Consciousness and Cognition 125 (2024) 103762

Contents lists available at ScienceDirect

Consciousness and Cognition

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

Full Length Article

Comparing third-party responsibility with intention attribution: An fMRI investigation of moral judgment

Eugenia Kulakova [a,b,*,1], Sofia Bonicalzi [c,d,1], Adrian L. Williams [e], Patrick Haggard [a,f,g]

[a] Institute of Cognitive Neuroscience, University College London, United Kingdom
[b] Charité – Universitätsmedizin Berlin, Klinik für Psychiatrie und Psychotherapie, Campus Benjamin Franklin, Berlin, Germany
[c] Dipartimento FilCoSpe, Roma Tre University, Rome, Italy
[d] Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft, Ludwig-Maximilians-Universität München, Germany
[e] Department of Life Sciences, Brunel University London, United Kingdom
[f] Laboratoire de Neurosciences Cognitives, École Normale Supérieure, Paris, France
[g] Institut d'études Avancées de Paris, Paris, France

ARTICLE INFO

Keywords:
Moral responsibility
Intention
Social cognition
Causality

ABSTRACT

Neuroimaging studies demonstrate that moral responsibility judgments activate the social cognition network, presumably reflecting mentalising processes. Conceptually, establishing an agent's intention is a sub-process of responsibility judgment. However, the relationship between both processes on a neural level is poorly understood. To date, neural correlates of responsibility and intention judgments have not been compared directly. The present fMRI study compares neural activation elicited by third-party judgments of responsibility and intention in response to animated pictorial stimuli showing harm events. Our results show that the social cognition network, in particular Angular Gyrus (AG) and right Temporo-Parietal Junction (RTPJ), showed stronger activation during responsibility vs. intention evaluation. No greater activations for the reverse contrast were observed. Our imaging results are consistent with conceptualisations of intention attribution as a sub-process of responsibility judgment. However, they question whether the activation of the social cognition network, particularly AG/RTPJ, during responsibility judgment is limited to intention evaluation.

1. Introduction

People hold each other morally responsible for the negative consequences of their behaviours, be they misdemeanours or serious wrongdoings. Although the conceptual debate on responsibility is ramified (Bonicalzi et al., 2022; Fischer & Tognazzini, 2011), attributing responsibility generally involves the recognition that the agent has misbehaved and is thereby an appropriate target for moral reproach (Alicke, 2000; Monroe & Malle, 2019). One central question during responsibility attribution is whether the moral violation was carried out intentionally, i.e., under the guidance of conscious intentions or plans (Cushman, 2008; Mele & Sverdlik,

* Corresponding author at: Charité Universitätsmedizin Berlin, Klinikum für Psychiatrie und Psychotherapie, Campus Benjamin Franklin, Hindenburgdamm 30, 12203 Berlin, Germany.
E-mail address: eugenia.kulakova@charite.de (E. Kulakova).
[1] First two authors contributed equally.

https://doi.org/10.1016/j.concog.2024.103762
Received 10 December 2023; Received in revised form 28 August 2024; Accepted 16 September 2024
Available online 18 September 2024
1053-8100/

1996). Responsibility and intention attributions play a major role in social cognition, with intention evaluations traditionally conceptualised as a sub-process of establishing responsibility. Occasionally, responsibility is attributed even in the absence of the intention to harm and vice versa. For example, when no bad intention can be acknowledged but things still go wrong due to the agent's behaviour, a negative outcome may suffice to attribute responsibility, although reduced. This is typical of cases of negligence due to unintentional, carelessness behaviour (Ames & Fiske, 2013). Reversely, when the agent's intention to harm is frustrated and the harmful outcome prevented, the bad intention can be enough to morally condemn a person (Yaffe, 2010; Young & Saxe, 2009b). Importantly, regardless of the actual *presence* or *magnitude* of a malevolent intention, responsibility judgments still require the *evaluation* of the underlying intention.

In addition to inferring and evaluating the wrongdoer's subjective intentional state, responsibility evaluations also take into account the objective, external nature of the behavioural sequence and its outcomes (Greene et al., 2001). Here, moral psychology resonates with the law, where legal responsibility and criminal liability depend on the joint presence of an act causing negative consequences (*actus reus*) and of a malevolent intent (*mens rea*) (Shute et al., 1993). Evaluating the *actus reus* involves different considerations about the action itself, its outcome, and the wrongdoer's causal control (Beecher-Monas and Garcia-Rill, 2017). Importantly, these aspects can be distinguished from intention as they relate to external/objective facts rather than psychological mental states. We will refer to this conceptual complement of intention evaluation during responsibility attribution in terms of the "external situation".

Although the distinction between intention and responsibility judgments holds conceptually (see also Guglielmo & Malle, 2010; Leslie, Knobe & Cohen, 2006; Waldmann, Nagel & Wiegmann, 2012), their relationship at the neural level remains poorly understood. Evidence from neuroimaging studies highlights how intention evaluations are central to responsibility judgments. Young and Saxe (2009a) showed that evaluations of social scenarios (e.g., a day care employee preparing meat loaf for a group of children) including morally relevant information (whether the employee used fresh meat) activated the brain's social cognition network (Maliske & Kanske, 2022), involving the Temporo-Parietal Junction (TPJ), the Precuneus (PC) and the Medial Prefrontal Cortex (MPFC). This network is central to the mental simulation of other agents' intentions, motives and beliefs, usually referred to as *Theory of Mind* (Fletcher et al., 1995; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014). The activation of this network was interpreted in terms of the spontaneous computation of, and reasoning about, the agent's intentions and beliefs, as no direct information about intentions was provided in the stimuli (Sommer et al., 2007). Similar activations in the social cognition network have been observed for the processing of other people's intentions, even in the absence of moral implications or negative outcomes (Bara, Ciaramidaro, Walter, & Adenzato, 2011; Grafton, 2009).

As such, the current state of evidence in moral neuroscience suggests that activity in the social cognition network during moral evaluations is mainly driven by the computation of agents' intentions and beliefs, which are conceptualised as sub-components of moral evaluations (Cushman & Young, 2011). It should be noted that these conclusions partially rely on reverse inferences, i.e., the demonstrably problematic identification of mental processes based on the activation of specific brain regions (Bzdok et al., 2012; Poldrack, 2006). In this study we were interested in the direct comparison of the neural correlates of both intention and responsibility evaluation. Such direct comparison can help to clarify whether neural data supports the conceptual relationship between intention and responsibility judgments and the role of the social cognition network.

So far, similar direct comparisons of different cognitive processes have only been carried out in meta-analyses of brain activations during moral tasks. Several such studies confirmed the reliable involvement of the social cognition network in different types of moral cognition. Sevinc & Spreng (2014) mapped moral processing into a network including the Posterior Cingulate Cortex (PCC), Temporal Pole (TP), Ventromedial and Medial Prefrontal Cortex, TPJ and Inferior Frontal Gyrus (IFG), which the authors attributed to a domain-general process of mental simulation and integration of social and contextual information. Passively watching stimuli with a moral content elicited stronger activations in the Lingual Gyrus, Amygdala and MPFC, while active moral judgment showed increased involvement in the TPJ, Angular Gyrus (AG), TP and PFC, demonstrating the influence of additional task demands. Another meta-analysis investigated the convergence of brain regions involved in moral cognition with both ToM and empathy, the socio-affective process of grasping others' feelings (Bzdok et al., 2012). They found that activation in the TPJ, MPFC and Medial Temporal Gyrus (MTG) were common to all three tasks, and concluded that all three cognitive processes engage a network involved in the automated prediction of social events that modulate behaviour. Another meta-analysis used a similar approach to investigate brain activations for moral evaluations (i.e., judging the appropriateness of other agents' moral evaluations) compared to active moral judgments (Garrigan, Adlam, & Langdon, 2016). While both processes showed overlap in core social cognition regions such as MTG, CG and MFG, active moral decisions activated the MTG, PC, MFG, IPG and CC to a stronger extent than more detached moral evaluations. Other meta-analyses reported similar results (Boccia et al., 2017; Eres et al., 2018; Han 2017).

In light of the accumulating support for the involvement of core social network regions in moral judgment, the hypothesis that it reflects implicit or explicit intention computation remains the most accepted (Buckholtz et al., 2008; Young & Dungan, 2012; but see Decety & Lamm, 2007). Surprisingly, however, no direct comparison of the neural basis of moral responsibility and intention computation has yet been reported. This task remains challenging, as it is difficult to develop experimental scenarios that modulate the extent of mental state reasoning required during moral judgment. In fact, implicit intention computations can occur irrespective of the explicitly available information about the scenario.

The present study sets out to fill this gap by directly comparing the neural basis of responsibility and intention judgments. For such evaluation, the neural correlates of responsibility and intention evaluations need to be compared directly within the same sets of stimuli and participants. This comparison was carried out in separate blocks but in response to identical visual stimuli depicting harm events.

This design allowed us to clarify whether any areas would show greater activation for intention compared to responsibility judgments, or vice versa. This allows investigating whether the conceptual analysis can be supported or called into question by neural data. On a more explorative level, assuming that subtracting activity related to one cognitive process from another removes its neural contribution, any areas that would show stronger activation for responsibility compared to intention judgments were expected to represent the complementary "external situation" driving moral evaluations. Evidence for or against these hypotheses will contribute to a clearer delineation of the central cognitive processes involved in moral judgment. It will further clarify the contribution of the social cognition network previously observed in moral reasoning and assumingly hosting explicit or implicit intention attributions.

A novel paradigm (Bonicalzi et al., 2022) was employed in the present imaging study. Moral scenarios were represented as animations of geometrical shapes (reminiscent of the animation employed by Heider & Simmel, 1944) delivering or receiving painful shocks. Such non-verbal, visual presentation avoided the semantic and social load of classic language-based vignettes. Additionally, they stripped moral interactions from naturalistic physical action to prevent any form of motor resonance (Grafton, 2009; Walter et al., 2004). Another advantage of visual stimuli is that animations are sufficiently similar, but not easily recognisable from trial to trial. This allows the necessary repetition of identical stimuli across blocks without the risk of recognition and memory-related effects. To provide a wide range of moral scenarios and variation in the respective judgments, the animations showed the same morally relevant negative outcome (one agent delivering a painful shock to another agent) but varied in the causal structure leading to the shock.

This causal variation allowed the explorative comparison of the neural processing of causation by action with causation by omission. Such manipulation of action initiation usually elicits the omission effect, leading to diminished responsibility and intention ratings for negative outcomes caused by omissions compared to direct actions (Ritov & Baron, 1990; Spranca, Minsk, & Baron, 1991). The exact mechanism of the omission effect remains debated. However, neuroimaging evidence suggests that ascribing more blame to active than passive moral harm is an automatic mechanism that requires conscious effort to be overcome (Cushman, Murray, Gordon-McKeon, Wharton, & Greene, 2011). The second introduced causal factor was outcome preventability, which refers to whether the agent's action or inaction had the full causal power to realise or prevent the outcome. Preventable outcomes have been associated with increased responsibility ratings compared to non-preventable outcomes, i.e., where the outcome would have happened anyway even if the agent had behaved otherwise (Wells & Gavanski, 1989; Bonicalzi et al., 2022). The direct comparison of neural activation for responsibility and intention judgments during a broad variety of visually depicted scenarios will help to clarify the relation between the two investigated cognitive processes during moral evaluations.

2. Methods

2.1. Participants

The sample included 30 healthy, right-handed volunteers (17 female) aged between 19 and 49 years (M=26, SD=7) with normal or

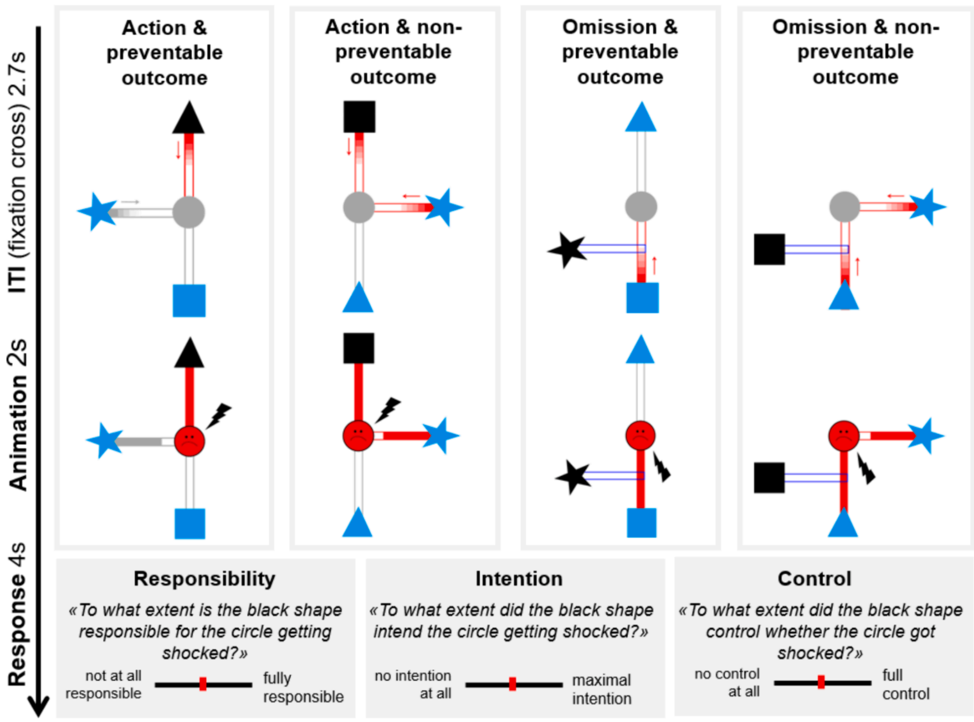

Fig. 1. Schematic illustration of the task structure. Shock events were presented in four different causal scenarios, followed by judgments of responsibility, intention or control (counterbalanced across blocks).

corrected-to-normal vision. All were fluent English speakers with no history of neurological disorders. Participants were recruited via online advertisement and gave written informed consent before scanning. They were paid £20 for participation. Testing was conducted at Brunel University London, and the study was approved by the local ethics review board and carried out in accordance with the Declaration of Helsinki.

2.2. Stimuli and procedure

The stimuli consisted of short (2sec) non-verbal animations showing different geometrical shapes interacting by sending harmful (red), neutral (grey) or inhibitory (blue) impulses via pipes that connected them in varying arrangements (see Fig. 1 for schematic examples and Supplementary materials for example animation). The stimuli were designed to depict different connections between agents while keeping contextual and semantic information to a minimum. Four different causal conditions were employed: (1) *Action & Preventable Outcome*: the black target shape intervenes to directly produce the outcome, which would not have occurred without its intervention, i.e., the target shape could have prevented the outcome; (2) *Omission & Preventable Outcome*: the black target shape does not prevent the outcome, which would not have occurred with its intervention, i.e., the target shape could have prevented the outcome; (3) *Action & Non-Preventable Outcome*: the black target shape intervenes to directly produce the outcome, which would have occurred even without its intervention, i.e., the target shape could not have prevented the outcome because another shape had already initiated a similar intervention towards the central shape; (4) *Omission & Non-Preventable Outcome*: the black target shape does not prevent the outcome, which would have occurred even with the intervention, i.e., the target shape could not have prevented the outcome (see Fig. 1). An animated example video can be found in the Supplementary Materials.

We presented the causal structures in varying spatial arrangement of the contributing shapes. Participants were familiarised with the animations prior to scanning in a 5-minute introduction video (available in Supplementary materials). They learned that the central shape (a grey circle) was always the passive receiver, while the surrounding light blue shapes were active and could send impulses to the circle. Participants were instructed to rate the active shape presented in black (one per trial). Depending on the colour of the pipe to which a light blue shape was attached, it could act differently:

(1) *Red shock*: Using a red pipe, a shape could send a painful stimulus to the circle, the circle then turned red, which represented pain.
(2) *Grey neutral*: Grey pipes could be used to transmit a grey impulse, which was neutral and did not affect the circle in any way (used for balancing the amounts of low-level visual motion).
(3) *Blue inhibition*: Some active shapes were attached to blue pipes which linked to another red pipe rather than directly to the circle. By using such blue pipes, the respective shape could transmit a blue impulse, which blocked any approaching red impulse of the pipe to which it was attached.

The introduction video further clarified that all involved shapes were ignorant regarding the activity of the other shapes, except for shapes with blue pipes, which had information about whether the red pipe to which they were attached was active or not. This experimental paradigm has previously been employed in a large online sample (Bonicalzi et al., 2022) which showed that participants did not have problems understanding and following the instructions, and carrying out the required evaluations.

Our behavioural measures were third-party ratings of responsibility and intention. We further included a (non-moral) causal control condition to ascertain that the introduced causal factors of action initiation and outcome preventability were registered by participants. We expected that behavioural responsibility ratings would align with causal responsibility judgments, as previously observed in a between subject design employing the same paradigm (Bonicalzi et al., 2022). All three ratings were collected in separate counterbalanced blocks. The instruction video introduced the rating scale that participants would use in the first block. The scales of the second and third session were presented inside the scanner. For the dependent measure of responsibility, the participants were asked "*To what extent was the black shape responsible for the circle getting shocked?*" Participants answered by adjusting the slider of a visual analogue scale with the extremes (coded numerically as values between 0–100) labelled as "*not at all responsible*"/"*fully responsible*", respectively. For the dependent measure of intention participants answered the question "*To what extent did the black shape intend the circle getting shocked?*" on a scale with the endpoints "*no intention at all*"/"*maximal intention*". The question for the measure of causal control was "*To what extent did the black shape control whether the circle got shocked?*" with the endpoints labelled "*no control at all*"/"*full control*".

Within each block, all active shapes (triangle, square, star) were assigned the role of the black shape in a counterbalanced way. Behavioural rating responses were collected with a two-button response box, which the subjects were holding with both hands and pressing with their thumbs. After each animation, participants had 4 s to adjust the position of the slider, after which the final slider position was registered, and the next trial started after an inter-trial-interval of 2.7 s. Offline pilot ratings showed that this time was sufficient to express the rating. Per block (i.e., dependent measure) participants viewed 100 trials (20 per causal condition), including 20 fillers depicting successful interventions in omission trials to avoid prior expectations about the probability of a shape sending an intervening blue impulse or not. The order of the animations within each block was randomised across participants. All visual stimuli were projected on a monitor located behind the participant's head and could be seen through a mirror apparatus on top of the head coil.

2.3. fMRI data acquisition

Functional and structural imaging was acquired with a Siemens 3 Tesla Tim-Trio Scanner, located at the Combined Universities Brain Imaging Centre (CUBIC). Functional images sensitive to the BOLD contrast were obtained with a T2*-weighted gradient echo-planar imaging (EPI) sequence (TR=2500 ms; TE=31 ms; matrix size = 64 × 64; voxel size = 3 × 3 × 3 mm; FOV=192 × 192 mm). 36 axial slices were imaged parallel to the bi-commissural (co-planar with AC-PC) line. Per subject, three sessions, each one comprising 406 whole head images including 5 dummy scans at the beginning, were acquired. In addition to functional scanning, a high-resolution structural scan (T1-weighted MP-RAGE sequence; TR=1.83 ms; TE=3.00 ms; voxel size 1 × 1 × 1 mm; FOV=256 × 256 mm; 160 slices) was acquired to facilitate normalisation and localisation of functional data.

2.4. fMRI data processing

The fMRI data were processed using Statistical Parametrical Mapping (SPM 12, Wellcome Department of Imaging Neuroscience, London, UK) software, implemented in a MATLAB 2016 (Mathworks, Natick, MA) runtime environment. Single subject data were pre-processed individually. First, images were pre-coregistered to an EPI template. Functional data were realigned for correction of movement artefacts. Structural data were segmented into white and grey matter and skull structures were separated from brain tissue. These steps were conducted to facilitate the normalization of the structural image to standard MNI space (Montreal Neurological Institute, McGill, Montreal, Canada). Last steps included co-registration of the mean functional scan to the structural data to normalise them to standard MNI space and smoothing the functional data with a 6 mm full-width at half-maximum Gaussian kernel.

The pre-processed data were analysed with a general linear model (GLM) approach. Per subject and session, each trial was modelled as a boxcar function with the duration of 6 s and convolved with a synthetic hemodynamic response function. Movement parameters were entered as additional regressors of no interest. Low-frequency noise was removed by a filter with a cutoff of 128 sec and serial correlations were taken into account using an autocorrelation model AR(1). On the individual level, contrasts for each condition relative to implicit fixation baseline were computed separately per session (dependent measure: responsibility, intention, control).

At the second level, a random effects analysis was used to allow for population inference. A within subject one-way ANOVA design was selected to address the main contrast between the three rating conditions (responsibility, intention, control). Whole brain results are reported at $p < 0.05$ with FWE peak level correction and a cluster extent of > 10 voxels. Anatomical labels were retrieved using the AAL3 toolbox. Note that no parametric modulation analysis was conducted because the main goal of the study was to compare the respective neural processes underlying the *evaluation* of the different concepts of responsibility/intention/control, rather than identifying brain regions that covary with the *magnitude* eventually ascribed to the respective agents as the outcome of such evaluation process.

To further analyse the causal factors (action vs. omission; preventable vs. non-preventable outcome) and their potential interaction with the rating conditions, we conducted another explorative 3 (rating condition: responsibility/intention/control) × 2 (action initiation: action/omission) × 2 (outcome preventability: preventable/non-preventable) ANOVA by means of a flexible-factorial design.

3. Results

3.1. Behavioural results

Fig. 2A shows the behavioural results of the causal manipulation on the three rating conditions: responsibility, intention, and causal control. When rating responsibility, participants judged agents as more responsible if they committed actions compared to omissions to

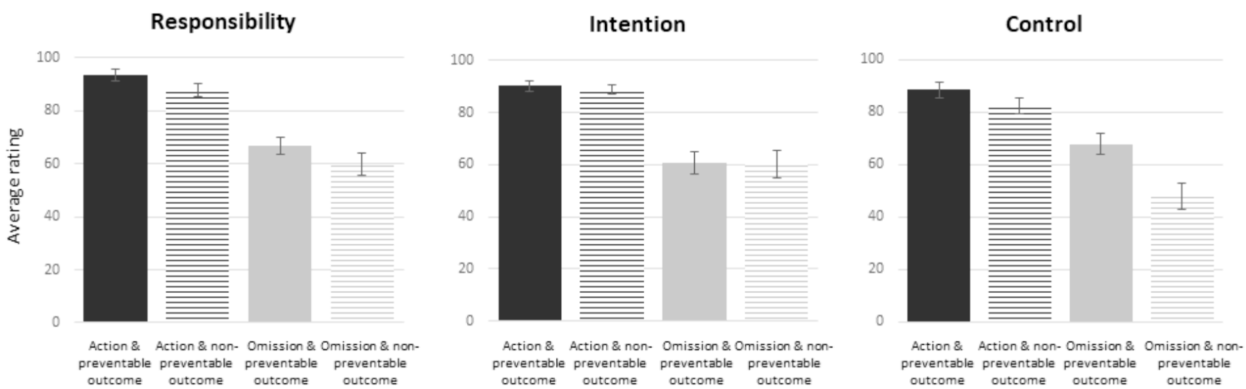

Fig. 2. Behavioural results. Behavioural ratings of responsibility, intention, and causal control, respectively. Error bars indicate 2 SE across participants.

act ($F(1,29) = 101.45$, $p < 0.001$, $\eta_p 2 = 0.78$), which replicates the well-known omission bias. Furthermore, more responsibility was attributed in trials showing preventable compared to non-preventable outcomes ($F(1,29) = 11.14$, $p < 0.01$, $\eta_p^2 = 0.28$). Intention ratings showed a main effect of action initiation with increased intention ratings for actions compared to omissions ($F(1,29) = 45.20$, $p < 0.001$, $\eta_p^2 = 0.61$), and no other significant effects. Control ratings showed a pattern comparable to that of responsibility judgments. A main effect of action initiation indicated increased control attributed to actions compared to omissions ($F(1,29) = 35.49$, $p < 0.001$, $\eta_p^2 = 0.55$). Furthermore, more control was attributed to preventable compared to non-preventable outcome trials ($F(1,29) = 28.79$, $p < 0.001$, $\eta_p^2 = 0.50$). The interaction between both factors was significant ($F(1,29) = 11.41$, $p < 0.01$, $\eta_p^2 = 0.28$), suggesting that the difference in control attributed to actions compared to omissions was larger in non-preventable ($t(29) = 6.09$, $p < 0.001$) compared to preventable conditions ($t(29) = 4.76$, $p < 0.001$). These analyses confirm that participants distinguished between different causal structures in presence of an identical outcome and are in line with our previous findings using this paradigm (Bonicalzi et al., 2022).

3.2. Imaging results

3.2.1. Moral judgments

Our main interest was the neural contrast between responsibility and intention judgments. Imaging results are reported in Table 1 and visualised in Fig. 3. Judging responsibility vs. intention showed significantly higher activation in bilateral Angular Gyri, with the right cluster located in the area corresponding to RTPJ (See Fig. 3A).

The inverse contrast did not reveal any significant clusters. When compared to non-moral causal control ratings, responsibility evaluations were associated with stronger activations in the right Cuneus, right Superior Parietal Gyrus and right Postcentral Gyrus. Intention vs. causal control ratings showed stronger activations in the right Supplementary Motor Area, right Inferior Frontal Gyrus, right Postcentral Gyrus, left Superior Occipital Cortex and right Putamen compared to control ratings. Control ratings did not show greater activations when compared with responsibility or intention, respectively.

3.2.2. Explorative region of interest analysis

According to our outlined hypothesis, regions with stronger activations for responsibility compared to intention judgments were assumed to underlie the evaluation of external, behaviour-based aspects of responsibility, rather than participants' inferences about the agent's internal mental states. It could be assumed that similar evaluations would be required in processing the causal factors. For instance, to evaluate omissions as opposed to actions, the simulation of the unrealised action – to which omission is the contrast – might be required. Similarly, appreciating the preventability of an outcome might require a comparison between factual and hypothetical actions. We thus conducted an explorative region of interest analysis to investigate if our main finding of bilateral AG would show sensitivity to any of the causal factors. For this purpose, the exported contrast estimates of both AG clusters from an explorative 3 × 2 × 2 full-factorial design were subjected to a repeated measures ANOVA.

The left AG cluster showed a main effect of rating condition ($F(2,58) = 4.9$, $p < 0.05$, $\eta_p{}^2 = 0.14$). Follow-up t-tests replicated the significant difference between responsibility and intention ratings ($t(29) = 3.3$, $p < 0.01$). Interestingly, the cluster also showed a strong effect of action initiation, with stronger activations in omission compared to action trials ($F(1,29) = 32.1$, $p < 0.001$, $\eta_p{}^2 = 0.53$). No further effects or interactions were significant. Similarly, the right AG/RTPJ cluster also showed a main effect of rating condition ($F(2,58) = 6.0$, $p < 0.01$, $\eta_p{}^2 = 0.17$). Here activation during responsibility judgments was significantly stronger compared to both intention ($t(29) = 3.1$, $p < 0.01$) and causal control evaluations ($t(29) = 2.9$, $p < 0.05$) (see Fig. 3B). An effect of action initiation

Table 1
Imaging results. Brain regions activated by the contrasts of interest (peak-level threshold $p < 0.05$, FWE-corrected, cluster extent > 10 voxels). Abbreviations: H=Hemisphere, k = cluster extent in voxel, L=left, R=right.

—	MNI coordinates			k	T	p_{FWE} peak-level	Anatomical label
	x	y	z				
Responsibility > Intention							
	−39	−67	11	10	5.9	< 0.001	**L Angular Gyrus**
	51	−70	20	11	5.8	< 0.001	**R Angular Gyrus (TPJ)**
Intention > Responsibility							
	No significant clusters						
Responsibility > Control							
	12	−91	5	14	6.5	< 0.001	**R Cuneus**
	21	−52	65	15	5.7	< 0.001	**R Superior parietal gyrus**
	21	−34	59	12	5.5	< 0.001	**R Postcentral gyrus**
Control > Responsibility							
	No significant clusters						
Intention > Control							
	12	17	62	10	5.9	< 0.001	**R SMA**
	51	38	−1	12	5.8	< 0.001	**R IFG**
	15	−31	53	15	5.8	< 0.001	**R Postcentral gyrus**
	−12	−88	8	12	5.8	< 0.001	**L Superior occipital cortex**
	27	−4	−7	11	5.4	0.002	**R Putamen**
Control > Intention							
	No significant clusters						

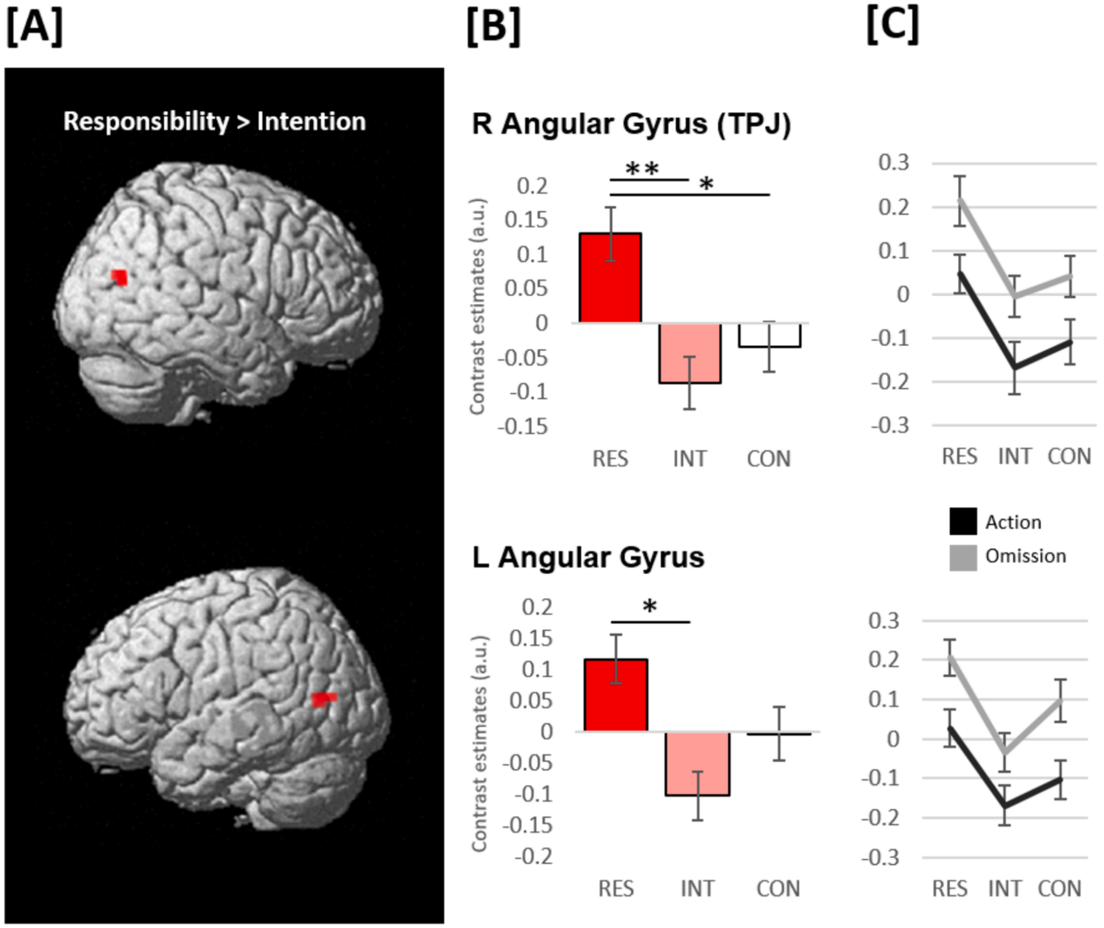

Fig. 3. Imaging results. (A) Significant activations projected on single brain surface. (B) Graphs show contrast estimates (arbitrary units, a.u.) for both significant clusters across judgment conditions. (C) Contrast estimates (arbitrary units) show main effect of the causal factor of action initiation, with stronger activations for omission compared to action.

emerged again, with stronger activations during omission compared to action trials ($F(1,29) = 16.1$, $p < 0.001$, ${\eta_p}^2 = 0.36$) (see Fig. 3C). No further main or interaction effects were significant. This analysis suggests that the bilateral Angular Gyri were not only activated more strongly by the evaluation of moral responsibility, but were also involved in causal evaluations, with stronger activations during the processing of omission compared to action trials.

3.2.3. Causal conditions

The study was designed to allow a direct comparison between the different rating conditions employed in the three blocks of the experiment (responsibility, intention, control). However, to further explore the relationship between rating conditions and the two causal factors in a post-hoc region of interest analysis, a flexible-factorial 3 × 2 × 2 ANOVA was computed. The exploratory nature of this model was due to the expectedly low power to detect a significant effect of either of the factors. Each cell ultimately contained only twenty trials to capture elaborate, high-level cognitive processes, while correlations between factors could not be excluded. As such, the exploratory model failed to show significant main effects for the factors rating condition and outcome preventability. No interaction effects were significant. Only the condition of action initiation showed a significant main effect, with omission trials showing stronger activations in bilateral Precuneus, left Parietal Inferior Gyrus and left and right Insula compared to action trials. Action trials did not show any excess activation compared to omission. A results table and figure of these results can be found in the Supplementary Material. Overall, none of the regions which showed sensitivity to causal manipulations overlapped with the main findings of rating condition on the whole brain level.

4. Discussion

The present study set out to investigate the neural basis of third-party moral judgments. We were particularly interested in the brain areas underlying the evaluations of responsibility and intention, and their relation. It is widely assumed that to judge responsibility, one has to take into account both the agent's intention (i.e., the mental states, motives and beliefs that guide their actions) as well as the relevant external factors involved in realising these actions (Mikhail, 2007; Krueger & Hoffman, 2016).

Two hypotheses were tested. Under the assumption that intention computation is a sub-set of establishing responsibility, no stronger activation for intention compared to responsibility judgments was expected. On this view, the neural activation that underlies intention computation should equally occur in both judgment tasks, and thus cancel out in the contrast. This hypothesis was tested by identifying brain regions with significantly stronger activations during intention compared to responsibility ratings. Reversely, however, more exploratively, brain areas that would show stronger activations for responsibility compared to intention judgments could be assumed to represent the complement of intention computation (i.e., the "external situation"), as assumed by conceptual theories of moral reasoning. A particularly interesting case would emerge if the regions showing such excess activation for responsibility evaluations would correspond with areas previously associated with implicit or explicit intention computations in the context of responsibility judgments within the social cognition network. This would question the previous assumption that such activations reflect ongoing intention computation and invite a clearer delineation of the cognitive processes involved in both types of judgments.

4.1. Intention vs. responsibility

Our results did not reveal stronger neural activation during intention compared to responsibility evaluations. This neural finding is in line with the conceptual view that intention computation is fully realised within responsibility attribution. This means that the neural processes related to evaluating an agent's intention or motivation to harm similarly occur when rating the agent's moral responsibility for the harm they caused. While this finding elegantly fits into prominent theoretical frameworks of moral cognition, an obvious methodological caveat is that the absence of evidence does not correspond to the evidence of absence. As such, our study might have lacked statistical power to detect a neural effect. The threshold employed to reduce false positive results could have reduced sensitivity for detecting a weak signal, leaving the possibility of false negative results. Nevertheless, the present null result is an important first approximation of the neural contribution of intention evaluation in the context of third-party moral responsibility judgments.

4.2. Responsibility vs. intention

Beyond incorporating the agent's intentions, responsibility evaluations take into account the agent's actual actions (or inactions) and their respective outcomes, how these are realised in the physical world, and how they relate to current social and moral norms, all aspects that we subsumed under the term "external situation", as opposed to internal, mental states. Thus, responsibility judgments strongly reflect the way the world actually works, and not just the contents of the agents' mind. Our second hypothesis exploratively investigated these external aspects of moral evaluations by identifying areas with stronger activations during responsibility compared to intention ratings. This contrast revealed stronger activation in sub-regions of the social cognition network, namely bilateral Angular Gyri, with the right cluster corresponding to RTPJ.

At first glance, this result replicates the commonly observed involvement of social cognition regions related to belief and intention computation in the context of moral judgment. However, these areas showed significantly stronger activation during responsibility compared to intention judgments. Therefore, it can be argued that they underlie cognitive processes that distinguish responsibility and intention evaluation, rather than serving both. In particular, a direct contrast between the rating conditions makes it highly unlikely to assume that these areas are involved only in explicit or implicit intention computations. Had these activations been related to intention computation only, then they should have been present both in intention and responsibility judgment and thus not appeared in the contrast comparing the two processes.

Instead, the stronger activation found in these areas during responsibility compared to intention judgment could reflect some additional cognitive component involved in responsibility judgment, but not in intention judgment. For example, do the excess activations for responsibility reflect some additional computations, which are qualitatively different from mentalising, or might these additional computations still be linked to the representation of others, and thus be social in nature? Some previous theoretical work suggests the latter. Indeed, responsibility is a social construct and can only be appreciated with reference to social group norms of condemnation or blame (Adolphs, 2009). Thinking from the viewpoint of a whole social group might draw on these areas to an extent that goes beyond individual belief and intention computation (Walter et al., 2004).

However, a second line of argument suggests that responsibility judgments do indeed involve additional processes that are qualitatively different from social cognition. In particular, responsibility judgments also require evaluations of the external causal events that bring about the morally negative outcome (Blos, Chatterjee, Kircher, & Straube, 2012; Woods et al., 2014). The default mode network – which strongly overlaps with the social cognition network (Mars et al., 2012) – has been involved in tracking causality, in terms of the hidden dependencies that go beyond the observable visual interactions between agents (Blakemore & Decety, 2001; Satpute et al., 2005). In turn, causal evaluations can be interpreted in terms of extracting and mentally simulating latent and unobservable relations, a process that has been conceptualised as a form of hypothetical or even counterfactual thinking (Kulakova et al., 2013; Van Hoeck et al., 2014, Van Hoeck, Watson, & Barbey, 2015). The default mode network has further been implied in the integration of different co-occurring computations for a coherent evaluation of a complex, multifaceted situation (Raichle 2015).

Our design included a non-moral causal control rating condition. This allowed us to clarify that RTPJ activation for responsibility judgments could not be attributed to causal reasoning alone, since it showed significantly higher activations for responsibility ratings compared to causal control evaluations in our region of interest analysis. Nevertheless, responsibility judgments might still pose *increased* causal processing demands. A study that used visual stimuli of balls, either colliding in a physical manner comparable to a game of marbles (physical causality) or by attracting each other over a distance comparable to social agents (social causality), showed that the direct comparison of both causal conditions revealed very similar RTPJ activations [45—66 22] for the social evaluations (Wende et al., 2013). Even though in this study only social but not physical causality required the participants to anthropomorphise the observed balls to make a judgment, the concept of social as opposed to physical causality might still offer an interesting distinction to describe different forms of complex rule-based evaluations. The notion of such causal factors resonates with theoretical assumptions that responsibility judgment goes beyond intention computation because it not only takes into account the contents of the agents' mind but also reflects the complex way the world actually works.

4.3. Region of interest analysis

What makes the observed Angular Gyrus activations even more intriguing is the finding that they both showed sensitivity for the causal manipulation of action initiation, with stronger activations for omissions compared to actions, even if this effect was only evident in the region of interest analysis. What could be the conceptual similarity between responsibility judgments and omissions as compared to intention ratings and cases of action, respectively? One tentative answer could be that both require increased processing demands for mentally representing different (and partly opposing) perspectives (Arora et al., 2020; Schurz et al., 2013). Responsibility judgment is a process that requires the evaluation of an agent's behaviour from the point of view of a social group. As such, responsibility seems to be a triadic relation (x is held responsible for y by z), while intentional relations are two-place (x intended y). As such, the additional cognitive component of responsibility, as compared to intention, need not solely be a socio-normative content, but can stem from its more complex semantic representation. Similarly, appreciating omissions might be argued to require the simultaneous evaluation of two states of affairs, i.e., the actual inaction and the alternative, yet unrealised, action. Compared to omissions, direct actions can be fully appreciated by the outwardly observed behaviour, with no need to simulate or hold in mind further alternatives. Another interpretation of higher AG activation in the evaluation of omission might stem from a lower degree of agency attributed to omitting compared to acting. In first-person judgments of agency, higher AG activation has been linked with the perception of low or even absent agency ('*Wasn't me!*'), establishing the basis for a functional self-other distinction (Chambon et al., 2013; Voss et al., 2017). It is possible that a similar judgment of non-agency, potentially followed by the detection of alternative agents, occurs during evaluation of third-party omissions. The need to distinguish between the relative contribution of different agents might also be relevant to establish their respective responsibility, as opposed to rating intention, where the agents' mere desire to do harm is more relevant than their actual contribution to doing so.

4.4. Processing causal structure

Besides contrasting different types of judgments, the present study introduced different causal sequences of observed unfolding harm events. These were varied along two factorial dimensions. The factor of action initiation determined whether a shock was caused by a direct action vs. by an omission to act and prevent it. The factor of outcome preventability manipulated whether the evaluated agent could have potentially prevented the outcome, or whether the outcome would have happened anyway independently of the agent's behaviour. At the whole brain level, the only significant finding with respect to the manipulation of the causal structure was increased activation in the bilateral Precuneus, left and right Insula and left Inferior Parietal Gyrus in omission compared to action trials, which is in line with previous studies of the omission effect (Cushman et al., 2011). Actions are characterised by the initiation of an observable physical or motoric impulse, while omissions are in fact non-events, as no observable activity occurs. To evaluate to what extent the agent's omission had an impact in the actual world, one has to consider possible but unrealised alternatives rather than simply observe unfolding factual events. Such mental imagination or simulation of potential events has previously been associated with activation in the Precuneus (Cavanna & Trimble, 2006). Comparable activation in the Precuneus has further been observed in reasoning about mental and hypothetical content (van Hoeck et al., 2014). The absence of significant activation of other regions previously observed for omission vs. action might stem from the reduced visual, semantic and social complexity of our stimuli. Participants did not have to hold complex social situations in mind, which might account for the reduced activity in frontal and temporal regions (Cushman et al., 2011). The observation that omissions were associated with increased brain activity in comparison to actions, while the inverse contrast remained insignificant, is in line with accounts that explain the omission bias in terms of increased difficulty to process the implications of omissions compared to actions (Byrne & McEleney, 2000).

4.5. Limitations of the present study

This study set out to contribute to a central debate within moral neuroscience regarding the relationship between responsibility and intention computations and their respective neural correlates. A central goal was to obtain both types of judgment for the same stimulus items, and for the same participants, so that any differences between judgment types could be unambiguously attributed to high-level cognitive processes related to factors such as moral, social, or causal thinking, rather than some confounding factor. This goal is inherently challenging, and both cognitive processes are potentially realised as a complex neural process within space and time, to which the reported activations can only be approximations. Of course, we cannot entirely exclude the possibility of potential confounds. For example, the Angular Gyrus activation of our central contrast could potentially represent increased general attentional resources required to fully appreciate a complex causal and social situation. Such increased working memory demands might be necessary to process the unrealised possibilities in omission and the social implications of responsibility judgments. We cannot rule out such low-level explanations, even though it would be still in line with our favoured interpretation that AG activations stem from a form of increased causal complexity.

In a similar vein, our causal control judgment condition might be argued to be very similar to responsibility evaluation, assuming the latter was interpreted in terms of causal rather than moral responsibility. Nevertheless, significant neural differences between both conditions emerged. Interestingly, the occipital cluster with stronger activation for responsibility judgments compared to control resembles the central finding of a study that compared counterfactual to hypothetical reasoning in visual and auditory stimuli (Kulakova et al., 2013). As such, this pattern nicely fits into our discussion of increased demands of mental simulation and representation when establishing moral responsibility as opposed to mere causal control.

Ultimately, however, it could be argued that all our rating conditions lacked substantive moral content, as participants merely rated artificial interactions of geometrical shapes rather than real human agents. Similarly, to judge responsibility, participants were not explicitly asked to consider the agent's blameworthiness, appropriate social reproach, or anticipated guilt. Those considerations may have emerged automatically in participants' judgments of responsibility but were not explicitly suggested. Social considerations of this kind could be assumed to involve brain areas for social cognition to an even higher extent. It can be questioned whether such stimuli were sufficiently emotionally engaging to elicit moral cognition comparable to the moral assessment of human behaviour. We need to admit that this might present a trade-off at the cost of ecological validity. However, even these minimalistic stimuli and instructions led to the activation of social cognition areas. Further, the activations we observed cannot be parsimoniously explained by any need to anthropomorphise the shapes featuring in our stimuli (Wheatley et al., 2007), as it is unlikely that our rating conditions invoked such need to a different extent. This non-obvious and uncued activation of social brain networks in our responsibility judgments favours our interpretation of the underlying cognitive demands of complex social and causal processing.

4.6. Summary

Taken together, our results suggest that evaluating the responsibility of an agent for harming another agent seems to include the cognitive processes that are involved in establishing the intention of this agent to do harm – but not only these processes. As such, intention ratings seem to be a subset of responsibility ratings, both at a conceptual and neural level. This is the first direct demonstration of this neural relationship and supports prominent theoretical assumptions which until now were based on indirect evidence.

Our study further supports that there is more to responsibility attribution than mere intention attribution. Specifically, attribution of responsibility includes both intention computation as well as the consideration of whether the agent's behaviour in fact made any difference to the course of events. Further, attribution of responsibility includes a range of social, normative and moral considerations that are not strictly related to computing intentions. In our neuroimaging data, responsibility judgments showed increased neural activation in bilateral Angular Gyri (including RTPJ), over and above activations of these regions in intention computation. These regions have previously been linked to intention computation in the context of moral reasoning. However, our findings call into question whether previously observed activations in these areas were indeed related to intention computation. Instead, they could have reflected some cognitive operation that correlates with intention computation, but in fact are more reliably linked to responsibility judgment than to intention judgment. We provided a discussion of the conceptual significance of these activations of the social cognition network, or default-mode-network, respectively. Our tentative but favoured conclusion is that RTPJ activation, in particular, might represent the increased demands of social causal evaluations, in particular related to mentally representing a higher number of social agents and ramifications in the case of responsibility judgments, compared to intention judgments. We see a connection between this increased causal complexity and the need to mentally simulate non-actual, counterfactual events when evaluating cases of omissions compared to direct actions. As such, the Angular Gyrus showed sensitivity to both moral and causal processes.

Our results are compatible with the assumption that moral judgment can potentially be reduced to basic, non-strictly moral, cognitive functions such as causal and mental state reasoning. This is in line with the idea that the cognitive basis for moral judgments might have evolved by recycling ontogenetically older neural structures (Adolphs, 2009; Decety & Lamm, 2007; Yoder & Decety, 2017). Future research is required to evaluate if our findings can be generalised beyond negative harm scenarios, and for instance hold for evaluations of morally positive or even non-moral outcomes. The assumption that there is no hard line between moral and non-moral events resonates with the philosophical notion that moral phenomenology is not unified, which might be traced back to the way our brains have evolved (Sinnott-Armstrong, 2008). Furthermore, extending the present findings towards more fine-grained analyses of the distributed neural patterns (e.g., by means of MVPA), which the present univariate approach cannot provide might be a promising next steps towards a better understanding of the differences and similarities of the two investigated cognitive processes.

5. Conclusion

In the present study, we used a novel, language-free experimental paradigm to investigate the neural basis of third-party responsibility and intention judgments in the context of moral harm scenarios. Responsibility judgments for morally relevant events are commonly assumed to rely both on the evaluation of the agent's intentions and motives, as well as the evaluation of external circumstances of the harmful action and its outcomes. Our results suggest that evaluating the intention of an agent to do harm does not show stronger activations compared to rating the responsibility of this agent. This first result suggests that on a neural level, explicit intention judgments might be a subset of responsibility evaluations. Furthermore, the Angular Gyrus (including a region corresponding to the RTPJ) was more strongly activated by responsibility judgment than by intention computation. This second result questions the assumption that previously observed activations of these regions were primarily driven by mentalising processes associated with intention evaluation. Other processes, such as counterfactual and causal thinking may also be involved. Our findings are relevant to neuroscientific accounts of morality, by clarifying the relation between the cognitive processes of responsibility and intention computation during moral evaluation.

Funding

The work was supported by the Summer Seminars in Neuroscience and Philosophy, Duke University. S.B. was further supported by the 'Subjectivity, Agency and Social Cognition' Grant to P.H. from the John Templeton Foundation, by the PRIN [grant 20175YZ855] from the Italian government, and by a fellowship at the Paris Institute for Advanced Study (France), with the financial support of the French State, programme "Investissements d'avenir" managed by the Agence Nationale de la Recherche [ANR-11-LABX-0027-01 Labex RFIEA+]. P.H. and S.B. were additionally supported by a AHRC Science in Culture grant to P.H. [Award number: 162746]. P.H. and E.K. were further supported by a European Research Council Advanced Grant [HUMVOL 323943] to P.H. P.H. was further funded by a Chaire de Recherche Jean D'Alembert Paris-Saclay/Institute of Advanced Studies, Paris, and benefitted from a fellowship at the Paris Institute for Advanced Study (France), with the financial support of the French State, programme "Investissements d'avenir" managed by the Agence Nationale de la Recherche (ANR-11-LABX-0027-01 Labex RFIEA+). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

CRediT authorship contribution statement

Eugenia Kulakova: Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sofia Bonicalzi:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Adrian L. Williams:** Writing – review & editing, Validation, Supervision, Software, Resources, Methodology, Data curation. **Patrick Haggard:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Data availability

Data will be made available on request.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.concog.2024.103762.

References

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*, 693–716.
Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556.
Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24*(9), 1755–1762.
Arora, A., Pletzer, B., Aichhorn, M., & Perner, J. (2020). What's in a hub? – Representing identity in language and mathematics. *Neuroscience*.
Bara, B. G., Ciaramidaro, A., Walter, H., & Adenzato, M. (2011). Intentional minds: A philosophical analysis of intention tested through fMRI experiments involving people with schizophrenia, people with autism, and healthy individuals. *Frontiers in Human Neuroscience, 5*, 7.
Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*(8), 561.
Blos, J., Chatterjee, A., Kircher, T., & Straube, B. (2012). Neural correlates of causality judgment in physical and social context—the reversed effects of space and time. *Neuroimage, 63*(2), 882–893.
Boccia, M., Dacquino, C., Piccardi, L., Cordellieri, P., Guariglia, C., Ferlazzo, F., & Giannini, A. M. (2017). Neural foundation of human moral reasoning: An ALE meta-analysis about the role of personal perspective. *Brain Imaging and Behavior, 11*(1), 278–292.
Bonicalzi, S., Kulakova, E., Brozzo, C., Gilbert, S., & Haggard, P. (2022). The dynamics of moral judgment: Joint role of dependence and transference causal explanations on responsibility attribution. *Philosophical Psychology, 9*.
Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron, 60*(5), 930–940.
Byrne, R. M., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1318.
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.
Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain, 129*(3), 564–583.
Chambon, V., Wenke, D., Fleming, S. M., Prinz, W., & Haggard, P. (2013). An online neural substrate for sense of agency. *Cerebral Cortex, 23*(5), 1031–1037.
Beecher-Monas, E., & Garcia-Rill, E. (2017). Actus reus, mens rea, and brain science: What do volition and intent really mean? *Kentucky Law Journal, 106*(2).
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.
Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2011). Judgment before principle: Engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience, 7*(8), 888–895.
Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*(6), 1052–1075.
Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist, 13*(6), 580–593.
Eres, R., Louis, W. R., & Molenberghs, P. (2018). Common and distinct neural networks involved in fMRI studies investigating morality: An ALE meta-analysis. *Social Neuroscience, 13*(4), 384–398.
Fischer, J. M., & Tognazzini, N. (2011). The Physiognomy of Responsibility. *Philosophy and Phenomenological Research, 82*(2), 381–417.
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.
Garrigan, B., Adlam, A. L., & Langdon, P. E. (2016). The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgments. *Brain and Cognition, 108*, 88–97.

Grafton, S. T. (2009). Embodied cognition and the simulation of action to understand others. *Annals of the New York Academy of Sciences, 1156*(1), 97–117.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.
Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin, 36*(12), 1635–1647.
Han, H. (2017). Neural correlates of moral sensitivity and moral judgment associated with brain circuitries of selfhood: A meta-analysis. *Journal of Moral Education, 46*(2), 97–113.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*(2), 243–259.
Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences, 39*(8), 499–501.
Kulakova, E., Aichhorn, M., Schurz, M., Kronbichler, M., & Perner, J. (2013). Processing counterfactual and hypothetical conditionals: An fMRI investigation. *NeuroImage, 72*, 265–271.
Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science, 17*(5), 421–427.
Maliske, L., & Kanske, P. (2022). The social connectome–moving toward complexity in the study of brain networks and their interactions in social cognitive and affective neuroscience. *Frontiers in Psychiatry, 13*, Article 845492.
Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. (2012). On the relationship between the "default mode network" and the "social brain". *Frontiers in Human Neuroscience, 6*, 189.
Mele, A., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies., 82*(3), 265–287.
Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.
Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*(2), 215.
Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.
Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience, 38*, 433–447.
Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making, 3*(4), 263–277.
Satpute, A. B., Fenker, D. B., Waldmann, M. R., Tabibnia, G., Holyoak, K. J., & Lieberman, M. D. (2005). An fMRI study of causal judgments. *European Journal of Neuroscience, 22*(5), 1233–1238.
Schurz, M., Aichhorn, M., Martin, A., & Perner, J. (2013). Common brain areas engaged in false belief reasoning and visual perspective taking: A meta-analysis of functional brain imaging studies. *Frontiers in Human Neuroscience, 7*, 712.
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.
Sevinc, G., & Spreng, R. N. (2014). Contextual and perceptual brain processes underlying moral cognition: A quantitative meta-analysis of moral reasoning and moral emotions. *PloS one, 9*(2), e87427.
Shute, S., Gardner, J., & Horder, J. (1993). *Action and Value in Criminal Law*. Oxford: Clarendon Press.
Sinnott-Armstrong, W. (2008). Is moral phenomenology unified? *Phenomenology and the Cognitive Sciences, 7*(1), 85–97.
Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *Neuroimage, 35*(3), 1378–1384.
Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*(1), 76–105.
Van Hoeck, N., Begtas, E., Steen, J., Kestemont, J., Vandekerckhove, M., & Van Overwalle, F. (2014). False belief and counterfactual reasoning in a social environment. *Neuroimage, 90*, 315–325.
Van Hoeck, N., Watson, P. D., & Barbey, A. K. (2015). Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience, 9*, 420.
Voss, M., Chambon, V., Wenke, D., Kühn, S., & Haggard, P. (2017). In and out of control: Brain mechanisms linking fluency of action selection to self-agency in patients with schizophrenia. *Brain, 140*(8), 2226–2239.
Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). Oxford University Press.
Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., & Bara, B. G. (2004). Understanding intentions in social interaction: The role of the anterior paracingulate cortex. *Journal of Cognitive Neuroscience, 16*(10), 1854–1863.
Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*(2), 161.
Wende, K. C., Nagels, A., Blos, J., Stratmann, M., Chatterjee, A., Kircher, T., & Straube, B. (2013). Differences and commonalities in the judgment of causality in physical and social contexts: An fMRI study. *Neuropsychologia, 51*(13), 2572–2580.
Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: Distinct roles for the social network and mirror system. *Psychological Science, 18*(6), 469–474.
Woods, A. J., Hamilton, R. H., Kranjec, A., Minhaus, P., Bikson, M., Yu, J., & Chatterjee, A. (2014). Space, time, and causality in the human brain. *Neuroimage, 92*, 285–297.
Yaffe, G. (2010). *Attempts: In the Philosophy of Action and the Criminal Law*. Oxford: Oxford University Press.
Yoder, K. J., & Decety, J. (2017). The neuroscience of morality and social decision-making. *Psychology, Crime & Law*, 1–17.
Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience, 7*(1), 1–10.
Young, L., & Saxe, R. (2009a). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*(7), 1396–1405.
Young, L., & Saxe, R. (2009b). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*(10), 2065–2072.