Journal of Cognitive Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pecp21

Alexithymia increases moral acceptability of accidental harms

Indrajeet Patil[a] & Giorgia Silani[a]
[a] Neuroscience Sector, Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy
Published online: 18 Jun 2014.

To cite this article: Indrajeet Patil & Giorgia Silani (2014) Alexithymia increases moral acceptability of accidental harms, Journal of Cognitive Psychology, 26:5, 597-614, DOI: 10.1080/20445911.2014.929137

To link to this article: http://dx.doi.org/10.1080/20445911.2014.929137

Journal of Cognitive Psychology, 2014
Vol. 26, No. 5, 597–614, http://dx.doi.org/10.1080/20445911.2014.929137

Alexithymia increases moral acceptability of accidental harms

Indrajeet Patil and Giorgia Silani

Neuroscience Sector, Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy

Previous research shows that when people judge moral acceptability of others' harmful behaviour, they not only take into account information about the consequences of the act but also an actor's belief while carrying out the act. A two-process model has been proposed to account for this pattern of moral judgements and posits: (1) a causal process that detects the presence of a harmful outcome and is motivated by empathic aversion stemming from victim suffering; (2) a mental state-based process that attributes beliefs, desires, intentions, etc. to the agent in question and is motivated by imagining personally carrying out harmful actions. One prediction of this model would be that personality traits associated with empathy deficits would find accidental harms more acceptable not because they focus on innocent intentions but because they have reduced concern for the victim's well-being. In this study, we show that one such personality trait, viz. alexithymia, indeed exhibits the predicted pattern and this increased acceptability of accidental harm in alexithymia is mediated by reduced dispositional empathic concern. Results attest to the validity of two-process model of intent-based moral judgements and emphasise key role affective empathy plays in harm-based moral judgements.

Keywords: Alexithymia; Belief; Empathy; Intention; Moral judgement.

1.1. Role of mental state and outcome in moral judgement

Moral judgement entails judging others' actions on the dimension of right and wrong, but this requires not only that we focus on the consequences of these actions but also the intent with which they were accomplished. A number of recent studies indeed demonstrate that mental state information (e.g. desire, belief, intention) is one of the crucial inputs into moral decision-making (for a review, see Young & Tsoi, 2013). This becomes especially conspicuous when beliefs are incongruent with the outcome. Consider the cases where the agent unintentionally ends up hurting someone (accidental harm: neutral belief, negative outcome) or cases of incompetent harm-doers who think they would harm someone, but fail in doing so (attempted harm: negative belief, neutral outcome). In such cases, moral judgements are influenced more by information about mental state of the agent than outcomes. People tend to judge accidental harms more leniently than intentional harms, both of which contain a negative outcome but only intentional harms possess negative intent (Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007). On the other hand, attempted harms are judged more severely than the neutral condition although both actions lead to neutral outcome but only the attempted harm has negative intent (Cushman, 2008; Young et al., 2007). In fact, attempted harms are judged even more severely

Correspondence should be addressed to Indrajeet Patil, Neuroscience Sector, International School for Advanced Studies (SISSA-ISAS), Via Bonomea 265, Trieste 34136, Italy. E-mail: ipatil@sissa.it

The authors gratefully acknowledge Giulia Elli for providing Italian translation of scenarios and three anonymous reviewers for their insightful comments on the earlier version of this manuscript.

than accidental harms, harmful outcome notwithstanding (Cushman, 2008).

Judgments about neutral and intentional harms tend to be straightforward, since the belief information tends to be congruent with the outcome information in such cases. But even when beliefs and outcomes are incongruent with each other, there are different ways that this incongruence can behaviourally manifest itself depending on the valence of the conflicting belief and outcome. Evidence from developmental psychology shows that children (even preverbal infants) start condemning negative intent that does not result in negative outcome (attempted harm) before they start exculpating agents who accidently produce negative outcome with benign intent (Baird & Astington, 2004; Cushman, Sheketoff, Wharton, & Carey, 2013; Hamlin, 2013); for example, children find the agent who accidently misdirects a traveller more blameworthy ("more naughty") than the agent who attempts to misdirect the traveller but mistakenly shows her the correct path (Baird & Astington, 2004). Thus, it is clear that there is something especially difficult in forgiving agents who produce negative outcome accidently. The two-process model explains how the processes underlying exculpation of accidental harms and condemnation of attempted harms are dissimilar, thus explaining why these two cases elicit different pattern of responses.

1.2. Two-process model for intent-based moral judgement

The two-process model for intent-based moral judgement (Cushman, 2013; Cushman et al., 2013; Young et al., 2007) proposes an arrangement where competitive interaction occurs between two processes (Figure 1 and Table 1): (1) a causal reasoning process that detects a harmful outcome and attributes condemnation to the causally responsible agent; (2) a mental state-based reasoning processes that represent the mental states (beliefs, intentions, etc.) with which the agent acted. It is important to note that in order to assess agent's behaviour on moral dimension (good/bad, right/wrong), these two processes cannot be restricted purely to information processing but also need to have some affective content that motivates competing behavioural tendencies (condemnation or exculpation) supported by respective processes (Hauser, 2006; Huebner, Dwyer, & Hauser, 2009) and indeed there is overwhelming evidence to support the causal role of emotions in moral condemnation of harmful behaviour (for a review, see Avramova & Inbar, 2013). Consequently, both of these processes possess some affective content which motivates them, but the psychological origin of this affect differs for each process. The emotional aversion that supports the causal reasoning-based condemnation results from aversion to harmful outcomes ("outcome aversion"), viz. mental simulation of distress in the *victim* affected by the act (Hoffman, 2001; Pizarro, 2000). On the other hand, the source of negative affect that motivates the mental state reasoning process stems from aversion to the harmful actions independent of the outcomes ("action aversion") because such actions have been associated with aversive reinforcement history (Blair, 2007, 2013; Cushman, 2013; Crockett, 2013). This negative affect can stem from concrete sensorimotor and perceptual representations of harmful actions (e.g. pushing, slapping; Cushman, Gray, Gaffey, & Mendes, 2012) or abstract categorical descriptions of such actions (e.g. "agent A intends to burn X"; Miller, Hannikainen, & Cushman, 2014) without further considering whether these actions result in harm. Thus, when people judge harmful third-party behaviour, they engage in a first-person evaluative simulation (Miller & Cushman, 2013) in which they imagine themselves in the stead of the *agent* and think about carrying out the harmful acts and experience a pang of negative emotions. Therefore, the emotional aversion stems from the sheer thought of performing a harmful action (Blair, 1995).

To sum up, emotion acts as a gain antecedent to moral judgements (Decety & Cacioppo, 2012): It amplifies the contribution of the cognitive process it supports to the ultimate output which represents the final moral judgement. More precisely, the negative affect arising from outcome (or empathic) aversion increases the moral condemnation based on the causal role of agent in producing the negative outcome, whereas action aversion-based negative affect amplifies the moral condemnation of the harmful intent. Thus, underlying cognitive processes determine whether the agent needs to be condemned or exculpated depending on the presence or absence of negative outcome or intent, whereas the negative affect supporting these inputs decide the direction (causal information or mental state information) in which the final moral judgement will be swayed (Figure 1 and Table 1). It needs to be underscored that integration of outputs from these two systems

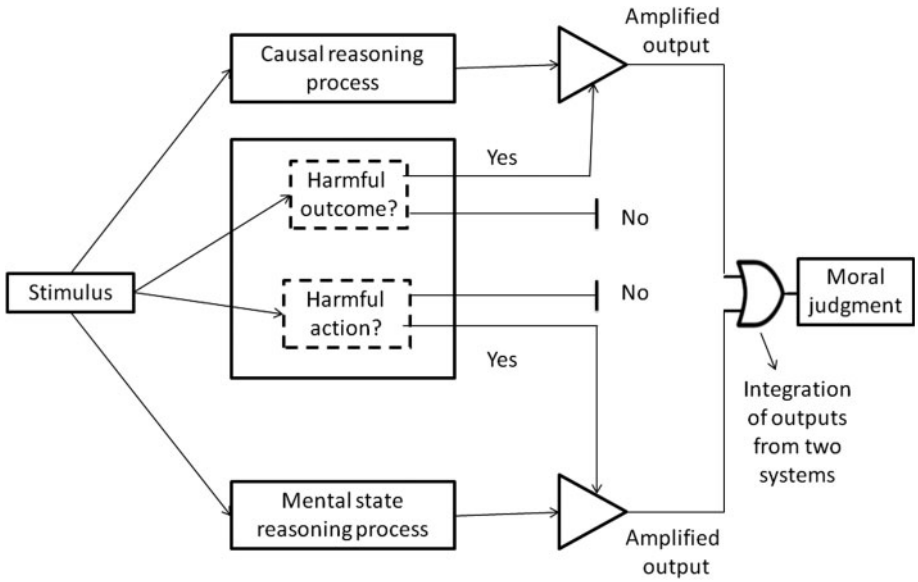

Figure 1. Schematic depiction of the two-process model for intent-based moral judgement. A causal reasoning process determines the causal role of the agent in the presence of harmful outcome and negative affect arising from empathy-based outcome aversion motivates condemnation of the agent based on his/her causal involvement in bringing about the negative outcome. A mental state reasoning process determines the beliefs on which the agent acted and negative affect based in action aversion drives the condemnation of the agent if negative intent is detected. Emotionally motivated outputs from both of these processes are finally integrated and ultimate moral judgement about wrongness/permissibility of agent's behaviour is computed.

is required if these outputs are to influence the final moral judgement and failure in integration can cause people to rely more on causal heuristics (Buon, Jacob, Loissel, & Dupoux, 2013), e.g. patients who have had complete and partial callosotomy surgery tend to rely more on the outcome information because information about mental states of the agent is not properly integrated while making the moral judgement (Miller et al., 2010).

1.3. Attempted versus accidental harm and evidence for the two-process model

Attempted harm and action aversion. According to two-process model, since cases of attempted harm do not involve any emotionally salient negative outcome, the causal processes remain silent in the absence of harmful outcome and there is no negative affect stemming from outcome aversion to motivate this process. Accordingly, when people

TABLE 1
Two-process model of intent-based moral judgement and its explanation for moral judgements about different cases. The model proposes competitive interaction between a causal reasoning-based process and a mental state reasoning-based process, which are, respectively, motivated by negative affect deriving from outcome and action aversion

Case	*Causal process output*	*Causal reasoning output motivated by which aversion?*	*Mental state reasoning output*	*Mental state reasoning output motivated by which aversion?*	*Cognitive conflict between systems?*	*Result*
Accidental	Condemn	Outcome	Exculpate	–	Yes	Mitigated exculpation
Attempted	–	–	Condemn	Action	No	Mitigated condemnation
Intentional	Condemn	Outcome	Condemn	Action	No	Unmitigated condemnation
Neutral	Exculpate	–	Exculpate	–	No	Unmitigated exculpation

face attempted harm scenarios, they do not show increased activity in the brain regions involved in detecting cognitive conflict due to predominant input from mental state reasoning process in the absence of robust input from causal processes (Young et al., 2007). Condemnation of attempted harm thus relies exclusively on harmful intent as inferred by the mental state processes which operate unabated in the absence of causal processes. Indeed, right temporoparietal junction (rTPJ, a brain region involved in encoding mental states with representational content, like beliefs, intentions, etc.; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014) exhibits highest activity when people reflect over attempted harm scenario as compared to any other type of harm scenario (Young et al., 2007; Young & Saxe, 2008) since deeming attempted harms requires robust representation of the malicious intention that agent harbours. Disrupting activity in rTPJ using non-invasive brain stimulation techniques interrupts the mental state processes that support condemnation of attempted harm and renders people more likely to deem attempted harm scenarios morally more permissible (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010).

Emotional aversion that motivates and subserves the mental state processes stems from aversion to harmful actions, i.e. aversion that stems from imagining putting oneself in the agent's shoes and carrying out the harmful action (Cushman et al., 2012; Miller & Cushman, 2013; Miller et al., 2014). This aversion results from evaluations of contextualised actions on the basis of their reinforcement histories that have come to associate negative value with the simulated actions and leads to behavioural suppression when people mentally simulate engaging in such acts (Crockett, 2013; Cushman, 2013). It needs to be stressed that individuals need not engage in such harmful behaviours first-hand for reinforcement learning to associate negative values to canonically harmful actions (not all of us learned that hitting is bad by personally hitting someone), these can also be learned via observation and simulation (Miller & Cushman, 2013).

The ventromedial prefrontal cortex (vmPFC) is an important brain region that holds information about negative reinforcement expectancy values associated with harmful actions learnt from prior reinforcement learning (Blair, 2007, 2013). Therefore, higher activity in this region would be expected to lead to higher aversion to harmful actions, which in its turn would boost contribution of output from the mental state reasoning process and lead to more severe condemnation of attempted harms. This is exactly what one neuroimaging study found (Young & Saxe, 2009b): More severe condemnation of attempted harm was positively correlated with activity in vmPFC. Decety and Cacioppo (2012) also show that intentional harms (which contain harmful intent) lead to increased activity in vmPFC as compared to accidental harms (which do not contain harmful intent), although both of these actions feature negative outcome. Also, imagining being the agent who intentionally harms others or exhibits aggressive behaviour towards a violent assailant elicits increased activity in vmPFC (Decety & Porges, 2011; King, Blair, Mitchell, Dolan, & Burgess, 2006). Importantly, people with damage to vmPFC tend to provide more lenient judgements for attempted harm cases with harmful intent (Ciaramelli, Braghittoni, & di Pellegrino, 2012; Young, Bechara, et al., 2010), arguably due to reduced action aversion in the absence of reinforcement expectancy information stored in vmPFC. These studies underscore the importance of mental state reasoning in the representation of harmful intent and motivating role of action aversion for amplifying contribution of this process to the final moral wrongness/permissibility judgement.

Accidental harm and outcome aversion. Accidental harms lead to negative outcome which intuitively invokes the causal processes (Buon, Jacob, et al., 2013) that holds the agent responsible for the negative outcome, whereas the empathic aversion to the victim's suffering motivates moral condemnation of this agent. But the mental state process countervails this response by exculpating the agent based on the innocent intentions. This leads to a cognitive conflict between these two processes, as shown by increased activity in the brain regions associated with detection of cognitive conflict (right inferior parietal cortex, bilateral middle frontal gyrus and bilateral anterior cingulate sulcus) when people contemplate over accidental harm cases (Young et al., 2007). This competitive account which underscores inverse relation between processing of salient negative outcome and processing of belief-related information is also supported by neuroimaging data which shows reduced activity in rTPJ when there is negative outcome present as compared to when it is absent (Young & Saxe, 2008, 2009a). Additionally, people show increased magnitude of neural activity (Decety & Cacioppo, 2012; Koster-Hale, Saxe, Dungan, & Young, 2013)

and different spatial pattern of neural activity (Koster-Hale et al., 2013) in rTPJ when they face accidental versus intentional harm cases. Also, the more the magnitude of this activity (Young & Saxe, 2009b) and the more pronounced difference in the spatial pattern of neural activity in rTPJ (Koster-Hale et al., 2013), the more people tend to exonerate accidental harm-doers by effectively counteracting the causal process-based condemnation by relying on robust mental state representations.

There is also evidence to implicate empathy-based negative affect that motivates causal processes that condemn the agent for his/her causal involvement in bringing about the negative outcome, neutral intention notwithstanding. When people evaluate situations involving negative outcome (intentional and accidental harm cases), they show increased activity in regions involved in empathy (e.g. insula) and show greater looking time (measured with eye-tracking) at the victim of harmful actions rather than the perpetrator (Decety, Michalska, & Kinzler, 2012). Individuals who are genetically predisposed to be more empathic due to variation of oxytocin receptor gene (Wu, Li, & Su, 2012) tend to be less forgiving of accidental harms (Walter et al., 2012), probably due to predominance of empathy-driven causal process in the final moral judgement. Psychopaths, who are well known to have deficient empathic skills (see, e.g., Marsh, in press), tend to judge accidental harms more leniently, arguably owing to reduced empathic aversion to victim distress (Young, Koenigs, Kruepke, & Newman, 2012). On the other hand, autistic people tend to judge accidental harms more harshly (Buon, Dupoux, et al., 2013; Koster-Hale, et al., 2013; Moran et al., 2011) due to weakened input from mental state processes, which are impaired in this population due to problems associated with the capacity to attribute mental states or theory of mind (Baron-Cohen, Leslie, & Frith, 1985). In other words, both psychopathic and autistic people have a reduced cognitive conflict when they judge accidental harms as compared to normal population but for diametrically opposite reasons: The former due to reduced empathic aversion to harmful outcomes that motivate causal processes, whereas the latter due to deficient theory of mind skills that impel mental state processes (cf. intentional and phenomenological stances; Theriault & Young, 2014). This underscores the important prediction of this model that people can judge accidental harms more leniently because: (1) they engage in enhanced mental state reasoning (as indexed by activity in rTPJ) or (2) they have reduced aversion to harmful outcomes due to deficient empathic response and thus causal processes lose their motivational impetus based on outcome aversion that usually demands that agent be held causally responsible for the harmful outcome and be condemned for the same. Although there is evidence to support the former route whereby people forgive accidental harm-doers based on enhanced mental state reasoning (Koster-Hale et al., 2013; Young & Saxe, 2009b), the latter route remains unexplored. Personality traits with known empathy deficits thus provide an opportunity to explore this path not taken. We now turn to one such personality trait.

1.4. Alexithymia and empathy deficits

Alexithymia is a sub-clinical, dimensional personality construct that is characterised by reduced capacity to experience emotions, absence of tendency to reflect on one's own emotions, difficulty in identifying feelings and bodily sensations associated with emotional arousal and describing these feelings to others, despite a basic awareness of bodily arousal and sensation (Lane, Ahern, Schwartz, & Kaszniak, 1997; Nemiah, Freyberger, & Sifneos, 1976). Given the importance of proper emotional processing for successful social behaviour, elevated levels of alexithymia are associated with impaired performance on a number of important social cognition skills, e.g. emotion recognition, emotion regulation, empathy, processing of affective language and prosody (for more extensive overview of these deficits, see Bird & Cook, 2013; Kano & Fukudo, 2013; Moriguchi & Komaki, 2013; van der Velde et al., 2013; Wingbermühle, Theunissen, Verhoeven, Kessels, & Egger, 2012). Of interest to the current study are problems associated with empathy in alexithymic personalities.

Empathy is composed of two different and equally important components: (1) cognitive empathy involves understanding others' emotional states by forming abstract mental representations of these states while maintaining self-other distinction; (2) affective empathy involves experiencing these emotional states (de Vignemont & Singer, 2006). Recent work in social neuroscience supports the shared network model of empathy which posits that the same brain regions that are involved in mapping the body's physiological states that inform us our subjective feelings states

are also involved when we try to predict the feeling states of others (for meta-analytic evidence, see Lamm, Decety, & Singer, 2011). In other words, when people try to understand emotional states of others and experience these states vicariously, they are guided by their own internally generated affective states (Hooker, Verosky, Germine, Knight, & D'Esposito, 2008). But this very ability to interocept on one's own affective states is compromised in alexithymia (e.g., Silani et al., 2008). Thus, shared network model of empathy predicts that impaired emotional awareness in the alexithymia would result in empathic deficits.

Indeed, high level of alexithymia is associated with reduced activity in the neural network supporting empathy while empathising with others who are experiencing pain (Bird et al., 2010; Moriguchi et al., 2007). They also report to feel less distress at others' suffering and are less motivated to act altruistically to relieve another's distress (FeldmanHall, Dalgleish, & Mobbs, 2012). Various self-report measures of empathy, e.g. interpersonal reactivity index (IRI: M. H. Davis, 1980, 1983), show that people with alexithymia report having less empathic concern (EC) for others and a reduced tendency for perspective taking (PT) in both community and psychiatric/clinical populations (for a review, see Bird & Cook, 2013; also see Grynberg, Luminet, Corneille, Grezes, & Berthoz, 2010; Guttman & Laporte, 2002; Jonason & Krause, 2013). They also show a reduced affective empathic response to emotional facial expressions (Lockwood, Bird, Bridge, & Viding, 2013). Thus, there is an overwhelming evidence that trait alexithymia is characterised by poor ability to understand what others feel (cognitive) and experience or share others' emotional states (affective).

1.5. Current study

The discussion so far explains the importance of a proper empathic response that drives emotional aversion to negative outcomes and it also elucidates impairment in this empathic ability in alexithymia. Naturally, the question arises as to how people with alexithymia would judge accidental harms where there is a conflict between negative outcome and innocent intention. The two-process model for intent-based moral judgement would predict that alexithymic personalities would find accidental harms more acceptable due to reduced EC for and/or reduced PT of the victim, which would pre-empt negative affect that acts as impetus for moral condemnation of the agent. Thus, we make three key predictions for the current study:

(1) Higher level of trait alexithymia will predict reduced EC, PT (subscales of IRI) and increased acceptability ratings for accidental harm.
(2) Reduced EC and PT will predict higher moral acceptability ratings for accidental harm.
(3) EC and/or PT will mediate the relation between trait alexithymia and acceptability of accidental harm.

Although there is negative outcome present even in the case of intentional harm, we would not expect trait alexithymia to be associated with higher acceptability for these cases since the agent is condemned on the basis of malicious intentions as well as negative outcome, so there is no conflict between two systems. Additionally, we do not expect significant association in any direction between alexithymia and attempted harms because the condemnation of the agent in this case does not rely on empathic resonance with the victim since there is no negative outcome and condemnation is driven exclusively by mental state reasoning (or theory of mind ability) which is not impaired in alexithymia (e.g. Bernhardt et al., 2013; Lockwood et al., 2013; but see Moriguchi et al., 2006). Furthermore, there is no evidence pertaining to whether trait alexithymia is impaired on reinforcement learning with consequent failure in developmentally acquiring aversion to harmful actions/intent, although this still needs to be experimentally demonstrated.

We note that in this study we are focusing only on effects of mental states (e.g. beliefs, intentions) and outcome (harmful or not) inputs on moral judgement and they represent *sufficient* but not *necessary* inputs to moral judgement (Inbar, Pizarro, & Cushman, 2012). Decisions about moral blame and permissibility depend on many contextual factors, including an agent's desires (Cushman, 2008), reasonableness of the beliefs (Young, Nichols, & Saxe, 2010), severity of the harm (Trémolière & De Neys, 2013), an agent's prior record (Kliemann, Young, Scholz, & Saxe, 2008), the means of harm (Greene, Nystrom, Engell, Darley, & Cohen, 2004), external constraints on the agent (e.g., coercion; Woolfolk, Doris, & Darley, 2006), wicked desire (Inbar et al., 2012),

an agent's control over his/her behaviour (Provencher & Fincham, 2000), type of moral transgression (Young & Saxe, 2011), etc.

2. MATERIALS AND METHODS

2.1. Participants

Two hundred and ninety-five (193 women) native Italian speakers between the ages of 18 and 60 (M = 24.96, 95% confidence interval [CI] [23.46, 24.61]) voluntarily logged on to fill a web survey. There was no monetary compensation for the participants. The survey web page was promoted through discussion on online forums, social network and via word of mouth. Exclusion criteria for participation included presence of a diagnosed psychiatric illness and/or history of psychiatric treatment, history of significant neurological illness or brain injury.

2.2. Measures and procedure

All participants gave an informed consent before starting the survey. They then progressed through a series of self-report measures that assessed variables of interest and gave judgements on four moral situations. The order in which participants completed the various questionnaires and different types of scenarios was randomised across participants. There was no time limit to answer any of the questionnaires. All the questionnaires and moral scenarios provided were in Italian and the translated documents are available upon request to the corresponding author.

Moral judgement. Stimuli consisted of four variations of four scenarios for a total of 16 stories (see Supplementary Text S1 for detailed description of the scenarios). A 2 × 2 within-subjects design was used for each scenario involving belief and outcome as factors with two levels (neutral, negative) which created four possible conditions: (1) neutral harm (protagonist holds true belief that she would not cause harm); (2) attempted harm (protagonist has false belief that she would cause harm); (3) accidental harm (protagonist falsely believes that she would not cause harm and ends up causing harm); and (4) intentional harm (protagonist correctly believes that she would cause harm and successfully does so). Each possible belief was true for one outcome and false for the other. Each participant saw one variation of each scenario, for a total of four stories. The order of scenarios from each condition was pseudorandomised such that the same condition appeared in all possible positions across subjects.

After each scenario, participants were asked to make judgements about moral acceptability of the protagonist's behaviour on a 9-point Likert scale (1: *completely morally inacceptable*, 9: *completely morally acceptable*). All scenarios were adapted from Young and Saxe (2008). We included only non-lethal harms (broken limbs, burned skin, etc.) in our scenarios to have more variation in the ratings, since lethal harms are judged more severely (Trémolière & De Neys, 2013). All scenarios provided information about beliefs with which the agents acted and the intent had to be inferred from these beliefs.

Alexithymia. To assess trait alexithymia, we used validated Italian version of Toronto Alexithymia Scale-20 (TAS-20) questionnaire (Bagby, Taylor, & Parker, 1994; Italian version: Bressi et al., 1996), which has been argued to be the best current measure overall for assessing alexithymia due to its sound reliability, validity and broad generalisability (Timoney & Holder, 2013). The TAS-20 is a 20-item scale that consists of three subscales: difficulty describing feelings (five items, e.g. "It is difficult for me to find the right words for my feelings."), difficulty identifying feelings (seven items, e.g. "When I am upset, I don't know if I am sad, frightened, or angry.") and externally oriented thinking (eight items, e.g. "I prefer to analyse problems rather than just describe them."). Items were rated using a 5-point Likert scale (1: *strongly disagree*, 5: *strongly agree*). Higher total scores indicate higher levels of alexithymia. Each factor has shown good internal consistency, with Cronbach's alpha of 0.78, 0.75, and 0.66, respectively. Test–retest reliability for the full scale is 0.77 (Bagby et al., 1994).

Empathy. Participants completed validated Italian version of IRI (M. H. Davis, 1980, 1983; Italian version: Albiero, Ingoglia, & Lo Coco, 2006), a 28-item self-report questionnaire with four 7-item subscales, assessing specific aspects of dispositional empathy. Participants reported agreement with statements on a 5-point Likert scale (0: *never true for me*, 4: *always true for me*). The four subscales consisted of: (1) fantasy scale (F), which measures the propensity to transpose oneself into imaginary lives of fictional characters and

understand their feelings and actions (e.g. "I really get involved with the feelings of the characters in a novel."); (2) PT scale, which measures the tendency to spontaneously adopt the psychological point of view of others (e.g. "I try to understand my friends better by imagining how things look from their perspective."); (3) EC scale, which measures the *other-oriented* tendency to experience feelings of warmth, compassion and concern for the unfortunate others (e.g. "When I see someone being taken advantage of, I feel kind of protective towards them."); (4) personal distress (PD) scale, which measures the *self-oriented* tendency to feel personal unease and discomfort in reaction to the emotions of others (e.g. "When I see someone who badly needs help in an emergency, I go to pieces."). All four subscales have good internal and test–retest reliabilities (M. H. Davis, 1980): internal reliabilities range from 0.71 to 0.77, whereas test–retest reliabilities range from 0.62 to 0.71.

2.3. Data analysis

Statistical analysis was conducted using SPSS 22.0 software. Given recent criticism of dichotomous logic of null-hypothesis testing and poor reproducibility of p values in psychological research (Ioannidis, 2005), we include recommended CIs for estimates and effect sizes (Cumming, 2014) generated using resampling and bootstrapping methods (Kirby & Gerlanc, 2013). Unless otherwise stated, all 95% bias-corrected and accelerated CIs for estimates were generated using 10,000 bootstrap samples. If present, asymmetry in 95% bias-corrected and accelerated CIs reflects asymmetry of the underlying sampling distribution of point estimates. We also include traditional p values, all of which are exact rather than based on asymptotic approximation and computed from two-tailed statistical tests.

Since the dependent variables of interest (scores on IRI subscales and acceptability ratings for different scenarios) did not follow a normal distribution (Shapiro–Wilk test: $p < .01$) and were ordinal variables, all the tests employed were non-parametric. The only parametric test used was repeated measures, analysis of variance (ANOVA) which is robust to violations of a normality assumption (Stevens, 1996). We used ordered logistic regression models when regression was of interest to us instead of linear regression. Test of parallel lines showed that none of the regression models violated the proportional odds assumption ($p > .05$). We report *unstandardised* logit coefficients (B) from which odds ratios (OR) can be computed using exponential function as e^B. We do not report and compute OR from *standardised* logit coefficient because there is no widely agreed upon definition of it, thereby preventing straightforward interpretation (Hosmer & Lemeshow, 2004). OR greater than 1 or less that 1 denote that increase in value of predictor variable is associated with increased likelihood for *higher* or *lower* value of criterion variable, respectively.

Also, for inter-group comparisons, Mann–Whitney U test was used and effect size (r) for these tests was computed as $r = Z/\sqrt{n}$, where Z is the standardised statistic (Fritz, Morris, & Richler, 2012). Additionally, we report Hodges–Lehmann (HL) estimator values for the median difference between groups being compared. Correlation analysis (presented in supplementary analysis) was done using Spearman rank correlations and, when necessary, partial Spearman rank correlations were computed using SPSS syntax (see: http://www-01.ibm.com/support/docview.wss?uid=swg21474822).

For mediation analysis, we did not use Sobel's test because: (1) it has small statistical power and is not recommended for small sample sizes ($n <$ 1000, MacKinnon, Lockwood, Hoffman, West, & Sheets (2002)) which was the case for our study (n = 295); (2) it evaluates samples on the assumption that indirect effects follow normal distribution, which is hardly true in practice. We instead used the non-parametric, Preacher–Hayes bootstrapping method to estimate indirect effects in mediation analysis because statistical power-wise it is more robust with small sample sizes ($n < 25$) and it does not assume normal distribution (Preacher & Hayes, 2004, 2008a).

Because sufficient power is required to claim meaningful null effects (i.e. TAS, subscales of IRI and ratings on moral scenarios are not correlated with each other), we conducted a sensitivity analysis using *G*Power 3.1* (Faul, Erdfelder, Lang, & Buchner, 2007) to determine how small an effect we could detect in each correlation analysis (exact two-tailed tests). With a sample size of 295, a Type II error probability of $\alpha = 0.05$, and a statistical power (Type I error probability) of $\beta = 0.80$, we could detect an effect size > 0.162 (i.e. a small effect; Cohen, 1992). We could not perform similar analysis for ordinal logistic regression analysis due to unavailability of this option in *G*Power*.

3. RESULTS

3.1. Descriptive statistics

All measurement scales and their subscales showed reliable internal consistency, as assessed by Cronbach's alpha (Table 2).

The mean alexithymia score in our sample was 43.92 (95% CI [42.64, 45.19]) with a minimum–maximum spread of 20–77, which is within the normative range, since alexithymia scores in the general population are believed to follow a normal distribution with a mean of 43.34 (SD = 8) (Loas et al., 2001).

Strong gender effects were observed on empathy subscales (Table 2); women reported to have more propensity to fantasise with fictional characters (r = .256, HL estimator = 3.000 [2.000, 4.000]), to possess more EC for others' welfare (r = .295, HL estimator = 3.000 [2.000, 4.000]) and to feel more personal discomfort and unease in stressful social situations (r = .264, HL estimator = 3.000 [1.000, 4.000]). There was no self-reported difference across genders in ability to take perspective of others.

Moral judgement of participants for different scenarios was susceptible to both belief and outcome information (Table 2 and Figure 2 for mean values). A 2 × 2 repeated measures ANOVA determined the exact influence of belief (neutral, negative) and outcome (neutral, negative) information on moral judgements. Results revealed that there was a main effect of belief ($F(1,294) = 313.5$, $p < .001$, $p\eta^2 = 0.516$) such that behaviour of agents who acted on a negative belief was condemned more severely than that of agents who acted on neutral or benign belief. Also, acts which resulted in harmful outcomes were condemned more severely compared to acts which did not ($F(1,294) = 223.0$, $p < .001$, $p\eta^2 = 0.431$). These main effects were mediated by an interaction effect between belief and outcome ($F(1,294) = 27.4$, $p < .001$, $p\eta^2 = 0.085$) which meant that moral acceptability of negative and neutral outcomes depended on whether these outcomes were produced with negative or neutral belief. *Post hoc* Bonferroni-corrected t tests showed that when there was no harmful outcome, behaviour of agents who acted with harmful intent was still judged to be less morally acceptable than agents who acted with neutral intent (mean difference = −3.180 (95% CI [−3.526, −2.833]), adjusted $p < .001$). Additionally, when there was a negative outcome, behaviour of agents who accidently produced this outcome acting on false belief was found to be more acceptable than when this outcome was intentionally produced (mean difference = 2.037 (95% CI [1.633, 2.412]), adjusted $p < .001$). Although there was more variation in ratings for attempted harm (coefficient of variation = 63.0%) as compared to accidental harm (coefficient of variation = 55.5%) scenarios, this difference was not significantly different (Levene's

TABLE 2
Means with 95% CIs, Cronbach's alphas, medians, minimum–maximum spread for moral judgements on different types of harm, alexithymia score and empathy IRI subscale scores

Item	*Cronbach's alpha*	*Mean [95% CI]*	*Median*	*Min, Max*	*Gender effects (Z)*
Neutral	–	7.17 [6.93, 7.40]	8	1, 9	0.307
Accidental	–	4.69 [4.40, 4.99]	4	1, 9	−0.403
Attempted	–	3.98 [3.69, 4.26]	4	1, 9	−1.425
Intentional	–	2.66 [2.42, 2.90]	2	1, 9	−1.732
DDF	0.661	12.25 [11.79, 12.68]	12	5, 24	−1.249
DIF	0.781	16.45 [15.77, 17.12]	15	7, 32	2.032*
EOT	0.628	15.23 [14.74, 15.72]	15	8, 29	−2.865**
TAS-total	0.816	43.92 [42.64, 45.19]	43	20, 77	−0.457
F	0.792	17.34 [16.80, 17.87]	18	4, 28	4.399***
PT	0.831	17.88 [17.36, 18.41]	18	3, 28	1.182
EC	0.783	18.37 [17.88, 18.86]	19	5, 28	5.065***
PD	0.801	11.14 [10.60, 11.68]	11	1, 26	4.535***
IRI-total	0.842	64.73 [63.31, 66.04]	66	28, 95	6.001***

DDF = difficulty describing feelings; DIF = difficulty identifying feelings; EOT = externally oriented thinking; TAS = Toronto alexithymia scale; *F* = fantasy; PT = perspective taking; PD = personal distress; EC = empathic concern; IRI = interpersonal reactivity index; "–" = not applicable. Z = standardised statistic from Mann–Whitney U-test. Positive value of Z signifies that women scored higher on this variable than men.

*$p < .05$; **$p < .01$; ***$p < .001$.

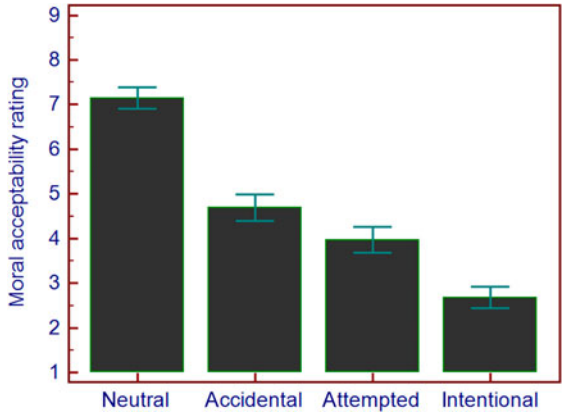

Figure 2. Moral judgements given by participants on a 9-point Likert scale (1: *completely moral inacceptable*, 9: *completely moral acceptable*) for different types of harms: neutral harm (neutral belief, neutral outcome), accidental harm (neutral belief, negative outcome), attempted harm (negative belief, neutral outcome), and intentional harm (negative belief, negative outcome). Error bars represent 95% CIs.

test: $F(1,588) = 3.011$, $p = .083$). Thus, people showed equal degree of disagreement for both accidental and attempted harm cases, both of which contained incongruent information about belief and outcome.

3.2. Regression analysis

Ordinal logistic regression (Table 3) showed that trait alexithymia was associated with reduced EC and PT with an OR of 0.9704 (95% CI [0.9493, 0.9881]) and 0.9724 (95% CI [0.9493, 0.9900]), respectively. In other words, people who scored high on alexithymia reported to have diminished tendency to adopt somebody else's psychological point of view and reduced warm and compassionate feelings for others. Elevated levels of alexithymia were also associated with increase in reported personal discomfort and unease at the misfortune of others with OR = 1.0434 (95% CI [1.0274, 1.0682]).

Crucially, people with higher alexithymia scores were more likely to find accidental harms morally more acceptable, OR = 1.0232 (95% CI [1.0020, 1.0449]). In other words, people with alexithymia judged behaviour of agents who hurt someone while acting on false belief more leniently. There was no other association between any other types of harm and severity of trait alexithymia (Table 3).

When acceptability ratings for different types of harms were regressed onto subscales of IRI, various interesting associations emerged (Table 4). Increased tendency to transform oneself into fictional world of stories and movies was associated with reduced acceptance of intentional harms with an OR of 0.9167 (95% CI [0.8711, 0.9627]), probable due to enhanced ability to experience subjective reality of the victim. Also, enhanced cognitive ability to spontaneously adopt point of view of others was associated with reduced acceptance of acts with negative outcomes, viz. accidental (OR = 0.9474 [0.8976, 1.0020]) and intentional harms (OR = 0.9455 [0.9039, 0.9900]), probably owing to viewing things from victims' point of view and understanding the resulting distress in the victim. There was no association between the PD subscale of IRI and ratings for any type of harm.

TABLE 3
Alexithymia (TAS) scores predicting moral judgements for different types of harm and scores on empathy IRI subscales

Predictor variable	*Categorical variable*	*Logit coefficient [95% CI]*[a]	*Wald's* χ^2	p *value*
	Neutral	0.010 [−0.012, 0.032]	1.089	.297
	Accidental	0.023 [0.002, 0.044]	6.090	.014
	Attempted	0.005 [−0.014, 0.024]	0.329	.566
	Intentional	0.015 [−0.006, 0.035]	2.502	.114
TAS	*F*	0.003 [−0.015, 0.020]	0.084	.772
	PT	−0.028 [−0.052, −0.010]	9.421	.002
	EC	−0.030 [−0.052, −0.012]	10.720	.001
	PD	0.043 [0.027, 0.066]	20.775	<.001

TAS = Toronto alexithymia scale; *F* = fantasy; PT = perspective taking; PD = personal distress; EC = empathic concern; CI = confidence interval. See Supplementary Table S1 for the same analysis with age and gender as additional predictor variables.

[a]Ninety-five per cent bias-corrected and accelerated CIs for logit coefficients were generated using 10,000 bootstrap samples. Positive or negative value of logit coefficient denote that increase in value of predictor variable is associated with increased odds for *higher* or *lower* value of criterion variable, respectively.

TABLE 4
IRI subscale scores predicting moral judgements for different types of scenarios

Predictor variable	*Categorical variable*	*Logit coefficient [95% CI]*[a]	*Wald's* χ^2	p *value*
F	Neutral	0.006 [−0.038, 0.051]	0.079	.778
	Accidental	−0.010 [−0.061, 0.038]	0.225	.635
	Attempted	−0.032 [−0.082, 0.018]	2.136	.144
	Intentional	−0.087 [−0.138, −0.038]	14.09	<.001
PT	Neutral	−0.034 [−0.079, 0.012]	2.152	.144
	Accidental	−0.054 [−0.108, 0.002]	5.823	.016
	Attempted	−0.021 [−0.073, 0.030]	0.929	.335
	Intentional	−0.056 [−0.101, −0.010]	5.890	.015
EC	Neutral	−0.056 [−0.110, −0.003]	5.073	.024
	Accidental	−0.073 [−0.125, −0.020]	9.218	.002
	Attempted	−0.096 [−0.152, −0.044]	15.425	<.001
	Intentional	−0.120 [−0.167, −0.075]	22.357	<.001
PD	Neutral	−0.012 [−0.061, 0.033]	0.296	.587
	Accidental	0.014 [−0.036, 0.065]	0.400	.527
	Attempted	−0.041 [−0.084, 0.005]	3.542	.060
	Intentional	0.010 [−0.038, 0.059]	0.198	.656

F = fantasy; PT = perspective taking; EC = empathic concern; PD = personal distress; CI = confidence interval. See Supplementary Table S2 for the same analysis with age and gender as additional predictor variables.

[a]Ninety-five per cent bias-corrected and accelerated CIs for OR were generated using 10,000 bootstrap samples. Positive or negative value of logit coefficient denote that increase in value of predictor variable is associated with increased odds for *higher* or *lower* value of criterion variable, respectively.

Interestingly, EC was associated with reduced acceptability for all types of harm, irrespective of the outcome or belief. Since EC taps into self-reported ability to feel concerned about the welfare of others and feeling protective of the unfortunate others, it is not surprising that increased EC predicted reduced acceptance of accidental harm (OR = 0.9296 [0.8825, 0.9802]), attempted harm (OR = 0.9085 [0.8589, 0.9569]) and intentional harm (OR = 0.8869 [0.8462, 0.9277]).

Surprisingly, there was a negative association between EC and ratings for neutral harm scenario in which there was neither harmful intent nor outcome (OR = 0.9455 [0.8958, 0.9970]). Additional exploration of the data showed that this was case only for the order in which neutral harm scenario was presented right after intentional harm scenarios (p = .033) but in no other order (ps > .3). Thus, it is possible that this association was a result of carry-over effects (Davis, 1995) from the intentional harm scenario to neutral harm. In other words, having just faced a scenario where agent intentionally hurt someone, people must have judged even a neutral scenario more severely.

Similar results were obtained after adding additional demographic predictor variables (age and gender) to the regression models (see Supplementary Tables S1 and S2). Correlation analysis revealed identical results to the regression analysis (see Supplementary Table S3).

3.3. Mediation analysis

Trait alexithymia was associated with reduced scores on PT and EC subscales of IRI which in their turn were negatively associated with higher ratings on accidental harm (Figure 3). Multiple mediation analysis was thus carried out to explore the extent to which increased tendency in trait alexithymia to pass lenient moral judgement for accidental harm was mediated through indirect effects stemming from decreased levels of EC and PT in this trait. Bootstrap estimation of 95% bias-corrected and accelerated CIs for the indirect effect was done implementing Preacher–Hayes' bootstrapping multiple mediation method for indirect effects using 20,000 bootstrap samples (Preacher & Hayes, 2008a). This analysis showed that EC (95% CI [0.0004, 0.0150]) but not PT (95% CI [−0.020, 0.0116]) mediated the relation between trait alexithymia and more lenient judgements for accidental harm (Figure 3). Index of mediation (Preacher & Hayes, 2008b) for EC, which is a standardised effect size measure of mediation, was 0.0274 (95% CI [0.0018, 0.0647]). Similar results emerged even after running the same mediation analysis controlling for age and gender (see Supplementary Figure S1).

Thus, trait alexithymia influenced moral judgements via EC such that it led to reduction in EC

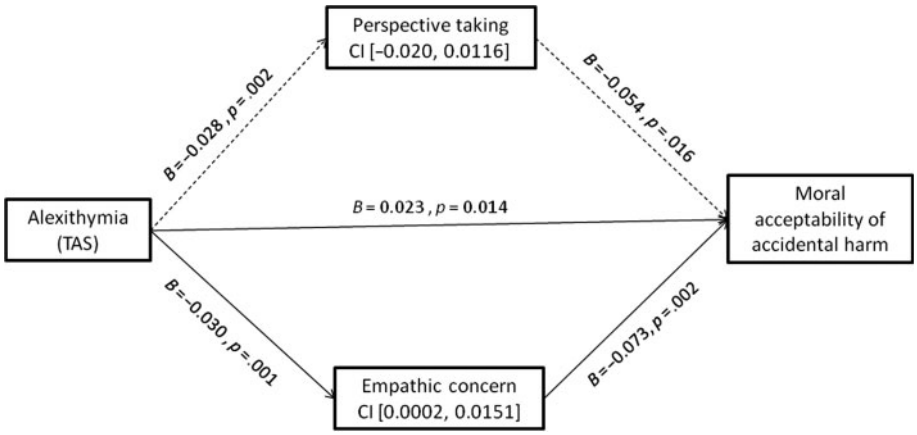

Figure 3. Mediation analysis results. Negative logit coefficient from ordinal regression denotes reduced EC, reduced perspective taking and increased moral acceptability of accidental harm. Bias-corrected and accelerated 95% CIs from 20,000 bootstrap samples are reported for specific indirect effects. The increased tendency to find accidental harms more acceptable in trait alexithymia was due to reduced EC. Continuous lines denote significant mediation path, whereas dashed lines denote insignificant mediation path. Direct effect of alexithymia on accidental harm judgement has also been shown for the sake of completeness.

for the victim which weakened the affective aversion to the harm that befell the victim and participants focused more on the innocent intentions when they assessed behaviour of the protagonist.

4. DISCUSSION

Mature moral judgements about complex situations from everyday life not only take into account the consequences of actions but also the intent with which these actions were pursued. An interesting situation arises when there is no homologous mapping of the valence of the belief onto the valence of the outcome, i.e. when acts undertaken with neutral or negative beliefs lead to negative or neutral outcomes, respectively. In such cases, lay people's judgements rely more on the specified mental states of the actors. Agents who accidentally cause harm are judged more leniently, whereas agents acting with negative intent to harm are judged more severely. The two-process model implicates different processes which underlie these intent-based moral judgements. A causal process tracks the valence of the outcome and is motivated by negative emotional response originating from affective resonance with the victim in the situation to condemn the agent in case the outcome is negative (i.e. harmful outcome). Another process engages in attributing mental states to the agents and assesses their behaviour based on this information and is motivated by emotional aversion arising from evaluative simulation of the harmful actions described in the story. In this study, we explored how reduction in the emotional aversion due to lack of EC for the victim associated with trait alexithymia affects intent-based moral judgements. We hypothesised and found that reduced empathic aversion to negative outcomes made the mental state processes dominate the decision-making and made accidental harms seem more acceptable to alexithymic personalities.

4.1. Empathy and harm-based moral judgements

Although it has been argued that empathy is not necessary for moral judgements (Prinz, 2011) where there is no clearly discernible victim, e.g. insider trading, its epistemic and motivational role in harm-based moral judgements is undeniable (Pizarro, 2000; Ugazio, Majdandžić, & Lamm, in press). Empathy has been argued to have led to evolution of harm/care-based morality (Decety, 2014) and is said to be essential for proper moral development (Hoffman, 2001), especially for harm/care-based morality (Blair, 1995). Empathy enables people to share the affective state of the victim of others' moral actions. If action that needs to be judged leads to distress in the victim, then empathic resonance with the victim's suffering leads to empathic arousal in the observer and informs her that a morally relevant event is taking place (epistemic role) and motivates the observer to either approach the victim to alleviate her suffering

or to withdraw from the situation to remove the source of PD (motivational role). Thus, empathy is implicated as a moral emotion because it is a moral marker by which people understand that moral norms are being violated and are motivated to deem those actions morally wrong which cause suffering in the victim.

But evidence to support these claims mainly comes from research done with moral dilemmas where the agent needs to *intentionally* harm one person in order to save a greater number of people from getting hurt (for a review, see Ugazio et al., in press). In this study, we broaden the scope of empathy for harm-based moral judgements by showing its role in situations where a victim gets hurt not only intentionally but also *accidentally*. Empathic aversion to negative outcomes is an important factor that motivates causal reasoning processes to hold the agent accountable for his/her actions that caused suffering in the victim, benign intentions notwithstanding. Accordingly, the self-reported tendency to spontaneously adopt others' point of view and experience their subjective reality was negatively associated with acceptability of both accidental and intentional harms (Table 4). Also, the tendency to feel warm and sympathetic feelings for others who are experiencing some kind of adversity was also associated with reduced acceptance of behaviour that led to harm, be this behaviour intentional or accidental (Table 4). Thus, both cognitive and affective aspects of empathy were involved in condemning behaviour that resulted in a negative outcome. It is important to note that only the contribution of EC was significant when both EC and PT were entered simultaneously into the regression model to predict acceptability ratings for accidental harm (EC: OR = 0.9408 [0.8896, 0.9900], Wald's χ^2 = 5.773, p = .016; PT: OR = 0.9656 [0.9121, 1.0263], Wald's χ^2 = 2.163, p = .141) and intentional harm (EC: OR = 0.8923 [0.8470, 0.9333], Wald's χ^2 = 18.005, p < .001; PT: OR = 0.9841 [0.9380, 1.0377], Wald's χ^2 = 0.444, p = .505). Therefore, it seems that even though it is important to adopt others' psychological point of view to *understand* the hardship they are going through, it is even more vital to actually *experience* compassionate feelings for their suffering if this understanding is to motivate any prosocial behaviour (which would be expressing outrage at the harmful outcome in this instance).

Alexithymia is characterised by reduction in this very feeling of concern for others' well-being which is why the victim suffering in case of accidental harm did not register on emotional level with these individuals and did not lead to a potent empathic arousal. This weakened empathic aversion failed to motivate the causal processes which would have condemned the agent based on his/her causal involvement in the act that led to the negative outcome. In the absence of causal processes, the mental state processes took over and these individuals focused more on the intentions of the agent. This made it more likely that people with higher scores on alexithymia found accidental harms morally more acceptable.

In a previous study, Young et al. (2012) found that criminal psychopaths were also more likely to find accidental harms more permissible. The question is whether this makes them more rational than normal population, since normatively we would expect it to be a rational choice to forgive someone for causing a bad outcome while acting on false beliefs. Young et al. (2012) refuted this assertion and posited that there are different routes that one can take to come at forgiveness: "On the one hand, enhanced theory of mind and consideration of the agent's innocent intention may lead to forgiving accidents; on the other hand, reduced sensitivity to the victim's experience of pain and suffering" (p. 664). They hypothesised that criminal psychopaths take the latter route, but this has not been tested yet. There exists evidence for the former route which involves people forgiving agents by focusing more on innocent intentions (Koster-Hale et al., 2013; Young & Saxe, 2009b), but this is the first study to demonstrate that reduced EC for the victim can also lead one to forgive accidental harms. Thus, people with alexithymia make normatively desirable judgement for normatively undesirable reasons (cf. Hertwig & Volz, 2013).

4.2. Alternative explanation

One issue that we have turned a blind eye to in the discussion so far is our key assumption that IRI subscales are tracking the cognitive and affective empathy for the *patient* (i.e. victim) and not the *agent* in the scenarios. In general, mental state reasoning for an independent observer (read participant) can happen at multiple levels such that mind is attributed to the agent who is acting, to the patient who is at the receiving end of these actions, to the other observers who are judging this independent observer (e.g. experimenter), etc.

(Young & Waytz, 2013). Thus, a case needs to be made that variance in moral judgements associated with trait alexithymia was due to variance in the empathic reasoning about the patient and *not* the agent, especially as measured via IRI. We restrict our discussion to PT and EC because they mediated the relation between trait alexithymia and moral judgements and were only subscales of IRI which primarily predicted moral judgements for different types of harm.

PT involves viewing things from others' point of view, but this can mean either from an agent's or patient's point of view. Since higher scores on PT were associated with higher condemnation of behaviours that resulted in harmful outcome (both accidental and intentional harm; Table 4), we contend that PT tapped into ability to adopt point of view of the *patient*. If it were associated with taking agent's perspective, PT would have exhibited negative association with attempted and intentional harm which contain agents with negative intents and positive association with accidental harm involving agent with benign intent, but these associations were not observed.

EC gauges how apprehensive someone gets for the welfare of others in negative situations and how warm and compassionate they feel towards them. Since there was no salient information concerning the affective state of the agent in the scenarios, it is not clear why this subscale would measure EC for the agent. On the other hand, each scenario explicitly referred to the actual or counterfactual harm that befell or would have befallen the patient (e.g. "Her sister's arm hits the iron, and she gets badly burned. She cries hysterically all the way the emergency room."). Accordingly, EC was negatively associated with all types of harm (Table 4).

This is the first study to explore how the empathic reasoning about the patient, in addition to the mental state reasoning about the agent, influences intent-based moral judgement. Future brain imaging studies should try to decouple these processes and investigate if they share convergent and/or divergent neural architecture.

4.3. Limitations

One statistical limitation is that some of the main results might be false positives. The sensitivity analysis (see Data Analysis) showed that our study could detect correlation coefficient greater than 0.162 and yet we detected significant correlations (see Supplementary Table S3) of 0.127 (TAS and rating on accidental harm) and −0.134 (EC and rating on accidental harm). Although we cannot completely rule out this possibility, we establish the stability of our results by computing CIs for estimators of interest (correlation coefficients, regression coefficients, etc.) using bootstrapping methods with as many as 10,000 samples (minimum recommended: 1000; IBM SPSS Bootstrapping 20.0 manual, 2011). Bootstrapping methods are valuable when the sample sizes are small and are not representative of the entire population because they are asymptotically more accurate than the standard intervals computed based on sample variance and normality assumptions (Adèr, Mellenbergh, & Hand, 2008). To sum up, although we did not have sufficient sample size to detect significant associations that we did detect, we prop these results up with more reliable statistical techniques.

5. CONCLUSION

This research provided additional evidence for a link between trait alexithymia and empathy deficits and explored how this disruption translated into behaviour in hypothetical moral situations. Our findings suggest that impairment in empathic skills, especially EC, contributes to reduction in harm aversion which leads to increased propensity in alexithymic population towards finding accidental harms more acceptable. Although previous research has implicated empathy in moral condemnation of intentional harms, this is the first study to explicitly explore the role of empathy in condemning accidental harms. Results are consistent with the predictions of the two-process model of intent-based moral judgements and provide further evidence for its validity.

Supplementary material

Supplementary (Figure/Table/content) is available via the "Supplementary" tab on the article's online page (http://dx.doi.org/10.1080/20445911.2014.929137).

Original manuscript received March 2014
Revised manuscript received May 2014
Revised manuscript accepted May 2014
First published online June 2014

REFERENCES

Adèr, H. J., Mellenbergh, G. J., & Hand, D. J. (2008). *Advising on research methods: A consultant's companion*. Huizen: Johannes van Kessel Publishing.

Albiero, P., Ingoglia, S., & Lo Coco, A. (2006). Contributo all'adattamento italiano dell'Interpersonal Reactivity Index [A contribution to the Italian validation of the Interpersonal Reactivity Index]. *Testing Psicometria Metodologia*, *13*, 107–125.

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

Bagby, R. M., Taylor, G. J., & Parker, J. D. A. (1994). The Twenty-Item Toronto Alexithymia Scale: II. Convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research*, *38*, 33–40. doi:10.1016/0022-3999(94)90006-X

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

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46. doi:10.1016/0010-0277(85)90022-8

Bernhardt, B. C., Valk, S. L., Silani, G., Bird, G., Frith, U., & Singer, T. (2013). Selective disruption of sociocognitive structural brain networks in autism and alexithymia. *Cerebral Cortex*. Advance online publication. doi:10.1093/cercor/bht182

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

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

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, *57*, 1–29. doi:10.1016/0010-0277(95)00676-P

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences*, *11*, 387–392. doi:10.1016/j.tics.2007.07.003

Blair, R. J. R. (2013). The neurobiology of psychopathic traits in youths. *Nature Reviews Neuroscience*, *14*, 786–799. doi:10.1038/nrn3577

Bressi, C., Taylor, G., Parker, J., Bressi, S., Brambilla, V., Aguglia, E., … Invernizzi, G. (1996). Cross validation of the factor structure of the 20-item Toronto Alexithymia Scale: An Italian multicenter study. *Journal of Psychosomatic Research*, *41*, 551–559. doi:10.1016/S0022-3999(96)00228-0

Buon, M., Dupoux, E., Jacob, P., Chaste, P., Leboyer, M., & Zalla, T. (2013). The role of causal and intentional reasoning in moral judgment in individuals with high functioning autism. *Journal of Autism and Developmental Disorders*, *43*, 458–470. doi:10.1007/s10803-012-1588-7

Buon, M., Jacob, P., Loissel, E., & Dupoux, E. (2013). A non-mentalistic cause-based heuristic in human social evaluations. *Cognition*, *126*, 149–155. doi:10.1016/j.cognition.2012.09.006

Ciaramelli, E., Braghittoni, D., & di Pellegrino, G. (2012). It is the outcome that counts! Damage to the ventromedial prefrontal cortex disrupts the integration of outcome and belief information for moral judgment. *Journal of the International Neuropsychological Society*, *18*, 962–971. doi:10.1017/S1355617712000690

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

Cushman, F. A. (2008). Crime and punishment: Differential reliance on causal and intentional information for different classes of moral judgment. *Cognition*, *108*, 353–380. doi:10.1016/j.cognition.2008.03.006

Cushman, F. A. (2013). Action, outcome and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*, 273–292. doi:10.1177/1088868313495594

Cushman, F. A., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*, 2–7. doi:10.1037/a0025071

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

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*, 363–366. doi:10.1016/j.tics.2013.06.005

Davis, M. H. (1980). A multi-dimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.

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

Davis, A. (1995). The experimental methods in psychology. In G. Breakwell, S. Hammond, & C. Fife Shaw (Eds.), *Research methods in psychology* (pp. 50–68). London: Sage.

Decety, J. (2014). The neuroevolution of empathy and caring for others: Why it matters for morality. In J. Decety & Y. Christen (Eds.), *New frontiers in social neuroscience* (pp. 127–151). New York: Springer.

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

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, *22*, 209–220. doi:10.1093/cercor/bhr111

Decety, J., & Porges, E. C. (2011). Imagining being the agent of actions that carry different moral consequences: An fMRI study. *Neuropsychologia*, *49*, 2994–3001. doi:10.1016/j.neuropsychologia.2011.06.024

de Vignemont, F., & Singer, T. (2006). The empathic brain: How, when and why? *Trends in Cognitive Sciences*, *10*, 435–441. doi:10.1016/j.tics.2006.08.008

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

FeldmanHall, O., Dalgleish, T., & Mobbs, D. (2012). Alexithymia decreases altruism in real social decisions. *Cortex*, *49*, 899–904. doi:10.1016/j.cortex.2012.10.015

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18. doi:10.1037/a0024338

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400. doi:10.1016/j.neuron.2004.09.027

Grynberg, D., Luminet, O., Corneille, O., Grezes, J., & Berthoz, S. (2010). Alexithymia in the interpersonal domain: A general deficit of empathy? *Personality and Individual Differences*, *49*, 845–850. doi:10.1016/j.paid.2010.07.013

Guttman, H., & Laporte, L. (2002). Alexithymia, empathy, and psychological symptoms in a family context. *Comprehensive Psychiatry*, *43*, 448–455. doi:10.1053/comp.2002.35905

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, *128*, 451–474. doi:10.1016/j.cognition.2013.04.004

Hauser, M. (2006). *Moral minds*. New York, NY: Harper Collins.

Hertwig, R., & Volz, K. G. (2013). Abnormality, rationality, and sanity. *Trends in Cognitive Sciences*, *17*, 547–549. doi:10.1016/j.tics.2013.08.011

Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*. New York, NY: Cambridge University Press.

Hooker, C. I., Verosky, S. C., Germine, L. T., Knight, R. T., & D'Esposito, M. (2008). Mentalizing about emotion and its relationship to empathy. *Social Cognitive and Affective Neuroscience*, *3*, 204–217. doi:10.1093/scan/nsn019

Hosmer, D. W., & Lemeshow, S. (2004). *Applied logistic regression* (2nd ed.). New York, NY: John Wiley & Sons.

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

Inbar, Y., Pizarro, D. A., & Cushman, F. A. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, *38*, 52–62. doi:10.1177/0146167211430232

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124. doi:10.1371/journal.pmed.0020124

Jonason, P. K., & Krause, L. (2013). The emotional deficits associated with the Dark Triad traits: Cognitive empathy, affective empathy, and alexithymia. *Personality and Individual Differences*, *55*, 532–537. doi:10.1016/j.paid.2013.04.027

Kano, M., & Fukudo, S. (2013). The alexithymic brain: The neural pathways linking alexithymia to physical disorders. *BioPsychoSocial Medicine*, *7*, 1. doi:10.1186/1751-0759-7-1

King, J. A., Blair, R. J. R., Mitchell, D. G. V., Dolan, R. J., & Burgess, N. (2006). Doing the right thing: A common neural circuit for appropriate violent or compassionate behavior. *NeuroImage*, *30*, 1069–1076. doi:10.1016/j.neuroimage.2005.10.011

Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, *45*, 905–927. doi:10.3758/s13428-013-0330-5

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*, 2949–2957. doi:10.1016/j.neuropsychologia.2008.06.010

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences USA*, *110*, 5648–5653. doi:10.1073/pnas.1207992110

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*, 2492–2502. doi:10.1016/j.neuroimage.2010.10.014

Lane, R. D., Ahern, G. L., Schwartz, G. E., & Kaszniak, A. W. (1997). Is alexithymia the emotional equivalent of blindsight? *Biological Psychiatry*, *42*, 834–844. doi:10.1016/S0006-3223(97)00050-4

Loas, G., Corcos, M., Stephan, P., Pellet, J., Bizouard, P., Venisse, J. L., … Jeammet, P. ; Réseau Inserm no. 494013. (2001). Factorial structure of the 20-item Toronto Alexithymia Scale. Confirmatory factorial analyses in nonclinical and clinical sample. *Journal of Psychosomatic Research*, *50*, 255–261. doi:10.1016/S0022-3999(01)00197-0

Lockwood, P. L., Bird, G., Bridge, M., & Viding, E. (2013). Dissecting empathy: High levels of psychopathic and autistic traits are characterised by difficulties in different social information processing domains. *Frontiers in Human Neuroscience*, *7*, 760. doi:10.3389/fnhum.2013.00760

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*, 83–104. doi:10.1037/1082-989X.7.1.83

Marsh, A. (in press). Empathic and moral deficits in psychopathy. In H. Maibom (Ed.), *Empathy in morality* (pp. 138–154). Oxford: Oxford Universitxy Press.

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

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*. Retrieved from http://dx.doi.org/10.1037/a0035361

Miller, M., Sinnott-Armstrong, W., Young, L., King, D., Paggi, A., Fabri, M., … Gazzaniga, M. (2010).

Abnormal moral reasoning in complete and partial callosotomy patients. *Neuropsychologia*, *48*, 2215–2220. doi:10.1016/j.neuropsychologia.2010.02.021

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

Moriguchi, Y., Decety, J., Ohnishi, T., Maeda, M., Mori, T., Nemoto, K., … Komaki, G. (2007). Empathy and judging other's pain: An fMRI study of alexithymia. *Cerebral Cortex*, *17*, 2223–2234. doi:10.1093/cercor/bhl130

Moriguchi, Y., & Komaki, G. (2013). Neuroimaging studies of alexithymia: Physical, affective, and social perspectives. *BioPsychoSocial Medicine*, *7*, 8. doi:10.1186/1751-0759-7-8

Moriguchi, Y., Ohnishi, T., Lane, R. D., Maeda, M., Mori, T., Nemoto, K., …Komaki, G. (2006). Impaired self-awareness and theory of mind: An fMRI study of mentalizing in alexithymia. *NeuroImage*, *32*, 1472–1482. doi:10.1016/j.neuroimage.2006.04.186

Nemiah, J. C., Freyberger, H., & Sifneos, P. E. (1976). Alexithymia: A view of the psychosomatic process. In O. W. Hill (Ed.), *Modern trends in psychosomatic medicine* (pp. 430–439). London: Butterworths.

Pizarro, D. A. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behavior*, *30*, 355–375. doi:10.1111/1468-5914.00135

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, *36*, 717–731. doi:10.3758/BF03206553

Preacher, K. J., & Hayes, A. F. (2008a). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891. doi:10.3758/BRM.40.3.879

Preacher, K. J., & Hayes, A. F. (2008b). Contemporary approaches to assessing mediation in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The Sage sourcebook of advanced data analysis methods for communication research* (pp. 13–54). Thousand Oaks, CA: Sage.

Prinz, J. (2011). Is empathy necessary for morality? In A. Coplan & P. Goldie. (Eds.), *Empathy: Philosophical and psychological perspectives* (pp. 211–229). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199539956. 003.0014

Provencher, H. L., & Fincham, F. D. (2000). Attributions of causality, responsibility, and blame for positive and negative symptom behaviours in caregivers of persons with schizophrenia. *Psychological Medicine*, *30*, 899–910. doi:10.1017/S0033291799002342

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. doi:10.1016/j.neubiorev.2014.01.009

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

Stevens, J. P. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Theriault, J., & Young, L. (2014). Taking an "intentional stance" on moral psychology. In J. Sytsma (Ed.), *Advances in experimental philosophy of mind* (pp. 76–101). London: Continuum Press.

Timoney, L. R., & Holder, M. D. (2013). *Emotional processing deficits and happiness: Assessing the measurement, correlates, and well-being of people with alexithymia* (pp. 35–39). Dordrecht: Springer.

Trémolière, B., & De Neys, W. (2013). Methodological concerns in moral judgment research: Severity of harm shapes moral decisions. *Journal of Cognitive Psychology*, *25*, 989–993. doi:10.1080/20445911.2013.841169

Ugazio, G., Majdandžić, J., & Lamm, C. (in press). Are empathy and morality linked? Insights from moral psychology, social and decision neuroscience, and philosophy. In H. Maibom (Ed.), *Empathy in morality* (pp. 155–171). Oxford: Oxford University Press.

van der Velde, J., Servaas, M. N., Goerlich, K. S., Bruggeman, R., Horton, P., Costafreda, S. G., & Aleman, A. (2013). Neural correlates of alexithymia: A meta-analysis of emotion processing studies. *Neuroscience & Biobehavioral Review*, *37*, 1774–1785. doi:10.1016/j.neubiorev.2013.07.008

Walter, N. T., Montag, C., Markett, S., Felten, A., Voigt, G., & Reuter, M. (2012). Ignorance is no excuse: Moral judgments are influenced by a genetic variation on the oxytocin receptor gene. *Brain Cognition*, *78*, 268–273. doi:10.1016/j.bandc.2012.01.003

Wingbermühle, E., Theunissen, H., Verhoeven, W. M. A., Kessels, R. P. C., & Egger, J. I. M. (2012). The neurocognition of alexithymia: Evidence from neuropsychological and neuroimaging studies. *Acta Neuropsychiatrica*, *24*, 67–80. doi:10.1111/j.1601-5215.2011.00613.x

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

Wu, N., Li, Z., & Su, Y. (2012). The association between oxytocin receptor gene polymorphism (OXTR) and trait empathy. *Journal of Affective Disorders*, *138*, 468–472. doi:10.1016/j.jad.2012.01.009

Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron*, *65*, 845–851. doi:10.1016/j.neuron.2010.03.003

Young, L., Camprodon, J., 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 USA*, *107*, 6753–6758. doi:10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National*

Academy of Sciences USA, *104*, 8235–8240. doi:10.1073/pnas.0701408104

Young, L., Koenigs, M., Kruepke, M., & Newman, J. (2012). Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology*, *21*, 659–667. doi:10.1037/a0027489

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*, 333–349. doi:10.1007/s13164-010-0027-y

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*, 1912–1920. doi:10.1016/j.neuroimage.2008.01.057

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*, 202–214. doi:10.1016/j.cognition.2011.04.005

Young, L., & Saxe, R. (2009a). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*, 1396–1405. doi:10.1162/jocn.2009.21137

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

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass*, *7*, 585–604. doi:10.1111/spc3.12044

Young, L., & Waytz, A. (2013). Mind attribution is for morality. In S. Baron-Cohen, H. Tager-Flusberg, & M. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (3rd ed., pp. 93–103). New York, NY: Oxford University Press.