Will the Real Moral Judgment Please Stand up? The Implications of Social Intuitionist Models of Cognition for Meta-Ethics and Moral Psychology

Author(s): Jeanette Kennett and Cordelia Fine

Source: *Ethical Theory and Moral Practice*, Feb., 2009, Vol. 12, No. 1, Empirically Informed Moral Theory (Feb., 2009), pp. 77-96

Published by: Springer

Stable URL: https://www.jstor.org/stable/40284273

Ethic Theory Moral Prac (2009) 12:77–96
DOI 10.1007/s10677-008-9136-4

Will the Real Moral Judgment Please Stand Up?

The Implications of Social Intuitionist Models of Cognition for Meta-ethics and Moral Psychology

Jeanette Kennett · Cordelia Fine

Accepted: 29 September 2008 / Published online: 13 November 2008

Abstract The recent, influential Social Intuitionist Model of moral judgment (Haidt, Psychological Review 108, 814–834, 2001) proposes a primary role for fast, automatic and affectively charged moral intuitions in the formation of moral judgments. Haidt's research challenges our normative conception of ourselves as agents capable of grasping and responding to reasons. We argue that there can be no 'real' moral judgments in the absence of a capacity for reflective shaping and endorsement of moral judgments. However, we suggest that the empirical literature indicates a complex interplay between automatic and deliberative mental processes in moral judgment formation, with the latter constraining the expression and influence of moral intuitions. We therefore conclude that the psychological literature supports a normative conception of agency.

Keywords Meta-ethics · Moral judgment · Automatic processing · Moral intuitions · Moral agency · Reason-responding · Reason-tracking

1 Introduction

The broad context of this paper is the ongoing debate in meta-ethics and moral psychology about the nature of moral judgment. Are moral judgments sentimental responses to agents, acts, or situations or are they the products of reason arrived at through the employment of our cognitive faculties? Are moral judgments intrinsically motivating or is the connection between moral judgment and motivation contingent on factors external to the judgment process itself?

J. Kennett (✉)
Centre for Applied Philosophy and Public Ethics,
Australian National University and Charles Sturt University,
Haydon Allen Building (Building 22), Canberra, ACT 0200, Australia
e-mail: jeanette.kennett@anu.edu.au

C. Fine
Centre for Applied Philosophy and Public Ethics,
University of Melbourne, Melbourne, Australia

In recent years there has been a surge of research into moral cognition and judgment in the social and cognitive sciences and the hope has been that this work might inform philosophical debates. Can the empirical turn in moral psychology resolve the question of which is the 'real' moral judgment—the best deserver of the title? Is it the judgment driven by sentiment or is it the product of rational reflective processes?

The focus of this paper is on the contribution that dual processing models of cognition might make to our understanding of moral judgment and more significantly to our understanding of the nature of moral agency. Do they—can they—offer empirical vindication of one or other philosophical position? Jonathan Haidt has argued from research which indicates that our moral attitudes are typically the product of fast, automatic, affectively charged processing to a version of sentimentalism about moral judgment. Haidt's research challenges not just a cluster of meta-ethical views of moral judgment but our normative conception of ourselves as agents capable of grasping and responding to reasons. If moral judgments must be made by agents so conceived then it may be that there are no 'real' moral judgments. We think that a closer examination of the interaction between automatic and controlled reflective processes in moral judgment provides some counter to scepticism about our agency and makes room for the view shared by rationalists and sophisticated sentimentalists alike that genuine moral judgments are those that are regulated or endorsed by reflection.

2 Automatic *Versus* Controlled Attitudes

Dual process models of attitudes, like other dual process models of cognition, contrast automatic processes (also referred to as implicit, reflexive, impulsive or System 1 processes) with controlled (or explicit, reflective, deliberative, or System 2) processes. Automatic processes can be understood as "control of one's internal psychological processes by external stimuli and events in one's immediate environment, often without knowledge or awareness of such control" (Bargh and Williams 2006, p.1). Thus, stimuli and events automatically activate a pattern of associations that "fill in information, quickly and automatically, about the characteristics that previously have been observed or affective reactions that previously have been experienced, in situations that resemble the current one." (Smith and DeCoster 2000, p.110). Characteristics stereotypically associated with such stimuli (e.g., for black men: aggressive, criminal, athletic) become activated, thus making characteristics that are consistent with the stereotype more accessible to consciousness (see Kunda and Spencer 2003). Research also suggests that most, if not all, stimuli also activate evaluative associations: an automatic attitude (positive or negative) that is immediate, unintentional and preconscious (e.g., Fazio 2001; Duckworth et al. 2002). We may even become aware of the evaluative tone of our automatic activations in the form of a positive or negative 'gut feeling'. For example, Haidt (2007, p. 998) notes that that "[w]hen we think about sticking a pin into a child's hand ... most of us have an automatic intuitive reaction that includes a flash of negative affect." However, in other cases (possibly those that are less strongly affectively charged) we may not have such introspective access (see Gawronski et al. 2006 for discussion of ways in which automatic attitudes can be 'unconscious').

In contrast to automatic processes, prototypical controlled processes are characterized as slow and effortful, and intentionally or consciously deployed. As Sherman et al. (2008) have pointed out, a central tenet of dual process models is that controlled processes represent resource-dependent self-regulatory processes that control the effects of automatically

activated implicit processes, inasmuch as their effects may conflict with consciously held beliefs or goals. Thus in the case of judgments, the role of self-regulatory control processes is to "overcome inappropriate automatic influences" (p. 320). The exertion of self-regulatory resources thus enables judgments and behaviours that are more accurate and/or in keeping with the individual's consciously endorsed values. Such controlled processes, however, are thought to take place only to the extent that the individual is both motivated and able to engage in them (e.g., Fazio 1990; Fazio and Olson 2003; Strack and Deutsch 2004.

One of the most popular paradigms for measuring 'automatic attitudes' is the Implicit Association Test (IAT: Greenwald et al. 1998). Commonly, this computerized test measures the relative speed at which people are able to pair two categories of object (e.g., Black faces and White faces) with pleasant and unpleasant stimuli. Participants are required to respond as quickly as possible on the test, thus limiting the extent to which their behavior can be modified by slower, controlled processing. The typical finding is that participants are slowed on 'incongruent' trials (in which Black faces are paired with pleasant stimuli and White faces are paired with unpleasant stimuli), relative to 'congruent' trials in which the opposite pairings are required (e.g., Greenwald and Krieger 2006). This is taken to indicate that Black people automatically activate associations that are more negative than those activated by White people; that is, they have an implicit bias against Black people relative to Whites. Some psychologists have made the important point that no behavior or judgment is the outcome of *purely* automatic or controlled processes (e.g., see Payne et al. 2005b); both will be involved to some degree. While acknowledging that even implicit measures are unlikely to be free of all controlled processing influence, as a convenient shorthand we will refer to such measures as reflecting 'automatic attitudes'.

The roles of automatic and controlled attitudes in judgment and behaviour have been valuably documented in the prejudice control literature. In line with the theorized importance of both motivation and available cognitive resources in overcoming unwanted automatic influence, the prejudice control literature often finds that automatic social attitudes have a greater influence when behaviours or judgments are made spontaneously (e.g., Dovidio et al. 1997; Rydell and McConnell 2006), under time pressure (e.g., Payne 2001) or when self-regulatory resources are either temporarily (e.g., Govorun and Payne 2006; Hofmann et al. 2007) or chronically low (e.g., Payne 2005). For example, Dovidio et al. (1997) found that automatic racial attitudes, but not self-reported attitudes, predicted eye gaze with a black person. Rydell and McConnell (2006) found that participants' choice of where to place their chair next to that supposedly about to be occupied by a (fictional) character called 'Bob' was predicted by participants' automatic attitudes towards Bob, but not their self-reported attitudes.

By contrast, studies in which people are asked to respond in less spontaneous ways, or in less cognitively demanding conditions, find that their behaviour is better predicted by self-reported attitudes rather than automatic ones. For example, Dovidio et al. (1997) found that self-reported racial attitudes, not automatic attitudes, were the best predictors of jury-like decisions about crimes made by black people. Similarly, Rydell and McConnell (2006) found no relation between people's automatic attitudes towards the fictional 'Bob' and their ratings of how much they would like him as a room-mate, friend, and so on. Rather, these were predicted by their self-reported attitudes towards Bob. Research also indicates that whether evaluations of others are based on activated stereotypes, or more accurate, individuating information, depends on motivation to be accurate, and availability of cognitive resources for controlled processing (see Kunda and Spencer 2003).

Social psychological research has tended to focus on attitude-objects for which there can be a low correspondence between implicitly and explicitly measured attitudes (see, for

example, Hofmann et al. 2005; Nosek 2007, for data and discussions of the parameters that influence discrepancy between the two types of measure). Of course this is not always, or even usually, the case. For example, Greenwald et al. (2003) found a very high correspondence between automatic and self-reported attitudes towards George Bush relative to Al Gore during the 2000 US Presidential election. We hypothesise that this will be the case for many morally-loaded attitudes. But for cases where there is conflict between judgments based primarily on automatic intuitions, and those that are the outcome of additional controlled processing, which attitude best constitutes or represents the agent's moral stance? Which is the 'real' moral judgment? Is it the one based on the ubiquitous automatic attitude that 'leaks' into the agent's socially significant behaviours, or that can become expressed verbally when the agent is unable to discount, cognitively elaborate, invalidate or disguise it? Or is it the attitude expressed in judgments at times when the agent is motivated and cognitively able to engage in self-regulatory control of automatic attitudes?

3 Moral Judgment, Normative Authority and Reasons for Action

There are a number of ways in which philosophers might approach the question of which is the 'real' moral judgment. First we might ask which process, automatic or controlled, is most *influential* in the determination of moral attitudes or moral judgments. Second we can ask which process best answers to our *concept* of moral judgment. Third, we can ask which judgment has *normative authority.*

Which process determines expressed moral judgment is a question that seems amenable to empirical enquiry and it is no surprise that social psychologists have focused their attention on this question. But the latter questions are of at least as much importance to philosophers. R Jay Wallace (1999) has argued that, whether or not this is a conceptual truth, morality is very widely taken to be a normative domain. Theories of moral judgment, whether emanating from philosophy or social psychology, thus need to account for the normative force of moral judgment. In this paper we adopt the view, argued for by philosophers such as Michael Smith (1994) and John Deigh (1995) that moral judgments are or invoke reasons claims. Says Smith: "It is a conceptual truth that claims about what we are morally required to do are claims about our reasons." (p.84) Deigh concurs: "One cannot coherently think, I ought to do such and such, though there is no reason for my doing it". (p. 748) We think that the view that moral claims are at least intended by those who make them to be reasons claims is not a philosopher's invention. It is implicit in everyday moral discourse where claims of right and wrong are clearly meant to identify or invoke considerations for or against courses of action.[1] As James Rachels says:

> If someone tells you that a certain action would be wrong ...you may ask why it would be wrong and if there is no satisfactory answer you may reject that advice as unfounded...moral judgements require backing by reasons. This is a point about the logic of moral judgement...One must have reasons or else one is not making a moral judgement at all. (Rachels 1993)

[1] We acknowledge that talk about 'our' concept of moral judgment is ambiguous between the concept actually held by the folk or a version that refines and systematises elements of the folk concept. Perhaps we could test the first empirically by doing a survey but we think it unlikely that this will deliver a clear answer. The accounts we favour do the second. As one of us has argued elsewhere (Kennett 2006) we think the concept that makes best sense of moral practices which are universal—of offering justifications, excuses, and of holding responsible—is expressed in the platitude that moral requirements have the status of reasons.

Moreover if this conceptual claim is correct, it identifies the source of the normative authority of moral judgment.

Smith and others have noted that the term 'reason' is ambiguous between explanation and justification (Woods 1972; Smith 1987). Smith distinguishes between reasons which motivate and explain action, and reasons which serve to justify action. Sometimes when we talk of someone's reason for doing something we are simply referring to their purpose in so doing. Why did Jack open the fridge? The answer 'to get a cold beer' gives us an explanation. Jack wants a cold beer and believes that there is one in the fridge. We can thus make sense of Jack's behaviour. We can similarly make sense of the behaviour of the family dog who scratches at the back door when he wants to go out. In both cases it is appropriate to say that Jack's reason for going to the fridge is that he wants a beer and the dog's reason for scratching at the back door is that he wants to go out. But these explanatory or motivating reasons are surely not the reasons invoked by moral claims. Moral judgments are judgments about justifying or normative reasons and these may come apart from motivating reasons. Jack's action may be explained by his desire for a cold beer without being justified by it. If it is Jack's tenth beer for the afternoon and he has sole care of an infant we would think that he ought not to get another beer—in other words we would think that his normative reasons speak against getting another beer.

When an individual makes a moral judgment, it is plausible to suppose that the reasons implied or adduced in support of the judgment must be reasons which the agent herself (albeit perhaps mistakenly) takes to justify and not merely to explain the judgment. Otherwise the judgment can have no normative authority for her. On the face of it, judgments determined by automatic processes that occur below the level of consciousness would seem to lack this authority and so could not count as the agent's 'real' moral judgments, at least absent a process of reflective endorsement through which the agent identifies and aligns herself with the considerations that she takes to justify the initial automatic moral attitude. This is the view we will defend. But recent work in both philosophy and social psychology casts doubt on this position.

4 Reason Tracking and Reason Responding

In her paper ' Emotion, Weakness of Will and the Normative Conception of Agency' Karen Jones (Jones 2003) outlines two distinct ways in which creatures might latch onto features of their circumstances which do in fact constitute reasons for them. Note that in setting up this distinction Jones' focus is on reasons which *in fact* justify a course of action rather than with the reasons *taken by the agent* to justify a judgment or course of action.

First, Jones says, a creature might be a *reason tracker.* Reason trackers are "capable of registering reasons and behaving in accordance with them but ... need not possess the concept of a reason nor have a self conception." Second, a creature might be a *reason responder.* Reason responders are "agents capable of tracking reasons in virtue of responding to them as reasons". Non human animals may be excellent reason trackers but so far as we know they do not have the cognitive capacity to respond to reasons, as reasons. But persons are both reason trackers and reason responders. The distinction Jones draws between reason tracking and reason responding thus appears to map onto the dual processing model of human cognition. Reason tracking is fast and automatic with only the results available to consciousness. Reason responding is a slower, controlled, deliberative process. What is significant for our purposes is the suggestion that either process may enable us to latch onto reasons which justify moral judgment and action. Jones suggests that

it's an empirical question which mechanism best latches onto reason giving considerations. If normative authority goes with the reasons which actually justify a judgment rather than with the reasons the agent takes to justify a judgment, then the question of which process delivers judgments that have normative authority is also an empirical question.

Jones' argument proceeds as follows. As practical agents, we are trying to latch onto considerations that *really are* reason-giving for us in a situation. Our affective responses sometimes key us to the presence of real and important reason giving considerations—that is they can be reason tracking. We may form judgments about situations or actions directly on the basis of our affective responses, without being able to articulate the reasons for the judgment. These judgments often enough run counter to those conscious deliberative reason responding judgments for which we can provide an explicit justification. Qua reason responders we may fail to be reason trackers. Jones discusses an example from Nomi Arpaly of a young woman who gives up a PhD program against her all things considered judgment and is unable until much later to identify the reasons which did indeed justify her decision to withdraw. The case of Huckleberry Finn is another where we think that the (objective) justifying reasons are all on the side of Huck helping Jim to freedom. Huck's access to those reasons however is wholly via his affective responses to Jim. His efforts to respond to reasons via deliberation about what he ought to do, lead him to a moral conclusion which is utterly at odds with the affective responses which ultimately determine his actions and which in practical terms might be argued to constitute his true though unarticulated moral stance. In such cases, Jones thinks we confront the possibility that it is more rational for us to act against our deliberative judgments. Our deliberative judgments may thus not have any special normative authority.

And indeed it seems that some people give normative authority to their gut feelings or intuitions over the deliverances of reason. If it feels wrong (or right) they take this as sufficient justification for judgment and action. Perhaps they are correct to do so. Perhaps 'real' or authoritative moral judgments—judgments that both best represent the agent's moral stance and best latch onto the reasons that do in fact justify them—are underpinned by automatic processes. Jonathan Haidt's influential social intuitionist model of moral judgment suggests that this is in fact the case.

5 Haidt's Social Intuitionist Model of Moral Judgment

Jonathan Haidt's social intuitionist model or SIM (e.g., Haidt 2001, 2007; Haidt and Bjorkland 2007a), draws strongly on the dual process models of social cognitive psychology. He has proposed that moral judgments are largely based on 'gut feelings' or 'moral intuitions' which he describes as:

> the sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like-dislike, good-bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion (modified from Haidt 2001 p. 818). (Haidt and Bjorkland 2007a, p. 188).

Thus moral intuitions are a subcategory of automatic attitudes. Haidt acknowledges the possibility of moral reasoning, but a feature of the SIM is the limited role ascribed to such reasoning in moral judgment:

> We often engage in conscious verbal reasoning too, but this controlled process can occur only after the first automatic process has run, and it is often influenced by the

> initial moral intuition. Moral reasoning, when it occurs, is usually a post-hoc process in which we search for evidence to support our initial intuitive reaction. (Haidt 2007, p. 998).

For Haidt, automatic moral intuitions are authoritative in determining moral judgment. Haidt (2001, see pp. 819–823) presents three lines of evidence to contest the causal importance of reason in moral judgment. First, he outlines dual process models of social attitudes and persuasion, emphasizing conditions under which people rely on automatic evaluations, automatically activated stereotypes, and intuitive (or 'heuristic' or 'peripheral') processing in evaluating arguments. Second, he discusses the number of ways in which reasoning has been shown to be biased and motivated by desires other than that of penetrating the truth. His third line of argument draws on evidence of our facility in constructing post hoc justifications for our behaviour when the true causes of it are unknown to us.

For example Haidt claims that the primacy of moral intuitions over moral reasoning is revealed in cases of 'moral dumbfounding' in which people are unable to justify the moral judgments they have made (Haidt and Hersh, 2001). Yet because the negative 'gut feeling' about the act remains (for example, that consensual incest is wrong, even if it has no negative consequences), they continue to insist that the act is morally wrong even though they are unable to rationally justify that position.[2] Wheatley and Haidt (2005) also demonstrate that moral transgressions are judged more harshly when exposure to them is accompanied by a hypnotically induced flash of disgust. They hypothesised that this disgust "would be interpreted by participants as a kind of information, specifically, as an intuition that the action in question was morally wrong." (p. 780) and conclude that this is an illustration of Hume's (1739/1969, p. 462) assertion that "reason is … the slave of the passions, and can pretend to no other office than to serve and obey them" (cited on p.783).

Although he acknowledges the possibility of private moral reasoning, Haidt argues that in reality this is a rare occurrence. Instead, revision of moral judgment occurs primarily through social interaction, in which other people's views evoke new moral intuitions in us. Thus according to the SIM, in most cases a person's moral judgment is based upon his or her currently evoked moral intuition, and controlled reasoning processes exert their effect (if any) *prior* to the triggering of the intuition on which the moral judgment is based.

Furthermore Haidt appears to ascribe normative authority to our moral intuitions, arguing that:

> … the roots of human intelligence, rationality, and ethical sophistication should not be sought in our ability to search for and evaluate evidence in an open and unbiased way … we should instead look for the roots of human intelligence, rationality, and virtue in what the mind does best: perception, intuition, and other mental operations that are quick, effortless, and generally quite accurate...(Haidt 2001, pp. 821–822)

In terms of pinpointing what we mean by the 'real' moral judgment, Haidt's position would appear to be, then, that to know a person's (current) moral intuition about a person or situation is to know virtually all that we need to know about their moral stance at that time. To ask them to reason about their position is merely to assess their facility either at

[2] We set aside a number of concerns about the dumbfounding experiments and Haidt's interpretation of the results here but see Kasachkoff and Saltzstein (2008) and Kennett (forthcoming).

confabulating reasons to support their intuitions, or their ability to trigger new moral intuitions in others. As Haidt puts it:

> Moral reasoning is often like the press secretary for a secretive administration—constantly generating the most persuasive arguments it can muster for policies whose true origins and goals are unknown … (Haidt 2007, p.1000).

Similarly, by characterising intuitive processes as "generally quite accurate" we can take him to mean that adding (one's own) moral reasoning to the moral judgment making process does not buy any great improvements in the quality of those judgments. So if, as suggested by the SIM, private moral reasoning processes are optional and post hoc elaborations on moral intuitions, it makes little sense to imbue our controlled judgments (that we mistakenly suppose to constitute reason responding) with any normative authority.

The implications of the SIM for meta-ethics seem straightforward. Simple sentimentalism (e.g., Prinz 2006) and emotivist accounts of moral judgment which see the twin function of moral judgment as the expression of sentiments and the influencing of others (e.g., Ayer 1936; Stevenson 1937) are in line with the SIM.[3] Some naturalist accounts of morality will also survive but rationalist accounts, as well as sophisticated sentimentalist views of moral judgment (e.g., Rachels 1993) which require reflective endorsement and shaping of the sentiments, are claimed to lack empirical support. It might be argued that these latter accounts do not require empirical support since they make conceptual, rather than descriptive claims about moral judgment, but we believe that rationalists and sophisticated sentimentalists should be concerned if it turned out that moral reasoning as they conceive of it is epiphenomenal.

Haidt's view of the impotence of controlled reflective processes in the formation and revision of moral judgment has far reaching implications, not just for rationalist and sophisticated sentimentalist accounts of moral *judgment,* but for the conception of *agency* presupposed by those accounts. In the next section we will lay out in more detail the normative or reflective conception of agency and evaluative judgment which is seemingly called into question by Haidt's data. We suggest that the controlled reflective processes integral to this conception are indeed necessary to moral agency and that there are no 'real' moral judgments in the absence of this capacity. Then in the final section of the paper we will suggest that this conception of ourselves and of our capacities is consistent with a broader data set.

6 Moral Judgments and Moral Agency

> As a reflective agent, I cannot view myself as merely a system—however well functioning—of sub-systems, that passively register and respond to environmental stimuli much as a thermostat registers and responds to changes in temperature.... To think of myself in this way is not to think of myself as an agent at all. It is to give up thinking of myself as rationally guiding my actions via reasons. (Jones 2003, pp. 188–189.)

[3] Haidt might not agree that his theory is emotivist but we think that his interpretation of public moral 'reasoning' as a process by which exposure to the responses of others may evoke new intuitions in us as clearly consistent with the emotivist claim that a function of moral judgment is to influence the feelings of others. Haidt has argued against charges of emotivism, noting that the moral intuitions on which moral judgments are thought to be based are not purely affective (Haidt and Bjorkland 2007b). However, our understanding of the SIM is that the moral judgment is based largely on the affective, evaluative component of the intuition.

An implication of Haidt's work is that our first personal experience of ourselves as reason responders is illusory. Haidt's research focuses on the cognitive processes which produce moral judgment and he argues that, overwhelmingly, automatic processes *determine* an individual's moral judgment. But if this is so then such judgments arguably fail to satisfy our *concept* of a moral judgment because our sense of the agency of the individual who produces the judgment is undermined. It will thus also fail to account for the *normativity* of moral judgment since such judgments can be normative only for reason responders. Haidt's data might mean that our moral concepts require drastic revision and that we need to take a much more modest deflationary view of our agency.

We do not think our moral concepts would survive such a drastic rewriting to bring them into line with the science. Our claim in this section is that genuine moral judgments must be made by moral agents and that moral agents must, as a matter of conceptual necessity, be reason responders and not merely reason trackers. If it should turn out that human agency is of the minimal kind suggested by Haidt then our moral concepts will lack application, and moral discourse and practice will be systematically in error since they are irreducibly predicated on the assumption that we are reason responders.

A normative or reflective conception of agency underpins both cognitivist and sophisticated non-cognitivist accounts of evaluative judgment. James Rachels, Harry Frankfurt, and Gary Watson each emphasise the intuitive distinction between mere desires or unmediated affective responses, and evaluations—a distinction that seems difficult to account for in terms of automatic processes alone. For Frankfurt and Watson the distinction between what we *merely* want and what we *really* want relies upon reflective self-awareness. It relies among other things on a consideration of what, if anything, our first order desires and affective responses give us reason to do. This is what can transform automatic or intuitive responses into judgments with normative authority. According to Frankfurt (1971, p.7) the capacity for reflective self evaluation '...is manifested in the formation of second order desires' in which the agent endorses certain of their first order desires and withdraws themselves from others. For Watson (1982, p.105), a person's values are 'that set of considerations which he—in a cool and non-deceptive moment—articulates as definitive of the good, fulfilling and defensible life'. This set of considerations frames or regulates 'judgments of the form: the thing for me to do is *a*...'

There are two things worth noting about this conception of value and normative judgment. The first is that on both these accounts one's values are in many ways informed by the deliverances of the automatic system. Reflection on the good and defensible life engages meta-cognitive processes with input from a wide range of sources including ones desires, emotions, moods, personal history, and principles. Second, it is consistent with this view that the particular value judgments we make in line with our reflective views could become fast and habitual over time, as noted by Saltzstein and Kasachkoff (2004). Learning to drive is initially cognitively demanding and effortful but as we all know once the skills are learned driving becomes automatic. It should be no surprise to learn that many of our ordinary moral judgments are like this; indeed on an Aristotelian view that is the point of moral education. It does not belie the central role of reflection in moral agency.[4]

[4] Hierarchical views like Frankfurt's have been criticised on a number of grounds. It is not immediately clear, for example, that second order desires need be a product of reflection—they might just happen to us—or that reflection will issue in second order, rather than first order desires. Frankfurt however clearly takes reflection be essential to personhood. He says " it is only in virtue of his rational capacities that a person is capable of becoming critically aware of his own will" (pp. 11–12). It is critical reflection on his desires that constitutes him as a person rather than a wanton. If this is right then a mere hierarchy of cognitive states will not do the trick. Controlled processes which take as their subject the question of where the agent herself should stand on a particular issue will be required.

For both Watson and Frankfurt reflective self-evaluation importantly constitutes us as persons rather than wantons. They share with philosophers including Jones, Velleman (2000), and Korsgaard (2002) a normative conception of agency according to which qua agents, we have a grounding commitment to guiding ourselves in accordance with our conception of our reasons.[5] If we lack such a commitment or such a conception, as wantons do, then we fail to meet the conditions of moral agency. To see that reasons responsiveness is indeed a necessary condition of moral agency, let us consider in turn cases of reasons tracking without reasons responsiveness and cases of reasons responsiveness where reason tracking is significantly impaired.

A bird, a reptile, or a dog, may respond effectively and efficiently to cues in their environments that constitute good reasons for their behaviour given their natural goals of reproduction and survival. Plainly however birds, reptiles and dogs are not moral agents. It might be thought that these examples are unfair to those who hold that moral judgments are the product of automatic affective systems, because such creatures lack other forms of responsiveness that many see as essential to morality; for example, sympathy or empathy, which do not depend upon reflective processes. It might be argued that they do not, therefore, demonstrate the importance of reasons responsiveness to moral judgment and agency. But small children and dogs and some non-human primates do have the capacity for empathy and yet they are not counted as moral agents. We suggest this is precisely because, and insofar as, they lack the developed rational capacities essential to such agency including the capacity to deliberate upon and regulate their automatic affective responses.

Reasons responsiveness is central to morality in part because without it we cannot tell a satisfying story that makes sense of our practices of holding responsible. We systematically exclude individuals and categories of creatures from moral responsibility when we judge them to lack the capacity to respond to reasons *as reasons* and to guide their behaviour accordingly. That we exclude animals and small children on these grounds is not controversial. Interestingly it is much more controversial as to whether we should exclude those whose automatic moral responses are defective, such as psychopaths who radically lack emotional empathy, from responsibility. When we do so exclude them it is arguably because we think that they are also deficient in other capacities that are essential to moral agency. Psychopaths don't appear to understand that moral claims are reasons claims and so it can be argued that they are not capable of genuine moral judgment.[6] Psychopaths display gross deficiencies in both reason tracking and reason responding, at least with respect to the moral domain. In such cases it makes no sense to try to identify the individual's moral stance or to hold them accountable for their moral failures.[7]

The more compelling cases for our purposes are cases of reasons responsiveness but impaired reason tracking. Here we are thinking of certain high functioning autistic

[5] Jones (2003, p.194) argues persuasively that "the commitment to rational guidance is the commitment to the on-going cultivation and exercise of whatever abilities it is that enable the agent to have and display the capacities that are characteristic of reason-responders." She thinks that these capacities will include various kinds of emotional responsiveness. In this respect we interpret her as providing an account of ideal rational agency.

[6] Psychopaths are also deficient in self-regulatory capacities but unlike Haidt we think these capacities require the kind of reflective self-awareness that is central to rationalist and sophisticated sentimentalist accounts.

[7] For more detailed arguments supporting this general conclusion see Fine and Kennett (2004), Kennett (2002, 2006) and Kennett and Matthews (2007).

individuals who suffer significant impairments of cognitive empathy[8] (that is, perspective taking) and social awareness but who nonetheless possess moral concepts and appear eminently capable of deliberating about what they should do and acting accordingly. They are concerned directly with what they have reason to do even if their access to those reasons via automatic and intuitive processes is significantly impaired. As one autistic man, Jim Sinclair (Sinclair, 1992, p.300) says:

> Even if I can tell what the cues mean, I may not know what to do about them. The first time I ever realized someone needed to be touched was during an encounter with a grief-stricken, hysterically sobbing person who was in no condition to respond to my questions about what I should do to help. I could certainly tell he was upset. I could even figure out that there was something I could do that would be better than nothing. But I didn't know what that something was.[9]

Sinclair's realisation that he should do something—that the other's distress provided a reason for action—and his eventual practical conclusion, is clearly not wholly or primarily dependent on the deliverances of the automatic reason tracking systems, but rather on the application of an explicit concern to do the right thing, whatever that should turn out to be.[10] Sinclair was concerned directly with what he had reason to do in the circumstances and he attempted to arrive at his reasons through explicit reflection. As such he surely counts as a moral agent albeit a somewhat clumsy one.

A defender of the view that judgments produced by automatic processes count as genuine moral judgments might be willing to concede that animals and small children are not moral agents but make the following objection. Maybe only individuals who are capable of a certain kind of reflection can count as moral agents, and maybe one must be such an agent before one's moral utterances get to count as moral judgments. Nevertheless it doesn't follow that moral judgments must be reasons-responsive. They could still emerge directly from automatic reasons tracking mechanisms. After all automatic attitudes may be explicitly reported as the agent's moral opinion and even when they are disavowed, as in the case of Huck Finn, we might judge them to be more expressive of the agent's 'real' values.

We think that the normativity of moral judgment is most plausibly cashed out in terms of reflective endorsement and regulation so our claim is that in cases of conflict the agent's considered view deserves the title of the 'real' or authoritative moral judgment even when

[8] Their impairments in empathy differ from those of the psychopath. Blair (1996) has shown that autistic children are somewhat sensitive to others' distress (emotional empathy) even though they have great difficulty with perspective taking (cognitive empathy). In addition, recent research with adult populations with Asperger's syndrome finds that although they are impaired in cognitive empathy, they score similarly to neurotypical controls on measures of emotional empathy (Dziobek et al. 2008; Rogers et al. 2007).

[9] This example also cited in Kennett (2002).

[10] We are not suggesting here that intuition plays no role whatsoever in the process of moral reasoning in the case of autistic individuals. Perhaps Sinclair's initial realisation that he should do something was in the nature of an intuition. But in this case it is plain that there is a reliance on explicit controlled processes in order to resolve a puzzle and reach a practical conclusion about what ought to be done. Moreover the process of reflection is reflexive, it does involve taking oneself to be responsive to reasons and Sinclair sees his efforts in this light.

we think their automatic responses better track the reasons that there are.[11] But as we have suggested above this does not require that our each and every moral response must be the product or subject of explicit effortful deliberation in order to so count. Plainly we do not have the cognitive resources to devote to such a task. Nevertheless, we think our reflective views of 'the good, defensible and fulfilling life' are capable of regulating our moral responses, directly or indirectly, so that spontaneous intuitions that do not accord with our reflectively endorsed evaluative framework may be modified, overridden or set aside. In the next section we return to a consideration of the empirical literature to support this view.

7 The Interplay Between Automatic Moral Intuitions and Controlled Processes

Haidt has made a valuable contribution to the moral psychology field by forcing a consideration of the role of automaticity in moral judgment. Research in social cognitive psychology has established the ubiquity and importance of such processes in judgment and action (e.g., Bargh and Chartrand 1999; Bargh and Ferguson 2000). It is certainly very plausible that much of the processing underpinning our everyday moral judgment and behaviour is done automatically and with little effortful cognition. Nonetheless, a close examination of the social cognitive psychology literature suggests that Haidt's (2001) claim that "moral reasoning is rarely the direct cause of moral judgment" (p. 815) is to overstate the primacy of automatic processes in social judgment, and to underplay the contribution of controlled processes. The SIM has been critiqued previously for the restricted role it allows controlled processes in moral judgment (e.g., Saltzstein and Kasachkoff 2004; Fine 2006). Here, we wish to argue that the strategic over-riding of moral intuitions based on automatically activated stereotypes or evaluations can occur in two ways. Govorun and Payne (2006, p.130) provide a useful terminology that distinguishes between "after-the-fact" and "up-front" mental control. 'After-the-fact' correction "focuses on mental undoing, whereas ['up-front' control] focuses on how thinking is done to begin with." (p.131). In line with this, we suggest that moral intuitions can be controlled in both these ways: by a slow, intentional, deliberative and effortful 'after-the-fact' correction; or with 'up-front' preconscious control of those activations, deployed in the absence of awareness or conscious volition.

7.1 After-the-fact Correction

Research suggests that people can and do engage in 'after the fact' adjustment of their initial social or moral judgments. For example while a person's incidental mood or emotional state can 'contaminate' her moral judgments (e.g., Forgas and Moylan 1987; Wheatley and Haidt 2005)—something no rationalist would deny—this bias can be corrected more or less accurately (see Wilson and Brekke 1994), when the individual's attention is drawn to their mood as a possible source of bias (e.g., Schwartz and Clore

[11] We believe this is consistent with the view held by Jones and also argued for in Bennett (1974) and Kennett (2001) that conflict between sympathy and considered judgments should sometimes prompt a reconsideration of the judgments themselves since we are not omniscient about value and our sympathies are often (though not infallibly) reason tracking. Such reconsiderations are part and parcel of ideal reasons responsiveness. Huck Finn clearly fell short of this ideal in making his moral judgments, nevertheless it seems reasonable to assume that his 'real' moral judgments were the judgments for which he could provide an explicit (though no doubt flawed) justification and with respect to which he regarded himself as weak.

1983), or she is motivated to be accurate (e.g., Lerner et al., 1998; although see Payne et al. 2005a for evidence that bias correction cannot always be successfully achieved).[12]

In the light of such findings, we can revisit the conclusions drawn by Wheatley and Haidt (2005) following their demonstration that moral transgressions are judged more harshly when exposure to them is accompanied by a hypnotically induced flash of disgust. The authors included in their analyses only participants who were successfully hypnotised to be 'amnesic' as to the source of their feeling of disgust. The authors' justification for this is that the participants' "lack of conscious memory for the true cause of their disgust affords the most stringent test of whether disgust informs moral judgment." (p. 781). In fact, their exclusion of participants who were aware that their feelings of disgust were morally irrelevant leaves open the possibility that these participants would discount its influence—in other words, rather than reason being the slave of the passions, reason would ignore the passions.

Other research demonstrates that when people become aware that they have a tendency to make certain types of judgments in a biased way (for example, due to the activation of negative stereotypes about a racial group) then, if they are motivated to be unprejudiced, they will effortfully over-ride their intuitively-based judgments, so long as they have the cognitive resources to do so (for examples, discussed with respect to the SIM, see Fine 2006). For example, volunteers interviewed by Monteith and colleagues reported going through a spontaneous process of consciously and effortfully over-riding their intuitive judgments:

> This summer my girlfriend and I were looking at the horses downtown, and ... this Black guy started walking toward us. Of course ... I immediately thought, here comes some homeless guy—he's going to ask us for money. But from my past experience [reflecting on the automatic assumptions he made about Black people] ... I had to stop and think to myself, "Maybe he's not homeless, maybe he's not going to ask me for money. He might not say anything to me." I stopped and I thought about the past experience and it made me change my decision to something I probably wouldn't have made. (from Monteith et al. 2002; pp. 1046–7).

A slightly different example comes from ratings of the funniness of racist and sexist jokes. Ratings of such jokes made under time pressure or distraction can be assumed to reflect relatively automatic 'acceptance' of racist or sexist attitudes. That such jokes are evaluated more negatively without time constraints or distraction suggests that effortful, controlled processes may be employed in an after-the-fact 'spoiling' of the joke (Eyssel and Bohner 2007; Monteith and Voils 1998).

Observational data provided by Haidt and Hersh (2001) suggests that even strongly affectively charged moral intuitions can be over-ridden when the individual considers them to be inappropriate. In their study of the moral judgments of conservative and liberal students, Haidt and Hersh (2001) found that while some students used their affective responses as justifications for moral condemnation (*it's disgusting so it must be wrong*), they also found liberals sometimes explicitly discounted their strong emotional reactions to

[12] While it might be argued that such correction requires external promptings rarely found outside the laboratory we would disagree. We do often draw each other's attention to mood as influencing our judgments and we often correct for mood unprompted. It is not uncommon for people to explain judgments and actions which they reflectively disavow in terms of the contaminating effects of emotion etc. "I was too angry (tired, upset, down, excited). I didn't mean it".

sexual acts. One student, for example, described how he would "quickly discard" the "warning bell" (p. 211) that came up when he considered the ethics of gay anal sex.

More formally, Gabriel et al. (2007) assessed automatic (using the IAT), cognitive and affective attitudes towards homosexuals in a sample of students. The cognitive attitude scale tapped beliefs about homosexuality (for example, that female homosexuality is a sickness, or that gay men should not be allowed to work with children). The affective attitude scale, by contrast, tapped emotional responses to homosexuality, including experiences that Haidt might term moral disgust or other moral emotions. In this questionnaire, the participant indicates how much discomfort he would feel if, for example, he learned that his son's teacher was gay, or if he saw two lesbians kissing. For participants who indicated only low internal motivation to control prejudiced responses, automatic attitudes correlated with both cognitive and affective attitudes, as the SIM would predict. But for participants with a high internal motivation to control prejudice, increasingly negative automatic attitudes towards homosexuality manifested neither in more negative affective attitudes nor more negative cognitive attitudes.

Participants were then given the opportunity to sign a petition to maintain funds and to donate money to a (real) local gay organization facing the prospect of having public funding discontinued. Contrary (we assume) to the predictions of the SIM, cognitive attitudes, but not affective attitudes, predicted helping behaviour:[13] "Individual differences on the cognitive attitude scale (i.e. equal rights for homosexuals) predicted support for a political plea of homosexuals, whereas affective attitudes (i.e. the affective reaction to imagined displays of homosexual behaviour) did not covary with support." (p. 373). While of course the affective and automatic attitudes measured in this study related to homosexuality *per se*, rather than to the particular gay organization, it seems problematic for the SIM that these affectively-charged attitudes regarding homosexuality did not appear to contribute to the moral decision-making process of deciding what level of support to offer to a gay organization. It suggests that participants were aware that their affective responses to homosexuality were morally irrelevant to the dilemma at hand. To put it in terms of the distinction between reason tracking and reason responding the participants did not accept that their emotions were reason tracking.

In line with evidence such as this, Haidt has recently modified his position by suggesting that the:

> "tight connection between flashes of intuition and conscious moral judgments … is not inevitable: Often a person has a flash of negative feeling, for example, towards stigmatized groups … yet because of one's other values, one resists or blocks the normal tendency to progress from intuition to consciously endorsed judgment." (Haidt and Bjorkland 2007a, p.818)

This resistance or blocking appears to involve processes that fall within Haidt's definition of moral reasoning; that is, "conscious mental activity that consists of transforming given information about people in order to reach a moral judgment" (Haidt 2001; p.818). Thus this amendment of the SIM would appear to be of some significance, acknowledging as it does that moral reasoning can disrupt the "tight connection" between intuition and judgment, the causal pathway that is a central tenet of the SIM.

[13] The study found complex and somewhat unexpected interactions between private *versus* public setting, motivation to control prejudiced responses, and implicit attitudes.

7.2 Ignoring Moral Intuitions

Thus while affect and judgment likely correlate very closely for many moral situations, the examples presented here suggest that in cases in which an individual's reflectively endorsed values lead her to regard her intuitive reactions as inappropriate, the intuitions are discounted rather than rationalized. The SIM does offer scope for cases in which a judgment based on moral intuition is over-ridden by moral reason, but these instances are "hypothesized to be rare, occurring primarily in cases in which the initial intuition is weak and processing capacity is high" (Haidt 2001, p. 819). This is an empirical question, but it seems plausible to us that moral judgments regarding stigmatized groups and sexual behavior, for example, may be ones in which initial intuitive reactions are not necessarily weak, and to which some individuals may be motivated to devote controlled processing capacity. Nonetheless, Haidt (2001) is right to emphasize the relative rarity of conscious, controlled processing in everyday life. So-called correction models generally propose that, in order to correct for unwanted biases, people must be aware of a potential source of bias, be motivated to correct for it, and have the controlled processing capacity to do so (see Wilson and Brekke 1994; Wegener and Petty 1995). Thus in the next section we turn to evidence that such correction or compensatory processes may take place even when these criteria aren't met.

7.3 Unconscious Volition

A growing body of evidence suggests that automatic processes can be 'preconsciously' controlled, in accord with consciously held goals. Glaser and Kihlstrom (2005, p. 171) refer to this as "unconscious volition" or "'compensatory automaticity'; strategic yet nonconscious compensations for unintended thoughts, feelings, or behaviors." A comprehensive review of these data is beyond the scope of this article, but we provide here two examples of the preconscious control of automatic processes by consciously held goals. First, we discuss here findings showing that control of the influence of unendorsed automatic attitudes need not always take place 'after-the-fact' but can also occur 'up-front', or preconsciously.

Keith Payne and colleagues have, in recent years, explored the role of the self-regulatory processes that enable people to limit the influence of automatically activated information. Payne uses a paradigm in which a sequence of guns and tools appear on the computer screen. The volunteer's task is to categorize the object with a key press, as quickly as possible. Before each trial a male face appears that is either black or white. Both automatic and controlled processes contribute to responses on the weapon-identification task, and the experimental design enables separation of automatic and controlled contributions to performance. For most participants, an automatic 'black man-danger' association will, in the absence of control, result in 'false positive' errors on trials in which a black face is followed by a harmless tool. However, Payne (2005) has found that people with a high capacity for self-regulatory control, and stronger motivation to control prejudice, show less behavioural expression of automatically activated associations; that is, they show fewer 'false gun' errors after seeing black faces. Importantly, this is not because negative automatic attitudes are any less strongly activated in these individuals. Rather, they exert greater cognitive control which, according to Payne (2005, p. 491), enables people to "constrain their processing to the relevant information rather than being driven by irrelevant but activated information."

How might this occur? Barrett et al. (2004, p. 564) speculated that:

> "... controlled processing may not be merely reversing the effects of automatic processing, but it may also prevent (or allow) the expression of attention on

> representations that were activated in a stimulus-driven way. As long as one has a processing goal (like an egalitarian goal to prevent stereotyping, for example), as well as the WMC [working memory capacity] to deploy goal-directed attentional effects, that processing goal can be enacted. ... For example, it may be that a property of the person (e.g., skin pigmentation) automatically activates both a stereotype and a goal to be egalitarian, and with sufficient WMC resources, the activation level of the stereotype can be suppressed before it influences subsequent processing, thereby allowing egalitarian outcomes with perceived ease."

Recent evidence seems consistent with this suggestion. Amodio et al. (2008), using Payne's weapon-identification task, have collected electroencephalographic (EEG) data suggesting that the conflict between an activated stereotype and the goal to avoid stereotyping is detected preconsciously. Amodio et al. found that a burst of "conflict monitoring" neural activity takes place about 100ms before volunteers successfully avoid making a mistaken call of "gun" following a black face. Furthermore, people who were motivated to control prejudice because of their own internal moral standards showed larger waves of conflict-monitoring EEG activity, and fewer false positive errors, than volunteers whose prejudice-control motivations stemmed more from concern about how they came across to others. Amodio et al. (2008, p. 72) suggest that "[t]hese findings show that effective response control may be deployed without a person's awareness that a race-biased response was averted."

It could be objected that these findings are compatible with a moral intuitionist position: that in such cases individuals are simply acting in accordance with an activated egalitarian intuition, rather than a conflicting negative intuition. However, we would argue against such an interpretation. First, the research presented by Payne and colleagues suggests that, although the deployment of 'up-front' control is neither consciously willed nor accessible, it is dependent on the availability of controlled processing resources. Thus, the weapon-identification judgments of individuals with chronically or temporarily lower executive control capacities are more strongly determined by their automatic attitudes (Payne 2005; Govorun and Payne 2006). Although Haidt's definition of moral reasoning includes a conscious component, Barrett et al. (2004) have argued that conscious experience is not diagnostic of controlled processing and that "people can engage in controlled processing even when they do not experience themselves as doing so." (p. 563). Glaser and Kilhstrom (2005) even suggest that "the unconscious is indeed capable of holding such meta-cognitive processing goals (e.g., accuracy) which it will pursue through self-monitoring, and that it will, under some conditions, compensate for anticipated threats to the attainment of those goals." (p. 190). There seems a case, then, for arguing that control of the effects of automatic activations does not represent a competition between two competing stimulus-driven automatic processes. Rather, it should be characterised as a preconscious, but controlled, over-riding of a moral intuition that threatens reflectively endorsed goals.

Thus, we would argue that this preconscious control can at least sometimes be the causal consequence of prior or current reflective endorsement of an egalitarian goal. Bargh and colleagues (e.g., Bargh and Ferguson 2000; Bargh et al. 2001) have argued that automatic goal pursuit ("the nonconscious activation and operation of goals" Bargh et al. 2001 p. 1014) occurs when a goal has been consciously selected in a particular situation. Over time, the goal representation becomes automatically associated with features of the situation and thus comes to be activated automatically in such situations. Bargh et al. (2001) argue that "on the basis of the assumption that goals become automated through their repeated selection in a given situation, such automatic goals should generally be in line with the

individual's valued, aspired-to life goals and purposes." (p. 1015). Although to our knowledge this assumption has not been tested against longitudinal data, it is consistent with research suggesting that the rapid, non-volitional over-riding of automatically activated associations is associated with consciously reported concerns to do so (e.g., Lepore and Brown 1997; Moskowitz et al. 1999; Kiefer and Sanchez 2007; Maddux et al. 2005).

It is possible that the automatic pursuit of a goal then leads to a conscious endorsement of that goal (i.e., that the direction of causality is contrary to that which we propose). However, Bargh et al. (2001) found that it was only in participants who had consciously activated the goal to cooperate that self-reports of intention to cooperate were related to actual cooperation during the experiment. Bargh et al. (2001, p. 1024) therefore suggested that:

> The construction of one's own degree of intentionality based on the amount of behavior exerted, therefore, would seem possible only when one knew all along what one was attempting to do ... For instance, the inference that one is a cooperative person may be more likely following conscious pursuit of the goal of cooperation than when the identical outcomes are produced on the basis of unconsciously operating cooperation goals.

Remarkably, very recent work with the weapon-identification task suggests that even what is automatically activated by a stimulus may be changed by prior conscious intentions. Stewart and Payne (2008) asked one group of volunteers to make the counter-stereotypical commitment that "Whenever I see a Black face on the screen, I will think the word, 'safe'." These volunteers made fewer false positive gun responses following a black face, compared with groups who made non counter-stereotypical commitments. Further analysis of the data revealed that, unlike the findings of Payne's previous work, this was not due to increased cognitive control. Rather, it arose from a reduction in biased automatic activations. Stewart and Payne (2008, p. 1344) note that this conscious commitment strategy provides "one way in which conscious strategies can be used to overcome automatic stereotyping even when all the right circumstances (e.g., opportunity for controlled thinking, awareness, etc.) are not in place."

7.4 Summary

Our concern with the SIM, then, is that it does not appear to fully present the complexity of the relationship between automatic and controlled processes in moral judgment. While we concur with Haidt that many of our moral judgments will be based on intuitive responses, we argue that in certain situations, in certain individuals, those very same moral judgments will *not* be based on automatic evaluations of events or people. We may effortfully over-ride judgments based on moral intuitions, discount moral emotions that we believe to be irrelevant or misplaced, and exert preconscious control such that the activated associations of our moral intuitions do not interfere with the processing of more relevant information. We have argued that this is best conceptualised as the preconscious influence of prior moral reasoning on the intuitive judgment link. In the light of these cases, we would argue that the moral judgment made intuitively in a distracted or tired moment does not deserve normative authority if it diverges from the judgment the agent would have made in a more reflective or cognitively resourced situation. The real moral judgment is ultimately the one that the agent can reflectively endorse.

8 Conclusion

We believe that the empirical challenge which Haidt's work presents to reflective conceptions of moral agency and moral judgment can be countered, and indeed that overall the evidence supports the view that there is a complex interplay between automatic and controlled processes in producing moral judgment which is best understood within a broadly reflectivist conception of agency. It is unlikely that a mature agent's settled moral judgments, many of which will have been revisited and modified over a prolonged period, could be shown to be the product solely of one or other system. If we are right, sophisticated sentimentalist and rationalist meta-ethical views answer better to all the available evidence (as well as to our moral concepts) than the simple sentimentalist and emotivist views Haidt's account initially seems to favour. Insofar as the data suggests a constant interplay between automatic and controlled processes in producing and modifying moral judgment it should also prompt a re-examination of the traditional opposition between reason and emotion and the effort to locate moral judgment wholly on one or other side of the divide. The reflective self-awareness that makes us agents who are capable of moral judgment and of the regulation of our moral responses relies on input from both processes and perhaps on the exercise of additional cognitive resources not encompassed by dual processing models. But that is a subject for another time.

Acknowledgments The authors would like to thank Neil Levy, Edward Hare, and audiences at the Australasian Association of Philosophy Conference 2007, Monash University, Radboud University and the University of Oxford for helpful comments on earlier versions of this paper. The authors acknowledge the support of the Australian Research Council for this project. We thank audiences at the Australasian Association of Philosophy Conference 2007 and at Monash University, Radboud University, and the University of Oxford for stimulating and helpful discussions. We owe particular thanks to Neil Levy and an anonymous referee for forcing us to clarify our argument at a number of points.

References

Amodio DM, Devine PG, Harman E (2008) Individual differences in the regulation of intergroup bias: the role of conflict monitoring and neural signals for control. J Pers Soc Psychol 94:60–74

Ayer AJ (1936) Language, truth, and logic (2nd edn). Gollancz, London (1946)

Bargh J, Chartrand TL (1999) The unbearable automaticity of being. Am Psychol 54:462–479

Bargh J, Ferguson MJ (2000) Beyond behaviourism: on the automaticity of higher mental processes. Psychol Bull 126:925–945

Bargh JA, Williams EL (2006) The automaticity of social life. Curr Dir Psychol Sci 15:1–4

Bargh JA, Gollwitzer PM, Lee-Chai A, Barndollar K, Trötschel R (2001) The automated will: nonconscious activation and pursuit of behavioral goals. J Pers Soc Psychol 81:1014–1027

Barrett LF, Tugade MM, Engle RW (2004) Individual differences in working memory capacity and dual-process theories of mind. Psychol Bull 130:553–573

Bennett J (1974) The conscience of Huckleberry Finn. Philosophy 49:123–34

Blair RJR (1996) Brief report: morality in the autistic child. J Autism Dev Disord 26:571–579

Deigh J (1995) Empathy and universalizability. Ethics 105:743–763

Dovidio JF, Kawakami K, Johnson C, Johnson B, Howard A (1997) On the nature of prejudice: automatic and controlled components. J Exp Soc Psychol 33:510–540

Duckworth KL, Bargh JA, Garcia M, Chaiken S (2002) The automatic evaluation of novel stimuli. Psychol Sci 13:513–519

Dziobek I, Rogers K, Fleck S, Bahnemann M, Heekeren HR, Wolf OT, Convit A (2008) Dissociation of cognitive and emotional empathy in adults with asperger syndrome using the multifaceted empathy test (MET). J Autism Dev Disord 38:464–473

Eyssel F, Bohner G (2007) The rating of sexist humor under time pressure as an indicator of spontaneous sexist attitudes. Sex Roles 57:651–660

Fazio RH (1990) Multiple processes by which attitudes guide behaviour: The MODE model as an integrative framework. In: Zanna MP (ed) Advances in experimental social psychology, vol 23. Academic, San Diego, pp 75–109
Fazio RH (2001) On the automatic activation of associated evaluations: an overview. Cogn Emot 15:115–141
Fazio RH, Olson MA (2003) Implicit measures in social cognition research: their meaning and use. Annu Rev Psychol 54:297–327
Fine C (2006) Is the emotional dog wagging its rational tail, or chasing it? Philos Explor 9:83–98
Fine C, Kennett J (2004) Mental impairment, moral understanding and criminal responsibility: psychopathy and the purposes of punishment. Int J Law Psychiatry 27: 425–443
Forgas JP, Moylan SJ (1987) After the movies: the effects of transient mood states on social judgments. Pers Soc Psychol Bull 13:478–489
Frankfurt H (1971) Freedom of the will and the concept of a person. J Philos 68:5–20
Gabriel U, Banse R, Hug F (2007) Predicting private and public helping behaviour by implicit attitudes and the motivation to control prejudiced reactions. Br J Soc Psychol 46:365–382
Gawronski B, Hofmann W, Wilbur CJ (2006) Are "implicit" attitudes unconscious? Conscious Cogn 15:485–499
Glaser J, Kihlstrom JF (2005) Compensatory automaticity: Unconscious volition is not an oxymoron. In: Hassin RR, Uleman JS, Bargh JA (eds) The new unconscious. Oxford University Press, Oxford, pp 171–195
Govorun O, Payne BK (2006) Ego depletion and prejudice: separating automatic and controlled components. Social Cogn 24:111–136
Greenwald AG, McGhee DE, Schwartz JKL (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 74:1464–1480
Greenwald AG, Nosek BA, Banaji MR (2003) Understanding and using the implicit association test I: an improved scoring algorithm. J Pers Soc Psychol 85:197–216
Greenwald AG, Krieger LH (2006) Implicit bias: scientific foundations. Calif Law Rev 94:945–967
Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814–834
Haidt J (2007) The new synthesis in moral psychology. Science 316:998–1002
Haidt J, Bjorkland F (2007a) Social intuitionists answer six questions about moral psychology. In: Sinnott-Armstrong W (ed) Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity. MIT, Boston, pp 181–218
Haidt J, Bjorkland F (2007b) Social intuitionists reason, in conversation. In: Sinnott-Armstrong W (ed) Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity. MIT, Boston, pp 241–254
Haidt J, Hersh MA (2001) Sexual morality: the cultures and emotions of conservatives and liberals. J Appl Soc Psychol 31:191–221
Hofmann W, Gschwendner T, Nosek BA, Schmitt M (2005) What moderates implicit-explicit consistency? Eur Rev Soc Psychol 16:335–390
Hofmann W, Rauch W, Gawronski B (2007). And deplete us not into temptation: automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. J Exp Soc Psychol 43:497–504
Jones K (2003) Emotion, weakness of will and the normative conception of agency. In: Hatzimoysis A (ed) Philosophy and the Emotions. Cambridge University Press, Cambridge, pp 181–200
Kasachkoff T, Saltzstein HD (2008) Reasoning and moral decision making: a critique of the social intuitionist model. EJDS 2:287–302
Kennett J (2001) Agency and responsibility: A common-sense moral psychology. Clarendon, Oxford
Kennett J (2002) Autism, empathy and moral agency. Philos Q 52:340–357
Kennett J (2006) 'Do psychopaths really threaten moral rationalism?'. Philos Explor 9:69–82
Kennett J, Matthews S (2007) Normative agency. In: Atkins K, MacKenzie C (eds) Practical Identity and Narrative Agency. Routledge, New York
Kennett J (forthcoming) Living with one's choices: moral reasoning in vivo and in vitro
Kiefer AK, Sanchez DT (2007) Men's sex-dominance inhibition: do men automatically refrain from sexually dominant behavior? Pers Soc Psychol Bull 33:1617–1631
Korsgaard CM (2002) Self-constitution: agency, identity, and integrity. The Locke lectures available at http://www.people.fas.harvard.edu/~korsgaar/
Kunda Z, Spencer SJ (2003) When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. Psychol Bull 129:522–544
Lepore L, Brown R (1997) Category and stereotype activation: is prejudice inevitable? J Pers Soc Psychol 72:275–287
Lerner JS, Goldberg JH, Tetlock PE (1998) Sober second thought: the effects of accountability, anger, and authoritarianism on attributions of responsibility. Pers Soc Psychol Bull 24:563–574

Maddux WW, Barden J, Brewer MB, Petty RE (2005) Saying no to negativity: the effects of context and motivation to control prejudice on automatic evaluative responses. J Exp Soc Psychol 41:19–35
Monteith MJ, Voils CI (1998) Proneness to prejudiced responses: toward understanding the authenticity of self-reported discrepancies. J Pers Soc Psychol 75:901–916
Monteith MJ, Ashburn-Nardo L, Voils CI, Czopp AM (2002) Putting the brakes on prejudice: on the development and operation of cues for control. J Pers Soc Psychol 83:1029–1050
Moskowitz GB, Gollwitzer PM, Wasel W, Schaal B (1999) Preconscious control of stereotype activation through chronic egalitarian goals. J Pers Soc Psychol 77:167–184
Nosek BA (2007) Implicit-explicit relations. Curr Dir Psychol Sci 16:65–69
Payne BK (2001) Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. J Pers Soc Psychol 81:181–192
Payne K (2005) Conceptualizing control in social cognition: how executive functioning modulates the expression of automatic stereotyping. J Pers Soc Psychol 89:488–503
Payne BK, Cheng CM, Govorun O, Stewart BD (2005a) An inkblot for attitudes: affect misattribution as implicit measurement. J Pers Soc Psychol 89:277–293
Payne B, Jacoby LL, Lambert AJ (2005b) Attitudes as accessibility bias: Dissociating automatic and controlled processes. In: Hassin RR, Uleman JS, Bargh JA (eds) The New Unconscious. Oxford University Press, Oxford, pp 393–420
Prinz J (2006) The emotional basis of moral judgments. Philos Explor 9:29–43
Rachels J (1993) Subjectivism. In: Singer P (ed) A Companion to ethics. Blackwell, Oxford, pp 432–441
Rogers K, Dziobek I, Hassenstab J, Wolf OT, Convit A (2007) Who cares? Revisiting empathy in asperger syndrome. J Autism Dev Disord 37:709–715
Rydell RJ, McConnell AR (2006) Understanding implicit and explicit attitude change: a systems of reasoning analysis. J Pers Soc Psychol 91:995–1008
Saltzstein HD, Kasachkoff T (2004) Haidt's moral intuitionist theory: a psychological and philosophical critique. Rev Gen Psychol 8:273–282
Schwartz N, Clore GL (1983) Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. J Pers Soc Psychol 45:513–523
Sherman JW, Gawronsko B, Gonsalkorale K, Hugenberg K, Allen TJ, Groom CJ (2008) The self-regulation of automatic associations and behavioral impulses. Psychol Rev 115:314–335
Sinclair J (1992) Bridging the gaps: An inside-out view of autism (or, do you know what I don't know?). In: Schopler E, Mesibov GB (eds) High-functioning individuals with autism. Plenum, New York, pp 294–302
Smith M (1987) The Humean theory of motivation. Mind 96:36–61
Smith M (1994) The moral problem. Blackwell, Oxford
Smith ER, DeCoster J (2000) Dual-process models in social and cognitive psychology: conceptual integration and links to underlying memory systems. Pers Soc Psychol Rev 4:108–131
Stevenson CL (1937) The emotive meaning of ethical terms. Mind, 46:14–31
Strack F, Deutsch R (2004) Reflective and impulsive determinants of social behaviour. Pers Soc Psychol Rev 8:220–247
Stewart BD, Payne BK (2008) Bringing automatic stereotyping under control: implementation intentions as efficient means of thought control. Pers Soc Psychol Bull 34:1332–1345
Velleman JD (2000) The possibility of practical reason. Oxford University Press, Oxford
Wallace RJ (1999) Moral cognitivism and motivation. Philos Rev 108(2):161–219
Watson G (1982) Free agency. In: Watson G (ed) Free will, 1st edition. Oxford University Press, Oxford, pp 96–110
Wegener DT, Petty RE (1995) Flexible correction processes in social judgment: the role of naïve theories in corrections for perceived bias. J Pers Soc Psychol 68:36–51
Wheatley T, Haidt J (2005) Hypnotic disgust makes moral judgments more severe. Psychol Sci 16:780–784
Wilson TD, Brekke N (1994) Mental contamination and mental correction: unwanted influences on judgments and evaluations. Psychol Bull 116:117–142
Woods M (1972) Reasons for action and desires. Suppl proc Aristot Soc 46:189–201