

# SCIENTIFIC REPORTS



OPEN

## The behavioral and neural basis of empathic blame

Indrajeet Patil<sup>1,2</sup>, Marta Calò<sup>3</sup>, Federico Fornasier<sup>3</sup>, Fiery Cushman<sup>1,2</sup> & Giorgia Silani<sup>1,4</sup>

Received: 28 October 2016

Accepted: 31 May 2017

Published online: 12 July 2017

Mature moral judgments rely both on a perpetrator's intent to cause harm, and also on the actual harm caused—even when unintended. Much prior research asks how intent information is represented neurally, but little asks how even unintended harms influence judgment. We interrogate the psychological and neural basis of this process, focusing especially on the role of empathy for the victim of a harmful act. Using fMRI, we found that the 'empathy for pain' network was involved in encoding harmful outcomes and integrating harmfulness information for different types of moral judgments, and individual differences in the extent to which this network was active during encoding and integration of harmfulness information determined severity of moral judgments. Additionally, activity in the network was down-regulated for acceptability, but not blame, judgments for accidental harm condition, suggesting that these two types of moral evaluations are neurobiologically dissociable. These results support a model of "empathic blame", whereby the perceived suffering of a victim colors moral judgment of an accidental harmdoer.

Law, philosophy and psychology agree on a basic two-part template for moral judgment. On the one hand, we judge actions by considering the mental states that prompted them—for instance, a malicious motive or a misguided belief<sup>1–3</sup>. On the other hand, we also judge actions by assessing whether they actually caused harm, and how much<sup>1,4,5</sup>. Combining these features, if we see a person fire a gun, we naturally ask, "why did she shoot?" and "did she hit anyone?" before passing judgment.

The contribution of mental state analysis to moral judgment is well studied. Past research identifies the brain networks responsible for mental state analysis in moral judgment<sup>6–12</sup>, characteristic deficits in clinical populations<sup>13,14</sup>, and so forth<sup>15</sup>. Indeed, the synthesis of behavioral and neural research on mental state analysis in moral judgment stands out as a remarkable case study of integrative social cognitive neuroscience.

In contrast, far less research interrogates the behavioral and neural basis of the second major contributor to moral judgment: Representations of the actual harm caused by an act<sup>16–18</sup>. These play a crucial role in "moral luck", the phenomenon of chance outcomes influencing moral judgments<sup>1,19–22</sup>. (For instance, two drunk drivers who fall asleep at the wheel face very different penalties if one runs into a tree and another runs into a person). Although chance outcomes only play a small role in adults' moral judgments of a person's character or conduct (i.e., *acceptability judgments*), they play a large role in the assignment of blame and punishment<sup>1,4,23,24</sup> (i.e., *blame judgments*) (for a more detailed discussion, see Supplementary Text S1). We aim to clarify the psychological and neural basis of this effect.

Existing behavioral research hints that a key feature of harm representation is our capacity for empathy with a victim<sup>18,25–31</sup>. Below, we report new behavioral findings that support and refine the "empathic blame" hypothesis, which states that the degree to which we condemn others for producing harmful outcomes, intentionally or unintentionally, depends on the degree to which we empathize with the victim's suffering. We then turn to our principle aim, which is to interrogate its neural basis. Specifically, we ask whether, and how, a network of brain regions associated with empathy for the victim contribute to the moral judgment of third party actions.

In order to accomplish this, we build on prior research<sup>32</sup> showing that a specific network of brain regions is activated when we witness others suffer, referred to here as the "empathy network". These brain regions overlap partially with the brain regions that are activated when we experience the sensation of pain ourselves, indicating that they may enable empathy for suffering (i.e., a congruent emotional response in an agent and an observer<sup>33</sup>).

<sup>1</sup>Scuola Internazionale Superiore di Studi Avanzati, Neuroscience Sector, Trieste, Italy. <sup>2</sup>Department of Psychology, Harvard University, Cambridge, USA. <sup>3</sup>University of Trieste, Trieste, Italy. <sup>4</sup>Department of Applied Psychology: Health, Development, Enhancement and Intervention, University of Vienna, Vienna, Austria. Fiery Cushman and Giorgia Silani contributed equally to this work. Correspondence and requests for materials should be addressed to I.P. (email: [patilindrajeet.science@gmail.com](mailto:patilindrajeet.science@gmail.com)) or F.C. (email: [cushman@fas.harvard.edu](mailto:cushman@fas.harvard.edu)) or G.S. (email: [giorgia.silani@univie.ac.at](mailto:giorgia.silani@univie.ac.at))

(We refer to this set of neural structures as the “empathy network” not because it is activated exclusively by empathy—indeed, it certainly is not<sup>34</sup>—but rather because it responds reliably in the presence of empathy for the suffering of others<sup>35</sup>). Past research shows that the activation of this network plays a role in promoting prosocial action aimed at alleviating victim suffering<sup>36–40</sup>. Yet, at the neural level, it has not previously been associated with the moral condemnation of those who cause others to suffer<sup>41</sup>.

Following prior research into the neural basis of moral judgment<sup>2, 7, 8, 17</sup>, we presented participants with vignettes describing third-party actions that varied both in their intent and their outcome towards a potential victim. For instance, a person might put a white powder they believe to be sugar (neutral intent) or poison (harmful intent) into a potential victim’s coffee, and the powder might be sugar (neutral outcome) or poison, causing the victim to die (harmful outcome).

We use these cases to test three novel predictions. First, we predict that the empathy network will show enhanced activation at the time when the outcome to a victim is described, compared with other periods of the stimulus presentation when no information about harm is provided. This response might be selective to cases in which harm occurs, or it might generalize even to cases in which harm does not occur, indicating a reliance on affective mirroring to process both harmful and benign outcomes. In either event, this prediction is consistent with the hypothesis that empathy underlies the moral sensitivity to the experiences of the victim of a potential transgression. However, it can also be seen as consistent with the possibility that people empathize with victims in a manner disconnected from their moral judgment of the harmdoer.

Thus, more specifically, we predict a positive correlation between activation of the empathy network when delivering their moral judgments—presumably reflecting the influence of empathy for the victim upon moral judgment—and condemnation of actions that result in harm. In contrast, we predict no such correlation for malicious acts that happen not to produce harm; for instance, when a person pours what they believe to be poison into a person’s coffee, but the powder turns out to be harmless sugar. In other words, we expect that the more one empathizes with the suffering victim, the more one is likely to condemn the perpetrator’s action, but only for actions that result in harm.

Finally, we anticipate that enhanced activation of the empathy network during moral judgment will be moderated by the category of judgment, i.e. whether acceptability or blame judgment needs to be made. As noted above, moral luck (i.e., moral judgments that are sensitive even to chance outcomes) is substantially stronger for blame judgments than for acceptability judgments. In other words, even when people’s behavior is judged morally acceptable, they will be blamed for accidental harms that arise from it. Consequently, we predicted greater activation of the empathy network during judgments of blame (which rely more on the assessment of harm) than during judgments of acceptability (which rely less on the assessment of harm). This final prediction allows for an especially strong inference that activation of the empathy network subserves the task of *moral judgment*, rather than arising as an incidental byproduct of the stimuli.

In summary, we hypothesize that empathy for victims contributes to the condemnation of harm and, therefore, that (1) neural activation consistent with affective sharing will occur when learning about outcomes to victims, (2) especially among people whose moral judgments are most sensitive to outcomes, and (3) especially for types of judgment that are most sensitive to outcomes.

## Behavioral Studies

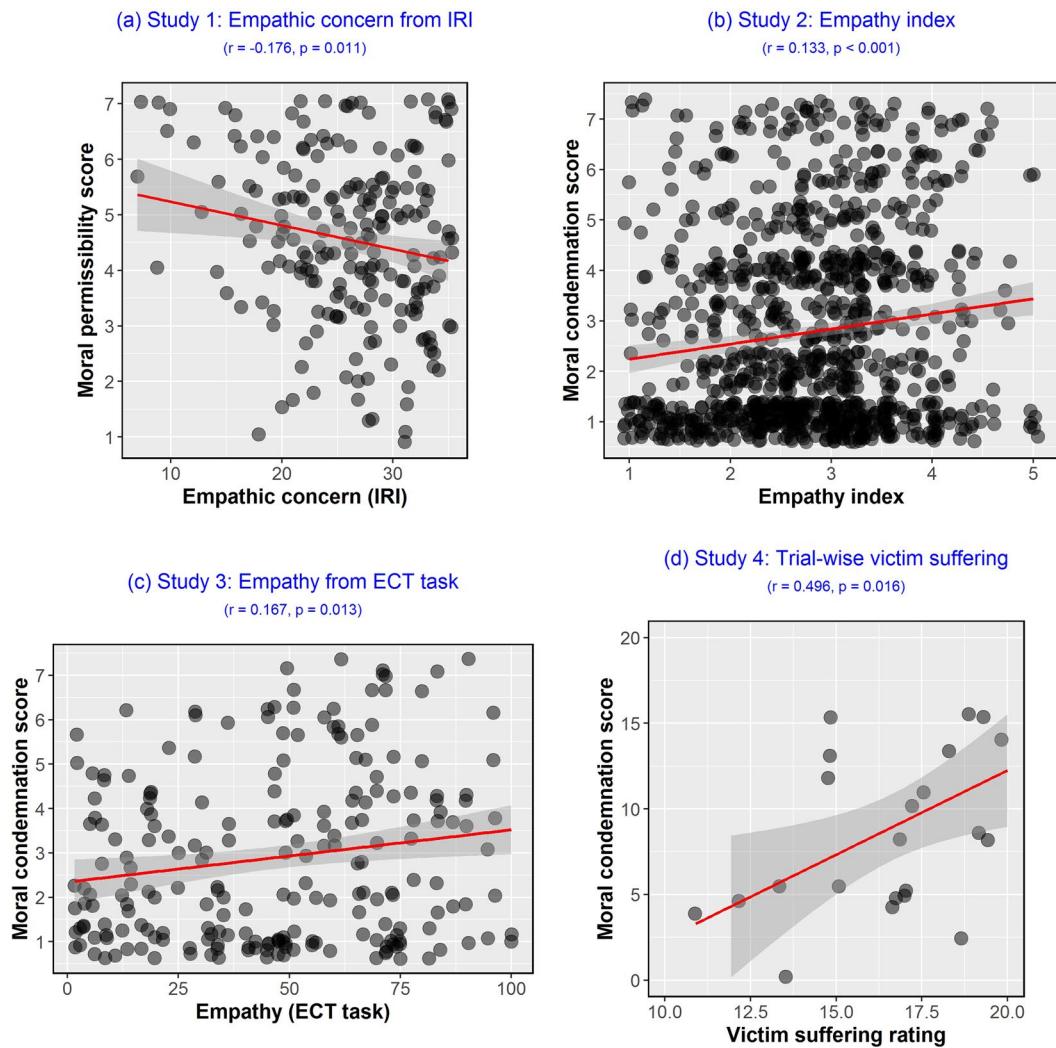
We begin by presenting the results of eight preliminary behavioral studies that motivate our neural investigation. These studies do not each precisely match our imaging study in terms of the content of the stimuli or manner of presentation, nor were they designed to. Rather, they establish a theoretical background against which our neuro-imaging design, analysis, and interpretation can best be understood.

### Studies (1–4): Empathy predicts the condemnation of accidents

The empathic blame hypothesis posits that we condemn actions based on the harm they cause because we empathize with the victim. This effect would be especially crucial for the condemnation of accidental harms, as compared to intentional harms. This is because intentional harms can be condemned for another salient reason: the agent’s malicious intent. In contrast, accidental harms are condemned solely based on harmful outcomes (perhaps due to empathy, as we hypothesize). A strong prediction, then, is that measure of empathy will correlate with measure of moral condemnation for accidents. We tested this prediction across four studies that utilized several trait and state measures of empathy (Fig. 1; full details in Supplementary Text S2–S5).

In Study 1, we found that participants scoring high on dispositional empathetic concern (EC, assessed with Interpersonal Reactivity Index or IRI<sup>42</sup>) judged accidental harms to be *less* morally permissible ( $r = -0.177$ , 95% CI  $[-0.306, -0.042]$ ,  $t(205) = -2.574$ ,  $p = 0.011$ ). We focused on EC subscale based on prior work showing that higher EC is associated with more severe assessment of accidental harms<sup>27, 29</sup>. In Study 2, we replicated this effect using another measure of trait empathy (Empathy index<sup>43</sup>) and found a positive association between higher self-report scores of empathy and condemnation of accidents ( $r = 0.133$ , 95% CI  $[0.072, 0.193]$ ,  $t(1002) = 4.234$ ,  $p < 0.001$ ). In Study 3, we borrowed a recently developed measure designed to dissociate empathy from compassion<sup>30</sup> and found that there was a positive correlation between empathy scores and severity of judgments for accidents ( $r = 0.167$ , 95% CI  $[0.036, 0.293]$ ,  $t(218) = 2.506$ ,  $p = 0.013$ ). In Study 4, we found that trial-by-trial ratings of perceived victim suffering for each moral vignette predicted harsher moral condemnation only for the accidental harms ( $r = 0.496$ , 95% CI  $[0.106, 0.754]$ ,  $t(21) = 2.618$ ,  $p = 0.016$ ). Note: Only simple linear correlations are shown here for illustrative purposes. To see results from ordinal mixed-effects regression analyses for these studies, see Supplementary Text S2–S5).

In summary, these studies supported the empathic blame hypothesis: individuals that either self-reported to have more empathic predisposition or reported to have stronger situational empathic engagement gave more



**Figure 1.** Positive relationship between empathy and condemnation of accidents. Behavioral studies 1–4 revealed that individuals who scored high on empathy, as assessed using (a) empathic concern (EC) subscale of Interpersonal Reactivity Index (IRI), (b) Empathy index measure, (c) empathy scores from the Empathy-Compassion Task (ECT), and (d) trial-by-trial victim suffering ratings, condemned accidental harm-doers more severely. Different questions were asked across studies 1–4: (a) moral permissibility, (b,c) wrongness and punishment, and (d) acceptability and blame.

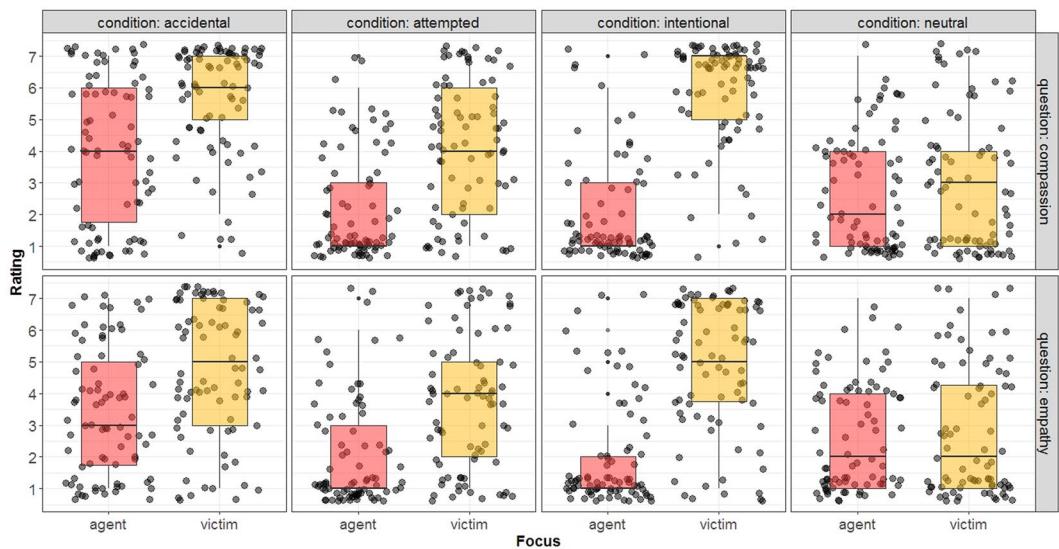
severe ratings for agents who unintentionally caused harm. We also note that all effect sizes were small to typical, according to contemporary guidelines<sup>44</sup>.

### Study 5: People empathize with victims more than perpetrators

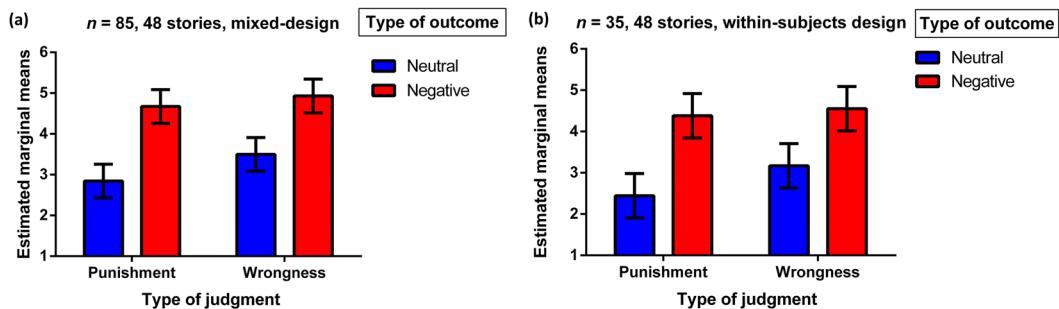
The preceding studies tacitly assumed that the relevant target of empathy in a moral vignette is the victim. In theory, however, unless properly tested, we cannot rule out the possibility that people are empathizing with the *perpetrator* of the harmful accident (“Oh dear, he must feel awful about what happened!”). The positive association found between assessment of victim suffering and blame for the agent speaks against this alternative, since it is difficult to see how sharing the internal states of the accidental harmdoer—which are blameless—would *increase* the severity of moral judgment. In Study 5 we test of this inference by directly asking people who they empathized with. Across all conditions of our design we find greater empathy for victims than agents ( $p < 0.001$ ; see Supplementary Text S6), although accidents do tend to elicit the greatest degree of empathy for agents, relative to the other conditions (Fig. 2).

### Studies 6–8: Punishment shows greater outcome effects than acceptability

As we have already discussed, prior research shows that judgments of moral blame show greater dependence on the harm caused to a victim than do judgments of moral acceptability<sup>1, 23, 24, 45</sup>. This effect is crucial to our neural analyses, and so we first successfully replicated it (Fig. 3)—both in a within-subjects design (similar, but not identical, to our neuroimaging paradigm) in Study 6, and then in a mixed-design (more similar to prior research) in Study 7 ( $p < 0.001$ ; see Supplementary Text S7 and S8 for full details).



**Figure 2.** Empathy is for the victims of harm, not the agent. Boxplots with jittered data-points represent participants' trial-by-trial ratings for empathy (affective sharing, i.e.) and compassion questions revealed that they empathized with and felt compassion for the victim, and not the agent, of the third-party dyad. Participants did report that they felt some empathy and compassion for agents involved in accidental harms.



**Figure 3.** Increased reliance on outcomes for punishment versus wrongness judgments. The outcome-by-judgment interaction effect was observed in (a) a mixed-design and (b) a within-subjects design such that the difference in ratings for negative versus neutral outcomes was greater for punishment judgments as compared to wrongness judgments. Error bars represent 95% confidence intervals.

To provide further convergent evidence, we conducted an additional study (Study 8) in which we explicitly instructed participants to share the victim's pain (vs. no instructions) and then provide either punishment or wrongness judgment. Again, our focus was cases of accidental harm. As expected, when instructed to share victim's pain, participants punished the accidental harm-doer more (Fig. 4), but their wrongness judgments remained unaffected (interaction effect:  $p = 0.06$ ; full details in Supplementary Text S9).

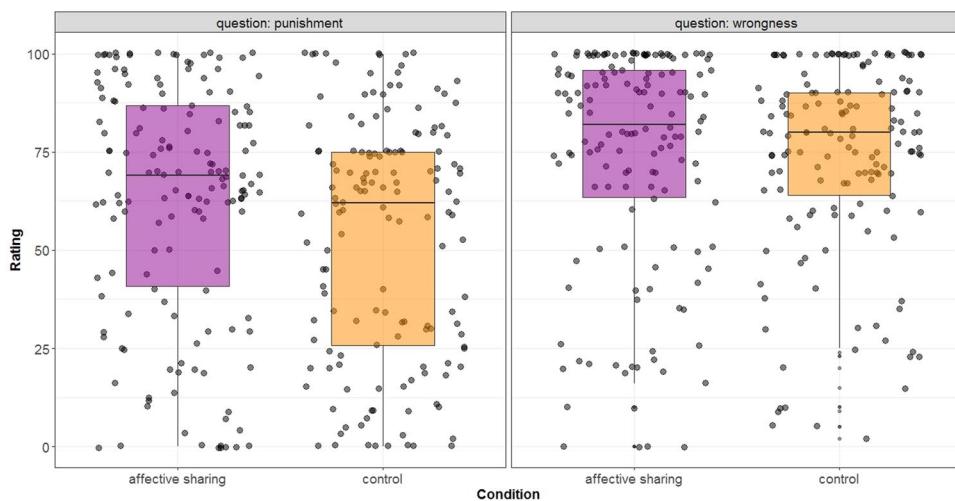
### fMRI Study (Study 9)

Having established the effects of interest at the behavioral level, we next turned to investigating the neural basis of empathic blame. We investigated if harmful outcomes are encoded and retrieved via the empathy network and whether this information is recruited differentially for different types of moral judgments.

### Methods and Materials

**Participants.** We tested 50 healthy community members (32 females; average age 23.06, SD = 3.08). We excluded all data from one participant because we discovered after testing that he had consumed clinically-prescribed psychoactive drugs. Functional data from two participants was removed due to excessive head motion, from four participants due to high collinearity among regressors (see below), and data from one additional participant could not be collected due to technical error. Thus, valid functional data was available for 42 participants, while behavioral data was available for 49 participants.

**Experimental stimuli and procedure.** *Moral judgment task.* Participants read and responded to 36 moral scenarios adapted from prior studies<sup>1,6</sup> and translated into Italian (see Supplementary Text S10 for more details). Across participants, each scenario appeared in four variations derived from a  $2 \times 2$  design crossing *belief*



**Figure 4.** Instructional affective sharing increases punishment for accidents. Boxplots with jittered data-points show participants' wrongness and punishment judgments in two conditions: when they were instructed to share the victim's pain (affective sharing group) and when no such instruction was provided (control group). Instructional affective sharing increased punishment ratings, but left wrongness judgments unaltered.

(neutral, negative) and *outcome* (neutral, negative). Each participant viewed 9 scenarios from each cell of this design.

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*: this segment provided information about whether the agent was acting with a neutral or harmful belief; (iv) *consequence*: this final segment described agent's action and its outcome. We use the terms *mental-state information* instead of *belief*, and *consequence* instead of *outcome*, to avoid confusion: the latter terms represent factors of the experimental design, while the former terms represent story segments containing information about the agent's beliefs and the nature of the outcome, respectively. All story text was then removed and replaced with the question and response scale (see Fig. 5).

After reading each story, participants provided two types of moral judgments<sup>1</sup> (presented in randomized order) on all trials:

- *acceptability*—“How morally acceptable was [the agent]’s behavior”? (1: *Totally acceptable* to 7: *Not at all acceptable*);
- *blame*—“How much blame does [the agent] deserve”? (1: *Not at all* to 7: *Very much*).

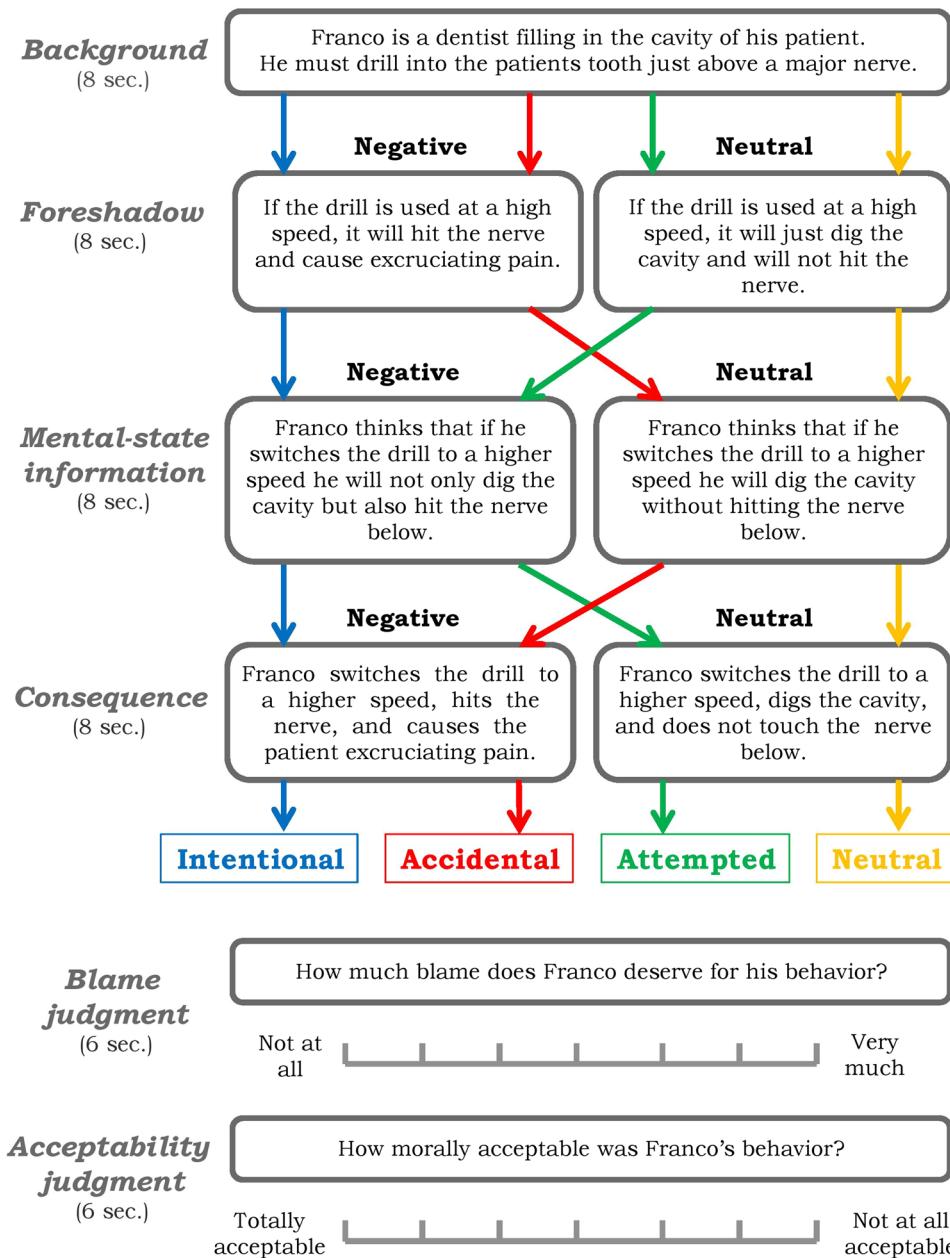
Note that although the online behavioral studies (6–7) included the terms “wrongness” and “punishment,” the fMRI study used the terms “acceptability” and “blame”. This is because the surveys were conducted in English with US participants, while the fMRI study was conducted in Italy and the scenarios were presented in Italian and introduced two additional concerns- (i) there is no direct translation of “wrongness” in Italian and the most natural analogue is “acceptability”; (ii) participants seem to adopt an internal punishment scale based on a legal metric of incarceration duration when the term “punishment” is used, even in the absence of explicit instructions to do so<sup>46</sup>, and we were not certain whether this would be the case with Italian participants and thus used the term “blame”. Crucially, prior research shows similar patterns of reliance on harm caused for both blame judgments and punishment judgments<sup>1</sup>. Indeed, the effect for punishment has been replicated across numerous studies<sup>1, 4, 23, 24</sup>.

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 initial location of the cursor was randomized across trials. 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 Supplementary Text S11.

In summary, the six segments of interest could be divided into two phases: four *reading phase* segments (*background*, *foreshadow*, *mental-state information*, *consequence*), when initial representations were formed, and *judgment phase* segments (*acceptability* and *blame*), when the encoded information was recruited in the service of making moral judgments.

*Functional localizer task.* Empathy involves the ability to understand and share others' affective states (emotions, pain, etc.) in isomorphic manner while maintaining self-other distinction<sup>32</sup>. It has been consistently shown that perceiving others in pain (empathy for pain) activates a set of brain regions that also encode nociceptive information while one is experiencing pain first-hand<sup>32, 35, 47, 48</sup>, the so-called “empathy network”.

To localize the empathy network in each participant, we used a modified version of a prior task<sup>49</sup>. Participants were shown 18 videos, each lasting for 3 s, of people experiencing painful auditory stimulation. After each video,

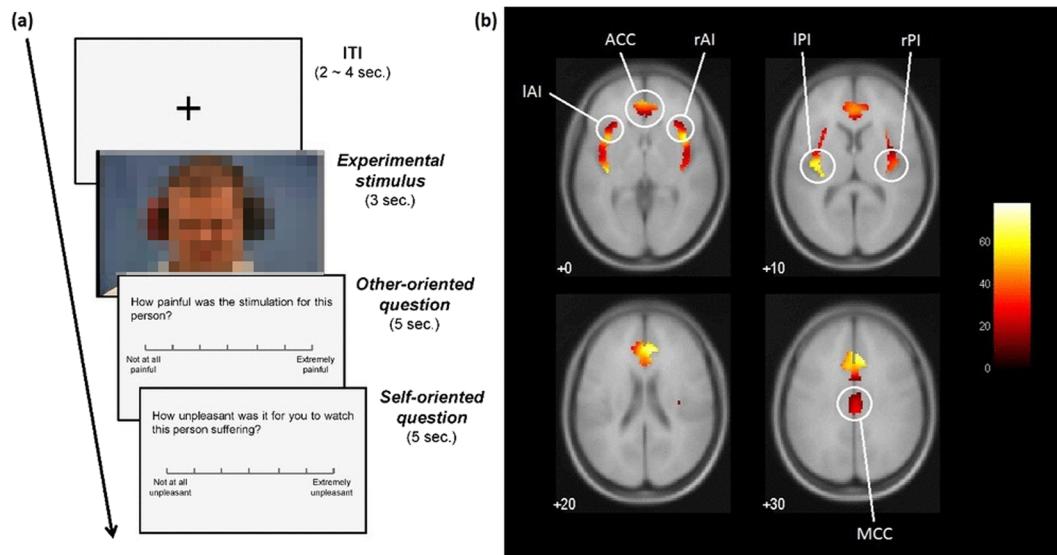


**Figure 5.** Experimental stimuli and design. Each moral vignette consisted of the following text segments: a *background* stem providing set-up for the story, a *foreshadow* segment that foreshadowed the nature of outcome, a *mental-state information* segment that provided information about actor's belief, a *consequence* segment that described action and its outcome. These segments were then followed by questions assessing acceptability and blame judgments.

participants provided ratings for the videos (see Fig. 6(a)). Mean inter-trial interval (ITI) was 2 s and was randomly jittered (jitter range: 0–2 s) to reduce predictability of the stimuli presentation.

**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, with standard parameters (see Supplementary Text S12). This study was conducted according to the principles in the Declaration of Helsinki, approved by the Ethics Committee of the Hospital ‘Santa Maria della Misericordia’ (Udine, Italy), and was carried out in accordance with the approved guidelines. All participants provided written informed consent before any study procedure was initiated.

Data were analyzed with SPM12. Each subject's data were motion-corrected (outliers were detected using the Art toolbox) and then normalized onto a common stereotactic space. Data were smoothed by using a Gaussian filter (full width half maximum = 6 mm at first-level) and high-pass-filtered (see Supplementary Text S13 for a more detailed description).



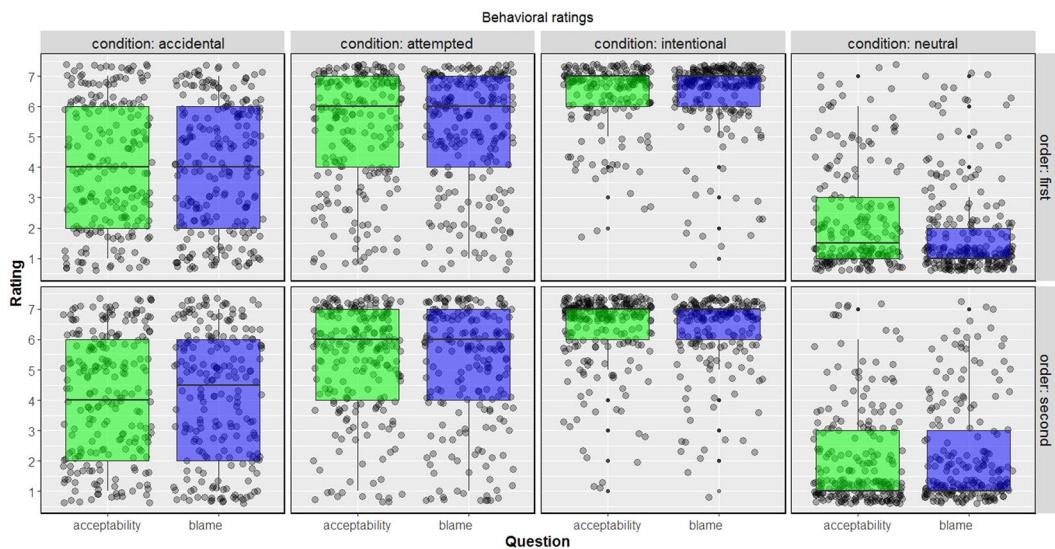
**Figure 6.** Task schematics and results for the empathy localizer task. (a) The videos showed individuals receiving auditory stimulation and displaying the transition from neutral facial expression (0.5 s) to exhibiting painful facial expressions (2.5 s). After presentation of each painful video, participants responded to two questions: one assessing *other-oriented* empathic response by gauging intensity of the experienced pain, while the other assessing *self-oriented* distress via experienced unpleasantness. The face of the target has been pixelated to protect the identity of the person in photo. (b) Brain regions where the BOLD signal was higher while watching painful videos as compared to baseline ( $n = 49$ , group-level random effects analysis,  $p < 0.05$ , FWE-corrected,  $k > 10$ ), masked with Neuromorphometrics anatomical atlas labels for ROIs. The slice numbers represent z-coordinate in MNI stereotactic space and the color bar denotes the F-statistic. Regions of Interest (ROIs) are highlighted and labeled. Abbreviations: ACC: anterior cingulate cortex, AI: anterior insula, MCC: midcingulate cortex, PI: posterior insula.

**fMRI data analysis at first-level.** For both moral judgment and empathy localizer tasks, the design matrices for fixed-effects General Linear Model were constructed by convolving a canonical hemodynamic response function (HRF) with the stimulus function for events (boxcar function) to create regressors of interest along with its temporal and dispersion derivatives. For a more detailed account, see Supplementary Text S13. Additionally, since we modelled each segment of the story and there was no jitter between these segments, there was possibility of high collinearity among regressors. Although we carried out collinearity diagnostic to address this issue (Supplementary Text S13), this still remains a limitation of the current study and future research should address this concern<sup>18</sup>.

**ROIs selection and localization at individual level.** At the first level, the following ROIs for empathy for pain were defined for each participant based on the localizer task (number of participants for whom an ROI was localized is given in parentheses): bilateral posterior insula (l-PI: 26/42, r-PI: 25/42), bilateral anterior insula (l-AI: 32/42, r-AI: 34/42), anterior cingulate cortex (ACC: 35/42), and midcingulate cortex (MCC: 38/42) (see Fig. 6(b)). The selection of ROIs was based on their consistency across quantitative and qualitative reviews<sup>32, 35, 50–52</sup> (Supplementary Text S11). ROI coordinates were derived from the contrast: experimental video > baseline (higher BOLD activity while watching painful videos as compared to baseline;  $p$ (uncorrected)  $< 0.001$ ,  $k > 10$ ). Not all ROIs could be localized for all participants (see Supplementary Text S14).

**ROI data extraction.** The data from spherical ROIs with a radius of 8 mm was extracted and analyzed using the MarsBar toolbox (v0.44) for SPM (<http://marsbar.sourceforge.net/>). Within the ROI, the average percent signal change (PSC) was computed relative to the adjusted mean of the time series (for more details, see Supplementary Text S15). 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. PSCs were also extracted for the *background* and *foreshadow* segments, although no information was available at this stage for any morally relevant evaluation to commence. As recommended<sup>53</sup>, data defining ROIs was independent from the data used in the repeated measures statistics. This helps us sidestep the nonindependence error<sup>54</sup> that can lead to spurious correlations and the observed results are thus unbiased and more trustworthy. Restricting analysis to a few ROIs also reduces Type-I error by drastically limiting the number of statistical tests performed<sup>55</sup>.

**fMRI data analysis at second-level.** Since the primary analyses centered on ROI analyses, details of the analysis at the group level are provided in Supplementary Text S13.



**Figure 7.** Behavioral results from the fMRI study. Boxplots with jittered data points for moral judgments given by participants in the fMRI study, displayed separately for each order. No outcome-by-judgment interaction was observed in either order. Note that higher ratings on acceptability question mean that the behavior was deemed to be *less* acceptable.

**Exploratory functional connectivity analysis.** Exploratory functional connectivity was carried out using standardized psychophysiological interaction (sPPI) analysis<sup>56–58</sup>. Specifically, we explored which brain regions showed changes in information exchange with the areas involved in decisions about blame (versus acceptability) for accidental harm cases. The ROI and whole-brain analysis revealed r-AI to be the only region that consistently tracked the outcome-by-judgment interaction (see Results) and thus this was chosen to be the seed region. We took the recommended precautions<sup>59,60</sup> while carrying out the PPI analysis (full details provided in Supplementary Text S16).

**Behavioral and ROI Data analysis.** Statistical analysis was conducted with R programming language. Given that both the behavioral data (items within conditions within participants) and PSC data from ROIs in the empathy network (conditions within segments within ROIs within participants) had multilevel or nested structure, we utilized linear mixed-effects models (LMM) to correctly handle the inherent dependencies in nested designs and to reduce probability of Type I error due to reduced effective sample size<sup>61–63</sup>. Additional advantage provided by the LMMs over the traditional ANOVA analyses is that they do not depend upon various assumptions (e.g., symmetric variance-covariance matrix) and can deal with missing data<sup>64</sup> (important for the current study as not all ROIs were localized for all participants). We note that there was a minimum of five observations per cell, a requirement for robust multilevel analysis<sup>65</sup>. Also, as recommended for confirmatory hypothesis testing<sup>62</sup>, all models included the maximal random effects structure (as long as no convergence issues were encountered). Maximum Likelihood (ML) estimation and variance components covariance structure were used.

**Data availability statement.** Unthresholded statistical maps of reported contrasts are available on Neurovault<sup>66</sup> at the following address: <http://neurovault.org/collections/1712/>. All the behavioral data are available at: <https://osf.io/893eh/>.

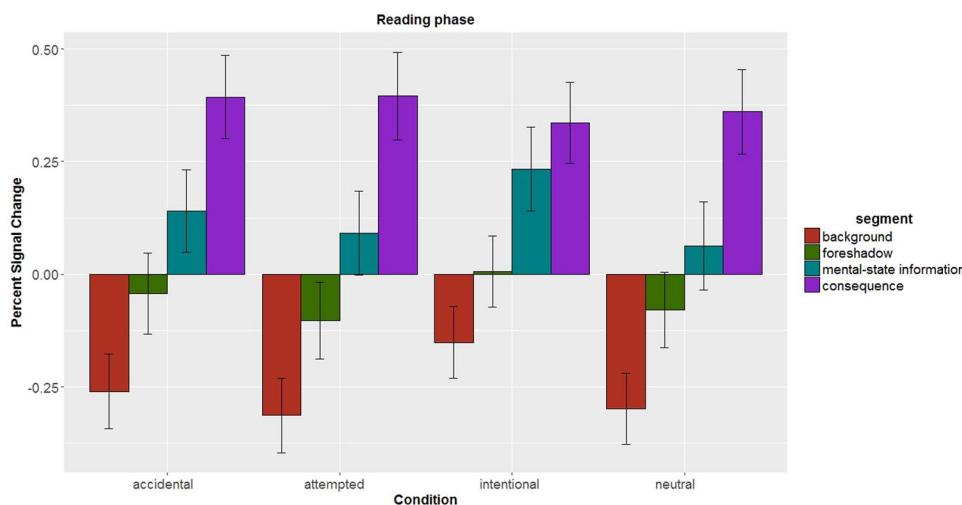
## Results

**Behavioral Results.** We used LMM that included fixed effects for all within-subjects factors: belief, outcome, type of question, and question order, as well as all possible two-way, three-way, and four-way interactions. We included random intercepts for participant. Given that previous work has revealed that presenting both types of moral judgments in within-subjects designs can lead to order effects<sup>67</sup>, we also included order in which questions were presented as one of the factors and present the results separately for each order (Fig. 7).

As expected from previous work, there was a main effect of belief and outcome and interaction between belief and outcome. But the fixed effect of interest to the current investigation, the interaction between outcome and the type of question/judgment, was not significant: estimate = 0.231, se = 0.233,  $df = 3370.585$ ,  $t = 0.993$ ,  $p = 0.321$ . Recall though that we did obtain this interaction in behavioral studies, reported as Studies 6–7, and this effect has also been observed in Italian sample<sup>23</sup> with similar design (For possible explanation of this null effect, see Supplementary Text S20). All other fixed effects and descriptive statistics are provided in Supplementary Text S17.

## fMRI data analysis results

**Greater empathy network response while encoding outcome information.** If the empathy network tracks the harm that a victim suffers during stimulus presentation, then the entire network should be



**Figure 8.** The empathy network encodes harmfulness of outcomes. The estimated PSC across all ROIs of the empathy network during the reading phase text, i.e. when the participants were provided with information about background, foreshadow, beliefs (*mental-state information*), and outcomes (*consequence*). The PSC in the network was significantly greater than the baseline when outcome information was provided, but not when other type of information was provided. Error bars correspond to 95% confidence intervals.

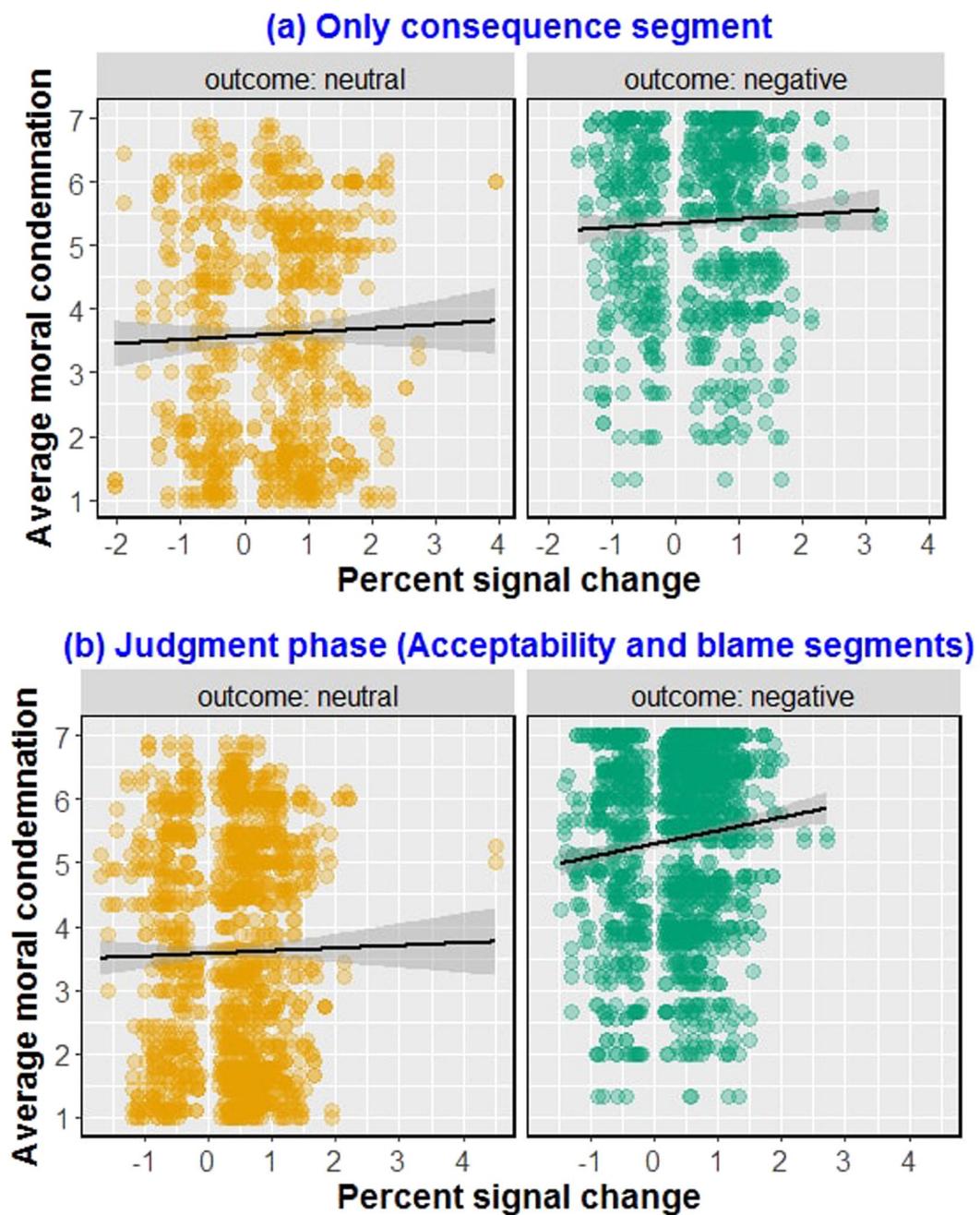
selectively more active for part of the story when outcomes to the victim are described (*consequence* segment) than for other parts of the story dominated by other types of information (e.g., background of the story (*background* segment), foreshadowing of the outcome (*foreshadow* segment), or information about belief states provided during the (*mental-state information* segment)). In other words, the response in the empathy network at encoding stage should be stimulus-bound<sup>16</sup>: modulated by the presence or absence of information about the victim's resulting state.

Indeed, LMM on the PSC data from reading phase segments revealed a main effect of *consequence* segment, such that the empathy network showed greater PSC during *consequence* segment than background (estimate = 0.655, se = 0.0553,  $z = 11.855$ ,  $p < 0.001$ ), foreshadow (estimate = 0.436, se = 0.0552,  $z = 7.910$ ,  $p < 0.001$ ), and mental-state information (estimate = 0.297, se = 0.0551,  $z = 5.390$ ,  $p < 0.001$ ) segment (see Fig. 8). All other fixed effects and graphical illustrations are provided in Supplementary Text S18. Thus, the empathy network was more active when information about harmfulness of outcomes was provided than when other type of information was presented. These results reveal neural basis of the empathic process by which harmfulness of outcomes is encoded.

**Empathy network response predictive of moral judgments for negative outcome conditions.** Next, we assessed the relationship between the activation of the empathy network and moral judgment. We predicted that the degree to which the entire empathy network responds to outcome information, both during reading phase (*consequence* segment, when this information is provided for the first time) and during the judgment phase (the *acceptability* and *blame* segments combined, when this information is recruited in the service of moral condemnation), will be predictive of the severity of moral condemnation, irrespective of the type of judgment. Additionally, this pattern of response should hold only for negative outcome conditions where a victim is harmed (accidental and intentional harms, i.e.) and not for neutral outcome conditions where no salient information about harm is present (neutral and attempted harms, i.e.).

As hypothesized, the mixed-effects regression on PSC in the empathy network during the judgment phase showed that moral condemnation was predicted by activity in the empathy network, more so for harmful outcomes than neutral outcomes (PSC  $\times$  outcome: estimate = 0.1959, se = 0.0628,  $df = 2947.94$ ,  $t = 3.1188$ ,  $p = 0.0018$ ). A similar effect was also found for the *consequence* segment in the reading phase (PSC  $\times$  outcome: estimate = 0.1477, se = 0.0710,  $df = 457.99$ ,  $t = 2.0798$ ,  $p = 0.0377$ ). In other words, both the initial encoding of the affective state of the victim during consequence segment, and the integration of this information while making judgments determined the degree to which empathic reaction towards the victim informs judgments (see Fig. 9). This result also underscores that the empathic response in this network was directed at the victim and not the perpetrator, since- (i) higher empathic response with the perpetrator would have predicted reduced moral condemnation; (ii) no information about affective state of the perpetrator was provided in the stimuli (for more detailed discussion, see Supplementary Text S19).

**Differential empathy network response to accidental harms for different judgments.** If it is true that the empathy network is encoding morally relevant harmful outcomes (as shown by preceding results) and harmful outcomes matter more for blaming as compared to assessing acceptability of third-party moral violations, the empathy network should be more active for blame judgments as compared to acceptability judgments. This should specifically be true for moral luck (accidental versus neutral) contrast, since at the behavioral level the severity of endorsed blame for an agent who accidentally produced harm while acting under false belief is

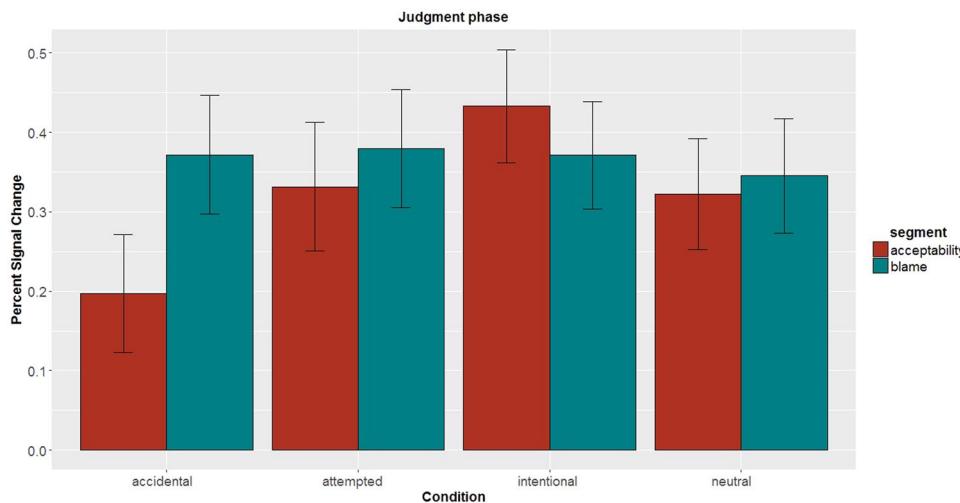


**Figure 9.** Brain-behavior correlation. The degree to which the empathy network was active (Percent Signal Change) while (a) reading information about harmfulness of outcomes and (b) providing ratings on judgments predicted the severity of moral judgments for actors who caused harmful outcomes.

increased significantly more than acceptability with reference to a neutral case<sup>1,23</sup>. Thus, we expected the empathy network to be active to a greater degree while people were assigning blame to accidents as compared to assessing the acceptability of an accident, but for no other condition.

The LMM carried out on PSC data from the judgment phase segments (*acceptability* and *blame*, i.e.) revealed that this was indeed the case. There was a significant fixed effect for the interaction belief × outcome × segment: estimate = -0.2622, se = 0.0841, df = 2940, t = -3.119, p = 0.002). All other fixed effects are provided in Supplementary Text S20. This interaction reflects that the empathy network was active to a greater degree for *blame* as compared to *acceptability* segments only for the case of accidental harm (neutral belief and negative outcome: estimate = 0.175, se = 0.042, t = 4.149, p < 0.001), but not otherwise (see Fig. 10).

To see which subregions of the empathy network showed the moral luck effect, we carried out exploratory *post hoc* analyses, which revealed that the pattern of greater PSC for blame versus acceptability judgments for accidents was seen in r-AI, r-PI, and ACC ( $p < 0.05$ ). But, at the group level in whole-brain analysis ( $p(\text{FWE}) < 0.05$ ), the moral luck effect was observed only in r-AI (full details in Supplementary Text S21).



**Figure 10.** The empathy network activity during the judgment phase. The estimated PSC across all ROIs of the empathy network during the judgment phase, i.e. when the participants made two types of moral judgments. The PSC was higher for blame as compared to acceptability judgments only for accidental harm scenarios ( $p < 0.001$ ). Error bars correspond to 95% confidence intervals. Abbreviations—PSC: percent signal change, ROI: region of interest.

Prior behavioral research on the judgment of accidental harm cannot distinguish two alternative regulation hypotheses: information about harmful outcomes may be either (i) selectively *down*-regulated while assessing acceptability of accidents, or instead (ii) selectively *up*-regulated while making blame judgments. We can address this question by interrogating the overall activation profile of the empathy network across our design. Comparing accidental harm condition with other conditions favors the former account because the activity for accidental harm condition is attenuated (as compared to other conditions) only for acceptability judgments but not for blame judgments. In other words, it appears that acceptability judgments are insulated from the influence of information about accidental harm due in part to the selective a down-regulation of the empathy network. To further probe this effect, we carried out functional connectivity analysis on r-AI, since this was the only ROI which consistently showed the moral luck effect across both ROI and whole-brain analyses.

**Functional connectivity results.** To investigate the neural regions that exhibited changes in functional connectivity with r-AI while making acceptability and blame judgments for accidental harm condition, exploratory psychophysiological interaction (PPI) analyses were conducted during judgment phase segments. This analysis revealed that r-AI exhibited decreased exchange of information (negative PPI effect, i.e.) with the left middle frontal gyrus or l-dlPFC (left dorsolateral prefrontal cortex;  $x = -34, y = 10, z = 36; \beta = -0.2480, t = -4.0582, k = 23, p(\text{uncorrected}) = 0.0003$ ) while making acceptability as compared to blame judgments. Notably, the dlPFC is associated with two functions of relevance to moral decision-making, information integration<sup>16,45</sup> and causal reasoning<sup>68</sup>. We return to consider the significance of this finding within our preferred account more fully below.

## Discussion

Our findings indicate that empathy for victims contributes to the moral condemnation of harmful acts (see Supplementary Text S22 for extended discussion). This role is particularly strong in cases of accidents, when a person causes harm that they did not intend, and thus provides an explanation for the phenomenon of “moral luck”. Several convergent behavioral and neural effects support this model of empathic blame. First, the condemnation of accidents is associated with increased perceptions of a suffering victim, and also with increased trait empathy. Second, among a network of brain regions that reliably activate during empathy for pain, we find increased activation as people learn about the harm caused by third-party action, and again when they make moral judgments of that action. Third, individuals who show the greatest neural response in the empathy network to harmful actions also show the greatest moral condemnation of the perpetrator. Finally, consistent with prior research showing that blame judgments are especially susceptible to moral luck effects, we show greater activation of the empathy network when people judge blame for an accidental harm, compared with judgments of acceptability. This effect is unique to the case of accidental harm. Together, these results indicate that people assign blame even to unintentional harmdoers in part because of their empathy for the suffering of a victim, and illuminate the neural basis of this effect.

Surprisingly, we observed identical overall activation levels in the empathy network during the presentation of information about harmful and neutral outcomes. Apparently, even in conditions where no harm was presented or implied in the stimulus, participants simulated the possibility of harm, or mirrored its absence, nonetheless<sup>69</sup>. We regard it as unlikely that people continuously evaluate the possibility of harm for all neutral events; rather, the presence of conditions with harmful outcomes in our experiment likely produced a relative contrast effect<sup>69</sup>. 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<sup>70</sup> and behavioral reactions towards a particular stimulus is contingent on

affective properties of other stimuli concomitant with it<sup>71,72</sup>. This finding indicates that in some contexts activation of the empathy network can be proactive (i.e., driven by the participants' expectations) rather than strictly reactive (i.e., driven by properties of the stimulus itself). A similar effect was also observed in the self-report data (Study 5), such that participants reported to have felt empathy for victims in situations where there was no harmful outcome (Fig. 2). By analogy, past work focusing on mentalizing during moral judgments which also shows that the mentalizing network tends to exhibit significant activation both in presence and absence of harmful intent<sup>2,6</sup>.

Similarly, we observed re-activation of the empathy network when participants made moral judgments of all four event types—a period during which no additional harm-relevant information was provided. Again, this finding has a natural analog in prior research on the neural basis of intent-based moral judgment: The mentalizing network reactivates during the judgment of cases both including and excluding a culpable mental state<sup>2,6</sup>. We propose that the re-activation of the empathy network reflects the retrieval of previously encoded representations of the harmfulness of outcomes, facilitating integration with other morally-relevant content<sup>73</sup>.

An analysis of functional connectivity revealed that r-AI, which exhibits the clearest neural instantiation of the moral luck effect, showed context-sensitive connectivity pattern with dlPFC. When judging accidents, dlPFC showed tighter coupling with putative harm representations in r-AI for blame judgments than for acceptability judgments, mirroring the greater behavioral influence of harm for blame than for acceptability. 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<sup>8,16,18,45</sup>, the so-called “integration-and-selection” function of dlPFC. In case of third party moral judgment, information about intent and outcome must be integrated into an overall judgment. Because acceptability judgments are dominated by the assessment of intent, while blame judgments rely additionally on the empathic representation of the harm to the victim<sup>1</sup>, the latter judgment type may invoke greater coupling with a brain region implicated in integration. This account makes a prediction, currently untested, that key nodes of the mentalizing network would also show coupling with dlPFC reflecting their role in representing the mental state of the harmdoer. An alternative and viable interpretation stems from prior research implicating the dlPFC in assigning causal responsibility to agents<sup>68</sup>. Thus, coupling between l-dlPFC and r-AI may reflect the integration of information about harm to a victim and the perpetrator causally responsible for that harm. Resolving these accounts, which are currently speculative, remains an important area for future research.

Although we refer to an “empathy network”, this is not intended to imply a network *specific* to empathy, but rather one that responds *reliably* to empathy for suffering<sup>32,35,47,52,74,75</sup>. Indeed, past research shows that these regions are involved in the processing of a variety of non-nociceptive, multimodal sensory inputs<sup>32,76</sup> and undergird a host of other cognitive and affective functions<sup>48,77–81</sup> and can be found to be active even in the absence of subjective sensation of pain<sup>82</sup>. Nevertheless, the tight correspondence between activation of this network, stimuli concerning victim pain (e.g., a rabid dog biting an older lady, skiers breaking their legs in accident, etc.), and participants’ moral judgments favor the conclusion that this network indexes empathy for suffering in our task.

In summary, we provide evidence that the empathy network supports the encoding and integration of harm representations during moral judgment. These results indicate a novel role for a well-studied functional network of brain regions, while also refining current theories of moral judgment. Finally, they clarify the psychological basis of moral luck—a longstanding matter of philosophical and legal concern<sup>17</sup>.

## References

- Cushman, F. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* **108**, 353–80 (2008).
- Young, L., Cushman, F., Hauser, M. & Saxe, R. The neural basis of the interaction between theory of mind and moral judgment. *Proc. Natl. Acad. Sci.* **104**, 8235–40 (2007).
- Malle, B. F., Guglielmo, S. & Monroe, A. E. A Theory of Blame. *Psychol. Inq.* **25**, 147–186 (2014).
- Cushman, F., Dreber, A., Wang, Y. & Costa, J. Accidental outcomes guide punishment in a ‘trembling hand’ game. *PLoS One* **4**, e6699 (2009).
- Alter, A. L., Kernochan, J. & Darley, J. M. Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. *Law Hum. Behav.* **31**, 319–335 (2007).
- Young, L. & Saxe, R. The neural basis of belief encoding and integration in moral judgment. *Neuroimage* **40**, 1912–20 (2008).
- Yu, H., Li, J. & Zhou, X. Neural Substrates of Intention–Consequence Integration and Its Impact on Reactive Punishment in Interpersonal Transgression. *J. Neurosci.* **35**, 4917–4925 (2015).
- Treadway, M. T. *et al.* Corticolimbic gating of emotion-driven punishment. *Nat. Neurosci.* **17**, 1270–5 (2014).
- Sellaro, R. *et al.* Increasing the role of belief information in moral judgments by stimulating the right temporoparietal junction. *Neuropsychologia* **77**, 400–408 (2015).
- Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A. & Saxe, R. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proc. Natl. Acad. Sci.* **107**, 6753–6758 (2010).
- Patil, I., Calò, M., Fornasier, F., Young, L. & Silani, G. Neuroanatomical correlates of forgiving unintentional harms. *Sci. Rep.* **7**, 45967 (2017).
- Margoni, F. & Surian, L. Children’s intention-based moral judgments of helping agents. *Cogn. Dev.* **41**, 46–64 (2017).
- Buon, M. *et al.* The role of causal and intentional judgments in moral reasoning in individuals with high functioning autism. *J. Autism Dev. Disord.* **43**, 458–70 (2013).
- Ciaranelli, E., Braghitta, D. & di Pellegrino, G. It is the outcome that counts! Damage to the ventromedial prefrontal cortex disrupts the integration of outcome and belief information for moral judgment. *J. Int. Neuropsychol. Soc.* **18**, 962–71 (2012).
- Chakroff, A. & Young, L. How the Mind Matters for Morality. *AJOB Neurosci.* **6**, 43–48 (2015).
- Buckholtz, J. W. & Marois, R. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* **15**, 655–661 (2012).
- Krueger, F. & Hoffman, M. The Emerging Neuroscience of Third-Party Punishment. *Trends Neurosci.* **39**, 499–501 (2016).
- Ginther, M. R. *et al.* Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. *J. Neurosci.* **36**, 9420–9434 (2016).
- Martin, J. W. & Cushman, F. In *The Blackwell Companion to Experimental Philosophy* (eds Sytsma, J. & Buckwalter, W.) 190–202 (Wiley Blackwell, 2016).

20. Lench, H. C., Domsky, D., Smallman, R. & Darbor, K. E. Beliefs in moral luck: When and why blame hinges on luck. *Br. J. Psychol.* **106**, 272–287 (2015).
21. Young, L., Nichols, S. & Saxe, R. Investigating the Neural and Cognitive Basis of Moral Luck: It's Not What You Do but What You Know. *Rev. Philos. Psychol.* **1**, 333–349 (2010).
22. Williams, B. & Nagel, T. Moral luck. *J. Aristot. Soc.* **50**, 115–151 (1976).
23. Patil, I., Young, L., Sinay, V. & Gleichgerrcht, E. Elevated moral condemnation of third-party violations in multiple sclerosis patients. *Soc. Neurosci.* **12**, 308–329 (2017).
24. Martin, J. W. & Cushman, F. Why we forgive what can't be controlled. *Cognition* **147**, 133–143 (2016).
25. Buon, M., Seara-Cardoso, A. & Viding, E. Why (and how) should we study the interplay between emotional arousal, Theory of Mind, and inhibitory control to understand moral cognition. *Psychon. Bull. Rev.* **23**, 1660–1680 (2016).
26. Gong, P. *et al.* The CAG polymorphism in androgen receptor (AR) gene impacts the moral permissibility of harmful behavior in females. *Psychoneuroendocrinology* **80**, 74–79 (2017).
27. Patil, I. & Silani, G. Alexithymia increases moral acceptability of accidental harms. *J. Cogn. Psychol.* **26**, 597–614 (2014).
28. Walter, N. T. *et al.* Ignorance is no excuse: moral judgments are influenced by a genetic variation on the oxytocin receptor gene. *Brain Cogn.* **78**, 268–73 (2012).
29. Trémolière, B. & Djériouat, H. The sadistic trait predicts minimization of intention and causal responsibility in moral judgment. *Cognition* **146**, 158–171 (2016).
30. Miller, R. & Cushman, F. Empathy, compassion, and moral judgment: The dissociable effects of other-oriented emotions across help and harm contexts. *PLoS One* (2017).
31. Piazza, J., Landy, J. & Goodwin, G. Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition* **131**, 108–124 (2014).
32. Lamm, C., Decety, J. & Singer, T. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage* **54**, 2492–502 (2011).
33. de Vignemont, F. & Singer, T. The empathic brain: how, when and why? *Trends Cogn. Sci.* **10**, 435–441 (2006).
34. Uddin, L. Q. Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* **16**, 55–61 (2014).
35. Lamm, C., Bukowski, H. & Silani, G. From shared to distinct self-other representations in empathy: evidence from neurotypical function and socio-cognitive disorders. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **371**, 20150083 (2016).
36. Tusche, A., Bockler, A., Kanske, P., Trautwein, F.-M. & Singer, T. Decoding the Charitable Brain: Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic Giving. *J. Neurosci.* **36**, 4719–4732 (2016).
37. Han, X. *et al.* Empathy for pain motivates actions without altruistic effects: Evidence of motor dynamics and brain activity. *Soc. Cogn. Affect. Neurosci.*, doi:[10.1093/scan/nsx016](https://doi.org/10.1093/scan/nsx016) (2017).
38. Patil, I. *et al.* Neuroanatomical basis of concern-based altruism in virtual environment. *Neuropsychologia*, doi:[10.1016/j.neuropsychologia.2017.02.015](https://doi.org/10.1016/j.neuropsychologia.2017.02.015) (2017).
39. Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E. & Roiser, J. P. Neurocomputational mechanisms of prosocial learning and links to empathy. *Proc. Natl. Acad. Sci.* **113**, 9763–9768 (2016).
40. Patil, I., Melsbach, J., Hennig-Fast, K. & Silani, G. Divergent roles of autistic and alexithymic traits in utilitarian moral judgments in adults with autism. *Sci. Rep.* **6**, 23637 (2016).
41. Ugazio, G., Majdandžić, J. & Lamm, C. In *Empathy and Morality* (ed. Maibom, H. L.) 155–171 (Oxford University Press, 2014).
42. Davis, M. H. Measuring individual differences in empathy: Evidence for a multidimensional approach. *J. Pers. Soc. Psychol.* **44**, 113–126 (1983).
43. Jordan, M. R., Amir, D. & Bloom, P. Are Empathy and Concern Psychologically Distinct? *Emotion* **16**, 1107–1116 (2016).
44. Gignac, G. E. & Szodorai, E. T. Effect size guidelines for individual differences researchers. *Pers. Individ. Dif.* **102**, 74–78 (2016).
45. Buckholtz, J. W. *et al.* From Blame to Punishment: Disrupting Prefrontal Cortex Activity Reveals Norm Enforcement Mechanisms. *Neuron* **87**, 1369–80 (2015).
46. Buckholtz, J. W. *et al.* The Neural Correlates of Third-Party Punishment. *Neuron* **60**, 930–940 (2008).
47. Betti, V. & Aglioti, S. M. Dynamic construction of the neural networks underpinning empathy for pain. *Neurosci. Biobehav. Rev.* **63**, 191–206 (2016).
48. Zaki, J., Wager, T. D., Singer, T., Keysers, C. & Gazzola, V. The Anatomy of Suffering: Understanding the Relationship between Nociceptive and Empathic Pain. *Trends Cogn. Sci.* **20**, 249–59 (2016).
49. Lamm, C., Batson, C. D. & Decety, J. The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *J. Cogn. Neurosci.* **19**, 42–58 (2007).
50. Bzdok, D. *et al.* Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Struct. Funct.* **217**, 783–96 (2012).
51. Fan, Y., Duncan, N. W., de Grecq, M. & Northoff, G. Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neurosci. Biobehav. Rev.* **35**, 903–911 (2011).
52. Iannetti, G. D. & Mouraux, A. From the neuromatrix to the pain matrix (and back). *Exp. brain Res.* **205**, 1–12 (2010).
53. Poldrack, R. Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci.* **2**, 67–70 (2007).
54. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect. Psychol. Sci.* **4**, 274–290 (2009).
55. Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localizers. *Neuroimage* **30**, 1088–96 (2006).
56. Friston, K. *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229 (1997).
57. Smith, D. V., Gseir, M., Speer, M. E. & Delgado, M. R. Toward a cumulative science of functional integration: A meta-analysis of psychophysiological interactions. *Hum. Brain Mapp.* **37**, 2904–17 (2016).
58. Di, X., Huang, J. & Biswal, B. B. Task modulated brain connectivity of the amygdala: a meta-analysis of psychophysiological interactions. *Brain Struct. Funct.* **222**, 619–634 (2016).
59. O'Reilly, J. X., Woolrich, M. W., Behrens, T. E. J., Smith, S. M. & Johansen-Berg, H. Tools of the trade: Psychophysiological interactions and functional connectivity. *Soc. Cogn. Affect. Neurosci.* **7**, 604–609 (2012).
60. Di, X., Reynolds, R. C. & Biswal, B. B. Imperfect (de)convolution may introduce spurious psychophysiological interactions and how to avoid it. *Hum. Brain Mapp.* **38**, 1723–1740 (2017).
61. Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V. & van der Sluis, S. A solution to dependency: using multilevel analysis to accommodate nested data. *Nat. Neurosci.* **17**, 491–496 (2014).
62. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013).
63. Judd, C. M., Westfall, J. & Kenny, D. A. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.* **103**, 54–69 (2012).
64. Baayen, R. H., Davidson, D. J. & Bates, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390–412 (2008).
65. Maas, C. J. M. & Hox, J. J. Robustness issues in multilevel regression analysis. *Stat. Neerl.* **58**, 127–137 (2004).
66. Gorgolewski, K. J. *et al.* NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform* **9**, 8 (2015).
67. Cushman, F., Sheketoff, R., Wharton, S. & Carey, S. The development of intent-based moral judgment. *Cognition* **127**, 6–21 (2013).

68. Operkalski, J. T. & Barbey, A. K. In *Oxford Handbook of Causal Reasoning* (ed. Waldmann, M.) (Oxford University Press, 2017).
69. Liljeholm, M., Dunne, S. & O'Doherty, J. P. Anterior insula activity reflects the effects of intentionality on the anticipation of aversive stimulation. *J. Neurosci.* **34**, 11339–48 (2014).
70. Elliott, R., Friston, K. & Dolan, R. J. Dissociable neural responses in human reward systems. *J. Neurosci.* **20**, 6159–65 (2000).
71. Mellers, B., Schwartz, A., Ho, K. & Ritov, I. Decision Affect Theory: Emotional Reactions to the Outcomes of Risky Options. *Psychol. Sci.* **8**, 423–429 (1997).
72. Patil, I., Cogoni, C., Zangrando, N., Chittaro, L. & Silani, G. Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Soc. Neurosci.* **9**, 94–107 (2014).
73. Cushman, F. Deconstructing intent to reconstruct morality. *Curr. Opin. Psychol.* **6**, 97–103 (2015).
74. Novembre, G., Zanon, M. & Silani, G. Empathy for social exclusion involves the sensory-discriminative component of pain: a within-subject fMRI study. *Soc. Cogn. Affect. Neurosci.* **10**, 153–64 (2015).
75. Sessa, P., Meconi, F. & Han, S. Double dissociation of neural responses supporting perceptual and cognitive components of social cognition: Evidence from processing of others' pain. *Sci. Rep.* **4**, 7424 (2014).
76. Liberati, G. et al. Nociceptive Local Field Potentials Recorded from the Human Insula Are Not Specific for Nociception. *PLoS Biol.* **14**, e1002345 (2016).
77. Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P. & Singer, T. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nat. Commun.* **7**, 10904 (2016).
78. Wager, T. D. et al. Pain in the ACC? *Proc. Natl. Acad. Sci.* **113**, E2474–E2475 (2016).
79. Craig, A. D. B. How do you feel-now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* **10**, 59–70 (2009).
80. Shackman, A. J. et al. The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nat. Rev. Neurosci.* **12**, 154–67 (2011).
81. Sevinc, G., Hakan Gurvit, I. & Nathan Spreng, R. Salience network engagement with the detection of morally laden information. *Soc. Cogn. Affect. Neurosci.*, doi:[10.1093/scan/nsx035](https://doi.org/10.1093/scan/nsx035) (2017).
82. Salomons, T. V., Iannetti, G. D., Liang, M. & Wood, J. N. The 'Pain Matrix' in Pain-Free Individuals. *JAMA Neurol.* **73**, 755–6 (2016).

## Acknowledgements

We would like to thank Justin Martin, Jens Melsbach, and Ryan Miller for their help with the behavioral data collection. The authors also gratefully acknowledge Marine Buon, Stephanie Campbell, Mina Cikara, Justin Martin, Federica Meconi, Giovanni Novembre, Livia Tomova, Bastien Trémolière, Hongbo Yu, Marco Zanon, and Micaela Maria Zucchelli for their helpful comments on the earlier version of the manuscript. We would also like to thank Jeanette Mumford for her invaluable advice regarding fMRI data analysis and Claus Lamm for providing the stimuli for the empathy localizer task. This work was supported by the Vienna Science and Technology Fund (WWTF; Project CS15-003) to GS. This article was supported by the Open Access Publishing Fund of the University of Vienna.

## Author Contributions

I.P. and G.S. designed fMRI research, while I.P. and F.C. designed behavioral studies; G.S. programmed the fMRI tasks; I.P., M.C., F.C., and F.F. acquired data; I.P. analyzed the data; I.P. and F.C. wrote the paper; G.S. provided critical revisions. All authors approved the submitted version of the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at doi:[10.1038/s41598-017-05299-9](https://doi.org/10.1038/s41598-017-05299-9)

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017