COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018, Suppl. 1) 4–37

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12517

Two Models of Moral Judgment

Shane Bretz, Ron Sun

Department of Cognitive Science, Rensselaer Polytechnic Institute

Received 19 March 2016; received in revised form 23 April 2017; accepted 26 May 2017

Abstract

This paper compares two theories and their two corresponding computational models of human moral judgment. In order to better address psychological realism and generality of theories of moral judgment, more detailed and more psychologically nuanced models are needed. In particular, a motivationally based theory of moral judgment (and its corresponding computational model) is developed in this paper that provides a more accurate account of human moral judgment than an existing emotion-reason conflict theory. Simulations based on the theory capture and explain a range of relevant human data. They account not only for the original data that were used to support the emotion–reason conflict theory, but also for a wider range of data and phenomena.

Keywords: Simulation; Cognitive modeling; Cognitive architecture; Implicit; Explicit; Motivation; Moral judgment

1. Introduction

While morality refers to principles that govern right or wrong actions, ethics studies these principles. Some systems of ethics are based on "rational" analysis of rightness of actions (Kant, 1780) or preferability of outcomes (Bentham, 1781), whereas others emphasize sensory experiences, emotions, or intuitions as the origin of moral sentiments (e.g., Hume, 1738). This classical distinction between rationalism and sentimentalism continues to influence contemporary studies of moral judgment. For instance, early psychological theories of morality exemplified the rationalist perspective, emphasizing explicit reasoning (Kohlberg, 1969; Piaget, 1965), whereas more recent theories of moral judgment acknowledge both kinds of influences, although they differ on the relative importance of each (e.g., Greene, 2007; Haidt, 2001).

The trolley problem (Foot, 1978) has been used for testing these conflicting theories. For example, a version of the problem presents a runaway trolley car speeding down a

Correspondence should be sent to Ron Sun, Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: rsun@rpi.edu

track toward five railway workers (Thomson, 1985). A bystander stands at a switch, which may be pulled to divert the trolley onto a side-track, killing an individual on that track but sparing the five workers. This is considered a "high-conflict" moral dilemma because it presumably pits different moral principles against each other (Koenigs et al., 2007). In this case, it pits utilitarian calculation (the greatest good for the greatest number) against deontological principles (which concern duty and obligation; e.g., in no circumstance may one kill another).

Experiments using such high-conflict moral dilemmas have revealed various factors that influence moral judgment, which often fail to enter explicit (conscious) awareness (Cushman, Young, & Hauser, 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Wheatley & Haidt, 2005). Greene et al. (e.g., Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene et al., 2009) have proposed an emotion–reason dual-process theory that views emotion and rational calculation as two separate, conflicting processes: In high-conflict moral dilemmas, rational calculation may prefer a "utilitarian" solution (e.g., saving the five by killing one[1]), but emotion may prefer a "non-utilitarian" solution, and thus reason and emotion conflict. In this conflict, differences in situational contexts (as will be described later) influence the rate at which subjects approve of taking the "utilitarian" action.

In the remainder of this paper, first, a review of human data on the trolley problem is provided. Second, a framework based on the Clarion cognitive architecture is outlined, which leads to two contrasting models for moral judgment. Then, simulation studies are described in which the two models are assessed by their abilities to account for human data. Finally, the models are compared to other existing theories (both conceptual and computational).

2. Some important data of human moral judgment

The trolley problem detailed below is a laboratory task that reduces the complexity of real-world situations down to a precisely defined scenario to enable precise studies of behavioral data. Although its realism is limited (see, e.g., Bennis, Medin, & Bartels, 2010; Bloom, 2011), some of the issues that it addresses are important and readily applicable to real-world situations (Foot, 1978; Thomson, 1985).

The experiments of Greene et al. (2009) tested eight variations of the trolley problem to identify factors that affected moral judgment. Factors were tested by comparing subjects' Likert-scaled judgments of moral appropriateness of taking the "utilitarian" action in different versions of the dilemma. Explanations of these versions can be found in Table 1, along with the factors tested.

For example, as shown in Table 1, the *standard footbridge* scenario places the subject on a footbridge overlooking a trolley track. A runaway trolley is speeding towards five workers at the end of the track, and the subject may push a large man standing on the footbridge onto the track, killing the man but stopping the trolley and saving the five workers (only the large man can stop the trolley; there is no alternative). Another version,

Table 1
The dilemmas used in Greene et al.'s (2009) experiments. Dilemma set 1 examines contact, proximity, and personal force. Dilemma set 2 examines personal force and intention

Eight Dilemmas	
(1) Standard Footbridge	A bystander is on a footbridge overlooking the trolley tracks. The bystander may push a large man on the footbridge onto the tracks, stopping a runaway trolley and saving five workers at the end of the track, but killing the large man
(2) Footbridge Pole	Similar to Standard Footbridge, but the bystander may use a pole to push the large man onto the tracks
(3) Footbridge Switch	Similar to Standard Footbridge, but the bystander may throw a switch to drop the large man onto the tracks
(4) Remote Switch	Similar to Footbridge Switch, though the bystander and the switch are located far away from the large man
(5) Loop	The bystander may throw a switch, diverting the trolley onto a side-track that reconnects with the main track. A large man on the side-track will be killed, but he will stop the trolley
(6) Loop Weight	Similar to Loop, though a large weight behind the large man will stop the trolley. The large man is still killed
(7) Obstacle Push	The bystander may throw a switch diverting the trolley onto a terminal side-track. To reach the switch, the bystander must push a man off the bridge, killing him
(8) Obstacle Collide	Similar to Obstacle Push, though the bystander merely incidentally collides with the man, sending him over the bridge, killing him

Factors Tested

	Dilemma Name (#)	Factors Tested
Dilemma Set 1 (1, 2, 3, and 4)	Standard Footbridge (1)	Contact, Proximity, Personal Force, Intention
	Footbridge Pole (2)	Proximity, Personal Force, Intention
	Footbridge Switch (3)	Proximity, Intention
	Remote Switch (4)	Intention
Dilemma Set 2 (5, 6, 7, and 8)	Loop (5)	Intention, No Personal Force
	Loop Weight (6)	No Intention, No Personal Force
	Obstacle Push (7)	Intention, Personal Force
	Obstacle Collide (8)	No Intention, Personal Force

Note. Explanations of the factors: Contact—Whether the actor must make physical contact with the *one* sacrificed; Proximity—Whether the *one* sacrificed is in close proximity to the actor; Personal Force—Whether the actor must emit bodily force upon the *one* sacrificed regardless of whether through direct contact or through transfer (e.g., using a pole, a bat, etc.); Intention—Whether the *one* sacrificed is killed on purpose, intended as a means of achieving saving the five, or the death of the *one* is simply a consequence of saving the five.

the *footbridge switch* scenario, is identical to the *standard footbridge* scenario except that the subject can drop the large man onto the track by pulling a switch (rather than pushing him). The difference in the subject's endorsement of the "utilitarian" action in these two different scenarios arguably represents the effect of using "personal force" (force emanating from one's own body applied onto the victim). Other factors may be similarly analyzed as shown in Table 1.

It has been found that if the action requires the subject to "intentionally" kill (intended as a means to an end, as opposed to as a side-effect or accident) and if the subject must use personal force (i.e., force from his/her own body, as opposed to through a switch or a button), then the subject is less likely to judge such an action appropriate. See Fig. 1.

Specifically, in the first set of dilemmas (1–4), they found a significant effect of personal force (dilemmas 2 vs. 3, $p = .006$), no effect of contact (dilemmas 1 vs. 2, $p = .23$), and no effect of spatial proximity (dilemmas 3 vs. 4, $p = .74$), as shown by planned pair-wise contrasts.

In the second set of dilemmas (5–8), they found an interaction of personal force and intention ($p = .006$). The dilemma with both personal force and intention (dilemma 7) differed significantly from the others ($p = .004$, .02, .0006), while the dilemmas with one or the other factor alone did not statistically differ from one another (dilemmas 5, 6, and 8; $p > .2$), as shown by planned contrasts.

Yet another effect (observed by Greene et al.) might be termed the effect of "locus of intervention," between the first (1–4) and the second (5–8) set of dilemmas: The first set was less acceptable than the second. In the first set, the victim was displaced and used as a direct means to stop the trolley, while in the second set, the victim is either not displaced or not used as a direct means.

According to Greene et al., the source of these effects was implicit (unconscious) emotional responses to moral violations: Subjects implicitly recognized violations (such as personal force and intention) without identifying them explicitly (Cushman et al., 2006; Hauser et al., 2007). They argued that "non-utilitarian" decisions in these dilemmas were driven by emotion, which was fast, ingrained, and often without explicit awareness of its exact trigger. By contrast, "utilitarian" decisions were driven by rational cost–benefit analysis, which was slow, conscious, and deliberate.

Separately, in Greene et al. (2008), the effect of cognitive load on moral judgment was examined. Subjects were asked if they would perform a "utilitarian" action (harming one

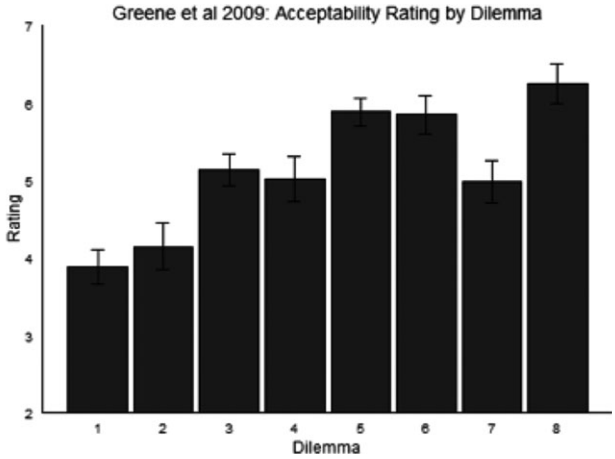

Fig. 1. Human data of acceptability ratings from Greene et al. (2009).

to save many) in a variety of dilemmas. To introduce cognitive load, they asked subjects to monitor a stream of digits presented concurrently with moral judgment and report the presence of a target digit.

For high-conflict "*personal*" moral dilemmas (a subset of high-conflict moral dilemmas described before[2]), without load, reaction times for utilitarian judgments did not differ from non-utilitarian judgments (p = .91). Under load, utilitarian judgments took significantly longer (p = .001). While reaction times for non-utilitarian judgments remained the same between no load and load conditions (p = .75), RTs for utilitarian judgments increased under load (p = .002). These findings pointed to a specific interfering effect of cognitive load on utilitarian judgments (see Fig. 2, bottom).

However, the results above were specific to high-conflict "personal" moral dilemmas. For "*impersonal*" moral dilemmas (Greene et al., 2008; supplementary materials), an

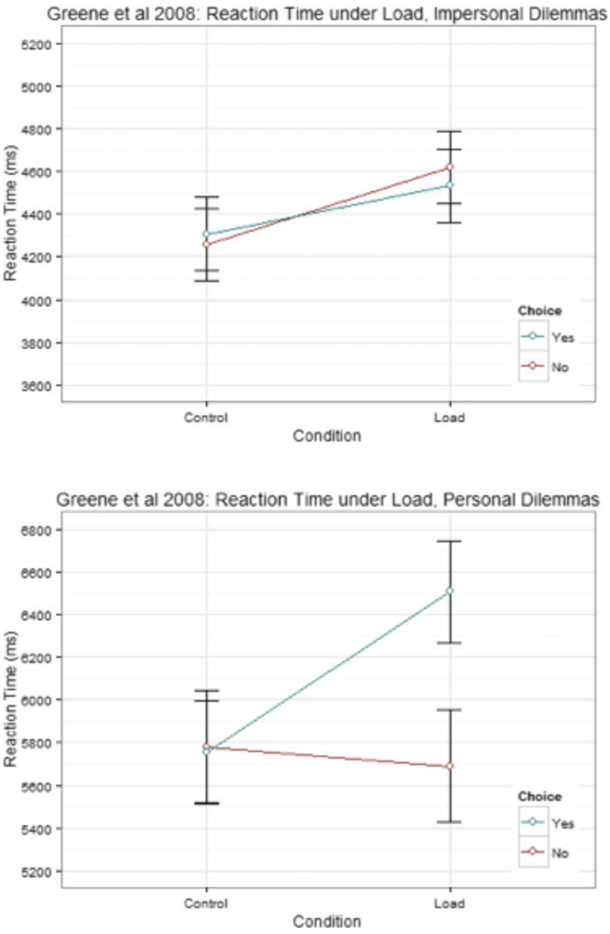

Fig. 2. Human reaction time data from Greene et al. (2008), for impersonal (top) and personal (bottom) moral dilemmas. "Yes" indicates "utilitarian" responses; "no" indicates "nonutilitarian" responses.

effect of load (p = .03), no effect of judgment (p = .85), and no interaction (p = .53) were found. In this case, load increased RTs for both utilitarian and non-utilitarian judgments, with no interaction (see Fig. 2, top).

Greene et al. (2008) argued that in personal dilemmas, load affected only utilitarian judgments because these judgments required explicit (rational) cognitive processes (i.e., explicit reasoning), which were slowed down by load. However, their emotion–reason conflict theory could not explain the RT data for impersonal dilemmas (Greene et al., 2008): In impersonal dilemmas, both utilitarian and non-utilitarian judgments were slowed down under load. (There are also other psychological data and phenomena that the emotion–reason conflict theory, as is, cannot easily explain; more on this later.)

Below we will look into how we can explain and computationally model these effects above, as well as a number of other relevant phenomena to be discussed later.

3. The Clarion cognitive architecture

Clarion is a cognitive architecture, that is, a relatively comprehensive and generic theory of psychological processes, with its corresponding computational instantiations (Sun, 2002, 2003, 2016). Thus, a variety of human data concerning moral judgment can be simulated computationally. Clarion has been well validated (e.g., Hélie & Sun, 2010; Sun, Merrill, & Petersen, 2001; Sun, Slusarz, & Terry, 2005). It consists of a number of subsystems in a dual-representation, dual-process architecture. In particular, Clarion accounts for basic human motivations that provide the underlying basis for behavior. This emphasis on motivation helps to integrate general cognitive capacities with motivational considerations (as well as personality, emotion, sociality, and culture; Sun & Mathews, 2012; Sun & Wilson, 2014; Wilson & Sun, 2014), which has significant implications for explaining human morality (Sun, 2013).

3.1. Basic assumptions

The Clarion theory includes a set of basic assumptions as follows (Sun, 2002, 2016). First, there is the distinction between implicit and explicit processes, which has been argued extensively before in the literature, both in relation to Clarion (e.g., Hélie & Sun, 2010; Sun, 2002, 2012; Sun et al., 2005) and to psychology generally (e.g., Evans & Frankish, 2009; Reber, 1989). In light of these extensive extant treatments, we will not repeat the arguments here. Generally speaking, explicit processes are consciously accessible, are more deliberative and controlled, and require more attention. Implicit processes, by contrast, are more automatic and less effortful, taking place outside of conscious awareness (Sun, 2002).[3]

Second, there is also a distinction between procedural (action-centered) and declarative (non-action-centered) processes (e.g., Anderson & Lebiere, 1998; Squire, 1987). Procedural processes denote those that select and carry out actions and skills, while declarative processes denote the utilization of general knowledge. Although some have claimed that procedural knowledge is implicit while declarative knowledge is explicit (or variations thereof), extensive arguments have been made for two orthogonal distinctions (see, e.g., Sun, 2012, 2016).

In Clarion, procedural and declarative processes are located in separate subsystems (Sun, 2012). Furthermore, each subsystem contains both an explicit and an implicit component (termed “level”; Sun et al., 2005). In this way, the distinctions between explicit and implicit processes and between procedural and declarative processes are separate and orthogonal.

A third assumption is that motivation involves explicit goals and implicit drives (Sun, 2003, 2009, 2016). Drives capture internally felt needs, both physiological and social, which are activated by internal or external contexts and contribute to the selection of goals and behaviors (Sun, 2009).[4] Drives may be categorized based on whether they are approach or avoidance oriented (Carver, 2006; Gray, 1987). On the basis of activation of these drives, explicit goals may be chosen. Explicit goals determine specific courses of actions. They also help to integrate other information (e.g., reasoning over situational contexts). Implicit and explicit motivations interact with each other and with other processes (Adams, Wright, & Lohr, 1996; Son Hing, Chung-Yan, Hamilton, & Zanna, 2008), allowing for complex, nuanced behavior. These points have been argued in prior work cited above. The relevance of motivation to cognition has been shown by, for example, Simon (1967), Markman and Maddox (2005), and Wilson, Sun, and Mathews (2009).

A fourth assumption concerns metacognitive regulation (based on motivation; Sun & Mathews, 2012; Sun, 2003, 2016). According to Clarion, metacognition includes a number of functions, such as goal selection and parameter setting, as well as the determination of cognitive “modes” (the degree of reliance on explicit or implicit processes; e.g., based on avoidance-oriented drives). According to Clarion, metacognitive and motivational processes are closely tied and motivational underpinnings are key factors in metacognitive regulation (as addressed in Sun, 2016).

3.2. Overview

In accordance with these assumptions, a general sketch of Clarion is shown in Fig. 3. Clarion consists of four major subsystems: the action-centered subsystem (ACS) for dealing with action selection involving procedural knowledge and processes, the non-action-centered subsystem (NACS) for reasoning and memory involving declarative knowledge and processes, the motivational subsystem (MS) for capturing drives and goals, and the metacognitive subsystem (MCS) for regulating other subsystems.

Each subsystem consists of two “levels” of representations (i.e., a dual-representational structure, as explained earlier). In each subsystem, the “top level” encodes explicit knowledge and carries out explicit processes; the “bottom level” encodes implicit knowledge and carries out implicit processes. The two levels interact, for example, by cooperating in action through generating integrated action recommendations (Sun et al., 2001, 2005).

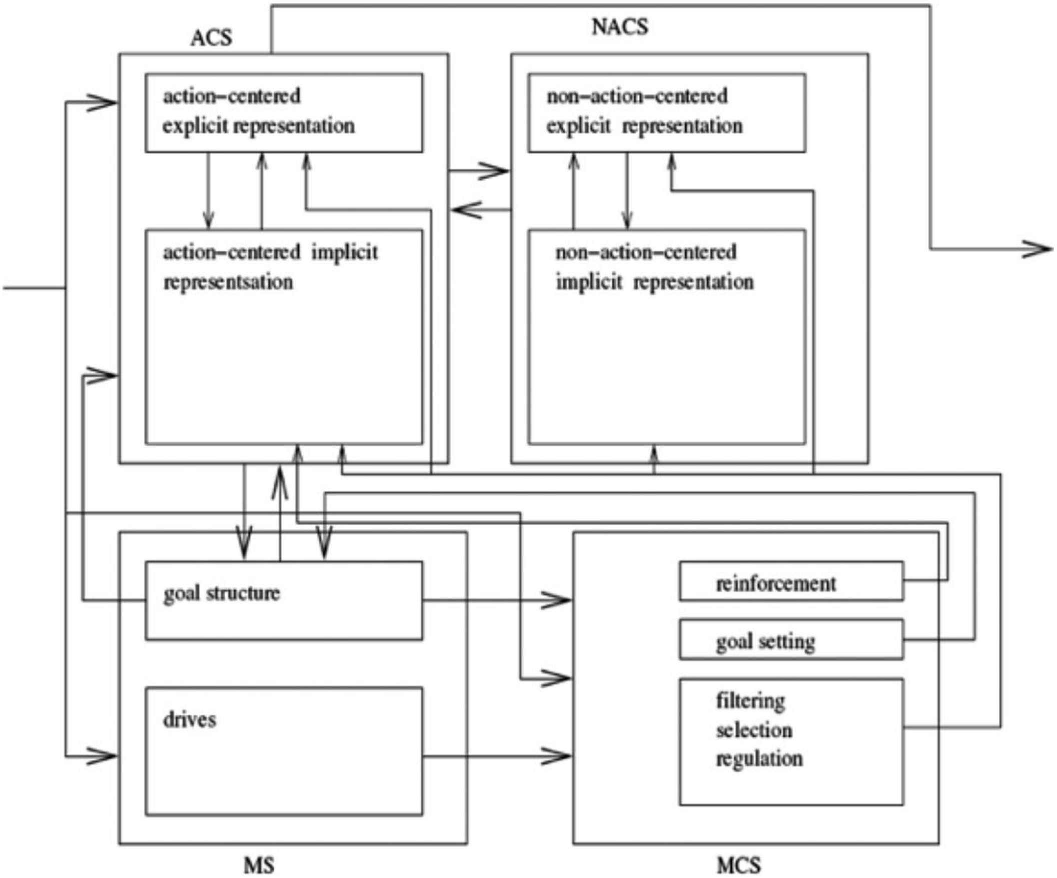
ACS
NACS
action–centered explicit representation
action–centered implicit representsation
non–action–centered explicit representation
non–action–centered implicit representation
goal structure
drives
reinforcement
goal setting
filtering selection regulation
MS
MCS

Fig. 3. The subsystems of the Clarion cognitive architecture. The major information flows are shown with arrows. See the text for explanations.

Clarion has thus far provided explanations for empirical psychological findings in a wide range of domains (see, e.g., Hélie & Sun, 2010; Sun & Wilson, 2014; Sun et al., 2001, 2005; Wilson et al., 2009). For example, a variety of well-known skill learning tasks have been simulated and explained using Clarion, some of which are highly implicit while others involve more explicit processes. Simulations have also been done with reasoning tasks, metacognitive and motivational tasks, and social simulation tasks. An important characteristic of Clarion is its focus on the cognition-motivation-environment interaction, as opposed to dealing only with cognition in the narrow sense. This focus is of importance to explaining moral judgment (as will be discussed).

3.3. Subsystems

Below, a general sketch of Clarion is provided; the Appendix provides some technical details (e.g., those needed for simulating moral judgment; for a full specification of

Clarion, see Sun, 2016). Each of the four subsystems will be described in turn. Their computational mechanisms and processes have been previously justified extensively on the basis of psychological data, which will not be repeated here (see Sun, 2002, 2003, 2016).

3.3.1. Action-centered subsystem

The action-centered subsystem captures procedural processes, in a dual representational form: Its bottom level carries out implicit procedural processes, whereas its top level carries out explicit procedural processes.

The algorithm for action selection may be described informally as follows (Sun, 2002, 2016): One first observes the current state of the world. The bottom level of the subsystem automatically computes a "value" for each possible action based on the current state. At the top level of this subsystem, rules determine possible actions (and their values of either 0 or 1) based on the current state. A final action is chosen by combining the values resulting from the two levels and performing a stochastic selection based on the combined values. Once the chosen action is performed, feedback is received and utilized for learning. The cycle of perception and action begins anew.

In this subsystem, the bottom level is implemented using a Backpropagation neural network, with distributed connectionist representation (Rumelhart & McClelland, 1986). The top level is implemented utilizing action rules with symbolic-localist representation (Sun, 1994). Action values from the two levels are combined using a weighted sum. Stochastic selection of an action is done using a Boltzmann distribution with a "temperature" parameter that determines stochasticity (see Appendix). Various types of learning (connectionist or symbolic) are carried out at each level or across levels. See Appendix for further details of this subsystem.

3.3.2. Non-action-centered subsystem

The non-action-centered subsystem captures declarative processes in a dual representational form. The top level of the subsystem contains explicit "associative rules" (in a symbolic-localist form), while the bottom level consists of implicit "associative memory" neural networks. The operation of this subsystem is directed by the ACS, which may request reasoning as part of its action. This subsystem is not needed for the simulations below, so its description is abbreviated here (see Sun, 2016).

3.3.3. Motivational subsystem

The motivational subsystem is concerned with why one does what one does. This subsystem focuses action in ways relevant to one's survival and functioning in the world (Sun, 2009). This subsystem has far-reaching effects: Its goals direct action selection in the ACS and reasoning in the NACS; it also influences regulatory functions in the MCS (Sun, 2003, 2016).

Dual representation is involved: Explicit goals derive from, and hinge upon, implicit drives. For example, an explicit goal "*find food*" may be generated based on an implicit drive "*hunger*" (Sun, 2009). See Fig. 4 for a sketch.

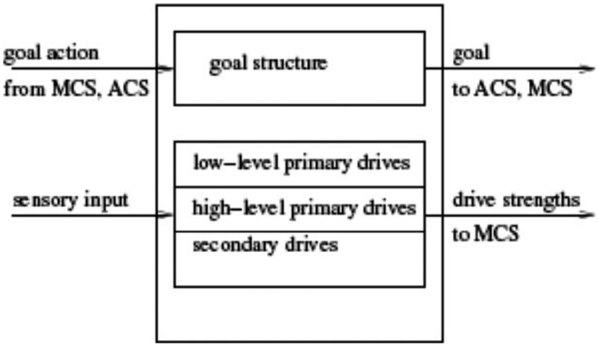

Fig. 4. The structure of the motivational subsystem. See the text for details.

Primary drives are essential to an individual (Maslow, 1943; Murray, 1938; Reiss, 2010). They represent basic physiological and psychological needs, mostly formed through evolution. Low-level primary drives concern mostly basic physiological needs (e.g., *Food, Water,* and *Reproduction*). High-level primary drives represent mostly socially oriented psychological needs (e.g., *Dominance and power, Fairness*, and *Similance*; Sun, 2009). The Appendix provides specifications of these drives. These drives have been justified based on work in social psychology and ethology regarding intrinsic needs and motives (see Sun, 2009, 2016; see also Maslow, 1943; Murray, 1938; Reiss, 2010).

A distinction between approach- and avoidance-oriented drives was alluded to earlier. It has been argued (e.g., Gray, 1987) that there exists a distinction between a behavioral approach system (BAS) and a behavioral inhibition system (BIS). Others have similarly argued for such distinctions (e.g., Cacioppo, Gardner, & Berntson, 1999; Clark & Watson, 1999). The BAS is sensitive to cues signaling reward, resulting in active approach and characterized by positive affect. The BIS is sensitive to cues of punishment, results in avoidance, and is characterized by anxiety or fear. The distinction between approach- and avoidance-oriented drives (see Appendix) provides an underlying basis for the distinction between the BAS and BIS.

Processing of these drives involves modules within the MS (Sun, 2016). Roughly, the activation (strength) of a drive is determined by the product of *stimulus* (the input to the drive) and *deficit* (the internal inclination toward activating the drive). Further details of drive processing are provided in the Appendix.

3.3.4. Metacognitive subsystem

In the metacognitive subsystem, metacognitive control and regulation are in the forms of: (1) setting goals (which are then used by the ACS) on the basis of drives, (2) generating reinforcement signals for learning (within the ACS) on the basis of drives and goals, (3) setting essential parameters of the ACS and the NACS, and so on (Sun, 2003, 2016).

Structurally, this subsystem is divided into a number of functional modules (Sun, 2016). For instance, the goal selection module, in order to select a new goal, first determines goal strengths, which are then turned into a probability distribution, from which a new goal is chosen stochastically (see Appendix for details).

For another instance, the processing mode module decides the weighting of the two levels of the ACS (for computing a weighted sum, as discussed earlier). Based on an inverted U curve (Fig. 5), the maximum strength of avoidance-oriented drives determines the weight of the top level (explicit processing). Note that the relationship between arousal and performance following an inverted U curve has long been noted (Yerkes & Dodson, 1908). Humphreys and Revelle (1984) explored such a curve relating anxiety to performance. Wilson et al. (2009) explored modulation of explicit processes (thus performance) by avoidance-oriented drives (anxiety) using the curve.

4. Two models of moral judgment based on Clarion

Below, based on Clarion, two models of moral judgment are sketched.

4.1. Sketch of Model 1

Our first model attempts to provide, within Clarion, a simple yet reasonably faithful implementation of Greene et al.'s (2008, 2009) emotion–reason conflict theory. We leverage Clarion's two-level, dual-process framework to capture the conflict between emotional and rational processes as stipulated by Greene et al.

According to Greene et al., "non-utilitarian" decisions in high-conflict dilemmas are driven by emotion that is fast and ingrained, and an individual may be unaware of the exact trigger of the emotion (i.e., it may be implicit). By contrast, "utilitarian" decisions

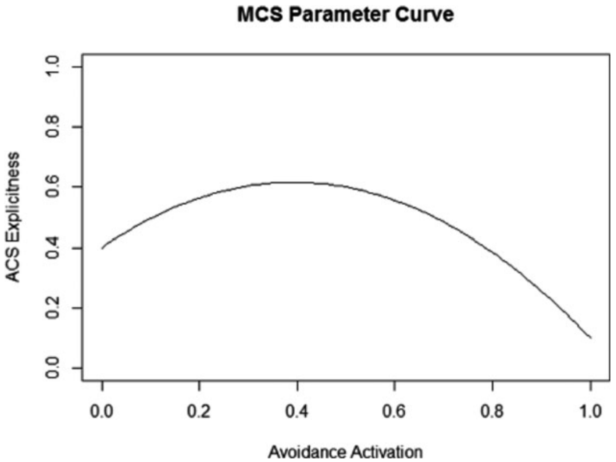

Fig. 5. The metacognitive subsystem determines the weight for the top level of the action-centered subsystem based on activation of avoidance-oriented drives. Bottom-level weight is 1—top-level weight.

are driven by rational cost–benefit analysis that is slow, conscious, and deliberate (i.e., explicit). For capturing this theory as it was described by Greene et al., the model needs only two types of processes within the action-centered subsystem (i.e., the top and the bottom level). Within the subsystem, the bottom level captures the emotion described by Greene et al., while the top level captures the rational processes. Thus, in correspondence with Greene et al.'s theory, when contextual factors favor emotion, processes at the bottom level are more prominent. Otherwise, processes at the top level are more prominent. The cross-level integration mechanism within the subsystem addresses the conflict and competition between them (see Section 3). No other component of Clarion is necessary, because the theory does not involve more than two types of processes. Further details will be presented in the next section.

Of course, this implementation of Greene et al.'s theory is, in a sense, forcing the theory into the Clarion cognitive architecture. Thus, some might argue that it falls short of providing an accurate computational instantiation of the theory. However, we could not find any substantial discrepancy between this model and the original theory (besides the fact that emotion maps to the bottom level and reason to the top level).

4.2. Sketch of Model 2

On the other hand, Model 2 is aimed at providing a more detailed, more nuanced, and more motivationally based account of moral judgment, directly derived from the basic assumptions and the basic framework of Clarion described earlier. Thus, more subsystems of Clarion are involved in Model 2.

One basic idea behind Model 2 is that the dynamics of moral judgment may be deeper and more complex than a simple emotion-reason conflict. A more detailed, motivationally based approach may be more suitable for accounting for nuances of this dynamics.

Specifically, somewhat similar to Model 1, in Model 2, moral judgment results from a combination of implicit and explicit processes. However, insofar as cognition in general is driven by motivation, moral judgment too is rooted in motivational processes: That is, moral judgment is the result of implicit and explicit cognitive processing guided by drives and goals. Motivations may be either approach- or avoidance-oriented. In particular, avoidance-oriented drives determine (in part) the degree of explicit processing within the action-centered subsystem, among other possible consequences such as selecting avoidance-oriented goals and actions (Section 3).

We can justify specifically the role of motivation in Model 2 (besides the general justifications in Section 3). Bloom (2011) objected to the lack of motivational and cultural foundations in many studies of morality. Hoffman (1979) believed that sentiments such as empathy, guilt, or anxiety provided "motive force" behind moral action. Moll, de Oliveira-Souza, and Zahn (2008) proposed a six-component framework underlying moral emotions; some of these components, such as attachment, aggressiveness, and dominance, closely resemble drives in Clarion (Sun, 2009). Tomasello and Vaish (2013) discussed the origin of morality as resulting from motivations for social cooperation, which competed with self-centered motivations for need fulfillment (Sun, 2009).

In terms of linkage between type of motivation and type of moral judgment, Janoff-Bulman, Sheikh, and Hepp (2009) found that subjects who were primed with reward seeking (approach) cues were more likely to make prescriptive moral declarations ("one *should*" help, care for, etc.), while those primed with punishment avoidance cues were more likely to make proscriptive declarations ("one *should not*" steal, cheat, etc.). Additionally, subjects' BAS scores (approach orientation) were significantly correlated with the extent to which subjects endorsed prescriptive moral actions, while BIS scores (avoidance orientation) were correlated with the extent to which subjects endorsed proscriptive moral actions.

We may also link implicit and explicit processes to types of moral judgments in a way that is more general than Greene et al.'s theory. In general, implicit processes may not be fully equipped to perform detailed cost–benefit ("utilitarian") analysis necessary to arrive at fully contextualized decisions. Therefore, implicit processes are more likely to represent deontological principles that are evolutionarily acquired or deeply culturally ingrained. For example, the trolley problem may require the subject to recognize the tradeoff between the lives of five railway workers and one bystander and, realizing this tradeoff, conclude that regardless of the moral acceptability of the action itself, the consequence of the action is preferable to that of not acting. Reasoning of this kind often places demands on attention and working memory and requires explicit processes (Evans & Frankish, 2009; Piaget, 1965), although it is possible that some simple forms of cost–benefit calculation might be performed implicitly.

On the other hand, for an individual, fundamental deontological principles may be biologically pre-endowed to some extent to begin with or inculcated over a long period of time, and therefore they tend to be implicit and embodied (Allchin, 2009; Sun, 2009). They tend to be fast in their application. They may be more concerned with characteristics of actions per se rather than nuances of contexts or consequences (Cushman, Gray, Gaffey, & Mendes, 2012).

However, moral principles need not be bound to implicit processes: They may be learned and applied explicitly (Kohlberg, 1969; Royzman, Landy, & Leeman, 2015). Thus, within explicit processes, deontological principles and cost–benefit calculations may compete (Kahane, 2012). However, given a rationalistic cultural milieu, rational analysis may often win at the explicit level (Kohlberg, 1969; Piaget, 1965).

Therefore, explicit processes need not be limited to rational cost–benefit analysis; for example, one may follow an explicitly stated deontological principle of not doing any harm, or adopt an explicit goal of not doing any harm (Kohlberg, 1969). Conversely, implicit processes need not be limited to "non-utilitarian" responses; for example, one may follow an internalized behavioral routine to help others, or act in response to a goal to help others (e.g., due to an intrinsic need to help others; Baron, Gürçay, Moore, & Starcke, 2012).[5] Various empirical results, for example, from Baron et al. (2012), Trémolière and Bonnefon (2014), Millar, Turri, and Friedman (2014), and Royzman et al. (2015), support such non-exclusivity.

Although this account of moral judgment by Model 2 (as a direct application of the Clarion framework) bears some resemblance to Greene et al.'s account, they differ in a few critical ways. First, Model 2 posits that moral cognition is driven by motivation, which influences both implicit and explicit cognitive processes (Sun, 2013). It thereby offers an account of motivational aspects of moral judgment. Moreover, it posits that the locus of moral judgment primarily lies in motivational dynamics (more later in Section 5).

The second major difference lies in the handling of relationship between two types of processes. As discussed above, both the Clarion theory and Greene et al.'s theory (both Model 1 and 2) are dual-process theories where one type is explicit (conscious) and the other implicit. While Greene et al. proposed direct competition between emotion and reason as the basic explanatory framework, we are against this simplistic, folk psychological notion. As discussed above, according to Clarion, explicit or rational processes do not necessarily lead to "utilitarian" responses, and implicit or emotional processes do not necessarily lead to "non-utilitarian" responses. In Model 2, both implicit and explicit processes reach a conclusion in accordance with an individual's motivations to some extent, and motivational dynamics can affect the amount of reliance on explicit processes (Section 3), which can actually account for empirical differences found between different situational contexts (e.g., those discussed in Section 2; more on this later).

Furthermore, in Greene et al.'s emotion–reason conflict theory, implicit processes relevant to morality consisted only of emotional responses; different responses to different dilemmas were explained by differences in emotional engagement (Greene et al., 2008, 2009). However, characterizing implicit moral judgment as the contribution of emotion alone falls short in terms of describing the wider scope of moral psychology. Some implicit intuitions and instincts relevant to morality may not be emotion based (see, e.g., Hélie & Sun, 2010; Kahane, 2012; Monroe, 2012; Sun & Wilson, 2014). As Monroe (2012) put it, implicit moral processes include "deep-seated instincts, predispositions, and habitual patterns of behavior," coming from "genetic predispositions, social roles, or culturally inculcated norms." Furthermore, although psychopaths possess greatly depressed harm-aversive affective reactions or feelings of guilt, they show almost normal moral judgment (while acting immorally; Blair, 1995; Herpertz & Sass, 2000).

Contrary to the folk psychology of a strict emotion–reason dichotomy, it has been known from relevant research that emotion involves both reactive ("physiological") processes (Zajonc, 1980) and "cognitive" appraisal (Frijda, 1986; Lazarus, 1991). In other words, emotion may involve reason and explicit processes to a significant extent. Conversely, reason is also affected often significantly by emotion and motivation (see, e.g., Blanchette & Caparos, 2013; Gubbins & Byrne, 2014; Kunda, 1990; Lodge & Taber, 2013), as well as by other implicit processes (Evans & Frankish, 2009; Reber, 1989). The upshot is that, contrary to the folk psychology, emotion and reason are not fully separable; nor do they map neatly onto implicit and explicit processes.

Moreover, emotion and reason are not elemental processes. They each involve a large set of mechanisms, often overlapping. Thus, they may not be the best, clearest explanatory constructs. Cognitive science seeks to replace such folk psychological notions with more precise, more fine-grained, and better grounded ones. The motivationally based account from Clarion may explain a broader range of empirical results while offering a more nuanced view (see Sections 5 and 6).

Finally, although motivation and emotion are often regarded as highly intertwined, they are not identical. For example, motivation may explain why psychopaths show almost normal moral judgment (while acting immorally), although they lack relevant affective reactions, while Greene et al.'s (2009) emotion–reason conflict theory cannot easily explain this (see Section 6).

5. Simulations of the human data of moral judgment

This section describes how the two models fare in simulating the human data discussed in Section 2.

5.1. Simulation setup

5.1.1. Details of Model 1

As described earlier, Model 1 involves only the action-centered subsystem of Clarion.[6] Specifically, the bottom level of the subsystem consists of a Backpropogation neural network responding to those salient factors identified by Greene et al. (2009): intention, personal force, and locus of intervention (Section 2). Implicit processing of these contextual factors occurs with distributed representation. This part of the model satisfies Greene et al.'s concept of "emotional" response that may not be fully accessible to an individual.

The input to the neural network are features, each of which corresponds to a factor identified before, represented by an individual node. The output consist of two nodes representing "utilitarian" and "non-utilitarian" choices, respectively, as responses to the question of whether it is morally acceptable to sacrifice one life to save five lives.

The network was trained based on the analysis of the human data of Greene et al. (2009) concerning contextual factors involved in these dilemmas. The output node corresponding to the non-utilitarian response was trained to approximate a relation to the factors affecting judgments as identified by Greene et al.[7] (Specifically, $S = ax + B$ where S was the output strength, x was the number of those factors present, B was the base activation, and a was the amount of activation for each factor; $a = .25$, $B = .5$.[8]) The activation of the node corresponding to the utilitarian response was trained to be at a low base level *.05* (because, according to Greene et al., utilitarian conclusions result from explicit processes and not from the bottom level). Trained as such, the bottom level of the ACS corresponds to Greene et al.'s notion of emotional response.

The top level of the ACS captures the rational process in Greene et al.'s theory. It contains explicit rules that respond to utilitarian cost–benefit analyses. With symbolic-localist representation, these rules capture a consciously accessible and deliberative process (see Appendix), as identified by Greene et al. For instance, a rule indicating an action is appropriate is activated if the condition of saving five by killing one is met.[9]

Thus, both processes necessary for capturing Greene et al.'s theory are present. The recommendations of the two levels are combined (Section 3). To simulate the Likert rating, the combined activation of the "utilitarian" response normalized by the Boltzmann distribution (see Appendix) is used. Response time is simulated using the standard Clarion equation[10] (Sun, 2003, 2016).

To capture rating data from Greene et al. (2009), 80 simulated subjects were run for each of the eight dilemmas. To capture reaction time data from Greene et al. (2008), 50 simulated subjects were run in the control condition and another 50 under the load condition. Of the eight dilemmas, dilemmas 1 and 2 constitute high-conflict personal dilemmas, while 5, 6, and 8 constitute impersonal dilemmas. Dilemmas 3, 4, and 7 fall somewhere in between (in which some of the features of high-conflict personal dilemmas are present but not all). The load condition is captured through activation of rules concerning the distractor task.[11]

Parameter values involved in Model 1 as well as related specifications and other details can be found in Appendix.

5.1.2. Details of Model 2

According to Model 2, however, moral psychology is driven by motivation. Its locus lies primarily in the motivational dynamics within the motivational subsystem, where drives are activated by situational factors and explicit goals are set based on competition among drives. The action-centered subsystem generates responses in some accordance with the goals: Both explicit and implicit processes of the subsystem are influenced by chosen goals, although each may prefer different conclusions based on context. The metacognitive subsystem decides the weighting of explicit/implicit processes in the ACS based on motivation (Section 3).

Specifically, the MS consists of a set of drives that compete to determine goals. To keep the model minimal, two drives were used in this simulation, one of the avoidance type (e.g., *Honor*; see Appendix) and another of the approach type (e.g., *Nurturance*). Drives were implemented in a Backpropogation network and trained. The activation of each drive was determined by *stimulus* × *deficit* (see Appendix). The *stimulus* to the avoidance drive was determined by factors identified by Greene et al. (2009), using the same relation described earlier (with $B = .5$ and $a = .25$). The *stimulus* to the approach drive was set to a constant (*.5*), because the lives saved versus lost remained constant over these dilemmas. Each simulated individual was given a different drive *deficit* value, drawn from a normal distribution (for the approach and the avoidance drive separately; see Appendix), to capture individual differences.

The avoidance drive supports the goal "avoid doing harm," and the approach drive supports the goal "save lives" (Janoff-Bulman et al., 2009). Each drive leads to activation of the corresponding goal (see Appendix for details).

Within the ACS, the goal associated with the avoidance drive ("avoid doing harm") favors the "non-utilitarian" action; the goal associated with the approach drive ("save lives") favors the "utilitarian" action. This subsystem is configured somewhat similarly to Model 1, but notably with variable goals that do not stay the same as in Model 1. The bottom level of the ACS computes an implicit reaction based on the situational context (as in Model 1), but it also considers the chosen goal. The top level contains those cost–benefit rules as used in Model 1, but also rules that recommend actions corresponding to the chosen goal.

Although the bottom level of the ACS in Model 2 is somewhat similar to that in Model 1, the interpretation is quite different. In Model 1, the bottom level represents emotional responses stipulated by Greene et al. In Model 2, responses to the context are generated mostly from the motivational (and the metacognitive) subsystem, and the bottom level of the ACS instead represents implicit action schemas. The redundancy between the motivational and the action-centered subsystem captures multiple types of influences: Motivation influences action selection by specifying goals (desired outcomes), but there exists also aversion to harmful actions themselves regardless of the outcomes of those actions (as argued by, e.g., Cushman et al., 2012). The redundancy also captures different kinds of implicit processes: A motivational response is implicitly computed in the motivational (and the metacognitive) subsystem, while the bottom level of the ACS follows an implicit action schema (either learned or innate). Moreover, in Model 2, the bottom level of the ACS takes into consideration the chosen goal.[12]

The top level of the ACS captures explicit procedural processes. It contains explicit rules along the lines of cost–benefit analyses (as in Model 1). However, in addition, it also contains explicit rules that link a goal to actions that achieve the goal. These different types of rules compete against each other in recommending actions.

The relative weight of each level of the ACS (i.e., the degree of explicitness) is set by the MCS based on avoidance drive activation (see Section 3).

For simulating Greene et al. (2009), 80 simulated subjects were run for each of the eight dilemmas. For simulating RT data of Greene et al. (2008), 50 simulated subjects were run in the load condition and another 50 in the control condition.

Parameter values involved in Model 2 as well as related specifications and other details can be found in the Appendix.

5.2. Simulation results

5.2.1. Results of Model 1

Fig. 6 presents the simulated Likert ratings by Model 1, alongside the corresponding human data from Greene et al. (2009). The model captures the major effects in the human data.

For the first set of dilemmas (1–4), simulated ratings differed significantly ($p < .001$). Planned pairwise contrasts revealed no effect of spatial proximity (dilemma 3 vs. 4, $p = .644$) and no effect of physical contact (dilemma 1 vs. 2, $p = .729$). There was a significant effect of personal force (dilemma 2 vs. 3, $p < .001$). These results were the same as the human data.

For the second set (5-8), an interaction between intention and personal force ($p < .001$) was found in the simulated data. A follow-up analysis comparing the cases with either force or intention alone (dilemmas 5, 6, and 8) found no significant difference

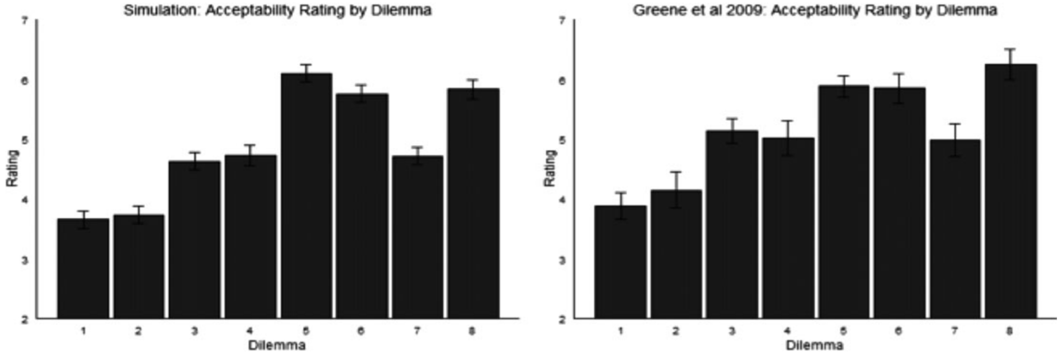

Fig. 6. Acceptability ratings from the Model 1 simulation data (left) and the human data (right).

($p > .3$), but the force plus intention case (dilemma 7) differed significantly from each of the other cases (all $p < .01$). Thus, the combined presence of intention and personal force lowered ratings significantly, as in the human data.

To analyze simulated reaction time data, three groups were considered: impersonal dilemmas (dilemmas 5, 6, & 8), personal dilemmas (dilemmas 1 & 2), and an intermediary group (dilemmas 3, 4, & 7, without human data to compare to).[13]

Fig. 7 (top left) shows the simulation data for impersonal dilemmas. There was a significant load by judgment interaction ($p < .001$). Planned contrasts showed that reaction times for utilitarian responses under load differed significantly from both utilitarian responses in the no load condition ($p < .001$) and non-utilitarian responses in the load condition ($p < .001$), different from the human data where utilitarian and non-utilitarian responses under load did not differ. Reaction times did not differ between non-utilitarian responses under load and no load ($p > .2$), again different from the human data. However, reaction times did not differ between utilitarian and non-utilitarian responses in the no load condition ($p > .2$), the same as in the human data.

Fig. 7 (bottom left) shows the simulation data for personal dilemmas. There was a significant load by judgment interaction ($p < .01$). Planned contrasts showed that reaction times for utilitarian responses under load differed significantly from both utilitarian responses in the no load condition ($p < .01$) and non-utilitarian responses in the load condition ($p < .01$), the same as in the human data. Reaction times did not differ between non-utilitarian responses under load and no load ($p > .2$), nor between utilitarian and non-utilitarian responses in the no load condition ($p > .2$), the same as in the human data.

5.2.2. *Discussion of Model 1 results*

While Model 1 captured some effects in the human data, a glaring discrepancy exists between the human and the simulated reaction time data for impersonal dilemmas. Model 1 predicted a similar pattern of reaction time data for personal and impersonal dilemmas, showing an effect of load on utilitarian, but not non-utilitarian, responses. This prediction was inconsistent with the human data, which showed that utilitarian and non-utilitarian

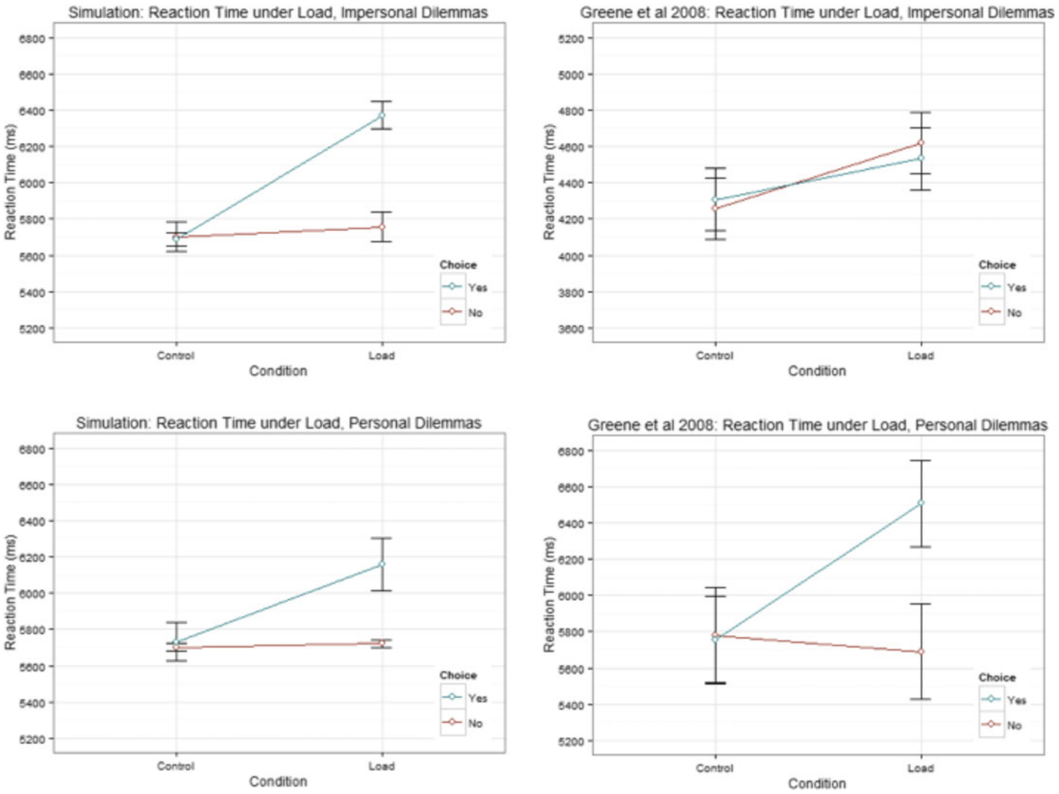

Fig. 7. Model 1 simulated (left) versus human (right) reaction time data for impersonal (top) and personal (bottom) moral dilemmas.

responses were similarly affected in impersonal dilemmas. The finding that both types of responses were affected by load in impersonal dilemmas in this experiment is important in our view (even though Greene et al., 2008, only considered personal dilemmas important).

If the human reaction time data could be explained by Greene et al.'s theory of emotional and rational processes, then we would expect one of three possibilities:

1 If the RT data could be explained by load affecting specifically explicit "rational" processes, then we would expect a similar effect in both personal and impersonal dilemmas and no effect of load on non-utilitarian responses (because non-utilitarian responses result from implicit "emotional" processes), as predicted by Model 1. Instead, in the human data on impersonal dilemmas, load affected RT generally.
2 If the RT data were to be explained by load affecting both rational and emotional processes, then we would expect load affecting non-utilitarian responses in personal dilemmas as well, which was not the case in the human data.
3 If the RT data were to be explained by *conflict* between two processes, with the conflict recruiting additional cognitive control (Greene, 2007), then we would expect no

effect of load in impersonal dilemmas, where such conflict was absent or significantly reduced. This was not the case in the human data.

Therefore, this discrepancy cannot be easily explained away within Greene et al.'s theory. Below, Model 2 provides a more nuanced explanation and a remedy for this discrepancy.

5.2.3. *Results of Model 2*

Simulation results from Model 2 were analyzed similarly. Fig. 8 presents the simulated ratings alongside the corresponding human data of Greene et al. (2009).

For the first set of dilemmas (1–4), simulated ratings differed significantly ($p < .001$). Planned pairwise contrasts revealed no effect of spatial proximity (dilemma 3 vs. 4, $p = .369$) and no effect of physical contact (dilemma 1 vs. 2, $p = .073$), the same as the human data. There was a significant effect of personal force (dilemma 2 vs. 3, $p < .001$), the same as the human data.

For the second set, an interaction between intention and personal force was found ($p < .001$), as in the human data. Follow-up analysis found no significant difference among the cases with force or intention alone (dilemmas 5, 6, & 8; $p > .3$), but the force plus intention case (dilemma 7) differed significantly from each of the other cases (all $p < .001$), the same as the human data.

Simulated reaction time results were also analyzed. Fig. 9 (top left) shows the results for impersonal dilemmas. There was a significant effect of load ($p < .001$), but no significant effect of judgment ($p > .2$). There was no significant load by judgment interaction ($p > .2$). Load increased reaction times for both utilitarian and non-utilitarian responses. This was the same as the human data in all aspects.

Fig. 9 (bottom left) shows the results for personal dilemmas. There was a significant load by judgment interaction ($p < .001$), the same as the human data. Planned contrasts showed that reaction times for utilitarian responses under load differed significantly from

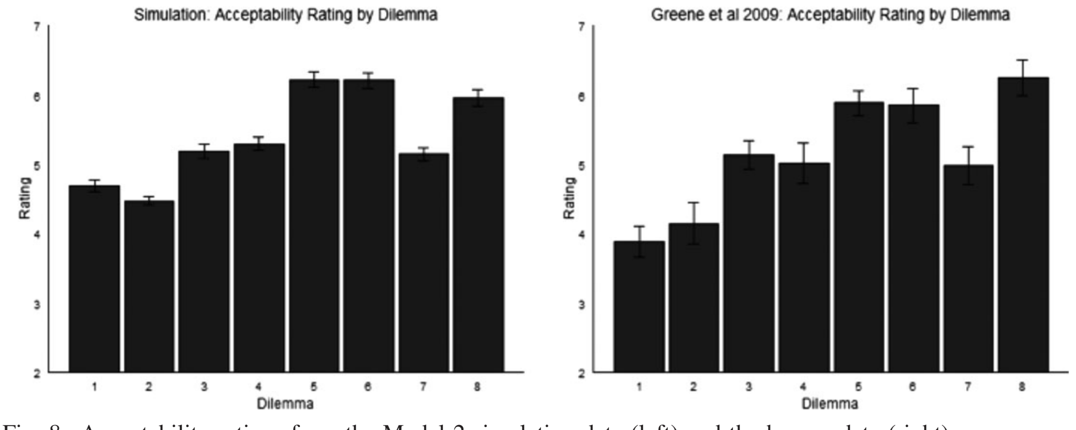

Fig. 8. Acceptability ratings from the Model 2 simulation data (left) and the human data (right).

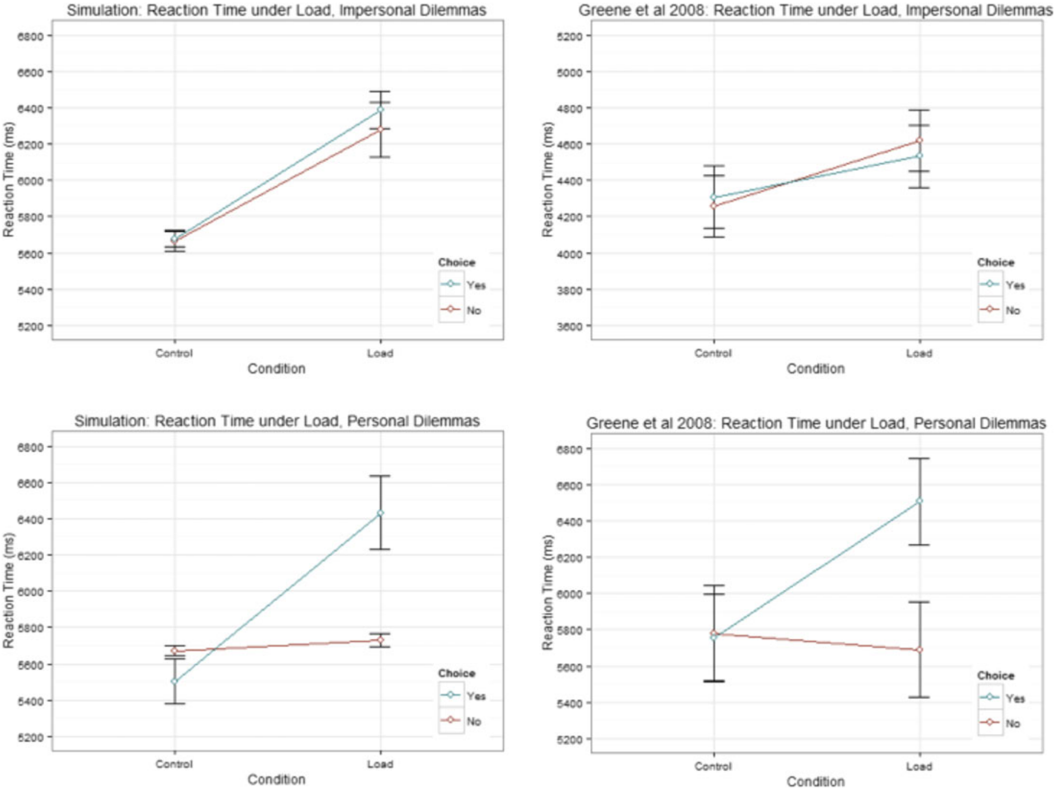

Fig. 9. Model 2 simulated (left) versus human (right) reaction time data for impersonal (top) and personal (bottom) moral dilemmas.

utilitarian responses in the no load condition (p < .001) and from non-utilitarian responses in the load condition (p < .001). Reaction times did not differ between non-utilitarian responses under load and no load (p > .2), nor between utilitarian and non-utilitarian responses in the no load condition (p > .2). All these effects match the human data.

5.2.4. Discussion of Model 2 results

Model 2 captures all the effects in the human data presented thus far (as well as other data discussed in the next section). It provides an explanation that is different from Model 1. Whereas it posits the distinction between implicit and explicit processes, it attributes the ultimate source of moral judgment to the underlying motivation. For example, a scenario requiring an individual to kill using personal force activates more strongly avoidance-oriented drives, which increase the likelihood that the individual produces, via avoidance goals, more negative judgments. In Model 2, implicit and explicit processes do not necessarily generate one or the other choice and each process may work in some accordance with activated motivations (even though explicit processes may have some utilitarian orientation and implicit processes may have some non-utilitarian orientation).

Model 2 succeeded in capturing all the significant effects in the reaction time data for impersonal dilemmas, where Model 1 notably failed. In Model 2, the RT differences were accounted for not solely by the tendency for utilitarian and non-utilitarian responses to come from different processes, but more by the effect of motivation on these processes. Specifically, metacognitive modulation of explicitness of cognitive processing occurred on the basis of motivation. Increased activation of avoidance-oriented drives led to a decrease in explicit processing (through the inverted U curve). In *impersonal dilemmas*, avoidance-oriented drives were relatively less activated (because these dilemmas involved fewer major moral violations). Thus, relatively more weight was given to explicit processes regardless of whether a utilitarian or a non-utilitarian response was chosen. With cognitive load, explicit cognitive processes were more likely to be affected (Sun et al., 2005). Thus, in impersonal dilemmas, both utilitarian and non-utilitarian responses were slowed down by cognitive load.

In *personal dilemmas*, however, relatively high avoidance-oriented drive activation (due to more major moral violations) caused more weight to be given to implicit processes. In these dilemmas, individuals with less inclination toward activating avoidance-oriented drives were more likely to generate utilitarian responses (Section 5.1.2; Janoff-Bulman et al., 2009). These individuals relied more on explicit processes (as determined by the inverted U curve, due to lower avoidance-oriented drive activation) and thus cognitive load affected their response times. Individuals with more inclination toward activating avoidance-oriented drives were more likely to generate non-utilitarian responses. These individuals relied less on explicit processes (as determined by the inverted U curve) and were less affected by cognitive load. Thus, in personal dilemmas, response times under load were increased for utilitarian responses and unchanged for non-utilitarian responses.[14]

Greene et al. (2008) argued that utilitarian responses resulted from explicit rational processes, which under load were slowed down, whereas non-utilitarian responses resulted from implicit emotional processes, which were not affected. In contrast, Clarion posits that implicit or explicit processes do not *exclusively* lead to non-utilitarian or utilitarian responses. The RT differences can be accounted for by metacognitive regulation based on motivation. In this way, Clarion resolves one major difficulty of the emotion–reason conflict theory.

A number of other differences between the Clarion theory and the emotion–reason conflict theory that have been alluded to earlier may be summarized here. First, Clarion offers an account of *motivational* aspects of moral judgment. It may be further argued that the locus of morality primarily lies in motivational dynamics, which influences implicit and explicit cognitive processes (Sun, 2013, 2016). Second, in terms of the *relationship* between two types of processes, while Greene et al. suggested competition and conflict between emotional and rational processes, we propose that two processes, implicit and explicit, work together to reach a conclusion (Kunda, 1990; Manfrinati, Lotto, Sarlo, Palomba, & Rumiati, 2013), taking into account an individual's motivation. Motivation also affects the degree of reliance on explicit or implicit processes.

Through Clarion, the notion of *emotion* may be spelled out as well. According to Clarion, emotions are generated primarily in response to the activation of motivations and the likelihood of satisfying the activated motivations; they likely lead to behaviors consistent with the activated motivations (Sun & Mathews, 2012; Wilson, 2012; Wilson & Sun, 2014). That is, emotions result from the evaluation of possibility of, and progress towards, satisfying outstanding basic needs or motives, both implicitly and explicitly (through reactive affect and deliberative appraisal, by the ACS and the NACS). However, the causal role of motivation is prior to emotion (i.e., motivation is a cause of emotion, as fleshed out in Wilson & Sun, 2014). Thus, motivational dynamics may be the ultimate determinant of both moral judgment and emotion. In that regard, Model 2 offers deeper explanations for moral judgment as well as for emotion (see Wilson, 2012; Wilson & Sun, 2014), besides capturing more accurately the experimental data discussed above.

Note that Model 2 is motivational because it relates moral judgment to essential motivational constructs (e.g., approach- vs. avoidance-oriented motivation). Furthermore, it can relate the motivational interpretation of moral judgment to similar interpretations of other motivational phenomena as well as emotion, personality, and other related aspects (see, e.g., Sun & Wilson, 2014; Wilson & Sun, 2014; Wilson et al., 2009).

6. Some further data and simulations

Model 2 accounts for many other moral judgment phenomena. Below a few examples are presented briefly (omitting details due to length considerations).

6.1. Effect of justification pressure

Justification pressure refers to the situation created by the requirement that subjects justify their moral judgments. Rai and Holyoak (2010) asked subjects to provide explicit justifications for choosing the “utilitarian” action in the standard trolley problem. Subjects who were asked to provide up to two justifications provided fewer justifications on average but, paradoxically, agreed more with the “utilitarian” action than did subjects asked to provide up to seven justifications ($p < .05$). Rai and Holyoak argued that this result contradicted the emotion–reason conflict theory of Greene et al., which, as is, would predict that more engagement of explicit rational processes elicited by increased explicit justification requirements would lead to more utilitarian judgments. Rai and Holyoak claimed that instead of an emotion–reason conflict, these results resembled domain-general effects (e.g., those associated with the availability heuristic).

As another test of the motivationally based theory (Model 2), we simulated the justification pressure effect. Model 2 accounts for this effect by viewing it as a motivational issue. That is, the requirement that a subject provides justifications gives rise to “anxiety” if the individual experiences difficulty in providing the solicited quantity or quality of justifications (analogous to the social evaluation pressure in Lambert et al., 2003 or Beilock & Carr, 2001; see Wilson et al., 2009). According to Clarion, such anxiety may increase the likelihood of non-utilitarian moral judgments in two ways. First, such anxiety, which is captured by activation of avoidance-oriented drives (Wilson et al., 2009), increases the likelihood of avoidance goals and thus avoidant (non-utilitarian) outcomes. Second, increased activation of avoidance-oriented drives may decrease explicitness of processing and thus utilitarian responses become less likely (as discussed in Section 4).

The Model 2 setup described before was used for this simulation. Increase of justification pressure was captured through an increase of the avoidance drive gain (see Appendix; Wilson et al., 2009), which led to increased activation of the avoidance-oriented drive and in turn led to increased avoidant behavior.

Fig. 10 shows the human and the simulation data together.[15] Statistical analysis showed that simulated ratings decreased significantly when more justifications were requested ($p < .001$), consistent with the human data. The effect, which was modest, was in line with Rai and Holyoak's modest findings (though effect size increased with increased parameter adjustment).

Greene et al.'s theory predicted that more engagement of rational processes through more justifications would elicit more utilitarian judgments (Rai & Holyoak, 2010). Thus, Model 1, as is, cannot capture this effect.

6.2. *Effect of mortality salience*

Trémolière, De Neys, and Bonnefon (2012) found, through a pair of studies, that reminding subjects of their mortality induced lower rates of utilitarian responses. They noted that emotion had been repeatedly found unassociated with mortality salience (e.g., Arndt, Allen, & Greenberg, 2001; Rosenblatt, Greenberg, Solomon, Pyszczynski, & Lyon, 1989), which contradicted Greene et al.'s theory. They offered an alternative explanation

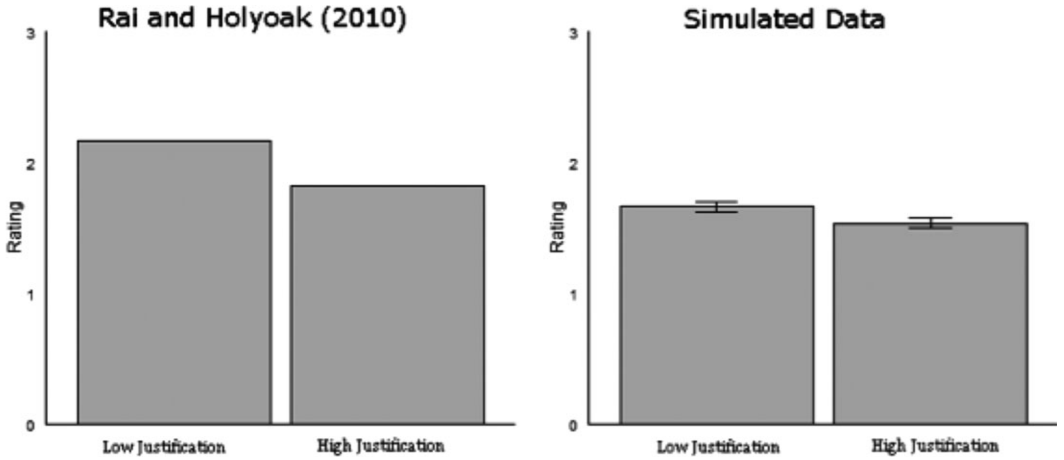

Fig. 10. The effect of justification pressure on ratings of acceptability. Human data and Model 2 simulation data are compared.

where the effect of mortality salience was to force a "switch from an analytic to an intuitive, experiential mindset," that is, to emphasize moral intuitions rather than explicit analysis.

The motivationally based theory (Model 2) accounts for the mortality salience effect. Similar to the account of justification pressure, it is posited that reminding subjects of their mortality increases activation of avoidance-oriented drives (e.g., the drive for avoiding danger, especially of an existential nature). Consequently, avoidance goals and avoidant ("non-utilitarian") actions or judgments are more likely to be chosen. Also, when activation of avoidance-oriented drives is high, the weight of explicit processes is reduced and thus utilitarian responses become less likely. The literature on mortality salience supports this interpretation, based on the close relationship between mortality salience manipulation and activation of the avoidance system (i.e., BIS; Strachan et al., 2007; Quirin et al., 2011; Tritt, Inzlicht, & Harmon-Jones, 2012).

We thus simulated this effect with Model 2. The details of the simulation setup were identical to those that captured justification pressure above and thus are omitted here. Model 2 successfully captured the mortality salience effect ($p < .001$).

As mentioned above, emotion has been found unassociated with mortality salience (Arndt et al., 2001; Rosenblatt et al., 1989). Therefore, Model 1 (Greene et al.'s emotion–reason conflict theory) could not capture this phenomenon.

6.3. Effect of psychopathic personality

Psychopathic individuals exhibit a number of affective deficiencies, including a lack of remorse, guilt, or affection, general callousness, and highly depressed response to stress cues (e.g., Blair, 1995; Herpertz & Sass, 2000). In terms of behavior, psychopathic individuals show abnormally aggressive and violent behavior and deficiencies in impulse control (Hare, 1999).

Yet, despite tendencies for immoral behavior, there is no clear evidence suggesting that such individuals lack normal moral judgment (Bartels & Pizarro, 2011; Blair, 1995; Cima, Tonnaer, & Hauser, 2010; Glenn, Koleva, Iyer, Graham, & Ditto, 2010; Tassy, Deruelle, Mancini, Leistedt, & Wicker, 2013). Generally speaking, studies examining psychopathic individuals' chosen or preferred *actions* tend to find abnormal preference for "utilitarian" actions, whereas those examining their *judgments* reveal no difference from normal subjects. For instance, Tassy et al. (2013) conducted a pair of studies finding precisely this distinction. It appears that "psychopaths know right from wrong but don't care" (Cima et al., 2010). Furthermore, their moral understanding is not tied to emotion: Psychopaths show a capacity for making normal moral distinctions, whereas a large literature shows impairment of their social and moral emotional capacity (Decety & Skelly, 2013), which contradicts Greene et al.'s theory.[16]

Model 2 accounts for the psychopathic effect, on the basis of motivation, in particular the distinction between approach- and avoidance-oriented drives. Psychopathic individuals' increased "utilitarianism" could be explained by abnormally low sensitivity of avoidance-oriented drives (e.g., for risk avoidance) and abnormally high sensitivity of approach-oriented drives. These dispositions also lead to reward- or stimulation-seeking behavior. Evidence for both dispositions exists in the literature on psychopathy (Newman, MacCoon, Vaughn, & Sadeh, 2005; Uzieblo, Verschuere, & Crombez, 2007).

According to the motivationally based theory (Model 2), due to low avoidance sensitivity, an individual tends to make *action* decisions with a higher degree of explicitness (Section 3). Low avoidance sensitivity also decreases the likelihood of selecting avoidance goals. Both tendencies lead to a higher proportion of utilitarian actions. Simulation by Model 2 with this setup (similar to the two previous simulations) was conducted. The simulation results confirmed this effect ($p < .001$).

On the other hand, psychopathic individuals' *judgments* may be made using compensatory strategies: Whereas *action* decisions may be centered on fulfillment of naturally occurring drives or needs (as discussed before), *judgments* may be based more on considerations for social norms (e.g., by employing reasoning within the NACS). This leads to taking a different stance when making *judgments*. Thus, in this case, outcomes are different from action choices. Based on these ideas, Model 2 captures the psychopathic effect.

There is no obvious way for Greene et al.'s theory (Model 1) to capture the psychopathic effect, because emotion is not significantly involved here (Decety & Skelly, 2013). Model 1 failed to simulate the effect.

7. General discussion

7.1. Comparisons

The comparison between the motivationally based theory of moral judgment (from Clarion) and Greene et al.'s (2008) emotion–reason conflict theory has been discussed at length in Sections 4–6, and thus will not be repeated here. However, some comparisons to other theories are provided below.

A theory of moral judgment was provided by Thagard and Finn (2011). In their model EMOCON, emotion is generated both somatically through bodily reactions to stimuli as well as through cognitive appraisal of the effect of stimuli (similar to the Clarion model of emotion; Wilson & Sun, 2014). Somatic bodily reactions may be motivationally based, preparing an individual for taking actions that further goals (as in Clarion). Appraisal is a slower process involving reasoning over how the state of the world affects progression toward goals (as in the Clarion model of emotion; Wilson & Sun, 2014). Thagard and Finn (2011) propose that moral intuitions result from a part of the EMOCON model with a subset of relevant motivations. This view of moral intuition is consistent with the Clarion theory of moral judgment, even though Clarion goes beyond emotion (as argued in detail in Section 4; see also White, 2010). In addition, Clarion provides a more detailed computational model addressing these aspects and beyond, in ways intrinsic to this cognitive architecture.

Another model of moral judgment suggests yet another dual-process distinction—a distinction between model-free and model-based reinforcement learning (Cushman, 2013).

Reinforcement learning gradually approaches an optimal action policy for obtaining the maximum long-term reward. Model-free algorithms make action decisions based only on the current state of the world. Model-based algorithms create an internal model of the world that action selection takes into consideration. Cushman (2013) mapped this distinction to the distinction between action-based and outcome-based moral decision making: Non-utilitarian judgments that focused on the proposed actions per se (such as pushing a man to his death) resulted from model-free processes, while utilitarian judgments that focused on outcomes (saving more people by sacrificing few) resulted from model-based processes.

The difference between the two levels in Clarion includes a distinction between model-free processes (e.g., in the form of neural networks) and model-based processes (e.g., rule-based reasoning, including that concerning future states). Thus, Cushman's model may be viewed as a subset of Clarion. However, while it is true that implicit processes in the ACS of Clarion are more concerned with the morality of actions themselves and explicit processes are more concerned with the best overall outcomes, neither precludes the reverse (see Section 4). For example, if the MS in Clarion sets a goal to avoid harmful actions, explicit processes of the ACS may choose an action in accordance with the goal, rather than based on the best outcome. (An individual might, for instance, rationalize this by emphasizing the importance of maintaining a consistent set of moral rules over maximizing the good in a particular context.) Furthermore, more complex forms of moral reasoning (beyond simple outcome assessment) may also be carried out in Clarion, while Cushman's model cannot capture these more complex forms (Bennis et al., 2010; Sun, 2013).

Finally, other computational approaches to morality have been suggested through the use of neural networks (e.g., Guarini, 2010) or through the use of logic (e.g., Arkin, Ulam, & Wagner, 2012; Bringsjord, Arkoudas, & Bello, 2006). Most of these approaches focus on imbuing robotic systems with ethical decision making rather than strictly modeling human moral judgment.

7.2. Further work

In general, morality is complex, involving many factors, mechanisms, and processes. The Clarion model needs to account for more empirical data. Many additional mechanisms in Clarion (see Sun, 2016) may also need to be involved, for example, when more elaborate reasoning is used (Cushman et al., 2006), or when explanation, justification, or communication is required (Hauser et al., 2007). In particular, attributing intention and agency, which is important to moral judgment, may require more elaborate reasoning. Bennis et al. (2010) also argued for imitation-based, advice-based, coherence-based, case-based, and identity-based processes in moral judgment. As pointed out in Sun (2013), reasoning of various types, social interaction in various contexts, and so on may all need to be taken into consideration in order to fully account for human morality. Many of these mechanisms are available in Clarion, although not used here, because the current models were meant to be as minimal as possible while accounting for a range of data.

In the same vein, the motivational aspect in the simulation was deliberately made as simple as possible. In the future, exactly what drives and goals are relevant in everyday moral contexts should be explored (Thagard & Finn, 2011). Clarion offers motivational mechanisms with the requisite depth to examine these questions (Sun, 2009). Clarion is also capable of large-scale social simulation (Sun, 2006), so it may also address social and cultural influences on morality (Bloom, 2011).

Acknowledgments

This research was supported by the ONR grants N00014-08-1-0068 and N00014-13-1-0342. Thanks to Paul Bello and Jeff White for discussions. Thanks to a number of students who participated in discussion, coding, or simulation. Thanks to Josh Greene for providing the draft versions of his papers and supplemental materials before their publication. Thanks are also due to the editor and the reviewers for their very useful comments.

Notes

1. Greene et al. labeled the two possible actions in the trolley problem as the “utilitarian” and “non-utilitarian” (or “deontological”) actions. We will keep this terminology when discussing the trolley problem in this work. However, do note that these labels may not be accurate descriptors when considering the underlying processes responsible for moral judgment.
2. “Personal” moral dilemmas were described by Greene et al. (2001) and Greene (2008) as those that were more emotionally involved, as established by independent coders. In these dilemmas, generally speaking, harm was done directly and served as a direct means to an end, although the category was not clearly delineated. Relatedly, moral dilemmas can also be classified as high conflict or low conflict, as described earlier.
3. Explicit processes involve explicit (readily accessible) knowledge, while implicit processes involve implicit knowledge. The distinction has been based on voluminous empirical findings in many domains, but it also involves many nuances and some controversies. See Sun (2002, 2016) for details.
4. The notion of drive utilized in Clarion is somewhat similar to, but more general than, those proposed by others (e.g., Hull, 1951). See Sun (2009).
5. Explicit rational decisions may be assimilated into implicit processes over time and thus become implicit, through “implicitation” (e.g., as captured by the “top-down learning” mechanism in Clarion; Sun, 2016). On the other hand, deontological rules may be explicitly followed, rather than resulting exclusively from implicit or emotional processes; such rules may be explicitly established through social processes or through internal “explicitation” (using, e.g., the “bottom-up learning” mechanism in Clarion; Sun et al., 2001; Karmiloff-Smith, 1986).

6. The MS and the MCS were assumed to provide a simple generic goal ("save lives"), constant throughout different dilemmas. Thus, no detail of the MS and the MCS was needed in Model 1.
7. It was assumed that the recognition of these factors was accomplished implicitly in a rapid, reflexive way, analogous to implicit visual pattern recognition. There was usually no elaborate explicit reasoning involved (which, however, might occur on top of implicit recognition).
8. This simplification was sufficient for this simulation. More generally, however, the activation should take into consideration nuances such as degree of deontological violation, which involves assessing, for example, significance of a moral rule, severity of violation of a moral rule, number of violated moral rules, and so on.
9. In a more complete implementation, a curve representing rule activation beginning at zero and asymptotically approaching maximum activation as the proportion of the saved versus the harmed rises may be used.
10. RT = PT + DT + AT, where PT was the perceptual time, AT was the actuation time, and DT was the decision time. PT + AT was set to 5,000 ms. DT varied according to a normal distribution (mean = 1,200 ms, *SD* = 250 ms). See Appendix.
11. Multiple cycles might be needed to reach a moral judgment due to attending to the distractor task. RT differences resulted (in part) from likelihood of attending to the distractor task, which was in turn determined by the activation of distractor task rules. See Appendix for details.
12. Specifically, if the goal "save lives" was set, the neural network output "Yes" (the "utilitarian" choice) at full activation (1.0); otherwise it output "Yes" with activation depending on contextual factors (as described before). If the goal "avoid doing harm" was set, the network output "No" at full activation; otherwise it output "No" with activation depending on contextual factors (as described before).
13. Note that the exact magnitude of each RT was not modeled. The differences between different types of answers within or across control and load conditions were modeled. This was due to the need to focus on main issues and to avoid over-fitting and other complications. There are many factors affecting RTs, not all of which were modeled even in a detailed process model.
14. Of course, similar individual differences existed in dealing with impersonal dilemmas. But, because of more explicit processing overall when dealing with impersonal dilemmas, both utilitarian and non-utilitarian responses were affected by load.
15. Ratings of appropriateness were changed from the original 9-point Likert scale to a 4-point Likert scale to correspond to the scale used in the human data.
16. To say psychopathic individuals possess completely normal moral judgment may be simplistic. For example, they may have greater difficulty making the distinction between conventional and moral violations (Blair, 1995) and view accidental moral violations more permissibly (Young, Koenigs, Kruepke, & Newman, 2012).

References

Adams, H., Wright, L., & Lohr, B. (1996). Is homophobia associated with homosexual arousal? *Journal of Abnormal Psychology*, *105*(3), 440–445.

Allchin, D. (2009). The evolution of morality. *Evolution: Education and Outreach*, *2*(4), 590–601.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, *100*(3), 571–589.

Arndt, J., Allen, J. J., & Greenberg, J. (2001). Traces of terror: Subliminal death primes and facial electromyographic indices of affect. *Motivation and Emotion*, *25*(3), 253–277.

Baron, J., Gürçay, B., Moore, A. B., & Starcke, K. (2012). Use of a Rasch model to predict response times to utilitarian moral dilemmas. *Synthese*, *189*(1), 107–117.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*(1), 154–161.

Beilock, S., & Carr, T. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, *130*, 701–725.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*, 187–202.

Bentham, J. (1781/2005). *Introduction to the principles of morals and legislation*. Available at https://archive.org/details/introductiontoth033476mbp (Original work published 1781). Accessed January 28, 2014.

Blair, R. J. (1995). A cognitive developmental approach to mortality: Investigating the psychopath. *Cognition*, *57*(1), 1–29.

Blanchette, I., & Caparos, S. (2013). When emotions improve reasoning: The possible roles of relevance and utility. *Thinking & Reasoning*, *19*(3–4), 399–413.

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *The Yale Review*, *99*(2), 26–43.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, *21*(4), 38–44.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, *76*(5), 839–855.

Carver, C. S. (2006). Approach, avoidance, and the self-regulation of affect and action. *Motivation and Emotion*, *30*, 105–110.

Cima, M., Tonnaer, F., & Hauser, M. D. (2010). Psychopaths know right from wrong but don’t care. *Social Cognitive and Affective Neuroscience*, *5*(1), 59–67.

Clark, L. A., & Watson, D. (1999). Temperament: A new paradigm for trait psychology. *Handbook of Personality: Theory and Research*, *2*, 399–423.

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2–7.

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.

Decety, J., & Skelly, L. (2013). The neural underpinnings of the experience of empathy: Lessons for psychopathy. In K. N. Ochsner, & S. M. Kosslyn (Eds.), *The Oxford handbook of cognitive neuroscience*, Vol. 2 (pp. 228–243). New York: Oxford University Press.

Evans, J., & Frankish, K. (Eds.) (2009). *In two minds: Dual processes and beyond*. Oxford, UK: Oxford University Press.

Foot, P. (1978). *The problem of abortion and the doctrine of the double effect in virtues and vices. Oxford review*. Oxford, UK: Basil Blackwell.

Frijda, N. (1986). *The emotion*. Cambridge UK: Cambridge University Press.

Glenn, A. L., Koleva, S., Iyer, R., Graham, J., & Ditto, P. H. (2010). Moral identity in psychopathy. *Judgment and Decision Making*, *5*(7), 497–505.

Gray, J. A. (1987). Perspectives on anxiety and impulsivity: A commentary. *Journal of Research in Personality*, *21*(4), 493–509.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323; author reply 323–324.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Volume 3* (pp. 35–80). Cambridge, MA: MIT University Press.

Greene, J. D., Cushman, F., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

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.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Guarini, M. (2010). Particularism, analogy, and moral cognition. *Minds and Machines*, *20*(3), 385–422.

Gubbins, E., & Byrne, R. M. J. (2014). Dual processes of emotion and reason in judgments about moral dilemmas. *Thinking & Reasoning*, *20*(2), 245–268.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Hare, R. D. (1999). Psychopathy as a risk factor for violence. *Psychiatric Quarterly*, *70*(3), 181–197.

Hauser, M., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language*, *22*(1), 1–21.

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: A unified theory and a connectionist model. *Psychological Review*, *117*(3), 994–1024.

Herpertz, S. C., & Sass, H. (2000). Emotional deficiency and psychopathy. *Behavioral Sciences & the law*, *18*(5), 567–580.

Hoffman, M. L. (1979). Development of moral thought, feeling, and behavior. *American Psychologist*, *34* (10), 958–966.

Hull, C. L. (1951). *Essentials of behavior*. New Haven, CT: Yale University Press.

Hume, D. (1738/2003). *A treatise of human nature*. Available at: Project Gutenburg Web site: http://www.gutenberg.org/ebooks/4705 (Original work published 1738). Accessed January 28, 2014.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review*, *91*(2), 153–184.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521–537.

Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind & Language*, *27*(5), 519–545.

Kant, I. (1780/2004). *The metaphysical elements of ethics* (Abbot, T., trans.). Available at Project Gutenberg Web site: http://www.gutenberg.org/cache/epub/5684/pg5684.html (Original work published 1780). Accessed January 30, 2014.

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. *Cognition*, *23*, 95–147.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*(7138), 908–911.

Kohlberg, L. (1969). Stage and Sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). New York: Rand McNally.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Lambert, A., Payne, B., Jacoby, L., Shaffer, L., Chasteen, A., & Khan, S. (2003). Stereotypes as dominant responses: On the social facilitation of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology*, *84*, 277–295.

Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, *46*(4), 352–367.

Lodge, M., & Taber, C. (2013). *The rationalizing voter*. New York: Cambridge University Press.

Manfrinati, A., Lotto, L., Sarlo, M., Palomba, D., & Rumiati, R. (2013). Moral dilemmas and moral principles: When emotion and cognition unite. *Cognition & Emotion*, *27*(7), 1276–1291.

Markman, A. B., & Maddox, W. T. (2005). The implications of advances in research on motivation for cognitive models. *Journal of Experimental and Theoretical Artificial Intelligence*, *17*, 371–384.

Maslow, A. (1943). A theory of human motivation. *Psychological Review*, *50*, 370–396.

Millar, J. C., Turri, J., & Friedman, O. (2014). For the greater goods? Ownership rights and utilitarian moral judgment. *Cognition*, *133*(1), 79–84.

Moll, J., de Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition: Sentiments, concepts, and values. *Annals of the New York Academy of Sciences*, *1124*, 161–180.

Monroe, K. R. (2012). Cognition and moral choice. In R. Sun (Ed.), *Grounding social sciences in cognitive sciences* (pp. 183–205). Cambridge, MA: MIT Press.

Murray, H. (1938). *Explorations in personality*. New York: Oxford University Press.

Newman, J. P., MacCoon, D. G., Vaughn, L. J., & Sadeh, N. (2005). Validating a distinction between primary and secondary psychopathy with measures of Gray's BIS and BAS constructs. *Journal of Abnormal Psychology*, *114*(2), 319–323.

Piaget, J. (1965). *The moral judgment of the child*. New York: Free Pass.

Quirin, M., Loktyushin, A., Arndt, J., Küstermann, E., Lo, Y. Y., Kuhl, J., & Eggert, L. (2011). Existential neuroscience: A functional magnetic resonance imaging investigation of neural responses to reminders of one's mortality. *Social Cognitive and Affective Neuroscience*, *7*(2), 193–8.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science*, *34*(2), 311–321.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*, 219–235.

Reiss, S. (2010). Skinny on Maslow's hierarchy: Is Maslow's hierarchy valid? *Psychology Today*, July 24, 2010.

Rosenblatt, A., Greenberg, J., Solomon, S., Pyszczynski, T., & Lyon, D. (1989). Evidence for terror management theory: I. The effects of mortality salience on reactions to those who violate or uphold cultural values. *Journal of Personality and Social Psychology*, *57*(4), 681–690.

Royzman, E. B., Landy, J. F., & Leeman, R. F. (2015). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cognitive Science*, *39*, 325–352.

Rumelhart, D., & McClelland, J., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructures of cognition*. Cambridge, MA: MIT Press.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review*, *74*(1), 29–39.

Son Hing, L. S., Chung-Yan, G. A., Hamilton, L. K., & Zanna, M. P. (2008). A two-dimensional model that employs explicit and implicit attitudes to characterize prejudice. *Journal of Personality and Social Psychology*, *94*(6), 971–987.

Squire, L. R. (1987). *Memory and brain*. New York: Oxford University Press.

Strachan, E., Schimel, J., Arndt, J., Williams, T., Solomon, S., Pyszczynski, T., & Greenberg, J. (2007). Terror mismanagement: Evidence that mortality salience exacerbates phobic and compulsive behaviors. *Personality and Social Psychology Bulletin*, *33*, 1137–1151.

Sun, R. (1994). *Integrating rules and connectionism for robust commonsense reasoning*. New York: John Wiley and Sons.

Sun, R. (2002). *Duality of the mind: A bottom-up approach toward cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sun, R. (2003). A detailed specification of CLARION 5.0. Technical report. RPI, Troy, NY.

Sun, R. (Ed.) (2006). *Cognition and Multi-Agent Interaction*. New York: Cambridge University Press.

Sun, R. (2009). Motivational representations within a computational cognitive architecture. *Cognitive Computation*, *1*(1), 91–103.

Sun, R. (2012). Memory systems within a cognitive architecture. *New Ideas in Psychology*, *30*, 227–240.

Sun, R. (2013). Moral judgment, human motivation, and neural networks. *Cognitive Computation*, *5*(4), 566–579.

Sun, R. (2016). *Anatomy of the mind*. New York: Oxford University Press.

Sun, R., & Mathews, R. C. (2012). Implicit cognition, emotion, and meta-cognitive control. *Mind & Society*, *11*(1), 107–119.

Sun, R., Merrill, E., & Petersen, T. (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science*, *25*, 203–244.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, *112*(1), 159–192.

Sun, R., & Wilson, N. (2014). A model of personality should be a cognitive architecture itself. *Cognitive Systems Research*, *29–30*, 1–30.

Tassy, S., Deruelle, C., Mancini, J., Leistedt, S., & Wicker, B. (2013). High levels of psychopathic traits alters moral choice but not moral judgment. *Frontiers in Human Neuroscience*, *7*, 229.

Thagard, P., & Finn, T. (2011). Conscience: What is moral intuition. In C. Bagnoli (Ed.), *Morality and the Emotions* (pp. 150–169). Oxford, UK: Oxford University Press.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Toates, F. (1986). *Motivational systems*. Cambridge, UK: Cambridge University Press.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, *64*, 231–255.

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, *40*, 923–930.

Trémolière, B., De Neys, W., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, *124*(3), 379–384.

Tritt, S. M., Inzlicht, M., & Harmon-Jones, E. (2012). Toward a biological understanding of mortality salience (and other threat compensation processes). *Social Cognition*, *30*(6), 715–733.

Tyrell, T. (1993). Computational mechanisms for action selection. Ph.D Thesis, University of Edinburgh, Edinburgh, UK.

Uzieblo, K., Verschuere, B., & Crombez, G. (2007). The Psychopathic Personality Inventory: Construct validity of the two-factor structure. *Personality and Individual Differences*, *43*(4), 657–667.

Watkins, C. (1989). Learning with delayed rewards. Ph.D. Thesis, Cambridge University, Cambridge, UK.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780–784.

White, J. (2010). Understanding and augmenting human morality: An introduction to the ACTWith model of conscience. In L. Magnani, W. Carnielli, & C. Pizzi (Eds.), *Model-based reasoning in science & technology* (pp. 607–621). Berlin: Springer.

Wilson, N. R. (2012). Towards a psychologically plausible comprehensive computational theory of emotion. Ph.D Thesis, Rensselaer Polytechnic Institute, Troy, NY.

Wilson, N., & Sun, R. (2014). Coping with bullying: A computational emotion-theoretic account. In P. Bello, et al. (Eds.), *Proceedings of the annual conference of cognitive science society*, Quebec City, Quebec, Canada (pp. 3119–3124). Austin, TX: Cognitive Science Society.

Wilson, N. R., Sun, R., & Mathews, R. C. (2009). A motivationally based simulation of performance degradation under pressure. *Neural Networks*, *22*(5), 502–508.

Yerkes, R., & Dodson, J. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482.

Young, L., Koenigs, M., Kruepke, M., & Newman, J. P. (2012). Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology*, *121*(3), 659–667.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inference. *American Psychologist*, *35*(2), 151–175.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix.