Do theories of implicit race bias change moral judgments?

C. Daryl Cameron

B. Keith Payne

Joshua Knobe

Abstract

Recent work in social psychology suggests that people harbor "implicit race biases," biases which can be unconscious or uncontrollable. Because awareness and control have traditionally been deemed necessary for the ascription of moral responsibility, implicit biases present a unique challenge: do we pardon discrimination based on implicit biases because of its unintentional nature, or do we punish discrimination regardless of how it comes about? The present experiments investigated the impact such theories have upon moral judgments about racial discrimination. The results show that different theories differ in their impact on moral judgments: when implicit biases are defined as unconscious, people hold the biased agent less morally responsible than when these biases are defined as automatic (i.e., difficult to control), or when no theory of implicit bias is provided.

Keywords: implicit bias, moral judgment, unconscious, automatic, stereotyping, responsibility

"If it were indeed the case... that stereotyping occurs without an individual's awareness or control, then the implications for society... are tremendously depressing. Most ominously, how could anyone be held responsible, legally or otherwise, for discriminatory or prejudicial behavior when psychological science has shown such effects to occur unintentionally?"

Bargh (1999)

"Unwitting or ingrained bias is no less injurious or worthy of eradication than blatant or calculated discrimination... the fact that some may have been unaware of that motivation, even within themselves, neither alters the fact of its existence nor excuses it."

Price Waterhouse v. Hopkins (1989), cf. Lane, Kang, & Banaji (2007)

On February 4, 1999, four white policemen shot a young man named Amadou Diallo nineteen times. They seem to have believed that that he was reaching for a gun when in fact he was only trying to pull out his wallet. Reactions to this shooting were polarized. Some people thought that the policemen had simply made an honest mistake, while others thought that this event was a symptom of a pervasive racial bias on the part of the New York Police Department. But suppose that the people watching these news reports had learned about recent findings in social psychology which suggest that racial biases can operate without conscious awareness or intentional control. How might knowing about these implicit race biases influence the moral judgments people make in a case like Amadou Diallo's?

Though overt racism has been in decline for decades, research suggests that more subtle forms of racial bias may be quite prevalent throughout the population (Nosek, 2007). These implicit biases may not be consciously recognized, and are often quite difficult to control (Bargh,

1999). Importantly, they are associated with discriminatory behavior, such as non-verbal negativity toward out-group members (Dovidio, Kawakami, & Gaertner, 2002; McConnell & Leibold, 2001), severity of criminal sentencing decisions (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006), and greater likelihood of mistaking a harmless tool for a gun when it is held by a Black man (Correll, Park, Judd, & Wittenbrink, 2002; Payne, 2001). Implicit race biases may thus have morally relevant outcomes that most people would not explicitly endorse (for reviews, see Jost et al., 2009; Payne & Cameron, 2010).

Aside from the huge impact that this research has had within the scientific community (Blasi & Jost, 2006; Gawronski, Lebel, & Peters, 2007; Payne, 2001), it is also receiving a great deal of attention for the difficult moral questions that it raises (Arkes & Tetlock, 2004; Banaji, Nosek, & Greenwald, 2004; Fiske, 1989; Fiske, 2005; Jolls & Sunstein, 2006; Kelly & Roedder, 2008; Lane et al., 2007; Mitchell & Tetlock, 2007). If people come to believe that racial discrimination is the result of unconscious and uncontrollable processes, will they conclude that individuals who engage in racial discrimination are not morally responsible or blameworthy for what they have done?

Two Views on Implicit Race Bias

Moral views on discrimination might depend on the specific interpretation of implicit race bias that becomes embedded in public consciousness. Though most theories agree that implicit race biases counteract intention and to some degree control, scientific opinion is more divided as to how much consciousness we have of their presence and influence. This difference of opinion can be traced to a broader tension running through the field as to what is "implicit" about implicit social cognition, and this tension has spawned different process accounts of implicit race bias (Payne & Gawronski, 2010).

One strand of research has defined implicit biases as primarily *unconscious* in nature. For instance, Greenwald and Banaji (1995) defined implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable thought, feeling, or action toward social objects" (p. 8). Wilson, Lindsey, and Schooler (2000) similarly argued for the existence of unconscious, implicit attitudes that are separate from and potentially at odds with conscious, explicit attitudes. On this unconscious view of implicit race bias, people have unconscious racial biases which cause discriminatory behavior surreptitiously.

By contrast, a second strand of research has defined implicit biases as primarily *automatic* in nature. Fazio and colleagues (1995) have argued that implicit biases are conscious attitudes that are activated automatically, and which influence behavior depending on how much behavioral control can be brought to bear. Many suggest that in the cognitively busy settings of everyday life, the motivation and ability to control implicit biases will be lacking (Bargh, 1999; Wilson & Brekke, 1994; but see Devine & Monteith, 1999 for a more optimistic reading). On this automatic view, people are aware of their racial biases but have difficulty controlling against their influence. Although there are more complex nuances that further distinguish specific theories, most existing theories of implicit bias tend to fall into one of these two thematic trends in the implicit social cognition tradition.

Implicit Race Bias and Moral Responsibility

Though psychologists and philosophers have speculated on the moral implications of implicit biases, no empirical studies have examined moral intuitions about implicit race bias. We begin by considering why we might expect implicit bias theories to reduce judgments of moral responsibility at all. And second, we consider whether there might be different effects for different theories of implicit race bias.

The general consensus in common sense and legal circles is that to be held morally responsible for an action, a person must have awareness of its implications and control over its execution (Kelly & Roedder, 2008; Machery, Faucher, & Kelly, 2009). Because discrimination resulting from implicit race biases defies intentional control, the control principle may be violated. This control principle is seen in classical and modern theories of moral responsibility attribution (Shaver, 1985; Weiner, 1995; Wigley, 2007) as well as lay intuitions (Pizarro, Uhlmann, & Salovey, 2003). If people are unable to prevent unwanted implicit race biases from influencing their decision making, then they might be held less responsible for discriminatory outcomes that follow.

Yet in addition to the question of whether implicit bias theories generally reduce judgments of moral responsibility, we can ask whether certain implicit bias theories might do so more than others. The critical feature distinguishing the two classes of implicit bias theories described above is conscious awareness. Should lacking conscious awareness of race attitudes matter for moral responsibility?

Unlike discrimination resulting from automatic bias, discrimination resulting from unconscious bias violates both the control and awareness conditions for the ascription of moral responsibility (Kelly & Roedder, 2008; Machery, Faucher, & Kelly, 2009). Without the consciousness of having an implicit race bias, it seems difficult or impossible to exert control to correct it (Hall & Payne, 2009; Levy, 2008; Nahmias, 2006; Wigley, 2007; Wilson & Brekke, 1994). Importantly, lay theories about the unconscious mind track this philosophical intuition. People acknowledge the existence of socially unacceptable unconscious impulses that are distinct from explicit moral beliefs (Moscovici, 1968/2008). They believe that these impulses can influence and interfere with the operation of conscious will, potentially compromising the

integrity of rational moral agency (Moscovici, 1968/2008; Tallis, 2002; Taslitz, 2007). People may also believe that once biases are made conscious, they become amenable to regulatory control (Moscovici, 1968/2008). Someone who discriminates on the basis of unconscious bias might thus be seen as lacking the capacity for moral judgment, whereas someone who discriminates on the basis of automatic bias might be seen as merely negligent or weak-willed. Explaining discrimination by an "unconscious" theory of implicit race bias might therefore reduce moral responsibility judgments more than explaining discrimination by an "automatic" theory of implicit race bias.

On the other hand, some have claimed that the unconsciousness of a bias does not warrant any additional reduction in moral responsibility (Nosek & Hansen, 2008; Sher, 2006; Smith, 2005). Suhler and Churchland (2009) have argued that consciousness and control are orthogonal: people can control their behavior even if it is driven by unconscious biases. Similarly, Bargh (2009) recently argued that "it is one's intentions that matter [for legal questions of personal culpability], not whether those intentions were unconscious or conscious." According to this perspective, explaining discrimination by an unconscious theory of implicit race bias should not reduce responsibility judgments any more than an explanation by an automatic theory.

Our studies were designed to answer two questions about implicit race bias and moral responsibility. First, does explaining discrimination as being due to implicit race bias lead to a general reduction in moral responsibility attribution, compared to when no such explanation is provided? Second, does explaining discrimination as being due to unconscious race bias reduce responsibility more than explaining it as being due to automatic race bias?

Experiment 1

Experiment 1 was designed to test whether people reduce their judgments of moral responsibility for discrimination when it results from implicit race bias compared to when no such explanation is given, and if so, whether one theoretical description of implicit race bias reduces responsibility more than the other. To answer these questions, we created scenarios to represent three different ways of explaining racially discriminatory behavior. The scenarios all began as follows:

John is in charge of promotions at a major company. He is supposed to decide between various candidates on the basis of merit. John is White.

The scenario representing the unconscious theory of implicit race bias continued with the following critical section:

Consciously, John thinks people should be treated equally, regardless of race. Despite this, John has a sub-conscious dislike for African Americans. He is unaware of having this dislike, but if he knew, he would disagree with this feeling because he sincerely believes in equality. This sub-conscious dislike drives his behavior in ways he does not know about.

When John decides whether or not to promote an employee, he tries to decide only on merit. But because he is unaware of this sub-conscious dislike, he is not always successful at preventing it from influencing his judgment. As a result, John sometimes unfairly denies African Americans promotions.

The scenario representing the automatic theory of implicit race bias included the following critical section. We did not stipulate racial bias as completely uncontrollable, because no current scientific theory of implicit race biases makes such claims. Rather, we presented the case as one in which the protagonist strongly desired to exert self-control over unwanted impulses:

Upon reflection, John thinks people should be treated equally, regardless of race. Despite this, John sometimes finds that he has a gut feeling of dislike toward African Americans. He is aware of having this dislike, but disagrees with this feeling because he sincerely believes in equality. This gut feeling of dislike drives his behavior in ways that he has difficulty controlling.

When John decides whether or not to promote an employee, he tries to decide only on merit. But because it is difficult to control these gut feelings, he is not always successful at preventing them from influencing his judgment. As a result, John sometimes unfairly denies African Americans promotions.

The third condition did not explain the protagonist's behavior using any theory of implicit race bias. Participants read only that the protagonist believes in equal treatment and that he or she discriminates. We refer to this condition as the "folk" view because participants had to rely on their own inferences to make judgments about the case. It seemed plausible that participants would view such agents as hypocritical and deem them the most morally responsible for their discriminatory actions. The critical section of the folk condition read as follows:

John says he thinks people should be treated equally, regardless of race. However, John sometimes unfairly denies promotions to African Americans.

We randomly assigned one of three content domains for generality: promotions within an organization, decisions to rent, and grading essay exams. In the renting scenario, the protagonist was named "Jane" and in the grading scenarios the protagonist was named "Jim." We did not expect any differences to emerge across content domains. Each participant read only one scenario.

Method

Participants

Ninety-two introductory psychology students at the University of North Carolina (60 females, 32 males) participated in the study for course credit. There were 3 Asian American, 17 African American, 75 Caucasian, and 7 Hispanic participants.

Design

Participants were randomly assigned to view a scenario describing the unconscious, automatic, or folk conditions. A renting, grading, or promotion scenario was randomly assigned in each condition. The dependent variable was the degree of moral responsibility attributed to the agent in the scenario.

Materials and Procedure

Participants were seated in front of a computer and informed that they would be reading a short story and asked to answer a number of questions. After reading the scenario, participants received four questions in random order, which together constituted a moral responsibility scale. Each question had 5-point Likert-type scaling (from 1 = Strongly disagree to 5 = Strongly agree). The scale included the following items: "John (or Jane or Jim, depending upon the assigned content domain of the scenario) is morally responsible for his treating African Americans unfairly", "John should be punished for treating African Americans unfairly", "John should not be blamed for treating African Americans unfairly" (reverse coded), and "John should not be held accountable for treating African Americans unfairly" (reverse coded). These were followed by questions about participant race and gender, and additional questions that will not be examined here.

Results

The moral responsibility scale had acceptable internal consistency (Cronbach's $\alpha = .65$). As predicted, a one-way Analysis of variance (ANOVA) showed a significant main effect of

theory condition, $F(2, 89) = 6.39$, $p = .003$, $\eta^2 = .13$ (see Figure 1).[1] In order to test the nature of this difference, we conducted post hoc analyses using Tukey's HSD. The automatic condition was not significantly different from the folk condition, $p = .80$. The unconscious condition, however, was significantly different from the automatic condition, $p = .03$. When discrimination was explained using the unconscious theory of implicit bias, participants were particularly unlikely to hold the agent morally responsible, and this accounted for the overall main effect of theory condition.[2]

Discussion

Having a theory of implicit race bias to explain discriminatory behavior significantly reduced judgments of moral responsibility. And it was not just any theory that had this effect, because subjects in the automatic and folk conditions did not make significantly different responsibility judgments. They blamed discrimination resulting from conscious but uncontrollable bias nearly as much as discrimination without any explanation. This might be seen as rather surprising, given that the agent in the scenario was stipulated as being genuinely egalitarian and having a great deal of difficulty controlling racial bias. Only discrimination resulting from unconscious bias was excused, suggesting that conscious awareness matters for judgments of moral responsibility. These findings are consistent with lay intuitions (Moscovici 2008/1968; Taslitz, 2007) and perspectives that emphasize the importance of conscious awareness (Levy, 2008; Nahmias, 2006), rather than perspectives which suggest that intent is critical regardless of consciousness (Bargh, 2009; Suhler & Churchland, 2009). In Experiment 2, we sought to replicate these initial findings as well as explore what motivated these differences in responsibility judgments.

Experiment 2

We learned in the first experiment that people ascribed less moral responsibility when discrimination was explained as a result of unconscious bias. In Experiment 2 we attempted to replicate these results and understand what motivated these differences in moral responsibility judgments. We examined four possible mediators: perceptions of intent to discriminate; perceptions that the bias reflected the actor's true self; anger and disgust toward the actor; and perceptions of the controllability of racial bias.

Much of the interest and controversy over implicit race bias research follows from the idea that these biases run counter to people's intentions (Banaji, Bazerman, & Chugh, 2003; Fiske, 1989; Fiske, 2005). Traditional theories of moral responsibility argue that responsibility judgments depend upon prior attributions of intent (Shaver, 1985; Weiner, 1995). Importantly, the folk concept of intent includes conscious awareness of an action's implications (Malle, 2005). If unconscious biases drive people's behavior in ways they do not know about, then one key criterion for intentional action is lacking. For intent to mediate the moral responsibility findings, people would have to perceive that the agent with unconscious bias had less intent to discriminate.

Other theories of moral responsibility have focused on whether an action reflects the true self of the person who performs it (Dan-Cohen, 1991; Frankfurt, 1969; Sripada, in press). True self is taken to reflect a person's core attitudes that have been solidified through prior acts of endorsement and identification (Sripada, in press). Social psychology researchers have also debated whether implicit biases are "personally endorsed" and reflective of a person's true self (Arkes & Tetlock, 2004; Gawronski, Peters, & Lebel, 2008; Nosek & Hansen, 2008). For true self to mediate the moral responsibility findings, people would have had to perceive that discrimination resulting from unconscious bias is less reflective of an agent's true self.

We also examined anger and disgust toward the protagonist. Emotions have been shown to influence many kinds of moral judgment (Damasio, 1994; Greene, 2008; Haidt, 2001; Nichols, 2004; Prinz, 2008). On one popular account of these effects, people use emotions as salient information about the severity of a moral violation (Goldberg, Lerner, & Tetlock, 1999; Schnall, Haidt, Clore, & Jordan, 2008; Sinnott-Armstrong, Young, & Cushman, 2010). People might feel less negative emotion toward unconscious bias because they view such biases as sabotaging self-control and autonomous agency (Moscovici, 1968/2008; Taslitz, 2007). By contrast, people might view discrimination due to automatic bias as due to negligence, and feel more negative emotion. For anger or disgust to mediate the moral responsibility findings, people would have had to feel less anger or disgust toward discrimination resulting from unconscious bias.

The final mediator is perceived controllability. People believe that unconscious biases are especially difficult to control but that once made conscious, they are regulated more easily (Moscovici 2008/1968; Taslitz, 2007). Thus, automatic biases might be seen as more controllable than unconscious biases. Although control and intent are often used together in everyday discourse (e.g., "intentional control"), we treat these as distinct constructs. Intent means the motivation to discriminate, whereas control means the capacity to regulate against discriminatory impulses. For perceived controllability to mediate our moral responsibility findings, participants would have had to indicate that discrimination resulting from unconscious race bias was less controllable.

Although each of these variables can be a precursor to moral judgment, they can also be a byproduct of moral judgment. Moral judgments have been shown to predict judgments of intent (Knobe, 2006), controllability (Alicke, 2000), core values (Knobe & Roedder, 2009), and

emotions (Huebner, Dwyer, & Hauser, 2009). It is therefore possible that any differences in moral responsibility judgment caused by implicit bias theories might produce corresponding downstream changes in each of these variables. Thus, in addition to examining these variables as mediators of implicit bias theory on moral judgment, we also examined whether moral judgment mediated the effect of implicit bias theory on each of these variables.

Method

Participants

Ninety-two introductory psychology students at the University of North Carolina (67 females, 25 males) participated for course credit. There were 5 Asian American, 13 African American, 65 Caucasian, 6 Hispanic, and 3 Native American participants. Data for 5 subjects who expressed confusion about the experimental procedures were excluded.

Design

The design was the same as Experiment 1, except for the inclusion of additional items in the questionnaire phase of the experiment. Participants were assigned randomly into the unconscious, automatic, or folk conditions. The same scenarios were used.

Materials and Procedure

The four items from the moral responsibility scale were presented in random order prior to the rest of the questions. Subsequent questions were randomized and came from separate scales (Intent, True Self, Anger, Disgust, Controllability) representing the mediation paths mentioned above. The Intent scale consisted of two items: “John (or Jane or Jim, depending on the content domain of the scenario) had an intention to discriminate against African Americans” and “John had an intention to treat African Americans fairly” (reverse coded). The True Self scale consisted of four items: “Do you believe that deep down, John is really prejudiced against

African Americans?" "Do you think that deep down, John really believes in racial equality? (reverse coded), "John's treatment of African Americans reflects the kind of person he truly is", and "John's actions cannot be used to judge the kind of person he truly is" (reverse coded). Anger and Disgust were measured using two items: "To what extent do you feel anger toward John" and "To what extent do you feel disgust toward John?" The Controllability scale included three items: "John can control the influence of his racial attitudes on his decisions", "John could have acted fairly toward African Americans if he had exerted more effort", and "John could not have controlled how he acted" (reverse-coded). The Intent, True Self, and Controllability measures used the same scale labels as the moral responsibility scale, whereas the Anger and Disgust measures used a different labeling (1 = Not at all to 5 = Very much). As in Experiment 1, there followed questions about participant race and gender, and other questions that will not be examined here.

Results

As in Experiment 1, the Moral Responsibility scale had adequate internal consistency (Cronbach's $\alpha = .65$). Replicating the results of Study 1, the theory used to explain discrimination significantly affected judgments of moral responsibility, $F(2, 89) = 10.07$, $p < .001$, $\eta^2 = .19$ (see Figure 2)[1]. Post hoc analyses showed that although the automatic condition elicited slightly lower judgments of responsibility than the folk condition, the difference was not significant $p = .18$. The unconscious condition, however, was significantly different from the automatic condition, $p = .04$. When people were led to understand the case using the unconscious theory of implicit bias, they were particularly unlikely to hold the agent morally responsible.[2] Because there was no significant difference in moral judgment between the automatic and folk conditions, these two conditions were combined together and compared against the unconscious

condition for all further inferential analyses.

We next examined the characteristics of our proposed mediating variables. Table 1 displays the means and standard deviations of Intent, True Self, Anger, Disgust, and Controllability for each condition. Table 2 presents the inter-correlations between these variables and Moral Responsibility. All scales had moderate internal consistency (two Intent items $r = .38$, four True Self items $\alpha = .75$, three Controllability items $\alpha = .73$). Unconscious theory did not significantly influence judgments of intent, $F(1, 90) = 2.64$, $p = .11$, $\eta^2 = .03$, or true self, $F(1, 90) = .66$, $p = .42$, $\eta^2 = .03$. However, unconscious theory reduced anger, $F(1, 90) = 8.77$, $p = .004$, $\eta^2 = .09$, but not disgust, $F(1, 90) = 3.12$, $p = .08$, $\eta^2 = .03$. Finally, participants in the unconscious theory condition reported that the agent was less able to control racial bias, $F(1, 90) = 16.96$, $p < .001$, $\eta^2 = .16$. Intriguingly, participants in the automatic condition – in which control over bias was stipulated as being low – did not discount judgments of controllability compared to the folk condition, $F(1, 56) = .03$, $p = .87$, $\eta^2 = .00$. These preliminary results suggest that anger and controllability might be associated with moral responsibility judgments.

To more formally test this supposition, we conducted a multiple mediation model (Preacher & Hayes, 2008). One virtue of such models is that, like regression models, they test for the indirect effect associated with a given mediator while controlling for all other mediators included in the model. Our model examined the influence of unconscious theory on moral judgment with four simultaneous mediators: intent, true self, anger, and perceived controllability. Figure 3 presents the multiple mediation model with coefficients for the influence of unconscious theory (compared to the other two conditions) on each of the four mediators and for the direct influence of each mediator on moral judgment. As illustrated in Table 3, only the indirect effects for anger and controllability were significant. In summary, participants

discounted moral responsibility for discrimination resulting from unconscious racial bias, and this was associated with feeling less anger and judging the bias to be less controllable.

In addition to these forward mediation results, we also examined reverse mediation. We ran four reverse single mediation models, with moral judgment mediating the influence of unconscious theory on intent, true self, anger, and controllability. As illustrated in Table 4, all reverse single mediations were significant. Figure 4 displays the combined path diagram for the four reverse single mediation models. Unconscious theory reduced judgments of moral responsibility, which reduced judgments of intent, true self, anger, and controllability.

These mediation models suggest that anger and controllability were associated with reductions in moral responsibility, but the results are inconclusive about the direction of causal influence. In contrast, for intent and true self, the results are clear: the unconscious theory of implicit bias reduced moral responsibility judgments, which in turn reduced judgments of intent and true self.

Discussion

Experiment 2 replicated the main findings of Experiment 1. Though explaining racial discrimination by a theory of implicit bias did produce a significant reduction in moral responsibility, this effect was driven almost entirely by the unconscious theory condition. Once again, participants did not discount moral responsibility for discrimination resulting from automatic racial biases. Rather, it was only when these biases were described as being unconscious that participants were willing to discount responsibility. There appears to be something morally significant about a person being unconscious of his or her racial biases, even if they eventuate in discriminatory decisions.

To examine what might underlie this difference, we examined judgments of intent, true

self, anger, and controllability. We found that reductions in moral responsibility were associated with all of these variables. In the cases of anger and controllability, both "forward" and "reverse" mediation models provided evidence of mediation. One possible interpretation is that the relationships are indeed bi-directional. However, these variables may simply be so collinear that they could be considered multiple indicators of moral responsibility. And so, we resist drawing firm conclusions about the direction of causal influence for these variables. Moreover, we found mediational evidence suggesting that subjects made intent and true self judgments in line with their prior judgments of moral responsibility. Although this finding may seem counterintuitive, it is consistent with previous research demonstrating that people often infer intent from morally harmful outcomes, rather than using intent to inform moral judgments (e.g., Knobe, 2006). Consistent with that research, judgments of intent and true self appeared to be consequences of moral judgments, rather than causes.

General Discussion

The current studies investigated the impact that different theories of implicit race bias have on judgments on moral responsibility. The results of both experiments indicate that such theories can reduce judgments of moral responsibility. When participants learned about acts of racial discrimination that were not explained by any psychological theory, they made the most severe moral judgments. When the discrimination was explained as the result of an automatic bias that was conscious but difficult to control, their moral judgments were not much changed. But when they learned that the discrimination resulted from an unconscious bias – an attitude that the agent didn't know existed – their moral judgments were significantly more lenient. This question has been a matter of speculation, leading to the diverse opinions reflected in the quotes that began our paper. Our studies have shown that explaining discrimination by theories of

implicit race bias influences moral responsibility judgments about racial discrimination. And it appears to matter *which* account of implicit race bias proves to be correct, as conscious and unconscious automaticity had different implications for assigning moral responsibility. Our results suggest that contrary to some recent philosophical arguments (e.g., Suhler & Churchland, 2009) consciousness mattered for moral responsibility judgments.

Our findings that perceptions of the actor's intent and true self followed, rather than influenced moral judgment suggests that these perceptions may have served as post-hoc justifications for the morality verdict. Consistent with intuitionist models of moral judgment (Haidt, 2001), subjects may have formed an immediate moral judgment and subsequently justified it by generating plausible reasons. It is as if subjects felt outraged by the discrimination, and therefore decided that the actor must have had bad intentions at some level.

That leaves us to speculate on why people discounted moral responsibility for discrimination due to unconscious bias, but not for discrimination due to automatic bias. It is possible that subjects perceived the failure to control a conscious prejudice as indicative of weakness of the will. In the context of racial discrimination, weakness of will – or neglecting to follow through on explicit moral principles – might be seen as especially blameworthy. In popular culture, the unconscious is sometimes perceived as diseased, irrational, and as sabotaging the ability to act in line with explicit moral principles (Moscovici, 1968/2008; Taslitz, 2007). Moreover, it seems intuitive that a person cannot exert control over a bias that they do not know exists. By contrast, people often assume that we can control habits and automatic psychological phenomena once they have been made conscious or explicit (Moscovici, 1968/2008). Explicit biases – even if they are difficult to control – might be seen as something that any truly moral person could overcome. Discrimination because of automatic (but not

"unconscious") bias might be taken to reflect weakness of will in the face of a telling moral challenge.

A second explanation for the difference between unconscious and automatic bias might be that when participants generated reasons to justify their moral judgments, they found more plausible reasons for expressing outrage in the automatic condition. That is, all subjects may have initially felt outraged by the discrimination, regardless of the cause. But participants in the unconscious condition may have found weaker reasons to sustain their initial moral outrage, leading them to perceive less moral responsibility and to judge the actor as less prejudiced. Although this is speculative, we see implicit bias as a rich context for conducting further research on the dynamics between immediate moral reactions and post hoc justifications.

The kind of old-fashioned prejudice defined by Allport (1954) as “conscious antipathy” is rather easy to assign blame for, given modern sensibilities. Implicit race biases are a trickier story, and participants in our studies wanted to hold a person less responsible for discrimination resulting from unconscious biases. Our studies are the first experimental point of contact between empirical research on implicit race biases (e.g., Payne, 2001) and conceptual research in applied ethics (e.g., Kelly & Roedder, 2008). Theories of implicit bias have received considerable attention in the popular press and in legal scholarship (Blasi & Jost, 2006; Krieger & Fiske, 2006; Lane et al., 2007), and there is reason to believe that they may continue to influence popular opinion and lay definitions of “prejudice” (Hodson & Esses, 2005; Sommers & Norton, 2006). Yet researchers are still debating whether implicit biases should be understood as unconscious or simply difficult to control (Fazio & Olson, 2003; Gawronski, Lebel, & Peters, 2007; Hall & Payne, 2009; Uhlmann, Pizarro, & Bloom, 2008). As the current research suggests, the stakes in this debate are doubly high. The winners may influence not only psychological

theory, but also the ordinary judgments people make about moral responsibility in cases of unequal treatment.

Authors' Note

C. Daryl Cameron and B. Keith Payne, University of North Carolina at Chapel Hill, Chapel Hill, NC. Joshua Knobe, Program in Cognitive Science and Department of Philosophy, Yale University, New Haven, CT.

This research was supported in part by a National Science Foundation Graduate Research Fellowship awarded to C. Daryl Cameron. We thank Lawrence J. Sanna, Paul Miceli, and Lindsay Kennedy for helpful comments on this research. We also thank everyone involved in the UNC Social Psychology Organizational Research Group who provided useful feedback during presentation of this research.

Correspondence concerning this article should be addressed to C. Daryl Cameron, Department of Psychology, Campus Box 3270, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599. E-mail: dcameron@email.unc.edu.

References

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126,* 556-574.

Allport, G. (1954). *The nature of prejudice.* Reading, MA: Addison-Wesley.

Arkes, H., & Tetlock, P.E. (2004). Attributions of implicit prejudice, or Would Jesse Jackson fail the Implicit Association Test? *Psychological Inquiry, 15,* 257-278.

Banaji, M.R., Bazerman, M.H., & Chugh, D. (2003). How (un)ethical are you? *Harvard Business Review, 81,* 56-64.

Banaji, M.R., Nosek, B.A., & Greenwald, A.G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry, 15,* 279-310.

Bargh, J.A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 361-382). New York: Guilford.

Bargh, J.A. (2009). The simplifier: A conversation with John Bargh. *Edge, 290.* Accessed online at http://www.edge.org/3rd_culture/bargh09/bargh09_index.html.

Blair, I.V., Judd, C.M., & Chapleau, K.M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15,* 674-679.

Blasi, G., & Jost, J.T. (2006). System justification theory and research: Implications for law, legal advocacy, and social justice. *California Law Review*, *94*, 1119-1168.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314-1329.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York:

Harper-Collins Publishers.

Dan-Cohen, M. (1991). Responsibility and the boundaries of the self. *Harvard Law Review, 105,* 959-1003.

Devine, P. G., & Monteith, M. J. (1999). Automaticity and control in stereotyping. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 339-360). New York: Guilford.

Dovidio, J., Kawakami, K., & Gaertner, S.L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82,* 62-68.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital sentencing outcomes. *Psychological Science, 17,* 383-386.

Fazio, R.H., Jackson, J.R., Dunton, B.C., & Williams, C.J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013-1027.

Fazio, R.H., & Olson, M.A. (2003). Implicit measures in social cognition: Their meaning and use. *Annual Review of Psychology, 54,* 297-327.

Fiske, S.T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J.S. Uleman & J. Bargh (Eds.), *Unintended thought* (pp. 253-283). New York: Guilford.

Fiske, S.T. (2005). What's in a category?: Responsibility, intent, and the avoidability of bias against outgroups. In A. Miller (Ed.), *The social psychology of good and evil* (pp. 127-140). New York: Guilford.

Frankfurt, H.G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy,*

66, 829-839.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2*, 181-193.

Gawronski, B., Peters, K. R., & LeBel, E. P. (2008). What makes mental associations personal or extra-personal? Conceptual issues in the methodological debate about implicit attitude measures. *Social and Personality Psychology Compass*, *2,* 1002-1023.

Goldberg, J., Lerner, J., & Tetlock, P. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology, 29,* 781-795.

Greene, J. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Volume 3. The neuroscience of morality: Emotion, disease, and development* (pp. 35-79). Cambridge, MA: MIT Press.

Greenwald, A.G., & Banaji, M.R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4-17.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 818-838.

Hall, D., & Payne, B. K. (2009). Unconscious attitudes, unconscious influence, and challenges to self-control. In Y. Trope, K. Ochsner, & R. Hassin (Eds.), *Social, cognitive and neuroscientific approaches to self-control.* New York, NY: Oxford University Press.

Hodson, G., & Esses, V.M. (2005). Lay perceptions of ethnic prejudice: Causes, solutions, and individual differences. *European Journal of Social Psychology, 35,* 329-344.

Huebner, B., Dwyer, S., & Hauser, M. (2008). The role of emotion in moral psychology. *Trends in Cognitive Sciences, 13,* 1-6.

Jolls, C., & Sunstein, C.R. (2006). The law of implicit bias. *California Law Review, 94,* 969-996.

Jost, J.T., Banaji, M.R., & Nosek, B.A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology, 25*, 881-919.

Jost, J.T., Rudman, L.A., Blair, I.V., Carney, D., Dasgupta, N., Glaser, J. & Hardin, C.D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior, 29,* 39-69

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass, 3*, 522-540.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 130,* 203-231.

Knobe, J., & Roedder, E. (2009). The ordinary concept of valuing. *Philosophical Issues, 19*, 131-147.

Krieger, L.H., & Fiske, S.T. (2006). Behavioral realism in employment discrimination law: Implicit bias and disparate treatment. *California Law Review, 94,* 997-1062.

Lane, K.A., Kang, J., & Banaji, M.R. (2007). Implicit social cognition and law. *Annual Review of Law and Social Science, 3,* 427-451.

Levy, N. (2008). Restoring control: Comments on George Sher. *Philosophia, 36,* 213-221.

Machery, E., Faucher, L., & Kelly, D. (2010). On the alleged inadequacy of psychological accounts of racism. *The Monist, 93,* 228-255.

Malle, B. (2005). Folk theory of mind: Conceptual foundations of human social cognition. In R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 225-255). New

York: Oxford University Press.

McConnell, A.R., & Leibold, J.M. (2001). Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435-442.

Mitchell, G., & Tetlock, P.E. (2007). Antidiscrimination law and the perils of mindreading. *Ohio State Law Journal, 67,* 1023-1122.

Moscovici, S. (2008). *Psychoanalysis: Its image and public.* Cambridge, UK: Polity Press. (Original work published 1968).

Nahmias, E. (2006). Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture, 6,* 215-237.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment.* New York: Oxford University Press.

Nosek, B.A. (2007). Implicit-explicit relations. *Current Directions in Psychological Science, 16,* 65-69.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265-292). New York: Psychology Press.

Nosek, B. A., & Hansen, J.J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion, 22*, 553-594.

Payne, B.K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181-192.

Payne, B.K., & Cameron, C.D. (2010). Divided minds, divided morals: How implicit social cognition underpins and undermines our sense of social justice. In B. Gawronski & B.K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications.* New York: Guilford.

Payne, B.K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B.K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications.* New York: Guilford.

Pizarro, D.A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14,* 267-272.

Preacher, K.H., & Hayes, A.F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879-891.

Prinz, J. (2008). *The emotional construction of morals.* Oxford: Oxford University Press.

Schnall, S., Haidt, J., Clore, G., & Jordan, A. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*, 1096-1109.

Shaver, K.G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness.* New York: Springer-Verlag.

Sher, G. (2006). Out of control. *Ethics, 116,* 285-301.

Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions. In J. Doris & the Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 246-272). Oxford: Oxford University Press.

Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics, 115,*

236-271.

Sommers, S.R., & Norton, M.I. (2006). Lay theories about White racists: What constitutes racism (and what doesn't). *Group Processes and Interpersonal Relations, 9,* 117-138.

Sripada, C. (in press). The "Deep Self" model and asymmetries in folk judgments about intentional action. *Philosophical Studies.*

Suhler, C., & Churchland, P. (2009). Control: Conscious and otherwise. *Trends in Cognitive Sciences, 13,* 341-347.

Tallis, F. (2002). *Hidden minds: A history of the unconscious.* New York: Arcade.

Taslitz, A.E. (2007). Forgetting Freud: The courts' fear of the subconscious in date rape (and other) cases. *Boston University Public Interest Law Review, 17,* 145-194.

Uhlmann, E.L., Pizarro, D.A., & Bloom, P. (2008). Varieties of unconscious social cognition. *Journal for the Theory of Social Behaviour, 38,* 293-322.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct* New York: Guilford.

Wigley, S. (2007). Automaticity, consciousness, and moral responsibility. *Philosophical Psychology, 20,* 209-225.

Wilson, T.D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117-142.

Wilson, T.D., Lindsey, S., & Schooler, T.Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101-126.

Table 1.

Descriptive Statistics, Experiment 2

Variable	Unconscious *M*	Automatic *M*	Folk *M*
Intent	1.97 (*SD* = .84)	2.09 (*SD* = .76)	2.40 (*SD* = .76)
True Self	3.13 (*SD* = .92)	3.23 (*SD* = .82)	3.07 (*SD* = .98)
Anger	1.94 (*SD* = 1.04)	2.43 (*SD* = .96)	2.80 (*SD* = 1.16)
Disgust	2.06 (*SD* = 1.10)	2.21 (*SD* = 1.03)	2.77 (*SD* = 1.28)
Controllability	3.03 (*SD* = 1.00)	3.77 (*SD* = .63)	3.74 (*SD* = .76)

Table 2.

Correlations between Variables, Experiment 2

Variable	Moral Responsibility	Intent	True Self	Anger	Disgust	Controllability
Moral Responsibility	1.00					
Intent	.35**	1.00				
True Self	.31**	.34**	1.00			
Anger	.50**	.33**	.45**	1.00		
Disgust	.36**	.25*	.26*	.70**	1.00	
Controllability	.48**	.44**	.23*	.36**	.14*	1.00

**$p < .01$

*$p < .05$

Table 3.

Multiple Mediation Model, Experiment 2

Mediator	Indirect Effect of Theory via each Mediator (*B*)	*SE*	*Z*	*p*
Intent	-.02	.03	-.78	.44
True Self	-.01	.02	-.57	.57
Anger	-.13	.06	-2.09	.04
Controllability	-.14	.07	-2.10	.04

Table 4.

Reverse Single Mediation Models, Experiment 2

Variable	Indirect Effect of Theory via Moral Judgment (*B*)	*SE*	*Z*	*p*
Intent	-.21	.09	-2.46	.01
True Self	-.24	.10	-2.41	.02
Anger	-.40	.13	-3.04	.002
Controllability	-.27	.10	-2.82	.005

*$p < .01$

Notes

1. To ensure that scenario (e.g. renting vs. grading vs. hiring) did not make a difference, we conducted a 3 x 3 ANOVA with scenario as one factor and theory condition as the second factor. In Experiment 1 there was neither a main effect of scenario, $F(2, 83) = 1.32$, $p = .27$, nor was there an interaction between scenario and theory condition, $F(4, 83) = .83$, $p = .51$. In experiment 2, neither the main effect of scenario, $F(2, 83) = 1.02$, $p = .36$, nor the interaction was significant, $F(4, 83) = .71$, $p = .59$. The effects thus do not depend on the particular scenario used.
2. We also investigated the interaction between theory condition and participant race for the subset of our study containing only African-American and Caucasian participants. There was a significant interaction between theory condition and participant race, $F(2, 76) = 4.54$, $p = .01$. Caucasian participants generally displayed the same pattern as the overall sample ($F(2, 62) = 8.81$, $p < .001$), with the unconscious condition eliciting lower responsibility judgments than the automatic ($p = .09$) and folk ($p < .001$) conditions. However, African-American participants only showed a marginal main effect of theory condition, $F(2, 14) = 3.03$, $p = .08$. This was driven by higher responsibility attributions in the automatic condition compared to the folk ($p = .08$) and unconscious ($p = .19$) conditions. Given the low number of African-American participants, these results should be interpreted with caution, especially because the effects of participant race did not replicate in Experiment 2. In Experiment 2 there was no interaction between theory condition and participant race, $F(2, 78) = .13$, $p = .88$.

Figure Captions

Figure 1. Mean judgment of moral responsibility for the unconscious, automatic, and folk conditions. Error bars show SE of the mean.

Figure 2. Mean judgment of moral responsibility for the unconscious, automatic, and folk conditions. Error bars show SE of the mean.

Figure 3. Multiple mediation model for Study 2. Anger and perceived controllability each partially mediate the effect of unconscious theory on moral responsibility. Unconscious theory represents the binary variable where the unconscious theory condition is contrasted against the composite of the automatic and folk conditions. Asterisks represent regression coefficients that are significant at either the $p < .05$ (*) or $p < .01$ (**) levels.

Figure 4. Combined path diagram for reverse single mediation models in Study 2. Moral responsibility mediates the influence of unconscious theory on each of the four outcome variables. Unconscious theory represents the binary variable where the unconscious theory condition is contrasted against the composite of the automatic and folk conditions. Asterisks represent regression coefficients that are significant at the $p < .01$ (**) level.

Figure 1.

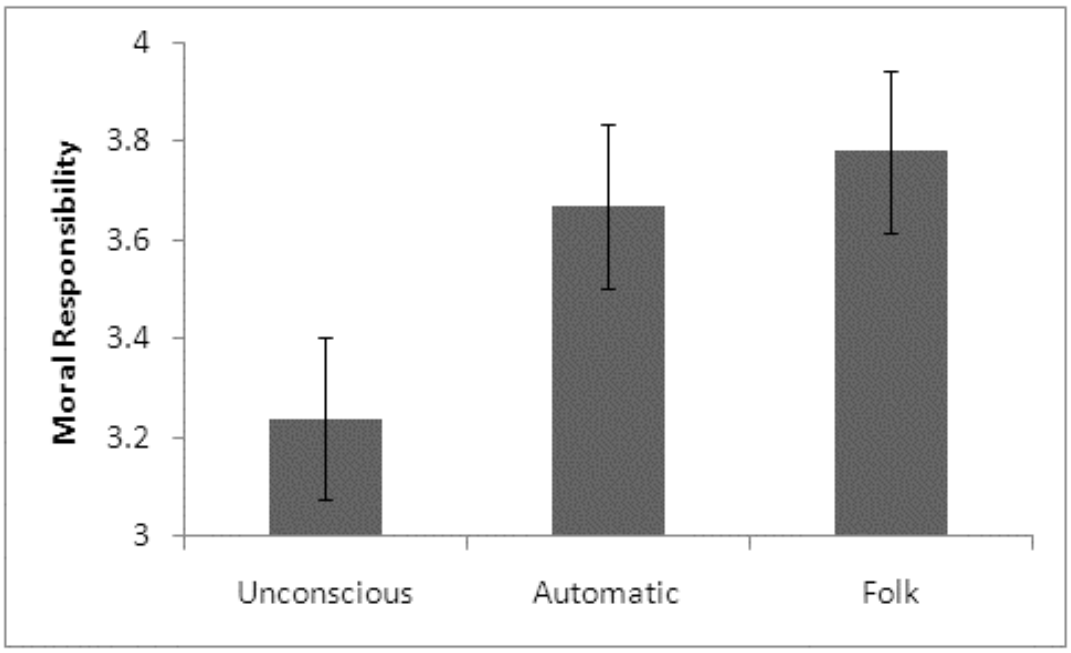

Moral Responsibility
4
3.8
3.6
3.4
3.2
3
Unconscious
Automatic
Folk

Figure 2.

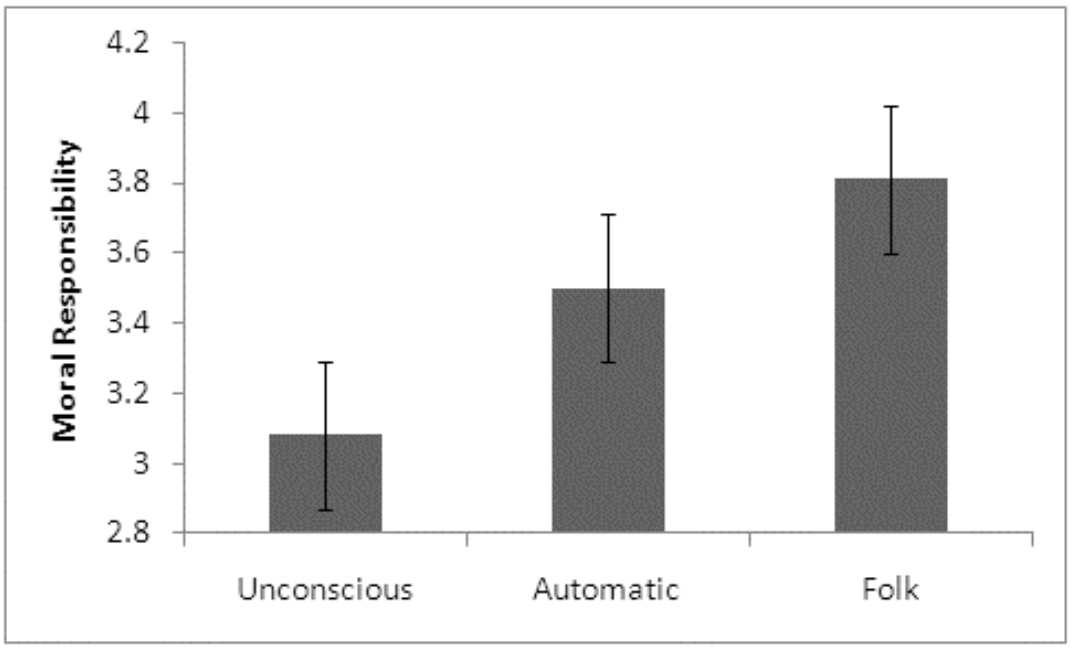

Moral Responsibility
4.2
4
3.8
3.6
3.4
3.2
3
2.8
Unconscious
Automatic
Folk

Figure 3.

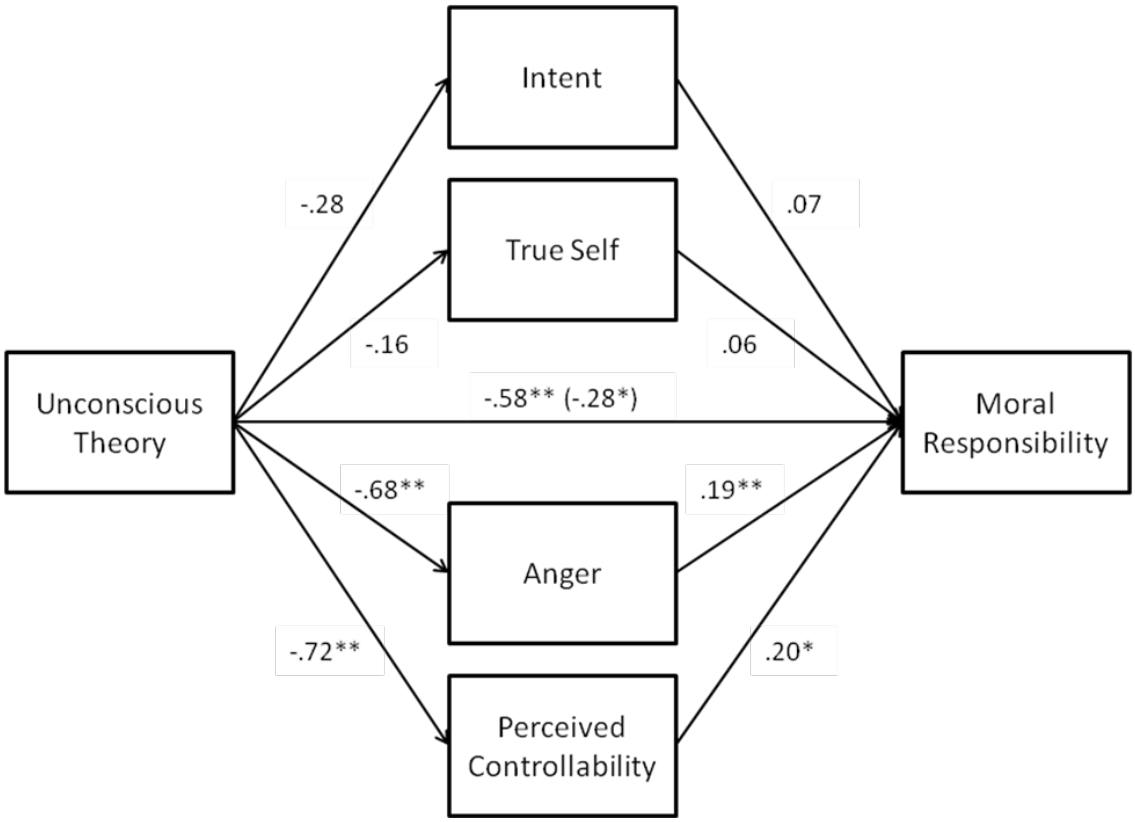
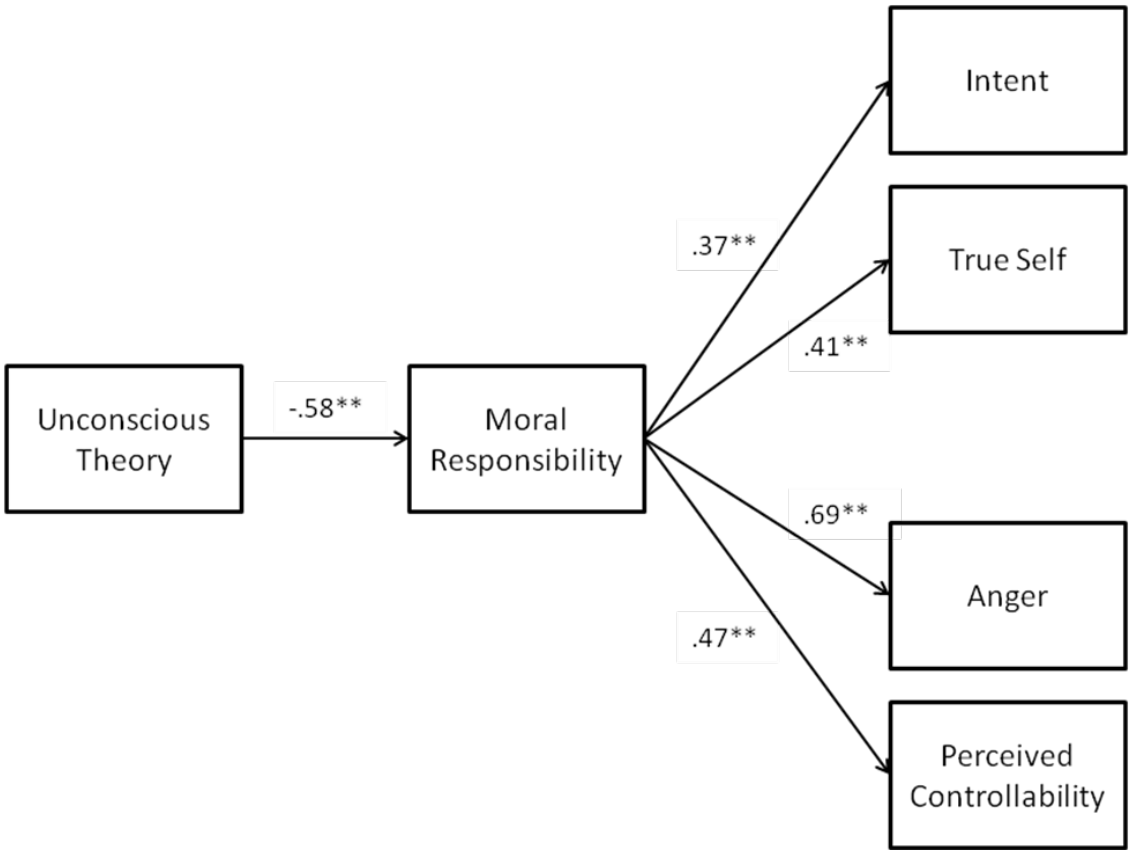

Intent
True Self
Anger
Perceived Controllability
Unconscious Theory
Moral Responsibility
-.28
-.16
-.68**
-.72**
.07
.06
.19**
.20*
-.58** (-.28*)

Figure 4.


Intent
True Self
.37**
.41**
Unconscious Theory
-.58**
Moral Responsibility
.69**
Anger
.47**
Perceived Controllability