Measuring Moral Vacillation

Revati Shivnekar (revatis@iitk.ac.in)
Department of Cognitive Science, IIT Kanpur
Uttar Pradesh, India

Nisheeth Srivastava (nsrivast@iitk.ac.in)
Department of Cognitive Science, IIT Kanpur
Uttar Pradesh, India

Abstract

Moral decision-making research is currently dominated by experimental studies that employ dilemmas, situations where more than one course of action may be justifiable. Humans almost characteristically vacillate between options before reaching a conclusion while reasoning on such problems. Current experimental designs disregard this vital aspect of moral decisions by only measuring judgments produced at the end of reasoning. We present an experimental paradigm for measuring moral conflict as a function of vacillations experienced by participants while deliberating. We conducted two experiments to correlate our measure with two different definitions of conflict prevalent in the literature. Across both experiments, we found that people vacillate more on conflicting problems and that vacillations correlate with their subjective feeling of conflict and confidence. We also found that the pattern of deliberation uncovered by these vacillations is inconsistent with currently favored models of moral reasoning and more consistent with a single accumulation to threshold process.

Keywords: moral vacillations; conflict measurement; moral decision-making; dual-process theory

Introduction

Moral dilemmas pit utilitarian and deontological principles against one another and are a staple of the contemporary study of moral decision-making. In these dilemmas, the actor must either take action that maximizes welfare, such as saving more people by killing some (the utilitarian principle), or the actor should do nothing (as per the 'do-no-harm' deontological principle). The experimental study of moral decision-making centrally revolves around asking lab participants to respond to descriptions or simulations of moral dilemmas of this nature, systematically changing the nature of the dilemmas presented, and measuring cohort-level changes in response proportions for either choice.

Such experimental paradigms have found conceptual consilience with the dual-process theory of decision-making (DPT), which has been extensively employed in the field of moral psychology in the last few decades (Bago & De Neys, 2019; Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2009). This approach argues for the role of two distinct cognitive mechanisms in moral judgments, with System 1 supporting quick responses based on intuitive inclinations and System 2 supporting delayed responses based on some utility calculation.

The corrective model of DPT, commonly used to frame moral cognition studies, states that in dilemmas where the deontological and utilitarian principles cue distinct choices, deontological judgments (henceforth, D) are supported by System 1 (Greene et al., 2004, 2009; Paxton, Ungar, & Greene, 2012). In contrast, utilitarian judgments (henceforth, U) are available after spending resources such as time and working memory capacity. Thus, conflict is typically cited as a mechanistic phenomenon resulting from both systems competing for control of the final judgment (see Greene et al. (2004)). To elaborate, in impersonal dilemmas in which the action operates indirectly through mechanistic mediations like diverting a trolley, shooting a gun, etc., System 1 has a weak D preference that can easily be overridden by System 2's strong preference for U, resulting in most people choosing U in such dilemmas. Alternatively, in case of personal dilemmas in which the harm comes to the victim by direct application of muscular force, most individuals typically select the D alternative (Cushman et al., 2006; Greene et al., 2009; Moore, Clark, & Kane, 2008). System 1's D response has a strong activation in such cases that System 2 cannot overcome, causing the majority of people to choose D. When people do select U in these scenarios, they require extra time to commit to it. Because of this resource reliance hypothesis, reaction times are frequently used as a conflict metric in experimental studies of moral decision-making (Greene et al. (2001); Paxton et al. (2012) but also see Baron and Gürçay (2017)).

Alternatively, Koenigs et al. (2007) have operationalized conflict as an agreement on final judgments on moral dilemmas among individuals. They segregated personal dilemmas into either high- or low-conflict groups. A dilemma was considered low-conflict when close to 100% participants disagreed with the proposed U action. On the other hand, the high-conflict dilemmas produced no such pattern in judgments with varying degrees of U endorsement at the cohort level.

Finally, Bago and De Neys (2019) define conflict in terms of choices cued by deontological and utilitarian principles. When these two principles cue distinct choices, there is a conflict in resolving such a dilemma. But when they cue the same choice, conflict is minimized (see also Białek and De Neys (2016, 2017)).

To summarise, experimental research into moral decision-making currently measures conflict either indirectly via re-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

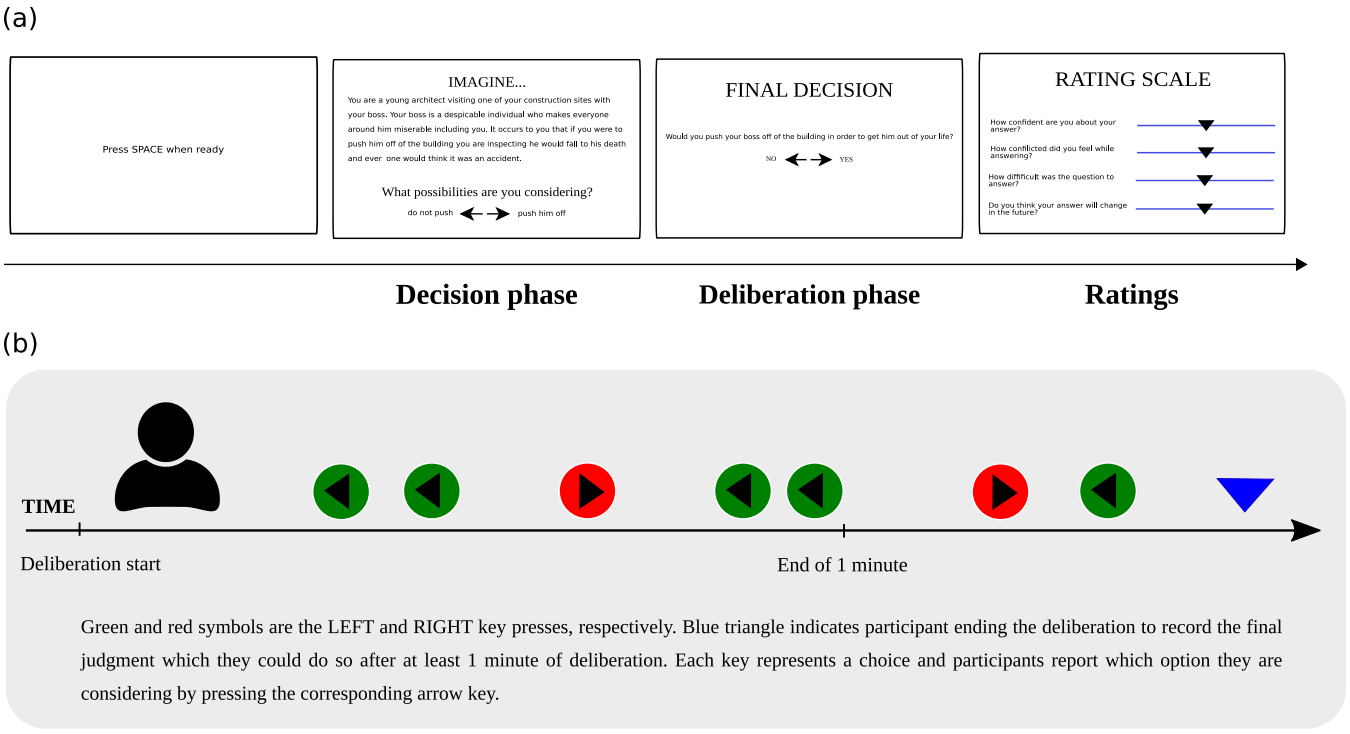

Figure 1:
(a) shows the trial structure in both experiments. (b) depicts representational key-presses during the deliberation phase of a trial over a time period. In moral trials, D and U corresponded to the left and right arrow keys, respectively.

action times, with support from corrective DPT models, as cohort-level disagreement on choices, or based on normative expectations of behavior emerging from alternative principles of choice. Although all these approaches are reasonable, they have certain limitations.

While it is certainly true that respondents may take longer to respond in situations where they are conflicted, the fact that someone is taking a long time to respond may not be because of problems in executive control between a fast System 1 and a slow System 2; it could be because sequential consideration of evidence is shifting their preferences below an evidence threshold. Thus, the strong theoretical commitments that accompany response time-based measurements of conflict require substantiation that is not yet apparent (see Gürcay and Baron (2017)). Further, while a response time measurement may summarize the magnitude of conflict experienced during a trial (compared with other trials from the same respondent), it cannot actually identify instances of conflict within a trial and is thus a summary measure with coarse granularity.

The operationalizations of dilemmas as conflicted based on cohort-level disagreement (Koenigs et al., 2007) as well as based on normative expectations of rationality (Bago & De Neys, 2019) attributes levels of conflict to specific presentations of moral dilemmas, not to an individual's experience of it; which are entirely different things. For instance, when asked to choose between tea and coffee, a group may split entirely down the middle, suggesting that beverage choice is potentially a high-conflict decision. However, each group member may have experienced no conflict in their beverage selection. We elaborate on this point further.

Measuring cognitive conflict as within-trial vacillation

A common experience in reasoning about moral dilemmas is one of 'vacillation' before we commit to a choice. When we deliberate over complex situations, we consciously analyze many arguments that lead to divergent options, possibly in an arbitrary order. Often we find ourselves switching gears while reasoning and reconsidering choices previously explored. People fluctuate in their preferences mentally, switching back and forth between options as different considerations reveal themselves sequentially during deliberation. Such an experience of conflict when the choice is not simple, and different arguments pull us in different directions is familiar and is, in fact, characteristic of 'dilemmas' (Cushman & Greene, 2012; Greene, 2007; Paxton et al., 2012).

Therefore, we suggest that measuring interim preference reversals or vacillations within choice trials is critical for realistically measuring cognitive conflict and differentiating reasonable theories of moral decision-making. To this end, we present an experimental paradigm that captures par-

ticipants' instantaneous preferences while reasoning about moral dilemmas. Participants reported which direction their thoughts were leaning at the time they were contemplating choices in dilemmas. They could report their thoughts whenever they felt a preference building and as often as they wished to do so. This gave us an insight into how individuals reason internally over time. Our objective in employing this paradigm was to establish the internal and external validity of our intra-trial vacillation measurements as a measure of cognitive conflict. To establish internal validity, it would be sufficient to show that people vacillate more when they subjectively feel conflicted during a choice. To establish external validity, it would be sufficient to show that vacillations are correlated systematically with earlier measurements of conflict defined in the literature, viz., response times and subjective confidence and conflict assessments of individuals.

In this paper, we report results from two experiments testing two operationalizations of conflict (by Koenigs et al. (2007) and Bago and De Neys (2019)). Experiment 1 assessed whether conflict, defined as the level of agreement in final judgments at the cohort level, translates to mental vacillations. Experiment 2 examined whether people vacillate more when deontological and utilitarian principles cue separate choices rather than the same one.

Experiment 1

Method

Participants Based on a pilot study (Cohen's d = 0.67, α = .05, power = .8), we collected data from 25 participants (13 females; mean age= 25.3 years).

Experiment Design and Procedure Participants deliberated on 16 problems from four conditions taken from Koenigs et al. (2007): non-moral (NM), impersonal (IM), low-conflict (LC), and high-conflict (HC). From the listed dilemmas in Koenigs et al. (2007), we ranked all moral problems (IM, LC, and HC) based on the mean emotionality ratings within the condition. We then short-listed four cases from each condition based on expected exposure to participants (unfamiliarity with the presented dilemmas was confirmed during the pilot).

Each of these problems required participants to make a two-alternative forced choice between an action and its omission. On NM trials, the problem contexts did not invoke any moral principles. The actions included scheduling appointments, deciding between two routes to take (2 problems had this action), and buying product A instead of B. Moral trials required participants to judge the rightness of the action in light of the circumstances. Particularly, actions in LC and HC conditions saved a bigger group by injuring or killing a small number of people. These actions were personal in the sense that they caused direct harm to another person or group of people (for in depth discussion of 'personal' actions used in this sense, see (Greene et al., 2001, 2004)). Six out of 8 of these trials involved death. In contrast, IM dilemmas did not have any trial where the victim died as a result of carrying out the action. These included actions that facilitated the actor's welfare, such as stealing cash from a wallet on the ground, bribing to win a case, enabling illicit financial operations, and lying on one's CV. These dilemmas can be found in Koenigs et al. (2007).

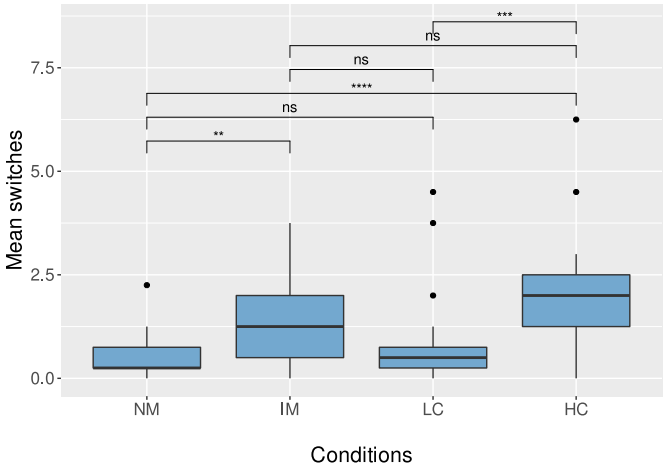

Figure 2: Box plot of mean switches across for participants in Experiment 1. Whiskers indicate the interquartile range. *p* values are Bonferroni corrected for multiple comparisons.

All trials were self-paced. A trial consisted of deliberation and decision, followed by participants rating their experience of reasoning on the problem (see Figure 1 for the trial structure). In the deliberation phase, participants read the context of the problem. At the bottom of the screen, choices were shown under the prompt "What possibilities are you considering?". Each choice was associated with either the LEFT (L) or RIGHT (R) arrow key. On moral trials, L was always deontological while R was utilitarian. During the deliberation phase, participants were asked to be attentive to their thoughts and indicate which option they preferred at that moment by pressing the appropriate key corresponding with that option. They could report their preferences during this time any number of times (but at least once) and whenever they wished to do so. We excluded trials when no key was pressed during this time. This phase lasted at least one minute to stimulate reasoning, although participants could take longer if needed. After 1 minute, they could proceed to the next screen to give their final decision. The arrow keys on this and the previous screen corresponded with the same option. Finally, they rated their experience of reasoning about the problem on four 5-point scales, viz., how confident they felt about their final answer, how conflicted they felt while deliberating, how difficult the question was to answer, and whether they think they will change their decision in the future.

Results

Conflict at the cohort level is reflected in switches in preferences. Our paradigm could capture participants' momentary preferences that were oftentimes modified as they deliberated. Going back and forth between the two op-

tions was frequently observed. A change of response while deliberating from left to right or vice versa counted as a switch. Repeated measures one-way ANOVA testing whether conditions predict the number of switches was significant, $F(3,72) = 14.908, p < .01, \eta^2 = .383$. We then conducted pairwise t-test comparisons using Bonferroni correction. As expected, though HC recorded the highest number of [$M = 2.04, SD = 1.39$] switches, there was no significant difference in switches recorded in HC and IM [$M = 1.38, SD = 1.01$], $t(24) = 2.14, p = .04, d = .5$. Switches recorded in NM [$M = 0.52, SD = 0.51$] were significantly lower than HC ($t(24) = 6.32, p < .001, d = 1.45$) and IM ($t(24) = 4.29, p < .001, d = 1.08$). Similarly, as expected, LC [$M = 0.82, SD = 1.1$] also recorded a significantly lower number of switches than HC ($t(24) = 5.11, p < .001, d = 0.97$) and IM ($t(24) = 2.14, p < .001, d = 0.53$). But, there was no difference switches between LC and NM ($t(24) = 1.46, p = .2, d = 0.35$). These results are graphically displayed in Figure 2.

This set of observations suggests that in accordance with our hypothesis, mental vacillations are diagnostic of discrepancies in cohort-level final judgments. In essence, we show that group disagreement on endorsing an action manifests in vacillations within an individual while thinking. Participants switched more frequently when there was considerable cohort-level conflict, as in HC than when there was broad agreement on the chosen action, as in LC. Interestingly, actions in IM dilemmas that differed from harm-inflicting actions in HC produced as many shifts in preferences as HC.

Temporal order of judgments is inconsistent with dual process theories. As in Koenigs et al. (2007), the LC condition elicited D final judgments on most trials (92%). Likewise, the HC and IM conditions received more mixed responses, with people preferring U over D on 57% and 37% of trials for these conditions, respectively. We then compared the first and the last preferences recorded by participants while they deliberated. There could be four different types of response transitions: DD (first and last responses were both D), DU (first was D and last U), UD (first was U and last D), and UU (first and last both U). Firstly, it is worth noting that all 4 response change types were reported in moral dilemmas (see Table 1), which is unexpected based on DPT. The DPT models emphasize that certain kinds of judgments precede some other judgments. For instance, the corrective model hypothesizes that D judgments are quickly generated, which may or may not be corrected to U. Hence, DU and DD are two possible response changes in a moral trial, with UU remaining a possibility if D judgments are covertly missed; however, UD transitions cannot be accommodated by DPT models. As is evident in Table 1, UD transitions, where respondents begin with a utilitarian judgment and then settle into a deontologically motivated position occur significantly more frequently than theoretical expectation across conditions (one-proportion z-test: $z = 46.52, p < .001, CI_{percent} = [11,19]$).

DD and UU were the most commonly observed response transitions in Experiment 1. However, even within such trials, preferences still shifted intermediately for a significant fraction of the trials. The percentages of UU and DD trials, when participants had switched at least twice in between, were significantly above floor at 58% (one proportion z-test: $z = 82.18, p < .001, CI_{percent} = [20,30]$) and 33% (one proportion z-test: $z = 202.93, p < .001, CI_{percent} = [45,57]$) for UU and DD, respectively.

Table 1: Number of instances of response changes in Experiment 1.

	DD	DU	UD	UU
IM	41	4	22	33
LC	80	4	12	4
HC	32	19	11	37

Table 2: Pairwise comparisons of dwell times in seconds. The main diagonal contains mean dwell times SD in parentheses. The remaining cells contain pairwise t-tests with Bonferroni corrected α and Cohen's d in parentheses. Note, ns: $p > .008, * : p < .008$

	NM	IM	LC	HC
NM	**0.903 [1.74]**	-	-	-
IM	6.04 (1.6)*	**6.21 [4.51]**	-	-
LC	3.12 (0.95)*	3.26 (0.76)*	**3.28 [3.08]**	-
HC	6.22 (1.72)*	1.08 (0.29) ns	3.97 (1.01)*	**7.60 [5.23]**

Dwell times are longer for high conflict dilemmas. Since we had instituted an obligatory one-minute period for deliberation, the interpretability of the total time taken to deliberate as a measure of conflict reduces as people could have finished making a decision before the one minute was over. Nonetheless, the duration of time it took participants to reach a decision was predicted by the type of dilemma (RM one-way ANOVA: $F(1.31, 31.36) = 6.749, P < .001, \eta^2 = 0.219$), with only the difference between NM and HC conditions remaining significant after controlling for alpha ($t(24) = 2.91, p = .008, d = .84$).

We also computed the dwell times, defined as the mean of time taken to switch between two choices in a condition for each participant. Dwell time varied depending on the type of dilemma. (RM one-way ANOVA: $F(3,72) = 17.533, P < .001, \eta^2 = 0.422$), such that the dwell times were highest in HC, followed by IM, LC, and then NM. The pairwise comparisons are tabulated in 2. This suggests that overcoming momentary preferences in highly conflicting dilemmas may be difficult and hence, may require more time.

Moral vacillations correlate with the subjective experience of conflict. The literature frequently defines conflict as a lack of confidence in one's own judgment (Frey, Johnson, & De Neys, 2018; Mevel et al., 2015; Pennycook, Fugelsang, & Koehler, 2015). Our data indicate that such an operational definition may be reasonable. The subjective ratings of confidence in the final answer and conflict while deliberating were strongly correlated in our experiment ($r = -0.76, p < .01$). In addition, when there were more switches, participants reported more conflict ($r = .63, p < .01$) and less confidence ($r = -.55, p < .01$). The number of switches experienced was also correlated to participants' subjective impression of how difficult the problem was to answer ($r = 0.58, p < .01$) and if they believe their answer will change in the future ($r = 0.62, p < .01$). Furthermore, the type of dilemma predicted the confidence (RM one-way ANOVA: $F(3,72) = 49.24, p < .01, \eta^2 = 0.67$) and conflict (RM one-way ANOVA: $F(3,72) = 63.97, p < .01, \eta^2 = 0.73$).

Thus, overall Experiment 1 shows that a cohort-level operationalization of conflict does translate to conflict experienced by individuals, supporting prior work (Koenigs et al., 2007). The discrepancy in judgments at this level appears to predict how individuals feel reasoning about them. If individuals in a group diverge on their opinions on an issue, they may also experience shifts in transitory preferences within them.

Experiment 2

In Experiment 2, we tested if vacillations are predictive of when deontological and utilitarian principles do not converge on a choice, a definition of conflict given by Bago and De Neys (2019). We have pre-registered this experiment on osf.io.

Method

Participants We calculated a sample size of 22, based on a moderate effect of 0.55, at α = .05, and power .8. We collected data from 23 students, and 1 failed to follow the instructions. We analyzed the data of 22 participants (8 females, mean age = 20.3 years).

Experiment design Participants deliberated on 9 problems from 3 conditions: non-moral (NM), no-conflict, and conflict dilemmas. Only the NM problems were taken from Koenigs et al. (2007) while the rest were from Bago and De Neys (2019). Utilitarian actions in all moral trials (conflict and non-conflict) were impersonal (action included flipping a switch, pressing a button etc.). The effect of these actions was such that they killed a group of people as a side-effect of saving another group. In conflict dilemmas, the choice was between an action that saved many by deflecting harm on few individuals. In non-conflict trials, the action caused death of many by saving few. Here the supposedly deontic and utilitarian principles converge on saving many by letting the few die. We call this converging choice U in non-conflict trials for aiding the discussion. In the rating phase of a trial, we only included confidence and conflict scales. Otherwise, the experiment's trial structure remained unchanged.

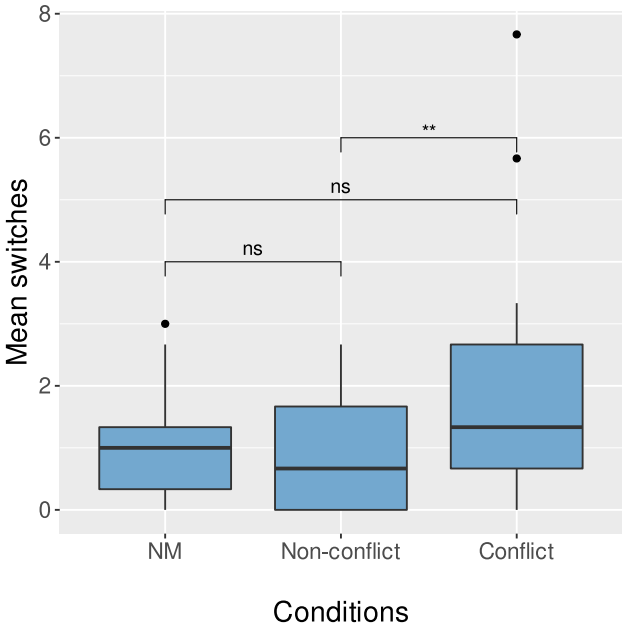

Figure 3: Box plot of mean switches across for participants in Experiment 2. Whiskers indicate the interquartile range. p values are Bonferroni corrected for multiple comparisons.

Table 3: Number of instances of response changes in Experiment 2.

	DD	DU	UD	UU
Non-conflict	4	8	4	50
Conflict	8	5	7	46

Results

People vacillate when ethical principles do not converge on one choice. Participants reported vacillations most frequently in the conflict [$M = 1.94, SD = 2.3$], followed by NM [$M = 1.11, SD = 1.53$] and then non-conflict cases [$M = 0.79, SD = 1.38$]. A repeated measures ANOVA determined that these means were significantly different, $F(1.18, 24.68) = 8.33, p = .006, \eta^2 = 0.28$. This difference remained significant even after removing outliers ($F(2,36) = 12.263, p < .001, \eta^2 = 0.41$). Pairwise comparisons revealed that only the mean switches between conflict and non-conflict cases were significant after adjusting for alpha ($t(20) = 3.87, p < .001, d = 0.8$). Together, these results indicate that people vacillate more when deontological and utilitarian principles do not cue the same choice. Furthermore, people equally switched in Experiment 1's HC condition and Experiment 2's conflict condition (Welch $t(36.2) = 0.18, p = .86, d = 0.05$).

Vacillations are more informative than response transitions. Just like Bago and De Neys (2019) report, participants' responses were largely utilitarian in non-conflict and conflict trials (87.9% and 77.27%, respectively). Participants also showed all 4 response transitions between the first and the last key pressed during deliberation like in Experiment 1 (see Table 3. But unlike in Experiment 1, most transitions were UU. Actions in moral dilemmas in Experiment 2 (conflict and non-conflict) were all impersonal, and they deflected harm coming to one group onto another. As a result, U was a lucrative choice for participants (see Greene et al. (2001, 2009); Moore et al. (2008)). Nonetheless, 35% of all UU trials had at least 2 switches and this proportion was significantly more than 0% (One proportion z test: $z = 40.39, p < .001, CI_{percent} = [26, 47]$). This suggests that even though U was expected to be a characteristic response to these dilemmas, participants' preferences momentarily shifted while deliberating.

Vacillations did not predict the reaction times or dwell times. The means of time taken to deliberate in a condition did not differ (RM one-way ANOVA: $F(1.38, 28.97) = 2.99, p = .08, \eta^2 = 0.12$). The dwell times, too, were not significantly different (RM one-way ANOVA: $F(1.56, 32.68) = 1.22, p = .33, \eta^2 = 0.05$). We discuss reasons for this null result in the Discussion section further below.

Conflict and confidence correlate with switches. When participants felt more conflicted during deliberation, they were underconfident in their final decision ($r = -.79, p < .001$). Further, when people vacillated more, they reported more conflict ($r = .41, p < .001$) and less confidence ($r = -.41, p < .001$). Like in Experiment 1, the type of dilemma predicted the conflict (RM one-way ANOVA: $F(2, 42) = 11.99, p < .001, \eta^2 = 0.34$) as well as the conflict ratings (RM one-way ANOVA: $F(2, 42) = 11.35, p < .01, \eta^2 = 0.35$).

Discussion

The primary contribution of this paper is methodological - we present a way to monitor vacillations during the deliberation process as a marker of cognitive conflict in moral decision-making, and show that these vacillations are in fact, reasonably well-correlated with classical measures of cognitive conflict (Cushman et al., 2006; Greene et al., 2001; Gürcay & Baron, 2017; Koenigs et al., 2007; Paxton et al., 2012). Process-tracking methods, such as mouse-tracking, have also been deployed in moral decision-making experiments recently (Gürcay & Baron, 2017; Koop, 2013). In such measurements, mouse trajectories demonstrate that changes in inclinations may be taking place in the same proportion toward either of the two options. However, it is impossible to state when in a given trial these adjustments occur while thinking, and thus, these assessments, like the classic instruments we discussed previously, also only measure the summary of conflict across a trial, not intra-trial fluctuations. Our method of measuring conflict enables experimenters to observe the entire time-course of a respondent's decision, opening up the possibility of more fine-grained analyses, including the use of gaze or neurophysiological markers for measuring the experience of conflict in future work.

Our experimental results empirically demonstrate some natural aspects of respondents' engagement with moral dilemmas. Across both our studies, we see that mental vacillations substantially validate the subjective experience of conflict. When people vacillate during a decision, they report feeling conflicted afterward. The more times they change their minds while making a decision, the less confident they become in their decisions, consistent with evidence-integration-to-threshold accounts of the choice process (Ratcliff & McKoon, 2008). In addition, our results add to the existing line of research questioning a strict adherence to dual-process models like in Greene et al. (2009) (see Koop (2013)). Respondents' interim preferences while reasoning uncovered multiple cases of them starting deliberations by considering the (ostensibly slow) utilitarian option and later switching to the (ostensibly fast) deontic option. A decision-making process like this is consistent with a single process account yielding both genres of decisions, possibly mediated by the number of choices under consideration (Srivastava & Vul, 2015).

From Experiment 2, we see that the experience of conflict may not always correspond to lengthened reaction times. One possible explanation for this, apart from the known poor predictivity of reaction times (Baron & Gürçay, 2017), could be the minimum deliberation period used in the design of our experiment, during which participants were instructed to keep thinking about the problem. Even if the choice was made earlier, participants had to wait until the end of this period to record their final verdict. Additionally, the stimuli used in Experiment 2 were considerably more homogeneous than the ones used in Experiment 1 (for a full list of dilemmas used in Experiment 2, see Study 1 from Bago and De Neys (2019)). As a result, Experiment 2 stimuli may have been more predictable than Experiment 1, resulting in more similar reaction times across conditions.

Thus, in sum, we have designed an experimental paradigm to measure how people navigate choices in dilemmas, and presented empirical evidence suggesting that this process is potentially more informative than can be captured by trial summary statistics like response times or cohort-disagreement levels. Importantly, our results also shows that currently *en vogue* theoretical accounts of moral reasoning do not explain the sequence of deliberations for moral dilemmas well. We expect the greater visibility into the deliberation process introduced by our paradigm to help improve such theoretical models.

Acknowledgments

We thank Prof Narayanan Srinivasan for providing insights throughout the project. RVS thanks Arjun Mitra, Divya Pathak, and Roshan Jayarajan for keeping her spirits up by

feeding her cauliflower sabji and organizing frequent movie nights. RVS also thanks Pepe for being herself.

References

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

Baron, J., & Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition*, *45*(4), 566–575.

Białek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology*, *28*(5), 631–639.

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

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social neuroscience*, *7*(3), 269–279.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*(12), 1082–1089.

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly journal of experimental psychology*, *71*(5), 1188–1208.

Greene, J. D. (2007). Why are vmpfc patients more utilitarian? a dual-process theory of moral judgment explains. *Trends in cognitive sciences*, *11*(8), 322–323.

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*(3), 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*(2), 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*(5537), 2105–2108.

Gürcay, B., & Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning*, *23*(1), 49–80.

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

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision making*, *8*(5), 527–539.

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227–237.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? individual differences in working memory capacity, executive control, and moral judgment. *Psychological science*, *19*(6), 549–557.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive science*, *36*(1), 163–177.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? a three-stage dual-process model of analytic engagement. *Cognitive psychology*, *80*, 34–72.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Srivastava, N., & Vul, E. (2015). Choosing fast and slow: explaining differences between hedonic and utilitarian choices. In *Cogsci.*