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/© 2016 Elsevier B.V. All rights reserved.

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.

1.2. Developing a formalized process model of moral judgment

Multinomial processing tree models formalize the latent cognitive processes that cause performance on a task (Batchelder & Riefer, 1999; Riefer & Batchelder, 1988; Sherman, Klauer, & Allen, 2011), with some of the most prominent examples in psychology including process dissociation (Jacoby, 1991; Payne, 2001, 2008; Payne & Cameron, 2014) and the Quadruple process model (Conrey et al., 2005). These models stipulate *a priori* how component processes interact to drive task performance, such that task performance can be used to estimate the probabilities of each process operating (Gawronski, Sherman, & Trope, 2014). Overall, the multinomial model is similar to a "Control-dominating" process dissociation model with a guessing parameter (Payne & Bishara, 2009). We selected this kind of model because of its wide use in social psychology and its ability to account for data on sequential priming tasks similar to the Moral Categorization Task (Bishara & Payne, 2009).

For the Moral Categorization Task, we stipulate three processes that drive judgment: Intentional Judgment (I), Unintentional Judgment (U), and Response Bias (B). Intentional Judgment is the ability to follow task instructions and intentionally evaluate whether target actions are morally wrong, operating with probability I. A person with high Intentional Judgment can follow task instructions and make accurate moral judgments about target actions. Intentional Judgment is similar to the Control parameter in process dissociation models.

Unintentional Judgment is the tendency to judge the morality of target actions in a prime-consistent manner, operating when Intentional Judgment fails with conditional probability $(1 - \mathrm{I}) \times \mathrm{U}$. A person with high Unintentional Judgment will have stronger implicit moral evaluations of prime actions. Once activated by prime actions, these should carry over and influence moral judgments about target actions. Unintentional Judgment is prime-consistent, with prime content reflecting two ends of a single continuum from typically wrong to typically not wrong: if the prime is a transgression then Unintentional Judgment should bias target judgments toward "wrong", and if the prime is a neutral action then Unintentional Judgment should bias target judgments toward "not wrong." In other words, Unintentional Judgment is not always defaulted to evaluate prime actions as wrong, but instead captures variation in the moral content of prime stimuli; as noted below, we estimate the tendency to always judge actions as wrong as a separate parameter (Response Bias). Unintentional Judgment is similar to the Automatic parameter in process dissociation models. We stipulate that Intentional and Unintentional Judgment differ on one feature—intentionality—because participants' task intentions are set to evaluate target actions while avoiding influence of prime actions. Prime influence directly counters task intentions, and so qualifies as strongly *counter*-intentional (Moors & De Houwer, 2006).

When both Intentional Judgment and Unintentional Judgment fail, Response Bias is the directional tendency to judge target actions as always wrong with conditional probability $(1 - \mathrm{I}) \times (1 - \mathrm{U}) \times \mathrm{B}$ or always not wrong with $(1 - \mathrm{I}) \times (1 - \mathrm{U}) \times (1 - \mathrm{B})$. When unable to judge the target action accurately, and in the absence of an implicit moral evaluation, a person might have a directional tendency to always judge target actions as morally wrong. By including Response Bias, we disentangle implicit moral evaluations from indiscriminate tendencies to moralize anything regardless of content.

Parameters are estimated from observed frequencies of responses in each condition of the task, using the equations in the process tree depicted in Fig. 1 and Appendix A. Each tree branch depicts the combination of processes stipulated to cause accurate or inaccurate responses on each trial type, such that accuracy on each trial type is the sum of probabilities across branches of the tree. For instance, on Wrong-Neutral trials, accurate responses occur when Intentional Judgment operates and Response Bias tends toward "not wrong" when Intentional Judgment and Unintentional Judgment fail, represented as the joint set of probabilities: $\mathrm{I} + (1 - \mathrm{I}) \times (1 - \mathrm{U}) \times (1 - \mathrm{B})$. An inaccurate response occurs when Intentional Judgment fails and there is prime-consistent Unintentional Judgment, and Response Bias tends toward "wrong" when Intentional Judgment and Unintentional Judgment fail, represented as the joint set of probabilities: $(1 - \mathrm{I}) \times \mathrm{U} + (1 - \mathrm{I}) \times (1 - \mathrm{U}) \times \mathrm{B}$. Parameter values are estimated through maximum likelihood estimation, iteratively changing values until optimal fit is reached between observed and model-predicted frequencies. Model fit is assessed with a likelihood-ratio G^2 statistic, with a non-significant result indicating model fit, and with the effect size *w*, with values less than 0.05 indicating acceptable fit (cf. Clerkin, Fisher, Sherman, & Teachman, 2014).

Testing specific hypotheses about parameters requires constraining them and comparing model fit against the baseline model. Effect sizes are derived for these comparisons using effect size *w*, with higher *w* values indicating greater influence of the manipulation on the parameter. If constraining any parameter to zero significantly reduces model fit, then it can be said to influence task performance. If constraining any parameter across conditions reduces model fit, then the manipulation influences the parameter in question. To examine individual differences, the model is estimated for each participant to derive individual-level parameter estimates, which are then correlated with individual scores on personality traits of interest.

1.3. Convergence with moral personality

If the current approach is capturing implicit moral evaluations, then Unintentional Judgment should converge with moral personality traits. We included four measures of moral personality across Experiments 1–4: psychopathic tendencies, moral identity, guilt proneness, and sacred value judgments. We mention these in the Method of each experiment, but examine relationships with parameter estimates in a later section (Section 7).

People with psychopathic tendencies are important for models of moral cognition because they can express normatively correct moral judgments despite acting immorally (Kiehl, 2008). Although psychopathic tendencies are not usually linked to changes in explicit moral judgment, they are linked to different patterns of brain activation (Aharoni, Antonenko, & Kiehl, 2011; Harenski, Harenski, Shane, & Kiehl, 2010; Marsh & Cardinale, 2012, 2014; for review, see Schaich Borg & Sinnott-Armstrong, 2013). Psychopathic individuals have reduced associations between harmful actions and negative affect (Blair, 2007; Gray et al., 2003), and exhibit callous affect in response to the suffering of others (Kiehl, 2008), which may in turn produce weaker implicit moral evaluations of moral transgressions.

Moral identity is the degree to which morality is important to a person's self-concept (Aquino & Reed, 2002). Internalized moral identity involves intrinsic importance to the self-concept, whereas symbolic moral identity involves extrinsic (i.e., reputational) importance. Internalized moral identity has been more consistently linked with moral behaviors (Aquino & Reed, 2002; Reed & Aquino, 2003). According to the social cognitive model of moral identity (Aquino, Reed, Freeman, Lim, & Felps, 2009), having stronger internalized moral identity makes information about morality more readily accessible, which may in turn produce stronger implicit moral evaluations of moral transgressions.

Guilt and shame proneness reflect the degree to which people experience guilt and shame about committing moral transgressions. Moral emotions are thought to inhibit commission of moral

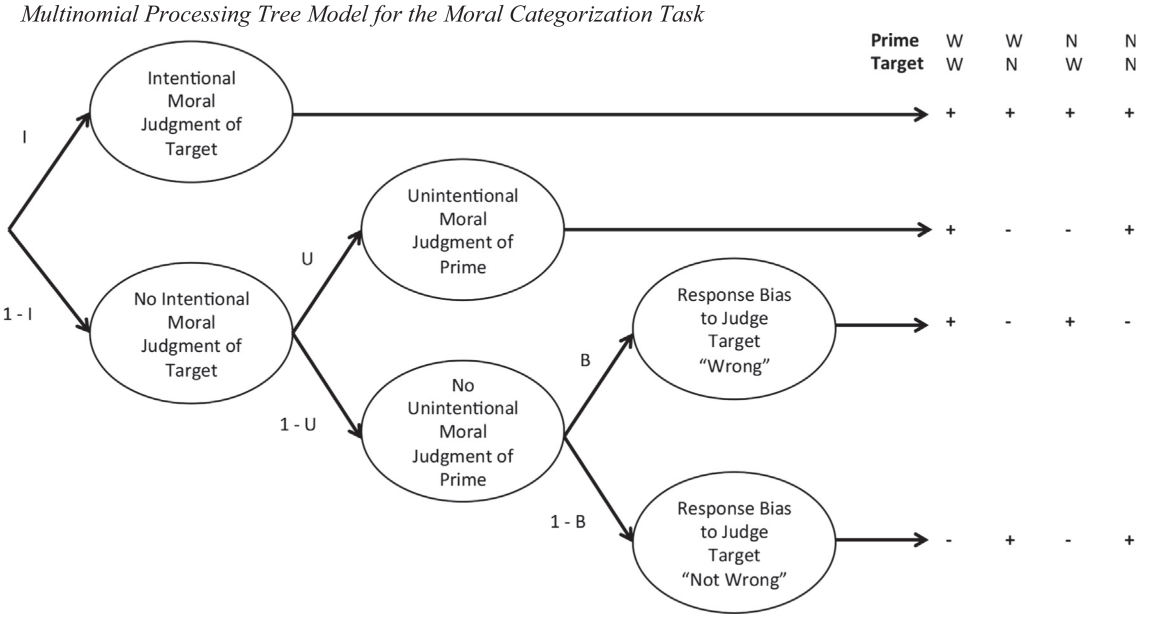
Multinomial Processing Tree Model for the Moral Categorization Task
Prime W W N N
Target W N W N
Intentional Moral Judgment of Target
I
+ + + +
No Intentional Moral Judgment of Target
1 - I
Unintentional Moral Judgment of Prime
U
+ - - +
No Unintentional Moral Judgment of Prime
1 - U
Response Bias to Judge Target "Wrong"
B
+ - + -
Response Bias to Judge Target "Not Wrong"
1 - B
- + - +

Fig. 1. Multinomial processing tree model for the moral categorization task. *Note.* The processing tree illustrates the formalized model of underlying component processes that lead to either accurate (+) or inaccurate (–) moral judgments of target actions on the Moral Categorization Task, for each of the four prime-target combinations ("W" = Wrong, "N" = Neutral). Each path name ("I", "U", "B") is the probability of that process operating. The tree depicts four possible paths: (a) Intentional Judgment—represented as I—drives the response, leading to accuracy on all trial types; (b) Unintentional Judgment, the tendency to morally judge target actions in a prime-consistent manner, drives response when Intentional Judgment fails, with probability $(1 - \mathrm{I}) \times \mathrm{U}$; (c) Response Bias to judge target actions as "wrong" when Intentional and Unintentional Judgment fail, with probability $(1 - \mathrm{I}) \times (1 - \mathrm{U}) \times \mathrm{B}$; and (d) Response Bias to judge target actions as "not wrong" when Intentional and Unintentional Judgment fail, with probability $(1 - \mathrm{I}) \times (1 - \mathrm{U}) \rtimes (1\text{-B})$.

transgressions (Bandura, 1999), and guilt proneness has been linked to moral behavior (Cohen, Panter, & Turan, 2012; Cohen, Wolf, Panter, & Insko, 2011). People high in guilt proneness have stronger aversion to transgressions—because they are more sensitive to their interpersonal costs—which may produce stronger implicit moral evaluations of transgressions. By contrast to guilt, shame is less consistently related to moral behavior (Cohen et al., 2011), and so we did not have a strong prediction that shame proneness would associate with implicit moral evaluations.

Finally, we included a measure of sacred value judgments, as the degree to which people would be willing to commit moral violations for money. To the degree that people perceive particular principles as sacred, they have stronger aversion to acts that violate those principles (Graham, Haidt, & Nosek, 2009; Tetlock, Kristel, Elson, Green, & Lerner, 2000), which should in turn produce stronger implicit moral evaluations of moral transgressions.

2. Experiment 1: Establishing the paradigm

In Experiment 1, we had three aims. First, we wanted to validate that we could capture a within-subjects priming effect in the Moral Categorization Task. Second, we used the multinomial model to disentangle influences of Intentional Judgment, Unintentional Judgment, and Response Bias. Third, we examined how a processing manipulation would influence these parameters. We predicted that imposing a fast response deadline would reduce Intentional Judgment, but not Unintentional Judgment or Response Bias.

2.1. Method

2.1.1. Participants

We recruited 65 undergraduate students (43 female, 19 male, 3 unreported) for course credit. Participants were randomly assigned to the 400-ms ($N = 23$), 500-ms ($N = 20$), or 800-ms ($N = 22$) deadline conditions. We used an outlier removal criterion for any participants whose overall error rates on the task were greater than 3 standard deviations above the sample mean, and only needed to implement this criterion in Experiment 5.

2.1.2. Materials and procedures

After being seated at individual computer workstations, participants completed the Moral Categorization Task. Participants were told:

> Each trial of this task will start with a fixation cross, +, that you should keep your eyes on. Then we will show you pairs of words flashed one after the other. Ignore the yellow words, which come first. Blue words will come second. Your job is to make a quick judgment of whether the blue words represent an action that is morally wrong. If the blue word represents an action that is morally wrong, press the M key on the keyboard. If the blue word does not represent an action that is morally wrong, press the Z key. Ignore the influence of the yellow words that come beforehand. Finally, please respond as fast as possible.

Word colors were chosen to be similar on brightness and were not counterbalanced; we have no *a priori* reason to believe that counterbalancing would significantly impact results. Each trial began with a fixation cross displayed in the center of the screen for 200 ms. The cross was followed by a prime word presented for 100 ms, followed by a blank screen for 75 ms, and then a target word which remained on screen until participants responded. The response deadline was constrained to 400, 500, or 800 ms depending on condition. We chose two fast deadlines because, as this was the first time we had used a deadline in this task, we did not know what speed would be sufficiently fast to cause a high error rate but still allow respondents to perform the task. If participants exceeded the response deadline, they saw a large red "X" and were instructed to "Please respond faster!" Participants completed 3 blocks of 40 trials each. Within each block, there were 10 trials per prime-target combination (Wrong-Wrong, Wrong-Neutral,

Neutral-Wrong, Neutral-Neutral), which were presented randomly. Prime and target words were selected to represent actions that are usually considered morally wrong or morally neutral, with neutral actions chosen to be lower on valence and arousal based on the Affective Norms for English Words Database (Bradley & Lang, 1999).

We focused on actions that involve harm to others, given that such concerns are deemed important to morality across populations (Graham et al., 2013; Gray, Schein, & Ward, 2014). There were two lists of morally wrong words (words that name kinds of acts that are morally wrong) and two lists of morally neutral words (words that name kinds of acts that are morally neutral). Word lists were counterbalanced so that one list of morally wrong words was used for prime words and the other list was used for target words, and similarly for the neutral word lists. Morally wrong items included: *murder*, *rape*, *racism*, *assault*, *lying*, *stealing*, *torture*, *betrayal*, *abuse*, *cheating*, *slaughter*, *genocide*, *terrorism*, *massacre*, *theft*, *cruelty*, *deception*, *molesting*, *killing*, and *robbery*. Morally neutral items included: *writing*, *farming*, *painting*, *baking*, *poetry*, *wondering*, *golf*, *leisure*, *modesty*, *agreement*, *travel*, *whistling*, *industry*, *reunion*, *nursing*, *listening*, *passage*, *watching*, *tennis*, and *exercise*.

2.1.2.1. Individual difference measures. Participants completed the Self-Reported Psychopathy Scale (Levenson, Kiehl, & Fitzpatrick, 1999), the Self-Importance of Moral Identity Scale (Aquino & Reed, 2002), the Guilt and Shame Proneness Scale (Cohen et al., 2011), and the Moral Foundations Sacredness Scale (Graham et al., 2009).

2.1.2.2. Exploratory measures and demographics. We also included the following exploratory measures: the Penn Inventory of Scrupulosity (Abramowitz, Huppert, Cohen, Tolin, & Cahill, 2002), Empathy Quotient Short-Form (Wakabayashi et al., 2006), and the Disgust Propensity and Sensitivity Scale (Olatunji, Cisler, Deacon, Connolly, & Lohr, 2007). Participants reported their gender, ethnicity, age, and political orientation (from 1 = *Extremely liberal* to 7 = *Extremely conservative*).

2.2. Results

2.2.1. Error rates

Responses were coded for accuracy, and all responses were used in analysis regardless of whether they exceeded response deadlines in order to maximize data used. We included the deadline to increase overall response speed; because the error rates are interpreted as means across conditions, missing a deadline on a particular trial does not invalidate that trial, as long as the deadline is effective in keeping the average speed within the desired range.

As predicted, a 3(Deadline: 400, 500, 800 ms) × 2(Prime: wrong, neutral) × 2(Target: wrong, neutral) mixed ANOVA revealed a Prime × Target interaction, $F(1,62) = 75.86$, $p < 0.001$, $\eta_p^2 = 0.55$, such that participants made more errors judging neutral targets after wrong vs. neutral primes, and more errors judging wrong targets after neutral vs. wrong primes. Table 1 displays error rates by prime, target, and deadline condition. There was a Prime × Target × Deadline interaction, $F(2,62) = 4.01$, $p = 0.023$, $\eta_p^2 = 0.12$, such that the Prime × Target interaction was stronger under the 400-ms deadline, $F(1,22) = 30.41$, $p < 0.001$, $\eta_p^2 = 0.58$, and 500-ms deadline, $F(1,19) = 33.22$, $p < 0.001$, $\eta_p^2 = 0.64$, than under the 800-ms deadline, $F(1,21) = 16.51$, $p = 0.001$, $\eta_p^2 = 0.44$. As expected, increasing time to respond reduced the impact of primes on responses.

2.2.2. Multinomial model

We conducted modeling analyses using MultiTree (Moshagen, 2010). From the behavioral data, observed frequencies of accurate and inaccurate responses were computed for each cell of the priming task and entered as data into MultiTree. The model was stipulated by writing equations that specify which response (accurate, inaccurate) on a given trial type (i.e., Wrong-Wrong, Wrong-Neutral, Neutral-Wrong, Neutral-Neutral) results from each path through the processing tree. The overall probability of a given response on a given trial type is then computed as the sum of each of these paths (displayed in Appendix A). Using maximum likelihood estimation, MultiTree iteratively changes parameter estimates to achieve optimal fit between observed and model-expected frequencies. Thus, the parameter estimates are logically defined in accordance with the model, and estimated relative to the behavioral data, such that the overall model fit and parameter estimates can be evaluated. For all experiments, we used random starting values for parameters and 5000 iterations (Moshagen, 2010). In this model fitting approach, a *p*-value greater than 0.05 indicates acceptable model fit, that the model can account for the observed behavioral frequencies. Establishing model fit is important in these analyses, because it validates that the underlying process model accounts for task performance. Failure to find model fit implies that the process model does not accurately account for task performance, and that alternative process model specifications may be preferable.

For each deadline condition, we estimated one parameter each for Intentional Judgment, Unintentional Judgment, and Response Bias. For all experiments, we estimated Unintentional Judgment

Table 1
Mean proportion of errors by prime, target, and deadline, Experiment 1.

Target	400 ms deadline condition			500 ms deadline condition			800 ms deadline condition		
	Wrong prime	Neutral prime	g	Wrong prime	Neutral prime	g	Wrong prime	Neutral prime	g
Wrong	0.22 (0.16)	0.34 (0.15)	−0.71**	0.21 (0.14)	0.29 (0.14)	−0.54**	0.08 (0.11)	0.12 (0.14)	−0.31*
Neutral	0.47 (0.20)	0.34 (0.24)	0.54**	0.43 (0.24)	0.32 (0.19)	0.49**	0.17 (0.14)	0.10 (0.12)	0.45**

Note. Standard deviations are in parentheses. Effect sizes are Hedges' g_{av} for simple effect comparisons between prime types within a target category. $^*p < 0.050$, $^{**}p < 0.010$.

Table 2
Parameter estimates, Experiment 1.

Parameter	400 ms deadline condition	500 ms deadline condition	800 ms deadline condition	Across conditions	
	Estimate [95% CI]	Estimate [95% CI]	Estimate [95% CI]	$\Delta G^2(2)$	*w*
Intentional	0.32 [0.28, 0.35]	0.38 [0.34, 0.42]	0.76 [0.74, 0.79]	443.96**	0.48
Unintentional	0.35 [0.26, 0.44]	0.31 [0.21, 0.42]	0.38 [0.23, 0.54]	0.64	0.02
Resp. Bias	0.51 [0.47, 0.54]	0.53 [0.48, 0.57]	0.46 [0.38, 0.53]	2.48	0.04

Note. $^\dagger p < 0.100$, $^*p < 0.050$, $^{**}p < 0.010$.

for wrong-prime trials while constraining Unintentional Judgment for neutral-prime trials to zero. An alternative modeling option would be to constrain Unintentional Judgment after wrong-prime and neutral-prime trials to be equal. We used the first approach for two reasons. First, variation in implicit moral evaluations about transgression primes was of primary theoretical interest. Second, in later experiments we aimed to compare Unintentional Judgment after wrong vs. negative primes, and so constraining across prime categories would be infeasible. Another alternative modeling option would be to estimate Unintentional Judgment as the tendency to rate target actions as morally wrong when any moral transgression as present—either as prime or target. Although such a model might be seen as capturing reactivity to moral stimuli, it does not capture the definition of implicit moral evaluations. We are specifically interested in how participants implicitly evaluate the moral content of primes, through their systematic and counter-intentional influence on target judgments. By contrast, moral content of target actions is defined as relevant input for target judgments, and its influence would not be counter to task intentions. Thus, modeling implicit moral evaluation as a tendency to rate any target as immoral when any transgression is present (as prime or target) does not fit the conceptual definition of implicit moral evaluations.

The model fit the data well, $G^2(3) = 1.52$, $p = 0.677$, $w = 0.03$. Table 2 displays parameter estimates by deadline condition. As predicted, imposing fast deadlines decreased Intentional Judgment, $\Delta G^2(2) = 443.96$, $p < 0.00001$, $w = 0.48$, but did not influence Unintentional Judgment, $\Delta G^2(2) = 0.64$, $p = 0.725$, $w = 0.02$, or Response Bias, $\Delta G^2(2) = 2.48$, $p = 0.289$, $w = 0.04$.

2.3. Discussion

Using a novel implicit measure, Experiment 1 revealed that participants displayed consistent biases in their moral judgments: when primed with words denoting morally wrong actions such as *murder*, they were more likely to mistakenly judge morally neutral target actions as wrong. To understand this systematic pattern of mistakes, we applied a multinomial model to understand processes underpinning moral judgments. Performance depended on how accurately participants could morally judge target actions (Intentional Judgment), implicit moral evaluations of prime actions (Unintentional Judgment), and a directional tendency to judge target actions as wrong (Response Bias). Imposing fast response deadlines reduced Intentional Judgment but not Unintentional Judgment or Response Bias, suggesting that Intentional Judgment has operating conditions typically ascribed to controlled processing.

3. Experiment 2: The role of negative affect

In Experiment 2, we aimed to replicate Experiment 1 while addressing an alternative explanation of task performance. Because the prior experiment did not include non-moral negative content, it is possible that the morally wrong primes influenced moral judgment due to negative affect, and not due to the moral content of the primes themselves. Although negative affect is likely to be a component of implicit moral evaluations, conceptual content about morality should play an additional role (Cameron et al., 2015; Nichols, 2004). We added non-moral negative stimuli that were matched on valence and arousal. If the Moral Categorization Task is about morality, rather than negative affect alone, then transgression primes should influence moral judgment above and beyond non-moral negative primes.

3.1. Method

3.1.1. Participants

We recruited 100 undergraduates (68 female, 32 male) for course credit.

3.1.2. Materials and procedures

3.1.2.1. Moral categorization task. Participants completed the same task as in Study 1. The task consisted of two blocks of 180 trials each. Within each block, there were 20 trials per prime by target combination. Response deadline was held constant at 500 ms. Morally wrong items were: *murder*, *rape*, *abuse*, *theft*, *killing*, *assault*, *torture*, *betrayal*, *slaughter*, and *terrorism*. Morally neutral items were: *writing*, *golf*, *baking*, *agreement*, *leisure*, *painting*, *tennis*, *farming*, *modesty*, and *exercise*. Non-moral negative items were: *cancer*, *disaster*, *trauma*, *distress*, *pain*, *rabies*, *accident*, *nightmare*, *bankruptcy*, and *rejection*. Morally wrong and non-moral negative items were chosen to be matched on valence ($M_{\text{Moral}} = 1.86$, $SD_{\text{Moral}} = 0.23$; $M_{\text{Negative}} = 1.73$, $SD_{\text{Negative}} = 0.27$; on 9-point scale with 1 representing most negative valence and 9 representing most positive valence) and arousal ($M_{\text{Moral}} = 6.45$, $SD_{\text{Moral}} = 0.42$, $M_{\text{Negative}} = 7.05$, $SD_{\text{Negative}} = 0.52$; on 9-point scale with 1 representing lowest arousal and 9 representing highest arousal) based upon the Affective Norms for English Words database (Bradley & Lang, 1999).

3.1.2.2. Individual difference measures. Participants completed the four moral personality measures: the Self-Reported Psychopathy Scale, the Self-Importance of Moral Identity Scale, the Guilt and Shame Proneness Scale, and the Moral Foundations Sacredness Scale.

3.1.2.3. Exploratory measures and demographics. Participants also completed 10 sacrificial dilemmas (the congruent and incongruent versions of the abortion, baby, car, torture, and vaccine dilemmas from Conway & Gawronski, 2013). For each stimulus used in the Moral Categorization Task, participants rated negative valence, positive valence, arousal, and moral wrongness. Lastly, participants rated a series of controversial moral issues, such as *abortion* and *euthanasia*, on negative and positive valence, arousal, and moral conviction. The sacrificial dilemmas and stimulus evaluation measures were exploratory and will not be discussed further. Finally, participants reported gender, ethnicity, age, religiosity, political orientation, and socioeconomic status using the MacArthur ladder (Adler & Ostrove, 1999).

3.2. Results

3.2.1. Error rates

For negative target trials, we coded responses as inaccurate if participants judged targets as wrong. As predicted, a 3(Prime: wrong, negative, neutral) × 3(Target: wrong, negative, neutral) within-subjects ANOVA revealed a Prime × Target interaction, $F(4,396) = 22.08$, $p < 0.001$, $\eta_p^2 = 0.18$. Table 3 displays error rates by prime and target condition. To understand this interaction, we examined the influence of prime type on judgment accuracy for each target type.

Prime type influenced judgments of neutral targets, $F(2,198) = 28.40$, $p < 0.001$, $\eta_p^2 = 0.22$, with greater errors after wrong vs. neutral primes, negative vs. neutral primes, and wrong vs. negative primes. Prime type influenced judgment of wrong targets, $F(2,198) = 10.57$, $p < 0.001$, $\eta_p^2 = 0.10$, with greater errors after neutral vs. wrong primes and neutral vs. negative primes. Lastly, prime type influenced judgment of negative targets, $F(2,198) = 18.23$,

Table 3
Mean proportion of errors by prime and target, Experiment 2.

Target	Wrong prime	Negative prime	Neutral prime	$g_{\text{WrongNegative}}$	$g_{\text{WrongNeutral}}$	$g_{\text{NegativeNeutral}}$
Wrong	0.22 (0.17)	0.24 (0.17)	0.27 (0.16)	−0.12	-0.30^{**}	-0.18^{*}
Negative	0.58 (0.28)	0.56 (0.29)	0.52 (0.29)	0.05	0.19^{**}	0.14^{**}
Neutral	0.41 (0.25)	0.37 (0.25)	0.33 (0.24)	0.12^{**}	0.29^{**}	0.17^{**}

Note. Standard deviations are in parentheses. Effect sizes are Hedges' g_{av} for simple effect comparisons between prime types within a target category. $^{\dagger}p < 0.100$, $^{*}p < 0.050$, $^{**}p < 0.010$.

Table 4
Parameter estimates, Experiment 2.

Parameter	Wrong prime	Negative prime	Neutral prime	Across conditions	
	Estimate [95% CI]	Estimate [95% CI]	Estimate [95% CI]	$\Delta G^2(1)$	w
Intentional	0.38 [0.37, 0.39]	0.38 [0.37, 0.39]	0.38 [0.37, 0.39]	–	–
Unintentional	0.21 [0.17, 0.26]	0.12 [0.08, 0.17]	0.00 (constant)	12.62^{**}	0.06
Resp. Bias	0.55 [0.53, 0.57]	0.55 [0.53, 0.57]	0.55 [0.53, 0.57]	–	–

Note. Intentional Judgment and Response Bias parameters constrained to be equal across prime conditions, and Unintentional Judgment constrained to zero in the Neutral prime condition. $^{\dagger}p < 0.100$, $^{*}p < 0.050$, $^{**}p < 0.010$.

$p < 0.001$, $\eta_p^2 = 0.16$, with greater errors after wrong vs. neutral primes and negative vs. neutral primes.

3.2.2. Multinomial model

As documented in Appendix A, parameters for negative-prime trials were computed identically to the parameters for wrong-prime trials. Prime-consistent responses to negative primes were expected to bias judgments of target actions toward "wrong", as with wrong primes. Thus, a comparison between Unintentional Judgment parameters for wrong-prime and negative-prime trials captures the difference across these prime conditions in the tendency to judge target actions as "wrong". Implicit moral evaluations should be stronger toward transgression primes (e.g., *murder*) than toward negative non-moral primes (e.g., *cancer*).

We estimated one Intentional Judgment parameter, two Unintentional Judgment parameters (for wrong and negative primes), and one Response Bias parameter. This initial model did not fit, $G^2(5) = 810.76$, $p < 0.00001$, $w = 0.45$. We noted that error rates on negative-target trials were substantially higher than on other trials. Participants may have had difficulty separating negative affective responses to targets from the focal task of morally judging targets, thus not performing as instructed and leading to higher errors. Because participants responded very differently on these target trials than on other target trials, we excluded these, leading to adequate fit, $G^2(2) = 2.43$, $p = 0.297$, $w = 0.02$. Table 4 displays parameter estimates. As predicted, Unintentional Judgment was stronger toward wrong primes than negative primes, $\Delta G^2(1) = 12.62$, $p = 0.0004$, $w = 0.06$, indicating that participants had a stronger tendency to judge target actions as wrong after being primed with moral transgressions. Were the effect due merely to incongruence created by negative affect, such specificity would not be observed.

3.3. Discussion

Experiment 2 built upon prior findings in multiple ways. First, the priming effect for morally wrong and neutral primes and targets directly replicated prior findings. Second, the addition of non-moral negative primes revealed that the original priming effect is not reducible to negative affect. Although negative affective primes such as *cancer* influenced performance on the priming task, this effect was weaker. Modeling revealed that Unintentional Judgment was stronger after wrong primes than negative primes, suggesting that implicit moral evaluations are stronger when they involve both negative affect and conceptual knowledge about morality.

4. Experiment 3: Idiographic moral judgments

Another alternative explanation is that the task is not capturing participants' personally endorsed implicit evaluations, but rather stereotypes about which actions society deems acceptable or unacceptable. To account for this possibility, in Experiment 3 we instructed participants to self-generate moral transgressions that they had strong personal opposition to. We expected that participant-generated moral transgressions would operate similarly to the researcher-generated moral transgressions used in previous experiments. Such results would generalize our approach to idiographic moral evaluations, which are of growing interest in moral psychology (Gray & Keeney, 2015; Meindl & Graham, 2014; Skitka, 2010).

4.1. Method

4.1.1. Participants

We recruited 58 undergraduates (38 female, 19 male, 1 unreported) for course credit.

4.1.2. Materials and procedures

4.1.2.1. Moral categorization task. At the beginning of the task, participants were instructed:

> Before getting started, please enter in an action or issue that YOU personally believe is morally wrong. This should be an action or issue that violates your core moral beliefs, values, and convictions. In order to provide valid data, please type in 1 or 2 words to describe this action or issue, and make sure it is in ALL CAPITAL LETTERS.

After typing in a response, participants were instructed:

> Next, please enter in a DIFFERENT action or issue that YOU personally believe is morally wrong. This should be an action or issue that violates your core moral beliefs, values, and convictions. In order to provide valid data, please type in 1 or 2 words to describe this action or issue, and make sure it is in ALL CAPITAL LETTERS.

The first idiographic moral issue was always used as the prime on idiographic-prime trials and the second idiographic moral issue was always used as the target on idiographic-target trials, such that the

Table 5
Mean proportion of errors by prime and target, Experiment 3.

Target	Wrong prime	Idio. prime	Negative prime	Neutral Prime	$g_{\text{Wrong Idio}}$	$g_{\text{Wrong Neg}}$	$g_{\text{Wrong Neut}}$	$g_{\text{Idio Neg}}$	$g_{\text{Idio Neut}}$	$g_{\text{Neg Neut}}$
Wrong	0.19 (0.17)	0.17 (0.17)	0.19 (0.17)	0.23 (0.16)	0.11	0.02	−0.21	−0.09	-0.33^{**}	−0.23
Idio.	0.15 (0.16)	0.15 (0.17)	0.17 (0.17)	0.21 (0.14)	−0.05	−0.11	-0.40^{**}	−0.06	-0.32^{**}	-0.27^{*}
Neg.	0.64 (0.26)	0.66 (0.25)	0.64 (0.26)	0.59 (0.26)	−0.08	0.00	0.17	0.08	0.25^{*}	$0.17^{\dagger}$
Neut.	0.49 (0.23)	0.49 (0.22)	0.45 (0.24)	0.39 (0.25)	0.01	$0.15^{\dagger}$	0.39^{**}	0.14	0.39^{**}	0.25^{*}

Note. Standard deviations are in parentheses. Effect sizes are Hedges' g_{av} for simple effect comparisons between prime types within a target category. $^{\dagger}p < 0.100$, $^{*}p < 0.050$, $^{**}p < 0.010$.

Table 6
Parameter estimates, Experiment 3.

Parameter	Wrong prime	Idiographic prime	Negative prime	Neutral prime	Across conditions	
	Estimate [95% CI]	Estimate [95% CI]	Estimate [95% CI]	Estimate [95% CI]	$\Delta G^2(2)$	w
Intentional	0.35 [0.34, 0.36]	0.35 [0.34, 0.36]	0.35 [0.34, 0.36]	0.35 [0.34, 0.36]	–	–
Unintentional	0.27 [0.21, 0.33]	0.28 [0.22, 0.34]	0.20 [0.14, 0.27]	0.00 (constant)	6.18^{*}	0.06
R. Bias-WN	0.63 [0.61, 0.65]	0.63 [0.61, 0.65]	0.63 [0.61, 0.65]	0.63 [0.61, 0.65]	–	–
R. Bias-Idio	0.68 [0.66, 0.70]	0.68 [0.66, 0.70]	0.68 [0.66, 0.70]	0.68 [0.66, 0.70]	–	–

Note. Intentional Judgment and the two Response Bias parameters (R. Bias-WN is for wrong-target and neutral-target trials, R. Bias-Idio is for idiographic-target trials) constrained to be equal across prime conditions, and Unintentional Judgment constrained to zero in the Neutral prime condition. $^{\dagger}p < 0.100$, $^{**}p < 0.010$, $^{*}p < 0.050$.

same idiographic stimulus was presented repeatedly. Because of this, the idiographic stimuli (one for prime, one for target) were repeated more frequently than the other stimulus categories. Many of the idiographic issues generated by participants were similar to the non-controversial moral stimuli, with the most common items being murder ($n = 18$), rape ($n = 15$), cheating ($n = 10$), and lying ($n = 10$). Participants completed 2 blocks of 320 trials each, with 40 trials per prime-target combination. The other word lists and response deadline were identical to Experiment 2.

4.1.2.2. Individual difference measures. Participants completed the four moral personality measures: the Self-Reported Psychopathy Scale, the Self-Importance of Moral Identity Scale, the Guilt and Shame Proneness Scale, and the Moral Foundations Sacredness Scale.

4.1.2.3. Exploratory measures and demographics. Participants completed exploratory measures of social distancing (Skitka, Bauman, & Sargis, 2005) and valence, arousal, and wrongness ratings for each stimulus in the Moral Categorization Task. Finally, participants reported gender, ethnicity, age, religiosity, political orientation, and subjective socioeconomic status.

4.2. Results

4.2.1. Error rates

Four participants did not provide priming data. We also excluded data for five participants who failed to follow instructions in generating their two idiographic transgression stimuli (for instance, by entering a non-word or entering the same stimulus twice, by reporting an act with a moral judgment, e.g., "STEALING IS WRONG", or an act with a justification for a judgment, e.g., "RAPE-VIOLATES CONSENT"); re-including these participants did not change results from what is reported below. As predicted, a 4 (Prime: wrong, idiographic, negative, neutral) × 4(Target: wrong, idiographic, negative, neutral) within-subjects ANOVA revealed a Prime × Target interaction, $F(9,432) = 11.10$, $p < 0.001$, $\eta_p^2 = 0.19$. Table 5 displays error rates by prime and target. To understand this interaction, we examined the influence of prime type on judgment for each type of target.

Prime type influenced judgment of neutral targets, $F(3,144) = 14.96$, $p < 0.001$, $\eta_p^2 = 0.24$, such that participants made more errors after wrong vs. neutral primes, idiographic vs. neutral primes, and negative vs. neutral primes, and marginally more errors after wrong vs. negative primes. Prime type influenced judgment of wrong targets, $F(3,144) = 5.16$, $p = 0.002$, $\eta_p^2 = 0.10$, with more errors after neutral vs. idiographic primes. Prime type influenced judgments about idiographic targets, $F(3,144) = 10.17$, $p < 0.001$, $\eta_p^2 = 0.18$, with more errors after neutral vs. wrong primes, neutral vs. idiographic primes, and neutral vs. negative primes. Lastly, prime type influenced judgments of negative targets, $F(3,144) = 5.47$, $p = 0.001$, $\eta_p^2 = 0.10$, with more errors after idiographic vs. neutral primes, and marginally more errors after negative vs. neutral primes.

The goal in this experiment was to test whether participant-generated idiographic stimuli were treated similarly to researcher-generated wrong stimuli. We compared whether wrong and idiographic primes were similarly distinct from neutral primes, and whether wrong and idiographic primes were similarly distinct from negative primes, separately for each type of target (wrong, idiographic, negative, neutral). To the degree that wrong and idiographic primes exert comparable influence relative to these contrast prime categories (negative, neutral), they can be inferred to be similar. As seen in Table 5, wrong and idiographic primes each elicited fewer errors than neutral primes, similarly on both wrong-target and idiographic-target trials. On neutral-target trials, wrong and idiographic primes both elicited more errors than neutral primes. Wrong and idiographic primes did not elicit different amounts of errors compared to negative primes, similarly on both wrong-target and idiographic-target trials. In all but one instance the descriptive directions of the wrong and idiographic prime vs. negative prime comparisons were negative (i.e., wrong and idiographic primes eliciting fewer errors than negative primes); furthermore, we hesitate to interpret any changing signs of these comparisons given that they were non-significant. In summary, when examining the relevant comparisons across prime categories (wrong and idiographic vs. neutral, wrong and idiographic vs. negative), the majority of these comparisons exhibit a similar direction to each other, suggesting that wrong and idiographic prime and target stimuli are being treated similarly by participants.

4.2.2. Multinomial model

For modeling analyses, we excluded negative-target trials as in Experiment 2. We estimated a baseline model with one Intentional Judgment parameter, three Unintentional Judgment parameters (after wrong, idiographic, and negative primes, with Unintentional

Judgment after neutral primes set to zero), and two Response Bias parameters. We estimated a unique Response Bias parameter for idiographic-target trials because this target stimulus was presented more frequently and thus may have elicited a stronger bias to judge target actions as wrong compared to other target types. This model fit the data adequately, $G^2(6) = 11.65$, $p = 0.070$, $w = 0.08$. Table 6 displays parameter estimates by prime condition. As expected, prime condition influenced Unintentional Judgment, $\Delta G^2(2) = 6.18$, $p = 0.045$, $w = 0.06$. Unintentional Judgment did not differ after wrong primes and idiographic primes, $\Delta G^2(1) = 0.17$, $p = 0.684$, $w = 0.01$, but was marginally weaker after negative primes than wrong primes, $\Delta G^2(1) = 3.70$, $p = 0.054$, $w = 0.04$, and was weaker after negative primes than idiographic primes, $\Delta G^2(1) = 5.41$, $p = 0.020$, $w = 0.05$. As expected, Response Bias was stronger for idiographic targets than for wrong and neutral targets, $\Delta G^2(1) = 18.00$, $p < 0.0001$, $w = 0.10$.

4.3. Discussion

Experiment 3 revealed that participant-generated moral transgressions operated similarly to the researcher-generated moral transgressions used in previous experiments, influencing judgment accuracy of neutral targets in a comparable manner. This result suggests that the current approach can be used to assess idiographic moral evaluations.

5. Experiment 4: Neurophysiology of moral judgment

In Experiment 4, we used event-related potentials to validate processing characteristics of Intentional Judgment. Event-related potentials allow temporally precise measurement of neural activity during stimulus presentation and behavioral response (Amodio, Bartholow, & Ito, 2014). We focused on the error-related negativity (Gehring, Goss, Coles, Meyer, & Donchin, 1993), a negative deflection on the electroencephalogram (EEG) that occurs within 100 ms of committing an error on a task, and that is thought to be produced by the anterior cingulate cortex (Dehaene, Posner, & Tucker, 1994). The error-related negativity is typically thought to indicate discrepancy between expected (correct) and actual (incorrect) outcomes (Holroyd & Coles, 2002), or conflict monitoring more generally (Yeung, Botvinick, & Cohen, 2004). The ERN may also reflect affective responses to errors (Cavanaugh & Shackman, in press; Gehring & Willoughby, 2002; Luu, Tucker, Derryberry, Reed, & Poulsen, 2003), and be a "distress signal" when performance does not meet expectations (Bartholow et al., 2005, p. 41; Proudfit, Inzlicht, & Mennin, 2013). In the Moral Categorization Task, we expected that Intentional Judgment would associate most strongly with error-related negativity on incongruent trials, particularly those that require an incompatible moral response to be inhibited (i.e., Wrong-Neutral trials).

5.1. Method

5.1.1. Participants

We recruited 58 college undergraduates from a large, urban Canadian campus (38 female, 20 male) for course credit.

5.1.2. Materials and procedures

5.1.2.1. Individual difference measures. In an online session prior to the experiment, participants completed three moral personality measures: the Self-Reported Psychopathy Scale, the Self-Importance of Moral Identity Scale, and the Guilt and Shame Proneness Scale.

5.1.2.2. Exploratory measures and demographics. In the online session, participants also completed the following exploratory measures: the Narcissistic Personality Inventory (Raskin & Terry, 1988), Mach-IV inventory of Machiavellianism (Christie & Geis, 1970), Free Will and Determinism Plus Scale (FAD-Plus; Paulhus & Carey, 2011), and Internal Control Index (Duttweiler, 1984). Finally, participants reported gender, ethnicity, age, current and childhood household income, subjective socioeconomic status, English as first language, and handedness.

5.1.2.3. Moral categorization task. Participants completed 3 blocks of 100 trials each. Within each block, there were 20 trials each for the 2(Prime: Wrong, Neutral) × 2(Target: Wrong, Neutral) design, as well as a set of 20 Negative-Neutral trials. To reduce task duration, we did not include a Negative-Wrong cell. Word lists and timing were identical to Experiment 2, and response deadline was held constant at 450 ms.

5.1.2.4. Neurophysiological recording and processing. EEG activity during the Moral Categorization Task was recorded using a stretch Lycra cap (Electro-Cap International, Eaton, OH) embedded with 32 tin electrodes. Recordings were taken using a midline recording montage, with concentration on the Fz, FCz, Cz, CPz, Pz, and Oz electrodes while using the mastoid M1 and M2 as reference electrodes and the vertical oculogram VEOG+ and VEOG− to help filter for eyeblinks. Recordings were digitized at 512 Hz using ASALab4 acquisition software (Advanced Neuro Technology B.V., Enschede, The Netherlands) with an average-electrode reference and forehead ground. EEG data was analyzed using Brain Vision Analyzer 2.0 (Brain Products GmbH, Munich, Germany). EEG data was re-referenced to the average of the two mastoid channels (M1 and M2), corrected for vertical-oculogram artifacts (Gratton, Coles, & Donchin, 1983), and digitally filtered offline between 0.1 and 30 Hz (24 dB IIR filter). An automatic procedure provided in Brain Vision Analyzer was used to reject the artifacts. The criteria applied were a voltage step of no more than 18 μV between sample points, a voltage difference of 150 μV within 150 ms intervals, voltages above 85 μV and below −85 μV, and a maximum voltage difference of less than 1.00 μV within 100 ms intervals. These intervals were rejected from individual channels in each trial. ERP epochs were time-locked to responses and created by examining continuous EEG from 200 ms pre-response and 800 ms post response, with −200 ms to 0 ms used for baseline correction. ERP averages were created separately for each prime-target condition (i.e., Wrong-Neutral, Wrong-Wrong, Neutral-Neutral, Neutral-Wrong, Negative-Neutral). Data for these conditions were averaged within participants independently for correct (correct-related negativity) and incorrect responses (error-related negativity), and then grand-averaged within the respective conditions. These were

Table 7
Mean proportion of errors by prime and target, Experiment 4.

Target	Wrong prime	Negative prime	Neutral prime	g_{WrongNeg}	$g_{\text{WrongNeut}}$	g_{NegNeut}
Wrong	0.30 (0.18)	–	0.39 (0.19)	–	-0.45^{**}	–
Neutral	0.33 (0.17)	0.29 (0.15)	0.24 (0.16)	0.24^{*}	0.51^{**}	0.30^{**}

Note. Standard deviations are in parentheses. Effect sizes are Hedges' g_{av} for simple effect comparisons between prime types within a target category. $^{\dagger}p < 0.100$, $^{*}p < 0.050$, $^{**}p < 0.010$.

defined at frontocentral sites (Fz, FCz, Cz), examining the mean amplitudes within a 0–100 ms time window.

5.2. Results

5.2.1. Error rates

This analysis focused on the 2 × 2 of wrong and neutral primes and targets, and did not incorporate Negative-Neutral trials. Because we did not include Negative-Wrong target trials, the design was unbalanced and so we excluded this set of trials. As predicted, a 2(Prime: wrong, neutral) × 2(Target: wrong, neutral) within-subjects ANOVA revealed a Prime × Target interaction, $F(1,57) = 35.63$, $p < 0.001$, $\eta_p^2 = 0.39$, such that participants made more errors judging neutral targets after wrong vs. neutral primes, and more errors judging wrong targets after neutral vs. wrong primes. Table 7 displays error rates by prime and target condition.

5.2.2. Multinomial model

In this experiment, we estimated four parameters: one Intentional Judgment parameter, two Unintentional Judgment parameters (for wrong primes and negative primes, with Unintentional Judgment toward neutral primes set to zero), and one Response Bias parameter. This model fit the data well, $G^2(1) = 0.04$, $p = 0.840$, $w = 0.00$. Table 8 displays parameter estimates. Replicating previous experiments, Unintentional Judgment was stronger after wrong primes than negative primes, $\Delta G^2(1) = 14.15$, $p = 0.0002$, $w = 0.06$.

5.2.3. Error-related negativity

Because there were 5 within-subjects conditions of the task, we needed a sufficient number of error trials in each condition. Given past work indicating that the reliability of error-related negativity values stabilizes after people commit about five errors (Olvet & Hajcak, 2009), we needed to exclude 14 participants who made too few errors (<5) on at least 1 of the 5 trial types. Difference waves were computed by subtracting the correct-related negativity from the error-related negativity, which controls for processes common to neural activation during accurate and inaccurate judgments (Luck, 2005).

5.2.4. Error-related negativity by electrode site and condition

First, we submitted difference waves to a 3(Electrode site: Fz, FCz, Cz) × 2(Prime: wrong, neutral) × 2(Target: wrong, neutral) repeated measures ANOVA. Fig. 2 displays error-related negativity difference wave amplitudes (0–150 ms post-response), derived from the difference between incorrect and correct trials at the FCz (Panel A) and Cz (Panel B) electrode sites, separately for each prime-target combination. There was a main effect of electrode site, $F(2,86) = 12.58$, $p < 0.001$, $\eta_p^2 = 0.23$. Error-related negativity was weaker on the Fz electrode site compared to the FCz site, $p < 0.001$, and the Cz site, $p = 0.008$, and did not differ on the latter two sites, $p = 0.998$. Given that past studies focus on the FCz frontocentral site (e.g., Amodio, Devine, & Harmon-Jones, 2008; Amodio et al., 2004; Inzlicht & Al-Khindi, 2012; Legault, Al-Khindi, & Inzlicht, 2012), we averaged across FCz and Cz sites and focused on these for further analyses. Additionally, main effects and higher-order interactions for prime and target type on error-related negativity were non-significant, $ps > 0.600$. Thus, the error-related negativity did not differ across trial types in the Moral Categorization Task, suggesting similar error-and performance monitoring across trial types.

Table 8
Parameter estimates, Experiment 4.

Parameter	Wrong prime	Negative prime	Neutral prime	Across conditions	
	Estimate [95% CI]	Estimate [95% CI]	Estimate [95% CI]	$\Delta G^2(1)$	w
Intentional	0.37 [0.36, 0.39]	0.37 [0.36, 0.39]	0.37 [0.36, 0.39]	–	–
Unintentional	0.22 [0.19, 0.26]	0.12 [0.07, 0.17]	0.00 (constant)	14.15^{**}	0.06
Resp. Bias	0.38 [0.37, 0.40]	0.38 [0.37, 0.40]	0.38 [0.37, 0.40]	–	–

Note. Intentional Judgment and Response Bias parameters constrained to be equal across prime conditions, and Unintentional Judgment constrained to zero in the Neutral prime condition. $^{\dagger}p < 0.100$, $^{*}p < 0.050$, $^{**}p < 0.010$.

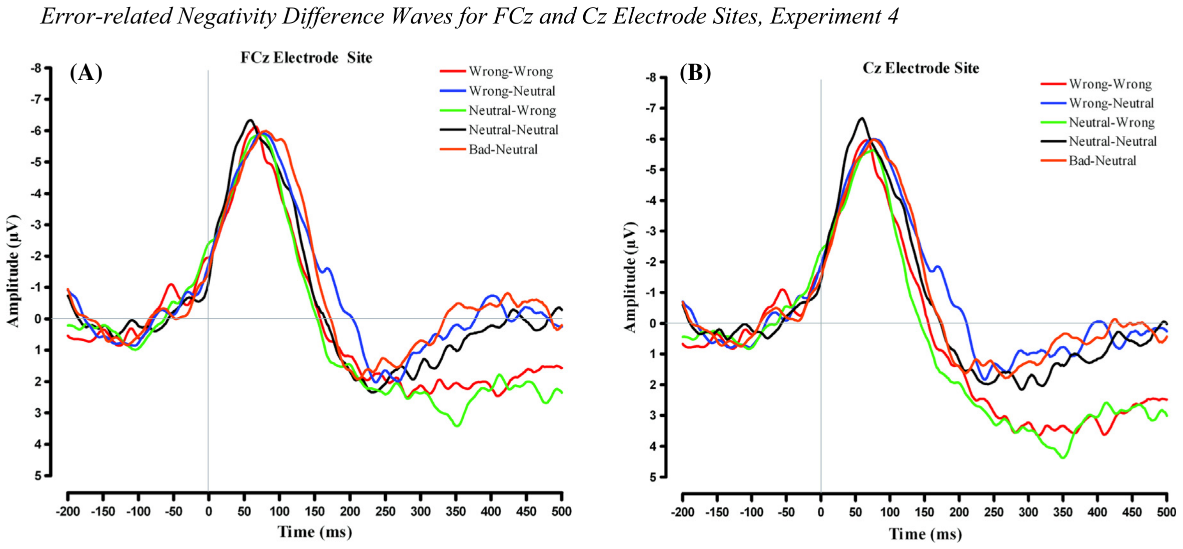

Fig. 2. Error-related negativity difference wave amplitudes (0–150 ms post-response), derived from the difference between incorrect and correct trials at the FCz (Panel A) and Cz (Panel B) electrode sites, Experiment 4. Difference waves are depicted for each prime-target combination (Wrong-Wrong, Wrong-Neutral, Neutral-Wrong, Neutral-Neutral, Negative-Neutral).

5.2.5. Error-related negativity and the multinomial model

Next, we examined the relationships between the error-related negativity and individual-level parameter estimates. First, we examined the relationship between error-related negativity on Wrong-Neutral trials with Intentional Judgment, while controlling for the error-related negativity on Wrong-Wrong trials. As expected, higher error-related negativity on Wrong-Neutral trials was associated with increased Intentional Judgment, $\beta = 0.49$, $t(41) = 3.85$, $p < 0.001$. To isolate prime variance, we predicted Intentional Judgment simultaneously from error-related negativity on Wrong-Neutral, Negative-Neutral, and Neutral-Neutral trials (cf. Amodio et al., 2004). Intentional Judgment was marginally associated with higher error-related negativity on Wrong-Neutral trials, $\beta = 0.35$, $t(40) = 2.02$, $p = 0.051$, but not Negative-Neutral trials, $\beta = 0.28$, $t(40) = 1.67$, $p = 0.103$, or Neutral-Neutral trials, $\beta = 0.12$, $t(40) = 0.71$, $p = 0.485$. The relationship between ERN and Intentional Judgment was strongest on incongruent trials, particularly trials in which incompatible moral responses needed to be inhibited (i.e., on Wrong-Neutral trials).

5.3. Discussion

Experiment 4 revealed that during the Moral Categorization Task, Intentional Judgment corresponded to error-related negativity, a neurophysiological indicator of cognitive control. This finding replicates prior work using event-related potential methodology with sequential priming and process dissociation (Amodio et al., 2004, 2008), and provides independent validation that the Intentional Judgment parameter has characteristics of controlled processing.

6. Experiment 5: Association with voting behavior

In the preceding experiments, we validated the Moral Categorization Task and corresponding multinomial model, finding that Intentional Judgment was sensitive to time pressure whereas Unintentional Judgment was not reducible to negative affect. In the final experiment, we aimed to establish the predictive validity of Unintentional Judgment, by examining its relationship with behavior in a field setting. We examined whether Unintentional Judgment about gay marriage would associate with real-world behavior: voting for or against Amendment One, a North Carolina constitutional amendment to legally define marriage as between one man and one woman. In the time leading up to the vote on May 8, 2012, debate about this amendment was heavily moralized. We conducted a field study of voters on the day of the referendum to test whether Unintentional Judgment of gay marriage as morally wrong would align with voting behavior. We adapted the task to include gay marriage stimuli as primes and targets alongside the normatively wrong and neutral stimuli. By examining judgment accuracy and parameter estimates separately for Amendment One supporters and opponents, we expected to validate implicit moral evaluations about a debated moral issue on the basis of known groups. Critically, we are *not* suggesting that voting in a particular direction indicates moral competence. Instead, we are suggesting that to the degree that some voters implicitly evaluate gay marriage as morally wrong, they will be more likely to vote in favor of an amendment that opposes gay marriage.

6.1. Method

6.1.1. Participants

We recruited 65 participants (41 female, 24 male, $M_{\text{age}} = 47.12$ - years, $SD_{\text{age}} = 13.82$ years) in Orange County, North Carolina. Experimenters approached voters after they had exited polling stations and asked if they would like to participate in a study on "voting and attitudes." Participants were compensated $5. We excluded data for 1 participant who reported a vision condition that made it difficult to see the screen, and 1 participant whose overall error rate was more than 3 standard deviations above the sample mean.

6.1.2. Materials and procedures

6.1.2.1. Voting. After being seated with laptops, participants read: "On today's ballot, you were asked to vote either for or against Amendment One. Amendment One is the Constitutional amendment to provide that marriage between one man and one woman is the only domestic legal union that shall be valid or recognized in the state of North Carolina. Did you vote for or against Amendment One? Please answer as honestly as possible." Participants reported whether they had voted for, voted against, or abstained from voting on Amendment One.

6.1.2.2. Moral categorization task. The task consisted of 3 blocks of 45 trials each. Within each block, there were 5 trials per prime-target combination. Response deadline was held constant at 600 ms. The morally wrong items included: *murder*, *rape*, *domestic assault*, *lying*, *stealing*, *genocide*, *molesting*, *domestic abuse*, *cheating*, and *robbery*. The morally neutral items included: *writing*, *golf*, *baking*, *agreement*, *leisure*, *painting*, *tennis*, *farming*, *modesty*, and *exercise*. The gay marriage word lists only had one item each: *gay marriage*, *same-sex marriage*. We used reduced word lists for this condition because few synonyms would be possible to use without sacrificing intended meaning or brevity. Gay marriage stimuli were used as both primes and targets.

6.1.2.3. Explicit moral judgment. After finishing the task, participants were asked "To what degree is gay marriage morally wrong?" (from 1 = *Not at all* to 5 = *Extremely*).

6.1.2.4. Demographics. Participants reported gender, age, ethnicity, political orientation separately for social and economic issues, political party affiliation, religiosity, religious affiliation, and whether the reason they were voting was because of Amendment One.

6.2. Results

6.2.1. Voting, political orientation, and explicit moral judgment of gay marriage

Within our sample, 34 participants voted against and 29 participants voted for Amendment One. Social and economic conservatism were very highly correlated, $r(63) = 0.91$, and were averaged together into an overall conservatism estimate, revealing a balanced political sample ($M = 3.69$, $SD = 1.93$). In terms of party affiliation, 26 participants self-identified as Democrats, 20 as Republicans, and 17 as Independents. Moral wrongness judgments of gay marriage ($M = 2.76$, $SD = 1.80$) followed a bimodal distribution. Nearly half of the sample (46.0%) judged that gay marriage was not at all wrong; the second largest percentage of the sample judged that gay marriage was extremely wrong (31.7%). Conservatism was associated with stronger explicit wrongness judgments about gay marriage, $r(63) = 0.63$, $p < 0.001$, and voting in favor of Amendment One, B = 0.95, S.E. = 0.22, Wald = 18.15, $p < 0.001$. Explicit wrongness judgments about gay marriage were associated with voting in favor of Amendment One, B = 1.65, S.E. = 0.36, Wald = 21.70, $p < 0.001$.

6.2.2. Error rates

Error rates were very low ($M = 0.04$, $SD = 0.05$), likely due to the fact that from behavioral observation, many participants exceeded

Table 9
Mean proportion of errors by prime, target, and amendment one vote, Experiment 5.

Target	Voted for amendment one					Voted against amendment one				
	Wrong prime	Gay marriage prime	Neutral prime	g_{WN}	g_{GN}	Wrong prime	Gay marriage prime	Neutral prime	g_{WN}	g_{GN}
Wrong	0.03 (0.05)	0.02 (0.06)	0.01 (0.02)	0.42	0.28	0.07 (0.10)	0.06 (0.10)	0.06 (0.09)	0.02	0.00
Neutral	0.08 (0.13)	0.08 (0.11)	0.03 (0.06)	0.41	0.44*	0.05 (0.07)	0.05 (0.06)	0.05 (0.11)	0.06	0.04

Note. Standard deviations are in parentheses. Effect sizes are Hedges' g_{av} for simple effect comparisons between prime types within a target category. $^{\dagger}p < 0.100$, $^{**}p < 0.010$, $^{*}p < 0.050$.

the response deadline. Low error rates can make modeling more difficult, and should thus license caution in interpreting results. We opted to examine the model here given the practical importance and unique nature of the field study. First, we conducted a 3(Prime: wrong, gay marriage, neutral) × 2(Target: wrong, neutral) × 2(Vote: for Amendment One, against Amendment One) mixed ANOVA. This analysis allowed us to examine whether gay marriage primes were exerting similar influence as moral transgression primes, and whether this differed based on voting decision. Table 9 displays error rates by prime and Amendment One vote.

There was a Prime × Vote interaction, $F(2,122) = 3.14$, $p = 0.047$, $\eta_p^2 = 0.05$, but this was not qualified by a Prime × Target × Vote interaction, $F(2,122) = 0.79$, $p = 0.456$, $\eta_p^2 = 0.01$. To better understand the Prime × Vote interaction, we estimated the influence of prime on judgment for Amendment One supporters and opponents. For ease of description, we present separate analyses for each target type (wrong, neutral), although these patterns are not different from one another given the lack of 3-way interaction. For wrong targets, there was not a Prime × Vote interaction, $F(2,122) = 0.40$, $p = 0.673$, $\eta_p^2 = 0.01$, such that prime type did not influence judgment of wrong targets for Amendment One supporters, $F(2,56) = 1.43$, $p = 0.249$, $\eta_p^2 = 0.05$, or opponents, $F(2,66) = 0.01$, $p = 0.989$, $\eta_p^2 = 0.00$. For neutral targets, there was a Prime × Vote interaction, $F(2,122) = 3.22$, $p = 0.043$, $\eta_p^2 = 0.05$, such that prime type influenced judgment of neutral targets for Amendment One supporters, $F(2,56) = 3.80$, $p = 0.028$, $\eta_p^2 = 0.12$, but not opponents, $F(2,66) = 0.11$, $p = 0.899$, $\eta_p^2 = 0.00$. Critically, Amendment One supporters made more errors judging neutral targets after gay marriage primes compared to neutral primes. Amendment One supporters showed a non-significant trend to make more errors judging neutral targets after wrong primes than neutral primes, and no difference in errors after wrong and gay marriage primes. In summary, among participants who voted for Amendment One, gay marriage primes acted similarly to morally wrong primes, causing reduced accuracy in moral judgments about neutral targets. However, given the low amount of errors and small sample of the field study, we suggest that these effects should be interpreted cautiously, as preliminary evidence.

6.2.3. Multinomial model

For the modeling analysis, we divided the sample into Amendment One opponents (who support gay marriage) and Amendment One supporters (who oppose gay marriage). We focused our modeling analysis on the contrast of key interest: the influence of prime type (wrong, gay marriage, neutral) on judgments of neutral targets. We had sufficient degrees of freedom to estimate five parameters with six cells of data. For each voter group, we estimated two Unintentional Judgment parameters (one each for wrong-prime and gay-marriage prime trials, with neutral-prime trials set to zero). Because we did not expect Intentional Judgment to differ across groups, we estimated one Intentional Judgment parameter constrained to be the same across voter groups. A single Response Bias parameter was constrained to chance (0.50) and held constant across voter groups, as we did not expect this to vary across conditions and we did not have additional degrees of freedom to estimate this parameter.

This model fit the data, $G^2(1) = 1.92$, $p = 0.166$, $w = 0.05$. Table 10 displays parameter estimates by prime condition and voting decision. Among Amendment One opponents, Unintentional Judgment did not differ from zero for wrong primes, $\Delta G^2(1) = 0.05$, $p = 0.820$, $w = 0.01$, or gay marriage primes, $\Delta G^2(1) = 0.16$, $p = 0.692$, $w = 0.01$, and these did not differ from each other, $\Delta G^2(1) = 0.02$, $p = 0.884$, $w = 0.00$. By contrast, among Amendment One supporters, Unintentional Judgment differed from zero for wrong primes, $\Delta G^2(1) = 6.77$, $p = 0.009$, $w = 0.08$, and gay marriage primes, $\Delta G^2(1) = 5.93$, $p = 0.015$, $w = 0.08$, and these did not differ from each other, $\Delta G^2(1) = 0.02$, $p = 0.899$, $w = 0.00$. Finally, Unintentional Judgment about gay marriage was marginally stronger among Amendment One supporters than Amendment One opponents, $\Delta G^2(1) = 2.93$, $p = 0.087$, $w = 0.06$, indicating that participants who voted against gay marriage had stronger implicit moral evaluations of gay marriage as morally wrong.

6.3. Discussion

In Experiment 5, we replicated the priming effect on the Moral Categorization Task with the specific issue of gay marriage and showed how it differed as a function of voting: among people who voted for Amendment One, seeing *gay marriage* as a prime made them more likely to mistakenly judge neutral target actions as morally wrong. Multinomial modeling revealed that compared to Amendment One opponents (who support gay marriage), Amendment One supporters (who oppose gay marriage) displayed marginally stronger Unintentional Judgment toward gay marriage primes. People who voted to abolish gay marriage had stronger implicit moral evaluations that gay marriage was morally wrong. This result indicates that the parameter estimates has predictive validity, connecting implicit moral evaluations to social behavior. The low error rates in this study make formal modeling more difficult, and license caution in interpretations of modeling results. Thus, we suggest the current finding is preliminary, and that it should inspire comprehensive tests of the predictive validity of these process parameters.

7. Individual differences analyses

In the final set of analyses, we examined associations between individual-level parameter estimates and four moral personality measures: psychopathic tendencies, moral identity, guilt proneness, and sacred value judgments. We predicted that Unintentional Judgment about moral transgressions would correlate negatively with psychopathic tendencies, and positively with moral identity, guilt proneness, and sacred value judgments.

7.1. Method

7.1.1. Participants

The total number of participants in Experiments 1–4 with complete data for individual-level parameter estimates and moral per-

Table 10
Parameter estimates by amendment one vote, Experiment 5.

Parameter	Wrong prime	Gay marriage prime	Neutral prime	Across conditions	
	Estimate [95% CI]	Estimate [95% CI]	Estimate [95% CI]	$\Delta G^2(1)$	w
Voted against					
Intentional	0.91 [0.88, 0.94]	0.91 [0.88, 0.94]	0.91 [0.88, 0.94]	–	–
Unintentional	0.06 [−0.46, 0.58]	0.10 [−0.43, 0.64]	0.00 (constant)	0.02	0.00
Response bias	0.50 (constant)	0.50 (constant)	0.50 (constant)	–	–
Voted for					
Intentional	0.91 [0.88, 0.94]	0.91 [0.88, 0.94]	0.91 [0.88, 0.94]	–	–
Unintentional	0.76 [−0.01, 1.53]	0.71 [−0.05, 1.46]	0.00 (constant)	0.02	0.00
Response bias	0.50 (constant)	0.50 (constant)	0.50 (constant)	–	–
Unintentional by voter group (ΔG^2, w)	3.93*, 0.06	2.93†, 0.06			

Note. Intentional Judgment constrained to be the same across voter groups, Response Bias estimated as a constant (0.50) across voter groups, and Unintentional Judgment constrained to zero in the Neutral prime condition. $^{**}p < 0.010$, $^\dagger p < 0.100$, $^* p < 0.050$.

sonality measures was $N = 245$. For Experiment 1, we used participants in the 400-ms and 500-ms deadline conditions. We excluded 9 participants whose individual models did not converge, and 1 participant whose parameter estimates were unstable due to a negative variance, leaving a final sample of 235 participants.[1] *A priori* power analysis revealed that to find an individual difference correlation of $r = 0.20$ with 80% power and a two-tailed α probability of 0.05, the required sample size was $N = 193$. Thus, collapsing across Experiments 1–4 provided sufficient power to detect modest correlations of parameter estimates with moral personality traits.

7.1.2. Materials and Procedures

7.1.2.1. Psychopathic tendencies. Participants completed Levenson's Self-Report Psychopathy Scale (Levenson, Kiehl, & Fitzpatrick, 1995). This 26-item scale captures psychopathic traits including callous affect, lack of empathy and guilt, manipulation, impulsivity, and antisocial lifestyle and behaviors (e.g., "For me, what's right is what I can get away with").

7.1.2.2. Moral identity. Participants completed the 10-item Self-Importance of Moral Identity Scale (Aquino & Reed, 2002). Participants were told to think of someone who embodied nine morally praiseworthy characteristics: caring, compassionate, fair, friendly, generous, hardworking, helpful, honest, and kind. Internalization items asked participants to report how much these moral traits were important to the self-concept (e.g., "I strongly desire to have these characteristics"), whereas the symbolization items asked participants to report how much they display these moral traits in social settings (e.g., "I am actively involved in activities that communicate to others that I have these characteristics").

7.1.2.3. Guilt proneness. Participants completed the Guilt and Shame Proneness Scale (GASP; Cohen et al., 2011). In this 16-item measure, participants anticipate emotional responses to moral transgressions (e.g., "After realizing you have received too much change at a store, you decide to keep it because the salesclerk doesn't notice. What is the likelihood that you would feel uncomfortable about keeping the money?") The four sub-scales include Guilt-Negative-Behavior-Evaluation, or guilt at one's own actions; Guilt-Repair, or guilt-based actions to repair relationships following transgressions; Shame-Negative-Self-Evaluation, or shame about oneself; and Shame-Withdraw, or shame-based actions to withdraw from public.

7.1.2.4. Sacred value judgments. Participants completed the 20-item Moral Foundations Sacredness Scale (Graham et al., 2009), which assesses willingness to violate moral standards for money. Participants are instructed: "Try to imagine actually doing the following things, and indicate how much money someone would have to pay you, (anonymously and secretly) to be willing to do each thing. For each action, assume that nothing bad would happen to you afterwards. Also assume that you cannot use the money to make up for your action." The scale contains four items for each of the five foundations posited by Moral Foundations Theory: harm (e.g., "kick a dog in the head, hard"), fairness (e.g., "cheat in a game of cards played for money with some people you don't know well"), in-group loyalty (e.g., "leave the social group, club, or team that you most value"), authority, (e.g., "make a disrespectful gesture at your boss, teacher, or professor") and purity (e.g., "get a blood transfusion of 1 pint of disease-free, compatible blood from a convicted child molester"). Response options are: $0 (I'd do it for free), $10, $100, $1000, $10,000, $100,000, a million dollars, and never for any amount of money.

7.2. Results

Table 11 displays correlations between parameter estimates and moral personality variables. Unintentional Judgment toward wrong primes correlated negatively with psychopathic tendencies ($\alpha = 0.82$), $r(235) = -0.20$, $p = 0.002$—with similar associations for primary psychopathy, $r(235) = -0.17$, $p = 0.011$, and secondary psychopathy, $r(235) = -0.19$, $p = 0.004$—and positively with internalized moral identity ($\alpha = 0.81$), $r(235) = 0.24$, $p < 0.001$, symbolic moral identity ($\alpha = 0.77$), $r(235) = 0.13$, $p = 0.049$, GASP Guilt Negative-Behavior-Evaluation ($\alpha = 0.62$), $r(235) = 0.17$, $p = 0.009$, and GASP Shame Negative-Self-Evaluation, ($\alpha = 0.61$), $r(235) = 0.17$, $p = 0.011$, but not GASP Guilt-Repair ($\alpha = 0.66$), $r(235) = 0.10$,

[1] In the individual difference analyses, there were also 21 participants whose individual-level models failed to have adequate fit. When excluding these participants, results were relatively unchanged. Unintentional Judgment about transgressions correlated negatively with psychopathic tendencies ($r = -0.21$, $p = 0.002$); positively with internalized moral identity ($r = 0.25$, $p < 0.001$), GASP Guilt-Negative-Behavior-Evaluation ($r = 0.17$, $p = 0.014$), GASP Shame-Negative-Self-Evaluation ($r = 0.18$, $p = 0.007$); marginally positively with symbolic moral identity ($r = 0.12$, $p = 0.079$) and GASP Guilt-Repair ($r = 0.12$, $p = 0.071$); and not with sacred values ($r = 0.11$, $p = 0.156$), or GASP Shame-Withdrawal ($r = -0.01$, $p = 0.905$). Regression analyses produced similar results. For psychopathic tendencies, there was a unique negative relationship with Unintentional Judgment after transgression primes ($\beta = -0.16$, $t = -2.35$, $p = 0.020$), which held when controlling for Unintentional Judgment after negative primes ($\beta = -0.19$, $t = -2.19$, $p = 0.030$). For internalized moral identity, there were positive relationships with Unintentional Judgment after transgression primes ($\beta = 0.17$, $t = 2.50$, $p = 0.013$) and Intentional Judgment ($\beta = 0.16$, $t = 1.95$, $p = 0.053$), which held when controlling for Unintentional Judgment after negative primes (Unintentional Wrong: $\beta = 0.19$, $t = 2.21$, $p = 0.028$; Intentional: $\beta = 0.22$, $t = 2.42$, $p = 0.017$). For GASP Guilt-Negative-Behavior-Evaluation, there was a unique positive relationship with Unintentional Judgment after transgression primes ($\beta = 0.14$, $t = 1.93$, $p = 0.055$), which held when controlling for Unintentional Judgment after negative primes ($\beta = 0.20$, $t = 2.27$, $p = 0.024$). Finally, there were no relationships of process parameters with sacred value judgments ($ps > 0.140$). Because results were unchanged when excluding these participants, we retained the fuller sample to maximize data used.

Table 11
Correlations between parameter estimates and moral personality traits, Experiments 1–4.

Moral personality	Intentional	Unintentional wrong	Unintentional negative	Response bias
Self-reported psychopathy	$-0.11^{\dagger}$	-0.20^{**}	-0.08	0.05
Internalized moral identity	0.21^{**}	0.24^{***}	0.12	-0.01
Symbolic moral identity	0.03	0.13^{*}	0.08	0.02
GASP guilt-NBE	0.14^{*}	0.17^{**}	0.09	-0.05
GASP guilt-repair	0.07	0.10	-0.03	0.08
GASP shame-NSE	$0.13^{\dagger}$	0.17^{*}	$0.12^{\dagger}$	0.05
GASP shame-withdraw	0.00	-0.01	-0.06	-0.08
Sacred value judgments	0.04	$0.15^{\dagger}$	0.11	0.02

Note. Guilt-NBE = Guilt-Negative-Behavior-Evaluation. Shame-NSE = Shame-Negative-Self-Evaluation. For all personality measures except Sacred Value Judgments, $N = 235$ for associations with Intentional, Unintentional Wrong, and Response Bias, and $N = 196$ for Unintentional Negative, which was not assessed in Experiment 1. For Sacred Value Judgments, assessed in Experiments 1–3, $N = 177$ for associations with Intentional, Unintentional Wrong, and Response Bias, and $N = 138$ for Unintentional Negative. $^{***}p < 0.001$, $^{**}p < 0.010$, $^{*}p < 0.050$, $^{\dagger}p < 0.100$.

$p = 0.125$, or GASP Shame-Withdraw ($\alpha = 0.77$), $r(235) = -0.01$, $p = 0.896$, and marginally positively with sacred value judgments ($\alpha = 0.89$), $r(177) = 0.15$, $p = 0.053$. By contrast, Unintentional Judgment toward negative primes did not correlate with any moral personality measures ($ps > 0.090$), with only a marginal correlation with GASP Shame Negative-Self-Evaluation, $r(196) = 0.12$, $p = 0.096$.

Intentional Judgment correlated positively with internalized moral identity, $r(235) = 0.21$, $p = 0.001$, and GASP Guilt Negative-Behavior-Evaluation, $r(235) = 0.14$, $p = 0.032$, marginally positively with GASP Shame Negative-Self-Evaluation, $r(235) = 0.13$, $p = 0.056$, and marginally negatively with psychopathic tendencies, $r(235) = -0.11$, $p = 0.089$, but did not correlate with symbolic moral identity, $r(235) = 0.03$, $p = 0.649$, GASP Guilt-Repair, $r(235) = 0.07$, $p = 0.308$, GASP Shame-Withdrawal, $r(235) = 0.00$, $p = 0.997$, or sacred values, $r(177) = 0.04$, $p = 0.610$. Response Bias did not correlate with any moral personality measures ($ps > 0.220$).

To examine the specificity of these relationships to Unintentional Judgment toward wrong primes and account for method variance across studies, we conducted a series of multiple regressions predicting each moral personality variable from Unintentional Judgment toward wrong primes, Intentional Judgment, and Response Bias. In these analyses, we included three contrast-coded variables to account for method variance across the four experiments. Because Unintentional Judgment toward negative primes was only assessed in Experiments 2–4, we first present regression analyses excluding this variable to allow for maximal sample size, and subsequently add this variable in a separate regression. This second analysis has a reduced sample size and only includes two contrast-coded variables across three studies.

First, psychopathic tendencies were associated with reduced Unintentional Judgment toward wrong primes, $\beta = -0.18$, $t = -2.61$, $p = 0.010$, and marginally increased Response Bias, $\beta = 0.13$, $t = 1.68$, $p = 0.095$, but not with Intentional Judgment, $\beta = 0.01$, $t = 0.06$, $p = 0.950$. When adding Unintentional Judgment toward negative primes, there was still a relationship for Unintentional Judgment toward wrong primes, $\beta = -0.20$, $t = -2.38$, $p = 0.018$, but not Unintentional Judgment toward negative primes, $\beta = 0.06$, $t = 0.76$, $p = 0.450$. As expected, participants higher in psychopathic tendencies showed reduced implicit moral evaluations of moral transgressions.

Second, internalized moral identity was independently associated with increased Unintentional Judgment toward wrong primes, $\beta = 0.17$, $t = 2.47$, $p = 0.014$, and increased Intentional Judgment, $\beta = 0.16$, $t = 2.03$, $p = 0.043$, but not Response Bias, $\beta = -0.00$, $t = -0.05$, $p = 0.960$. When adding Unintentional Judgment toward negative primes, there were still significant relationships for Unintentional Judgment toward wrong primes, $\beta = 0.19$, $t = 2.24$, $p = 0.026$, and Intentional Judgment, $\beta = 0.20$, $t = 2.32$, $p = 0.022$, but not Unintentional Judgment toward negative primes, $\beta = -0.06$, $t = -0.71$, $p = 0.476$. On the other hand, symbolic moral identity was only marginally positively associated with increased Unintentional Judgment toward wrong primes, $\beta = 0.13$, $t = 1.80$, $p = 0.073$, but not with Intentional Judgment, $\beta = -0.03$, $t = -0.31$, $p = 0.758$, or Response Bias, $\beta = -0.04$, $t = -0.48$, $p = 0.633$. When adding Unintentional Judgment toward negative primes, there were no unique relationships ($ps > 0.130$). As expected, participants for whom morality was more intrinsically important to their self-concepts showed stronger implicit moral evaluations of moral transgressions. These individuals also showed increased Intentional Judgment, which could be due to having moral information more readily accessible to make target judgments accurately (Aquino et al., 2009), or to being more generally experienced with engaging in moral behaviors. These unique relationships of moral identity with Unintentional Judgment and Intentional Judgment indicate the advantage of the formal modeling approach, as this method is one of the only attempts to measure these component processes simultaneously.

Third, Guilt-Negative-Behavior Evaluation was marginally positively associated with increased Unintentional Judgment toward wrong primes, $\beta = 0.13$, $t = 1.88$, $p = 0.061$, but not Intentional Judgment, $\beta = 0.07$, $t = 0.92$, $p = 0.358$, or Response Bias, $\beta = -0.06$, $t = -0.72$, $p = 0.472$. When adding Unintentional Judgment toward negative primes, there was a positive relationship for Unintentional Judgment toward wrong primes, $\beta = 0.19$, $t = 2.16$, $p = 0.032$, but not Unintentional Judgment toward negative primes, $\beta = -0.05$, $t = -0.60$, $p = 0.553$. For Guilt-Repair, there were not any significant relationships with process parameters in the initial analysis ($ps > 0.480$) or when also including Unintentional Judgment toward negative primes ($ps > 0.100$). For Shame-Negative-Self-Evaluation, there were not any significant relationships with process parameters in the initial analysis ($ps > 0.140$) or when also including Unintentional Judgment toward negative primes ($ps > 0.180$). For Shame-Withdraw, there was a marginal positive relationship for Response Bias, $\beta = 0.11$, $t = 1.68$, $p = 0.094$, and when including Unintentional Judgment toward negative primes, there were marginal positive relationships for Unintentional Judgment toward wrong primes, $\beta = 0.12$, $t = 1.72$, $p = 0.088$, and Response Bias, $\beta = 0.12$, $t = 1.67$, $p = 0.096$. As expected, participants more prone to experience guilt about acting unethically showed increased implicit moral evaluations of moral transgressions.

Finally, there were no significant relationships between any process parameters and sacred value judgments ($ps > 0.080$), with only a marginal positive relationship for Unintentional Judgment toward wrong primes, $\beta = 0.15$, $t = 1.76$, $p = 0.081$. There were no relationships when including Unintentional Judgment toward negative primes ($ps > 0.150$). In contrast to the results for psychopathic tendencies, moral identity, and guilt proneness, there was no evidence for unique relationships with implicit moral evaluations of moral transgressions.

Across three measures of moral personality (psychopathic tendencies, moral identity, guilt proneness, but not sacred value judgments), participants who cared more about morality exhibited stronger implicit moral evaluations about moral transgressions. These results provide convergent support for the construct validity of implicit moral evaluations.

8. General discussion

In everyday experience, the gap between moral belief and moral behavior can range from small to severe. Psychopaths may lurk among us, but so do people who cheat on board games, taxes, and lovers. Hypocrisy abounds in everyday life, as people can readily say that they believe certain actions to be morally wrong, while lacking the implicit moral evaluations that would prevent them from engaging in such behaviors. To understand variability in moral personality and moral behavior, it is useful to quantify implicit moral evaluations using rigorous measurement techniques that help disentangle underlying component processes.

Across five experiments, we developed the Moral Categorization Task and multinomial model to understand component processes involved in moral judgment. In the task, participants are instructed to judge target actions as morally wrong or not, while avoiding prime action words that were wrong or neutral. Consistently across studies, transgression primes such as *murder*—compared to neutral primes—biased moral judgments about neutral targets such as *baking*, suggesting that the primes were moralizing target actions.

A task dissociation approach would equate this performance bias with the underlying process of interest (implicit moral evaluations). By contrast, performance bias could be due to multiple processes. Our multinomial model captures three processes: Intentional Judgment (moral judgment of target actions), Unintentional Judgment (an evaluative tendency to morally judge target actions in a prime-consistent manner when Intentional Judgment fails), and Response Bias (a directional tendency to judge target actions as "wrong" when both Intentional Judgment and Unintentional Judgment fail). The primary advantage of the multinomial model is capturing these distinct component processes within the same paradigm, rather than assuming that specific measures only capture specific processes.

The multinomial model quantifies distinct component processes, but does not license the conclusion that Intentional and Unintentional Judgment differ in terms of operating conditions such as consciousness or efficiency. In the current experiments, we used modeling to test theoretically derived predictions about when processes would be more or less pronounced under different experimental conditions. Across studies, Unintentional Judgment was insensitive to fast response deadlines and unrelated to neurophysiological signals of control (the error-related negativity). By contrast, Intentional Judgment was impaired by fast deadlines and associated with error-related negativity. This dissociation by relevant manipulations validates that these processes are independent and specifies conditions under which they operate (Payne, 2008).

We stipulate that Intentional and Unintentional Judgment are distinct on the basis of one feature: intentionality. This theory-driven distinction is based upon what participants are instructed to do: evaluate target actions while avoiding the influence of prime actions. On the assumption that participants are following task instructions, then their task intentions should be to morally judge target actions; any moral judgment of prime actions should conflict with task intentions. Because Unintentional Judgment is posited to influence performance only when Intentional Judgment fails, prime influence can be defined as unintentional. By formally specifying the counter-intentional nature of implicit moral evaluations, this work can advance the study of moral cognition by providing greater precision about one of its central constructs.

In our experiments, we validated that Unintentional Judgment in response to moral transgressions on the Moral Categorization Task is not simply a negative affective response. Unintentional Judgment toward wrong primes was stronger than Unintentional Judgment toward non-moral negative primes (Experiments 2–4). Given that these primes are matched on affective dimensions of valence and arousal but differ in moral content, this result suggests that Unintentional Judgment toward wrong primes is attuned to moral content in particular. Further supporting this claim, Unintentional Judgment about transgressions uniquely associated with moral personality. Importantly, we do not claim a firm distinction between "the moral mind" and non-moral psychological processes. Consistent with the emerging consensus in social neuroscience, moral evaluations recruit a host of domain-general processes including affect, conceptual knowledge, attention, and others (Cameron et al., 2015; Cushman & Young, 2011; Decety & Cowell, 2014; Greene, 2015; Young & Dungan, 2012).

Finally, we validated the Unintentional Judgment on the Moral Categorization Task can capture idiographic moral evaluations. In Experiment 4, participants listed moral transgressions that they were strongly opposed to, which were then incorporated as stimuli. Participant-generated moral transgressions produced similar priming effects to researcher-generated moral transgressions. Experiment 5 examined responses to a moral issue on which personal opinions vary: gay marriage. Implicit moral evaluations of gay marriage as morally wrong were stronger for participants who voted in favor of a state constitutional amendment against gay marriage, establishing that the process parameters can align with real-world behavior.

8.1. Modeling moral cognition

Formal models of moral cognition can increase theoretical precision in moral psychology by more clearly specifying which processes shape moral judgment, under what conditions, and why (Crockett, 2016). Multinomial modeling has many theoretical and methodological advantages in understanding implicit moral evaluations. A primary advantage is that instead of inferring underlying processes from task performance, it stipulates *a priori* how distinct processes interact as mediators between experimental inputs (e.g., prime stimuli, response deadline) and behavioral outputs (e.g., judgments on the Moral Categorization Task; cf. Gawronski & Bodenhausen, 2015; Gawronski et al., 2014). Our multinomial model parallels the "Control-dominating" process dissociation model: the causal influence of Unintentional Judgment on task performance is logically contingent on Intentional Judgment failing.

The multinomial model does not assume the existence of dual systems, such as "System I" and "System II" (Kahneman, 2003). Dual systems approaches have been critiqued because many of the features meant to cluster within a given system—such as consciousness, intentionality, and efficiency—do not reliably co-occur (Kruglanski & Gigerenzer, 2011; Moors & De Houwer, 2006). Intentional Judgment and Unintentional Judgment reflect different processes operating within the Moral Categorization Task, and formally positing these processes does not require theoretical assumptions of dual systems models. This approach does not assume different processing modes, such that people alternate between an "automatic mode" and a "controlled mode" (Greene, 2013). Rather, these component processes operate simultaneously to influence moral judgment. Most importantly, multinomial modeling avoids the task dissociation assumption that equates a given task with a given process. Rather, performance on any task—ranging from sequential priming to self-report—is the net effect of many interacting processes.

Finally, this approach is useful because it specifies a different task domain than previous multinomial models of moral judgment (e.g., Conway & Gawronski, 2013). Whereas prior work involves moral decision-making in ambiguous situations with competing principles in play, the current work focuses on immediate responses to unambiguous stimuli. Our approach assumes that moral judgments, regardless of the principles on which they are based, can occur in response to target actions (Intentional Judgment) or in response to prime actions (Unintentional Judgment). In our view, a comprehensive approach to understanding variability in moral cognition requires assessing moral evaluations using a variety of measurement techniques which capture explicit moral evaluations, implicit moral evaluations, and moral behavior.

9. Conclusion

Implicit moral evaluations—unintentional moral assessments of the actions and characters of others—are central to many accounts of moral cognition. Argued to be the psychological fulcrum of moral competence, these evaluations need to be assessed rigorously in a way that separates signal from noise. The current work makes two advances toward this goal. First, we developed a new implicit measure of moral judgment, that moves beyond self-report to capture moral reactivity that is relatively unfiltered by socially desirable response correction. Second, we developed a multinomial model of task performance, dissociating variation in implicit moral evaluations from co-activated processes such as intentional moral evaluations and moralistic response biases. Our approach advances theorizing in moral cognition by formally specifying implicit moral evaluations—as counter-intentional and comprised of affect and conceptual knowledge about morality—and precisely quantifying who has them and who does not. By improving theoretical and methodological precision, we can better predict moral behavior and understand the boundaries of moral competence.

Author note

We thank the many research assistants who have helped with data collection for this project, with particular gratitude to Amanda Anderson, Edward Bartels, Joseph Heffner, Kellyn Platek, and Erika Scheib for helping to collect voting data in the field in Experiment 5. We thank members of the Moral Attitudes and Decision-Making Lab at Duke University, Mind Perception and Morality Lab at University of North Carolina, and the Moral Psychology Research Group for providing helpful feedback on this research.

Supplementary materials

The data for all studies can be accessed at the Open Science Framework: osf.io/34bpt.

Appendix A. Multinomial model equations

$p(\text{accurate}|\text{wrong prime \& wrong target}) = \mathrm{I} + (1-\mathrm{I}) \times \mathrm{U} + (1-\mathrm{I}) \times (1-\mathrm{U}) \times \mathrm{B}$

$p(\text{inaccurate}|\text{wrong prime \& wrong target}) = (1-\mathrm{I}) \times (1-\mathrm{U}) \times (1-\mathrm{B})$

$p(\text{accurate}|\text{wrong prime \& neutral target}) = \mathrm{I} + (1-\mathrm{I}) \times (1-\mathrm{U}) \times (1-\mathrm{B})$

$p(\text{inaccurate}|\text{wrong prime \& neutral target}) = (1-\mathrm{I}) \times \mathrm{U} + (1-\mathrm{I}) \times (1-\mathrm{U}) \times \mathrm{B}$

$p(\text{accurate}|\text{neutral prime \& wrong target}) = \mathrm{I} + (1-\mathrm{I}) \times (1-\mathrm{U}) \times \mathrm{B}$

$p(\text{inaccurate}|\text{neutral prime \& wrong target}) = (1-\mathrm{I}) \times \mathrm{U} + (1-\mathrm{I}) \times (1-\mathrm{U}) \times (1-\mathrm{B})$

$p(\text{accurate}|\text{neutral prime \& neutral target}) = \mathrm{I} + (1-\mathrm{I}) \times \mathrm{U} + (1-\mathrm{I}) \times (1-\mathrm{U}) \times (1-\mathrm{B})$

$p(\text{inaccurate}|\text{neutral prime \& neutral target}) = (1-\mathrm{I}) \times (1-\mathrm{U}) \times \mathrm{B}$

References

Abramowitz, J. S., Huppert, J. D., Cohen, A. B., Tolin, D. F., & Cahill, S. P. (2002). Religious obsessions and compulsions in a non-clinical sample: The Penn Inventory of Scrupulosity (PIOS). *Behaviour Research and Therapy, 40*, 825–838. http://dx.doi.org/10.1016/S0005-7967(01)00070-5.

Adler, N. E., & Ostrove, J. M. (1999). Socioeconomic status and health: What we know and what we don't. *Annals of the New York Academy of Sciences, 896*, 3–15. http://dx.doi.org/10.1111/j.1749-6632.1999.tb08101.x.

Aharoni, E., Antonenko, O., & Kiehl, K. A. (2011). Disparities in the moral intuitions of criminal offenders: The role of psychopathy. *Journal of Research in Personality, 45*, 322–327. http://dx.doi.org/10.1016/j.jrp.2011.02.005.

Amodio, D. M., Bartholow, B. D., & Ito, T. A. (2014). Tracking the dynamics of the social brain: ERP approaches for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience, 9*, 385–393. http://dx.doi.org/10.1093/scan/nst177.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology, 94*, 60–74. http://dx.doi.org/10.1037/0022-3514.94.1.60.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15*, 88–93. http://dx.doi.org/10.1111/j.0963-7214.2004.01502003.x.

Aquino, K., & Reed, A. II, (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*, 1423–1440. http://dx.doi.org/10.1037//0022-3514.83.6.1423.

Aquino, K., Reed, A., II, Freeman, D., Lim, V. K., & Felps, W. (2009). Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality. *Journal of Personality and Social Psychology, 97*, 123–141. http://dx.doi.org/10.1037/a0015406.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*, 193–209.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*, 154–161. http://dx.doi.org/10.1016/j.cognition.2011.05.010.

Bartholow, B. D., Pearson, M. A., Dickter, C. L., Sher, K. J., Fabiani, M., & Gratton, G. (2005). Strategic control and medial frontal negativity: Beyond errors and response conflict. *Psychophysiology, 42*, 33–42. http://dx.doi.org/10.1111/j.1469-8986.2005.00258.x.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86. http://dx.doi.org/10.3758/BF03210812.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*(9), 536–554. http://dx.doi.org/10.1111/spc3.12131.

Bishara, A. J., & Payne, B. K. (2009). Multinomial process tree models of control and automaticity in weapon misidentification. *Journal of Experimental Social Psychology, 45*, 524–534. http://dx.doi.org/10.1016/j.jesp.2008.11.002.

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences, 11*, 387–392. http://dx.doi.org/10.1016/j.tics.2007.07.003.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* Technical Report C-1. The Center for Research in Psychophysiology, University of Florida.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*, 330–350. http://dx.doi.org/10.1177/1088868312440047.

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*. http://dx.doi.org/10.1177/1088868314566683.

Christie, R., & Geis, F. L. (1970). *Studies in machiavellianism*. New York: Academic Press.

Cima, M., Tonnaer, F., & Lobbestael, J. (2007). Moral emotions in predatory and impulsive offenders using implicit measures. *Netherlands Journal of Psychology, 63*, 133–142. http://dx.doi.org/10.1007/BF03061076.

Clerkin, E. M., Fisher, C. R., Sherman, J. W., & Teachman, B. A. (2014). Using the Quadruple Process model to evaluate change in implicit attitudinal responses during therapy for panic disorder. *Behaviour Research and Therapy, 52*, 17–25. http://dx.doi.org/10.1016/j.brat.2013.10.009.

Cohen, T. R., Panter, A. T., & Turan, N. (2012). Guilt proneness and moral character. *Current Directions in Psychological Science, 21*, 355–359. http://dx.doi.org/10.1177/0963721412454874.

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP Scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology, 100*, 947–966. http://dx.doi.org/10.1037/a0022641.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469–487. http://dx.doi.org/10.1037/0022-3514.89.4.469.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*, 216–235. http://dx.doi.org/10.1037/a0031021.

Crockett, M. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current Directions in Psychological Science, 25*, 85–90. http://dx.doi.org/10.1177/0963721415624012.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*, 1052–1075. http://dx.doi.org/10.1111/j.1551-6709.2010.01167.x.

Decety, J., & Cowell, J. M. (2014). Friends or foes: Is empathy necessary for moral behavior? *Perspectives on Psychological Science, 9*, 525–537. http://dx.doi.org/10.1177/1745691614545130.

Degner, J. (2009). On the (un-) controllability of affective priming: Strategic manipulation is feasible but can possibly be prevented. *Cognition and Emotion, 23*, 327–354. http://dx.doi.org/10.1080/02699930801993924.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, 303–305. http://dx.doi.org/10.1111/j.1467-9280.1994.tb00630.x.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin, 139*, 477–496. http://dx.doi.org/10.1037/a0029065.

Duttweiler, P. C. (1984). The internal control index: A newly developed measure of locus of control. *Educational and Psychological Measurement, 44*, 209–221.

Erdfelder, E., Auer, T. S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology, 217*, 108–124. http://dx.doi.org/10.1027/0044-3409.217.3.108.

Fazio, R., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238. http://dx.doi.org/10.1037/0022-3514.50.2.229.

Gawronski, B., & Bodenhausen, G. V. (2015). Social-cognitive theories. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology*. New York, NY: Guilford Press.

Gawronski, B., Sherman, J. W., & Trope, Y. (2014). Two of what? A conceptual analysis of dual-process theories. In B. Gawronski, J. W. Sherman, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 3–19).

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science, 4*, 385–390. http://dx.doi.org/10.1111/j.1467-9280.1993.tb00586.x.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295*, 2279–2282. http://dx.doi.org/10.1126/science.1066893.

Graham, J., Englander, Z., Morris, J. P., Hawkins, C. B., Haidt, J., & Nosek, B. A. (2016). Warning bell: Liberals implicitly respond to group morality before rejecting it explicitly. Available at SSRN 2071499.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47*, 55–130.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neuropsychology, 55*, 468–484. http://dx.doi.org/10.1016/0013-4694(83)90135-9.

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science, 6*, 859–868. http://dx.doi.org/10.1177/1948550615592241.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology, 3*, 405–423. http://dx.doi.org/10.1007/s13164-012-0112-5.

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General, 143*, 1600–1615. http://dx.doi.org/10.1037/a0036149.

Gray, N. S., MacCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003). Forensic psychology: Violence viewed by psychopathic murderers. *Nature, 423*, 497–498. http://dx.doi.org/10.1038/423497a.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.). *Moral psychology* (Vol. 3, pp. 35–79). Cambridge, MA: MIT Press.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin.

Greene, J. D. (2015). The rise of moral cognition. *Cognition, 135*, 39–42. http://dx.doi.org/10.1016/j.cognition.2014.11.018.

Greenwald, A. G., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464.

Greenwald, A. G., Uhlmann, E. L., Poehlman, T. A., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41. http://dx.doi.org/10.1037/a0015575.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814. http://dx.doi.org/10.1037/0033-295X.108.4.814.

Harenski, C. L., Harenski, K. A., Shane, M. S., & Kiehl, K. A. (2010). Aberrant neural processing of moral violations in criminal psychopaths. *Journal of Abnormal Psychology, 119*, 863–874. http://dx.doi.org/10.1037/a0020979.

Hofmann, W., & Baumert, A. (2010). Immediate affect as a basis for intuitive moral judgement: An adaptation of the affect misattribution procedure. *Cognition and Emotion, 24*, 522–535. http://dx.doi.org/10.1080/02699930902847193.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*, 1369–1385. http://dx.doi.org/10.1177/0146167205275613.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*, 679–709. http://dx.doi.org/10.1037/0033-295X.109.4.679.

Inzlicht, M., & Al-Khindi, T. (2012). ERN and the placebo: A misattribution approach to studying the arousal properties of the error-related negativity. *Journal of Experimental Psychology: General, 141*, 799–807. http://dx.doi.org/10.1037/a0027586.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541. http://dx.doi.org/10.1016/0749-596X(91)90025-F.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about moral judgment. *Social Neuroscience*. http://dx.doi.org/10.1080/17470919.2015.1023400.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209. http://dx.doi.org/10.1016/j.cognition.2014.10.005.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697–720. http://dx.doi.org/10.1037/0003-066X.58.9.697.

Kiehl, K. (2008). Without morals: The cognitive neuroscience of criminal psychopaths. In W. Sinnott-Armstrong (Ed.). *Moral psychology: The neuroscience of morality* (Vol. 3, pp. 119–149). Cambridge, MA: MIT Press.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*, 97–109. http://dx.doi.org/10.1037/a0020762.

Legault, L., Al-Khindi, T., & Inzlicht, M. (2012). Preserving integrity in the face of performance threat: Self-affirmation enhances neurophysiological responsiveness to errors. *Psychological Science, 23*, 1455–1460. http://dx.doi.org/10.1177/0956797612448483.

Leuthold, H., Kunkel, A., Mackenzie, I. G., & Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. *Social Cognitive and Affective Neuroscience*. http://dx.doi.org/10.1093/scan/nsu151.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology, 68*, 151–158. http://dx.doi.org/10.1037/0022-3514.68.1.151.

Luck, S. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., & Blair, R. J. R. (2006). The neural basis of implicit moral attitude—An IAT study using event-related fMRI. *Neuroimage, 30*, 1449–1457. http://dx.doi.org/10.1016/j.neuroimage.2005.11.005.

Luu, P., Tucker, D. M., Derryberry, D., Reed, M., & Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of action regulation. *Psychological Science, 14*, 47–53. http://dx.doi.org/10.1111/1467-9280.01417.

Marsh, A. A., & Cardinale, E. M. (2012). Psychopathy and fear: Specific impairments in judging behaviors that frighten others. *Emotion, 12*, 892–898. http://dx.doi.org/10.1037/a0026260.

Marsh, A. A., & Cardinale, E. M. (2014). When psychopathy impairs moral judgments: Neural responses during judgments about causing fear. *Social Cognitive and Affective Neuroscience, 9*, 3–11. http://dx.doi.org/10.1093/scan/nss097.

Meindl, P., & Graham, J. (2014). Know thy participant: The trouble with nomothetic assumptions in moral psychology. In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 233–252). London: Bloomsbury.

Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology, 11*, 99–111. http://dx.doi.org/10.1037/1089-2680.11.2.99.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132*, 297–326. http://dx.doi.org/10.1037/0033-2909.132.2.297.

Moshagen, M. (2010). MultiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42–54. http://dx.doi.org/10.3758/BRM.42.1.42.

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

Olatunji, B. O., Cisler, J. M., Deacon, B. J., Connolly, K., & Lohr, J. M. (2007). The disgust propensity and sensitivity scale-revised: Psychometric properties and specificity in relation to anxiety disorder symptoms. *Journal of Anxiety Disorders, 21*, 918–930. http://dx.doi.org/10.1016/j.janxdis.2006.12.005.

Olvet, D. M., & Hajcak, G. (2009). The stability of error-related brain activity with increasing trials. *Psychophysiology, 46*, 957–961. http://dx.doi.org/10.1111/j.1469-8986.2009.00848.x.

Paulhus, D. L., & Carey, J. M. (2011). The FAD–Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment, 93*, 96–104. http://dx.doi.org/10.1080/00223891.2010.528483.

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. http://dx.doi.org/10.1037/0022-3514.81.2.181.

Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass, 2*, 1073–1092. http://dx.doi.org/10.1111/j.1751-9004.2008.00091.x.

Payne, B. K., & Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. *European Review of Social Psychology, 20*, 272–314. http://dx.doi.org/10.1080/10463280903162177.

Payne, B. K., & Cameron, C. D. (2014). Dual-process theory from a process dissociation perspective. In B. Gawronski, J. Sherman, & Y. Trope (Eds.), *Dual-process theories of the social mind*. Guilford Press.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277–293. http://dx.doi.org/10.1037/0022-3514.89.3.277.

Perugini, M., & Leone, L. (2009). Implicit self-concept and moral action. *Journal of Research in Personality, 43*, 747–754. http://dx.doi.org/10.1016/j.jrp.2009.03.015.

Proudfit, G. H., Inzlicht, M., & Mennin, D. (2013). Anxiety and error monitoring: The importance of motivation and emotion. *Frontiers in Human Neuroscience, 7*, 636. http://dx.doi.org/10.3389/fnhum.2013.00636.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology, 54*, 890–902. http://dx.doi.org/10.1037/0022-3514.54.5.890.

Reed, A., II, & Aquino, K. F. (2003). Moral identity and the expanding circle of moral regard toward out-groups. *Journal of Personality and Social Psychology, 84*, 1270–1286. http://dx.doi.org/10.1037/0022-3514.84.6.1270.

Riefer, D., & Batchelder, W. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*, 318–339. http://dx.doi.org/10.1037/0033- 295X.95.3.318.

Schaich Borg, J., & Sinnott-Armstrong, W. (2013). Do psychopaths make moral judgments? In K. Kiehl & W. Sinnott-Armstrong (Eds.), *Handbook on psychopathy and law* (pp. 107–128). NY: Oxford University Press.

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*. http://dx.doi.org/10.1177/0146167215591501.

Sherman, J. W., Klauer, K. C., & Allen, T. J. (2011). Mathematical modeling of implicit social cognition. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications*. Guilford Press.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass, 4*, 267–281. http://dx.doi.org/10.1111/j.1751-9004.2010.00254.x.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*, 895–917. http://dx.doi.org/10.1037/0022-3514.88.6.895.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*, 853–870. http://dx.doi.org/10.1037/0022- 3514.78.5.853.

Van Bavel, J. J., Xiao, Y. J., & Cunningham, W. A. (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social and Personality Psychology Compass, 6*, 438–454. http://dx.doi.org/10.1111/j.1751-9004.2012.00438.x.

Wakabayashi, A., Baron-Cohen, S., Wheelwright, S., Goldenfeld, N., Delaney, J., Fine, D., ... Weil, L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences, 41*, 929–940. http://dx.doi.org/10.1016/j.paid.2006.03.017.

Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95–116). Guilford Press.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111*, 931–959. http://dx.doi.org/10.1037/0033-295X.111.4.931.

Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience, 7*, 1–10. http://dx.doi.org/10.1080/17470919.2011.569146.