Cognition 158 (2017) 224–241

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Implicit moral evaluations: A multinomial modeling approach

C. Daryl Cameron [a,b,*], B. Keith Payne [c], Walter Sinnott-Armstrong [d,e,f,g], Julian A. Scheffer [a], Michael Inzlicht [h]

[a] Department of Psychology, The Pennsylvania State University, United States
[b] Rock Ethics Institute, The Pennsylvania State University, United States
[c] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, United States
[d] Department of Philosophy, Duke University, United States
[e] Center for Cognitive Neuroscience, Duke University, United States
[f] Kenan Institute for Ethics, Duke University United States
[g] Duke Institute for Brain Sciences, Duke University, United States
[h] Department of Psychology, University of Toronto, Canada

ARTICLE INFO

Article history:
Received 19 October 2015
Revised 13 October 2016
Accepted 25 October 2016

Keywords:
Morality
Moral judgment
Ethics
Implicit cognition
Modeling

ABSTRACT

Implicit moral evaluations—i.e., immediate, unintentional assessments of the wrongness of actions or persons—play a central role in supporting moral behavior in everyday life. Yet little research has employed methods that rigorously measure individual differences in implicit moral evaluations. In five experiments, we develop a new sequential priming measure—the Moral Categorization Task—and a multinomial model that decomposes judgment on this task into multiple component processes. These include implicit moral evaluations of moral transgression primes (Unintentional Judgment), accurate moral judgments about target actions (Intentional Judgment), and a directional tendency to judge actions as morally wrong (Response Bias). Speeded response deadlines reduced Intentional Judgment but not Unintentional Judgment (Experiment 1). Unintentional Judgment was stronger toward moral transgression primes than non-moral negative primes (Experiments 2–4). Intentional Judgment was associated with increased error-related negativity, a neurophysiological indicator of behavioral control (Experiment 4). Finally, people who voted for an anti-gay marriage amendment had stronger Unintentional Judgment toward gay marriage primes (Experiment 5). Across Experiments 1–4, implicit moral evaluations converged with moral personality: Unintentional Judgment about wrong primes, but not negative primes, was negatively associated with psychopathic tendencies and positively associated with moral identity and guilt proneness. Theoretical and practical applications of formal modeling for moral psychology are discussed.

1. Introduction

Imagine that you open your morning newspaper and read that a school of children overseas has been bombed as part of a terrorist attack. Innocent children were killed, and it is likely that more will die as a result of the attack. Before you have engaged in reflective thought, you have an immediate flash of negative affect and a moral intuition: this is *wrong*. If someone asked you to justify your reaction, you might reason that the bombing violates the inherent dignity of human life, or you might appeal to the consequences that it has wrought. You also might wonder why you are even being asked this question, and whether it reveals a disturbing *lack* of morality that makes your conversation partner seem less trustworthy. Such implicit moral evaluations seem to be the beating heart of human morality, and it is important to know who has them and who does not.

In the current work, we use tools from cognitive science to develop a new measure and formal model of implicit moral evaluations. We define *implicit moral evaluations* as immediate, unintentional assessments of the moral wrongness of actions or persons. Prominent accounts of moral cognition verbally describe features of implicit moral evaluations (e.g., Greene, 2008; Haidt, 2001), but little research has formally specified their processing characteristics. We stipulate that implicit moral evaluations are strongly counter-intentional: not only can they arise spontaneously without any intention (i.e., weak unintentionality), but they can also influence moral judgments and behaviors *in opposition to* contrary intentions (i.e., strong unintentionality; cf. Moors & De Houwer, 2006). In order to provide a test of whether implicit moral evaluations are counter-intentional, we need to utilize measurement techniques that are designed to capture unintentional influence,

* Corresponding author.
E-mail address: cdc49@psu.edu (C.D. Cameron).

http://dx.doi.org/10.1016/j.cognition.2016.10.013
0010-0277/

as well as formal models that disentangle unintentional influences from other co-activated processes. Our research is the first to *a priori* formalize implicit moral evaluations with this conceptual precision and test whether they are strongly counter-intentional.

The present work advances the field of moral cognition by specifying the operating conditions of implicit moral evaluations. Moreover, this work speaks to the relationship between moral cognition and other, non-moral forms of evaluative processing. We stipulate that implicit moral evaluations are not merely reducible to affective evaluations. Instead, we suggest that implicit moral evaluations require both core affect (i.e., valence and arousal) and accessible conceptual knowledge about relevant moral rules (Cameron, Lindquist, & Gray, 2015; Nichols, 2004). What makes an implicit moral evaluation different from other implicit affective evaluations is this conceptual content related to morality. That said, we do not consider implicit moral evaluations to be a natural kind, categorically distinct from non-moral evaluations (Cameron et al., 2015). Instead, the difference is likely to be one of degree, with many of the same domain-general processes, such as affect, shared between moral and non-moral evaluations. This is also important given that what is deemed to be morally relevant can vary substantially across individuals (Graham et al., 2013), and within the same individual across different situations (Van Bavel, Xiao, Cunningham, 2012). Because moral relevance is idiographic and dynamic, it is likely that processes comprising implicit moral evaluations overlap substantially with non-moral cognition (see also Decety & Cowell, 2014; Young & Dungan, 2012).

In the current paper, we suggest that implicit moral evaluations are but one of many cognitive processes activated in response to morally relevant situations. Just as people can engage in unintentional moral evaluations, they can also intentionally morally evaluate the actions and characters of others. One theoretically novel aspect of our approach is that we suggest that intentional and unintentional forms of moral evaluation can operate simultaneously within the same moral context. Moreover, some people may be habitual "moralizers", biased to respond to most actions and people as morally wrong regardless of the situation or moral content involved. In order to dissociate implicit moral evaluations from intentional moral evaluations and response biases, we draw upon formal models. Although formal models are well used across cognitive science (for reviews, see Batchelder & Riefer, 1999; Erdfelder et al., 2009; Payne & Bishara, 2009; Riefer & Batchelder, 1988), they have been applied only sparingly in moral psychology to understand component processes of moral cognition (Crockett, 2016). Modeling variation in implicit moral evaluations—in a way that disentangles this latent process from others that may be activated in response to moral transgressions—can lead to more refined theoretical predictions about who will engage in moral behavior. The present research develops an implicit measure of moral judgment called the Moral Categorization Task, and a formal model for decomposing moral judgments on this task into their underlying component processes.

1.1. Developing an implicit measure of moral judgment

Despite the prevalence of claims about automaticity within moral psychology, little research has used tools from social cognition to model variability in implicit moral evaluations. Implicit measures—such as the implicit association test (Greenwald, McGhee, & Schwartz, 1998), affect misattribution procedure (Payne, Cheng, Govorun, & Stewart, 2005), and evaluative priming task (Fazio, Sanbonmatsu, Powell, & Kardes, 1986)—capture automatically activated evaluations while bypassing self-report (for review, see Wentura & Degner, 2010), and can predict explicit attitudes and behaviors (Cameron, Brown-Iannuzzi, & Payne, 2012; Greenwald, Uhlmann, Poehlman, & Banaji, 2009; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). Although limited, some work has attempted to use implicit measures to assess variation in implicit moral evaluations (e.g., implicit association test: Aquino & Reed, 2002; Cima, Tonnaer, & Lobbestael, 2007; Gray, MacCulloch, Smith, Morris, & Snowden, 2003; Luo et al., 2006; Perugini & Leone, 2009; affect misattribution procedure: Graham et al., 2016; Hofmann & Baumert, 2010).

Two concerns can be raised about prior uses of implicit measures of moral judgment. First, in the paradigms listed above, the target judgment is not moral judgment: it is the speed of relative associations (implicit association test) or the proportion of pleasant/unpleasant judgments (affect misattribution procedure). One goal of the current research is to develop an implicit measure that directly requires making moral judgments. Second, prior uses of implicit measures have taken a task dissociation approach, which assumes that implicit measures only capture automatic processes and explicit measures only capture controlled processes (cf. Payne, 2008). However, neither implicit nor explicit evaluation measures are "process-pure" (e.g., Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Payne, 2001): performance on both types of measures can result from automatic evaluations, executive control, or both. We present an alternative, multinomial modeling approach that does not make the task dissociation assumption, but rather dissociates multiple processes contributing to performance on the same task.

We developed a novel implicit measure of moral judgment: the Moral Categorization Task. On each of a series of trials, participants see two words in quick succession—a prime and a target—each of which can depict actions that are typically considered morally wrong (e.g., *murder*) or morally neutral (e.g., *baking*). Participants are instructed to judge whether the target word names a kind of act that is morally wrong or not, while avoiding the influence of the prime word. To allow for multinomial modeling, the judgment is binary (wrong vs. not wrong). Because the targets of judgment are normatively wrong or neutral, accuracy can be computed. To obtain sufficient errors for modeling, a response deadline is imposed on target judgment (e.g., Degner, 2009). This task is modeled on sequential priming tasks that have been used with process modeling, such as the weapon identification task (Payne, 2001) and affect misattribution procedure (Payne et al., 2005).

We constructed the Moral Categorization Task to capture immediate responses to moral transgressions, through their influence on categorization of acts as morally wrong or not wrong. Such reactions are among the most highly studied phenomena in moral psychology (though not typically under time pressure; Monin, Pizarro, & Beer, 2007), and are an important everyday feature of moral cognition that likely invoke different processes than those engaged by moral dilemma stimuli (e.g., reasoning to decide between competing moral principles; Monin et al., 2007). Given debate over the use of sacrificial dilemmas in moral psychology (Bartels & Pizarro, 2011; Bauman, McGraw, Bartels, & Warren, 2014; Greene, 2013; Kahane, 2015; Kahane, Everett, Earp, Farias, & Savulescu, 2015; Gray & Schein, 2012), we believe that our approach possesses a methodological advantage. Recent theory and evidence suggests that moral judgment operates by categorizing whether a particular act (e.g., *murder*) is a member of the set of acts that is immoral (DeScioli & Kurzban, 2013; Schein & Gray, 2015). This approach converges with neuroscience studies of moral evaluation: e.g., "everyday situations involving moral transgressions are likely to be evaluated on the basis of matching personal experiences and social knowledge stored in episodic and semantic memory" (Leuthold, Kunkel, Mackenzie, & Filik, 2015, p. 1021). In summary, the Moral Categorization Task is designed to capture a within-subjects priming effect on moral judgment, which can be formally modeled as resulting from individual differences in implicit moral evaluations, among other processes.