Routledge
Taylor & Francis Group

Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

Brains, trains, and ethical claims: Reassessing the normative implications of moral dilemma research

Michael T. Dale & Bertram Gawronski

To cite this article: Michael T. Dale & Bertram Gawronski (2022): Brains, trains, and ethical claims: Reassessing the normative implications of moral dilemma research, Philosophical Psychology, DOI: 10.1080/09515089.2022.2038783

To link to this article: https://doi.org/10.1080/09515089.2022.2038783

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2022.2038783

Routledge
Taylor & Francis Group

ARTICLE

Brains, trains, and ethical claims: Reassessing the normative implications of moral dilemma research

Michael T. Dale [a] and Bertram Gawronski [b]

[a]Department of Philosophy, University of Texas at Austin, Austin, TX, USA; [b]Department of Psychology, University of Texas at Austin, Austin, TX, USA

ABSTRACT

Joshua Greene has argued that the empirical findings of cognitive science have implications for ethics. In particular, he has argued (1) that people's deontological judgments in response to trolley problems are strongly influenced by at least one morally irrelevant factor, personal force, and are therefore at least somewhat unreliable, and (2) that we ought to trust our consequentialist judgments more than our deontological judgments when making decisions about unfamiliar moral problems. While many cognitive scientists have rejected Greene's dual-process theory of moral judgment on empirical grounds, philosophers have mostly taken issue with his normative assertions. For the most part, these two discussions have occurred separately. The current analysis aims to remedy this situation by philosophically analyzing the implications of moral dilemma research using the CNI model of moral decision-making – a formalized, mathematical model that decomposes three distinct aspects of moral-dilemma judgments. In particular, we show how research guided by the CNI model reveals significant conceptual, empirical, and theoretical problems with Greene's dual-process theory, thereby questioning the foundations of his normative conclusions.

ARTICLE HISTORY
Received 18 August 2021
Accepted 1 February 2022

KEYWORDS
Utilitarianism; consequentialism; deontology; CNI model; dual-process theory; ethics; moral psychology; neuroscience

1. Introduction

Joshua Greene has long argued that the empirical findings of psychology and cognitive neuroscience have implications for ethics (Greene, 2003, 2008, 2014). The basis for his case comes from the results of two influential studies that measured neural activity in participants making different types of moral judgments. The central finding of these studies is that participants who made characteristically utilitarian judgments showed greater activity in regions of the brain claimed to signify abstract reasoning and cognitive control, while those who made characteristically deontological judgments showed greater activity in regions of the brain claimed to signify emotional processing (Greene et al., 2004, 2001).[1]

CONTACT Michael T. Dale michaeldale@utexas.edu University of Texas at Austin, Austin, USA

As Greene originally defines the terms, a characteristically utilitarian judgment is one that maximizes overall welfare, whereas a characteristically deontological judgment is one that is congruent with moral rules, norms, and/or duties (Greene et al., 2004).[2] For example, in the classic trolley dilemma (Foot, 1967), participants are said to have made a characteristically utilitarian judgment if they find it acceptable to kill one person to save the lives of five others. Conversely, participants are said to have made a characteristically deontological judgment if they find it unacceptable to kill one person to save the lives of five others.

Greene goes on to argue that the associations between utilitarian judgments and cognitive reasoning and between deontological judgments and emotional processing leads to the *normative* conclusion that utilitarian judgments are more reliable than deontological judgments (Greene, 2003, 2008, 2013, 2014).[3] Needless to say, Greene's arguments caused a significant amount of controversy in both cognitive science and philosophy. Many cognitive scientists – focusing primarily on the empirical findings – have rejected Greene's dual-process (DP) theory of moral judgment in light of evidence that seems difficult to reconcile with the theory (e.g., Bago & De Neys, 2019; Cohen & Ahn, 2016; Holyoak & Powell, 2016). Philosophers, on the other hand, mostly took issue with Greene's normative assertions (for some notable exceptions, see Bluhm, 2014; Klein, 2011; Saunders, 2016). Some rejected the idea that empirical findings could inform ethics. Others argued that, even if it is possible to bridge the divide between science and philosophy, the normative upshot is quite different from Greene's conclusions (Allman & Woodward, 2008; Berker, 2009; Bruni et al., 2013; Dale, 2020; Dean, 2010; Heinzelmann, 2018; Königs, 2018; Kumar & Campbell, 2012; Paulo, 2019).

For the most part, these two discussions have occurred separately. Because cognitive scientists primarily focus on Greene's empirical arguments, they rarely engage with his normative conclusions. Conversely, many philosophers have neglected to read and respond to the most up-to-date empirical critiques of Greene's DP theory. The objective of this article is to remedy this situation by discussing one of the most substantial challenges to Greene's DP theory in recent years: the CNI model of moral decision-making (Gawronski et al., 2017).[4] In particular, we will show how research guided by the CNI model reveals significant conceptual, empirical, and theoretical problems with Greene's DP theory, thereby questioning the foundations of his normative conclusions. Our hope is that this analysis will be of interest to both philosophers and cognitive scientists, and that it will help foster more dialogue between the two disciplines.

The specific plan is as follows. In Section 2, we explain Greene's DP theory of moral judgment and how Greene arrives at his normative conclusions. Section 3 describes the CNI model of moral decision-making,

a formalized, mathematical model that resolves two major confounds in traditional moral dilemma research. Section 4 reviews empirical evidence obtained with the CNI model, outlining why this evidence poses a challenge to Greene's DP theory. Section 5 discusses the implications of the conclusions presented in Section 4 for Greene's normative arguments.

2. The dual process theory of moral judgment and its normative implications

2.1. The direct route

The original empirical foundation of Greene's DP theory is research that compared participants' responses to different instances of the trolley problem (Greene, 2008; Greene et al., 2004, 2001). In the *switch* dilemma, participants have to indicate if they find it acceptable to flip a switch to redirect the trolley to another track where it would kill only one person instead of the five it was originally heading toward (Foot, 1967). In the *footbridge* dilemma, participants have to indicate if they find it acceptable to push a large man onto the tracks to stop the trolley from hitting five people (Thomson, 1976). A well-replicated finding is that the majority of participants presented with the *switch* dilemma find it acceptable to flip the switch (making a characteristically utilitarian judgment), whereas the majority of participants presented with the *footbridge* dilemma find it unacceptable to push the large man (making a characteristically deontological judgment; Greene et al., 2004, 2001). As Greene argues, though, the two dilemmas are asking the same fundamental question: would you sacrifice one person in order to save five? So, why are people consistently responding to these dilemmas in different ways?

One explanation of these findings is that deontological intuitions about these moral scenarios are being influenced by factors such as personal force and spatial proximity (Greene, 2008, 2014). However, most people (and philosophers) would agree that it is morally irrelevant whether harm comes about from up-close and personal contact or a more removed vantage point. The harm is still the same, which means that people making characteristically deontological judgments in the trolley dilemma are influenced by morally irrelevant factors. For Greene, this conclusion is important because, if we know that some of our intuitions about a moral dilemma are influenced by morally irrelevant factors, it suggests that we should discount those intuitions (Greene, 2014, p. 713). One way to construct this argument is as follows (see Berker, 2009, p. 321):

> P1. The emotional processes that give rise to deontological intuitions respond to factors that make a dilemma personal rather than impersonal.

P2. The factors that make a dilemma personal rather than impersonal are morally irrelevant.

C1. Therefore, the emotional processes that give rise to deontological intuitions respond to factors that are morally irrelevant.

C2. Therefore, deontological intuitions do not have any genuine normative force.

This way of structuring the argument seems to capture what Greene is contending in his earlier writings (e.g., Greene, 2003, 2008) and, at one point, he explicitly agreed with this characterization of the argument (Greene, 2010, p. 12). However, this argument – while valid – is not sound (Paulo, 2019). Recall that Greene's primary findings center around people's judgments about the trolley dilemma. However, the fact that people's deontological judgments in trolley dilemmas are responding to a morally irrelevant factor does not mean that the processes giving rise to deontological judgments in *other* cases are also responding to morally irrelevant factors. Indeed, Dean (2010) discusses a number of cases in which deontological judgments are likely not tracked by "unreliable" emotional processing (e.g., keeping promises, not shoplifting, and praying at specified times).[5] Therefore, P1 is false (without further empirical data). In light of this, the argument needs to be modified, one possibility being the following (see Paulo, 2019, pp. 7–10):

P1. In situations like the *switch* dilemma, people tend to judge in characteristically utilitarian ways.

P2. In situations like the *footbridge* dilemma, people tend to judge in characteristically deontological ways.

P3. Situations like the *footbridge* dilemma have a high level of personal involvement.

P4. Situations like the *switch* dilemma have a low level of personal involvement.

P5. People's judgments in response to trolley dilemmas are strongly influenced by the level of personal involvement.

P6. The level of personal involvement is morally irrelevant.

C1. In situations like the *footbridge* dilemma, people's judgments are strongly influenced by the level of personal involvement.

C2. In trolley-like cases, people's characteristically deontological judgments are strongly influenced by at least one morally irrelevant factor, personal involvement, and are therefore at least somewhat unreliable.

However, this argument – which we will call DIRECT(DEONTOLOGICAL) – is not valid. In particular, the premises do not lead to C1. As mentioned, most people (and philosophers) agree that level of personal involvement is not morally relevant. So, we know that *something* is going wrong *somewhere*.

However, we do not know exactly where. One possibility is that the high level of personal involvement in the *footbridge* dilemma is influencing people's judgments. However, another equally plausible possibility is that the *low* level of personal involvement in the *switch* dilemma is influencing people's judgments (see, Kumar & Campbell, 2012). Greene himself acknowledges this. In his own words: "Are we oversensitive to personal [involvement] in response to *footbridge*, or undersensitive in response to *switch*" (Greene, 2014, p. 713)? As the premises are only explicit about there being differing levels of personal involvement between the two judgments – but never explicit about *which* level of personal involvement is actually having influence – it would be an unwarranted jump to state that it is in fact people's judgments in the *footbridge* dilemma that are being influenced. Which is all to say, C1 cannot be derived from P1-P6.[6]

The second option for Greene would be to pull back and restructure the argument in the following way (see Greene, 2010, p. 16; Paulo, 2019, p. 7):

> P1. People's judgments in response to trolley problems are strongly influenced by the presence/absence of personal force.
>
> P2. The presence/absence of personal force is morally irrelevant to the moral acceptability of actions such as these.
>
> C. People's judgments in response to trolley problems are strongly influenced by at least one morally irrelevant factor, personal force, and are therefore at least somewhat unreliable.

This argument – which we will refer to as DIRECT(GENERAL) – is both valid and sound; and, importantly, it is the route Greene ultimately decides to take (Greene, 2014, p. 13). However, it does not tell us much, and it does not have any specific normative implications. Greene knows this, though, and – in the latter part of Greene (2014)—attempts to put forward a broader argument with further reaching ethical implications, including implications about which types of judgments should be viewed as unreliable in DIRECT(GENERAL). This "indirect route" centers on a more general theory of moral judgment, which provides the basis for Greene's normative arguments about when we should trust our moral judgments and when we should doubt them.

2.2. The Indirect Route

Greene's DP theory of moral judgment belongs to a broader class of similar theories, generically referred to as the dual-systems view of the human mind (Evans & Stanovich, 2013; Gawronski et al., in press; Kahneman, 2011). A shared assumption of these theories is that human behavior is guided by two overarching neural systems. One is claimed to be associated with emotional processing, generating automatic responses to specific stimuli.

The other is claimed to be associated with conscious reasoning, which can override automatic responses given sufficient time and cognitive resources. According to Greene, the emotional (or "automatic") system is helpful due to its efficiency, but sometimes more careful deliberation is required, and in these situations, the conscious reasoning (or "manual") system is called upon (Greene, 2014, pp. 696–708).[7]

Greene further argues that, because our emotional system can only adequately respond to situations that it is either conditioned to respond to or genetically programmed to respond to, we ought to use our conscious reasoning system when dealing with unfamiliar* problems,[8] which he defines as those "with which we have inadequate evolutionary, cultural, or personal experience" (Greene, 2014, p. 714). Because natural selection and cultural inheritance almost certainly have not instilled any automatic responses to the trolley problem – because it was not only thought up recently but it is also bizarre and unrealistic – it qualifies as an unfamiliar* problem, which means that it would be best addressed by our conscious reasoning system. And so would, according to Greene, many other difficult moral issues we are dealing with today, such as global poverty, climate change, and terrorism (Greene, 2014, p. 716). These are all new phenomena for the human species, and there is little reason to think that we have automatic responses that can reliably address them. Thus, we should trust the responses from our conscious reasoning system more than those from our emotional system when responding to them.

But this only gets Greene so far. We still need to know which responses are generated by our conscious reasoning system, and which responses are generated by our emotional system. In a series of influential studies, Greene and colleagues measured participants' neural activity while they made moral-dilemma judgments (Greene et al., 2004, 2001). They found that participants who made characteristically utilitarian judgments showed greater activity in regions of the brain associated with abstract reasoning and cognitive control (e.g., dorso-lateral prefrontal cortex, anterior cingulate cortex, parietal lobe), whereas participants who made characteristically deontological judgments showed greater activity in regions of the brain associated with emotional processing (e.g., medial frontal gyrus, posterior cingulate gyrus, angular gyrus). In the years following these publications, there have been many follow-up studies, and Greene argues that, for the most part, the evidence supports a dual-process view that treats utilitarian judgments as the outcome of controlled cognitive analyses of costs and benefits and deontological judgments as the outcome of automatic emotional responses (for a detailed review of the evidence, see Greene, 2014, pp. 701–706).

If this is true, one can construct the following argument (see Dale, 2020):

P1. We ought to trust our manual system more than our automatic system when facing unfamiliar* problems.

P2. Many important moral problems are unfamiliar* problems.

P3. Deontological judgments are generated by an emotional neural network, while utilitarian judgments are generated by a conscious reasoning neural network.

P4. Our automatic system is an emotion-based system, while our manual system is a conscious reasoning-based system.

C1. Deontological judgments are an output of our automatic system, while utilitarian judgments are an output of our manual system.

This, then, leads to:

C2. We ought to trust our utilitarian judgments more than our deontological judgments when facing unfamiliar* moral problems (Greene, 2014, pp. 716-725).

Importantly, if this "indirect route" – which we will refer to as INDIRECT – is successful, then we have more reason to believe that it is the deontological judgments (and not the utilitarian judgments) that are responding to a morally irrelevant factor in DIRECT(GENERAL).

3. The CNI model of moral decision-making

Our central argument is that empirical findings obtained with the CNI model (Gawronski et al., 2017) pose a challenge to INDIRECT, because these findings question the soundness of P3. The CNI model is a formalized, mathematical model that has been developed to resolve two methodological confounds in research using the trolley problem (and structurally similar sacrificial dilemmas). One confound is rooted in the fact that characteristically utilitarian and characteristically deontological judgments are measured in a non-independent manner, such that acceptance of the utilitarian option requires rejection of the deontological option, and vice versa (see Conway & Gawronski, 2013). Although this approach may not seem problematic for philosophical debates about utilitarian and deontological ethics, it does pose an interpretational problem for psychological research on the mental processes underlying moral-dilemma judgments. To illustrate this problem, consider a hypothetical study in which an experimental manipulation influences the likelihood that participants prefer characteristically utilitarian over characteristically deontological judgments in the trolley problem. Assuming that utilitarian and deontological judgments are the products of two functionally independent processes (as proposed by Greene's DP theory), the observed effect on moral judgments may reflect either (a) an effect on the process underlying utilitarian judgments or (b) an effect on the process underlying deontological judgments (or both). This confound

leads to major ambiguities in the interpretation of findings obtained with the traditional dilemma paradigm, especially regarding the mental processes underlying moral-dilemma judgments.

A second confound is rooted in the fact that characteristically utilitarian judgments are typically conflated with action, while characteristically deontological judgments are conflated with inaction (see, Crone & Laham, 2017). Again, this approach may not seem problematic for philosophical debates about utilitarian and deontological ethics, but it does pose an interpretational problem for psychological research on the mental processes underlying moral-dilemma judgments. To illustrate this problem, imagine two participants A and B who approve of pulling the lever in the *switch* dilemma. The modal approach in moral dilemma research would suggest that both participants have made a characteristically utilitarian judgment. Now imagine that participant A approves of pulling the lever only when doing so would save the lives of five but not when it would save the life of just one, while participant B approves of pulling the lever regardless of whether it would save the lives of five or one. While the response pattern shown by participant A could be reasonably described as utilitarian in the sense that A's judgments are influenced by the consequences for the greater good, it would seem problematic to say the same for participant B, because B's judgments are unaffected by the consequences for the greater good. Instead, participant B's judgments are better understood as reflecting outcome-independent approval of the focal action that has nothing to do with maximizing welfare. Research using the trolley problem (and structurally similar sacrificial dilemmas) is unable to capture this important difference, because it conflates utilitarian judgments with general preference for action and deontological judgments with general preference for inaction.

The CNI model is a formalized, mathematical model that addresses these ambiguities by resolving the two confounds of traditional dilemma research. To this end, the CNI model requires observations of responses to four kinds of moral dilemmas: (1) dilemmas where a proscriptive norm prohibits a focal action and the benefits of this action for overall well-being are greater than the costs, (2) dilemmas where a proscriptive norm prohibits a focal action and the benefits of this action for overall well-being are smaller than the costs, (3) dilemmas where a prescriptive norm prescribes a focal action and the benefits of this action for overall well-being are greater than the costs, and (4) dilemmas where a prescriptive norm prescribes a focal action and the benefits of this action for overall well-being are smaller than the costs. By comparing participants' responses across the four types of dilemmas, the CNI model allows researchers to quantify three distinct aspects of moral-dilemma judgments: (a) sensitivity to consequences, (b) sensitivity to moral norms, and (c) general preference for inaction versus action (Gawronski et al., 2017). Sensitivity to consequences is captured by the CNI model's *C* parameter,

which reflects the extent to which participants prefer action when the benefits of action are greater than its costs and inaction when the benefits of action are smaller than its costs (see first row in Table 1). Sensitivity to moral norms is captured by the CNI model's *N* parameter, which reflects the extent to which participants prefer inaction when a proscriptive norm prohibits action and action when a prescriptive norm prescribes action (see second row in Table 1). Finally, general preference for inaction versus action is captured by the CNI model's *I* parameter, which reflects the extent to which participants show a general preference for either inaction or action regardless of the specifics of the four types of moral dilemmas (see third and fourth row in Table 1).[9] Based on the acronyms of the three parameters, the model is called the CNI model of moral decision-making.

Conceptually, scores on the *C* parameter quantify the extent to which participants' responses across the four types of dilemmas are influenced by the consequences of a given action for the greater good. Scores on the *N* parameter quantify the extent to which participants' responses across the four types of dilemmas are influenced by moral norms prohibiting or prescribing a focal action. Finally, scores on the *I* parameter quantify the extent to which participants show a general preference for either inaction or action across the four types of dilemmas. Each parameter is distinct from the other two in the sense that they can vary independently. For example, a participant's sensitivity to consequences on the *C* parameter has no a priori implications for that person's sensitivity to moral norms on the *N* parameter, in that participants may show (1) a high score on *C* and a low score on *N*, (2) a low score on *C* and a high score on *N*, (3) high scores on both *C* and *N*, or (4) low scores on both *C* and *N*. The same is true for scores on the *I* parameter, which can vary independently of *C* and *N*.

To obtain quantitative estimates for the three parameters, the CNI model utilizes a multinomial modeling approach (Batchelder & Riefer, 1999; Hütter & Klauer, 2016). This approach provides four mathematical

Table 1. Response patterns captured by the three parameters of the CNI model of moral decision-making as a function of type of moral norm (proscriptive vs. prescriptive) and consequences for the greater good (benefits greater vs. smaller than costs). The first row depicts the response pattern captured by the *C* parameter (sensitivity to consequences); the second row depicts the response pattern captured by the *N* parameter (sensitivity to moral norms); the third and fourth rows depict the response pattern captured by the *I* parameter.

Proscriptive Norm		Prescriptive Norm	
Benefits of Action Greater than Costs	Benefits of Action Smaller than Costs	Benefits of Action Greater than Costs	Benefits of Action Smaller than Costs
action	inaction	action	inaction
inaction	inaction	action	action
inaction	inaction	inaction	inaction
action	action	action	action

equations that include the empirically observed probabilities of action (vs. inaction) responses on the four kinds of dilemmas as known numerical values and the three model parameters as unknowns. For example, the probability of showing an action response on dilemmas where a proscriptive norm prohibits the focal action and the benefits of this action for overall well-being are greater than the costs is captured by the equation:

$$p(\text{action} \mid \text{proscriptive norm, benefits} > \text{costs}) = C + [(1 - C) \times (1 - N) \times (1 - I)]$$

Similar equations capture responses on the other three kinds of dilemmas. Using maximum likelihood statistics, the model aims to identify specific values for the three parameters, so that the probabilities of action (vs. inaction) responses to the four kinds of dilemmas predicted by the model equations by means of these values come as close as possible to the empirically observed probabilities of action (vs. inaction) responses on the four kinds of dilemmas. To the extent that the discrepancies between predicted and observed probabilities are small, it can be inferred that the model provides an accurate description of participants' patterns of responses to the four kinds of dilemmas (i.e., the model "fits" the data). Conversely, to the extent that the discrepancies between predicted and observed probabilities are large, the model would not provide an accurate description of participants' responses to the four kinds of dilemmas (i.e., the model does not "fit" the data). Decisions regarding the two cases are based on goodness-of-fit statistics, in that the probabilities of responses predicted by the model should not significantly deviate from the probabilities of observed responses in the data.[10] To the extent that the model fits the data, further tests can be conducted to investigate whether a given factor (e.g., cognitive load, personal involvement) influences moral judgments by influencing sensitivity to consequences, sensitivity to moral norms, or general preference for inaction versus action (or a combination of the three). Similarly, analyses can be conducted to investigate correlations between a measured variable (e.g., individual differences in cognitive reflection) and the three model parameters (e.g., Körner et al., 2020).

The value of the CNI model for providing a more nuanced understanding of responses to traditional dilemmas pitting moral norms against consequences for the greater good can be illustrated with data on the relation between the three CNI parameters and traditional dilemma judgments. Using the terminology of the CNI model, traditional dilemma judgments can be conceptualized as preference for action (vs. inaction) on dilemmas where a proscriptive norm prohibits a focal action and the benefits of this action for overall well-being are greater than the costs (similar to the trolley problem). Analyses by Gawronski et al. (2020) have shown that each parameter of the CNI model meaningfully predicts responses on traditional dilemmas even when controlling for the

respective other two. First, the *C* parameter has been found to be positively related to preference for action in traditional dilemmas, indicating that greater sensitivity to consequences is associated with a greater tendency to prefer characteristically utilitarian judgments in traditional dilemmas. Second, the *N* parameter has been found to be negatively related to preference for action in traditional dilemmas, indicating that greater sensitivity to moral norms is associated with a greater tendency to prefer characteristically deontological judgments in traditional dilemmas. Finally, the *I* parameter has been found to be negatively related to preference for action in traditional dilemmas, indicating that greater preference for inaction versus action is associated with a greater tendency to prefer characteristically deontological judgments in traditional dilemmas.

Although the CNI model may seem in competition with Greene's DP theory, it is important to emphasize that there is no a priori conflict between the two, because they address fundamentally different levels of analysis (for a discussion, see, Gawronski et al., 2018). Whereas the CNI model is a formalized, mathematical model that *describes* patterns of responses in a quantitative manner without making any assumptions about the mental processes underlying the observed response patterns, the DP theory is a mechanistic theory that aims to *explain* responses to moral dilemmas by identifying their underlying mental processes (see De Houwer, 2011; Gawronski & Bodenhausen, 2015). Put differently, the CNI model provides a mathematical tool to quantify the extent to which different factors (e.g., consequences for the greater good) influence moral-dilemma judgments, but the model remains silent on the mental processes by which these factors influence moral-dilemma judgments. Thus, there is no a priori reason why empirical findings of research using the CNI model should be in conflict with the mental process assumptions of the DP theory. By extension, the same applies to Greene's normative conclusions. However, to the extent that empirical findings of research using CNI model turn out to be inconsistent with the assumptions of Greene's DP theory, the soundness of P3 (which is essentially the DP theory) would seem questionable, posing a challenge for Greene's normative conclusions. In the following section, we explain why empirical findings obtained with the CNI model pose a major challenge to Greene's DP theory, and thereby his normative conclusions.

4. Empirical evidence

4.1. Sensitivity to moral norms as deontological responding

An important question for this analysis is how the three parameters of the CNI model map onto the two processes hypothesized by the DP theory. One potential answer is that sensitivity to consequences captured by the

C parameter reflects a pattern of utilitarian responding, whereas sensitivity to moral norms captured by the *N* parameter reflects a pattern of deontological responding. Moreover, general preference for inaction versus action could be interpreted as a domain-independent response bias that may influence moral judgments in a manner that has no direct moral relevance (e.g., a general acquiescence bias to show affirmative responses on self-report measures). This conceptualization requires scrutiny of two hypotheses:

H1a: Sensitivity to consequences (as captured by the CNI model's *C* parameter) is rooted in controlled cognitive analyses of costs and benefits.

H2a: Sensitivity to moral norms (as captured by the CNI model's *N* parameter) is rooted in automatic emotional responses to the idea of causing harm.

Both hypotheses conflict with the available empirical evidence. First, to the extent that sensitivity to consequences is rooted in controlled cognitive analyses of costs and benefits (H1a), disrupting such analyses via cognitive load should reduce scores on the CNI model's *C* parameter. This prediction conflicts with the findings of two studies by Gawronski et al. (2017, Studies 2a and 2b), which poses a challenge to H1a. Consistent with the results of earlier work using the trolley problem (e.g., Greene et al., 2008; Suter & Hertwig, 2011; but see, Baron et al., 2015; Tinghög et al., 2016), the authors found that cognitive load reduced participants' acceptance of action responses in dilemmas where the described actions conflict with a proscriptive norm and the benefits of action are greater than the costs. However, further analyses with the CNI model revealed that this effect was *not* driven by reduced sensitivity to consequences under cognitive load, as suggested by H1a. Instead, cognitive load influenced moral-dilemma judgments by increasing participants' general preference for inaction versus action. In other words, participants simply became more action averse under cognitive load, and increased action aversion was not associated with a lower sensitivity to consequences.[11] That is, moral judgments were influenced by the described consequences to the same extent regardless of cognitive load. Together, these results question the DP hypothesis that utilitarian judgments are rooted in controlled cognitive analyses of costs and benefits.[12]

Second, to the extent that sensitivity to norms is rooted in automatic emotional responses (H2a), enhancing such emotional responses via increased personal involvement should increase scores on the CNI model's *N* parameter. This prediction conflicts with the findings of two studies by Gawronski et al. (2017, Studies 3a and 3b; see also Körner et al., 2020),

which poses a challenge to H2a. Consistent with the results of earlier work using the trolley problem (e.g., Greene et al., 2001), the authors found that personal involvement reduced participants' acceptance of action responses in dilemmas where the described actions conflict with a proscriptive norm and the benefits of action are greater than the costs. However, further analyses with the CNI model revealed that this effect was *not* driven by a reduced sensitivity to moral norms under conditions of increased personal involvement, as suggested by H2a. Instead, personal involvement increased participants' general preference for inaction versus action, and increased action aversion was not associated with a greater sensitivity to moral norms. In fact, sensitivity to moral norms was significantly reduced (rather than increased) by personal involvement.[13] Together, these results conflict with the DP hypothesis that deontological judgments are rooted in automatic emotional responses.[14]

4.2. Generalized inaction as deontological responding

A potential way to reconcile the reviewed findings with the DP theory is to interpret general preference for inaction versus action on the CNI model's *I* parameter as an instance of deontological responding. Consistent with this idea, Baron and Goodwin (2020) argued that a bias against action is an explanation of deontological judgments rather than an alternative process. Although the CNI model's terminology suggests that moral norms are relevant only for the response pattern captured by the *N* parameter, the general norm *first, do no harm* can lead to a bias against action like the general response bias captured by the *I* parameter (see, Gawronski et al., 2020). Indeed, a deontological interpretation of generalized inaction seems much closer to Greene's hypothesis that deontological responses are rooted in automatic emotional responses to the idea of causing harm (see Gawronski et al., 2018). This conceptualization requires scrutiny of two hypotheses:

H1b: Sensitivity to consequences (as captured by the CNI model's *C* parameter) is rooted in controlled cognitive analyses of costs and benefits.

H2b: Generalized inaction (as captured by the CNI model's *I* parameter) is rooted in automatic emotional responses to the idea of causing harm.

A deontological interpretation of the *I* parameter helps to reconcile at least some of the reviewed findings with Greene's DP theory. In line with the propositions that (a) general inaction on the CNI model's *I* parameter reflects a pattern of deontological responding, (b) deontological judgments are rooted in automatic emotional responses, and (c) increased personal

involvement enhances automatic emotional responses, Gawronski et al. (2017, Studies 3a and 3b) found that scores on the *I* parameter were greater under high (vs. low) personal involvement. To reconcile the observed effect of cognitive load on the *I* parameter, one could argue that cognitive effort is required for the suppression of automatic emotional reactions rather than the analysis of costs and benefits. Such an assumption would reconcile the DP theory with the finding that cognitive load increased scores on the *I* parameter without affecting the *C* parameter (Gawronski et al., 2017, Studies 2a and 2b). Based on these arguments, one could argue that the DP theory is actually in line with the findings obtained with the CNI model.

However, a more thorough analysis renders such a conclusion premature. If generalized inaction on the *I* parameter is interpreted as an indicator of deontological responding (rather than a morally irrelevant response bias), one still has to say something about the meaning of norm-congruent responses captured by the *N* parameter. We are not aware of any potential interpretation of norm-congruent responding that would conflict with a deontological view. Yet, if both the *N* and the *I* parameter are interpreted as instances of deontological responding, the DP theory fails to explain why personal involvement *increases* one kind of deontological responding (i.e., generalized inaction on the *I* parameter) and *decreases* the other kind of deontological responding (i.e., sensitivity to moral norms on the *N* parameter), as shown by Gawronski et al. (2017, Studies 3a and 3b; see also Körner et al., 2020). Moreover, if automatic emotional responses are claimed to underlie the patterns of deontological responding captured by the *I* parameter, but not the patterns of deontological responding captured by the *N* parameter, the DP theory would be unable to account for findings by Gawronski et al. (2018), showing that incidental happiness reduces scores on the *N* parameter without affecting scores on the *I* parameter.[15] A previously suggested interpretation of such mood effects is that happiness influences moral-dilemma judgments by dampening negative emotional reactions to the idea of causing harm (Valdesolo & DeSteno, 2006). If that was the case, the above arguments imply that incidental happiness should reduce scores on the *I* parameter, not the *N* parameter. Yet, if the effects of incidental happiness are explained by assuming that negative emotional reactions drive sensitivity to moral norms on the *N* parameter, the proposed explanation would directly contradict the rejection of the very same link in the post-hoc explanation of the finding that personal involvement reduced (rather than increased) sensitivity to moral norms on the *N* parameter.[16]

For a DP account to remain coherent, any ad hoc assumptions have to be applied consistently to different sets of findings. Claiming one thing to explain one set of findings (e.g., automatic emotional reactions underlie deontological response patterns captured by the *I* parameter, but not the *N* parameter) and the opposite to explain a different set of findings (e.g., automatic emotional reactions underlie deontological response patterns

captured by the *N* parameter, but not the *I* parameter) makes a DP account of these findings logically incoherent, rendering Greene's DP theory conceptually and empirically implausible.

5. Implications for Greene's normative arguments

5.1. Implications for DIRECT

Because it is the stronger version of the argument, let's first consider the implications of these findings for DIRECT(DEONTOLOGICAL). Recall the argument:

> P1. In situations like the *switch* dilemma, people tend to judge in characteristically utilitarian ways.
>
> P2. In situations like the *footbridge* dilemma, people tend to judge in characteristically deontological ways.
>
> P3. Situations like the *footbridge* dilemma have a high level of personal involvement.
>
> P4. Situations like the *switch* dilemma have a low level of personal involvement.
>
> P5. People's judgments in response to trolley dilemmas are strongly influenced by the level of personal involvement.
>
> P6. The level of personal involvement is morally irrelevant.
>
> C1. In situations like the *footbridge* dilemma, people's judgments are strongly influenced by the level of personal involvement.
>
> C2. In trolley-like cases, people's characteristically deontological judgments are strongly influenced by at least one morally irrelevant factor, personal involvement, and are therefore at least somewhat unreliable.

Because DIRECT(DEONTOLOGICAL) does not depend on the soundness of Greene's DP theory, it avoids any direct attack from the findings obtained with the CNI model. If it is the case that people's judgments are strongly influenced by the level of personal involvement *in trolley-like scenarios*, then it still can be claimed that *in trolley-like scenarios*, deontological judgments are influenced by a morally irrelevant factor. Thus, if one is to accept validity of DIRECT(DEONTOLOGICAL), then the findings obtained through the CNI model do no major damage to it. However, as already discussed in Section 2.1., the validity of DIRECT(DEONTOLOGICAL) is in serious doubt. As a result, DIRECT(GENERAL) is probably the best Greene can do with regard to his "direct route."

Here, again, is DIRECT(GENERAL):

P1. People's judgments in response to trolley problems are strongly influenced by the presence/absence of personal force.

P2. The presence/absence of personal force is morally irrelevant to the moral acceptability of actions such as these.

C. People's judgments in response to trolley problems are strongly influenced by at least one morally irrelevant factor, personal force, and are therefore at least somewhat unreliable.

As we explained, this argument has no specific normative conclusions. It only states that something is going wrong somewhere when people make judgments about trolley dilemmas. Hence, the findings obtained with CNI model do not add anything for the evaluation of DIRECT(GENERAL). With that said, though, recall that the second (and arguably the primary) goal of Greene's (2014) project is to give us a more general theory (i.e., INDIRECT) about when we should trust our moral judgments and when we should doubt them, such that he could then apply it back onto DIRECT(GENERAL) to reveal that it is in fact the deontological judgments that we should discount. If this is the case, then the success of DIRECT(GENERAL) depends on the soundness of INDIRECT.

5.2. Implications for INDIRECT

Here, again, is INDIRECT:

P1. We ought to trust our manual system more than our automatic system when facing unfamiliar* problems.

P2. Many important moral problems are unfamiliar* problems.

P3. Deontological judgments are generated by an emotional neural network, while utilitarian judgments are generated by a conscious reasoning neural network.

P4. Our automatic system is an emotion-based system, while our manual system is a conscious reasoning-based system.

C1. Deontological judgments are an output of our automatic system, while utilitarian judgments are an output of our manual system.

C2. We ought to trust our utilitarian judgments more than our deontological judgments when facing unfamiliar* moral problems (Greene, 2014, pp. 716-725).

The findings obtained with the CNI model have significant implications for this argument. As explained in Section 4.2., if utilitarian responding is equated with sensitivity to consequences on the *C* parameter and deontological responding is equated with sensitivity to moral norms on the *N* parameter, P3 suggests that (1) cognitive load should influence moral-dilemma judgments by reducing sensitivity to consequences and (2)

personal involvement should influence moral-dilemma judgments by increasing sensitivity to moral norms. Both predictions conflict with the available evidence in that (1) cognitive load was found to influence moral-dilemma judgments via increased action aversion rather than reduced sensitivity to consequences and (2) personal involvement was found to influence moral-dilemma judgments via increased action aversion and reduced (rather than increased) sensitivity to moral norms. Moreover, attempts to reconcile these inconsistencies by interpreting generalized inaction on the *I* parameter as an instance of deontological responding render explanations of different findings by means of P3 logically incoherent, in that they require claiming one thing to explain one set of findings and the opposite to explain a different set of findings. Specifically, a mapping of emotional processes and deontological responses on the *I* parameter fails to explain why incidental happiness influences moral judgments by reducing deontological responses on the *N* parameter rather than the *I* parameter. These inconsistencies suggest that the impact of emotional processes on moral judgments is much more complex than that suggested by P3.

Greene might respond to this concern by reminding us of his fMRI data (Greene et al., 2004, 2001) indicating that participants who made deontological judgments showed heightened activity in areas of the brain associated with emotion. However, as many have already pointed out (e.g., Berker, 2009; Bluhm, 2014; Christensen & Gomila, 2012; Dale, 2020; Dean, 2010; Klein, 2011; Moll & De Oliveira-Souza, 2007; Prinz, 2016), the neuroscientific evidence is much more complicated. For example, some of the regions claimed to be associated with conscious reasoning showed heightened activity during the *footbridge* dilemma, while some of the regions claimed to be associated with emotional processes showed heightened activity during the *switch* dilemma (Greene et al., 2004, 2001). These complexities fit well with the findings obtained the CNI model, which suggest that the role of emotional processes in moral-dilemma judgments is much more complicated than claimed by Greene's DP theory.

Furthermore, the CNI data give us good reason to question Greene's DP model as a whole. Of course, it may be the case that there are two overarching neural systems: one associated with emotional processing, generating automatic responses to specific stimuli; and another associated with conscious reasoning, which allows for more deliberate and reflective processing. However, it can no longer be claimed that this general DP system maps onto human moral judgment in the way that Greene claims. Indeed, maintaining that deontological judgments are associated with one particular overarching system (e.g., the automatic system) requires proponents of the DP theory to make contradictory assumptions in explaining findings with the CNI model.

As a result of all this, C1 is no longer tenable, which means that we do not have any compelling reason to discount our deontological judgments. This conclusion not only calls C2 of INDIRECT into question; it also means that Greene is not going to get the result he wanted with DIRECT(GENERAL).

Now, Greene still may want to hold onto half of INDIRECT's C2. That is, even if it is no longer tenable to discount (or distrust) our deontological judgments, he still may want to claim that we have good reason to trust our utilitarian judgments, because they are the result of our manual system. However, based on our analysis, we should be suspicious of any attempt to associate a particular class of moral judgments with a specific neural system. Indeed, it now seems plausible that more evidence would render such an association as untenable as the presumed link between deontological judgments and automatic emotional processes.

Moreover, even if one were to give Greene the benefit of the doubt, the proposed link between utilitarian judgments and conscious reasoning still would not have any significant normative implications. It is certainly possible that a particular type of moral judgment is associated with a cognitive system that is liable to make reliable judgments, but that does not mean that it makes *more* reliable judgments than the system underlying a different type of moral judgments. Thus, even if utilitarian judgments were generated by conscious reasoning in a manual system, such a link would have no implications for the stand-off between utilitarianism and deontology, which has been the primary goal of Greene's normative project.

6. Conclusion

In this article, we argued that empirical findings obtained with the CNI model have significant implications for Greene's DP theory. Although there is a no a priori conflict between the two theoretical approaches, the findings obtained with the CNI model render the DP theory conceptually and empirically implausible, which poses a challenge to Greene's normative conclusions derived from the DP theory. To be clear, we do not want to say that (cognitive) neuroscience and/or psychology cannot weigh in on the debate between utilitarianism and deontology. Indeed, as this article has implied, the possibility is there. Yet, our analysis suggests that, counter to Greene's arguments, the available evidence does not support a moral superiority of utilitarianism.

Notes

1. We deliberately chose the expression "claimed to signify" because the conclusions in these studies are based on inferences of a particular mental process from the observation of neural activity in a particular brain area (i.e., neural activity X, therefore mental process Y) based on evidence showing that the mental process is associated with neural activity in that area (i.e., mental process Y, therefore neural activity X). In cognitive neuroscience, this logical fallacy is known as the reverse inference problem (Beer, 2015; Poldrack, 2006); philosophers know it as the fallacy of affirming the consequent.
2. The qualifier *characteristically* is important, because it does not stipulate specific mechanisms underlying the two kinds of judgments (Conway et al., 2018). For example, a characteristically utilitarian judgment is one that increases overall welfare regardless of the psychological processes that have led to the judgment (i.e., characteristically utilitarian judgments may or may not result from deliberate reasoning about overall welfare). Similarly, a characteristically deontological judgment is one that is congruent with moral rules, norms, and/or duties regardless of the psychological processes that have led to the judgment (i.e., characteristically deontological judgments may or may not result from deliberate reasoning about moral rules, norms and/or duties). From this perspective, the two kinds of judgments are defined by properties of the judgments themselves, rather than the psychological processes underlying the two kinds of judgments, the latter of which is treated as an empirical question. It is important to acknowledge – as Greene (2014, p. 699) does – that such a behavioral interpretation of utilitarian and deontological judgments is at odds with the use of the concepts of *utilitarianism* and *deontology* in much of the philosophical literature. For example, Kahane (2012) and Paulo (2019) contend that a philosophical understanding of deontology is much more nuanced than Greene's behavioral understanding, and Rosas and Koenigs (2014) and Hennig and Hütter (2020) argue that many of the dilemmas derived from Greene's conceptualization confound utilitarianism with egoism. While we admit that a behavioral interpretation is controversial, discussing this controversy in detail would be outside the scope of the current project (for a detailed discussion, see, Conway et al., 2018).
3. The term *reliable* may seem somewhat vague. Because of its significance for Greene's arguments, we discuss it in detail later in this article. For now, it should be understood as something along the lines of "unaffected by morally irrelevant factors."
4. The letters *C, N*, and *I* refer to the three parameters of the model, capturing sensitivity to consequences (*C*), sensitivity to moral norms (*N*), and general preference for inaction versus action (*I*). Section 3 explains the model in detail.
5. One of the reasons Greene's findings do not generalize here is because the trolley dilemma is concerned with causing physical harm to an individual, yet many of the moral decisions we make in our everyday lives are not concerned with physically harming individuals. For such everyday moral decisions, factors that Greene claims are morally irrelevant in the trolley dilemma could very well be relevant.
6. One way to respond to this criticism is to claim that personal involvement is causally efficacious in both dilemmas, just in differing amounts (i.e., there is low personal involvement in *switch* and high personal involvement in *footbridge*). However, this line of argument is not very promising, as we still would not have any indication about which level of personal involvement is morally superior.
7. Cushman (2013) presented a dual-system theory that has some resemblance to Greene's DP theory. However, Cushman's theory differs from Greene's theory by emphasizing the learning mechanisms underlying evaluations of actions and

evaluations of outcomes. Although it would be interesting to analyze the normative implications of Cushman's theory and its relation to the CNI model, the current article focuses primarily on Greene's DP theory and his argument about the normative implications his theory.

8. Following Dale (2020), we use an asterisk to indicate that we are referring to Greene's technical notion of familiarity/unfamiliarity.
9. Scores on each parameter can range between 0 and 1. Whereas greater scores on the *C* and the *N* parameters reflect relatively greater sensitivity to consequences and relatively greater sensitivity to moral norms, respectively, scores greater than 0.5 on the *I* parameter reflect a general preference for inaction and scores lower than 0.5 reflect a general preference for action.
10. Another important criterion is the magnitude (or effect size) of the observed discrepancy, because even negligible discrepancies can be statistically significant in studies with large numbers of observations (see, Gawronski et al., 2017, Footnote 6).
11. According to Gawronski et al. (2017), participants under cognitive load might have become more action averse, because they felt that they do not have the cognitive capacity to make a well-informed decision. Because harm caused by action is psychologically perceived as more severe than the same amount of harm caused by inaction (Cushman et al., 2006; Spranca et al., 1991), they may prefer not to engage in any action regardless of consequences and norms.
12. Different from the effect of cognitive load in Gawronski et al.'s (2017) studies, Kroneisen and Steghaus (2021) found that time pressure reduced sensitivity to consequences on the CNI model's *C* parameter. However, this effect failed to replicate in a follow-up study reported in the same article. Thus, while it is unclear why cognitive load and time pressure might have different effects, there is little evidence for reduced sensitivity to consequences under suboptimal processing conditions.
13. According to Gawronski et al. (2017), high personal involvement may increase action aversion, because high involvement leads to more concrete visualizations of potential harm (Trope & Liberman, 2010), and harm caused by action is psychologically perceived as more severe than the same amount of harm caused by inaction (Cushman et al., 2006; Spranca et al., 1991). Conversely, low personal involvement may increase sensitivity to norms, because low involvement permits greater psychological distance (Trope & Liberman, 2010), and thus abstract reasoning about moral norms (Körner & Volk, 2014).
14. Gawronski et al. (2017) manipulated personal involvement by comparing judgments of moral acceptability to judgments of whether participants would perform the described action. Although the two kinds of judgments are associated with different degrees of personal involvement, it is worth noting that they differ in other ways that go beyond personal involvement. Thus, it seems possible that personal involvement is driving one of the two observed effects, while the other effect is driven by a factor that is unrelated to personal involvement. Although this ambiguity poses a challenge to interpretations of the obtained differences, it does not save H2a because neither potential case involves increased sensitivity to moral norms under conditions of high personal involvement. On the one hand, personal involvement might decrease sensitivity to moral norms and a confounded factor might increase action aversion. On the other hand, personal involvement might increase action aversion and a confounded factor might decrease sensitivity to moral norms.
15. According to Gawronski et al. (2018), incidental happiness may reduce sensitivity to norms by dampening negative emotional reactions to the idea of violating moral norms (Nichols & Mallon, 2006). Although this interpretation may seem similar to

the DP hypothesis that deontological judgments are rooted in negative emotional reactions to the idea of causing harm, it is different in that it includes conscious considerations of norms as an antecedent of emotional reactions, an idea that is explicitly rejected by Greene (see Greene et al., 2001).

16. The independence of sensitivity to moral norms and general preference for inaction versus action is also supported by various other findings. For example, Zhang et al. (2018) found that chronic stress increased action aversion on the CNI model's *I* parameter without affecting sensitivity to moral norms on the model's *N* parameter. Conversely, Białek et al. (2019) found that reading moral dilemmas in a foreign language reduced sensitivity to moral norms on the *N* parameter without affecting general action tendencies on the *I* parameter. Similarly, Gawronski and Brannon (2020) found that recalling autobiographical memories involving social power reduced sensitivity to moral norms on the *N* parameter without affecting general action tendencies on the *I* parameter.

Acknowledgments

For helpful discussion, we would like to thank Miriam Schoenfield, Julia Driver, and David Sosa.

Disclosure Statement

The authors report that there are no competing interests to declare.

Notes on contributors

Michael T. Dale is a Ph.D. candidate in the Department of Philosophy at the University of Texas at Austin.

Bertram Gawronski is a Professor in the Department of Psychology at the University of Texas at Austin.

ORCID

Michael T. Dale http://orcid.org/0000-0001-7827-5248
Bertram Gawronski http://orcid.org/0000-0001-7938-3339

References

Allman, J., & Woodward, J. (2008). What are moral intuitions and why should we care about them? A neurobiological perspective. *Philosophical Issues*, *18*(1), 164–185. https://doi.org/10.1111/j.1533-6077.2008.00143.x

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

Baron, J., & Goodwin, G. P. (2020). Consequences, norms, and inaction: A critical analysis. *Judgment and Decision Making*, *15* (3) , 421–442 doi:10.31234/osf.io/9zsnr.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284. https://doi.org/10.1016/j.jarmac.2014.09.003

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86. https://doi.org/10.3758/BF03210812

Beer, J. S. (2015). Cognitive neuroscience of social behavior. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 183–204). Guilford Press.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy and Public Affairs*, *37*(4), 292–329. https://doi.org/10.1111/j.1088-4963.2009.01164.x

Białek, M., Paruzel-Czachura, M., & Gawronski, B. (2019). Foreign language effects on moral dilemma judgments: An analysis using the CNI model. *Journal of Experimental Social Psychology*, *85*, 103855. https://doi.org/10.1016/j.jesp.2019.103855

Bluhm, R. (2014). No need for alarm: A critical analysis of Greene's dual-process theory of moral decision-making. *Neuroethics*, *7*(3), 299–316. https://doi.org/10.1007/s12152-014-9209-0

Bruni, T., Mameli, M., & Rini, R. (2013). The science of morality and its normative implications. *Neuroethics*, *7*(2), 159–172. https://doi.org/10.1007/s12152-013-9191-y

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Cohen, D., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359–1381. https://doi.org/10.1037/xge0000210

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

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, *179*, 241–265. https://doi.org/10.1016/j.cognition.2018.04.018

Crone, D. L., & Laham, S. M. (2017). Utilitarian preferences or action preferences? De-confounding action and moral code in sacrificial dilemmas. *Personality and Individual Differences*, *104*, 476–481. https://doi.org/10.1016/j.paid.2016.09.022

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x

Dale, M. (2020). Neurons and normativity: A critique of Greene's notion of unfamiliarity. *Philosophical Psychology*, *33*(8), 1072–1095. https://doi.org/10.1080/09515089.2020.1787972

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science*, *6*(2), 202–209. https://doi.org/10.1177/1745691611400238

Dean, R. (2010). Does neuroscience undermine deontological theory? *Neuroethics*, *3*(1), 43–60. https://doi.org/10.1007/s12152-009-9052-x

Evans, J., & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15. doi:10.1093/0199252866.003.0002.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

Gawronski, B., & Bodenhausen, G. V. (2015). Social-cognitive theories. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 65–83). Guilford Press.

Gawronski, B., & Brannon, S. M. (2020). Power and moral dilemma judgments: Distinct effects of memory recall versus social roles. *Journal of Experimental Social Psychology*, *86*, 103908. https://doi.org/10.1016/j.jesp.2019.103908

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2018). Effects of incidental emotions on moral-dilemma judgments: An analysis using the CNI model. *Emotion*, *18*(7), 989–1008. https://doi.org/10.1037/emo0000399

Gawronski, B., Conway, P., Hütter, M., Luke, D. M., Armstrong, J., & Friesdorf, R. (2020). On the validity of the CNI model of moral decision-making: Reply to Baron and Goodwin (2020). *Judgment and Decision Making*, *15* (6) , 1054–1072.

Gawronski, B., Luke, D. M., & Creighton, L. A. in press. Dual-process theories. D. E. Carlston, K. Johnson, & K. Hugenberg Eds., *The Oxford handbook of social cognition* 2nd. Oxford University Press.

Greene, J. (2003). From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology? *Nature Reviews Neuroscience*, *4*(10), 846–849. https://doi.org/10.1038/nrn1224

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

Greene, J. (2010). *Notes on 'The normative insignificance of neuroscience' by Selim Berker*. Unpublished Manuscript, Harvard University.

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

Greene, J. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, *124*(4), 695–726. https://doi.org/10.1086/675875

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J., Nystrom, L., Engell, A., Darley, J., & Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Heinzelmann, N. (2018). Deontology defended. *Synthese*, *195*(12), 5197–5216. https://doi.org/10.1007/s11229-018-1762-3

Hennig, M., & Hütter, M. (2020). Revisiting the divide between deontology and utilitarianism in moral-dilemma judgment: A multinomial modeling approach. *Journal of Personality and Social Psychology*, *118*(1), 22–56. https://doi.org/10.1037/pspa0000173

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, *142*(11), 1179–1203. https://doi.org/10.1037/bul0000075

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, *27*(1), 116–159. https://doi.org/10.1080/10463283.2016.1212966

Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind and Language*, *27*(5), 519–545. https://doi.org/10.1111/mila.12001

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Klein, C. (2011). The dual track theory of moral decision-making: A critique of the neuroimaging evidence. *Neuroethics*, *4*(2), 143–162. https://doi.org/10.1007/s12152-010-9077-1

Königs, P. (2018). On the normative insignificance of neuroscience and dual-process theory. *Neuroethics*, *11*(2), 195–209. https://doi.org/10.1007/s12152-018-9362-y

Körner, A., Deutsch, R., & Gawronski, B. (2020). Using the CNI model to investigate individual differences in moral-dilemma judgments. *Personality and Social Psychology Bulletin*, *46*(9), 1392–1407. https://doi.org/10.1177/0146167220907203

Körner, A., & Volk, S. (2014). Concrete and abstract ways to deontology: Cognitive capacity moderates construal level effects on moral judgment. *Journal of Experimental Social Psychology*, *55*, 139–145. https://doi.org/10.1016/j.jesp.2014.07.002

Kroneisen, M., & Steghaus, S. (2021). The influence of decision time on sensitivity for consequences, moral norms, and preferences for inaction: Time, moral judgments, and the CNI model. *Journal of Behavioral Decision Making*, *34*(1), 140–153. https://doi.org/10.1002/bdm.2202

Kumar, V., & Campbell, R. (2012). On the normative significance of experimental moral psychology. *Philosophical Psychology*, *25*(3), 311–330. https://doi.org/10.1080/09515089.2012.660140

Moll, J., & De Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, *11*(8), 319–321. https://doi.org/10.1016/j.tics.2007.06.001

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542. https://doi.org/10.1016/j.cognition.2005.07.005

Paulo, N. (2019). In search of Greene's argument. *Utilitas*, *31*(1), 38–58. https://doi.org/10.1017/S0953820818000171

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Science*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Prinz, J. (2016). Sentimentalism and the moral brain. In S. M. Liao (Ed.), *Moral brains: The neuroscience of morality* (pp. 45–73). Oxford University Press.

Rosas, A., & Koenigs, M. (2014). Beyond "utilitarianism": Maximizing the clinical impact of moral judgment research. *Social Neuroscience*, *9*(6), 661–667. https://doi.org/10.1080/17470919.2014.937506

Saunders, L. (2016). Reason and emotion, not reason or emotion in moral judgment. *Philosophical Explorations*, *19*(3), 252–267. https://doi.org/10.1080/13869795.2016.1212395

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76–105. https://doi.org/10.1016/0022-1031(91)90011-T

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458. https://doi.org/10.1016/j.cognition.2011.01.018

Thomson, J. J.(1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217. https://doi.org/10.5840/monist197659224

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., Västfjäll, D., & Espinosa, M. (2016). Intuition and moral decision-making: The effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PLOS ONE*, *11*(10), e0164012. https://doi.org/10.1371/journal.pone.0164012

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychologic distance. *Psychological Review*, *117*(2), 440–463. https://doi.org/10.1037/a0018963

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476–477. https://doi.org/10.1111/j.1467-9280.2006.01731.x

Zhang, L., Kong, M., Li, Z., Zhao, X., & Gao, L. (2018). Chronic stress and moral decision-making: An exploration with the CNI model. *Frontiers in Psychology*, *9*, 1702. https://doi.org/10.3389/fpsyg.2018.01702