

NOTICE: This is an Author's Original Manuscript of an article whose final and definitive form, the Version of Record, has been published in **Social Neuroscience** [January 2014, copyright Taylor & Francis], available online at: [www.tandfonline.com/doi/abs/10.1080/17470919.2013.870091](http://www.tandfonline.com/doi/abs/10.1080/17470919.2013.870091)

Please cite as:

Patil I., Cogoni C., Zangrando N., Chittaro L., Silani G. Affective basis of Judgment-Behavior Discrepancy in Virtual Experiences of Moral Dilemmas, *Social Neuroscience*, 9(1), 2014, pp. 989–993

## Affective basis of Judgment-Behavior Discrepancy in Virtual Experiences of Moral Dilemmas

Indrajeet Patil<sup>1</sup>, Carlotta Cogoni<sup>1</sup>, Nicola Zangrando<sup>2</sup>, Luca Chittaro<sup>2</sup>, Giorgia Silani<sup>1</sup>

<sup>1</sup>Scuola Internazionale Superiore di Studi Avanzati, Neuroscience Sector, Trieste, Italy

<sup>2</sup>Human-Computer Interaction Lab. Dept. of Mathematics and Computer Science, University of Udine, Udine, Italy

**Keywords:** Moral judgment, virtual reality, moral dilemmas, decision making, skin conductance

### Abstract

Although research in moral psychology in the last decade has relied heavily on hypothetical moral dilemmas and has been effective in understanding moral judgment, how these judgments translate into behaviors remains a largely unexplored issue due to the harmful nature of the acts involved. To study this link, we follow a new approach based on a desktop virtual reality environment. In our within-subjects experiment, participants exhibited an order-dependent judgment-behavior discrepancy across temporally-separated sessions, with many of them *behaving* in utilitarian manner in virtual reality dilemmas despite their non-utilitarian *judgments* for the same dilemmas in textual descriptions. This change in decisions reflected in the autonomic arousal of participants, with dilemmas in virtual reality being perceived more emotionally arousing than the ones in text, after controlling for general differences between the two presentation modalities (virtual reality vs. text). This suggests that moral decision-making in hypothetical moral dilemmas is susceptible to contextual saliency of the presentation of these dilemmas.

## Introduction

Hypothetical moral dilemmas have been a useful tool in understanding moral decision-making, especially in elucidating the affective and cognitive foundations of moral judgment (Cushman & Greene, 2011; Waldmann, Nagel, & Wiegmann, 2012; Christensen & Gomila, 2012). A typical example of such dilemmas is the trolley dilemma (Thomson, 1985):

“A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Is it appropriate for you to turn the trolley in order to save five people at the expense of one?”

Psychological investigation of people’s moral judgments has relied on the way people respond to these dilemmas. Affirmative response to this dilemma is said to be utilitarian, since it agrees with John Stuart Mill’s utilitarianism which argues that those moral actions are good which maximize the wellbeing of the maximum number of agents involved in the situation (Mill, 1998). On the other hand, negative response is said to be non-utilitarian or deontological, referring to Kantian deontology which evaluates the moral status of an action based not on its consequences but based on the features of the act itself, relative to the moral rules regarding rights and duties of the agents involved in the situation (Kant, 1785/2005). Moral psychologists are concerned with the cognitive processes mediating these responses and the appraisal mechanisms that motivate these processes. The aim of studying moral judgments has primarily been about understanding how people distinguish between right and wrong, but the issue of how these moral judgments translate into behavior remains still unclear: would someone who judges switching the trolley as morally appropriate actually resort to this course of action when the full repertoire of contextual features come into play?

A recent study (Tassy, Oullier, Mancini, & Wicker, 2013) showed that there is a discrepancy between judgments people make and the choice of action they endorse in moral dilemmas. People were more likely to respond in a utilitarian manner to the question “Would you do....?” (which was a probe question for choice of moral action) than to the question “Is it acceptable to....?” (which was a probe question for moral judgment). Or, in other words, people showed a tendency to choose actions they judged to be wrong. Another study (Tassy, et al., 2012) showed that objective evaluative judgment and subjective action choice in moral dilemmas about harm might rely on distinct cognitive processes. These studies are suggestive of the hypothesis that the selection of moral behavior and endorsement of an abstract moral judgment in a moral dilemma are mediated by partially distinct neural and psychological processes. But shortcoming of these studies was that they relied completely on self-report questionnaire data and thus could not ascertain if what participants considered their choice of moral action on paper would indeed be their actual action if they were to face the same situation in more salient situations.

In a more realistic setting, a recent study (FeldmanHall et al., 2012) used a pain-versus-gain paradigm to show that in the face of contextually salient motivational cues (like monetary gain) people were ready to let others get physically hurt, which contrasts starkly with the previous research showing that aversion to harming others is one of the most deeply-ingrained of moral intuitions (Cushman, Young, & Hauser, 2006; Haidt, 2007). They also showed that the behavior of participants in real life increasingly deviated away from the judgment they made as the presentation of moral situations became increasingly contextually impoverished. As the experimental setup became progressively estranged from real-life setting, people had to rely more and more on the mental simulation of the situation and had to make decisions without the context-dependent knowledge which would otherwise have been available to them in the real-life setting (Gilbert and Wilson, 2007). Qualitatively, the pain-versus-gain paradigm

differs from the trolley dilemmas, the former pitting self-benefit against welfare of others while the latter pitting welfare of two sets of strangers. Nevertheless, it is legitimate to assume that the same concerns apply to hypothetical moral dilemmas, which are usually presented in text format with all the non-essential contextual information stripped away (Christensen & Gomila, 2012), leading participants to rely more on the abbreviated, unrepresentative, and decontextualized mental simulations of the considered situations (Gilbert & Wilson, 2007).

The advantage of relying on text- or graphic-based questionnaires is its great experimental controllability, but the downside is that it greatly simplifies the issue at hand by removing all the non-essential contextual features of the dilemmas, raising issue of generalizability of the obtained results. The impoverished and unrealistic experimental stimuli limit participant's engagement and thus cannot affect participants with the targeted experimental manipulation. On the other hand, more elaborate experimental designs engender increases in cost and may cause loss in experimental control. This trade-off has been a hallmark feature of research in experimental social psychology (Blascovich et al., 2002).

Moral dilemmas are especially difficult to create realistically in laboratory settings because of the ethical problems associated with violent and harmful experimental situations. Virtual reality (VR) helps to take a step forward in studying such situations in a more ecologically valid manner. A number of studies have investigated behavior in situations containing elements of violence rendered using VR and show that people respond realistically to such situations (for a review, see Rovira et al., 2009). This is an indication that VR can provide a good middle ground in terms of experimental realism and control to study social situations involving physical harm.

To the best of our knowledge, only one study (Navarrete, McDonald, Mott, & Asher, 2012) used contextually rich, immersive VR reconstructions of trolley dilemmas to address the relationship between moral judgment and moral behavior. They compared the behavior (proportion of utilitarian decisions taken) of participants in VR with judgments of participants from previous studies which relied on the text-based scenarios (Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2007; Valdesolo & DeSteno, 2006). They found that the behavior of participants in VR (proportion of utilitarian decisions: 88.5-90.5%) was congruent with the judgment-data from previous research, which led to the conclusion that there was not a significant difference between judgment and behavior in situations where an individual is harmed for the greater good, at least so far as the decision-making goes in situations involving salient sensory input in the absence of real-life consequences. One shortcoming of the study is that the decisions taken by participants were not compared with their own judgments but with the judgments of people who participated in previous experiments, making it a between-subject design. As a result, the experiment could not address the relation between judgments and behavior *for the same individual*.

Our study tries to address this issue and differs from Navarrete *et al.* (2012) in some crucial aspects: (a) we use a within-subject design, as opposed to between-subject design; (b) we use desktop VR hardware (a common LCD monitor), as opposed to immersive VR hardware (Head-Mounted Display); (c) we use four different moral dilemmas involving harm, as opposed to just one; (d) we focus just on action conditions, instead of both action and omission conditions; (e) we record skin conductance in order to characterize physiological responses associated with moral judgments and moral behavior (after controlling for the general differences in VR and text scenarios).

Contextual saliency refers to the ability of the experimental stimuli to supply contextual information which is available in real-life situations, rather than being limited to just necessary and sufficient amount of information. In the current study, we observed differences in the contextual saliency between the

two modes of presentation of the moral dilemmas and a resultant differential capacity of these modes to engage affective processing. We therefore expected that people would respond differently in judging text dilemmas (which are limited in emotional engagement) as compared to acting in VR situations (which are more life-like and hence could be more emotionally arousing). We also expected any difference between the judgments people make in text dilemmas and their actions in VR dilemmas sessions to be due to the putative differential propensity of the two modes of presentation of the moral dilemmas to engage emotions and would thus reflect in the skin conductance data (Dawson, Schell, & Filion, 2007), because it would index the ability of the presentation modality to engage emotional processing. Although it remains controversial if emotions are necessary and/or sufficient for moral judgments (Huebner, Dwyer, & Hauser, 2009), it is well-established that emotions either co-occur or ensue from moral judgments (Avramova & Inbar, 2013). Thus, our first prediction was that the observed judgment-behavior discrepancy would have an affective basis, as indexed by SCR activity.

Further, participants could show judgment-behavior discrepancy in two ways: by making either more or less number of utilitarian decisions in VR as compared to text session. To predict in which way emotions would influence this discrepancy, we relied on Greene's dual process model (Greene et al., 2001, 2004, 2008). This model posits two types of computational processes to explain the observed behavioral pattern in moral dilemmas: intuitive emotional processes that automatically evaluate the stimulus on the moral dimensions of right/wrong to come up with a judgment and support non-utilitarian decision and controlled reasoning processes that rely on deductive reasoning and cost-benefit analysis to arrive at a judgment and support utilitarian decision. Additionally, these two processes contribute to the final decision differently, depending upon the nature of the dilemma and its ability to engage emotional processing. For example, personal moral dilemmas (e.g. footbridge dilemma in which the agent in the scenario can save maximum number of lives by pushing a large man standing next to him/her off of a footbridge) are found to be more emotionally engaging than the impersonal moral dilemmas, as shown by both neuroimaging data (Greene et al., 2001, 2004) and skin conductance activity (Moretto, Ladavas, Mattioli, & di Pellegrino, 2009), and elicit more non-utilitarian judgments. In the current study, we focused exclusively on impersonal moral dilemmas. Since we expected VR dilemmas to engage emotional processing more than their textual counterparts, we predicted that a smaller proportion of utilitarian responses will be observed for VR than text dilemmas.

## **Methods**

### **Participants**

In this study, we recruited 40 healthy participants (24 female, 16 male) between ages of 18 and 28 ( $M=22.8$ ,  $SD=2.6$  years). Each participant was paid €15 as a compensation for his/her travel and time. All participants were native Italian speakers and had normal or corrected-to-normal vision. Except for one participant, all of them were right-handed. The study was approved by the ethics committee of the hospital "Santa Maria della Misericordia" (Udine, Italy). The experiment was carried out at the Human-Computer Interaction Laboratory (HCI Lab), Department of Mathematics and Computer Science (University of Udine, Italy).

### **Experimental stimuli**

In each (text/VR) session, subjects faced 8 moral dilemmas, divided equally into 4 experimental conditions and 4 control conditions, for a total of 16 dilemmas in the two sessions. Control conditions controlled for the general differences across text and VR presentation modalities: length of the trial for a given session, attention deployment, visual complexity of the stimuli, etc. Experimental condition

*dilemmas* pitted welfare of one individual against welfare of 2 or 5 individuals, while the control condition scenarios pitted welfare of one individual against damage to empty boxes and thus posed no dilemma between different moral ideologies. Hence, the experimental conditions specifically tapped into the decision-making in dilemmatic situations, while this was not the case for control conditions. For example, in the train dilemma, a train was directed towards 2 or 5 humans walking on the track and participants had to switch the train onto an alternative track if they wanted to save this group of people by sacrificing a single human walking on the alternative track. In the control condition version of the same dilemma, the train was directed towards one human and participants could divert it on the alternative track on which there were just empty boxes and no humans. To summarize, control conditions were used not only to control for differences in the presentation modalities, but also to study the emotional response which was specific to decision-making in moral *dilemmas*. In any session, the experimental and control conditions were presented randomly. We included variation in number of victims in the dilemmas so as to avoid the dilemmas becoming too predictable, which could have resulted in subjects premeditating the response even before they read or saw the dilemma. It needs to be mentioned that though the number of victims in each dilemma was randomized, the total number of victims for each session was same for both text and VR sessions and for all participants. There were always two experimental dilemmas with two number of victims, while the other two experimental dilemmas with five number of victims. All the dilemmas used in this study were impersonal moral dilemmas (Greene et al., 2001, 2004).

The virtual environments were implemented using the C# programming language and the Unity3D game engine, see Figure 1 for a film-strip of the VR version of the train dilemma and Supplementary Information for the video clips of experimental VR dilemmas and description of text dilemmas. For each VR dilemma, a textual version of the same dilemma was written for use in the text session. One aspect of VR scenarios that needs to be stressed here is that participants had to witness highly salient consequences of their actions, e.g. in the train dilemma, participants saw virtual agents getting hit and run over by the train and their bleeding corpses lying on the track afterwards (Figure 1C).

(Figure 1 about here)

## Procedure

We followed a within-subjects design, whereby each participant had to face the same dilemmas in the text session that employed textual descriptions and in a VR session that presented the dilemmas as interactive virtual experiences. The order in which participants performed the task was counterbalanced: half participants performed the text session first, the other half the VR session first. Participants were randomly assigned to a particular order. Participants performed the second session after a variable number of days in order to avoid spillover effects of decisions made in the previous session. The average interval between two sessions was 102 days ( $SD= 53$ ) and did not differ for the two orders ( $t(32) = -1.028, p = 0.31$ ). Large variation in the interval between two sessions was due to the practical concern of availability of different participants.

### *Behavioral task*

After the participants arrived in the laboratory, they were told that the study concerned decision-making in social settings. To address concerns about social desirability bias, the computer console for the participants was separated from experimenters using curtains. All scenarios in the experiment were displayed on a 30-in. LCD computer monitor with speakers. Subjects were seated in a semi-dark room at

a viewing distance of 100 cm from the screen. Responses were recorded using a Nintendo Nunchuck joystick.

Before beginning with the experiment, participants were familiarized with the virtual experiences and the text scenarios, using training sessions. For the text scenarios, participants were trained to use joystick in an example situation containing non-meaningful verbal text, and were instructed about how to use the response button in order to change the screen and select the response. For the VR training sessions, we used four parts of tutorial environments, each of them introducing the virtual environment which would later be presented in experimental trials. Participants were instructed about the meaning of different visual signals present in all the scenarios and how to use the response button in order to make a choice. For example, in the tutorial for the train dilemma (see Figure 1), they were explained that the presence of a green or red light indicates the track available for the train to continue on (green: pass, red: no pass); while a yellow-black striped line marked the point till which it was possible for them to make a choice by switching the red and green lights via the joystick. After the training session, all participants were asked to operate these tutorials without experimenter's help. After making sure that they understood the procedure, they were presented with the actual experimental stimuli.

In the text session, the trial started with a period of silence for 1 minute with fixation cross on the screen and then the text of the scenario appeared. The dilemma description remained on the screen for the rest of the trial. A second press on the same button presented the question asking for the judgment from the participant ("Is it appropriate for you to [nature of the action]?") and lasted for 12 seconds (see Figure 2). By default, the option highlighted was non-utilitarian (*no*) and participants had to press again the same button to change it to utilitarian (*yes*) if they wanted to endorse a utilitarian outcome. Once the response was made, it could not be changed. After the response, the text faded and was replaced by fixation cross.

In the VR session, participants were presented with the VR versions of the dilemmas on the same computer screen and asked to respond with the same button of the joystick used in the text session. The trial started with a period of silence for 1 minute with fixation cross on the screen and then the virtual scenarios appeared. Each experimental and control scenario lasted for 18 seconds and participants had to respond within 10 seconds from the beginning of the scenario (see Figure 2), after which it was not possible for them to make a choice. Participants could keep track of the time limit by a pre-specified event (as explained during the familiarization phase using training environment), e.g. in the train dilemma, they had to respond before the train crossed the yellow-black striped line (indicated with red circle in Figure 1). In all the VR scenarios, the threat was by default directed towards the maximum number of virtual humans (2 or 5), e.g. in the train dilemma, the signal was green for the track on which two/five virtual humans were walking (see Figure 1). Thus, participants had to press the button on the joystick to change the signal from green to red for the track on which there were five virtual humans, which automatically gave a green signal for the train to pass on the alternative track on which there was one virtual human walking (of course, only if they wanted to achieve a utilitarian outcome in this situation).

(Figure 2 about here)

In the post-experiment debriefing, we explicitly asked participants about any difficulties or technical snags they faced during the session. None of them mentioned of failure to respond due to unavailability of sufficient time or having pressed a wrong button in confusion. This gives us more confidence to conclude that participants' responses were a result of their true moral choices rather than failure to respond in time or confusion.

### *Electrodermal activity recording*

While participants performed the task, their electrodermal responses were monitored as an index of arousal and somatic state activation (Dawson, Schell, & Fillion, 2007). For each participant, prewired Ag/AgCl electrodes were attached to the volar surfaces of the medial phalanges of the middle and index fingers of the non-dominant hand, which left the dominant hand free for the behavioral task. The electrode pair was excited with a constant voltage of 0.5 V and conductance was recorded using a DC amplifier with a low-pass filter set at 64 Hz and a sample frequency of 256 Hz. As subjects performed the task seated in front of the computer, SCR was collected continuously using a Thought Technology Procomp Infiniti encoder and stored for off-line analysis on a second PC. Each trial (experimental or control) was preceded by a 1-minute baseline recording period during which participants rested in the chair, while their SCR activity returned to baseline. Presentation of each dilemma was synchronized with the sampling computer to the nearest millisecond, and each button press by the subjects left a bookmark on the SCR recording. Subjects were asked to stay as still as possible in order to avoid any introduction of noise in the data due to hand movements. SCR activity was not recorded for the familiarization/training phase of the VR session.

### *Questionnaire*

At the end of the experiment, a recall questionnaire asked participants about how much could they remember about their decisions in the previous session. Participants had to qualitatively describe what they could recall, instead of reporting it on a scale. This data was later quantified by two referees blind to the purpose of the experiment. The responses were categorized into a 5-point Likert scale ranging from -2 (can't remember anything) to 0 (remember something) to 2 (remember everything).

## **Results**

### *Behavior*

For each participant, we computed the proportion of utilitarian decisions by calculating the number of experimental dilemmas in which a utilitarian decision was taken divided by the total number of dilemmas (which was four for all the participants), e.g. if the participant made utilitarian decision for 2 out of 4 dilemmas, the score was 0.5 for that participant for that particular session. Control condition data was not analyzed for this dependent variable because it did not pose any dilemma. Indeed, all the participants saved the virtual human over the empty boxes in the control condition. The proportions of utilitarian decisions were computed for each participant for each session separately. The average of these proportions was computed across subjects for each session and compared between the two sessions to check for the discrepancy between judgment and behavior. The data was analyzed for 34 participants for the reasons described in the Electrodermal Activity results section. Statistical Analysis was carried out using SPSS 11 Software (SPSS Inc., Chertsey UK).

In the text session, the average proportion of judgments endorsing utilitarian outcome was 0.76 ( $SD = 0.32$ ); while for the VR session, the average proportion of actions that endorsed utilitarian outcome was 0.95 ( $SD = 0.14$ ) (see Figure 3).

(Figure 3 about here)

The distribution of utilitarian proportions did not follow normal distribution for both sessions (Shapiro-Wilk test:  $p_s < 0.01$ ). Thus, we compared mean ranks of these proportions from two sessions using related-samples Wilcoxon signed rank test and found a significant difference:  $Z = -3.35$ ,  $p = 0.001$  (two-tailed). Therefore, the difference between the proportions of utilitarian decisions taken in the two

sessions was significant, with people acting in more utilitarian manner in VR session than they judged in text session. Unexpectedly, this effect was dependent on the order (see Table 1) in which participants carried out the sessions (text-first [ $n = 19$ ]:  $Z = -2.98$ ,  $p = 0.003$ ; VR-first [ $n = 15$ ]:  $Z = -1.52$ ,  $p = 0.13$ ). To further investigate the order effects, we computed a discrepancy index for each participant as the difference between proportion of utilitarian decisions taken in VR and text session. One-sample Wilcoxon signed rank test (two-tailed) showed that median of discrepancy index was significantly different from zero for text-first order ( $Z = 2.98$ ,  $p = 0.003$ ), but not for VR-first order ( $Z = 1.86$ ,  $p = 0.063$ ). Additionally, chi-square test for independence with order of sessions (dummy coded 0: text-first and 1: VR-first) and judgment-behavior discrepancy (dummy coded as 0: no discrepancy and 1: exhibited discrepancy) as numerical variables gave a marginally significant result ( $\chi^2(1) = 3.348$ ,  $p = 0.06$ ,  $\phi = -0.323$ ). In other words, ratio of participants who exhibited to who did not exhibit judgment-behavior discrepancy was dependent on the order in which participants faced the sessions.

(Table 1 about here)

Hence, participants behaved in more utilitarian manner in the VR session as compared to the text session, but the effect was strongest when they faced text first. Our prediction about inconsistency between judgments and actions was thus borne out by these results.

#### *Response Time*

Since the non-utilitarian response was the default choice, subjects did not have to press any button to take a non-utilitarian decision, which meant that we could not collect data regarding response time for these decisions. The response time data could only be recorded for the utilitarian responses. In the text session, the reaction time for the utilitarian decision was taken to be the time difference between the appearance of the question on the screen and participant's response, while in the VR session, it was the interval between the time at which the virtual scenarios started and the time at which response was given. Since the two sessions featured different presentation modalities with different cognitive requirements, one requiring language comprehension while the other requiring visual perception of the situation, the elicited response times were not directly comparable. We harnessed control conditions from the respective sessions for this purpose. We computed a response time (RT) index for each subject by computing the difference between response time for utilitarian decisions in experimental condition and control conditions (in control condition, utilitarian decision was saving virtual human over empty boxes), denoted by RT (uti-con). Two subjects did not take any utilitarian decision in experimental condition of one of the sessions, so the sample size for this analysis was 32. The distribution of response time indices for both sessions followed normal distributions (Shapiro-Wilk test:  $ps > 0.2$ ).

Paired-samples  $t$ -test showed that the difference in RT (uti-con) for VR ( $M = 0.72s$ ,  $SD = 1.50$ ) and text ( $M = 0.21s$ ,  $SD = 1.33$ ) dilemmas was not significant ( $t(31) = 1.547$ ,  $p = 0.132$ ). This result was independent of the order in which sessions were performed by participants: for text-first,  $t(16) = 1.027$ ,  $p = 0.32$ ; while for VR-first,  $t(14) = 1.240$ ,  $p = 0.24$ . Thus, controlling for the differences in the presentation of the dilemmas in two sessions, subjects who endorsed utilitarian options did not differ in the amount of time they required to respond in text and in VR.

#### *Electrodermal Activity*

For the VR session, skin conductance data was analyzed for the entire length of the trial (which lasted for 18 seconds since the beginning of the scenario). For the text session, the skin conductance data was analyzed for a window of  $[-53, +5]$  seconds, centered on the appearance of the question. This particular window was selected because 53 seconds was the average time required by participants to read the description of the dilemma, after which the question appeared, and 5 seconds was the average response time. These two time segments were comparable across two sessions, since they included the



time period in which participants comprehended and contemplated over available options, formed a preference, and executed the response. But there was one difference between the two SCR windows analyzed for two sessions: only the window in VR session included witnessing distressing consequences for 8 seconds, while no such condition (e.g. reading the consequences) was present for the window in text session (See Figure 2).

Skin conductance data of three participants was removed for being outliers (2 SD away from mean value). Additionally, skin conductance data could not be recorded from one participant during the VR session and from two participants during the text session due to a temporary malfunction in the recording device. Skin conductance data were thus analyzed for both sessions for 34 participants. For the analysis of skin conductance data, we used Ledalab software (<http://www.ledalab.de/>) on Matlab (v 7.12.0) platform. Ledalab performs a continuous decomposition analysis to separate the phasic and tonic components. We defined SCR as the maximal conductance increase obtained in the SCR window of 1s to 3 s relative to the onset of the analysis window. To avoid false positive signals, the minimum threshold for SCR to be valid was 0.02  $\mu$ S. We then computed SCRs for all the trials as “area under curve” (Moretto et al., 2009). The “area under curve” measurement is the time integral of phasic driver within response window with straight line between the end points of the window taken as baseline rather than zero. The area is expressed in terms of amplitude units (microsiemens,  $\mu$ S) per time interval (sec). Area bounded by the curve thus captures both the amplitude and temporal characteristics of an SCR and therefore is a more valid indicator than either aspect alone (Figner & Murphy, 2010). All SCRs were square-root-transformed to attain statistical normality (Shapiro-Wilk test:  $p > 0.2$ ).

We carried out repeated-measures ANOVA on SCRs with session (text, VR) and condition (experimental, control) as within-subjects factors (see Figure 4).

(Figure 4 about here)

The ANOVA revealed a main effect of session ( $F(1,33) = 65.15, p < 0.001, \eta^2 = 0.67$ ) which was independent of the order of sessions (for order VR-first:  $F(1,14) = 26.45, p < 0.001$  and for order text-first:  $F(1,18) = 41.07, p < 0.001$ ). Thus, the moral dilemmas were more emotionally arousing when presented in VR than when presented in textual format, irrespective of the condition. The ANOVA also revealed a main effect of condition ( $F(1,33) = 11.28, p = 0.002, \eta^2 = 0.26$ ), which meant that the moral dilemmas in experimental conditions were perceived to be more emotionally arousing than the control conditions. This effect was independent of the order; for order VR-first:  $F(1,14) = 7.65, p = 0.016, \eta^2 = 0.37$  and for order text-first:  $F(1,17) = 5.44, p = 0.032, \eta^2 = 0.24$ . Post-hoc t-tests revealed that the experimental conditions were more arousing than control conditions only for VR session:  $t(33) = 3.68, p = 0.001, \text{Cohen's } d = 1.28$  (for order VR-first:  $t(14) = 3.58, p = 0.003, \text{Cohen's } d = 1.91$  and for order text-first:  $t(18) = 2.28, p = 0.036, \text{Cohen's } d = 1.07$ ). But experimental conditions were no more arousing than the control condition for the text session:  $t(33) = 0.67, p = 0.51$  (for order VR-first:  $t(14) = -.05, p = 0.96$  and for order text-first:  $t(18) = 1.40, p = 0.18$ ). This is consistent with our hypothesis: because of the contextually impoverished nature of the text dilemmas, the experimental conditions failed to push the emotional buttons and, thus, making decisions in experimental conditions was no more arousing than in control conditions. But this was not the case for (possibly ecologically more valid) VR dilemmas; for VR dilemmas, making choices in experimental dilemmas was more emotionally arousing than in control conditions. Finally, we observed a robust interaction effect between session and condition:  $F(1,33) = 12.72, p = 0.001, \eta^2 = 0.28$ . This interaction effect was independent of the order in which participants faced the two sessions (for order VR-first:  $F(1,14) = 10.28, p = 0.007$ , while for order text-first:  $F(1,18) = 4.31, p = 0.052$ ). Thus, taking decisions in experimental moral dilemmas was more emotionally arousing in the VR session as compared to the text session, after controlling for the differences in these two presentation modalities.

In the preceding analysis, we have not analyzed the data for utilitarian and non-utilitarian decisions separately and thus it can be argued that the SCRs for non-utilitarian decisions might have confounded the results. Thus, we performed another analysis using only the experimental conditions (from both sessions) in which utilitarian decisions were taken and removed the trials in which non-utilitarian decisions were taken. This led to reduction in the sample size, since three subjects had not taken any utilitarian decision in one of the sessions. All the previous results were replicated in this analysis; main effect of session ( $F(1,29) = 73.74, p < 0.001, \eta^2 = 0.73$ ), main effect of condition ( $F(1,29) = 9.20, p = 0.005, \eta^2 = 0.25$ ), and interaction ( $F(1,29) = 11.50, p = 0.002, \eta^2 = 0.29$ ). Additionally, these results were true for both order VR-first (session:  $p < 0.0001$ , condition:  $p = 0.03$ , session by condition:  $p = 0.05$ ) and text-first (session:  $p < 0.0001$ , condition:  $p = 0.016$ , session by condition:  $p = 0.007$ ). A similar ANOVA model could not be constructed for non-utilitarian decisions because there was not enough SCR data for VR session; non-utilitarian decision was taken only in 5% of experimental trials.

### *Questionnaire*

Recall questionnaire data showed that participants could recall ( $M = 0.77, SD = 0.77$ ) their decisions from the previous sessions fairly well (one-sample Wilcoxon signed rank test:  $Z = 3.758, p < 0.000$ ) in both sessions orders (VR-first:  $p = 0.014$ , for text-first:  $p = 0.006$ ). This could potentially have confounded the main behavioral result: participants who could remember better would show less discrepancy to remain consistent as compared to participants who could not. This explanation seems unlikely because there was no significant correlation between recall and discrepancy index ( $\rho(32) = 0.13, p = 0.50$ ) for both session orders (VR-first:  $p = 0.77$ , text-first:  $p = 0.36$ ). Additionally, there was no correlation between session gap (in number of days) and discrepancy index ( $\rho(32) = 0.06, p = 0.79$ ; VR-first:  $p = 0.34$ , text-first:  $p = 0.75$ ) or recall ( $\rho(32) = -0.07, p = 0.71$ ; VR-first:  $p = 0.83$ , text-first:  $p = 0.96$ ).

## **Discussion**

In this experiment, we showed that a change in contextual saliency in the presentation of dilemmas led to differences in autonomic arousal and endorsement of utilitarian principle in hypothetical moral dilemmas, but these differences were dependent on the order in which dilemmas were presented. In the following sections, we discuss various aspects of the observed results.

### *Judgment-behavior discrepancy and order effects*

Moral dilemmas create a decision space which pits the utilitarian rule dictating preference for lives of many over lives of few against the deontological rule prohibiting actively or passively killing innocent few to save many. We predicted that the choice people would make in this dilemma would depend on the contextual saliency of the presentation of the dilemma; in the contextually more salient presentation of the dilemmas, people would have to rely less on the abridged and unrepresentative mental simulations of the dilemma (Gilbert & Wilson, 2007). As per this prediction, we found that participants exhibited judgment-behavior discrepancy by endorsing utilitarian principle more in contextually salient VR dilemmas as compared to the same dilemmas presented using relatively arid text format. To put it differently, even though some of the participants judged sacrificing one to save many as morally inappropriate in text dilemmas, when full spectrum of contextual cues was provided using VR environment, they resorted to act in utilitarian fashion contradicting their earlier endorsement of deontological principle.

Interestingly, these results were dependent on the order of sessions (see Table 1) such that only the participants who completed the text dilemmas first and then faced VR dilemmas exhibited the judgment-behavior discrepancy. In the VR-first order, participants were consistent in the moral principle

they endorsed. In other words, participants exhibited more discrepancy (or less equivalency) in endorsing utilitarian principle across text and VR dilemmas only when the text dilemmas were presented first.

These results raise a number of questions: Why are the same dilemmas treated differently when presented in two different modalities? Why do people show judgment-behavior discrepancy in a particular direction? Why is this discrepancy dependent on the order in which the dilemmas are presented? We posit that answers to all these questions are connected via a common element of emotional processes to which we turn next.

### *Role of emotions in judgment-behavior discrepancy*

We had predicted that the superior contextual saliency of the VR environments would elicit higher emotional arousal in participants. Accordingly, we found that VR trials were indeed emotionally more arousing than text trials. We found that the experimental conditions (containing dilemmas) were emotionally more arousing than the control conditions (no dilemmas), but post-hoc comparisons showed that this was true only for VR dilemmas. Thus, the text dilemmas were no more arousing than the control conditions without any dilemmas as a result of reliance on abstract, abridged, mental simulations of the text scenarios that left participants affectively cold (Gilbert & Wilson, 2007). But the heightened skin conductance activity in VR with respect to text dilemmas could have been due to the general differences in the two presentation modalities, thus we checked if VR dilemmas were more emotionally arousing than the text dilemmas controlling for these differences using control conditions from the respective sessions. Control conditions were matched with the experimental conditions in a given presentation modality for most of the cognitively important aspects of the stimulus that can elicit SCR activity, e.g. length of the trial, cognitive load, stimulus novelty, surprise, etc. (Dawson, Schell, & Fillion, 2007), except for the dilemmatic aspect. Thus, we interpreted any difference in skin conductance activity between the two conditions as a gauge of emotional arousal in decision-making in dilemmatic situations. This dilemmatic emotional arousal was significantly higher for VR dilemmas (VR[experimental-control]) than text dilemmas (Text[experimental-control]):  $t(33) = 3.57, p = 0.001$ , Cohen's  $d = 1.24$ . We maintain that the observed judgment-behavior discrepancy was a direct result of differential ability of these two presentation modalities to effectively engage affective processing.

Based on Greene's dual process model (Greene et al. 2001, 2004, 2008), we had predicted that this increase in affective arousal would be associated with decrease in proportion of utilitarian responses. But we found exactly the opposite result; higher emotional processing led to more utilitarian responding. Previous studies using either just text dilemmas (for a review, see Greene, 2009) or just virtual dilemmas (Navarrete et al., 2012) overwhelmingly support predictions of the dual process model: increase in emotional processing/arousal was associated with lower likelihood of a utilitarian response and higher likelihood of a non-utilitarian response. This is the first study involving both text and VR dilemmas investigating the role of emotion in judgment as well as behavior. Additionally, we did not have enough skin conductance data for non-utilitarian responses in VR session (only 5% trials) to conduct any meaningful statistical analysis on skin conductance data for non-utilitarian choices. Thus, implications of results of this study for Greene's dual process model are unclear.

One possible explanation for our results in this framework is the following. The dual process model posits that intuitive emotional processes support non-utilitarian decisions, while deliberative reasoning processes support utilitarian decisions. Although these processes agree most of the time with the responses they come up with (e.g. a negative response to the question "Is it morally appropriate to torture people for fun?"), sometimes they can conflict (e.g. in the trolley dilemma, where there is an intense pang of emotions at the prospect of sacrificing someone, while the cost-benefit analysis is

demanding it). This cognitive conflict is detected by anterior cingulate cortex (ACC), resolved with the help of dorsolateral prefrontal cortex (dlPFC) (Greene et al., 2004). But it has been shown that cognitive conflict resolution is accompanied by autonomic arousal (Kobayashi, Yoshino, Takahashi, & Nomura, 2007). Thus, it is possible that the association between increase in utilitarian responding in VR dilemmas and heightened autonomic arousal in VR with respect to text actually represent the greater demand for cognitive conflict resolution in VR dilemmas, which are perceived to be more difficult than the text dilemmas (as shown by both objective SCR data) and might elicit stronger cognitive conflict. This explanation makes a testable prediction that considering VR dilemmas will lead to higher activity in ACC and dlPFC, as compared to text dilemmas. Future studies should investigate if this is indeed the case.

That said, we think that our results fit with the predictions of Cushman's version of the dual-process model (Cushman, 2013). In this model, the two processes that compete with and (sometimes) complement each other depend upon different value-representation targets. One process assigns value directly to the actions (e.g. negative value to the representation of pushing someone off the bridge or positive value to the representation of giving food to a beggar), while the other process assigns value to the outcome (e.g. negative value to the representation of physical harm to the person pushed off the bridge or positive value to the representation of content face of a beggar). Given that deontological decisions focus more on the nature of actions, while utilitarian decisions focus more on consequences of an action, it follows that this model associates utilitarian decisions with a cognitive process dependent on outcome-based value representations while deontological decisions with a cognitive process dependent on action-based value representations. The model contends that both processes have some affective content and are responsible for motivating the respectively endorsed behavioral responses.

In the light of this model, we hypothesize that in VR participants could have been more sensitive to outcomes because they witnessed distressing consequences (gory deaths of virtual humans) of their actions and emotions motivated them to act in order to minimize the distress by choosing the best of two emotionally aversive options in which either one or numerous (2 or 5) deaths occur. We posit that outcome-based value representation for not acting to save numerous innocent individuals from harm and seeing them die has more negative value than choosing to act and see the death of one innocent individual. With textual descriptions, people need to rely more on mental simulation of the situation and, given the paucity of the contextual features (audio and visual representations) which are accessible to people during such mental simulation, they cannot access context-dependent knowledge important for decisions that would otherwise be accessible to them in a more ecologically valid situation (Gilbert & Wilson, 2007). As a result, they tend to focus more on their basic duty of not being responsible for the death of any individuals. This attributes more negative value to the representation of an agent's action which is responsible for the harm than to the representation of an agent's inaction which is responsible for the harm. Thus, in the text session, people judge that actions maximizing aggregate welfare at the expense of physical harm to someone are inappropriate.

Outcomes are made more salient by the VR session in at least two ways: (i) the number of bodies that are going to be harmed are easily comparable on the screen before making a choice. This would predict increased utilitarian choice beginning with the very first experimental VR dilemma; (ii) since participants watch somebody get harmed in a violent and gory way after making a choice, this might influence their subsequent choices, making them more sensitive to outcomes. This would predict that participants' first choices in the VR dilemmas would be similar to their text choices, but that subsequent choices in the VR dilemmas would be more utilitarian. In order to arbitrate between these two possibilities, we carried out a new analysis. We noted that out of 33 participants only 3 (out of which 2 later changed to utilitarian choices) made a non-utilitarian decision on their first dilemma in VR, while 10 made a non-utilitarian decision on their first dilemma in the text session. Binary logistic regression with categorical predictor variables (VR, text) and response on the first dilemma as dependent variable (dummy coded as

0: non-utilitarian and 1: utilitarian) showed that participants were highly more likely to give a utilitarian response from the very beginning of the session in VR session than the text session ( $OR = 7.75$ , Wald's  $\chi^2 = 6.27$ ,  $p = 0.012$ ). This analysis supports the first hypothesis that the outcomes are made more salient due to the foregrounding of the virtual humans on the screen and not due to watching the gory deaths in the first non-utilitarian decision in VR. It could also be that the foregrounding of the virtual humans invokes the prospect of watching gory deaths, which motivates people to minimize the distress by choosing a utilitarian option. But this is just a speculation with no data from the current experiment to support it.

### *Role of emotions in order effects*

As mentioned above, observed asymmetric order-dependent judgment-behavior discrepancy was due to more labile judgments on the text dilemmas across orders (Mann-Whitney U test (2-tailed):  $p = 0.08$ ), while actions in the VR dilemmas were relatively more stable across orders (Mann-Whitney U test (2-tailed):  $p = 0.39$ ). This response pattern is reminiscent of the finding that when people face the trolley dilemma after considering the footbridge dilemma, they are significantly less likely to endorse utilitarian resolution, but making a judgment about the trolley dilemma has little to no effect on judgments about the footbridge dilemma (Schwitzgebel & Cushman, 2012). Schwitzgebel & Cushman (2012) suggest that participants' desire to maintain consistency between their responses (Lombrozo, 2009) is upheld when the emotionally more arousing case (e.g. footbridge) comes first and exerts influence on the emotionally less arousing case (e.g. trolley) so that these two cases are judged in a consistent manner, but overridden when the emotionally less arousing case comes first and fails to exert influence on the emotionally more arousing case and the two cases are judged in an inconsistent manner.

Similarly, in our experiment, when the participants acted in the emotionally salient VR dilemmas in the first session, these choices influenced the judgments in the text session and no discrepancy was observed. On the other hand, when the participants first judged emotionally flat text dilemmas in the first session and then faced the VR dilemmas, the desire to be consistent with responses from previous session was overridden by emotional impact of VR dilemmas. It is important to note that there was no significant difference in the ability to recall choices from the previous session for the group of participants in these two orders ( $Z = -0.57$ ,  $p = 0.62$ ). Therefore, variation in the ability to recall choices can't explain the observed pattern of order effect.

Thus, we assert that the differences in the inherent ability of the dilemma presentation modalities to elicit emotions were responsible for the observed asymmetric order effect.

### *Alternative explanations*

An alternative explanation for our behavioral results can be that the change in decisions is due to the different amount of time available for deliberation decisions in the two sessions, which can affect moral judgments (Suter & Hertwig, 2011; Paxton, Ungar, & Greene, 2012). Since the text session was self-paced, people had ample amount of time to ponder over the nature of the dilemma and then decide in 12 seconds. On the other hand, in the VR session, people had to comprehend and respond to these dilemmas within 10 seconds. It can thus be argued that people depended on quick affective processes while acting in the VR session but relied on slower, conscious reasoning processes when they made judgments in the text session. However, this seems unlikely because people took an equal amount of time in both sessions for endorsing the utilitarian option once controlled for differences specific to modality of presentation. Additionally, Suter and Hertwig (2011) showed that people, when pressured to give a response as quickly as possible, gave a smaller number of utilitarian responses but only in case of high-conflict moral dilemmas. There was no effect of available deliberation time on the likelihood of

making a utilitarian response on impersonal and low-conflict moral dilemmas. The same reasoning holds for the study by Paxton et al. (2012) which focused on moral judgments about sibling incest. In our experiment, we exclusively focused on impersonal dilemmas. This bolsters our contention that differences in the available time budget to make a decision cannot explain the observed pattern of discrepancy.

Another explanation can be that differences in cognitive load (reading vs. watching) intrinsic to the presentation modalities can explain this pattern of results, because cognitive load can modulate utilitarian decisions (Greene et al., 2008). However, effects of cognitive load cannot account for our results for three reasons. First, Greene et al.'s study showed that cognitive load affects utilitarian decisions but just in case of personal, high-conflict moral dilemmas (our study involved only impersonal dilemmas). Second, more importantly, the same study showed that there was a significant difference in the reaction time for utilitarian decisions in two conditions (load and no-load), but there was no change in the proportion of utilitarian decisions in these two conditions. So, although participants took more time to come to a utilitarian resolution under cognitive load, they made utilitarian decision nonetheless. Third, in our study, we controlled for the general differences in the presentation modalities using appropriate control conditions which were matched for most of the cognitive aspects except for the dilemmatic one. These considerations together with our reaction time data (people took equal amount of time to make utilitarian decisions in two sessions) make it highly unlikely that differences in cognitive load can explain the observed discrepancy.

#### *Shortcomings of the study*

Relying on impersonal moral dilemmas might have reduced the discrepancy. A significant percentage (53%) of the sample did not show any judgment-behavior discrepancy due to ceiling effect. It has been consistently found (Greene et al., 2001, 2004; Hauser et al., 2007; Mikhail, 2007) that there is a wide agreement among lay people that the best action in impersonal dilemmas is the one that allows an innocent individual to be physical harmed to achieve the maximum welfare for the maximum number of agents involved, with as many as 90% people endorsing this utilitarian outcome. However, there is a wide disagreement (Greene et al., 2001, 2004; Hauser et al., 2007; Mikhail, 2007) over the best course of action in case of personal moral dilemmas where an agent needs to be intentionally harmed as a mean to achieve the end of aggregate welfare, with proportion of people endorsing utilitarian outcomes varying widely depending on the context of the dilemmas at hand. Thus, it was not surprising that out of the 18 people who did not change their decisions, 17 had endorsed utilitarian actions in all the moral dilemmas in both sessions. Since this group of participants endorsed the maximum number of utilitarian decisions in both sessions, there was no room for judgment-behavior discrepancy to manifest. Future studies should extend current findings by using VR renditions of personal moral dilemmas. We speculate that the discrepancy would be greater for these dilemmas.

Another drawback of this study was that the moral behavior was investigated using virtual situations, which, although perceptually more salient and ecologically more valid, were still improbable. This poses limitations on the generalizability of these results to real-life setting. But we would like to note that predicting real-life behavior was not the primary objective of this study (cf. Mook, 1983).

## Conclusion

To summarize, in this study we have demonstrated that people show an order-dependent judgment-behavior discrepancy in hypothetical, impersonal moral dilemmas. This discrepancy was a result of the differential ability of contextual information to evoke emotions which motivate behavior, as indicated by the difference in SCR between the two modalities (VR vs. text). People judged in less utilitarian (or more action-based) manner in emotionally flat and contextually impoverished moral dilemmas presented in text format, while they acted in more utilitarian (or more outcome-based) manner in the emotionally arousing and contextually rich versions of the same dilemmas presented using virtual environments.

## Acknowledgments

We are thankful to the three anonymous reviewers for their invaluable comments and suggestions. We would also like to thank Eva-Maria Seidel for providing the Matlab script to analyze the skin conductance data and Riccardo Sioni for supervision during SCR recordings. The authors declare no conflict of interest.

## References

- Avramova, Y. R., & Inbar, Y. (2013). Emotion and moral judgment. *WIREs Cognitive Science*, 4, 169-178.
- Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13, 103–124.
- Christensen, J.F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. *Neurosci. Biobehav. Rev.* Apr; 36(4):1249-64.
- Cushman, F. A., (2013). Action, outcome and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17(3), 273-292.
- Cushman, F. A., Young, L., & Hauser, M. D. (2006). The Role of Reasoning and Intuition in Moral Judgments: Testing three principles of harm. *Psychological Science*, 17(12): 1082-1089.
- Cushman, F.A. , & Greene, J.D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, 7(3-4), 269-279.

- Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 159–181). Cambridge, UK: Cambridge University Press.
- FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navardy, L., & Dalgleish, T. (2012). What we say and what we do: the relationship between real and hypothetical moral choices. *Cognition*, 123, 434–441.
- Figner, B., & Murphy, R. O. (2010). Using skin conductance in judgment and decision making research. In M. Schulte-Mecklenbeck, A. Kuehberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research*. New York, NY: Psychology Press.
- Gilbert, D.T., & Wilson, T.D. (2007). Propection: experiencing the future. *Science*, 317(5843), 1351–4.
- 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*, Vol. 293, Sept. 14, 2001, 2105-2108.
- 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*, Vol. 44, 389-400.
- Greene, J.D., Morelli, S.A., Lowenberg, K., Nystrom, L.E., Cohen, J.D. (2008) Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, Vol. 107, 1144-1154.
- 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*, Vol. 111 (3), 364-371.
- Greene, J.D. (2009). The cognitive neuroscience of moral judgment, in *The Cognitive Neurosciences IV*, M.S. Gazzaniga, Ed. MIT Press, Cambridge, MA.
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316 , 998 – 1002.
- Hauser, M., Cushman, F., Young, L., Jin, R., & Mikhail, J. (2007). A dissociation between moral judgment and justification. *Mind and Language*, 22(1), 1-21.
- Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, 13, 1–6.
- Kant, I. (1785/2005). *The Moral Law: Groundwork of the Metaphysics of Morals* (Routledge; 2nd edition).
- Kobayashi, N., Yoshino, A., Takahashi, Y., & Nomura S. (2007). Autonomic arousal in cognitive conflict resolution. *Auton. Neurosci.*, 132, pp. 70–75.
- Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Science*, 11(4), 143–152.
- Mill, J.S. (1998). *Utilitarianism*, R. Crisp, ed. (New York: Oxford University Press).
- Mook, D.G. (1983). In defense of external invalidity. *Am. Psychol.*, 38:379–387.
- Moretto, G., Làdavas, E., Mattioli, F., & di Pellegrino, G. (2009). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience*, 22(8), 1888–1899.



- Navarrete, C.D., McDonald, M., Mott, M., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated 3-D "trolley problem." *Emotion*, 12(2): 364-70.
- Paxton, J.M., Ungar, L., Greene, J.D., (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36(1) 163-177.
- Rovira, A., Swapp, D., Spanlang, B., & Slater, M. (2009). The use of virtual reality in the study of people's responses to violent incidents. *Frontiers in Behavioral Neuroscience*. 3:59.
- Schwitzgebel, E. & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, 27, 135–53.
- Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*. 119(3):454-8.
- Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., et al. (2012). Disrupting the right pre-frontal cortex alters moral judgment. *Soc. Cogn. Affect. Neurosci.* 7, 282–288.
- Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Front. Psychol.* 4:250.
- Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, 94, 1395–1415.
- Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science*, 17(6), 476-477.
- 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). New York: Oxford University Press.

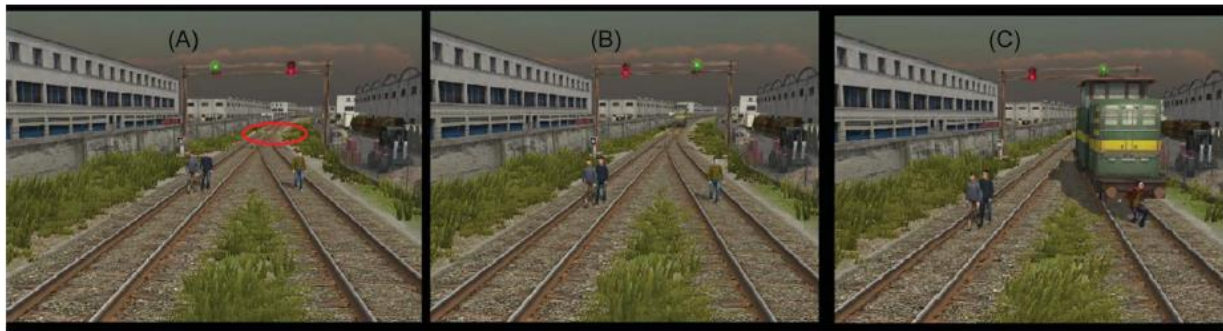
## Footnotes

<sup>1</sup>In addition to other differences mentioned in the Introduction section, our study also differed in this crucial aspect from the study of Navarrete et al. (2012), since in their study participants did not witness death of any virtual agent: “Screams of distress from either one or five agents became audible depending on the direction of the boxcar and the placement of the agents. Screaming was cut short at the moment of impact, and the visual environment faded to black.” (p. 367)

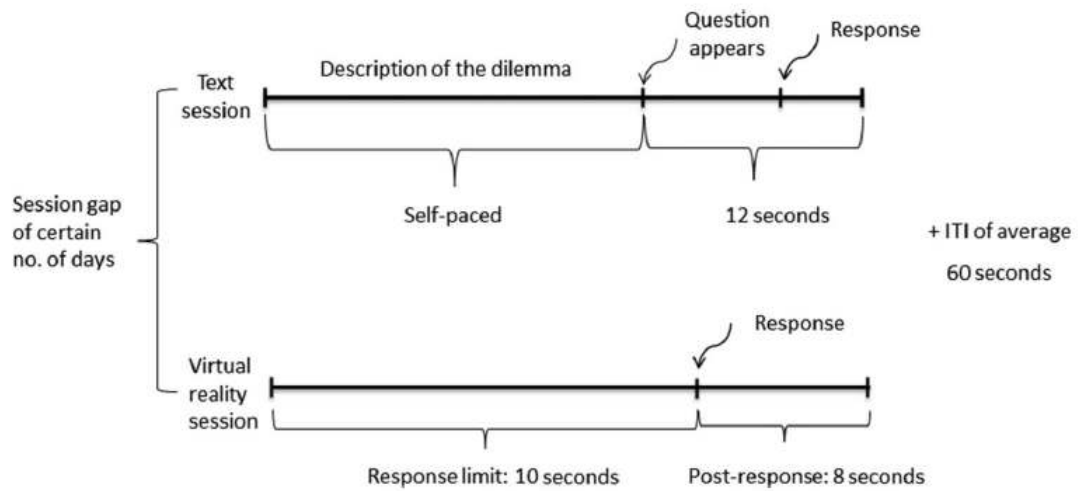
Table 1.

The judgment-behavior discrepancy between two sessions was dependent on the order in which participants performed sessions. (VR: virtual reality)

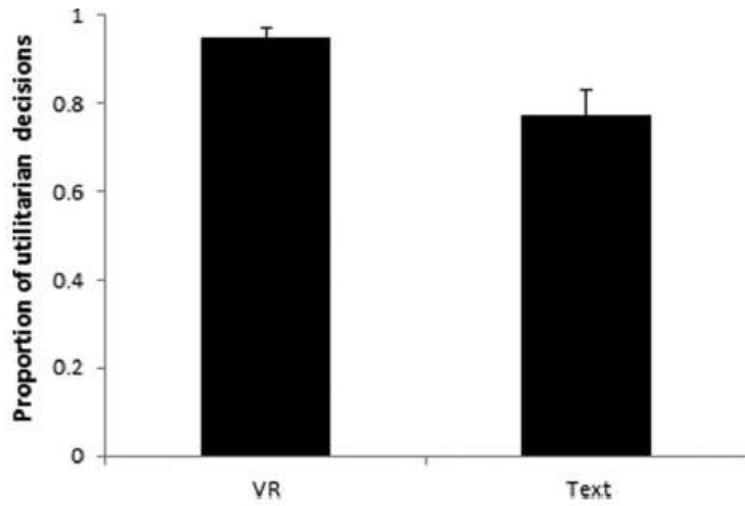
Order	Sample size	Change in proportion of utilitarian decisions (VR-text)		
		0	>0	<0
Text-first	19	8	<b>11</b>	0
VR-first	15	10	<b>4</b>	1



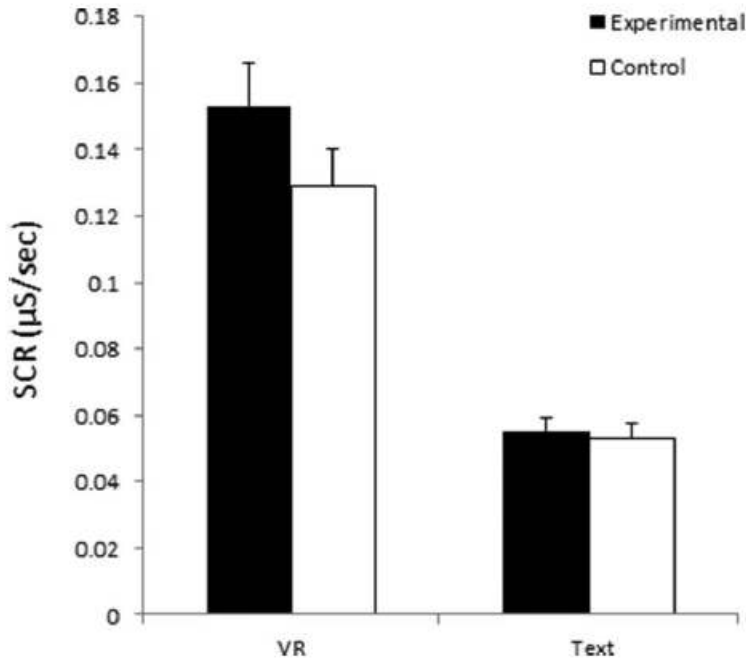
*Figure 1.* Film-strip of one representative dilemma from virtual reality session: the train dilemma. Participants had to make a decision in 10 seconds, before the train crossed the yellow-black striped line (highlighted in red circle). Train was by default directed at the maximum number of virtual agents, as shown by the green signal for the respective rail-track. (A). If participants wanted to achieve an utilitarian outcome, they had to change the signal for the track where two people were walking from “green” to “red” by pressing a button on the joystick, which automatically turned the signal to “green” for the alternative track where one virtual human was walking (B). After 10 seconds, the response buttons were automatically disabled and participants could not change their decision. After this, participants witnessed consequences of their actions (for 8 seconds) as the train progressed on the selected course and ran over the virtual human(s) (C). In this particular instance, participant endorsed a utilitarian outcome by choosing to actively divert the train on the alternative track.



*Figure 2.* Design of the experiment. Participants completed the task in two sessions, separated by a variable number of days. In the text session, participants read the dilemmas at their own pace and then gave their judgments in 12 seconds, while in the VR session they had to act within 10 seconds since the beginning of the virtual environment and witnessed consequences of their actions afterwards for 8 seconds.



*Figure 3.* Proportion of utilitarian decisions made in the two sessions differed significantly with people acting in more utilitarian manner in virtual reality (VR) dilemmas as compared to their judgments in the same dilemmas presented with text. Error bars indicate standard errors.



*Figure 4.* Skin conductance responses in the two sessions for making decisions in experimental and control conditions. Taking decisions in virtual reality (VR) dilemmas was more emotionally arousing than in textual dilemmas, even after controlling for general differences in the two modalities of presentation using the respective control conditions. Only in VR session taking decisions in experimental conditions was more emotionally arousing than control conditions. Error bars indicate standard errors.

## SUPPLEMENTAL MATERIALS - Descriptions of text dilemmas

The detailed descriptions (translated into English) of the moral dilemmas used in the text session are provided below. The original descriptions used in the experiment were in Italian.

### **Burning Car - Experimental**

You are the operator of a bridge in a huge parking complex. The bridge you operate connects two different floors. You see that a car out of control is approaching the bridge and is on fire. You don't know what is going on but you still need to operate the bridge.

You see that on the top floor there are five people walking and on the bottom floor, there is one person walking. The floors they are walking on are so narrow that only one car can pass through at any given time. Right now, the bridge is in the upward position. So if you don't do anything, the car would continue its course on the top floor and kill those five people in its way. But if you move the bridge in the downward position, the car would go on the bottom floor and would kill the one person walking there.

Is it appropriate for you to move the bridge in order to avoid the death of the five people, killing one person?

### **Burning Car - Control**

You are the operator of a bridge in a huge parking complex. The bridge you operate connects two different floors. You see that a car is approaching the bridge and is on fire. You don't know what is going on but you still need to operate the bridge.

You see that on the top floor there is one person walking and on the bottom floor, there are five empty boxes. The floor the person is walking on is so narrow that only one car can pass through at any given time. Right now, the bridge is in the upward position. So if you don't do anything, the car would continue its course on the top floor and kill the one person in its way. But if you move the bridge in the downward position, the car would go on the bottom floor and would crush all the boxes there.

Is it appropriate for you to move the bridge in order to avoid the death of that one person, destroying the boxes?

### **Lifting Magnet - Experimental**

You are the controller of a lifting magnet in a junkyard. Lifting magnets lift the heavy magnetic objects at some height and transport them to another place and drop them. The magnet, in automatic mode, is moving a car at some height from one place to another on the right side of the platform. Suddenly, you realize that if the magnet continues on its course, the magnet would drop the car on five people standing below. On the left, there is one person standing.

You can take control of the magnet. If you do nothing, the magnet would proceed to the right and drop the car attached to it, causing the death of the five people standing below. You can turn the magnet to left side of the platform, causing the death of the single person standing there.

Is it appropriate for you to turn the magnet in order to avoid the death of the five people, killing one person?

### **Lifting Magnet - Control**

You are controller of a lifting magnet in a junkyard. Lifting magnets lift the heavy magnetic objects at some height and transport them to another place and drop them. The magnet, in automatic mode, is moving, a car at some height from one place to another on the right side of the platform. Suddenly, you realize that if the magnet continues on its course, the magnet would drop the car on one person standing below. On the left, there are five empty boxes.

You can take control of the magnet. If you do nothing the magnet would proceed to the right and drop the car attached to it, causing the death of the person standing. You can turn the magnet to left side of the platform, causing the destruction of the boxes.

Is it appropriate for you to turn the magnet in order to avoid the death of that one person destroying the boxes?

### **Pier - Experimental**

You are in charge of operating an automatic coast-guard boat. From your operating station, you can see that there are five swimmers on your right who are being approached by sharks. But you also see that there is one swimmer on the left who is also being approached by sharks.

Right now, the boat you are operating is moving towards the person on the left. If you don't do anything, it can reach that one swimmer and he can be saved, but then the five swimmers on the right would get killed by sharks. You can save these five swimmers, only if you turn the boat to the right, but then the swimmer on the left would be killed.

Is it appropriate for you to turn the boat in order to avoid the death of the five swimmers letting one person die?

### **Pier - Control**

You are in-charge of operating an automatic coast-guard boat. From your operating station, you can see that there is a swimmer on your right who is being approached by sharks. But you also see that there are five empty boxes floating on the left.

Right now, the boat you are operating is moving towards the boxes on the left. You can save that swimmer, only if you turn the boat to the right, but then the boxes on the left would drown.

Is it appropriate for you to turn the boat in order to avoid the death of the swimmer, letting the boxes drowning?



**Train- Experimental**

You are standing on a railway track where a single track divides into two tracks. There is a switch to control the track of the train. You see a train out of control approaching rapidly. On the track extending to the left is a group of five railway workers. On the track extending to the right is a single railway worker.

If you do nothing, the train will proceed to the left, causing the death of the five workers. The only way to avoid the death of these workers is to hit a switch on your dashboard that will cause the train to proceed to the right, causing the death of the single worker. If you don't do this, those workers will be killed but one worker on the right track would remain safe.

Is it appropriate for you to hit the switch in order to avoid the death of the five workers, killing one person?

**Train - Control**

You are standing on railway track where a single track divides into two tracks. There is a switch to control the track of the train. You see a train out of control approaching rapidly. On the track extending to the left is a collection of five empty boxes. On the track extending to the right is a single railway worker.

If you do nothing the train will proceed on the right track and would kill the worker. You can avoid this by hitting a switch and turning the train on left track. But this would destroy the boxes.

Is it appropriate for you to hit the switch in order to avoid the death of the worker, destroying the boxes?

**SUPPLEMENTAL MATERIALS – Details on the 10-second response limit in VR (for both experimental and control scenarios).**

- For the burning car dilemma, participants had to respond before the car hit the ramp (shown in red circle):



- For the lifting magnet dilemma, participants had to respond before the magnet crossed the yellow-black striped line (shown in red circle):



- For the pier dilemma, participants had to respond before the coast-guard boat crossed the end of the floating objects (shown in red circle):



- For the train dilemma, participants had to respond before the train crossed the yellow-black striped line on the track (shown in red circle):

