are likelier to survive and reproduce. In '*reciprocal altruism*' (Trivers 1971), having a disposition to scratch another's back in return for having one's own back scratched tends to pay off for individuals when they're likely enough to interact again, as long as one is alert to cheating by the other and willing to break off relations in response to it. To do so is to employ a version of what evolutionary game theorists call the 'tit-for-tat' strategy, which has been shown to be evolutionarily stable in repeated interactions in suitable environments (Axelrod and Hamilton 1981). So, in certain contexts, altruistic behavior towards specific individuals is fitness-enhancing (that is, such as to increase the frequency of genes that program for it in future generations).

The question that interests us concerns the proximate mechanisms of altruistic behavior, which does not seem to issue from sheer ant-like instinct in species capable of intelligent agency. Instead, many believe that emotions have a crucial role. Why? To begin with, even if we remain as neutral as possible on the nature of emotions, it is relatively uncontroversial that affective responses (such as fear) prepare us for more or less broad types of action (such as fleeing) in response to specific changes in the environment (such as the presence of a predator), and that they have some degree of control over attention and motivation as well as independence from conscious deliberation (Frijda 1986, Scarantino 2014). Having such psychological mechanisms makes good evolutionary sense. For example, because eating rotten food is bad for survival, it's easy to see why a disgust response that *immediately* and *insistently motivates* us to avoid it is fitness-enhancing. At the same time, emotions, unlike hard-wired reflexes, are *flexible*: fear can result in many different kinds of situationally appropriate behavior to escape the salient threat.

The suggestion, then, is that prosocial emotions targeting kin and cooperative partners are an important proximate mechanism for fitness-enhancing altruistic behavior, which probably explains why they were selected for. For example, a mother's love binds her to her offspring, and causes her to feel empathic concern at the distress of her offspring (Sober and Wilson 1998). It is not easily ignored or silenced, arguably in virtue of its imperative phenomenal character (Kauppinen 2021). Such empathic feelings are strong motivators for non-instrumental helping (Batson 2010), which may benefit the offspring at net cost to self. Reciprocal helping and sharing, in turn, could in principle be motivated by strategic reasoning, but evidence from other primates and small children suggests that it is in the first instance based on affect. When it comes to primates, reciprocity occurs mostly in the context of long-term social relationships, such as among coalition partners, whose cooperation is strengthened by friendly feelings (Kitcher 1998). As Michael Tomasello

summarizes it, "great ape patterns of reciprocity on the behavioral level are underlain […] only by interdependence-based sympathy operating in both directions" (2016, 25). Similarly, Frans de Waal dubs the likeliest mechanism for limited reciprocal sharing in capuchin monkeys "attitudinal reciprocity", in which positive attitude of toleration of food-taking generates a similar response (de Waal 2000, Brosnan and de Waal 2002).

Emotions also plausibly play a key role in motivating *negative reciprocity*, that is, reacting to those who fail to do their part (Jensen 2010). Anger, in particular, motivates us to attack those who have cheated us, respectively, even when calculations of short-term self-interest would suggest otherwise. On individual instances, this can be costly, but as Robert Frank (1988) points out, if we succeed in developing a reputation as someone who will go after cheaters come what may, people will be less likely to cheat on us in the first place. For anger to play its role as such a 'commitment device', it must be to some extent unresponsive to incentives that are advantageous in the short term, and thus in a sense 'irrational' (Frank 1988, 51–56).

Does the fact that (broadly) prosocial emotions are originally fitness-enhancing show that they are ultimately selfish? No. We shouldn't confuse the adaptive 'purpose' of a trait with a person's purposes, which may be ultimately other-directed (Sober and Wilson 1998). To be sure, in some circumstances, reciprocally altruistic behavior is in the long-term interest of the agent, so she could, at least in principle, arrive at a decision to help another by cool self-interested calculation. Cynics have long suggested that such Machiavellian reasoning underlies all altruistic behavior, but this is empirically implausible for several reasons. First, animals seem to lack such calculative capabilities, and even people are not particularly good at such reasoning or following through in the face of temptation. An affective mechanism that does the job independently of such reasoning is thus more reliable (Sober and Wilson 1998, 312–321). Second, though the debate is on-going, psychological studies led by Daniel Batson

(1991; 2010) provide evidence that empathic concern can motivate people to help others even when failure to do so carries no psychological or other cost, suggesting that people with such motives genuinely desire that good of others for its own sake and not just, say, in order to feel better themselves. So, there are good reasons to think that affectively motivated psychological altruism is a genuine phenomenon.

1.2 Emotions and Evolution of Moral Agency

While some forms of altruistic behavior are widespread in our closest relatives, the motivation provided by altruistic emotions and desires is limited in scope and strength. Co-operation is intermittent, restricted to small groups, and vulnerable to temptation when selfish opportunities arise (e.g. Kitcher 2011), and maintaining peace requires constant effort that could be spent on other pursuits (de Waal 1982). Humans, in contrast, are *ultrasocial*, as it has become common to say. We cooperate intensively and extensively, relying on division of labor for many tasks from provision of food to provision of entertainment, deliberately teaching each other skills and share information that benefits recipients, respecting each other's property and dividing resources fairly among collaborators, providing some aid to strangers who we know will never reciprocate, paying taxes that will benefit people we may dislike, and so on (e.g. Tomasello and Vaish 2013; Henrich and Muthukrishna 2021, 219). Consequently, we're able to live fairly peacefully together in large societies in which we personally know only a small fraction of those we interact with.

What is it, then, that makes human ultrasociality possible? The consensus answer is that it is above all the existence of social norms, including moral norms, and the corresponding psychological ability for normative guidance, that enables us to overcome the limitations of altruistic desire and emotion in order to reap the benefits of broader and more reliable cooperation (e.g. Joyce 2006, Henrich and Henrich 2007, Kitcher 2011, Tomasello

2016). *Social norms* are shared standards for behavior that are enforced by sanctions meted out to those who fail to meet them. For example, if there is a strict social norm in a group to offer food to visitors even if one doesn't particularly feel like it, someone who fails to offer food incurs at least a risk of withdrawal of cooperation or communal protection, if not some type of punishment from publicly expressed disapproval to aggression and ostracism. The crucial thing is that once people who fail to act cooperatively are punished by *third parties* who do not themselves suffer harm, and cooperators rewarded, the evolutionary calculation changes – it no longer pays to withhold food from a visitor (or take someone else's things, or cheat on an exchange), if the price is, say, social exclusion (Boyd and Richerson 1992). Empirical studies have found that *altruistic punishment* is widespread across cultures, though its forms vary (Henrich et al. (eds.) 2004). For example, in experimental economics there is a robust finding that people are willing to pay money to have money taken away from people who have grabbed a lion's share of a windfall for themselves and left a stranger with little or nothing (Fehr and Gächter 2002).

For social norms to get off the ground, then, it is essential to explain why people are willing to engage in altruistic punishment that may not benefit themselves. Part of the explanation is that some forms of it may not be so costly – after all, if someone has a *reputation* as a cheater, withholding cooperation from them is also in your self-interest, just as it is for you to build a good reputation as a partner (Alexander 1987). What's more, punishing norm violators at a cost to yourself, like notable contributions to public good in general, is a good *signal* to others that you're a high-quality partner, which may open up new doors for you (Gintis et al. 2001, 116). And if there are other altruistic punishers in the neighborhood, they may be willing to punish those who are not willing to punish norm violators in spite of being in a position to do so (Blackburn 1998).

Still, people do engage in altruistic punishment even when it has a net cost to them – and of course, even when it is in their self-interest, they don't generally engage self-interested calculation before punishing. Rather, they punish thieves, for example, because they themselves consider theft to be wrong, and also refrain from stealing for the same reason. It is this capacity for what is sometimes called *normative guidance* that is essential for full-blown moral agency. And emotions seem to play a key role in its development. Philip Kitcher (2011) argues that normative guidance originates in obeying commands, which is initially motivated by *fear* of a dominant individual (see also Korsgaard 2010). In the case of humans, the ultimate dominant individual is a god, a watchful and powerful supernatural entity whose commands derive authority not only from fear but also the *awe* and *reverence* he or she inspires, even if belief in him or her is as a matter of fact most likely a false positive instance of our ability to detect purposeful agency around us (Atran and Noranzayan 2004). Obedience evidently requires a degree of self-control and inhibition of contrary desires in the subordinate. But there is of course more to normative guidance than pre-empting aggression, whether human or divine. At some point there must have been a transition from obeying a command by a stronger individual – "Don't eat the rabbit I killed!" – to obeying a norm, which is after all a kind of generalized command issued and enforced in the name of the group – "Don't eat animals killed by others!" – and doing so because one regards it as authoritative.

There are various theories about how and why this crucial step of developing a sense of obligation took place. To briefly sketch one influential (and controversial) account, Michael Tomasello (2016; 2020) proposes that moral agency emerged in two stages.[1] First, some change in the environment forced individuals to collaborate in foraging for food. This increased interdependence led to selection pressure in favor of *joint intentionality*, which

[1] For critical comments on Tomasello's account, see especially Roughley (ed.) 2018.

involves the ability to coordinate attention and take mutually interlocking steps towards a common goal (e.g. hunting an antelope). Tomasello argues that successful joint action requires mutual understanding of how each partner is supposed to perform their role and the ability to take each other's perspective. The hypothesis is that these cognitive achievements lead to both role-specific proto-normative expectations that are independent of particular individuals and a recognition that oneself and the other are in a sense equivalent: just as each partner is useful for me when they do their bit (e.g. chase the antelope toward me), I'm useful for them when I do my bit (e.g. throw a spear), and we both know that we both know this.

Particularly important for Tomasello's account is the idea that these new perspective-taking skills enabled the parties to think of themselves as part of a "we" and form attitudes towards others and themselves from that perspective (Gilbert 2006, Tuomela 2007). The benefits of committing to a shared project and, so to speak, identifying with the team then favor the development of primitive forms of the normative emotions of *resentment* toward those who fail to do their part or grab more than their share of the spoils, and *guilt* (or perhaps more plausibly *shame*, as Christoph Boehm (2012) emphasizes) for failing to do one's own. These emotions mesh in that the latter is triggered especially by a partner's resentment-fueled protest (cf. Darwall 2006). What is arguably crucial for these emotions is that they are *felt from a shared point of view that transcends one's own*: part of what it is to resent you for failing to do your bit (rather than just be angry or frustrated) is to take it that you, too, would be upset with me if I did what you have done, having indicated commitment to the joint project, and indeed that I'd then resent myself in your shoes. Since partners who are ashamed of failing to do their part are more reliable, potential cooperators who are sensitive to motives will prefer them to shameless ones, arguably leading to adaptive benefits (Rosas 2007, 690–691).

The second stage in Tomasello's story involves inter-group competition, which further increased interdependence within groups and hostility between them. This generated pressure for conformity within groups – doing things "our way", independently of any particular individual's beliefs or desires – and favored individuals who could adopt the group's perspective on their own actions. When I feel indignation, it's not just the feeling that you would feel the same towards yourself in my shoes but that *anyone*, or at least any of 'us', would feel. We are here at least approaching feelings felt from the point of view of what the 18th century sentimentalist Adam Smith (1790/2002) labeled the 'impartial spectator'. For Smith, imagining how any disinterested and well-informed third party would feel about our actions both sets a standard for what we should do and motivates us to do it even if it clashes with our individualistic preferences (for details, see Roughley 2018 and Kopajtic 2020).

Whatever the precise details of how we came to have fully moral emotions, it is important that the story explains how individuals genuinely *internalize* cooperative roles and social norms, that is, regard them as authoritative independently of punishment. As Kitcher puts it, even if conscience begins in fear, it "may later be dominated by shame or guilt, pride or hope, emotions available only in social environments where normative guidance, in some cruder form, has already taken hold" (2011, 94). The way these self-conscious emotions (Tangney and Dearing 2003) guide behavior differs sharply from fear. We don't refrain from violating an internalized norm because we anticipate feeling guilty and thus suffer a kind pain, but are rather repelled by the thought of performing the prohibited act, or simply fail to see the appeal – as it is sometimes put, whatever might speak in favor of doing it is *silenced* in deliberation (McDowell 1998).[2]

Moral agency, then, may well have its roots in the ability to feel self-conscious emotions from something like an impartial spectator's perspective. To be sure, as

[2] I emphasize this point, since appeal to anticipatory guilt is surprisingly common (e.g. Prinz 2007, Boehm 2012).

philosophers in the Kantian tradition emphasize, it hardly suffices for full-blown moral agency and responsibility to be able to govern oneself in light of *social* norms. Rather, for Kantians, it is only when we are able to *reflect* on what we have most *reason* to do and act in ways we could *justify* to any rational being that we become genuine moral agents (Kant 1785/1998, Korsgaard 1996, Scanlon 1998). It is certainly plausible that genuine moral agency requires the possibility of adopting a critical stance on actual social norms. The dispute between Kantians and sentimentalists concerns whether moving from internalizing the actual norms of a group to thinking about which norms we would endorse in somehow idealized conditions suffices for critical distance, as the latter argue (Kauppinen 2014a).

1.3 Emotions and Cultural Evolution

It is relatively uncontroversial that people have a disposition to internalize norms current in their society. It is, after all, an instance of more general cultural transmission, in which people model their beliefs and behavior on those around them, especially those with prestige (Richerson and Boyd 2005). Children will easily grasp what is okay around here and what's not, and distinguish between moral and merely conventional ones (Turiel 2006). It is a further question whether we're innately biased to pick up *specific norms*, including those that we now regard as stereotypically moral. Anthropology provides some evidence here. While some, like Jesse Prinz (2007), emphasize the diversity of moral systems, others, like Boehm (2012), note that at a deeper level there is striking uniformity. Many if not all contemporary hunter-gatherer bands actively propagate norms to reduce egoistic and nepotistic bias, such as norms against murder, theft, adultery, sorcery, and dishonesty, and for generosity and sacrifice for the group. (In-group bias is common, of course.) Recognizable variants of them can be found in the earliest historical documents as well as contemporary societies, though as Joseph Henrich (2020) emphasizes, modern large-scale Western societies are in many ways outliers

in virtue of the abstract, universalistic, and individualist cast of their versions of these moral norms.

Why do such cross-cultural similarities exist? One possibility would be that there are relatively tight innate constraints on moral norms (Mikhail 2011). But a more modest suggestion is that once the capacity to be guided by norms is in place, the norms that are most likely to survive over time will be those that match independent affective reactions like sympathetic concern and disgust (Nichols 2004; Nichols 2005, Kitcher 2011, 100–102) and the possibly prior normative stances of resentment and guilt or shame (Tomasello 2016). A further kind of explanation appeals to selection pressures at the group level. Since groups that have prosocial norms favoring reduced aggression, costly group defense, and mutual aid tend to fare better than those that don't, some argue that cultural group selection has favored the tendency to internalize norms with such content (Boyd and Richerson 1992, Joyce 2006, Henrich & Henrich 2007). Joseph Henrich, for example, argues that cumulative gene-culture co-evolution plays a role here. Roughly, once prosocial norms are in force, those with innate dispositions to violate them lose allies, partners, and opportunities via bad reputation and punishment, generating further selection pressure for norm-recognition, self-control, and reduced aggression (Henrich 2016, 79–80). This kind of co-evolution may explain some of the prosocial emotions, like sense of fairness, are already found in small children (Hamann et al 2011).

In sum, then, emotions may be essential for explaining altruistic behavior in humans (and close relatives), our capacity for normative guidance, and the survival of some norms over time. When people sacrifice for others, it may be out of love, out of feeling that they must do so to live with their conscience, or possibly out of pride for contributing to the community. Even though the ultimate explanation for why such emotions exist may be that they are

fitness-enhancing in an ultrasocial species, it does not follow that acting out of them is in any way selfishly motivated.

2. Emotions as Causes of Moral Judgment

Even if emotions have played a crucial role in the emergence of moral thought and behavior, it might be that we have now transcended such influences and make our judgments about right and wrong by reasoning, at least as adults. Indeed, one might be forgiven for getting such an impression from much 20th century theorizing (Kohlberg 1976). But recently winds have changed, and psychologists have come to emphasize the role of affect in moral judgment. Here it is good to bear in mind that talk of moral judgment can refer to many different responses, including *evaluations* something as good or bad or admirable, *deontic beliefs* about rightness or wrongness of actions, *hypological appraisals* of praise- or blameworthiness of agents, and *character assessments* as virtuous or vicious (cf. Malle 2020). A closer analysis might reveal that emotions are differentially involved in these different varieties of judgment.

An orthogonal distinction can be made between affects that are *incidental* or *integral* to judgment, where the former are not *about* the object of judgment and the latter are. For example, my evaluation of someone might be influenced by an incidental melancholy mood, or by an integral anger at what they've done.

To examine this issue, it is good to first distinguish between two types of psychological process, often called Type I (or 'intuitive') and Type II (or 'reflective'). The following table lists some of their characteristic features:

Type I	*Type II*
Doesn't require working memory	Working memory taxed
Automatic	Effortful, allows hypothetical thought

Only outcome is conscious	Conscious process
Fast	Slow
Parallel processing	Serial processing
Evolved early, similar to animal cognition	Evolved late, distinctively human
Often affective	Often draws on explicit rules, may override affective response

As Jonathan Evans and Keith Stanovich (2013) argue, these features don't always cluster together. Type I processing, in particular, is diverse, including plausibly evolved modules like estimation of distances and learned habits like multiplying single-digit numbers.

Drawing on a broad body of research, Jonathan Haidt (2001) argues that *most* moral judgments of *most* people are the result of intuitive rather than reflective processes. More specifically, he holds that intuitive moral judgments result from "a subclass of automatic processes that always involve at least a trace of 'evaluative feeling'" (Haidt and Kesebir 2010), where such feelings may be either full-blown emotions like disgust or unnamed flashes of affect.

There are many sources of evidence for this affective primacy hypothesis. In general, people seem to evaluate and categorize each other automatically and instantly, without being conscious of what they're doing, and these evaluations have a lasting effect (e.g. Zajonc 1980). Many neuroscientific studies show that areas of the brain associated with social emotions, such as the ventromedial prefrontal cortex, are consistently active when people make moral judgments (Pascual, Rodrigues, and Gallardo-Pujol 2013). Unsurprisingly, people whose emotional capacities are damaged struggle with moral evaluations (Anderson et al. 1999). Relatedly, there are persistent questions about the ability of psychopaths to make moral judgments even if their rational capacities are intact, making the absence of affective

empathy the likeliest candidate explanation (Aaltola 2014). Affective responses such as disgust predict people's moral judgments even in the absence of concrete harm (Haidt, Koller, and Dias 1993). Manipulating incidental affect, for example by creating a disgusting environment, has been found by some to make a difference to people's judgments (Schnall et al. 2008, Valdesolo and DeSteno 2006) – though such results have not replicated well, and the effect may disappear entirely when publication bias towards noteworthy results is taken into account (Landy and Goodwin 2015). And when people do make wrongness judgments on such affective basis, they sometimes hold on to them even if they can't articulate a harm-based rationale – a phenomenon that Haidt labels 'moral dumbfounding' (Haidt 2001).

When it comes to explaining these feelings, Haidt endorses a form of moral nativism, according to which a variety of moral affect programs have been selected for. He emphasizes that it is not only emotions like empathy and anger that are fitness-enhancing, but also feelings of reverence and contempt, which promote the smooth functioning of hierarchical authority in groups, and feelings of disgust, which play an important role in the social control of bodies and waste, or more abstractly purity (Haidt and Joseph 2004). Different cultures and subcultures encourage different kinds of moral response and tamp down others – most notably, Haidt maintains that liberal Westerners shun responses based on disgust and contempt, while conservatives are comfortable with feelings linked to authority, purity, and sanctity (Graham, Haidt, and Nosek 2009; cf. Shweder, Mahapatra, and Miller 1987).

Haidt acknowledges that explicit reasoning has a role in moral thinking, but holds that it is typically a *post hoc* attempt to rationalize the judgments that one has already made because of affective reactions. People look for confirming evidence, and can't be argued out of identity-defining moral convictions. For him, the function of reasoning is largely to convince others that we're morally good and those we dislike are morally bad (Haidt and Kesebir 2010). So it is unsurprisingly typically motivated and biased. Some convergent

evidence for this is provided by studies of political polarization. For example, Dan Kahan and colleagues found that "reliable employment of more effortful, conscious information processing will *magnify* the polarizing effects of identity-protective cognition" (Kahan 2013, 410) – that is to say, the more people engage in reasoning, the further apart they drift if their emotional starting points differ.

For Haidt, we may also try to persuade others to do what we want by offering what look like reasons, but are actually designed to simply push the right buttons. Only in exceptional circumstances some individuals, such as philosophers, may engage in dispassionate private reflection and rational argument (Haidt 2001, 829).

Haidt's social intuitionism is a bold synthesis that has come under criticism from many directions. Generally speaking, while the idea that both intuitive and reflective processes play a role in moral judgment is now widely accepted (Cushman, Young, and Greene 2010), Haidt's particular interpretations of their nature and roles are challenged. For example, the so-called 'moral grammar' approach (Huebner, Dwyer, and Hauser 2009; Mikhail 2011) holds that the relevant unconscious Type I processes involve complex analysis of the morally relevant features of the situation one is judging (such as whether the action was harmful, whether the harm was intentional or a side effect, and so on). This is followed by an unconscious application of innate moral principles operating on the outputs of the analysis, which yields a moral judgment, such as condemnation of harming another as a mere means to one's own end. On this view, then, emotions may contingently arise as a *consequence* rather than the cause of an already formed judgment. This type of rejoinder is also offered by more traditional rationalists like Josh May (2018), who appeals to the sensitivity of moral judgment to subtle differences in intentionality and outcomes of an action as evidence for *unconscious reasoning*. However, as he concedes, even if moral judgments *accord with* principles of the sort that philosophers articulate explicitly, it doesn't follow that they *result from* reasoning

with them (2018, 70). May also notes that the observed effects of incidental emotion in particular are weak for moral judgment. Even if true, this does not bear much weight against versions of affectivism that highlight *integral* emotions.

Others who endorse a form of affective primacy nevertheless emphasize that affective reactions may reflect social learning and *sensitivity to reasons*, forming what Peter Railton calls a "flexible, experience-based information-processing system quite capable of tracking statistical dependencies and of guiding behavioral selection via the balancing of costs, benefits, and risks" (Railton 2014, 833). For example, Railton suggests that negative affective reactions toward harmless sibling incest may track the *risk* of psychological harm, parallel to negative reactions towards harmless Russian roulette (Railton 2014, 848–849; cf. Jacobson 2012).

On the reasoning side, many have found Haidt's emphasis on rationalization excessively cynical. Paul Bloom and David Pizarro (2003) emphasize that people can and sometimes do take conscious steps to shape their affective reactions, and that engaging in explicit deliberation is necessary when faced with complex situations in which gut reactions conflict. Fiery Cushman and colleagues (Cushman, Young, and Hauser 2006) found that people's judgments do sometimes match the principles they consciously endorse, leaving it open that conscious reasoning plays a causal role in judgment. Indeed, priming people to engage in reflective thought prior to moral judging makes it more likely that they will go against their initial intuition in moral dilemmas (Paxton, Ungar, and Greene 2011). And some processes that may make a difference to moral judgments, such as deliberate emotion regulation, seem to straddle the divide between emotion and reason, since they involve overriding one's initial affective response to achieve aims like interpersonal coordination (Helion and Pizarro 2015, Kauppinen 2014b).

A different type of challenge maintains that different types of judgment involve different kinds of process. Joshua Greene and colleagues studied people's responses to the so-called trolley cases (Thomson 1976), in which the moral permissibility of actions that result in the same outcome seems to depend on *how* it is brought about. Briefly, in the Switch case, a person must choose between letting a runaway trolley kill five people and saving the five by turning it to a side track, which has the unfortunate side effect of killing one person stuck there. Most people think it is permissible to turn the trolley nonetheless. In the Footbridge Case, a person must choose between letting a runaway trolley kill five people and saving the five by pushing a large person from a footbridge onto the tracks to stop the trolley at the cost of their life. Most people think it is impermissible to push, even though the outcome of saving five at the cost of one life is the same as in Switch. Most people's verdicts about the Footbridge case thus align with so-called *deontological* views in normative ethics, according to which bringing about the most good is not always permissible, for example if it involves using someone as a mere means (see Kamm 2016 for an up-to-date account). Alternative *utilitarian* views hold that we must always do as much good as we can, and thus sacrifice one person with a life worth living to save five other people with lives worth living, other things being equal (Singer 2005).

Greene and colleagues found that in Footbridge, emotion-related areas of the brain are active when people decide against pushing (Greene et al. 2001), that only the verdicts of those who decide to push are slowed down when cognitive load is increased (and thus reflection made harder) (Greene et al. 2008), and that people with emotion deficits are more likely to sacrifice the one person (Koenigs et al 2007). From such data, Greene (2014) concludes that the deontological verdict of refusing to sacrifice one to save many is the result of an 'alarm-bell like' negative affective response aroused by the thought of 'up-close and personal' violence (or personal use of force), which would have been adaptive in our

evolutionary past. In contrast, the minority utilitarian verdict (pushing one is required) is the result of Type II reasoning from cost/benefit principles, which are themselves rooted in 'currency-like' affective responses that allow for tradeoffs and are not sensitive to factors like physical distance.

However, research by Guy Kahane and colleagues (Kahane et al. 2012, Kahane 2012) has seriously called into question the hypothesis that it is the deontological character of some judgments that explains Greene's data. Their findings suggest that it is when judgments are *counterintuitive* (which both utilitarian and deontological judgments can be in suitable contexts) that making them is associated with controlled processing, which overrides the initial intuition. For example, the deontological philosopher Immanuel Kant held that one should never lie, even to someone seeking to murder a person in hiding (Kant 1795/1996). The (very few) people who agree with Kant manifest the same sort of relatively slow and deliberate processing as those who make a utilitarian judgment about the Footbridge case in Greene's studies. The association of utilitarian verdicts with controlled processing in trolley cases has itself been called into question, as some studies find that time pressure actually increases the tendency to sacrifice one to save many (Rosas and Aguilar-Pardo 2019). Moreover, the decision to sacrifice one to save many is linked with caring less about harm to the one rather than caring more about the five, suggesting that it is not the utilitarian concern to bring about the best possible outcome that explains people's willingness to sacrifice one to save five but rather the absence of empathy for the one person (Gleichgerrcht and Young 2013).

In conclusion, it is fair to say that there is no consensus at present regarding the extent of affective influence on moral judgment, though it is certainly larger than many used to think. Since our evolved emotional responses, like sympathy and sense of fairness, may

conflict, it is hard to deny that reasoning must also play a role in at least some moral judgments – but the starting points of such reasoning may still be explained by emotion.

3. Emotions as Contents or Constituents of Moral Judgment

In the preceding section, I discussed debates about how affective responses *influence* moral judgment. But for sentimentalists in *metaethics*, the branch of philosophy that studies the nature of moral language, thought, truth, and knowledge, emotions have a more fundamental role. On many such views, values themselves are in some way dependent on human sentiments. In this section, I examine arguments for and against claims that emotions either are referred to in the *content* of moral judgment (in which case their truth or falsity, and thus moral facts, may depend on our affective dispositions) or at least in part *constitute* moral judgments. The moral judgments in question may involve general moral concepts, like *wrong* and *bad*, or response-specific ones like *admirable, desirable, contemptible*, and *blameworthy*.

3.1 Emotions as Contents of Moral Judgment: Dispositionalism and Neo-Sentimentalism

A natural way into a sentimentalist view of moral judgments begins with response-specific concepts like *admirable* or *blameworthy*, whose contents evidently make some kind of reference to emotional responses of admiration and blame, respectively. If more general concepts like *good* and *wrong* can be analyzed as making a more indirect reference to such responses, the sentimentalist analysis will extend to them as well. It is, after all, plausible that what is good merits *some kind of* positive attitude (Broad 1944), and what is wrong is typically also blameworthy (though when the agent has an excuse like invincible ignorance, they may also do wrong blamelessly) (Gibbard 1992, Kauppinen 2017). In this vein, the German philosopher Franz Brentano suggested in 1889 that "the good is that which is worthy of love, that which can be loved with a love that is correct" (1889/1969, 18).

Dispositionalism

Moral sentimentalists often appeal to an analogy to aesthetic properties like *being funny* (e.g. McDowell 1998, Wiggins 1998). Being funny is evidently somehow tied to human responses – it hardly makes sense to say that Chaplin's early short films would have been funny, had they somehow come into being in a world without human beings, whose senses of humor they as a matter of fact happen to tickle. Perhaps, then, when we say that something is funny, we are tacitly saying that it is such as to amuse us when we attend to it in suitable circumstances (in which we are not in any condition that gets in the way of amusement, like depressed or preoccupied). This is a claim about the *concept* of being funny, and thus about the *content* of our judgments about being funny, whether in thought or talk. If it is correct, then something *is* funny if (and only if) and because it amuses us when we attend to it and don't happen to be feeling down (etc.) – after all, apart from amusing us, what else is common to all the funny things from puns and memes to wry observations and slapstick? This second claim concerns the *metaphysics* of being funny, in particular about how facts about funniness relate to facts about other things. According to it, for Chaplin's films to be funny is nothing other than their being such as to amuse people in suitable conditions – a perfectly natural fact that depends on human beings having the affective dispositions they happen to have.

What would moral judgments look like on this model? One influential suggestion was made in the 18th century by David Hume, who said: "An action, or sentiment, or character is virtuous or vicious; why? because its view causes a pleasure or uneasiness of a particular kind" (Hume 1739–40/1978, 471). This basic idea can be developed in various ways. On *subjectivist* and *cultural relativist* views of moral language, if I *say* that something is wrong, I'm saying that it is such as to give rise to disapproval in me or in my culture, or that it goes against the norms accepted in our group (Westermarck 1903, Harman 1975, Prinz 2007).

Correspondingly, something *is* wrong for me or relative to us when it really is such as to cause disapproval in me or us. Because these views refer to people's dispositions to have affective reactions, they are sometimes called *dispositionalist* forms of moral sentimentalism. Just as people with different tastes and cultural backgrounds find different things amusing, so that different things really are funny in different cultures, subjectivists and relativists hold that different things really are right or wrong for different people or cultures, if they feel differently about them. Relative to us, cannibalism is morally wrong, but relative to cultures that approve of it, it is not. On these views, then, there are moral facts and truth, which are reducible to natural facts about people's emotional dispositions, and thus relative to individuals or cultures.

Many philosophers think, however, that such simple dispositionalist views fail to capture what we mean when we think and talk about right and wrong. One famous challenge, forcefully articulated by the influential early 20th century British moral philosopher G.E. Moore (1912/2005, 50–52), concerns the *possibility of moral disagreement*. Suppose that I'm having a chat with a cannibal. If I say "Cannibalism is wrong" and she says "Cannibalism is not wrong", we seem to disagree. Consequently, at least one of us must be wrong. Yet if what I'm saying is really that I am (or that people in my culture are) disposed to disapprove of cannibalism, and what she's saying is that she isn't (or that people in her culture aren't) disposed to disapprove of cannibalism, we *don't* disagree. After all, I agree that she doesn't disapprove of cannibalism – that's precisely my problem with her – and she agrees that I do disapprove of it. So if relativism were right, we could both be correct at the same time. Since we do seem to disagree, however, it seems we must each be making a claim about something more than our own response-dispositions.

Some dispositionalists try to avoid such problems by saying that moral facts are after all not like facts about being funny, but rather like facts about *color*, which are at the same

time both response-dependent and objective. After all, canaries are yellow, regardless of what any of us think, but at the same time, being yellow is a prime example of a dispositional property that depends on human response-tendencies, or what the 17th century philosopher John Locke (1690/2008) termed a *secondary quality*. Plausibly (and as a first approximation), for canaries to be yellow just is for them to be such as to give rise to experiences of yellowness in human observers with normal color vision in normal daylight – generally speaking, in circumstances that are suitable for making color judgments. So maybe when we say that an act is morally wrong, we're saying that it is such as to cause disapproval in any statistically normal person when they're in circumstances that are suitable for moral judgment – for example, when they're not biased, ignorant, or suffering from some incidental emotional disturbance. This is one way to interpret Hume, who explicitly made the analogy with color. If so, moral truths depend on the right kind of responses in the relevant circumstances. For Hume and many others, these circumstances are those in which people are *well-informed* about the facts of the case and *impartial* in the sense of setting aside their own particular interests and attachments before responding emotionally (Hume 1751/1948, Smith 1790/2002, Firth 1951, Cohon 2008). If there is sufficient uniformity in statistically normal sentimental responses in the suitable circumstances (which is far from obvious), moral truths will nevertheless be as objective as truths about color.

Dispositionalism and Normativity

Perhaps, then, there are forms of dispositionalism that avoid problems with disagreement and objectivity. But they still face a second major challenge, which is accounting for the seeming *normative purport* of moral talk and, correspondingly, the normative force of purported moral facts. When philosophers talk about normativity, the decidedly do not mean that something is *normal* in a group, as psychologists sometimes use the term, or even that it is regarded as

right by someone. Instead, if something is genuinely normative, it is something there is *reason* to do or something that *justifies* an action or something that we *should* or *ought* to do, regardless of whether anyone actually does so or thinks it's right (e.g. Parfit 2011). The challenge, then, is that when we say that cannibalism is morally wrong, it seems we're saying that *should* disapprove of and avoid it. But the mere fact that someone, or even everyone, *does* disapprove of cannibalism doesn't entail that anyone should disapprove of or avoid it.

Here are two famous arguments for this claim about moral concepts and corresponding facts. The first is a version of what is known as the Open Question Argument, also found in G.E. Moore's (1903) seminal work on modern metaethics. It's starting point is the simple assumption that if a claim about something that has feature F having the feature G as well is true by virtue of meaning of the terms for F and G, there is no conceptual room for asking of something that is F whether it is also G – the question is 'closed', as it were. For example, since "vixen" just means "female fox", it betrays conceptual confusion to grant that Rosa is a vixen and then go on ask "But is Rosa a female fox?". But it *does* make sense to grant that stealing from a beggar is apt to give rise to disapproving moral sentiments in me or us or any normal human observer and nevertheless ask "But is it morally wrong to steal from a beggar?". Someone who asks such a question does not betray conceptual confusion (although they may qualify as obtuse) in the way that someone who wonders whether vixens are female foxes does, so it can't be that the *concept* of being disposed to arouse disapproval is the same as the concept of being wrong. Generalizing from this, it seems that whatever disposition to arouse sentimental responses we attribute to an action, it always remains an open question whether it is right or wrong. Why? According to the critic of dispositionalism, it's because our moral talk is about something other than our having a disposition to feel one way or another (Shafer-Landau 2003). To put it differently, it seems it always makes sense to

ask for *reasons* for having a sentimental response towards something, and it won't suffice for an answer to say that we're disposed to feel that way (Brady 2013).

The second argument against dispositionalism has its roots in Plato's dialogue *Euthryphro*, in which Socrates asks whether something is pious because the Gods love it, or whether the Gods love it because it's pious, settling on the latter. Applied to dispositionalism, the argument says that it can't be the case that something is morally right because we're disposed to approve of it, since that would mean that were we to change so that we acquire a disposition to approve of, say, torturing children, it would become right to torture children. But, the argument goes, torture *wouldn't* become right. Instead, in this scenario *we* would have become horrible people (Blackburn 1984). Consequently, what is right or wrong must be independent of our attitudes, since variation in right and wrong doesn't track variation in attitudes. This is one point at which the analogy between ethics and humor may fail – perhaps things really are funny just because we as a matter of fact tend to be amused by them.

Error Theory

At this point, there's various directions one might go. *Moral realists* of various stripes simply reject sentimentalism in all its forms, and hold that there are objective moral facts independently of our response-dispositions (Shafer-Landau 2003, Enoch 2011). A more radical option is taken by *error theorists* about ethics (Mackie 1977, Joyce 2001), who accept our moral judgments *purport* to describe mind-independent and irreducibly normative properties that would be reason-giving for everyone, but deny that there *are* such properties. They would have to be 'queer', utterly unlike anything science can vouch for. For them, we have no reason to believe there are such 'non-natural' properties. Consequently, all (positive) moral judgments are systematically in error.

For error theorists, emotions enter the story by way of explaining why we make this error: we have a tendency to *project* our subjective responses onto the world, "gilding and staining with colours borrowed from sentiment", as Hume put it in one of his moods (1751/1948, Appendix I). Mackie and Joyce maintain that this projection is a *useful fiction*, which we should maintain for practical reasons even after we've realized there are no moral facts. For example, because we mistakenly think there are objective reasons against actions like cheating or killing weak competitors independently of our interests or desires, we avoid doing things that would be against our interests or desires in the long run. As discussed in the first section, having a sense of moral obligation promotes cooperation that is generally beneficial in the long run. So, error theorists say, we should keep pretending that some things really are wrong and inculcate dispositions to morally condemn them, even though there is no right or wrong.

Neo-Sentimentalism

However, there is also a distinctively sentimentalist line of response to the normativity challenge. According to *neo-sentimentalists*, moral concepts are concepts of *fitting* or *merited* sentimental responses (McDowell 1998, D'Arms and Jacobson 2000, Tappolet 2016). For example, to think of someone as admirable is to think that admiration is fitting towards her. Correspondingly, someone is admirable if and only if admiration *is* fitting towards her. More generally, as David Wiggins puts it, "x is good/right/beautiful if and only if x is such as to make a certain sentiment of approbation appropriate" (1998, 187). Views of this kind aim to dispel the alleged queerness of moral properties by showing how they mesh with fitting sentimental responses.

The obvious question for neo-sentimentalists is what it is for a sentimental response to be fitting or merited. Early forms of neo-sentimentalism articulated by John McDowell

(1998) and David Wiggins (1998) tended to analyze fittingness in terms of a *virtuous person's* responses. They draw in particular on Aristotle, who emphasized in the *Nicomachean Ethics* that virtue is manifest in having the right emotional responses:

> For example, fear, confidence, appetite, anger, pity, and in general pleasure and pain can be experienced too much or too little, and in both ways not well. But to have them at the right time, about the right things, towards the right people, for the right end, and in the right way, is the mean and best; and this is the business of virtue. (*NE* 1106b)

On such a view, then, it is fitting to, say, admire someone if and only if a virtuous person would admire them when suitably informed. The question that critics are apt to ask at this point is just *why* a virtuous person would admire someone like Nelson Mandela – after all, for Aristotle, virtue seems to involve *recognizing* what is good or bad independently of us. McDowell and Wiggins are perhaps surprisingly relaxed about this issue, holding that neither virtuous responses nor values have priority, but are rather 'siblings' (McDowell 1998). The aim isn't to reduce values to dispositions to give rise to virtuous emotional responses, but rather to 'elucidate' their nature by displaying the connection (Wiggins 1987). This may well not be satisfactory response to someone who finds the purported objectivity of values puzzling.

Instead of appealing to virtuous responses, neo-sentimentalists often suggest that responses are fitting when there is *sufficient reason* to have them (D'Arms and Jacobson 2000). The basic challenge for this type of view is to distinguish between what are called *right and wrong kind of reasons* for attitudes (Rabinowicz and Rønnow-Rasmussen 2004). For example, suppose that you would be tortured unless you admire a terrible dictator, who possesses the latest in emotion-recognition technology, so that he can very reliably tell

whether you are faking it. On the face of it, this is a sufficient reason for you to admire the dictator. But it doesn't seem to make it *fitting* to admire the dictator, since the dictator isn't admirable because he will torture you if you don't admire him – quite the opposite! That's what makes threats and rewards the wrong kind of reasons for attitudes. The problem, then, is to distinguish between right and wrong kinds of reason on principled grounds.

One plausible line of response is to say that the dictator's threat is only a reason to *want* to admire him, but not a reason to admire him, and therefore not a wrong kind of reason to do so either (Gibbard 1992, 37; Way 2012). Instead, only those reasons to which we could directly respond to by admiring count as right kind of reasons for doing so – for example, that the person has courageously rescued someone is a right kind of reason to admire, because it is a reason *for which* we can admire someone. Avowedly *anthropocentric* theorists like D'Arms and Jacobson (2014) highlight the 'we' part here. For them, features like being associated with the Nazi doctor Josef Mengele really are reasons to be disgusted with an object just because they reflect a deep and wide human concern with the histories of objects around us. While such contingent concerns are not beyond criticism in light of our other values, they do have rational significance.

Another possibility is to analyze fittingness in terms of *correctness*, so that Christine Tappolet (2016) develops what she calls *representational neo-sentimentalism*, which embraces this feature. For her, emotions have representational content that is akin to perceptual content, so that admiration, for example, represents its target as admirable. Consequently, it is fitting in the sense of being *correct* if and only if, and because, something is independently admirable (Tappolet 2016, 87). On such a view, while (at least some) evaluative *concepts* make reference to sentimental responses, evaluative *properties* or *facts* like who is or isn't admirable do not in any way depend on our response tendencies. They

thus presuppose some form of metaphysical realism about value, and face the challenge that error theory poses to the existence of such irreducible normative properties.

3.2 Emotions as Constituents of Moral Judgment: Emotivism and Expressivism

One response to the challenges that arise for both dispositionalist and neo-sentimentalist views is to make a radical move regarding moral language: perhaps it's not in the business of describing how things are in the world in the first place. If it has some non-descriptive function, it won't assume the existence of any mysterious normative and reason-giving properties. What is more, there is independent motivation for thinking this way: the whole point of making and uttering moral judgments seems to be guiding action, including responses to the actions of others. Moral discourse seems to be essentially *practical*, and thus crucially different from judgments simply representing how things as a matter of fact *are*. For such reasons, *emotivists* like A. J. Ayer (1936) held that moral utterances are not true or false, but only evince attitudes, rather like saying "boo!" to torture or "hurrah!" to charity. Charles Stevenson (1944) emphasized that such emotional expressions also serve the crucial function of influencing the attitudes of others, famously suggesting that calling something good just means "I approve of this; do so as well!". And Richard Hare (1952) suggested that moral assertions are disguised *imperatives*, serving to commend or command something rather than describe them.

As accounts of moral language, however, emotivism and its cousins are highly limited. They can't account for the contribution that moral terms make to the meaning of sentences that don't involve booing or hurrahing anything. Consider conditional sentences, as Peter Geach (1965) did. If I say "If Clinton gets the most electoral college votes, she will be elected President", I'm not asserting that Clinton gets the most electoral college votes – I'm entirely neutral on that. Similarly, if I say "If paying taxes is wrong, I'm going to hell", I'm

not taking any stance on paying taxes, and certainly I'm not booing paying taxes. Generally speaking, when moral predicates are used in such *embedded contexts*, they clearly don't serve to express attitudes. Yet they must have the same meaning they do in standalone uses, since we can construct perfectly valid inferences with embedded moral propositions, as Geach emphasized. For example, if the above moral conditional is true and it really *is* wrong to pay taxes, then it follows that I'm going to hell, by application of the basic *modus ponens* inference rule (which says that for any two propositions *p* and *q*, *p* and *if p, then q* together entail *q*).

Contemporary *expressivism* aims to do justice to the practicality of moral thought while accounting for the seemingly descriptive features of moral language. Allan Gibbard (1990), for example, holds that to think something is morally wrong is (roughly) to think that it is rational or appropriate to feel guilt for doing it and angry with others who do it, unless they have an excuse. To think that something is rational, in turn, is to accept norms that allow or prescribe it, or plan to do so in certain circumstances. In the same vein as the views discussed in Section 1, Gibbard holds that our evolutionary history explains why such attitudes of norm-acceptance are defeasibly motivating.

When it comes to moral language, in saying "Cheating is wrong", I am *expressing* such a complex attitude rather than *reporting* it. Moral utterances aren't *about* our attitudes, but rather vehicles for manifesting them to others for guidance and discussion. They have a propositional form, because this allows us to exploit the resources we already have for discussing the truth and implications of non-normative claims (Blackburn 1998). Indeed, contemporary expressivists tend to agree that moral sentences can be true, and that there are moral facts. On the background is a deflationary view of truth and facts, according to which it doesn't add anything to saying that torture is wrong to say that it is *true* that torture is wrong – both utterances express commitment to blaming for torture, or some such. This doesn't by

itself, however, suffice to solve Geach's problem of how to account for the meaning of moral terms in contexts in which they don't serve to express attitudes. It is a topic of lively debate, with a wide range of possible solutions offered (Schroeder 2010, Ridge 2014).

4. Does Emotional Influence Undermine the Justification of Moral Judgements?

In ordinary life, we sometimes question each other's beliefs by saying things like "You only think so because you're angry" and recognize that emotions can lead us astray. In this vein, some philosophers have argued that if our moral beliefs result from affective influence, they are not justified, or that they don't amount to moral knowledge, even if we assume that there is such a thing as moral truth. Others, however, defend the optimistic view that emotions can be guides to moral truth and even sources of moral knowledge.

4.1 Affective Debunking Arguments in Ethics

Suppose you believe that Jennifer is a bad person. If asked for reasons why, you'll say that there's something sinister about the way she speaks. But then you discover that you have been primed to feel incidental disgust at the thought of her, and that this is apt to give rise to the kind of thoughts you have regardless of what the person is actually like. Such news about the causal history of your belief plausibly undermines your justification for thinking that Jennifer is a bad person. Some would say that you never had any justification to begin with, because your belief was as a matter of fact formed as a result of an unreliable process (e.g. Goldman 1979). What does the epistemic work here isn't simply that your belief results from an emotion, but that the incidental emotional response itself doesn't in any way *track* the way things are, either from your own perspective or as a matter of fact. Instead, it results from what is called an 'off-track' process (Kahane 2011).

People who endorse so-called *debunking arguments* in ethics take a similar line of thought to support skepticism about the justification of moral beliefs. Here is a version of this type of argument, modelled on Vavova (2021):

Affective Evolutionary Debunking

1. Influence: Some of our key moral beliefs are the result of unreflective (Type I) affective reactions, or reasoning strongly biased by such reactions. (See Section 2.)
2. Off-track: Type I affective reactions are strongly shaped by evolutionary and cultural selection pressures that promote fitness-enhancing rather than truth-tracking responses. (See Section 1.)
3. Off-track influence: So, some of our key moral beliefs are influenced by processes that aim at fitness rather than moral truth (1, 2).
4. Gap: True moral beliefs and adaptive moral beliefs come wide apart. (This assumption is necessary for the argument, because sometimes it is the case that beliefs are adaptive *because* they are for the most part true – there isn't much of a gap between adaptive and true perceptual beliefs.)
5. Accidentality: So, some of our key moral beliefs are influenced by processes that lead to true beliefs at best accidentally (3, 4).
6. Bad influence: If a belief results from a process that leads to true belief at best accidentally, it is likely to be mistaken.
7. Skeptical conclusion: So, at least some of our key moral beliefs are likely to be mistaken (3, 4, 5). Consequently, they are not justified, and we should give them up.

Note that as I've formulated the argument, it doesn't cover only beliefs that are directly caused by affective responses, but also those that result from reasoning that is biased by such responses, for example because affective evaluative tendencies explain why some of its

premises are endorsed. Maybe I consciously reason to the conclusion that what Julie and Mark did is morally wrong from the premises that Julie and Mark engaged in incest and incest is always morally wrong; still, if I only believe that incest is always wrong because of an evolved affective response to incest, the argument says that my belief about Julie and Mark is likely to be mistaken, because it would have to be a massive coincidence that a process that favors genetic or cultural survival resulted in true moral belief. Its scope depends on how wide the influence described in the first premise is – either it applies selectively to only some of our moral beliefs, or to all of them.

Selective debunking. Joshua Greene and Peter Singer (2005) take the experiments described in Section 2 to show that only *deontological* judgments, such as the verdict that we shouldn't use one person as means to save many in the trolley cases, are the result of 'alarm-bell like' affective responses or misfiring heuristics, that is, affectively encoded rules of thumb that work or at least used to work in most cases (Sunstein 2005). Consequently, only they are likely to be mistaken and thus are unjustified, at least now that we know the etiology of our beliefs. In contrast, they believe that *utilitarian* judgments result from unbiased reasoning, specifically impersonal cost-benefit analysis, which allows us to transcend our Pleistocene legacy (Singer and de Lazari-Radek 2012). This would support utilitarianism's claim to correctness.

Of course, as I've already pointed out, there are empirical doubts about Greene's and Singer's construal of the first premise. What is more, it is far from clear that utilitarian reasoning is independent from indirect evolutionary influence. Negative feelings towards causing suffering and positive feelings towards causing happiness have a similar adaptive rationale as other evaluative attitudes, and so does the sympathy that might extend our attitudes beyond our own circle. So if there's an epistemic problem with evolved emotions, it

may also undermine the justification of utilitarian reasoning that relies on beliefs based on them (Kahane 2014).

Comprehensive debunking. Others, such as Richard Joyce (2006), take the debunking argument further, and maintain that *all* of our moral judgments are susceptible to evolved biases, and therefore unjustified. The basic principle he draws on is that if we can explain our belief that S is P without assuming that any S actually is P, the belief is unjustified. For example, if you discover that you only believe that Napoleon lost at Waterloo because you were some time in the past given a pill that causes people to have that belief regardless of what actually happened at Waterloo, you would lose your justification for believing that Napoleon lost at Waterloo (2006, 179). (This is not evidence that Napoleon won at Waterloo, of course.) Since Joyce believes that all our moral beliefs can directly or indirectly be explained in terms of fitness-enhancing emotional responses without assuming their truth – natural selection is, as it were, a pill that gets us to believe that it is wrong to cheat our partners, among other things, regardless of whether or not it *is* wrong to cheat our partners – he takes moral skepticism to follow across the board: none of our positive moral beliefs are justified.

Non-Skeptical Responses

Debunking arguments have been rejected for many reasons. Some reject premise 1 and maintain that moral reasoning can transcend its evolutionary origins (see again Singer and de Lazari-Radek 2012). Another common response is to reject premise 4, the Gap assumption. For example, David Enoch (2010) argues that in evaluating whether fitness-enhancing processes are truth-tracking, we must make some assumptions about evaluative truths – indeed, as Katia Vavova (2021) notes, the Gap premise itself makes such an assumption, namely that whatever the evaluative truths are, they come wide apart from what is adaptive.

One putative evaluative truth is that promoting your own and your loved ones' survival is at least somewhat good. If that's the case, when naturally selected attitudes lead us to place value on our children's survival, they lead us in the right direction, because there isn't a major gap between adaptive and true moral beliefs. Of course they're not the same thing – it might be adaptive to think that the neighboring tribe should be eliminated if there aren't enough resources for both, but few would claim it's morally right to do so – but there is enough overlap between what is adaptive and what is true for us to get started in corrective reasoning, even if its outcome reflects its starting points (Enoch 2010, 428). This is a natural approach for those who think that there are objectively normative and therefore non-natural moral facts independently of our opinions.

Since these responses to debunking are not distinctively sentimentalist, I'll set them aside and focus on premise 2, Off-Track. How could it be false? Well, seemingly the only option if the pervasive influence of evolved affects is granted is to hold that moral truth itself is shaped by evolution. And indeed, Street defends a form of metaethical constructivism, according to which evaluative truths are determined by our attitudes, more specifically by our values (Street 2010, 370–374). However, it is not our actual, evolution- and culture-shaped attitudes that make something right or wrong, but those that survive a process of seeking coherence among them. As she puts it:

> [E]valuative truth is a function of how all the evaluative judgements that selective pressures (along with all kinds of other causes) have imparted to us stand up to scrutiny in terms of each other; it is a function of what would emerge from those evaluative judgements in reflective equilibrium (Street 2006, 154).

So helping our neighbor is morally right (at least for *us* in our community), because a favorable attitude toward it survives reflective scrutiny in the light of our other commitments. On this view, then, debunking arguments only target mind-independent moral realism.

4.2 How Emotions Might Justify Moral Judgments

Suppose that one of the non-skeptical responses is right, and that judgments that are influenced by our evolved affective tendencies are not likely to be false. Could it actually be the case that emotions can positively *inform* us of what is right or wrong or good or bad? While many schools of philosophy, such as the ancient Stoics and contemporary Kantians are famously skeptical, there has long been a minority epistemic sentimentalist tradition. Notably, early 18th century British philosophers Shaftesbury and Hutcheson argued that we have an innate moral sense, analogous to other senses, which gives rise to positive affect toward motives that (roughly speaking) aim at the good of others for its own sake, and thus leads us to *correctly* approve of such motives (Shaftesbury 1699–1714/2001; Hutcheson 1725/2004). Hume and Smith also used the expression 'moral sense', adding an account of its operation in terms of empathy (Slote 2010).

Contemporary sentimentalist epistemologists similarly emphasize parallels between emotion and perception. It has become common to think that emotions, like perceptions, have *representational content*: they are *about* something. The best indication of this is that it seems they can be appropriate or inappropriate depending on how things are in the world. For example, if you mistakenly believe that someone intentionally slighted you, it is in one sense inappropriate for you to be angry with the person, while it would be appropriate if they had done so. For this to be the case, anger must in some way represent the person as having slighted you (as Aristotle already noted). It does not seem that we must *believe* that someone slighted us to be angry with them – after all, we can be afraid of something we believe to be

safe (Greenspan 1988). In this, emotions resemble perceptual experiences, which are similarly somewhat independent of our beliefs: even if you believe that the two lines in a Müller-Lyer illusion are the same length, they still *look* different.

Given these analogies, many have argued that emotions are (or are like) fallible *perceptual experiences* of evaluative properties – ways of being open to the way things are, evaluatively speaking (Döring 2007, Kauppinen 2013, Tappolet 2016, Cowan 2018, Milona and Naar 2020). Of course, emotions might be misleading, when they are based on false beliefs about the world, or when they are influenced by inappropriate background desires like greed. But epistemic sentimentalists hold that at their best, emotions provide the most direct access we have to the value of something – you might *know* that Chaplin's boxing fight in *City Lights* is funny by way of reading about it in a book, even if you have no sense of humor whatsoever, but surely watching it and being amused is the most direct access you can have to its funniness! (Cf. Pelser 2014)

In spite of the plausibility of emotions as ways of accessing value, critics argue that emotions are importantly disanalogous to perception. One important challenge is that unlike perceptual experiences, emotions are states for which we can ask for and give reasons, and can be criticizable for having (Helm 2001). This suggests that they can't be unjustified justifiers, as sources of non-inferential knowledge are supposed to be. Michael Brady (2013) holds that they are instead "useful stand-ins" that lead us to attend to the features that actually justify attributing evaluative properties. Debate on these issues remains open (see Tappolet 2016 and Cowan 2018 for sentimentalist responses). Much work remains to be done in specifying the conditions in which emotions can be expected to be reliable or what kind of moral truths they might provide access for – perhaps emotions are better guides to goodness than justice, for example, since justice often requires principled balancing of competing claims.

5. Conclusion

This chapter has surveyed some (though certainly not all) of the roles that emotions play in our moral lives. While a minority of thinkers relegate them to the margins, empirical evidence from biology, psychology, and neuroscience all points to a significant impact on our moral behavior and thought. From a normative perspective, philosophers have identified many ways in which emotions might be constitutive of moral truths and may supplement or guide reasoned reflection. However, given that our innate emotional tendencies are in-group biased and potentially in tension with each other, it is plausible that for emotional guidance to be desirable, our feelings must be trained, regulated, or in some other way enlightened in order to result in morally justifiable behavior.

References

Aaltola, Elisa 2014. Affective empathy as core moral agency: psychopathy, autism and reason revisited. *Philosophical Explorations* 17 (1): 76–92.

Alexander, Richard 1987. *The Biology of Moral Systems*. New York: Aldine de Gruyter.

Anderson, Elizabeth 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Anderson SW, Bechara A, Damasio H, Tranel D, Damasio AR. 1999. Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience* 2 (11): 1032–7.

Aristotle 2000. *Nicomachean Ethics*. Edited and translated by Roger Crisp. Cambridge: Cambridge University Press.

Atran, Scott and Norenzayan, Ara 2004. Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioral and Brain Sciences* 27: 713–770.

Axelrod, Robert and Hamilton, William D. 1981. The evolution of cooperation. *Science* 211(4489), 1390–1396.

Ayer, A. J. 1936. *Language, Truth, and Logic*. London: Gollancz.

Batson, Daniel 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Lawrence Erlbaum.

Batson, C. Daniel 2010. Empathy-induced altruistic motivation. In M. Mikulincer and P. R. Shaver (eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature*. American Psychological Association, pp. 15–34.

Blackburn, Simon (1984). *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Clarendon Press.

Blackburn, Simon 1998. *Ruling Passions*. Oxford: Clarendon Press.

Bloom, Paul 2016. *Against Empathy*. New York: HarperCollins.

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

Boesch, Christophe 1994. Cooperative hunting in wild chimpanzees. *Animal Behaviour* 48(3): 653–667.

Boyd, Robert and Richerson, Peter 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology & Sociobiology* 13(3): 171-195.

Brady, Michael 2013. *Emotional Insight*. Oxford: Oxford University Press.

Brentano, Franz 1889/1969. *The Origin of Our Knowledge of Right and Wrong*. Translated by Roderick Chisholm and Elizabeth Schneewind. London: Routledge.

Broad, C. D. 1944. Some Reflections on Moral–Sense Theories in Ethics. *Proceedings of the Aristotelian Society* 45: 131–166.

Brosnan, Sarah and de Waal, Frans 2002. A proximate perspective on reciprocal altruism. *Human Nature* 13(1): 129–152.

Cohon, Rachel 2008. *Hume's Morality: Feeling and Fabrication*. Oxford: Oxford University Press.

Cowan, Robert 2018. Epistemic sentimentalism and epistemic reason-responsiveness. In Anna Bergqvist & Robert Cowan (eds.), *Evaluative Perception*. Oxford: Oxford University Press, pp. 219–236.

Cushman, Fiery, Young, Liane, and Greene, Joshua 2010. Our multi-system moral psychology: Towards a consensus view. In J. Doris et al. (eds.), *Oxford Handbook of Moral Psychology*. New York: Oxford University Press.

Cushman, Fiery, Young, Liane, and Hauser, Mark 2006. The Role of Conscious Reasoning and Intuition in Moral Judgment. *Psychological Science* 17 (12), 1082–1089.

D'Arms, Justin and Jacobson, Daniel 2000. Sentiment and Value. *Ethics* 110(4): 722–48.

Darwall, Stephen 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

Darwin, Charles. 1871. *The Descent of Man*. London: John Murray.

de Waal, Frans 1982. *Chimpanzee Politics*. New York: Harper and Row.

de Waal, Frans 2000. Attitudinal reciprocity in food sharing among brown capuchin monkeys. *Animal Behaviour* 60(2): 253–261.

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

Döring, Sabine 2007. Seeing what to do: affective perception and rational motivation. *dialectica* 61: 363–393.

Enoch, David 2010. The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it. *Philosophical Studies* 148: 413–38.

Enoch, David 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press

Evans, Jonathan and Stanovich, Keith 2013. Dual-Process Theories of Higher Cognition. *Perspectives on Psychological Science* 8 (3): 223-241.

Fehr, Ernst and Gächter, Simon 2002. Altruistic Punishment in Humans. *Nature* 415, 137–140.

Firth, Roderick 1952. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research* 12: 317–345.

Frank, Robert 1988. *Passions Within Reason*. New York: W. V. Norton.

Frijda, Nico H. 1986. *The Emotions*. Cambridge: Cambridge University Press.

Geach, Peter 1965. Assertion. *Philosophical Review* 74 (4):449–465.

Gibbard, Allan 1990, *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

Gilbert, Margaret 2006. *A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society*. Oxford: Oxford University Press

Gintis, Herbert, Smith, Eric A. and Bowles, Samuel 2001. Costly signaling and cooperation. *Journal of Theoretical Biology* 213(1): 103–19.

Goldman, Alvin 1979. What is Justified Belief? In George Pappas (ed.), *Justification and Knowledge*. Boston: D. Reidel, pp. 1–25.

Gomes, C. M., Mundry, R., and Boesch, C. (2009). Long-term reciprocation of grooming in wild West African chimpanzees. *Proceedings of the Royal Society B* 276, 699–706.

Graham Jesse, Haidt Jonathan, and Nosek, Brian. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96 (5), 1029–46.

Greene, Joshua 2013. *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. London: Penguin.

Greene, Joshua 2014. Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics* 124 (4):695-726.

Greene, Joshua, Morelli, Sylvia, Kelly Lowenberg, Leigh E. Nystrom & Jonathan D. Cohen 2008. Cognitive Load Selectively Interferes with Utilitarian Moral Judgment. *Cognition* 107(3): 1144-1154.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 2105–2108.

Greenspan, Patricia 1988. *Emotions and Reasons: An Enquiry into Emotional Justification*. London: Routledge.

Gleichgerrcht, Ezequiel and Young, Liane 2013. Low levels of empathic concern predict utilitarian moral judgment. *PloS One*, 8(4).

Hamann, K., Warneken, F., Greenberg, J. R., and Tomasello, M. 2011. Collaboration encourages equal sharing in children but not in chimpanzees. *Nature* 476(7360): 328–331.

Haidt, Jonathan 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108, 814–834.

Haidt, J., and Joseph, C. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, Fall, 55-66.

Haidt, Jonathan and Kesebir, Selin 2010. Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.) *Handbook of Social Psychology*, 5th Edition. Hoboken, NJ: Wiley, 797–832.

Haidt, Jonathan, Koller Silvia, and Dias, Maria 1993. Affect, Culture, and Morality, Or Is It Wrong to Eat Your Dog? *Journal of Personality and Social Psychology* 65 (4):613-28.

Hamilton W. D. 1964. The genetical evolution of social behaviour I. *Journal of Theoretical Biology* 7(1): 1–16.

Hare, R. M. 1952. *The Language of Morals*. Oxford: Clarendon Press.

Harman, Gilbert 1975. Moral relativism defended. *Philosophical Review* 84: 3–22.

Helion, Chelsea and Pizarro, David 2015. Beyond dual-processes: The interplay of reason and emotion in moral judgment. In N. Levy & Clausen, J. (Eds.) *Springer Handbook for Neuroethics*.

Helm, Bennett 2001. *Emotional Reason*. Cambridge: Cambridge University Press.

Henrich, Joseph, Boyd, Robert, Bowles, Samuel, Camerer, Colin, Fehr, Ernst and Gintis, Herbert (eds.) 2004. *Foundations of Human Sociality - Economic Experiments and Ethnographic: Evidence From Fifteen Small-Scale Societies*. Oxford: Oxford University Press.

Henrich, Natalie and Henrich, Joseph 2007. *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. New York: Oxford University Press.

Henrich, Joseph 2020. *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. New York: Farrar, Strauss, and Giroux.

Henrich, Joseph and Muthukrishna, Michael 2021. The origins and psychology of human cooperation. *Annual Review of Psychology*, 72, 204–240.

Huebner, Bryce, Dwyer, Sue, and Hauser, Mark 2009. The role of emotion in moral psychology. *Trends in Cognitive Science*, 13(1).

Hume, David 1739–40/1978. *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge, 2nd rev. edn., P. H. Nidditch. Oxford: Clarendon Press, 1978.

Hume, David 1751/1948. *An Enquiry Concerning the Principles of Morals*. In H.D. Aiken (ed.), *Hume's Moral and Political Philosophy*. New York: Hafner Press, 1948.

Hutcheson, Frances 1725/2004. *An Inquiry Into the Original of Our Ideas of Beauty and Virtue in Two Treatises*. Edited by W. Leidhold. Indianapolis, IN: Liberty Fund.

Jensen Keith 2010. Punishment and spite, the dark side of cooperation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365(1553): 2635–2650.

Joyce, Richard 2001. *The Myth of Morality*. Cambridge: Cambridge University Press.

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

Kahan, Dan M. 2013. Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making* 8(4): 407–424.

Kahane, Guy 2012. On the Wrong Track: Process and Content in Moral Psychology. *Mind and Language* 27 (5): 519–545.

Kahane, Guy, Katja Wiech, Nicholas Shackel, Miguel Farias, Julian Savulescu & Irene Tracey 2012. The Neural Basis of Intuitive and Counterintuitive Moral Judgement. *Social Cognitive and Affective Neuroscience* 7 (4):393–402.

Kahane, Guy 2014. Evolution and impartiality. *Ethics* 124 (2):327–341.

Kamm, Frances 2016. *The Trolley Problem Mysteries*. New York: Oxford University Press.

Kant, Immanuel 1785/1998. *Groundwork for the Metaphysics of Morals*. Tr. Mary Gregor. Cambridge: Cambridge University Press.

Kant, Immanuel 1797/1996. On a supposed right to lie from philanthropic motives. In Mary Gregor (tr. and ed.), *Practical Philosophy*. Cambridge: Cambridge University Press, pp. 611–615.

Kauppinen, Antti 2013. A Humean theory of moral intuition. *Canadian Journal of Philosophy* 43(3): 360–381.

Kauppinen, Antti 2014a. Fittingness and idealization. *Ethics* 124 (3): 572–588.

Kauppinen, Antti 2014b. Empathy and emotion regulation. In Heidi Maibom (ed.) *Empathy and Morality*. New York: Oxford University Press, pp. 97–121.

Kauppinen, Antti 2017. Sentimentalism, blameworthiness, and wrongdoing. In Karsten Stueber and Remy Debes (eds.), *Ethical Sentimentalism*. Cambridge: Cambridge University Press, pp. 133–152.

Kauppinen, Antti 2021. Relational imperativism about affective valence. *Oxford Studies in the Philosophy of Mind* 1, 341–371.

Kitcher, Philip 1998. Psychological altruism, evolutionary origins, and moral rules. *Philosophical Studies* 89 (2-3): 283–316.

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

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F. A., Hauser, M. D., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature* 446: 908–911.

Kohlberg, Lawrence 1976. Moral stages and moralization: The cognitive-developmental approach. In T. Lickona (ed.), *Moral development and behavior: Theory, research, and social issues*. New York: Holt, Rinehart & Winston, pp. 31–53.

Kokkonen, Tomi 2021. *Evolving in Groups : Individualism and Holism in Evolutionary Explanation of Human Social Behaviour*. Dissertation, University of Helsinki.

Kopajtic, Lauren 2020. Adam Smith's Sentimentalist Conception of Self-Control. *Adam Smith Review* 12: 7–27.

Korsgaard, Christine 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Korsgaard, Christine 2010. Reflections on the evolution of morality. *The Amherst Lecture in Philosophy* 5: 1–29.

Landy, Justin F. and Goodwin, Geoffrey P. 2015. Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science* 10(4), 518–536.

de Lazari-Radek, Katarzyna and Singer, Peter 2012. The objectivity of ethics and the unity of practical reason. *Ethics* 123 (1): 9–31.

Locke, John 1690/2008. *An Essay Concerning Human Understanding*. Ed. Pauline Phemister. Oxford: Oxford University Press.

Mackie, John 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.

Malle, Bertram F. 2021. Moral judgments. *Annual Review of Psychology* 72: 293–318.

May, Joshua 2018. *Regard for Reason in the Moral Mind*. New York: Oxford University Press.

McDowell, John 1998. *Mind, Value, and Reality*. Cambridge, MA: Harvard University Press.

Mencius 1970. *Mencius*. Tr. D. C. Lau. Harmondsworth: Penguin.

Mikhail, John 2011. *Elements of Moral Cognition*. Cambridge: Cambridge University Press.

Milona, Michael and Naar, Hichem 2020. Sentimental Perceptualism and the Challenge from Cognitive Bases. *Philosophical Studies* 177(10): 3071–3096.

Moore, George Edward 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Moore, George Edward 1912. *Ethics*. Oxford: Oxford University Press.

Nichols, Shaun 2004. *Sentimental Rules*. New York: Oxford University Press.

Nichols, Shaun 2005. Innateness and Moral Psychology. In *The Innate Mind*, eds. P. Carruthers, S. Laurence, and S. Stich. New York: Oxford University Press.

Parfit, Derek 2011. *On What Matters* Vols. 1-2. Oxford: Oxford University Press.

Pascual, Leo, Rodrigues, Paulo, and Gallardo-Pujol, David 2013. How does morality work in the brain? A functional and structural perspective of moral behavior. *Frontiers in Integrative Neuroscience* 7, 65.

Paxton, J.M., Ungar, L., Greene, J.D. 2011. Reflection and reasoning in moral judgment. *Cognitive Science* 36, 163–77.

Pelser, Adam 2014. Emotion, evaluative perception, and epistemic justification. In Sabine Roeser and Cain Todd (eds.), *Emotion and Value*. Oxford: Oxford University Press, pp. 107–123.

Pizarro, David and Bloom, Paul 2003. The Intelligence of the Moral Intuitions: A Comment on Haidt. *Psychological Review* 110 (1):193-196.

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

Prinz, Jesse 2011. Against Empathy. *Southern Journal of Philosophy* 49, Supplementary Volume, 214–233.

Rabinowicz, Wlodek and Rønnow-Rasmussen, Toni 2004. The strike of the demon: on fitting pro-attitudes and value. *Ethics* 114(3): 391–423.

Railton, Peter 2014. The Affective Dog and Its Rational Tale. *Ethics* 124 (4): 813-859.

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

Ridge, Michael 2014. *Impassioned Belief*. Oxford: Oxford University Press.

Rosas, Alejandro 2007. Beyond the sociobiological dilemma: social emotions and the evolution of morality. *Zygon* 42 (3): 685–700.

Rosas, Alejandro and Aguilar-Pardo, David 2020. Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking & Reasoning* 26 (4): 534–551.

Roughley, Neil (2018). From shared intentionality to moral obligation? Some worries. *Philosophical Psychology* 31 (5): 736–754.

Scanlon, Thomas 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scarantino, Andrea 2014. The motivational theory of emotions. In Justin D'Arms & Daniel Jacobson (eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*. New York, Oxford University Press, pp. 156–185.

Scheffler, Samuel 2010. *Equality and Tradition*. New York: Oxford University Press.

Schnall S, Benton J, Harvey S. With a clean conscience: cleanliness reduces the severity of moral judgments. *Psychological Science* 19 (12): 1219–22.

Schroeder, Mark 2010. *Non-Cognitivism in Ethics*. London: Routledge.

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

Shaftesbury, Third Earl of (Anthony Ashley Cooper), 1699–1714. *An Inquiry Into Virtue and Merit*. In D. Den Uyl (ed.), *Characteristicks of Men, Manners, Opinions, Times*, Vol.2. Indianapolis, IN: Liberty Fund, pp. 1–100.

Shweder, R. A., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children*, Chicago: University of Chicago Press, 1–83.

Singer, Peter 2005. Ethics and Intuitions. *Journal of Ethics* 9 (3-4): 331-352.

Silk, Joan B. 2002. Kin selection in primate groups. *International Journal of Primatology* 23(4): 849–875.

Smith, Adam 1758/2002. *The Theory of Moral Sentiments*. Ed. Knud Haakonssen. Cambridge: Cambridge University Press.

Slote, Michael 2010. *Moral Sentimentalism*. New York: Oxford University Press

Sober, Elliot and Wilson, David Sloan 1998. *Unto Others*. Cambridge, MA: Harvard University Press.

Stevenson, Charles 1944. *Ethics and Language*. New Haven: Yale University Press.

Strawson, Peter 1962. Freedom and Resentment. *Proceedings of the British Academy* 48, 1-25.

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

Street, Sharon 2010. What is constructivism in ethics and metaethics? *Philosophy Compass* 5 (5): 363–384.

Sunstein, Cass 2005. Moral Heuristics. *Behavioral and Brain Sciences* 28(4), 531-573.

Tangney, June Price and Dearing, Ronda L. 2003. *Shame and Guilt*. New York: Guilford Press.

Tappolet, Christine 2016. *Emotions, Value, and Agency*. New York: Oxford University Press.

Thomson, Judith Jarvis (1976). Killing, letting die, and the trolley problem. *Monist* 59 (2): 204–217.

Tofilski, Adam, Couvillon, Margaret J, Evison, Sophie E. F., Helanterä, Heikki, Robinson, Elva J. H, Ratnieks, Francis L. W. 2009. Preemptive defensive self-sacrifice by ant workers. *American Naturalist* 172 (5): E239-E243.

Tomasello, Michael and Vaish, Amrisha 2013. Origins of human cooperation and morality. *Annual Reviews of Psychology* 64: 231–255.

Tomasello, Michael 2016. *A Natural History of Human Morality*. Cambridge, MA: Harvard University Press.

Tomasello, Michael 2020. The moral psychology of obligation. *Behavioral and Brain Sciences* 43: 1-33.

Trivers, Robert 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46 (1): 35–57.

Tuomela, Raimo 2007. *The Philosophy of Sociality: The Shared Point of View*. New York: Oxford University Press.

Turiel, Elliot 2006. The development of morality. In N. Eisenberg, W. Damon, & R. M. Lerner (eds.) *Handbook of child psychology: Social, emotional, and personality development*. Hoboken, NJ: Wiley, pp. 789–857.

Valdesolo, Piercarlo and DeSteno, David 2006. Manipulations of emotional context shape moral judgment. *Psychological Science*, 17(6), 476-477.

Vavova, Katia 2021. The limits of rational belief revision: A dilemma for the Darwinian debunker. *Nôus* 55 (3): 717–734.

Way, Jonathan 2012. Transmission and the Wrong Kind of Reason. *Ethics* 122 (3): 489–515.

Westermarck, Edward 1906. *The Origin and Development of Moral Ideas* Vol. 1. London: Macmillan.

Wiggins, David 1998. *Needs, Values, and Truth*. Oxford: Blackwell.

Zajonc, Robert 1980. Feeling and thinking: Preferences need no inferences. *American Psychologist* 35(2): 151–175.