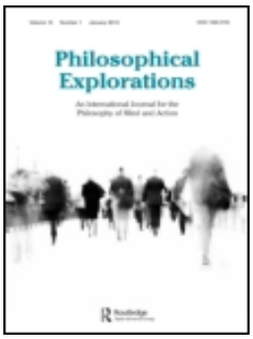

Routledge
Taylor & Francis Group

Philosophical Explorations

ISSN: 1386-9795 (Print) 1741-5918 (Online) Journal homepage: https://www.tandfonline.com/loi/rpex20

Is the emotional dog wagging its rational tail, or chasing it?

Reason in moral judgment

Cordelia Fine

To cite this article: Cordelia Fine (2006) Is the emotional dog wagging its rational tail, or chasing it?, Philosophical Explorations, 9:1, 83-98, DOI: 10.1080/13869790500492680

To link to this article: https://doi.org/10.1080/13869790500492680

Philosophical Explorations, Vol. 9, No. 1, March 2006

Routledge
Taylor & Francis Group

IS THE EMOTIONAL DOG WAGGING ITS RATIONAL TAIL, OR CHASING IT? Reason in moral judgment

Cordelia Fine

According to Haidt's (2001) social intuitionist model (SIM), an individual's moral judgment normally arises from automatic 'moral intuitions'. Private moral reasoning—when it occurs—is biased and post hoc, serving to justify the moral judgment determined by the individual's intuitions. It is argued here, however, that moral reasoning is not inevitably subserviant to moral intuitions in the formation of moral judgments. Social cognitive research shows that moral reasoning may sometimes disrupt the automatic process of judgment formation described by the SIM. Furthermore, it seems that automatic judgments may reflect the 'automatization' of judgment goals based on prior moral reasoning. In line with this role for private moral reasoning in judgment formation, it is argued that moral reasoning can, under the right circumstances, be sufficiently unbiased to effectively challenge an individual's moral beliefs. Thus the social cognitive literature indicates a greater and more direct role for private moral reasoning than the SIM allows.

Overview of the Social Intuitionist Model

The central claim of the social intuitionist model (SIM) (Haidt 2001) is that moral judgments are largely based on 'gut feelings':

> . . .moral judgments are like aesthetic judgments. They are gut feelings or intuitions. . . (Haidt 2002, 54)

These 'moral intuitions' he defines as, 'the sudden appearance in consciousness of a moral judgment, including an affective valence (good–bad, like–dislike), without any conscious awareness of having gone through the steps of searching, weighing evidence, or inferring a conclusion' (Haidt 2001, 818). These intuitions are thus a form of 'automatic process', the characteristics of which are presented in Table 1. Moral intuitions, according to the model, lead directly to moral judgments, which Haidt defines as 'evaluations (good versus bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture' (Haidt 2001, 817). Moral

ISSN 1386-9795 print/1741-5918 online/06/010083-16

DOI: 10.1080/13869790500492680

TABLE 1

Differences between automatic and controlled cognitive processes (adapted with permission from Haidt 2001, 818)

Automatic processes	Controlled processes
Fast and effortless	Slow and effortful
Unintentional, runs automatically	Intentional and controllable
Inaccessible, only results enter awareness	Consciously accessible and viewable
Does not demand attentional resources	Uses attentional resources

reasoning—a conscious, effortful and intentional (or 'controlled') process (see Table 1)—occurs only *after* the judgment has been made, if at all:

> [I]f somebody asks us to explain our judgment we search for reasons why our judgment is correct. Our moral reasoning works like a lawyer seeking evidence, not like a judge seeking truth . . . We make up justifications post hoc, which we present as though they were the causal reasons that led to our initial judgment. (Haidt 2002, 54)

The proposal that judgment precedes reasoning is represented by Links 1 and 2 in Haidt's (2001) model (see Figure 1).

While the emphasis of the SIM is on interactive social processes, as Figure 1 indicates, Haidt (2001) does acknowledge that an individual's private reasoning can sometimes influence their moral judgments. Link 5 represents the scenario in which a person overrides a weak intuition through 'sheer force of logic' (2001, 819). Link 6 represents the situation in which thinking about a moral situation triggers a new intuition that conflicts with the initial judgment. The individual must then decide between the contradictory intuitions, either by choosing the strongest intuition, or by using a reasoned principle or rule to decide between the alternatives. However, Haidt (2001, 2004) hypothesises that these

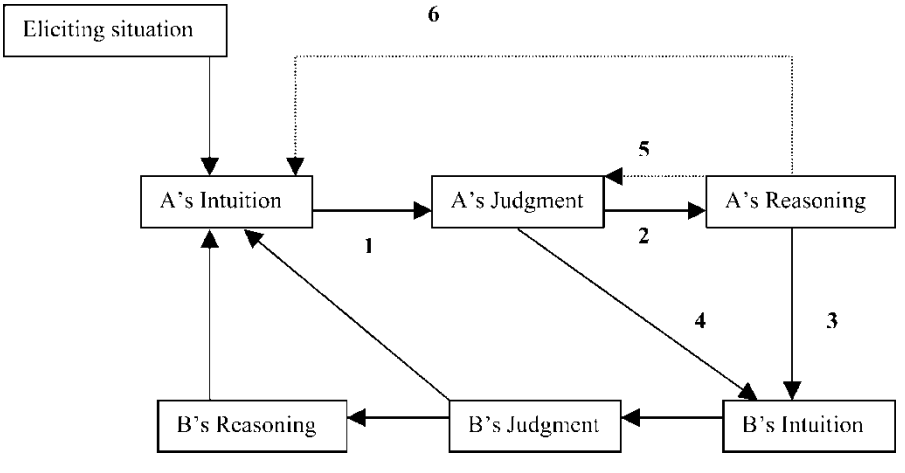

FIGURE 1

The social intuitionist model (SIM). Reprinted with permission from Haidt (2001, 815)

processes usually only occur in unusual circumstances (e.g., during formal moral judgment interviews) or in unusual individuals (e.g., philosophers):

> The core of the model gives moral reasoning a causal role in moral judgment but only when reasoning runs through other people [Link 3]. It is hypothesised that people rarely override their initial intuitive judgments just by reasoning privately to themselves because reasoning is rarely used to question one's own attitudes or beliefs. (Haidt 2001, 819)

Rather, one's own reasoning usually only acts indirectly via the intuitions, judgment and reasoning of other individuals. Haidt also argues that moral reasoning is biased and post hoc:

> moral reasoning is not left free to search for truth but is likely to be hired out like a lawyer by various motives, employed only to seek confirmation of preordained conclusions. (2001, 822)

It is thus a central tenet of the SIM that in everyday moral situations, private reasoning is slave to the intuitions: the emotional dog wags the rational tail.

In contrast to the SIM, I suggest that a closer examination of the processes involved in the formation of social and moral judgments reveals that private moral reasoning may potentially play a more important role in the development of an individual A's moral judgments than is currently allowed by the SIM. First, I present evidence that Link 1—the link from an automatic evaluation to a corresponding judgment—can be disrupted by controlled processes, in accordance with the individual's consciously held personal motivations or values. This suggests, in other words, that reasoning may sometimes be directly involved in the formation of moral judgements. Second, I describe work in stereotyping and other social judgment processes that suggests that automatic social processes can come to be importantly constrained by prior controlled cognitive processes. That is, the automaticity of a moral judgment does not imply a lack of prior input from controlled reasoning processes. Instead, rapid and easy moral judgments of the sort described by Haidt may reflect the 'automatization' of reasoned judgment. Finally I argue that, under the right circumstances, our reasoning can and does question our moral attitudes and beliefs.

Automatic Social Judgment

Haidt's moral intuitions—from which moral judgments are formed—are a form of automatic social judgment. Much of the last few decades of research in social cognitive psychology has focused on what has been called the 'automaticity of being' (Bargh and Chartrand 1999), with many researchers drawing the conclusion that rapid, effortless, automatic processes perform many of the duties of our social cognitive life. Controlled processes, in contrast, are more cognitively 'expensive' and—being slow, effortful and depleting of our limited attentional resources—are relatively rarely used. In line with this conclusion, Haidt (2001, 820) claims that 'the intuitive process is the default process, handling everyday moral judgments in a rapid, easy, and holistic way'. For example, he argues that 'people categorise other people instantly and automatically, applying stereotypes that often include morally evaluated traits (e.g., aggressiveness for African Americans; Devine, 1989)' (Haidt 2001, 820).

It is worth examining the process of stereotyping in greater detail, since it provides a useful example from the social judgment literature of how moral judgments are formed.

The effects of stereotypes on people's evaluations of others, to which Haidt refers, are often investigated using stereotype priming. Priming is a process which activates a mental category, such as 'dogs', or 'homosexuals' (an example of a mental category that is also a stereotype). Experimentally induced priming is assumed to simulate what happens when a person categorises someone in a natural setting. The *activation* of a stereotype refers to a state of increased accessibility of stereotypic attributes for use in judgments such as identification, categorization and inference. Stereotype activation is manifested in, for example: faster identification of stereotypic words; an increase in stereotypic completions of word fragments; and faster pronunciation of stereotypic words (for review, see Kunda and Spencer 2003). Stereotype *application*, in contrast, refers to the application of activated stereotyped traits when judging a person's behavior.

In the classic reference Haidt cites, Devine (1989) demonstrated the application of an activated stereotype. She subliminally primed participants with a mixture of words related to the negative Black stereotype, and neutral words. In the 'strong prime' group, 80 per cent of the words related to the negative Black stereotype and 20 per cent of the words were neutral. In the control 'weak prime' group, these ratios were reversed. Then, in a supposedly unrelated task, the participants were asked to rate the behavior of a story protagonist called Donald. Donald behaved in a series of ambiguously hostile ways (for example, he refused to pay his rent until his flat was repainted). Devine (1989) found that participants strongly primed by the Black stereotype applied it to their evaluations; they rated Donald as significantly more hostile than did the weakly primed group.

Haidt argues that evidence of automatic stereotyping 'illustrate[s] the operation of the intuitive judgment link (Link 1), in which the perception of a person or an event leads instantly and automatically to a moral judgment without any conscious reflection or reasoning' (2001, 820). This is represented in Figure 2.

The SIM can thus neatly account for the findings of Devine (1989) and other similar studies (e.g., Lepore and Brown 1997). However, what is problematic for the SIM is evidence that people do not inevitably base their judgments on their automatic 'intuitions'—there is not always a direct link between intuitions and judgments. Research suggests that 'personal motivations moderate the extent to which activated constructs [such as stereotypes] are applied' (Monteith, Sherman, and Devine 1998, 73). Furthermore, there is evidence that it is controlled cognitive processes that interrupt the route between intuitions and

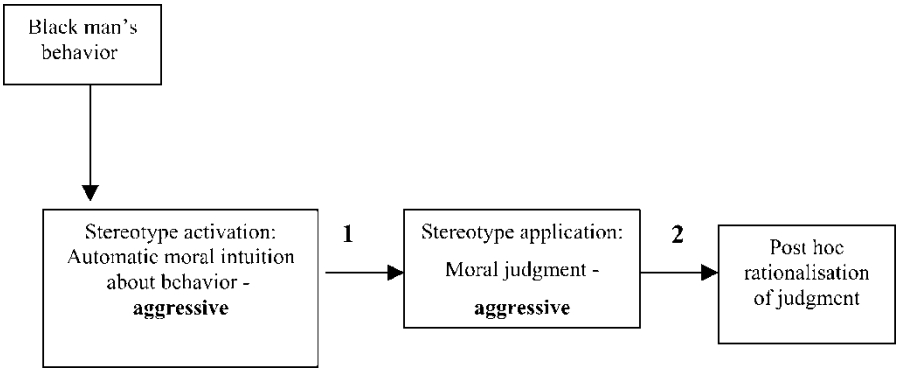

FIGURE 2
Automatic stereotype activation and application, represented in the context of the SIM

judgments. That is, conscious, effortful and intentional processes can have a causal effect *prior* to judgment formation. This evidence appears to challenge Haidt's claim that moral judgments follow automatically from the perception of a person or event.

Motivations arising from a desire to be accurate, a dependency on the person being judged for an outcome, and concerns about stereotyping have all been shown to disrupt Link 1 of the SIM. With regard to accuracy motivation, Thompson et al. (1994) primed participants with either positive or negative personality traits. The participants were then asked to perform a supposedly unrelated impression formation task. Some were motivated to be accurate in their impressions; others were not. Thompson et al. found that the 'low accuracy' condition participants interpreted information about the person they were evaluating in terms of the primed traits (a process known as assimilation); they automatically applied the activated constructs to their impressions, in line with the SIM. However, the 'high accuracy' condition participants did not show assimilation of the primed traits in their impressions. Further experimentation showed that 'resisting' assimilation effects required attentional resources. Thompson et al. (1994) repeated the experiment, but this time experimentally manipulated their participants' attentional resources (low versus high attentional depletion). A digit rehearsal task was used as a cognitive load. They found that only the participants who were both motivated to be accurate, *and* who had undepleted attentional resources available, were able to interpret the behavior of the target in a prime-inconsistent fashion. The other groups, by contrast, showed prime-consistent assimilation. Since only controlled, but not automatic, processes require attentional resources, controlled processes must have been involved in resisting assimilation during judgment formation.

Being dependent on another person for an outcome has also been shown to motivate people to make accurate judgments (e.g., Neuberg and Fiske 1987). Using the manipulation of 'outcome dependency', Pendry and Macrae (1994) asked participants to listen to a self-description of 'Hilda', an elderly woman. The 'outcome-dependent' participants were told that they would be working with her on a problem-solving task, and that a cash prize would be given to the best performing team. While listening to Hilda's personal profile, participants had to switch off a light bulb (using the space bar on a computer keyboard) every time it came on. The rationale behind this secondary task is that as the attentional resources used on the primary task (listening to Hilda) increases, performance on the lightbulb task slows (see Pendry and Macrae 1994, 314–15). The participants were then asked to evaluate Hilda on a number of stereotype consistent and inconsistent dimensions. 'Outcome-dependent' participants provided the least stereotypic evaluations of Hilda, compared with participants who thought that their chances for the cash prize did not depend on Hilda. Moreover, the 'outcome-dependent' participants responded significantly more slowly to the light bulb. Their processing of the information about Hilda—which involved disrupting the automatic link between the 'elderly woman' stereotype and their judgment about her—required controlled processing.

It also seems that people do not apply activated stereotypes to stereotyped groups if they believe that it is unacceptable (Pressly and Devine 1997, cited in Monteith, Sherman, and Devine 1998). Pressly and Devine found that their participants believed that stereotyping African Americans was unacceptable, but held no strong norms against stereotyping skinheads. The experimenters then activated either the skinhead, or the African American stereotype in their participants. In a subsequent impression formation task, the group membership (i.e., skinhead or not; African American or not) of the person being judged was either stated, or left unspecified. In the skinhead condition, judgments were more

stereotypic when the protagonist was identified as a skinhead—in other words, the participants freely applied the activated stereotype. In contrast, in the African American condition, judgments were *less* stereotypic when the protagonist was identified as being African American, compared with the condition in which his race was left unspecified. In fact, participants in the African American race-specified condition provided judgments that were even *less* stereotypic than judgments of a race-unspecified protagonist made by a control group who had had no stereotype activation. In other words, the former group of participants bent over backwards to avoid applying the activated stereotype to their judgments. As in the previous two experiments, it appears that controlled cognitive processes are necessary in order to prevent the participants from applying an activated stereotype. Research using a similar paradigm (Wyer et al. 1997, cited in Monteith, Sherman, and Devine 1998) found that only when attentional resources were available did participants stop themselves from applying an activated stereotype (African American or Asian American) to their impressions of a second target belonging to the same social group.

These findings suggest that, at least under certain circumstances (the presence of appropriate motivation and adequate attentional resources), Haidt (2001) may be premature in his assertion that '. . . perception of a person or event leads instantly and automatically to a moral judgment without any conscious reflection or reasoning'. As represented in Figure 3, it appears that Link 1 can in fact be disrupted by controlled, effortful, cognitive processes.

Also relevant to this claim are recent findings from neuroimaging research. Greene et al. (2004) measured brain activity using functional magnetic resonance imaging (fMRI) while participants performed either 'difficult' or 'easy' moral judgments (categorised according to the amount of time the subject required in order to make the judgment). The difficult moral dilemmas generally involved a conflict between deciding that it was either 'appropriate' or 'inappropriate' to perform a moral violation that would maximise overall utilitarian welfare. (For example, is it appropriate to smother your crying baby to death in order to prevent enemy soldiers from overhearing the child, and finding and killing you, your child, and your fellow townspeople?) The 'easy' moral dilemmas, by contrast, involved no such conflicts. Greene et al. (2004) found that the difficult, compared

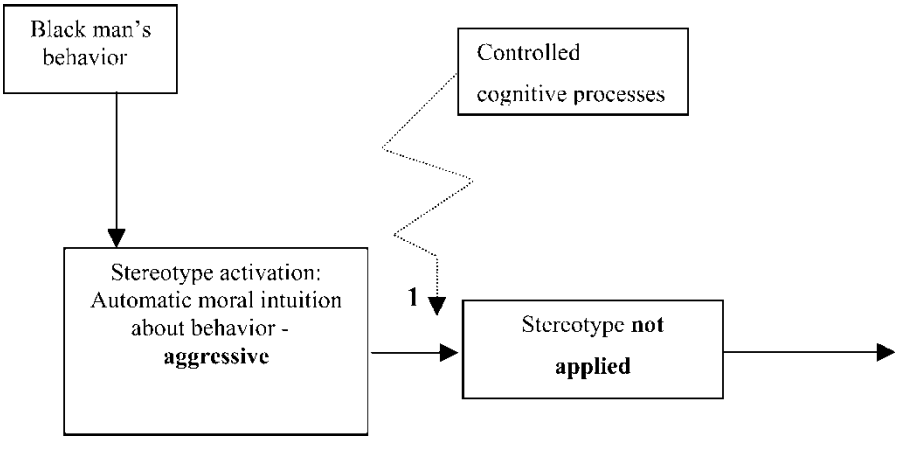

FIGURE 3
Disruption of Link 1 by controlled cognitive processes

with easy moral judgments, involved significant activity in the anterior dorsolateral prefrontal cortex (anterior DLPFC), a region associated with abstract reasoning processes. They suggest that the activation of this region during difficult moral judgments reflects the engagement of utilitarian reasoning and cognitive control over automatic 'intuitive' judgments (see also Greene et al. 2001). In line with this interpretation, greater anterior DLPFC activity was observed on trials in which a utilitarian response was considered to be appropriate (e.g., deciding to smother the baby). Increased activity was also observed in the anterior cingulate cortex (ACC) during difficult, compared to easy, moral judgments. The ACC is thought to be involved in 'cognitive conflict' (situations in which two or more possible responses are in competition with one another). The researchers' plausible interpretation of activity in the ACC is that it reflects the conflict between an automatic emotion-based response, and the alternative response to the moral dilemma based on a controlled utilitarian cost–benefit analysis.

Given the relatively low temporal resolution of fMRI it is possible that, in accordance with the SIM, increased neural activity in the anterior DLPFC during consideration of difficult moral dilemmas reflects not deliberation, but the process of *post hoc* justification of the participant's intuitive response. (This interpretation implies that a utilitarian response could be the automatic one.) However, there are two reasons why this seems a less plausible interpretation of the data. First, since participants were not asked to justify their responses it seems more likely that the neural activity corresponded to deliberation rather than justification. Secondly, the observation of greater activity in the anterior DLPFC on trials in which utilitarian responses were decided to be appropriate is consistent with the interpretation that during this processing time, an intuitive, emotion-based response was being 'mentally challenged'. Were this neural activity a reflection of *post hoc* justification processes, there would be no reason to expect to see greater activity for utilitarian than non-utilitarian judgments. Thus overall, as Greene et al. (2004, 397) note, the long deliberation times required by the volunteers to make their judgments, in combination with neural activity observed in brain regions associated with controlled cognitive processes, suggests that 'high level cognitive processes are marshaled in the resolution of difficult moral dilemmas and stand in tension with the social intuitionist claim that in nearly all cases moral judgments are more akin to perception than episodes of reasoning or reflection'. These neuroimaging findings therefore provide convergent support for the proposal that, via the use of attention-depleting, controlled cognitive processes, the formation of automatic judgments can be inhibited.

What are the Controlled Cognitive Processes Involved in Disrupting Link 1, and How Do They Act?

Haidt's (2001, 818) definition of moral reasoning is that it is a controlled cognitive process of 'transforming given information about people in order to reach a moral judgment'. The controlled processes identified in the activated stereotype studies would appear to fulfil these criteria. However, it is interesting to explore the functional role of people's conscious reflections on their moral judgments in more detail. Monteith and colleagues, for example, have found that many people report a conflict between their consciously held nonprejudiced beliefs and their persistent stereotypical responses. When asked how they *would* respond to situations involving stereotyped groups, and how they *should* respond, many people report a discrepancy. That is, they are aware that

their automatic responses fall short of their personal standards (e.g., Monteith and Voils 1998). In low-prejudice individuals, these discrepancies are accompanied by feelings of guilt and self-criticism (Monteith 1993).

Monteith and colleagues have begun to investigate how automatic processes can be controlled and changed in the context of stereotyping. They have proposed that 'the experience of discrepancy-related guilt should serve to establish cues for control that, when present in future situations in which prejudiced responses are possible, should instigate a variety of self-regulatory mechanisms that allow one to provide nonprejudiced responses' (Voils, Ashburn-Nardo, and Monteith 2002, 30). These processes are represented in Figure 4.

In a test of this model, Monteith et al. (2002) experimentally induced 'should/would' discrepant responses in low-prejudice participants by giving them false psychophysiological feedback. They were induced to believe that they were unable to control negative physiological reactions to pictures of Black people. In support of the model,

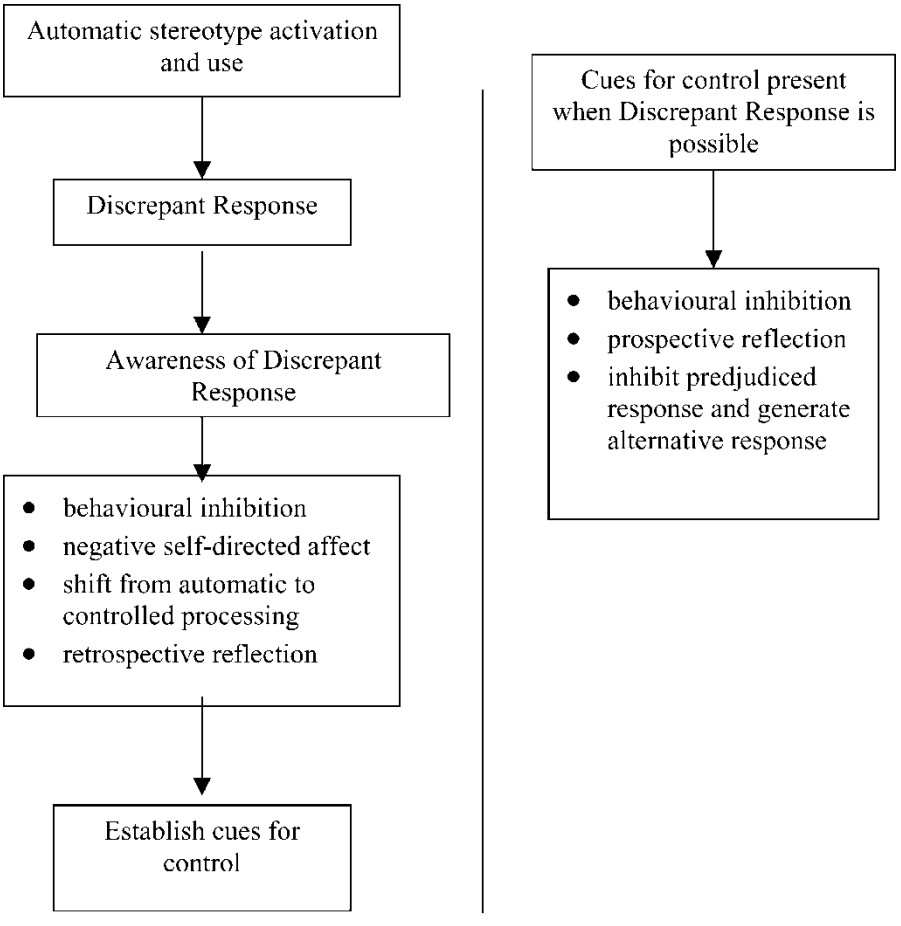

FIGURE 4
Model of the development of self-regulatory mechanisms for prejudice control. Reprinted with permission from Monteith et al. (2002, 1031)

they found evidence of behavioral inhibition, negative self-directed affect (e.g., guilt), and retrospective reflection on the discrepant response.

Monteith et al. (2002) then tested the part of the model concerned with the operation and effects of established cues for control of responses in situations in which a prejudiced response is possible. In one experiment, White, low-prejudice participants were given the racial Implicit Association Test (IAT). This test is sensitive to the affective valence associated with typical Black names, versus typical White names. (The participant is presented with a series of words and names, and asked to categorise each word as 'pleasant'/'unpleasant', and each name as 'White name'/'Black name'. In the 'incongruent' condition, participants use the same keyboard key to categorise words as 'pleasant' and names as 'Black'. In the 'congruent' condition, the same key is used to categorise words as 'unpleasant' and names as 'Black.' A reliable finding is that White participants are slower to categorise in the 'incongruent' condition, demonstrating a negative affective association with Black names.) Following the IAT, participants were shown their score (all were slower on 'incongruent' trials), and the implications of this in terms of racial prejudice were discussed with them. Then, in a supposedly unrelated task, they were presented with words and names, and asked to say whether they liked them. Some of the names had been used in the IAT task. Participants who had felt guilty about their performance on the IAT showed behavioral inhibition in response to Black names on this latter task, and were less likely to generate racially biased responses (that is, they were more likely to say that they liked the name). Monteith et al. (2002, 1046) suggested that 'because the participants had established cues for control, prospective reflection could occur, and racial biases were less likely to be manifested'. Again, this work suggests the view that controlled cognitive processes can intervene *prior* to social judgment formation.

Furthermore, interviews with low-prejudice participants reveal that they are aware of their automatic prejudicial tendencies, concerned about them, and make conscious efforts to adapt their responses to stereotyped groups (see Monteith 1993; Monteith et al. 2002).

> When I was younger . . . my brother, who I idolized, used 'fag' as an insulting comment. I picked it up but as I aged and slowly began to assimilate knowledge about homosexuality . . . I began to feel bad about myself. I felt that this was not how I wanted to respond toward minorities. (Monteith 1993, 484)

The 'Automatization' of Prejudice Control

Research into the nature of the controlled cognitive processes involved in disrupting automatic judgments is still in early stages. Nonetheless, the work of Monteith and colleagues suggests that, at least in the domain of judgments about racial groups, people do in fact question their automatic moral judgments and attitudes. Indeed, the model proposed by Monteith et al. (2002, 912) suggests that an individual's conscious reflection on their automatic responses may eventually lead to the successful 'automatization' of prejudice control: at first, stereotype application is controlled, and eventually, stereotype activation itself. Investigating this idea, Monteith and Voils (1998) compared two groups of low-prejudice participants: a 'No Discrepancy' group who reported no discrepancies in their 'should' and 'would' responses to stereotyped groups; and a 'Discrepancy' group who did report discrepancies. The participants were asked to rate a series of jokes, some

of which were racist. They were also given a task to perform concurrently that was either high or low in attentional distraction. In line with previous findings that attentional resources are necessary to inhibit automatic responses, the 'Discrepancy' group evaluated the racist jokes more favorably under high distraction, compared with low distraction. In contrast, the 'No Discrepancy' group provided equally unfavorable ratings of the racist jokes under conditions of both high and low distraction. This suggests that their non-prejudicial responses may have become 'automatized.'

More to Automatic Judgments than Meets the Eye?

This work raises the possibility that there may be more to automatic judgment processes than initially meets the eye. Indeed, this conclusion is also suggested by a closer examination of the research into automatic social judgment processes themselves. Bargh and Chartrand (1999, 476) have described automatic processes as '"mental butlers" who know our tendencies and preferences so well that they anticipate and take care of them for us, without having to be asked'. According to the influential auto-motive model, when a goal is consciously acted upon repeatedly and consistently in a particular situation, the goal becomes automated through its repeated selection (Bargh 1990). The goal is automatically triggered by the situation, in the absence of conscious intent. Bargh et al. (2001, 1015) 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'.

In line with the auto-motive model, Fishbach, Friedman, and Kruglanski (2003) reasoned that—with repeated exertion of self-control—the processing of tempting stimuli should unconsciously activate the high-priority goal that the temptation potentially threatens. They asked participants to rate how concerned they were about watching their weight and staying slim, and how successful they were in achieving this aim. They were then given a lexical decision task (identifying letter strings as words or non-words). The target words were related to weight watching (e.g., *diet*, *slim*, *thin*, *fit)* and were preceded by subliminal primes: either relevant temptation words (e.g., *cake* and *chocolate*), or irrelevant temptation words. The researchers found that the more important weight watching was to participants, the faster successful (but not unsuccessful) food self-regulators were in identifying diet-related words following the fattening food primes. Thus, temptation stimuli automatically activated the 'diet' category—but only in people who successfully exerted self-control with regard to food. A further experiment showed that 'temptation primed' weight-watching participants were more likely to select an apple than a chocolate bar, compared with control participants. Thus while these experiments demonstrate unconscious goal activation, as the authors note, this 'automatic goal activation may, in fact, . . . promote the personal control of behavior, if by that we mean action congruent with one's own system of subjective values and priorities' (Fishbach, Friedman, and Kruglanski 2003, 306).

Other research extends demonstration of such automatic goals in action to domains particularly relevant to moral judgment making. Moskowitz et al. (1999) hypothesised, in line with the auto-motive model, that 'volition, in the form of chronic egalitarian goals, leads to the passive and preconscious control of stereotype activation . . . Holding a chronic egalitarian goal can lead one to strive repeatedly for attainment of the goal, and it can lead to [automatic] activation of the goal whenever a goal-relevant person is encountered. Thus, the goal of being egalitarian would operate preconsciously—it need not

require awareness or effort' (167–68). To test this hypothesis, they compared men high or low in their motivation to judge women in a fair and egalitarian manner. The men were then primed with male and female faces, and the degree of stereotype activation was assessed using a word pronunciation task. The target words were attributes either consistent or inconsistent with the stereotype of women. Men unmotivated to judge women in a fair and egalitarian manner were quicker to respond to stereotyped attributes than to non-stereotyped attributes, following female faces. However, there were no such signs of stereotype activation in the egalitarian participants. A follow-up experiment suggested that this was because the egalitarian participants were inhibiting the primed stereotype.

Furthermore, Gollwitzer et al. (1999, cited in Gollwitzer 1999) have demonstrated that even a conscious commitment to a particular goal made shortly before priming can inhibit stereotype activation. In their account of 'strategic automaticity' in goal pursuit, the individual consciously forms an 'implementation intention' of the form 'When I encounter *x*, I will perform behaviour *y*'. This single act of will is sufficient to hand over control of the implementation of their goal to automatic processes that are controlled by situational cues. Gollwitzer and colleagues asked participants to form an implementation intention not to display stereotyped and prejudicial judgments (regarding the elderly, women, and homeless people). People who formed such intentions did not show the usual stereotype activation effects.

What are the implications for the SIM of these findings that subjectively held values can exert influence on stereotype activation and social judgments? While the evidence cited here does not directly contradict the claim that 'the perception of a person or an event leads instantly and automatically to a moral judgment without any conscious reflection or reasoning' (Haidt 2001, 820), it nonetheless suggests the SIM overlooks some of the important subtlety of how at least some automatic processes arise. It seems that at least some automatic processes reflect the action of prior controlled processes. Thus the fact that a process is currently automatic does not necessarily imply that conscious reflection or reasoning did not take place at an earlier point in the individual's history.

Distinguishing the Current Proposal from the SIM

Thus far, I have suggested two modifications to the processes involved in Link 1 of the SIM (see Figure 5). I have argued that research in stereotyping and other social judgment processes suggests that moral judgments are not necessarily formed automatically in the absence of controlled cognitive processes. Moreover, there is support for the idea that even when moral judgments follow the automatic route suggested by Link 1 in the SIM, the output of this automatic process may reflect the action of prior controlled processes.

Haidt (2001, 2004) has stressed that the SIM does include a role for reasoning. However, according to his account reasoning almost always acts via a social process involving one or more other individuals. Thus according to the SIM, individual A's (*post hoc*) reasoning will normally only influence A's judgment via the very indirect route of Link 3 (the 'reasoned persuasion' route). A's moral reasoning can affect the intuitions of another individual, B, whose subsequent automatic judgments and/or *post hoc* reasoning may then affect A's intuitions, in an iterative process. In contrast, the claim made here is that A's own private reasoning processes may play a direct and prior role in the formation of A's moral judgments.

Can the private moral reasoning links (Links 5 and 6) in the SIM account for at least some of the data offered here as a challenge to the model? In particular, can SIM account for

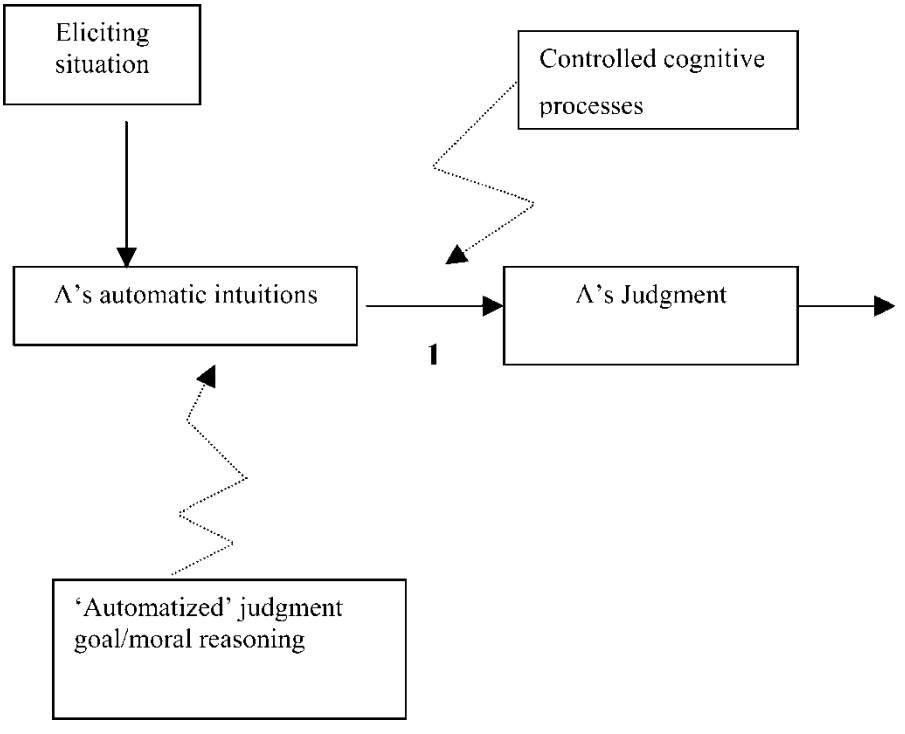

FIGURE 5
Modified model showing potential effects of controlled cognitive processes and 'automatized' judgment processes in the formation of moral judgments

situations in which individuals come to over-ride their automatic moral judgments? Link 5 appears to offer the best chance of accounting for these findings. According to the SIM, Link 5 represents circumstances in which:

> people . . . reason their way to a judgment by sheer force of logic, overriding their initial intuition. In such cases reasoning truly is causal . . . However, such reasoning is hypothesized to be rare, occurring primarily in cases in which the initial intuition is weak and processing capacity is high. In cases where the reasoned judgment conflicts with a strong intuitive judgment, a person usually has a 'dual attitude' in which the reasoned judgment may be expressed verbally yet the intuitive judgment continues to exist under the surface. (Haidt 2001, 819)

There are thus two important claims here with regard to Link 5: that it is relatively infrequent, and that it is relatively ineffective. With regard to the frequency with which automatic intuitions (or the judgments based on them) are overcome, this remains an important empirical question. However, it is interesting to note that the vast majority of individuals (over 90 per cent) report discrepancies between their privately experienced 'should' *versus* 'would' responses to stereotyped groups, and the conscious recognition of such discrepancies is independent of factors such as educational level, age and income (Voils, Ashburn-Nardo, and Monteith 2002). This would not appear to sit easily with the SIM claim that, in the absence of social pressure or persuasion, people provide *post hoc* moral justifications for

their automatic moral judgments. Haidt's other claim that, at best, private moral reasoning creates a 'dual attitude' (when the intuition is strong) may also be prematurely pessimistic. As indicated by the responses of Low Prejudice/No Discrepancy individuals to racist jokes (Monteith and Voils 1998), it seems possible that unwanted attitudes may be over-ridden, and the consciously preferred attitude may become 'automatized'.

Is Reasoning Up to the Job?

I now address concerns raised by Haidt (2001) that our social reasoning may be inadequate for the job of questioning our moral judgments. As part of his case for the SIM, Haidt provides evidence of two characteristics of our reasoning: the motivated nature of our reasoning ('the reasoning process is more like a lawyer defending a client than a judge or scientist seeking truth' (2001, 820); and the *post hoc* nature of reasons ('the reasoning process readily constructs justifications of intuitive judgments, causing the illusion of objective reasoning' (2001, 822)). Haidt appears to imply that reasoning is so biased toward supporting or justifying motivated or preordained conclusions that it is unlikely to play a questioning role in our moral judgments:

> . . . 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 look instead 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. (2001, 821–22)

Social psychologists have identified three main categories of motivation potentially at work whenever a person processes social or personally relevant information (e.g., Chaiken, Giner-Sorolla, and Chen 1996). Haidt (2001) refers to two of these categories: 'impression motivation' (the desire to create a good impression to the person with whom one is interacting); and 'defense motivation' (the desire to hold attitudes and beliefs that are consistent with one's self-concept). However, we are also importantly motivated by the desire to hold objectively true beliefs and attitudes ('accuracy motivation'). Chaiken, Giner-Sorolla, and Chen (1996) suggest that one or more motives may be simultaneously active in a given situation, and beliefs and attitudes will reflect their interaction. Research suggests that certain features of the situation are important in facilitating one or other motivation. For example, as we've seen, increasing the accountability of a person's judgment, outcome dependency, or a concern not to stereotype can motivate increased accuracy of social judgment formation, invoking the use of controlled processing strategies at some point during the judgment formation process (see Monteith, Sherman, and Devine 1998). Moreover, even when the biasing motivations of impression management and self-defense are at work, reasoning is not impervious to evidence. For example, participants motivated to rate themselves as more introverted because they have been told that introverts are more likely to be successful, are nonetheless sensitive to their actual score on this measure (Kunda and Sanitioso 1989; see also Kunda 1987, for a similar finding).

As a further part of his claim that reasoning is unlikely to effectively challenge our moral beliefs, Haidt (2001, 822) cites evidence showing that '[w]hen people are asked to explain the causes of their judgments and actions, they frequently cite factors that could not have mattered and fail to recognise factors that did matter'. (See, for example,

Nisbett and Wilson 1977). Haidt suggests that our explanations for our moral judgments may be similarly misguided and *post hoc*:

> [b]ecause the justifications that people give are closely related to the moral judgments that they make, prior researchers have assumed that the justificatory reasons caused the judgments. But if people lack access to their automatic judgment processes then the reverse causal path becomes more plausible. (2001, 822)

However, as Haidt (2001) himself is careful to note, a direction of causality from judgment to reasoning is only plausible if the individual does indeed lack access to the processes that caused their moral judgment. For this reason, we must be careful not to over-interpret evidence that we sometimes make *post hoc* and erroneous justifications of our judgments. This evidence does not provide independent support for the specific claim that moral reasons might—or are even likely to—fall into this category. If a moral judgment is, *contra* Haidt, in fact largely determined by conscious controlled processes (either moments before, or on prior occasions), then we will presumably be correct in supposing that those reasons were causal in our moral judgment.

It is, however, likely that our moral reasoning will often be biased by factors that act on us automatically. Mood, for example, can have a significant effect on social judgments (e.g., Forgas 1994). Indeed, Lerner, Goldberg, and Tetlock (1998) found that participants angered by a film showing the brutal beating of a teenager were subsequently harsher in their moral judgments of the responsibility of defendants in negligence cases, and were less sensitive to mitigating circumstances than participants who had watched a neutral film before making their judgments. Nonetheless, this does not imply that the entire process leading to their moral judgments must also have been automatic and unreasoned—we cannot justify tarring *all* of an individual's moral reasons with the same brush, just because biasing automatic processes can be shown to be at work. (Interestingly, angry participants who thought that they would have to explain their judgments to a researcher afterwards were able to overcome the biasing effects of 'carry-over' anger.)

Summary and Implications

The SIM performs the important service of placing the discussion of moral judgments within an updated context of contemporary cognitive social psychology. Haidt's model also forces a timely recognition of the extent to which our social (and presumably our moral) judgments are based on automatic processes. If we can extrapolate from the social cognitive literature to the domain of moral judgment, this has interesting practical implications with regard to the factors likely to influence the quality of our moral judgments. As the research described here suggests, moral judgments are likely to be improved by factors such as an emotionally neutral state, accountability, motivation to be accurate, possession of egalitarian values, availability of attentional resources, and a past history of conscious effort to make moral judgments in line with consciously held values. In contrast, lack of accountability, distraction, particular mood states, or threats to self-esteem (e.g., Sinclair and Kunda 2000) may impair the quality of our moral judgments.

However, I have suggested that Haidt's representation of the social judgment literature is too selective, giving rise to the misleading impression that controlled reasoning processes play little direct role in our moral judgments, but normally only act indirectly via other people. The research presented here suggests that automatic moral judgment

formation is not inevitable; controlled cognitive processes can interfere. Further research is necessary to ascertain the precise nature of these controlled processes; when in the judgment formation process they act, and how frequently they occur in everyday moral reasoning. However, research gathered by Monteith and colleagues does suggest that private moral reasoning may play an important role in moral judgments. Furthermore, in line with the idea that our social cognitive processes can become more efficient through the automatization of goals (e.g., Bargh 1990), there is preliminary evidence that judgment goals, arising from reasoning processes, can become 'automatized' (e.g., Monteith and Voils 1998; Moskowitz et al. 1999).

Thus, it appears that perhaps our social reasoning is not always as ignobly motivated and *post hoc* as the SIM implies. Where the individual is motivated to form accurate judgments, and has the attentional resources available to do so, automatic intuitions can be over-ridden and accurate judgments formed. It remains an important empirical question how often in everyday moral reasoning these criteria are met. But for the time being there remains the possibility that sometimes the emotional dog is not wagging the rational tail, but chasing it.

REFERENCES

BARGH, J. A. 1990. Auto-motives: Preconscious determinants of social interaction. In *Handbook of motivation and cognition: Foundations of social behavior.* Vol. 2, edited by E. T. Higgins and R. M. Sorrentino. New York: Guilford Press.

BARGH, J. A., and T. L. CHARTRAND. 1999. The unbearable automaticity of being. *American Psychologist* 54: 462–79.

BARGH, J. A., P. M. GOLLWITZER, A. LEE-CHAI, K. BARNDOLLAR, and R. TRÖTSCHEL. 2001. The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology* 81: 1014–27.

CHAIKEN, S., R. GINER-SOROLLA, and S. CHEN. 1996. Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In *The psychology of action: Linking cognition and motivation to behavior*, edited by P. M. Gollwitzer and J. A. Bargh. New York: Guilford Press.

DEVINE, P. G. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* 56: 5–18.

FISHBACH, A., R. S. FRIEDMAN, and A. W. KRUGLANSKI. 2003. Leading us not into temptation: Momentary allurements elicit overriding goal activation. *Journal of Personality and Social Psychology* 84: 296–309.

FORGAS, J. P. 1994. Sad and guilty? Affective influences on the explanation of conflict in close relationships. *Journal of Personality and Social Psychology* 66: 56–68.

GOLLWITZER, P. M. 1999. Implementation intentions: Strong effects of simple plans. *American Psychologist* 54: 493–503.

GREENE, J. D., L. E. NYSTROM, A. D. ENGELL, J. M. DARLEY, and J. D. COHEN. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44: 389–400.

GREENE, J. D., R. B. SOMMERVILLE, L. E. NYSTROM, J. M. DARLEY, and J. D. COHEN. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 2105–98.

HAIDT, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108: 814–34.

——. 2002. 'Dialogue between my head and my heart': Affective influences on moral judgment. *Psychological Inquiry* 13: 54–56.

——. 2004. The emotional dog gets mistaken for a possum. *Review of General Psychology* 8: 283–90.

KUNDA, Z. 1987. Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* 53: 636–47.

KUNDA, Z., and R. SANITIOSO. 1989. Motivated changes in the self-concept. *Journal of Experimental Social Psychology* 25: 272–85.

KUNDA, Z., and S. J. SPENCER. 2003. When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin* 129: 522–44.

LEPORE, L., and R. BROWN. 1997. Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology* 72: 275–87.

LERNER, J. S., J. H. GOLDBERG, and P. E. TETLOCK. 1998. Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin* 24: 563–74.

MONTEITH, M. J. 1993. Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology* 65: 469–85.

MONTEITH, M. J., L. ASHBURN-NARDO, C. I. VOILS, and A. M. CZOPP. 2002. Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology* 83: 1029–50.

MONTEITH, M. J., J. W. SHERMAN, and P. G. DEVINE. 1998. Suppression as a stereotype control strategy. *Personality and Social Psychology Review* 2: 63–82.

MONTEITH, M. J., and C. I. VOILS. 1998. Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies. *Journal of Personality and Social Psychology* 75: 901–16.

MOSKOWITZ, G. B., P. M. GOLLWITZER, W. WASEL, and B. SCHAAL. 1999. Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology* 77: 167–84.

NEUBERG, S. L., and S. T. FISKE. 1987. Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology* 53: 431–44.

NISBETT, R. E., and T. D. WILSON. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84: 231–59.

PENDRY, L. F., and C. N. MACRAE. 1994. Stereotypes and mental life: The case of the motivated but thwarted tactician. *Journal of Experimental Social Psychology* 30: 303–25.

SINCLAIR, L., and Z. KUNDA. 2000. Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin* 26: 1329–42.

THOMPSON, E. P., G. B. MOSKOWITZ, S. CHAIKEN, and J. A. BARGH. 1994. Accuracy motivation attenuates covert priming: The systematic *re*processing of social information. *Journal of Personality and Social Psychology* 66: 474–89.

VOILS, C. I., L. ASHBURN-NARDO, and M. J. MONTEITH. 2002. Evidence of prejudice-related conflict and associated affect beyond the college setting. *Group Processes and Intergroup Relations* 5: 19–33.

Cordelia Fine, Centre for Applied Philosophy and Public Ethics, Department of Philosophy, University of Melbourne, Parkville, Victoria 3010, Australia. E-mail: cfine@unimelb.edu.au