Judgment and Decision Making (2026), **vol. 21**:e7 1–61
doi:10.1017/jdm.2026.10030

EMPIRICAL ARTICLE

Norm conflicts and morality: The CNIS Conflict model of moral decision-making

Niels Skovgaard-Olsen and Karl Christoph Klauer

Psychology, University of Freiburg, Germany

Corresponding author: Niels Skovgaard-Olsen; Email: niels.skovgaard.olsen@psychologie.uni-freiburg.de

Received: 30 September 2025; **Revised:** 30 January 2026; **Accepted:** 31 January 2026

Keywords: moral judgment; norm conflict; MPT modeling; Deontology; Utilitarianism; individual variation

Abstract

The goal of this paper is to study individual variation in participants' adherence to conflicting moral views. To do this, we elicit participants' reflective attitudes in an argumentative task and introduce a new Conflict model of moral decision-making. This Conflict model builds on the widely used CNI model of moral judgments (Gawronski et al. [2017, *Journal of Personality and Social Psychology*, 113, 343–376]) but improves it in several respects. First, we follow Skovgaard-Olsen and Klauer (2024, *Personality and Social Psychology Bulletin*, 50(9), 1348–1367) in extending the model to investigate invariance violations of the models' parameters. Second, we model cases in which participants are conflicted between utilitarian and deontological response tendencies. In Experiment 1, we employ an argumentative paradigm to elicit commitments for moral views from participants to estimate latent classes in participants' moral views. We then measure a range of egoistic and altruistic covariates used in Kahane et al. (2015, *Cognition*, 134, 193–209) and Conway et al. (2018, *Cognition*, 179, 241–265) to investigate whether participants' acceptance of instrumental harm is associated with a genuine concern for the greater good or whether it is rather driven by antisocial character traits (Bartels and Pizarro [2011, *Cognition*, 121, 154–161]). Next, we report two validation studies of our new Conflict model. In a preregistered experiment, the discriminant validity of the conflict detection/resolution path of the Conflict model and the construct validity of its conflict parameter are tested. Finally, in a second validation study, we contrast response formats of dilemma judgments and find evidence in favor of using a format in which participants can opt out of difficult moral dilemmas when they feel conflicted, over the traditional format in moral psychology that lacks this possibility. We show that the CNI model is challenged by the finding of asymmetries in experienced conflict across conditions.

1. Introduction

To end the Second World War, President Truman faced a choice between at least two costly options: launching a conventional invasion of Japan with many civilian casualties as side-effects or using atomic bombs directly on the civilian population (Hiroshima and Nagasaki) with fewer total casualties but directly targeting civilians. Truman chose the latter on August 6 and 9, 1945. From a utilitarian perspective focused on aggregate outcomes, targeting civilians might be justified if it minimizes total deaths. From a deontological perspective emphasizing moral constraints on the means we use, directly killing civilians violates core moral prohibitions to kill innocent people. This dilemma illustrates an extreme version of the moral conflicts commonly used to contrast Utilitarianism and Deontology (Parfit, 2017, Ch. 56). When Hamas attacked Israel on October 7, 2023 and hid in a densely populated area

with over 2 million civilians in the Gaza strip with tunnels under mosques and hospitals, another such difficult dilemma of weighing civilian casualties against measures to prevent further terrorist attacks occurred.

The goal of this paper is to introduce a new Conflict model of moral decision-making, which builds on the CNI model of moral judgments (Gawronski et al., 2017) but extends it to model cases in which participants are conflicted by moral dilemmas. Based on this model and an experimental design developed to study norm conflicts (Skovgaard-Olsen et al., 2019), the present paper investigates individual differences in participants' moral views from the first-person perspective of a deliberating agent deciding what to do. In moral psychology, there has been much interest in investigating whether participants adopt utilitarian or deontological views in moral decision-making (Greene, 2013; Lombrozo, 2009; Waldmann et al., 2012; Holyoak & Powell, 2016), which contrast participants' tendency to maximize the well-being of all people involved with moral principles based on general considerations about the dignity and autonomy of human beings. Much of this work has been centered on Trolley dilemmas, which were originally used in philosophy to pinpoint a tension between these two ethical views. On the one hand, Utilitarianism permits instrumental harm, for example, in sacrificing 1 innocent person to save the lives of 5 innocent people for the sake of promoting the greater good. Kantian deontological principles, on the other hand, restrict our actions to avoid the use of coercion, deception, or direct harm to achieve our aims (Foot, 1967; Kamm, 1989; Kamm and Rakowski, 2019; Quinn, 1989; Thomson, 1976, 1985).

Yet, it is still controversial how best to measure participants' adherence to these contrasting moral views (Conway et al., 2018a; Kahane et al., 2015, 2018). On a narrow operationalization (found, e.g., in Greene, 2008; Greene et al., 2004), a utilitarian response is *defined* as selecting action in sacrificial dilemma like the Trolley, in which a run-away trolley is on a path to run over 5 people and the action is either to divert it to an alternative track with 1 person (*Switch*) or push a large person in front to activate its automatic brakes (*Footbridge*). A deontological response is *defined* by selecting inaction. Using similar dilemmas, Greene et al. (2001, 2004) and Greene (2008, 2013) report neuroscientific evidence of increased activity in areas relating to cognitive control (e.g., the dorsolateral prefrontal cortex) to argue that the utilitarian option is chosen based on a deliberative, rational process. Similarly, evidence of increased activity in emotional areas (e.g., the amygdala and the ventromedial prefrontal cortex) is used to make a case that Deontology is a *post hoc* rationalization of an evolutionary ancient emotional, alarm-like response to direct personal harm.

However, as Kahane et al. (2015, 2018) argue, there is a positive core to Utilitarianism that goes beyond the narrow operationalization, which both consists in an impartial concern with the greater good and motivates contemporary utilitarians' involvement in programs of effective altruism and animal rights movements (Singer, 2015). For utilitarians, like Jeremy Bentham, John Stuart Mill, Henry Sidgwick, and Peter Singer, the pleasure and pain of sentient beings are intrinsically good and bad states of the world, which contribute to the net sum of happiness which should be maximized across individuals (Darwall, 2003a, 2003b). In contemporary versions, less hedonistic components are considered. Moreover, the agent need not calculate the total utility of each individual act but should follow general rules that impartially maximize well-being (Brandt, 1965; Parfit, 2011).

In contrast, deontological views place a respect for agent autonomy at the center of their ethics. For example, by requiring that we never treat a person as a means only and by refraining from imposing burdens on others, which they would not themselves give their consent to. Historically, this notion contributed to the introduction of inalienable human rights (Bayefsky, 2013; Demenchonok, 2009; Herman, 2011; Rauber, 2009; Scanlon, 2011; Wolf, 2011; Wood, 2011), as well as to the notion of a sacrosanct human dignity, which was written into the German constitution after WW2. Deontologists have traditionally evaluated our actions by their intentions/action plans, instead of striving to produce the best outcomes. Whereas utilitarians see harm inflicted upon sentient beings as an *outcome* to be minimized, deontologists focus on the responsibility of agents and view harm as a morally evaluable *action* that is done to some being (Darwall, 2003b, p. 4). Viewed from this agent-perspective:

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

'Doing harm is worse (. . .) than failing to prevent it' (ibid.)

[*The Doctrine of Doing and Allowing*].

'directly intending harm is worse than causing it as an unintended side effect' (ibid.)

[*the Doctrine of Double Effect*, DDE].

DDE will figure centrally in our experiments below.[1] DDE is often applied to cases of intended harm on civilians versus harm to civilians as a foreseeable side-effect (e.g., due to military campaigns), as in our opening examples. These diverging categorizations of harm (as outcome vs. as action) lay the basis for the different attitudes toward sacrificial dilemma. For utilitarians, the total amount of suffering should be minimized. For deontologists, the actions taken to achieve this aim are decisive. By intending to kill 1 person as a necessary means to save 5 people without this person's consent, the one person is treated as a means only in a way that would be prohibited by a respect for this person's autonomy and individual rights (Thomson, 1985). By banning this, DDE 'gives each person some veto power over a certain kind of attempt to make the world a better place at his expense' (Quinn, 1989, 207).

Aside from difficulties with capturing the positive core of Deontology and Utilitarianism, their narrow operationalization in traditional sacrificial dilemmas has also been criticized on methodological grounds. In reply, process-dissociation (PD) approaches have been developed to disentangle multiple psychological processes that can give rise to the same observed response (Conway and Gawronski, 2013; Gawronski et al., 2017). For instance, the CNI model (Gawronski et al., 2017) posits three latent processes underlying moral judgments: a consequence-based response tendency (**C**), a norm-based response tendency (**N**), and a general bias toward inaction (**I**). These approaches are implemented using multinomial processing tree (MPT) models and applied to real-world moral dilemmas designed to vary consequences and norms independently.

1.1. The CNI model

Process-dissociation models do not assume that experimental tasks are process-pure. Instead, they model how different latent processes jointly determine categorical responses (Hütter and Klauer, 2016). In moral judgment tasks, the same action or inaction response may reflect sensitivity to outcomes, sensitivity to moral norms, or a general response tendency. Traditional sacrificial dilemmas confound these factors. Since a deontological response always requires inaction in the standard sacrificial dilemmas (e.g., where victims are fixed to the tracks of a run-away trolley), it cannot be separated from a *general inaction bias*. Moreover, since the utilitarian response always requires action in the standard dilemma (even when it goes against moral considerations), it cannot be separated from *antisocial tendencies* toward sacrifice.

The PD model of Conway and Gawronski (2013) was a first development in this direction with its separation of cases in which the benefits of action can be either smaller or greater than the cost of the outcome. What is distinctive about its further development into the CNI model (Gawronski et al., 2017) is the separation of a reluctance from causing harm through action (which may be based on deontological principles) from a simple response bias to select inaction in general (independent of the moral content). Figure 1 illustrates the processing structure assumed by the CNI model and how different response patterns arise from these three latent processes. To estimate these parameters, the model relies on a factorial design that crosses (a) whether consequences favor action or inaction (benefits greater vs. smaller than costs) with (b) whether moral norms prohibit an action or prescribe an intervention to prevent harm (proscriptive vs. prescriptive norms). This yields conditions in which Deontology and Utilitarianism either make the same predictions (*congruent* trials) or different predictions (*incongruent* trials), and in which both moral views sometimes predict action and sometimes predict inaction (see Figure 1).

[1] Sometimes DDE includes further proportionality constraints (McIntyre, 2019).

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

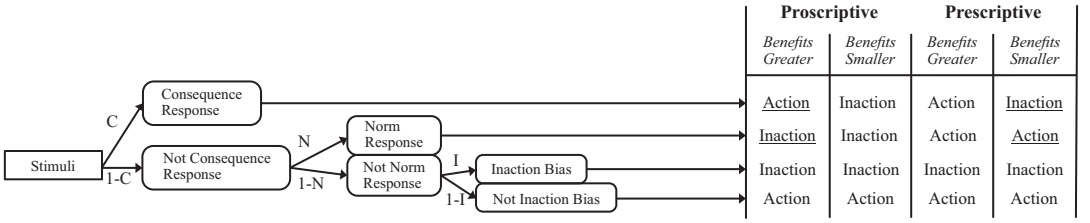


Figure 1. *The CNI model.*

Note: Gawronski et al. (2017). The tree diagram shows how three latent processes (C, N, I) generate observed responses. From left to right, each branch represents a decision path: stimuli first trigger either a consequence-based response (probability C) or not (1−C), then either a norm-based response (N) or not (1−N), and finally either an inaction bias (I) or not (1−I). The right panel shows which response results for 4 trial types, crossing 2 factors: proscriptive vs. prescriptive norms, and whether benefits are greater vs. smaller than the costs. The first two rows distinguish incongruent trials (where C and N favor different responses; underlined text) from congruent trials (where C and N favor the same response; plain text).

In this framework, a norm-based response is expressed when participants refrain from prohibited actions or intervene to prevent harm, depending on the action context. A consequence-based response is expressed when participants choose the option with better outcomes overall, regardless of action or inaction. The inaction-bias parameter captures a general tendency to favor inaction across conditions, independent of both norms and consequences. Participants' responses are analyzed using an MPT model to estimate the probability that each latent process contributed to the observed action or inaction decision (Batchelder and Riefer, 1999; Erdfelder et al., 2009; Hütter and Klauer, 2016; Schmidt et al., 2025).

While previous studies with traditional sacrificial dilemmas report a positive correlation between psychopathic traits and utilitarian choices (Marshall et al., 2018), it has been found using the CNI model that its parameters tend to be negatively correlated with psychopathic traits (Gawronski et al., 2017; Körner et al., 2020; Luke and Gawronski, 2021; Luke et al., 2021; Paruzel-Czachura et al., 2024). These findings suggest that individuals high in subclinical psychopathy display less restraint in sacrificing human life and also tend to be less influenced by the difference between whether the sacrifice occurs when the benefits for the greater good are larger or smaller than the costs of the outcome. They also illustrate that the CNI model has been extended to permit the study of individual differences by assessing its parameters at the individual level (e.g., Körner et al., 2020; Kroneisen and Heck, 2020; Skovgaard-Olsen and Klauer, 2024). In these and many other recent papers, the CNI model has thus fruitfully been employed to study moral judgments. Nevertheless, the goal of the present paper is to improve upon this model in two crucial respects.

1.2. The invariance assumption

Figure 2 presents the CNIS model, which is an extension of the CNI model that adds a skip option and permits the estimation of separate C and N parameters across conditions. The skip option allows participants to skip a dilemma when uncertain. This model extension was motivated by the finding in Klauer et al. (2015) that influential versions of process-dissociation models from cognitive psychology (Stroop task, cued recall) and social psychology (racial bias in the weapon task) violated the assumption that their parameters were invariant across congruent and incongruent trials.

As the comparison between Figures 1 and 2 shows, 2 assumptions of the CNI model of this kind stand out: (1) that the N parameter stays *invariant* (i.e., remains the same) across conditions with proscriptive and prescriptive norms and (2) that the C parameter stays invariant across all 4 CNI conditions (see Table 1).

As a result, the CNI model assumes that the utilitarian contrasts between action and inaction remain constant across conditions. Yet, by design, congruent and incongruent dilemmas differ in the *strength of the benefit to others* against which the cost of sacrificing one individual is weighed. In congruent

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

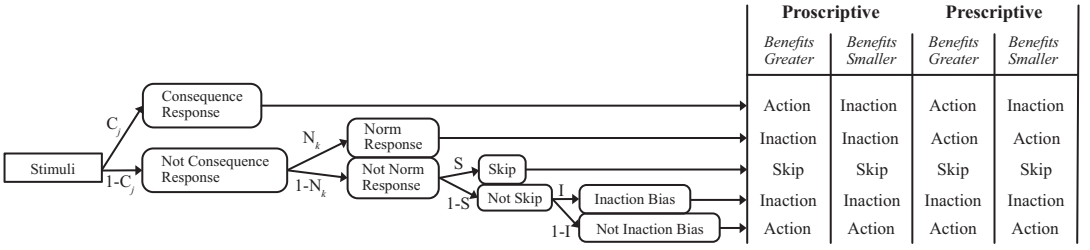


Figure 2. *The CNIS model.*

Note: Skovgaard-Olsen and Klauer (2024). The CNIS model extends the CNI model by adding a skip option (S) for participants who cannot decide. Unlike the original CNI model, different C parameters are estimated for each of 4 scenario conditions (C_j, j = 1, . . . ,4) and different N parameters for 2 norm types (N_k, k = 1, 2).

Table 1. *Abbreviations and conditions.*

Conditions	Abbreviations		Parameters	Invariance
Proscriptive, benefits > costs	ProGreater,	Pro>	$C_{Pro>}$, N_{pro}	C, N
Proscriptive, benefits < costs	ProSmaller,	Pro<	$C_{Pro<}$, N_{pro}	
Prescriptive, benefits > costs	PreGreater,	Pre>	$C_{Pre>}$, N_{pre}	
Prescriptive, benefits < costs	PreSmaller	Pre<	$C_{Pre<}$, N_{pre}	

Note: 'benefits > costs' means that the benefit of the sacrifice is greater than the costs of the sacrifice. In the last column, the invariance assumption is illustrated as the modeling of 4 conditions via just 2 parameters instead of 6.

dilemmas, the benefit to others is intentionally limited (e.g., avoiding mild discomfort), such that it is normatively insufficient to justify sacrifice, whereas in incongruent dilemmas the benefit is substantial (e.g., saving 5 lives). This asymmetry is used by the CNI model to construct conditions in which Deontology and Utilitarianism converge in their predictions, but it contradicts the assumption that a single C parameter can capture consequence-driven response tendencies across item types. Treating the C parameter as invariant within the CNI model would imply that limited and substantial utilitarian contrasts activate the consequence-based response with equal probability.

A further complication arises for the CNI model's assumption that the deontological norm forbidding a questionable action (e.g., killing someone) is as strong as the norm prescribing to interfere with someone else's actions to prevent that same action from occurring (e.g., preventing someone else in killing someone). In the deontological literature, proscriptive norms (e.g., prohibitions on causing harm) are often contrasted with duties to aid; with the former negative duties being treated as stricter than the latter positive duties (Kamm, 2007). However, in the CNI-type moral dilemmas introduced below, prescriptive trials do not introduce a new positive moral duty. Instead, they present the same deontological prohibition against sacrificial harm but embed it within a different action context (i.e., a situation in which the harmful action has been pre-selected and the participant must decide whether to overturn it). This manipulation thus changes the choice architecture rather than the underlying norm. But it is a questionable assumption that the norm should be activated with the same probability when embedded in different choice architectures, and indeed empirical work has shown that N parameters differ across conditions (Skovgaard-Olsen and Klauer, 2024).[2]

[2]A further point concerns the direction of this difference. If interpreted in light of the distinction between negative and positive duties, then one might expect $N_{pro} > N_{pre}$. But this reasoning does not apply to CNI-type dilemma under the current implementation. As we show below, our prescriptive items do not involve a positive duty to aid but rather the same negative duty not to harm but in an action context in which an assistant has pre-selected instrumental harm for the participant. Under such conditions, participants may experience reactance toward having a harmful action planned in their name and therefore show a stronger activation of the deontological norm in the prescriptive case, which corresponds to the pattern found in earlier work (Skovgaard-Olsen and Klauer, 2024) and replicated here.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

The addition of the extra skip response enriches the data structure so as to permit the estimation of separate C and N parameters (see Table 1). This then allows one to test empirically whether the C and N invariance assumptions are justified. Modeling the skip option is straightforward in the logic of the CNI model. According to this model, participants can either enter the latent state of judging scenarios based on consequences (with probability C) or enter a latent state of judging scenarios based on norms, when they do not enter the first state (with probability $(1-C) \times N$). If neither state is entered, participants enter a state of uncertainty (with probability $(1-C) \times (1-N)$) in which responses are no longer deterministic. In this state of uncertainty, participants, metaphorically speaking, throw a loaded dice with the faces 'inaction' and 'action', with probability (I) and (1−I), respectively. Since participants are instructed to use the skip option in the case of uncertainty, this state of uncertainty, reached with probability $(1-C) \times (1-N)$, is the only place in the model in which skipping can enter (Skovgaard-Olsen and Klauer, 2024). The extension of the CNI model in Figure 2 provides participants with a third face on their loaded dice, which now shows the faces 'action', 'inaction', and 'skip'. Because in the state of uncertainty, neither latent states based on norms nor consequences are activated, it also makes sense that the I parameter and the S parameter in the CNIS model do not depend upon type of dilemma. The reason is that the 4 CNI conditions are distinguished solely in terms of differences w.r.t. norms and consequences. In Skovgaard-Olsen and Klauer (2024), it was found that models of the type displayed in Figure 2 provide a superior balance of fit to the data and parsimony if separate C and N parameters are estimated. Instead of having 4 parameters (C, N, I, S), the best fitting model thus had 8 parameters (I, S, N_{pro}, N_{pre}, $C_{Pro>}$, $C_{Pro<}$, $C_{Pre>}$, $C_{Pre<}$), with parameter labels such as $C_{Pro>}$ explained in Table 1. For these reasons, the models considered here do not impose the problematic and empirically falsified invariance assumption (see Appendix B).

1.3. Response conflicts

The second change to the CNI model in the Conflict model introduced below concerns what happens if participants have not yet reached the state of uncertainty. According to the CNI model, when participants enter the latent state of judging based on consequences (with probability C), the response is determined by consequences with probability one, whether or not they may also be aware of the norms manipulated and whether or not the moral item is congruent or incongruent.

Suppose that participants do not enter the latent state of judging the scenario based on consequences (with probability 1−C) but they enter the latent state of judging the scenario based on norms (with probability N). In this case, the response is determined by norms with probability one—whether or not the moral item is congruent (with no conflict between deontological norms and utilitarian cost-benefit analysis) or incongruent (where deontological norms and utilitarian values conflict). And thus, when consequences or norms are activated, the response is deterministically captured by all-or-none processes in the CNI model with consequences dominating norms. As this shows, the CNI model does not have a mechanism by which participants can be sensitive to the difference between congruent and incongruent items. Conflicts between norms and consequences are simply pre-empted by the dominance of consequences over norms built into the model: Whenever norms and consequences are activated, consequences compel the response. In particular, the probability of reaching the state of uncertainty always equals $(1-C) \times (1-N)$, whether or not the item is congruent or incongruent. Against this, it could, however, be argued that a state of uncertainty is reached more often for incongruent than congruent moral items given that incongruent dilemma give rise to the conflict of whether to base one's judgment on the manipulated consequences or the manipulated norms. In this case, participants' skip responses might be asymmetrically distributed such that participants more often choose to skip a scenario in the incongruent conditions than in congruent conditions.

Similarly, it has been argued in Cohen and Ahn (2016) that the process-dissociation model of Conway and Gawronski (2013), on which the CNI model builds, makes a *deterministic response processes* assumption, which they find problematic. Once activated, a process is predicted to produce a given response without error. Yet, if participants feel the pull of both the deontological and utilitarian

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 2. *Sequence of process dissociation models.*

PD model	CNI model	CNIS model	Conflict model
			
Consequence manipulation Conway and Gawronski (2013)	Norm manipulation, inaction bias estimation Gawronski et al. (2017)	No C/N invariance assumption Skovgaard-Olsen and Klauer (2024)	Conflicting processes, latent classes

Note: A fifth model is provided by Liu and Liao's (2021) DNA model. To narrow our focus, we here concentrate on these four models, however. Models to the right incorporate the advantages of models to the left while incorporating new features.

response tendencies in the incongruent scenarios, there may be a stochastic process of how to resolve the conflict which needs to be modeled.

More generally, within dual-process theories there is a discussion about the nature of conflict detection and resolution (see, e.g., Bago and De Neys, 2018; Baron and Gürçay, 2017; Białek and De Neys, 2017; De Neys, 2023; Gürçay and Baron, 2016). But this issue has not yet been addressed within the framework of the PD and CNI models, despite their widespread applications to moral dilemma, in which participants often report feeling torn between response options (Conway et al., 2018b). The absence of conflict in the CNI model is especially problematic in the usual setting in which the model is applied: in which 4 variations of several scenarios are presented within-participant.[3] The reason is that participants may have applied utilitarian principles to maximize the greater good in a previous scenario—or in a previous condition of the same scenario. Yet, when presented with a further item, they may now feel the intuitive pull of, for example, responding according to Deontology and thus find themselves conflicted over the opposing forces applied to their responses.

This then motivates the CNIS Conflict model of moral judgments presented in Figure 3.[4] This Conflict model retains the model structure of Figure 2 for congruent items. For incongruent items, however, it permits that participants can be in state of conflict between Utilitarianism and Deontology. In this case, they can either resolve the detected conflict based on the relative strength of the latent states $\left(C_{res} = \frac{C}{(C+N)}\right)$, where consequences or norms determine the answer, or fail to resolve it, which then produces a skip response. The model moreover follows the CNIS model (Figure 2) in its solution to the first desideratum by using a skip option to introduce further degrees of freedom into the data to accommodate different C and N parameters in the different conditions. Model comparison will then decide empirically when and whether these specific C and N parameters (N_{pro}, N_{pre}, $C_{Pro>}$, $C_{Pro<}$, $C_{Pre>}$, $C_{Pre<}$) can be set equal across conditions, instead of constraining them *a priori* to take the same value. Further details on the statistical implementation of both the CNIS model (Figure 2) and the Conflict model (Figure 3) can be found in Appendix A.

Below we will keep referring to these four process-dissociation models. To provide an overview, Table 2 illustrates the relationship between these 4 types of models. All models in Table 2 apply process-dissociation methods to improve upon the narrow operationalization of Deontology as inaction and Utilitarianism as action in incongruent sacrificial dilemma.

The 4 models build sequentially on each other such that models to the right incorporate the advantages of models to the left but implement new features. In what follows, we employ the new Conflict model to establish moral profiles of participants (Experiment 1) and investigate the intended

[3] In our studies, participants saw the same scenarios in the intended means and side-effect conditions – but separated. They never saw the four CNI conditions in the same scenario within one session, as in the other CNI studies.

[4] As a short form, we will refer to this model as 'the Conflict model', although this label has also been used for a Rasch model before (Gürçay and Baron, 2016; Baron and Gürçay, 2017).

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

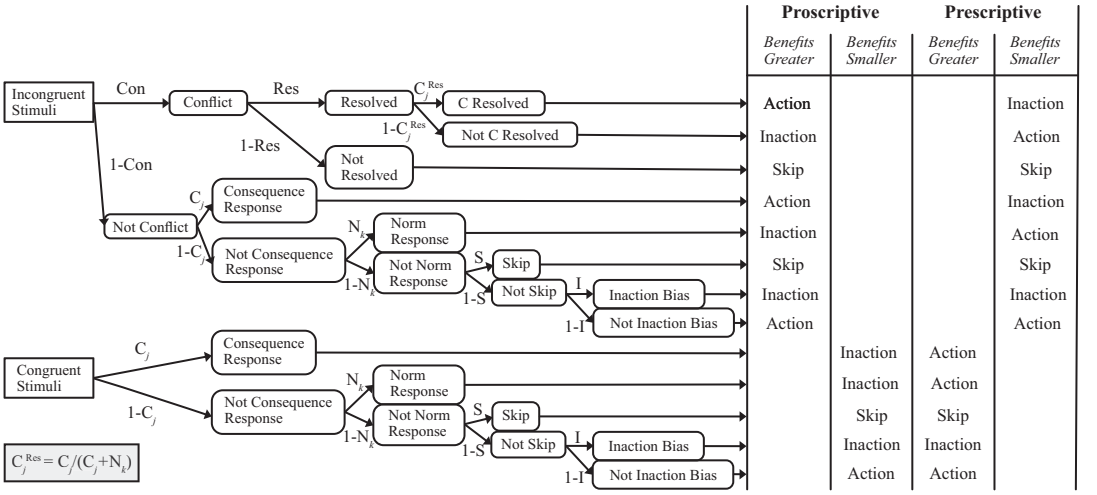


Figure 3. *The Conflict model.*

Note: The Conflict model distinguishes 2 processing routes for incongruent trials (top) versus congruent trials (bottom). On incongruent trials, the model hypothesizes that participants may first detect conflict (Con) between consequence- and norm-based considerations. If conflict is detected, they may resolve it (Res) in favour of consequences ($C_{Res} = C_j/(C_j+N^k)$) or norms ($1-C_{Res}$), or leave it unresolved and skip. If conflict is not detected ($1-$Con), processing proceeds as in the CNIS model (Figure 2). No conflict detection occurs on congruent trials and processing follows the standard CNIS paths. Different C and N parameters are estimated across the 4 CNI conditions (C_j, $j = 1,\ldots,4$ conditions; N_k, $k = 1, 2$ norms), while conflict detection (Con), resolution (Res), skip tendency (S), and inaction bias (I) parameters remain invariant across scenario types. The model is fitted to both scenarios where harm is a foreseeable side-effect versus an intended means with separate C and N parameters in each.

psychological interpretation of the model's parameters through 2 validation studies (Experiments 2 and 3). Readers with a stronger interest in moral cognition and individual variation can focus on Experiments 1 and 3. Experiment 2 is of more technical interest in validating the Conflict model.

2. Experiment 1: Establishing moral profiles via the Conflict model

Experiment 1 served multiple purposes which we briefly list here while deferring the full explanation to below:

(1) Testing the new Conflict model (Figure 3) through a model comparison with CNI and CNIS type models (Figures 1 and 2).

(2) Extending the CNI scenarios with a within-participant manipulation of intended means vs. foreseeable side-effects.

(3) Using a scorekeeping task to establish latent classes in participants' reflective attitudes (Skovgaard-Olsen and Cantwell, 2023; Skovgaard-Olsen et al., 2019). This analysis served the purpose of analyzing participants' moral judgments as coming from a mixture distribution of, e.g., Utilitarianism and Deontology allowing us to classify participants accordingly.

(4) Measuring a range of altruistic and egoistic covariates to further investigate the latent classes and assess 3 motivational hypotheses introduced below.

The introduction has already explained and motivated the first and second purpose of Experiment 1. Below we elaborate on the third and fourth purpose in more detail.

2.1. Norm conflict experiments and the problem of arbitration

Our research question concerning individual variation and latent classes of diverging moral commitments is motivated by a more general problem affecting the use of norms in empirical studies

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

of rationality. Since multiple norms often apply to a given task, Elqayam and Evans (2011) argue that the problem of arbitrating between competing norms is one of the main problems in cognitive psychology preventing the application of norms of judgment and decision making in experimental research. In contrast, the present experiment aims to show that by classifying participants according to different morality profiles based on divergent norms, the possibility of competing norms (e.g., Utilitarianism vs. Deontology) can be utilized for studying individual variation. For this purpose, participants' reflective attitudes are elicited via the Scorekeeping Task, which puts participants in the position of a scorekeeper (Skovgaard-Olsen, 2026; Skovgaard-Olsen and Cantwell, 2023; Skovgaard-Olsen et al., 2019). The scorekeeper assesses the performance of fictive participants, who have produced incompatible responses to the task that the participants have just completed. Through this task, participants commit to a moral outlook by criticizing and sanctioning their fictive peers based on the fictive participants' criticism of each other. A comparison is then made between participants' reflective attitudes and their own case judgments to investigate how well they match. Using this experimental task together with Bayesian latent class and hierarchical MPT models, permits us to study detailed patterns of individual variation in moral perspectives. Below and in Appendix A, we go into further details of how the Scorekeeping Task was implemented and how the latent classes were estimated. In the general discussion, we return to the issue of how to interpret participants' commitments to an ethical theory like Utilitarianism in light of recent discussions about different degrees to which participants' responses may match a given ethical theory (Conway et al., 2018a).

2.2. Instrumental harm and a concern for the greater good

Once latent classes are established, we use these to probe participants' motivation and whether they act for the right moral reasons (May, 2018).

Kahane et al. (2015, 2018) find evidence for a dissociation between accepting instrumental harm in sacrificial dilemma and having an impartial concern for the greater good. Kahane et al. (2015, 2018) therefore doubt that performance on sacrificial dilemma tells us much about participants' adherence to the positive core of Utilitarianism, which nowadays is associated with the effective altruism program. To further motivate this view, Kahane et al. (2015) can point to findings like the following. Phinea Gage-like patients with damage to the ventromedial prefrontal cortex, who display emotional deficits like absence of guilt, shame, and empathy, respond at higher rates of utilitarian judgments in traditional sacrificial dilemma (Koenigs et al., 2007; see also Mendez et al., 2005). Positive associations with both sub-clinical psychopathy (Marshall et al., 2018) and Machiavellianism (Bartels and Pizarro, 2011) have been found along with diminished emphatic concern among utilitarian respondents (Choe and Min, 2011, but see also Baron et al., 2018). These lines of evidence question whether endorsement of sacrificial harm is a valid indicator of an impartial concern with the greater good; it may just indicate a decreased aversion to causing the death of a person (Bartels and Pizarro, 2011). Conway et al. (2018a) replied to Kahane et al. (2015), as discussed further below. From their discussion, we extract the following competing hypotheses, which relate the C parameters of the Conflict model and the latent class of Utilitarianism to a range of egoistic and altruistic covariates.

> **H_1 (concern for greater good)**: Acceptance of instrumental harm reflects a genuine concern for the greater good. In this case, we should expect this tendency to generalize to responses in other contexts in which an impartial utilitarian concern competes with self-interest and other moral concerns. Predictions: 1) a negative correlation between C and moral egoism and subclinical psychopathy, 2) a positive correlation between C and identification with all of humanity (IWAH), 3) a higher proportion of self-sacrifice for the latent class of Utilitarianism in sacrificial trilemma that permit self-sacrifice in addition to sacrificing an innocent stranger, and 4) higher rates of altruistic behavior for the latent class of Utilitarianism in Greater Good scenarios that contrast self-interest and altruistic choices.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

H_2 (restricted concern for greater good): Acceptance of instrumental harm reflects a moral disposition to minimize harm in the interest of the greater good, which does not generalize to an unrestricted altruism toward strangers. That is, in the context of the sacrificial dilemma, participants show a genuine moral disposition, but it is one that does not generalize to other contexts. Predictions: 1) no association between C and IWAH, 2) negative associations between C and both egoism and subclinical psychopathy, and 3) no higher rates of self-sacrifice in moral trilemma that include the self-sacrifice option as a third option, and of altruistic behavior on greater good scenarios, for the latent class of Utilitarianism.

H_3 (antisocial motivation): Acceptance of instrumental harm reflects a narrow disposition largely driven, not by a concern for the greater good, but by antisocial tendencies. Predictions: 1) a *negative* association between C and markers of genuine concern for the greater good such as IWAH, 2) lower proportion of self-sacrifice in moral trilemma that include this third option and altruistic behavior in Greater Good scenarios in the latent class of Utilitarianism than the other latent classes, and 3) a *positive* association between C and selfish moral views and dispositions (i.e., a positive association between C and both primary psychopathy and moral egoism).

These research hypotheses mention egoistic and altruistic indicators (e.g., IWAH), which participants completed in our studies and which are described further below. In Experiment 1, we investigate these 3 competing motivational hypotheses through the lens of the latent classes we estimate. Our goal will be to investigate which of these 3 hypotheses best characterizes a latent class of Utilitarianism, if such a class can be found.

2.3. Method

2.3.1. Participants

The experiment was conducted over the Internet to obtain a large and demographically diverse sample. A total of 1,178 people completed the experiment. The participants were sampled through the Internet platform Mechanical Turk via CloudResearch from the USA, UK, Canada, and Australia. They were paid a small amount of money for their participation (on average $6 per hour) and told that there would in addition be a $1 bonus, if they answered accurately and participated in the second half 1 week later. The following exclusion criteria were used: not having English as native language, completing the task in more than 2 standard deviations below or above the mean completion time, failing to answer at least one of 2 simple SAT comprehension questions correctly in a warm-up phase, and answering 'not seriously at all' to the question 'How seriously do you take your participation' at the beginning of the study. The final sample for Session 1 consisted of 977 participants. Of these 977 participants, 787 (80.56%) participated in Session 2, 1 week later. To fit the Conflict model, data from both sessions are required. Thus, our analyses are focused on the latter participants.

Mean age of these 787 participants was 42.5 years, ranging from 19 to 91.[5] Among these, 40.53% self-identified as male; 57.18% as female; 11 participants indicated that they were non-binary, and 7 participants preferred not to reveal their gender. 70.65% indicated that the highest level of education that they had completed was an undergraduate degree or higher. Applying the exclusion criteria had a minimal effect on the demographic variables. For the analysis, we focus on the 769 participants who had participated in both sessions and who self-identified as either male or female to compare with previously reported gender effects.

2.3.2. Design

Each session had a within-participants design with the following factors varying within participant: consequence (smaller vs. greater) and norm (proscriptive vs. prescriptive). To allow for 3 trial replications for each of the 4 CNI conditions (3×4), each participant in total worked through 12

[5]We ignore two values below 18 given that MTurk limits the participation to adults.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

scenarios within a session. Across the 2 sessions, the following factor was varied within participant: causal structure (intended means vs. foreseeable side-effect). The same 12 background scenarios appeared in both sessions, but each scenario was presented in the intended means version in one session and the foreseeable side-effect version in the other session. In total, there were 24 items (12 scenarios × 2 causal structures) and 8 within-participant conditions (4 CNI conditions × 2 causal structures).

2.3.3. Materials and procedures

Participants were presented with the 4 CNI conditions across 12 background scenarios in random order and with random assignment of condition to scenario. Six of these scenarios (the Transplant, Vaccine, Bishop, Immune Deficiency, Dialysis, and Rwanda) were based on Gawronski et al. (2017) and Körner et al. (2020) but further modified. Two scenarios (Hospital, Sinking Boat) were based on Moore et al. (2008) but further modified and 4 scenarios were new (Welfare, Carcrash, Undercover, Avalanche Disaster).[6]

An example of the scenarios is displayed in Table 3. While the scenarios followed the form of the scenarios in Gawronski et al. (2017) and Körner et al. (2020) to implement the CNI conditions, we chose to replace some scenarios because the proposed action had the flavor of murder (e.g., shoving a person from the edge of a building in the Construction site scenario, or adding peanut oil that would kill a person with an allergy to a dish in the Peanuts scenario). In these cases, the proposed action did not fall within the purview of the professional duties of the agent described (construction worker, the head of a restaurant). In other cases, scenarios from Gawronski et al. (2017) and Körner et al. (2020) were replaced, because they depicted situations involving special norms (e.g., like torturing a guilty victim in the Torture scenario, and not negotiating with terrorists in the Abduction scenario). Others were replaced, because they did not permit us to parallelize the ratio of victims saved and sacrificed (e.g., the Assisted Suicide scenario and the Mother scenario), or because the victim had personal relationships to the agent deciding over the sacrifice (the Mother scenario).

Rosas and Koenigs (2014) criticize past sacrificial dilemma research for introducing confounds like *fated victims*, who would die irrespectively of the sacrifice, *guilty victims*, who deserve to be sacrificed due to their own immoral behavior, and *selfish reasons* for selecting the 'utilitarian option'. We carefully modified the scenarios to avoid these confounds. Across all scenarios, we further parallelized (1) the ratio of victims saved and sacrificed (5 vs. 1), (2) the probability terms describing how likely the outcomes of action/inaction would be, and (3) the way in which the smaller benefits were implemented in the ProSmaller and PreSmaller conditions, so that the gain of sacrificing an innocent person was trivially less than the cost of the sacrifice across all scenarios. Finally, we implemented a foreseeable side-effect and an intended means version of each scenario described below.

As already mentioned, the scenarios involve a manipulation of norms (proscriptive vs. prescriptive). Baron and Goodwin (2020) criticized the implementation of prescriptive norms in Gawronski et al. (2017) for requiring participants to interfere with the preselected action of another person, who in some cases was depicted as the participants' colleague or superior, which in turn raises unintended prospects of hurting this person's feelings, violating lines of authority, or risking retaliatory actions. To circumvent this problem, we chose to implement the prescriptive norm condition differently (see also Henning and Hütter, 2020, 2021 for a related solution). In our implementation, we stipulated that a person lower in the professional hierarchy (e.g., an assistant to the participant) had preselected the presented action in the prescriptive condition *under the assumption that this was what the participant wanted*. If unhappy with this preselection, participants could reverse the planned action without going against the will of another agent. Note that while deontologists, such as Kamm (2007), weigh negative rights (not to be harmed) stronger than positive rights (to be aided), this distinction does not carry over to the contrast between proscriptive and prescriptive norms under the present implementation. For what we label as 'prescriptive norms' here are not new deontological norms (e.g., norms to aid), but

[6]The complete list of scenarios can be found at the OSF project page: https://osf.io/9fer2. In BSc and MSc theses supervised, we calibrated these scenarios through pilot studies.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 3. *CNI conditions as implemented in one of the scenarios.*

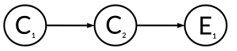
ProGreater	ProSmaller	PreGreater	PreSmaller
		Shared cover story	
You are the director of a medical welfare program which supports poor people financially. You are treating seriously ill patients under this program. One of them has a rare condition which can only be treated with some very expensive drugs.	You are the director of a medical welfare program which supports poor people financially. You are treating seriously ill patients under this program. One of them has a rare condition which can only be treated with some very expensive drugs.	You are the director of a medical welfare program which supports poor people financially. You are treating seriously ill patients under this program. One of them has a rare condition which can only be treated with some very expensive drugs.	You are the director of a medical welfare program which supports poor people financially. You are treating seriously ill patients under this program. One of them has a rare condition which can only be treated with some very expensive drugs.
		Intended means	
You could end the patient's treatment and redirect the funds to cover the treatment of 5 new patients. If you stop the treatment of the one patient with the rare condition, this patient will die, but you will save the lives of 5 other patients.	You could end the patient's treatment and redirect the funds to pay for better beds for 5 new patients to improve their lives. If you stop the treatment of the one patient with the rare condition, this patient will die, but the 5 other patients will sleep better.	You could end the patient's treatment and redirect the funds to pay for better beds for 5 new patients to improve their lives. If you stop the treatment of the one patient with the rare condition, this patient will die, but the 5 other patients will sleep better.	You could end the patient's treatment and redirect the funds to cover the treatment of 5 new patients. If you stop the treatment of the one patient with the rare condition, this patient will die, but you will save the lives of 5 other patients.
		Foreseeable side-effect	
You could replace the patient's treatment with a cheaper alternative and redirect the funds to cover the treatment of 5 new patients. If you replace the treatment of the one patient with the rare condition with a cheaper alternative, this patient will highly likely die, but you will save the lives of 5 other patients.	You could replace the patient's treatment with a cheaper alternative and redirect the funds to pay for better beds for 5 new patients to improve their lives. If you replace the treatment of the one patient with the rare condition with a cheaper alternative, this patient will highly likely die, but the 5 other patients will sleep better.	You could replace the patient's treatment with a cheaper alternative and redirect the funds to pay for better beds for 5 new patients to improve their lives. If you replace the treatment of the one patient with the rare condition with a cheaper alternative, this patient will highly likely die, but the 5 other patients will sleep better.	You could replace the patient's treatment with a cheaper alternative and redirect the funds to cover the treatment of 5 new patients. If you replace the treatment of the one patient with the rare condition with a cheaper alternative, this patient will highly likely die, but you will save the lives of 5 other patients.
		Shared choice	
The action: to end the treatment of the one patient with the rare condition. [/replace with cheaper alternative] Is it morally acceptable in this case to carry out the action?	The action: to end the treatment of the one patient with the rare condition. [/replace with cheaper alternative] Is it morally acceptable in this case to carry out the action?	The action: to end the treatment of the one patient with the rare condition. [/replace with cheaper alternative] Assuming that this is what you wanted, your staff has planned to carry out the action in order to buy better beds for the 5 new patients. However, you could prevent this. Is it acceptable in this case to prevent the action from being carried out?	The action: to end the treatment of the one patient with the rare condition. [/replace with cheaper alternative] Assuming that this is what you wanted, your staff has planned to carry out the action in order to cover the treatment of 5 new patients. However, you could prevent this. Is it acceptable in this case to prevent the action from being carried out?
		Yes, it is acceptable vs. No, it is not acceptable vs. Skip	

Note: Example of a CNI scenario in both the intended means and foreseeable side-effect conditions. First, a shared cover story is presented. Next, the continuations implementing the intended means vs. foreseeable side-effect conditions are displayed. Finally, the invariant choice is depicted along with the 3 response options (Action vs. Inaction vs. Skip). The words 'the action' were colored in blue. The ProGreater, ProSmaller, PreGreater, and PreSmaller abbreviations are defined in Table 1. In the Proscriptive condition, what Figures 1–3 label as "Action" ["Inaction"] corresponds to responding "Yes" ["No"] to whether it is acceptable to carry out the action; in the Prescriptive condition, it corresponds to responding "Yes" ["No"] to whether it is acceptable to prevent the action.

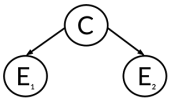
https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

just the old deontological norms against sacrificial harm, which are now applied in a context in which accepting the harm is a default set by an assistant. As a validation, Appendix B presents an analysis of action/inaction responses in the congruent trials, which are neither predicted by Utilitarianism nor Deontology. In Baron and Goodwin (2020, 2021) and Gawronski et al. (2020), such responses are denoted as 'perverse', but we here prefer the less value-laden term 'abnormality rates', which indicates that the response pattern goes against the expectations of 'normality' and is in this sense a form of pathological response–while remaining neutral on its etiology. In this analysis it is found that the modified scenarios do indeed reduce the rates of abnormal responses and that further modeling can reveal one mechanism leading to abnormal responses (see Appendix B).

Table 3 illustrates how the CNI conditions were implemented. Intended means scenarios follow the structure of a causal chain, in which the sacrifice of 1 person (C_2) as an act is presented as the only causal means for saving the 5 people (E_1).



Foreseeable side-effect scenarios follow the causal structure of a common cause, where the intended action has 2 effects; the intended outcome (E_1 = saving 5 people) and an unintended but foreseen side-effect (E_2 = killing 1 person).



The distinction is conceptually similar to the direct versus indirect harm distinction in Royzman and Baron (2002) and the distinction between instrumentality and non-instrumentality in Fahrenwaldt et al. (2025). In the intended means condition, the harm is direct since there would be no way to get the desired outcome if the harm did not occur. In this case, the victim's mind or body is used as an instrument to save others. In the foreseeable side-effect condition, the harm is indirect and obtaining the desired outcome is not required to occur. Rather, harm to the victim is a side-effect of the agent's action.

Note, the direct form of harm in our scenarios does not take the form of direct use of personal force (e.g., pushing, smothering) as in Greene et al. (2001) and Greene (2013). Often these two factors (i.e., use of the victim as an instrument and direct personal harm) are coupled in studies (Fahrenwaldt et al., 2025). Yet, our manipulation *does* involve a coupling of causal structure with a difference in outcome certainty. In intended means scenarios, we state that the sacrificed person 'will die' if the intervention succeeds, because the sacrifice is presented as the only available means (if the goal is achieved, the sacrifice must have occurred). In foreseeable side-effect scenarios, we state the sacrificed person 'will highly likely die', since the sacrifice is a predicted consequence of a common cause intervention (which permits that the intervention succeeds but the prediction fails because the side-effect is prevented).[7]

[7]We consider this coupling of causal structure with outcome certainty to be appropriate. In intended means scenarios, the sacrifice (C_2) is a required intermediate step in the causal chain from the agent's intervention (C_1) to the goal (E_1). When an agent commits to achieving E_1 through this chain, C_2 must occur, thus the victim "will die". In side-effect scenarios, the sacrifice (E_2) is a predicted but not required consequence of a common-cause structure ($E_2 \leftarrow C \rightarrow E_1$). Here E_1 can occur even if E_2 does not, since E_2 is not on the same causal path as E_1, thus the victim "highly likely will die". Correspondingly, classic examples of intended means (direct bombings of civilians, active euthanasia) involve certain harm if the action succeeds, while classic examples of foreseen side-effects (tactical bombing of military targets near civilians, passive euthanasia) involve likely but uncertain harm.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 4. *Overview of Experiment 1.*

	Session 1	Session 2
CNI scenarios	Intended means/foreseeable side-effect	Foreseeable side-effect/intended means
Altruistic covariates$_1$	Trilemma self-sacrifice	–
Egoistic covariates$_1$	Trilemma other-sacrifice	–
Scorekeeping task	With CNI items	With CNI Items
Altruistic covariates$_2$	Empathy	Greater good scenarios + Scorekeeping task
Egoistic covariates$_2$	Primary psychopathy	Rational egoism, moral egoism, psychological egoism
Demographic variables	Gender, age, education, language, nationality	Gender, age, education, language, nationality
Altruistic covariates$_3$	–	IWAH (with IWAA and IWAC as control).

Note: Session 1 and Session 2 were separated by one week. Participants who were randomly assigned to intended means in Session 1 saw the CNI scenarios in the foreseeable side-effects condition in Session 2 as a within-participant manipulation and *vice versa*. Blocks with covariates$_2$ were presented in random order. 'Greater good scenarios' = scenarios from Kahane et al. (2015). 'IWAH' = identification with all of humanity. 'IWAC' = identification with local community. 'IWAA' = identification with own nationality, which in previous studies was restricted to Americans but is here tailored to each participant's home country.

2.3.3.1. Presentation of scenarios

The order of the scenarios and the CNI conditions within scenarios were randomized for each participant anew.[8] Following Gawronski et al. (2017), participants were given an instruction asking participants to pay close attention to small details in scenarios that seem similar and warning them that some of the scenarios may be unpleasant to think about since they related to difficult real-life issues. In addition, participants were given the following instructions for use of the skip option: 'There is now also the option to "skip" a moral decision for cases, where you are undecided about whether the described action is morally acceptable or inacceptable. Please do not use the "skip" option more than 2 times'.

In each of the 2 sessions, participants first completed 2 practice items and then completed the moral judgments either for the intended means or foreseeable side-effect condition (with the order of these 2 causal structures randomized for each participant). The 12 background scenarios were presented in random order with random assignment of scenario to CNI condition. Experiment 1 followed Kahane et al. (2015) and Conway et al. (2018a) in measuring a range of covariates, which are categorized as egoistic or altruistic in Table 4.

2.3.4. Altruistic vs. egoistic covariates

The self versus other sacrifice options in Table 4 refer to a switch version of the run-away-Trolley (Huebner and Hauser, 2011; Thomson, 2008) in which participants are presented with the third option of sacrificing themselves in addition to the 2 classical options of action and inaction. Participants' level of psychopathy was probed via Levenson et al.'s (1995) subscale for primary psychopathy in a noninstitutionalized population. The subscale consists of 16 Likert-scaled items (e.g., 'For me, what's right is whatever I can get away with'), some of which are reverse-coded (α = .91). Participants' empathy was measured via the *Empathic Concern Subscale* of the *Interpersonal Reactivity Index* (Davis, 1980), which consists of 7 Likert-scaled items (e.g., 'I often have tender, concerned feelings for people less fortunate than me'), some of which are reverse-coded (α = .89).

[8]Gawronski et al. (2017) used a pseudo-random order, which is fixed to be the same for each participant. In pilot studies, we did not find differences between this procedure and the more rigorous randomized order and chose the latter instead.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 5. *Scorekeeping judgments.*

	Approve HIT	Accept criticism
Intended means	Deontology vs. Utilitarianism	Deontology vs. Utilitarianism
Foreseeable side-effect	Deontology vs. Utilitarianism	Deontology vs. Utilitarianism
Greater Good	Permitted vs. Obligatory	Permitted vs. Obligatory

Note: The table provides an overview of the scorekeeping judgments that participants made across 2 sessions: approving HITs and accepting criticism for each member of the 3 pairs. In Session 1 (week 1), participants made scorekeeping judgments for the CNI scenarios in the intended means or foreseeable side-effect condition (depending on random assignment). In Session 2 (week 2) participants made scorekeeping judgments for the CNI scenarios in the remaining foreseeable side-effect or intended means condition and for the Greater Good scenarios. On Mechanical Turk, tasks for participants are called 'HITs' ('Human Intelligence Tasks'), which is a label that is familiar to the participants.

The *Identification with All Humanity Scale* from McFarland et al. (2012) was used, which consists of 9 items (e.g., 'How much would you say you care (feel upset, want to help) when bad things happen to a) people in my community, b) people from your home country, and c) people anywhere in the world'). In each case, participants indicate how much they identify with all of humanity (IWAH), their local community (IWAC), or with people from their home country (IWAA). The IWAH (α = .91), IWAC (α = .88), and IWAA (α = .91) items were then presented as the final task with participants' preselected home country used for the IWAA items. In Table 4, IWAH is listed as falling under the 'altruistic covariates'. The IWAC and IWAA items were used as control items.

Participants' adherence to psychological, moral, and rational egoism was measured with 3 items from Kahane et al. (2015). *Psychological egoism* is the view that deep down people are always motivated by their self-interests in the end. *Rational Egoism* is the view that promoting one's own self-interest is the only rational thing to do. *Moral Egoism* is the view that promoting one's own self-interest is the only moral thing to do. The Greater Good scenarios were also based on Kahane et al. (2015). In these scenarios participants are presented with 7 dilemmas involving conflicts between self-interested and altruistic actions (α = .70). In a typical scenario, participants have the choice of either using money on their own luxury needs or on saving the lives of strangers in remote, developing countries. The task asks participants to judge how wrong they find the self-interested choice, which either privileges themselves, their family, or country over the urgent needs of distant strangers, on a 5-point Likert scale from ('Not at all wrong') to ('Very wrong').

2.3.5. The scorekeeping task

To elicit commitments to competing norms, a scorekeeping task was employed (following Skovgaard-Olsen, 2026; Skovgaard-Olsen and Cantwell, 2023; Skovgaard-Olsen et al., 2019). Across the 2 sessions, participants encountered 3 pairs of fictive participants who had given competing solutions to the tasks that the participants had just completed and who were criticizing one another. For each of the 2 causal structures, one of these fictive participants was defending the deontological solution and the other was defending the utilitarian solution. In addition, participants encountered a pair of fictive participants disagreeing over whether performing the self-sacrificing greater good deed in the Greater Good scenarios from Kahane et al. (2015) was morally permissible (but supererogatory) or morally obligatory. For each of the 3 pairs, participants judged whether the 2 presented criticism were compelling ('yes' vs. 'no') and selected the fictive peer whose task completion should be approved (see Table 5).

2.3.5.1. Procedure

Participants were randomly assigned to a pair of fictive peers presented via pictorial avatars and first names with the avatar's gender (male vs. female) and race (Caucasian vs. person of color) held constant within a pair. To illustrate the Scorekeeping Task, we here consider its application to the CNI scenarios. Participants were told that a pair of fictive peers had completed the same task as the one they just

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 6. *The scorekeeping task, the 2 conflicting responses.*

	Response	Criticism of peer
	Natalie: 'Yes, it is acceptable'	How can you say that it is not acceptable to carry out the action and save the greatest number of people? This makes no sense. Overall, the consequences would be better!
	Emma: 'No, it is not acceptable'	How can you say that it is acceptable to sacrifice the life of an innocent person? This makes no sense. The end does not justify the means!

Natalie and Emma cannot both be right!
In the following, you will see their responses repeated along with their criticism of each other.
Based on their mutual criticism, please help us decide whose response is most adequate and whose HIT should be approved.

Note: In the upper half, the left side shows the fictive participants' response; the right side shows his/her criticism of his/her peer opponent. The avatars and names were randomly varied such that gender (male vs. female) and race (Caucasian vs. person of color) were kept constant within a pair. This pair illustrates the opposition between Utilitarianism and Deontology.

finished but that the 2 peers responded very differently. Participants were then presented with one of the scenarios either in intended means or foreseeable side-effect condition (depending on which they had just completed). Then followed (on separate pages) the information illustrated in Table 6.

On Mechanical Turk, tasks are described as 'HITs' ('Human Intelligence Tasks') to participants and the approval of HITs determine whether a participant is paid. Since participants are financially rewarded by approvals of HITs and build up a reputation by completing these tasks, the approval of HITs was used as an ecologically valid sanctioning measure in Skovgaard-Olsen et al. (2019) that participants are motivated to reason about. Here we adopted this measure as well. After the presentation of the criticism, participants were asked whether they found the given criticism compelling ('yes' vs. 'no'). Finally, participants were presented with a page in which both responses were repeated with the instruction that they should indicate whose 'HIT' on Mechanical Turk should be approved after having seen, e.g., Natalie's and Emma's mutual criticism.

2.4. Results

The data analysis proceeded in several steps. First, participants' responses in the Scorekeeping Task were used to establish latent classes. Second, the Conflict model was fitted to the data in competition with other models, and the latent classes from the Scorekeeping Task were used to estimate latent group-specific means in the Conflict model. Finally, the relationships between the model parameters and the egoistic and altruistic covariates were investigated using structural equation modeling (SEM).[9]

A highest density interval (HDI) is an interval of the posterior distribution in which all points inside the interval have a higher probability density than points outside the interval. Below we report effects as credible when their 95% HDs excludes zero, meaning all values in the 95% most credible region agree on the direction of the effect.

2.4.1. Latent class analysis

In Appendix A, a latent class model is outlined, which was fitted to the scorekeeping judgments to classify participants into latent classes. This analysis was repeated for 2, 3, and 4 class solutions,

[9]R-scripts and data for all experiments have been uploaded to the OSF project page for this paper: https://osf.io/9fer2.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 7. *Latent class models.*

No. of classes	WAIC	LOOIC	Δelpd (SE)	Weight
2	7334.2	7435.3	−927.36 (37.65)	0
3	6317.6	6460.7	−440.05 (20.48)	0
4	5456.1	5580.6	–	1.00

Notes. The number of classes refers to the number of distinct qualitative patterns identified by the latent class analysis. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. Note that information criteria can take both positive and negative values and that the lowest value indicates the best performance.

each representing a distinct qualitative pattern. Table 7 compares the model fits of these solutions via information criteria. Substantially, the model favored by these criteria is evaluated on whether it yields theoretically meaningful class profiles.

The information criteria which assess the expected out-of-sample predictive accuracy of the fitted models indicate that a model with 4 classes best fit the data (since it has lower values on the information criteria). Figure 4 displays the posterior median and 95% HDI of participants' scorekeeping performance as a function of their assigned latent class.

As Figure 4 shows, it was possible to separate the following latent classes. First, a class of Deontology (N = 426, 54.1%), whose members accepted the HIT of Deontology across both the intended means and the foreseeable side-effect conditions, and had the highest posterior probability of accepting the criticism of Utilitarianism. Second, a class of Utilitarianism (N = 171, 21.7%), with the converse behavior. Third, a class of participants (N = 82, 10.4%), whose members sided with Deontology for intended means and Utilitarianism for foreseeable side-effect was identified. Due to this distinctive switching behavior, this third latent class was interpreted as following the Doctrine of DoubleEffect. Finally, a fourth latent class (N = 108, 13.7%) was identified exhibiting neither strong inclinations toward Utilitarianism nor toward Deontology. Instead, members of this latent class stood out by accepting that the altruistic deed was morally obligatory in the Greater Good scenarios. In contrast, the other 3 latent classes merely viewed the altruistic deed as morally permissible in the Greater Good scenarios. For this reason, we selected the label 'Altruism' for this fourth latent class.

To further validate the designation of 'Altruism' to the fourth latent class, participants' performance on all the greater good scenarios were investigated as a function of their latent class (see Figure 5). For this analysis, a Bayesian linear regression model was fitted with the latent classes as predictor and a sum-score for the performance on the Greater Good scenarios as dependent variable, which had been transformed to range within the interval [0,1], with higher values indicating a preference for the altruistic deed over the self-interested option. Bayesian regression models here and throughout the paper were fitted using the R package brms (Bürkner, 2017). The R-package tidybayes (Kay, 2023) was used to calculate the 95% HDI and display the posterior distributions in Figure 5.

As shown in Figure 5, the latent class of Altruism had a higher posterior probability of finding the choice of the self-interested action over the altruistic deed wrong across Greater Good scenarios. Among the remaining 3 latent classes, Utilitarianism did not differ from the latent class of DoubleEffect but had a higher posterior probability of finding the self-interested action wrong than Deontology.

In Appendix B, a self-sacrifice version of the switch Trolley scenario was analyzed and it is found that the participants in the latent classes had differential preferences for sacrificing a stranger, do nothing, or sacrifice themselves to save 5 people. Participants in the latent class of Utilitarianism and Doctrine of Double Effect showed the strongest preference for sacrificing a stranger, whereas participants in the latent class of Deontology preferred to do nothing and participants in the Altruism latent class preferred to sacrifice themselves.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

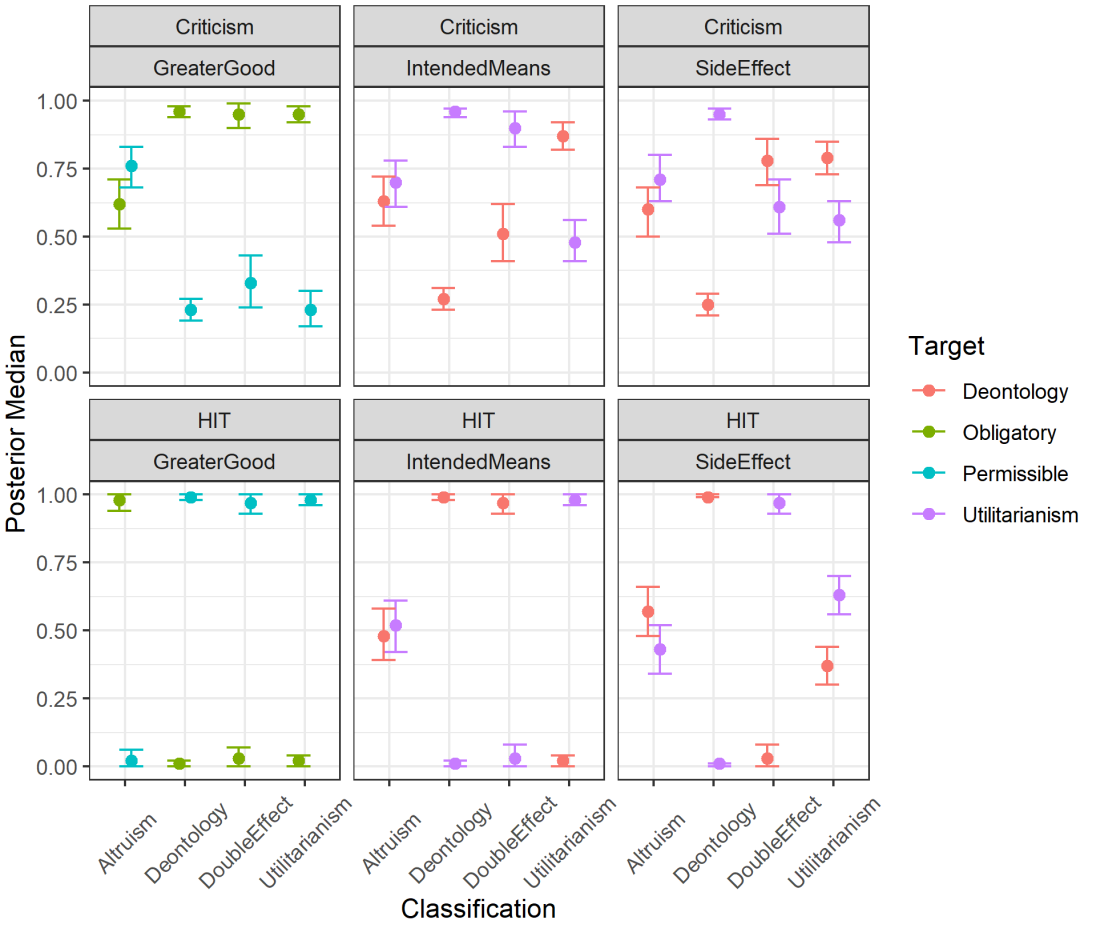


Figure 4. *Latent classes and the scorekeeping task.*

Note: Posterior median estimates with 95% HDI error bars for parameter estimates of the 4-class model. Rows: Criticism (top) vs. HIT approval (bottom). Columns: GreaterGood, IntendedMeans, and SideEffect avatar pairs. Colors (evaluation targets): Deontological avatar; Obligatory GreaterGood avatar; Permissible GreaterGood avatar; Utilitarianism avatar. X-axis (latent classes): Altruism accepts obligatory GreaterGood deed; Deontology always rejects sacrifices; DoubleEffect rejects sacrifices for intended means but accepts sacrifices for side-effects; Utilitarianism always accepts sacrifices.

2.4.2. Multinomial processing tree models

Next, participants' moral judgments were analyzed via 4 models (Table 8). These models were fitted to estimate individual MPT parameters but they aggregate over items and trial replications to be less sensitive to scenario-specific content effects. While the hierarchical model could be extended to include random scenario effects (Matzke et al., 2013) to assess interactions between scenarios and experimental conditions, such analyses would require substantially larger within-participant designs.

$CNIS_6$ and $CNIS_{14}$ extend the CNI model by a skip option. $CNIS_{14}$ differs from $CNIS_6$ in that it does not impose an invariance assumption for the C and N parameters across the CNI conditions (ProGreater, ProSmaller, PreGreater, PreSmaller) implicit in the CNI model of Gawronski et al. (2017). In each case, separate C and N parameters were estimated for the intended means and foreseeable side-effect conditions. Appendix A contains further details on their implementation in R and the Bayesian estimation of their model parameters. In Skovgaard-Olsen and Klauer (2024), $CNIS_{14}$ was found to be superior to $CNIS_6$.[10] Here, these 2 models are further contrasted with a new Conflict model which

[10] Yet, those studies only used the intended-means scenarios.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

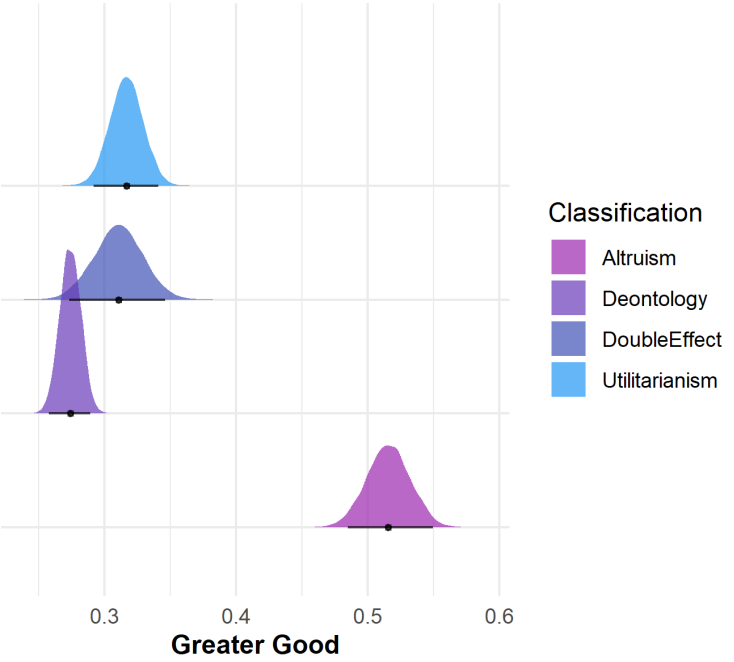


Figure 5. *Greater good judgments across latent classes.*
Note: The black dot and line indicates the posterior median and 95%-HDI. The sum score for the greater good scenarios was transformed to range within the unit interval.

Table 8. *Model comparison.*

Model	Parameters
$CNIS_6$	C_{intend}, C_{side}, N_{intend}, N_{side}, I, S
$CNIS_{14}$	$C^{Pro>}_{intend}$, $C^{Pro>}_{side}$, $C^{Pro<}_{intend}$, $C^{Pro<}_{side}$, $C^{Pre>}_{intend}$, $C^{Pre>}_{side}$, $C^{Pre<}_{intend}$, $C^{Pre<}_{side}$, N^{Pre}_{intend}, N^{Pre}_{side}, N^{Pro}_{intend}, N^{Pro}_{side}, I, S
$CNIS\text{-}Conflict_{14a}$ and $CNIS\text{-}Conflict\text{-}latent_{14a}$	$C^{Pro>,Pre>}_{intend}$, $C^{Pro>}_{side}$, $C^{Pro<}_{intend}$, $C^{Pro<}_{side}$, $C^{Pre>}_{side}$, $C^{Pre<}_{intend,side}$, N^{Pre}_{intend}, N^{Pre}_{side}, N^{Pro}_{intend}, N^{Pro}_{side}, I, S, Res, Con

Note: 'Pro>' = ProGreater, 'Pro<' = ProSmaller, 'Pre>' = PreGreater, 'Pre<' = PreSmaller. See Appendix A.

was implemented in different variations in a model fitting exercise. In Table 8, we only report the best performing version. In $CNIS\text{-}Conflict_{14a}$, it is tested whether adding a conflict architecture to the CNI model results in a better performing model.[11] $CNIS\text{-}Conflict\text{-}latent_{14a}$ is identical to this model but allows for latent group-specific means of the estimated MPT parameters based on the 4 latent classes in the scorekeeping phase.

Figure 6 compares the observed response frequencies and the posterior predictive behavior of the models. As shown, severe misfits between the observed response proportions and $CNIS_6$ were found. One example is that the model predicts an equal observed rate of skip responses based on $(1-C) \times (1-N) \times S$, in each type of scenario within the intended means and foreseeable side-effect conditions. It is thereby unable to account for the asymmetries between rates of skip responses between congruent and incongruent items. In addition, forcing C and N parameters to remain constant across conditions limits the model's ability to fit the data. In contrast, the 3 other models perform increasingly

[11]The results in Skovgaard-Olsen and Klauer (2024) were used assess for which of the C and N parameters invariance violations were most likely to occur.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 9. *Model comparison.*

	WAIC	LOOIC	Δelpd (SE)	p_{T1}	p_{T2}
CNIS-Conflict-latent$_{14a}$	15655.25	16358.42	–	.13	.0003
CNIS-Conflict$_{14a}$	15695.94	16465.91	−53.75 (14.6)	.12	.004
CNIS$_{14}$	16001.89	16620.29	−130.94 (24.7)	.00	.00
CNIS$_{6}$	17693.23	18034.22	−837.90 (42.2)	.00	.00

Note: Model comparison indices. LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. The test statistics T_1 and T_2 represent Bayesian p values and are based on the posterior predictive model checks in Klauer (2010).

well. All 3 models are also able to capture the asymmetry between skip responses across congruent and incongruent conditions. But unlike CNIS$_{14}$, the conflict models can capture this asymmetry by separately modeling skip responses that arise as the result of conflict detection, whereas CNIS$_{14}$ is forced to adjust the C_j and N_k parameters to make the activation of the $(1-C_j) \times (1-N_k) \times S$ differ between conditions. The best fit is obtained with CNIS-Conflict-latent$_{14a}$.

2.4.3. Model comparison

In Table 9, a formal model comparison is made of these 4 models based on (1) the expected out-of-sample predictive accuracy (assessed via information criteria) and (2) a test of whether there are statistically significant misfits of the models as indicated by the posterior predictive checks (T_1, T_2) proposed in Klauer (2010), where a small (Bayesian) p value for each indicates that the model fails to capture an aspect of the data.

As in Skovgaard-Olsen and Klauer (2024), it was found that the CNI model of Gawronski et al. (2017) extended with a skip parameter (CNIS$_6$) had a worse fit to the data than a version of the same model (CNIS$_{14}$), which permitted violations of the invariance assumption by estimating separate C and N parameters for the different CNI conditions. In addition, Table 9 shows that this model in turn was outperformed by models with the Conflict architecture.

LOOIC is an information criterion which is based on the expected out-of-sample predictive accuracy of the models. Lower LOOIC values indicate a better fit in the light of the parsimony vs. fit trade-off. Using this criterion, it was found that despite its extra structure, the Conflict model with the latent classes provides the best trade-off between parsimony and data fit. Similarly, it was found that only the conflict models were able to pass the posterior predictive check for the aggregate outcome frequencies (T_1) proposed in Klauer (2010).[12]

2.4.4. Analysis of parameter estimates

Figure 7 displays the posterior median parameter estimates of the CNIS-Conflict-latent$_{14}$ model, with C and N parameters averaged across the CNI conditions.

The posterior parameters follow expected qualitative patterns such as $N_{intend}^{Deontology} > N_{intend}^{Utilitarianism}$, $N_{intend}^{DoubleEffekt} > N_{intend}^{Utilitarianism}$, $N_{side}^{Deontology} > N_{side}^{Utilitarianism}$, $N_{side}^{Deontology} > N_{side}^{DoubleEffect}$, $C_{intend}^{Deontology} < C_{intend}^{Utilitarianism}$. In words: the latent classes Deontology and DoubleEffect had higher averaged N parameters for intended means than Utilitarianism and Deontology had higher averaged N parameter for foreseeable side-effect than Utilitarianism. Conversely, the latent class of Utilitarianism had higher averaged C parameters than Deontology for intended means. Table 10 reports for each parameter displayed in Figure 7 those pairwise contrasts between latent classes for which the 95% HDI of the contrasts excluded zero, that is, for which credible differences between the two classes were found.

[12]None of the models passed the posterior predictive check for the variability (T_2) across individuals, indicating that these models also failed to capture aspects of the data. This is, however, not unusual for large data sets such as the present, and in pilot studies with smaller sample sizes, CNIS-Conflict-latent$_{14a}$ was able to pass both posterior predictive checks.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 10. *Contrasts in class-specific means.*

Parameter	Altruism–Utilitarianism	Altruism–Deontology	Altruism–DoubleEffect	Utilitarianism–Deontology	Utilitarianism–DoubleEffect	Deontology–DoubleEffect
N_{intend}^{avg}	–	−.49 [−.68, −.30]	−.34 [−.56, −.12]	−.42 [−.62, −.23]	−.27 [−.50, −.06]	–
N_{side}^{avg}	–	−.32 [−.45, −.19]	–	−.28 [−.42, −.16]	–	.32 [.19, .47]
C_{intend}^{avg}	−.19 [−.29, −.09]	−.05 [−.10, −.01]	−.12 [−.22, −.03]	.14 [.04, .25]	–	–
C_{side}^{avg}	−.24 [−.36, −.12]	−.16 [−.28, −.04]	−.22 [−.37, −.07]	–	–	–
Con	–	–	–	.28 [.04, .49]	–	–
Res	–	–	–	.32 [.09, .57]	–	−.25 [−.52, −.02]
S	–	–	–	–	–	–
I	–	−.16 [−.27, −.06]	–	−.23 [−.35, −.11]	–	–

Note: The posterior median of the contrast effect is reported along with the 95% HDI in square brackets. Only contrasts effects for which the 95% HDI does not include zero are reported. The contrasts were calculated based on the differences in 10,000 random posterior draws. 'avg' = average of the parameters.

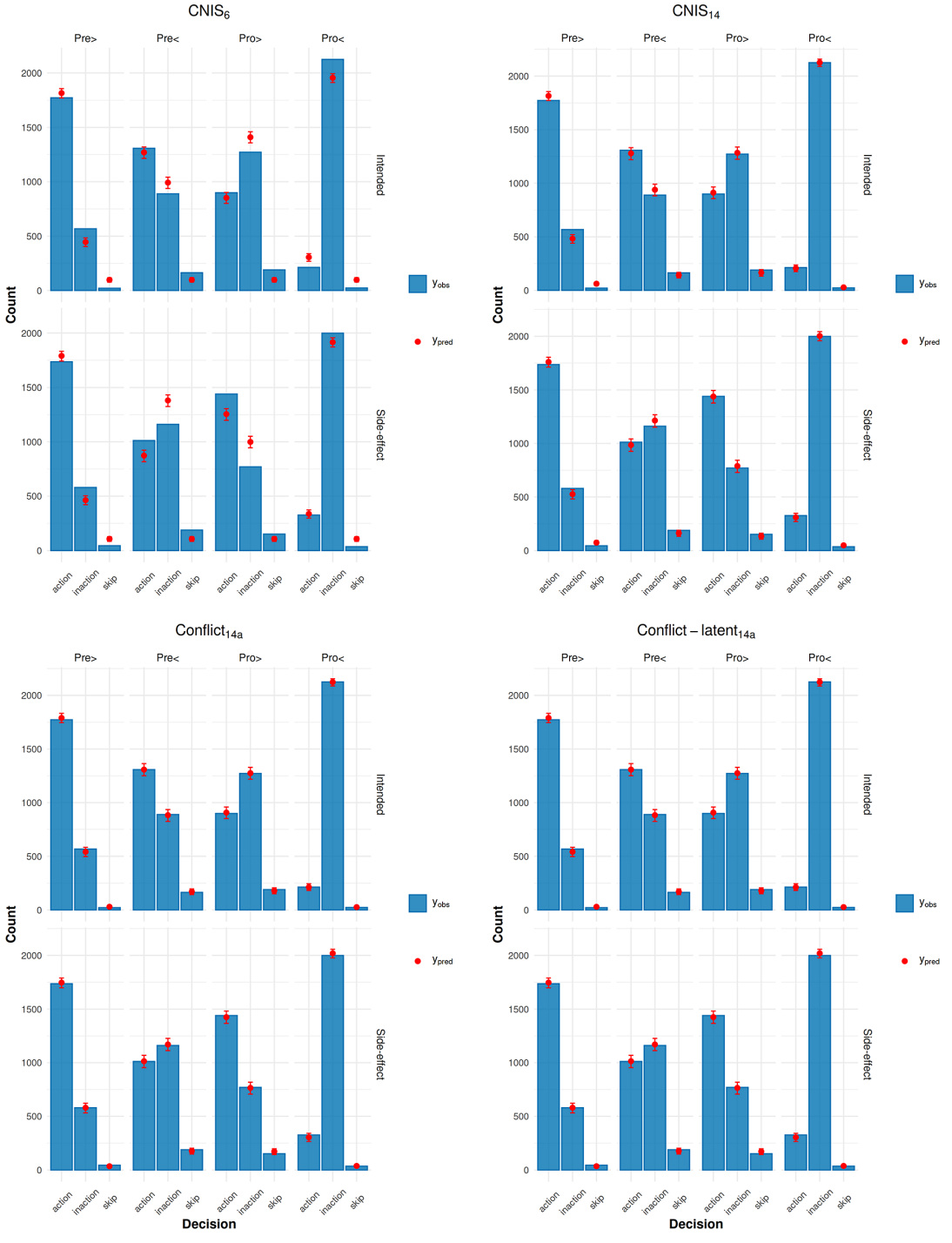


Figure 6. *Posterior predictive predictions.*

Note: Observed response frequencies (blue bars) and posterior predictive frequencies (red points) for 4 models: $CNIS_6$ and $CNIS_{14}$ (top row), CNIS-$Conflict_{14a}$ and CNIS-Conflict-$latent_{14a}$ (bottom row). Columns correspond to the 4 CNI scenarios: Pre<, Pre>, Pro>, Pro<, where > indicates that the benefit of the sacrifice is greater than the costs. Columns within panels show response type (action, inaction, skip), and panels are split by causal structure (intended means vs. foreseeable side-effect).

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

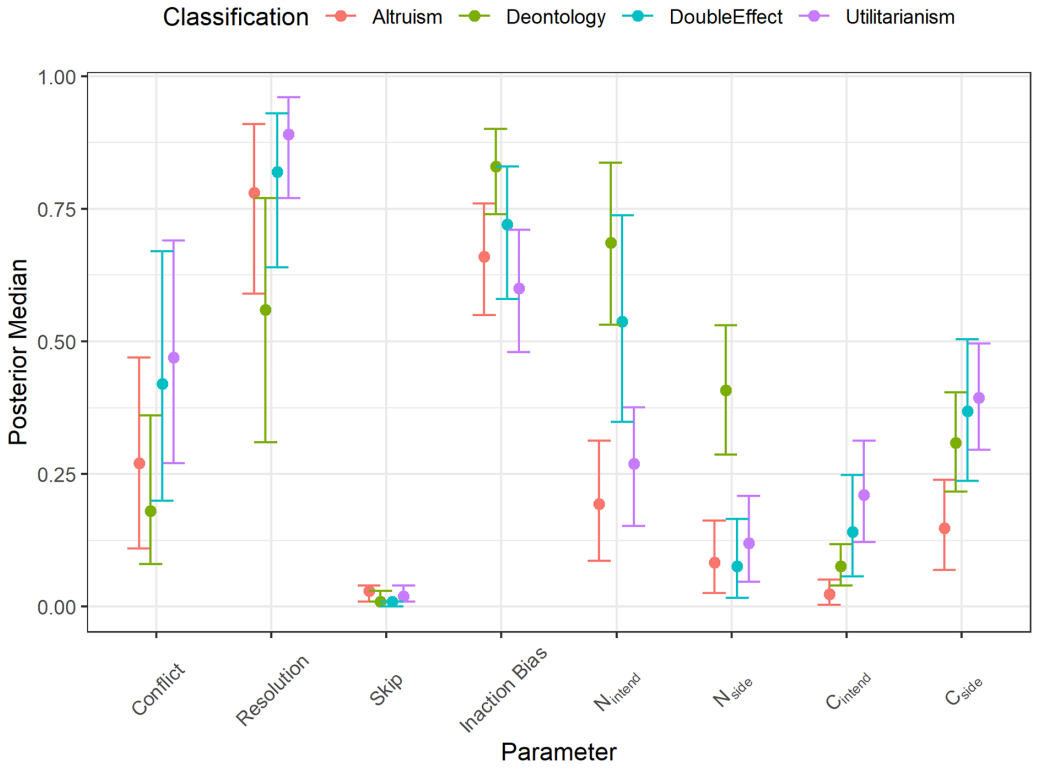


Figure 7. *Model parameters by latent classes.*

Note: Posterior median estimates with 95% HDI error bars for the parameter estimates of CNIS-Conflict-latent$_{14a}$ by latent class. *Conflict* indicates detection of response conflict in incongruent dilemmas; *Resolution* indicates successful conflict resolution toward action or inaction; *Skip* indicates a guessing response that is not sensitive to the presence/absence of detected conflicts. $N_{intend} = N^{avg}_{intend}$. $N_{side} = N^{avg}_{side}$. $C_{intend} = C^{avg}_{intend}$. $C_{side} = C^{avg}_{side}$.

Table 11. *Contrasts in the CNIS-Conflict-latent$_{14a}$ parameters.*

Contrast	Altruism	Utilitarianism	Deontology	Double effect
		Intended means vs. foreseeable side-effect		
$C^{avg}_{intend} - C^{avg}_{side}$	−.12 [−.21, −.04]	−.18 [−.29, −.08]	−.23 [−.33 −.13]	−.22 [−.36, −.08]
$N^{avg}_{intend} - N^{avg}_{side}$	–	.14 [.04, .24]	.28 [.10, .45]	.45 [.27, .65]

Note: The contrasts were calculated based on the differences in 10,000 random posterior draws of the mean parameters. In each case, the posterior median of the contrast effect is reported along with the 95% HDI in square brackets. Only contrasts effects where the 95% HDI crosses zero are reported. 'avg' = average of the parameters.

In addition to these effects in the differences of the C and N parameters across the latent classes, Table 10 also shows that the latent class of Deontology had a higher inaction bias than Utilitarianism and Altruism and that strong differences in the Conflict parameter emerged between Utilitarianism and Deontology. Table 11 reports effects of causal structure (intended means vs. foreseeable side-effect).

To simplify the interpretation of model parameters, Table 11 presents the averaged C and N parameters. In Appendix B, we show, however, that the data replicated the pattern reported in Skovgaard-Olsen and Klauer (2024) of showing strong violations of the invariance assumption in the individual C and N parameters, in violation of the CNI model.

As Table 11 shows, the manipulation of the intended-means and foreseeable side-effect in the CNI scenarios influenced the estimated C and N parameters within all latent classes albeit to

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

different extents. This stands in tension with the fact that only the latent class of DoubleEffect explicitly committed to a difference between intended means and foreseeable side-effect in their scorekeeping behavior (Figure 4). In the general discussion, we will return to this issue and discuss the repercussions of this finding for the psychological interpretation of the Doctrine of Double Effect.

As a distinguishing feature, the Conflict model has a conflict detection parameter (Con) for the case that both the latent states represented by the C and N parameters are activated at the same time. As a first step in validating whether this parameter indeed captures experienced conflict, we tested whether the estimated Con parameters could predict participants' self-criticism in the Scorekeeping Task. Self-criticism in the Scorekeeping Task occurs when participants accept the criticism of the position that they endorse in their HIT approvals. We fitted a Bayesian linear regression model with the median conflict parameter as predictor and self-criticism (the median posterior probability of accepting criticism of a given position while endorsing it via a HIT approval) as the outcome variable. It was found that conflict detection credibly predicts self-criticism, b = 1.0, 95% HDI [.20, 1.77].[13]

2.4.5. Gender distribution

A total of 787 participants were classified according to these 4 latent classes, but the analysis concentrated on the 769 participants who self-identified as either male or female. Past research with the CNI model and the PD models has shown that women tend to be more deontological than men and that there are no differences between women and men in their utilitarian tendencies (Conway and Gawronski, 2013; Gawronski et al., 2017). To investigate the distributions of gender within the latent classes, a Bayesian logistic regression model was fitted with self-reported gender as dependent variable (with '0' = male; '1' = female) and the latent classes as predictor. The median proportions of females within the latent classes matched previous results.[14] For the Altruism and DoubleEffect classes, an equal proportion of males and females remained within the 95% highest density interval.[15] To investigate differences in whether males and females resolve conflicts between Utilitarianism and Deontology in incongruent items, gender was used a predictor of the estimated Res parameter of the Conflict model. It was found that male participants had a higher posterior probability of resolving conflicts than female participants, $\tilde{\Delta}^{Res}_{Male-Female} = .05$, 95% HDI [.03, .06].

2.4.6. Structural equation modeling

To test $\mathbf{H_1}$ (concern for the greater good), $\mathbf{H_2}$ (restricted concern for the greater good), and $\mathbf{H_3}$ (antisocial motivation), we finally used structural equation modeling with the R-package blavaan (Merkle and Rosseel, 2018) to analyze relationships between altruistic and egoistic covariates and the model parameters. We included the 4 latent classes to investigate which of the 3 motivations best characterize the behavior of participants classified by these latent classes.

Structural equation modeling (SEM) is a generalization of regression models. Some of its benefits are that direct and indirect effects of explanatory variables can be estimated and that conditional independence constraints from a causal model can be incorporated (Kline, 2016; Shipley, 2016). Figure 8 illustrates the relationships between the covariates and the model parameters of the Conflict model across latent classes.

This multi-group structural equation model investigates whether relationships between altruistic and egoistic covariates and model parameters differ across the independently established latent classes. Each participant has individual-level parameter estimates (C, N, Con, Res, I) from the Conflict model,

[13] Since Altruism is not opinionated towards either Deontology or Utilitarianism (Figure 4), and the Con parameter measures conflict between the latter two within the CNI scenarios, Altruism was excluded from this analysis.

[14] $\tilde{p}_{Deontology} = .66$, 95% HDI [.62, .71], $\tilde{p}_{Utilitarianism} = .51$, 95% HDI [.44, .59], $\tilde{\Delta}_{Deontology-Utilitarianism} = .15$, 95% HDI [.06, .24].

[15] $\tilde{p}_{Altruism} = .44$, 95% HDI [.34, .53], $\tilde{p}_{DoubleEffect} = .51$, 95% HDI [.44, .59].

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

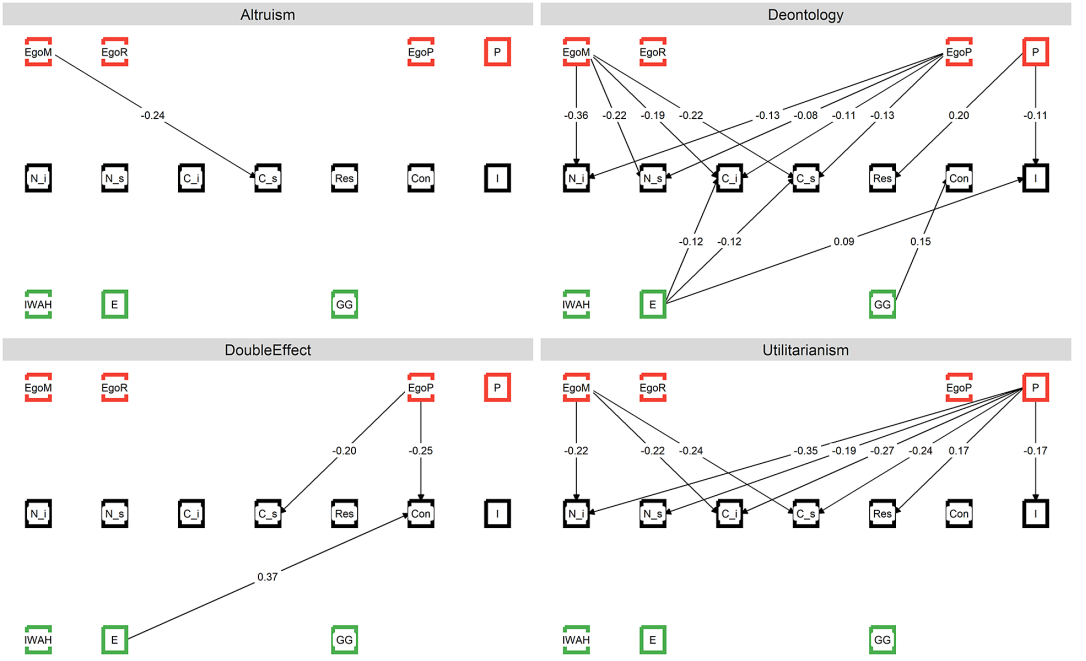


Figure 8. *Multigroup SEM analysis of altruistic and egoistic covariates.*

Note: Multi-group SEM analysis estimating how altruistic covariates (in green) and egoistic covariates (in red) affect the MPT parameters within each latent class (Altruism, Deontology, DoubleEffect, Utilitarianism). Only path coefficients with the 95% HDI not containing zero are shown. The SEM model included estimated variances and covariances which are not included in this graph due to lack of space. For convergence reasons, only empathy and primary psychopathy were permitted to affect the I (Inaction) parameter. Covariates: P = primary psychopathy; IWAH = identification with all of humanity; E = empathy; EgoM/EgoP/EgoR = moral, psychological, and rational egoism; GG = greater good scenarios. MPT Parameters: N_i/N_s = averaged norm-based response for intended means and side-effect; C_i/C_s = averaged consequence-based response for intended means and side-effect; I = general inaction bias; Con = conflict detection; Res = conflict resolution.

representing their response tendencies across different moral dilemmas. Using latent class membership as a grouping variable, the multi-group SEM estimates how altruistic and egoistic covariates predict these Conflict model parameters within each of the 4 classes (Altruism, Deontology, DoubleEffect, and Utilitarianism). This approach thus tests whether the covariates operate differently depending on the assigned latent class; a question that requires comparing patterns across classes rather than pooling them. The paths shown represent credible effects (95% HDI excluding zero) between covariates and parameters conditional on class membership.

As shown in Figure 8, moral egoism tended to be negatively associated with C and N parameters across 3 (Altruism, Deontology, and Utilitarianism) of the 4 latent classes. In the Deontology class, psychological egoism was also found to be negatively associated with the averaged C and N parameters. In the latent class of DoubleEffect class, psychological egoism was found to be negatively associated with C_{side}. In the Utilitarianism class, primary psychopathy was found to be negatively associated with the averaged C and N parameters. In the Deontology and Utilitarianism classes, primary psychopathy was negatively associated with the I parameter. Moreover, primary psychopathy tended to be positively associated with the Res parameter for Deontology and Utilitarianism.

Turning to the Altruistic covariates, in most cases it was found that there were no credible relationships with the model parameters, when controlling for the other variables. Exceptions were negative associations between empathy and the averaged C parameters for Deontology, a positive association between Empathy and the Con parameter for DoubleEffect, positive associations between GreaterGood (GG) and the Con parameter as well as between Empathy and inaction bias (I) in the class of Deontology. We discuss the import of these patterns for $\mathbf{H_1}$ to $\mathbf{H_3}$ below.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

2.5. Discussion

2.5.1. Model comparisons

We found (Table 9) that a model that permits violations of the invariance assumption for the C and N parameter ($CNIS_{14}$) outperforms a version of the CNI model without this feature ($CNIS_6$). Moreover, models that had the conflict architecture while accounting for invariance violations outperformed both latter models.

Based on participants' performance on the Scorekeeping Task, it was possible to identify 4 latent classes. Of these, the class of deontologists was by far the largest (N = 426, 54.1%). Participants' latent class memberships were then incorporated into a version of the Conflict model, such that separate group-level latent parameter means were estimated for each class. This class-informed version of the Conflict model outperformed models that did not incorporate latent class information (see Table 9).

2.5.2. Construct validity of the latent classes

The group-specific latent means of the best-fitting model had parameter estimates that conformed qualitatively to the interpretation of the latent classes (see Figure 7). Thus, participants adhering to Utilitarianism in the Scorekeeping Task were among those with the highest C parameters and lowest N parameters. Conversely, participants adhering to Deontology had the highest N parameters and were among those with the lowest C parameters. Furthermore, participants adhering to the Doctrine of Double Effect had the strongest effect of the manipulation of causal structure (intended means vs. foreseeable side-effect) corresponding to the pattern $N_{intend} > N_{side}$ and $C_{intend} < C_{side}$. In contrast, participants adhering to Altruism were among the ones with the lowest estimates for both the C and N parameter and neither showed a preference for Utilitarianism nor Deontology in the Scorekeeping Task (Figure 4). However, these participants stood out in judging the altruistic behavior in the greater good scenarios to be morally obligatory, and they had the highest ratings of finding the self-interested actions morally wrong across greater good scenarios (Figure 5). Moreover, they were the only participants who had a higher posterior probability of sacrificing themselves over the other choices in a self-sacrifice version of the Switch Trolley dilemma (Appendix B). Since we chose to couple the causal structure manipulation to a difference in outcome certainty (whether the sacrifice 'will occur' or 'is highly likely to occur'), we will consider the alternative hypothesis whether the resulting difference in expected values could account for the latent class of DoubleEffect in the General Discussion.

2.5.3. Primary psychopathy and CNI parameters

Studies that have reported positive associations between Utilitarianism and primary psychopathy (Bartels and Pizarro, 2011; Marshall et al., 2018) have relied on the conventional analysis of sacrificial dilemma. Compared to the CNI scenarios, this conventional analysis only investigates the ProGreater condition, in which norm-based and consequence-based responses are directly pitted against one another. The self-sacrifice scenario (Appendix B) replicates this effect that antisocial traits like primary psychopathy and rational egoism are associated with a higher probability of sacrificing an innocent person, whereas the prosocial traits like empathy and IWAH are positively associated with sacrificing oneself. For this condition only, the latent class of Utilitarianism also had the highest posterior probability of sacrificing an innocent person. However, the predictions for this particular incongruent condition are not diagnostic for the full predictions of Utilitarianism according to the Conflict model. Indeed, Figure 8 shows that the C parameters of the Conflict model are negatively associated with primary psychopathy and moral egoism in the latent class of Utilitarianism.

In addition, it was found that for 2 of the 4 latent classes there were credible effects of a negative association between primary psychopathy and the I parameter of the Conflict model. That is, across the CNI conditions, participants within these classes who are pronounced in subclinical psychopathy show a bias to selecting 'action' irrespectively of whether the benefits outweigh the costs, proscriptive or prescriptive norms are being tested, or whether the sacrifice is an intended means or merely a

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

foreseeable side-effect. Possibly, the positive associations between the Res parameter, representing conflict resolution, and primary psychopathy within the latent classes of Deontology and Utilitarianism are also to be interpreted in this light. If a conflict is detected by members of these classes, participants more pronounced in primary psychopathy tend to find it easier to resolve it.

2.5.4. Underlying moral motivations

Using this traditional approach of only contrasting Utilitarian and Deontological responses in the ProGreater condition, Kahane et al. (2015) found positive, bivariate correlations of utilitarian responses with primary psychopathy, rational egoism, psychological egoism, and negative associations with empathy and IWAH. On this basis, Kahane et al. (2015) conclude that acceptance of instrumental harm in sacrificial dilemma may be an expression of a calculating, selfish mindset prone to transgress conventional standards of morality ($\mathbf{H_3}$) rather than an impartial concern with the greater good ($\mathbf{H_1}$).

As a third alternative, the discussion between Kahane et al. (2015) and Conway et al. (2018a) introduces the possibility that acceptance of instrumental harm reflects a moral disposition to minimize harm in the interest of the greater good, but one that is more restricted in scope ($\mathbf{H_2}$). According to $\mathbf{H_2}$, participants may be morally motivated to select the outcomes that lead to the least suffering, but will not necessarily act against their own interest to make altruistic sacrifices that promote the greater good for strangers in distant countries.

Of the 3 hypotheses, $\mathbf{H_2}$ (Restricted Concern for the Greater Good) is best supported by our results based on structural equation analysis of partial correlations with the Conflict model parameters. It was thus found that moral egoism and primary psychopathy are negatively correlated with the C and N for participants adhering to Utilitarianism (against $\mathbf{H_3}$: antisocial motivation). Similarly, it was found that both moral and psychological egoism were negatively correlated with both the C and N parameters for participants adhering to Deontology, which made up the largest group of participants. Furthermore, while participants adhering to Altruism have the highest probability of sacrificing themselves in the self-sacrifice dilemma (Appendix B), tend to find the altruistic behavior in the greater good scenarios obligatory, and have the highest scores on the greater good items, participants adhering to Utilitarianism did not exhibit this pattern (against $\mathbf{H_1}$).

Finally, empathic concern was negatively correlated with the averaged C_{intend} and C_{side}[16] in the latent class of Deontology. This latter negative association could be interpreted in light of $\mathbf{H_3}$ (antisocial motivation). Yet, in view of the negative associations between the C parameters and both moral egoism and primary psychopathy, another interpretation more favorable to $\mathbf{H_2}$ (restricted concern for the greater good) presents itself. To be able to sacrifice innocent people to achieve the overall best outcomes, reduced empathy may be needed, for as Choe and Min (2011, p. 587) point out: 'Once people feel themselves to be in the victim's shoes, they are not able to sacrifice an innocent person no matter how great a profit is obtained'. We may thus conclude that the C parameters of the Conflict model do not measure a reduced aversion toward taking the lives of innocent people due to psychopathic traits (Bartels and Pizarro, 2011) or egoistic tendencies. But they are still a measure of approval of instrumental harm in conditions where the benefits are greater than the costs of the sacrifice. As such, the C parameters may be decreased by emphatic concern with the victim.

3. Experiment 2: Validation study

Having thus used the Conflict model to establish individual moral profiles of participants, we now turn to 2 validation studies which sought to investigate the intended psychological interpretation of its model parameters. We preregistered Experiment 2 on the Open Science Framework.[17] It had 3 main goals. The first goal was to selectively influence the conflict detection/resolution path of the Conflict model Con × (1−Res) by introducing a between-participants manipulation of the type of instruction (Anchor 10%

[16] I.e., the separate C parameters estimated for the intended means and foreseeable side-effect conditions, respectively.
[17] The preregistration can be found under: https://osf.io/vjnk4.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

vs. Anchor 25%) targeting the skip option. This then gives rise to the following hypothesis: (**H_4**) the contrast between the 2 instructions for the skip option (Anchor 10% vs. Anchor 25%) affects the Con × (1−Res) processing path in the Conflict model (see Figure 3). In contrast, the S parameter represents a guessing response in which participants select skip without being conflicted. The anchor manipulation was designed so as not to affect the S parameter. To this end, participants were instructed to use the skip option whenever they were undecided because several arguments for or against the action came to mind.

In Skovgaard-Olsen and Klauer (2024), a related anchor manipulation in the CNIS model was applied. Here, we modified the instructions to target more specifically the situation in which participants are undecided and consider using the skip option because several arguments for or against the action come to mind. The rationale for this change is that unlike the CNIS model in Skovgaard-Olsen and Klauer (2024), the Conflict model predicts an asymmetry in the frequency of participants' use of the skip option across the incongruent and congruent conditions. It permits that participants use the skip option for incongruent items more frequently because they detect a conflict between competing moral perspectives. In the Anchor 25% (/Anchor 10%) condition, participants were encouraged to use (/discouraged from using) the skip option in such cases thereby lowering (/raising) the bar for expressing experienced, but unresolved conflict. It was predicted that (**H_5**) higher resolution parameters would be found in the Anchor 10% than in the Anchor 25% condition, $\text{Res}_{\text{Anchor10\%}} > \text{Res}_{\text{Anchor25\%}}$, since we assumed that participants would be less likely to resolve the conflict between Utilitarianism and Deontology in the Anchor 25% condition. In addition, pilot studies had indicated for the conflict parameter that $\text{Con}_{\text{Anchor 25\%}} > \text{Con}_{\text{Anchor 10\%}}$. On an exploratory basis, we therefore set out to investigate the hypothesis (**H_6**) that credible directed effects would be found for the Con parameter as well, although we do not predict such an effect on theoretical grounds.

The second goal was to establish construct validity of the Con parameter by correlating the Con parameter with an extraneous conflict detection index. This conflict detection index was inspired by the use of a comparison of participants' confidence ratings in Białek and De Neys (2017) between incongruent and congruent conditions, as a measure of participants' conflict detection. However, instead of asking for confidence we follow Conway et al. (2018b) in asking how torn participants felt between action and inaction for each presented scenario. To quantify the treatment effect of the contrast between conflict-present (1) vs. conflict-absent (0) for the feeling-torn dependent variable, we calculated the following conflict detection index, for each participant, *i*:

$$\text{DetectionIndex}_i = \overline{\text{feeling torn}_i}^{\text{incongruent}} - \overline{\text{feeling torn}_i}^{\text{congruent}}$$

Given that each participant was presented with multiple incongruent trials (ProGreater, PreSmaller) and congruent trials (ProSmaller, PreGreater), an average was first computed for the 'feeling-torn' dependent variable in congruent as well as in incongruent trials, before calculating their difference. It was predicted (**H_7**) that the Con parameter would be positively correlated with this conflict detection index.

The third goal was to compare the Conflict model and a model that reduces the complexity of the processing tree of incongruent items (ProGreater, PreSmaller) by removing the Res vs. (1−Res) branches of the Conflict model and restricting the Con parameter to governing skip responses alone. In this model, the Con parameter receives the interpretation of 'expressed conflict' with conflicts invariably giving rise to skip responses. This simplified version is illustrated in Figure 9.

The motivation for the reduction is this. For the Conflict model, the Con parameter (conflict detection) is only indirectly associated with distinct behavioral outcomes via the (1−Res) path leading to additional skip responses when a detected conflict remains unresolved. While conflict detection and conflict resolution are distinct psychological processes, it is possible that a MPT model which collapses these 2 parameters by reinterpreting the Con parameter as 'expressed conflict' would perform better empirically (**H_8**).

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

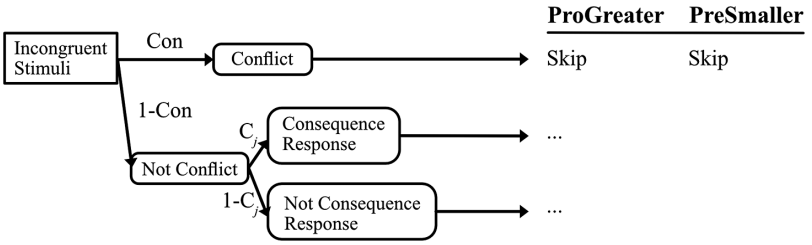


Figure 9. *Simplified Conflict model (CNIS-Conflict$_{simp}$).*
Note: The diagram depicts the part of the Conflict Model (Figure 3) that was simplified by removing the Res and C_{Res} paths and replacing it with a skip selection whenever a conflict is detected as an indication of expressed conflict.

3.1. Method

3.1.1. Participants

As per our preregistration, we aimed for 500 participants in each of the 2 between-participants conditions (explained below) after applying the following exclusion criteria: not having English as native language, completing the task in more than 2 standard deviations below or above the mean completion time, failing to answer at least one of 2 simple SAT comprehension questions correctly in a warm-up phase, and answering 'not seriously at all' to the question 'How seriously do you take your participation' at the beginning of the study. A total of 1,253 people completed the experiment over the Internet on Mechanical Turk via CloudResearch. The data analysis is based on a final sample of 1,013 participants, with 498 participants randomly assigned to the Anchor 25% condition and 515 participants in the Anchor 10% condition, in accordance with our preregistration. Applying the exclusion criteria had a minimal effect on the demographic variables. Mean age was 40.8 years, ranging from 18 to 94. 33.7 % of the participants self-identified as male, 64.7% self-identified as female, 21 participants indicated that they were non-binary and 1 participant preferred not to reveal their gender. 70.4% indicated that the highest level of education that they had completed was an undergraduate degree or higher.

3.1.2. Design

The experiment has a mixed design with the 4 CNI conditions (ProGreater, ProSmaller, PreGreater, and PreSmaller) and causal structure (intended means vs. foreseeable side-effect) varied within participants. Type of skip instruction is varied (Anchor 10% vs. Anchor 25%) as a between-participants factor.

3.1.3. Materials and procedures

The same stimulus materials were used as in Experiment 1. But unlike in Experiment 1, the 2 causal structures were displayed to the participants within one experimental session. Moreover, Experiment 2 neither included the Scorekeeping Task nor the covariates.

The crossing of the 4 CNI conditions with the two causal structures gives rise to eight versions of each dilemma. To obtain three independent replications of each CNI condition within each causal structure (i.e., three different dilemma scenarios instantiating each cell), 24 trials were presented to each participant. 12 of these trials were presented in one block with intended means and 12 were presented in one block with foreseeable side-effect contents. The order of the two blocks and the order of trials within a given block was random. The pairing of scenario (from 12 possible background scenarios that each implement the eight variations) to the randomly selected variation shown on a given trial was constrained by the following conditions: (1) participants did not see a given scenario (e.g., Vaccine) paired with the same CNI condition (e.g., ProGreater) across the two causal structures; (2) participants did not see scenarios in the same order across the 2 blocks; and (3) the last scenario in Block 1 was not the same as the first scenario shown in Block 2.

For each item, participants were presented with the three response options: action, inaction, and skip. The left-right location of the action and inaction response options on the screen varied randomly from

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

trial to trial and the location of the skip response button was fixed to the right-most position across trials (to highlight its different role from that of the action and inaction responses).

After each moral judgment, participants were asked whether they agree or disagree with the statement 'I feel torn between action and inaction' on a 7-point Likert scale from 1 ('completely disagree') to 7 ('completely agree'), following Conway et al. (2018b).

For the Anchor 10% condition, the instruction for the skip option was as follows:

> Please only use the skip option whenever you are undecided because several arguments for or against the action come to mind.
>
> It is better if you answer action or inaction than completely skip a decision! As a guideline, for a typical scenario in which people are undecided, we observe that at most 10% of the responses are skip options.

For the Anchor 25% condition the instruction for use of the skip option was as follows:

> Please only use the skip option whenever you are undecided because several arguments for or against the action come to mind.
>
> It is better if you skip a decision than mistakenly answer action or inaction! As a guideline, for a typical scenario in which people are undecided, we observe that at least 25% of the responses are skip options.

Between blocks, participants were first informed that they had reached 50% of the study and instructed that they should consider the moral scenarios of Block 2 with a fresh mind, since they may differ in important ways from the previous scenarios. The instructions concerning the skip option were then shown once more. At the end of the study, participants were asked demographic questions concerning their age, level of education, primary language, and which gender they identify with.

3.2. Results

As a manipulation check, a Bayesian logistic regression model was fitted with a skip indicator variable (skip = 1, action/inaction = 0) as outcome and the type of trials (incongruent vs. congruent), anchor condition (25 vs. 10), and their interaction as predictors, with participants as random intercepts. The median proportions of skip responses within incongruent trials for the 2 anchor conditions were $\tilde{p}_{incongruent,Anchor10} = .02$, 95% HDI [.01, .02], $\tilde{p}_{incongruent,Anchor25} = .13$, 95% HDI [.11, .15]. It was found that the median proportion of predicted skip responses was higher in the Anchor 25% condition than in the Anchor 10% condition for incongruent trials, $\tilde{\Delta}^{incongruent}_{Anchor10-Anchor25} = -.11$, 95% HDI [−.13, −.09], and much less so for congruent trials $\tilde{\Delta}^{congruent}_{Anchor10-Anchor25} = -.01$, 95% HDI [−.02, −.01].

In addition, it was found that the probability of skipping an individual trial increased for incongruent trials and the more torn a participant felt between action and inaction across both anchor conditions, based on a logistic regression that also includes feeling-torn ratings as predictors: $b^{Anchor10\%}_{FeelingTorn} = .71$, 95% HDI [.56, .85], $b^{Anchor25\%}_{FeelingTorn} = .88$, 95% HDI [.79, .96], $b^{Anchor10\%}_{Incongruent} = 1.58$, 95% HDI [.57, 2.56], $b^{Anchor25\%}_{Incongruent} = 1.30$, 95% HDI [.70, 1.88]. These simple effects were qualified by interaction effects, $b^{Anchor10\%}_{FeelingTorn:Incongruent} = -.19$, 95% HDI [−.36, −.02], $b^{Anchor25\%}_{FeelingTorn:Incongruent} = -.12$, 95% HDI [−.22, −.02]. Figure 10 shows the effects.

To test the hypothesis ($\mathbf{H_4}$) that the contrast between the 2 instructions for the skip option (Anchor 10% vs. Anchor 25%) affects the Con × (1−Res) processing path in the Conflict model, we calculated a contrast effect based on the posterior samples of Con × (1−Res) across both between-participants conditions. As expected, it was found that $\text{Con} \times (1-\text{Res})_{\text{Anchor 25\%}} > \text{Con} \times (1-\text{Res})_{\text{Anchor 10\%}}$ and that a 95% HDI of this contrast excluded zero (see Table 12).

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

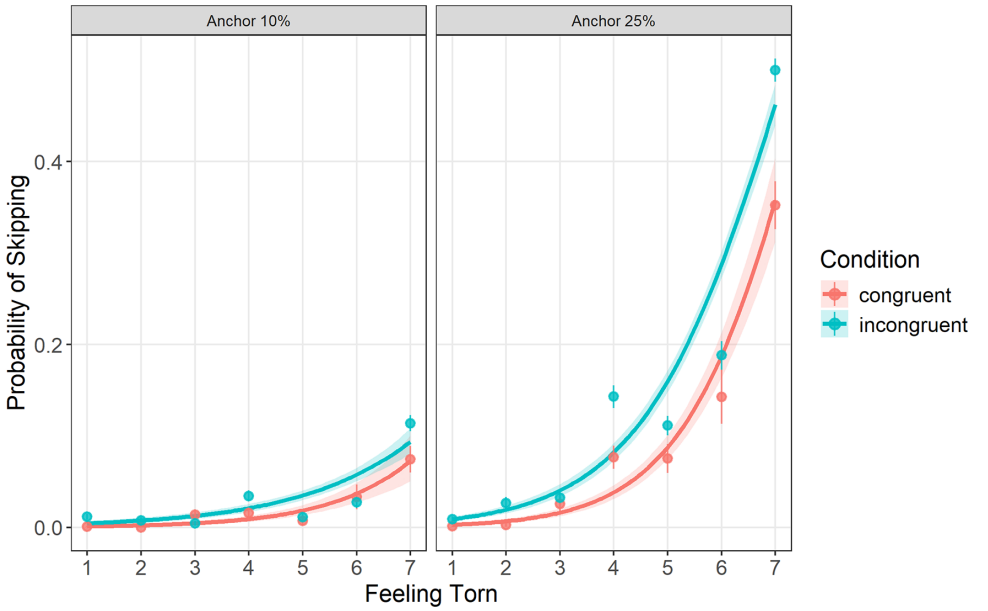


Figure 10. *Predicting skip responses in individual trials.*
Note: Probability of skipping as a function of feeling-torn ratings (1–7 scale) and trial congruency (congruent vs. incongruent) for 2 anchor conditions (10% vs. 25%). Points represent observed mean proportions of skip responses at each feeling-torn level, with error bars indicating ±1 standard error. Solid lines show predicted probabilities from Bayesian logistic regression models with 95% credible intervals (shaded regions).

Table 12. *Contrasts between Anchor 25% and Anchor 10%, CNIS-Conflict$_{14a}$.*

Contrast	$\tilde{\Delta}$	95% HDI
$\mathbf{Con} \times (1 - \mathbf{Res})^{25\%} - \mathbf{Con} \times (1 - \mathbf{Res})^{10\%}$	.14	[.11, .18]
$\mathbf{Con}^{25\%} - \mathbf{Con}^{10\%}$	.14	[−.06, .35]
$\mathbf{Res}^{25\%} - \mathbf{Res}^{10\%}$	−.28	[−.41, −.17]
$\mathbf{C}_{\mathbf{avg}}^{25\%} - \mathbf{C}_{\mathbf{avg}}^{10\%}$	−.07	[−.21, .10]
$\mathbf{N}_{\mathbf{avg}}^{25\%} - \mathbf{N}_{\mathbf{avg}}^{10\%}$	−.12	[−.35, .11]
$\mathbf{S}^{25\%} - \mathbf{S}^{10\%}$	.04	[.02, .06]
$\mathbf{I}^{25\%} - \mathbf{I}^{10\%}$	−.10	[−.26, .05]
$\mathbf{C}_{\mathbf{avg\_side}}^{25\%} - \mathbf{C}_{\mathbf{avg\_side}}^{10\%}$	.00	[−.10, .10]
$\mathbf{N}_{\mathbf{avg\_side}}^{25\%} - \mathbf{N}_{\mathbf{avg\_side}}^{10\%}$	−.13	[−.27, .01]

Note: The contrasts were calculated based on the differences in 10,000 random posterior draws of the mean model parameters of the CNIS-Conflict$_{14a}$ model in the Anchor 25% and Anchor 10% conditions.

Next, we predicted that $\text{Res}_{\text{Anchor 25\%}} < \text{Res}_{\text{Anchor 10\%}}$ and that a 95% HDI for the contrast of Res parameter across these 2 conditions would exclude zero ($\mathbf{H_5}$). As expected, it was found that participants had a higher posterior probability of resolving the conflict in the Anchor 10% than in the Anchor 25% condition and a 95% HDI of this contrast excluded zero (see Table 12). Pilot studies had indicated that $\text{Con}_{\text{Anchor 25\%}} > \text{Con}_{\text{Anchor 10\%}}$. On an exploratory basis, we next tested whether such an effect was credible ($\mathbf{H_6}$). The effect did go in the same direction as in the pilot studies and the contrast was not credible since its 95% HDI did not exclude zero (see Table 12), however. Table 12 moreover shows that

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 13. *Model comparison.*

	WAIC	LOOIC	Δelpd (SE)	p_{T1}	p_{T2}
			Anchor 25%		
CNIS-Conflict$_{14a}$	11316.91	11741.94	–	2.07e−05	.001
CNIS-Conflict$_{simp}$	11819.19	12164.25	−211.16 (24.47)	.00	.00
CNIS$_{14}$	12030.58	12401.16	−329.61 (30.27)	.00	.00
CNIS$_{6}$	13702.86	13895.08	−1076.57 (45.76)	.00	.00
			Anchor 10%		
CNIS-Conflict$_{14a}$	9394.10	9852.02	–	.01	.06
CNIS$_{14}$	9761.52	10167.38	−157.68 (22.23)	.00	.00
CNIS-Conflict$_{simp}$	9839.55	10194.64	−171.31 (20.82)	2.96e−06	.00
CNIS$_{6}$	10717.18	10958.08	−553.03 (36.09)	.00	.00

Note: LOOIC = leave-one-out cross-validation information criterion. WAIC = Watanabe-Akaike information criterion. 'elpd' = expected log predictive density is a measure of the expected out-of-sample predictive accuracy. The test statistics T_1 and T_2 represent Bayesian p values and are based on the posterior predictive model checks in Klauer (2010). Imposing the same constraint on the CNIS model as for the Conflict model on the C parameters, ($C^{Pro>}_{intend} = C^{Pre>}_{intend}$ and $C^{Pre<}_{intend} = C^{Pre<}_{side}$), also did not make the former competitive with the latter: $LOOIC_{10\%,\ CNIS12} = 10363.59$, $LOOIC_{25\%,\ CNIS12} = 12811.40$.

credible contrasts effects on neither the I parameter nor the averaged C and N parameters were found. Yet, a small effect on the S parameter did occur.

Next, we predicted that the conflict detection index below would be positively correlated with the Con parameter (**H_7**). While the Con parameter is estimated based on participants' moral judgments, this conflict detection index is an external measure that reveals whether participants notice the presence/absence of conflicts by comparing how torn they feel across conditions. Fitting a Bayesian linear regression model, it was found that the 95% HDI for the median conflict detection index excluded zero for both the Anchor 25% condition, $\overset{\sim Anchor25}{x_{DetectionIndex}} = 2.20$, 95% HDI [2.07, 2.32] and the Anchor 10% condition, $\overset{\sim Anchor10}{x_{DetectionIndex}} = 1.99$, 95% HDI [1.87, 2.12]. Participants gave higher ratings of feeling torn in incongruent than in congruent items indicating that they detected the conflict built into the incongruent items. As predicted, there was a positive correlation between the Con parameter and the conflict detection index for both conditions, $r^{Anchor25}_{DetectionIndex,Con} = .34$, 95% HDI [.26, .41], $r^{Anchor10}_{DetectionIndex,Con} = .13$, 95% HDI [.05, .22].

3.2.1. Model comparison

Next, we investigated in a model-selection contest whether the Conflict model could be simplified to the CNIS-Conflict$_{simp}$ model depicted in Figure 9 (**H_8**), in which the Res parameter, representing conflict resolution, is removed (Table 13). Based on pilot studies, we predicted that the Conflict model without this simplification would perform better in terms of the fit versus parsimony trade-off, as quantified by the WAIC and LOOIC information criteria (Vehtari et al., 2017). Going beyond the preregistration, we also sought to test whether the CNIS-Conflict$_{14a}$ outperformed the CNIS model from Skovgaard-Olsen and Klauer (2024), with 14 parameters for the intended means and side-effect conditions.

Across both conditions, the information criteria clearly preferred CNIS-Conflict$_{14a}$ over CNIS-Conflict$_{simp}$ and the 14-parameter and 6-parameter versions of the CNIS model. A visual illustration of the substantial improvement in predictive performance of the Conflict model compared to the CNIS models is presented in Appendix B.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

3.3. Discussion

Experiment 2 had 3 main goals. The first goal was to influence the conflict detection / resolution path of the Conflict model selectively. It was found that our Anchor manipulation had a credible effect on the Con × (1−Res) path of the Conflict model. Aside from a slight spill-over effect on the S parameter, Table 12 shows that the manipulation had the predicted selective influence on Res rather than Con. The second goal was to investigate the construct validity of the Con parameter by investigating correlations with a conflict detection index. This index encodes whether participants register the presence of a norm conflict between Utilitarianism and Deontology in their 'feeling-torn' judgments. It was found across the 2 anchor conditions that participants reliably detected the presence of conflict in the incongruent items and that the Con parameter of the Conflict model was positively correlated with this conflict detection index, thereby corroborating its construct validity. A final goal of Experiment 2 was to investigate whether a simplification of the Conflict model following Figure 9 was supported by the data. As Table 13 shows, the full Conflict model (Figure 3) was found to have a much better balance of fit and parsimony than both the simplified model (Figure 9) and the CNIS model of Skovgaard-Olsen and Klauer (2024).

4. Experiment 3: Response format

Experiments 1 and 2 were based on a 3-options response format, which through a skip option allows the model to capture when participants feel too conflicted to choose action or inaction. A possible objection, however, is that the original CNI model was developed for a 2-options format and might perform adequately when used in its intended setting. The purpose of Experiment 3 was therefore to directly compare these 2 response formats in a between-participants comparison to investigate the effects of including the skip option.

By design the CNI model cannot predict differences between congruent and incongruent dilemmas in participants' experience of conflict. In the CNI structure, the probability of entering a state of uncertainty is fixed by (1−C) × (1−N), regardless of whether dilemmas pit norms and consequences against each other (incongruent) or align them (congruent). In contrast, the Conflict model makes a qualitative distinction between the latent processes in congruent and incongruent trials. It thereby explicitly models conflict detection and resolution in incongruent items for which Deontology and Utilitarianism require incompatible responses. When comparing the 2 models on model-extrinsic indicators of experienced conflict (like response times, feeling torn ratings, and confidence in selected responses) their predictions diverge. The CNI model predicts a null effect: no differences between congruent and incongruent items. By contrast, the Conflict model predicts that incongruent items should show increased signs of conflict (longer RTs, feeling more torn, lower confidence in action/inaction responses) and that skip responses should be produced more frequently in incongruent conditions because participants experience greater conflict.[18] In Experiment 3, we test these predictions to find out how experienced conflicts are expressed in the 2-responses format and the 3-responses format, respectively.

[18]Some reviewers proposed that the CNI model might still allow an implicit measure of conflict detection, for example by using the product C × N, or the absolute difference |C−N| with small values as an indicator of conflict detection, as a proxy (see, e.g., Mata, 2019 for the latter suggestion applied to the PD model). Yet, the hierarchical structure of the CNI model makes such operationalizations conceptually incoherent. In the CNI model, the N-path is only reached when the C-path does *not* activate. Thus, the probability of activating the norm process is (1 − C) × N, not simply N. As a consequence, the model never allows simultaneous activation of both processes. Situations in which C and N are both high (e.g., C = N = 1) do not reflect "maximal conflict". Rather, they deterministically produce a C-response with *no* activation of the N-path. Conversely, cases in which C = N = 0 would be treated as high conflict by min(|C−N|), even though neither process activates at all. Due to these problems, neither C × N nor |C−N| can serve as an indicator of conflict in the CNI framework, which lacks a mechanism for genuine joint activation.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

4.1. Method

4.1.1. Participants

We aimed for ca. 300 participants in each of the 2 between-participants conditions (explained below) after applying the following exclusion criteria: not having English as native language, completing the entire task in more than 2 standard deviations below or above the mean completion time, finishing individual trials in zero seconds or more than 5 standard deviations of the response time of individual trials, failing to answer at least one of 2 simple SAT comprehension questions correctly in a warm-up phase, and answering 'not seriously at all' to the question 'How seriously do you take your participation' at the beginning of the study, and not showing non-respondence by responding to feeling-torn and confidence ratings in the same way in more than 80% of all trials across conditions.

A total of 813 people completed the experiment over the Internet on Mechanical Turk via CloudResearch. The data analysis is based on a final sample of 605 participants, with 308 participants in the 3 options condition and 297 participants in the 2 options condition. Applying the exclusion criteria had a minimal effect on the demographic variables. Mean age was 43.69 years, ranging from 18 to 87.[19] 37.56% of the participants self-identified as male, 60.29% self-identified as female, 9 participants indicated that they were non-binary and 4 participant preferred not to reveal their gender. 74.47% indicated that the highest level of education that they had completed was an undergraduate degree or higher.

4.1.2. Design

The experiment has a mixed design with the CNI conditions (ProGreater, ProSmaller, PreGreater, and PreSmaller) and causal structure (intended means vs. foreseeable side-effect) varied within participants. Response format is varied (2 response options [action, inaction] vs. 3 response options [action, inaction, skip]) as a between-participants factor.

4.1.3. Materials and procedures

The experiment followed the procedure of Experiment 2 with 3 exceptions. One was that response time was recorded for each individual trial as the interval from participants' click on a continue button (after having read the scenario), which revealed the response options, to the response. The second change was that after each scenario response, participants were asked to rate both how torn they felt and how confident they were in their response on a 7-point Likert scale on a second page in this fixed order. The third exception was that participants in the 3-responses condition were instructed that they could 'skip' a decision whenever they were undecided about whether the described action was morally acceptable or not, with no further guidelines about the frequency of its use.

4.2. Results

4.2.1. Three-responses format

To analyze the data, Bayesian linear regression models were used. First, we repeated the analysis from Experiment 2 showing that skip responses were predicted by feeling-torn ratings on individual trials, $b_{Skip}^{FeelingTorn} = .51$, 95% HDI [.39, .63], and higher rates of skip responses were found for incongruent trials, $b_{Skip}^{Incongruent} = 1.17$, 95% HDI [.41, 1.95]. Yet, the 95% HDI for their interaction did not exclude zero.

Second, we analyzed the 3 response options to investigate how skip responses differed from action/inaction responses in terms of (log) reaction times. It was found that there was a simple effect of inaction, $b_{Inaction} = -.08$, 95% HDI [−.12, −.03], which was qualified by credible interactions between the type of trial and the dilemma response, $b_{Inaction}^{Incongruent} = .14$, 95% HDI [.07, .20], $b_{Skip}^{Incongruent} = .23$, 95% HDI [.01, .44]. Thus, the main pattern is captured by the contrasts illustrated in Figure 11.

[19] We are ignoring one value of '3' here due to MTurk's restriction to adult participants.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

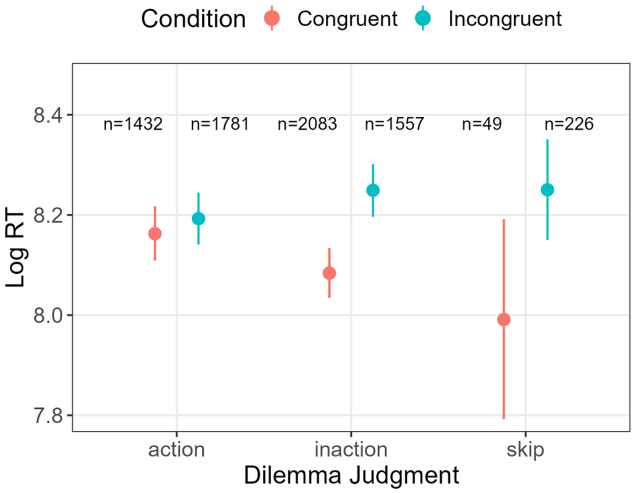


Figure 11. *Response times.*
Note: The figure shows the log reaction times as a function of the dilemma judgment and trial type. For each condition, the number of measurements within the condition is displayed.

For both inaction and skip responses, credible contrasts were found such that the response was quicker in congruent than incongruent trials, $\tilde{\Delta}^{Inaction}_{Congruent-Incongruent} = -.17$, 95% HDI [−.20, −.12], $\tilde{\Delta}^{Skip}_{Congruent-Incongruent} = -.26$, 95% HDI [−.47, −.05]. Overall, the skip responses in the congruent condition were the quickest but skip responses in both the congruent and incongruent conditions also had the widest 95% HDI and so credible contrasts with action and inaction responses were not found. Taken together, skip responses were more frequent and required more time in incongruent than congruent trials.

4.2.2. Conviction scale: Validating the psychological interpretation of skip responses

Next, a comparison was made between the different dilemma judgments and both feeling-torn and confidence ratings to assess whether the effects reflect differences in experienced conflict. Since feeling-torn and confidence ratings were negatively correlated ($r = -.73$), we treated them as complementary indicators of a single underlying conviction dimension by reverse coding the feeling-torn ratings and taking the average of these with the confidence ratings. Higher conviction values indicate greater decisional certainty.

We first analyzed the 3-response format to validate that skip responses appropriately reflect low conviction, particularly on incongruent trials in which participants experience moral conflict. Differences between response types were analyzed using ordinal regression models. One model analyzed each outcome on the conviction scale separately, and a second analyzed binned categories (FeelingTorn: conviction < 4; Neutral: conviction = 4; Confidence: conviction > 4). Since the binned categories are simpler to interpret, we focus on those here while plotting results of both models (Figure 12).

For the binned conviction outcomes, simple effects of skip relative to action, $b_{Skip} = -3.07$, 95% HDI [−3.67, −2.46] and trial type, $b_{Incongruent} = -2.54$, 95% HDI [−2.75, −2.34], were found which were qualified by credible interactions between the type of trial and the dilemma response, $b^{Incongruent}_{Inaction} = .35$, 95% HDI [.08, .61], $b^{Incongruent}_{Skip} = 1.26$, 95% HDI [.59, 1.92]. Figure 12 illustrates these interactions.

To further investigate how conviction differed across response types, we examined posterior contrasts comparing the relative strength of confidence and feeling torn within each condition. In incongruent trials, credible contrasts were found such that there was a higher posterior probability

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

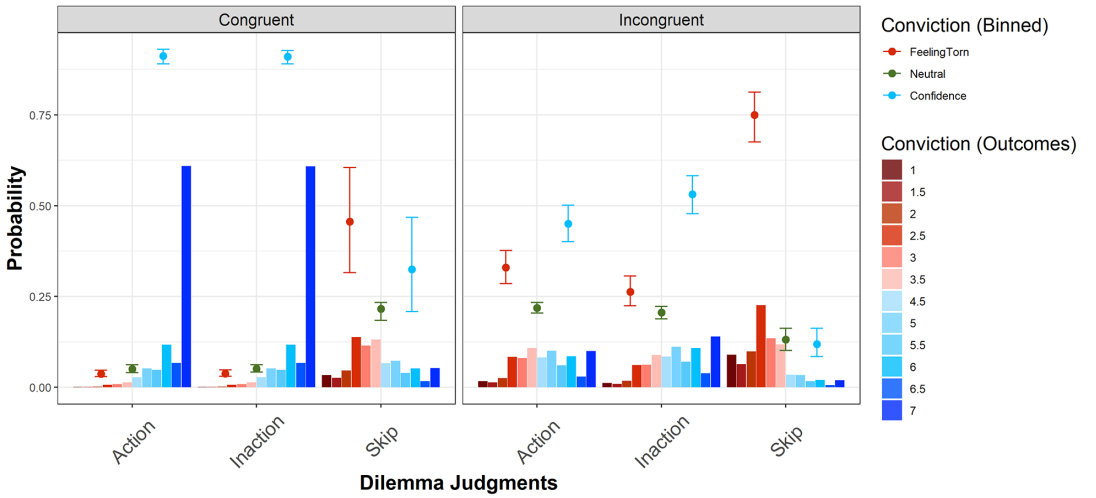


Figure 12. *Conviction scale, 3-responses format.*

Note: This figure shows participants' dilemma judgments (Action, Inaction, Skip) as a function of their conviction levels, separately for congruent and incongruent trials. Conviction was measured by combining feeling-torn and confidence ratings into a single scale: since these measures were negatively correlated ($r = -.73$), we reverse-coded feeling-torn ratings and averaged them with confidence ratings to create a unified conviction measure. The visualization displays both: (a) the probability distribution of specific conviction values for each judgment type (shown as colored histograms with the height representing probability), and (b) aggregated probabilities for 3 conviction categories (shown as dots with error bars representing 95% Bayesian credible intervals): FeelingTorn (conviction < 4), Neutral (conviction = 4), and Confidence (conviction > 4). Higher conviction values indicate greater certainty in one's judgment.

that participants' confidence outweighed how torn they felt than *vice versa* when selecting action or inaction, $\Delta^{\sim Incongruent,Action}_{Confidence-FeelingTorn} = .12$, 95% HDI [.03, .22], $\Delta^{\sim Incongruent,Inaction}_{Confidence-FeelingTorn} = .27$, 95% HDI [.18, .36]. In contrast, when selecting skip responses, there was a higher posterior probability that participants felt more torn than confident than *vice versa*, $\Delta^{\sim Incongruent,Skip}_{FeelingTorn-Confidence} = .63$, 95% HDI [.52, .73]. For congruent trials, credible contrasts were found such that there was a higher posterior probability that participants' confidence outweighed how torn they felt than *vice versa* when selecting action or inaction, $\Delta^{\sim Congruent,Action}_{Confidence-FeelingTorn} = .88$, 95% HDI [.85, .90], $\Delta^{\sim Congruent,Inaction}_{Confidence-FeelingTorn} = .87$, 95% HDI [.84, .90]. In contrast, no credible differences were found in posterior probabilities for differences in relative strength of confidence and feeling-torn for skip responses with congruent trials. Taken together, skip responses were not only more frequent and slower in incongruent trials but also reflected a substantially lower confidence and stronger experienced conflict. This confirms that the skip option works as intended: as a low-conviction response that participants select when experiencing moral conflict.

4.2.3. Effects of response formats on conviction for action and inaction judgments

Having ensured that our conviction measure worked as expected in the 3-responses format, we next probed whether the skip option affected the conviction levels associated with definitive action or inaction judgments. To address this question, we compared conviction levels between the 2-responses format (where participants could only choose action or inaction) and the 3-responses format (where skip was also available) for just action or inaction responses. This allows a direct comparison of how the presence versus absence of a skip option affects action and inaction judgments. Credible differences were not found for reaction times. Since feeling-torn and confidence ratings were negatively correlated ($r = -.71$), a conviction scale was again created as above. On this scale, differences between the 2- and 3-responses formats were analyzed using an ordinal regression model (Figure 13).

For incongruent items, differences between the 2- and 3-responses formats emerge in that (1) for action selections, there is no credible difference in the posterior probabilities of the binned feeling-torn

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

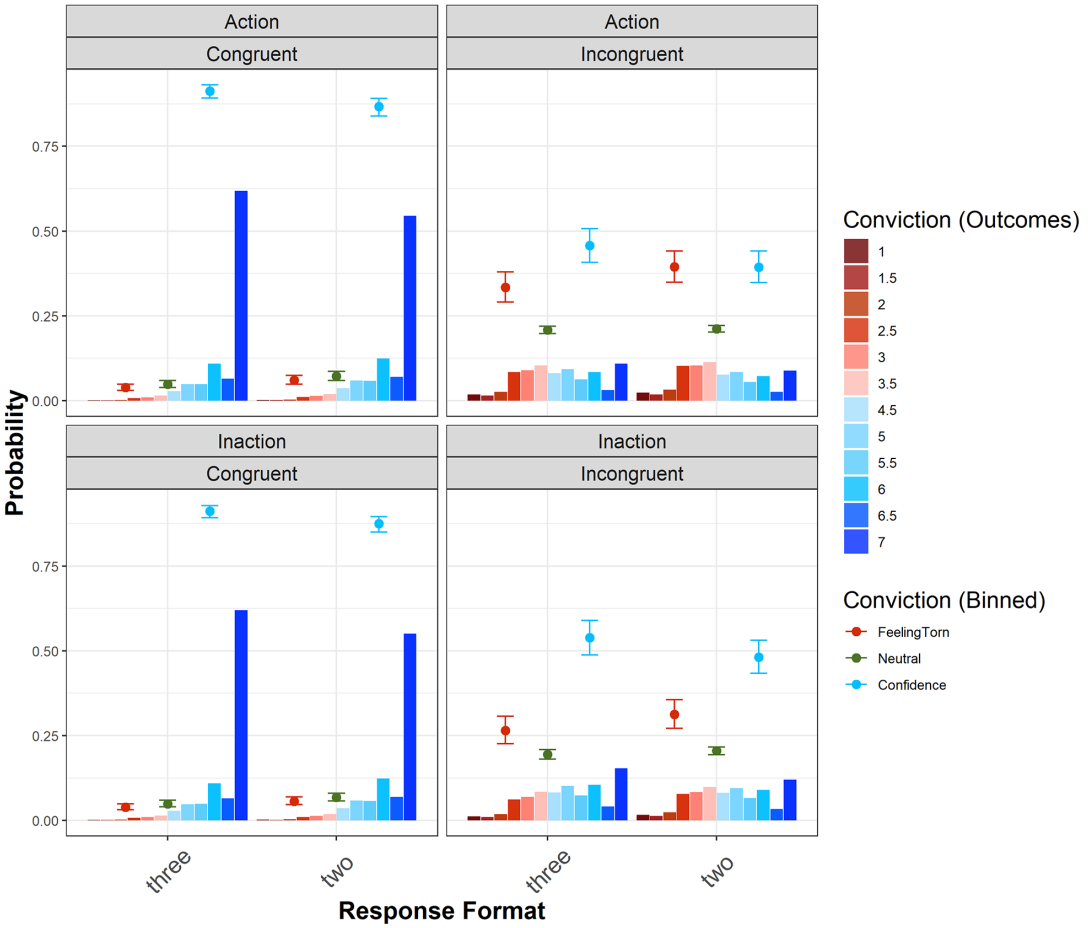


Figure 13. *Conviction. Response options.*

Note: This figure compares Action and Inaction judgments as a function of conviction levels across 2 types of response formats: a 2-responses format (left panels: Action vs. Inaction only) and a 3-responses format (right panels: Action, Inaction, or Skip). Only Action and Inaction responses are shown to directly compare the formats. Conviction was measured by reverse-coding feeling-torn ratings and averaging them with confidence ratings ($r = -.71$). The visualization shows: (a) probability distributions of specific conviction values for each judgment type (colored histograms) and (b) aggregated probabilities for 3 conviction categories (dots with 95% Bayesian credible intervals): FeelingTorn (conviction < 4), Neutral (conviction = 4), and Confidence (conviction > 4).

and confidence categories for the 2-responses format, $\Delta^{\sim Incongruent, Two}_{Confidence-FeelingTorn} = .001$, 95% HDI [−.09, .09], whereas a credible difference was found for the 3-responses format, $\Delta^{\sim Incongruent, Three}_{Confidence-FeelingTorn} = .12$, 95% HDI [.03, .22], (2) for inaction selections, a credible difference was found for both the 2- and 3-responses formats, but the difference was larger for 3 response options, $\Delta^{\sim Incongruent, Two}_{Confidence-FeelingTorn} = .17$, 95% HDI [.08, .26], $\Delta^{\sim Incongruent, Three}_{Confidence-FeelingTorn} = .27$, 95% HDI [.18, .37]. Taken together, the presence of the skip option led to increased decisional certainty in action and inaction responses relative to the 2-responses format.

4.2.4. Comparing the CNI model with the Conflict model

The Conflict model differs from the CNI model in representing the latent processes giving rise to the dilemma responses in congruent and incongruent items as qualitatively different. On the former octopus model, 8 latent processes are activated when participants are presented with incongruent items.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

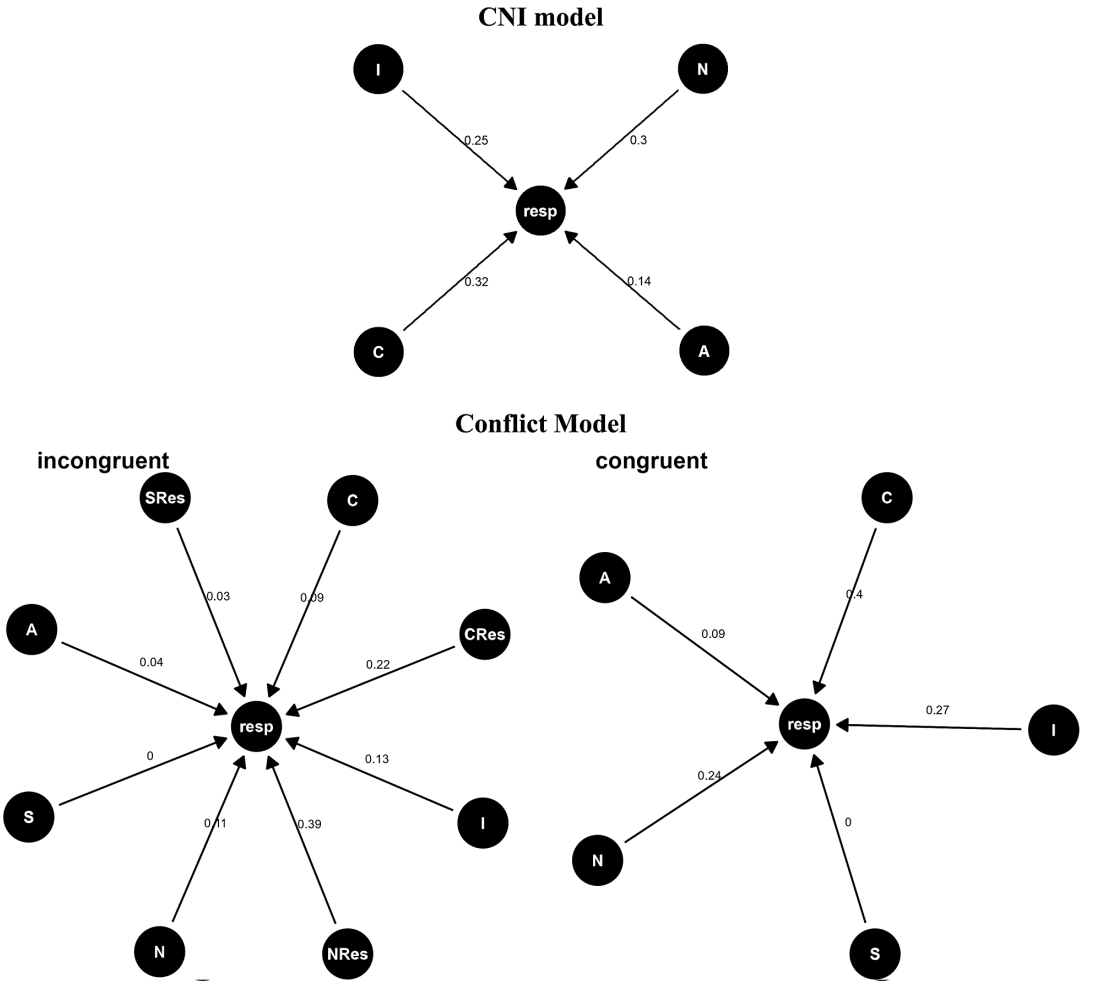


Figure 14. *Probability of path activations based on median MPT estimates.*

Note: The figure compares the cognitive process paths of 2 models for moral dilemma judgments. The CNI model (top) assumes the same 4 processes operate for all trials. The Conflict model (bottom) assumes different processes for congruent trials (where norms and consequences align) versus incongruent trials (where they conflict). Each node labels a path in the model. A 'path' refers to a latent process that outputs the observed response following stimulus presentation; internally, this process may be determined by a sequence of steps. In this figure, each path is represented by a single arrow from a latent-process node to the response node, with numbers indicating the probability that this path is activated. Probabilities are calculated based on the median estimates of the model parameters of the CNI model (top row) and the Conflict model (second row), aggregating over both the intended means and side-effect conditions. 'I' = activation of the inaction-bias path. 'A' = activation of the action-bias path. 'NRes' = activation of the path of resolving a detected conflict in a norm-based way. 'CRes' = activation of the path of resolving a detected conflict in the consequence-based way. 'SRes' = activation of the path of skipping a detected conflict.

Figure 14 illustrates the structural differences between these models using path diagrams. In these diagrams, each 'path' represents a latent cognitive process that can generate the observed response. The numbered arrows indicate the probability that each path is activated based on median model estimates.

The CNI model (Figure 14, top) proposes that all dilemma judgments result from activations of the consequence-based response, the norm-based response, or a response bias toward action or inaction. The Conflict model (Figure 14, bottom) proposes that incongruent items (where norms and consequences conflict) activate additional conflict-resolution processes not present in congruent items: (a) C_{Res}: Resolving conflict by prioritizing consequences, (b) N_{Res}: Resolving conflict by prioritizing norms, (c) S_{Res}: Resolving conflict by skipping.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

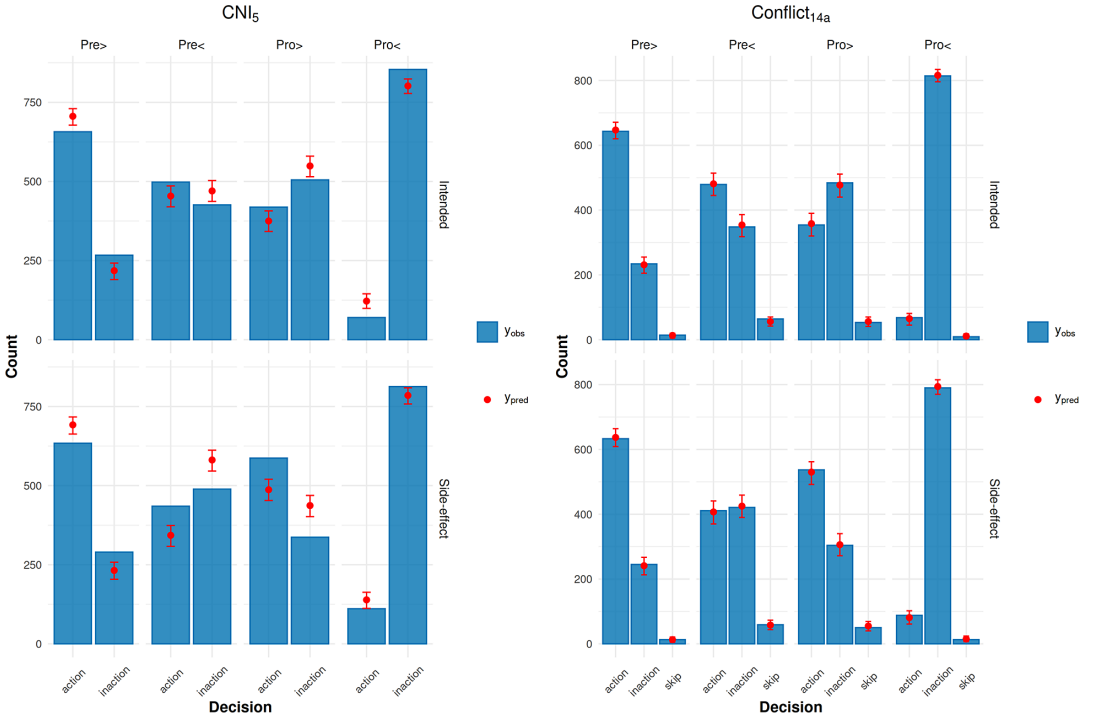


Figure 15. *Posterior predictive performance of the 2 models.*
Note: Observed response frequencies (blue bars) and posterior predictive frequencies (red points) for 2 models: CNI_5 (left) and CNIS-$Conflict_{14a}$ (right). Columns correspond to the 4 CNI scenarios: Pre<, Pre>, Pro>, Pro<, where > indicates that the benefit of the sacrifice is greater than the costs. Columns within panels show response type (action, inaction, skip), and panels are split by causal structure (Intended means vs. Foreseeable side-effect).

To compare the 2- and 3-responses formats, the proportion of C/N activation, conditional on action/inaction responses was computed. The complement of this proportion is the activation of the A or I path representing response biases, and so the proportion of C/N activation indicates the percentage of morally interpretable action/inaction responses.

It was found that there was a credible difference, such that a higher proportion of CN-activation was found for the Conflict model ($\tilde{x}$ = .70) than the CNI model ($\tilde{x}$ = .62), $\tilde{\Delta}^{C/N\ activation}_{CNI-CNISC}$ = −.08, 95% HDI [−.13, −.03]. That is, fewer action/inaction responses were attributed to response biases based on the Conflict model than the CNI model.

Since the CNI and Conflict models were fitted to different parts of the data (the 2- vs. the 3-responses format), they cannot be directly compared on information criteria like WAIC or LOOIC. Instead, their ability to fit their respective data sets can be compared by the T_1 and T_2 posterior model checks proposed in Klauer (2010). T_1 measures whether the models succeed in capturing the mean observed frequencies of dilemma judgments. T_2 measures whether the models capture the variability (variances and covariances) among the observed response frequencies. A small (Bayesian) *p* value for each indicates that the model fails to capture an aspect of the data, because this aspect of the actual observations is unlikely to be predicted by the model. It was found that $p_{T1} = p_{T2} = 0$ for the CNI model fitted to the 2-responses format and that $p_{T1} = .52$ and $p_{T2} = .02$ for the Conflict model fitted to the 3-responses format. Figure 15 illustrates the posterior predictive performance of each model.

As Figure 15 shows, the Conflict model is better able to fit the 3-responses format than the CNI model is able to fit the 2-responses format. In this comparison, the CNI model was evaluated under favorable conditions: it was tested in its original 2-responses format and allowed to estimate separate C

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

and N parameters for the intended-means and foreseeable side-effect cases. Nevertheless, the invariance assumption that fixes C and N across the 4 CNI conditions (ProGreater, ProSmaller, PreGreater, PreSmaller) still produces substantial misfits in the mean action and inaction response frequencies, indicating that one source of its problems lies in this invariance constraint.

4.3. Discussion

What happens when moral decision-makers experience conflict? In the CNI model, 2 latent processes may be activated: a consequence-driven process (with probability C) and a norm-driven process (with probability N). But because the model is structured hierarchically, the consequence-driven process always takes priority: if C activates, norms cannot influence the response. As a result, the model contains no mechanism by which conflicts between norms and consequences can be represented, detected, or resolved. The probability of entering an 'uncertainty' state is therefore fixed at $(1 - C) \times (1 - N)$ for *both* congruent and incongruent items, and the CNI model predicts *no differences* on external indicators of conflict such as response times, feeling torn, confidence, or the frequency of abstaining from choice.

The Conflict model is different. It posits 2 *qualitatively different* sets of latent processes, depending on whether dilemmas evoke a contrast between deontological and utilitarian considerations. For incongruent items, it includes processes of conflict detection and conflict resolution. Congruent items, by contrast, recruit only a unified response on which both moral views converge or its absence. The Conflict model thereby predicts that incongruent dilemmas should elicit more signs of conflict (longer RTs, higher torn ratings, lower confidence) and greater use of the skip option. The data supported these predictions. Skip responses were markedly more frequent for incongruent dilemmas. On congruent trials, skip responses appear to be occasional fast guesses, produced more quickly than action or inaction responses (Figure 11). On incongruent trials, skip responses were slower and occurred when participants reported higher degrees of conflict. This qualitative shift is also visible in participants' introspective ratings: feeling torn dominates confidence for skip responses in incongruent dilemmas only (Figure 12). Together, these findings indicate that the skip option is selectively used when participants experience genuine conflict between competing moral considerations.

The advantage of the 3-responses format appears not only in the pattern of skip responses, but also in the quality of action and inaction judgments. Participants reported *higher* confidence in these judgments when they had the option to opt out (Figure 13), suggesting that the 2-responses format forces decisions in trials in which participants would otherwise refrain because they feel conflicted.

Finally, in comparing the 2 models, the CNI model fitted to the 2-responses data failed to capture the observed mean frequencies of dilemma judgments and attributed a disproportionately large share of action/inaction responses to response-bias paths. The Conflict model, fitted to the 3-responses data, provided a substantially better account of the data (Figure 15) and attributed a higher proportion of these responses to the morally interpretable C and N processes. This pattern is what one would expect if the 2-responses format forces conflicted participants to choose action or inaction rather than expressing their conflict. Taken together, the results of Experiment 3 show that (1) congruent and incongruent dilemmas elicit systematically different levels of conflict; (2) skip responses are selectively used when participants experience that conflict; (3) the 3-responses format improves the measurement of moral decision-making by allowing participants to express conflict rather than suppressing it; and (4) the CNI model shows difficulties in adequately modeling action and inaction judgments, even in its intended 2-responses format.

5. General discussion

The CNI model (Gawronski et al., 2017) has advanced the computational modeling of moral judgments by systematically uncoupling factors that are normally confounded in traditional research on moral

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

judgment via Trolley dilemmas. Using multinomial processing trees and scenarios with 4 contrast cases, the CNI model attempts to dissociate responses according to Utilitarianism and Deontology in participants' case judgments.

5.1. Extending the CNI model

At the same time, we argued that the model would profit from further model development with regard to a couple of issues. The first concerns an invariance assumption commonly made in process-dissociations models. Because the degrees of freedom in a data set puts an upper limit to how many parameters can be estimated by the data, it is common to set MPT parameters equal across conditions. Yet, as Klauer et al. (2015) showed, this invariance assumption is often violated empirically for process-dissociation models in social and cognitive psychology. Building on these results, it was found in Skovgaard-Olsen and Klauer (2024) across 2 experiments that a CNIS model (Figure 2) that avoids the invariance assumption for the C and N parameters provides a much better fit relative to its complexity than a CNIS model which makes these assumptions. In Skovgaard-Olsen and Klauer (2024), it was furthermore shown how these invariance assumptions were implicated in recent methodological discussions over the assumptions of the CNI model (Baron and Goodwin, 2020, 2021; Gawronski et al., 2020). Accordingly, avoiding these invariance assumptions is the first desideratum motivating the Conflict model of moral judgments.

The second desideratum derives from the observation that conflicts between response tendencies are pre-empted by the CNI model through its architecture. According to the CNI model, when both the consequence-driven and the norm-driven process suggest a response and there is thus the potential for conflict, the consequent-driven process as the dominant process in the CNI model simply and invariably trumps the norm-driven process. Therefore, conflict detection and resolution have no place in the CNI model and the CNI model predicts a null effect on model-extrinsic indicators of conflict (e.g., response times, feeling torn, and confidence) across congruent and incongruent items. Yet, there is evidence that participants reliably detect conflict if and when it occurs. For example, in Experiment 2, participants felt reliably more torn for incongruent items that have the potential for conflict than in congruent items that do not have that potential. In Experiment 3, strong shifts occurred in participants' feeling-torn and confidence ratings between congruent and incongruent items such that participants confidently produced action/inaction judgments for the former and felt more torn through the detected conflict in the latter. Moreover, the asymmetry in the frequency and latency of skip responses indicates that such responses are produced by qualitatively different processes in incongruent compared to congruent items.

Similarly, in Conway and Gawronski (2013), participants found incongruent scenarios more difficult than congruent ones and required more time to respond to them. In Conway et al. (2018b), feeling-torn was measured and used in a mediation analysis for the effect of a trustful versus distrustful mindset manipulation concerning its effect on the U and D parameters of the PD model (Conway and Gawronski, 2013). Proponents of the PD model thus acknowledge verbally the issue of response conflicts. But it remains the case that neither the PD model in Conway and Gawronski (2013) nor the CNI model in Gawronski et al. (2017) implement a mechanism whereby participants can be in a state of conflict and thus neither model is able to account for the findings in Experiment 3, when considered as substantive theories of moral judgments rather than as measurement models. As a second desideratum, we thus suggest that a model for moral judgments should have a mechanism for dealing with response conflicts and cease to treat congruent and incongruent items symmetrically.

5.2. The Conflict model

Being able to satisfy both these desiderata while preserving the insights of the CNI model then motivates the Conflict model of moral judgments (Figure 3). This Conflict model retains the model structure of the CNIS model (Figure 2) for congruent items. For incongruent items, however, it permits

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

that participants can be in a state of conflict between Utilitarianism and Deontology, which they can either resolve based on the relative strengths of the consequence-driven process and the norm-driven process, or fail to resolve, which then produces a skip response (see Figure 3). In addition, the Conflict model estimates individual differences by estimating MPT parameters at the individual level and by estimating group-specific latent means in its MPT parameters based on the Scorekeeping Task (see Appendix A). As a distinguishing feature, the Conflict model assumes that qualitatively different processes operate in congruent and incongruent items due to the possibility of conflict detection and resolution in incongruent items.

To fit the Conflict model, we modified existing stimulus materials and introduced a manipulation of whether instrumental harm was an intended means or a foreseeable side-effect. It was found across all experiments that the rate of abnormal responses, which cannot be interpreted morally, was lower in our modified stimulus materials than in published CNI studies (Appendix B). In model comparisons (Tables 9 and 13 and Figure 6), it was furthermore found that this new Conflict model provides a better expected out-of-sample predictive accuracy than both the CNI model extended with a skip option ($CNIS_6$) and a version of the CNIS model without the invariance assumption for the C and N parameters ($CNIS_{14}$). The quantitative model comparisons across experiments show that the new Conflict model strongly outperformed these other models both in terms of information criteria (Vehtari et al., 2017) and posterior predictive checks (Klauer, 2010). In Appendix B and Experiment 1, visual illustrations of the predictive performance of these 3 models are presented, which show that only the new Conflict model is able to capture the mean frequencies in participants' responses across the intended means and foreseeable side-effect conditions.

In Experiment 3, a direct comparison between the 2-responses CNI model and the 3-responses Conflict model was made. Again, it was found that the Conflict model was better able to fit the mean response frequencies for both the intended means and foreseeable side-effect conditions than the CNI model (Figure 15). The evidence in Experiment 3, moreover, shows that without the skip option, participants felt less confident in their action/inaction responses than when they additionally had the third option of opting out of the moral dilemma. The detected conflict in incongruent items measured by feeling-torn and confidence ratings is an aspect of the mechanism underlying moral decision making which is not captured by the processing tree of the CNI model. The consequence of not representing this feature is that the CNI model forces participants to choose between an action and an inaction response despite feeling conflicted, which may explain why the CNI model attributed more action/inaction responses to response biases than the Conflict model in Experiment 3. We also addressed suggestions that conflict detection might be approximated indirectly within the CNI framework by algebraic combinations of its parameters (such as products like C×N or differences like |C−N|). However, because the CNI model's hierarchical structure never permits simultaneous activation of the C- and N-paths, such quantities cannot meaningfully represent conflict, reinforcing the need for an explicit mechanism as implemented in the Conflict model.

In Experiment 2, evidence of selective influence of a manipulation of skip instructions (Anchor 10% vs. Anchor 25%) was presented for the conflict detection / resolution path. In a between-participants comparison, the manipulation had credible effects on Res and Con × (1−Res) in the predicted directions, but not on the I and averaged C and N parameters. However, although the instructions targeted cases in which participants are undecided between competing arguments for or against the action, there was a slight spillover of the anchor manipulation on a difference in the S parameter as well. This might indicate that some participants experienced conflict in some of the items that are nominally congruent or that some participants generally felt more entitled to use the skip option in the Anchor 25% condition than in the Anchor 10% condition, given a higher baseline of skip uses was set.

In Experiment 2, evidence was also obtained for construct validity of the Con parameter in that it correlated positively with a conflict detection index registering whether participants noticed the presence/absence of conflicts in their feeling-torn evaluations. This evidence adds to the finding in Experiment 1 that the median Con parameters within the latent classes of Deontology, Utilitarianism,

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

and DoubleEffect were positively associated with the posterior degree of self-criticism within these groups.

5.3. Implications for dual-process theory

Experiment 1 goes beyond the dual-process theory by modeling a third latent class of participants following the Doctrine of Double Effect as a compromise between Utilitarianism and Deontology (discussed below). The Conflict model in turn goes beyond Greene's (2008, 2013) dual-process theory in modeling participants' responses as produced by 8 latent processes paths that differ between congruent and incongruent items (Figure 14).

According to Greene's (2008, 2013) dual-process theory, deontological respondents are supposed to respond automatically based on affective responses without detecting and resolving the conflict in incongruent trials. Białek and De Neys (2017) presented evidence challenging this by showing that deontological respondents also detect the conflict. Converging evidence for this is found in Experiment 1 (Figure 6), where it was shown that members of the latent class of Deontology also detected and resolved the conflict. At the same time, a credible contrast was also found such that participants in the latent class of Utilitarianism had a higher posterior probability of both detecting and resolving the conflict (Table 10), in partial support of the attribution of a deliberative response style to Utilitarianism (Patil et al., 2021).

It should be noted, however, that participants in Experiment 1 were exposed to arguments for both sides in the Scorekeeping Task, in both experimental sessions after completing the CNI items. Yet, this exposure to counterarguments, which was used as a device for measuring participants' reflective attitudes, did not shift participants toward a utilitarian response style. Instead, the majority of participants were captured by Deontology, when aggregating across items. This finding converges with the finding in Ng et al. (2023) that asking participants to think about the reasons for their judgments (as opposed to responding intuitively) can increase deontological response tendencies, against the predictions of the dual-process theory.

5.4. Participants' moral commitments

When previous studies have investigated participants' moral commitments, the focus has typically been on principles that they explicitly endorse (Robinson et al., 2015; Royzman et al., 2015), on principles that they can articulate (McHugh et al., 2020), or on principles that can be identified based on the moral reasoning extracted from their justifications (Crain, 1985; Kohlberg and Hersh, 1977).[20] With Haidt (2001) and Greene (2008, 2013), a more skeptical stance was introduced into moral psychology which holds that participants' justifications for common sense morality may to a wide extent be based on *post hoc* rationalizations and that the majority response to moral dilemma may be based on intuitive, emotional reactions rather than being caused by reasons. Yet, the empirical basis for such wide-ranging claims remains controversial (May, 2018, 2019).

As a middle course between assuming conscious application of moral principles in moral judgments and declaring moral commitments the product of *post hoc* rationalization, Lombrozo (2009) proposes that moral commitments may causally mediate moral judgments, by, for example, influencing which aspects of dilemmas are categorized as morally relevant. However, instead of relying on participants' explicitly held moral commitments, it may be worth investigating methods for implicitly eliciting participants' moral commitments to pursue this mediation hypothesis. Accordingly, instead of assuming that moral principles are consciously accessible for verbal reports, the guiding idea behind the

[20]Note that Kohlberg's approach has been challenged and supplemented by work showing that moral judgments can be assessed without requiring articulate justifications, e.g., in the social-domain tradition (Turiel, 1983) and the Defining Issues Test (Rest et al., 1999). Our focus here is, however, not on the ability to make moral–conventional distinctions but on attributing commitments to moral views.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table 14. *Taxonomy of Utilitarianism.*

Level 1	Sacrificial judgments that coincide with what Utilitarianism predicts.
Level 2	Utilitarian sacrificial judgments because participants engage in aggregate cost-benefit reasoning.
Level 3	Utilitarian sacrificial judgments because participants have a genuine moral concern for the greater good *within the context of the sacrificial dilemma*.
Level 4	Utilitarian sacrificial judgments because participants have a general impartial concern for the greater good which makes them care about maximizing the well-being of socially distant humans and animals.
Level 5	Utilitarian sacrificial judgments because participants have explicit commitments to Utilitarian principles which they can articulate.

Note: Taxonomy based on Conway et al. (2018a). Level one is included in all of the subsequent levels. What differs between levels 2–5 is the explanation for the response pattern.

Scorekeeping Task is that participants' sanctioning behavior and acceptance of criticism implicitly reveal which norms they adhere to (Skovgaard-Olsen, 2026; Skovgaard-Olsen and Cantwell, 2023; Skovgaard-Olsen et al., 2019). To implement this idea, latent classes were estimated based on participants' performance on the Scorekeeping Task and these latent classes were used in the Conflict model to estimate group-specific latent means in the MPT parameters.

As we have seen, Kahane et al. (2015, 2018) argue that acceptance of instrumental harm in sacrificial dilemma is not diagnostic of participants' acceptance of Utilitarianism understood as the moral view that strives to maximize aggregate well-being of all persons (or sentient beings) from a radical impartial perspective that gives equal weight to the interests of all. Conway et al. (2018a) reply by presenting the taxonomy of participants' adherence to Utilitarianism in Table 14.

Conway et al. (2018a) interpret Kahane et al.'s (2015) argument in favor of $\mathbf{H_3}$ (accepting instrumental harm in sacrificial dilemmas reflects antisocial tendencies) as a rejection of the claim that participants achieve level 3 or higher on the taxonomy in Table 14. In this case, the aggregate cost-benefit reasoning that participants engage in is interpreted as being the result of a calculating, selfish mind that does not aspire to impartially weigh everybody's interests to promote the greater good. Against this, Conway et al. argue that Kahane et al.'s (2015) empirical evidence in favor of $\mathbf{H_3}$ is based on their use of the conventional analysis, which contrasts action/inaction judgments only in the incongruent condition (ProGreater) where Utilitarianism and Deontology conflict. Conway et al. (2018a) revisited the results of Kahane et al. (2015) using the PD model of Conway and Gawronski (2013), a predecessor to the CNI model (Gawronski et al., 2017) without the feature of estimating inaction bias (see Table 2). However, by separating sacrifices with greater and smaller benefit, the PD model allows one to assess whether instrumental harm reflects aggregate cost-benefit reasoning. Unlike the conventional analysis, these process-dissociation methods (and with them the Conflict model) can discriminate between Levels 1 and 2.

Like Kahane et al. (2015), Conway et al. (2018a) find that participants permitting instrumental harm in sacrificial dilemma do not show strong altruistic tendencies that generalize across contexts (e.g., as operationalized by the Greater Good scenarios). Thus, like Kahane et al., they also doubt that ordinary people reach Level 4 and make, for example, the utilitarian connections between acceptances of instrumental harm in sacrificial dilemma and general views on foreign aid or animal rights. In the present Experiment 1, participants with utilitarian tendencies in sacrificial dilemma did not in general find the altruistic deeds on the Greater Good scenarios morally obligatory, which again suggests that these participants may be concerned with minimizing immediate harm, but do not have general tendencies to promote happiness among strangers at personal costs. Yet, on the other hand, Utilitarianism as assessed by the PD model does not associate positively with antisocial tendencies, and Conway et al. (2018a) thus argue that the participants do show a moral concern for the greater good, but one that is restricted to the contexts of the dilemma (Level 3). These findings are in general

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

agreement with the results of Experiment 1. In Table 9 and Figure 6, it was shown that the Conflict model outperforms the CNI model (extended with a S parameter), which in turn was introduced as a successor of the PD model (Gawronski et al., 2017), and so the analyses and findings of Experiment 1 can be seen as extending and confirming Conway et al. (2018a). As seen, the results of Experiment 1 were found to be consistent with $\mathbf{H_2}$ (restricted concern for Greater Good) and thus with the claim that participants' level of commitment amounts to Conway et al.'s (2018a) Level 3. But whereas previous findings were based on participants' immediate reactions to moral dilemmas, Experiment 1 shows that the Level 3 category of Utilitarianism persists once participants' reflective attitudes are elicited in a mini longitudinal study that confronts participants with different perspectives on the moral dilemma in an argumentative discourse.

To argue that utilitarian participants' concern with aggregate cost-benefit reasoning is not just the result of a reduced aversion to harming others, but a genuine moral concern within the context of the task, Conway et al. (2018a, study 6) include further measures relating to participants' moral identity, moral convictions about the wrongness of harm, and participants' concern with the group well-being of the innocent people saved. A further source of evidence of such a moral concern is Choe and Min's (2011) investigation into the negative emotions accompanying moral judgments in sacrificial dilemma. Of 6 different emotions, guilt was the most frequently reported emotion following sacrificial decisions.

5.5. The Doctrine of Double Effect (DDE)

Cushman (2016) argues the DDE matters for moral judgments not because it is a normative principle that participants consciously follow in their moral judgments but rather because it is a distinction that influences implicit information processing moderately, among other factors. A meta-analysis in Feltz and May (2017) based on more than 100 studies supports the idea of a real effect of the distinction between intended means and foreseeable side-effect that is however influenced by moderators such as whether the harm is brought about via use of direct personal force (see also Greene et al., 2009).

This idea of an implicit influence of the DDE fits with the findings in Experiment 1. While only a minority of participants explicitly commit themselves to the DDE in the scorekeeping task, participants committed to Deontology and Utilitarianism are still influenced by the Principle of Double Effect such that $\overline{\mathrm{C}}_{\mathrm{intend}} < \overline{\mathrm{C}}_{\mathrm{side-effect}}$ and $\overline{\mathrm{N}}_{\mathrm{intend}} > \overline{\mathrm{N}}_{\mathrm{side-effect}}$.

There is a discussion over the psychological nature of this implicit influence. Some argue that the difference between intended means and foreseeable side-effects affects causal reasoning by shifting the locus of intervention which directs attention (Waldmann and Dietrich, 2007; Waldmann et al., 2012, 2017; Wiegmann and Waldmann, 2014). Others argue that its effect on moral judgment is mediated by shifting attributions of intention (Cushman, 2016; Cushman and Young, 2011). In both cases, the distinction of intended means versus side-effect would affect moral judgments, not because the principle is built implicitly into an innate module of moral reasoning (as argued, e.g., in Mikhail, 2007) but because it affects information-processing in non-moral domains (attributions of causation or intention), which are recruited in moral judgment.

Cushman (2016) argues that DDE looks like a psychological mistake that is driven by non-moral factors that we are not consciously aware of and influence moral judgment. To evaluate the merits of this argument, it is useful to consider how cases of implicit influences on responses are assessed as possible psychological mistakes in the psychology of reasoning. Common arguments for treating an influence on behavior as a bias include showing that it conflicts with normative theory (e.g., base-rate neglect), that it potentially has bad consequences (e.g., wishful thinking, hindsight bias), or that it is an influence on judgment that we would try to avoid if we were made aware of it (e.g., the matching bias in the Wason selection task). Yet, in Cushman (2016), none of these arguments are made.

We thus find that no bad consequences of the Switch/Push asymmetry have been identified and that denials of a difference between Push and Switch when explicitly presented with contrasting pairs in Cushman et al. (2006) was visible only in 17% of the provided justifications. On these 2 counts then, the argument in favor of treating DDE as a psychological mistake fails. In contrast, its basis

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

in normative theory both has its defenders (McIntyre, 2019; Quinn, 1989; Wedgewood, 2011) and descenders (Bennett, 1966). In moral psychology, one common argument against is that participants do not consistently apply the same reasoning from the asymmetry between Switch and Footbridge cases to loop cases (Greene, 2013; Thomson, 1985). Yet, these different versions are not matched for their complexity. There is also an ongoing investigation of successor principles to DDE which cover such cases (Kamm, 2007).

For deontologists, the DDE is an optional principle, which can be used to soften behavioral restrictions to justify certain cases of instrumental harm, but it is not a principle that has been defended on Utilitarian grounds. Thus, for participants committed to Utilitarianism, the implicit influence of the Principle of Double Effect on their estimated C and N parameters could be viewed as signs of inconsistency.

5.5.1. Expected value vs. causal structure

Often the intended means versus foreseeable side-effect distinction is manipulated in terms of indirect harm versus direct use of personal force (Greene et al., 2001). In this case, the effects of instrumentality and personal force cannot be analyzed separately (Fahrenwaldt et al., 2025). We manipulated the distinction differently such that the direct form of harm does not take the form of direct use of personal force. However, because our manipulation coupled causal structure and outcome certainty (whether the sacrifice 'will occur' or 'is highly likely to occur'), an alternative hypothesis merits consideration. Perhaps the latent class of DoubleEffect is positioned between Deontology and Utilitarianism due to differences in expected values rather than causal structure.

Research on verbal probability expressions (Renooij and Witteman, 1999; see their Figure 1) indicates that 'highly likely' corresponds to approximately 90–95% probability. Under this interpretation, the expected values would be: intended means EV = +4, foreseeable side-effect EV = +4.05 to +4.10, doing nothing EV = −5. For participants solely focused on maximizing expected values, the small difference of 0.05–0.10 between the intended means and foreseeable side-effect conditions would not justify different moral judgments. The Utilitarian latent class accepts the sacrifice in both conditions, and the Deontology latent class rejects it in both conditions; neither pattern is explained by this small, expected value difference. These 2 classes already make up most of our participants. Could it then be that the behavior of the smaller DoubleEffect class (who switch their judgments between conditions) is accounted for by this expected value difference? If so, we would have to assume that the large difference in expected values in the intended means condition between accepting the sacrifice (+4) and doing nothing (−5) is insufficient to motivate acceptance but that adding 0.05–0.10 crosses a critical threshold in the side-effects condition. Such a non-linear response to expected values would be surprising.

We find it more plausible that deontological considerations relating to the difference in causal structure and concerns about agent responsibility account for this pattern. In the intended means condition, a deontological prohibition against using the death of the victim instrumentally to achieve a goal discourages the sacrifice. In the side-effects condition, attention is drawn to the fact that there are causal factors affecting the victim's fate outside the agent's control which reduce their responsibility. We manipulated the 2 dimensions of causal structure and outcome certainty together because classic intended means cases (active euthanasia, terror bombing) involve certain harm, while classic side-effect cases (passive euthanasia, tactical bombing with collateral damage) involve probable but uncertain harm. The conceptual coupling between causal structure and outcome uncertainty is reflected in how the Doctrine of Double Effect is applied in philosophical discussions and real-world moral dilemmas. We conjecture that it is this conceptual coupling rather than small differences in expected values that drives the responses of the DoubleEffect latent class. But future research could examine the expected value hypothesis by systematically varying the ratio of lives saved and sacrificed and their probabilities to determine the importance of this factor.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

6. Conclusion

In this paper, a new Conflict model of moral decision-making was proposed to address 2 concerns with the CNI model of moral judgments (Gawronski et al., 2017). First, the CNI model makes an invariance assumption in its model parameters that was found violated in Skovgaard-Olsen and Klauer (2024). Second, the CNI model does not have a built-in mechanism for handling cases where participants are conflicted between deontological and utilitarian response tendencies.

Originally, the PD model (Conway and Gawronski, 2013) was presented as an implementation of the dual-process theory (Greene, 2008; Greene et al., 2001, 2004, 2009). The CNI model built on this work, but did not adopt all of the assumptions of the dual-process theory (Gawronski et al., 2017). Both the PD and the CNI model are widely applied and continue to yield new insights into moral psychology. Yet, the multinomial processing tree model family that they are part of has not previously been extended to implement a mechanism for conflict detection and resolution, although recent work on the dual-process framework has identified these processes as an important frontier (see, e.g., Bago and De Neys, 2018; Baron and Gürçay, 2017; Białek and De Neys, 2017; De Neys, 2023; Gürçay and Baron, 2016). The CNIS Conflict model was developed to fill this gap within the framework of the PD and CNI models (see Table 2). We did this both by conserving their key advantages in modeling tasks that are not process pure and by implementing new features to handle the 2 problems of invariance violations and modeling conflict processes in moral dilemma. To the best of our knowledge, process dissociation models have not previously been employed to model the psychologically important processes of conflict detection and resolution, although discussions of dual-process theories in different domains of judgment and decision making clearly demonstrate the need for such models (De Neys, 2023).

Using Klauer's (2010) hierarchical latent trait model, the parameters of the Conflict model were estimated at the individual level (Experiment 1). In 2 validation studies (Experiments 2 and 3), the conflict structure of the new model was supported.

Experiment 1 applied the model to study participants' moral commitments. The experiment adopted an experimental task from Skovgaard-Olsen et al. (2019), Skovgaard-Olsen and Cantwell (2023), and Skovgaard-Olsen (2026) to study individual variation in cases of norm conflict in form of latent classes in the Conflict model's parameters. Four latent classes could thereby be identified with broadly consistent moral stances with respect to which participants differ. A structural equation analysis was used to contribute to a debate over whether participants' utilitarian response tendencies are an expression of a genuine moral concern or rather anti-social tendencies toward sacrifice of a calculating, selfish mind (Kahane et al., 2015). Our study shows that a plurality of moral stances coexist in different participants' reflective attitudes after exposure to conflicting moral views in a mini-longitudinal study. These moral stances are characterized by distinct covariates as well as distinct response patterns across diverse moral scenarios. This led to the investigation of a separate class following the DDE, which is not normally separated when analyzing differences between Utilitarianism and Deontology. Another novel conclusion is that there are more latent processes involved in shaping sacrificial dilemma responses than previously modeled, ranging from processes underlying responses biases and abnormal responses (Appendix B), over processes capturing norm-driven and consequence-driven processes to processes involved in the detection and resolution of conflict. On our octopus model, 8 such latent processes come together when participants are presented with incongruent items (Figure 15).

Across experiments, the Conflict model consistently performed better than rival models based on the CNI model or directly implementing it. In a direct comparison, it was found that the 3-responses format used to fit the CNIS Conflict model was better able to capture the conflict that participants experience in the 2-responses format used for fitting the CNI model (Experiment 3). Through its hierarchical structure, the CNI model preempts conflicts and it is unable to model differences in the decision-making process between congruent and incongruent items. The CNI model considered as implementing a substantive theory of moral judgment thus predicts a null effect on model-extrinsic indicators of conflicts (e.g., response times, feeling torn, and confidence) across congruent and incongruent items. In contrast, the Conflict model postulates 2 qualitatively different set of processes for congruent and

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

incongruent items. The stark asymmetries in model-extrinsic indicators of experienced conflict and the higher proportions of skip responses produced in a state of conflict in Experiment 3 strongly support the latter prediction over the former.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/jdm.2026.10030.

Funding statement. For this work, Niels Skovgaard-Olsen was supported by the research grants (497332489) and (560519952) from the German Research Council (DFG). The author furthermore acknowledges support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

Bago, B., & De Neys, W. (2018). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801.

Baron, J. & Goodwin, G. J (2021). Consequences, norms, and inaction: Response to Gawronski et al. *Judgment and Decision Making*, *16*(2), 566–595.

Baron, J., & Goodwin, G. P. (2020). Consequences, norms, and inaction: A comment. *Judgment and Decision Making*, *15*(3), 421–442.

Baron, J., & Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition*, *45*, 566–575.

Baron, J., Gürçay, B., & Luce, M. F. (2018). Correlations of trait and state emotions with utilitarian moral judgments. *Cognition and Emotion*, *32*(1), 116–129.

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

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Bayefsky, R. (2013). Dignity, honour, and human rights: Kant's perspective. *Political Theory*, *41*(6), 809–837.

Bennett, J. (1966). Whatever the consequences. *Analysis*, *26*(3), 83–102.

Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment & Decision Making*, *12*(2), 148–167.

Brandt, R. B. (1965). Toward a credible form of utilitarianism. In H.-E. Castaedañ, & G. Nakhnikian (Eds.), *Morality and the language of conduct* (pp. 107–143). Detroit, MI: Wayne State University Press.

Bürkner, P. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Choe, S. Y., & Min, K.-H. (2011). Who makes utilitarian judgments? The influences of emotions on utilitarian judgments. *Judgment and Decision Making*, *6*, 580–592.

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

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

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

Conway, P., Weiss, A., Burgmer, P., & Mussweiler, T. (2018b). Distrusting your moral compass: The impact of distrust Mindsets on moral dilemma processing and judgments. *Social Cognition*, *36*(3), 345–380.

Crain, W. C. (1985). Kohlberg's stages of moral development. In W. C. Crain (Ed.), *Theories of development* (pp. 118–136). London: Prentice-Hall.

Cushman, F. A. (2016). The psychological origins of the doctrine of double effect. *Criminal Law and Philosophy*, *10*(4), 763–766.

Cushman, F. A., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*(6), 1052–1075.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychol Sci.*, *17*(12), 1082-9.

Darwall, S. (Ed.). (2003a). *Consequentialism*. Oxford: Blackwell.

Darwall, S. (Ed.). (2003b). *Deontology*. Oxford: Blackwell.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*(85), 209–219.

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, e111, 1 –e171.

Demenchonok, E. (2009). The universal concept of human rights as a regulative principle: Freedom versus paternalism. *The American Journal of Economics and Sociology*, *68*(1), 273–302.

Elqayam, S., & Evans, J. St. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral & Brain Sciences*, *34*, 233–290.

Erdfelder, E., Auer, T., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift fur Psychologie / Journal of Psychology*, *217*, 108–124.

Fahrenwaldt, A., Olsen, J., Rahal, R.-M., & Fiedler, S. (2025). Intuitive deontology? A systematic review and multivariate, multilevel meta-analysis of experimental studies on the psychological drivers of moral judgments. *Psychological Bulletin*, *151*(4), 428–454.

Feltz, A., & May, J. (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition*, *166*, 314–327.

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

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

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

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. *47*, pp. 55–130). Cambridge, MA: Academic Press.

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

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

Greene, J. D., Cushman, F. A., 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*, 364–371.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400.

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*, 2105–2108.

Gürçay, B., & Baron, J. (2016). Challenges for the sequential two-system model of moral judgment. *Thinking & Reasoning*, *23*(1), 49–80.

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

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*, 264–284.

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

Hennig, M., & Hütter, M. (2021). Consequences, norms, or willingness to interfere: A proCNI model analyses of the foreign language effect in moral dilemma judgment. *Journal of Experimental Social Psychology*, *95*, 104148.

Herman, B. (2011). A Mismacth of methods. In S. Scheffler (Ed.), *On what matters* (Vol. *2*, pp. 83–115). Oxford: Oxford University Press.

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

Huebner, B., & Hauser, M. (2011). Moral judgments about altruistic self-sacrifice: When philosophical and folk intuitions clash. *Philosophical Psychology*, *24*(1), 73–94.

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, *27*, 116–159.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*, 131–164.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209.

Kamm, F. M. (1989). Harming some to save others. *Philosophical Studies*, *57*, 227–260.

Kamm, F. M. (2007). *Intricate ethics*. Oxford: Oxford University Press.

Kamm, F. M., & Rakowski, E. (Eds.) (2019). *The trolley problem mysteries*. Oxford: Oxford University Press.

Kay, M. (2023). tidybayes: Tidy Data and Geoms for Bayesian Models. R package version 3.0.3. https://mjskay.github.io/tidybayes/.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98.

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: The Guildford Press.

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

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory and Practice*, *16*(2), 53–59.

Körner, A., Deutsch, R., & Gawronski, B. (2020). Using the CNI model to investigate individual differences in moral judgments. *Personality and Social Psychology Bulletin*, 46(9), 1–16.

Kroneisen, M., & Heck, D. W. (2020). Interindividual differences in the sensitivity for consequences, moral norms, and preferences for inaction: Relating basic personality traits to the CNI model. *Personality and Social Psychology Bulletin*, *46*(7), 1013–1026.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*, *68*, 151–158.
Liu, C., & Liao, J. (2021). CAN Algorithm: An Individual Level Approach to Identify Consequence and Norm Sensitivities and Overall Action/Inaction Preferences in Moral Decision-Making. *Front. Psychol. 11*, 547916.
Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*, 273–286.
Luke, D. M., & Gawronski, B. (2021). Psychopathy and moral dilemma judgments: A CNI model analysis of personal and perceived societal standards. *Social Cognition*, *39*(1), 41–58.
Luke, D. M., Neumann, C. S., & Gawronski, B. (2021). Psychopathy and moral-dilemma judgment: An analysis using the four-factor model of psychopathy and the CNI model of moral decision-making. *Clinical Psychological Science*, *10*(3), 1–17.
Marshall, J., Watts, A. L., & Lilienfeld, S. O. (2018). Do psychopathic individuals possess a misaligned moral compass? A meta-analytic examination of psychopathy's relations with moral judgment. *Personality Disorder*, *9*(1), 40–50.
Mata, A. (2019). Social metacognition in moral judgment: Decisional conflict promotes perspective taking. *Journal of Personality and Social Psychology: Attitudes and Social Cognition*, *117*(6), 1061–1082.
Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2013). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*(1), 205–235.
May, J. (2018). *Regard for reason in the moral mind*. Oxford: Oxford University Press.
May, J. (2019). Précis of regard for reason in the moral mind. *Behavioral and Brain Sciences*, *42*, e146, 1 –e160.
McFarland, S., Webb, M., & Brown, D. (2012). All humanity is my ingroup: A measure and studies of identification with all humanity. *Journal of Personality and Social Psychology*, *103*(5), 830–853.
McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2020). Reasons or rationalizations: The role of principles in the moral dumbfounding paradigm. *Journal of Behavioral Decision Making*, *33*(3), 376–392.
McIntyre, A. (2019). Doctrine of double effect. In E. N. Salta (Ed.), *The Stanford Encyclopaedia of philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/double-effect/.
Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*, 193–197.
Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30.
Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Science*, *11*(4), 143–152.
Moore, A., Clark, B., & Kane, M. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgement. *Psychological Science*, *19*(6), 549–557.
Ng, N. L., Luke, D. M., & Gawronski, B. (2023). Thinking about reasons for one's choices increases sensitivity to moral norms in moral-dilemma judgments. *Personality and Social Psychology Bulletin*, *51*(1), 33–48.
Parfit, D. (2011). *On what matters, volume one, two*. Oxford: Oxford University Press.
Parfit, D. (2017). *On what matters, volume three*. Oxford: Oxford University Press.
Paruzel-Czachura, M., & Farny, Z. (2024). Psychopathic traits and utilitarian moral judgment revisited. *Personality and Social Psychology Bulletin*, *50*(9), 1368–1385.
Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., . . . Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*, *120*(2), 443–460.
Quinn, S. W. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy and Public Affairs*, *18*, 334–351.
Rauber, J. (2009). The United Nations – A Kantian dream come true? Philosophical perspectives on the constitutional legitimacy of the world organization. *The E-Journal on European, International and Comparative Law*, *5*(1), 49–76.
Renooji, S., & Witteman, C. L. M. (1999). Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, *22*, 169–194.
Rest, J.R., Narvaez, D., Bebeau, M. & Thoma, S. (1999). *Postconventional moral thinking: A neo-Kohlbergian approach*. Mahweh, NJ: Lawrence Erlbaum Associates.
Robinson, J. S., Joel, S., & Plaks, J. E. (2015). Empathy for the group versus indifference to the victim: Effects of anxious and avoidant attachment on moral judgment. *Journal of Experimental Social Psychology*, *56*, 139–152.
Rosas, A., & Koenigs, M. (2014). Beyond "utilitarianism": Maximizing the clinical impact of moral judgment research. *Social Neuroscience*, *9*(6), 661–667.
Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.
Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, *10*(4), 296–313.
Scanlon, T. M. (2011). How I am not a Kantian. In S. Scheffler (Ed.), *On what matters* (Vol. *2*, pp. 116–139). Oxford: Oxford University Press.
Schmidt, O., Erdfelder, E., & Heck, D. W. (2025). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*, *30*(4), 720–743.
Shipley, B. (2016). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R*. (2nd ed.) Cambridge: Cambridge University Press.
Singer, P. (2015). *The most good you can do: How effective altruism is changing ideas about living ethically*. New Haven: Yale University Press.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Skovgaard-Olsen, N. (2026). Causal Markov violations and hidden mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: https://doi.org/10.1037/xlm0001597
Skovgaard-Olsen, N., & Cantwell, J. (2023). Norm conflicts and epistemic modals. *Cognitive Psychology*, *145*, 101591.
Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review*, *126*(5), 611–633.
Skovgaard-Olsen, N., & Klauer, K. C. (2024). Invariance violations and the CNI model of moral judgments. *Personality and Social Psychology Bulletin*, *50*(9), 1348–1367.
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*, 204–217.
Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.
Thomson, J. J. (2008). Turning the trolley. *Philosophy and Public Affairs*, *36*(4), 359–374.
Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.
Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247–253.
Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). Oxford: Oxford University Press.
Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J.-F. Bonnefon & B. Trémolière (Eds.), *Moral inferences* (pp. 37–55). London: Routledge/Taylor & Francis Group.
Wedgwood, R. (2011). Defending double effect. *Ratio*, *24*(4), 384–401.
Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*, 28–43.
Wolf, S. (2011). Hiking the range. In S. Scheffler (Ed.), *On what matters* (Vol. *2*, pp. 33–57). Oxford: Oxford University Press.
Wood, A. (2011). Humanity as end in itself. In S. Scheffler (Ed.), *On what matters* (Vol. *2*, pp. 58–82). Oxford: Oxford University Press.

Appendix A: Multinomial processing tree models

The model equations for the 14-parameter version of the CNIS model ($CNIS_{14}$) arise by applying the following model equations separately for the intended means and the foreseeable side-effect conditions. In one joint model, different C and N parameters are then estimated for the 2 conditions.

$$P(\text{action}|\text{ProGreater}) = C_{Pro>} + (1-C_{Pro>}) \times (1-N_{Pro}) \times (1-S) \times (1-I)$$
$$P(\text{inaction}|\text{ProGreater}) = (1-C_{Pro>}) \times N_{Pro} + (1-C_{Pro>}) \times (1-N_{Pro}) \times (1-S) \times I$$
$$P(\text{skip}|\text{ProGreater}) = (1-C_{Pro>}) \times (1-N_{Pro}) \times S$$

$$P(\text{action}|\text{ProSmaller}) = (1-C_{Pro<}) \times (1-N_{Pro}) \times (1-S) \times (1-I)$$
$$P(\text{inaction}|\text{ProSmaller}) = C_{Pro<} + (1-C_{Pro<}) \times N_{Pro} + (1-C_{Pro<}) \times (1-N_{Pro}) \times (1-S) \times I$$
$$P(\text{skip}|\text{ProSmaller}) = (1-C_{Pro<}) \times (1-N_{Pro}) \times S$$

$$P(\text{action}|\text{PreGreater}) = C_{Pre>} + (1-C_{Pre>}) \times N_{Pre} + (1-C_{Pre>}) \times (1-N_{Pre}) \times (1-S) \times (1-I)$$
$$P(\text{inaction}|\text{PreGreater}) = (1-C_{Pre>}) \times (1-N_{Pre}) \times (1-S) \times I$$
$$P(\text{skip}|\text{PreGreater}) = (1-C_{Pre>}) \times (1-N_{Pre}) \times S$$

$$P(\text{action}|\text{PreSmaller}) = (1-C_{Pre<}) \times N_{Pre} + (1-C_{Pre<}) \times (1-N_{Pre}) \times (1-S) \times (1-I)$$
$$P(\text{inaction}|\text{PreSmaller}) = C_{Pre<} + (1-C_{Pre<}) \times (1-N_{Pre}) \times (1-S) \times I$$
$$P(\text{skip}|\text{PreSmaller}) = (1-C_{Pre<}) \times (1-N_{Pre}) \times S$$

For the *i*th participant, a data vector, $\boldsymbol{y_i}$, consisting of counts of each of these 3 response categories (action, inaction, skip) for each of the 4 CNI conditions (ProGreater, ProSmaller, PreGreater, PreSmaller) crossed with the 2 action type conditions (intended means vs. foreseeable side-effect) is formed. Via the CNIS model equations, these counts are modeled through a vector of 14 parameters for each participant, θ_i. In the 6-parameter version, separate C and N parameters are also estimated across the intended means vs. foreseeable side-effect conditions but the invariance assumption is made, whereby $N_{pre} = N_{pro} = N$ and $C_{Pro>} = C_{Pro<} = C_{Pre>} = C_{Pre<} = C$, resulting in a vector of 6 parameters

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

for each participant, θ_i. In the standard CNI model, the inaction bias corresponding to the I parameter governs responses when neither moral cue (norms or consequences) compels a response. Similarly, in the extended CNIS model, the skip option comes into play, if participants have no guidance as to their response from norms and consequences and thus, in the $(1-C_j) \times (1-N_k)$ cases. This dovetails with the instruction to be permitted to skip in case participants are undecided about whether the described action is morally acceptable or inacceptable. In the original model, participants have the choice between action and inaction in this state of uncertainty (cases with $(1-C) \times (1-N)$) with preferences governed by parameter I. One consequence is that although the skip parameter S is constant, the actual frequency of the use of the skip option can differ between the 4 types of dilemmas to the extent that C and N differ between them. In a sense what the model does is that it offers participants 3 choices instead of only two (i.e., action or inaction in the original model) in the case of reaching the uncertainty state with probability $(1-C_j) \times (1-N_k)$: They can then skip, choose action, or choose inaction with probabilities S, $(1-S) \times (1-I)$ and $(1-S) \times I$. The S parameter can also vary between persons.

A.1. Model equations for the Conflict model

As part of testing the Conflict model, different variants were compared in model selection exercises to identify the equality constraints on the C and N parameters leading to the optimal model. The following variants were considered: In CNIS-Conflict-latent$_{12}$, $C_{\text{intend}}^{\text{Pro}>} = C_{\text{intend}}^{\text{Pro}<} = C_{\text{intend}}^{\text{Pre}>}$ and $C_{\text{side}}^{\text{Pro}>} = C_{\text{side}}^{\text{Pro}<} = C_{\text{side}}^{\text{Pre}>}$. In CNIS-Conflict-latent$_{14}$, $C_{\text{intend}}^{\text{Pro}>} = C_{\text{intend}}^{\text{Pre}>}$ and $C_{\text{side}}^{\text{Pro}>} = C_{\text{side}}^{\text{Pre}>}$. In CNIS-Conflict-latent$_{14a}$, $C_{\text{intend}}^{\text{Pro}>} = C_{\text{intend}}^{\text{Pre}>}$ and $C_{\text{intend}}^{\text{Pre}<} = C_{\text{side}}^{\text{Pre}<}$.

Intended means

$$P(\text{action}|\text{ProGreater}) = \text{Con} \times \text{Res} \times (C_{\text{Pro}>}/(C_{\text{Pro}>} + N_{\text{Pro}})) + (1 - \text{Con}) \times C_{\text{Pro}>} + (1 - \text{Con}) \times (1 - C_{\text{Pro}>}) \times (1 - N_{\text{Pro}}) \times (1 - S) \times (1 - I)$$

$$P(\text{inaction}|\text{ProGreater}) = \text{Con} \times \text{Res} \times (1 - (C_{\text{Pro}>}/(C_{\text{Pro}>} + N_{\text{Pro}}))) + (1 - \text{Con}) \times (1 - C_{\text{Pro}>}) \times N_{\text{Pro}} + (1 - \text{Con}) \times (1 - C_{\text{Pro}>}) \times (1 - N_{\text{Pro}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{ProGreater}) = \text{Con} \times (1 - \text{Res}) + (1 - \text{Con}) \times (1 - C_{\text{Pro}>}) \times (1 - N_{\text{Pro}}) \times S$$

$$P(\text{action}|\text{ProSmaller}) = (1 - C_{\text{Pro}<}) \times (1 - N_{\text{Pro}}) \times (1 - S) \times (1 - I)$$

$$P(\text{inaction}|\text{ProSmaller}) = C_{\text{Pro}<} + (1 - C_{\text{Pro}<}) \times N_{\text{Pro}} + (1 - C_{\text{Pro}<}) \times (1 - N_{\text{Pro}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{ProSmaller}) = (1 - C_{\text{Pro}<}) \times (1 - N_{\text{Pro}}) \times S$$

$$P(\text{action}|\text{PreGreater}) = C_{\text{Pre}>} + (1 - C_{\text{Pre}>}) \times N_{\text{Pre}} + (1 - C_{\text{Pre}>}) \times (1 - N_{\text{Pre}}) \times (1 - S) \times (1 - I)$$

$$P(\text{inaction}|\text{PreGreater}) = (1 - C_{\text{Pre}>}) \times (1 - N_{\text{Pre}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{PreGreater}) = (1 - C_{\text{Pre}>}) \times (1 - N_{\text{Pre}}) \times S$$

$$P(\text{action}|\text{PreSmaller}) = \text{Con} \times \text{Res} \times (1 - (C_{\text{Pre}<}/(C_{\text{Pre}<} + N_{\text{Pre}}))) + (1 - \text{Con}) \times (1 - C_{\text{Pre}<}) \times N_{\text{Pre}} + (1 - \text{Con}) \times (1 - C_{\text{Pre}<}) \times (1 - N_{\text{Pre}}) \times (1 - S) \times (1 - I)$$

$$P(\text{inaction}|\text{PreSmaller}) = \text{Con} \times \text{Res} \times (C_{\text{Pre}<}/(C_{\text{Pre}<} + N_{\text{Pre}})) + (1 - \text{Con}) \times C_{\text{Pre}<} + (1 - \text{Con}) \times (1 - C_{\text{Pre}<}) \times (1 - N_{\text{Pre}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{PreSmaller}) = \text{Con} \times (1 - \text{Res}) + (1 - \text{Con}) \times (1 - C_{\text{Pre}<}) \times (1 - N_{\text{Pre}}) \times S$$

Foreseeable Side-effect

$$P(\text{action}|\text{ProGreater}) = \text{Con} \times \text{Res} \times (C_{\text{Pro}>,\text{side}}/(C_{\text{Pro}>,\text{side}} + N_{\text{Pro,side}}) + (1 - \text{Con}) \times C_{\text{Pro}>,\text{side}} + (1 - \text{Con}) \times (1 - C_{\text{Pro}>,\text{side}}) \times (1 - N_{\text{Pro,side}}) \times (1 - S) \times (1 - I)$$

$$P(\text{inaction}|\text{ProGreater}) = \text{Con} \times \text{Res} \times (1 - (C_{\text{Pro}>,\text{side}}/(C_{\text{Pro}>,\text{side}} + N_{\text{Pro,side}}))) + (1 - \text{Con}) \times (1 - C_{\text{Pro}>,\text{side}}) \times N_{\text{Pro,side}} + (1 - \text{Con}) \times (1 - C_{\text{Pro}>,\text{side}}) \times (1 - N_{\text{Pro,side}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{ProGreater}) = \text{Con} \times (1 - \text{Res}) + (1 - \text{Con}) \times (1 - C_{\text{Pro}>,\text{side}}) \times (1 - N_{\text{Pro,side}}) \times S$$

$$P(\text{action}|\text{ProSmaller}) = (1 - C_{\text{Pro}<,\text{side}}) \times (1 - N_{\text{Pro,side}}) \times (1 - S) \times (1 - I)$$

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

$$P(\text{inaction}|\text{ProSmaller}) = C_{\text{Pro}<,\text{side}} + (1 - C_{\text{Pro}<,\text{side}}) \times N_{\text{Pro,side}} + (1 - C_{\text{Pro}<,\text{side}}) \times (1 - N_{\text{Pro,side}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{ProSmaller}) = (1 - C_{\text{Pro}<,\text{side}}) \times (1 - N_{\text{Pro,side}}) \times S$$

$$P(\text{action}|\text{PreGreater}) = C_{\text{Pre}>,\text{side}} + (1 - C_{\text{Pre}>,\text{side}}) \times N_{\text{Pre,side}} + (1 - C_{\text{Pre}>,\text{side}}) \times (1 - N_{\text{Pre,side}}) \times (1 - S) \times (1 - I)$$

$$P(\text{inaction}|\text{PreGreater}) = (1 - C_{\text{Pre}>,\text{side}}) \times (1 - N_{\text{Pre,side}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{PreGreater}) = (1 - C_{\text{Pre}>,\text{side}}) \times (1 - N_{\text{Pre,side}}) \times S$$

$$P(\text{action}|\text{PreSmaller}) = \text{Con} \times \text{Res} \times (1 - (C_{\text{Pre}<,\text{side}} / (C_{\text{Pre}<,\text{side}} + N_{\text{Pre,side}}))) + (1 - \text{Con}) \times (1 - C_{\text{Pre}<,\text{side}}) \times N_{\text{Pre,side}} + (1 - \text{Con}) \times (1 - C_{\text{Pre}<,\text{side}}) \times (1 - N_{\text{Pre,side}}) \times (1 - S) \times (1 - I)$$

$$P(\text{inaction}|\text{PreSmaller}) = \text{Con} \times \text{Res} \times (C_{\text{Pre}<,\text{side}} / (C_{\text{Pre}<,\text{side}} + N_{\text{Pre,side}})) + (1 - \text{Con}) \times C_{\text{Pre}<,\text{side}} + (1 - \text{Con}) \times (1 - C_{\text{Pre}<,\text{side}}) \times (1 - N_{\text{Pre,side}}) \times (1 - S) \times I$$

$$P(\text{skip}|\text{PreSmaller}) = \text{Con} \times (1 - \text{Res}) + (1 - \text{Con}) \times (1 - C_{\text{Pre}<,\text{side}}) \times (1 - N_{\text{Pre,side}}) \times S$$

A.2. Bayesian hierarchical implementation

To estimate the MPT parameters of the Conflict model for each participant separately, we here follow Klauer's (2010) hierarchical latent trait method. In Skovgaard-Olsen and Klauer (2024), this approach was used to fit the CNIS model, and it has also been implemented in the TreeBUGS R-package by Heck et al. (2018). In this approach, a probit link function is used to transform MPT parameters (representing probabilities between 0 and 1) to the real line, $\Phi^{-1}(\theta)$. The transformed parameters are then modeled via a multivariate normal distribution while estimating mean, μ, and covariance matrix, Σ, from the data. The advantage of this approach is that heterogeneity in parameter estimates across participants and correlations among MPT parameters can be accommodated while allowing for partial aggregation of statistical information across participants in the posterior parameters of the multivariate normal distribution (Klauer, 2010). Accordingly, for each participant, i, the probit-transformed parameters are additively decomposed into a group mean, μ, and a random effect, $\Phi^{-1}(\theta) = \mu + \delta_i$. Table A1 illustrates CNIS_{14} for the intended means condition, whereby a distinct C parameter is estimated for each of the $j = 1, \ldots, 4$ CNI conditions, and a distinct N parameter is estimated for each of the $k = 1, 2$ types of norms. In addition, 6 parallel C and N parameters are estimated by CNIS_{14} for the foreseeable side-effect condition. For CNIS_6, one shared C parameter is estimated ($j = 1$) together with one shared N ($k = 1$) parameter for each the intended means vs. foreseeable side-effect conditions.

The models were fitted in a Bayesian framework through a Gibbs sampler, which estimates the posterior distributions of model parameters by means of Monte Carlo-Markov chains. In pilot studies, different versions of the CNIS-Conflict model were compared using the WAIC and LOOIC information criteria (Vehtari et al., 2017) and the posterior predictive checks (T_1, T_2) proposed in Klauer (2010). It was repeatedly found that introducing the following equality constraints on C parameters of the CNIS-Conflict model improved its performance and so we adopted this version of the model throughout the manuscript ($\text{CNIS-Conflict}_{14a}$): $C^{\text{Pro}>}_{\text{intend}} = C^{\text{Pre}>}_{\text{intend}}$ and $C^{\text{Pre}<}_{\text{intend}} = C^{\text{Pre}<}_{\text{side}}$.[21]

A.3. Latent class analysis of scorekeeping judgments

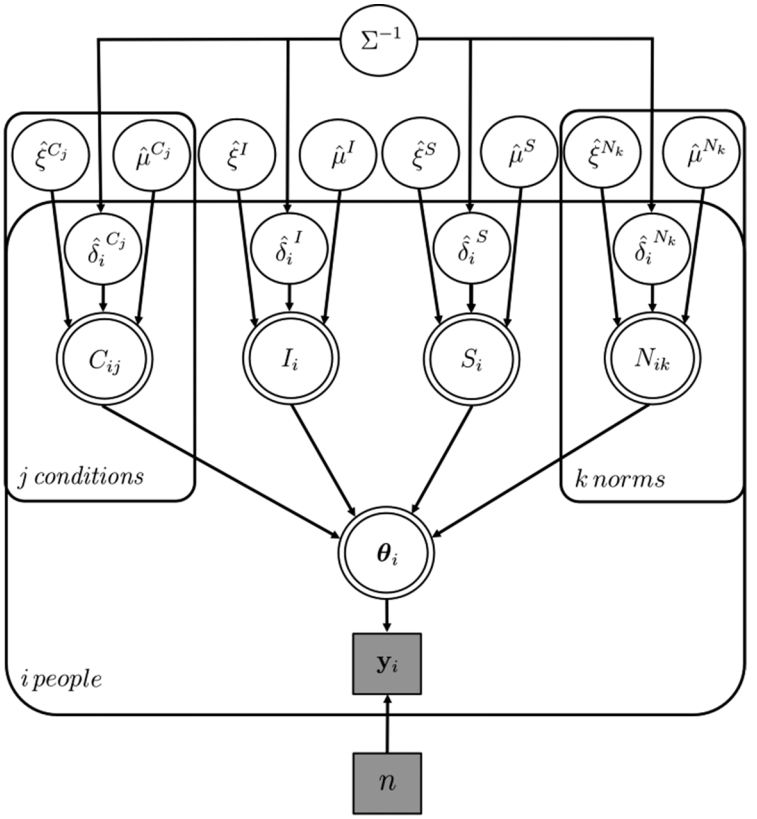
The latent class analysis in Table A2 was applied to the scorekeeping judgments.

This latent class analysis was repeated for 2, 3, and 4 group solutions. The categorical labels estimated by LCA_4 (Table 7) were used to estimate different MPT parameters of the $\text{CNIS-Conflict}_{14a}$,

[21] When producing posterior predictive plots for Conflict_{14a}, issues with arithmetic underflow for probabilities smaller than 1×10^{-10} were encountered when monitoring one of the variables used to calculate p_{T1}. Fixing these probabilities to 1×10^{-10} solved the problem.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table A1. *Hierarchical latent trait MPT model.*



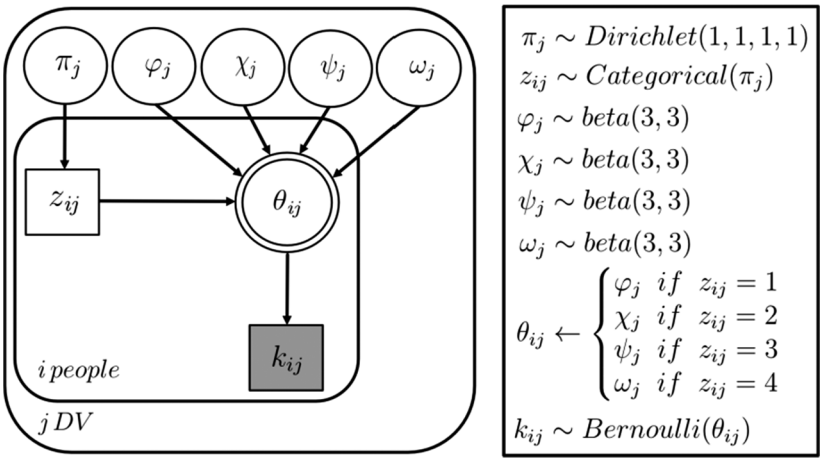
$$
\begin{aligned}
\hat{\mu}^{C_j}, \hat{\mu}^{I}, \hat{\mu}^{S}, \hat{\mu}^{N_k} &\sim Gaussian(0,1) \\
\hat{\xi}^{C_j}, \hat{\xi}^{I}, \hat{\xi}^{S}, \hat{\xi}^{N_k} &\sim Uniform(0,10) \\
\Sigma^{-1} &\sim Wishart(\mathbf{I}, df) \\
\left(\hat{\delta}_i^{C_j}, \hat{\delta}_i^{I}, \hat{\delta}_i^{S}, \hat{\delta}_i^{N_k}\right) &\sim MvGaussian(\mathbf{0}, \Sigma^{-1}) \\
C_{ij} &\leftarrow \Phi\left(\hat{\mu}^{C_j} + \hat{\xi}^{C_j}\hat{\delta}_i^{C_j}\right) \\
I_i &\leftarrow \Phi\left(\hat{\mu}^{I} + \hat{\xi}^{I}\hat{\delta}_i^{I}\right) \\
S_i &\leftarrow \Phi\left(\hat{\mu}^{S} + \hat{\xi}^{S}\hat{\delta}_i^{S}\right) \\
N_{ik} &\leftarrow \Phi\left(\hat{\mu}^{N_k} + \hat{\xi}^{N_k}\hat{\delta}_i^{N_k}\right) \\
\mathbf{y_i} &\sim Multinomial(\boldsymbol{\theta}_i, n)
\end{aligned}
$$

Note: There are 4 CNI conditions with 3 categorical responses (action, inaction, skip). Via the CNIS model equations displayed above, the outcome probabilities of the responses in the data vector, y_i, are represented by 14 theta parameters (here only illustrated for the 8 parameters in the intended means conditions). For each participant, a vector of 14 theta parameters, θ_i, is estimated. The inverse Wishart distribution has 14+1 degrees of freedom, *df*, and a 14×14 identity matrix, I, as scale matrix.

for participants who were classified as members of these 4 latent classes. This resulted in the CNIS-Conflict-latent$_{14a}$ model reported in Experiment 1, which is identical to the CNIS-Conflict$_{14a}$ model adds latent group-specific means to the estimated MPT parameters based on the 4 latent classes in the scorekeeping phase.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table A2. *Latent class analysis.*



Note: Table A2 displays the latent class analysis with 4 groups. In addition, versions with 2 and 3 groups solutions were fitted to the *j* scorekeeping judgments in a model comparison. The categorical HIT approval variable was expanded into a set of one-hot indicators and included alongside the criticism indicators in the LCA. This specification intentionally increases the influence of HIT approvals on class formation for purposes of diagnostic clustering but makes the resulting indicators structurally dependent. However, the indicator structure was held constant across all models, such that relative comparisons among models differing only in the number of latent classes remain internally comparable.

Appendix B: Violations of invariance

Table B1 shows a test of the invariance assumption across latent classes.

The invariance assumption in the CNI model involves setting C and N parameters equal across the 4 CNI conditions. Table B1 shows that this assumption is violated: credible differences in both C and N parameters emerge across conditions for each latent class, corroborating earlier findings (Skovgaard-Olsen and Klauer, 2024). Like in these past findings, when a difference in the N parameters appear, we again observe $N_{Pre} > N_{pro}$. Although some reviewers have found this pattern counterintuitive, the prescriptive trials in our materials do not involve a distinct positive duty (e.g., a duty to aid). Instead, they present the same deontological prohibition on sacrificial harm, but embed it in a context in

Table B1. *Contrasts in the CNIS-Conflict-latent$_{14a}$ parameters.*

Contrast	Altruism	Utilitarianism	Deontology	Double effect
		Invariance Assumption		
$\mathbf{N}^{\mathbf{pro}}_{\mathbf{intend}} - \mathbf{N}^{\mathbf{pre}}_{\mathbf{intend}}$	–	−.49 [−.72, −.21]	−.33 [−.69, −.02]	–
$\mathbf{C}^{\mathbf{Pro>,Pre>}}_{\mathbf{intend}} - \mathbf{C}^{\mathbf{Pre<}}_{\mathbf{intend}}$	–	–	–	–
$\mathbf{C}^{\mathbf{Pro<}}_{\mathbf{intend}} - \mathbf{C}^{\mathbf{Pre<}}_{\mathbf{intend}}$	.28 [.01, .58]	.46 [.10, .73]	.77 [.40, .95]	–
$\mathbf{N}^{\mathbf{pro}}_{\mathbf{side}} - \mathbf{N}^{\mathbf{pre}}_{\mathbf{side}}$	–	–	–	–
$\mathbf{C}^{\mathbf{Pro>}}_{\mathbf{side}} - \mathbf{C}^{\mathbf{Pre<}}_{\mathbf{side}}$	.30 [.10, .48]	.62 [.42, .78]	.45 [.33, .56]	.47 [.18, .72]
$\mathbf{C}^{\mathbf{Pro<}}_{\mathbf{side}} - \mathbf{C}^{\mathbf{Pre<}}_{\mathbf{side}}$	–	–	–	–
$\mathbf{C}^{\mathbf{Pre>}}_{\mathbf{side}} - \mathbf{C}^{\mathbf{Pre<}}_{\mathbf{side}}$	–	–	.41 [.12, .67]	—

Note: The contrasts were calculated based on the differences in 10,000 random posterior draws of the mean parameters. In each case, the posterior median of the contrast effect is reported along with the 95% HDI in square brackets. Only contrasts effects where the 95% HDI crosses zero are reported.

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

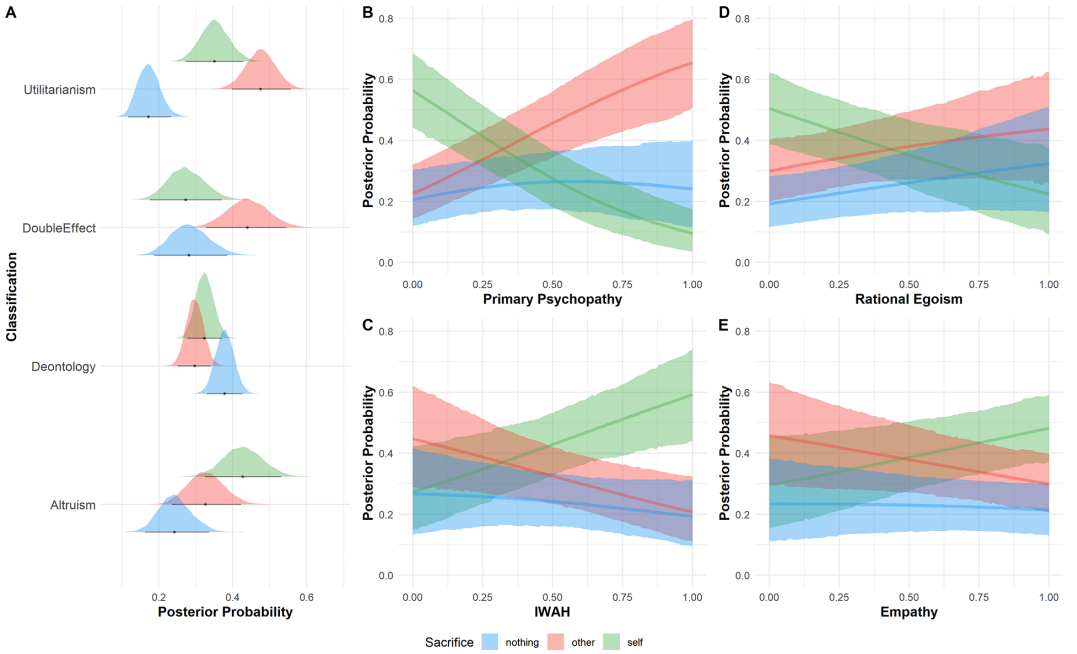


Figure B1. *Posterior predictive probabilities of categorical sacrificial responses.*
Note: Panels A, B, and C were produced based on the posterior predictions of one categorical regression model, which controlled for IWAC and IWAA. Panels D and E were produced by the posterior predictions of a second categorical regression model, which controlled for psychological egoism and moral egoism. All covariates were rescaled to range between 0 and 1. The black dots and lines in Panel A indicate the median and the 95% HDI. The ribbons in Panels B-E indicate 95% HDIs.

which instrumental harm has been preselected by an assistant on behalf of the participant. Under such circumstances, participants may experience reactance to having an assistant planning sacrificial harm in their name, which they do not condone. This could produce a stronger activation of the deontological norm in the prescriptive condition and thus explain the direction of the observed effects both in these and in previous studies.

B.1 Self-sacrifice

To investigate relationships between the latent classes, covariates, and the categorical outcomes of the modified runaway Trolley trilemma (sacrifice self vs. sacrifice other vs. do nothing), 2 categorical regression models were fitted. The first model included primary psychopathy, the latent classes, and IWAH ('identification with all of humanity') as predictors. In line with previous research (see Kahane et al., 2015), IWAA ('identification with home country') and IWAC ('identification with local community') were used as covariates to control for their influence when assessing the predictive value of IWAH.

Since primary psychopathy is negatively correlated with empathy ($r = -.56$, 95% HDI [−.61, −.51]) and positively correlated with rational egoism ($r = .52$, 95% HDI [.47, .57]), a second regression model included just these 2 predictors and psychological egoism and moral egoism as covariates. Figure B1 plots the posterior predictions of these 2 regression models and shows how sacrificial responses differ across latent classes and covariates.

Panels B–E of Figure B1 show that higher scores on primary psychopathy and rational egoism were associated with higher posterior probability of sacrificing the life of an innocent other person, whereas higher scores on IWAH and empathy were associated with higher posterior probability of self-sacrifice. Regarding the latent classes juxtaposed in Panel A, note that the presented switch version of

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

the Trolley scenario with the option of self-sacrifice corresponds to a foreseeable side-effect item in the ProGreater condition. For such items, members of Utilitarianism and DoubleEffect are expected to sacrifice the life of an innocent stranger, deontologists are expected to do nothing and altruists are expected to sacrifice themselves. These expectations were largely born out. To investigate these differences further, we analyzed pairwise contrasts between the latent classes to identify credible differences for which the 95% HDI do not include zero. Participants adhering to Utilitarianism had credibly higher posterior probability of sacrificing another innocent person than participants adhering to Altruism ($\Delta^{\sim other}_{Altruism-Utilitarianism} = -.15$, 95% HDI [−.28, −.03]) and Deontology ($\Delta^{\sim other}_{Utilitarianism-Deontology} = .18$, 95% HDI [.08, .26]) but participants adhering to the DDE did not ($\Delta^{\sim other}_{Utilitarianism-DoubleEffect} = .03$, 95% HDI [−.10, .17]). Participants adhering to Altruism had higher posterior probability of sacrificing themselves than participants adhering to the DDE ($\Delta^{\sim self}_{Altruism-DoubleEffect} = .15$, 95% HDI [.01, .29]) and lower posterior probability of doing nothing than participants adhering to Deontology ($\Delta^{\sim nothing}_{Altruism-Deontology} = -.14$, 95% HDI [−.23, −.03]). Furthermore, participants adhering to Deontology also had higher posterior probability of doing nothing than participants adhering to Utilitarianism ($\Delta^{\sim nothing}_{Utilitarianism-Deontology} = -.21$, 95% HDI [−.28, −.13]). Finally, participants adhering to the DDE had a higher posterior probability of sacrificing other than participants adhering to Deontology ($\Delta^{\sim other}_{Deontology-DoubleEffect} = -.14$, 95% HDI [−.26, −.03]).

Thomson (2008) introduced the third option of altruistic self-sacrifice in the Trolley dilemma to draw attention to the circumstance that if we do not feel morally obliged to sacrifice ourselves to save 5 strangers, then we cannot view it as morally permissible to sacrifice the life of a stranger without her consent. Our results indicate that participants were split over this third option of self-sacrifice. While participants adhering to Utilitarianism and the DDE continued to have the highest posterior probability of sacrificing someone else, the larger group of participants adhering to Deontology had the highest posterior probability of doing nothing (following Thomson, 2008). In contrast, only the minority of participants adhering to Altruism had self-sacrifice as the preferred choice.

B.2 Posterior predictive plots, Experiment 2

In Table B2, a visual comparison is shown between the descriptive response frequencies and the posterior predictive performance of the models contrasted in Experiment 2.

B.3 Abnormal responses

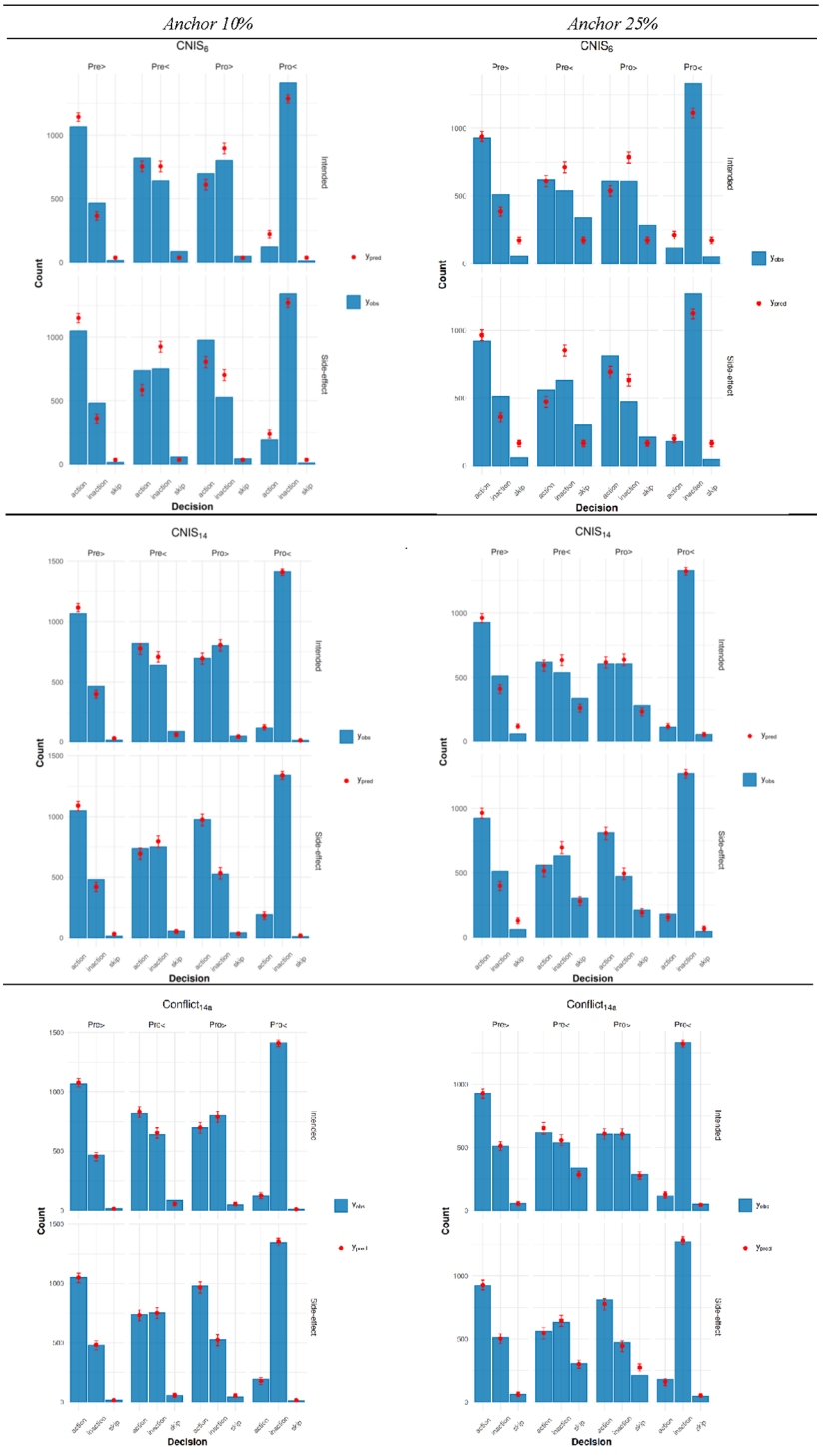
A final analysis concerns action/inaction responses in the congruent trials ($Pro_<$, $Pre_>$), which are neither predicted by Utilitarianism nor Deontology. In Baron and Goodwin (2020, 2021) and Gawronski et al. (2020) such responses are denoted as 'perverse', but we here prefer the less value-laden term 'abnormality rates', which indicates that the response pattern goes against the expectations of 'normality' and is in this sense a form of pathological response—while remaining neutral on its etiology. We investigate the rates of abnormal responses in our updated stimulus materials as compared to the original CNI stimulus materials (Gawronski et al., 2017; Körner et al., 2020) and whether there is a subset of participants who are predominantly responsible for producing abnormal responses.

To analyze response patterns of action/inaction responses in the congruent trials ($Pro_<$, $Pre_>$), which are neither predicted by Utilitarianism nor Deontology, we report the abnormality rates from 2 representative CNI studies with the original stimulus materials presented in a pseudo-random order (Gawronski et al., 2017; Körner et al., 2020) and compared these to the abnormality rates from Experiments 1–3, with the new updated stimulus materials presented in a random order (Figure B2).

To analyze the results, a Bayesian regression analysis was applied with the abnormality rates as outcome and the Type of experiment (CNI vs. Conflict) and the Condition (PreGreater vs. ProSmaller)

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

Table B2. *Posterior predictive plot, Experiment 2.*



Note: Observed response frequencies (blue bars) and posterior predictive frequencies (red points) for 3 models: $CNIS_6$ (first row), $CNIS_{14}$ (second row), and CNIS-$Conflict_{14a}$ (third row). Columns correspond to the 4 CNI scenarios: Pre<, Pre>, Pro>, Pro<, where > indicates that the benefit of the sacrifice is greater than the costs. Columns within panels show response type (action, inaction, skip), and panels are split by causal structure (Intended means vs. Foreseeable side-effect).

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

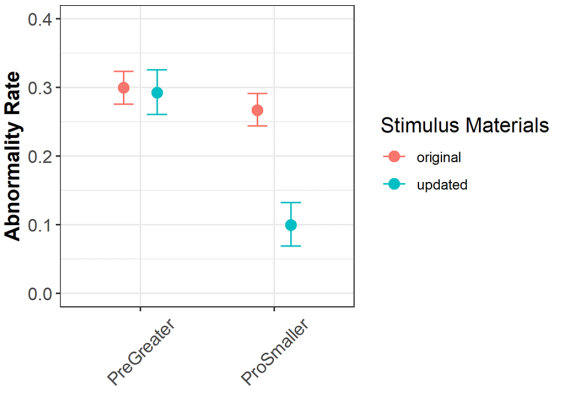


Figure B2. *Abnormality rates across studies.*
Note: The abnormality rates for the CNI studies were collected Gawronski et al., (2017, Table 2) and Körner et al. (2020, Table 5) and the abnormality rates for the updated stimulus materials were from Experiments 1, 2, and 3.

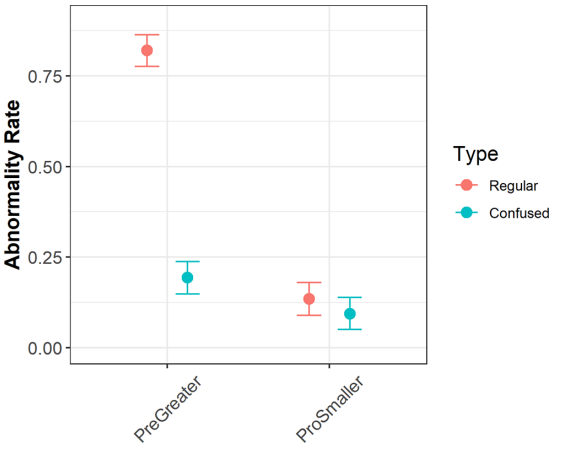


Figure B3. *Abnormality rates, confused versus non-confused.*
Note: The abnormality rates for the 3-responses option data of Experiments 1–3 based on whether participants were captured by the mixture component that confused action and inaction in the Prescriptive condition ("Confused") or the mixture component with the correct assignment of Action and Inaction response options.

as well as their interaction as predictors. It was found that there was evidence of a 2-way interaction, −.16, 95% HDI [−.11, −.22]. Pairwise contrasts revealed that there was no credible difference between the abnormality rates between Experiments 1, 2, and 3 and the CNI studies for the PreGreater condition ($\tilde{\Delta}_{original-updated}$ = .01, 95% HDI [−.03, .05]), but smaller abnormality rates were found for the updated stimulus materials in the ProSmaller condition ($\tilde{\Delta}_{original-updated}$ = .17, 95% HDI [.13, .21]). Further analysis indicated that abnormal responses in the PreGreater condition were disproportionately concentrated on a minority of participants. To investigate the hypothesis that high abnormality rates were produced by a confusion of the action and inaction response options among these participants, the models from Experiments 1–3 were refit with 2 mixture components; one component corresponding to the models reported in Experiments 1–3 and a second mixture component with the same models which reversed action and inaction response options for the prescriptive condition. Details on these models can be found in supplementary materials uploaded to the OSF project page.

We then compared the abnormality rates of the participants who were captured by these 2 mixture components (Figure B3).

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

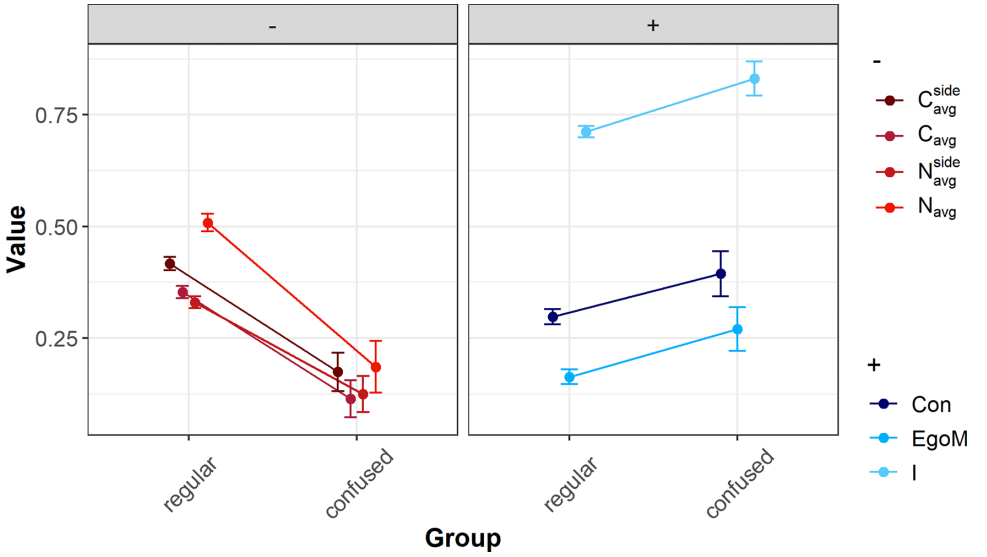


Figure B4. *Differences between confused and regular participants.*
Note: The figure shows differences in MPT parameters and moral egoism of the Conflict model for the confused subset of the participants compared to the remaining sample. 'EgoM' = Moral egoism. Only credible differences are plotted where the 95% HDI excludes zero. All the effects plotted are simple effects of group membership on the indexed parameters. The separation into increase (+) and decrease (−) is for visualization.

It was found that there was evidence of a 2-way interaction, .59, 95% HDI [.50, .68], such that there was no credible difference between the abnormality rates between participants captured by the 2 mixture components across studies for the PreSmaller condition ($\tilde{\Delta}_{Confused-Regular}$ = .04, 95% HDI [−.02, .11]), but larger abnormality rates were found for the PreGreater condition for the confused participants ($\tilde{\Delta}_{Confused-Regular}$ = .63, 95% HDI [.56, .69]). In other words, the elevated abnormality rate found in our studies in the PreGreater condition can be accounted for by a minority of participants who confused action and inaction response options for prescriptive norms.

To investigate whether confused participants differed from other participants, a multivariate Bayesian regression model with correlated effects was fitted and credible differences from this model are plotted based on the data from Experiment 1 in Figure B4.

As Figure B4 shows, credible differences were found such that the confused subset of participants had lower values on the averaged C and N parameters and higher values on the parameters Con and I as well on the moral egoism covariate than the remaining sample. In contrast, credible differences were not found for the covariates empathy, primary psychopathy, IWAH, and Greater Good.

Regarding abnormality rates in Experiments 1–3, it was found that the rates were about half the size of the rates found in previous published CNI studies for ProSmaller trials. Aside from effects due to our modified stimulus materials, one explanation for these reduced rates in the ProSmaller trials may be that we randomly assigned the CNI conditions to different scenarios instead of presenting all 4 versions of the scenarios within subject as previous studies with higher abnormality rates did (Gawronski et al., 2017; Körner et al., 2020). This procedural change prevents participants from confusing the ProSmaller condition with one of the previous conditions that they have encountered the same scenario in. For PreGreater trials, comparable rates as in previous published studies were found (Figure B2) and those rates were due to a minority of participants who confuse the action and inaction response options in the Prescriptive scenarios. Given that the rates for the abnormality rates for the PreGreater trials were comparable to those found in previous published studies (Gawronski et al., 2017; Körner et al., 2020), which used a similar design for the Prescriptive manipulation, it is likely that this response bias also affects those studies. When examining individual differences between the confused participants and the

https://doi.org/10.1017/jdm.2026.10030 Published online by Cambridge University Press

remaining sample, it was found that the confused participants had lower averaged C and N parameters, higher inaction bias, and more pronounced tendencies toward moral egoism than the remaining sample (Figure B4). By excluding such participants from the analysis, future studies would be able to lower the abnormality rates substantially (Figure B3), beyond the reduction in abnormality rates already achieved by the updated stimulus materials and their random presentation used in the present studies. This in turn would permit future studies to resolve one of the standing issues in the critical discussion between Baron and Goodwin (2020, 2021) and Gawronski et al. (2020) over the foundations of the CNI model.

Cite this article: Skovgaard-Olsen, N. and Klauer, K. C. (2026). Norm conflicts and morality: The CNIS Conflict model of moral decision-making. *Judgment and Decision Making*, e7. https://doi.org/10.1017/jdm.2026.10030