

A Sentimentalist Solution to the Moral Attitude Problem

Antti Kauppinen

Final draft, September 15, 2009

Abstract

Expressivists have long struggled with a non-circular individuation of the attitudes of approval and disapproval they argue constitute moral judgment. Since there seems to be no special moral feeling, they have in general tried to define moral attitudes in terms of their special functional role. Recently, Alexander Miller has presented strong arguments against Simon Blackburn's various attempts, and Justin D'Arms and Daniel Jacobson have formulated a persuasive challenge to Allan Gibbard's view. I suggest that expressivists are better off abandoning the focus on functional role and instead following the lead of classical sentimentalists like Hume and Smith. Their views can be construed as distinguishing moral attitudes in terms of the special process from which they characteristically result, namely impartial sympathizing with the emotional reactions of those affected by the action or character traits. I call this type of view *historical sentimentalism*, and argue it not only solves the moral attitude problem but also explains why moral attitudes typically have the functional features expressivists have picked up on. In the final section, I defend historical sentimentalism against the challenge raised by the plain fact that we do not always arrive at our moral judgments by way of impartial sympathy.

A Sentimentalist Solution to the Moral Attitude Problem

Antti Kauppinen

‘Tis only when a character is considered in general, without reference to our particular interest, that causes such a feeling or sentiment, as denominates it morally good or evil.

Hume, *Treatise* III.1.2, 472

Before you accuse, criticize and abuse, walk a mile in my shoes.

Elvis Presley¹

Metaethical expressivists believe that thinking that something is morally right or wrong consists essentially in having an attitude of approval or disapproval toward it.² But not just any approval or disapproval will do – we can, of course, approve of something without thinking that it is morally good, for example if it is to our own advantage or simply feels good. The attitude involved in moral judgment must be specifically *moral* approval or disapproval. An obvious way to distinguish specifically moral attitudes would be to do so in terms of moral content – to disapprove of something morally is to disapprove of it *because* it is morally wrong, as one sees it. But this avenue is closed for the expressivist, since it presupposes an independent grasp of thinking that something is morally wrong, which is precisely what the expressivist is trying to explain in terms of moral disapproval. As John McDowell put it, “Surely it undermines a projective account of a concept if we cannot home in on the subjective state whose projection is supposed to result in the seeming feature of

¹ ‘Walk a Mile in My Shoes’, music and lyrics by Joe South.

² Some expressivists believe that non-moral beliefs of one kind or another are also involved. This is an irrelevant complication for my purposes.

reality in question without the aid of the concept of that feature” (McDowell 1981/1998: 158).³

This gives rise to what Alexander Miller (2003) has usefully termed the ‘moral attitude problem’ for expressivists. To have a genuinely explanatory, naturalistic account of moral judgment, expressivists must be able to isolate specifically moral attitudes in non-moral terms. Miller believes that such attempts fall victim to an analogue of the Open Question Argument. That is, it is possible for someone to hold the sort of complex attitude that the expressivist proposes toward something without necessarily judging that it is morally right or wrong, in which case the expressivist account of moral thinking fails.⁴ My plan here is to first sharpen the problem by explaining the open feel in terms of the distinctive authority of moral judgments, and then propose a bold solution drawing on the classics of sentimentalism, David Hume and Adam Smith. Very briefly, *moral* disapproval is an attitude of disapproval that characteristically *results from* impartially placing oneself in the shoes of those primarily affected by the action and sharing their negative reactive attitudes – a process that involves information-seeking, sympathy, and reflective correction for typical perspectival distortions. This kind of *historical*, rather than phenomenological or functional, account of moral attitudes solves the moral attitude problem and as a bonus helps explain why moral attitudes typically have the sort of functional role expressivists identify. However, it faces a serious challenge of its own. Plainly, as common experience and psychological studies both show, we do not always arrive at moral judgments in this manner. Sometimes our moral judgments are *hot*, resulting from immediate affective reactions. At the other extreme, some people seem to make exclusively *cold* moral judgments without any kind of emotional process. But as I will argue, this problem can be avoided as long as we have grounds to think of moral attitudes

³ See also Wiggins (1987), D’Arms and Jacobson (1994), Merli (2008).

⁴ Miller is more sanguine towards Gibbard’s 1990 view, but as we will see, it is not exempt from the problem.

resulting from hot and cold processes as *parasitic* on those that meet the sentimentalist conditions.

1. The moral attitude problem

Early emotivists like Ayer were content to describe the attitude expressed by moral judgments as a special “ethical feeling” (Ayer 1946: 108). But as many have pointed out, it is simply false that there is some phenomenologically distinct feeling that must be present whenever we sincerely judge that something is bad or depraved or wrong. (From now on, I will focus for simplicity’s sake on the sort of disapproval associated with thinking that something is wrong.) Contemporary expressivists have consequently characterized moral attitudes in terms of their special *functional role* instead, given that both special phenomenal quality and moral content are ruled out. Simon Blackburn offers several related suggestions, all somewhat tentative. The simplest appeals to stability as the mark of moral attitudes: “[I]f we imagine the general field of an agent’s concerns, his or her values might be regarded as those concerns that he or she is also concerned to preserve.” (Blackburn 1998: 67) Moral disapproval, then, would seem to be disapproval that is sustained by a higher-order attitude of approval – we resist change to an attitude of ours that we approve of, and whose absence we might even disapprove of.⁵

This meshes well with what Blackburn says elsewhere of ‘emotional ascent’:

[A]t the bottom we start with pure preferences. Rising up we come to preferences that we prefer others to share. Rising further we come to preferences that we ‘demand’ of others; that is, if they do not share them we find ourselves averse or in opposition to them. Here, according to me, we begin to enter the territory of ethics. (Blackburn 2002: 125)

⁵ A terminological note: when I use the term ‘disapproval’ in this paper without qualification, it is to be taken as a flavourless booing that might be moral, aesthetic, gustatory, or whatever. This contrasts with a different use picked up by D’Arms and Jacobson who say that “disapproval is already a moral or quasi-moral notion” (1994, 760).

Moral disapproval, then, is disapproval the absence of which we disapprove in others, and presumably ourselves (which gives us the beginnings of understanding guilt). Yet further ascent is possible: we may disapprove of those who fail to disapprove of those who lack the first-order disapproval. That is, we not only disapprove of paedophiles, but also people who lack this disapproval, and even people who tolerate the paedophile-tolerant. This, on Blackburn's picture, amounts to *morally* disapproving of paedophiles, as opposed to holding some other kind of negative attitude toward them.

Alexander Miller rejects these suggestions on the basis of a version of the Open Question Argument. The claim is simple: for any proposed definition of a moral attitude in terms of non-moral attitudes, we can grant that someone holds the non-moral attitudes in question but yet intelligibly ask "But is she moralizing?" (Miller 2003: 48) Miller implicitly assumes that this open feel is not explained away by the unobviousness of the conceptual truth in question, lack of competence, performance error, or any other sort of condition that would in general endanger the inference from the existence of an open question to lack of identity. His argument comes in two varieties, which I will label direct and indirect. On the direct version, an expressivist account is rejected because it is intuitively possible to hold the attitudes and yet question whether the disapproval is moral rather than aesthetic, gustatory, or some other kind. The indirect strategy relies on a putative conceptual truth about moral judgment: an agent who genuinely judges that x is wrong is, *ceteris paribus*, disposed to demand that others share that very attitude (Miller 2003: 49). Miller then argues that it is possible to grant that an agent holds the attitudes the expressivist regards as sufficient and yet question whether the agent is disposed to demand that others share those very attitudes. The basic problem in each case is that the expressivist accounts are *too permissive*: they allow attitudes to count as moral even when it is intuitively implausible. The functional role they specify may be necessary, but it is not sufficient for moral disapproval.

Miller deploys the direct strategy against stable attitude and higher-order approval accounts of moral disapproval. The argument is straightforward: we can imagine someone having a higher-order approval toward (the maintenance of) her disapproval of *x* without *morally* disapproving of *x*. For example, someone who aesthetically disapproves of Wagner's music may approve of this disapproval and resist pressures to give it up. So, stability and higher-order approval do not suffice to distinguish moral from other attitudes (Miller 2003: 90–93). The indirect strategy, in turn, appears effective against Blackburn's emotional ascent story. That story does itself incorporate the idea that moral disapproval involves demanding others to disapprove as well. But, Miller argues, merely demanding that others share *some* disapproving attitude is not sufficient to distinguish a moral sentiment: “[W]hen I judge that murder is morally wrong, I express a non-cognitive sentiment towards murder, and I approve of everyone sharing that *same type* of non-cognitive sentiment: it wouldn't be enough, for example, for others to find murder merely aesthetically displeasing” (Miller 2003: 89). So Blackburn should say that A morally disapproves of *x* if and only if A disapproves of *x* and approves of others *morally* disapproving of *x* – but this is objectionably circular, since the right-hand side makes use of the very concept we are trying to capture. Nor does there seem to be any obvious fix. So, Miller concludes that Blackburn's attempts to account for moral attitudes in terms of non-moral attitudes fails.

This is where Miller stops, but I think we can strengthen the argument. Those who accept the classical Open Question Argument against naturalistic definitions of goodness are not content with the mere appearance of openness, but also try to explain *why* the question remains open. So, for example, non-cognitivists may argue that judgments about goodness are inherently motivating, while no judgments made in naturalistic terms are such, which is why we should expect a gap between the two (Darwall, Gibbard, and Railton 1992). Some further feature of the concept of goodness explains why it cannot be identical with any naturalistic

concept. This is the strategy I will adopt. But what could be the further feature in the case of moral attitudes? My suggestion is that what at bottom explains the open feel Miller picks up on is the distinctive *authority* associated with moral judgments.⁶ When we think that something is morally demanded of us, it is not just another concern. It not only potentially outweighs other considerations, but may *silence* them, rule them out of further deliberation. As Richard Joyce puts it, they have a special ‘practical clout’ – they cannot be legitimately ignored or evaded in deciding what to do (Joyce 2006: 57–64). We think that we, and others, are seriously at fault if we do not take these demands into account. We also take this to be the case regardless of our interests and desires, as well as (at least fundamentally) regardless of what external authorities think about the matter. Many moral judgments thus either are or entail what we might call self-directed categorical imperatives. The challenge for the expressivist, then, is to account for this practical clout in terms of attitudes.

With this explanation of openness at hand, we can see that Gibbard’s account of moral attitudes, which Miller believes to survive the challenge he raises to Blackburn, is also vulnerable to the objection. On Gibbard’s 1990 version, thinking that something is wrong is taking it to be rational to feel *guilt* or anger for doing it, where taking something to be rational is a non-cognitive attitude of accepting norms that permit or require it.⁷ (In his later (2003) version, it is a matter of *planning* to feel guilt or anger for the action, but the difference is not relevant to my point.) The question then becomes: is it possible for someone to accept norms that require guilt (and anger) for ϕ -ing without thinking that ϕ -ing is morally wrong? Does accepting a norm for guilt have the right kind of subjective normative authority? To answer this, it is essential to recall that guilt must itself be understood in non-moral terms here. As Gibbard acknowledges, his account would be circular if he accepted a judgmental theory of

⁶ Miller’s appeal to demanding that others share one’s attitude is a move in this direction.

⁷ More precisely, Gibbard correctly notes that we might think something is wrong without thinking it rational to be angry with the agent, if the agent does not meet the conditions of blameworthiness. He regards fixing this as straightforward, and I will ignore the complication in the following.

guilt, according to which it involves thinking that one has done something wrong, since, again, he is trying to understand what it is to think something is wrong in terms of guilt (Gibbard 1990: 148). So the norms one accepts must be taken to require a *non-moral* feeling or ‘syndrome’ (tendency to respond to situations with reactions).

And this, I believe, opens up the question again. It does seem possible that someone could accept norms that prescribe this sort of non-moralized feeling for certain kinds of actions – perhaps because she regarded this as advantageous to herself – without thereby thinking that such actions were morally wrong (or blameworthy), and without thinking that they and others would be seriously at fault if they did not. Consider Bill Clinton in the aftermath of the Monica Lewinsky scandal. Could he not have considered it appropriate, warranted perhaps by concern for his family’s feelings, to feel the guilt-feeling, perhaps in order to be able to express sincere contrition and mend things, while in his heart of hearts remaining convinced that what he did was just a morally harmless fling that was blown way out of proportion by the Independent Counsel investigation and the media? Or take one of the cases discussed by Justin D’Arms and Daniel Jacobson: “If you fire your foreign housekeeper because his presence makes the kids unhappy, it makes sense to feel guilty for what you have done, and doubly so for having pandered to the children’s naïve bigotry” (D’Arms and Jacobson 1994: 751). You might nevertheless think firing the housekeeper was the right thing to do, all things considered.⁸ If so, thinking that something is guilt-apt is not the same as thinking that it is wrong. To adapt a phrase familiar from criticisms of the buck-passing account of value, you can have the *wrong kind of reasons* for planning to feel guilt, in which case the connection to moral disapproval is broken.⁹

⁸ As this case suggests, the issue is related to the problem that expressivists have in explaining *reason* judgments in contrast to overall verdicts (see Dancy 2004). Sometimes guilt may be appropriate when there is strong reason against doing something that is nevertheless overall best. If, as Dancy argues, expressivists have a general problem in accounting for judgments about contributory reasons, it is no surprise that these cases cause problems.

⁹ Similar issues arise for Mark Schroeder’s (2008) view, according to which to think that something is morally wrong is to be for blaming for it. We can be for blaming for something without thinking that it is morally wrong.

Even if these cases are not convincing, I will argue in section III that there is a further problem for any view that focuses merely on functional features of moral judgment. Such views allow one to count as a moralizer even if one always arrives at a judgment – in Gibbard’s case, accepts a norm for non-moral guilt – on the basis of uncorrected gut reactions. This, I will argue, is too permissive as well. Not just everybody who has strong convictions counts as a participant in moral discourse and practice. In other words, even if Gibbard’s account were to solve the *authority* problem, it would still be faced with the *genealogical* problem. So Gibbard’s view is still too permissive, and the moral attitude problem remains open.

2. The sentimentalist solution

The first step toward solving the moral attitude problem is looking back. Moral attitudes, I want to suggest, can be distinguished by their characteristic history, rather than phenomenal feel or functional role. The basic thesis is thus simple:

The Historicist Thesis

An attitude is moral only if it characteristically results from a process of moral judging.

An account of this type will obviously be circular if the process of moral judging involves adopting moral attitudes, and incompatible with expressivism if it involves forming moral beliefs.¹⁰ This need not be the case, however, as the classic works of the sentimentalist tradition show. If we are able to delineate moral judgment in its process sense (the activity of moral judging) without reference to moral judgment in the product sense (the output of such an activity), we can use the former to individuate the latter. The kind of view that Hume and

¹⁰ Miller briefly raises the possibility of a historicist account, but dismisses it immediately on the assumption that it must inevitably be circular (Miller 2003: 46).

Adam Smith defended, I will argue, offers a promising way of cashing out moral judging in non-circular terms. Roughly, the sentimentalist account is something like the following:

Sentimentalism about Moral Judging

A distinctively moral process of forming an attitude transcends one's egocentric perspective by way of simulating the non-moral reactive attitudes that any informed and unbiased participant would have in the circumstances of those affected by the action.

Non-moral reactive attitudes are attitudes toward the manifest attitudes of others that do not as such involve moral evaluation. They include such emotions as anger and gratitude. Putting historicism and sentimentalism together, we get an initial statement of what I will call historical sentimentalism:

Historical Sentimentalism (HS)

An attitude is moral only if it characteristically results from a process of simulating the non-moral reactive attitudes that any informed and unbiased participant would have in the circumstances of those affected by the action.

The qualifier "characteristically" is essential in order to capture the fact that we can arrive at our moral judgments rashly or partially and still count as moralizing. I will discuss this in detail in the next section.

These are the bare bones of the sentimentalist solution. Superficially, it resembles ideal observer and other response-dependence accounts of moral rightness. It is, however, essential to bear in mind that here we are constructing a theory of what it is to *think* something is morally wrong, not a theory of what *is* morally wrong. Expressivist sentimentalists are not committed to thinking that whatever any impartial but otherwise normal person would disapprove merits disapproval or is morally bad. Nor is it part of the *content* of the attitude that it would be had by an impartial spectator. Rather, the view is that a distinctively moral attitude is one that characteristically results from a particular kind of imaginative exercise. First-order questions about the fittingness of such attitudes or truth of moral thoughts are downstream from this thesis, and logically independent of it.

Sympathy and Anger

How do we characteristically arrive at moral judgments, according to sentimentalists? To begin with, we plausibly approve of some things, actions, and characters because they give us pleasure or joy, and disapprove of others because they cause us pain or discomfort.

Consequently, we feel grateful or angry about them. Moralizing, whether good or bad, involves transcending this purely egocentric perspective somehow; that much is agreed on all sides. Let us take a concrete case to examine what sentimentalists have to say about it.

Moving House

John has recently moved to a new apartment. He enlisted the help of his friend Paul to carry his things, and Paul did as asked. Now Paul is moving, too, and asks John to return the favour. But John declines, preferring to stay at home and watch his new TV, which Paul dragged up the stairs. He has nothing more important to do.

I take it we disapprove of John in Moving House. But why, if his behaviour is nothing to us personally? For the sentimentalist, the answer involves *sympathy*, sharing the feelings and reactions of others. For Hume, this is just a matter of our natural capacity to share in the pleasures and pains of others – to feel pleasure at the signs of pleasure in others, to feel pain when it appears that others are in pain (*Treatise* III.3.1: 575–576). Sympathy extends our sentiment of disapproval to those actions and traits that give displeasure to others, in this case Paul. But this can't be enough. If circumstances were different – say if Paul hadn't helped John and wasn't a particularly close friend with him – we might still be displeased by his displeasure and realize it was caused by John's declining, but not disapprove of John. We might think he should not complain, even if we felt sorry for him. On the other hand, even if Paul was particularly good-natured and just happily shrugged off John's declining, we might still disapprove of John ourselves and think it would have been *appropriate* for Paul to resent him in the situation.

Adam Smith's more sophisticated notion of sympathy is designed to deal with such cases. It is not limited to sharing pleasure or pain. For him, we approve of a motive or reaction if we take it that we would have the same in the agent's situation. As a first pass, I disapprove of John's desire to stay at home, and derivatively his corresponding action, because I believe I wouldn't have the same desire were I in his circumstances. I compare John's actual reaction with my own hypothetical reaction in his situation, and find that the two do not match. If they were to match, I would, in Smith's sense, sympathize with him, and consequently approve of him.¹¹ As Smith puts it,

When the original passions of the person principally concerned are in perfect concord with the sympathetic emotions of the spectator, they necessarily appear to this last just and proper, and suitable to their objects; and, on the contrary, when, upon bringing the case home to himself, he finds that they do not coincide with what he feels, they necessarily appear to him unjust and improper, and unsuitable to the causes which excite them. (*TMS* I.i.3.1: 16)

Sympathy can lead us to approve or disapprove of actions that are not in our own interest, but it is not yet sufficient to distinguish specifically moral approval, nor explain its authority over us. After all, I can also sympathize in Smith's sense with John's shrewd business decisions or fail to share his taste in wines. That is, I can find that were I in John's shoes, I would invest in the same equipment, or wouldn't choose the same bottle to go with fish. For Smith, this either is, or causes me, to approve (or disapprove) of John, but surely this is not yet moralizing.

The next step to moral attitudes, then, is to focus not on whether we sympathize with the agent but with the reactions of those who are 'principally affected' by the action – would we, in their place, feel *reactive attitudes* like gratitude or anger toward the agent? These reactive attitudes should be themselves non- or pre-moral to avoid introducing circularity. Strawson, who introduces the term, lists resentment, gratitude, forgiveness, anger, and love as 'personal' (as opposed to moral or 'vicarious') reactive attitudes. They are all reactive in that

¹¹ I will make nothing of the distinction between empathy and sympathy here. 'Empathy' is often used to mean taking on the actual emotion of the other person (for example, Prinz forthcoming). As I see it, empathy in this sense plays only a minor role in the sentimentalist argument, *pace* Prinz's criticism, which thus misses its target.

they only make sense when adopted toward someone who is also a subject of attitudes. They are reactions to the “quality of others’ wills towards us, as manifested in their behaviour” (Strawson 1963/2003: 83). I will focus here on anger, since it can uncontroversially be felt toward someone without already making a moral judgment. (Both Smith and Strawson make use of resentment, but resenting someone may already involve a thinking they are morally blameworthy.¹²) For sentimentalists, it is *sympathy with anger of those affected* that gives rise to disapproval that is a step closer to a moral one, in the narrow sense of ‘what we owe each other’. Were I in Paul’s shoes, I would be angry with John’s failure to return the favour, his failure to reciprocate when it is expected of him. Since I sympathize with Paul’s anger, I disapprove of John – the more so, since I would not share John’s reaction to the request.

What if I sympathized with both – that is, what if I saw myself doing as John did in his situation and feeling as Paul does in his situation? This could, first, lead to an attempt to transcend my potentially idiosyncratic perspective – would *anyone* in John’s situation share his preferences, or in Paul’s share his anger, or is either just a reflection of my particular emotional constitution? (More on this move very soon.) Second, if this move leaves the issue unresolved, we might end up distinguishing between attitudes toward the agent and attitudes toward the action – perhaps John is blameless, though his action is wrong.¹³

Internalization and Subjective Authority

Disapproval based on sympathizing with the anger of those affected still lacks the distinctive subjective authority of moral judgments, however. We are aware of the fallibility of our

¹² Can you resent someone without already thinking that the person has done something wrong? Perhaps it is possible, but I will try to make the historical sentimentalist case without assuming it. I can, surely, resent the *fact* that someone else got the job I wanted without thinking that my competitor did anything wrong. But then I am not resenting the *person* who got the job. Maybe I think that the search committee did something wrong, however. This ties up with the sense in which resentment is a form of powerless anger toward superiors (think of Nietzsche’s *ressentiment*).

¹³ Smith refuses to make this distinction, claiming that we cannot sympathize with gratitude, for example, unless we sympathize with the motive of the agent. For him, consequently, “the sense of merit seems to be a compounded sentiment, and to be made up of two distinct emotions; a direct sympathy with the sentiments of the agent, and an indirect sympathy with the gratitude of those who receive the benefit of his actions.” (*TMS* II.ii.1.23)

reactions – our frequent ignorance of relevant facts, our partiality to ourselves and our friends, the influence of mood and hurry, the difficulty of putting ourselves in the shoes of others. We become aware of this, typically, when others fail to share our sympathy for Paul’s anger, that is, fail to sympathize with *us*. Since, on this account, to fail to share attitudes is to disapprove, we find others looking down on us on such occasions. Perhaps George and Ringo hear the same story about John and Paul, but have a laugh instead of a frown. *They* would not resent John in Paul’s situation, and so they find my attitude incorrect. Instead of facilitating cooperation, sympathetic attitudes sow discord. This forces us to examine both our own reactions, and those of others, as well, since we know they are not immune of the same sources of fallibility. To assess their merit we need a yardstick that is independent of the actual reactions of either.

Such a measure is provided by the reactions of an informed and impartial agent who is placed in the relevant situation:

[Gratitude and resentment], as well as all the other passions of human nature, seem proper and are approved of, when the heart of every impartial spectator entirely sympathizes with them, when every indifferent by-stander entirely enters into, and goes along with them. [...] To us, surely, that action must appear to deserve reward, which every body who knows of it would wish to reward, and therefore delights to see rewarded: and that action must as surely appear to deserve punishment, which every body who hears of it is angry with, and upon that account rejoices to see punished. (TMS I.ii.1.11–12)

As Stephen Darwall (1999) emphasizes, here we must be careful and not confuse Smithian sentimentalism with an ideal observer theory. We are not to imagine how someone viewing the situation from the outside – like Hare’s archangel taking in everyone’s preferences and then maximizing satisfaction¹⁴ – would react. Rather, in moral judging, we imagine how someone (indeed, anyone) free of the sources of distraction – informed, focused, unbiased, and so on, but otherwise like us in terms of emotional reactions – would feel were she in the position, in turn, of the agent and the patients of the action. In Darwall’s terms, we

¹⁴ See Hare 1981, esp. chapters 5 and 6. This sort of theory is decisively rejected by Rawls (1971: 184–190).

imaginatively project into the participants' situation "not as ourselves, but impartially, as any one of us" (Darwall 1999: 142). It would thus be less misleading to speak of an *impartial participant* theory (however oxymoronic that may sound) than of an impartial spectator.

Roughly, if we take that any impartial participant in Paul's shoes who knew what we take to be the relevant facts would share Paul's anger of John's behaviour, we not only disapprove of John's behaviour, but take this disapproval to be merited. This is not a trivial imaginative feat. It requires abstracting away our own peculiar tastes and preferences, and discerning what our reactions – the reactions of any normal human being, or perhaps of anyone with whom we would care to associate with, or maybe of anyone whose advice we would take in practical choices – would then be.¹⁵

We can make this proposal more concrete in terms of simulation theory in cognitive science. I will take as my point of departure Alvin Goldman's (2006) version of the theory, since it is the most fully developed and empirically informed. The core idea of simulation theory, as an empirical hypothesis, is that when we attribute mental states to others, we run our own psychological mechanisms off-line, using features of the target person's situation rather than our own as inputs. For example, if I know what someone wants and believes, I can predict that they will decide in one way rather than another by imagining that I want and believe that way (in Goldman's terms, *enacting* the other's desires and beliefs) and feeding these pretend desires and beliefs into my own decision-making system. Normally (on this very simple picture), the inputs to my decision making-system are my own beliefs and desires, and the output an intention to act in a certain way. When the system is run off-line, the output is not an intention, but a pretend-intention. To reach this result, the subject needs not only to use the target's beliefs and desires as input, but also needs to quarantine her own

¹⁵ Jonathan Schaffer, who emphasized the need to qualify the 'anyone' in the original formulation, suggested focusing on the reactions anyone I admire (personal communication). The trouble with this suggestion is that although admiration as such is a non-moral attitude (and thus avoids introducing circularity), the relevant sort of admiration here is inevitably the moral kind. This can easily be seen if we considering people I admire, for example, for their achievements in sport – surely I need not expect them to concur with my resentment to consider it warranted!

potentially different beliefs and desires from interfering with the process.¹⁶ This pretend-intention is then used to form a genuine belief about what the target's actual decision will be. Though simulation theorists differ in details, they are united in rejecting that (all or most) attribution processes involve inference on the basis of folk psychological laws. This includes generalizations of the type 'I would decide to ϕ in X's position, X is relevantly like me, so X will decide to ϕ '.¹⁷ The inferential step is not needed, because in the context of simulation, the pretend-state will itself cause the desired prediction to arise.¹⁸

Whether or not some version of simulation theory is true of everyday mindreading, it can be used to make the idea of sympathy more specific. I will sketch how such moral simulation might go. What is simulated here is not an actual decision, but a hypothetical emotional reaction to a situation. Since my purpose is not to construct a theory of emotion, I will simply assume a black box I will call the emotional response system (ERS) that takes as (one) input a factual representation. This is not to rule out other inputs such as bodily changes and desires, but I take it to be relatively uncontroversial that emotions that have intentional objects are aroused at least in part by situation-representations. For example, the belief that there is a wolf at the door will give rise fear, given a particular configuration of ERS. The input to ERS can also be make-believe: imagining a wolf at the door can also activate the system.¹⁹ It is a matter of much dispute in aesthetic psychology what the output is in such a case – is it (perhaps irrational) real fear or some sort of make-believe emotion?²⁰ I will here assume without further argument along the latter lines that the output is a *quasi-emotion*, a state that phenomenologically and physiologically resembles the real thing without having the

¹⁶ The fact that our attributions are often biased in the direction of our own actual states is nicely explained by simulation theory as a result of flawed quarantine, as Goldman (2006, section 7.7) notes.

¹⁷ See Goldman 2006: 30–1 on the 'resemblance-to-self premise'.

¹⁸ This is particularly emphasized by Robert Gordon (1995: 734–735).

¹⁹ Of course, imagination *need* not activate any kind of emotional response. Film-makers and theatre designers expend a lot of effort to create a suitable environment for fictional representations to engage the system.

²⁰ See, for example, Walton 1997.

same motivational consequences.²¹ Whether it will have any further consequence depends on the purpose we have for running the system off-line.

Assuming ERS can take other things than our own beliefs as input, it can be run to simulate someone else's reaction. Here it is not enough to change the input, however. As a rule, we need to *recalibrate* ERS to match that of the target. For example, if I am considering whether a child will find *The Fellowship of the Ring* scary, it is not enough for me to imagine what it will be like to see it, but what it will be like for *him* to see orcs and nazguls. I need to, as it were, import the child's settings and only then run the system off-line, possibly yielding quasi-fear as a result. Quasi-fear will not motivate me to flee but, in this context, perhaps helps decide against showing the film to the child.

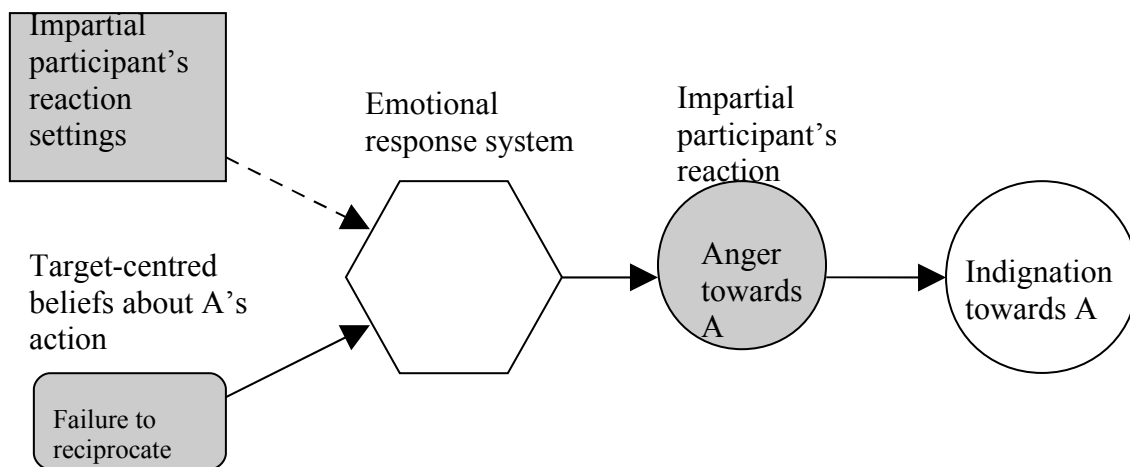
The simulation involved in the characteristic generation of moral sentiments, I venture, has notable similarities to the aesthetic simulation case. The input representation is the situation of those principally affected, as we see it. It is not relativized to the target's beliefs (which could be false), but consists rather in our view of the facts as they would look like from an informed target's standpoint.²² If Paul has false beliefs about John's action or motives, I ignore them in moralizing, but rather input the situation as it would appear to someone who knew the facts (as I take them to be). Finding out how things actually are is generally a prerequisite for moral simulation. Finally, as in the case of aesthetic simulation, I need to recalibrate my ERS, this time so that biases resulting from my particular experiences, relationships, tastes, preferences, and background beliefs are eliminated as far as possible. In short, I must import the emotional response settings of an impartial participant. It is important that these settings are not turned to zero, so to speak: some things, like hitting a crying baby, will still give rise to reactive attitudes. With these inputs and settings, the off-line output of

²¹ The term is Walton's. His use may differ from mine in details.

²² Cruz and Gordon (2003: 11) term this sort of recentering of the egocentric map 'counterindexical pretending'. In David Lewis's (1979) terms, the *de se* content of the readjusted belief-input is centered on the target, while the *de dicto* content remains the same.

the emotional response system is my view of the impartial participant's reaction in the shoes of those affected. This quasi-emotion, in turn, in the context of moralizing, gives rise to my own moral attitude towards the action or agent, such as moral disapproval in the form of indignation (or other kind of specifically *moral* anger). As in mindreading simulation, this need not involve explicit belief and inference. Diagram 1 presents an outline of how such simulation might work. The dashed line represents recalibration, square shape emotional response settings, and circle emotion. The target is the principally affected person (or, if large numbers of people are affected, a somehow representative person).

Diagram 1. Simulation as impartial participant.



The move to simulating an impartial participant's responses involves transcending our own immediate reactions, as it were, *vertically* as well as horizontally, since we do not just imagine how we would feel in someone else's situation, but how we would feel in that situation *were we* free of the potential distortions deriving from our particular position and condition. This vertical transcendence is what makes possible the moral assessment of our own conduct, especially correcting for "the natural misrepresentations of self-love" (*TMS*

III.3.5: 137). We can place ourselves in our own shoes impartially, as well as in the shoes of those affected by our actions. If we take that any impartial spectator, an imaginary “representative of mankind” or “man within the breast”, would fail to sympathize with our actions or actual sentiments in our situation, we come to morally disapprove of ourselves (*TMS* III.1: 109–113). As Smith’s terms like ‘the representative of mankind’ suggest, he takes the imaginary impartial spectator within to be the product of internalizing the reactions of actual other people, presumably those who have de facto authority over us, like our parents and friends. (This meshes nicely with social psychology and the less zany aspects of Freud.) To be sure, coming to grasp the point of specifically moral thinking involves aiming to transcend the potential parochiality of internalized authorities, but it does appear to be a matter of fact that our conception of how an impartial person would react is coloured by the reactions of actual peers.

Disapproval Beyond Anger and Indignation

The Moving House case involves a particular kind of moral violation, namely failure of reciprocity. Injustice is done, a particular individual is a victim we can sympathize with, and punishment of some sort, a negative sanction, is in order. But as moral psychologists emphasize, people morally condemn many things that have little to do with justice or harm, particularly outside the enclave of Westerners with high socioeconomic status. For example, Rozin et al (1999) develop a model in which contempt, anger, and disgust (‘CAD’) correspond to violations of community, autonomy, and the divine (which is also the ‘natural’) order, respectively.²³ Clearly, the historical sentimentalist model has to be modified to account for these varieties of moral disapproval. We cannot simulate a victim’s anger if there is no victim. Yet people who condemn, say, cruelty toward animals or masturbating with a dead

²³ See also Prinz 2007 for a refined version of the CAD model. Prinz argues that contempt is a mix of anger and disgust rather than a basic moral emotion.

chicken (Haidt et al. 1993) are surely moralizing. A natural suggestion is that in these cases we characteristically simulate the reactions of a potential trustee (in the case of animals and others whose reactions lie beyond our simulation capacity) or onlooker in the guise of an impartial spectator and find we would feel other non-moral reactive attitudes than anger or resentment.²⁴

I will focus here on the case of disgust, which has received much attention lately.²⁵ What Rozin calls ‘core disgust’ is clearly neither a reactive nor a moral attitude – it is a reaction to contaminated food and other things we wouldn’t put into our mouth. It is readily generalized to regulate other things to do with bodily functions like defecating and sex. As Rozin et al. (1993) and Nussbaum (2004), among others, have noted, such extended disgust marks the difference between what we consider human and what we consider part of our animal nature. It is no surprise that it plays an important role in religious ethics: when we behave like animals, we violate the divine (natural) order of things. In keeping with its psychological origin, extended disgust motivates not so much to punish as to avoid and reject its object in order to maintain (metaphorical) purity. Can it be fit into the historical sentimental model of moral thought? As long as we can distinguish between mere disgust and disgust we imagine anyone whose advice we would take to share, something like this seems possible. Someone who finds homosexuality morally wrong is not merely disgusted by (the thought of) gay sex. Instead, he thinks that this is not just a fact about him, but a reaction that every normal or non-defective person would share, a reaction that is merited by its object and is thus authoritative. Call this sort of disgust *repugnance*.

Repugnance comes close to moral disapproval, but it is not yet enough, since we can imagine finding something repugnant and yet not morally wrong. Rotting flesh may be

²⁴ I will not discuss the case of non-persons further here. It seems to me that the trustee view developed by Scanlon 1998, suitably modified, would do the work needed.

²⁵ In addition to the authors already mentioned, see e.g. Nichols 2004, Nussbaum 2004, and van Willigenburg (MS).

repugnant to us without our morally disapproving anyone or anything. For moral repugnance, we need to qualify the kind of disgust at issue. It must be a kind of a quasi-reactive attitude, a reaction to the quality of someone's manifest will. This leaves room for a distinction between the liberal who finds masturbating with a dead chicken repugnant but not morally wrong (since no one is harmed) and the conservative who finds not only the behaviour but also the 'unnatural' desire it manifests repugnant, and so morally disapproves of it. With these qualifications, it seems disgust-driven disapproval may be close enough to paradigmatic moral cases. This is not, of course, to make the first-order moral judgment that we should be guided by disgust in moralizing; perhaps we should keep oral and moral disapproval strictly apart.²⁶ Given the nature of disgust, in particular the 'logic of contamination' that Rozin emphasizes, it may be impossible to quarantine one's own initial reaction from the simulation process. But in an account of what it is to moralize, we should aim to include even those whose disapproval has such lowly origins, as long as they are disposed to regard their initial reactions as subject to correction for bias and ignorance.

There are further varieties of disapproval and approval to account for, but I take it that it is reasonably clear how to extend the story to cover them. (This will be particularly easy if we embrace the taxonomy of the CAD model and Prinz's suggestion that contempt is derivative from both anger and disgust.) Including these extensions to the canonical sentimentalist account yields the following more precise statement of historical sentimentalism:

Historical Sentimentalism+

An attitude is moral only if it characteristically results from a process of simulating the non-moral (quasi-)reactive attitudes like anger, gratitude, and reactive disgust that any unbiased and informed participant whose advice on the topic the subject is

²⁶ Nussbaum argues that disgust embodies "magical ideas of contamination, and impossible aspirations to purity, immortality, and nonanimality, that are just not in line with human life as we know it" (2004: 14), so that we'd be better off not basing our laws or moral norms on disgust.

disposed to take would have a) in the circumstances of those affected by the action or b) as a spectator or trustee, in case no subject is directly affected.²⁷

I emphasize again that on the expressivist sentimentalist view, to think that something is wrong is simply to disapprove of it, characteristically after the sort of process described. It is not to *believe* that any impartial participant would react to it negatively; one may engage in simulation without ever having such thoughts. This is important, among other things, for making sense of moral disagreement. On the expressivist view, this is just disagreement in attitude. It is not disagreement about what such-and-such participant would think (which would be an empirical question), or what would make a participant worth listening to (which would itself be a normative question).²⁸

A Note on Other Varieties of Moral Judgment

I have formulated historical sentimentalism as a thesis about all-out moral judgments, in particular wrongness-judgment. Can it be extended to other kinds of moral judgment? I have no room here for a full discussion, but will quickly sketch how sentimentalism about pro tanto reason judgments would go.²⁹

Practical reasons, as everyone agrees, are considerations that count in favour of actions and attitudes, including emotions. They may be outweighed by other reasons, and plausibly add up to overall oughts. Jonathan Dancy has recently argued that even if expressivists can account for all-things-considered moral judgment, they cannot make sense of judgments about pro tanto or contributory reasons. Thinking that one has *some* reason to ϕ certainly cannot involve finding guilt or anger for failing to ϕ appropriate, as a simple case will show. Suppose George has promised to help John move, but ends up having to choose

²⁷ 'Or' here should be read inclusively. When we say that someone is a rotten character or that an attack on the civilian population is foul, we are plausibly expressing sympathy with both anger and disgust.

²⁸ This point was prompted by comments by an anonymous referee for OSME.

²⁹ This, again, is a point that was pressed by an anonymous referee for OSME.

between keeping the promise and jumping in the water to save a drowning Ringo, and does the latter. In this situation, I may well judge that George had some reason to help John, but more reason to jump in the water. I will not think that it is appropriate for George to feel guilt for failing to help John, since I believe that he did the right thing in the circumstances. But in what, then, can the thought that he had (an outweighed) reason to help John consist in on the sentimentalist account? In Dancy's words, "The difficulty is to make sense of the idea that one might approve of S's ϕ -ing to some extent while being entirely against it." (Dancy 2006: 57)

The basics of an answer can be found in Gibbard, according to whom "[t]o say that R is some reason for S to ϕ in C is to express acceptance of a system of norms that directs us to award some weight to R in deciding whether to ϕ in C" (Gibbard 1990: 163). The core idea is that thoughts about pro tanto reasons consist in all-out attitudes that are not directed toward *actions* but rather toward the *way in which* the agent arrives at the action. Gibbard's original formulation, however, leaves ambiguous what it is to award weight in deciding³⁰, and only applies to reasons that play a role in deciding and deliberation. To avoid these problems, I propose the following tentative formulation in the same spirit:

Sentimentalism about Reason Judgment

To judge that r is a pro tanto reason for A to ϕ in C is to approve, characteristically as a result of an objectivity-conducive process, of A's knowing³¹ r having causal influence in the direction of ϕ -ing on the mechanism(s) responsible for A's ϕ -ing or not in C.

Historical sentimentalist simulation is one kind of objectivity-conducive process, and thus characteristic of moral reason judgments. (Prudential (dis)approval deserves its own sentimentalist story, which I will not attempt to tell here.) Some such process is needed to make sense of the normative authority of reason judgments. After all, in the special case of first person thoughts about reasons, to think that r is a pro tanto reason for me to ϕ is not just

³⁰ If awarding weight involves finding something a favouring consideration, the analysis has gone nowhere, as Scanlon (1998: 58) points out.

³¹ This could be 'being aware of' or some other factive psychological verb. For simplicity, I talk of knowing.

to be motivated to ϕ , but to regard r 's influence as somehow appropriate. As formulated, the thesis defines appropriateness in permissive terms, but it could easily be modified to account for thinking that something is a requiring or peremptory reason: it is to *disapprove* of A 's knowing that r *not* influencing the mechanism responsible for A 's ϕ -ing. Being able to distinguish between thoughts about more or less insistent reasons is a nice feature of the proposal.³² It is formulated in terms of influencing the mechanism leading to action in order to make it as general as possible. It doesn't just apply to deliberation, but also to reasons for feeling a certain way and non-deliberate actions. For example, to think that her son's recovery from illness is a reason for Anna to be happy is to approve of Anna's knowing that her son has recovered tweaking her hedonic level-setting mechanism in the direction of happiness.

How does this work for the test case? I think the fact that George has promised to help John is a pro tanto moral reason for him to help. On the current proposal, this means I morally approve of George's knowing that he has promised to help influencing the mechanism that is responsible for his either helping or not in the direction of helping. Perhaps all this amounts to is that there are more counterfactual scenarios (nearby possible worlds) in which George helps than otherwise. I can have this attitude even if a) George does not know that he has promised to help (and hence won't help), b) George's knowing that he has promised does not in fact make a difference, c) I think that all things considered, George should not help John, and c) I think that George shouldn't actually be motivated to help John (since he should rather wholeheartedly embrace the project of saving Ringo's life). All of these are desirable consequences for an account of reason judgment. The last is one reason why I formulate the analysis in terms of influencing the relevant mechanism rather than motivation. I take it that knowledge of r can influence the mechanism responsible for ϕ -ing in the direction of ϕ -ing without motivating one to ϕ , for example by influencing one's deliberation about whether to ϕ so that in a slightly different situation, one would be

³² The need to make such a distinction among reasons is emphasized by Joshua Gert (2004), among others.

motivated to or even decide to ϕ , or by placing ϕ -ing on the reserve options file in the deliberative system, as something to be considered if the situation changes. There is no need to assume an isomorphism between pro tanto reasons and patterns of motivation.

The account can be straightforwardly extended to comparative and overall reason judgments. To think that r is stronger reason to ϕ than r' is to not ϕ is to approve of the following scenario: knowledge of r influences the mechanism responsible for ϕ -ing in the direction of ϕ -ing to a greater degree than knowledge of r' influences the mechanism in the direction of not ϕ -ing. If the mechanism is deliberation, it is to approve of knowledge of r having a stronger causal influence toward deciding to ϕ than knowledge of r' has toward deciding not to ϕ . To think that the fact that Ringo is about to drown is a stronger peremptory reason for George than the fact that George has promised to help John is to disapprove of its not having more influence on George's decision. To think that George has most or overall moral reason to jump in the water is to morally disapprove of some set of considerations not having decisive influence in leading George to jump in the water – in short, to morally disapprove of George not jumping (or not trying to jump) in the water. This way we get from comparative reason judgments to overall ought-judgments, meeting an important desideratum for an account of normative thinking.

Solving the Moral Attitude Problem

This is still a rough sketch of historical sentimentalism, but I trust we can already see how it solves the moral attitude problem. According to it, my disapproval of John is moral only if it (characteristically, but I will bracket this here) results from simulating the sentiments that any informed and non-defective impartial participant would have in Paul's situation and finding that she would resent John's motive and action. On the expressivist view, I voice this disapproval by saying that John's action was morally wrong. Is it an open question whether I

morally disapprove of John's action? Miller deployed two arguments against contemporary expressivists, direct and indirect. Against the direct argument, it does not seem conceivable that someone who disapproves of an action as a result of a sentimentalist process disapproves of it aesthetically or in some other non-moral sense, for the wrong kind of reasons. Simulating the anger or resentment of an impartial participant just does not figure in arriving at that kind of disapproval.³³

Miller's indirect argument, in turn, relies upon the putative conceptual truth that moral disapproval involves being disposed to demand others to share the very same (type of) attitude. This is just what one would expect on the historical sentimentalist picture. When we find that any impartial participant in Paul's shoes would feel anger for John's action and consequently disapprove of it, we are out of sympathy with those who do not share that reaction, and to be out of sympathy is to disapprove. This is a simple and plausible way of cashing out "demanding others to share the very same attitude".³⁴ Nor do cases like the Clinton contrition (as imagined above) cause a problem, since the judgment of appropriateness of guilt does not result from impartially simulating the reactions of all those affected to the action but rather (we may now stipulate) simulating the reactions of the family to the presence or absence of guilt itself. Thus, the account survives the further challenge D'Arms and Jacobson pose to Gibbard. So, in short, it seems historical sentimentalism solves the moral attitude problem.

But this is not all. Having the canonical sentimentalist aetiology is a *diachronic* property of moral disapproval. However, the historical sentimentalist is by no means committed to denying that moral disapproval does not have characteristic *synchronic* features as well, such as a functional role and perhaps even a typical phenomenal feel. The only

³³ If someone disapproves of a work of art, say, as a result of such a process, that counts as a moral disapproval on this account. And so it should – people who condemn, say, Andres Serrano's Piss Christ on account of (perhaps mistakenly) believing that any impartial spectator would share the resentment (or reactive disgust) of many Christians *are* moralizing.

³⁴ One need not be disposed to demand that others share the attitude as a result of the sentimentalist process, but just that they adopt an attitude that plays the same functional role in their mental economy.

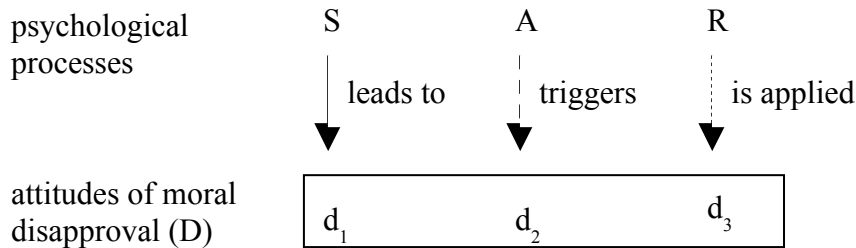
essential thesis is that the functional role and phenomenal feel are not sufficient or (in the case of the latter) necessary to distinguish it from other types of disapproval. Indeed, I believe that the right kind of functional role and characteristic sentimental etiology are individually necessary and jointly sufficient conditions of moral judgment. But in fact, the historical sentimentalist can go further. The synchronic functional role of the moral attitude – its characteristic authority and motivational impact – is *explained* by its origin. It makes a lot of sense that a disapproval that results from imagining how anyone in the patient’s situation (or any trustee or spectator) would feel carries the full weight of public blame and shame, and moreover blame and shame one takes to be merited. It is no surprise that competing inclinations are silenced. It makes sense that we would expect others – indeed anyone else – to hold the same attitude toward the object of evaluation. Thus, a sentimentalist aetiology makes intelligible something very much like Blackburn’s emotional ascent and the practical clout of moral judgment, and definitively closes the open question.³⁵

3. The challenge of hot and cold moral judgments

But surely common sense suggests that we can and often do morally disapprove of people and actions without going through any process of the sort the sentimentalists describe? Is not there such a thing as rash or self-serving moral judgment? Certainly we do, and there is. And if common sense is not enough, there are also a host of psychological studies pointing to the same. Sometimes we form what might be called *hot* moral judgments as a result of immediate affective reactions. At the other extreme, sometimes our judgments are *cold*, formed without any notable emotional process. Both of these types of judgment pose a *prima facie* challenge

³⁵ I discuss the relationship between diachronic and synchronic properties of moral judgment further in work in progress.

to the historical sentimentalist account. Using S for the canonical sentimentalist process, A for affect-laden reactions, and R for explicit rules, we can diagram the situation as follows:



The question for the historical sentimentalist is: how is it that token attitudes of disapproval triggered by affective reactions or formed as a result of applying rules count as moral ones, if moral disapproval is individuated by its sentimentalist aetiology? Why, and when, do attitudes like d_2 and d_3 belong to the same class of moral disapproval D as d_1 ? In this final section, I look at some of the psychological evidence and modify the historical sentimentalist thesis to accommodate it.

To begin with hot judgments, recent psychological literature is rife with studies purporting to show that people's moral judgments are influenced by brute affective reactions and seemingly irrelevant situational cues. To take just a few examples, Haidt, Koller, and Diaz (1993) found that especially people with a low socioeconomic status found disgust-arousing actions morally wrong without being able to give any account of why they did so; Wheatley and Haidt (2005) found that hypnotizing susceptible participants to experience disgust at the sight of random words resulted in a difference to their moral judgments about written scenarios³⁶; Valdesolo and DeSteno (2006) had people watch five minutes of comedy (Saturday Night Live) to put them in a good mood, and found that it made people more likely to judge that it is morally appropriate to push a fat man in front of a trolley to save five others.

³⁶ It is important that the hypnotized disgust was entirely rationally irrelevant to the moral status of the events of the story; for example, in one case, it was aroused by reading the word 'often'. The actions themselves were innocent, like 'fostering good discussions'.

This sort of results have led psychologists like Jonathan Haidt champion an affectivist (or ‘social intuitionist’, as he misleadingly calls it) theory of moral judgment, according to which most moral judgments of most people are triggered by a fast, effortless, and automatic affective system, whose outputs are subsequently ‘rationalized’ by confabulated stories that have nothing to do with their causal history, in case others demand justification (Haidt 2001, Haidt and Björklund 2008).³⁷ Clearly, in these cases we don’t engage in the canonical sentimentalist simulation.

The other empirical challenge comes from the opposite direction. First, we sometimes seem to make moral judgments simply as a result of categorizing some behaviour as falling under some rule we have internalized. We take it for granted that careless bombing of civilians is wrong without bothering with sympathy. More seriously yet, there appear to be people, such as autists, who are incapable of *ever* simulating other people’s reactions to situations, but nevertheless capable of distinguishing between violations of moral and conventional rules, and thus making moral judgments of at least some sort. Their judgments must always be *cold*, based on cool reasoning rather than sentiments resulting from sympathy.

Finally, small children, too, have limited ‘mindreading’ abilities, and consequently are not able to simulate the reactions of other people, or their own reactions in the place of another. Nevertheless, they are able to distinguish between violations of conventional and non-conventional rules. The latter are taken to be valid regardless of conventions, authorities, or place. Children as young as 3 ½ years old (Smetana 1981), judge behaviour that involves hurting others, such as throwing sand on another child’s face, to be wrong regardless of what an authority figure says – even if the authority figure is God and the community in question is

³⁷ It is worth noting that neither Haidt nor other affectivists such as Shaun Nichols offer a criterion for distinguishing moral from non-moral attitudes, or (in the case of Haidt) even from mere liking and disliking. I discuss this further in work in progress.

very religious (Nucci 1986). Whether children's judgments are hot, cold, or even hard-wired³⁸, they clearly don't involve the use of complex sympathy.

Asymmetric Dependence

If I accept the message of common sense and experimental results, how can I maintain that the distinctive feature of moral disapproval is its characteristic aetiology? The important thing is that all I am claiming is that the sentimentalist process is *characteristic* of moral disapproval, not that every token moral attitude results from it, or even that such aetiology is statistically normal. To begin with, it may well be that most of our moral attitudes are triggered by simple affective cues. But there is still a sense in which sentimentalist history is characteristic of moral attitudes. According to me, there is an *asymmetric logical dependence* between attitudes that result from sympathy and reflective correction, on the one hand, and gut reactions, on the other. What this means is roughly that to count as having moral attitudes at all, a person must at least *sometimes* arrive at approval and disapproval by way of a process that involves adopting (or trying to adopt) an impartial participant's perspective. It is not possible to always arrive at moral disapproval as a result of gut reactions and uncorrected sentiments.

In contrast, it *is* in principle possible to always arrive at moral disapproval as a result of sympathetic simulation of impartial resentment. In this sense, affect-based attitudes are *parasitic* on those that result from the sentimentalist process.³⁹ Further, the would-be moralizer must be disposed to *recognize* that this is the case. This need not, of course, be a matter of propositional knowledge, but simply acknowledging in practice that gut reactions do not entitle one to invest one's moral judgments with the sort of authority they have. This

³⁸ I leave out here theories on which moral attitudes are outputs of a dedicated moral module engaged in non-conscious, rule-governed computation much like the postulated 'language module'. See Hauser, Young, and Cushman 2008.

³⁹ I owe the idea of asymmetric logical dependence to Evan Simpson (1999: 202), who deploys the notion in a different context. Fodor's asymmetric dependence account of representation is another source of inspiration.

practical acknowledgement can just be a matter of seeking more information or waiting for oneself to calm down before making a judgment, or willingness to reconsider one's judgments in response to criticism that one has not looked at the issue from someone's perspective or that one has let one's cranky mood or instinctive disgust influence one's judgment.

To see why uncorrected attitudes are parasitic, imagine Bob, who feels outrage at everyone who beats him at bingo or outbids him at an auction. He also rages at the rain, when it causes him to have unpleasant experiences. Imagine, further, that his state is functionally and phenomenologically indistinguishable from the state of someone who disapproves of something as a result of the sentimentalist process. Bob, however, never engages in such an exercise of imagination, sympathy, and reflection – perhaps he lacks one or more of these capacities for some reason. In Haidtian terms, he is a complete social intuitionist.

Nevertheless, he expects everyone else to share in his disapproval, disapproves of those who do not feel guilt for beating him in a fair contest, and so on, and expresses this by calling people who do not do so 'vicious', 'odious', or 'depraved'. Is his disapproval moral? No! He is, at best, mimicking moralizing or trying to moralize. He has failed to grasp the point of the moral language game, so to speak. As Hume famously wrote,

When a man denominates another his *enemy*, his *rival*, his *antagonist*, his *adversary*, he is understood to speak the language of self-love, and to express sentiments peculiar to himself and arising from his particular sentiments and situation. But when he bestows on any man the epithets of *vicious* or *odious* or *depraved*, he then speaks another language, and expresses sentiments in which he expects all his audience are to concur with him. He must here, therefore, depart from his private and particular situation and must choose a point of view common to him with others; he must move some universal principle of the human frame and touch a string to which all mankind have an accord and symphony. (Hume, *Enquiry Concerning the Principles of Morals*, IX: 252)

Bob, who, recall, *never* departs from his 'private and particular situation', is not really moralizing at all, because he fails to recognize the demand to control his attitudes by way of

trying to reach the common point of view, the perspective of the impartial participant, and never does so.⁴⁰ At best, he *schmoralizes*.

The point is not that Bob is making a moral mistake, for that would already require being a moralist. To make this clearer, contrast him with two types of people who moralize *badly*. Sarah is a kind of ethical egoist, a person who thinks that everyone ought to do what most benefits her. She thinks she's morally special, and disapproves of the egalitarian approach of others. Nevertheless, she does try to put herself in the shoes of others and acknowledges the existence of perspectival distortions, but (no doubt mistakenly) thinks that any impartial spectator would sympathize with her sentiments.⁴¹ Her attitudes are still moral, even if we don't share them – she just moralizes badly, as seen from our normative perspective. Jane, in turn, is by temperament a rash and irresponsible moralizer. Most of the time, she makes knee-jerk judgments about people. She sees Mahmoud Ahmadinejad mocked on TV and immediately judges that he is evil and dangerous, and is willing to go high up on the emotional ladder on the issues. But sometimes, especially when her attitudes really matter, she is capable of putting herself in other people's place and considering whether everyone else would share her reactions, and these thoughts influence her approval and disapproval and reactions to others. When challenged on her gut reactions, she recognizes she needs to submit them to the sort of scrutiny sentimentalists describe. She acknowledges that it is a platitude that accusing someone without trying to 'walk a mile in their shoes' is not giving them their due. Jane is a bad moralizer in the sense that many of her attitudes do not merit the authority she gives to them, but she is still a participant in moral practice. It would not be a surprise if most of us turned to be a lot like her.

⁴⁰ Among other things, Bob is an additional counterexample to Gibbard's functionalist view – he illustrates the genealogical problem for all merely synchronic accounts of moral attitude.

⁴¹ Sarah thus manifests what is called an egocentric bias – she fails to quarantine properly her own attitudes and preferences from the simulation process, failing to adjust the settings of her emotional response mechanism. See Goldman 2006: 41–42, 164–173 for discussion of similar phenomena and references to empirical studies.

Thus, the fact that we make many of our moral judgments on the basis of affective reactions, without going through the sort of process of simulation and correction that historical sentimentalism requires, does not threaten the claim that it is just some such process that essentially and characteristically differentiates moral attitudes from others. Not all moral judgments could be hot – and if they were, it would be a mystery why they have the sort of intrapersonal and interpersonal authority they do.

How about cold judgments? As Hume and Smith themselves point out, some of our judgments are directly based on *rules* and principles rather than occurrent sentiments. For example, one may adopt the principle that embezzlement is morally wrong and make the corresponding judgment in a particular case without engaging one's imaginative and sympathetic capacities. But for the sentimentalist, such a rule is always an inductive generalization from sentimental verdicts on particular cases, and is adopted for pragmatic reasons (sympathetic simulation is costly in terms of time and effort) or to check likely distortions of self-love.⁴² For Hume, this is particularly important in the case of 'artificial virtues' like justice, since they may require overriding our natural sympathy in particular cases – for example, justice may demand resolving an estate dispute in favour of a rich, foolish bachelor who is my enemy instead of my poor friend who is a man of sense (*Treatise* III.II.6: 532). So sentimentalism can rather handily accommodate a broad class of cold rule-based, non-sentimental judgments, as long as they are parasitic on sentimental ones in the sense that the rules are ultimately based on them.

Autists, however, still remain a challenge, since they plausibly *never* themselves simulate the reactions of others. Shaun Nichols has argued that autists show that what he calls perspective-taking accounts of moral judgment must be empirically inadequate (Nichols 2004: 10–11). I agree that they pose a challenge to historical sentimentalism, which is a kind of perspective-taking account. However, I want to argue, autists' ability to make moral

⁴² See *TMS* III.4: 159–161, VII.iii.2: 319–320.

judgments, I want to suggest, is parasitic on the ability of normal moral judges much as the ability of colour-blind people to make colour-judgments is parasitic on the ability of people with normal vision. They must *defer* to others in judging when, say, guilt is appropriate. As Jeannette Kennett puts it, “In the case of many, perhaps most, autistic people, rules of conduct are not self-developed, or are somewhat naïve ... [they] are further handicapped by the literal nature of their thinking in recognizing when and how the rules apply in novel situations” (Kennett 2002: 351). Similarly, Victoria McGeer suggests that it is true of many autists that while they are capable of following moral rules, and even construct elaborate systems of them, they fail to grasp the underlying point of the rules precisely because they are lacking in sympathy (McGeer 2008). This is manifest, for example, in rigid application to cases that cry out for exception in light of the *raison d’être* of the rule. Since such autists depend on others who do grasp the point of having such rules in order to engage in the practice, their approval or disapproval counts as moral only in the same sort of way as a normal moralizer’s attitudes resulting from short-sighted and partisan verdicts count as moral, except here the asymmetric dependence is *social* rather than merely psychological in nature.

Kennett does argue that some autists, such as the noted animal scientist Temple Grandin, are capable of independent moral thinking, and claims that this supports Kantian over Humean theories of moral judgment. There is no space to discuss this here properly, but a closer examination of her own main example suggests that a Smithian version of sentimentalism is well placed to explain them. First, the judgements of conscientious autists are based on the explicit application of rules and the “hard-won realization that other people have needs and feelings different from [their] own” (Kennett 2002: 352), which is compatible with Smith’s view on which sympathy need not involve taking on the *actual* feelings of others. The autist still has feelings, so she should in principle be capable of running her emotional response system offline to generate new rules after realizing that they reflect a

common point of view. Further, Kennett's main example of an autistic 'moral innovator', Temple Grandin, is an unusually high-functioning autist and clearly capable of empathy, particularly with animals and other autists. For example, as Oliver Sacks reports, she is capable of consciously deceiving others, which low-functioning autists are not able to do precisely because of their inability to adopt other people's point of view.⁴³ I thus conclude that the empirical case of autism does not show that cold moral judgments could stand on their own.

The Kantian challenge is not quite so easily dispelled, however. For could there not be a dedicated Kantian who *never* sympathized with anyone, someone "cold in temperament and indifferent to the sufferings of others" (Kant 1785/1948: 398) who nevertheless concluded, on the basis of reasoning alone, that it is wrong to break a promise? Would that not be a genuine moral judgment rather than a schmoral one?⁴⁴ If so, sympathy cannot be necessary. Let us grant for the sake of argument something that is far from obvious, namely that some Kantian procedure indeed yields unique answers for what to do in particular situations. Since the hypothetical Kantian would be an independent, non-parasitic moralizer, the responses given above to the empirical challenge of autism do not suffice to counter this. The sentimentalist has basically two alternative strategies. The first is simply to bite the bullet. If the sentimentally dead Kantian's judgment had the functional role distinctive of moral judgment, it would be a freak coincidence. She would still have failed to grasp the point of the exercise – her actions would be guided by universalizable principles, but not because such principles facilitate coordination, cooperation, and mutual happiness, but out of some fetishism for avoiding practical self-contradiction for its own sake. Consequently, she would not really be moralizing, but only schmoralizing. Of course, the Kantian would say that this is

⁴³ Sacks relates an incident in which Grandin smuggles him into a meat-packing plant in a disguise, and notes: "I was astonished at this, for autistic people, it is said, have no pretend-play, and here Temple had, very coolly, and without the slightest hesitation, determined on a subterfuge and was all set to smuggle me into the plant." (Sacks 1995: 265)

⁴⁴ This point was made by several members of the audience at the Metaethics Workshop.

no fetishism, but proper reverence for reason and consequently respect for rational agents, which is what morality is all about. Unsurprisingly, this line of response thus leads to the general debate about the possibility of pure practical reason and its role in morality, which I cannot hope to adjudicate here. The second sentimentalist response strategy is to say that insofar as the pure Kantian's judgment is motivationally authoritative, that is because she is already tacitly engaging her sentimental resources. If the hypothetical Kantian is actually of this sort, she invests the reason-based judgment forbidding the use of another person as a mere means with a special authority because she antecedently sympathizes with the negative reaction of any impartial participant against using others as mere means. Again, this leads to a broader debate about the motivational authority of reason, which cannot be pursued further here.⁴⁵

Finally, what about small children? Their attitudes may be based either on affective reactions or explicit rules, but clearly not on a complex sentimentalist process. My basic response is to deny that their judgments (or corresponding attitudes) are moral. First, let us note that the justifications children offer for non-conventional judgments often *do* refer to other people's mental states – the reason why it is wrong, regardless of what authorities say, to throw sand in the face is that “it hurts” (Smetana 1981). So to some extent, children do seem to exercise sympathy precisely when they make non-conventional judgments. Second, there are non-question-begging reasons to think that children are just *moralizers-in-training*. To begin with, they do not appear to feel guilt either, perhaps until the age of seven.⁴⁶ Further,

⁴⁵ Similar considerations arise in the case of divine command theory, or any first-order theory that tells us not to base our moral judgments on sympathy but some given rules. (This objection was also raised in Madison.) Either people who follow such rules are schmoralizers (because they don't grasp the point of the rules, but make a fetish out of following perceived divine orders as such) or they tacitly engage their sympathetic abilities (which is supported by the fact that actual religious believers tend to obey the very rules that are most likely to receive sympathetic backing (thou shalt not kill) rather than rules that are arbitrary from the perspective of the common point of view).

⁴⁶ Nichols 2004: 90–95. Nichols takes this to be an argument against expressivism: since children make moral judgments without feeling guilt, guilt is not essential to making moral judgments. See also Merli 2008. I'm suggesting it is rather the other way around – since children don't feel guilt, they don't really make moral judgments either. This is not begging the question, since historical sentimentalism, unlike Gibbard, doesn't appeal to guilt in individuating moral judgments.

when faced with a more complex situation, like punishment, children do not seem to consider, or recognize the need to consider, the back story or the intentions of the punishers, which suggests that many judgments are merely triggered by affective reactions. As Gibbard suggests, they do not distinguish between something feeling reprehensible and its actually being such (the reprehension being warranted in terms of an impartial perspective), which is an essential part of mature moral competence (cf. Gibbard 2006: 205 ff.) Finally, it is not clear that morality has practical clout for them – which is not surprising if they are incapable of reactive attitudes like guilt or indignation. Moral judgments do not play the same functional role in their motivational economy as they do with adults. In short, independently of sentimentalism, there is good reason to believe that children’s judgments are not yet part of full-blown moral thinking – as Gibbard puts it, their judgments are ‘near-moral’ (Gibbard 2006: 203). Indeed, the sentimentalist story predicts precisely these shortcomings, given their underdeveloped simulation capacities and conceptual repertoire. Children’s judgments may be *non-conventional* without yet being fully *moral*.

The historical sentimentalist thesis needs to be refined to account for these alternative, secondary ways of arriving at moral disapproval. Here is a first pass:

Refined Historical Sentimentalism (RHS)

An agent A’s token attitude of disapproval *d* is moral only if either a) *d* results from the canonical sentimentalist process *S* or b) *d* belongs to a functional type *D* whose features and presence in A’s psychology are explained by A’s at least occasionally engaging in *S*, or by A’s emulating others who at least occasionally engage in *S*.

RHS states a disjunctive necessary condition for an attitude of moral disapproval. The first disjunct is essentially HS+. The second draws on the observation that any attitude triggered by a feeling or resulting from rule application could not possibly be moral unless it belongs to a functional type whose authority for the agent is explained by her at least occasionally engaging in complex sympathy. People who form such attitudes exclusively on the basis of affective reactions or application of rules have missed the point of moralizing.

In sum, it seems that the fact that not all moral judgments are made as a result of sentimental impartial sympathy does not undermine a historical sentimentalist account of moral attitudes. I have here presented an expressivist-friendly version of such an account, but it is worth noting that it need only be slightly modified to arrive at a version of cognitivism, according to which moral judgment consists in belief in the appropriateness of sentiments. The battle between expressivist and cognitivist versions, I believe, must be settled on other than psychological grounds. In any case, since historical sentimentalism not only solves the moral attitude problem but also explains the typical functional features of moral judgments, it has much to recommend it. So Hume was, broadly speaking, correct in the passage I use as my epigraph: a sentiment is moral only if it results from consideration of its object in general, without reference to our particular interest.⁴⁷

⁴⁷ Comments, questions, and suggestions by Matthew Chrisman, Frans Jacobs, Lilian O'Brien, Jonathan Schaffer, Jussi Suikkanen, Theo van Willigenburg, participants at the April 2008 Expressivism Workshop at St Andrews and the September 2008 Metaethics Workshop in Madison, and two anonymous referees for Oxford University Press materially improved this paper, and I believe any impartial observer would share in my gratitude.

References

- Blackburn, Simon (1998) *Ruling Passions: A Theory of Practical Reason* (Oxford: Clarendon Press).
- ____ (2002) 'Précis of *Ruling Passions*,' *Philosophy and Phenomenological Research*, 65: 122–35.
- Cruz, Joe and Gordon, Robert M. (2003) 'Simulation Theory,' *Nature Encyclopedia of Cognitive Science* (London: Macmillan), 9–14.
- Dancy, Jonathan (2004) *Ethics Without Principles* (Oxford: Clarendon Press).
- ____ (2006) 'What Do Reasons Do?', in T. Horgan and M. Timmons (eds.) *Metaethics After Moore* (Oxford: Oxford University Press), 39–59.
- D'Arms, Justin and Jacobson, Daniel (1994) 'Expressivism, Morality, and the Emotions,' *Ethics*, 104: 739–63.
- Darwall, Stephen (1999) 'Sympathetic Liberalism: Recent Work on Adam Smith,' *Philosophy and Public Affairs*, 28: 139–64.
- ____, Gibbard, Allan, and Railton, Peter (1992) 'Toward *Fin de siècle* Ethics: Some Trends,' *The Philosophical Review*, 101: 115–89.
- Gert, Joshua (2004) *Brute Rationality: Normativity and Human Action* (Cambridge: Cambridge University Press).
- Gibbard, Allan (1990) *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge, Mass.: Harvard University Press).
- ____ (2003) *Thinking How to Live* (Cambridge, Mass.: Harvard University Press).
- ____ (2006) 'Moral Feelings and Moral Concepts,' in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics Vol. 1* (New York: Oxford University Press), 195–215.
- Goldman, Alvin (2006) *Simulating Minds* (Oxford: Oxford University Press).

- Gordon, Robert M. (1986) 'Folk Psychology as Simulation,' *Mind and Language*, 1: 158–71.
- ____ (1995) 'Sympathy, Simulation, and the Impartial Spectator,' *Ethics*, 105: 727–42.
- Haidt, Jonathan, Koller, Silvia H., and Dias, Maria G. (1993) 'Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?', *Journal of Personality and Social Psychology*, 65: 613–28.
- Haidt, Jonathan (2001) 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment,' *Psychological Review* 108: 814–34.
- ____ and Björklund, Frederik (2008) 'Social Intuitionists Answer Six Questions About Moral Psychology,' in Sinnott-Armstrong (ed.) (2008), 181–217.
- Hare, Richard Mervyn (1981) *Moral Thinking: Its Levels, Methods, and Point* (Oxford: Oxford University Press).
- Hauser, Marc, Cushman, Fiery, and Young, Liane (2008), 'Reviving Rawls's Linguistic Analogy,' in Sinnott-Armstrong (ed.) (2008), 107–43.
- Hume, David (1739-1740/1978) *A Treatise of Human Nature*. Ed. L. A. Selby-Bigge, 2nd, rev. ed. P. H. Nidditch (Oxford: Clarendon Press).
- Hume, David (1751/1948) *Enquiry Concerning the Principles of Morals*, in H. D. Aiken (ed.) *Hume: Moral and Political Philosophy* (New York: Hafner Press), 171–291.
- Joyce, Richard (2006) *The Evolution of Morality* (Cambridge, Mass.: MIT Press).
- Kant, Immanuel (1785/1948). *Groundwork of the Metaphysic of Morals*. Tr. H. J. Paton (London: Hutchinson).
- Kennett, Jeannette (2002), 'Autism, Empathy, and Moral Agency,' *Philosophical Quarterly*, 52: 340–57.
- Lewis, David (1979) 'Attitudes De Dicto and De Se,' *The Philosophical Review*, 88 (4): 513–43.

- McDowell, John (1981/1998) 'Projection and Truth in Ethics,' reprinted in *Mind, Value, and Reality* (Cambridge, Mass.: Harvard University Press), 151–66.
- McGeer, Victoria (2008) 'Varieties of Moral Agency,' in W. Sinnott-Armstrong (ed.), *Moral Psychology. Vol. 3* (Cambridge, Mass.: MIT Press), 227–57.
- Merli, David (2008) 'Expressivism and the Limits of Moral Disagreement,' *Journal of Ethics*, 12: 25–55.
- Miller, Alexander (2003) *An Introduction to Contemporary Metaethics* (Cambridge: Polity Press).
- Nichols, Shaun (2004) *Sentimental Rules. On the Natural Foundations of Moral Judgment* (Oxford: Oxford University Press).
- Nucci, Lawrence (1986) 'Children's Conceptions of Morality, Social Conventions and Religious Prescription,' in C. Harding (ed.), *Moral Dilemmas* (Chicago: Precedent Press).
- Nussbaum, Martha (2004) *Hiding from Humanity. Disgust, Shame, and the Law* (Princeton, Princeton University Press).
- Prinz, Jesse (2007) *The Emotional Construction of Morals* (Oxford: Oxford University Press).
- Prinz, Jesse (forthcoming), 'Is Empathy Necessary for Morality?'
- Rawls, John (1971) *A Theory of Justice* (Oxford: Oxford University Press).
- Rozin, Paul, Haidt, Jonathan, and McClauley, C. R. (1993) 'Disgust,' in M. Lewis and J. Haviland (eds.) *Handbook of Emotions* (New York: Guilford).
- Rozin, Paul., Lowery, Laura, Imada, Sumio, and Haidt, Jonathan (1999) 'The CAD Hypothesis: A Mapping between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity),' *Journal of Personality and Social Psychology*, 76 (4): 574-86.
- Sacks, Oliver (1995) *An Anthropologist on Mars* (New York: Picador).

- Scanlon, Thomas (1998) *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press).
- Schroeder, Mark (2008) *Being For* (Oxford: Oxford University Press).
- Simpson, Evan (1999) 'Between Internalism and Externalism in Ethics,' *Philosophical Quarterly* 49: 201–14.
- Sinnott-Armstrong, Walter (ed.) (2008). *Moral Psychology, Vol. 2. The Cognitive Science of Morality: Intuition and Diversity* (Cambridge, Mass.: MIT Press).
- Smetana, Judith G. (1981) 'Preschool Children's Conceptions of Moral and Social Rules,' *Child Development*, 52: 1333–6.
- Smith, Adam (1759-1790/1976) *The Theory of Moral Sentiments*, ed. D. D. Raphael and A. L. Macfie (Oxford: Oxford University Press). (TMS)
- Strawson, Peter (1963/2003) 'Freedom and Resentment,' reprinted in Gary Watson (ed.), *Free Will* (Oxford University Press), 72–93.
- Valdesolo, Piercarlo and DeSteno, David (2006) 'Manipulations of Emotional Context Shape Moral Judgment,' *Psychological Science*, 17: 476–7.
- Walton, Kendall (1997) 'Spelunking, Simulation, and Slime,' in M. Hjort and S. Laver (eds.), *Emotion and the Arts* (Oxford: Oxford University Press), 37–49.
- Wheatley, Thalia and Haidt, Jonathan (2005) 'Hypnotically Induced Disgust Makes Moral Judgments More Severe,' *Psychological Science*, 16: 780–4.
- Wiggins, David (1987) 'A Sensible Subjectivism?', in *Needs, Values, and Truth* (Oxford: Blackwell), 185–214.
- Willigenburg, Theo van (MS) 'The Influence of Emotions on Moral Judgment: Disgust as a Poignant Case'