SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI

INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Neuroscience Area

Cognitive Neuroscience Curriculum

A multimodal investigation of moral decision making in harmful contexts

Thesis submitted for the degree of

Doctor Philosophiæ

Supervisor:

Dr. Giorgia Silani

Candidate:

Indrajeet Patil

January 2016

Thesis Committee

Externals:

Elisa Ciaramelli

Dipartimento di Psicologia

Università di Bologna

Fiery Cushman

Department of Psychology

Harvard University

Internals:

Davide Crepaldi

Neuroscience Area

International School for Advanced Studies (SISSA)

Raffaella Rumiati

Neuroscience Area

International School for Advanced Studies (SISSA)

Davide Zoccolan

Neuroscience Area

International School for Advanced Studies (SISSA)

To,

Aai ani Papa,

for their unconditional love and support.

"Two things fill the mind with ever new and increasing admiration and awe, the more often and steadily we reflect upon them: the starry heavens above me and the moral law within me…. I see them before me and connect them immediately with the consciousness of my existence."

- Immanuel Kant, *Critique of Practical Reason*

Contents

Acknowledgments

I started my PhD during what was possibly the worst dip in my academic career, with a fading interest in physics - the only subject I had studied thus far - and only a vague idea about possible future direction in my life. And here I am now, at the end of my PhD, feeling much more confident as a researcher with a strong affinity for continuing in academics with a number of different ideas to explore. Needless to say, I have many people to thank for this transformation.

First of all, I owe immensely to my adviser **Giorgia**, who took a significant risk in hiring a student without any background in psychology or neuroscience, for putting her trust in me and for being extremely patient while I tried to find my feet in the initial stages. I really admire her breadth of understanding and her ability to zoom in on the most crucial aspects of research and their broad implications and this is something that I will try to carry forward with me in future. I also appreciated the freedom I was given in terms of choosing the projects to work on and the encouragement for being self-sufficient enough to be an independent researcher. In addition to training me to become a good researcher, she has also imprinted on me what it means to be a good adviser and, more generally, a decent human being. I would consider myself lucky if I can manage to be even half as kind and caring towards my students as she was towards me.

Thanks are also due to amazing colleagues that I have had the opportunity to work with – either in person at my home institution or abroad via the magic of internet. First and foremost, I would like to thank **Marta** and **Fede** who were extremely patient while working on what seemed to be a never-ending project. Their sweet and innocent company definitely helped ameliorate frustrations brought upon by the complications that ensued during the fMRI study and I hope that I was a helpful mentor for them. It was a pleasure working with **Carlotta**, **Luca**, and **Nicola** on

what was a logistically and technically demanding project, but their committed efforts really smoothed the wrinkles and made it into an exciting and successful project. I gratefully acknowledge **Lisa Burger**, **Marco Zanon**, and **Emanuela** for their help during data collection of few other projects and **Alessio Isaja** for the technical support.

Thanks to **Jens** for being such an inspiring collaborator on the Vienna project! I have really learned a lot from your discipline and acumen when it comes to executing experiments. Hopefully we will continue our collaborative efforts in future with success.

I am greatly indebted to **Ezequiel** for giving me an opportunity to work on some exciting data from neurological populations and for being a terrific mentor. I found interactions with you and (the inimitable) **Liane** both intellectually stimulating and highly inspiring. I hope to continue this productive collaboration in future. Thanks are also due to **Fiery**, whose work has been a great inspiration for me, for insightful conversations during our collaborative effort and hopefully we will have better luck with our future projects.

Thanks are also due to the broader community of friends in Trieste for providing me with all the precious memories and warm company that made me feel at home in a city thousands of miles away from my actual home. I would especially like to thank **Manu** for her constant love, support, and encouragement. The times we had remained and continue to remain at the locus of my sanity and make me a better person. Also, I will perennially remain grateful to you for transmitting your OCD (Obsessive Cat Disorder) to me.

Big hugs to **Georgette** and **Julia** for having continuously been by my side no matter what! I have really enjoyed our aperitifs, fancy dinners at restaurants, late-night conversations about diverse topics, weekend getaways, movie nights, etc. It has been a lot of fun to know you both

and your company has played a big part in me feeling at home in Trieste and has prevented me from coming undone during some of the acutely frustrating phases of my stay in Trieste.

Thanks to **Wasim**, **Shital**, **Merlyn**, **Sunil**, and the larger community of Indian friends in Trieste for their companionship and for helping me sustain integrity of my umbilical cord to *swades* and the authentic Indian food.

Thanks are also due to my funny colleague and office-mate **Giovanni** (aka Giacomo) for all the laughs and for suggesting some alternative professions for me (involving roses) in case academia doesn't work out for me!

Abhilash has continued his role as my sounding board from a separate continent and he knows how much I appreciate it! Thanks to my dear friends **Abhijit** and **Sandesh** for providing me with havens for retreat in Germany whenever I got sick of Trieste.

For the lack of space, I can't thank every individual in Trieste and back in Pune who has helped me in one way or another to make this thesis possible. I am indebted to all of you.

I would be remiss if I did not acknowledge the thousands of YouTube uploaders who share amusing cat videos. Without the intermittent positive stimulation provided by these videos, it would have been very difficult for me to cope with anxieties and uncertainties that are characteristic of academic life.

My gratitude also goes out to my thesis committee (**Fiery** and **Elisa**) for the time and effort they have put in to help me improve the thesis matter.

Last but not the least, I would like to thank my family, **Aai**, **Papa**, and **Vishu**, whose support has never wavered whether my academic career was in waxing or waning phase. They have been

extremely patient with me and truly understand my passions and have tried to help in any way they can in supporting them. I would also like to apologize to them for being so absent for the last four years having turned into a workaholic and promise to do better in future.

General overview of the thesis

Since the two landmark publications in moral psychology (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001), the field has experienced an affective revolution that has put emotions at the center of the stage. Although work on exploring role of emotions in assessing morality of various types of moral acts (impure, unfair, etc.; Haidt, 2007) abounds, studying its role in harmful behaviors presents a unique challenge. The aversion to harming others is an integral part of the foundations of human moral sense and it presents itself in the form of deeply ingrained moral intuitions (Haidt, 2007). Since creating laboratory situations to investigate harm aversion raises ethical issues, research has primarily relied on studying hypothetical cases. In the current thesis, we utilize hypothetical vignettes to explore role of emotions in both moral judgment and behavior in harmful contexts, both when harm is carried out intentionally or produced accidentally.

Study 1 investigates the role of emotion in motivating utilitarian *behavior* in moral dilemmas when presented in contextually salient virtual reality format as compared to *judgment* about the same cases for their textual versions.

Study 2 investigates divergent contributions of two different sources of affect, one stemming from self-focused distress and the other focused on other-oriented concern, on utilitarian moral judgments in autistics.

Study 3 investigates the role of empathic arousal in condemning agents involved in unintentional harms and why harmful outcomes have a greater bearing on blame as compared to acceptability judgments.

Chapter 1

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

*This chapter is based on the following published article:

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*(1), 94-107. doi:10.1080/17470919.2013.870091

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.

1. 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 (Christensen & Gomila, 2012; Cushman & Greene, 2012; Waldmann, Nagel, & Wiegmann, 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, 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 & 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, Loomis, & Beall, 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, Swapp, Spanlang, & Slater, 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 et al., 2006; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001; 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, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene et al., 2004, 2001). 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., 2004, 2001) and skin conductance activity (Moretto, Làdavas, Mattioli, & di Pellegrino, 2010), 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.

2. Methods and Materials

2.1 *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).

2.2 *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., 2004, 2001).

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 Appendix S1 for description of text dilemmas (videos of VR scenarios can be downloaded from here: http://www.sissa.it/cns/cescn/SupInfos/vr_scenarios.zip). 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: 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.

2.3 *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 (also see Appendix S2 for details on how participants kept track of available time). 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).

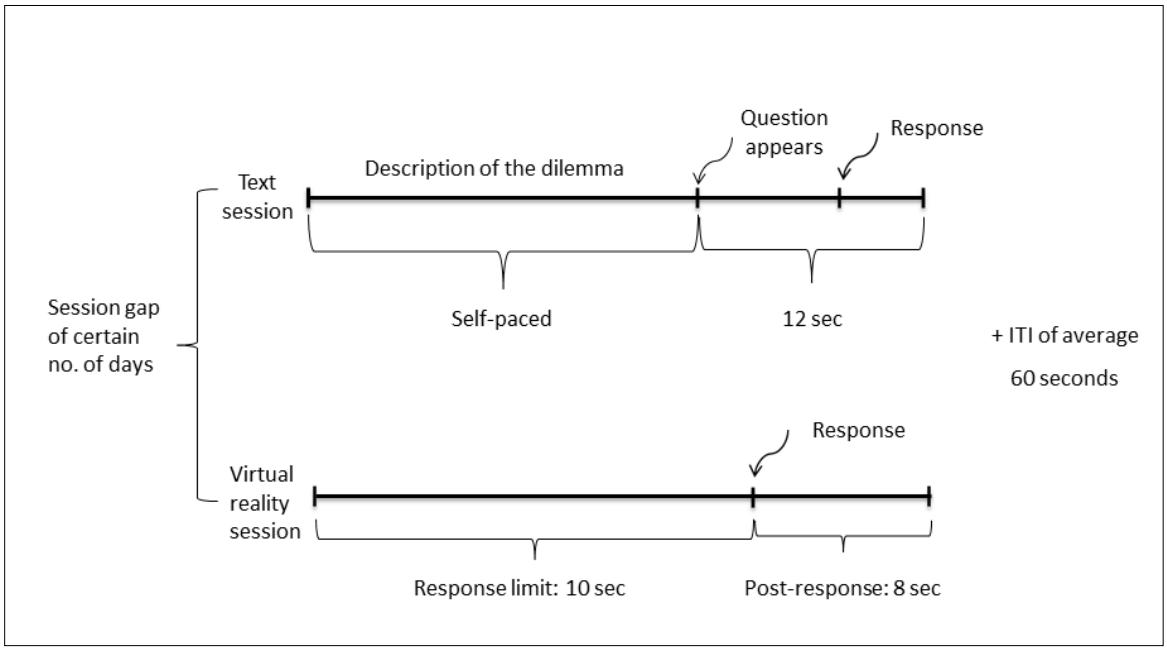


Figure 2: Design of the experiment. Participants completed the task in two sessions, separated by a variable number of days. In the text session, participants read the dilemmas at their own pace and then gave their judgments in 12 seconds, while in the VR session they had to act within 10 seconds since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

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 et al., 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).

3. Results

3.1 *Responses*

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).

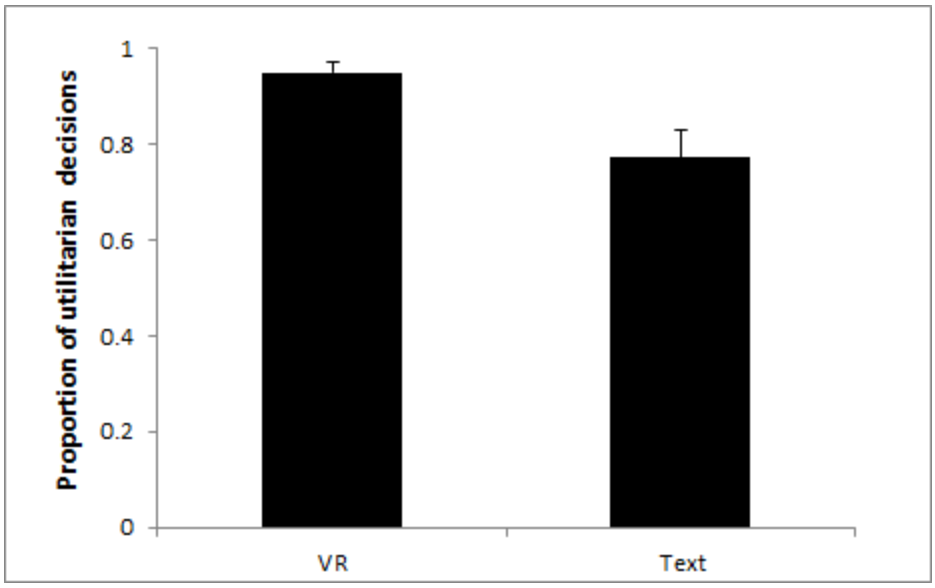


Figure 3: Proportion of utilitarian decisions made in the two sessions differed significantly with people acting in more utilitarian manner in virtual reality (VR) dilemmas as compared to their judgments in the same dilemmas presented with text. Error bars indicate standard errors.

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 (χ^2(1) = 3.348, p = 0.06, φ = -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: 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	11	0
VR-first	15	10	4	1

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.

3.2 *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: *p*s > 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.

3.3 *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[1] 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 µS. We then computed SCRs for all the trials as "area under curve" (Moretto et al.,

[1] 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)

2010). 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, μS) 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: ps > 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). The ANOVA revealed a main effect of session ($F(1,33) = 65.15$, $p < 0.001$, $p\eta^2 = 0.67$) 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 ($F(1,33) = 11.28$, $p = 0.002$, $p\eta^2 = 0.26$), 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: $F(1,14) = 7.65$, $p = 0.016$, $p\eta^2 = 0.37$ and for order text-first: $F(1,17) = 5.44$, $p = 0.032$, $p\eta^2 = 0.24$.

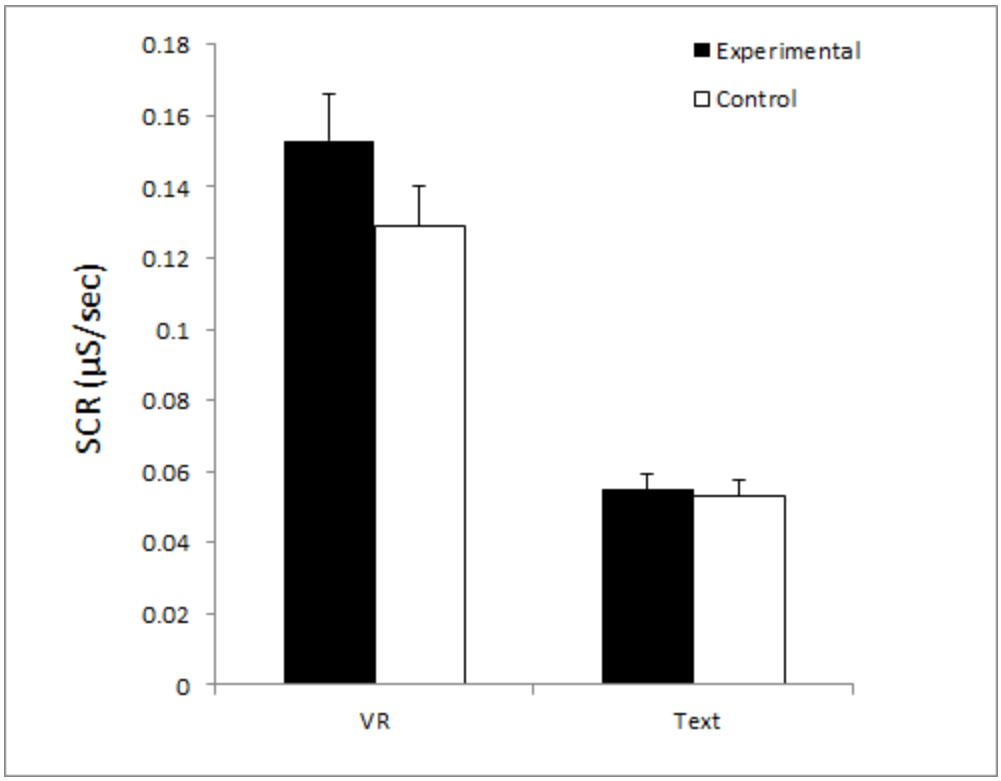


Figure 4: Skin conductance responses in the two sessions for making decisions in experimental and control conditions. Taking decisions in virtual reality (VR) dilemmas was more emotionally arousing than in textual dilemmas, even after controlling for general differences in the two modalities of presentation 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.

Post-hoc t-tests revealed that the experimental conditions were more arousing than control conditions only for VR session: $t(33) = 3.68$, $p = 0.001$, Cohen's $d = 1.28$ (for order VR-first: $t(14) = 3.58$, $p = 0.003$, Cohen's $d = 1.91$ and for order text-first: $t(18) = 2.28$, $p = 0.036$, 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: $F(1,33) = 12.72$, $p = 0.001$, $p\eta^2 = 0.28$. 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 ($F(1,29) = 73.74$, $p < 0.001$, $p\eta^2 = 0.73$), main effect of condition ($F(1,29) = 9.20$, $p = 0.005$, $p\eta^2 = 0.25$), and interaction ($F(1,29) = 11.50$, $p = 0.002$, $p\eta^2 = 0.29$). 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.

3.4 *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).

4. 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.

4.1 *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.

4.2 *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 et al., 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., 2008, 2004, 2001), 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, 2013a). 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 χ^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.

4.3 *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.

4.4 *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 (Paxton, Ungar, & Greene, 2012; Suter & Hertwig, 2011). 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 at 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.

4.5 *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., 2004, 2001; 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., 2004, 2001; 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).

5. 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.

Chapter 2

Divergent roles of autistic and alexithymic traits in utilitarian moral judgments in adults with autism*

*This chapter is a slightly modified version the following article submitted for publication:

Patil, I.[§], Melsbach, J.[§], Hennig-Fast, K., & Silani, G. (revision submitted). Divergent roles of autistic and alexithymic traits in utilitarian moral judgments in adults with autism. *Scientific Reports*. [§]Equal contribution.

Abstract

This study investigated hypothetical moral choices in adults with high-functioning autism and role of empathy and alexithymia in such choices. We used a highly emotionally salient moral dilemma task to investigate autistics' hypothetical moral evaluations about personally carrying out harmful utilitarian behaviours aimed at maximizing welfare. Results showed that they exhibited a normal pattern of moral judgments despite the deficits in social cognition and emotional processing. Further analyses revealed that this was due to mutually conflicting biases associated with autistic and alexithymic traits after accounting for shared variance: (*a*) autistic traits were associated *reduced* utilitarian bias due to elevated personal distress of demanding social situations, while (*b*) alexithymic traits were associated with *increased* utilitarian bias on account of reduced empathic concern for the victim. Additionally, autistics relied on their non-verbal reasoning skills to rigidly abide by harm-norms. Thus, utilitarian moral judgments in autism were spared due to mutually conflicting influence of autistic and alexithymic traits and compensatory intellectual strategies. These findings demonstrate the importance of empathy and alexithymia in autistic moral cognition and have methodological implications for studying moral judgments in several other clinical populations.

"[Autistic people are] cold, calculating killing machines with no regard for human life!"

- Facebook post by "Families Against Autistic Shooters" in response to the mass-shooting incident at Umpqua Community College, Oregon (as reported in *The New York Times* Op-Ed article "The Myth of the 'Autistic Shooter'" by Andrew Solomon, October 12, 2015)

1. Introduction

Harmful behaviours are inherently dyadic, comprising of an agent who harms and a victim who gets harmed (Gray & Schein, 2012). Accordingly, moral evaluations in healthy individuals about such behaviours hinges on two different routes to the understanding of other minds (Bzdok et al., 2012): a cognitive route that represents agent's beliefs and goals (called theory of mind (ToM) or *sociocognitive* route), while an affective route that identifies feeling states in the victim and elicits isomorphic feeling states (e.g., pain) in the observer (called empathy or *socioaffective* route).

Autism spectrum disorder (ASD) is characterized by problems with reciprocal social interaction, impaired communication, repetitive behaviours/narrow interests and impairments in these very aspects of social cognition and emotional processing necessary for proper moral reasoning (Bird & Cook, 2013). Although past work has investigated impact of ToM deficits on moral judgments, the effect of empathy deficits remains to be thoroughly investigated. Furthermore, recent body of work shows that only ToM deficits are inherent to the autistic phenotype and the empathy deficits are due to co-occurring alexithymia (Bird & Cook, 2013) (a subclinical condition characterized by difficulty in identifying and describing subjective feeling states, difficulty in differentiating feelings from bodily sensations, and diminished affect-related fantasy (Lane, Weihs, Herring, Hishaw, & Smith, 2015; Sifneos, 1973)). Thus, role of alexithymia in

moral evaluations in autism is to date largely unexplored (Brewer et al., 2015). The current study explores these issues further.

1.1 *Moral cognition in autism: an overview*

A number of prior studies have utilized variety of moral cognition tasks to explore if the capacity to judge third-party harmful behaviours is intact in ASD in the light of the deficits in social cognition and emotional functioning. This research shows that the distinction between intentional moral transgressions (that involve a suffering victim whose personal rights are violated; e.g. hitting others) and conventional transgressions (characterized by infraction of normative prohibitions but with no consequence for others' welfare; e.g. talking out of turn) is substantially intact in children and adults with ASD (Blair, 1996; Leslie, Mallon, & DiCorcia, 2006; Shulman, Guberman, Shiling, & Bauminger, 2012; Zalla, Barlassina, Buon, & Leboyer, 2011). These studies underscore that ASD population (both children and adults) can distinguish between *intentional* good and bad actions and have preserved moral knowledge (Gleichgerrcht et al., 2013; Li, Zhu, & Gummerum, 2014).

Although autistics do not seem to be impaired in evaluating intentional third-party harm-doings, they exhibit more enduring deficits on more complex intent-based moral judgment tasks that require integration of information about mental states of the agents with the information about outcomes of these acts. In particular, they judge accidental harms more harshly, arguably due to their inability to form a robust representation of agent's benign intentions (due to ToM deficits (Fletcher-Watson & McConachie, 2014)) that can be weighted up against a strong negative emotional response stemming from the victim suffering (Buon, Dupoux, et al., 2013; Koster-Hale, Saxe, Dungan, & Young, 2013; Moran et al., 2011; Roge & Mullet, 2011) (but see Baez et

al., 2012). Thus, this work is consistent with the profile of ASD featuring preserved psychophysiological/emotional response to others' affective states (affective empathy) but reduced cognitive understanding about others' internal states (ToM) and demonstrates how sociocognitive/ToM deficits in ASD modulate their moral judgments about third-party moral violations, but only when these processes need to operate in tandem with other processes (e.g., harm assessment) that provide conflicting contextual information that needs to be integrated for a final moral judgment (Baez & Ibanez, 2014; Zalla & Leboyer, 2011).

Despite an abundance of work focusing on role of ToM deficits on performance on intent-based judgment tasks that involve conflict between intent and consequences, there is a paucity of literature exploring how empathy deficits in ASD translate into behavioural choices in hypothetical scenarios.

1.2 *Empathy and moral condemnation of harmful behaviour*

Emotions play a pivotal role in condemnation of harmful behaviours (Avramova & Inbar, 2013) and empathy is a social emotion that plays a crucial role in such moral evaluations (Decety & Cowell, 2014; Ugazio, Majdandžić, & Lamm, 2014). This is because (real or hypothetical) harmful encounters include a suffering victim and empathy allows moral judges to understand their suffering and use the resulting "gut-feelings" to either approve or disapprove of such moral actions (Ugazio et al., 2014). But empathy is a multidimensional construct (Davis, 1983) consisting of a cognitive component that is involved in merely understanding the emotional states in others, while affective empathy enable observers to share these feeling states in an isomorphic manner. Accordingly, affective empathy has been found to be more consequential in motivating behaviour (for a review, see Ugazio et al., 2014). But affective empathy itself has two

disparate facets that are associated with different motivational tendencies (Decety & Cowell, 2014; Ugazio et al., 2014): (*i*) other-oriented *empathic concern* involves intuitions about protecting physical integrity of others and being apprehensive of any actions that result in harm to others and is associated with *appetitive* motivation to prevent harm to others; (*ii*) *self-oriented personal distress* reflects aversive feeling contingent on vicarious sharing of the others' emotional and physical distress and sense of loss of control in emotionally charged harmful situations and is associated with *avoidance* motivation to escape such distressful situation.

Given this crucial role of empathy in moral condemnation of harmful behaviour, ASD would be expected to have impairments in moral judgments in situations that harness these processes. But this simplistic picture is further complicated in light of the new insights provided by the alexithymia hypothesis (Bird & Cook, 2013) which postulates that only the deficits observed in the *sociocognitive* domain are unique to the autism phenotype, while the deficits associated with *socioaffective* domain are due to the co-occurring alexithymic phenotype and is not a feature of autism *per se* (Bernhardt et al., 2014). Although the preponderance rate of clinical levels of alexithymia in healthy population is at 10%, it is unusually prevalent (40-65%) in adults and children with ASD (Berthoz & Hill, 2005; Griffin, Lombardo, & Auyeung, 2015; Hill, Berthoz, & Frith, 2004; Salminen, Saarijärvi, Äärelä, Toikka, & Kauhanen, 1999). Therefore, it is important to account for its effects in emotional processing deficits observed in ASD, especially because trait alexithymia itself has been associated with impaired emotional processing (e.g., empathy (Grynberg, Luminet, Corneille, Grèzes, & Berthoz, 2010), emotion regulation(Swart, Kortekaas, & Aleman, 2009), emotional interoception (Silani et al., 2008), etc.). Thus, it is likely that, when observed, the emotional processing deficits in ASD are due to the presence of elevated levels of alexithymia. Indeed, after accounting for co-occurring alexithymia, autism is

no longer associated with aberrant neural activation while empathizing with others' pain (Bird et al., 2010), self-reported deficits on dispositional empathy (Aaron, Benson, & Park, 2015), or deficits in interocepting on one's own emotional states (Silani et al., 2008).

Thus, any investigation gauging effects of aberrant emotional skills on moral cognition in ASD should also account for effects of prevalent alexithymia. Indeed a number of recent studies have begun to explore role of alexithymia in moral judgments in both clinical (Gleichgerrcht, Tomashitis, & Sinay, 2015; Patil, Young, Sinay, & Gleichgerrcht, 2016) and non-clinical populations (Koven, 2011; Patil & Silani, 2014a, 2014b), but only one study thus far has investigated this issue (Brewer et al., 2015) in the ASD population and found limited support for the alexithymia hypothesis. In the current study, we further investigate role of emotional processing deficits and alexithymia in autistics' moral cognition with a well-validated moral judgment task.

1.3.2 *Utilitarian moral judgments on moral dilemmas*

One widely used task that assesses role of emotional processing in first-party, hypothetical harmful behaviours is the moral dilemma task (Christensen & Gomila, 2012; Greene et al., 2004). Moral dilemmas are situations where two moral principles conflict with each other, e.g. "do not do harm unto others" against "act in a way so that maximum number of people will be better off". In the harm domain, these dilemmas are instantiated by creating scenarios where the agent needs to act in order to produce the least harmful of possible outcomes (e.g., killing one to save many), i.e. situations where inaction would lead to more people getting hurt, but acting requires actively harming someone. These moral dilemmas are further divided into two classes based on the nature of harmful actions and their causal-intentional structure (Mikhail, 2007) (see

Figure 1 for examples): (*i*) moral dilemmas that require agents to harm someone in up close and personal manner, i.e. by executing a motor act (Greene et al., 2009), and where the victim needs to harmed as a *means* to achieve the greater good are called *personal* moral dilemmas (e.g., pushing someone to their death to save greater number of lives); (*ii*) moral dilemmas that feature harms that carried out not by physical force but by mechanical means and where the harm that befalls the victim is a *side-effect* of harmful act are called *impersonal* moral dilemmas (e.g., switching course of a trolley that kills someone to save more number of lives). Although the net outcome of choosing to act in both types of dilemmas can be the same (e.g., one life lost but five lives saved), most people endorse acting (which is said to be an utilitarian response) in cases of impersonal dilemmas but refuse to do so on personal dilemmas (which is said to be a deontological/non-utilitarian response (Greene et al., 2004)).

The dual-process model posits two types of processes that support each type of response in respective dilemma-contexts (Greene et al., 2004): (*i*) automatic, affect-laden intuitions that surface as a reflex to aversive nature of the proposed harm and subserve non-utilitarian moral judgment; (*ii*) controlled, deliberative reasoning processes that engage in cost-benefit analysis and support utilitarian solution. Therefore, according to this model, individuals endorse utilitarian moral judgments more frequently on impersonal but not personal moral dilemmas because personal cases lead to a stronger negative affect in response to severe physical harm that needs to be carried out using personal force. There is plenty of evidence to support this claim (Greene, 2014): neuroimaging (Greene et al., 2004), psychophysiological (Moretto et al., 2010), and behavioural (Szekely & Miu, 2015) measures corroborate this model by revealing that indeed personal moral dilemmas elicit a more pronounced emotional response than the impersonal cases. Of interest to the current investigation, this negative emotional arousal

partially stems from the harmful outcome, *viz.* empathic concern for the (to be sacrificed) victim's pain which causes personal distress in the moral judge (Ugazio et al., 2014).

Condition	*Non-moral*	*Impersonal*	*Personal*
Text description	You have a very bad headache. You go to the pharmacy looking for your favorite brand of headache medicine. When you get there, you find that the pharmacy is out of the brand that you are looking for. You have known the pharmacist at this store for a long time, and you trust him. He says he has a generic medicine that is "exactly the same" as the name-brand medicine that you wanted. In the past, he has always given you good advice.	You are the driver of a runaway trolley approaching a fork in the tracks. On the tracks going to the left is a group of five railway workers. On the tracks going to the right is a single railway worker. If you do nothing, the trolley will go to the left, causing the five workers to die. The only way to avoid the deaths of these five workers is to hit a switch on your dashboard that will make the trolley go to the right, leading to the death of the single worker.	A runaway trolley is heading down the tracks toward five workers, and will kill them if it keeps going. You are on a footbridge over the tracks, in between the approaching trolley and the five workers. Next to you on this footbridge is a stranger who is very large. The only way to save the lives of the five workers is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workers will be saved.
Behaviour	Would you [nature of action] in order to [outcome of the proposed action]? (yes/no)		
Arousal	How emotionally arousing did you find this scenario? (0 = not at all arousing; 20 = extremely arousing)		

Figure 1: *Stimulus examples*. Three conditions from the moral dilemma task with representative examples from each category. Each type of dilemma was followed by two questions: behaviour and emotional arousal. Impersonal and personal conditions involved moral content (implications for others' wellbeing), while the non-moral cases involved only pragmatic issues.

Despite extensive use of this task in healthy controls, very little work has been carried out with the autistic population. Extensive prior work has focused on investigating moral cognition in clinical populations (e.g., patients with damage to the prefrontal cortex) and subclinical traits (e.g., psychopathy) characterized by social cognition and emotional processing disturbances using the moral dilemma task. These studies have consistently revealed that these populations have increased rate of utilitarian judgments on emotionally charged personal dilemmas as compared to control brain-damaged or neurotypical individuals (Chiong et al., 2013; Ciaramelli, Sperotto, Mattioli, & di Pellegrino, 2013; Djeriouat & Trémolière, 2014; Gleichgerrcht,

Torralva, Roca, Pose, & Manes, 2011; Koenigs et al., 2007; Moretto et al., 2010; Patil, 2015; Taber-Thomas et al., 2014). Drawing on this prior work, one would expect that ASD would also beget a similar utilitarian moral profile due to similar sociocognitive and socioaffective problems.

Accordingly, one previous study has shown that ASD individuals are more willing to sacrifice someone for the greater good on personal moral dilemmas and report to perceive such situations to be less emotionally distressing as compared to controls, arguably due to reduced perspective-taking (cognitive empathy) that normally enables individuals to see things from the perspective of the person that needs to be sacrificed (Gleichgerrcht et al., 2013). But this study used only one moral dilemma per condition and thus generalizability of these results remains to be assessed. This finding is also surprising in the light of evidence for prevalent negative hyperarousal in autistic individuals (Capps, Kasari, Yirmiya, & Sigman, 1993; Samson, Hardan, Lee, Phillips, & Gross, 2015; Smith, 2009), which would make it less likely that they would make utilitarian moral judgments (Greene et al., 2004; Greene, 2014). Indeed, another unpublished study did not find any evidence for such increased utilitarian proclivity in ASD (Dr. Geoffrey Bird, personal correspondence).

The alexithymia hypothesis provides a plausible explanation for these conflicting findings in the past work. Recent research shows that elevated level of subclinical alexithymia is associated with utilitarian profile on personal moral dilemma (Koven, 2011), arguably due to reduced empathic concern (which stands for feelings of compassion and sympathy for the unfortunate others) for the victim that needs to be sacrificed (Patil & Silani, 2014b). Thus, it is possible that the prior finding about increased willingness to personally sacrifice someone for the greater good in ASD (Gleichgerrcht et al., 2013) was due to presence of greater number of alexithymics in the ASD

group as compared to healthy controls, since alexithymia is associated with both reduced perspective-taking and empathic concern for others (Grynberg et al., 2010). Thus, increased tendency to endorse harmful sacrificial behaviours on moral dilemmas might have resulted from failure to empathize with the victim that needs to be sacrificed due to co-occurring alexithymia in ASD. Alternatively, it is also possible that utilitarian inclination due to alexithymic traits was counterbalanced by non-utilitarian inclination due to autistic traits. Severity of autism is associated with increased personal distress during demanding social situations (Gu et al., 2015; Smith, 2009), which persists even after accounting for co-occurring alexithymia (Y.-T. T. Fan, Chen, Chen, Decety, & Cheng, 2014), and this increased personal distress leads to withdrawal from engaging in personally carrying out harmful actions (Sarlo, Lotto, Rumiati, & Palomba, 2014). Thus, the nature of *between-group* differences in utilitarian moral judgment in a given study may depend on these *within-ASD-group* interactions between autistic and alexithymia traits that exert mutually opposite influence on utilitarian moral judgments.

Past work in autism also shows that autistics develop compensatory strategies from early childhood to counteract their lack of social intuitions (Frith, 2004) whereby they strictly adhere to explicitly learned social rules and conventions in an inflexible or stereotyped manner (Baron-Cohen, Richler, Bisarya, Gurunathan, & Wheelwright, 2003). This can also be garnered from overreliance on rule-based thinking while making distinction between (third-party) conventional and moral norm transgressions (Shulman et al., 2012; Zalla et al., 2011), which are usually justified by healthy controls on the basis of considerations about victim suffering. Additionally, they rely less on emotional information and more on rule-based norm obedience while evaluating their own hypothetical choices about moral and prosocial behaviours (Brewer et al., 2015; Jameel, Vyas, Bellesi, Roberts, & Channon, 2014). Thus, it is possible that autistics rely on their

intellectual abilities to form strategies that help them deal with complexities of distressing social environments and make adaptive decisions in such settings. This important aspect of their cognition has gone understudied in the past work and we explore its role in utilitarian moral judgments in the current study in concert with other personality traits.

1.3.3 *Predictions*

Although we did not expect any group differences for utilitarian judgments on impersonal dilemmas based on prior work (Gleichgerrcht et al., 2013), we did not have any *a priori* predictions regarding the between-group difference for utilitarian judgments on personal dilemmas in light of the conflicting findings from past studies. Indeed, in our framework, this difference can vary from study-to-study depending on the intricate web of mutually conflicting inputs from a composite of personality traits in the ASD sample (autism, alexithymia, intelligence measures, etc.).

We made following predictions for moral judgments in autistics on personal moral dilemmas: (*i*) alexithymic traits in the ASD sample would be associated with increased utilitarian inclination (Koven, 2011; Patil & Silani, 2014b) to endorse harmful sacrificial actions due to reduced empathic concern (Aaron et al., 2015; Conway & Gawronski, 2013; Gleichgerrcht & Young, 2013; Grynberg et al., 2010; Guttman & Laporte, 2002; Miller, Hannikainen, & Cushman, 2014; Robinson, Joel, & Plaks, 2015; Royzman, Landy, & Leeman, 2015; Wiech et al., 2013); while (*ii*) autistic traits would be associated with reduced tendency to endorse utilitarian solution due to increased negative emotional arousal stemming from personal distress (Sarlo et al., 2014; Spino & Cummins, 2014) experienced by autistics while facing demanding social environments (Dziobek et al., 2008; Y.-T. T. Fan et al., 2014; Gu et al., 2015; Rogers, Dziobek, Hassenstab,

Wolf, & Convit, 2007; Smith, 2009). Note that although one may expect affective empathy (empathic concern and personal distress, i.e.) to predict *greater* endorsement for the utilitarian solution on personal dilemma due to greater empathizing with the many (Robinson et al., 2015) - who would die in case of inaction – this is not observed because the utilitarian course of action features causal intervention on an identifiable and singular victim (Wiss, Andersson, Slovic, Västfjäll, & Tinghög, 2015) that needs to be sacrificed and thus the other set of victims are pushed to the background in the causal model and does not elicit a robust empathic response (Waldmann & Dieterich, 2007; Wiegmann & Waldmann, 2014). Additionally, we note that although autism is associated with increased personal distress even after accounting for co-occurring alexithymia (Y.-T. T. Fan et al., 2014; Patil, Melsbach, Hennig-Fast, & Silani, 2016), trait alexithymia itself is also associated with greater personal distress but this association seems to be due to prevalent anxiety and is not characteristic of the alexithymic phenotype (Grynberg et al., 2010).

Additionally, we expected there to be a negative correlation between intelligence measure and utilitarian moral judgments in ASD representing rigid rule-based norm abidance, but we were agnostic as to which component of IQ (verbal or non-verbal) would be implicated as a compensatory strategy and made this decision based on the exploratory correlation analysis.

Although recently a number of criticisms have surfaced that challenge interpreting affirmative response on moral dilemma as *utilitarian* (Kahane, 2015), we use utilitarian to mean "characteristically utilitarian" as a function of the response content and not the underlying motivation (Greene, 2014). Thus, if a given individual responds affirmatively on a moral dilemma, we do not take this response to denote explicit endorsement of utilitarian moral principle ("those acts are better that save more number of lives") on her part, but only to mean

that this response coincides with a response that would be endorsed by a typical, card-carrying utilitarian moral philosopher (Greene, 2014).

2. Methods

2.1 *Participants*

The study sample consisted of 17 subjects (6 females) with a diagnosis of autism spectrum disorder (ASD group), who were recruited from autism-specific organizations, associations and internet communities via various information materials (e.g., print flyers and posters, digital flyers, and Facebook advertisings) and had undergone a screening for any current comorbid psychiatric or medical condition. Importantly, we did not exclude ASD participants who were on medication - 7 subjects were consuming psychoactive drugs, primarily for depression. The medicated ASD group did not differ on any of the variables of interest from the non-medicated ASD group. The diagnosis was carried out by experienced clinicians according to the internationally accepted ICD-10 diagnostic criteria (World Health Organization, 1992). In line with a prior study (Schneider et al., 2013) and DSM-V (American Psychiatric Association, 2013), we do not further divide 'ASD group' into 'high-functioning autism' and 'Asperger's Syndrome' subgroups. We use the terms 'autism', 'on the autism spectrum', 'autistic,' and 'autism spectrum disorder' to refer to the ASD group as these terms are preferred by this population (Kenny et al., 2015).

Seventeen age-, gender- and level of education-matched participants (4 females; $\chi^2(1) = 0.567$, $p = 0.452$) were also included in the healthy controls (HC) group after an interview to ensure absence of history of drug abuse, neurological or neuropsychiatric disorders. We note that although the final ASD group consisted of high-functioning autistic individuals with IQ comparable to the control group, the highest educational degrees that autistic individuals

possessed tended to be slightly lower than the healthy controls (see Table 1; presented after References).

All participants were financially compensated for their time and travel expenses and gave written informed consent. The study was approved by the local Ethics Committee (University of Vienna) and conducted in accordance with the declaration of Helsinki.

2.2 *Questionnaires*

Various questionnaires (German-validated versions) were administered to assess individual differences in various aspects of the socioaffective processing: (*i*) Autism Spectrum Quotient (AQ) to assess severity of autistic traits (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001; Freitag et al., 2007); (*ii*) Toronto Alexithymia Scale (Bagby, Parker, & Taylor, 1994; Kupfer, Brosig, & Brähler, 2000) (TAS) to assess severity of alexithymic traits; (*iii*) Interpersonal Reactivity Index (Davis, 1983; Paulus, 2009) (IRI) as a self-report measure of trait empathy and Multifaceted Empathy Test (Dziobek et al., 2008) (MET; revised version provided by I. Dziobek, personal correspondence) as a performance measure of state empathy; (*iv*) Emotion Regulation Questionnaire (Abler & Kessler, 2009; Gross & John, 2003) (ERQ) to assess emotion regulation profile; (*v*) Beck Depression Inventory (Beck, Steer, & Brown, 1996; Hautzinger, 1991) (BDI) to assess severity of depression; (*vi*) short version of Raven's Standard Progressive Matrices (Bilker et al., 2012; Raven, Raven, & Court, 1998) (SPM) and Mehrfachwahl-Wortschatz-Intelligenztest-B (Lehrl, Triebig, & Fischer, 1995; Lehrl, 1995) (MWT-B; Multiple Choice Vocabulary Intelligence Test) to assess non-verbal and verbal intelligence, respectively.

Good internal reliability was observed for subscales of questionnaires (see Table 1). For more detailed discussion about the questionnaires and their internal reliability analyses, see Appendix (Text S1).

2.4 *Moral dilemma judgments*

Stimuli: Experimental stimuli were text-based scenarios. There were three conditions representing each class of scenario: non-moral practical dilemmas (n = 6), impersonal moral dilemmas (n = 6), and personal moral dilemmas (n = 6) (see Figure 1 for representative examples and Appendix (Text S2) for detailed description of the scenarios). All scenarios featured first-person narrative.

Personal dilemmas featured situations that demanded agents (read participants) to carry out actions using personal force that violated others' personal rights (Greene et al., 2009). Compared to personal dilemmas, impersonal cases featured actions which were less emotionally salient and implicated the agent in the scenarios in less personal manner. The common denominator between moral dilemmas was that they pitted the normative injunction against violating someone's individual rights by harming them in personal or impersonal manner against the utilitarian option of saving greater number of lives.

Non-moral scenarios posed practical questions and lacked any moral content. Data from non-moral scenarios are included in every model as a control condition. Thus, if any systematic differences are observed for moral dilemmas on any dependent variable, we can ascertain that this effect is specific to the moral domain by checking if the same effect is observed also for prudential, non-moral dilemmas.

Procedure: All participants were individually tested in a quiet room at the Faculty of Psychology of the University of Vienna. The experiment was carried out in two sessions separated on average by a week (M_{ASD} = 5.87±3.02 days, M_{HC} = 6.13±2.00 days, t(24.046) = -0.279, p = 0.783). In one session, participants completed the moral dilemma task; while in the other session, they completed another task (data not reported here). Similarly, in one session, participants completed AQ, IRI, TAS, and MET; while in the other session, participants completed ERQ and two other questionnaires (data not reported here). The moral tasks and questionnaire set pairings were randomized across sessions and participants. For the moral judgment task, before starting the actual experiment, each participant took part in one practice trial to ensure that they had understood all the instructions.

Moral judgment task and MET were administered on a computer, while the questionnaires were administered in paper-and-pencil format. The stimuli for the moral judgment tasks were presented using Cogent 2000 (Wellcome Department of Imaging Neuroscience, http://www.vislab.ucl.ac.uk/cogent.php) running on MATLAB platform. The text of the stories was presented in a black 21-point Arial font on a white background with a resolution of 800 × 600 pixels. MET task was presented using OpenSesame 2.8.1 program (Mathôt, Schreij, & Theeuwes, 2012) with a resolution of 1920 × 1080 pixels.

For the moral judgment task, the order of presentation of scenarios from each condition was randomized within subjects. Each dilemma description was presented in a single screen. Participants could read this screen at their own pace and move to the questions, by pressing the spacebar on the keyboard. The next two screens, presented in the same order for all participants, contained questions assessing: behavioural choice and emotional arousal (for exact wording, see Figure 1). The behaviour and arousal questions lasted for as long as the participants needed. The

affirmative answer on the behaviour question always corresponded to commission of sacrificial action. The spatial location (left or right arrows on the keyboard) of two options (yes or no) was constant across scenarios and subjects in order to avoid confusion and reduce working memory demands, especially for the ASD group. The emotional arousal ratings were recorded using a computerized visual analog scale (VAS), implemented as horizontal on-screen bar and responses were later converted to standardized scores with [min, max] of [0, 20].

We focused on behavioural choice of action ("Would you do it?") over appropriateness of action ("Is it appropriate for you to do it?") because: (*i*) it tends to be more emotionally arousing (Patil, Cogoni, Zangrando, Chittaro, & Silani, 2014), (*ii*) it tends to elicit more egocentric/self-focused (versus allocentric/other-focused) frame of reference because of potential self-relevant consequences (Tassy et al., 2013), and (*iii*) perceived appropriateness of utilitarian course of action on moral dilemmas does not differ in ASD (Gleichgerrcht et al., 2013) (as compared to healthy controls). Thus, the behavioural choice of action provides a more sensitive measure to tap into moral cognition in autism.

Two ASD participants did not complete the moral dilemma task due to their unavailability for the second session, while data from one control participant could not be collected due to technical problems with MATLAB. The descriptive statistics for measures other than moral dilemma task thus include data from these additional participants. All results remain identical after excluding this data and thus they are retained in the current analysis.

2.6 *Statistical analysis*

All statistical analysis was carried out using JASP 0.7.1.12 (https://jasp-stats.org/). Effect size measures are reported as per prior recommendations (Lakens, 2013). All tests are two-tailed,

unless otherwise stated. As recommended (Weissgerber, Milic, Winham, & Garovic, 2015), we provide univariate scatter-plots instead of bar graphs, especially given the small sample sizes in the current study. We follow recommended guidelines (Nimon, 2012) to ensure that our data met the statistical assumptions associated with the general linear model-based statistical tests.

Correlation analysis was carried out using Spearman's *rho* as it is more robust to univariate outliers (Pernet, Wilcox, & Rousselet, 2013) than Pearson's *r*. To compare significance of between-group differences in correlations, we used Fisher's *Z*-test as implemented in FZT-computator (http://psych.unl.edu/psycrs/statpage/FZT_backup.exe).

2.7 *Path analysis*

In order to study complex web of interactions between different personality variables for utilitarian moral judgments, we conducted path analysis. Path analysis was performed in SPSS Amos 22 using maximum likelihood estimation (Arbuckle, 2013). Path analysis is a multivariate technique that requires formal specification of a model to be estimated and tested based on prior research and hypothesis. It involves specifying relationships between study variables and multiple equations denoting these relationships are solved simultaneously to test model fit and estimate parameter estimates (Arbuckle, 2013). Note that path analysis is concerned only with *testing* the validity of theoretically-inspired models by fitting them to the observed data and not with *building* models (Streiner, 2005). As such, it cannot arbitrate as to whether the given model is correct or not, but only whether it fits the observed data. In the current study, path analysis was used to study divergent contributions of personality traits in utilitarian moral judgments in ASD. To this effect, models were constructed based on past work in the field and our theoretical predictions. The model fit was further improved by reducing model misspecification error with

the inclusion of variables based on their correlation pattern with the variables of interest. As recommended (Streiner, 2005), model fit was not improved based on modification indices, but based on drawing paths that were theoretically meaningful.

All variables were standardized and centred before the analysis. Presence of multivariate outliers was investigated using Mahalanobis distance (none found). Since all paths represent linear relationships with a theoretically predicted direction, the significance threshold for regression coefficients associated with each path was determined based on one-tailed tests. Although there was a possibility of mediation effect involving some of the paths, no formal mediation analysis was carried out because the sample size was insufficient to carry out such analyses (Fritz & MacKinnon, 2007).

In order to assess goodness of model fit, we chose indices that have been found to be least susceptible to effects of sample size, model misspecification, and parameter estimates. Following guidelines provided by Hooper and colleagues (Hooper, Coughlan, & Mullen, 2008), we used - (*i*) model chi-square and the root mean square error of approximation (RMSEA), along with the associated *p*-value for close fit, as the absolute fit indices (which measure the model fit in comparison to no model at all), (*ii*) comparative fit index (CFI) along with its parsimony index (PCFI) as the incremental fit indices (which gauge the model fit with respect to null model where all variables are uncorrelated). We do not report the standardized root mean square residual (SRMR) as Amos does not produce this index in the presence of missing data. The recommended cut-off values are (Hooper et al., 2008): RMSEA $\leq$ 0.07 (good), 0.07 < RMSEA $\leq$ 0.10 (moderate), *p* for close fit > 0.05, CFI $\geq$ 0.95. There is no recommended cut-off for PCFI.

3. Results

3.1 *Elevated levels of alexithymia in ASD*

As expected, ASD group had higher alexithymia score than the HC group (see Table 1). There were 8 autistics (out of 17 or 47%) who were also clinically alexithymic (Bagby et al., 1994) (≥ 54), while no participant from the control group scored above the clinical cut-off. The frequency of alexithymics differed significantly across groups ($\chi^2(1) = 10.462$, $p = 0.001$, $\phi = 0.555$).

3.2 *Emotional processing deficits in ASD*

As expected autistics were impaired (as compared to controls) on a number of emotional processing measures (see Table 1): (*i*) they reported to have reduced dispositional tendency to adopt others' perspective and to experience increased personal distress in interpersonal interactions; (*ii*) they also exhibited maladaptive emotion regulation profile that relied more on suppressing emotion-expressive behaviour rather than reappraising emotional response; (*iii*) they did not exhibit any impairment on performance measures of empathy but did take longer to complete this task, arguable by relying on compensatory mechanisms; (*iv*) they exhibited increased levels of depression.

Table 1: Descriptive statistic and group differences for various demographic, clinical, and experimental variables of interest.

Variable	Cronbach's alpha	HC (n = 16)		ASD (n = 15)		Welch's t-test			
		Mean	SD	Mean	SD	t	df	p	Cohen's d
Clinical and demographic									
Age	-	32.03	9.44	37.35	13.02	-1.295	25.43	0.207	-0.470
Education	-	4.50	1.41	3.40	1.92	1.807	25.67	0.083	0.656
SPM	-	7.44	1.32	7.53	1.64	-0.179	26.84	0.86	-0.065
MWT-B	-	29.94	2.82	31.13	4.21	-0.924	24.24	0.365	-0.336
BDI	-	3.25	2.35	9.53	7.81	-2.992	16.37	0.008	-1.106
AQ-k	0.954	5.69	3.00	24.87	3.44	-16.49	27.88	< .001	-5.951
SIS	0.945	1.06	1.34	9.00	1.89	-13.41	25.10	< .001	-4.873
IC	0.861	2.19	2.23	8.53	1.55	-9.25	26.85	< .001	-3.286
CR	0.842	2.44	1.41	7.33	2.16	-7.42	23.89	< .001	-2.701
SPF-IRI	0.658	50.31	6.10	50.80	8.32	-0.19	25.59	0.855	-0.067
Fantasy	0.683	13.00	2.68	10.87	3.40	1.93	26.65	0.064	0.700
Empathic Concern	0.748	13.94	3.23	13.40	3.02	0.48	29.00	0.636	0.172
Perspective-taking	0.756	14.38	2.68	11.73	2.91	2.62	28.36	0.014	0.945
Personal distress	0.804	9.00	1.93	14.80	3.55	-5.60	21.32	< .001	-2.049
TAS	0.863	34.75	3.96	53.60	8.63	-7.74	19.37	< .001	-2.841
DIF	0.888	9.63	1.86	20.13	5.01	-7.64	17.56	< .001	-2.817
DDF	0.844	11.38	2.19	20.20	2.51	-10.40	27.84	< .001	-3.755
EOT	0.473	13.75	2.52	13.27	3.37	0.45	25.87	0.656	0.163
ERQ									
ERQ - Reappraisal	0.873	27.13	7.08	20.53	8.41	2.35	27.47	0.026	0.851
ERQ - Suppression	0.726	12.69	3.20	15.87	6.70	-1.67	19.78	0.111	-0.613
MET									
Cognitive - positive	-	16.50	3.18	15.53	1.68	1.066	23.09	0.298	0.376
Cognitive - positive - RT (in ms)	-	5563.28	1540.90	8609.84	3002.64	-3.519	20.59	0.002	-1.290
Cognitive - negative	-	14.38	2.39	15.07	3.39	-0.653	25.02	0.52	-0.237
Cognitive - negative - RT (in ms)	-	6103.65	2012.07	7979.87	2861.18	-2.099	24.98	0.046	-0.763
Affective - positive	-	5.56	1.55	4.11	1.44	2.691	29	0.012	0.965
Affective - positive - RT (in ms)	-	2933.40	1176.21	4663.17	2052.09	-2.855	22	0.009	-1.043
Affective - negative	-	5.47	1.02	4.82	1.86	1.207	21.36	0.241	0.442
Affective - negative - RT (in ms)	-	3796.17	1241.32	4819.58	2255.86	-1.551	21.46	0.136	-0.567

Note that results from emotional processing measures are only briefly described here as data from these measures were ancillary to the main objective of the study. These results will be discussed in greater depth elsewhere (Patil, Melsbach, et al., 2016).

3.3 *Moral dilemma task*

The descriptive statistics for all variables associated with this task have been tabulated in Appendix (Text S3). Although we had response time data, we do not draw any inferences about underlying psychological processes from analysis of this data as this practice of reverse inference has recently been demonstrated to be problematic (Krajbich, Bartling, Hare, & Fehr, 2015).

Accordingly, analysis of response time data is provided in the Appendix (Text S4). Suffice it to note here that there were no group differences for any condition for any type of response (utilitarian or non-utilitarian).

3.3.1 *No group differences in behavioural choice on moral dilemmas*

A 3 (condition: non-moral, impersonal, personal) × 2 (group) mixed ANOVA for behaviour question revealed a main effect of condition ($F(1.536,44.534) = 31.736$, $p < 0.001$, $p\eta^2 = 0.523$, $\omega^2 = 0.494$), but there was neither a main effect of group ($F(1,29) = 0.293$, $p = 0.593$) nor a group-by-condition interaction effect ($F(1.536,44.534) = 1.032$, $p = 0.347$). Thus, autistics and controls did not differ in terms of their willingness to act in utilitarian manner on moral dilemmas. Of interest to us was personal moral dilemma on which autistics reported to be slightly less utilitarian than controls (see Figure 2), although this difference was not significant ($t(28.65) = 1.572$, mean difference = -0.117, 95% CI [-0.268, 0.035], p(uncorrected) = 0.127, d = 0.566).

Decomposing the main effect of condition with planned Bonferroni-corrected comparisons revealed expected pattern of judgment for both groups: participants were more likely to be utilitarian on impersonal moral dilemmas as compared to personal moral dilemmas (HC: $t(15) = 4.652$, mean difference = 0.302, 95% CI [0.180, 0.424], $p < 0.001$, d =1.163; ASD: $t(14) = 8.000$, mean difference = 0.444, 95% CI [0.318, 0.571], $p < 0.001$, $d = 2.066$) (see Figure 2).

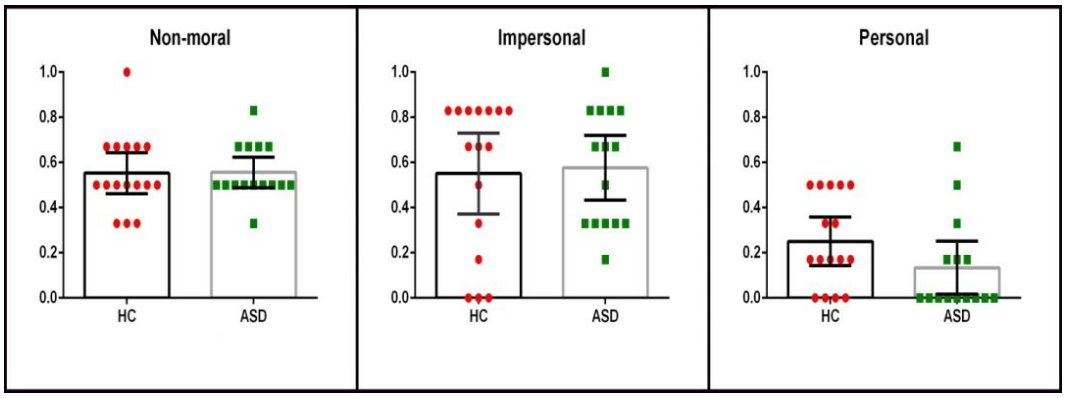


Figure 2: *Summary of results for the behaviour question*. Univariate scatter-plots (and corresponding bar-graphs) for proportion of affirmative responses on each type of scenario for each group for the behaviour question. For impersonal and personal moral dilemmas, higher scores indicate increased utilitarian tendency. Error bar represents 95% confidence intervals.

3.3.2 *Group differences in emotional arousal while facing moral dilemmas*

A 3 (condition: non-moral, impersonal, personal) × 2 (group) mixed ANOVAs for the arousal question revealed was a main effect of condition ($F(1.578,45.756) = 104.700$, $p < 0.001$, $p\eta^2 = 0.783$, $\omega^2 = 0.771$) but no condition-by-group interaction ($F(1.578,45.756) = 0.250$, $p = 0.727$). Planned comparisons revealed that both groups felt more emotionally aroused while facing scenarios from impersonal (HC: $t(15) = 9.517$, mean difference = 10.419, 95% CI [8.085, 12.750], $p < 0.001$, $d = 2.379$; ASD: $t(14) = 9.203$, mean difference = 11.495, 95% CI [8.816, 14.170], $p < 0.001$, $d = 2.376$) and personal (HC: $t(15) = 7.096$, mean difference = 8.476, 95% CI [5.930, 11.020], $p < 0.001$, $d = 1.774$; ASD: $t(14) = 6.161$, mean difference = 9.336, 95% CI [6.086, 12.590], $p < 0.001$, $d = 1.591$) dilemma conditions as compared to non-moral conditions. But both types of moral dilemmas were rated to be equally emotionally arousing (HC: mean

difference = -1.942, p = 0.144; ASD: mean difference = -2.518, p = 0.096). Thus, autistics were not impaired in decoding emotional saliency of different types of scenarios.

Interestingly, there was also a main effect of group (F(1,29) = 16.720, p < 0.001, $p\eta^2$ = 0.366, ω^2 = 0.336). Bonferroni-corrected post-hoc comparisons revealed that ASD individuals found all scenarios to be more emotionally arousing than controls (non-moral: t(18.92) = 3.690, mean difference = 3.736, 95% CI [1.616, 5.855], p = 0.006, d = 1.357; impersonal: t(28.81) = 3.552, mean difference = 4.812 , 95% CI [2.040, 7.583], p = 0.003, d = 1.270; personal: t(27.88) = 2.556, mean difference = 4.596, 95% CI [0.912, 8.279], p = 0.048, d = 0.923; see Figure 3). Note that the emotional arousal was not specific to the moral domain, but was domain-general as would be expected based on prior studies (Samson et al., 2015).

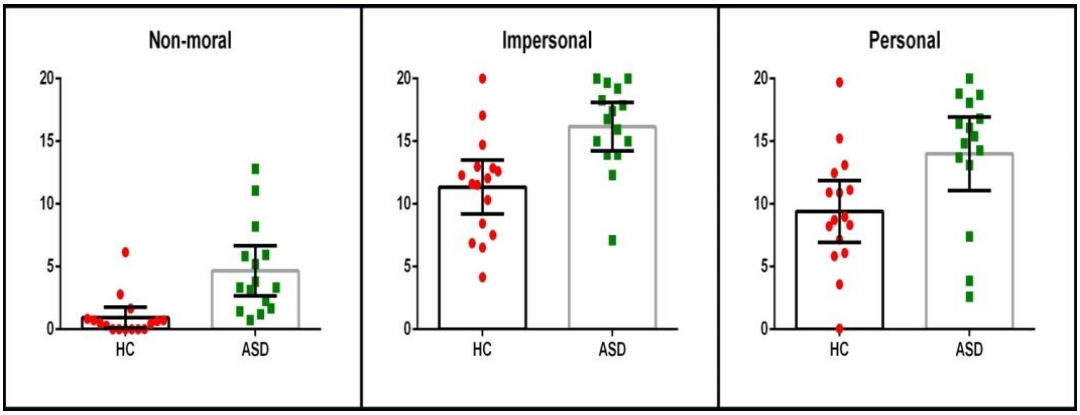


Figure 3: *Summary of results for the emotional arousal question*. Univariate scatter-plots (and corresponding bar-graphs) for self-reported emotional arousal (higher ratings denote more emotional arousal) while facing each type of scenario for each group. Error bar represents 95% confidence intervals.

3.4 Correlations analyses for utilitarian moral judgments on moral dilemmas

Correlations between moral judgments, arousal ratings, empathy, emotion regulation, and intelligence measures were computed. Additionally, between-group differences in correlation patterns were investigated. Full details of these analyses are provided in Appendix (Text S5-10).

In addition to the variables of a priori interest (AQ, TAS, EC, and PD), we used this correlation analyses to select additional variables that may have an influence on utilitarian moral judgments in ASD group. Interestingly, MWT-B was correlated negatively with utilitarian judgments on personal dilemmas in ASD (ρ = -0.739, p = 0.002), while SPM showed a marginally significant negative correlation (SPM: ρ = -0.459, p = 0.085). This pattern did not differ from the pattern observed in controls for MWT-B (ρ = -0.521, p = 0.039; Z = 0.926, p = 0.354), but it did differ for SPM (ρ = 0.392, p = 0.134; Z = 3.606, p < 0.001). Thus, while higher general non-verbal intellectual abilities were associated with higher endorsement for utilitarian option on personal dilemmas in healthy controls, the pattern was exactly opposite in ASD participants such that higher SPM scores were predictive of reduced tendency to behave in utilitarian manner, although the correlation was only marginally significant (see Figure 4; also see Appendix (Text S11) for a similar scatterplot for MWT-B). No such group difference was observed for a measure of verbal intelligence. Thus, we selected SPM as a measure of non-verbal intelligence in our path model that we suspected was utilized by autistics as a compensatory strategy to cope with arousing social situations. We note that non-verbal IQ was chosen to represent a possible compensatory strategy not based on where it was significant or not, but based on the fact that the correlation between non-verbal IQ and moral judgment differed across groups.

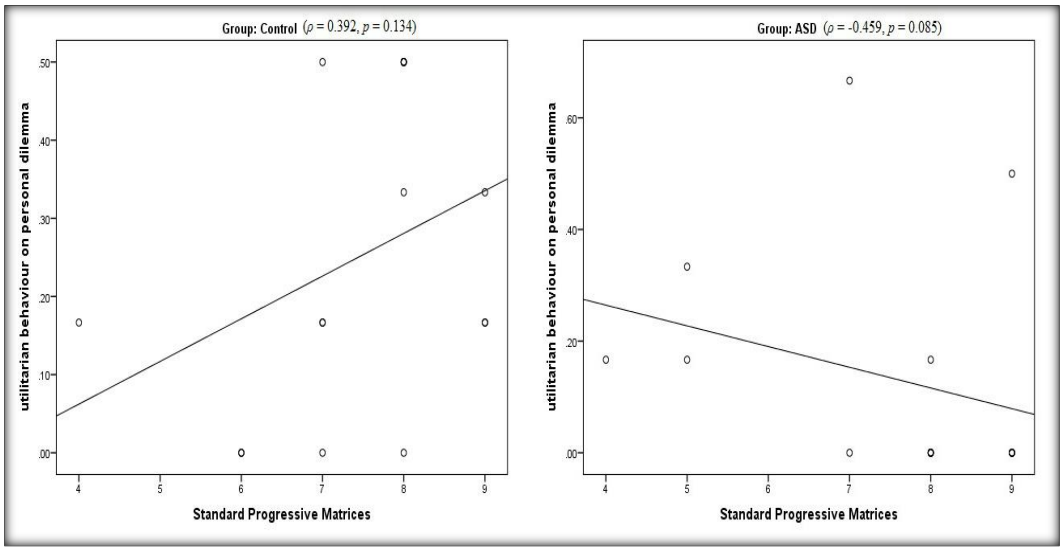


Figure 4: *Non-verbal reasoning skills and moral judgments.* The relation observed between non-verbal intelligence scores (as assessed by Raven's Standard Progressive Matrices) and utilitarian moral judgment on personal moral dilemmas was diametrically opposite for the two groups ($Z = 3.606$, $p < 0.001$). In controls, higher SPM scores were associated with a greater tendency to make utilitarian judgments, while autistics with higher SPM scores exhibited less favourable position for utilitarian option. Note that the number of data-points in the scatterplot seems to be less than the sample sizes due to overlap between data-points (denoted by circles with thicker circumference). Reported *p*-values are two-tailed.

3.5 Path analysis of utilitarian moral judgments in ASD

In order to assess why utilitarian moral judgments were preserved on personal moral dilemmas in ASD despite the prevalent deficits in social cognition and emotional processing associated with this disorder, we formulated a path model for the different processes that were predicted to mediate mutually conflicting influences to leave the final moral judgment intact.

As mentioned before, alexithymic traits were predicted to be associated with increased utilitarian profile (Koven, 2011) due to reduced empathic concern (Patil & Silani, 2014b), while autistic traits were expected to be associated with reduced utilitarian tendency on account of increased personal distress (Y.-T. T. Fan et al., 2014; Gu et al., 2015; Sarlo et al., 2014; Smith, 2009). Additionally, we included SPM as a measure of intelligence since our correlation analyses showed that higher SPM scores were associated with reduced utilitarian tendency in the ASD group and thus may index rule-based compensatory strategy to evaluate moral behaviour on hypothetical cases in ASD (Brewer et al., 2015; Jameel et al., 2014). We also accounted for possible effects of medication (Price, Cole, & Goodwin, 2009) status (dummy-coded as ON = 1, OFF = 0) on mediating variables; all effects of interest are observed even after exclusion of this variable and hence this variable was retained based on the improvement of the model fit. Although perspective-taking subscale of IRI has been implicated in increased utilitarian moral judgments on personal dilemmas in a prior ASD study (Gleichgerrcht et al., 2013), we did not include it in the path analysis because - (*i*) none of the previous studies investigating predictive ability of different aspects of empathy (using IRI) in utilitarian moral judgments reveal any association between these two variables (Gleichgerrcht & Young, 2013; Patil & Silani, 2014b; Sarlo et al., 2014), and (*ii*) inclusion of this variable led to a poor model fit ($p < 0.05$). Additionally, although we had both trait (IRI) and state (MET) measures of empathy we included only trait measures since a past study reveals that trait measures are better predictors of moral judgments on moral dilemmas than state measures (Choe & Min, 2011). Additionally emotion regulation measures were not incorporated in the path model because they were not correlated with moral judgments (Appendix (Text S6)).

The final model created with the inclusion of these variables is shown in Figure 5.

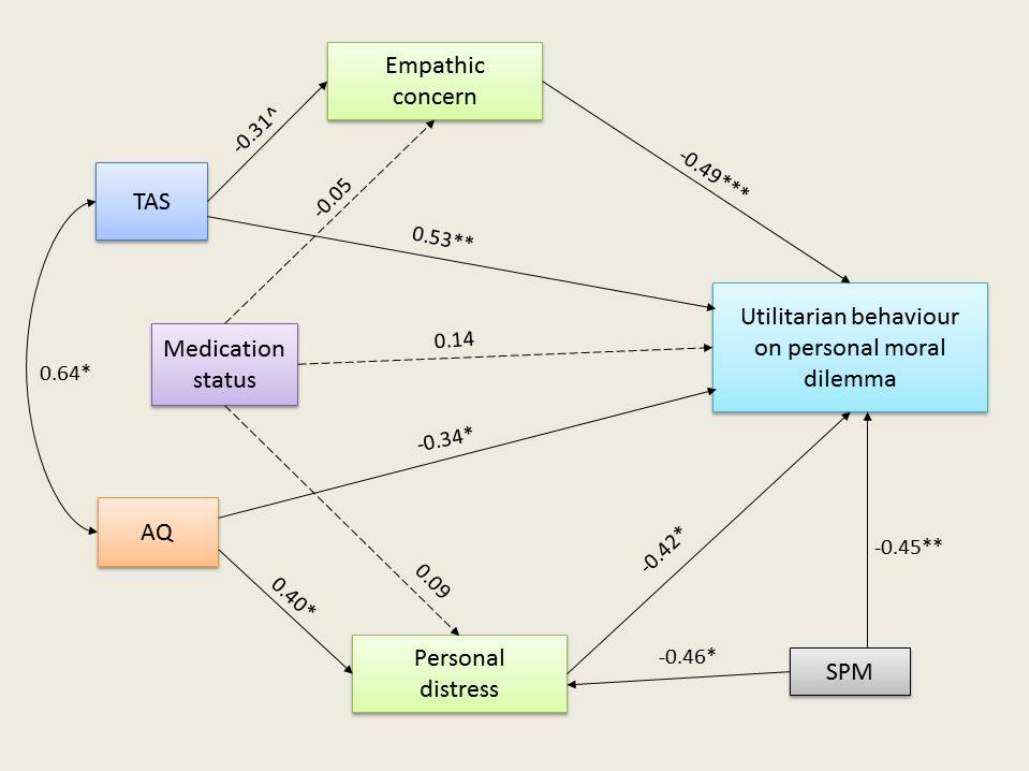


Figure 5: *Path diagram from the path analysis model for utilitarian moral judgment.*

The path analysis model showing the divergent influences of autistic and alexithymic traits on utilitarian moral judgments on personal moral dilemmas in the ASD group, mediated by empathic concern and personal distress components of trait empathy. Additional variables accounted for effects of medication status (some autistics were consuming medication (= 1), while some were not (= 0)) and non-verbal reasoning scores (as assessed by Raven's SPM). Values shown are standardized parameter estimates (betas). Although not shown in the figure, all endogenous variables are associated with errors. Solid lines represent significant relationships between predictors and the criterion variables, while dotted lines represent no significant relationship. Asterisks indicate significance of paths (^$p < 0.1$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, all one-tailed).

The model for behaviour question had a moderate fit ($\chi^2(9) = 10.007$, $p = 0.350$, $\chi^2/df = 1.112$, RMSEA = 0.089, 90% CI [0, 0.322], p for close fit = 0.378, CFI = 0.960, PCFI = 0.411). Together, the independent variables accounted for 69.5% of all variance (R^2) in utilitarian moral judgments (for more details about betas from path analysis, see Appendix (Text S12)).

As predicted, we found that once shared variance between autistic and alexithymic traits was accounted for, alexithymic traits exhibited increased affinity for personally carrying out the necessary harmful actions and autistic traits were associated with reduced tendency to endorse the utilitarian option. Furthermore, the influence of these two traits on moral judgments was mediated by dissociable components of empathy: (*a*) increased alexithymia score was associated with reduced dispositional empathic concern for others' welfare, which itself was associated with increased tendency to endorse utilitarian solution; (*b*) greater severity of autistic traits was associated with empathic hyperarousal in response to demanding social situation, which itself predicted reduced tendency to engage in harmful behaviour. Furthermore, greater capacity to reason non-verbally was also associated with reduced utilitarian behaviour, arguably due to developmentally acquired compensatory strategy of rigid norm-compliance.

Note that we did not carry out a similar path analysis with the control population because there was less amount of variation in personality traits (as compared to the ASD population; see Table 1) to detect such subtle array of interactions between these traits (as assessed by Levene's test, e.g., TAS: $F(1,32) = 5.359$, $p = 0.027$; personal distress: $F(1,32) = 6.424$, $p = 0.016$). Future studies can explore the same path model in a large control population with enough variation in the data to detect such interactions.

Since the estimates of the parameters are unstable in path analysis (Streiner, 2005) when the sample sizes are too small (like in the current study), we also assessed validity of the key results using a simpler model in a hierarchical regression analysis (Brewer et al., 2015) (for full details, see Supplementary Data (Text S13)). This analysis also revealed that after controlling for age, gender, and depression and after accounting for shared variance between autistic and alexithymic traits, severity of autism was associated with reduced utilitarian tendency (β = -0.701, p = 0.019), while alexithymia was predictive of increased utilitarian inclination (β = 0.840, p = 0.006).

4. Discussion

Despite a large body of work investigating role of alexithymia in emotional processing deficits in autism (Bird & Cook, 2013), its role in autistics' moral cognition remains to be thoroughly explored. Moral cognition lies at the heart of interpersonal interactions and thus it in important to investigate this aspect of autistic cognition. In the current study, we explored moral evaluations in autistic participants on hypothetical, emotionally charged moral dilemmas that assessed their behavioural tendency to physically carry out harmful actions to avoid greater harm from occurring. Three primary results emerged from the current investigation. First, adults with ASD could properly distinguish between emotionally aversive personal dilemmas from impersonal dilemmas and endorsed behavioural choices that were comparable to controls. Second, autistic and alexithymia traits were associated with opposite utilitarian inclinations due to dissociable roles of self-oriented unease and other-oriented feelings of concern. Third, autistics relied on their intact non-verbal reasoning skills while making normative choices, probably to compensate for their other deficits in the interpersonal domain.

Preserved utilitarian moral judgments in autism

As in healthy controls, autistic participants perceived making hypothetical choices on morally dilemmatic situations to be more emotionally arousing than finding solutions to practical problems and were more ready endorse utilitarian option on impersonal as compared to personal moral dilemmas. Moreover, ASD participants found all conditions to be more arousing than controls, which comports well with prevalent negative arousal states reported in literature on autism (Capps et al., 1993; Samson et al., 2015; Smith, 2009). Remarkably, this elevated negative emotional arousal and social and emotional processing deficits notwithstanding, not

only did the autistic participants not show previously observed (Gleichgerrcht et al., 2013) utilitarian bias, they exhibited increased tendency to *reject* the utilitarian option on emotionally salient dilemmas that required direct physical harm to a victim (e.g. pushing someone to their death). Our proposed framework premeditated such pattern of response based on a web of mutually conflicting influences of various subdimensions of autistic personality on first-hand, hypothetical moral choices.

Dissociable empathy-utilitarianism associations between autistic and alexithymic traits

There is plenty of evidence to support the claim that emotions motivate individuals to reject harmful transgressions, even if such actions are necessary to stave off harm of bigger magnitude (Greene, 2014). Recent research also sheds light on the exact nature of psychological processes that constitute this negative affect (Miller et al., 2014): aversion to harmful outcome (e.g. victim suffering) and aversion to the nature of harmful action itself (e.g. sensorimotor properties of the action). But the motivations subserving rejection of actions with harmful outcomes are of two varieties (Sarlo et al., 2014): self-oriented personal distress and other-oriented empathic concern. Accordingly, since autistic traits are associated with increased personal distress during demanding interpersonal interactions (as shown by both self-reported ratings (Dziobek et al., 2008; Rogers et al., 2007) and hemodynamic responses (Y.-T. T. Fan et al., 2014; Gu et al., 2015)), we reasoned that their moral judgments would be influenced by this emotional bias *against* the utilitarian option(Sarlo et al., 2014). On the other hand, since alexithymic traits are associated with reduced empathic concern for others' wellbeing (as shown by both self-report (Aaron et al., 2015; Grynberg et al., 2010; Patil & Silani, 2014b) and neuroimaging evidence (Bird et al., 2010; FeldmanHall, Dalgleish, & Mobbs, 2013)), they would be more likely to evaluate prospect of personally harming someone in a hypothetical scenario *in favour of* the

utilitarian solution (Gleichgerrcht & Young, 2013). Thus, given the prevalence (Berthoz & Hill, 2005; Griffin et al., 2015; Hill et al., 2004; Salminen et al., 1999) of alexithymia in ASD (in the current sample: 47%), we expected these dissociable empathic motivations mediating influences of autistic and alexithymia traits to cancel each other out leaving the final moral judgment unimpaired. This is exactly what was observed in the data, as shown by its fit to this theoretically-constructed path model (Figure 5): egoistic motivation to reduce personal distress led to reduced utilitarian tendency for autistic traits, while reduced altruistic motivation to prevent harming others led to increased utilitarian proclivity for alexithymic traits. This model reveals that the spared moral capacity in autism to evaluate hypothetical harmful behaviours was a result of cancellation of opposite influences that are scaffolded on emotional biases introduced by dissociable empathic profiles of autistic and alexithymic traits. Thus, the current findings shed light not only on the different aspects of emotional empathy that autistic and alexithymic traits are associated with but also on how these traits relate to moral judgments.

We note that the current findings are in conflict with a prior study (Brewer et al., 2015) that investigated role of alexithymia in moral acceptability of emotion-evoking statements (e.g., "I could easily hurt you" (fear), "I never wash my hands" (disgust), etc.) and found that alexithymia was predictive of acceptability judgments only in controls but not in ASD and concluded that autistics' judgments were based on complying with social rules and were less susceptible to emotional biases. It is possible that these differences stem from emotional saliency of the stimuli used across studies; moral dilemmas involve situations where the individuals have to mull over behavioural choice of directly harming or even killing someone for the benefit of the many and are, thus, inherently highly emotionally evocative (Greene, 2014), while providing more objective acceptability judgments about emotional sentences may not engage emotional

processes to the same extent (Patil et al., 2014; Tassy et al., 2013). Another possibility is that there was not enough variation in alexithymia scores in the ASD group to detect an effect (indeed, variance in alexithymia scores in the control group was higher than in the ASD group in this study).

Compensatory intellectual strategies in autism

Despite their social impairments, both children and adults with autism still manage to acquire knowledge about normative canon consisting of appropriateness of various moral behaviours (Gleichgerrcht et al., 2013; Zalla et al., 2011). For example, they can properly distinguish between moral norms that relate to suffering in victims from social conventions that are context-bound societal rules (Blair, 1996; Leslie et al., 2006; Shulman et al., 2012; Zalla et al., 2011). Although neurotypical individuals justify such distinction by referring to considerations about emotional consequences for the victim, the justifications provided by autistics tend to lack such empathic discourse and involve more rule-based rationale (Shulman et al., 2012; Zalla et al., 2011). It is possible that in the absence of recourse to strong moral intuitions, autistics developmentally acquire compensatory strategies (Frith, 2004) that rely on spared intellectual abilities; indeed research in moral development showing that children with intellectual disabilities lag behind their typically developing peers in terms of moral reasoning (Langdon, Clare, & Murphy, 2010) provides circumstantial evidence for this claim. These can enable them to make such normatively significant distinctions by conforming to normative rules, sometimes in an inflexible and stereotyped manner (Baron-Cohen et al., 2003) which can make them adopt even harsher criterion for moral evaluations (Li et al., 2014; Zalla et al., 2011). Accordingly, prior studies show that autistics exhibit a more rigid, rule-based profile to *justify* their moral choices on such tasks (Shulman et al., 2012; Zalla et al., 2011) and enhanced verbal intelligence

is predictive of quality of such justifications (Barnes, Lombardo, Wheelwright, & Baron-Cohen, 2009; Grant, Boucher, Riggs, & Grayson, 2005), but these studies did not investigate role of such intellectual capabilities in moral *judgments*.

In the current study, we found that even after accounting for variance associated with autistic and alexithymic traits, non-verbal IQ was negatively predictive of utilitarian moral judgments. Thus, it is possible that autistics relied on non-verbal reasoning to reject the proposition of directly causing harm to others. For example, instead of retrieving semantic representations (e.g. for personal dilemma (Greene et al., 2004), it can be "ME HURT YOU = WRONG"), they can rely on visual imagery of the same rule, which has indeed been shown to support non-utilitarian moral judgments in healthy individuals (Amit & Greene, 2012). Prior studies support this line of reasoning, e.g., a previous neuroimaging study (Carter, Williams, Minshew, & Lehman, 2012) showed that typically developing children automatically encode their social knowledge into language while assessing behaviour of others in paradigms with minimum verbal requirements, but no such pattern is found in autistic children. Anecdotal reports from autistic individuals also note that they primarily rely on non-verbal thoughts (Hurlburt, Happé, & Frith, 1994) (as one autistic noted (Carter et al., 2012): "I think in pictures. Words are like a second language to me…. When somebody speaks to me, his words are instantly translated into pictures."). The current findings are also consistent with the prior findings that show - (*i*) verbal IQ is correlated with *justifications* but not the moral *judgments* in children with ASD (Barnes et al., 2009; Grant et al., 2005), (*ii*) no correlation between verbal IQ and utilitarian moral judgments in ASD (Gleichgerrcht et al., 2013), and (*iii*) some moral principles operative in moral evaluations seem to be inaccessible during conscious moral reasoning and seem to operate intuitively and are, thus, difficult to verbalize (Cushman et al., 2006).

Therefore, we maintain that the current findings hint at non-verbal intelligence as a compensatory strategy that high-functioning autistics rely on while endorsing moral choices that are in line with prevalent socio-moral norms. Although a prior study implicated intellectual abilities in forming compensatory strategies to perform a task in the perceptual domain (Rutherford & Troje, 2012), no study thus far has investigated the same for the social domain and future hypothesis-driven studies should investigate the effect observed in the current study further.

5. *Implications*

Current investigation underscores the importance of studying various aspects of cognition in clinical populations, even if they do not exhibit any visible deficits on the task being studied. More specifically, the current study raises a methodological concern for studies investigating moral cognition (especially in the harm domain) in clinical populations that have unusually high incidence rate of alexithymia (Bird & Cook, 2013) (e.g., schizophrenia (de Achával et al., 2013), multiple sclerosis (Gleichgerrcht et al., 2015), Parkinson's disease (Fumagalli et al., 2015), etc.): all such studies should account for effects of co-occurring alexithymia on moral evaluations.

6. *Limitations*

Validity of the conclusions drawn from the current study is contingent upon the following limitations. The primary limitation of the current study was the sample size, which was relatively small for the complexity of the statistical model investigated. Although we demonstrated validity of the main results in a separate regression analyses, future studies can explore various hypotheses stemming from the current investigation in a bigger sample (even in healthy population). Another limitation of the current study is the use of IRI to measure various

components of empathy since the IRI items measuring empathic concern and personal distress do not seem to map well onto recent social neuroscience conception of empathy (Ugazio et al., 2014) and also has psychometric problems (Koller & Lamm, 2015). Thus, the current findings should be replicated with other empathy measures. Additionally, the moral dilemma task has recently been criticized (Kahane, 2015) to have contexts that are too contrived and extreme to provide any cues about social behaviour in everyday life-like situations. We note though that such unfamiliar settings are especially helpful to shed light on processes that may not be robustly recruited while judging more mundane situations that can be resolved by easily accessible social rules (Christensen & Gomila, 2012). Future studies can explore the role of alexithymia in reduced prosocial sentiments in autism using a more ecologically valid paradigm (e.g. 'Above and Beyond' task (Jameel et al., 2014)), since this reduction in prosocial behaviour can be due to alexithymia (FeldmanHall et al., 2013). Another limitation is that the current study used a single moral judgment parameter that treats utilitarian and deontological tendencies as inversely related to each other and conflate disregard for deontic prohibitions and endorsement of utilitarian principles and future studies should use process dissociation approach to study these separable appraisals (Conway & Gawronski, 2013). Lastly, the diagnosis of autism was partially based on gold standard diagnostic instruments for ASD such as the Autism Diagnostic Interview – Revised (Lord, Rutter, & Le Couteur, 1994) (ADI-R) or the Autism Diagnostic Observation Schedule (Lord et al., 2000) (ADOS) because these documents were not available for all participants and, therefore, an additional inclusion criterion was based on AQ-k. Future studies should attempt to include these standard diagnostic instruments as well.

Chapter 3

The role of empathy in moral condemnation of accidental harms and moral luck: An fMRI investigation

Abstract

A number of past studies have investigated neural basis of the capacity for mental state reasoning (i.e., reasoning about beliefs and intentions) in drawing a moral distinction between intentional and accidental harms and also condemning agents unsuccessfully acting with intent to harm. Less attention has been paid to the role of empathic reasoning (i.e., reasoning about pain and emotions) in condemning accidental harm-doers. Additionally, past work shows that mere presence of harmful outcome amplifies condemnation (known as moral luck effect) and this effect is stronger for blame than acceptability judgments and possible role of empathy in this phenomenon remains unstudied. The current investigation focused on these questions revealed two important results. One, participants who exhibited greater magnitude of activity in posterior insula while reading information about harmful outcome (reflecting encoded intensity of victims' pain) condemned accidental harm-doer more severely. Second, judgments about attributing blame to agents relied more on empathic assessment of the victim than making acceptability judgments, as reflected in the anterior insula activity and its increased functional connectivity with dorsolateral prefrontal cortex while making blame judgments. Thus, the current results shed light on the role of empathy in condemning unintentional harms and its role on mediating influence on greater relevance of harmful outcomes for blame as compared to other types of judgments.

1. Introduction

On 15 February 2012, the Italian oil tanker *MT Enrica Lexie* was travelling in international waters, off the Indian coast, when the two Italian marines aboard noticed another ship nearby. They falsely believed it to be a pirate ship and opened fire, killing two Indian fishermen on board. Unsurprisingly, this incidence led to big diplomatic fallout between the two countries involving complicated legal jurisdiction and functional immunity. But, more interestingly from a psychological perspective, the public opinion in India differed widely with some focusing on the innocent intentions and mistaken beliefs of the marines, while others focusing on the disastrous outcome involving loss of two lives. Additionally, although the citizens of India understood that it was morally acceptable to shoot at another vessel in self-defense under the false belief about threat, they still could not curb their punitive instinct to see the marines punished for negligently killing the fishermen.

This incidence nicely illustrates two features of the human moral mind we will be focusing on in the current study: (*i*) there are inter-individual differences, on a behavioral and neurophysiological level, in forgiving third-party unintentional harms; (*ii*) the presence of harmful outcomes has a greater influence on punishment than wrongness judgments.

1.1 *Neural basis of two-process model for third-party moral judgments*

The two-process model for intent-based moral judgments (Cushman, Sheketoff, Wharton, & Carey, 2013; Cushman, 2008, 2015a) posits two independent computational processes, each of which is capable of providing judgment based on separate analyses of the situation: (*i*) a causal reasoning process active in the presence of a harmful outcome (victim suffering) that provides

evaluation based on the analysis of the agent's causal role in producing such outcome ("causal responsibility = bad"); (*ii*) an intent-based reasoning process that condemns the agent in the presence of a culpable mental state ("malicious belief/desire/intent = bad").

Based on past research, it can be argued that these two systems can rely for their inputs from two distinct routes involved in understanding other minds with dedicated neurocognitive mechanisms: empathy and theory of mind.

On the other hand, *empathy* involves ability to understand and share others' affective states (emotions, pain, etc.) in isomorphic manner while maintaining self-other distinction[2] (de Vignemont & Singer, 2006). It has been consistently shown that perceiving others in pain activates a cluster of brain regions, known as the pain matrix, that encode nociceptive information while one is experiencing pain first-hand (Lamm, Decety, & Singer, 2011). The pain matrix consists of two distinct yet interacting areas (Peyron, Laurent, & García-Larrea, 2000) coding for the sensory-discriminative component of the painful *stimulus* (location, intensity, and duration) and the affective-motivational component of the painful *experience* (unpleasantness, negative affect). The former primarily consists of the somatosensory cortices (S1, S2) and the bilateral posterior insula (PI), while the latter consists of the bilateral anterior insula (AI), the dorsal anterior cingulate cortex (dACC), and the anterior middle cingulate cortex (aMCC). Meta-analytic evidence shows that witnessing or imagining someone else's pain activates neural representations coding primarily for affective-motivational, but not the sensory-discriminative, feature of the experience during direct pain perception (Lamm et al., 2011). In other words,

[2] Note that this definition of empathy distinguishes it from other related but distinct concepts like empathic concern, sympathy, or compassion (Gonzalez-Liencres et al., 2013) as these constructs represent prosocial feelings - driven by empathy - that are not congruent with another's affective state and induce caring and comforting behavior (e.g., we can be concerned about a grieving friend without sharing her grief).

empathizing with others leads to sharing affective consequences (like subjective unpleasantness) of this experience but not the full-blown nociceptive episode.

On the other hand, *Theory of mind* (ToM) entails abstract inferential process via which we think about others' thoughts that have some representational content, e.g. beliefs, desires, knowledge, intentions, etc. (Koster-hale & Saxe, 2013). This capacity is neurally implemented in a specific network consisting primarily of bilateral temporoparietal junction (TPJ), sections of medial prefrontal cortex (mPFC), and precuneus (PC).

Although output from these systems do not conflict with each other when both pieces of information are mutually coherent (for neutral cases and intentional harm cases), conflict arises when the two systems provide different judgments (in case of accidental harm). The final judgment for such conflictual cases is the result of competitive interaction between these antecedent evaluations and it depends on the relative weight (which is itself determined by underlying personality traits; Prehn et al., 2008) given to the output from each system (Buckholtz et al., 2015; Young, Cushman, Hauser, & Saxe, 2007).

Mental-state reasoning process: There is plenty of evidence which shows that, when faced with third-party harm-norm violations, both older children and healthy adults assessments overwhelmingly predicate on the information about mental state (Alter, Kernochan, & Darley, 2007; Baird & Astington, 2004; Cushman, 2008; Gummerum & Chu, 2014). Behaviorally, individuals tend to forgive accidental harm-doers based on their benign intentions, while they condemn attempted harms based on malicious intentions despite the non-occurrence of harmful outcome (Cushman, 2008). At the neural level, rTPJ has been shown to be the most important ToM node for mediating mental state attributions during moral judgment (for a review, see

Young & Tsoi, 2013). The rTPJ exhibits greatest magnitude of hemodynamic response while condemning attempted harm cases, where the perpetrator intends but fails to harm someone and thus the condemnation relies heavily on intent information (Young et al., 2007; Young & Saxe, 2008), and disrupting activity in rTPJ leads to reduced severity of condemnation for attempted harms (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). Additionally, neurological patients (Baez et al., 2014; Ciaramelli, Braghittoni, & di Pellegrino, 2012; Young, Bechara, et al., 2010) and sadistic individuals (Trémolière & Djeriouat, 2016) with a reduced emotional response to harmful intent tend to have more favorable assessment of attempted harm cases. Note that no conflict between the two systems arise in case of condemning attempted harm cases because the causal reasoning process remains silent in the absence of harmful outcome and the intent-based process operates unabated to condemn the victim (Cushman, 2008; Young et al., 2007).

Compared to condemning attempted harm cases, a more difficult situation arises when one needs to forgive unintentional harms, where the harm-doer causes bad outcome while acting under false belief. Forgiving accidental harm thus requires a robust representation of innocent intent that can counteract prepotent tendency to condemn the actor based on harm assessment. Accordingly, individuals with higher magnitude of activity (Young & Saxe, 2009) and greater differentiation in spatial pattern of activity (Koster-Hale et al., 2013) in rTPJ distinguishing intentional from unintentional harms – which can happen as early as 62 ms post-stimulus (Decety & Cacioppo, 2012) - tend to forgive accidental harms more by down-regulating emotional arousal in response to harm (Treadway et al., 2014). Additionally, stimulating this patch of cortex increases role of belief information in moral judgments, as shown by reduced severity of evaluations for accidental harms (Sellaro et al., 2015; also see Ye et al., 2015), while

reducing the role of mental state information by cognitively exhausting participants leads to the opposite pattern (Buon, Jacob, Loissel, & Dupoux, 2013). Populations showing impaired mental state inference, like autism spectrum disorder, show abnormal pattern in rTPJ (Koster-Hale et al., 2013) and are thus less likely to exculpate accidental harm-doers (Baez et al., 2012; Buon, Dupoux, et al., 2013; Moran et al., 2011).

Causal reasoning process: Although the role of mental state inference in condemnation of attempted harm and exculpation of accidental harm abounds, much less attention has been paid to the neural basis of causal-system based condemnation of accidental harm cases. Causal analyses of accidental harm begins with the detection of harmful outcome and searches for an agent who can be held responsible in the causal model of the event (Sloman, Fernbach, & Ewing, 2009) and be condemned for causing suffering in the victim. Thus, it is possible that the degree to which individuals pay attention to the causal role of actors who accidentally produce negative outcome is in turn determined by the extent to which they empathize with the victim. In other words, understanding and feeling victim distress can motivate individuals to condemn accidental harm-doers more by implicating them based on their causal involvement. Prior studies do reveal that people's causal judgments are impacted by motivational factors. An agent who acts in a way judged to be immoral is ascribed more causality than one who abides by moral norms (Alicke, 1992), e.g. a driver who runs over someone in an accident is held to be less causally responsible when he is rushing home to hide gift for his parents (not blameworthy) than to hide stash of drugs (blameworthy).

There is some indirect evidence to support this claim. While evaluating harmful situations, participants spend greater time looking at the victim than the perpetrator and exhibit increased activity in the empathy network (Decety, Michalska, & Kinzler, 2012). Individuals who score

high on self-report measures of dispositional empathy are more inclined to condemn accidental harms (Trémolière & Djeriouat, 2016, Study 1). Also, individuals with a certain genetic variation of oxytocin receptor gene that predisposes them to be more empathic are more reluctant to exculpate accidental harm-doers (Walter et al., 2012). Subclinical (e.g., alexithymia; Patil & Silani, 2014) and clinical (e.g., psychopathy; Young, Koenigs, Kruepke, & Newman, 2012) personalities characterized by reduced empathic concern for others also exhibit increased tendency to forgive accidents, arguably because they are less motivated to hold the agent causally responsible in the absence of strong empathic aversion. The current study explores the neural basis of this behaviorally observed empathic condemnation of unintentional harms and explores if empathic arousal can be another motivational factor that influences causal system output.

1.2 *Acceptability vs. blame judgments and moral luck*

A further complication is introduced by recent work suggesting that not all types of moral judgments are equivalent and they can be distinguished based on their evaluative foci (Cushman, 2015a; Malle, Guglielmo, & Monroe, 2014). The wrongness/acceptability/permissibility (henceforth, represented only by the term *acceptability*) represent a class of judgments that are concerned with the evaluation of *actions* (or *action plans*) with respect to a norm system and functions as a way to *declare* that the behavior is incongruent with mutually agreed upon moral norms (e.g. "Do not harm others"). On the other hand, blame and punishment together represent a class of judgments that focus on evaluating *agents* for their involvement in norm violating events and functions as a social mechanism to *regulate* their behavior in order to deter repetition of such behavior in future.

Although adult human moral judgment is primarily modulated by information about an intent, outcomes play a substantial role too (Cushman, 2008) and plenty of psychological research provides evidence for this "outcome bias" in lay judgments (Berg-Cross, 1975; Cushman, Dreber, Wang, & Costa, 2009; Cushman, 2008; Mazzocco, Alicke, & Davis, 2004). More importantly, outcomes matter to a different degree for different classes of moral judgments: acceptability judgments exhibit lesser sensitivity to outcome information as compared to blame/punishment judgments (Cushman, 2008). A recent study attests to this putative differential reliance of acceptability[3] and punishment judgments on subcomponent processes; in particular, the authors found that reducing cortical excitability of an area involved in integration of belief and outcome information (dorsolateral prefrontal cortex, dlPFC) selectively impaired punishment but not acceptability judgments (Buckholtz et al., 2015). Thus, evaluation of acceptability of agent's moral behavior primarily relies on assessment of actor's mental state during the act and on determining culpability of this state with respect to normatively acceptable code of conduct, while the blame/punishment for the agent additionally involves appraisal of whether harm occurred, severity of harm caused, and actor's causal involvement in production of harm (Cushman, 2015a). More concretely, in the example discussed above, although people are willing to deem behavior of marines who fired shots at another vessel - mistakenly believing it to be a pirate ship - as acceptable, they expect marines to be blamed and punished only when this action leads to death of innocent fishermen but not otherwise.

[3] Note that although the authors in the original study use the term "blameworthy" to describe the judgment we are referring to here as "acceptability," we note that the question participants answered was "'Please indicate how morally responsible [the agent] is for his actions described in the scenario." and not "How much blame does the agent deserve?". Additionally, past research shows that lay judgments do not treat responsibility and blame judgments the same way (Guglielmo, 2015).

Philosophers call the phenomenon whereby the mere presence of bad outcome contributes to moral evaluations of otherwise identical actions as *moral luck* (Nagel, 1985). Although there is a wide variation in the degree to which lay individuals endorse moral luck as a normative moral principle, their moral judgments are found to be amenable to it nonetheless (Lench, Domsky, Smallman, & Darbor, 2015). Thus, this body of research reveals that moral luck matters more for blame/punishment judgments than acceptability judgments because the causal system contingent on harmful outcome has a greater bearing on such judgments. This asymmetric reliance on outcomes while deciding on blame/punishment (vis-à-vis acceptability) has convincingly been argued to be an upshot of the ultimate evolutionary function of blame/punishment (Cushman, 2013b, 2015b; Martin & Cushman, 2016), which is to utilize the learning capacity of social partners to modify their harmful behavior - even if it was unintended – by being more careful in the future. At the mechanistic level, however, this is implemented via inflexible moral outrage at the harm-doer for the harm s/he caused without any conscious computation of its adaptive value (Martin & Cushman, 2016).

Although this literature demonstrates that outcomes - and, in turn, moral luck - matter more for blame judgments than acceptability judgments, it remains to be studied how this is implemented at the neural level and the exact psychological chassis that supports this effect. One previous study (Young, Nichols, & Saxe, 2010) revealed that people are driven to judge accidental harm-doers to be more blameworthy not because someone got hurt, but primarily because false beliefs held by actors are deemed to less justified ("it is not reasonable to believe that the other ship is a pirate ship when there are no overt signs to suggest so"). Note that although this study shows that moral luck partly stems from mental state assessments, it does not explain why blame judgments are more susceptible to influence of harmful outcomes than acceptability judgments. After all,

agent's false beliefs are as unjustified while assessing acceptability of their behavior as attributing blame to them. Thus, none of the existing data sheds light on the neural substrates that mediate influence of moral luck on different types of judgments. One possible source of this effect is empathy: neurobiological models of punishment posit that a suffering conspecific is a source of negative arousal in the observer (aversive excitator) and elicits an inflexible, Pavlovian-like response to blame/punish the agent (Cushman, 2013b; Seymour, Singer, & Dolan, 2007). Thus, it is possible that moral luck matters more for blame/punishment as compared to acceptability judgments because such judgments rely to a greater degree on empathic assessment of the victim. A recent behavioral study (Patil, Young, et al., 2016) provides some preliminary evidence to support this hypothesis. This study found that when inter-individual differences in self-reported cognitive empathy (perspective-taking subscale of IRI, to be precise) in healthy adults is accounted for, moral luck no longer influences punishment judgments to a greater degree than acceptability judgments.

Combining these insights with the two-process framework, we hypothesize that people are more likely to blame accidents than judge them to be wrong as compared to neutral cases because of the greater reliance on output from causal analyses of perpetrator's involvement, which itself is provoked by empathic assessment of the victim. In other words, we predict that areas involved in empathizing with others' suffering would be more active while assigning blame to the accidental harm-doers than while evaluating acceptability of their behavior.

2. Methods and Materials

2.1 *Participants*:

A total of 50 healthy community members (32 female) without any history of neurological problems were recruited to participate in this study and were financially compensated for their time and travel. Average age was 23.06 years (SD = 3.08), with a range of 18 to 35. All participants provided written informed consent and the study was approved by the local ethics committee. All data from one participant was excluded from the final analysis as he was consuming clinically-prescribed psychoactive drugs and did not divulge this information in pre-scanning telephone interview. Functional data from two participants was removed due to excessive head motion (see below) and data from one additional participant could not be collected due to technical error. Thus, functional data was available for 46 participants, while behavioral for 49.

2.2 *Experimental stimuli and procedure*:

Moral judgment task: Experimental stimuli were text-based scenarios. Stimuli consisted of four variations of 36 unique scenarios for a total of 144 stories. All scenarios were primarily taken from previous studies (Cushman, 2008; Young, Camprodon, et al., 2010) and were adapted in Italian (see Appendix Text S1 for more details). The four variations were the result of a 2 × 2 within-subjects design where the factors *belief* (neutral, negative) and *outcome* (neutral, negative) were independently varied such that agents in the scenario produced either a neutral outcome or a harmful outcome while acting with the belief that they were causing either a neutral outcome or a harmful outcome. Each possible belief was true for one outcome and false for the other. Each participant saw one variation of each scenario, for a total of 36 stories. All scenarios

were equivalent in word count across all four variations (*p*s > 0.05). A number of factors that have been varied in previous studies were hold constant in the current 2-by-2 design (for more, see Appendix Text S2).

Each scenario lasted for 32 s and consisted of four cumulative segments (each lasting for 8 s): (i) *background:* this stem was common to all variations and provided settings in which the story took place; (ii) *foreshadow:* this segment foreshadowed whether the outcome will be neutral or harmful; (iii) *mental-state information*[4]*:* this segment provided information about whether the agent was acting with a neutral or harmful belief; (iv) *consequence*[5]*:* this final segment described agent's action and its outcome. All story text was then removed and replaced with the question and response scale (see Figure 1). Note that all scenarios provided information only about beliefs with which the agents acted and the intent (harmful or neutral) had to be inferred. After reading each story, participants provided two types of moral judgments (Cushman, 2008) which were presented in randomized order:

[1] *acceptability* - "How morally acceptable was [the agent]'s behavior?" (1: *Completely acceptable* to 7: *Not at all acceptable*);

[2] *blame* "How much blame does [the agent] deserve?" (1: *None at all* to 7: *Very much*).

Each question lasted for 6 s and participants could provide their judgment using a 7-point Likert scale on which cursor could be moved using two fingers. The location at which the cursor

[4] We use the term *mental-state information* instead of *belief* to avoid confusion as the latter term represents one of the factors of the experimental design while the former represents a story segment containing information about the mental state of the agent while acting.

[5] We use the term *consequence* instead of *outcome* to avoid confusion as the latter term represents one of the factors of the experimental design while the former represents a story segment containing information about the nature of the outcome.

initially appeared on the scale was chosen at random on each trial to make sure that there were no systematic differences across conditions in terms of the required cursor movement, as this could have confounded effects of interest with movement-related activity (especially in ROIs like r-AI, cf. Mutschler et al., 2009). The response buttons were active as long as the question remained on the screen and so participants could move the cursor to one position and could later change it again to a new position. Note that this meant that we could not collect any meaningful response time data. After each scenario, participants viewed a fixation cross on the screen for a jittered ITI of 2-4 seconds. Additional details about the experimental protocol are provided in Appendix Text S3.

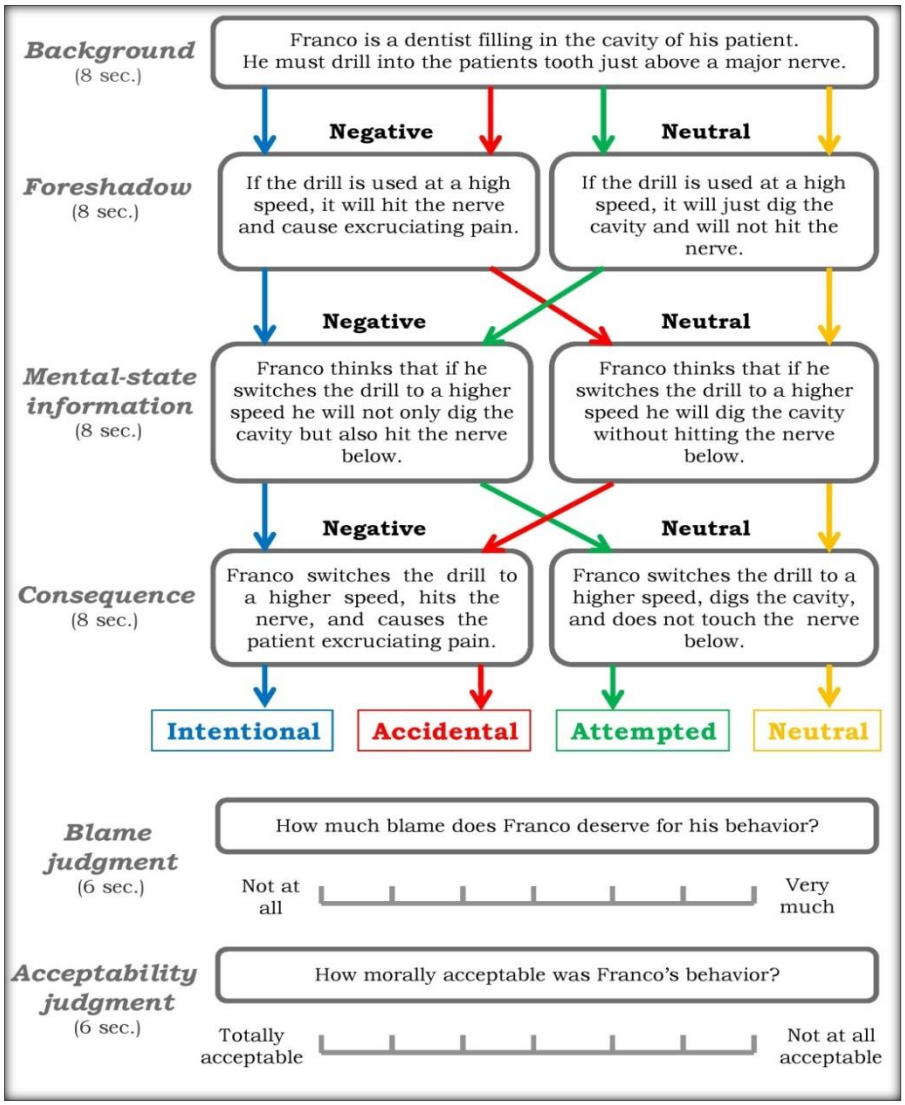


Figure 1. Experimental stimuli and design. Each moral vignette consisted of the following text segments (each lasting for 8 seconds): a *background* stem providing set-up for the story, a *foreshadow* segment that foreshadowed the nature of outcome, a *mental-state information* segment that provided information about actor's belief, a *consequence* segment that described action and its outcome. These segments were then followed by questions assessing acceptability and blame judgments (each lasting for 6 seconds and presented in random order) that participants had to respond to using a 7-point Likert scale.

*Functional localizer task***:** To localize functional empathy network in participants, we used the task from a prior study (Lamm, Batson, & Decety, 2007). Participants were told that they would be witnessing videos of people experiencing painful auditory stimulation. As opposed to the instructions provided in the original study, we did not tell participants that this stimulation was part of a medical treatment, as we suspected that this could have led to down-regulation of the empathic response (Lamm et al., 2007). Each participant was shown 18 videos, each lasting for 3 s, featuring one individual (male or female) wearing headphone. The video showed these individuals displaying the transition from neutral facial expression (0.5 s) to exhibiting painful facial expressions triggered by auditory stimulation (2.5 s). After each video, participants responded to two questions: one assessing *other-oriented* empathic response by gauging intensity of the experienced pain ("How painful was this stimulation for this person?"; -3: *not at all painful* to 3: *extremely painful*), while the other assessing *self-oriented* distress via experienced unpleasantness ("How unpleasant was it for you to watch this person suffering?"; -3: *not at all unpleasant* to 3: *extremely unpleasant*) on a 7-point Likert scale. Mean inter-trial interval (ITI) was 2 s and was randomly jittered (jitter range: 0-2 s) to reduce predictability of the stimuli presentation (for schematics of the task design, see Appendix Text S4).

2.3 *fMRI data acquisition and preprocessing*:

All fMRI scans were acquired using a 3T Philips Achieva scanner at the Hospital 'Santa Maria della Misericordia' (Udine, Italy), equipped with an 8-channel head coil. High-resolution structural images were acquired as 180 T1-weighted transverse images (0.75 mm slice thickness). Functional images were acquired in interleaved manner using a T2*-weighted echoplanar imaging (EPI) sequence with 33 transverse slices covering the whole brain with the following parameters: slice thickness = 3.2 mm; interslice gap = 0.3 mm; repetition time (TR) =

2000 ms, echo time (TE) = 35 ms; flip angle = 90º, field of view = 230 × 230 mm^2; matrix size = 128 × 128, SENSE factor 2. The slices were oriented at a 30º oblique angle to the AC-PC. This slice prescription was selected for optimization of BOLD signal (by reducing drop-out effects caused by the air-tissue interface) in the orbitofrontal cortex (based on recommendations by Weiskopf, Hutton, Josephs, & Deichmann, 2006).

Data were analyzed with SPM12 (www.fil.ion.ucl.ac.uk/spm/software/spm12). Each subject's data were motion-corrected (outliers were detected using Art toolbox; see Appendix Text S5) and then normalized onto a common stereotactic space (the Montreal Neurological Institute template). Data were then smoothed by using a Gaussian filter (full width half maximum = 6 mm at first-level), and high-pass-filtered.

2.4 *fMRI data analysis at first-level*:

For each participant and for each task, the design matrices for fixed-effects General Linear Model were constructed by convolving a canonical hemodynamic response function or HRF with the stimulus function for events (boxcar function) to create regressors of interest along with its temporal and dispersion derivatives. For more details, see Appendix Text S6.

Moral judgment task: For the main task, there were 72 regressors of interest (with additional nuisance regressors) from a 6 (text segment: background, foreshadow, mental-state information, consequence, acceptability question, blame question) × 2 (belief: neutral, negative) × 2 (outcome: neutral, negative) × 3 (type of HRF: canonical, time derivative, dispersion derivative).

Functional localizer task: In the first-level design matrix for empathy localizer task, there were 3 regressors of interest corresponding to the informed basis set convolved with the event of

witnessing empathy-eliciting videos. There were 6 additional regressors for events involving ratings for perceived pain in others and experienced unpleasantness.

ROIs selection: At first level, for each participant, the following ROIs for empathy for pain, based on the localizer task, were defined[6] for both sensory-discriminative and affective-motivational components(Bzdok et al., 2012; Y. Fan et al., 2011; Lamm et al., 2011): bilateral PI, bilateral AI, dACC, and aMCC. (see Table 1; also see Appendix Text S7 for figure) At individual level, not all ROIs could be localized for all participants. For a list of coordinates for all ROIs for each individual, see Appendix Text S8.

Table 1. ROI coordinates from the localizer experiments.

Type of ROI	ROI	Individual ROIs				Whole-brain contrast		
		n	*x*	*y*	*z*	*x*	*y*	*z*
Empathy (> baseline)	dACC	41/49	2	32	23	10	28	26
	l-AI	39/49	-36	13	2	-36	10	0
	r-AI	43/49	37	16	1	40	12	0
	l-PI	31/49	-37	-16	-10	-34	-22	4
	r-PI	28/49	38	-10	4	38	-20	6
	aMCC	44/49	3	5	38	-2	12	42

Note: Average peak voxels for ROIs are in MNI coordinates (in mm). The "Individual ROIs" columns show the average peak voxels for individual subjects' ROIs. The "Whole-brain contrast" columns show the peak voxel in the same regions in the whole-brain random effects group analysis. Results at both subject-and group-level were masked anatomically by neuromorphometrics atlas.

[6] A prior meta-analysis shows that somatosensory cortices (S1, S2) show increased hemodynamic activity in contralateral regions during experiencing painful stimulation administered to one's own hand, while in ipsilateral regions while observing the same body part in pictures (Lamm et al., 2011). Since the current task of interest did not feature any salient information regarding laterality of body parts in pain, we did not include S1 and S2 in our ROIs.

Each subjects' whole brain *F*-contrast image (experimental videos > baseline) was masked with anatomical atlas labels provided by Neuromorphometrics, Inc. (http://neuromorphometrics.com/) under academic subscription. Each ROI was defined by peak voxel of cluster containing more than 10 significantly active contiguous voxels ($p < 0.001$, uncorrected).

ROI data analysis: The data from spherical ROIs with a radius of 8mm was extracted and analyzed using the MarsBar toolbox (v0.44) for SPM (http://marsbar.sourceforge.net/) (Brett, Anton, Valabregue, & Poline, 2002). Within the ROI, the average percent signal change (PSC) was computed relative to the adjusted mean of the time series (for more details, see Appendix Text S6). The responses of ROIs were measured while participants read the *mental-state information* (8 s) and *consequence* (8 s) segments of the moral stories and gave *acceptability* (6 s) and *blame* (6 s) judgments. ROI analyses were not performed for the *background* and *foreshadow* segments from the stories as insufficient information was available at this stage for any morally relevant evaluation to commence. The current study is unique in investigating neural processes subserving moral judgments not only while the belief/intent information is presented (Young et al., 2007; Yu, Li, & Zhou, 2015), but also while information about consequences is provided and integrated into making acceptability and blame judgments.

As recommended (Poldrack, 2007), data defining ROIs was independent from the data used in the repeated measures statistics. Restricting analysis to a few ROIs thus reduced Type-I error by drastically limiting the number of statistical tests performed (Saxe, Brett, & Kanwisher, 2006).

2.5 *Brain-behavior correlations*:

Correlational analysis was carried out to assess relationship between neural activity in each ROI (PSC) for story segments of interest and behavioral response (moral judgments) for all four types

of stories. To avoid false positive brain-behavior correlations, we followed recommended steps (Pernet et al., 2013; Rousselet & Pernet, 2012; Schwarzkopf, De Haas, & Rees, 2012) and computed Spearman's rho as a correlation measure and ran a robustness check (for more, see Appendix Text S9). Note that these correlations were computed between PSC extracted from ROIs which were selected based on independent functional data. This helps us sidestep the nonindependence error (Vul, Harris, Winkielman, & Pashler, 2009) that can lead to spurious correlations and the observed results are thus unbiased and more trustworthy.

2.6 *fMRI data analysis at second-level*:

Moral judgment task: The group-level random effects analyses were conducted for each segment by contrasting the (canonical HRF) beta-weights from each subject's first-level analyses in a single full factorial design generated using a 4 (segment) × 2 (belief) × 2 (outcome) design matrix.

Functional localizer task: The empathy network at group-level was localized by entering beta-weights from all HRF contrasts from first-level in a full factorial design (*F*-contrast).

Whole-brain analyses were thresholded at $p < 0.05$, Family-wise Error (FWE) corrected at the threshold level (primary threshold: $p < 0.001$, extent threshold: $k > 10$). For additional details about second-level analyses, see Appendix Text S5.

2.7 *Psychophysiological interaction analysis*:

Functional connectivity was assessed using standardized psychophysiological interaction (sPPI) analysis (Friston et al., 1997); specifically, we explored which brain regions showed changes in information exchange with the areas involved in decisions about blame (vs acceptability) for

accidental harm cases. The ROI analysis revealed r-AI to be the only region which tracked outcome-by-judgment interaction (see Results) and thus this was chosen to be the seed region. We took the recommended precautions (O'Reilly, Woolrich, Behrens, Smith, & Johansen-Berg, 2012) while carrying out PPI analysis (full details provided in Appendix Text S10).

3. Results

3.1 *Behavioral data*:

3.1.1 *Effect of belief and outcome on behavioral ratings*

Descriptive statistics for moral judgments is provided in Appendix Text S11. A 2 (belief) × 2 (outcome) repeated measure ANOVA carried out separately for acceptability and blame judgments showed that both main effects of belief (acceptability: $F(1,48) = 211.55$, $p < 0.001$, $p\eta^2 = 0.815$, $\omega^2 = 0.808$; blame: $F(1,48) = 203.72$, $p < 0.001$, $p\eta^2 = 0.809$, $\omega^2 = 0.802$) and outcome (acceptability: $F(1,48) = 114.34$, $p < 0.001$, $p\eta^2 = 0.704$, $\omega^2 = 0.694$; blame: $F(1,48) = 119.67$, $p < 0.001$, $p\eta^2 = 0.714$, $\omega^2 = 0.704$) and also their interaction (acceptability: $F(1,48) = 22.76$, $p < 0.001$, $p\eta^2 = 0.322$, $\omega^2 = 0.303$; blame: $F(1,48) = 29.14$, $p < 0.001$, $p\eta^2 = 0.378$, $\omega^2 = 0.360$) were significant.

As expected, participants assessed agents who acted with negative belief more severely (less acceptability and more blame) than agents who acted with neutral belief (see Figure 2). Also, agents who produced harmful outcome were condemned more severely than those who did not. But this condemnation was modulated by the information about agent's beliefs, as shown by interaction between belief and outcome factors. For example, accidental harms were forgiven (compared to intentional harms) based on the innocent intentions, while attempted harms were condemned (as compared to neutral and accidental cases) based on the malicious intentions (ps < 0.001).

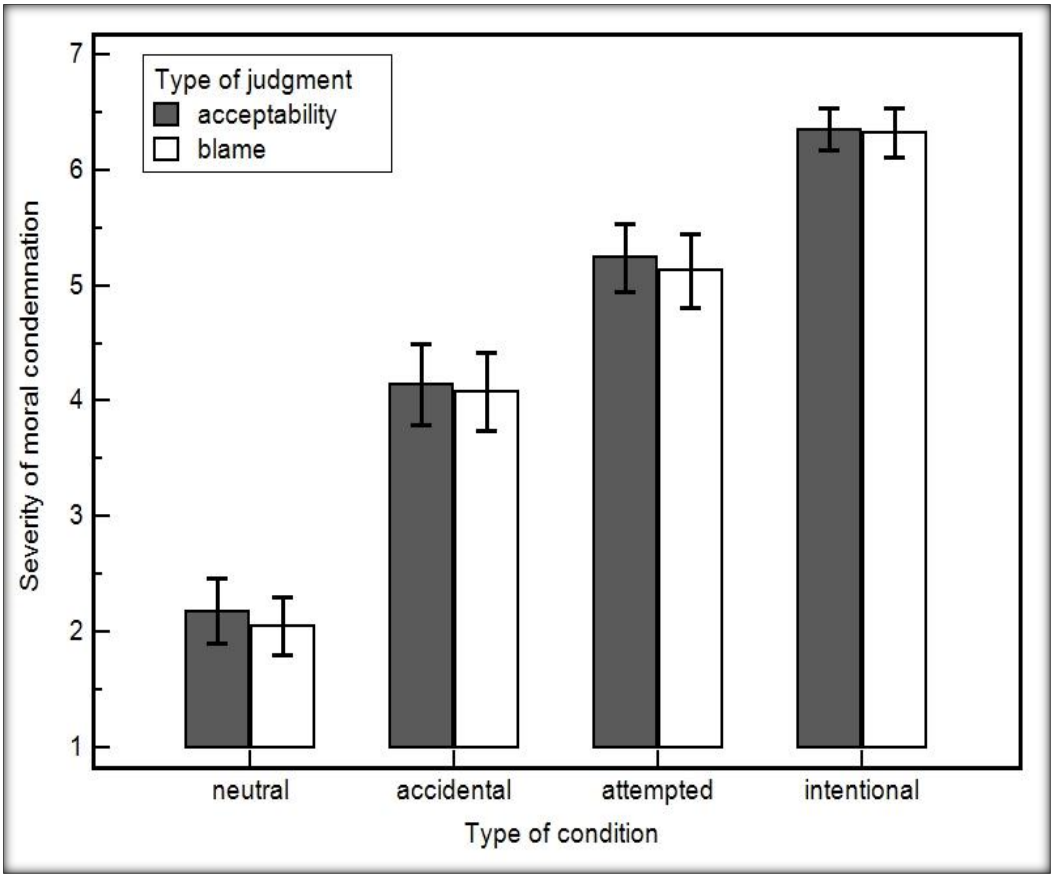


Figure 2. Moral judgments about acceptability of behavior and blame for moral agents for different types of harms: neutral case (neutral belief, neutral outcome), accidental harm (neutral belief, negative outcome), attempted harm (negative belief, neutral outcome), and intentional harm (negative belief, negative outcome). Error bars represent 95% confidence intervals. Higher scores represent more severe condemnation (less acceptable, more blame).

3.1.2 *Moral luck: acceptability versus blame judgments*

In order to assess if there was a difference in the degree to which participants relied on outcome information while judging acceptability of agent's behavior versus deciding on severity of blame, we carried out a 2 (belief) × 2 (outcome) × 2 (type of question: acceptability, blame)

repeated measure ANOVA. We were interested in outcome-by-question interaction which was not significant ($F(1,48) = 1.869$, $p = 0.178$). Therefore, we split this 3-way ANOVA into separate ANOVAs for each type of belief, but still the outcome-by-question interaction was not significant (neutral belief: $F(1,48) = 0.642$, $p = 0.427$; negative belief: $F(1,48) = 1.015$, $p = 0.319$).

We argue that absence of this effect in the current study was due to small sample size in the current study ($n = 49$ versus $n > 1000$ in (Cushman, 2008). Alternatively, this discrepancy might be due to cultural differences as the current study was conducted in Italy, while study sample from Cushman (2008) consisted of Americans. But this is unlikely because another study conducted in Italy found this effect with a bigger sample size ($n = 113$, see Appendix in Patil et al., 2016). It is also possible that this effect is a result of differences between study designs: within-subjects (current study) vs. between-subjects (Cushman, 2008). In future analysis, we will be focusing only on each participant's response on the first question[7] in order to eliminate possible order effects (Cushman et al., 2013).

3.2 *Functional localizer results*

During the localizer task, the participants' ratings revealed that although they recognized that the noxious stimulation was really painful for the protagonist in the video ($M = 1.393$, $SD = 0.810$), this did not elicit self-oriented unpleasantness in proportional manner ($M = 0.244$, $SD = 1.599$), although there was more variation in the latter than former ratings.

Correlating ratings provided on questions probing other- and self-oriented empathic response with moral judgments revealed that only self-oriented experience of unpleasantness was

[7] We would like to thank Fiery Cushman for this suggestion.

predictive of endorsed moral condemnation, but solely for the scenarios with harmful intent (attempted and intentional harm, i.e.). Full details are provided in Appendix Text S12.

3.3 *ROI analysis results*:

The ROI analysis was carried out on PSC data from four conditions (neutral, accidental, attempted, intentional) and four text segments (mental-state information, consequence, acceptability, blame) with the objective of answering following questions: (*i*) if activity in ROIs was greater than baseline *within* each condition for each segment, (*ii*) if there were any systematic differences in activity *across* different conditions for each segment, (*iii*) if there were any differences in activity *across* both segment *and* conditions indexing effect of moral luck, and (*iv*) if activity *within* each segment for a given condition was correlated with moral judgment for that condition.

3.3.1 *Average ROI activity during consequence segment*

We assessed if the average PSCs during each segment of interest was greater than zero (with respect to rest) by carrying out one-sample *t*-tests (two-tailed). We predicted that empathy ROIs, especially the affective-motivational component, will exhibit increased activity during the *consequence* segment, when information about affective state of the victim is revealed, but not during the *mental-state information* segment that provides cues about actor's intentions. Accordingly, we found that the bilateral AI, dACC, and aMCC all showed increased activity (*p*s < 0.05) during the consequence segment across all conditions, but not the bilateral PI (all statistical details along with bar graphs are provided in Appendix Text S13). Additionally, this response was primarily restricted to when outcome information was revealed and not when belief information was provided. Thus, we found a selective increase in activity while participants were

reading part of the moral vignette that provided information about harmfulness of the outcome, irrespective of whether someone was harmed (accidental/intentional) or not (neutral/attempted).

If it is true that these empathy ROIs are tracking subjective unpleasantness or salience of harmful outcomes, then they would be expected to exhibit increased activity only for conditions with negative outcomes (accidental and intentional, i.e.), but not for other conditions. Instead we observed significant activity across all conditions, which would mean that affective encoding of victim state happened even in the absence of any information about harm. One possible explanation for this peculiar response pattern is that presence of conditions with harmful outcome produced a relative contrast effect such that even in conditions where no harm was expected, aversive outcome was anticipated nonetheless (Liljeholm, Dunne, & O'Doherty, 2014). This is in line with prior studies which show that neural responses to events can be modulated by the overall contextual setting in which these events take place (e.g., anterior insula response to reward and punishment; Elliott, Friston, & Dolan, 2000) and behavioral reactions towards a particular stimulus is contingent on affective properties of other stimuli concomitant with it (Mellers, Schwartz, Ho, & Ritov, 1997).

3.3.2 *Across-condition differences in PSC*

In order to assess how activity in ROIs varied across different conditions, we carried out analysis on averaged PSC values only for the text segments that were of *a priori* interest to us: mental-state information, consequence, acceptability, and blame (see Figure 1). For each segment, we ran a 2 (belief) × 2 (outcome) repeated measures ANOVA for each ROI. For the sake of brevity, we report here only those ANOVAs which revealed significant effects (main and/or interaction)

and the post-hoc tests probing these effects were significant. Full details about the rest of the analyses have been provided in Appendix Text S14-S15.

The only empathy ROI in which main effect of outcome was found was in dACC for the acceptability segment ($F(1,38) = 5.769$, $p = 0.021$). Bonferroni-corrected post-hoc tests showed that this was due to greater PSC in response to the intentional as compared to attempted harm condition (mean difference = 0.296, 95% CI [0.095, 0.498], $p = 0.010$).

Interestingly, in l-AI, a main effect of belief ($F(1,35) = 5.120$, $p = 0.030$) and a belief-by-outcome interaction ($F(1,35) = 5.226$, $p = 0.028$) was found for the consequence segment. Post-hoc tests carried out to investigated these effects further revealed only one significant comparison (see Appendix Text S15 for figure): magnitude of PSC was greater while reading consequence information about the attempted harm condition as compared to neutral case (mean difference = 0.317, 95% CI [0.126, 0.508], $p = 0.004$). In other words, although no explicit information about possible victim suffering was provided, participants still exhibited increased neurohemodynamic response in this empathy region while reading about outcomes in the attempted harm scenarios, possibly denoting counterfactual reasoning about harm that could have befallen the victim while reading the outcome information. This line of reasoning is supported by a previous study which showed that, when asked, participants provide downward counterfactuals (how things could have been worse) most frequently for attempted harm cases, while priming participants with downward counterfactual leads to more severe evaluation of attempted harms (Lench et al., 2015). Thus, it is possible that condemnation of attempted harm relies not only on intent-based reasoning process, as argued in previous research (Cushman, 2008; Young et al., 2007), but also on causal-based reasoning process, motivated by counterfactual empathic reasoning about possible harmful outcomes. Since this effect was

neither predicted nor expected, we do not discuss it further, but it raises an interesting possibility that can be explored further in future studies.

No other significant differences were found in post-hoc comparison for any of the other ROIs or segments.

3.3.3 *Moral luck in PSC: acceptability versus blame judgments*

We carried out a 2 (outcome: neutral, negative) × 2 (judgment: acceptability, blame) repeated measures ANOVA separately for neutral (accidental versus no-harm condition) and negative (intentional versus attempted condition) belief for each ROI to investigate neural basis of moral luck and only the outcome-by-judgment interaction was of interest to us. Note that this analysis focused only on the PSCs from the two text segments when participants provided acceptability and blame judgments. For the sake of brevity, we report here only those ANOVAs which revealed interaction effect *and* the post-hoc tests probing these effects were significant. Full details about the rest of the analyses have been provided in Appendix Text S16.

The ANOVAs carried out to investigate moral luck found the outcome-by-judgment interaction only in r-AI ($F(1,37) = 5.750$, $p = 0.022$) and none of the other ROIs. Post-hoc comparisons showed (see Figure 3) that the PSC was higher when participants were deciding on blame for accidental harms as compared to when they were deciding about acceptability of behavior of accidental harm-doers (mean difference = 0.188, 95% CI [0.044, 0.333], $p = 0.024$), but no such differentiation in response was observed for neutral cases ($p = 0.967$). Thus, blame evaluations about accidents relied to a greater degree on the information about victim suffering than acceptability judgments but no such asymmetry was observed for neutral cases where there was no negative outcome.

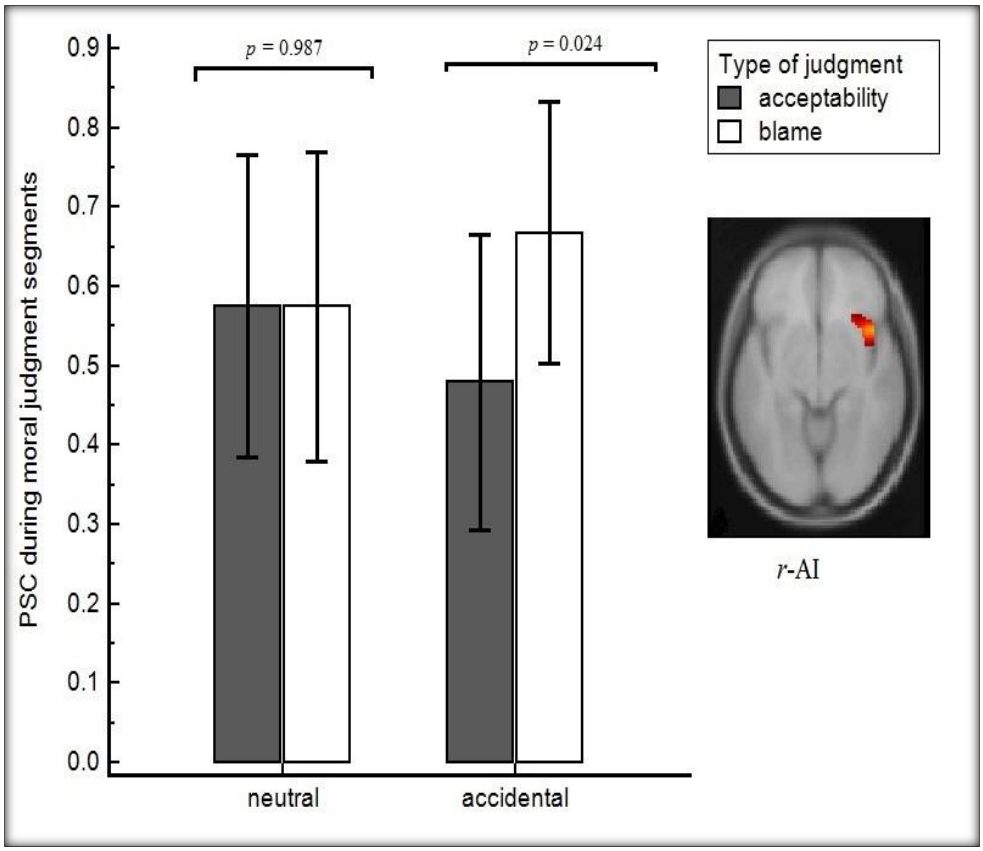


Figure 3. The PSC in the right anterior insula (r-AI) during the story segments when the participants provided moral judgments. The PSC was higher for blame compared to acceptability judgments only for accidental harm scenarios, but not for no-harm scenarios. The displayed *p*-values have been corrected for multiple comparisons and error bars correspond to 95% confidence intervals.

This result begs the question as to why the moral luck effect was observed in the PSC data, but not in the behavioral data. One explanation can be that there was less amount of variation in behavioral data (coefficient of variation for accidental condition: blame = 29.1%, acceptability = 29.3%) as compared to PSC data (coefficient of variation for accidental condition: blame = 54%, acceptability = 58.9%) due to restricted range of ratings that could be recorded in the scanner.

3.3.4 *Brain-behavior correlations*:

Correlating PSC during various segments for each type of scenario with acceptability and blame judgments revealed various significant correlations, but only two of these results survived robustness checks and are reported here (for full details, see Appendix Text S17-18). There was a positive correlation between PSC in l-PI while reading consequence segment of accidental harm scenarios for both acceptability (ρ = 0.427, p = 0.021; robust correlations: $\rho_{skipped}$ = 0.524, 95%CI [0.215, 0.730]; pi = 0.52, p = 0.010) and blame (ρ = 0.428, p = 0.021; robust correlations: $\rho_{skipped}$ = 0.523, 95%CI [0.193, 0.734]; pi = 0.52, p = 0.012) judgments for accidental harm (see Figure 4). In other words, individual differences in encoding victim's pain, possibly the sensory component of the pain, were predictive of severity of moral condemnation for accidental harm-doers with more empathic individuals endorsing harsher moral judgments.

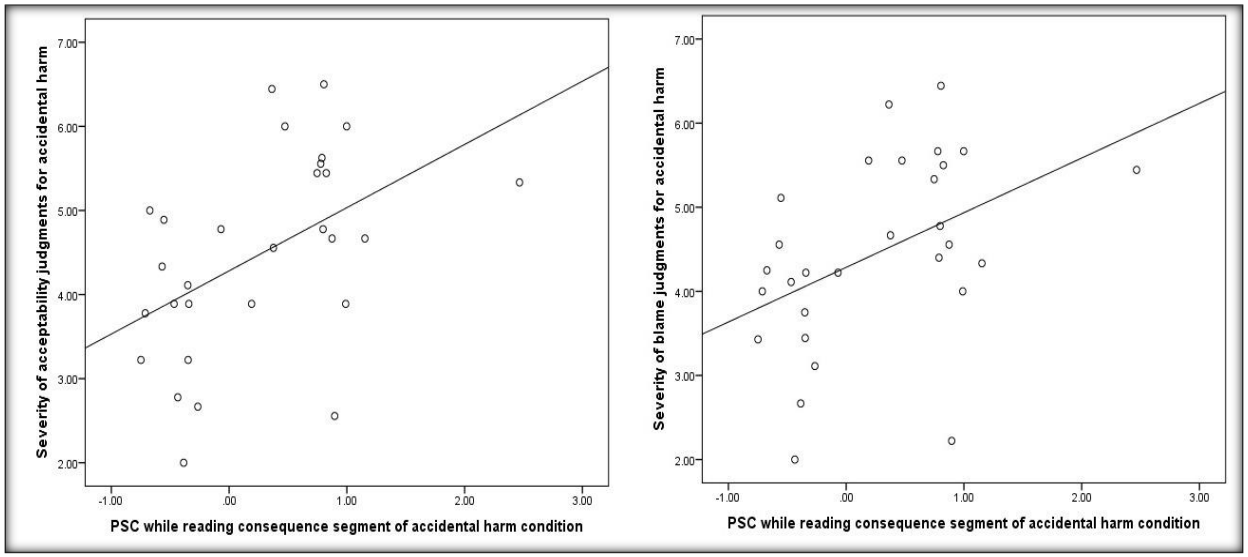


Figure 4. Individual differences in average percent signal change (PSC) in the left posterior insula (l-PI) while reading information about consequences was positively correlated with average acceptability and blame judgments for accidental harm scenarios.

3.4 *Functional connectivity results*:

In order to investigate the neural regions that exhibited changes in functional connectivity with r-AI while making acceptability and blame judgments for accidental harm condition, PPI analyses were conducted during these segments. This analysis revealed that r-AI exhibited increased exchange of information (positive PPI effect, i.e.) with the left middle frontal gyrus or l-dlPFC (x = -34, y = 10, z = 36; p(uncorrected) < 0.001, k > 10) while making blame as compared to acceptability judgments (see Figure 5). Setting the extent threshold to zero voxels (k = 0) revealed a similar cluster also in the right middle frontal gyrus or r-dlPFC (x = 36, y = 10, z = 50). No brain region showed negative PPI with r-AI. In other words, the r-AI exhibited increased functional connectivity with bilateral dlPFC during blame as compared to acceptability judgments and did not show decreased functional connectivity with any region across judgment contexts.

3.5 *Whole-brain results*:

No effects of interest were observed at the whole-brain level in second-level analysis at the corrected thresholds. Details are reported in Appendix Text S19. These results are consistent with the higher power of functional ROI analyses to detect subtle but systematic response profiles (Saxe et al., 2006).

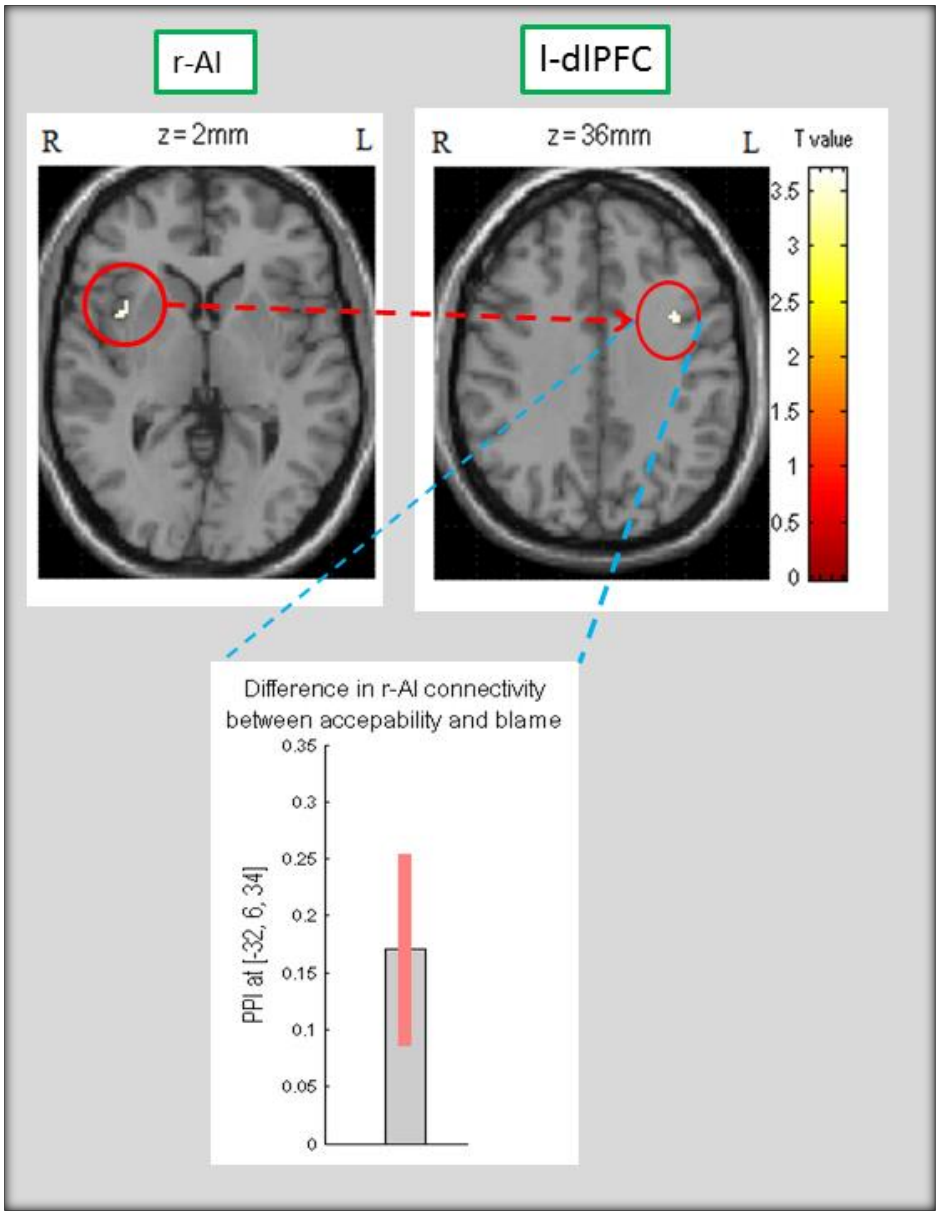


Figure 5. Regions showing increased functional connectivity with r-AI during blame judgments for accidental harm (compared with acceptability judgments) and plot of parameter estimate for difference in functional connectivity. Error bars represent 90% confidence interval. The *z*-coordinate is in MNI-coordinates. Abbreviations - PPI: psycho-physiological interaction; r-AI: right anterior insula; l-dlPFC: left dorsolaternal prefrontal cortex.

4. Discussion

Current study investigated the role of empathic reaction to victim suffering in condemning agents who accidently produced harmful outcome. The results revealed that greater neural activity in posterior insula, a node in the sensory-discriminative part of the pain matrix, was predictive of greater severity of moral assessment. In other words, the degree to which people rely on causal assessment of accidental harm-doer's role in bringing about the bad outcome is motivated by empathic aversion. Additionally, the current study also found that blame judgments relied to a greater degree on empathic assessment of the victim as compared to acceptability judgments during evaluations about accidental harms. Thus, the current findings support the prior observation that moral luck (greater reliance on outcome information) has greater bearing on blame/punishment judgments than on acceptability/wrongness judgments and localizes the source of this effect to differential integration of information about victim suffering during these two types of evaluations.

Although a burgeoning body of research demonstrates the complex relationship between empathy and morality (for reviews, see Decety & Cowell, 2014; Ugazio, Majdandžić, & Lamm, 2014), there is consensus that one of the primary moral domains where it exerts its greatest influence is harm. Harmful behaviors feature a clearly delineated victim and empathy enables people to share their affective state while evaluating moral valence of actions and plenty of prior research supports this claim. For example, empathic aversion constitutes strong negative emotional response to personally harming others for the greater good in moral dilemmas (Gleichgerrcht & Young, 2013; Patil & Silani, 2014b; Wiech et al., 2013), empathic computations also undergird other-oriented justice sensitivity for victims of harmful behavior (Decety & Yoder, 2015; Yoder & Decety, 2014) and drives altruistic behavior that comes at cost

to self (FeldmanHall, Dalgleish, Evans, & Mobbs, 2015), and active intergroup harm involves down-regulation of empathic concern for the outgroup members (Hein, Silani, Preuschoff, Batson, & Singer, 2010). But this preceding work has primarily focused on harmful acts carried out with intent to harm and, thus far, the neural basis of condemnation of unintentional harmful acts - where there is no homologous mapping between intended and realized outcome - remains sparsely studied. The current study investigated such situations in third-party settings and, consistent with prior research, found that increased empathic arousal stemming from sharing victim's pain led to more severe condemnation for accidental harm-doers. Furthermore, vicariously shared unpleasantness of this painful experience also motivated individuals to increase blame judgments for accidental harm-doers significantly more than the perceived acceptability of such behavior.

4.1 *Role of the insular cortex in moral condemnation*

The insular cortex has been known to play a key role in basic emotions and emotional processing related to social interactions (for a review, see Lamm & Singer, 2010). Insular cortex is a viscerosensory region that underpins neural representations about internal body states (e.g., pain, hunger, etc.) and represents subjective affective states (e.g., arousal, feelings, etc.). In particular, there is a posterior-to-anterior gradient in the complexity of representations of interoceptive signals that map psychophysiological states (Craig, 2002, 2009) such that PI maps only primary interoceptive information (e.g., location and intensity of painful stimulation), while AI, especially r-AI, re-represents these signals where they become consciously accessible and constitutes a subjective emotional experience ("I am feeling pain."). While empathizing with others in pain, these insular representations of bodily states are harnessed in two ways (Singer, Critchley, & Preuschoff, 2009) - (*i*) to form predictive representation of physiological response

to painful stimuli from self-centered perspective based on nociceptive information (i.e., subjective feeling state), and (*ii*) to simulate self and other-oriented subjective painful experience based on these predictions (i.e., empathic feeling state). In other words, our sensitivity to others' painful experience activates the same - primarily affective dimension of - neural representations that represent this state during first-hand painful experience.

In the backdrop of this model for the functional role of insular cortex, it is possible to interpret the observed brain-behavior correlation in l-PI during the *encoding* phase (*consequence* segment) and outcome-by-judgment interaction in r-AI during the *integration* phase (*acceptability* and *blame* segments).

Posterior insula and condemnation of accidental harms: In the current study, we found that the higher activity in l-PI while reading about harmful consequences in unintentional harm condition was predictive of severity of both acceptability and blame moral judgments endorsed by participants (Section 3.3.4). Given that PI is fundamental in representing intensity of nociceptive stimulation in the self (Segerdahl, Mezue, Okell, Farrar, & Tracey, 2015), it is possible that participants shared the sensory-discriminative component of the protagonist's pain (e.g., reading about outcome where a rabid dog bites an older lady may invoke sensory representation of dog's teeth puncturing the skin). Such bottom-up mapping of sensory-discriminative aspect of others' pain determines intensity of shared pain and this personally aversive experience drives moral condemnation for the third-party actors (Miller & Cushman, 2013). We thus take the activity in l-PI to denote *empathic arousal* (also known as emotional sharing), which represents the most rudimentary component of empathy involving duplication of another's affective state in the observer in an automatic manner without conscious awareness as to the source of this arousal (Decety & Cowell, 2014; Gonzalez-Liencres, Shamay-Tsoory, & Brüne, 2013). Indeed, past

research has shown that empathy for pain does not solely rely on affective nodes of the pain matrix, but can also automatically recruit fine-grained somatic representations to extract sensory aspects of others' pain (e.g. source and intensity) and map them onto the observer's sensorimotor system (Avenanti, Bueti, Galati, & Aglioti, 2005).

At first blush, the current finding seems to contradict the earlier evidence showing that only affective-motivational and not sensory component of empathy is copied in the observer. But note that the l-PI was not significantly activated with respect to baseline while reading about harmful consequences and thus, *on average*, information about harmful outcome did not elicit activity in sensory areas of pain matrix, as would be expected based on prior work. What we are arguing is that the cross individual variation in the *degree* to which sensory component is shared seems to impact their perception of intensity of harmfulness of the outcome, which, in its turn, is used to calibrate the severity of moral condemnation of accidents. Also, note that we are not arguing that l-PI tracks the emotional arousal in response to assessed intensity of others' pain, rather only the intensity itself. The emotional arousal stemming from pain perception is likely to be encoded in amygdala (Buckholtz et al., 2008; Hesse et al., 2015; Shenhav & Greene, 2014; Treadway et al., 2014; Yu et al., 2015).

Although a number of previous studies have shown that trait levels of various dimensional aspects of empathy are predictive of severity of moral judgments for accidental harms (Patil & Silani, 2014a; Trémolière & Djeriouat, 2016), none thus far have investigated neural correlates of this association. We resist the temptation to draw any link between prior self-report measures and activity in l-PI (e.g., activity in l-PI represents empathic concern, etc.) because the vast majority of previous studies have failed to find any association between self-report measures of

dispositional empathy (like IRI) and context-specific neural response in empathy-eliciting situations (for a review, see Decety, 2011).

Thus, based on the current findings, we posit that inter-individual differences in the severity of third-party moral evaluations of social agents who accidentally harm others stem from empathic arousal originating in vicarious encoding of somatosensory aspect of victim's pain.

Anterior insula and moral luck: Activity in r-AI featured the outcome-by-judgment interaction during integration phase such that it exhibited greater activity during blame as compared to acceptability judgment for unintentional harms, although no such difference was found for neutral cases (Section 3.3.3). The r-AI belongs the core empathy network (along with aMCC and dACC; Bzdok et al., 2012; Fan et al., 2011; Lamm et al., 2011) and indexes affective-motivational aspect of other-oriented empathic sensitivity that involves sharing subjective unpleasantness of the target's painful experience. Thus, the current data suggest that presence of harmful outcomes has a greater influence on the blame than acceptability judgments for accidental harms because the information about victim's affective state is integrated to a greater degree during blame than acceptability judgments.

From another perspective, the pain matrix has also been reconceptualized more broadly to be a part of the salience network which is involved in detecting and orienting attention towards sensory stimuli that are crucial for homeostatic balance and pain represents one such salient aspect of the internal and external environment (Uddin, 2014). Converging evidence demonstrates that activity in r-AI correlated with subjective salience across diverse task domains (Uddin, 2014). Additionally, the r-AI forms the central node of the salience network and coordinates activity of other large-scale neurocognitive networks by causally influencing activity

in central hubs of such networks (Menon & Uddin, 2010). For example, detection of salient event like experience of pain leads to r-AI-induced changes in activity of the dlPFC, a central hub in the central executive network that orchestrates externally oriented cognition and allocates attentional resources to attend to salient event.

Thus, from the r-AI as salience processing hub perspective, the selectively greater activation in r-AI during blame versus acceptability judgments for accidental harm cases could be interpreted to mean that subjective salience of information about harmfulness of the outcome is greater when one needs to decide on how much blame to attribute to the agent as compared to when agent's behavior needs to be evaluated on right-wrong dimension.

Irrespective of the perspective one subscribes to, we would expect r-AI to change its functional connectivity across different judgments domains with the brain region that plays central role in integration of different inputs (dlPFC) and, indeed, this is what was observed.

4.2 *Moral luck and differential integration of empathy inputs in dlPFC*

The functional connectivity analysis designed to explore the region that exhibited context-sensitive changes in exchange of information with r-AI revealed only one region (Section 3.4): the bilateral dlPFC showed increased connectivity with r-AI during blame as compared to acceptability judgments. This finding sits comfortably with the emerging consensus regarding the role of dlPFC as a superordinate, integrative node in decision making system that combines representations of inputs from multiple subprocesses to reach a final output that biases response selection (Buckholtz & Marois, 2012; Buckholtz et al., 2015; Treadway et al., 2014).

The dlPFC has been observed across diverse social and non-social decision making contexts: moral decision making in dilemmatic contexts (Cushman, Murray, Gordon-McKeon, Wharton,

& Greene, 2012; Kuehne, Heimrath, Heinze, & Zaehle, 2015), second-party (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) and third-party norm enforcement (Buckholtz et al., 2008), norm compliant behavior (Ruff, Ugazio, & Fehr, 2013), goal-directed planning and model-based computations (Gęsiarz & Crockett, 2015), organizing and integrating working memory content (De Pisapia, Slomski, & Braver, 2006; Owen, McMillan, Laird, & Bullmore, 2005), analogical reasoning (Bunge, Wendelken, Badre, & Wagner, 2005), etc. Although initially it was suggested that the recurrent occurrence of dlPFC across such wide variety of social and non-social decision making contexts reflects its involvement in cognitive control that inhibits prepotent responses, recent work casts doubt that this is its sole function and seems to support the alternative "integration-and-selection" function of dlPFC (for a review, see Buckholtz, 2015). In this framework, dlPFC is specifically recruited when the task demands involve holding abstract representations synthesized in multiple information processing streams in working memory and then integrating them depending on the adaptive requirements imposed by the task context. For example, in case of intent-based moral judgments, the information processing streams consist of abstract inferential processes that evaluate (*i*) presence of culpable mental states and (*ii*) assess severity of harm and are later integrated to form the final moral judgment. But note that different categories of final moral judgments have different adaptive demands: the acceptability judgments are primarily a product of mental state evaluations, while blame judgments rely additionally on the empathic evaluation of the victim (Cushman, 2008). Thus, the dlPFC would be expected to integrate information about the empathic assessment of the victim to a different degree across decision contexts due to its sensitivity to adaptive demands inherent to each task domain. Indeed, disrupting activity in bilateral dlPFC results in maladaptive performance on punishment but not wrongness judgments

because punitive judgments rely to a greater degree on integrative ability of the dlPFC (Buckholtz et al., 2015). Thus, we argue that the observed decision-context-dependent change in functional connectivity between r-AI and l-dlPFC reflects increased integration of harm assessment during blame judgments that biases selection of magnitude of blame for agents causally responsible for the harmful outcome.

4.3 *Conclusions*

In summary, the current findings expand the role of empathy in condemning harmful acts from intentional to unintentional. It also sheds light on the distinct contributions of different components of empathy and their neural correlates. The current study also highlights the motivational role of interpersonal sensitivity of third-party judges to enforce widely shared sentiments about appropriate behavior. Although the current study assessed attribution of blame to moral agents, the same neurocognitive architecture is expected to underpin punishment judgments since the punishment arises from the attribution of blame (Cushman, 2008; Fincham & Roberts, 1985; Shultz, Schleifer, & Altman, 1981). Thus, conclusions derived from the current study also inform current neurobiological models of punishment as well.

4.4 *Future work*

Current study was hypothesis-driven in terms of the proposed role of empathy in condemnation of accidental harms and differential influence of moral luck on different categories of moral evaluations. As such, we did not explore roles of other processes and activity in regions mediating these processes: mental state reasoning in rTPJ, affective arousal encoding in amygdala, emotion regulation in dACC, etc. In future connectivity analysis study, we would explore complex interactions between these critical nodes during moral judgments.

Additionally, we adopted an ultra-conservative approach which could have significantly reduced degrees of freedom at the first-level analysis (due to high number of regressors) and reduced sample size at the group level (focusing only on participants in which ROIs were localized) both of which could have contributed to loss of power that prevented us from detecting more subtle effects in both ROI and whole-brain analyses. Future work will focus on re-analysis of this data with more liberal methodological approach.

Bibliography

Aaron, R. V., Benson, T. L., & Park, S. (2015). Investigating the role of alexithymia on the empathic deficits found in schizotypy and autism spectrum traits. *Personality and Individual Differences*, *77*, 215–220. doi:10.1016/j.paid.2014.12.032

Abler, B., & Kessler, H. (2009). Emotion Regulation Questionnaire - Eine Deutschsprachige Fassung des ERQ von Gross und John. *Diagnostica*, *55*(3), 144–152. doi:10.1026/0012-1924.55.3.144

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368–378. doi:10.1037//0022-3514.63.3.368

Alter, A. L., Kernochan, J., & Darley, J. M. (2007). Morality Influences How People Apply the Ignorance of the Law Defense. *Law & Society Review*, *41*(4), 819–864. doi:10.1111/j.1540-5893.2007.00327.x

American Psychiatric Association. (2013). *DSM 5*. *American Journal of Psychiatry*. doi:10.1176/appi.books.9780890425596.744053

Amit, E., & Greene, J. (2012). You See, the Ends Don't Justify the Means: Visual Imagery and Moral Judgment. *Psychological Science*, *23*(8), 861–868. doi:10.1177/0956797611434965

Arbuckle, J. (2013). IBM SPSS Amos 22 user's guide. *Crawfordville, FL: Amos Development Corporation*.

Ashby, F. G. (2011). *Statistical analysis of fMRI data* (1st ed.). Cambridge, Massachusetts: MIT press.

Avenanti, A., Bueti, D., Galati, G., & Aglioti, S. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. *Nature Neuroscience*, *8*(7), 955–60. doi:10.1038/nn1481

Avramova, Y. R., & Inbar, Y. (2013). Emotion and moral judgment. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 169–178. doi:10.1002/wcs.1216

Baez, S., Couto, B., Torralva, T., Sposato, L. a, Huepe, D., Montañes, P., … Ibanez, A. (2014). Comparing moral judgments of patients with frontotemporal dementia and frontal stroke. *JAMA Neurology*, *71*(9), 1172–6. doi:10.1001/jamaneurol.2014.347

Baez, S., & Ibanez, A. (2014). The effects of context processing on social cognition impairments in adults with Asperger's syndrome. *Frontiers in Neuroscience*, *8*(September), 1–9. doi:10.3389/fnins.2014.00270

Baez, S., Rattazzi, A., Gonzalez-Gadea, M. L., Torralva, T., Vigliecca, N. S., Decety, J., … Ibanez, A. (2012). Integrating intention and context: assessing social cognition in adults with Asperger syndrome. *Frontiers in Human Neuroscience*, *6*(November), 302. doi:10.3389/fnhum.2012.00302

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale--I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23–32. doi:10.1016/0022-3999(94)90005-1

Baird, J. a, & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development*, (103), 37–49. doi:10.1002/cd.96

Baldner, C., & McGinley, J. J. (2014). Correlational and exploratory factor analyses (EFA) of commonly used empathy questionnaires: New insights. *Motivation and Emotion*, *38*(5), 727–744. doi:10.1007/s11031-014-9417-2

Barnes, J. L., Lombardo, M. V, Wheelwright, S., & Baron-Cohen, S. (2009). Moral dilemmas film task: A study of spontaneous narratives by individuals with autism spectrum conditions. *Autism Research* , *2*(3), 148–56. doi:10.1002/aur.79

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., & Wheelwright, S. (2003). The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autsm, and normal sex differences. *Phiosophical Transactions: Biological Sciences*, *358*, 361–374.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. doi:10.1023/A:1005653411471

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck depression inventory-II. *San Antonio, TX: Psychological Corporation*, 1–82.

Berg-Cross, L. (1975). Intentionality, degree of damage, and moral judgments. *Child Development*, *46*(4), 970–974. doi:10.2307/1128406

Bernhardt, B., Valk, S., Silani, G., Bird, G., Frith, U., & Singer, T. (2014). Selective Disruption of Sociocognitive Structural Brain Networks in Autism and Alexithymia. *Cerebral Cortex* , *24*(12), 3258–67. doi:10.1093/cercor/bht182

Berthoz, S., & Hill, E. L. (2005). The validity of using self-reports to assess emotion regulation abilities in adults with autism spectrum disorder. *European Psychiatry*, *20*(3), 291–298. doi:10.1016/j.eurpsy.2004.06.013

Bilker, W. B., Hansen, J. a, Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354–69. doi:10.1177/1073191112446655

Bird, G., & Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. *Translational Psychiatry*, *3*(7), e285. doi:10.1038/tp.2013.61

Bird, G., Silani, G., Brindley, R., White, S., Frith, U., & Singer, T. (2010). Empathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain*, *133*(5), 1515–1525. doi:10.1093/brain/awq060

Blair, R. J. R. (1996). Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders*, *26*(5), 571–579. doi:10.1007/BF02172277

Blascovich, J., Loomis, J., & Beall, A. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*, 103–124.

Brett, M., Anton, J., Valabregue, R., & Poline, J. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage*, *16*(2), S497.

Brewer, R., Marsh, A. A., Catmur, C., Cardinale, E. M., Stoycos, S., Cook, R., … Al, B. E. T. (2015). The Impact of Autism Spectrum Disorder and Alexithymia on Judgments of Moral Acceptability, *124*(3), 589–595.

Bruneau, E., Dufour, N., & Saxe, R. (2013). How we know it hurts: item analysis of written

narratives reveals distinct neural responses to others' physical pain and emotional suffering. *PloS One*, *8*(4), e63085. doi:10.1371/journal.pone.0063085

Buckholtz, J. W. (2015). Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences*, *3*(June), 122–129. doi:10.1016/j.cobeha.2015.03.004

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The Neural Correlates of Third-Party Punishment. *Neuron*, *60*(5), 930–940. doi:10.1016/j.neuron.2008.10.016

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661. doi:10.1038/nn.3087

Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., & Marois, R. (2015). From Blame to Punishment: Disrupting Prefrontal Cortex Activity Reveals Norm Enforcement Mechanisms. *Neuron*, *87*(6), 1369–80. doi:10.1016/j.neuron.2015.08.023

Bunge, S., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, *15*(3), 239–49. doi:10.1093/cercor/bhh126

Buon, M., Dupoux, E., Jacob, P., Chaste, P., Leboyer, M., & Zalla, T. (2013). The role of causal and intentional judgments in moral reasoning in individuals with high functioning autism. *Journal of Autism and Developmental Disorders*, *43*(2), 458–70. doi:10.1007/s10803-012-1588-7

Buon, M., Jacob, P., Loissel, E., & Dupoux, E. (2013). A non-mentalistic cause-based heuristic

in human social evaluations. *Cognition*, *126*(2), 149–55. doi:10.1016/j.cognition.2012.09.006

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function*, *217*(4), 783–96. doi:10.1007/s00429-012-0380-y

Capps, L., Kasari, C., Yirmiya, N., & Sigman, M. (1993). Parental perception of emotional expressiveness in children with autism. *Journal of Consulting and Clinical Psychology*, *61*(3), 475–484. doi:10.1037/0022-006X.61.3.475

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fmri experiments. *Frontiers in Neuroscience*, *6*(OCT), 1–13. doi:10.3389/fnins.2012.00149

Carter, E. J., Williams, D. L., Minshew, N. J., & Lehman, J. F. (2012). Is he being bad? Social and language brain networks during social judgment in children with autism. *PloS One*, *7*(10), e47241. doi:10.1371/journal.pone.0047241

Chiong, W., Wilson, S. M., D'Esposito, M., Kayser, A. S., Grossman, S. N., Poorzand, P., … Rankin, K. P. (2013). The salience network causally influences default mode network activity during moral reasoning. *Brain : A Journal of Neurology*, *136*(Pt 6), 1929–41. doi:10.1093/brain/awt066

Choe, S. Y., & Min, K. (2011). Who makes utilitarian judgments ? The influences of emotions on utilitarian judgments. *Judgment and Decision Making*, *6*(7), 580–592.

Christensen, J., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1249–64. doi:10.1016/j.neubiorev.2012.02.008

Ciaramelli, E., Braghittoni, D., & di Pellegrino, G. (2012). It is the outcome that counts! Damage to the ventromedial prefrontal cortex disrupts the integration of outcome and belief information for moral judgment. *Journal of the International Neuropsychological Society*, *18*(6), 962–71. doi:10.1017/S1355617712000690

Ciaramelli, E., Sperotto, R. G., Mattioli, F., & di Pellegrino, G. (2013). Damage to the ventromedial prefrontal cortex reduces interpersonal disgust. *Social Cognitive and Affective Neuroscience*, *8*(2), 171–80. doi:10.1093/scan/nss087

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216–35. doi:10.1037/a0031021

Craig, A. D. B. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8), 655–666. doi:10.1038/nrn894

Craig, A. D. B. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews. Neuroscience*, *10*(1), 59–70. doi:10.1038/nrn2555

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–80. doi:10.1016/j.cognition.2008.03.006

Cushman, F. (2013a). Action, outcome, and value: a dual-system framework for morality.

Personality and Social Psychology Review, *17*(3), 273–92. doi:10.1177/1088868313495594

Cushman, F. (2013b). The role of learning in punishment, prosociality, and human uniqueness. In *Signaling, Commitment and Emotion, Vol. 2: Psychological and Environmental Foundations of Cooperation*. Cambridge, Mass.: MIT press.

Cushman, F. (2015a). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97–103. doi:10.1016/j.copsyc.2015.06.003

Cushman, F. (2015b). Punishment in Humans: From Intuitions to Institutions. *Philosophy Compass*, *10*(2), 117–133. doi:10.1111/phc3.12192

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS One*, *4*(8), e6699. doi:10.1371/journal.pone.0006699

Cushman, F., & Greene, J. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*. doi:10.1080/17470919.2011.614000

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. (2012). Judgment before principle: engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience*, *7*(8), 888–95. doi:10.1093/scan/nsr072

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. doi:10.1016/j.cognition.2012.11.008

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. doi:10.1111/j.1467-9280.2006.01834.x

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. doi:10.1037/0022-3514.44.1.113

Dawson, M., Schell, A., & Filion, D. (2007). The Electrodermal System. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of of psychophysiology* (pp. 159–181). Cambridge: Cambridge University Press.

de Achával, D., Villarreal, M. F., Salles, A., Bertomeu, M. J., Costanzo, E. Y., Goldschmidt, M., … Guinjoan, S. M. (2013). Activation of brain areas concerned with social cognition during moral decisions is abnormal in schizophrenia patients and unaffected siblings. *Journal of Psychiatric Research*, *47*(6), 774–82. doi:10.1016/j.jpsychires.2012.12.018

De Pisapia, N., Slomski, J. A., & Braver, T. S. (2006). Functional Specializations in Lateral Prefrontal Cortex Associated with the Integration and Segregation of Information in Working Memory. *Cerebral Cortex*, *17*(5), 993–1006. doi:10.1093/cercor/bhl010

de Vignemont, F., & Singer, T. (2006). The empathic brain: how, when and why? *Trends in Cognitive Sciences*, *10*(10), 435–41. doi:10.1016/j.tics.2006.08.008

Decety, J. (2011). Dissecting the Neural Mechanisms Mediating Empathy. *Emotion Review*, *3*(1), 92–108. doi:10.1177/1754073910374662

Decety, J., & Cacioppo, S. (2012). The speed of morality: a high-density electrical neuroimaging study. *Journal of Neurophysiology*, *108*(11), 3068–3072. doi:10.1152/jn.00473.2012

Decety, J., & Cowell, J. M. (2014). Friends or Foes: Is Empathy Necessary for Moral Behavior? *Perspectives on Psychological Science*, *9*(5), 525–537. doi:10.1177/1745691614545130

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, *22*(1), 209–20. doi:10.1093/cercor/bhr111

Decety, J., & Yoder, K. J. (2015). Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict sensitivity to injustice for others. *Social Neuroscience*, *11*(1), 1–14. doi:10.1080/17470919.2015.1029593

Devlin, J., & Poldrack, R. (2007). In praise of tedious anatomy. *NeuroImage*, *37*(4), 1033–41; discussion 1050–8. doi:10.1016/j.neuroimage.2006.09.055

Djeriouat, H., & Trémolière, B. (2014). The Dark Triad of personality and utilitarian moral judgment: The mediating role of Honesty/Humility and Harm/Care. *Personality and Individual Differences*, *67*, 11–16. doi:10.1016/j.paid.2013.12.026

Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., & Convit, A. (2008). Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). *Journal of Autism and Developmental Disorders*, *38*(3), 464–473. doi:10.1007/s10803-007-0486-x

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. doi:10.1016/j.neuroimage.2004.12.034

Elliott, R., Friston, K., & Dolan, R. J. (2000). Dissociable neural responses in human reward systems. *The Journal of Neuroscience*, *20*(16), 6159–65.

Fan, Y., Duncan, N. W., de Greck, M., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience and Biobehavioral Reviews*, *35*(3), 903–911. doi:10.1016/j.neubiorev.2010.10.009

Fan, Y.-T. T., Chen, C., Chen, S.-C. C., Decety, J., & Cheng, Y. (2014). Empathic arousal and social understanding in individuals with autism: Evidence from fMRI and ERP measurements. *Social Cognitive and Affective Neuroscience*, *9*(8), 1203–13. doi:10.1093/scan/nst101

Fedorenko, E., & Kanwisher, N. (2011). Functionally Localizing Language-Sensitive Regions in Individual Subjects With fMRI: A Reply to Grodzinsky's Critique of Fedorenko and Kanwisher (2009). *Linguistics and Language Compass*, *5*, 78–94. doi:10.1111/j.1749-818X.2010.00264.x

Fehse, K., Silveira, S., Elvers, K., & Blautzik, J. (2014). Compassion, guilt and innocence: An fMRI study of responses to victims who are responsible for their fate. *Social Neuroscience*, *10*(3), 243–252. doi:10.1080/17470919.2014.980587

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *NeuroImage*, *105*, 347–356. doi:10.1016/j.neuroimage.2014.10.043

FeldmanHall, O., Dalgleish, T., & Mobbs, D. (2013). Alexithymia decreases altruism in real social decisions. *Cortex*, *49*(3), 899–904. doi:10.1016/j.cortex.2012.10.015

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: the relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–41. doi:10.1016/j.cognition.2012.02.001

Figner, B., & Murphy, R. O. (2010). Using skin conductance in judgment and decision making research. In *A Handbook of Process Tracing Methods for Decision Research: A Critical Review and User's Guide* (pp. 163–184). New York: Psychology Press.

Fincham, F. D., & Roberts, C. (1985). Intervening Causation And The Mitigation Of Responsibility For Harm Doing. *Journal of Experimental Social Psychology*, *21*(2), 178–194. doi:10.1016/0022-1031(85)90014-9

Fletcher-Watson, S., & McConachie, H. (2014). Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD). *The Cochrane Library*, *3*, CD008785. doi:10.1002/14651858.CD008785.pub2

Freitag, C. M., Retz-Junginger, P., Retz, W., Seitz, C., Palmason, H., Meyer, J., … von Gontard, A. (2007). German adaptation of the autism-spectrum quotient (AQ): Evaluation and short version AQ-k. *Zeitschrift Fur Klinische Psychologie Und Psychotherapie: Forschung Und Praxis*, *36*, 280–289. doi:10.1026/1616-3443.36.4.280

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender Differences in Responses to Moral Dilemmas: A Process Dissociation Analysis. *Personality and Social Psychology Bulletin*, *41*(5), 696–713. doi:10.1177/0146167215575731

Friston, K., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, *6*(3), 218–229. doi:10.1006/nimg.1997.0291

Friston, K., Rotshtein, P., Geng, J. J., Sterzer, P., & Henson, R. N. (2006). A critique of functional localisers. *NeuroImage*, *30*(4), 1077–87. doi:10.1016/j.neuroimage.2005.08.012

Frith, U. (2004). Emanuel Miller lecture: Confusions and controversies about Asperger syndrome. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. doi:10.1111/j.1469-7610.2004.00262.x

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*(3), 233–239. doi:10.1111/j.1467-9280.2007.01882.x

Fumagalli, M., Marceglia, S., Cogiamanian, F., Ardolino, G., Picascia, M., Barbieri, S., … Priori, A. (2015). Ethical safety of deep brain stimulation: A study on moral decision-making in Parkinson's disease. *Parkinsonism & Related Disorders*, *21*(7), 709–716. doi:10.1016/j.parkreldis.2015.04.011

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience*, *9*(May), 1–18. doi:10.3389/fnbeh.2015.00135

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: experiencing the future. *Science* , *317*(5843), 1351–4. doi:10.1126/science.1144161

Gitelman, D. R., Penny, W. D., Ashburner, J., & Friston, K. (2003). Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *Neuroimage*, *19*(1), 200–207. doi:10.1016/S1053-8119(03)00058-2

Glaser, D., & Friston, K. (2004). Variance components. In J. Ashburner, K. Friston, & W. Penny (Eds.), *Human brain function* (2nd ed., pp. 781–93).

Gleichgerrcht, E., Tomashitis, B., & Sinay, V. (2015). The relationship between alexithymia, empathy and moral judgment in patients with multiple sclerosis. *European Journal of Neurology*, *22*(9), :1295–303. doi:10.1111/ene.12745

Gleichgerrcht, E., Torralva, T., Rattazzi, A., Marenco, V., Roca, M., & Manes, F. (2013). Selective impairment of cognitive empathy for moral judgment in adults with high functioning autism. *Social Cognitive and Affective Neuroscience*, *8*(7), 780–788. doi:10.1093/scan/nss067

Gleichgerrcht, E., Torralva, T., Roca, M., Pose, M., & Manes, F. (2011). The role of social cognition in moral judgment in frontotemporal dementia. *Social Neuroscience*, *6*(2), 113–122. doi:10.1080/17470919.2010.506751

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PloS One*, *8*(4), e60418. doi:10.1371/journal.pone.0060418

Gonzalez-Liencres, C., Shamay-Tsoory, S. G., & Brüne, M. (2013). Towards a neuroscience of empathy: Ontogeny, phylogeny, brain mechanisms, context and psychopathology. *Neuroscience and Biobehavioral Reviews*, *37*(8), 1537–1548. doi:10.1016/j.neubiorev.2013.05.001

Grant, C. M., Boucher, J., Riggs, K. J., & Grayson, A. (2005). Moral understanding in children with autism. *Autism : The International Journal of Research and Practice*, *9*(3), 317–331. doi:10.1177/1362361305055418

Gray, K., & Schein, C. (2012). Two Minds Vs. Two Philosophies: Mind Perception Defines Morality and Dissolves the Debate Between Deontology and Utilitarianism. *Review of Philosophy and Psychology*, *3*(3), 405–423. doi:10.1007/s13164-012-0112-5

Greene, J. (2009). The Cognitive Neuroscience of Moral Judgment. *The Cognitive Neurosciences IV*, *4*, 987–999.

Greene, J. (2014). Beyond Point-and-Shoot Morality : Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, *124*(4), 695–726. doi:10.1086/675875

Greene, J., Cushman, F., Stewart, L., Lowenberg, K., Nystrom, L., & Cohen, J. (2009). Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–71. doi:10.1016/j.cognition.2009.02.001

Greene, J., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive Load Selectively Interferes with Utilitarian Moral Judgment. *Cognition*, *107*(3), 1144–1154. doi:10.1016/j.cognition.2007.11.004

Greene, J., 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. doi:10.1016/j.neuron.2004.09.027

Greene, J., 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. doi:10.1126/science.1062872

Griffin, C., Lombardo, M. V, & Auyeung, B. (2015). Alexithymia in children with and without autism spectrum disorders. *Autism Research*. doi:10.1002/aur.1569

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*(2), 348–362. doi:10.1037/0022-3514.85.2.348

Grynberg, D., Luminet, O., Corneille, O., Grèzes, J., & Berthoz, S. (2010). Alexithymia in the interpersonal domain: A general deficit of empathy? *Personality and Individual*

Differences, *49*(8), 845–850. doi:10.1016/j.paid.2010.07.013

Gu, X., Eilam-Stock, T., Zhou, T., Anagnostou, E., Kolevzon, A., Soorya, L., … Fan, J. (2015). Autonomic and brain responses associated with empathy deficits in autism spectrum disorder. *Human Brain Mapping*, *36*(9), 3323–38. doi:10.1002/hbm.22840

Gu, X., & Han, S. (2007). Attention and reality constraints on the neural processes of empathy for pain. *NeuroImage*, *36*(1), 256–267. doi:10.1016/j.neuroimage.2007.02.025

Guglielmo, S. (2015). Moral judgment as information processing: an integrative review. *Frontiers in Psychology*, *6*, 1637. doi:10.3389/fpsyg.2015.01637

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition*, *133*(1), 97–103. doi:10.1016/j.cognition.2014.06.001

Guttman, H., & Laporte, L. (2002). Alexithymia, empathy, and psychological symptoms in a family context. *Comprehensive Psychiatry*, *43*(6), 448–55. doi:10.1053/comp.2002.35905

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834. doi:10.1037/0033-295X.108.4.814

Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science*, *316*, 998–1002. doi:10.1126/science.1137651

Hauser, M., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A Dissociation Between Moral Judgments and Justifications. *Mind & Language*, *22*(1), 1–21. doi:10.1111/j.1468-0017.2006.00297.x

Hautzinger, M. (1991). Das Beck-Depressions-Inventar (BDI) in der Klinik. *Der Nervenarzt*, *62*, 689–96.

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron*, *68*(1), 149–160. doi:10.1016/j.neuron.2010.09.003

Henson, R., Rugg, M., & Friston, K. (2001). The choice of basis functions in event-related fMRI. *Neuroimage*, *13*(6), 149. doi:10.1016/S1053-8119(01)91492-2

Hesse, E., Mikulan, E., Decety, J., Sigman, M., Garcia, M. del C., Silva, W., … Ibanez, A. (2015). Early detection of intentional harm in the human amygdala. *Brain*. doi:10.1093/brain/awv336

Hill, E., Berthoz, S., & Frith, U. (2004). Brief report: Cognitive processing of own emotions in individuals with autistic spectrum disorder and in their relatives. *Journal of Autism and Developmental Disorders*, *34*(2), 229–235. doi:10.1023/B:JADD.0000022613.41399.14

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*(1), 1–6. doi:10.1016/j.tics.2008.09.006

Hurlburt, R. T., Happé, F., & Frith, U. (1994). Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological Medicine*, *24*(02), 385. doi:10.1017/S0033291700027367

Inbar, Y., Pizarro, D. a., & Cushman, F. (2012). Benefiting from misfortune: when harmless

actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, *38*(1), 52–62. doi:10.1177/0146167211430232

Inglis, B. (2015). A checklist for fMRI acquisition methods reporting in the literature. *The Winnower*. doi:10.15200/winn.143191.17127

Jameel, L., Vyas, K., Bellesi, G., Roberts, V., & Channon, S. (2014). Going "above and beyond": are those high in autistic traits less pro-social? *Journal of Autism and Developmental Disorders*, *44*(8), 1846–58. doi:10.1007/s10803-014-2056-3

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560. doi:10.1080/17470919.2015.1023400

Kant, I. (2005). *The moral law: Groundwork of the metaphysic of morals* (2nd ed.). London: Routledge.

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2015). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*. doi:10.1177/1362361315588200

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829–832. doi:10.1126/science.1129156

Kobayashi, N., Yoshino, A., Takahashi, Y., & Nomura, S. (2007). Autonomic arousal in cognitive conflict resolution. *Autonomic Neuroscience: Basic and Clinical*, *132*(1-2), 70–75. doi:10.1016/j.autneu.2006.09.004

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. doi:10.1038/nature05631

Koller, I., & Lamm, C. (2015). Item Response Model Investigation of the (German) Interpersonal Reactivity Index Empathy Questionnaire. *European Journal of Psychological Assessment, 1*(-1), 1–11. doi:10.1027/1015-5759/a000227

Koster-Hale, J., Bedny, M., & Saxe, R. (2014). Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition, 133*(1), 65–78. doi:10.1016/j.cognition.2014.04.006

Koster-hale, J., & Saxe, R. (2013). Functional neuroimaging of theory of mind. In S. Baron-Cohen, M. Lombardo, & H. Tager-Flusberg (Eds.), *Understanding Other Minds* (3rd ed., pp. 132–163). Oxford University Press.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America, 110*(14), 5648–53. doi:10.1073/pnas.1207992110

Koven, N. S. (2011). Specificity of meta-emotion effects on moral decision-making. *Emotion*. doi:10.1037/a0025616

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications, 6*, 7455. doi:10.1038/ncomms8455

Kuehne, M., Heimrath, K., Heinze, H.-J., & Zaehle, T. (2015). Transcranial direct current

stimulation of the left dorsolateral prefrontal cortex shifts preference of moral judgments. *PloS One*, *10*(5), e0127061. doi:10.1371/journal.pone.0127061

Kupfer, J., Brosig, B., & Brähler, E. (2000). Überprüfung und Validierung der 26-Item Toronto Alexithymie-Skala anhand einer repräsentativen Bevölkerungsstischprobe. *Zeitschrift Für Psychosomatische Medizin Und Psychotherapie*, *46*(4), 368–384.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV). doi:10.3389/fpsyg.2013.00863

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience*, *19*(1), 42–58. doi:10.1162/jocn.2007.19.1.42

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–502. doi:10.1016/j.neuroimage.2010.10.014

Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Structure & Function*, *214*(5-6), 579–91. doi:10.1007/s00429-010-0251-3

Lane, R. D., Weihs, K. L., Herring, A., Hishaw, A., & Smith, R. (2015). Affective Agnosia: Expansion of the Alexithymia Construct and a New Opportunity to Integrate and Extend Freud's Legacy. *Neuroscience & Biobehavioral Reviews*, *55*, 594–611. doi:10.1016/j.neubiorev.2015.06.007

Langdon, P. E., Clare, I. C. H., & Murphy, G. H. (2010). Developing an understanding of the

literature relating to the moral development of people with intellectual disabilities. *Developmental Review*, *30*(3), 273–293. doi:10.1016/j.dr.2010.01.001

Laurent, S. M., Nuñez, N. L., & Schweitzer, K. a. (2015). The influence of desire and knowledge on perception of each other and related mental states, and different mechanisms for blame. *Journal of Experimental Social Psychology*, *60*(APRIL), 27–38. doi:10.1016/j.jesp.2015.04.009

Lehrl, S. (1995). *Mehrfachwahl-Wortschatz-Intelligenztest: MWT-B.* Balingen: PERIMED-spitta.

Lehrl, S., Triebig, G., & Fischer, B. (1995). Multiple choice vocabulary test MWT as a valid and short test to estimate premorbid intelligence. *Acta Neurologica Scandinavica*, *91*(5), 335–345. doi:10.1111/j.1600-0404.1995.tb07018.x

Lench, H. C., Domsky, D., Smallman, R., & Darbor, K. E. (2015). Beliefs in moral luck: When and why blame hinges on luck. *British Journal of Psychology*, *106*(2), 272–287. doi:10.1111/bjop.12072

Leslie, A. M., Mallon, R., & DiCorcia, J. a. (2006). Transgressors, victims, and cry babies: is basic moral judgment spared in autism? *Social Neuroscience*, *1*(3-4), 270–283. doi:10.1080/17470910600992197

Li, J., Zhu, L., & Gummerum, M. (2014). The relationship between moral judgment and cooperation in children with high-functioning autism. *Scientific Reports*, *4*, 4314. doi:10.1038/srep04314

Lieberman, M. D., & Cunningham, W. a. (2009). Type I and Type II error concerns in fMRI

research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience*, *4*(4), 423–428. doi:10.1093/scan/nsp052

Liljeholm, M., Dunne, S., & O'Doherty, J. P. (2014). Anterior insula activity reflects the effects of intentionality on the anticipation of aversive stimulation. *The Journal of Neuroscience*, *34*(34), 11339–48. doi:10.1523/JNEUROSCI.1126-14.2014

Lindquist, M. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, *23*(4), 439–464. doi:10.1214/09-STS282

Lindquist, M., Meng Loh, J., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *NeuroImage*, *45*(1 Suppl), S187–98. doi:10.1016/j.neuroimage.2008.10.065

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*(2), 273–86. doi:10.1111/j.1551-6709.2009.01013.x

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., Dilavore, P. C., … Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223. doi:10.1023/A:1005592401947

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*(5), 659–685. doi:10.1007/BF02172145

Lotto, L., Manfrinati, A., & Sarlo, M. (2014). A New Set of Moral Dilemmas: Norms for Moral

Acceptability, Decision Times, and Emotional Salience. *Journal of Behavioral Decision Making*, *27*(1), 57–65. doi:10.1002/bdm.1782

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, *25*(2), 147–186. doi:10.1080/1047840X.2014.877340

Martin, J. W., & Cushman, F. (2016). The adaptive logic of moral luck. In J. Sytsma & W. Buckwalter (Eds.), *The Blackwell Companion to Experimental Philosophy*.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*. doi:10.3758/s13428-011-0168-7

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, *5*(4), 434–58. doi:10.1037/1082-989X.5.4.434

Mazzocco, P., Alicke, M., & Davis, T. (2004). On the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology*, *26*(2-3), 131–146. doi:10.1080/01973533.2004.9646401

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *NeuroImage*, *61*(4), 1277–1286. doi:10.1016/j.neuroimage.2012.03.068

Mellers, B., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision Affect Theory: Emotional Reactions to the Outcomes of Risky Options. *Psychological Science*, *8*(6), 423–429. doi:10.1111/j.1467-9280.1997.tb00455.x

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model

of insula function. *Brain Structure & Function*, *214*(5-6), 655–67. doi:10.1007/s00429-010-0262-0

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. doi:10.1016/j.tics.2006.12.007

Mill, J. S. (1998). *Utilitarianism*. (R. Crisp, Ed.). New York: Oxford University Press.

Miller, R., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass*, *7*(10), 707–718. doi:10.1111/spc3.12066

Miller, R., Hannikainen, I., & Cushman, F. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion , 14*(3), 573–87. doi:10.1037/a0035361

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379–387. doi:10.1037/0003-066X.38.4.379

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. doi:10.1111/j.1467-9280.2008.02122.x

Moran, J. M., Young, L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2688–2692. doi:10.1073/pnas.1011734108

Moretto, G., Làdavas, E., Mattioli, F., & di Pellegrino, G. (2010). A psychophysiological

investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience, 22*(8), 1888–1899. doi:10.1162/jocn.2009.21367

Mumford, J., Poline, J.-B., & Poldrack, R. (2015). Orthogonalization of Regressors in fMRI Models. *Plos One, 10*(4), e0126255. doi:10.1371/journal.pone.0126255

Mutschler, I., Wieckhorst, B., Kowalevski, S., Derix, J., Wentlandt, J., Schulze-Bonhage, A., & Ball, T. (2009). Functional organization of the human anterior insular cortex. *Neuroscience Letters, 457*(2), 66–70. doi:10.1016/j.neulet.2009.03.101

Nagel, T. (1985). Moral Luck. *Philosophy, 19*(226), 544. doi:10.1017/S0031819100066729

Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: emotion and action in a simulated three-dimensional "trolley problem". *Emotion , 12*(2), 364–70. doi:10.1037/a0025561

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*. doi:10.3389/fpsyg.2012.00322

O'Reilly, J. X., Woolrich, M. W., Behrens, T. E. J., Smith, S. M., & Johansen-Berg, H. (2012). Tools of the trade: Psychophysiological interactions and functional connectivity. *Social Cognitive and Affective Neuroscience, 7*(5), 604–609. doi:10.1093/scan/nss055

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25*(1), 46–59. doi:10.1002/hbm.20131

Patil, I. (2015). Trait psychopathy and utilitarian moral judgement: The mediating role of action aversion. *Journal of Cognitive Psychology, 27*(3), 349–366.

doi:10.1080/20445911.2015.1004334

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*(1), 94–107. doi:10.1080/17470919.2013.870091

Patil, I., Melsbach, J., Hennig-Fast, K., & Silani, G. (2016). Role of alexithymia in emotional processing deficits in adults with autism spectrum disorder. *In Prep.*

Patil, I., & Silani, G. (2014a). Alexithymia increases moral acceptability of accidental harms. *Journal of Cognitive Psychology, 26*(5), 597–614. doi:10.1080/20445911.2014.929137

Patil, I., & Silani, G. (2014b). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology, 5*, 501. doi:10.3389/fpsyg.2014.00501

Patil, I., Young, L., Sinay, V., & Gleichgerrcht, E. (2016). Elevated moral condemnation of third-party violations in multiple sclerosis patients. *Under Revision*.

Paulus, C. (2009). Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity.

Paxton, J. M., Ungar, L., & Greene, J. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–77. doi:10.1111/j.1551-6709.2011.01210.x

Pernet, C. (2014). Misconceptions in the use of the General Linear Model applied to functional MRI: a tutorial for junior neuro-imagers. *Frontiers in Neuroscience, 8*, 1. doi:10.3389/fnins.2014.00001

Pernet, C., Wilcox, R., & Rousselet, G. (2013). Robust correlation analyses: False positive and

power validation using a new open source matlab toolbox. *Frontiers in Psychology*, *3*, 606. doi:10.3389/fpsyg.2012.00606

Peyron, R., Laurent, B., & García-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis. *Neurophysiologie Clinique*, *30*(5), 263–88. doi:10.1016/S0987-7053(00)00227-6

Phillips, J., & Shaw, A. (2015). Manipulating Morality: Third-Party Intentions Alter Moral Judgments by Changing Causal Reasoning. *Cognitive Science*, *39*(6), 1320–47. doi:10.1111/cogs.12194

Poldrack, R. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, *2*(1), 67–70. doi:10.1093/scan/nsm006

Poldrack, R., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage*, *40*(2), 409–414. doi:10.1016/j.neuroimage.2007.11.048

Poldrack, R., Mumford, J., & Nichols, T. (2011). *Handbook of functional MRI data analysis* (1st ed.). New York: Cambridge University Press.

Power, J. D., Barnes, K. a., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. doi:10.1016/j.neuroimage.2011.10.018

Prehn, K., Wartenburger, I., Mériau, K., Scheibe, C., Goodenough, O. R., Villringer, A., … Heekeren, H. R. (2008). Individual differences in moral judgment competence influence neural correlates of socio-normative judgments. *Social Cognitive and Affective*

Neuroscience, *3*(1), 33–46. doi:10.1093/scan/nsm037

Price, J., Cole, V., & Goodwin, G. M. (2009). Emotional side-effects of selective serotonin reuptake inhibitors: qualitative study. *The British Journal of Psychiatry* , *195*(3), 211–217. doi:10.1192/bjp.bp.108.051110

Raichle, M., & Mintun, M. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, *29*, 449–76.

Raven, J., Raven, J. C., & Court, J. (1998). *Manual for Raven's progressive matrices and vocabulary scales*.

Ridgway, G. R., Henley, S. M. D., Rohrer, J. D., Scahill, R. I., Warren, J. D., & Fox, N. C. (2008). Ten simple rules for reporting voxel-based morphometry studies. *NeuroImage*, *40*(4), 1429–1435. doi:10.1016/j.neuroimage.2008.01.003

Ridgway, G. R., Litvak, V., Flandin, G., Friston, K., & Penny, W. D. (2012). The problem of low variance voxels in statistical parametric mapping; a new hat avoids a "haircut." *NeuroImage*, *59*(3), 2131–2141. doi:10.1016/j.neuroimage.2011.10.027

Robinson, J. S., Joel, S., & Plaks, J. E. (2015). Empathy for the group versus indifference toward the victim: Effects of anxious and avoidant attachment on moral judgment. *Journal of Experimental Social Psychology*, *56*, 139–152. doi:10.1016/j.jesp.2014.09.017

Roge, B., & Mullet, E. (2011). Blame and forgiveness judgements among children, adolescents and adults with autism. *Autism*, *15*(6), 702–712. doi:10.1177/1362361310394219

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., & Convit, A. (2007). Who cares? Revisiting empathy in Asperger syndrome. *Journal of Autism and Developmental Disorders*, *37*(4),

709–715. doi:10.1007/s10803-006-0197-8

Rousselet, G., & Pernet, C. (2012). Improving standards in brain-behavior correlation analyses. *Frontiers in Human Neuroscience*, *6*, 119. doi:10.3389/fnhum.2012.00119

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*(December), 59. doi:10.3389/neuro.08.059.2009

Royzman, E. B., Landy, J. F., & Leeman, R. F. (2015). Are Thoughtful People More Utilitarian? CRT as a Unique Predictor of Moral Minimalism in the Dilemmatic Context. *Cognitive Science*, *39*(2), 325–352. doi:10.1111/cogs.12136

Ruff, C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, *342*(6157), 482–4. doi:10.1126/science.1241399

Rutherford, M. D., & Troje, N. F. (2012). IQ predicts biological motion perception in autism spectrum disorders. *Journal of Autism and Developmental Disorders* , *42*(4), 557–65. doi:10.1007/s10803-011-1267-0

Salminen, J. K., Saarijärvi, S., Äärelä, E., Toikka, T., & Kauhanen, J. (1999). Prevalence of alexithymia and its association with sociodemographic variables in the general population of Finland. *Journal of Psychosomatic Research*, *46*(1), 75–82. doi:10.1016/S0022-3999(98)00053-1

Samson, A. C., Hardan, A. Y., Lee, I. a., Phillips, J. M., & Gross, J. J. (2015). Maladaptive Behavior in Autism Spectrum Disorder: The Role of Emotion Experience and Emotion Regulation. *Journal of Autism and Developmental Disorders*, *45*(11), 3424–32.

doi:10.1007/s10803-015-2388-7

Sanfey, A., Rilling, J., Aronson, J., Nystrom, L., & Cohen, J. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*(5626), 1755–1758. doi:10.1126/science.1082976

Sarlo, M., Lotto, L., Rumiati, R., & Palomba, D. (2014). If it makes you feel bad, don't do it! Egoistic rather than altruistic empathy modulates neural and behavioral responses in moral dilemmas. *Physiology & Behavior*, *130*, 127–34. doi:10.1016/j.physbeh.2014.04.002

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *NeuroImage*, *30*(4), 1088–96; discussion 1097–9. doi:10.1016/j.neuroimage.2005.12.062

Schneider, K., Pauly, K. D., Gossen, A., Mevissen, L., Michel, T. M., Gur, R. C., … Habel, U. (2013). Neural correlates of moral reasoning in autism spectrum disorder. *Social Cognitive and Affective Neuroscience*, *8*(6), 702–10. doi:10.1093/scan/nss051

Schwarzkopf, D. S., De Haas, B., & Rees, G. (2012). Better Ways to Improve Standards in Brain-Behavior Correlation Analysis. *Frontiers in Human Neuroscience*, *6*(July), 1–6. doi:10.3389/fnhum.2012.00200

Schwitzgebel, E., & Cushman, F. (2012). Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind and Language*, *27*(2), 135–153. doi:10.1111/j.1468-0017.2012.01438.x

Segerdahl, A. R., Mezue, M., Okell, T. W., Farrar, J. T., & Tracey, I. (2015). The dorsal posterior insula subserves a fundamental role in human pain. *Nature Neuroscience*, *18*(4),

499–500. doi:10.1038/nn.3969

Sellaro, R., Güroğlu, B., Nitsche, M. A., van den Wildenberg, W. P. M., Massaro, V., Durieux, J., … Colzato, L. S. (2015). Increasing the role of belief information in moral judgments by stimulating the right temporoparietal junction. *Neuropsychologia*, *77*, 400–408. doi:10.1016/j.neuropsychologia.2015.09.016

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*, *8*(4), 300–311. doi:10.1038/nrn2119

Shenhav, A., & Greene, J. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience*, *34*(13), 4741–9. doi:10.1523/JNEUROSCI.3390-13.2014

Shulman, C., Guberman, A., Shiling, N., & Bauminger, N. (2012). Moral and social reasoning in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *42*(7), 1364–76. doi:10.1007/s10803-011-1369-8

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of Causation, Responsibility, and Punishment in Cases of Harm-Doing. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *13*(3), 238–253. doi:10.1037/h0081183

Sifneos, P. E. (1973). The prevalence of "alexithymic" characteristics in psychosomatic patients. *Psychotherapy and Psychosomatics*, *22*(2), 255–262. doi:10.1159/000286529

Silani, G., Bird, G., Brindley, R., Singer, T., Frith, C., & Frith, U. (2008). Levels of emotional awareness and autism: an fMRI study. *Social Neuroscience*, *3*(2), 97–112. doi:10.1080/17470910701577020

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13*(8), 334–340. doi:10.1016/j.tics.2009.05.001

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* , *303*(5661), 1157–1162. doi:10.1126/science.1093535

Sloman, S., Fernbach, P. M., & Ewing, S. (2009). Causal Models: The Representational Infrastructure for Moral Judgment. In *Psychology of Learning and Motivation* (Vol. 50, pp. 1–26). doi:10.1016/S0079-7421(08)00401-5

Smith, A. (2009). The empathy imbalance hypothesis of autism: a theoretical approach to cognitive and emotional empathy in autistic development. *The Psychological Record*, *59*, 489–510.

Spino, J., & Cummins, D. D. (2014). The Ticking Time Bomb: When the Use of Torture Is and Is Not Endorsed. *Review of Philosophy and Psychology*, *5*(4), 543–563. doi:10.1007/s13164-014-0199-y

Streiner, D. L. (2005). Finding our way: An introduction to path analysis. *Canadian Journal of Psychiatry*, *50*(2), 115–122.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–8. doi:10.1016/j.cognition.2011.01.018

Swart, M., Kortekaas, R., & Aleman, A. (2009). Dealing with feelings: Characterization of trait Alexithymia on emotion regulation strategies and cognitive-emotional processing. *PLoS*

ONE, *4*(6). doi:10.1371/journal.pone.0005751

Szekely, R., & Miu, A. C. (2015). Incidental emotions in moral dilemmas: The influence of emotion regulation. *Cognition & Emotion*, *29*(1), 64–75. doi:10.1080/02699931.2014.895300

Taber-Thomas, B. C., Asp, E. W., Koenigs, M., Sutterer, M., Anderson, S. W., & Tranel, D. (2014). Arrested development: early prefrontal lesions impair the maturation of moral judgement. *Brain* , *137*(Pt 4), 1254–61. doi:10.1093/brain/awt377

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., … Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Social Cognitive and Affective Neuroscience*, *7*(3), 282–8. doi:10.1093/scan/nsr008

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between Judgment and Choice of Action in Moral Dilemmas. *Frontiers in Psychology*, *4*(May), 250. doi:10.3389/fpsyg.2013.00250

Theriault, J., & Young, L. (2014). Taking an "Intentional Stance"on Moral Psychology'. In J. Systma (Ed.), *Advances in Experimental Philosophy of Mind* (pp. 101–124). Continuum Press.

Thomson, J. J. (1985). The Trolley Problem. *Yale Law Journal*, *94*(6), 1395. doi:10.1119/1.1976413

Timoney, L. R., & Holder, M. D. (2013). *Emotional Processing Deficits and Happiness: Assessing the Measurement, Correlates, and Well-Being of People with Alexithymia*. *Emotional Processing Deficits and Happiness*. doi:10.1007/978-94-007-7177-2

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., … Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience*, *17*(9), 1270–5. doi:10.1038/nn.3781

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient Kill-Save Ratios Ease Up the Cognitive Demands on Counterintuitive Moral Utilitarianism. *Personality & Social Psychology Bulletin*, *40*(7), 923–930. doi:10.1177/0146167214530436

Trémolière, B., & Djeriouat, H. (2016). The sadistic trait predicts minimization of intention and causal responsibility in moral judgment. *Cognition*, *146*, 158–171. doi:10.1016/j.cognition.2015.09.014

Uddin, L. Q. (2014). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, *16*(1), 55–61. doi:10.1038/nrn3857

Ugazio, G., Majdandžić, J., & Lamm, C. (2014). Are Empathy and Morality Linked? Insights from Moral Psychology, Social and Decision Neuroscience, and Philosophy. In H. L. Maibom (Ed.), *Empathy and Morality* (pp. 155–171). Oxford University Press.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476–7. doi:10.1111/j.1467-9280.2006.01731.x

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Psychological Science*, *4*(3), 274–290. doi:10.1111/j.1745-6924.2009.01132.x

Waldmann, M., & Dieterich, J. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247–253.

doi:10.1111/j.1467-9280.2007.01884.x

Waldmann, M., Nagel, J., & Wiegmann, A. (2012). Moral Judgment . *The Oxford Handbook of Thinking and Reasoning*, (19), 274–299.

Walter, N. T., Montag, C., Markett, S., Felten, A., Voigt, G., & Reuter, M. (2012). Ignorance is no excuse: moral judgments are influenced by a genetic variation on the oxytocin receptor gene. *Brain and Cognition*, *78*(3), 268–73. doi:10.1016/j.bandc.2012.01.003

Weiskopf, N., Hutton, C., Josephs, O., & Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. *Neuroimage*, *33*(2), 493–504.

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology*, *13*(4), e1002128. doi:10.1371/journal.pbio.1002128

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition*, *126*(3), 364–72. doi:10.1016/j.cognition.2012.11.002

Wiegmann, A., & Waldmann, M. (2014). Transfer effects between moral dilemmas: a causal model theory. *Cognition*, *131*(1), 28–43. doi:10.1016/j.cognition.2013.12.004

Wilke, M. (2014). Isolated Assessment of Translation or Rotation Severely Underestimates the Effects of Subject Motion in fMRI Data. *PLoS ONE*, *9*(10), e106498. doi:10.1371/journal.pone.0106498

Wiss, J., Andersson, D., Slovic, P., Västfjäll, D., & Tinghög, G. (2015). The influence of identifiability and singularity in moral decision making. *Judgment and Decision Making*, *10*(5), 492–502.

Woo, C. W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419. doi:10.1016/j.neuroimage.2013.12.058

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301. doi:10.1016/j.cognition.2005.05.002

World Health Organization. (1992). The ICD-10 Classification of Mental and Behavioural Disorders. *International Classification*, *10*, 1–267.

Yamada, M., Camerer, C. F., Fujie, S., Kato, M., Matsuda, T., Takano, H., … Takahashi, H. (2012). Neural circuits in the brain that are activated when mitigating criminal sentences. *Nature Communications*, *3*, 759. doi:10.1038/ncomms1757

Ye, H., Chen, S., Huang, D., Zheng, H., Jia, Y., & Luo, J. (2015). Modulation of neural activity in the temporoparietal junction with transcranial direct current stimulation changes the role of beliefs in moral judgment. *Frontiers in Human Neuroscience*, *9*(659). doi:10.3389/fnhum.2015.00659

Yoder, K. J., & Decety, J. (2014). The Good, the bad, and the just: justice sensitivity predicts neural response during moral evaluation of actions performed by others. *The Journal of Neuroscience*, *34*(12), 4161–6. doi:10.1523/JNEUROSCI.4648-13.2014

Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010). Damage to Ventromedial Prefrontal Cortex Impairs Judgment of Harmful Intent. *Neuron*, *65*(6), 845–851. doi:10.1016/j.neuron.2010.03.003

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(15), 6753–6758. doi:10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(20), 8235–40. doi:10.1073/pnas.0701408104

Young, L., Koenigs, M., Kruepke, M., & Newman, J. P. (2012). Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology*, *121*(3), 659–67. doi:10.1037/a0027489

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the Neural and Cognitive Basis of Moral Luck: It's Not What You Do but What You Know. *Review of Philosophy and Psychology*, *1*(3), 333–349. doi:10.1007/s13164-010-0027-y

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*(4), 1912–20. doi:10.1016/j.neuroimage.2008.01.057

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072. doi:10.1016/j.neuropsychologia.2009.03.020

Young, L., & Tsoi, L. (2013). When Mental States Matter, When They Don't, and What That Means for Morality. *Social and Personality Psychology Compass*, *7*(8), 585–604. doi:10.1111/spc3.12044

Yu, H., Li, J., & Zhou, X. (2015). Neural Substrates of Intention–Consequence Integration and Its Impact on Reactive Punishment in Interpersonal Transgression. *The Journal of Neuroscience*, *35*(12), 4917–4925. doi:10.1523/JNEUROSCI.3536-14.2015

Zalla, T., Barlassina, L., Buon, M., & Leboyer, M. (2011). Moral judgment in adults with autism spectrum disorders. *Cognition*, *121*(1), 115–26. doi:10.1016/j.cognition.2011.06.004

Zalla, T., & Leboyer, M. (2011). Judgment of Intentionality and Moral Evaluation in Individuals with High Functioning Autism. *Review of Philosophy and Psychology*, *2*(4), 681–698. doi:10.1007/s13164-011-0048-1

Appendices

Appendix: Chapter 1

Appendix S1: 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?

Appendix S2:

Here we explain the exact details of how participants kept track of the 10-second response limit in VR scenarios (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):

Appendix: Chapter 2

Text S1: Questionnaires used and their internal reliability

Autism Spectrum Quotient (AQ-k): Autistic traits were assessed in all participants with the shortened, German-validated, 33-item version of the Autism Spectrum Quotient self-report questionnaire (Baron-Cohen et al., 2001; Freitag et al., 2007) designed for both clinical and community samples. This scale is further divided into three subscales: social interaction and spontaneity (SIS, 11 items; e.g. "I enjoy meeting new people"), imagination and creativity (IC, 12 items; e.g. "When I'm reading a story, I can easily imagine what the characters might look like"), communication and reciprocity (CR, 10 items; e.g. "I frequently find that I don't know how to keep a conversation going"). All questions were rated on a 4-point Likert scale from "Definitely Agree" to "Definitely Disagree" and were later recoded to 0 and 1.

Depression: Depressive symptoms in both groups were measured using Beck Depression Inventory (Beck et al., 1996; Hautzinger, 1991).

Mehrfachwahl-Wortschatz-Intelligenz-Test (Multiple choice vocabulary test, *MWT*): MWT-B is the most commonly used version of MWT and is considered a measure of verbal intelligence (Lehrl et al., 1995; Lehrl, 1995). It consists of 37 items and each item consists of five words (e.g. nesa - naso - nose - neso - nosa), out of which one authentic word needs to be recognized by the participants. Familiarity of the words varies widely and each correctly recognized word gives a point (thus possible scores range from

0 to 37). MWT-B has been shown to have good test-retest reliability and tends to be highly correlated with the other widely used measure of global IQ, viz. WAIS-Full-IQ.

Raven's Standard Progressive Matrices (SPM): SPM is considered a nonverbal estimate of fluid intelligence and the abbreviated version consisting of nine items (Form-A) was administered in the current study (Bilker et al., 2012; Raven et al., 1998). SPM items involve increasingly difficult pattern matching tasks and rely to a little degree on language abilities. Each correct answer is allotted one point and thus possible scores range from 0 to 9. This nine-item version has been shown to have good test-retest reliability and a high correlation with the full form SPM (Bilker et al., 2012).

Toronto Alexithymia Scale (TAS): Interindividual differences in subclinical alexithymia were evaluated using the German-validated 18-item Toronto Alexithymia Scale (Bagby et al., 1994; Kupfer et al., 2000) consisting of three subscales: Difficulty Identifying Feelings (DIF, 7 items; e.g., "When I am upset, I don't know if I am sad, frightened, or angry"), Difficulty Describing Feelings (DDF, 5 items; e.g., "It is difficult for me to find the right words for my feelings"), and Externally-Oriented Thinking (EOT, 6 items; e.g., "I prefer to analyze problems rather than just describe them"). Each item consisted of statements about emotional awareness and participants reported their agreement with these statements using a 5-point-Likert scale (1: *strongly disagree*, 5: *strongly agree*). TAS has been argued to be the best current measure overall for assessing alexithymia due to its sound reliability, validity, and broad generalizability (Timoney & Holder, 2013).

Interpersonal Reactivity Index (IRI): The Interpersonal Reactivity Index (Davis, 1983; Paulus, 2009) was used to assess specific aspects [fantasizing, empathic concern (EC), perspective-taking (PT), and personal distress (PD)] of dispositional empathy. The scale consisted of 16-items (four per subscale) and participants reported agreement with statements on a 5-point Likert scale (1: *never true for me*, 5: *always true for me*). Based on recent psychometric assessments of the IRI questionnaire (Baldner & McGinley, 2014), we *a priori* decided not to explore the fantasy subscale beyond descriptive statistics, as it does not map well onto the current neuroscientific discourse on empathy. Additionally, we focus on individual components of empathy rather than focusing on the entire construct of empathy since this approach provides more fine-grained understanding about empathy deficits in clinical populations.

Emotion Regulation Questionnaire (ERQ): Frequency of cognitive reappraisal and expressive suppression strategies to regulate emotions in everyday life was assessed with the Emotion Regulation Questionnaire (Abler & Kessler, 2009; Gross & John, 2003). Participants reported agreement with each statement using a 7-point Likert scale (1: *strongly disagree*, 7: *strongly agree*). Cognitive reappraisal is a cognitive strategy involving reinterpretation of events to reduce their emotional impact (6 items; e.g., ‘‘I control my emotions by changing the way I think about the situation I’m in.’’) and expressive suppression includes response-focused regulation involving inhibition of emotion-expressive behavior (4 items; e.g., ‘‘I control my emotions by not expressing them.’’).

Multifaceted Empathy Test (MET): To compliment the self-report paradigm (i.e., IRI), we also used a more naturalistic and ecologically valid performance measure, Multifaceted Empathy Test (Dziobek et al., 2008), to assess both cognitive and affective component of empathy for positive and negative emotions. We used the new, improved version MET-CORE-II (condensed and

revised; Isabel Dziobek, personal correspondence) which includes 20 negative and 20 positive photographic stimuli (presented in blocks of either positive or negative emotional valence which consisted of 10 pictures randomized within each block) that depict people in emotionally charged contexts. In cognitive empathy condition (i.e., block 1, 4, 6, and 7: "What is this person feeling?"), participants had to choose an appropriate emotion from four available options (e.g., scared, despaired, confused, impatient.) and their accuracy and response times were recorded. In emotional empathy condition (i.e., block 2, 3, 5, and 8: "How much do you feel with this person?"), the degree of empathic concern participants felt for the person in the picture was assessed on a 9-point Likert scale (1: *not at all* , 9: *very much*) and response time data was also recorded.

- ***Internal reliability***

Group differences in Cronbach's alphas were investigated using `cocron` package in R (http://comparingcronbachalphas.org/) which implements inferential statistics on alphas.

Scale	HC	ASD	Item coun	$\chi^2(1)$	*p*
AQ-k (*n* = 17)					
SIS	0.592	0.887	11	5.040	0.025
IC	0.635	0.753	12	0.507	0.477
CR	0.423	0.718	10	1.616	0.204
SPF-IRI (*n* = 17)					
FS	0.490	0.818	4	2.327	0.127
EC	0.765	0.564	4	0.863	0.353
PT	0.715	0.720	4	0.001	0.979
PD	0.545	0.741	4	0.720	0.396
TAS (*n* = 17)					
DDF	0.519	0.811	5	2.167	0.141
DIF	0.547	0.694	7	0.449	0.503
EOT	0.509	0.597	6	0.109	0.742
ERQ (*n* = 17)					
Reappraisal	0.865	0.819	6	0.239	0.625
Suppression	0.569	0.782	4	1.046	0.306
Moral dilemma task (*n* = 15)					
Impersonal - behavior	0.857	0.645	6	1.934	0.164
Personal - behavior	0.566	0.664	6	0.159	0.691

Text S2: Textual description of moral dilemmas

Impersonal and personal moral dilemmas were chosen from previously published batteries of moral dilemmas (Greene et al., 2004; Lotto, Manfrinati, & Sarlo, 2014; Moore, Clark, & Kane, 2008; Patil et al., 2014) and posed a conflict between actively harming less number of individuals for the welfare of many. In both impersonal and personal moral dilemmas, a number of factors that previous research has shown to affect moral judgments varied freely in order to increase heterogeneity and thus decrease predictability of experimental stimuli, e.g. whether the sacrificial actions benefited self or other (Lotto et al., 2014), whether the victim's death was inevitable (Moore et al., 2008), kill-save ratios (Trémolière & Bonnefon, 2014) (e.g., 1:6, 1:100s, etc.), etc.

Full descriptions of the scenarios used in the moral dilemma task are provided here. The German translations of the scenarios are available from the authors on request.

Non-moral scenarios

1. Two trips

You are bringing home some plants from the store. You have lined the trunk of your car with plastic to catch the mud from the plants, but your trunk will not hold all of the plants you have bought.

You could bring all of the plants home in one trip, but you would need to put some of the plants in the back seat. If you put the plants in the back seat, the mud from the plants will ruin your fine leather upholstery, which would cost thousands of dollars to replace.

2. Mutual Fund

You are at home one day when the mail arrives. You receive a letter from a company that provides financial services. You have heard of this company, which has a good reputation. They have invited you to invest in a mutual fund. The minimum investment for this fund is €1000.

You already know a lot about this particular mutual fund. It has performed poorly over the past few years. Based on what you know, there is no reason to think that it will perform any better in the future.

3. Jogging

You intend to accomplish two things this afternoon: going for a jog and doing some paperwork. In general you prefer to get your work done before you exercise.

The weather is nice at the moment, but the weather forecast says that in a couple of hours it will start to rain. You very much dislike jogging in the rain, but you don't care what the weather is like while you do paperwork.

4. Pharmacy

You have a very bad headache. You go to the pharmacy looking for your favorite brand of headache medicine. When you get there, you find that the pharmacy is out of the brand that you are looking for.

You have known the pharmacist at this store for a long time, and you trust him. He says he has a generic medicine that is "exactly the same" as the name-brand medicine that you wanted. In the past, he has always given you good advice.

5. Meeting

You need to travel to a nearby city in order to attend a meeting that starts at 2:00 PM. You can either take the train or the bus. The train will get you there just in time for your meeting no matter what.

The bus is scheduled to arrive an hour before your meeting. However, the bus sometimes is several hours late because of traffic. It would be nice to have an extra hour before the meeting, but it is very important that you arrive on time.

6. Scenery

An old friend invites you to spend the weekend at his summer home. This house is up the coast from where you live. You plan to drive, and you can take either the highway or the coastal road.

The highway will get you there in about three hours, but the scenery along the highway is very boring. The coastal road will get you to your friend's house in about three hours and fifteen minutes, and the scenery along the coastal road is very beautiful.

Moral impersonal

7. Hospital fumes

You are the late-night watchman in a hospital. Due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. In a certain room of the hospital are three patients. In another room there is a single patient. If you do nothing the fumes will rise up into the room containing the three patients and cause their deaths.

The only way to avoid the deaths of these patients is to hit a certain switch, which will cause the fumes to bypass the room containing the three patients. As a result of doing this the fumes will enter the room containing the single patient, causing his death.

8. Trolley

You are the driver of a runaway trolley approaching a fork in the tracks. On the tracks going to the left is a group of five railway workers. On the tracks going to the right is a single railway worker.

If you do nothing, the trolley will go to the left, causing the five workers to die. The only way to avoid the deaths of these five workers is to hit a switch on your dashboard that will make the trolley go to the right, leading to the death of the single worker.

9. Motor boat

You are driving your motor boat in a small bay when your attention is drawn to cries of help from five people who are drowning at the end of a very narrow channel which is right in front of you. Between you and the people who are drowning, to one side of the channel, is another person who is calmly swimming.

If you do nothing, these five people would drown soon. You can steer towards the end of the channel at high speed. But you know that the person who is swimming in between will be hit by the motorboat and die, but the other five people will be saved.

10. Burning Car

You are the operator of a bridge in a huge parking complex. The bridge you operate connects two different floors. You see that an out of control car is approaching the bridge and is on fire. 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.

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.

11. Lifting magnet

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.

12. Nurse

You are a nurse who is in charge of a machine which controls drug dosage levels in patients' blood. Because of a technical failure, the machine is supplying a lethal dose of the drug to four patients. Another patient, in a single room, is hooked up to the same machine and has not undergone any variation in dosage.

If nothing is done, these four patients would die due to lethal poisoning caused by drug overdose. You can press the button to block the drug supply to the four patients. You know that the overdose of drug will be redirected to the patient in the single room, who will die, but the other four will be saved.

Moral personal

13. Footbridge

A runaway trolley is heading down the tracks toward five workers, and will kill them if it keeps going. You are on a footbridge over the tracks, in between the approaching trolley and the five workers. Next to you on this footbridge is a stranger who is very large.

The only way to save the lives of the five workers is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workers will be saved.

14. Miners

You are leading a rescue team for seven miners that are stuck in an underground mine, which is flooding. Six miners are trapped at the bottom and will drown if not rescued soon. One miner is trapped higher in the elevator shaft and will not drown.

The only way to rescue the six at the bottom is to quickly send down the mine elevator. If you do this, the one miner in the shaft will be crushed to death. If you do not send down the elevator, you will have enough time to rescue the one miner in the shaft.

15. Vitamins

You are leading a group that is lost in the wilderness. Your group includes a family of six with a genetic vitamin deficiency. A few people's kidneys contain large amounts of this vitamin. There is one such person in your group.

The only way to save the lives of the six family members is to remove one of this man's kidneys and take the necessary vitamins from it. He will not die if you do this, but his health will get worse. He does not want to give his kidney, but you have the power to do what you choose.

16. Epidemic

An epidemic has spread worldwide killing millions of people. You have developed two substances in your underground shelter. One of them is a cure but the other one is deadly. You don't know which is which.

Two people have run downstairs to your shelter trying to avoid the epidemic. The only way to identify the cure is to inject each of these people with one of the two substances. One person will live but the other will die. Then you will be able to start saving lives with the cure.

17. Savannah

You are carrying out research into the behaviour of lions in the Savannah of Central Africa. From an observation tower, you can see four people hiding behind a bush. They are being followed by a lion and are trying to get away from it. The lion has seen them and is heading for the bush. Meanwhile, another person has been able to climb the observation tower and is standing near you.

If you do nothing, the lion is going to find the four people hiding in the bush and kill them. You can push off the person standing near you so that the lion is drawn towards him. You know that this person will be mauled to death, but the other four will have time to escape.

18. Transplant

You are a young doctor working in the emergency surgery ward of your local hospital. Four of the patients under your care in your ward are dying from grave injury to vital organs as a result of a serious car accident. You have another patient in the ward who is now in good health and ready to be sent home.

If nothing is done, these four patients are going to die. You can anaesthetize the patient in good health and remove his vital organs to transplant them in the patients who are dying. You know that this patient will die, but the other four can be saved.

Text S3: Descriptive statistics for behavioral ratings for moral judgment tasks

Moral dilemma task									
Question	**Condition**	**HC (*n* = 16)**				**ASD (*n* = 15)**			
		Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
behaviour	non-moral	0.55	0.17	0.33	1.00	0.56	0.12	0.33	0.83
	impersonal	0.55	0.34	0	0.83	0.58	0.26	0.17	1.00
	personal	0.25	0.20	0	0.50	0.13	0.21	0	0.67
emotional arousal	non-moral	0.93	1.58	0	6.16	4.66	3.61	0.72	12.80
	impersonal	11.35	4.05	4.16	20.00	16.16	3.49	7.09	20.00
	personal	9.40	4.64	0.05	19.71	14.00	5.32	2.61	20.00

Text S4: Details for response time data

Descriptive statistics and group differences for response time data from the moral dilemma task. No response time data was available for the arousal ratings. Note that although we had response time data for the moral dilemma task, we do not carry out any analysis on this data to make an inference about underlying psychological processes, as this practice of reverse inference has been recently demonstrated to be problematic (Krajbich et al., 2015).

scenario	Type of response	HC					ASD					*t*	*df*	*p*
		n	**Min**	**Max**	**Mean**	**SD**	***n***	**Min**	**Max**	**Mean**	**SD**			
non-moral	average	16	3.80	11.25	6.11	2.07	15	4.38	21.65	7.95	4.36	-1.485	19.71	0.153
impersonal		16	3.12	32.02	7.01	7.14	15	3.78	24.04	7.26	5.01	-0.114	26.94	0.91
personal		16	3.06	24.30	6.81	5.20	15	3.46	12.84	6.35	2.64	0.315	22.56	0.756
impersonal	utilitarian	16	0	12.02	4.23	3.19	15	3.12	30.59	8.20	6.89	-2.037	19.45	0.055
	non-utilitarian	17	0	32.02	6.90	7.72	14	3.18	20.77	7.39	4.81	-0.213	27.21	0.833
personal	utilitarian	16	3.34	9.34	5.55	2.12	11	2.87	11.06	6.68	2.76	-1.141	17.79	0.269
	non-utilitarian	17	0	12.20	5.70	3.02	15	2.93	18.21	7.91	4.00	-1.747	25.9	0.092
non-moral	utilitarian	12	3.27	44.85	11.05	13.13	10	3.33	7.93	5.78	1.58	1.378	11.38	0.195
	non-utilitarian	16	2.50	16.91	5.99	3.75	15	3.82	36.70	8.41	8.12	-1.051	19.43	0.306

Note: The reported *p*-values are uncorrected for multiple comparisons and would have to be adjusted for factorial design of the study.

Text S5

Correlation (Spearman's rho) between arousal ratings and moral judgments on moral dilemma task. *$p < 0.05$ (two-tailed)

Correlation pair	Control ($n = 16$)		ASD ($n = 15$)		Fisher's *Z*-test
	ρ	*p*	ρ	*p*	
Non-moral behaviour-arousal	-0.261	0.330	-0.397	0.142	0.382
Impersonal behaviour-arousal	0.219	0.414	**-0.583**	**0.023**	**2.222***
Personal behaviour-arousal	0.384	0.142	-0.150	0.594	1.389

Text S6

Correlation (Spearman's rho) between moral judgments on the moral dilemma task and arousal ratings and ERQ. *$p < 0.05$ (two-tailed)

Threshold of significance = 0.0167						Fisher's *Z*-test	
		HC (*n* = 16)		ASD (*n* = 15)			
variable	statistic	ERQ - reappraisal	ERQ - suppression	ERQ - reappraisal	ERQ - suppression	ERQ - reappraisal	ERQ - suppression
non-moral affirmative behaviour	ρ	.441	-.449	-.055	.214	1.32	1.751
	p	.087	.081	.845	.443		
impersonal utilitarian behaviour	ρ	.068	.454	.382	.100	0.835	0.973
	p	.802	.077	.160	.724		
personal utilitarian behaviour	ρ	.088	.278	.302	.450	0.558	0.498
	p	.745	.296	.274	.092		
non-moral emotional arousal	ρ	-.449	-.308	.070	-.250	1.383	0.157
	p	.081	.246	.805	.368		
impersonal emotional arousal	ρ	.386	-.320	.132	-.001	0.685	0.826
	p	.139	.228	.640	.997		
personal emotional arousal	ρ	.276	-.309	.317	.107	0.112	1.066
	p	.301	.244	.250	.703		

Text S7

Correlation (Spearman's rho) between moral judgments on the moral dilemma task and arousal ratings and alexithymia. *$p < 0.05$ (two-tailed)

Threshold of significance = 0.0167				
variable	**statistic**	**HC (*n* = 16)**	**ASD (*n* = 15)**	**Fisher's *Z*-test**
non-moral affirmative behaviour	ρ	-.008	-.283	0.707
	p	.977	.307	
impersonal utilitarian behaviour	ρ	-.159	.030	0.476
	p	.556	.917	
personal utilitarian behaviour	ρ	-.088	.246	0.848
	p	.746	.378	
non-moral emotional arousal	ρ	.164	**0.765**	**2.105***
	p	.543	**.001**	
impersonal emotional arousal	ρ	-.181	.022	0.512
	p	.502	.937	
personal emotional arousal	ρ	-.059	-.165	0.268
	p	.827	.557	

Text S8

Correlation (Spearman's rho) between moral judgments on the moral dilemma task and arousal ratings and SPF-IRI. *$p < 0.05$ (two-tailed)

Threshold of significance = 0.0167										
variable	statistic	HC (n = 16)			ASD (n = 15)			Fisher's Z-test		
		EC	PT	PD	EC	PT	PD	EC	PT	PD
non-moral affirmative behaviour	ρ	.483	.258	.382	-.377	-.062	-.014	**2.307***	0.814	1.04
	p	.058	.334	.144	.166	.827	.959			
impersonal utilitarian behaviour	ρ	-.173	.121	-.331	-.331	-.093	.212	0.423	0.537	1.397
	p	.522	.655	.210	.229	.742	.447			
personal utilitarian behaviour	ρ	.035	.212	-.193	-0.573	-.329	-.315	1.716	1.391	0.326
	p	.896	.431	.474	.026	.232	.253			
non-moral emotional arousal	ρ	-.313	-.388	-.088	-.018	-.018	.341	0.764	0.978	1.108
	p	.238	.138	.746	.949	.949	.213			
impersonal emotional arousal	ρ	.381	.322	.263	**0.641**	-.042	.227	0.896	0.939	0.096
	p	.146	.224	.325	**.010**	.881	.416			
personal emotional arousal	ρ	.151	.364	.070	0.523	-.178	.251	1.07	1.402	0.466
	p	.576	.166	.796	.045	.525	.367			

Text S9

Correlation (Spearman's rho) between moral judgments on dilemma task and arousal ratings and MET performance. *$p < 0.05$ (two-tailed)

		Threshold of significance = 0.0167								**Fisher's *Z*-test**			
		HC (*n* = 16)				**ASD (*n* = 15)**							
variable	**statistic**	**Cognitive empathy positive correct answers**	**Cognitive empathy negative correct answers**	**Emotional empathy positive average**	**Emotional empathy negative average**	**Cognitive empathy positive correct answers**	**Cognitive empathy negative correct answers**	**Emotional empathy positive average**	**Emotional empathy negative average**	**Cognitive empathy positive correct answers**	**Cognitive empathy negative correct answers**	**Emotional empathy positive average**	**Emotional empathy negative average**
non-moral affirmative behaviour	ρ	-.178	-.232	-.175	-.129	-.260	-.159	.270	-.008	0.215	0.19	1.133	0.304
	p	.509	.386	.516	.635	.350	.572	.330	.977				
impersonal utilitarian behaviour	ρ	.332	.381	-.275	0.000	-.152	-.272	.487	.225	1.245	1.699	**2.034***	0.572
	p	.209	.146	.303	1.000	.588	.326	.066	.421				
personal utilitarian behaviour	ρ	.361	.158	.040	-.015	-.025	-.147	.008	-.227	1.007	0.768	0.08	0.54
	p	.170	.560	.884	.957	.930	.600	.977	.417				
non-moral emotional arousal	ρ	-.286	-.309	0.573	-.195	.375	-.289	-.100	.150	1.72	0.055	1.879	0.871
	p	.283	.244	.020	.470	.168	.296	.723	.594				
impersonal emotional arousal	ρ	.261	.141	-.109	.237	.167	.423	-.148	.304	0.246	0.773	0.099	0.181
	p	.329	.603	.688	.376	.552	.116	.597	.271				
personal emotional arousal	ρ	-.028	-.012	.012	.071	-.226	.107	-.011	.236	0.505	0.298	0.057	0.423
	p	.918	.964	.966	.794	.418	.705	.970	.398				

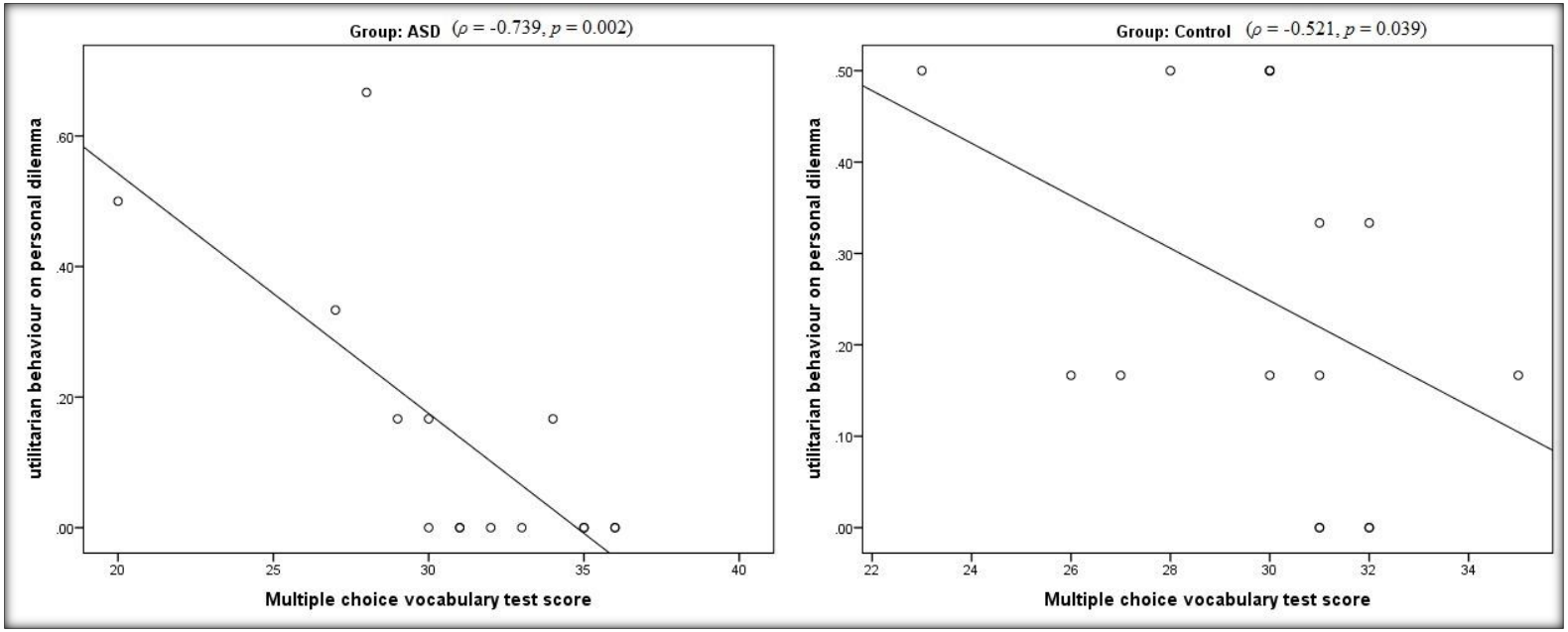
Text S10

Correlation (Spearman's rho) between moral judgments on dilemma task and arousal ratings and personality traits. *$p < 0.05$ (two-tailed)

variable	statistic	Threshold of significance = 0.0167								Fisher's Z-test			
		HC (n = 16)				ASD (n = 15)							
		AQ-k	SPM	MWT-B	BDI	AQ-k	SPM	MWT-B	BDI	AQ-k	SPM	MWT-B	BDI
non-moral affirmative behaviour	ρ	-.060	-.042	.102	-.418	-0.595	-.122	-.176	.129	1.562	0.201	0.7	1.436
	p	.825	.876	.707	.107	.019	.665	.531	.646				
impersonal utilitarian behaviour	ρ	-.358	**0.718**	-.047	.306	-.423	-.493	-.358	.041	0.192	**3.606***	0.818	0.687
	p	.173	**.002**	.862	.249	.116	.062	.190	.885				
personal utilitarian behaviour	ρ	-.281	.392	-0.521	.358	-.352	-.459	**-0.739**	-.096	0.197	**2.274***	0.926	1.176
	p	.293	.134	.039	.174	.198	.085	**.002**	.734				
non-moral emotional arousal	ρ	.057	-.135	-.325	.364	**0.715**	-.006	-.059	.212	**2.099***	0.324	0.695	0.415
	p	.833	.618	.219	.166	**.003**	.984	.834	.448				
impersonal emotional arousal	ρ	.163	.406	-.368	-.046	.497	-.052	.408	.161	0.951	1.206	**2.047***	0.521
	p	.546	.118	.160	.865	.060	.854	.131	.567				
personal emotional arousal	ρ	-.111	.248	-.440	.058	.123	-.111	.292	.239	0.587	0.911	1.931	0.464
	p	.682	.355	.088	.831	.663	.693	.291	.391				

Text S11

Verbal reasoning skills and moral judgments: The relation observed between global/verbal intelligence scores (as assessed by MWT-B) and utilitarian moral judgment on personal moral dilemmas was similar for the two groups (Z = 0.926, p = 0.354). In both controls and ASD group, higher MWT-B scores were associated with a reduced tendency to make utilitarian judgments. Note that the number of data-points in the scatterplot seems to be less than the sample sizes due to overlap between data-points (denoted by circles with thicker circumference). Reported p-values are two-tailed.


Group: ASD (ρ = -0.739, p = 0.002)
utilitarian behaviour on personal dilemma
.00
.20
.40
.60
20
25
30
35
40
Multiple choice vocabulary test score
Group: Control (ρ = -0.521, p = 0.039)
utilitarian behaviour on personal dilemma
.00
.10
.20
.30
.40
.50
22
24
26
28
30
32
34
Multiple choice vocabulary test score

Text S12: Details for the standardized regression coefficients for each path from the path analysis model

Since we had directional hypotheses for most paths, one-tailed p-values have been provided. *Note*: S.E. – Standard Error, C.R. – Critical Ratio, MJ – moral judgment for the behavior question on the personal moral dilemma

			β	p (1-tailed)	S.E.	C.R.
EC	<---	TAS	-0.313	0.100	0.254	-1.234
PD	<---	AQ	0.396	0.030	0.214	1.881
PD	<---	SPM	-0.461	0.015	0.214	-2.186
PD	<---	Medication status	0.091	0.334	0.422	0.429
EC	<---	Medication status	-0.047	0.426	0.501	-0.186
MJ	<---	AQ	-0.338	0.049	0.175	-1.651
MJ	<---	TAS	0.535	0.004	0.168	2.712
MJ	<---	EC	-0.491	0.001	0.132	-3.157
MJ	<---	PD	-0.418	0.013	0.157	-2.234
MJ	<---	SPM	-0.452	0.004	0.146	-2.646
MJ	<---	Medication status	0.142	0.171	0.250	0.952

Text S13

Hierarchical regression analyses (conducted separately in the control and ASD groups) determined whether alexithymia or ASD symptom severity predicted utilitarian responses on the behavior question for personal moral dilemmas once age, gender, and depression were controlled for, and whether each could predict the dependent variables after the other was controlled for. As noted in a previous study (Brewer et al., 2015), it is necessary to perform hierarchical regressions with alexithymia and ASD symptom severity entered in both possible orders to independently investigate the effect of each, after controlling for the other, because of collinearity. We acknowledge that limitation of the following analysis is again that the sample size is smaller than recommended (Maxwell, 2000). Note: All reported *p*-values are from two-tailed tests.

As mentioned in the main text, there was less amount of variation in trait alexithymia in control sample as compared to ASD sample (SD_{ASD} = 8.091, SD_{HC} = 3.890; Levene's test: $F(1,32) = 5.359$, $p = 0.027$), but not for AQ-k scores (SD_{ASD} = 3.238, SD_{HC} = 3.182; Levene's test: $F(1,32) = 0.046$, $p = 0.831$). This was a result of our methodological choice to not match the two groups for alexithymia scores (in contrast to prior recommendations (Bird & Cook, 2013)), since this requires oversampling the control group and consequently is unrepresentative with respect to distribution of alexithymic traits in this population. Given that there was not enough variation in TAS scores in control group with respect to the ASD group, but equivalent variation in AQ scores, we expected analogous result in the HC group only for the AQ but not TAS traits. We did not carry out moderation analysis to see if factor group moderated relationship between alexithymia severity and utilitarian judgments (like in a previous study (Brewer et al., 2015)), because the two groups were not matched for alexithymia scores.

- For ASD group (n = 15) with order ASD severity first, alexithymia second

Step	Predictor	ASD group : behavior ratings for personal dilemma				
		β	t	p	R^2	ΔR^2 (p)
1	(Constant)		2.704	0.021	31.2%	31.2% (0.232)
	Gender	-0.193	-0.751	0.468		
	Age	-0.498	-1.915	0.082		
	Depression	-0.186	-0.725	0.483		
2	(Constant)		1.621	0.136	32.6%	1.4% (0.657)
	Gender	-0.164	-0.599	0.563		
	Age	-0.472	-1.712	0.118		
	Depression	-0.166	-0.614	0.553		
	ASD severity	-0.126	-0.457	0.657		
3	(Constant)		1.120	0.292	72.2%	39.6% (0.006)
	Gender	-0.170	-0.919	0.382		
	Age	-0.319	-1.664	0.130		
	Depression	-0.073	-0.394	0.703		
	ASD severity	-0.701	-2.846	0.019		
	Alexithymia	0.840	3.583	0.006		

- For ASD group (n = 15) with order alexithymia first, ASD severity second

Step	Predictor	ASD group : behavior ratings for personal dilemma				
		β	t	p	R^2	ΔR^2 (p)
1	(Constant)		2.704	0.021	31.2%	31.2% (0.232)
	Gender	-0.193	-0.751	0.468		
	Age	-0.498	-1.915	0.082		
	Depression	-0.186	-0.725	0.483		
2	(Constant)		-0.080	0.938	47.2%	16% (0.112)
	Gender	-0.259	-1.087	0.303		
	Age	-0.481	-2.011	0.072		
	Depression	-0.186	-0.788	0.449		
	Alexithymia	0.406	1.743	0.112		
3	(Constant)		1.120	0.292	72.2%	25% (0.019)
	Gender	-0.170	-0.919	0.382		
	Age	-0.319	-1.664	0.130		
	Depression	-0.073	-0.394	0.703		
	Alexithymia	0.840	3.583	0.006		
	ASD severity	-0.701	-2.846	0.019		

As can be noted from the regression coefficients, both autistic and alexithymic traits were significant predictors of the utilitarian moral behavior on personal dilemmas, but with opposite signs. That is, greater severity of autistic traits was associated with increased endorsement of the option of inaction, while higher scores on alexithymia were characterized by greater endorsement of the utilitarian choice. Thus, the pattern revealed by complex path analysis was also observed in this simpler regression analysis.

Next, we investigated if the same pattern was also observed in the control sample.

- <u>For HC group (n = 16) with order ASD severity first, alexithymia second</u>

Step	Predictor	HCgroup : behavior ratings for personal dilemma				
		β	t	p	R^2	ΔR^2 (p)
1	(Constant)		5.369	< 0.001	69.4%	69.4% (0.002)
	Gender	-0.217	-1.285	0.223		
	Age	-0.788	-4.728	< 0.001		
	Depression	0.234	1.422	0.180		
2	(Constant)		8.511	< 0.001	86.3%	16.9% (0.004)
	Gender	-0.299	-2.493	0.030		
	Age	-0.848	-7.216	< 0.001		
	Depression	0.254	2.204	0.050		
	ASD severity	-0.423	-3.693	0.004		
3	(Constant)		4.405	0.001	86.4%	0.1% (0.890)
	Gender	-0.300	-2.383	0.038		
	Age	-0.844	-6.709	< 0.001		
	Depression	0.254	2.108	0.061		
	ASD severity	-0.418	-3.378	0.007		
	Alexithymia	-0.017	-0.142	0.890		

As can be seen from the final model, gender emerged as a significant negative predictor of utilitarian judgment, i.e. women were less likely to judge harming the few for the greater good than men, which agrees with a recent large-scale meta-analysis (Friesdorf, Conway, & Gawronski,

2015). Additionally, it was also observed that older people were less likely to endorse utilitarian judgment, but this result might be an artifact of small sample size as prior surveys with study sample large enough to investigate age-related variation on moral judgments revealed null results (Hauser et al., 2007).

- For HC group (n = 16) with order alexithymia first, ASD severity second

Step	Predictor	HC group : behavior ratings for personal dilemma				
		β	t	p	R^2	ΔR^2 (p)
1	(Constant)		5.369	< 0.001	69.4%	69.4% (0.002)
	Gender	-0.217	-1.285	0.223		
	Age	-0.788	-4.728	< 0.001		
	Depression	0.234	1.422	0.180		
2	(Constant)		3.060	0.011	70.8%	1.4% (0.481)
	Gender	-0.227	-1.313	0.216		
	Age	-0.767	-4.444	0.001		
	Depression	0.241	1.434	0.179		
	Alexithymia	-0.122	-0.729	0.481		
3	(Constant)		4.405	0.001	86.4%	15.6% (0.007)
	Gender	-0.300	-2.383	0.038		
	Age	-0.844	-6.709	< 0.001		
	Depression	0.254	2.108	0.061		
	Alexithymia	-0.017	-0.142	0.890		
	ASD severity	-0.418	-3.378	0.007		

As expected, we found evidence for decreased non-utilitarian tendency with autistic traits also in the control sample, but no evidence for alexithymic traits due to lack of enough variation in these traits. A prior study done with healthy sample did reveal utilitarian bias in trait alexithymia (Patil & Silani, 2014a), but this study did not investigate the role of autistic traits. Future studies should investigate divergent contributions of these two traits in a large sample consisting of healthy adults.

Appendix: Chapter 3

Text S1: Scenario details

Scenario type by version breakdown. Red and green cells denote scenarios taken from Cushman (2008) and Young, Camprodon, Hauser, Pascual-Leone, & Saxe (2010), respectively.

Note: The exact wording of the details can be found in the original papers or can be requested from the corresponding author. Italian translations are also available on request.

No.	scenario	v1	v2	v3	v4
1	Popcorn	neu	att	int	acc
2	Malaria Pond/African pond	att	int	acc	neu
3	Spinach	int	acc	neu	att
4	Peanut allergy	acc	neu	att	int
5	Rabies/Rabid dog	neu	att	int	acc
6	Meatloaf	att	int	acc	neu
7	Seatbelt/Amusement park	int	acc	neu	att
8	Teenagers/Skiing	acc	neu	att	int
9	Ham sandwich	neu	att	int	acc
10	Safety Cord/Rock climbing	att	int	acc	neu
11	Sesame seeds	int	acc	neu	att
12	Coffee/Chemical Plant	acc	neu	att	int
13	Bridge	neu	att	int	acc
14	Pool	att	int	acc	neu
15	Mushrooms	int	acc	neu	att
16	Latex	acc	neu	att	int
17	Motorboat	neu	att	int	acc
18	Asthma	att	int	acc	neu
19	Veterinarian/Dog poison	int	acc	neu	att
20	Zoo	acc	neu	att	int
21	Sushi	neu	att	int	acc
22	Cayo/Monkeys	att	int	acc	neu
23	Wet floor	int	acc	neu	att
24	Lab	acc	neu	att	int
25	Vitamin	neu	att	int	acc
26	Airport	att	int	acc	neu
27	Chairlift	int	acc	neu	att
28	Bike	acc	neu	att	int
29	Safety Town/Fire drill	neu	att	int	acc
30	Parachute	att	int	acc	neu
31	Sculpture	int	acc	neu	att
32	Dentist	acc	neu	att	int
33	Iron	neu	att	int	acc
34	Tree House	att	int	acc	neu
35	Jellyfish/Ocean	int	acc	neu	att
36	Laptop	acc	neu	att	int

Text S2: Additional details on nature of stimuli used

When faced with possible harmful situations, human judges tend to perceive them in terms of a moral dyad consisting of (*a*) a moral agent with capacity for purposeful action and goal-directed behavior who is attributed *moral responsibility* for intending to cause or causing harm and (*b*) a moral patient/victim with capacity for sensations and feelings and is attributed *moral rights* that need to be defended (Gray & Schein, 2012; Theriault & Young, 2014). In other words, while assessing behavior of a perpetrator, judges need to simulate both epistemic (beliefs, knowledge, desires, etc.) and feeling (pain, suffering, etc.) states in others.

But we note that intent and harmfulness inputs represent *sufficient* but not *necessary* inputs to moral judgment (Inbar, Pizarro, & Cushman, 2012). Additionally, a number of other factors that have been shown to influence moral judgments about third-party violations were held constant across scenarios. In none of the scenarios, victims were responsible for their own fate since such scenarios tend to elicit reduced empathic reasoning about victims (Fehse, Silveira, Elvers, & Blautzik, 2014). Also, none of the scenarios systematically manipulated information about how reasonable the agent's belief was (Young, Nichols, et al., 2010) or the nature of agent's desires (Cushman, 2008; Laurent, Nuñez, & Schweitzer, 2015). Additionally, all scenarios were formulated in such a way that the agent was in control of his/her own behavior (Martin & Cushman, 2016). The agent was causally responsible for the outcome and no information that would diminish agent's perceived responsibility for the outcome was presented (apart from belief information), e.g. mitigating circumstances (Buckholtz et al., 2008; Yamada et al., 2012) or external constraints on the agent by third-parties (Phillips & Shaw, 2015; Woolfolk, Doris, & Darley, 2006). Moreover, when present, the nature of harmful outcome was

described in a plain rather than graphic language (Treadway et al., 2014). Importantly, all protagonists in scenarios had an obligation towards victims (due to their role in relational context) and possessed the capacity to foresee and prevent the event (Malle et al., 2014).

Text S3: Experimental protocol

The study was approved by the ethics committee of the hospital "Santa Maria della Misericordia" (Udine, Italy) and the data were collected at the same hospital. There was no restriction on handedness of participants (8 left-handed, as assessed using self-report) and all participants had normal or corrected-to-normal vision. Rule-out criteria for participation included Italian as a secondary language, presence of a diagnosed psychiatric illness and/or history of psychiatric treatment, history of significant neurological illness or brain injury, and current usage of psychoactive drugs.

Subjects were completely agnostic to the purpose of the experiment and did not receive any information about the nature of the experiment apart from the fact that it involved decision-making in social context. There were no practice trials before the actual experiment as the experimental protocol employed was easily comprehensible and participants were given general instructions about the nature of stimuli and handling the response pad before they entered the scanner. For all tasks, the stimuli were presented in a rapid event-related design.

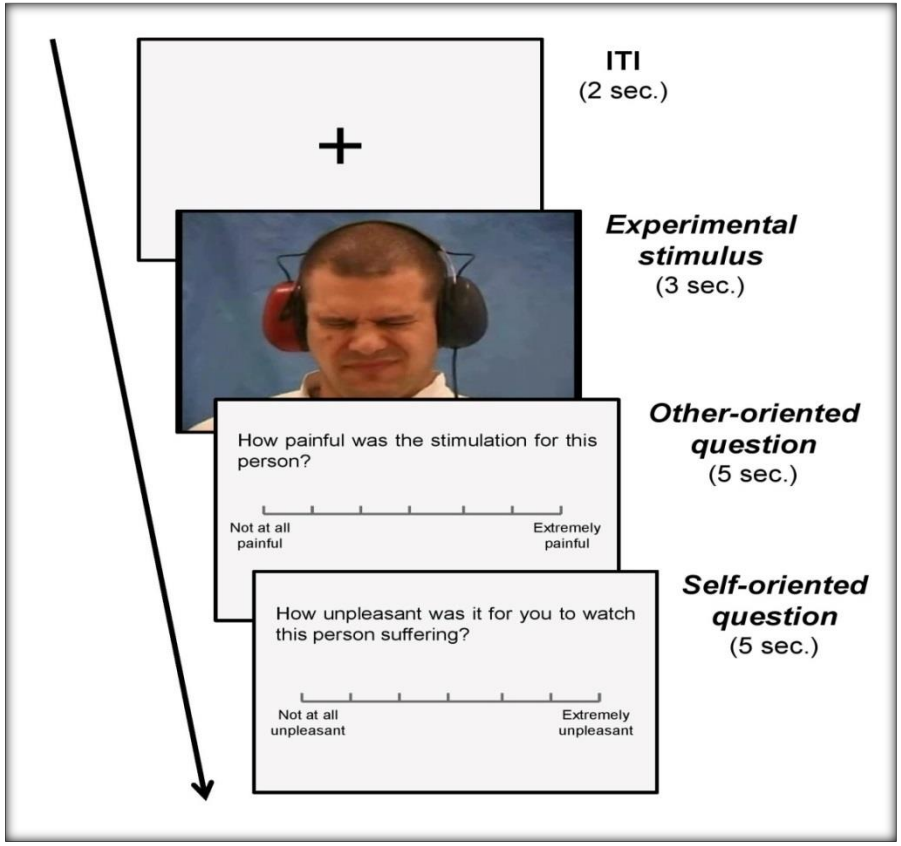
Scenarios were presented in the scanner using a visual display presented on an LCD panel and back-projected onto a screen positioned at the front of the magnet bore. Subjects were positioned supine in the scanner so as to be able to view the projector display using a

mirror above their eyes. The behavioral data were collected using a Lumina response box (LP-400, Cedrus Corporation, San Pedro, USA). The stimuli were presented using Cogent 2000 (Wellcome Department of Imaging Neuroscience, http://www.fil.ion.ucl.ac.uk/Cogent2000.html) running on MATLAB platform. The text of the stories was presented in a black 21-point Arial font on a white background with a resolution of 800 × 600.

In the same session, participants completed both the moral judgment task and the empathy localizer task. The order in which participants performed moral judgment task and empathy localizer task was counterbalanced across participants.

Text S4: Task designs for localizer tasks

Task schematics for the empathy localizer task (Lamm et al., 2007). For more details, see the main text.



Text S5: Additional details about fMRI preprocessing and data visualization

Given the huge variation in possible preprocessing pipelines and flexibility in methodological choices (Carp, 2012), we provide extensive details about our choices and rationale for the same here. To report acquisition and preprocessing details, we have followed a prior set of guidelines (Inglis, 2015; Poldrack et al., 2008).

Preprocessing:

Data were analyzed with SPM12 (www.fil.ion.ucl.ac.uk/spm/software/spm12; Wellcome Department of Imaging Neuroscience, London, UK). First three scans were discarded to avoid T1-equilibration effects. The scans were not slice timing corrected because for relatively short TR (2 seconds or less), it can lead to artifacts (Poldrack, Mumford, & Nichols, 2011, p.42, 48). All functional volumes were realigned in two steps: initially to the first volume and then to the mean realigned image. The estimation of realignment parameters was carried out using a 6-parameter affine (rigid body) transformation such that the cost function comprising of difference in voxel intensities between images was minimized. The voxel intensities from old images were then resampled using higher-order interpolation (B-spline basis functions) to create new motion-corrected voxel intensities in resliced images. The average of the motion-corrected images was co-registered to each individual's structural MRI scan using a 9-parameter affine transformation such that a suitable between-modality cost function (normalized mutual information) was minimized. The realigned functional images were then normalized to the ICBM-space template (2 mm × 2 mm × 2 mm voxels) for European brains by applying nonlinear deformation field estimated from the best overlay of the atlas image on the individual subjects' co-registered structural image. The normalized images

were then smoothed by convolving an isotropic Gaussian kernel with full width at half maximum (FWHM) of 10 mm (= $\sqrt{6^2 + 8^2}$, 6 mm at first and 8 mm at second level) in order (*i*) boost signal-to-noise ratio to ease the detection of large clusters, (*ii*) overcome imperfections remaining from inter-subject registration, and (*iii*) validate assumptions of Gaussian random field theory (RFT) applied later to correct for multiple comparisons during statistical analysis (Poldrack et al., 2011, pp.50-52).

Motion and artefact analysis:

In order to avoid false positive activations owing to head movement, the following data quality checks were employed for each participant and for each task. Data from a participant for a particular task was removed without further analysis if TR-to-TR head movement exceeded 5 mm at any point during the task (none removed).

After this check, the artefact detection analysis was carried out using the Art toolbox (www.nitrc.org/projects/artifact_detect). For each task, outlier scans were identified based on two measures (cf. Koster-Hale, Bedny, & Saxe, 2014): (*a*) if the TR-to-TR composite motion was more than 2mm and/or (*b*) if the scan-to-scan global BOLD signal normalized to *z*-scores deviated from mean more than *z* = 3. Each time-point identified as an outlier was regressed out as a separate nuisance covariate in the first-level design matrix. Note that the motion outliers were identified based on composite motion parameter as this is a more comprehensive measure that outperforms individual motion parameters (Wilke, 2014). Any participant with more than 20% outliers scans were excluded from the analysis. Scanning data for the moral judgment task was discarded for two participants due to excessive head motion (outlier scans > 20%), but their behavioral data was retained.

Using the Art toolbox, we also ensured that there were no systematic correlations between any of the task-related parameters, realignment parameters, and global BOLD-signal, which can lead to artifactual activation or loss of task-related signal after removing motion-related signal (Poldrack et al., 2011, p.44). Since we regressed out scans with excessive movement and the task regressors were not correlated with BOLD activity, we did not unwarp the realigned images to remove variance associated with susceptibility-by-movement interactions (B0 distortions).

In the table below, we tabulate percentage of outlier scans from motion and artifact analysis:

Note: Green cells represent missing data. Yellow cells represent discarded data - ID14 was on medication (for more, see Methods and Materials section in the main text). Red cells denote outlier data.

ID	Percent of outlier scans (glabal intensity: z > 3 and motion: > 2mm)	
	Empathy	Intent
1	0.90%	3%
2	0%	1.90%
3	0%	2.40%
4	3.70%	6.60%
5	0%	0.20%
6	0%	3.50%
7	5.00%	
8	1.20%	3.40%
9	1.20%	0.50%
10	1.20%	1.20%
11	0%	0.70%
12	1.20%	0.90%
13	6.80%	3.90%
14	2.50%	5.40%
15	0%	2.30%
16	0%	7.50%
17	2.50%	21.30%
18	0%	4.20%
19	3.70%	5.00%
20	1.90%	7.10%
21	10.50%	25%
22	4.30%	2%
23	5.60%	15.10%
24	6.80%	8.30%
25	0%	0.80%
26	5.60%	5.50%
27	1.20%	0.60%
28	3.70%	0.50%
29	6.80%	7.80%
30	8.60%	9.90%
31	0%	1.30%
32	6.90%	4.40%
33	8.00%	4.70%
34	8.60%	6.70%
35	4.30%	3.30%
36	9.90%	15.70%
37	11.10%	2.90%
38	0%	3.00%
39	11.10%	10.90%
40	3.70%	0.80%
41	1.20%	8.30%
42	3.70%	0.80%
43	2.50%	1.80%
44	0%	0.90%
45	5.60%	4.20%
46	0%	3.20%
47	11.70%	12.50%
48	1.20%	2.20%
49	3.70%	0.90%
50	6.80%	5.08%

Data reporting and visualization:

Combination of the Anatomy Toolbox v2.1 (Eickhoff et al., 2005) and Neuromorphometrics atlas was used for anatomical interpretation. All peaks of activations reported are in MNI-coordinates but no Brodmann Area (BA) labels have been reported as assigning functional activations to cytoarchitectotonically defined BAs can be inaccurate in the absence of probabilistic maps of underlying cytoarchitectotonic variability (Devlin & Poldrack, 2007). All statistical parametric maps are displayed on smoothed, representative scans (average of 305 T1 images, provided in SPM12) and not on a single brain as this can deceive the reader into thinking that anatomical localization is more precise than is actually possible (Ridgway et al., 2008).

Text S6: Additional details for fMRI data analysis

First-level analysis:

For each participant and for each task, the design matrices for a fixed-effects General Linear Model were constructed by convolving a canonical hemodynamic response function or HRF (double gamma, including a positive γ function and a smaller, negative γ function to reflect the BOLD undershoot) with the stimulus function for events (boxcar function) to create regressors of interest. Even a minor misspecification in hemodynamic model can lead to biased estimators and loss of power, possibly inflating the type I error rate (Lindquist, Meng Loh, Atlas, & Wager, 2009). Thus, in order to account for subject-to-subject and voxel-to-voxel variation in evoked BOLD response, the stimulus function was also convolved with partial derivative of canonical HRF with respect to onset latency

(which allows for delay in peak response) and dispersion (which relaxes assumption about width of the response) to form the informed basis set (Henson, Rugg, & Friston, 2001).

Note that the inclusion of temporal derivative of HRF also reduces impact of slice timing differences by allowing some degree of timing misspecification, which is crucial for our study since we did not do slice timing correction (Ashby, 2011, pp.47-51). The convolution was performed in a higher resolution time-domain than TR (16 time-bins per TR). As a default, SPM orthogonalizes HRF derivatives on canonical HRF and not on the rest of the design matrix (Pernet, 2014). The orthogonality of other regressors of interest was also visually inspected in design matrices since collinearity between regressors can lead to highly unstable parameter estimates and loss of statistical power (Mumford, Poline, & Poldrack, 2015). High-pass temporal filtering with a cut-off of 128s was used to remove low-frequency drifts and power spectra were visually inspected to ascertain that signals of interest were not being filtered out. Temporal autocorrelations in fMRI time series data were modelled using an autoregressive AR(1) model. Since in the current study neither the ITI was less than 1 second nor was the stimulus exposure duration less than 3 seconds, we were confident that the BOLD-response did not exhibit significant nonlinearities and thus a second-order Volterra series was not modelled in the design matrix for any of the tasks (Ashby, 2011, pp.33-34).

Second-level analysis:

Heterogeneity of variance between different levels of factors and non-sphericity in the data was accounted for by estimating parameters using Weighted Least Squares (pre-whitening the data using estimated non-sphericity and then applying Ordinary Least

Squares; SPM12 Manual, pp.277-78). Not assuming sphericity was especially important for our design since we included informed basis set at first-level that leads to stronger assumption about sphericity (Glaser & Friston, 2004), although only canonical HRF contrasts were retained for the whole-brain analysis for the moral judgment task because of the complexity of design (cf. http://imaging.mrc-cbu.cam.ac.uk/imaging/DealingWithDifference).

Whole-brain analyses were thresholded at $p < 0.05$, Family-wise Error (FWE) corrected at the threshold level (primary threshold: $p < 0.001$, extent threshold: $k > 10$). The cluster-level inference has greater overall sensitivity over more stringent voxel-level inference, but the primary limitation of the former approach is that one can only claim that there is true signal *somewhere* in the large clusters (which can span many anatomical regions) that are found and thus is ill-suited to investigate question about overlapping or distinct activations across conditions and this limitation should be kept in mind while interpreting the results (Woo, Krishnan, & Wager, 2014).

Overall grand mean scaling was applied to the data, but no global normalization was used as this procedure has been known to introduce bias in the results (Ashby, 2011, p.97). Also, no implicit threshold masking was applied. Activations lying outside of the brain (due to low variance problem) (Ridgway, Litvak, Flandin, Friston, & Penny, 2012) were weeded out using explicit threshold mask formed by averaging first-level masks for respective task from each participant.

Text S6: Additional details for ROI analysis

To carry out ROI analysis, we investigated functional specificity (and *not* specialization) (Friston, Rotshtein, Geng, Sterzer, & Henson, 2006) of empathy network for each individual participant using functional localizer task. We note that the ROIs were not tailored to be the same for all participants and were determined on an individual basis for both tasks following individual-subjects functional localization approach (Fedorenko & Kanwisher, 2011).

The data from spherical ROIs with a radius of 8mm was extracted and analyzed using the MarsBar toolbox (v0.44) for SPM (http://marsbar.sourceforge.net/) (Brett et al., 2002). The GLM model set to time-series of summary statistic (mean signal) from each ROI was similar to that in the whole-brain analysis except that autocorrelations in the time-series were modelled using fmristat AR(2) (http://www.math.mcgill.ca/keith/fmristat/) processes instead of AR(1) processes, since second order autoregressive model is the most parsimonious way to model signal due to aliased physiological artefacts (Lindquist, 2008). Note that since ROI analysis was carried out at the first-level, the smoothing kernel applied to data was 6 mm and not 10 mm.

Within the ROI, the average percent signal change (PSC) was computed relative to the adjusted mean of the time series. Quality check was performed by reviewing if any of the PSC values were extreme (> 5%) as these can be indicative of artefacts in the data (Mazaika, 2009: http://cibsr.stanford.edu/documents/FMRIPercentSignalChange.pdf; Raichle & Mintun, 2006) and, when found, data from that particular ROI was excluded.

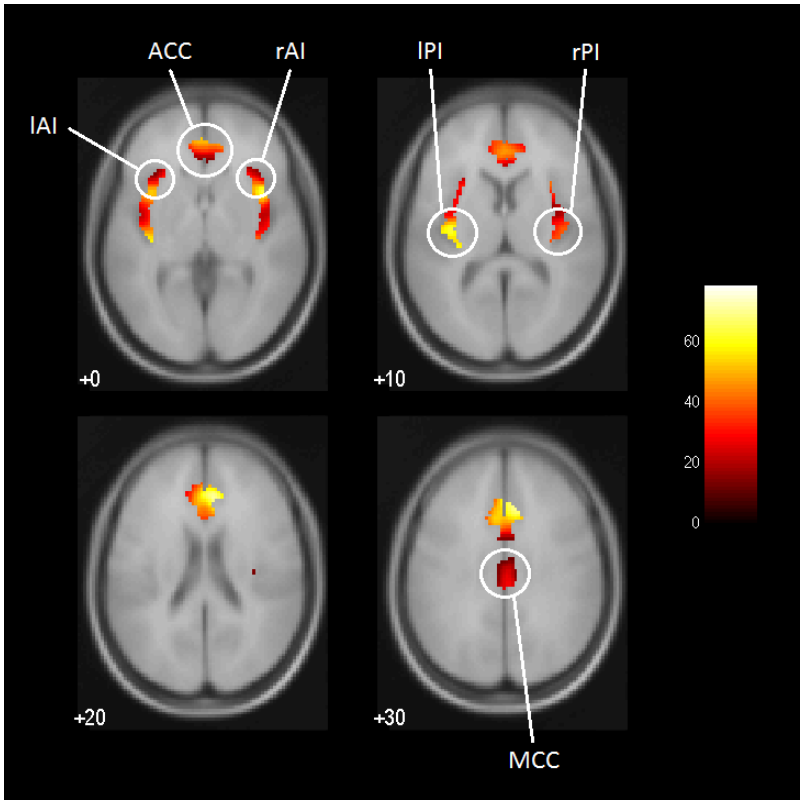
The computation of PSC was based on GLM parameter estimates for canonical HRF and its derivatives, which provides a better estimate of the true PSC (Pernet, 2014). As recommended (Poldrack, 2007), data defining ROIs was independent from the data used in the repeated measures statistics. Restricting analysis to a few ROIs thus reduced Type-I error by drastically limiting the number of statistical tests performed (Saxe et al., 2006).

Note that although ToM can also be expected to be recruited while thinking about others' emotional experiences as well, a prior study shows that physical pain recruits primarily the empathy network (AI-aMCC) while processing of emotional suffering (without physical pain) recruits regions overlapping with ToM network (Bruneau, Dufour, & Saxe, 2013). As such, none of the ToM regions were included in our list of ROIs for empathy for pain.

Text S7: Empathy ROIs localized using functional localizer task at whole brain level

Brain regions where the BOLD signal was higher while watching painful videos as compared to baseline (n = 49, random effects analysis, $p < 0.05$, FEW-corrected, $k > 10$), masked with Neuromorphometrics anatomical atlas in random effects analysis carried out at the second-level. The coordinates for peak activations are reported in Table 1 in the main text.

Note: The slice numbers represent *z*-coordinate in MNI stereotactic space and the color bar denotes the *F*-statistic.



Text S8: ROI coordinates at individual level from empathy localizer

Note: "-" not localized

ID	dACC	aMCC	L-AI	R-AI	L-PI	R-PI
1	[-2;42;-6]	[-6;10;34]	[-30;18;4]	[34;10;8]	[-38;-18;14]	[38;-6;2]
2	[-4;40;24]	[4;20;36]	[-36;22;0]	[44;12;-6]	[-42;-8;-8]	[42;-6;-10]
3	[4;30;30]	[6;8;44]	[-36;16;2]	[40;14;-4]	-	-
4	[12;30;24]	[10;24;30]	[-34;20;4]	[38;22;2]	[-32;-26;14]	[38;-20;14]
5	[4;26;34]	[-2;6;46]	[-32;14;8]	[42;14;-2]	[-38;-14;8]	[38;-6;0]
6	[-2;24;28]	[12;-26;38]	[-32;16;6]	[34;26;-4]	[-40;-10;10]	[42;-14;2]
7	[-8;36;22]	[-8;14;38]	[-30;20;4]	[40;16;0]	[-36;-16;14]	[40;0;-16]
8	[2;32;22]	[10;22;32]	[-34;18;-2]	[34;16;-4]	[-36;-20;10]	-
9	[-2;34;28]	[10;16;34]	[-32;18;8]	[32;28;-2]	-	[40;-16;12]
10	[6;42;-4]	[4;-2;48]	-	-	-	-
11	[4;36;26]	[12;22;28]	[-32;26;2]	[34;24;-8]	[-34;-20;2]	-
12	[-4;48;8]	[6;6;38]	[-32;14;4]	[36;26;-2]	[-38;-16;16]	[36;-14;16]
13	[2;46;0]	[10;6;38]	-	[32;28;4]	[-36;-16;10]	[36;-18;14]
14	[8;26;26]	[-2;6;46]	[-34;24;-2]	[44;14;-4]	[-36;-12;14]	-
15	[0;40;0]	[8;22;32]	-	[40;2;8]	-	-
16	[2;26;32]	[2;16;38]	[-40;10;-4]	[44;16;-6]	-	[42;-10;8]
17	[10;32;24]	[8;16;34]	[-44;14;-4]	[34;22;0]	[-38;-10;8]	[36;-10;16]
18	[4;26;34]	[2;18;38]	[-32;16;2]	[40;14;-2]	[-34;-18;12]	[44;-10;6]
19	[12;40;8]	[8;-26;44]	-	[32;18;6]	[-38;-20;12]	[36;-12;12]
20	[-4;34;20]	[4;2;42]	[-42;16;-4]	[40;10;2]	[-42;-16;10]	-
21	-	-	[-36;-2;10]	-	-	-
22	[4;36;26]	[8;20;32]	[-38;6;4]	[38;16;-4]	[-38;-14;16]	[34;-22;12]
23	[0;32;26]	[-6;20;34]	[-42;16;-4]	[32;28;-2]	-	-
24	-	[10;-10;40]	-	-	-	-
25	[4;26;34]	[-2;18;38]	[-36;14;2]	-	-	-
26	-	-	-	[42;8;0]	-	[38;-16;12]
27	[-10;30;24]	[8;10;40]	[-32;20;8]	[32;20;8]	[-36;-16;14]	[38;-16;16]
28	[6;32;26]	[4;10;42]	-	[42;16;-2]	-	-
29	[2;38;22]	[-2;-6;40]	[-30;24;6]	[32;18;6]	[-36;-18;14]	-
30	[2;28;30]	[-10;0;40]	[-32;16;4]	[34;18;6]	[-40;-14;8]	[36;-12;12]
31	-	-	-	-	-	-
32	[10;26;28]	[6;-2;48]	[-40;8;2]	[34;22;6]	[-38;-14;16]	[36;6;-20]
33	[-2;28;32]	[-8;16;32]	[-42;6;0]	[44;2;-2]	[-34;-18;12]	[44;-6;-6]
34	[2;24;28]	[2;22;28]	[-42;0;6]	[32;16;6]	[-42;-14;10]	[40;-6;-4]
35	[2;34;26]	[-2;0;46]	[-44;12;-4]	[34;22;6]	[-36;-22;14]	-
36	[-2;30;20]	-	[-38;10;0]	[34;16;6]	-	-
37	[4;24;32]	[2;22;34]	[-34;10;6]	-	-	-
38	[2;26;22]	[-6;20;30]	[-34;16;4]	[34;14;8]	[-42;-12;-2]	-
39	-	[2;-2;48]	-	[40;14;0]	[-42;-16;10]	-
40	-	[4;-2;44]	-	[32;20;8]	-	[36;6;-20]
41	-	-	[-40;-8;8]	[40;-2;0]	[-40;-10;6]	[40;-4;0]
42	[-2;12;42]	[0;-22;42]	[-30;18;2]	[42;4;2]	[-34;-20;2]	[36;-18;-2]
43	[4;28;32]	[2;-24;38]	[-40;12;-2]	[32;26;8]	-	[42;-10;8]
44	[-6;24;32]	[0;-18;38]	[-44;12;-8]	[36;18;4]	-	[36;-18;12]
45	[-6;38;26]	[0;-24;38]	[-38;18;-10]	[44;18;-8]	[-34;-16;18]	-
46	[2;46;8]	[6;-16;42]	[-42;2;0]	[44;8;-2]	[-32;-24;8]	[40;-12;4]
47	[12;36;20]	[8;-14;44]	[-38;8;2]	[36;10;8]	[-44;-10;6]	-
48	-	[-6;20;30]	[-42;10;-2]	[42;2;8]	-	[38;-12;12]
49	[-4;44;10]	[4;18;36]	[-32;12;8]	[32;18;4]	[-38;-14;14]	-
50	[12;32;22]	[4;-18;36]	[-32;12;10]	[36;10;6]	[-34;-16;12]	[36;-12;12]
Avg.	[2;32;23]	[3;5;38]	[-36;13;2]	[37;16;1]	[-37;-16;-10]	[38;-10;4]

Text S9: Additional details for brain-behavior correlation analysis

To avoid false positive brain-behavior correlations, we followed recommended steps (Rousselet & Pernet, 2012; Schwarzkopf et al., 2012): (*i*) In order to avoid undue influence of univariate outliers on the overall results (false positive correlation or masking), Spearman's rank correlation coefficient was preferred over Pearson's *r*. (*ii*) Significant correlations ($p < 0.05$) found from this analysis were further investigated using Robust correlation toolbox (Pernet et al., 2013), since Spearman's *rho* is not robust to multivariate normality violation or bivariate outliers or heteroscedasticity. Skipped Spearman ($\rho_{skipped}$) correlations were used as robust correlations (standard Spearman correlation on data cleaned up for bivariate outliers). (*iii*) All significant Skipped correlations ($p < 0.05$) were reported with robust confidence intervals computed by bootstrapping (1000 resamples) the cleaned data to emphasize their likely unreliability. (*iv*) If the nominal confidence interval differed from the bootstrapped confidence interval for significant correlations, Shepherd's *pi* correlation (Spearman's *rho* after removing potential bivariate outliers identified through the bootstrapped Mahalanobis distance (10,000 resamples) and adjusting the *p*-value; Schwarzkopf et al., 2012) was used as an additional robust test. As 24 correlation tests were run for every ROI, the Bonferroni-corrected threshold for statistical significance would have resulted in a very stringent threshold (0.05/24 = 0.0020) increasing the risk of false negatives (Lieberman & Cunningham, 2009) and thus we did not use such stringent threshold (as recommended by Rousselet & Pernet, 2012).

Text S10: Psychophysiological Interaction analysis details

Given our *a priori* hypothesis on the role of empathy on blame judgment, time series were extracted from a seed voxel in the r-AI (at coordinates given by the localizer task) that showed an increase in BOLD signal during blame (versus acceptability) judgments for accidental harm cases at an uncorrected threshold of $p < 0.99$ within 8 mm of this voxel, for each subject individually. Note that such liberal threshold was chosen to ensure all voxels in the ROI are used to compute the connectivity (McLaren, Ries, Xu, & Johnson, 2012). The time series from seed region was summarized by the first eigenvariate across all suprathreshold voxels. The resulting time series were adjusted for effects of no interest by demeaning the eigenvector by all effects not included in that contrast. This BOLD time series was deconvolved to estimate a neuronal time series for this region using the PPI-deconvolution parameter defaults in SPM12 (Gitelman, Penny, Ashburner, & Friston, 2003). The PPI regressor was calculated as the element-by-element product of the ROI neuronal time series and a vector coding for the main effect of task (contrast vectors: *blame* = 1, *acceptability* = -1). This product was then re-convolved with the canonical HRF.

At first-level analysis, we included the PPI as a regressor of interest in a GLM. The task vectors and the extracted time series were modelled as additional regressors, in order to assess the PPI estimates over and above shared functional activation and task-independent correlations in BOLD signal between the seed and other regions (O'Reilly et al., 2012). These regressors were convolved with a canonical HRF and high-pass filtered (128 s). Since functional connectivity results have been shown to be severely affected by movement artifacts (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012), we also included realignment parameters and regressors for

outlier in the first-level PPI models. Subject-wise PPI models were run, and contrast images were generated. Regions with positive or negative PPI denote region with greater or lesser context-dependent change in connectivity with the seed region. These subject-wise contrast images were then entered into second-level GLM analyses to generate *t*-maps on which statistical inference was carried out using uncorrected threshold of $p < 0.001$, $k > 10$. We did not choose FWE-correction for this analysis because PPI analyses tend to lack power (O'Reilly et al., 2012) and thus wanted to avoid greater risk of false negatives (Lieberman & Cunningham, 2009).

Text S11: Descriptive statistics for moral judgments

Type of judgment	Condition	Mean	SD	Min	Max
acceptability	neutral	2.18	0.98	1.00	5.57
	accidental	4.14	1.21	2.00	6.50
	attempted	5.25	1.03	1.67	7.00
	intentional	6.35	0.62	4.44	7.00
blame	neutral	2.05	0.88	1.00	4.67
	accidental	4.09	1.19	1.22	6.44
	attempted	5.13	1.10	1.67	6.78
	intentional	6.33	0.75	3.40	7.00

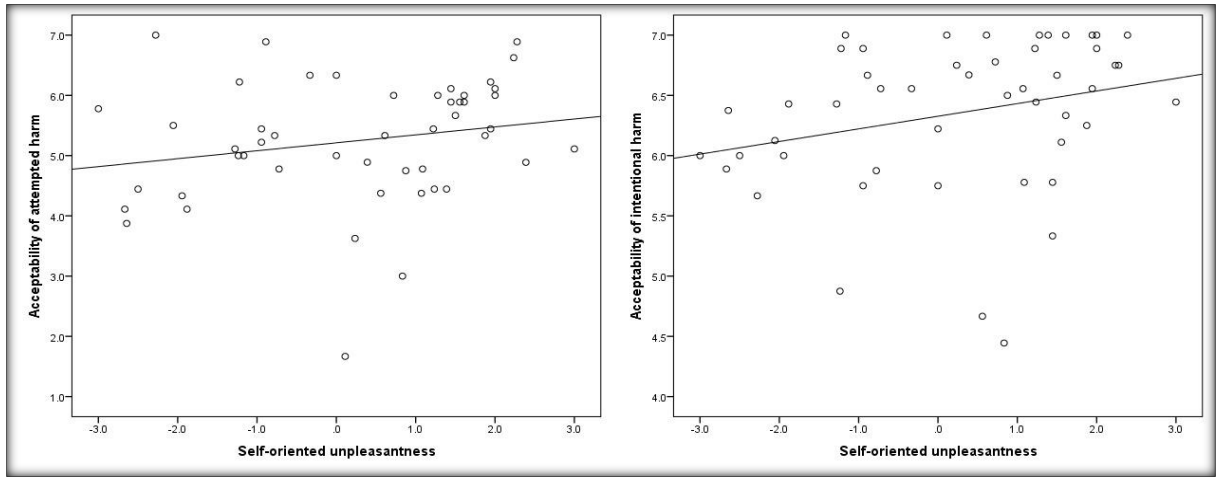
Text S12: Correlations between empathy (performance measures scores) and behavioral ratings with robustness check

(significant results denoted by yellows cells)

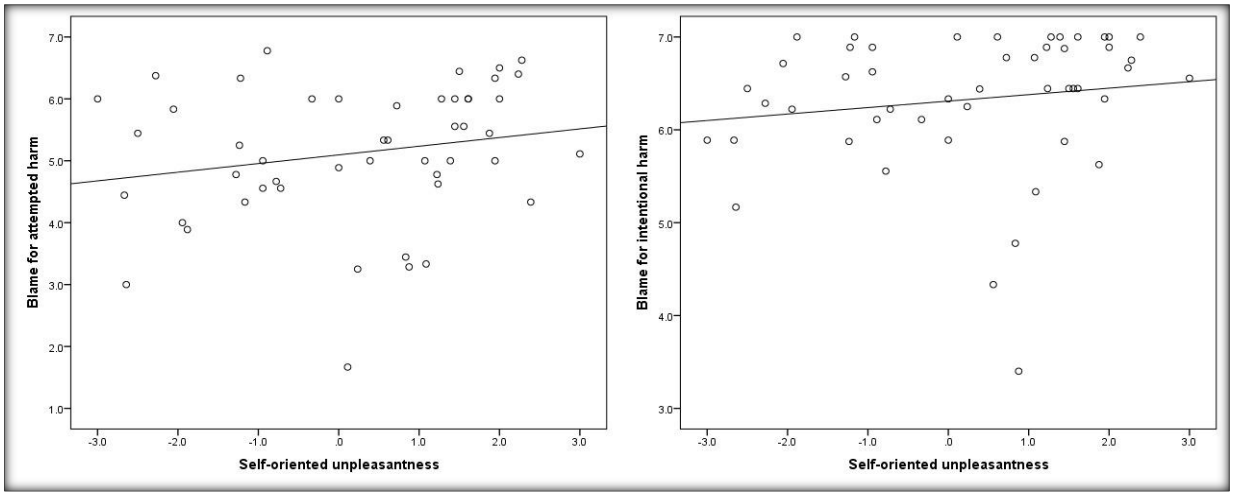
Correlating the average of ratings provided by participants for the other- and self-oriented empathy questions of the localizer task revealed that only the self-oriented unpleasantness while watching videos of others receiving painful stimuli were predictive of intent-based moral judgments. In particular, the more unpleasant was the subjective experience, the more severe was the condemnation (less acceptable, more blame) for attempted and intentional harms (scatterplots are provided below). This raises the question as to why the self-oriented distress was not predictive of condemnation for accidental harm scenarios, similar to the pattern detected in brain-behavior correlations (see Results in the main text). One possibility is that the differences between the experimental contexts may result in divergent results (cf. Decety, 2011). We would like to note that it is not unusual to find such inconsistent results between state and trait measures and neural activity (for a review of such discrepancies, see Supplementary Information in Lamm, Decety, & Singer, 2011).

Type of judgment	condition	statistic	Other-oriented estimation (*n* = 49)	Self-oriented unpleasantness (*n* = 49)	Robustness check (Spearman's skipped correlation)
acceptability	neutral case	ρ	0.128	0.119	-
		p	0.380	0.416	
	accidental harm	ρ	0.238	0.038	-
		p	0.099	0.795	
	attempted harm	ρ	0.042	**0.300**	r = 0.375, 95% CI [0.104, 0.645]
		p	0.773	**0.036**	
	intentional harm	ρ	0.165	**0.379**	r = 0.431, 95% CI [0.157, 0.632]
		p	0.257	**0.007**	
blame	neutral case	ρ	0.255	**0.319**	r = 0.281, 95% CI [-0.013, 0.526]
		p	0.077	**0.025**	
	accidental harm	ρ	0.259	0.189	-
		p	0.072	0.194	
	attempted harm	ρ	0.026	**0.302**	r = 0.351, 95% CI [0.043, 0.613]
		p	0.859	**0.035**	
	intentional harm	ρ	0.293	**0.293**	r = 0.333, 95% CI [0.0793, 0.576]
		p	0.041	**0.041**	

- Scatterplot with acceptability judgments for attempted and intentional harm cases and self-oriented unpleasantness ratings



- Scatterplot with blame judgments for attempted and intentional harm cases and self-oriented unpleasantness ratings

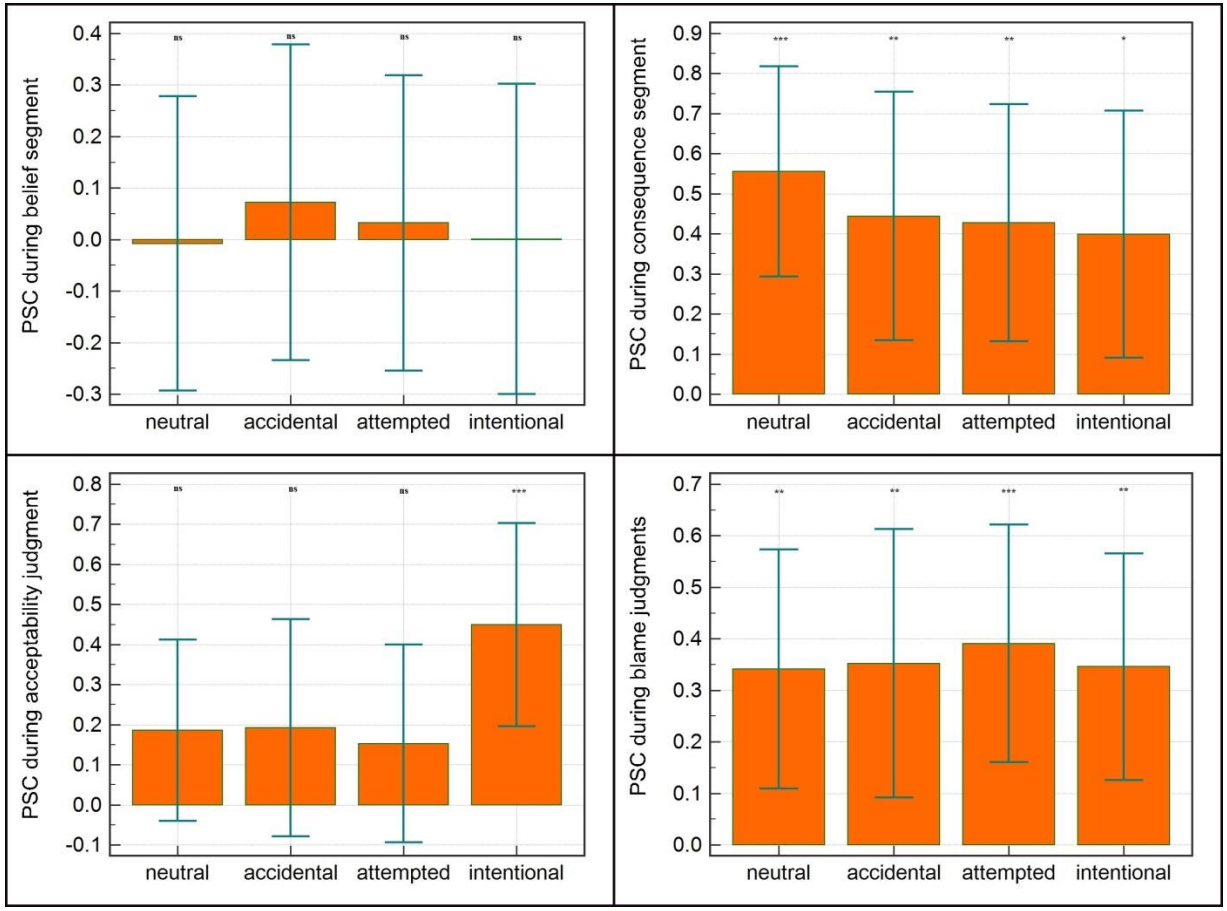


Text S13: One-sample *t*-tests on ROI data for each segment and for each condition

For all figures in this section, the error bars represent 95% confidence interval.

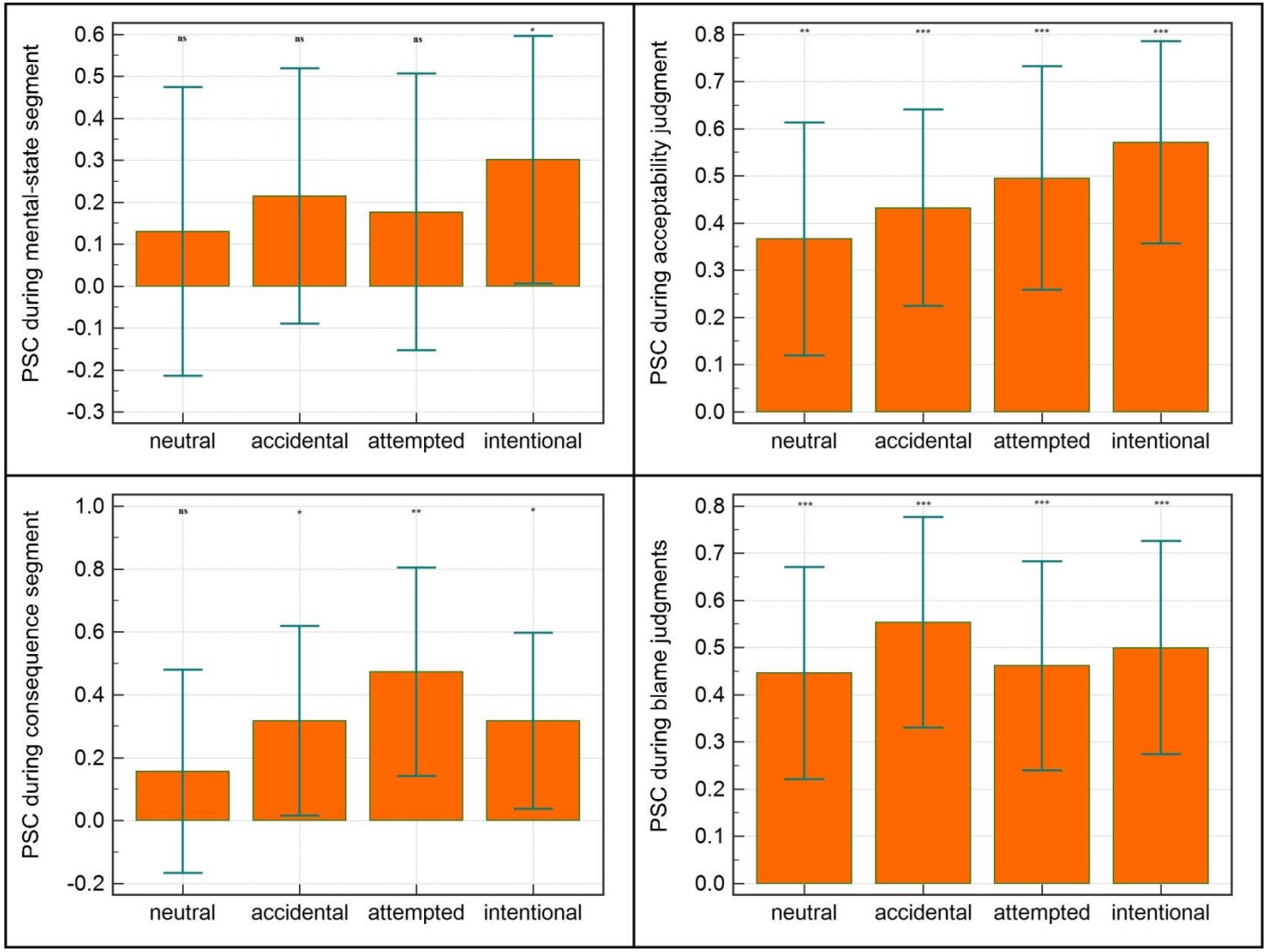
- **dACC:**

Segment	Condition	M	SD	*t*(38)	*p*
mental-state information	neutral	-0.008	0.882	-0.055	0.957
	accidental	0.072	0.945	0.479	0.635
	attempted	0.032	0.886	0.227	0.822
	intentional	0.001	0.929	0.007	0.994
consequence	neutral	0.555	0.809	4.284	< 0.001
	accidental	0.444	0.957	2.899	0.006
	attempted	0.428	0.913	2.923	0.006
	intentional	0.399	0.950	2.622	0.012
acceptability	neutral	0.186	0.698	1.664	0.104
	accidental	0.192	0.835	1.437	0.159
	attempted	0.153	0.761	1.255	0.217
	intentional	0.449	0.781	3.591	0.001
blame	neutral	0.342	0.715	2.982	0.005
	accidental	0.352	0.804	2.736	0.009
	attempted	0.391	0.712	3.429	0.001
	intentional	0.346	0.680	3.178	0.003


PSC during belief segment
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
ns
neutral
accidental
attempted
intentional
PSC during consequence segment
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
***
**
**
*
neutral
accidental
attempted
intentional
PSC during acceptability judgment
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
ns
ns
ns
***
neutral
accidental
attempted
intentional
PSC during blame judgments
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
**
**
***
**
neutral
accidental
attempted
intentional

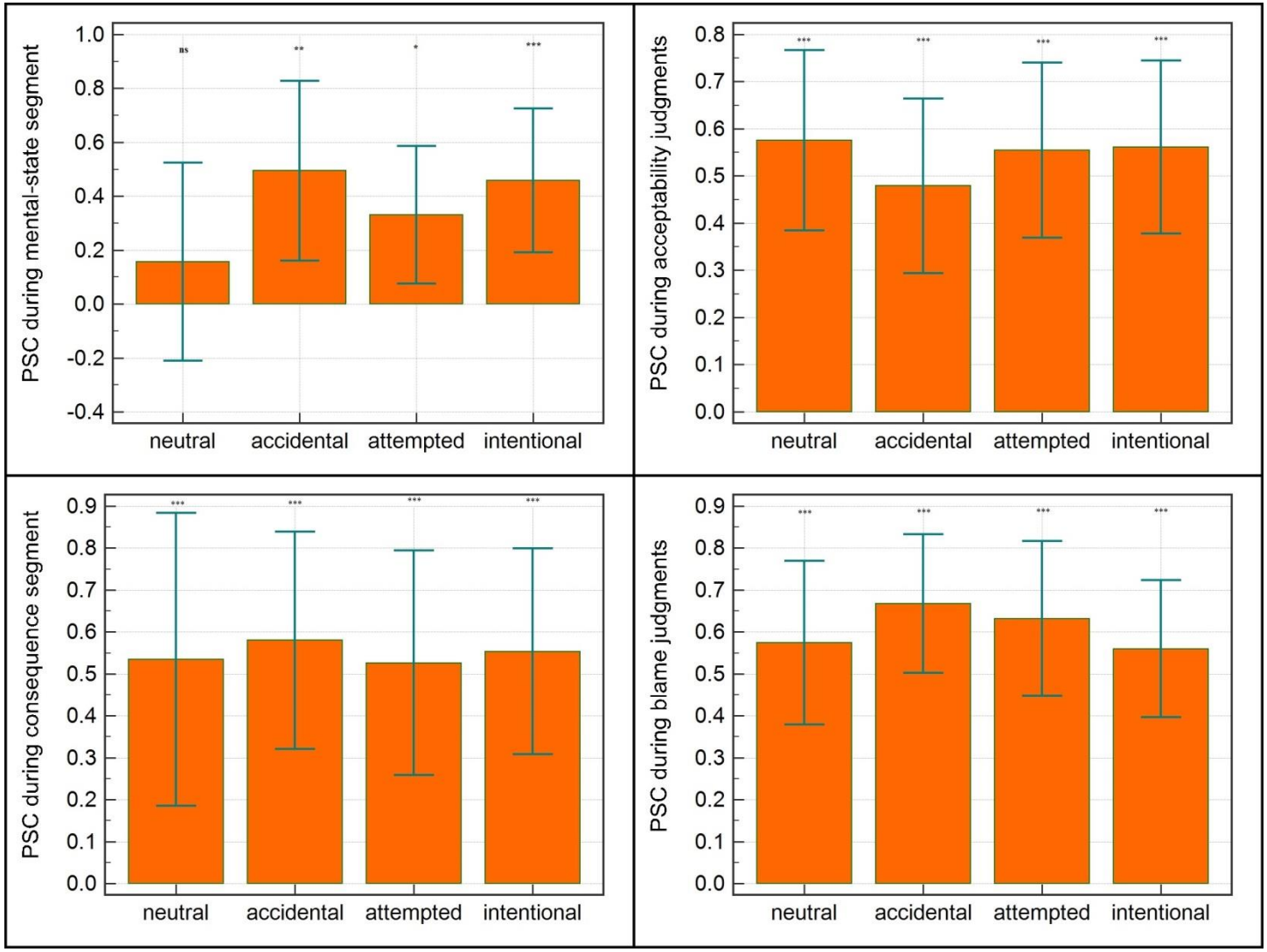
- **l-AI**

Segment	Condition	M	SD	*t*(35)	*p*
mental-state information	neutral	0.130	1.018	0.768	0.448
	accidental	0.215	0.900	1.431	0.161
	attempted	0.176	0.975	1.084	0.286
	intentional	0.301	0.872	2.071	0.046
consequence	neutral	0.156	0.956	0.979	0.334
	accidental	0.317	0.893	2.132	0.040
	attempted	0.473	0.979	2.897	0.006
	intentional	0.317	0.826	2.301	0.027
acceptability	neutral	0.367	0.730	3.013	0.005
	accidental	0.433	0.615	4.222	< 0.001
	attempted	0.495	0.702	4.237	< 0.001
	intentional	0.571	0.633	5.415	< 0.001
blame	neutral	0.446	0.665	4.023	< 0.001
	accidental	0.553	0.659	5.040	< 0.001
	attempted	0.461	0.654	4.234	< 0.001
	intentional	0.500	0.668	4.489	< 0.001


PSC during mental-state segment
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
ns
ns
ns
*
neutral
accidental
attempted
intentional
PSC during acceptability judgment
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
**
***
***
***
neutral
accidental
attempted
intentional
PSC during consequence segment
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
ns
*
**
*
neutral
accidental
attempted
intentional
PSC during blame judgments
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
***
***
***
***
neutral
accidental
attempted
intentional

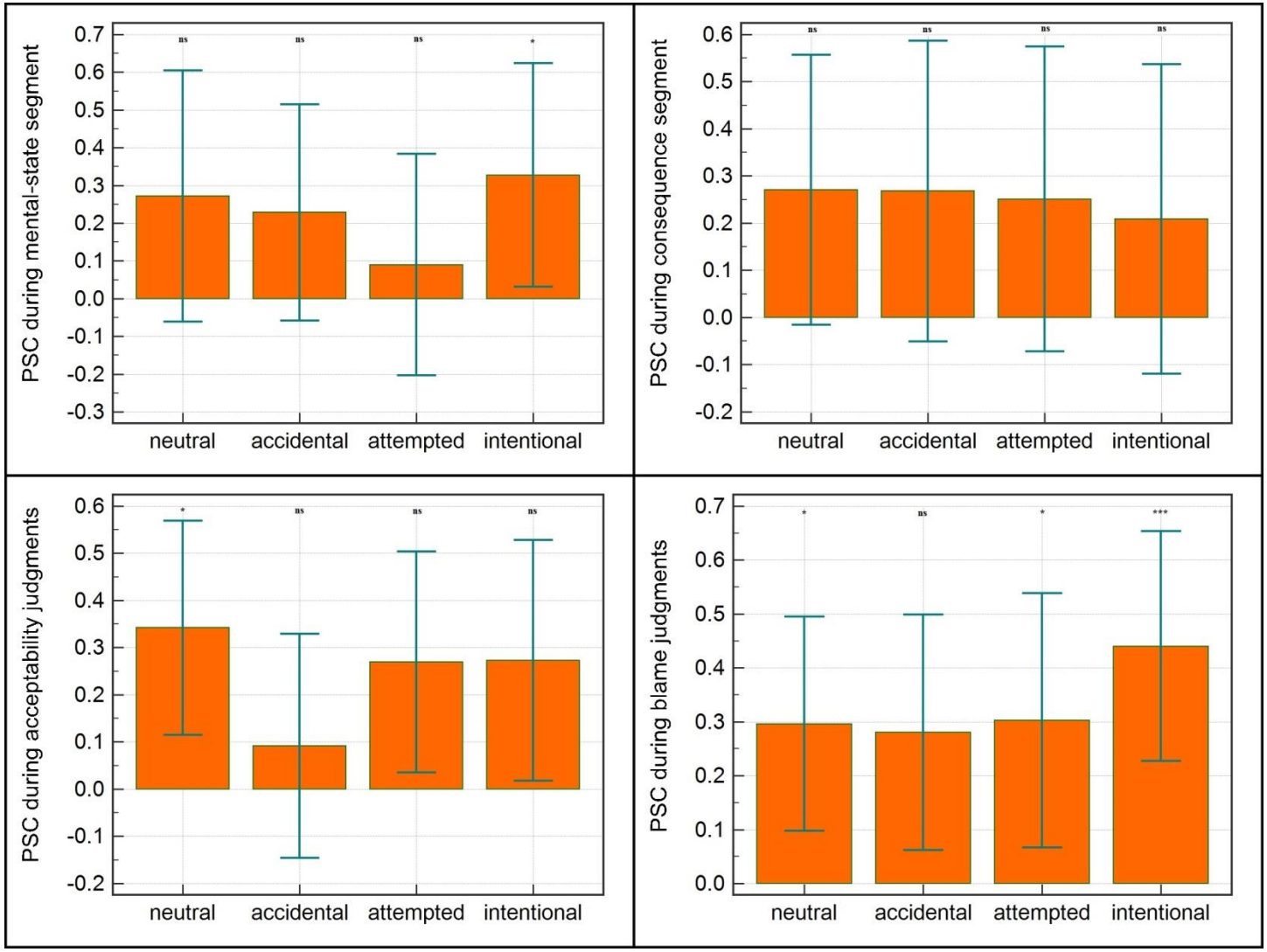
- **r-AI**

Segment	Condition	M	SD	*t*(37)	*p*
mental-state information	neutral	0.157	1.117	0.868	0.391
	accidental	0.495	1.016	3.002	0.005
	attempted	0.330	0.776	2.622	0.013
	intentional	0.459	0.810	3.487	0.001
consequence	neutral	0.535	1.062	3.101	0.004
	accidental	0.580	0.789	4.532	< 0.001
	attempted	0.526	0.816	3.978	< 0.001
	intentional	0.554	0.748	4.565	< 0.001
acceptability	neutral	0.575	0.582	6.100	< 0.001
	accidental	0.479	0.564	5.236	< 0.001
	attempted	0.555	0.565	6.055	< 0.001
	intentional	0.561	0.559	6.192	< 0.001
blame	neutral	0.574	0.594	5.966	< 0.001
	accidental	0.668	0.502	8.197	< 0.001
	attempted	0.632	0.562	6.929	< 0.001
	intentional	0.560	0.498	6.925	< 0.001


PSC during mental-state segment
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
ns
**
*
***
neutral
accidental
attempted
intentional
PSC during acceptability judgments
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
***
***
***
***
neutral
accidental
attempted
intentional
PSC during consequence segment
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
***
***
***
***
neutral
accidental
attempted
intentional
PSC during blame judgments
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
***
***
***
***
neutral
accidental
attempted
intentional

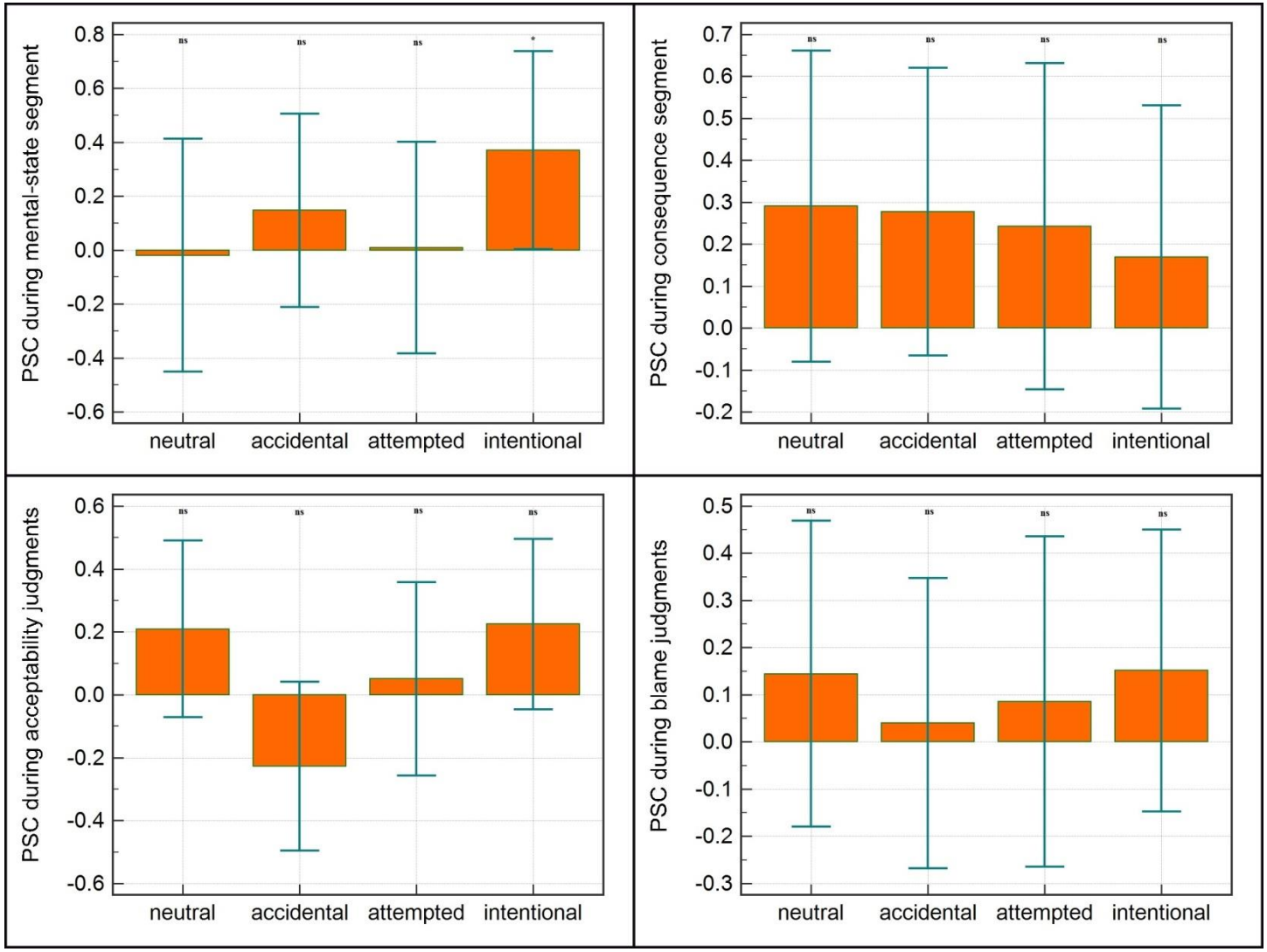
- **l-PI**

Segment	Condition	M	SD	*t*(28)	*p*
mental-state information	neutral	0.207	0.808	1.378	0.179
	accidental	0.169	0.704	1.293	0.207
	attempted	0.064	0.714	0.479	0.635
	intentional	0.323	0.707	2.462	0.020
consequence	neutral	0.157	0.707	1.197	0.242
	accidental	0.262	0.765	1.845	0.076
	attempted	0.188	0.792	1.278	0.212
	intentional	0.185	0.767	1.297	0.205
acceptability	neutral	0.244	0.603	2.175	0.038
	accidental	0.036	0.605	0.321	0.751
	attempted	0.225	0.602	2.012	0.054
	intentional	0.219	0.626	1.887	0.070
blame	neutral	0.245	0.522	2.528	0.017
	accidental	0.189	0.590	1.727	0.095
	attempted	0.254	0.597	2.291	0.030
	intentional	0.356	0.540	3.551	0.001


PSC during mental-state segment
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
ns
ns
ns
*
neutral
accidental
attempted
intentional
PSC during consequence segment
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
ns
ns
ns
ns
neutral
accidental
attempted
intentional
PSC during acceptability judgments
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
*
ns
ns
ns
neutral
accidental
attempted
intentional
PSC during blame judgments
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
*
ns
*
***
neutral
accidental
attempted
intentional

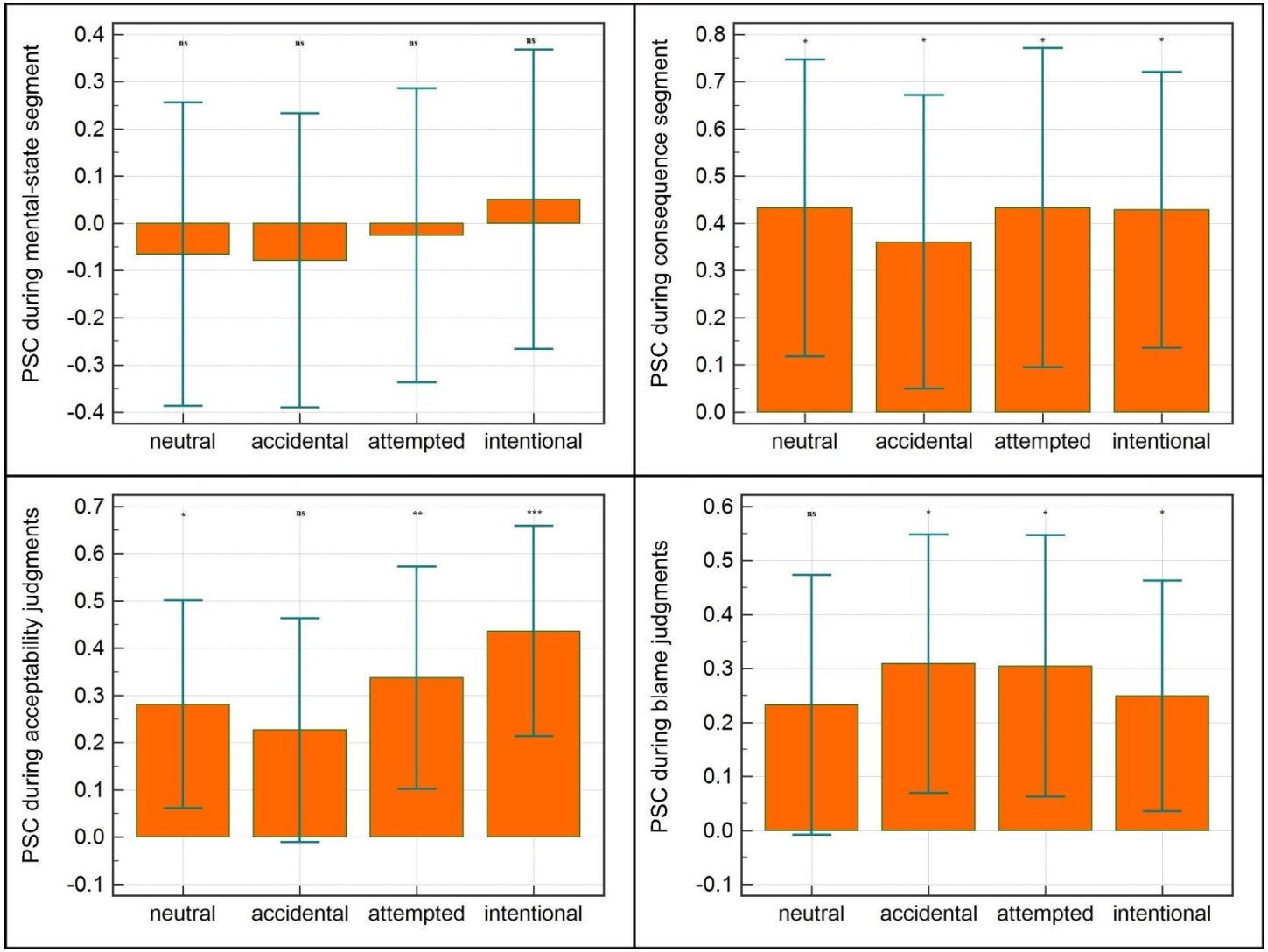
- **r-PI**

Segment	Condition	M	SD	*t*(25)	*p*
mental-state information	neutral	-0.019	1.068	-0.091	0.928
	accidental	0.148	0.888	0.849	0.404
	attempted	0.010	0.972	0.052	0.959
	intentional	0.371	0.909	2.083	0.048
consequence	neutral	0.290	0.918	1.613	0.119
	accidental	0.277	0.850	1.663	0.109
	attempted	0.243	0.964	1.283	0.211
	intentional	0.169	0.897	0.961	0.346
acceptability	neutral	0.209	0.694	1.538	0.137
	accidental	-0.228	0.665	-1.746	0.093
	attempted	0.051	0.760	0.341	0.736
	intentional	0.225	0.672	1.706	0.100
blame	neutral	0.144	0.803	0.916	0.368
	accidental	0.040	0.762	0.269	0.790
	attempted	0.086	0.868	0.504	0.618
	intentional	0.151	0.740	1.043	0.307


PSC during mental-state segment
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
ns
ns
ns
*
neutral
accidental
attempted
intentional
PSC during consequence segment
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
ns
ns
ns
ns
neutral
accidental
attempted
intentional
PSC during acceptability judgments
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
ns
ns
ns
ns
neutral
accidental
attempted
intentional
PSC during blame judgments
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
ns
ns
ns
ns
neutral
accidental
attempted
intentional

- **aMCC**

Segment	Condition	M	SD	*t*(41)	*p*
mental-state information	neutral	-0.065	1.033	-0.408	0.685
	accidental	-0.078	1.000	-0.507	0.615
	attempted	-0.026	0.999	-0.166	0.869
	intentional	0.051	1.017	0.325	0.747
consequence	neutral	0.433	1.008	2.781	0.008
	accidental	0.360	0.998	2.339	0.024
	attempted	0.433	1.085	2.588	0.013
	intentional	0.428	0.939	2.957	0.005
acceptability	neutral	0.281	0.705	2.580	0.014
	accidental	0.227	0.760	1.932	0.060
	attempted	0.338	0.754	2.900	0.006
	intentional	0.436	0.715	3.957	0.001
blame	neutral	0.232	0.772	1.950	0.058
	accidental	0.309	0.767	2.607	0.013
	attempted	0.304	0.778	2.536	0.015
	intentional	0.249	0.686	2.358	0.023


PSC during mental-state segment
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
ns
ns
ns
ns
neutral
accidental
attempted
intentional
PSC during consequence segment
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
*
*
*
*
neutral
accidental
attempted
intentional
PSC during acceptability judgments
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
*
ns
**
***
neutral
accidental
attempted
intentional
PSC during blame judgments
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
ns
*
*
*
neutral
accidental
attempted
intentional

Past studies have revealed activity in the shared empathy network (consisting primarily of AI and aMCC) using pictures of body parts sustaining injuries (Gu & Han, 2007) or visual stimuli of facial expressions in response to such injuries (Lamm et al., 2007) or abstract cues representing administration of pain to another person physically present in the room (Singer et al., 2004), but very few studies have utilized verbal narratives to convey information about physical pain (e.g., Bruneau, Dufour, & Saxe, 2013) and the results presented above additionally demonstrates validity of this modality.

Text S14: ROI analysis across-conditions with empathy ROIs

Results from a 2(belief: neutral, negative) × 2(outcome: neutral, negative) repeated measures ANOVA in each ROI for average percent signal change (PSC) extracted for each text segment in each condition. Yellow cells represent significant ($p < 0.05$) values.

- dACC ($n = 39$)

Text segment	main effect of belief	main effect of outcome	interaction
mental-state (8s)	F (1,38) = 0.022, p = 0.883	F (1,38) = 0.081, p = 0.778	F (1,38) = 0.377, p = 0.543
consequence (8s)	F (1,38) = 1.172, p = 0.286	F (1,38) = 0.507, p = 0.481	F (1,38) = 0.182, p = 0.672
acceptability (6s)	F (1,38) = 2.670, p = 0.110	F (1,38) = 5.769, p = 0.021	F (1,38) = 2.781, p = 0.104
blame (6s)	F (1,38) = 0.102, p = 0.752	F (1,38) = 0.039, p = 0.845	F (1,38) = 0.171, p = 0.681

- l-AI (n = 36)

Text segment	main effect of belief	main effect of outcome	interaction
mental-state (8s)	F (1,35) = 0.551, p = 0.463	F (1,35) = 1.642, p = 0.209	F (1,35) = 0.055, p = 0.816
consequence (8s)	F (1,35) = 5.120, p = 0.030	F (1,35) = 0.001, p = 0.975	F (1,35) = 5.226, p = 0.028
acceptability (6s)	F (1,35) = 3.464, p = 0.071	F (1,35) = 1.449, p = 0.237	F (1,35) = 0.015, p = 0.905
blame (6s)	F (1,35) = 0.103, p = 0.750	F (1,35) = 3.344, p = 0.076	F (1,35) = 0.470, p = 0.497

- l-PI (n = 29)

Text segment	main effect of belief	main effect of outcome	interaction
mental-state (8s)	F (1,28) = 0.003, p = 0.959	F (1,28) = 2.082, p = 0.160	F (1,28) = 1.934, p = 0.175
consequence (8s)	F (1,28) = 0.077, p = 0.784	F (1,28) = 0.383, p = 0.541	F (1,28) = 0.259, p = 0.615
acceptability (6s)	F (1,28) = 1.784, p = 0.192	F (1,28) = 2.329, p = 0.138	F (1,28) = 2.053, p = 0.163
blame (6s)	F (1,28) = 1.248, p = 0.273	F (1,28) = 0.145, p = 0.706	F (1,28) = 1.648, p = 0.210

- r-AI (n = 38)

Text segment	main effect of belief	main effect of outcome	interaction
mental-state (8s)	F (1,37) = 0.448, p = 0.508	F (1,37) = 2.538, p = 0.120	F (1,37) = 0.394, p = 0.534
consequence (8s)	F (1,37) = 0.029, p = 0.866	F (1,37) = 0.265, p = 0.610	F (1,37) = 0.011, p = 0.917
acceptability (6s)	F (1,37) = 0.311, p = 0.580	F (1,37) = 0.759, p = 0.389	F (1,37) = 0.641, p = 0.429
blame (6s)	F (1,37) = 0.252, p = 0.619	F (1,37) = 0.081, p = 0.778	F (1,37) = 2.178, p = 0.148

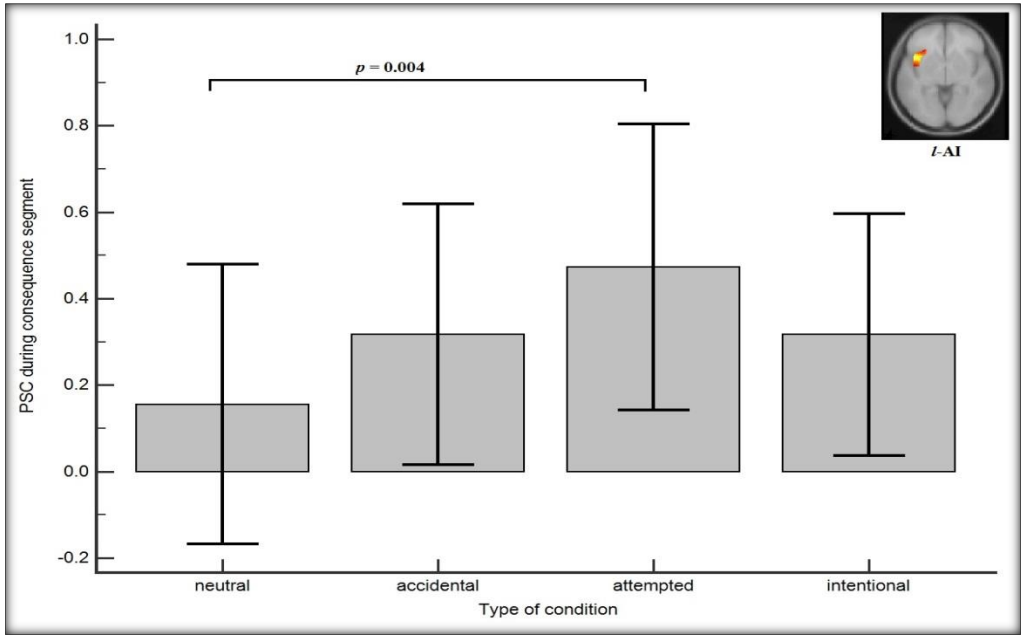
- r-PI (n = 26)

Text segment	main effect of belief	main effect of outcome	interaction
mental-state (8s)	F (1,25) = 0.793, p = 0.382	F (1,25) = 3.782, p = 0.063	F (1,25) = 0.256, p = 0.618
consequence (8s)	F (1,25) = 0.858, p = 0.363	F (1,25) = 0.214, p = 0.648	F (1,25) = 0.112, p = 0.741
acceptability (6s)	F (1,25) = 3.103, p = 0.090	F (1,25) = 4.256, p = 0.050	F (1,25) = 24.820, p < 0.001
blame (6s)	F (1,25) = 0.138, p = 0.714	F (1,25) = 0.129, p = 0.723	F (1,25) = 1.430, p = 0.243

- aMCC (n = 42)

Text segment	main effect of belief	main effect of outcome	interaction
mental-state (8s)	F (1,41) = 0.726, p = 0.399	F (1,41) = 0.154, p = 0.697	F (1,41) = 0.294, p = 0.591
consequence (8s)	F (1,41) = 0.211, p = 0.648	F (1,41) = 0.188, p = 0.667	F (1,41) = 0.152, p = 0.699
acceptability (6s)	F (1,41) = 5.407, p = 0.025	F (1,41) = 0.108, p = 0.744	F (1,41) = 1.187, p = 0.282
blame (6s)	F (1,41) = 0.007, p = 0.933	F (1,41) = 0.026, p = 0.873	F (1,41) = 1.407, p = 0.242

Text S15: Belief-by-interaction effect in l-AI for consequence segment



Text S16: ROI analysis with empathy ROIs for moral luck

Results from a 2(outcome: neutral, negative) × 2(type of judgment: acceptability, blame) repeated measures ANOVA, carried out separately for neutral and negative belief, in each ROI for average percent signal change (PSC) extracted for moral judgment segments. Only interaction effects were of interest. Yellow cells represent significant ($p < 0.05$) values for the interaction term and green cells represent interaction terms *with* significant post-hoc comparisons for the judgment factor.

ROI	Comparison	main effect of consequence	main effect of judgment	interaction
dACC ($n = 39$)	Acc vs. Neu	$F\ (1,38) = 0.009, p = 0.925$	$F\ (1,38) = 8.019, p = 0.007$	$F\ (1,38) = 0.001, p = 0.974$
	Int vs. Att	$F\ (1,38) = 3.659, p = 0.063$	$F\ (1,38) = 1.024, p = 0.318$	$F\ (1,38) = 4.508, p = 0.040$
l-AI ($n = 36$)	Acc vs. Neu	$F\ (1,35) = 2.950, p = 0.095$	$F\ (1,35) = 3.067, p = 0.089$	$F\ (1,35) = 0.173, p = 0.680$
	Int vs. Att	$F\ (1,35) = 1.623, p = 0.211$	$F\ (1,35) = 0.771, p = 0.386$	$F\ (1,35) = 0.159, p = 0.692$
l-PI ($n = 29$)	Acc vs. Neu	$F\ (1,28) = 3.503, p = 0.072$	$F\ (1,28) = 2.286, p = 0.142$	$F\ (1,28) = 1.131, p = 0.297$
	Int vs. Att	$F\ (1,28) = 0.615, p = 0.440$	$F\ (1,28) = 1.033, p = 0.318$	$F\ (1,28) = 0.834, p = 0.369$
r-AI ($n = 38$)	Acc vs. Neu	$F\ (1,37) = 0.001, p = 0.977$	$F\ (1,37) = 5.365, p = 0.026$	$F\ (1,37) = 5.750, p = 0.022$
	Int vs. Att	$F\ (1,37) = 0.452, p = 0.506$	$F\ (1,37) = 0.501, p = 0.484$	$F\ (1,37) = 0.349, p = 0.558$
r-PI ($n = 26$)	Acc vs. Neu	$F\ (1,25) = 23.975, p < 0.001$	$F\ (1,25) = 1.418, p = 0.245$	$F\ (1,25) = 1.315, p = 0.258$
	Int vs. Att	$F\ (1,25) = 3.036, p = 0.094$	$F\ (1,25) = 0.077, p = 0.784$	$F\ (1,25) = 1.139, p = 0.296$
aMCC ($n = 42$)	Acc vs. Neu	$F\ (1,41) = 0.030, p = 0.862$	$F\ (1,41) = 0.112, p = 0.740$	$F\ (1,41) = 1.315, p = 0.258$
	Int vs. Att	$F\ (1,41) = 0.141, p = 0.709$	$F\ (1,41) = 3.052, p = 0.088$	$F\ (1,41) = 0.938, p = 0.338$

Text S17: Brain-behavior correlations for empathy ROIs

Spearman's correlations between average percent signal change (PSC) in each ROI in each condition. The four text segments investigated were: mental-state information, consequence, acceptability, blame. Yellow cells represent significant ($p < 0.05$) values.

- dACC ($n = 39$)

Accidental harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.195, p = 0.235	ρ = -0.135, p = 0.412	ρ = -0.210, p = 0.200	-
blame	ρ = -0.130, p = 0.429	ρ = -0.135, p = 0.411	-	ρ = 0.005, p = 0.974
Attempted harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.178, p = 0.279	ρ = 0.055, p = 0.739	ρ = 0.082, p = 0.620	-
blame	ρ = -0.155, p = 0.345	ρ = 0.101, p = 0.542	-	ρ = 0.156, p = 0.344
Intentional harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.047, p = 0.777	ρ = -0.081, p = 0.623	ρ = 0.006, p = 0.872	-
blame	ρ = 0.027, p = 0.869	ρ = 0.127, p = 0.440	-	ρ = 0.163, p = 0.323
Neutral harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.020, p = 0.902	ρ = -0.014, p = 0.930	ρ = -0.070, p = 0.671	-
blame	ρ = -0.078, p = 0.636	ρ = -0.217, p = 0.185	-	ρ = -0.295, p = 0.068

- l-AI ($n = 36$)

Accidental harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.281, p = 0.097	ρ = 0.028, p = 0.871	ρ = 0.351, p = 0.036	-
blame	ρ = 0.235, p = 0.167	ρ = 0.002, p = 0.990	-	ρ = 0.214, p = 0.209
Attempted harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.104, p = 0.547	ρ = -0.001, p = 0.997	ρ = 0.097, p = 0.575	-
blame	ρ = 0.111, p = 0.518	ρ = -0.080, p = 0.644	-	ρ = -0.018, p = 0.916
Intentional harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.124, p = 0.472	ρ = 0.003, p = 0.986	ρ = -0.175, p = 0.308	-
blame	ρ = 0.134, p = 0.436	ρ = 0.021, p = 0.905	-	ρ = 0.189, p = 0.269
Neutral case	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.020, p = 0.906	ρ = -0.024, p = 0.888	ρ = 0.034, p = 0.845	-
blame	ρ = -0.020, p = 0.908	ρ = -0.060, p = 0.729	-	ρ = -0.138, p = 0.421

- l-PI (n = 29)

Accidental harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = 0.319, p = 0.092$	$\rho = 0.427, p = 0.021$	$\rho = 0.153, p = 0.428$	-
blame	$\rho = 0.352, p = 0.061$	$\rho = 0.428, p = 0.021$	-	$\rho = 0.289, p = 0.128$
Attempted harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = 0.015, p = 0.938$	$\rho = -0.078, p = 0.687$	$\rho = -0.076, p = 0.694$	-
blame	$\rho = -0.065, p = 0.739$	$\rho = -0.260, p = 0.173$	-	$\rho = -0.076, p = 0.694$
Intentional harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = 0.025, p = 0.896$	$\rho = 0.081, p = 0.674$	$\rho = 0.022, p = 0.909$	-
blame	$\rho = 0.335, p = 0.101$	$\rho = 0.042, p = 0.844$	-	$\rho = 0.074, p = 0.727$
Neutral case	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = 0.209, p = 0.317$	$\rho = 0.385, p = 0.058$	$\rho = 0.245, p = 0.238$	-
blame	$\rho = 0.347, p = 0.090$	$\rho = 0.229, p = 0.271$	-	$\rho = 0.055, p = 0.795$

- r-AI (n = 38)

Accidental harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = 0.310, p = 0.058$	$\rho = 0.196, p = 0.237$	$\rho = 0.174, p = 0.296$	-
blame	$\rho = 0.322, p = 0.049$	$\rho = 0.108, p = 0.517$	-	$\rho = 0.153, p = 0.358$
Attempted harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = -0.007, p = 0.965$	$\rho = 0.034, p = 0.842$	$\rho = -0.173, p = 0.299$	-
blame	$\rho = -0.101, p = 0.548$	$\rho = -0.128, p = 0.445$	-	$\rho = -0.037, p = 0.827$
Intentional harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = -0.483, p = 0.200$	$\rho = -0.302, p = 0.065$	$\rho = -0.379, p = 0.19$	-
blame	$\rho = -0.172, p = 0.302$	$\rho = -0.050, p = 0.766$	-	$\rho = -0.015, p = 0.931$
Neutral case	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	$\rho = 0.002, p = 0.989$	$\rho = 0.102, p = 0.543$	$\rho = 0.082, p = 0.625$	-
blame	$\rho = -0.127, p = 0.446$	$\rho = -0.118, p = 0.479$	-	$\rho = -0.114, p = 0.495$

- r-PI (n = 26)

Accidental harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.179, p = 0.381	ρ = -0.297, p = 0.141	ρ = -0.262, p = 0.196	-
blame	ρ = 0.068, p = 0.742	ρ = -0.181, p = 0.377	-	ρ = -0.130, p = 0.527
Attempted harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.189, p = 0.355	ρ = 0.019, p = 0.927	ρ = 0.111, p = 0.589	-
blame	ρ = 0.154, p = 0.453	ρ = -0.034, p = 0.868	-	ρ = 0.033, p = 0.875
Intentional harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.074, p = 0.719	ρ = 0.021, p = 0.920	ρ = -0.038, p = 0.854	-
blame	ρ = 0.086, p = 0.677	ρ = -0.042, p = 0.840	-	ρ = -0.042, p = 0.838
Neutral case	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.232, p = 0.254	ρ = -0.102, p = 0.620	ρ = -0.166, p = 0.419	-
blame	ρ = -0.127, p = 0.537	ρ = -0.297, p = 0.140	-	ρ = -0.277, p = 0.170

- aMCC (n = 42)

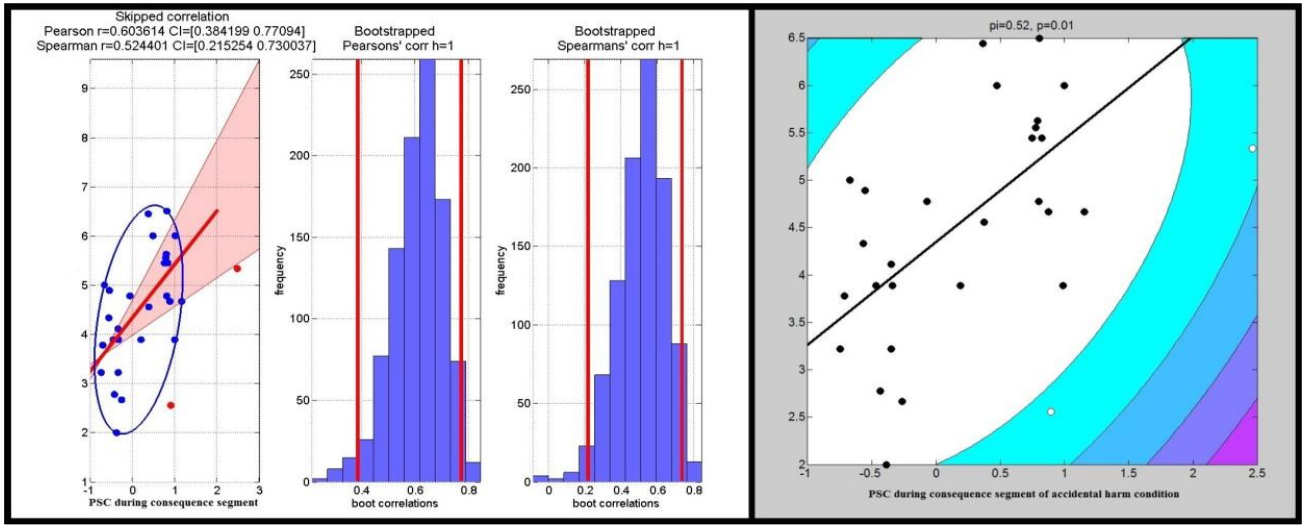
Accidental harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.207, p = 0.189	ρ = 0.119, p = 0.454	ρ = 0.132, p = 0.406	-
blame	ρ = 0.286, p = 0.067	ρ = 0.261, p = 0.095	-	ρ = 0.126, p = 0.428
Attempted harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.002, p = 0.989	ρ = 0.077, p = 0.629	ρ = 0.003, p = 0.985	-
blame	ρ = -0.091, p = 0.566	ρ = 0.023, p = 0.884	-	ρ = -0.068, p = 0.670
Intentional harm	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = -0.239, p = 0.128	ρ = -0.078, p = 0.622	ρ = 0.058, p = 0.715	-
blame	ρ = -0.107, p = 0.498	ρ = 0.085, p = 0.594	-	ρ = 0.119, p = 0.454
Neutral case	**mental-state**	**consequence**	**acceptability**	**blame**
acceptability	ρ = 0.141, p = 0.372	ρ = 0.057, p = 0.718	ρ = 0.059, p = 0.710	-
blame	ρ = 0.094, p = 0.554	ρ = 0.041, p = 0.798	-	ρ = 0.116, p = 0.464

Text S18: Brain-behavior correlations for empathy ROIs – robustness check

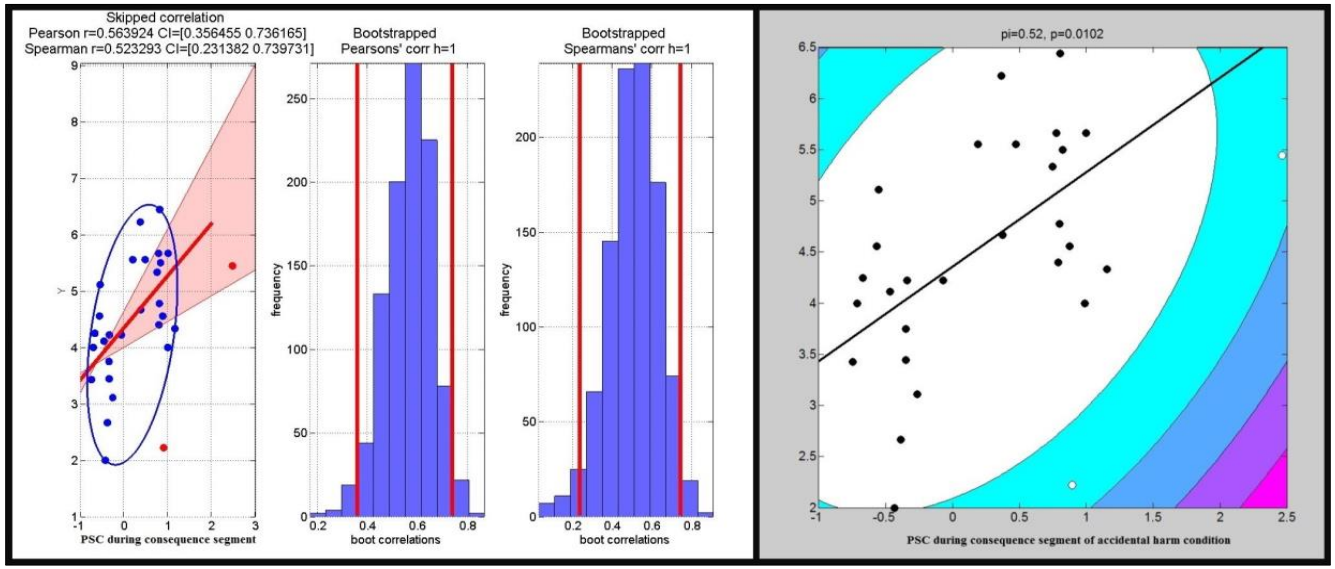
Robust Spearman's skipped correlations between average percent signal change (PSC) in each ROI in each condition, computed only for correlations with Spearman's $\rho < 0.05$. Yellow cells represent significant ($p < 0.05$) values.

ROI	Behavioral rating	PSC during which segment	Spearman Skipped correlation	Shepherd's Pi
l-AI	acceptability for accidental harm	acceptability judgment	$r = 0.318$, 95% CI [-0.036, 0.619]	-
l-PI	acceptability for accidental harm	consequence	$r = 0.524$, 95%CI [0.215, 0.730]	$Pi = 0.52$, $p = 0.0100$
l-PI	blame for accidental harm	consequence	$r = 0.523$, 95%CI [0.193, 0.734]	$Pi = 0.52$, $p = 0.0102$
r-AI	blame for accidental harm	mental-state information	$r = 0.292$, 95%CI [-0.045, 0.568]	-

- **Robust correlation between acceptability rating and PSC in l-PI during consequence segment**



- **Robust correlation between blame rating and PSC in l-PI during consequence segment**



Text S19: Whole-brain results

Random-effects analyses of the whole brain were also conducted for the intent task for the same text segments ($p < 0.05$, FWE-corrected, $k > 10$) to explore neural correlates of moral luck effect [(blame: accidental > neutral) > (acceptability: accidental > neutral)], but no suprathreshold activation was detected. Similar results were found for brain-behavior correlation analyses at the whole-brain level. These results are consistent with the higher power of functional ROI analyses to detect subtle but systematic response profiles (Saxe et al., 2006).

The only effect that survived multiple comparisons was interaction effect between belief and outcome observed in rTPJ for the mental-state information segment (x = 46, y = -54, z = 26; t = 4.88, p(FWE-corrected) = 0.011, k = 37). This result is in agreement with prior evidence that rTPJ is crucial in encoding belief information (Young et al., 2007).