Journal of Experimental Social Psychology 94 (2021) 104102

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

When my wrongs are worse than yours: Behavioral and neural asymmetries in first-person and third-person perspectives of accidental harms

Joshua Hirschfeld-Kroen, Kevin Jiang, Emily Wasserman, Stefano Anzellotti, Liane Young *

Boston College, 140 Commonwealth Avenue Chestnut Hill, MA 02467, USA

ARTICLE INFO

Keywords:
Morality
Accident
Agency
Harm
Theory of mind
fMRI

ABSTRACT

Research on third-party moral judgments highlights two mechanisms as central to moral judgments of accidental harms: the inference of intent and the perception of harm. However, little is known about how these mechanisms are recruited when people evaluate themselves for harm that they have accidentally caused. Here we explore how a person's perspective — as either actor or observer — influences their moral judgments of accidental harm. We use fMRI to investigate how brain regions involved in the inference of intent and the perception of harm differentially respond when participants either cause (first-person) or observe (third-person) accidental harm. First, we find that people judge their own accidental harms more harshly than they judge others' accidents, and hold themselves more responsible for the unintended harmful outcomes of their choices. Second, we find that regions responding to the first-hand experience of pain are also more sensitive to first-person harms relative to third-person harms, and brain-behavior relationships in a subset of these regions suggest that the tendency to judge oneself more harshly may be supported by a greater sensitivity to the victim's experience of harm. Third, though we find that first-person harms recruit regions for mental state inference to a lesser extent than third-person harms, this difference does not appear to account for the behavioral differences in moral judgment between first-person and third-person harms. The results of this experiment suggest that accidental harms are an important context for broadening our understanding of the relationship between agency, empathy, and moral judgments about the self.

1. Introduction

In 2017, an article in the New Yorker profiled people who had, in a tragic stroke of bad luck, accidentally caused the death of another person (Gregory, 2017). One woman, given the pseudonym Patricia, described her frustration with loved ones who tried to comfort her in the days and weeks after she killed a motorcyclist with her car: "Yes, it was an accident, and in a certain sense we were both to blame, but, at the end of the day, *I* hit him, *I* took his life," she said. "No matter how much you want to dismiss it as an accident, I still feel responsible for it, and I am." She cried, "I hit him! Why does nobody understand this?"

This unsettling quote points to an asymmetry between Patricia's feeling that she had done something deeply wrong and the expressed feelings of others, who seemed to view her behavior more mercifully than she possibly could. The discrepancy mirrors a phenomenon that was first described by the philosopher, Bernard Williams, as 'agent-regret' — the first-person feeling of remorse and responsibility that is distinguished, in severity and kind, from what might be felt by a mere spectator to accidental harm (Williams, 1981). Williams tells a story much like Patricia's, of a lorry driver who, through no fault of his own, accidentally runs over a child who has suddenly darted out into the road. Williams proposes that we would rightly expect the driver to "feel differently from any spectator, even a spectator next to him in the cab".

In his discussion of this case, Williams makes an empirical claim about a discrepancy in moral judgments between a person who has caused harm accidentally (the actor) and a person who has merely witnessed the same event (the observer). Specifically, he predicts that the actor will judge themselves more harshly, and hold themselves more responsible, than an observer of the same situation would. This claim is rather surprising in light of a large body of research on self-serving positivity biases in social and moral attributions (Mezulis, Abramson, Hyde, & Hankin, 2004), such as the fundamental attribution error (Hewstone, 1990; Ross, 1977). And yet, while the prescriptive implications of the phenomenon of 'agent-regret' have been explored at length in the philosophical literature (Jacobson, 2013; MacKenzie, 2017; Sussman, 2018; Williams, 1981), no prior psychological research

* Corresponding author.
E-mail address: liane.young@bc.edu (L. Young).

https://doi.org/10.1016/j.jesp.2021.104102
Received 17 February 2020; Received in revised form 23 September 2020; Accepted 9 January 2021
Available online 2 February 2021

has tested for the descriptive existence of such a discrepancy in moral judgments of accidental harm, or the psychological mechanisms that might account for such a difference.

In the present study, we sought to fill these gaps in the literature. We began by asking whether people make harsher (or more lenient) moral judgments about accidental harm when they are actors, relative to mere observers. Next, we sought to uncover the possible neural mechanisms that might differentiate the moral evaluation of accidental harms between first-person and third-person perspectives. Extensive work on moral psychology highlights two features in particular – the inference of intent and the perception of harm – as central to moral judgments of accidents, and harms more generally (Cushman, 2008; Gray, Young, & Waytz, 2012). However, little is known about how people infer intent and perceive harm differently in the context of first-person accidental harm. Thus, in addition to testing for a behavioral asymmetry between first- and third-person moral judgments of accidental harm, the present study aims to answer the following questions:

1) After accidentally harming another person, do actors tend to think *more* about their own mental states than an observer would, or *less*?
2) After accidentally harming another person, do actors focus *more* on the victim's experience of harm, relative to an observer, or *less*?

Below, we review prior research that speaks to these questions, and ultimately informs the hypotheses of the present study. For the sake of simplicity, we divide our discussion of 'inference of intent' and 'perception of harm' into separate sections, treating them as independent psychological mechanisms that operate in parallel. We note, however, that these mechanisms may interact, such that the inference of an agent's intent could influence the perception of how much harm they caused to the victim(s), for example, and vice versa (Ames & Fiske, 2015; Gray, Schein, & Ward, 2014).

1.1. Inference of intent

The judgment that an accidental harm is *accidental* depends crucially on the ability to recognize that the actor did not have the intention to cause harm, which can in turn reduce the severity of one's judgment and increase the likelihood of forgiveness (Cushman, 2008). This capacity to incorporate mental state information in the construction of moral judgments is supported by a set of regions known as the Theory of Mind (ToM) network, which includes the bilateral temporoparietal junction (TPJ), dorsomedial prefrontal cortex (dmPFC), and the precuneus (PC) (Fletcher et al., 1995; Saxe & Kanwisher, 2003). Regions in this network differentiate intentional harm from accidental harm by representing the relevant mental states and integrating them for the eventual formation of moral judgments (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; Young & Saxe, 2009).

However, less is known about how ToM processes support how people think about themselves, and in particular, how people represent their own mental states when they make moral judgments about the harm that they have accidentally caused to others. Are we more or less likely to form moral judgments of our accidental harms by considering our innocent mental states, relative to when we judge others for the same type of violation? Intuitively, it would seem that we have more direct access to our own mental states than the mental states of others, which should ultimately facilitate the process of self-directed forgiveness. Further, a body of social psychological work on the fundamental attribution error (Hewstone, 1990; Ross, 1977), and other self-serving biases in social judgment, might lead us to expect that people would be uniquely attentive to their own innocent intentions, to the extent that consideration and expression of their own innocence could yield more favorable judgments by others. Critically, both of these accounts predict that people will judge themselves *less* harshly than others for harms that are accidental.

Alternatively, if a model of 'agent-regret' is correct and actors do judge themselves more harshly for accidental harm than observers, one possible explanation for such a phenomenon is that actors are *less* likely to reason about their own mental states. Contrary to the prior suggestion that people may have more direct access to their own mental states, Daryl Bem famously proposed that we infer our mental states just as we infer the states of others, as if we were third parties bearing witness to our own actions and inferring, after the fact, what we were thinking and feeling (Bem, 1972; see Cushman, 2018 on why such post-hoc rationalization is rational). A still more radical hypothesis is that we not only infer our own mental states like we infer the states of others, as Bem proposed, but that we actually reason about our own mental states *less* frequently, and with a more *limited* capacity. Indeed, there is some evidence that we are actually worse at representing our own past false beliefs than the false beliefs of others (Gweon, Young, & Saxe, 2011).

Such a deficit could be understood in light of the observation that we are constantly evaluating and predicting the behaviors of others by inferring their mental states, whereas we may not need to explicitly represent our own mental states *qua mental states* nearly as often (Young & Tsoi, 2013). When I sweeten my coffee by adding sugar, it may not be necessary for me to represent my desire explicitly, insofar as my habit of pouring sugar into a cup can be coordinated without metacognition. By contrast, if I see someone else do the same thing, it may be far more likely that I reason about why they're doing what they're doing (e.g. she wants her coffee to be sweet). In this way, mental state inferences about other people may feature more regularly and effortlessly in social cognition than mental state inferences about oneself, leading to potential deficits when we form judgments that could require us to spontaneously reason about our own intentions, beliefs, and desires, as is the case with accidental harms.

1.2. Perception of harm

Moral judgments are not only about the intent of an agent, but also about the severity of harm that is experienced by a victim who suffers (Gray, Waytz, & Young, 2012). Research on moral luck suggests that the perceived severity of harmful outcomes can influence moral judgments independently of inferences that people make about a harmful actor's beliefs or desires (Cushman, 2008; Martin & Cushman, 2016). The representation of harm to others is thought to be supported, in part, by regions in an 'empathic pain network' or 'salience network', including the anterior insula (AI) and the anterior cingulate cortex (ACC) (Singer & Lamm, 2009). These regions are reliably activated during the first-hand experience of pain and the vicarious observation of another person in pain, suggesting a shared neural substrate for the representation of pain and discomfort in both self and other (Krueger & Hoffman, 2016; Lamm, Decety, & Singer, 2011; Singer et al., 2004). Consistent with this view, activity in these regions has been found to correlate with self-reported empathy (Singer et al., 2004) and the facilitation of costly prosocial behavior towards suffering victims (Crockett & Lockwood, 2018; Tusche, Böckler, Kanske, Trautwein, & Singer, 2016). Most relevant to the present study, recent work has supported a model of 'empathic blame' in which greater empathy for the victim of accidental harm, encoded and represented in this same set of 'empathic pain' regions, enhances *third-party* condemnation of the actor (Patil, Calò, Fornasier, Cushman, & Silani, 2017).

How might the sensitivity to a victim's suffering differ in the first-person case, when we ourselves are responsible for accidentally causing harm to an innocent victim? Prior work on dehumanization might suggest that people selectively *reduce* their empathy for the victims of harm, or otherwise underestimate the degree of harm, when they (or their group) are responsible (Castano & Giner-Sorolla, 2006; Lee, Hardin, Parmar, & Gino, 2019; Leidner, Castano, Zaiser, & Giner-Sorolla, 2010). However, this phenomenon has typically been observed in the context of intentional harms, and is interpreted as a motivated strategy for avoiding perceived threats to one's moral character (Leidner et al., 2010).

In the case of accidental harms, a model of 'agent-regret' might lead us to predict that people will be motivated to focus *more* on the severity of harm when they are responsible, relative to when they are observers, and that this asymmetry would in turn contribute to the relatively harsher moral judgments they would make about themselves. Prior research on the neural correlates of empathy has provided evidence that the activity in 'empathic pain' regions, such as the ACC and the AI, can be modulated by a variety of socially relevant contexts and motivations. Activity in 'empathic pain' regions is reduced when the observed victim is more socially distant (Cheng, Chen, Lin, Chou, & Decety, 2010), and when the victim belongs to a different group or race (Azevedo et al., 2013; Hein, Silani, Preuschoff, Batson, & Singer, 2010). Most relevant to the present hypothesis, 'empathic pain' regions have been found to be sensitive to the agency of accidental harm, such that these regions are *more* active when the observer is fully responsible for having accidentally caused harm to the victim, compared to when that responsibility is partially shared with the victim (Cui, Abdelgabar, Keysers, & Gazzola, 2015; Yu, Hu, Hu, & Zhou, 2014), and when the victim is fully responsible for accidentally harming themselves (Koban, Corradi-Dell'Acqua, & Vuilleumier, 2013).

Critically, in these studies the actor's responsibility for harm is always inversely related to the victim's responsibility for harm, leaving open the possibility that the neural markers of empathy were *reduced* to the degree that the victim was more responsible for their own suffering, as has been shown in other work (Decety, Echols, & Correll, 2010). In the present study, we seek to more directly test the role of first-person agency by investigating the contrast between actors and observers in their relative recruitment of 'empathic pain' regions following accidental harm, as well as the role of these regions in producing subsequent moral judgments about accidental harm. Although 'empathy for pain' is one process among many processes that might constitute the broader construct of empathy (Bzdok et al., 2012), we chose to focus on 'empathic pain' regions given robust meta-analytic evidence that they overlap with regions involved in the first-hand experience of pain (Lamm et al., 2011), as well as recent evidence that 'empathic pain' regions are also recruited for moral judgments of accidental harms (Patil et al., 2017).

1.3. Overview of present research

In the present study, we designed a novel fMRI task in which participants believed that either they themselves or another player were making choices that could result in accidental harm being caused to another person. By isolating regions of the brain involved in ToM and 'empathic pain' processing, respectively, we provide a preliminary test of the theories laid out above.

First, we predicted that people would judge accidental harms more harshly when they themselves were the agent of harm, as opposed to mere observers. Second, we predicted that this pattern could be produced by a reduction in ToM in the first-person case, such that people would be less likely to spontaneously represent their own, relative to others', innocent intentions, and that this deficit would be reflected in reduced ToM network activity when participants were actors, relative to observers. Third, we predicted that people might also be more focused on the severity of the victim's pain when they had caused the pain themselves, relative to merely witnessing it. On this hypothesis, we predicted that regions of the putative 'empathic pain' network (AI and ACC) would show enhanced activity when participants were causing harm themselves, compared to when they were passively observing someone else cause harm. (See Fig. 1.)

2. Method

2.1. Participants

Twenty participants (N = 7 female, mean age 24.55 years) were recruited for the study via flyers and a recruitment email list. All participants were screened beforehand via an online survey to ensure that MRI scanning posed no physical risk. In addition, participants were screened for their sensitivity to emotional and physical pain, as well as their comfort with the experimental procedures (see Supplementary Materials); respondents with high pain sensitivity, or who did not feel comfortable with the experiment, were not contacted for further participation. Each participant was compensated at the rate of $25/h (for time spent in the scanner) and $10/h (for time spent outside the scanner). After collecting data from all twenty participants, we pre-registered and ran our primary analyses without collecting any further data.

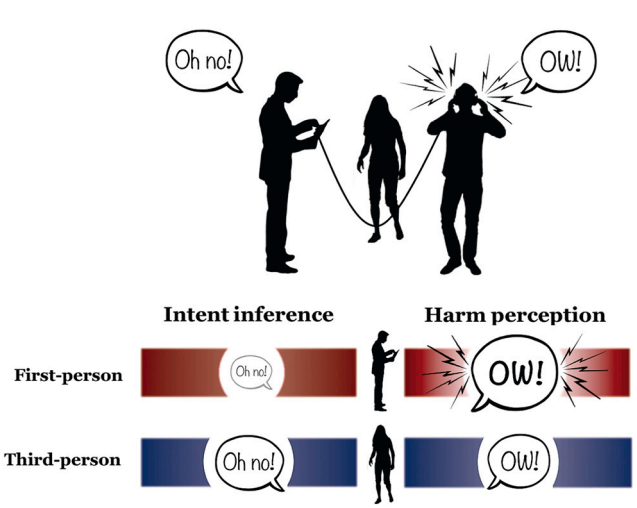

Fig. 1. Visual diagram of hypotheses. The silhouettes depict a hypothetical scenario in which the person on the left (the actor) accidentally blasts music, harming the listener on the right, while the person in the middle (the observer) witnesses the accident. We propose that the actor, relative to an observer, may judge their own accidental harm more harshly because 1) they are less likely to consider their own innocent intent, and/or 2) they are more likely to perceive harm when they are the cause. Critically, these mechanisms may be compatible, in that they could jointly contribute to an asymmetry between moral judgments of first-person and third-person accidental harms.

2.2. Procedure

Each participant went through the pre-scan consent and preparation with two confederate research assistants, who arrived at the scanning center independently and acted as though they were other participants in the study. The actual participant went through the main experimental tasks either in the order *ToM localizer → Card Choice task* or *Card Choice task → ToM localizer*; task order was alternated across participants. These two tasks were either preceded by or followed by a third task; this third task was not analyzed as part of this study and will not be discussed any further. The tasks were first described to the participant, after which the participant and two confederates viewed sample trials from each of the tasks.

2.3. Card choice task

The researcher explained that the three people would play a game together in the research study, with only one person being scanned, while the other two played the game on laptops placed outside of the scanner. There were two different roles in the game — Active player and Passive player — that were divided among the three players. Two of the three people would play as Active players, while one person would play as the Passive player. The actual participant was always assigned to the role of Active player in a way designed to appear random to the

participant, while the other confederate players were "randomly" assigned to the roles of Active player and Passive player, respectively. In addition, participants were told that a computer program would also play as the Active player on a portion of trials.

Participants were told that, on each trial of the task, the Active player (either the participant themselves, another player, or a random computer program) would have an opportunity to choose one of two differently colored cards (e.g. a red card vs a blue card). Participants believed that this card choice would in turn affect the Passive player, whose role was simply to passively experience the outcomes of the card choice (Fig. 2). The researcher explained that the Active player's goal was to attempt to learn a hidden relationship between card choices and outcomes, and that the Active player and Passive player should think of one another as partners, emphasizing the cooperative nature of the task. In reality, there was no hidden pattern that participants could learn: in all trials, the pairing of card colors was randomized, as was the mapping between card colors and outcomes.

The set of possible outcomes that could arise from a card choice were different depending on whether the Active player was a human or a computer. In trials where the Active player was a human (the participant or the other player), participants believed that the card choice could lead to either a harmful outcome for the Passive player (administered as a noise blast through headphones), or a neutral outcome (no noise blast). These outcomes were displayed to participants in the form of a "live" four-second video feed of the Passive player. In reality, the order of harmful and neutral trials was randomly shuffled, and the Passive player (a confederate) had been filmed in advance expressing 36 unique facial responses for harmful trials, and 36 unique responses for neutral trials. A colored border around the video feed (red for Harm, green for Neutral), combined with the Passive player's observed response, ensured that the participant was aware of the outcome of the card choice.

In trials where the Active player was the computer, a harmful noise blast would always result from the card choice, but it would be administered either to the Passive player or to the participant themselves. In this case, participants would either view a 4-s video feed of the Passive player receiving the noise blast, or a blurred photo of themselves at the same time as they received the noise blast during a 4-s period. Participants were aware that the relationship between the computer's card "choices" and the subsequent outcomes was completely random; this enabled us to collect data on the participants' neural responses to physical discomfort.

Before the participant entered the scanner, the Passive confederate was calibrated for sensitivity to uncomfortable noise blasts on a laptop, in view of the participant. This was done so that participants would believe that the Passive player was receiving uncomfortable noise blasts during the game. Separately, participants underwent the same calibration procedure with the scanner's audio headset. During calibration, the volume of the white noise was increased at a regular interval, and participants were instructed to respond when the noise was uncomfortably loud but not painful; this noise level was used for trials in which the computer administered a noise blast to the participant, so as to ensure that participants would find the noise blasts subjectively uncomfortable.

In every trial, after viewing or experiencing the outcome of a card choice, participants provided a wrongness judgment, i.e., How wrong was your/their/the computer's action? 1 (Not at all wrong) – 4 (Very wrong). In sum, the task was made up of six possible trial conditions:

1. Self Agent + Harm Outcome (to Passive player)
2. Self Agent + Neutral Outcome (to Passive player)
3. Other Agent + Harm Outcome (to Passive player)
4. Other Agent + Neutral Outcome (to Passive player)
5. Computer Agent + Harm Outcome (to Self)
6. Computer Agent + Harm Outcome (to Other)

These conditions are broken down by three *Agent* conditions, referring to the three possible Active players – Self, Other, and Computer. *Agent* was a cross-run condition, meaning that there was only ever one Active player making card choices within a given scanner run. Participants went through two contiguous scanner runs (blocks) for each Agent, and the order of *Agent* blocks was counterbalanced across participants. Self agent and Other agent runs each contained a total of 24 trials, divided equally into two possible *Outcome* conditions (12 Harm and 12 Neutral), for a total of 48 Self trials and 48 Other trials per participant. The order of *Outcome* type was randomly shuffled within runs. Computer agent runs also contained 24 trials, but these runs were instead broken down into two *Patient* conditions (12 self and 12 other), for a total of 48 Computer trials. The order of *Patient* type was randomly shuffled within runs.

In the scanner, each trial was presented in four sequential segments – *prompt* (2 s), *card choice* (4 s), *video* (4 s), and *judgment* (4 s). Although there were jittered fixations between trials, there was no jitter between sections of a trial, which leaves open the possibility of high collinearity among percent signal change values at contiguous time points. Given this limitation, our primary neural analyses were conducted by averaging neural activity across the entire trial time course, and we conservatively interpret exploratory analyses in which trials were broken up into the *card choice*, *video* and *judgment* sections.

2.4. Theory of Mind (ToM) localizer

Participants also completed a ToM functional localizer task

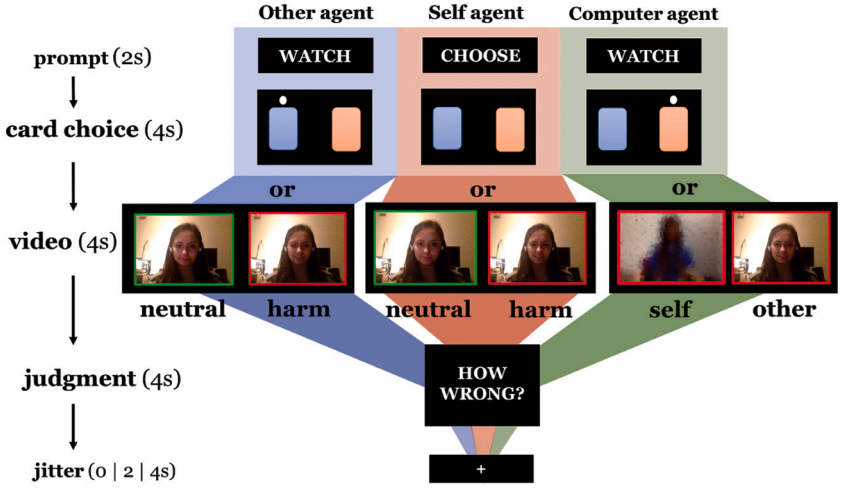

Fig. 2. Outline of Card Choice task trial structure. Trials began with a prompt (2 s), followed by a card choice (4 s), made by the other player, the participant, or a computer. A white dot informed participants of the card that was "chosen" by the other player or the computer on each trial. The card choice was followed by a video showing the outcome of the choice (4 s). A colored border around the video feed (red for harm, green for neutral), combined with the Passive player's observed facial expressions, ensured that participants knew whether a harmful outcome had occurred. After viewing the video, participants made a moral wrongness judgment (4 s) about the agent's action on a scale from 1 (Not at all wrong) to 4 (Very wrong). Each trial was separated from the previous trial by a jittered fixation of either 0, 2, or 4 s. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

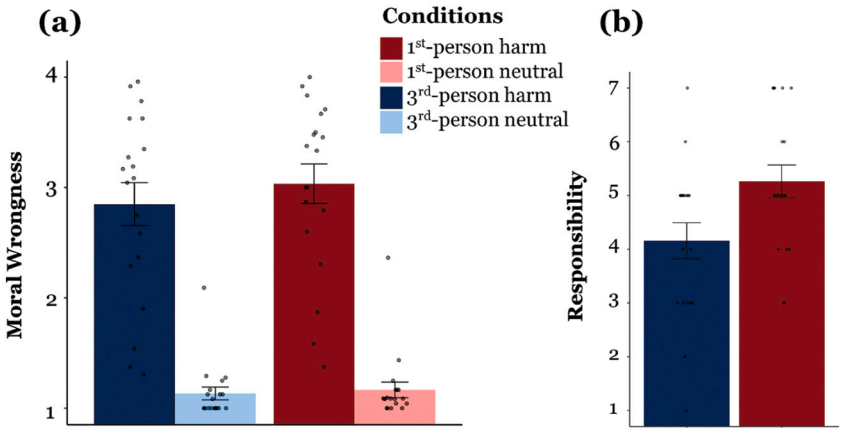

Fig. 3. Moral wrongness (a) and moral responsibility (b) judgments of accidental harms, in the form of noise blasts, are more severe when caused by oneself versus by another. Participants judged their own actions as more wrong (1, not at all wrong; 7, very wrong) than the other person's actions, but only when the card choice yielded a harmful noise blast for the Passive player. Error bars indicate standard error of the mean, calculated between-subjects, after averaging over all trials per subject per condition.

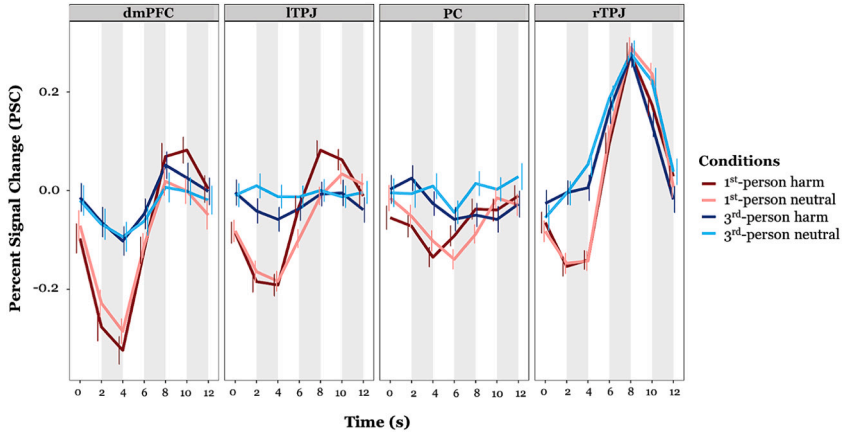

Fig. 4. Percent signal change (PSC) time courses for each condition across ToM ROIs: dorsomedial prefrontal cortex (dmPFC), left temporoparietal junction (lTPJ), precuneus (PC), and right temporoparietal junction (rTPJ). Each trial is broken up by the following stimulus bound sections, denoted by the shaded columns: *card choice* (from $t = 2$–4 s), *video* (from $t = 6$–8 s), and *judgment* (from $t = 10$–12 s). Error bars indicate standard error of the mean.

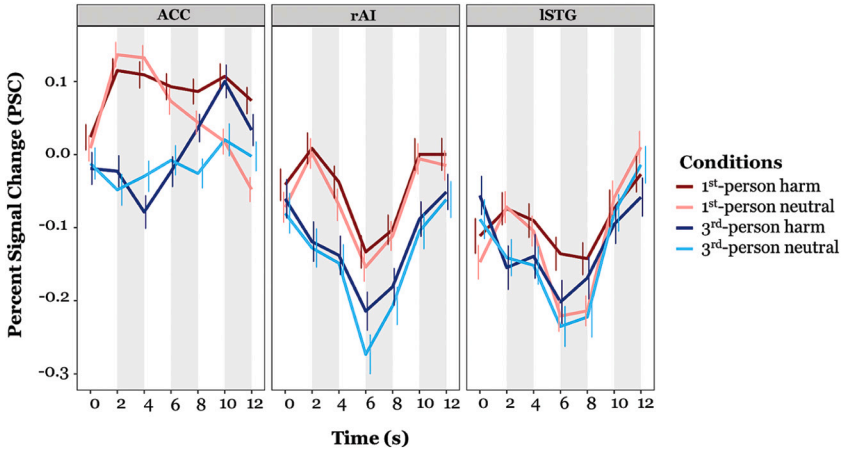

Fig. 5. Percent signal change (PSC) time courses for each condition across pain ROIs: anterior cingulate cortex (ACC), right anterior insula (rAI), and left superior temporal gyrus (lSTG). Each trial is broken up by the following stimulus bound sections, denoted by the shaded columns: *card choice* (from $t = 2$–4 s), *video* (from $t = 6$–8 s), and *judgment* (from $t = 10$–12 s). Error bars indicate standard error of the mean.

consisting of 10 stories about mental states (e.g. false-belief condition) and 10 stories about physical representations (e.g. false-photograph condition; see https://saxelab.mit.edu/saxelab-resources for the task files). The task was presented in two 4.5 min runs. We chose this task because it has been employed extensively in prior research on moral judgment and ToM (e.g., Dungan & Young, 2019; Gaesser, Hirschfeld-Kroen, Wasserman, Horn, & Young, 2019; Kim, Mende-Siedlecki, Anzellotti, & Young, 2021; Park & Young, 2020; Theriault, Waytz, Heiphetz, & Young, 2020), and because it consistently isolates a set of regions (e.g. RTPJ, LTPJ, dmPFC, PC) whose activity generalizes reliably

across the set of stories (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011).

2.5. Post-scan

Participants filled out a short post-scan survey with measures meant to capture a range of perceptions and judgments about themselves, the other Active player, the Computer agent, and the Passive player (see Supplementary Materials for all post-scan measures). In particular, participants were asked to report their feelings about how responsible they, the other player, and the computer were, on a scale from 1 (Not at all responsible) to 7 (Completely responsible), when harm was caused to the Passive player ("When your partner received a noise blast, did you feel that [you were/the other person was/the computer was] responsible?").

They also completed the Interpersonal Reactivity Index (IRI), which differentiates among four distinct subcomponents of empathy: empathic concern, perspective taking, personal distress, and fantasy (Davis, 1983). Each subscale contained 7 items with high inter-item reliability (all Cronbach's alpha >0.80), and all items were anchored on a scale from 1 (Does not describe me well) to 5 (Describes me very well). The empathic concern (EC) subscale measures a person's tendency to feel compassion or sympathy for the suffering of another (ex: "I often have tender, concerned feelings for people less fortunate than me"). The perspective taking (PT) subscale measures an individual's tendency to spontaneously adopt the perspective of other people (ex: "Before criticizing somebody, I try to imagine how I would feel if I were in their place"). The personal distress (PD) subscale measures the tendency to experience discomfort in response to the distress of others (ex: "When I see someone who badly needs help in an emergency, I go to pieces"). We did not make any predictions about the Fantasy subscale, and it is not discussed further.

After completing the survey, participants were debriefed on the purpose of the experiment and fully informed of the deception used in the task (see debriefing script in experimental materials: https://osf.io/3hq89/). None of the participants reported that they had realized the deception (either involving the confederates, or the randomness of the cards) during the experiment.

2.6. MRI data collection and analysis

Participants were scanned on a Siemens 3T Prisma scanner at the Martinos Imaging Center in Cambridge, MA, using a 32-channel coil. Images were acquired in 36 slices (3-mm isotropic voxels), TR = 2 s, TE = 30 ms, flip angle = 90°. All fMRI data were preprocessed using fMRIPrep (Esteban et al., 2019), a Nipype based tool (Gorgolewski et al., 2011). Each T1w (T1-weighted) volume was corrected for intensity non-uniformity (Tustison et al., 2010) and skull-stripped (with the OASIS template) using ANTs (Avants, Epstein, Grossman, & Gee, 2008). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov, Evans, McKinstry, Almli, & Collins, 2009) was performed through nonlinear registration with the antsRegistration tool of ANTs, using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed with FSL on the brain-extracted T1w (Zhang, Brady, & Smith, 2001). Functional data was motion corrected using FSL's mcflirt (Jenkinson, Bannister, Brady, & Smith, 2002). This was followed by co-registration to the corresponding T1w using boundary-based registration with six degrees of freedom, using flirt (FSL) (Greve & Fischl, 2009). Motion correcting transformations, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step with ANTs using Lanczos interpolation. Images were smoothed with a 5 mm-FWHM Gaussian kernel.

All fMRI analyses were conducted within regions of interest (ROIs) by extracting percent signal change (PSC). The onset of each trial was defined as the onset of the card-choice section, and the duration of the entire trial was specified as the 12 s (6 TRs) of the trial (excluding the prompt section), with all time points shifted by 4 s from presentation time to adjust for hemodynamic lag. Averaging across all voxels in the ROI, percent signal change (PSC) relative to baseline was calculated for each time point (TR) in each condition, where PSC (at time t) = 100 × [(average magnitude response for condition at time t – average magnitude response at baseline)/average magnitude response at baseline]. Specifically, baseline for each condition and each ROI was calculated on a runwise basis as the average response in that ROI at all time points of the 'prompt' section within a run, excluding the first 4 s after the offset of each trial (to allow the hemodynamic response to decay). Primary analyses were conducted on the entire trial duration (12 s), and exploratory analyses were broken down by sections of the trial (*card choice*, *video*, and *judgment*). All reported analyses were conducted on a subset of the data that excludes the Computer agent conditions (although see the *Empathic pain ROIs* section below for the relevance of these conditions in ROI selection).

Framewise displacement (FD) was calculated for each functional run using the implementation of Nipype (Power et al., 2014). We originally pre-registered a motion scrubbing procedure that involved removing all volumes with more than 1 mm of FD, and entire runs in which more than 10% of volumes passed this 1 mm threshold. The latter criteria would have yielded excessive data loss, leading to the establishment of a more conservative procedure of removing all volumes with more than 1 mm of FD and subsequently removing any trials that contained more than one image passing this threshold. One participant was excluded because we could not localize any ToM ROIs, and they moved by more than 1 mm of FD in over 50% of trials.

2.7. Theory of Mind ROIs

We defined individually tailored functional ROIs in the RTPJ, LTPJ, dmPFC, and PC based on a whole-brain contrast of false-belief stories over false-photograph stories in the ToM localizer (Dodell-Feder et al., 2011; Supplementary Table 1). The GLM used to define ToM ROIs included six additional noise regressors for the anatomical CompCor variant (aCompCor) (Behzadi, Restom, Liau, & Liu, 2007). The ROIs were defined as all contiguous voxels within a 9-mm radius of the peak voxel that survived the contrast threshold ($p < 0.001$, uncorrected, $k > 16$). This extent threshold of 16 voxels was computed via 1000 iterations of a Monte Carlo simulation (Slotnick et al., 2003).

2.8. Empathic pain ROIs

We used the conditions from the Card Choice task in which players were harmed by the Computer agent to isolate ROIs involved in processing the experience of pain or discomfort caused by the noise blast. Crucially, these conditions (Conditions 5–6) are distinct from those used for the main analyses (Conditions 1–4). An event in the Card Choice task was defined within the GLM framework, implemented in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/), as a single trial with an onset starting at the card choice section and a duration of 12 s, modeled as boxcar regressors convolved with a standard hemodynamic response function (HRF). The GLM also included six noise regressors for the anatomical CompCor variant (aCompCor) (Behzadi et al., 2007).

A whole-brain group-level contrast of *computer harms self* > *computer harms other* ($p < 0.001$, uncorrected, $k > 16$, extent threshold set by permutation testing) yielded a pair of regions that emerge consistently across qualitative and quantitative reviews of the neural overlap between nociceptive and 'empathic pain' processing: the right Anterior Insula (rAI) and the Anterior Cingulate Cortex (ACC) (Lamm et al., 2011). This contrast also yielded a large cluster with a peak coordinate in the left superior temporal gyrus (STG), which is a structure that is implicated in auditory processing. We include this region as part of the 'empathic pain' network because it has also been implicated via meta-

analyses in the anticipation of first-person pain (Palermo, Benedetti, Costa, & Amanzio, 2015), and responds to the observation of painful facial expressions in others (Botvinick et al., 2005). ROIs were drawn as 9-mm radius spheres around the peak coordinates that came out of the second-level whole-brain contrast (see Supplementary Table 1 for peak MNI coordinates; see Supplementary Materials for brain-behavior analyses that validate the involvement of these ROIs in first-person pain processing).

2.9. Analysis plan

All statistical analyses were conducted in R (version 3.5.1) programming language, unless otherwise specified. For primary analyses, we implemented linear mixed-effects models (LMMs) with the *lmer* package (Bates, Mächler, Bolker, & Walker, 2015), and obtained *p* values for fixed effects via Satterthwaite's degrees of freedom method in *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2015). While both primary and exploratory analyses were only conducted with a subset of measures and participants, we report all measures, manipulations and exclusions. One participant was excluded from fMRI analyses as a result of excessive motion (see MRI data collection and Analysis). Another participant was excluded from all further analyses (behavioral and fMRI) because they did not finish either run of the Other Agent condition. These exclusions yielded a final sample size of 19 participants for behavioral analyses, and 18 participants for fMRI analyses, consistent with recent neuroimaging research examining ToM regions ($N = 18$, Gaesser et al., 2019; $N = 16$, Niemi, Wasserman, & Young, 2018; $N = 18$, Tsoi, Dungan, Waytz, & Young, 2016). Behavioral data, percent signal change (PSC) data, analysis scripts, and a pre-registration of primary analyses are all available on OSF (https://osf.io/3hq89/).

First, we conducted behavioral analyses of in-scanner judgments of moral wrongness, to test whether these judgments were more severe for first-person harms relative to third-person harms. Our full model predicted moral wrongness judgments from fixed effects of *Agent* (self vs other), *Outcome* (harm vs neutral), their interaction, and included by-subject random intercepts and slopes for the effects of agent, outcome, and their interaction. Ratings of moral wrongness were made on a scale from 1 (not at all wrong) to 4 (very wrong), a range that has yielded meaningful patterns in prior fMRI studies (Niemi et al., 2018; Theriault et al., 2020; Tsoi, Dungan, Chakroff, & Young, 2018). Effect sizes (Cohen's *d*) for the effects in this model were estimated by dividing the mean difference between conditions by the square root of the summed variance components (as described in Brysbaert & Stevens, 2018). We also ran a paired samples *t*-test of post-scan responsibility judgments to test whether participants felt that they were more responsible than the other Active player for the harmful trials in which the Passive player received a noise blast. Ratings of moral responsibility were made on a scale from 1 (not at all responsible) to 7 (completely responsible). A sensitivity analysis determined that the behavioral sample size of 19 participants yields 80% power to detect a minimum effect size of 0.68 for a paired-sample t-test.

Second, we conducted univariate ROI analyses to determine whether ToM ROIs and 'empathic pain' ROIs, respectively, differed in their relative response magnitudes across conditions. This analysis allowed us to test our hypotheses about the distinct roles that ToM and 'empathic pain' regions might play in accounting for a behavioral asymmetry between first- and third-person moral judgments of accidental harms. The full model for each ROI predicted PSC from fixed effects of *Agent* (self vs other), *Outcome* (harm vs neutral), their interaction, and included by-subject random intercepts and slopes for the effects of agent, outcome, and their interaction. In our primary analyses, we tested the effects of experimental conditions on PSC across the entire trial time course, and then in exploratory analyses, we broke PSC down by the stimulus-bound sections of each trial: *card choice*, *video*, and *judgment*.

Finally, we conducted LMMs for each ROI to explore whether activity in any of these regions could predict the wrongness judgments participants made in a given trial, and whether this relationship depended on the agent type, the outcome type, or their interaction. Models predicted wrongness judgments from fixed-effects terms of *Agent* (self vs other), *Outcome* (harm vs neutral), averaged PSC for the entirety of the trial (averaging across *card*, *video*, and *judgment* sections), and the interactions of these effects, as well as random slopes to account for by-subject variability in PSC. In exploratory analyses, we tested whether wrongness judgments could be predicted from PSC broken down by the stimulus-bound sections of each trial: *card choice*, *video*, and *judgment*.

3. Results

3.1. Behavioral analyses

As predicted, we observed a main effect of *Agent* ($t(21.16) = -2.85$, $p = 0.01$, $d = 0.14$ [95% CI = 0.04, 0.23]), such that wrongness judgments were slightly harsher overall for Self relative to Other. We also observed a main effect of *Outcome* ($t(19.01) = 10.02$, $p < 0.0001$, $d = 2.22$ [1.78, 2.66]), such that harmful outcomes were judged worse than neutral outcomes. These effects are qualified by a marginally significant interaction between *Agent* and *Outcome* ($t(20.25) = -1.80$, $p = 0.09$), with a planned comparison showing that wrongness judgments were harsher for first-person harms (M = 3.03, S.E. = 0.18) relative to third-person harms (M = 2.85, S.E. = 0.20) ($t(20.16) = -2.59$, $p = 0.02$, $d = 0.22$ [0.05, 0.39]), but there was no significant difference in wrongness judgments between first-person (M = 1.17, S.E. = 0.07) and third-person perspectives (M = 1.13, S.E. = 0.06) for neutral outcomes ($t(18.03) = -0.82$, $p = 0.42$, $d = 0.04$ [−0.06, 0.13]) (Fig. 3, see Supplementary Fig. 2 for wrongness plots by individual).

Looking at post-scan judgments of responsibility, we find even stronger evidence for a model of 'agent-regret', with participants reporting that they felt significantly more responsible (M = 5.26, S.E. = 0.30) than they thought the other player was (M = 4.16, S.E. = 0.34) for harmful outcomes that occurred during the game ($t(18) = 2.96$, $p = 0.008$, $d = 0.79$ [0.11, 1.47]). As expected, average wrongness judgments were positively related to post-scan responsibility judgments, for both first-person and third-person harms ($r(16) = 0.53$, $p = 0.02$).

All behavioral results were also tested in a pre-registered, follow-up vignette-based study using Amazon Mechanical Turk, where, notably, the key interaction between *Agent* and *Outcome* was found to be significant (see Supplementary Study).

3.2. ROI analyses – Theory of Mind

We hypothesized that regions involved in ToM might be recruited less when participants were in the position to accidentally harm the Passive player, as opposed to merely observing another Active player as a potential agent of accidental harm. We find some evidence for this hypothesis: we observe a main effect of *Agent* that is significant in the dmPFC (dmPFC: $t(21.15) = 2.76$, $p = 0.01$) and marginal in other ToM ROIs (RTPJ: $t(16.85) = 1.82$, $p = 0.09$; LTPJ: $t(18.34) = 1.91$, $p = 0.07$; PC: $t(18.12) = 1.88$, $p = 0.08$), such that all regions show a pattern of reduced activity when participants themselves were the agent, relative to the other player.

This pattern appears to be driven primarily by significant differences in activity in all ToM ROIs during the *card choice* section, when participants were either making a card choice themselves or watching the other agent making a choice (RTPJ: $t(17.67) = 4.75$, $p = 0.0002$; LTPJ: $t(18.19) = 5.66$, $p < 0.0001$; dmPFC: $t(19.07) = 5.46$, $p < 0.0001$; PC: $t(17.86) = 2.35$, $p = 0.03$) (Fig. 4).We did not observe significant differences between conditions in univariate ToM activity during the *video* section or the *judgment* section.

3.3. ROI analyses – Empathic pain

We hypothesized that regions involved in processing participants'

own experience of pain would be recruited more when participants themselves were responsible for causing pain to someone else, relative to merely observing pain that had been caused by another agent. To test this hypothesis, we conducted ROI analyses centered on coordinates in the right anterior insula (rAI), the anterior cingulate cortex (ACC), and the left superior temporal gyrus (lSTG), all of which were active while participants received an uncomfortable noise blast that was caused by a Computer agent (Fig. 5).

Looking at PSC averaged over the entire trial, we find support for our hypothesis, observing a main effect of *Agent* in the right AI ($t(17.12) = -2.73$, $p = 0.01$) and the ACC ($t(18.36) = -5.15$, $p < 0.0001$), such that these regions were more active when participants were the agent relative to the other player. Although we find a similar trend in the left STG, this pattern does not reach significance ($t(15.69) = -1.73$, $p = 0.10$) when looking at the whole trial.

In exploratory analyses, we broke PSC down by sections of the trial to explore the possible divergence of functional roles across ROIs. During the *card choice* section, we continue to find the same main effect of *Agent* in the rAI ($t(17.17) = -2.66$, $p = 0.02$), the ACC ($t(18.70) = -6.24$, $p < 0.0001$), and now the left STG ($t(16.68) = -2.44$, $p = 0.03$), such that these regions were more active when participants were choosing the cards relative to observing the other agent making a card choice.

During the *video* section, we find a main effect of *Agent* in the right AI ($t(16.30) = -2.94$, $p = 0.009$) and the ACC ($t(18.12) = -3.84$, $p = 0.001$). We also find a significant main effect of *Outcome* in the left STG ($t(22.74) = 2.40$, $p = 0.02$), showing that this region was more sensitive to harmful outcomes relative to neutral outcomes. Although we observe a similar trend during the *video* section whereby activity in AI and ACC appear higher for harmful outcomes relative to neutral outcomes, this trend does not reach significance in either ROI (AI: $t(22.16) = 1.62$, $p = 0.12$; ACC: $t(73.60) = 1.49$, $p = 0.14$).

During the *judgment* section, we observe a marginal main effect of Agent in the right AI ($t(17.16) = -1.98$, $p = 0.06$), with higher activity for Self agent trials relative to Other agent trials. In the ACC, we now find a main effect of Outcome ($t(18.14) = 3.13$, $p = 0.006$), with harmful outcomes eliciting more ACC activity than neutral outcomes across both Self and Other agent conditions. We do not observe any significant condition differences in the left STG during this section.

3.4. Brain behavior relationships in ToM ROIs

Across ToM ROIs, we find a consistent interaction between neural activity and Outcome (harm vs neutral), such that the relationship between trial-by-trial activity and wrongness judgments was more strongly *positive* for harmful outcomes relative to neutral outcomes (RTPJ: $t(1472.68) = 3.14$, $p = 0.002$; dmPFC: $t(1362.09) = 5.56$, $p < 0.0001$; LTPJ: $t(1280.74) = 2.87$, $p = 0.004$; PC: $t(1504.54) = 1.83$, $p = 0.07$). In other words, greater neural activity in ToM ROIs was associated with harsher moral judgments of accidental harms.

3.5. Individual differences in ToM activity

Exploratory analyses reveal that this positive relationship between moral judgments and ToM activity is also reflected in individual differences in ToM activity. That is, participants who had higher ToM activity during harm trials, averaged across ROIs during the *card choice* section, 1) made harsher wrongness judgments both about their own choices ($r(16) = 0.63$, $p = 0.005$) and the choices of the other agent ($r(16) = 0.61$, $p = 0.008$), and 2) assigned more responsibility for harmful outcomes both to themselves ($r(16) = 0.54$, $p = 0.02$), and marginally to the other player ($r(16) = 0.42$, $p = 0.08$). However, asymmetries in ToM activity between first-person and third-person trials (Self – Other) do not appear to explain the observed behavioral asymmetries in wrongness judgments ($r(16) = -0.08$, $p = 0.74$) or responsibility judgments ($r(16) = -0.11$, $p = 0.66$).

When we break these analyses down for each ROI, we see some evidence that ToM ROIs may actually diverge in their functional roles for processing accidental harms. In the RTPJ, we find that higher activity during the *card choice* section is associated with harsher moral judgments of accidental harms *only* for first-person harms (*wrongness*: $r(16) = 0.55$, $p = 0.02$), but not for third-person harms (*wrongness*: $r(16) = 0.27$, $p = 0.28$). By contrast, we find the opposite pattern in the PC, such that higher activity is associated with harsher wrongness judgments about third-person harms ($r(16) = 0.70$, $p = 0.001$), and not first-person harms ($r(16) = 0.31$, $p = 0.21$). In the dmPFC, higher activity is associated with harsher wrongness judgments for both first-person harms ($r(16) = 0.59$, $p = 0.01$) and third-person harms ($r(16) = 0.53$, $p = 0.02$).

3.6. Brain behavior relationships in empathic pain ROIs

In the 'empathic pain' ROIs, we hypothesized, based on prior work on empathic blame (Patil et al., 2017), that neural activity would have a stronger positive relationship with wrongness judgments for harmful outcomes, for both first- and third-person conditions, relative to neutral outcomes.

In the right AI we did not find such a relationship between neural activity and wrongness judgments when looking at neural activity averaged over the entire trial. However, when looking just at rAI activity during the *judgment* section, we find that AI activity was indeed more positively related to wrongness judgments in harm trials than neutral trials, regardless of agent type (PSC × Outcome: $t(1457.31) = 2.49$, $p = 0.01$).

We find the opposite of this predicted interaction in the ACC, however, where higher average activity across the entire trial is more *negatively* related to moral judgments for harmful outcomes relative to neutral outcomes, regardless of agent type (PSC × Outcome: $t(1504.79) = -2.29$, $p = 0.02$). This interaction holds when looking just at ACC activity during the *judgment* section (PSC × Outcome: $t(1544.78) = -3.19$, $p = 0.001$).

In the STG, we find a three-way interaction between PSC, Agent type, and Outcome (PSC × Agent × Outcome: $t(1502.74) = -3.31$, $p = 0.001$), such that STG activity positively predicts wrongness judgments for self-caused harms, but not for other-caused harms or neutral trials. We continue to observe this three-way interaction if we look only at PSC during the *judgment* section (PSC × Agent × Outcome: $t(1502.45) = -3.53$, $p = 0.0004$).

3.7. Individual differences in empathic pain activity

Exploratory analyses of individual differences in AI activity reveal that participants with more activity in the right AI tended to hold themselves more responsible following first-person harms ($r(16) = 0.46$, $p = 0.05$), and we observe a trend in the same direction for third-person harms ($r(16) = 0.35$, $p = 0.16$). Despite finding a positive relationship between trial-by-trial wrongness ratings and activity in the right AI, we do not observe a parallel relationship between individual differences in AI activity and average wrongness judgments. Furthermore, we did not observe a negative relationship between individual differences in ACC activity and wrongness judgments or responsibility judgments, despite what was observed in trial-by-trial analyses.

In the STG, we find results that converge with trial-by-trial analyses, showing that individuals who had more STG activity when they caused harm tended to judge their actions as more wrong ($r(16) = 0.46$, $p = 0.05$) and held themselves more responsible for harm ($r(16) = 0.56$, $p = 0.02$), whereas individual differences in STG activity while participants were observing the other player causing harm did not positively correlate with wrongness judgments ($r(16) = -0.36$, $p = 0.14$) or responsibility judgments ($r(16) = 0.31$, $p = 0.21$).

In sum, the brain-behavior relationships we observe in the right AI and STG, but not the ACC, are broadly consistent with a model of empathic blame, which suggests that higher activity corresponds with harsher moral judgments about the agent of accidental harm (Patil et al.,

2017). In the STG, we see a more nuanced version of this model, in which the positive relationship between neural activity and moral judgments is present only for first-person harms, but not for third-person harms, suggesting the possibility that differences in empathy for the victim between first-person and third-person perspectives may partly account for the observed behavioral asymmetries in moral judgments between actors and observers. We find some support for this hypothesis, showing that individuals who recruited more STG activity while viewing the outcomes of their own harms versus the harms of another (Self – Other) were marginally more likely to assign more responsibility to themselves than the other agent ($r(16) = 0.42$, $p = 0.08$).

To further test the assumption that the AI and STG are involved in empathic processing, we turned to participants' post-scan scores on the Interpersonal Reactivity Index (IRI), which provides a measure of dispositional empathy broken down by multiple dissociable subcomponents (Davis, 1983). To the extent that AI and STG activity are tracking a consideration of the victim's experience of pain, we would expect individual differences in AI and STG activity to relate positively to scores on either the empathic concern (EC) or the perspective taking (PT) subcomponents, or both. An alternative possibility is that AI and STG activity are tracking with a more egoistic distress that participants might feel while witnessing another person in pain, a response that is thought to be motivationally distinct from a concern for the suffering of others (Batson, O'Quin, Fultz, Vanderplas, & Isen, 1983). If this were the case, then individual differences in AI and STG activity would be more likely to show a positive relationship with individual differences on the personal distress (PD) subscale of the IRI.

We find that participants with higher scores in PT tended to have higher activity in the right AI during the *video* section of harm trials (when the Passive player was observed experiencing pain) after participants had caused the pain themselves ($r(16) = 0.62$, $p = 0.006$), but not when the other player had caused harm ($r(16) = 0.16$, $p = 0.53$). Interestingly, we find the opposite pattern in the STG: individuals with higher scores in PT tended to have higher activity in the STG during the *video* section when they were observing harm that had been caused by the *other* player ($r(16) = 0.48$, $p = 0.04$), but not when they were observing harm that they themselves had caused ($r(16) = 0.23$, $p = 0.35$). Critically, we do not find evidence for a positive relationship between PD and individual differences in activity in either region ($p > 0.5$).

4. Discussion

The present study provides empirical support for the hypothesis that people judge the wrongness of their own accidental harms more harshly than those caused by someone else, and hold themselves more responsible for the unintended harmful outcomes of their choices. We also provide preliminary neural evidence that the observed actor-observer asymmetry is at least partly rooted in a greater sensitivity to the victim's experience of harm from the first-person perspective relative to the third-person perspective. Building on prior work on the neural correlates of empathy, we show that regions involved in the first-hand experience of pain (rAI, ACC, and lSTG) also respond during the anticipation and observation of harm to a victim, and, crucially, that these regions are more responsive when that harm is caused by participants themselves, relative to another person. We replicate work showing that neural correlates of 'empathic pain' predict moral condemnation of accidental harms (Patil et al., 2017), finding that the degree of activity in the rAI positively predicts moral judgments for harmful outcomes. We identify a novel brain-behavior relationship in the lSTG, such that this region contributes only to moral judgments of first-person harms, but not third-person harms. In support of a role for both the right AI and left STG in representing the victim's experience of harm, we find that participants who are higher in a trait measure of perspective taking (PT) tend to have higher activity in these regions while they are watching the Passive player wincing in pain.

The behavioral asymmetry we observed is somewhat surprising in light of a wide body of research on self-serving positivity biases in social and moral attributions (Mezulis et al., 2004), such as the fundamental attribution error (Hewstone, 1990; Ross, 1977). However, while the specific pattern we document is novel, it shares similarities with other behavioral asymmetries between Self and Other that have been identified in prior research on moral judgments and behavior. For instance, people are more inclined to pay to reduce others' pain than their own, and they require more compensation to increase others' pain relative to their own (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014). Similarly, people are more likely to punish fairness violations on behalf of others relative to themselves (FeldmanHall, Sokol-Hessner, Van Bavel, & Phelps, 2014). One mechanism that may connect these findings with what we observed in the present study was initially proposed by Crockett et al. (2014): that under conditions of uncertainty about their *degree* of harmful impact on others, people may "systematically err on the side of reducing others' pain at their own expense". Recent research has supported this claim, finding that while people exploit uncertainty about *whether* harm has occurred in order to behave selfishly, they behave more prosocially when they are uncertain about the *degree* of harm they may have caused (Kappes et al., 2018). In the present study, participants would not have been uncertain about whether they had caused harm, but may have experienced some uncertainty about precisely how much pain the Passive player was experiencing as a result of the noise blast. One possibility is that this impact uncertainty was somehow magnified, or made more salient, by the experience of having caused the harm oneself, leading participants to err on the side of caution by judging themselves more harshly and holding themselves more responsible. Such a difference in impact uncertainty between first-person and third-person harms may also help to explain the differences in neural activity that we observed in 'empathic pain' regions.

Although we had predicted that regions involved in representing the first-hand experience of pain would also be more sensitive to self-caused harms, we did not expect to find that the STG in particular reflected this pattern. Given that the STG is located within the auditory cortex, one possibility is that it is involved in representing the aversive auditory properties of the noise blast, both when they are experienced firsthand, and when they are simulated vicariously (Bastiaansen, Thioux, & Keysers, 2009). On this account, our results suggest that first-person harms may be differentiated from third-person harms by the degree to which they recruit lower-level embodied simulations of the victim's experience of pain, and the degree to which such information is ultimately integrated into the formation of moral judgments about harm. Moreover, our finding that individual differences in PT correlate with STG activity following third-person harms but not first-person harms may be attributed to greater individual variability during third-person harms in the degree to which embodied simulations of pain are engaged at all. We note that this embodied account of empathy is speculative: the present study does not enable us to distinguish it from alternative accounts in which judgments of wrongness do not rely on simulation per se (Mahon & Caramazza, 2008; see also Caramazza, Anzellotti, Strnad, & Lingnau, 2014 for a critique of strong claims of embodied cognition).

We were surprised to find that, although the ACC was more active for first-person relative to third-person trials, higher activity in the ACC contributed to more *lenient* wrongness judgments for harmful outcomes. This brain-behavior relationship is the opposite of what was observed in the AI, and it is also at odds with a recent study finding that higher activity in both the ACC and bilateral AI predicts *harsher* third-person wrongness judgments about the agent of accidental harm (Patil et al., 2017). One key difference between the present study and the study by Patil et al. (2017) is that, while participants in their study were merely observers who formed third-person judgments of accidental harm, participants in the present study occupied the roles of both actor and observer. This evaluative context may have altered the functional role of the ACC, which has also been implicated in error monitoring and egoistic social distress (Bastin et al., 2017; Eisenberger, Lieberman, &

Williams, 2003), and may therefore have been involved in detecting the conflict between one's moral status and the observation of having harmed someone else (FeldmanHall, Dalgleish, Evans, & Mobbs, 2015).

This explanation points to a broader possibility that the divergent patterns in the AI and ACC may reflect the implementation of guilt and shame, respectively, as different emotional responses to interpersonal harm that also have distinct behavioral signatures. Guilt is thought to arise from the empathic recognition of another person's distress (Hoffman, 1982), thereby enabling actors to remediate harm and communicate mutual concern following (often accidental) transgressions (Baumeister, Stillwell, & Heatherton, 1994). By contrast, shame is theorized as a more egoistic form of distress that inclines the ashamed person to withdraw from the situation (Tangney, Miller, Flicker, & Barlow, 1996) and does not therefore consistently motivate prosocial behavior (De Hooge, Zeelenberg, & Breugelmans, 2007). This model is broadly compatible with recent computational modeling which demonstrates that the ACC and AI can be dissociated in implementing distinct moral strategies across individuals, finding that ACC activity tracks inequity aversion and social hierarchy, whereas the AI appears to be implicated in the guilt that arises from not living up to the expectations of another person (van Baar, Chang, & Sanfey, 2019). Critically, the involvement of the AI in a process of guilt is not mutually exclusive with the view that it is a key node in an empathy network, as a number of prior studies suggest that feelings of guilt are constituted in part by a process of perspective-taking and empathy towards the victim of harm (Leith & Baumeister, 1998; Zebel, Doosje, & Spears, 2009). Furthermore, our finding that individual differences in PT correlate with AI activity following first-person harm but not third-person harm is consistent with this account, as the relationship between perspective taking towards the victim and the subsequent experience of guilt is likely to be specific to first-person harms, given the theorized role of guilt in remediating interpersonal harms in particular (Baumeister et al., 1994).

The present findings also highlight the key role that the intentionality of harm can play in modulating the actor's sensitivity to the victim's suffering. When people cause harm intentionally, they do so with the desire to cause harm, and with the prior awareness that their actions will produce a harmful outcome. Therefore, at least in cases where people harm others for instrumental purposes, they are not only motivated to rationalize the harm they have caused, thereby reducing their empathic response to the victim (Rai, Valdesolo, & Graham, 2017), but they also have the time to do so (McGraw, 1987; Tsang, 2002). By contrast, accidental transgressions often produce outcomes that are unexpected *and* undesirable, which may limit the actor's ability and motivation, respectively, to reduce their empathy for the victim. The nature of these transgressions may additionally enhance actor attentiveness to victim experiences, which prior work has shown can increase empathy, even in cases of intentional harms (Tang & Harris, 2015). Moreover, the present results provide the first evidence that, specifically for accidental harms, the agent of harm may be motivated to selectively *amplify* their empathy for victims. Although we focus solely on accidental harms in the present study, future research should directly investigate the possibility of an interaction between an actor's intentionality in causing harm and their degree of empathy for the victim who suffers.

We found mixed support for the hypothesized role of ToM regions in distinguishing first-person harms from third-person harms. On the one hand, we observed reduced activity in all ToM regions while participants themselves were making a card choice, as opposed to observing the other player making a card choice. This finding is consistent with the prediction that people are *less* likely to reason about the mental states that produced their own choices, relative to those that informed the choices of others. However, activity in ToM regions did not appear to influence moral judgments in the direction that was predicted. On the contrary, we observed a robust *positive* relationship between neural activity in all ToM ROIs and moral judgments following harmful outcomes. This finding is inconsistent with prior research on third-party moral judgments of accidental harm, which has found that greater consideration of an actor's innocent mental states, reflected in increased recruitment of ToM, correlates with more *lenient* moral judgments (Young & Saxe, 2009).

ToM activity in this task may have stemmed from a variety of potential sources beyond the direct consideration of the agent's own intentions to cause harm. One possibility is that participants are recruiting ToM to consider how the Passive player, the other Active player, or even the experimenters, are evaluating the participants' intentions, which may in turn lead participants to judge themselves more harshly in order to communicate their innocence. Alternatively, participants may be reasoning about mental states other than intent that are nevertheless relevant to forming moral judgments about the harm that was caused. For instance, participants may recruit ToM to reason about the agent's *effort* (e.g. "how much thought did she put into deciphering the pattern?") or *ability* (e.g. "I should have learned this pattern by now"). This possibility is consistent with the 'Path Model' of blame, which suggests that, once an event is judged to be unintentional, agents are subsequently evaluated more harshly in proportion to their capacity to prevent harm (Malle, Guglielmo, & Monroe, 2014). We recognize that these proposals are speculative, and encourage future work to test them directly (though see Supplementary Materials for exploratory analyses that provide a preliminary test of the *preventability* hypothesis).

A strength of the present study is that it provides a context in which "real" accidental harm can be caused and observed. Participants were led to believe that they were actually causing minor discomfort to other people, thereby side-stepping the potential concern that people might only judge themselves more harshly in hypothetical scenarios, when the perceived cost of self-condemnation could be relatively low (FeldmanHall et al., 2012). At the same time, it is unclear whether these behavioral and neural asymmetries between first-person and third-person perspectives would persist in cases of much more severe accidental harm. For instance, how might people respond differently if they were instead found responsible for accidentally running over another person with their car? Although the opening anecdote in the introduction might suggest that people would judge themselves more harshly even in these extreme cases, it remains an open question as to whether this pattern would actually persist for more severe instances of harm. However, obvious ethical concerns limit the experimental investigation of this possibility.

The observed behavioral and neural patterns may also be moderated by the particular type of moral judgment that people are tasked with making. Prior work has shown that people are sensitive to different features of a moral context depending on the category of moral judgment they are making, with wrongness and permissibility judgments showing the greatest sensitivity to intent, and blame and punishment judgments more sensitive to assessments of causal responsibility and the severity of harm (Cushman, 2008). This distinction may help to explain why responsibility judgments, which are sensitive to the severity of harm (Robbennolt, 2000), showed a larger asymmetry between first- and third-person accidents than wrongness judgments. That is, to the extent that differences in the assessment of harm contribute to the asymmetry between moral judgments of first-person and third-person accidents, as our neural findings suggest, we might expect that judgments like responsibility and blame would show larger behavioral asymmetries. Furthermore, a recent study found that regions in the empathy network (AI and ACC) responded more when participants made blame judgments about accidental harms relative to permissibility judgments (Patil et al., 2017). Along these lines, we might also predict that the observed neural asymmetries in 'empathic pain' regions would be more pronounced if participants were asked to evaluate first-person and third-person blameworthiness or responsibility on a trial-by-trial basis.

Why would people be more sensitive to the harmful consequences of their own accidents, relative to the accidents of others? One possibility is that people are, by default, more sensitive to the consequences of their

own actions because such a bias would be useful for learning adaptively. Alternatively, research on construal theory might lead us to predict that first-person harms, which are more psychologically proximal (i.e., less abstract) than third-person harms, are subsequently more likely to generate a greater focus on the consequences of those actions (e.g. harm) relative to the causes (e.g. intent) (Rim, Hansen, & Trope, 2013). Finally, participants may judge themselves more harshly in anticipation of, or as a way of mitigating, the victim's negative evaluation of them. For instance, prior work has shown that victims overestimate the intentionality of transgressions, leading them to underestimate how much transgressors desire forgiveness (Adams & Inesi, 2016). One intriguing possibility is that actors might judge themselves more harshly following an accidental transgression in order to compensate for the potential uncertainty that victims could have about the actor's innocent intentions. This explanation is consistent with Bernard Williams' claim, in reference to a driver who has accidentally run over a child, that "people will try, in comforting him, to move the driver from this state of feeling, move him indeed from where he is to something more like the place of a spectator, but it is important that this is seen as something that should *need* to be done, and indeed some doubt would be felt about a driver who too blandly or readily moved to that position." (Williams, 1981). Future work could join these two approaches by testing how victims respond to actors who judge their own accidental harm as severely as an observer, or more severely. The framework outlined above also makes the prediction that the behavioral asymmetry between actors and observers would be largest in a context where actors' judgments are being observed by the victim and nearby spectators, relative to a more private context.

4.1. Limitations

Finally, we turn to several limitations of the present work. First, we acknowledge that a key limitation of the present study is its small sample size. While behavioral effects were replicated in a pre-registered vignette-based experiment (see Supplementary Study), neural analyses were restricted to the data collected from scanner participants. We note though that an advantage of the present approach is that general conclusions are drawn not only from neural data, but also from behavioral data, as well as the relationship between the two. Moreover, recent fMRI research has relied on similarly sized samples (Gaesser et al., 2019; Niemi et al., 2018; Tsoi et al., 2016). Nevertheless, we emphasize that future work should prioritize recruiting larger fMRI populations, which could increase model sensitivity to additional neural effects with potential relevance to the brain-behavior relationships observed in the present work.

Second, due to hardware limitations (i.e. MR-safe button box), wrongness judgments in the scanner were measured on a 4-point Likert scale, which may have limited our ability to detect behavioral patterns. Crucially, though, an online follow-up study provided additional evidence for key wrongness effects using a 7-point Likert scale (see Supplementary Study). These behavioral findings should be further confirmed using an in-person game paradigm.

Third, we acknowledge that, while the difference between first- and third-person wrongness ratings in cases of accidental harm was robust, the effect size was modest. Interestingly, the effect is larger in between-subjects analyses of the supplemental study, using data from the first two conditions that each participant saw (i.e., with *Agent* as a between-subjects variable). This finding suggests that participants may feel particularly inclined in a within-subjects paradigm to maintain consistent ratings across Self and Other *Agent* conditions. Future work could explore ways to counteract this tendency by, for instance, incorporating *Agent* as a true between-subjects condition, or utilizing an unmarked slider scale for ratings.

5. Conclusion

In contrast with a wide body of social psychological work demonstrating self-serving biases in social and moral judgment, we have found evidence that people judge their own accidental harms more harshly than the same harms committed by another person. The present results also diverge from prior work on empathy and dehumanization by suggesting that empathy for the victim of harm can be enhanced in cases where one is morally responsible for harm that was unintended. Taken together, these results suggest that accidental harms are a unique context for clarifying theories about the relationship between responsibility for harm, empathy for the victim of harm, and moral judgments about the self.

Open Practices

Data, analysis scripts, experimental materials, and a pre-registration of the primary analyses are all available on OSF (https://osf.io/3hq89/). Raw fMRI data is available in BIDS compatible format on OpenNeuro (https://openneuro.org/datasets/ds002222).

Acknowledgments

The authors thank Grace Elliott, Julia Napoli, and Robert Wright for their help in running the study; Minjae Kim, Ryan McManus, BoKyung Park, Justin Martin, Gordon Kraft-Todd, Bertram Malle, and Oriel FeldmanHall for their helpful feedback and comments. This work was supported by the John Templeton Foundation (61061 to L.Y.); the National Science Foundation (1627157 to L.Y.); the Boston College Virtue Project (to L.Y.); and a shared instrumentation grant from the National Institutes of Health (S10-OD021569 to MIT).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2021.104102.

References

Adams, G. S., & Inesi, M. E. (2016). Impediments to forgiveness: Victim and transgressor attributions of intent and guilt. *Journal of Personality and Social Psychology, 111*(6), 866.

Ames, D. L., & Fiske, S. T. (2015). Perceived intent motivates people to magnify observed harms. *Proceedings of the National Academy of Sciences, 112*(12), 3599–3605.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis, 12*(1), 26–41.

Azevedo, R. T., Macaluso, E., Avenanti, A., Santangelo, V., Cazzato, V., & Aglioti, S. M. (2013). Their pain is not our pain: Brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Human Brain Mapping, 34*(12), 3168–3181.

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications, 10*(1), 1483.

Bastiaansen, J. A., Thioux, M., & Keysers, C. (2009). Evidence for mirror systems in emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2391–2404.

Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., … Jerbi, K. (2017). Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cerebral Cortex (New York, NY: 1991), 27*(2), 1545–1557.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Batson, C. D., O'Quin, K., Fultz, J., Vanderplas, M., & Isen, A. M. (1983). Influence of self-reported distress and empathy on egoistic versus altruistic motivation to help. *Journal of Personality and Social Psychology, 45*(3), 706.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin, 115*(2), 243.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage, 37*(1), 90–101.

Bem, D. J. (1972). Self-perception theory. In , *Vol. 6. Advances in experimental social psychology* (pp. 1–62). Academic Press.

Botvinick, M., Jha, A. P., Bylsma, L. M., Fabian, S. A., Solomon, P. E., & Prkachin, K. M. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *NeuroImage, 25*, 312–319.
Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1).
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 and Function, 217*(4), 783–796.
Caramazza, A., Anzellotti, S., Strnad, L., & Lingnau, A. (2014). Embodied cognition and mirror neurons: A critical assessment. *Annual Review of Neuroscience, 37*, 1–15.
Castano, E., & Giner-Sorolla, R. (2006). Not quite human: Infrahumanization in response to collective responsibility for intergroup killing. *Journal of Personality and Social Psychology, 90*(5), 804.
Cheng, Y., Chen, C., Lin, C. P., Chou, K. H., & Decety, J. (2010). Love hurts: An fMRI study. *Neuroimage, 51*(2), 923–929.
Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences, 111*(48), 17320–17325.
Crockett, M. J., & Lockwood, P. L. (2018). Extraordinary altruism and transcending the self. *Trends in Cognitive Sciences, 22*(12), 1071–1073.
Cui, F., Abdelgabar, A. R., Keysers, C., & Gazzola, V. (2015). Responsibility modulates pain-matrix activation elicited by the expressions of others in pain. *NeuroImage, 114*, 371–378.
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.
Cushman, F. (2018). Rationalization is rational. *Behavioral and Brain Sciences*, 1–69.
Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113.
De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion, 21*(5), 1025–1042.
Decety, J., Echols, S., & Correll, J. (2010). The blame game: The effect of responsibility and social stigma on empathy for pain. *Journal of Cognitive Neuroscience, 22*(5), 985–997.
Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *Neuroimage, 55*(2), 705–712.
Dungan, J. A., & Young, L. (2019). Asking "why?" Enhances theory of mind when evaluating harm but not purity violations. *Social Cognitive and Affective Neuroscience, 14*(7), 699–708.
Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302*(5643), 290–292.
Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Oya, H. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods, 16*(1), 111.
FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage, 105*, 347–356.
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), 434–441.
FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Communications, 5*, 5306.
Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109–128.
Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage, 47*, S102.
Gaesser, B., Hirschfeld-Kroen, J., Wasserman, E. A., Horn, M., & Young, L. (2019). A role for the medial temporal lobe subsystem in guiding prosociality: The effect of episodic processes on willingness to help others. *Social Cognitive and Affective Neuroscience, 14*(4), 397–410.
Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics, 5*, 13.
Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology. General, 143*(4), 1600–1615.
Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry, 23*(2), 206–215.
Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124.
Gregory, A. (2017, September 11). *The Sorrow and Shame of the Accidental Killer*. The New Yorker. Retrieved from https://www.newyorker.com/magazine/2017/09/18.
Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage, 48*(1), 63–72.
Gweon, H., Young, L., & Saxe, R. R. (2011). Theory of mind for you, and for me: Behavioral and neural similarities and differences in thinking about beliefs of the self and other. In *Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 33, No. 33)*.
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*, 149–160.
Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology, 20*(4), 311–335.
Hoffman, M. L. (1982). Development of prosocial motivation: Empathy and guilt. In *The development of prosocial behavior* (pp. 281–313). Academic Press.
Jacobson, D. (2013). Regret, agency, and error. *Oxford Studies in Agency and Responsibility, 1*, 95–125.
Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage, 17*(2), 825–841.
Kappes, A., Nussberger, A. M., Faber, N. S., Kahane, G., Savulescu, J., & Crockett, M. J. (2018). Uncertainty about the impact of social decisions increases prosocial behaviour. *Nature Human Behaviour, 2*(8), 573.
Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. *Cerebral Cortex, 31*(2), 884–898.
Koban, L., Corradi-Dell'Acqua, C., & Vuilleumier, P. (2013). Integration of error agency and representation of others' pain in the anterior insula. *Journal of Cognitive Neuroscience, 25*(2), 258–272.
Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences, 39*(8), 499–501.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models [Computer software manual]. Retrieved from http://CRAN.R-project.org/packagelmerTest.
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), 2492–2502.
Lee, J. J., Hardin, A. E., Parmar, B., & Gino, F. (2019). The interpersonal costs of dishonesty: How dishonest behavior reduces individuals' ability to read others' emotions. *Journal of Experimental Psychology: General, 148*(9), 1557–1574. https://doi.org/10.1037/xge0000639.
Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin, 36*(8), 1115–1129.
Leith, K. P., & Baumeister, R. F. (1998). Empathy, shame, guilt, and narratives of interpersonal conflicts: Guilt-prone people are better at perspective taking. *Journal of Personality, 66*(1), 1–37.
MacKenzie, J. (2017). Agent-regret and the social practice of moral luck. *Res Philosophica, 94*(1), 95–117.
Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris, 102*(1–3), 59–70.
Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.
Martin, J. W., & Cushman, F. (2016). The adaptive logic of moral luck. In *The Blackwell companion to experimental philosophy* (pp. 190–202).
McGraw, K. M. (1987). Guilt following transgression: An attribution of responsibility approach. *Journal of Personality and Social Psychology, 53*(2), 247.
Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin, 130*(5), 711.
Niemi, L., Wasserman, E., & Young, L. (2018). The behavioral and neural signatures of distinct conceptions of fairness. *Social Neuroscience, 13*(4), 399–415.
Palermo, S., Benedetti, F., Costa, T., & Amanzio, M. (2015). Pain anticipation: An activation likelihood estimation meta-analysis of brain imaging studies. *Human Brain Mapping, 36*(5), 1648–1661.
Park, B., & Young, L. (2020). An association between biased impression updating and relationship facilitation: A behavioral and fMRI investigation. *Journal of Experimental Social Psychology, 87*, 103916.
Patil, I., Calò, M., Fornasier, F., Cushman, F., & Silani, G. (2017). The behavioral and neural basis of empathic blame. *Scientific Reports, 7*(1), 5200.
Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage, 84*, 320–341.
Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences, 114*(32), 8511–8516.
Rim, S., Hansen, J., & Trope, Y. (2013). What happens why? Psychological distance and focusing on causes versus consequences of events. *Journal of Personality and Social Psychology, 104*(3), 457.
Robbennolt, J. K. (2000). Outcome severity and judgments of "responsibility": A meta-analytic review 1. *Journal of Applied Social Psychology, 30*(12), 2575–2609.
Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *, Vol. 10. Advances in experimental social psychology* (pp. 173–220). Academic Press.
Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*(4), 1835–1842.
Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences, 1156*(1), 81–96.
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.
Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart, J., Jr (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Cognitive Brain Research, 17*(1), 75–82.
Sussman, D. (2018). Is agent-regret rational? *Ethics, 128*(4), 788–808.

Tang, S., & Harris, L. (2015). Construing a transgression as a moral or a value violation impacts other versus self-dehumanisation. *Revue Internationale de Psychologie Sociale, 28*(1), 95–123.

Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology, 70*(6), 1256.

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (2020). Theory of Mind network activity is associated with metaethical judgment: An item analysis. *Neuropsychologia*, 107475.

Tsang, J. A. (2002). Moral rationalization and the integration of situational factors and psychological processes in immoral behavior. *Review of General Psychology, 6*(1), 25–50.

Tsoi, L., Dungan, J., Waytz, A., & Young, L. (2016). Distinct neural patterns of social cognition for cooperation versus competition. *NeuroImage, 137*, 86–96.

Tsoi, L., Dungan, J. A., Chakroff, A., & Young, L. L. (2018). Neural substrates for moral judgments of psychological versus physical harm. *Social Cognitive and Affective Neuroscience, 13*(5), 460–470.

Tusche, A., Böckler, A., Kanske, P., Trautwein, F. M., & Singer, T. (2016). Decoding the charitable brain: Empathy, perspective taking, and attention shifts differentially predict altruistic giving. *Journal of Neuroscience, 36*(17), 4719–4732.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging, 29*(6), 1310.

Williams, B. (1981). *Moral luck: Philosophical papers 1973–1980*. Cambridge University Press.

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, 107*(15), 6753–6758.

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

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.

Yu, H., Hu, J., Hu, L., & Zhou, X. (2014). The voice of conscience: Neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience, 9*(8), 1150–1158.

Zebel, S., Doosje, B., & Spears, R. (2009). How perspective-taking helps and hinders group-based guilt as a function of group identification. *Group Processes & Intergroup Relations, 12*(1), 61–78.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging, 20*(1), 45–57.