Supplementary materials

Supplementary study

This study was designed to conceptually replicate behavioral findings from the scanner task in an online vignette-based experiment with a well-powered sample. All hypotheses and associated analyses were pre-registered and are available on OSF (https://osf.io/3hq89/).

Participants

Participants were recruited online using Amazon Mechanical Turk at an approximate rate of $8/hour. The total sample consisted of 250 adults (110 female, 138 male, 2 non-binary; M_{Age} = 38.02, SD_{Age} = 12.39, $Range_{Age}$ = 19-81 years).

Procedure

Participants were asked to imagine themselves participating in a game that matched the experimental design of the original study. Each participant played through four rounds of the Card Choice game, with each round corresponding to one of the four *Agent* x *Outcome* conditions (i.e., Self-Harm, Self-Neutral, Other-Harm, Other-Neutral). Condition order was counterbalanced across all participants such that each participant always saw either the two Self *Agent* conditions first, or the two Other *Agent* conditions first to enable additional between-subjects analyses of the initial two conditions only. After each round of the game, participants rated (a) wrongness ("How wrong was [your action/the other Active player's] action?") on a 7-point scale with anchors at 1 ("not at all wrong") and 7 ("very wrong") and (b) responsibility ("How responsible did you feel [you were/the other Active player was] for the

Passive player's experience of [the noise blast/no sound]?") on a 7-point scale with anchors at 1 ("not at all responsible") and 7 ("completely responsible").

Results

The full model predicts moral wrongness judgments from fixed effects of *Agent*, *Outcome*, their interaction, and includes by-subject random intercepts and random slopes for the effect of *Outcome*. Using data from all four conditions for each participant in a within-subjects analysis, we observed no significant main effect of *Agent*, and a main effect of *Outcome* ($t(250) = 15.74$, $p < 0.001$, $d = 1.27$ [1.11, 1.43]), such that harmful outcomes are judged as more wrong than neutral outcomes.

As predicted, we observed a significant interaction between *Agent* and *Outcome* ($t(500) = -3.29$, $p = .001$). This interaction is robust across multiple exclusion criteria, including removing all participants who failed an attention check ($N = 206$, $t(412) = -2.38$, $p = .018$), and additionally removing all participants who didn't treat the scenario as morally relevant, as indicated by rating both self-harm and other-harm conditions at floor ($N = 156$, $t(312) = -2.58$, $p = .010$). Wrongness judgments were harsher for first-person harms ($M = 3.60$, S.E. = 0.13) relative to third-person harms ($M = 3.42$, S.E. = 0.13) ($t(502) = -2.613$, $p = .009$, $d = 0.11$ [0.03, 0.19]), an effect that is consistent across exclusion criteria ($N = 206$, $t(414) = -2.340$, $p = .0197$, $d = 0.09$ [0.01, 0.18]; $N = 156$, $t(314) = -2.376$, $p = .0181$, $d = 0.15$ [0.03, 0.27]). While third-person wrongness judgments ($M = 1.50$, S.E. = 0.07) were greater than first-person judgments ($M = 1.36$, S.E. = 0.07) in cases of neutral outcomes ($t(502) = 2.032$, $p = .0427$, $d = 0.09$ [0.00, 0.17]), this difference is no longer significant with the application of more extensive exclusion criteria ($N = 206$, $t(414) = 1.024$, $p = .3065$; $N = 156$, $t(314) = 1.262$, $p = .2079$).

A sensitivity analysis (estimated by simulation using the simr package, R version 3.6.3; Green & MacLeod, 2016) indicated that the 2-way interaction between *Agent* and *Outcome* could be detected at a minimum effect size 10% below the effect size observed in the present work, while retaining ~80% power (Arend & Schafer, 2019; Bloom, 1995). All fixed effects in the model were multiplied by .9, and a Monte Carlo simulation was used to compare the model above to an alternative omitting the 2-way interaction, (power = 85.20%, 95% CI = [82.85%, 87.34%], 1000 simulations).

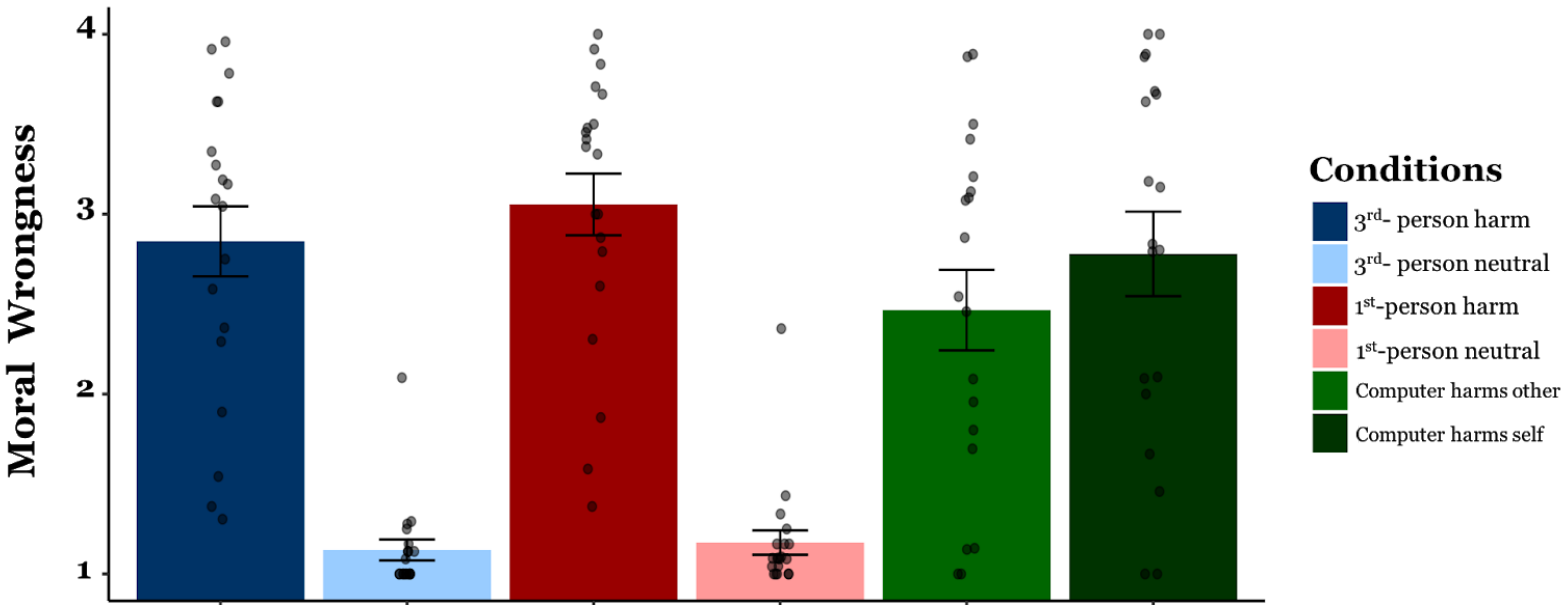
Additionally, we ran a between-subjects analysis, using data from the first two conditions that each participant saw (either the Self-Harm and Self-Neutral conditions, or the Other-Harm and Other-Neutral conditions). Using the full model without random slopes, which could not be estimated due to the limited number of observations per subject, we observed a main effect of *Agent* ($t(250) = -2.86$, $p = .0046$, $d = 0.27$ [0.09, 0.46]), such that judgements of the self were harsher than judgments of other people. There was also a main effect of *Outcome* ($t(250) = 15.27$, $p < 0.001$, $d = 1.28$ [1.11, 1.44]), such that harmful outcomes are judged as more wrong than neutral outcomes.

We again observed a significant interaction when including all participants ($t(250) = -4.42$, $p < .001$), removing participants who failed an attention check ($N = 206$, $t(206) = -4.50$, $p < .001$), and additionally removing participants who didn't treat the scenario as morally relevant ($N = 156$, $t(156) = -4.66$, $p < .001$). Wrongness judgments were harsher for first-person harms ($M = 4.06$, S.E. = 0.15) than third-person harms ($M = 3.02$, S.E. = 0.15) ($t(496) = -5.049$, $p < .0001$, $d = 0.64$ [0.39, 0.89]), which is consistent across exclusion criteria ($N = 206$, $t(412) = -5.149$, $p < .0001$, $d = 0.72$ [0.45, 1.00]; $N = 156$, $t(316) = -4.896$, $p < .0001$, $d = 0.80$ [0.48,

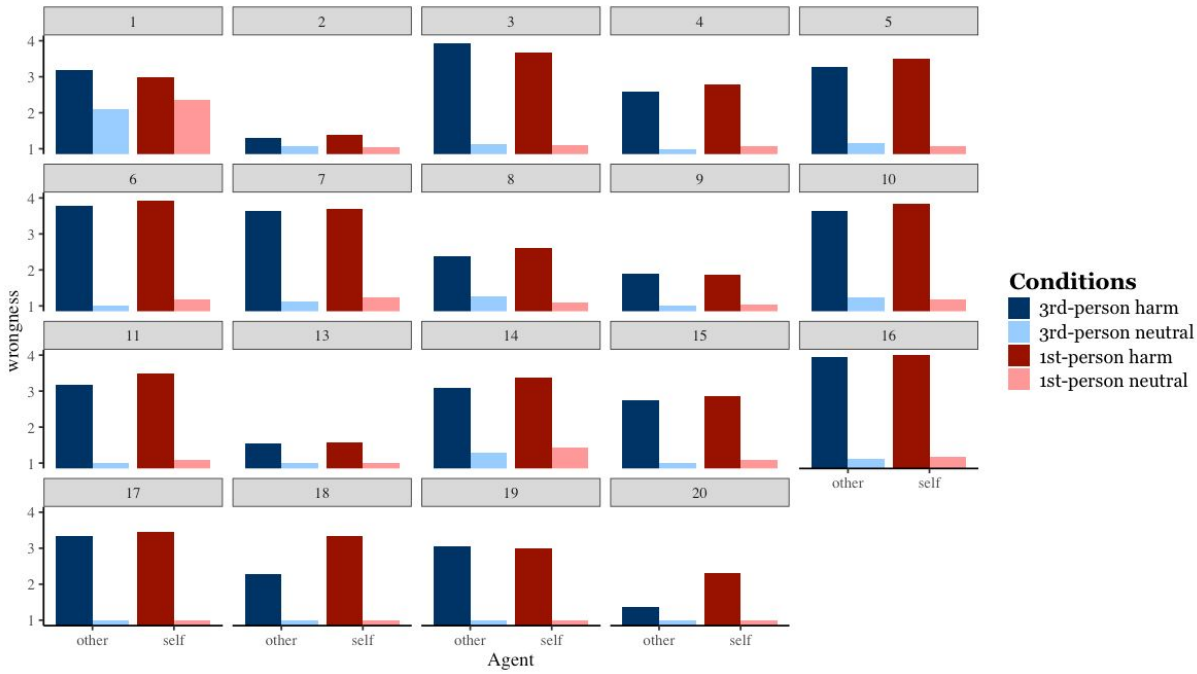
1.11]). There is no significant difference between first-person (M = 1.23, S.E. = 0.16) and third-person judgments (M = 1.41, S.E. = 0.16) for neutral outcomes (t(496) = 0.777, p = .4376), which is also consistent across exclusion criteria (N = 206, t(412) = 0.859, p = .3910; N = 156, t(316) = 1.526, p = .1281).

A sensitivity analysis, similar to the one conducted above, indicated that this 2-way interaction between *Agent* and *Outcome* could be detected at a minimum effect size 30% below the observed effect size while retaining ~80% power (power = 86.00%, 95% CI = [83.69%, 88.09%], 1000 simulations).

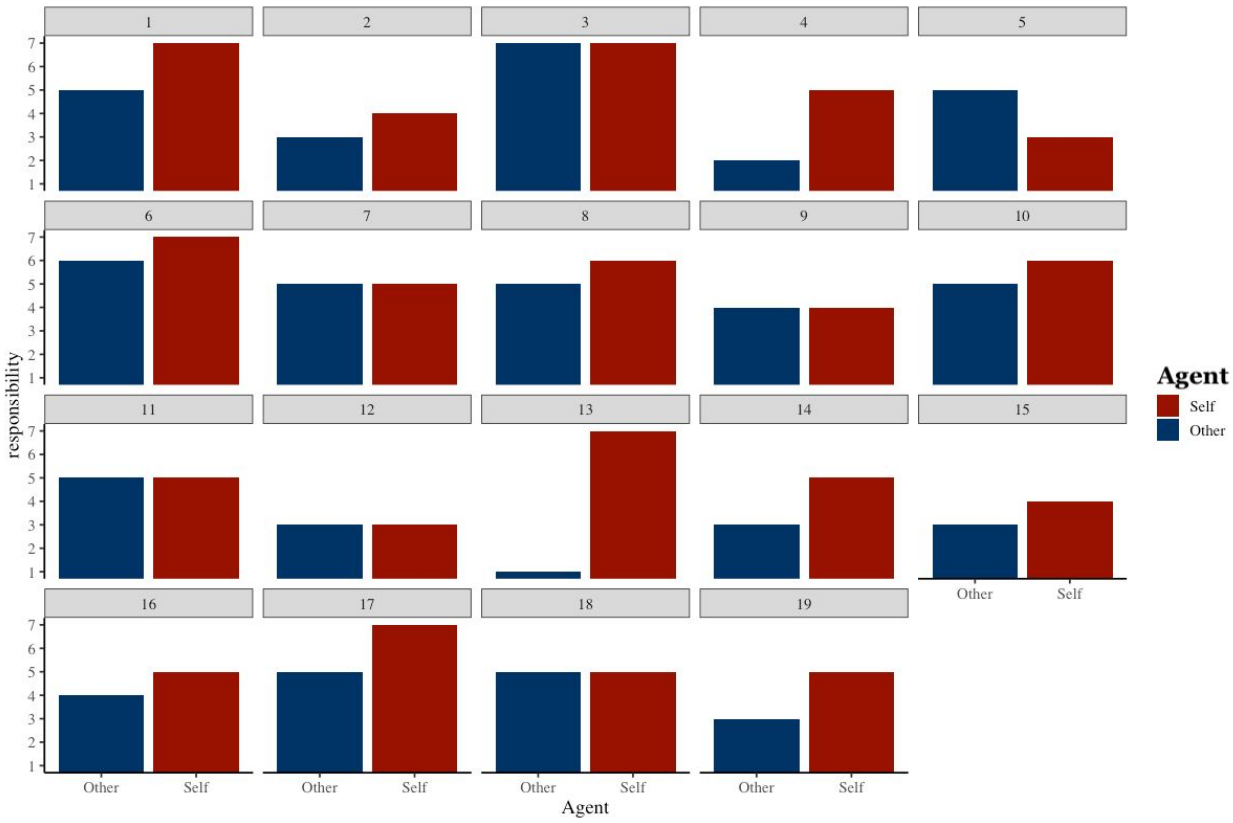
Looking at responsibility judgments, we also replicate our findings from the scanner task, with participants reporting that they themselves felt more responsible (M = 4.60, S.E. = 0.12) than they thought the other player was (M = 3.91, S.E. = 0.13) for harmful outcomes ($t(249) = 5.8$, $p < .001$, $d = 0.34$ [0.17,0.52]). A sensitivity analysis determined this sample had 80% power to detect a minimum effect size of 0.18 for a paired-sample t-test. The responsibility effect is consistent across exclusion criteria (N = 206, t(205) = 5.9250, p < .001, $d = 0.37$, [0.18, 0.57]; N = 155, t(156) = 5.8594, p < .001, $d = 0.50$ [0.27, 0.72]). As expected, wrongness judgments were positively related to responsibility judgments, for both first-person (r(250) = .5302, p < .001) and third-person harms (r(250) = .4036, p < .001).



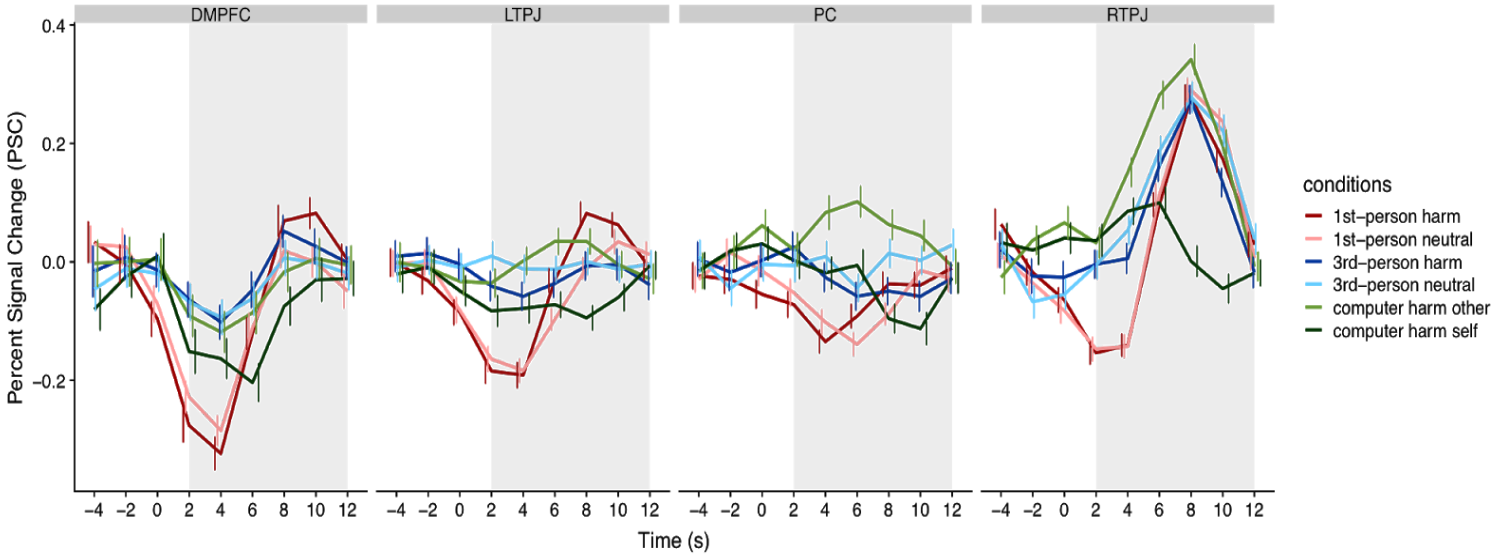
Supplementary Figure 1. Average moral wrongness for all conditions. Moral wrongness judgments were made on a scale from 1, not at all wrong, to 4, very wrong. Error bars indicate standard error of the mean.



Supplementary Figure 2. Moral wrongness judgments by subject. Moral wrongness judgments were made on a scale from 1, not at all wrong, to 4, very wrong. All subjects differentiated between harm and neutral outcomes such that harm outcomes were always rated as more morally wrong than neutral outcomes. Fifteen subjects also rated first person harm as more morally wrong than third person harm, while four subjects showed the opposite pattern.



Supplementary Figure 3. Responsibility judgments by subject. Ratings of moral responsibility were made on a scale from 1, not at all responsible, to 7, completely responsible. Twelve subjects judged themselves more responsible for harm, one subject judged the other player as more responsible, and six subjects judged self and other equally.



Supplementary Figure 4. Percent signal change (PSC) time courses for all conditions across Tom ROIs: dorsomedial prefrontal cortex (dmPFC), left temporoparietal junction (LTPJ), precuneus (PC), and right temporoparietal junction (RTPJ) . Each trial is broken up by the following stimulus bound sections: *card choice* (from t = 2-4s), *video* (from t = 6-8s), and *judgment* (from t = 10-12s). Error bars indicate standard error of the mean.

Supplementary Table 1 Peak MNI coordinates for ToM and Pain ROIs

		MNI coordinates				
Region name	t value	x	y	z	k	# of subjects
ToM ROIs						
rTPJ	10.60	57	-55	34	1498	19/19
lTPJ	12.46	-48	-61	28	1342	19/19
precuneus	13.82	-12	-49	40	862	19/19
dmPFC	4.99	6	62	28	92	19/19
Pain ROIs						
rAI	6.33	51	2	1	426	
ACC	6.51	6	26	31	322	
lSTG	5.41	-60	-34	16	400	

Supplementary Table 2. Linear mixed effects models for group ROI analyses

		card choice			video			judgment		
ROI	Fixed effects	*t*	*df*	*p*	*t*	*df*	*p*	*t*	*df*	*p*
<u>ToM network</u>										
RTPJ	Agent	4.75	17.67	**<0.001**	0.72	17.95	0.48	-1.03	15.88	0.32
	Outcome	-0.80	108.16	0.43	-0.99	679.8	0.32	-1.96	38.20	0.06
	Agent x Outcome	-0.55	18.86	0.59	-0.24	108.1	0.81	-0.97	106.9	0.33
LTPJ	Agent	5.66	18.19	**<0.001**	0.14	18.28	0.89	-1.96	119.05	0.05
	Outcome	-1.38	206.04	0.16	1.63	752.3	0.10	-0.34	869.73	0.74
	Agent x Outcome	-0.83	147.63	0.41	-1.54	18.52	0.14	-0.26	24.04	0.80
DMPFC	Agent	5.46	19.07	**<0.001**	0.53	19.53	0.60	-0.79	29.56	0.44
	Outcome	-1.28	37.97	0.21	0.60	20.45	0.32	1.22	16.53	0.24
	Agent x Outcome	0.45	23.84	0.45	0.25	61.97	0.80	-0.81	29.69	0.43
PC	Agent	2.35	17.86	**0.03**	1.29	18.23	0.21	0.86	20.03	0.40
	Outcome	-0.80	472.68	0.42	0.34	37.71	0.74	-1.53	84.79	0.13
	Agent x Outcome	0.44	50.46	0.66	-2.05	18.96	0.06	-0.97	35.97	0.34
<u>Pain network</u>										
rAI	Agent	-2.66	17.17	**0.02**	-2.94	16.30	**0.009**	-1.98	17.16	0.06
	Outcome	0.92	131.71	0.34	1.62	22.16	0.11	0.66	81.59	0.51

	Agent x Outcome	-0.53	82.07	0.60	0.49	127.4	0.63	0.38	447.95	0.70
ACC	Agent	-6.24	18.70	**<0.001**	-3.84	18.12	**0.001**	-0.29	19.15	0.77
	Outcome	-1.36	218.64	0.18	1.49	73.60	0.14	3.13	18.14	**0.006**
	Agent x Outcome	0.09	183.09	0.93	-0.64	66.36	0.53	-1.92	89.57	0.06
lSTG	Agent	-2.44	16.68	**0.03**	-1.41	16.67	0.18	-0.39	16.73	0.70
	Outcome	0.21	115.05	0.83	2.40	22.74	**0.02**	-1.65	18.28	0.12
	Agent x Outcome	-0.13	106.40	0.89	-0.76	25.14	0.45	-0.23	24.75	0.82

Note: Significant effects ($p < 0.05$) are highlighted in bold.

Supplementary Table 3. Means and Standard Deviations for DVs

	1st-person		**3rd-person**	
DV	***M***	***SD***	***M***	***SD***
Wrongness	3.03	0.78	2.85	0.85
Responsibility	5.26	1.33	4.16	1.46
Preventability	4.21	1.75	3.63	1.38
Controllability	4.37	1.30	4.16	1.26
Difficulty	3.16	1.92	3.05	1.78
Real person	6.21	1.03	5.84	1.34
Random outcomes	4.11	1.52	4.32	1.63

Note: Moral wrongness ratings were made on a scale from 1 to 4. All other DVs in the table were made on a scale from 1 to 7.

Relating individual differences in neural activity during first-person pain to self-report of experienced discomfort

The empathic pain ROIs (rAI, ACC, lSTG) were isolated by conducting a whole-brain contrast *computer harms self* (*CHS*) > *computer harms other* (*CHO*) ($p < 0.001$, uncorrected, $k > 16$, extent threshold set by permutation testing). In theory, this contrast was conducted to isolate brain regions that are involved in processing the participants' first-person experience of pain from the noise blast. To test this assumption, we conducted ROI analyses within each empathic pain ROI, investigating whether by-subject variation in the difference in PSC between *CHS* and *CHO* trials (CHS – CHO) correlates with individual differences in post-scan ratings of experienced discomfort from receiving a noise blast. These ROI analyses were conducted on averaged PSC values extracted from the *video* section, the section in *CHS* trials when participants were immediately experiencing the pain from the noise blast themselves.

We find that participants who reported experiencing greater discomfort while receiving the noise blast also tended to show a larger difference in neural activity in all empathic pain ROIs between *CHS* trials and *CHO* trials (STG: $r(16) = 0.62$, $p = 0.006$; rAI: $r(16) = 0.61$, $p = 0.007$; ACC: $r(16) = 0.56$, $p = 0.02$). These results provide strong evidence for the assumption that the STG, the right AI, and the ACC were involved in processing the first-hand pain that participants experienced from the noise blast.

Exploratory analyses relating individual differences in ToM activity to preventability judgments

In brain-behavior analyses of ToM regions, we consistently observed an unexpected *positive* relationship between neural activity in all ToM ROIs and moral judgments following harmful outcomes. In our discussion, we proposed that ToM activity in this context may reflect the consideration of morally relevant mental states *other* than intent to cause harm, such as the agent's *effort* (e.g. "how much thought did she put into deciphering the pattern?") or *ability* (e.g. "I should have learned this pattern by now"). This possibility is consistent with the 'Path Model' of blame, which suggests that, once an event is judged to be unintentional, agents are subsequently evaluated more harshly in proportion with their capacity to prevent harm (Malle, Guglielmo & Monroe, 2014). In order to provide a preliminary test of this hypothesis, we looked at the relationship between individual differences in ToM activity and post-scan judgments of preventability ("When your partner received a noise blast, did you feel that [you/the other person] could have prevented that from happening?").

When we average across ToM regions, we find that individual differences in activity over the course of the entire trial correlates with the extent to which participants felt that they themselves could have prevented harm ($r(16) = 0.53$, $p = 0.02$), but not the extent to which they felt the other person could have prevented harm ($r(16) = 0.10$, $p = 0.69$). When we break the relationship down by ToM ROIs, we find a similar pattern is significant in the RTPJ ($preventability_{self}$: $r(16) = 0.51$, $p = 0.03$), the LTPJ ($preventability_{self}$: $r(16) = 0.48$, $p = 0.04$), and a trend in the dmPFC ($preventability_{self}$: $r(16) = 0.39$, $p = 0.11$), but we do not identify this pattern in the PC ($preventability_{self}$: $r(16) = 0.07$, $p = 0.79$). This finding provides preliminary

support for the possibility that, at least in some ROIs, ToM activity was recruited in order to represent mental states that were relevant to determining the preventability of harm, and that this relationship may have been more specific to self-caused harms.

Pre-scan screening

Pain sensitivity and comfort pre-scan screening

1. In your estimation, how sensitive are you to physical pain? (1-7; not at all sensitive, extremely sensitive)

2. In your estimation, how sensitive are you to emotional pain? (1-7; not at all, extremely)

3. If you choose to participate in this study, you will experience noise blasts. We will determine your individual threshold for noise discomfort using a ramp-up procedure, starting at a low level and progressing until you feel moderate discomfort. In your estimation, are you comfortable with the prospect of receiving noise blasts? (Y/N)

4. If you choose to participate in this study, you will occasionally cause noise blasts to your partner. We will determine their individual threshold for noise discomfort using a ramp-up procedure, starting at a low level and progressing until they feel moderate discomfort. Neither you nor any other participant will ever cause *intentional* harm to your partner. In your estimation, are you comfortable with the prospect of causing noise blasts to your partner? (Y/N)

5. Do you have any known hearing issues, or ear-related issues, that might prevent you from being able to participate in this study? (Y/N/If yes, explain:)

Post-scan questionnaire

Interpersonal Reactivity Index (IRI)

EC – empathic concern, PT – perspective taking, PD – personal distress, FS - fantasy

1) I often have tender, concerned feelings for others less fortunate than me. (EC_1)

2) I daydream and fantasize, with some regularity, about things that may happen to me. (FS_1)

3) Sometimes I find it difficult to see things from the "other guy's" point of view. (PT_1) (R)

4) Sometimes I don't feel very sorry for other people when they are having problems. (EC_2) (R)

5) I get really involved in the feelings of the characters in a novel. (FS_2)

6) In emergency situations, I feel apprehensive and ill-at-ease. (PD_1)

7) I am usually objective when I watch a movie or play, and I often don't get completely caught up in it. (FS_3) (R)

8) I try to look at everybody's side of the disagreement before I make a decision. (PT_2)

9) When I see someone being taken advantage of, I feel kind of protective towards them. (EC_3)

10) I sometimes feel helpless when I am in the middle of a very emotional situation. (PD_2)

11) I sometimes try to understand my friends better by imagining how things would look from their perspective. (PT_3)

12) Becoming extremely involved in a good book or movie is somewhat rare for me. (FS_4) (R)

13) Other people's misfortunes do not usually disturb me a great deal. (EC_4) (R)

14) If I'm sure I'm right about something, I don't waste much time listening to other people's arguments. (PT_4) (R)

15) After seeing a play or movie, I have felt as if I were one of the characters. (FS_5)

16) Being in a tense emotional situation scares me. (PD_3)

17) When I see someone being treated unfairly, I sometimes don't feel very much pity for them. (EC_5) (R)

18) I am usually pretty effective in dealing with emergencies. (PD_4) (R)

19) I am often quite touched by things that I see happen. (EC_6)

20) I believe that there are two sides to every question and I try to look at them both. (PT_5)

21) I would describe myself as a pretty soft-hearted person. (EC_7)

22) When I watch a good movie, I can very easily put myself in the place of the leading character. (FS_6)

23) I tend to lose control during emergencies. (PD_5)

24) When I'm upset at someone, I usually try to "put myself in his shoes" for a while. (PT_6)

25) When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me. (FS_7)

26) When I see someone who badly needs help in an emergency, I go to pieces. (PD_6)

27) Before criticizing somebody, I try to imagine how I would feel if I were in their place. (PT_7)

Difficulty

You - When you were playing the game, how difficult was it for you to rate your actions? [1 Not at all difficult – 7 Very difficult]

Other person - When the other person was playing the game, how difficult was it for you to rate their actions? [1 Not at all difficult – 7 Very difficult]

Computer program - When the computer program was playing the game, how difficult was it for you to rate its actions? [1 Not at all difficult – 7 Very difficult]

Control

You - How much control did you feel you had over the outcome of the game? [1 No control at all – 7 Complete control]

Other person - How much control did you feel the other person had over the outcome of the game? [1 No control at all – 7 Complete control]

Computer program - How much control did you feel the computer program had over the outcome of the game? [1 No control at all – 7 Complete control]

Responsibility

You - When your partner received a noise blast, did you feel that you were responsible? [1 Not at all responsible – 7 Completely responsible]

Other person - When your partner received a noise blast, did you feel that the other person was responsible? [1 Not at all responsible – 7 Completely responsible]

Computer program - When your partner received a noise blast, did you feel that the computer program was responsible? [1 Not at all responsible – 7 Completely responsible]

Preventability

You - When your partner received a noise blast, did you feel that you could have prevented that from happening? [1 Definitely could not have prevented – 7 Definitely could have prevented]

Other person - When your partner received a noise blast, did you feel that the other person could have prevented that from happening? [1 Definitely could not have prevented – 7 Definitely could have prevented]

Computer program - When your partner received a noise blast, did you feel that the computer program could have prevented that from happening? [1 Definitely could not have prevented – 7 Definitely could have prevented]

Perceived discomfort

How uncomfortable do you think the noise blasts were for your partner? [1 Not at all uncomfortable – 7 Very uncomfortable]

Experienced discomfort

How uncomfortable were the noise blasts for you? [1 Not at all uncomfortable – 7 Very uncomfortable]

Real person

You - Did you believe that you were playing with a real person? [1 Not at all – 7 Completely]

Other person - Did you believe that the other person was playing with a real person? [1 Not at all – 7 Completely]

Computer program - Did you believe that the computer program was playing with a real person? [1 Not at all – 7 Completely]

Random outcomes

You - Did you believe the outcomes of the cards were random when you played the game? [1 Not at all random – 7 Completely random]

Other person - Did you believe the outcomes of the cards were random when the other person played the game? [1 Not at all random – 7 Completely random]

Computer program - Did you believe the outcomes of the cards were random when the computer program played the game? [1 Not at all random – 7 Completely random]