Help or flight: Neural defensive circuits promote helping under threat in humans

Joana B. Vieira & Andreas Olsson

Department of Clinical Neuroscience, Karolinska Institutet, Sweden

This is a working paper and has not been subjected to peer-review.

Abstract

Helping of conspecifics under threat has been observed across species. In humans, the dominant view proposes that empathy is the key proximal mechanism driving helping motivation in a threatening context, but little is known about how one's own defensive responses to the threat may guide helping decisions. In this pre-registered study, we manipulated threat imminence to activate the entire defensive brain circuitry, and assess the impact of different defensive responses on risky helping behaviour. Forty-nine participants underwent fMRI scanning while making trial-by-trial decisions about whether or not to help a co-participant avoid aversive shocks at the risk of receiving a shock themselves. Helping decisions were prompted under imminent and distal threat, based on the spatiotemporal distance to the administration of the shock to the co-participant. We found that greater engagement of reactive fear circuits (insula, ACC, PAG) during the threat presentation led to helping decisions, whereas engagement of cognitive fear circuits (hippocampus and vmPFC) preceded decisions not to help. Relying on representational similarity analysis, we identified how the defensive circuitry uniquely represented the threat to oneself, and the distress of the co-participant during the task. Importantly, we found that the strength with which the amygdala represented the threat to oneself, and not the other's distress, predicted decisions to help. Our results demonstrate that defensive neural circuits coordinating fast escape from immediate danger may also facilitate decisions to help others, potentially by engaging neurocognitive systems implicated in caregiving across mammals. Taken together, our findings provide novel insights into the proximal basis of altruistic responding, suggesting that defensive responses may play a more important role in helping than previously understood.

Main

Helping someone in danger (for example, by saving a person who fell on the train tracks, or running into a building in flames to rescue someone inside) may entail exposing oneself to health and life-threatening risks. Nevertheless, such helping behaviors are observed across species (Hernandez-Lallement et al., 2020; Oliner, 2004; Preston, 2013; Rand & Epstein, 2014; Waal & Lanting, 1997). Risky helping differs from other altruistic actions in that it occurs in the simultaneous presence of two highly salient cues: the distress of a conspecific in need, and a potential threat to the self. In humans, a wealth of research has been dedicated to the former, explaining how perceiving distress in others may trigger the motivation to help (Batson et al., 1987; de Waal & Preston, 2017), particularly if the helper is not under threat themselves (Preston, 2013). But virtually nothing is known about how, in a threatening situation, one's own responses to the threat may drive decisions to help. More so, animal research suggests defensive brain mechanisms may in fact be implicated in aiding or protecting conspecifics (Ben-Ami Bartal et al., 2016; Bosch, 2013; Rickenbacher et al., 2017). Understanding the neurocognitive processes underlying the motivation to both safeguard oneself and help others is critical to explain inter-individual behaviour in dangerous contexts. The overarching goal of our study was thus to determine how one's own defensive responses to threat guide decisions to help others in dangerous situations.

In humans, defensive responses to threat are graded as a function of the proximity or imminence of the threatening stimulus, paralleling predatory avoidance responses in other mammals (Fanselow & Lester, 1988; Mobbs et al., 2020). Distal and unpredictable threats are typically associated with risk assessment and intermittent anxiety, allowing for slower and more flexible escape decisions. As threat imminence increases and an attack becomes more likely, fixed and species-specific responses are triggered, such as such as freezing or, if immediate avoidance is necessary, fight-or-flight. Some behavioural reports indicate that different states along the defensive continuum may have dissociable effects on prosocial behaviour. For example, it has been demonstrated that, following acute social stress, participants behave more prosocially in economic games (e.g., Tomova et al., 2017; von Dawans et al., 2012, 2019), make more hypothetical moral decisions (Singer et al., 2017), and show greater empathy for others (Tomova et al., 2017). Importantly, it has been shown that individuals were more likely to help a co-participant avoid aversive electrical shocks when the threat of shock was imminent rather than distal (Vieira et al., 2020). This behavioural pattern was accompanied by faster reaction times and heart rate during imminent compared to distal threats, paralleling what has been found in response to imminent self-directed threats (Hashemi et al., 2019; Roelofs, 2017). Consistent with these laboratorial studies, higher danger in real life situations (captured via public surveillance footage) has been associated with higher likelihood of bystander intervention (Lindegaard et al., 2021). Taken

together, these findings suggest that defensive states triggered by high threat imminence may not only enable fast avoidance of self-directed threats, but also facilitate helping when others are under threat. Yet, the neural basis of these effects are unknown. Indeed, although a general defensive brain system has been identified in both animal and human studies, it is unknown how the activation of specific sub-circuits underlying different defensive states (e.g., freezing versus fight-or-flight) impact decisions towards others in a threatening context.

We aimed to characterize the involvement of different defensive neural responses on helping decisions under threat. To do so, we used an fMRI paradigm adapted from Vieira et al. (2020), in which participants make helping decisions at different stages of threat imminence (details in Figure 1A). Briefly, a participant is asked to decide on a trial-by-trial basis whether or not to help a co-participant (in reality, a confederate) avoid aversive electrical shocks. In each trial, the participant watches a supposedly live video-feed of the co-participant, and a visual cue signaling an upcoming shock. The participant is asked to decide whether or not to help the co-participant avoid the shock at the risk of receiving a shock her/himself. These decisions are prompted in some trials in the beginning of the trial (*distal threat*), and in others immediately prior to the shock delivery (*imminent threat*). If the participant decides not to help, the co-participant always receives a shock; if the participant decides to help, there is a fixed probability both participant and co-participant will receive a shock.

Our hypotheses were guided by previous work on neural responses to the imminence of self-directed threats. It has been shown that the response to distal threats (*i.e*., unpredictable, spatially distant, retreating, slow moving) is coordinated by so-called "cognitive fear" circuits, which include the ventromedial prefrontal cortex (vmPFC) and hippocampus (Mobbs et al., 2020; Qi et al., 2018). Conversely, imminent threats (*i.e.,* predictable, spatially close, looming, fast moving) predominantly engage "reactive" fear circuits, which include the dorsal anterior cingulate (dACC), insula, and periaqueductal gray (PAG) (Mobbs et al., 2020; Qi et al., 2018). The amygdala plays a central role in both circuits, namely by coordinating adaptive switches between defensive states as a function of threat imminence (for example, from freezing to fight-or-flight), through oxytocin-mediated communication between its central (CeA) and basolateral nucleus (BLA) (Terburg et al., 2018; Tovote et al., 2016). According to these findings, in our paradigm we expected neural activation within the full defensive circuitry to respond to the threat level of the trial (i.e., safe, 1 shock and 2 shocks), with higher engagement of brain regions previously included in cognitive fear circuits (i.e., vmPFC and hippocampus) in response to distal threats, and higher engagement of regions included in "reactive" fear circuits (i.e., insula, dACC and PAG) in response to imminent threats. At the behavioural level, prior findings (Vieira et al., 2020) showed that helping decisions were more frequent under imminent than distal trials, suggesting that the activation of "reactive" fear circuits would facilitate helping behaviour.

Therefore, we predicted that helping decisions would be associated with higher engagement of brain regions previously included in reactive fear circuits (*i.e*. amygdala, insula, ACC and PAG).

One important aspect of our paradigm was that, as in most real-life dangerous situations, the threat and the conspecific in need were simultaneously presented. To dissociate representations of threat and other's distress within the defensive circuitry, and assess their unique association with helping behaviour, we introduced a second task in the experimental protocol (Figure 1B). After the scan, participants were presented with all clips of the co-participant shown during the scan (without the visual threat cue), and asked to rate how distress (i.e., "signs of discomfort, anxiety or uneasiness") he was feeling in each clip. Additionally, participants were presented with images of the visual threat cues in both distal and imminent positions, and asked to rate how threatened they felt *themselves* when they saw those images during the scan. These behavioural ratings were used as behavioural models in a representational similarity analysis (RSA) to identify neural representations of other's distress and threat, and determine their association with helping behaviour.

It has been proposed that helping of conspecifics is a form of altruistic responding that has evolved from offspring care in mammals, is triggered by signals of distress and vulnerability, and is especially likely to occur when the helper is not threatened and has the competence to help (Preston, 2013). Some have also suggested that helping may involve some form of down-regulation (Hortensius & de Gelder, 2018) or buffering (Taylor, 2006) of one's threat responses. These distress-focused accounts presuppose a certain degree of conflict between other-oriented care and self-defence in threatening situations. However, there is some evidence in rodents that defensive responses may be required for helping and caregiving (e.g., anxious rat mothers display enhanced maternal behavior after pharmacological activation of defensive brain circuits (Bosch, 2011, 2013; Bosch et al., 2005), whereas mice bred to have low anxiety display significant defects in maternal behaviors (Taylor et al., 2017); helping behavior in rats is compromised following treatment with anxiolytic drugs that suppress defensive circuits (Ben-Ami Bartal et al., 2016)). It is thus possible that, like demonstrated in other animals, defensive mechanisms responsible for individual defense play a bigger role on helping behavior than previously understood. Based on these earlier findings, two competing predictions can be made in our study. The demonstration that neural representations of other's distress are positively associated with helping decisions would support existing distress/empathy-based explanations of helping. Conversely, the demonstration that neural representations of threat to oneself predicts helping would suggest that one's own defensive state may promote helping of others (in line with Vieira et al., 2020). In sum, we predicted that, during a threatening situation, the activation of reactive fear circuits, and the strength of the neural representation of the threatening stimulus would facilitate helping behaviour.

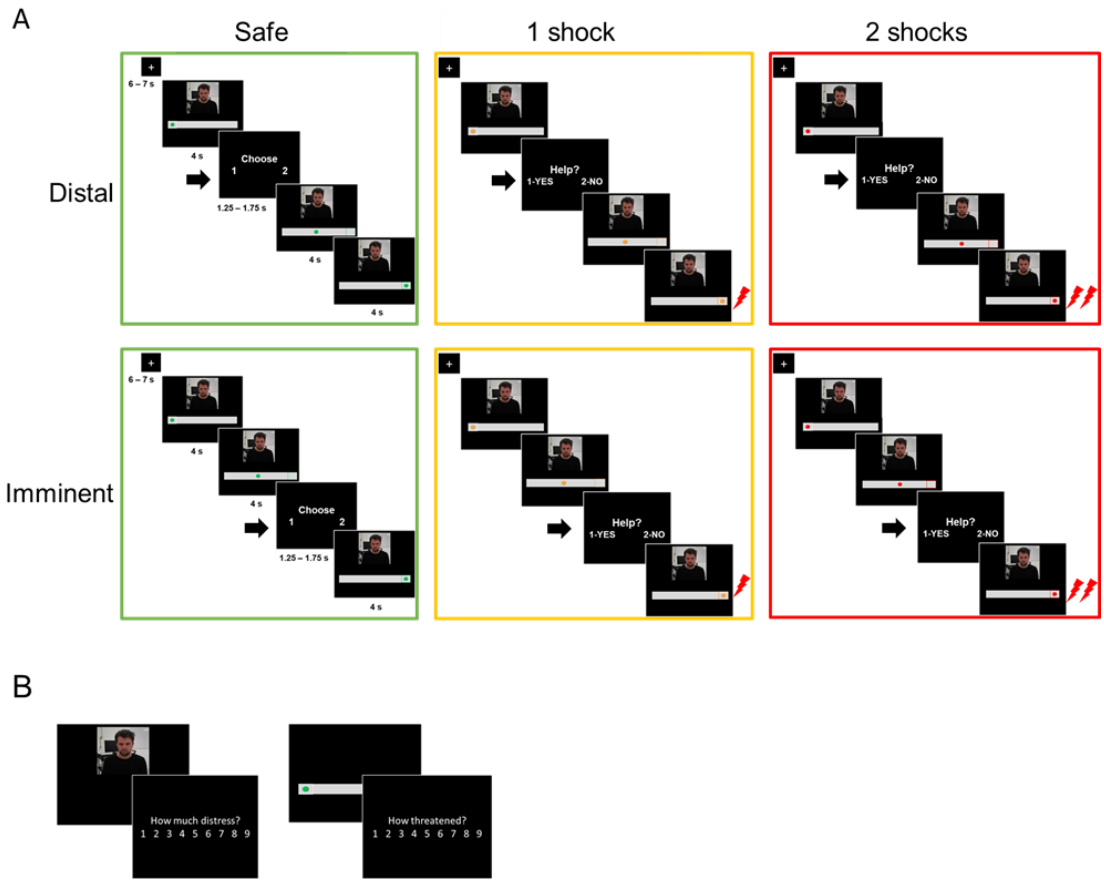

Figure 1. Outline of the experimental tasks. A. fMRI helping under threat task. Participants saw the co-participant on the screen, together with a visual cue signaling threat (an upcoming shock). There were three threat levels: safe (0 shocks, green circle), moderate threat (1 shock, yellow circle) and high threat (2 shocks, red circle). In each trial of the task, the circle started static on the left, and then moved to the right. Participants were prompted to decide whether they wanted to help the co-participant or not either in the beginning of trial (distal) or right before the moment of shock delivery (imminent). If they decided to help, there was a 70% chance both participant and co-participant would receive shocks; if they decided not to help, the co-participant would always receive a shock, and the participant would not. Decisions prompted on safe trials were to arbitrarily choose to press 1 or 2, since no shocks would be administered. B. After the fMRI task, outside the scanner, participants re-watched clips of the co-participant presented during the scan, and were asked to rate how much "discomfort, anxiety or uneasiness" he was experiencing in each clip on a 9-point scale. They were also presented images of the threat cues and asked to rate, on the same scale, how threatened they felt themselves when they saw those images during the scan.

Results

Helping behaviour

Helping behaviour was operationalised as the percentage of helping responses throughout the task. Using Generalized Linear Mixed Models (GLMMs), we found no significant effect of either threat imminence (β=.017, *se*=.014, *t*=1.24, *p*=.218), threat level (β=-.068, *se*=.053, *t*=-1.282, *p*=.205), nor a threat imminence*level interaction (β=-.099, *se*=.019, *t*=-.497, *p*=.622) on helping behaviour (Figure 2A). Despite the lack of statistically significant group-level effects of threat imminence, individual data revealed that the number of participants helping more during imminent than distal threats was higher (n=27) than those helping more during distal (n=11), or the same amount during imminent and distal (n= 11) (Figure 2B).

Previous work suggested the impact of threat imminence on helping behaviour might be especially strong in individuals with high empathic tendencies (Vieira et al., 2020). We thus checked whether there was an interaction between the empathic concern scale of the Interpersonal Reactivity Inventory (Davis, 1983) and threat imminence on helping behaviour. Results showed no significant effects of threat imminence (β=.018., *se*=.014, *t*=1.249, *p*=.215) and threat level (β=-.050, *se*=.051, *t*=-.985, *p*=.330), and no significant interaction between imminence and empathic concern was found (β=.018, *se*=.014, *t*=1.292, *p*=.20). However, a significant association emerged between empathic concern and helping behaviour (β=-.101, *se*=.041, *t*=-2.473, *p*=.02), indicating those higher in empathic concern displayed less frequent helping behaviour.

Also, to account for the possibility that decisions varied throughout the experiment, we also performed a mixed effects logistic regression on single trial dichotomous responses (help or no help), including the trial number as a fixed effect. This analysis revealed no significant effects (available in Supplementary material), indicating the individuals did not respond differently as time passed.

Finally, analysis of reaction times showed faster decisions during imminent versus distal trials (β=-173.30, *se*=12.09, *t*=-14.33, *p*<.0001), and for shock versus safe trials (β[1 shock]=-43.26, *se*=12.76, *t*=-3.39, *p*=.001; β[2 shock]=-53.35, *se*=13.43, *t*=-3.97, *p*=.0002). No significant threat imminence*level interaction was detected (Figure 2C).

Post-scan ratings of co-participant distress and threat to the self

After the scan, participants were asked to re-watch all clips of the co-participant during the scan, and rate the level of "discomfort, anxiety or uneasiness" they thought *he* was experiencing in each clip. Of note, these clips were shown without the threat cues (see Figure 1B) that were also present during the scan, in order to isolate the response to distress and threat cues. Results showed participants rated the distress of the co-participant being significantly higher during imminent than distal clips (β=.418,

se=.089, *t*=4.676, *p*<.0001), and progressively higher across the three levels of threat (1 Sh: *β*=.279, *se*=.077, *t*=3.611, *p*=.0005; 2 Sh: *β*=.559, *se*=.081, *t*=6.853, *p*<.0001; reference class was safe). No significant threat level*imminence interaction was found (Figure 2F). These results suggest the video clips used in the scan successfully portrayed subtle variations in cues of distress by the co-participant.

Participants were also presented isolated images of the threat cues used during the scanning task (namely, the green, yellow and red circles, both in the distal and imminent positions; Figure 1B), and asked to rate how threatened they felt *themselves* when they saw those visual cues during the scan. Results showed participants rated threat stimuli as more threatening (1 Sh: *β*=2.255, *se*=.297, *t*=7.598, *p*<.0001; 2 Sh: *β*=3.92, *se*=.297, *t*=13.89, *p*<.0001; reference class was safe; Figure 2G). Additionally, imminent cues were rated as more threatening, but only for 1 and 2 shocks, and not for safe trials (imminence*1 Sh: *β*=.894, *se*=.420, *t*=2.129, *p*=.034.; imminence*2 Sh: *β*=.936, *se*=.420, *t*=2.23, *p*=.027).

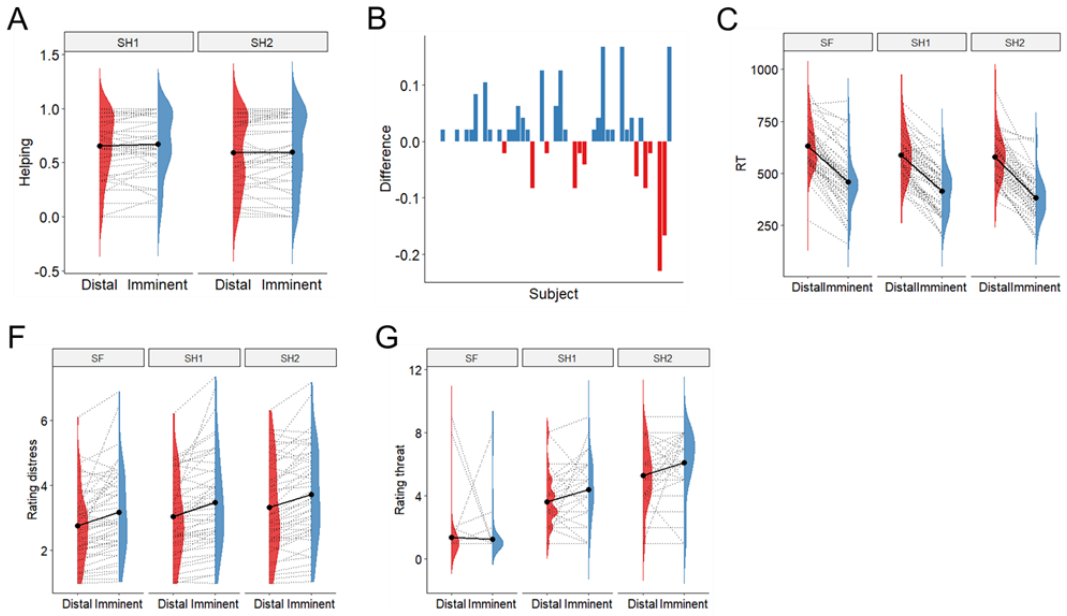

Figure 2. A. There was no evidence of differential helping during imminent and distal threats, nor during 1-shock and 2-shock trials. B. Difference between proportion of helping in imminent and distal trials (y axis) across subjects (x axis); 27 participants helped more during imminent than distal threats, 11 helped more during distal, and 11 helped the same amount. C. Responses were faster during imminent than distal trials across threat levels. D. Participants rated the co-participant's distress as higher during imminent than distal trials, and as progressively higher across the three threat levels. E. Participants reported feeling more threatened when watching 2-shock cues (red circle), followed by 1-shock cues (yellow circle) and safe cues (green circle), and when watching cues signaling imminent than distal threat.

Neural responses

Analysis of brain responses focused on a liberally defined set of brain regions that integrate the brain's defensive system, namely the vmPFC and vlPFC/IFG (Mobbs et al., 2009, 2010; Wendt et al., 2017), the hippocampus (Qi et al., 2018), the insula (Mobbs et al., 2010; Wendt et al., 2017), the ACC (Mobbs et al., 2009, 2010), the amygdala (Mobbs et al., 2009; Terburg et al., 2018) and the midbrain PAG (Mobbs et al., 2007, 2009, 2010; Qi et al., 2018; Wendt et al., 2017) (Figure 3).

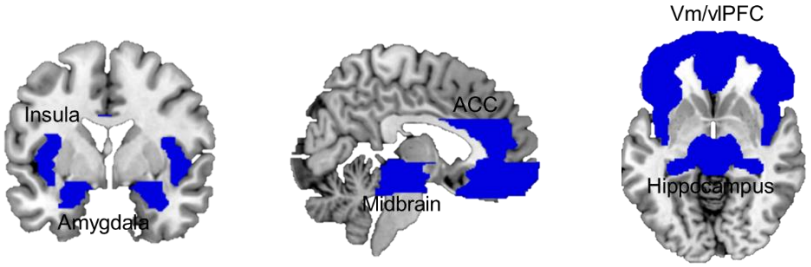

Figure 3. Combined ROI mask including bilateral ventral and lateral medial frontal cortex, dorsal ACC, insula, hippocampus, amygdala and midbrain.

a. Neural coding of threat imminence and threat level

Using a Searchlight cross-classification algorithm (12-radius sphere mm), we first identified, at the individual level, brain sites in which multivoxel patterns were distinguishable between distal and imminent threats. A group-level one-sample t test on the individual accuracy maps showed neural patterns in the bilateral insula and dorsal medial prefrontal gyrus dissociated between distal and imminent threats (*FWE* < .05, *k* > 10; Figure 4A).

We also performed Support Vector Machine regression analysis to identify sites in which multivoxel activation was linearly associated with increasing threat level (*i.e.,* from safe, to 1 shock and 2 shocks). Group-level results showed that within the amygdala, insula, ACC, hipoccampus, and regions within the orbitofrontal cortex, activation patterns tracked with threat level (*FWE* < .05, *k* > 10; Figure 4B).

Finally, we ran univariate analyses to identify differences in average activation when responding under distal and imminent threats. We performed a threat imminence (distal, imminent) by threat level (safe, 1 shock, 2 shocks) ANOVA, in which we were especially interested in identifying brain regions with a significant interaction that would indicate differential responses to imminence for threatening stimuli specifically (full ANOVA results available in Supplementary material). Brain regions displaying a significant interaction included the bilateral insula, OFC and IFG and ACC (Table 1; Figure 4D). In shock trials (both 1 and 2 shocks), the bilateral insula and ACC presented higher activation for distal compared to imminent threats, whereas the bilateral IFG and right OFC showed higher activation in imminent compared to distal threats. These results are opposite to our predictions that brain regions

246 previously implicated in reactive fear circuits (*i.e.,* insula and ACC) would be more active during
247 imminent threats, and regions implicated in cognitive fear circuits (*i.e.,* IFG and OFC) during distal
248 threats.

249 b. Helping decisions during distal and imminent threats

250 We ran a Searchlight analysis to localize dissociable neural patterns guiding decisions under distal and
251 imminent threats. Results showed that, prior to helping decisions, patterns of activation in the insula,
252 hippocampus and dorsal cingulate were distinguishable between distal and imminent threats (*FWE*
253 <.05, *k* > 10; Figure 4C).

254 Additionally, we performed a univariate decision type (help, not help) by threat imminence (distal,
255 imminent) ANOVA, and focused on brain regions displayed a significant interaction between the two,
256 which would indicate activation differences when making decisions under distal and imminent threat
257 conditions. Of note, due to the reduced number of not helping trials, 'no help' decisions in this analysis
258 included not only threat trials in which participants did not help, but also responses made in safe trials
259 (details in the methods section). Results showed a significant interaction in the midbrain PAG, bilateral
260 insula, right hippocampus, dorsal ACC, OFC and vmPFC, which was driven by the distal condition (Table,
261 2; Figure 4E). Indeed, during distal threats, higher activation in the hippocampus, vmPFC and OFC was
262 followed by decisions not to help, whereas higher activation in the dACC, PAG and insula led to helping
263 decisions. These results are in line with our prediction that the engagement of reactive fear circuits
264 facilitates helping behaviour.

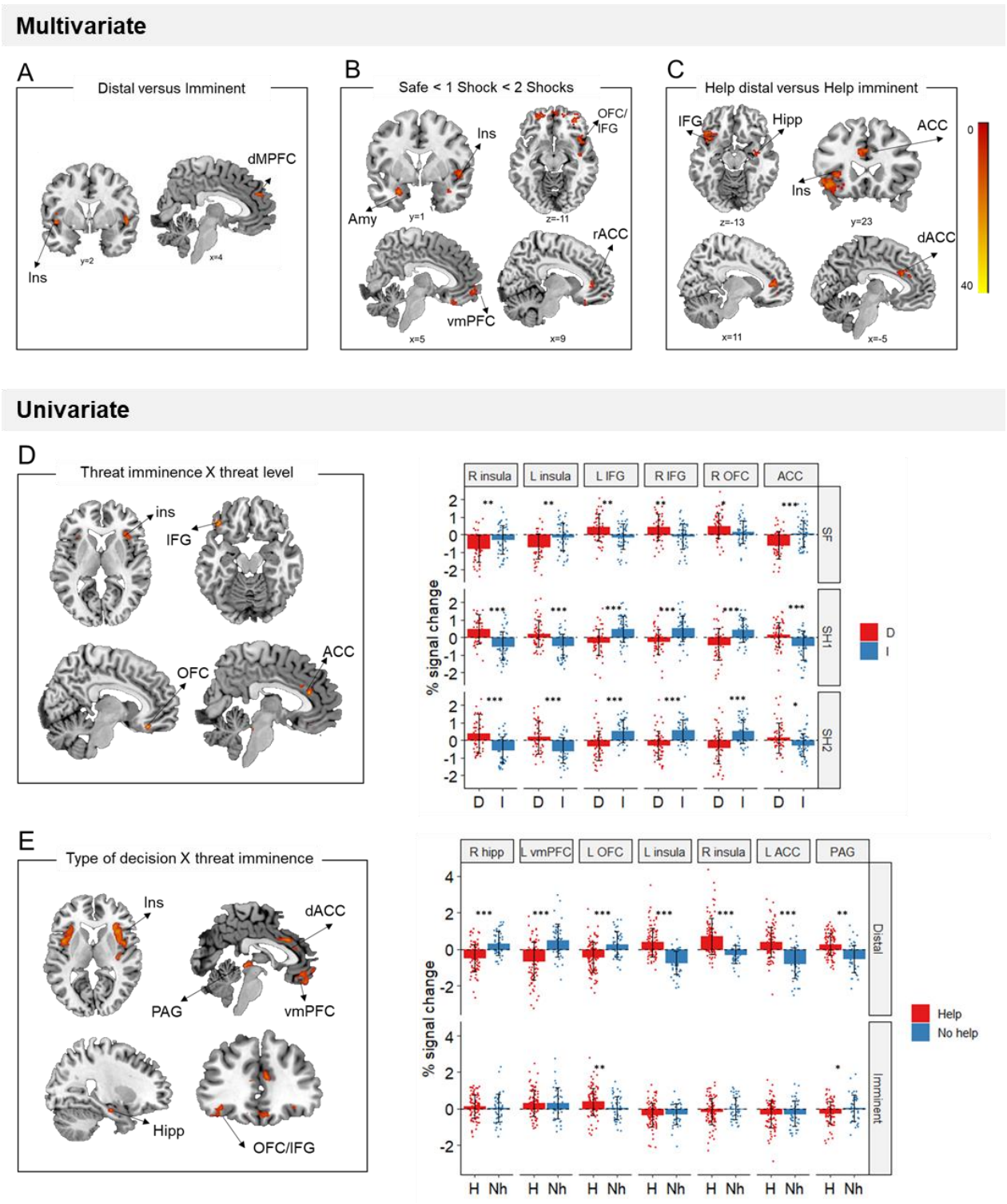

Figure 4. A. Local multivoxel activation patterns (identified by searchlight analysis) in the insula and dmPFC were distinguishable between distal and imminent threats, irrespective of helping decisions. B. Local multivoxel activation patterns (identified by support vector regression) in the amygdala, insula, OFC/IFG, vmPFC, and ACC were linearly associated with varying threat level. C. Local multivoxel activation patterns (identified by searchlight) in the insula, IFG, hippocampus and ACC were distinguishable when making helping decisions under distal and imminent threat. D. Clusters in the

insula, IFG, OFC and ACC displayed a significant threat imminence*threat level interaction. E. Clusters in the insula, ACC, IFG/OFC, vmPFC, hippocampus and PAG displayed a significant type of decision*threat imminence interaction. Ins: insula; ACC: anterior cingulate cortex; AMY: amygdala; OFC: orbitofrontal cortex; IFG: inferior frontal gyrus; vmPFC: ventromedial prefrontal cortex; PAG: periaqueductal gray; Hipp: hippocampus; dmPFC: dorsomedial prefrontal cortex. *p<.05, **p<.01, ***p<.001.

282 **Table 1.** Multivariate results based on threat imminence (distal – imminent) and level (safe, 1 shock, 2
283 shocks; *FWE* < .05)

Searchlight Distal versus Imminent					
	R/L	**k**	**x, y, z**	**T**	**BA**
Insula, superior temporal gyrus	R	40	46, -4, -8	7.43	22
Insula	L	17	-46, 2, -4	6.64	13
		18	-42, 4, 8	7.15	
Medial prefrontal cortex	R	20	10, 48, 28	7.13	9
SVM regression SF - 1 SH - 2SH					
Hippocampus	R	12	24, -10, -20	7.56	
Insula	R	25	42, -8, -8	8.34	13
Rolandic operculum	R	139	52, 0, 0	7.12	22, 47
Superior temporal gyrus, amygdala	R	11	38, 2, -24	7.73	
Amygdala	L	37	-24, 4, -24	7.96	
Rectus	R	12	6, 30, -24	7.07	11
Anterior cingulate	R	13	8, 38, 8	7.02	
Middle orbital frontal gyrus	R	42	32, 44, -14	7.32	11
Middle orbital frontal gyrus	L	11	-4, 48, -10	6.75	11
		44	-28, 54, -10	7.59	
		41	40, 56, -4	7.18	
Superior frontal orbital gyrus	R	19	16, 54, -14	6.78	10
Medial frontal gyrus	R	30	6, 58, -8	6.97	10
Searchlight Help during distal versus imminent threats					
	R	29	36, -8, -16	8.04	
Hippocampus	R	21	50, 4, -2	7.01	22
Insula	R	23	32, 14, 14	7.18	13
Inferior frontal gyrus	L	207	-48, 20, -6	8.81	47, 38, 22, 13, 45
Mid cingulate, dorsal anterior cingulate	R	43	2, 22, 30	7.25	32, 9, 24, 6
Insula	L	45	-32, 26, 0	7.83	47, 13, 45
Anterior cingulate	R	53	10, 40, 8	7.92	32, 10

284

285

286 **Table 2.** Results of the univariate ANOVAS (*FWE* <.05)

Univariate threat level*imminence interaction					
	R/L	**k**	**x, y, z**	**F**	**BA**
Insula	R	768	38, 20, 6	28.14	13
Insula	L	521	-30, 24, -4	20.58	13
IFG, OFC	L	593	-40, 32, -16	23.72	11
OFC	R	407	8, 32, -20	23.43	11
ACC	R	278	4, 32, 20	18.29	24
OFC, IFG	R	52	34, 40, -12	13.18	11, 47
Response*imminence interaction					
	R/L	**k**	**x, y, z**	**F**	**BA**
Midbrain		60	2, -32, -4	16.85	
Insula	R	192	38, -16, -2	20.62	13, 47, 22, 44, 6, 45, 21
Hippocampus	R	38	30, -14, -20	15.03	
Insula	L	595	-36, -8, -4	18.68	13, 22, 44, 6, 47, 45
Dorsal anterior cingulate	L	151	-2, 14, 30	16.13	24, 6, 32, 5, 4
Insula	R	748	30, 20, -8	18.18	47, 13, 22, 44, 6, 45, 21
Inferior frontal/orbital gyrus	L	46	-34, 36, -10	11.47	11
Rectus, ventral medial frontal gyrus	L	238	0, 46, -20	19.94	11, 10, 25

287

288 c. Helping and neural representations of threat and distress

289 One of our goals was to determine whether helping decisions were predominantly driven by the
290 response to another person's distress (in line with empathy-based explanations), by one's own
291 defensive state (in line with Vieira et al., 2020), or by both. To this end, we performed Representational
292 Similarity Analysis (RSA; Diedrichsen & Kriegeskorte, 2017). This analysis was done separately for
293 imminent and distal trials, and comprised 3 steps (Figure 5). First, we modelled each trial in the first-
294 level analysis, in order to estimate one beta coefficient per trial. Then, for each ROI, we extracted betas
295 from each voxel in each trial to estimate the correlation (expressed in r values) between all voxels per
296 trial. Resulting r values were used to construct a representational similarity matrix across all trials, that
297 reflects the correlation between all voxels in each ROI, in each trial. This matrix was then transformed

(1-r) to reflect dissimilarity instead of similarity (*i.e.,* representational dissimilarity matrix, RDM). In the second step, distress ratings provided after the scan on the unique video clips shown in each trial were used to construct an RDM that reflected the difference in perceived distress ("discomfort, anxiety, uneasiness") of the co-participant between trials (expressed in Euclidean distances). Post-scan threat level ratings were used in an identical manner to construct an RDM that reflects between-trial differences in how threatened the participant felt themselves during the scan. Finally, in the third step, we estimated the second-order similarity (kendall's τ) between neural and behavioural RDMs. This similarity metric allowed us to assess, for each ROI, whether trial-by-trial multivoxel patterns during the scan primarily represented the co-participant's distress, or the threat to oneself. Importantly, it allowed us to determine whether and to what degree neural representations of other's distress and of threat to oneself were associated with helping behaviour. To do so, second-order similarity values were entered in a linear model predicting average helping percentage during the scan. Predictors in this model were the similarity between neural and threat RDM, the similarity between neural and distress RDM, as well as threat imminence. Results show that, regardless of threat imminence, the similarity between neural and threat RDMs in the left amygdala (β=4.41, *se*=1.35, *t*=3.27, *p*=.006) and left insula (β=2.46, *se*=0.97, *t*=2.53, *p*=.047) was positively associated with helping behaviour (Figure 6; Table S3). In other words, the more strongly these brain regions, especially the amygdala, represented the threat to oneself, the more frequently the participant decided to help.

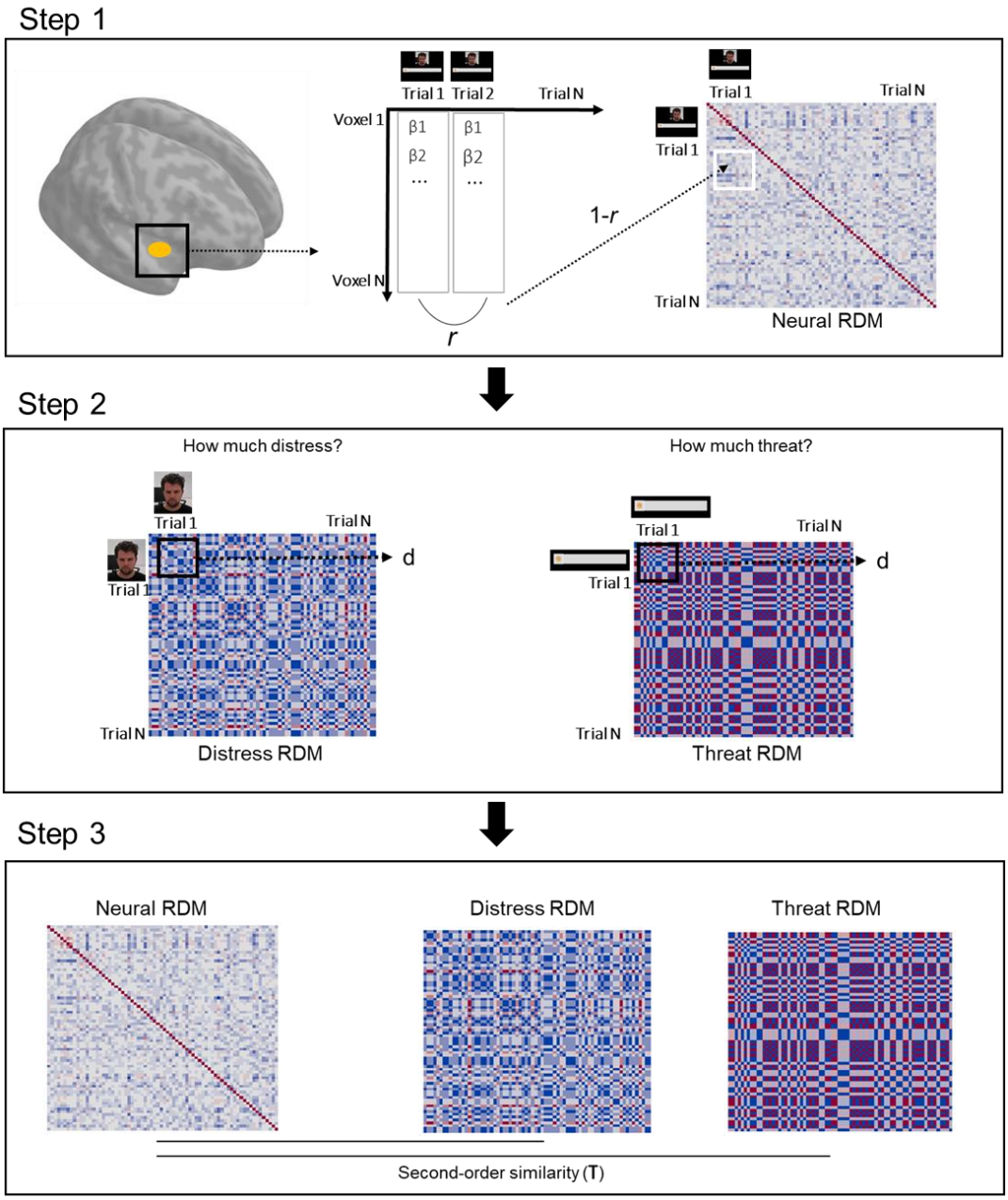

Figure 5. Schematic of the RSA pipeline. On step 1, we extracted the vector of trial-by-trial betas for each voxel in a given ROI. We then calculated the correlation (Pearson *r*) between all trial pairs. These correlation values were inverted (1-r) and used to create a trial by trial matrix, wherein each cell represents how correlated activation across all voxels of the ROI was in each trial pair (neural representational dissimilarity matrix, RDM). On step 2, post-scan ratings of the *co-participant's distress* in each unique clip were used to construct a trial-by-trial matrix, wherein each cell contained the Euclidean distance between the rating of each pair of clips (distress RDM). A similar method was used with the ratings of *threat to the participant* (threat RDM). On step 3, the second-order similarity between the neural RDM and distress RDM, and between the neural RDM and threat RDM were calculated using a ranked correlation method (Kendall's tau).

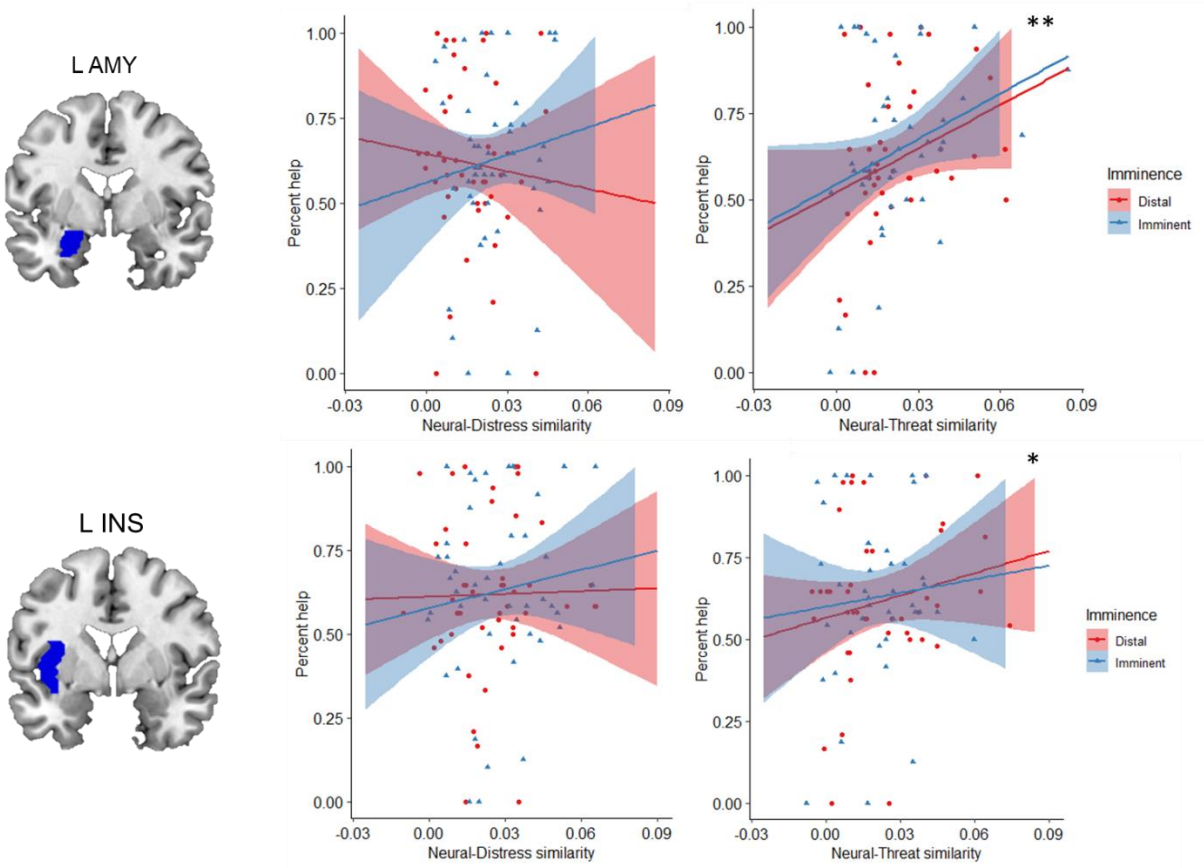

Figure 6. Regardless of threat imminence, the similarity between neural and threat RDMs in the left amygdala and insula predicted higher frequency of helping decisions. *p = .047; **p = .006

Discussion

Our overarching goal with this study was to determine how one's own defensive responses influence decisions to help others under threat. Our findings strongly suggested that neural circuits that coordinate fast avoidance responses from self-directed threats (reactive fear circuits; Mobbs et al., 2020) also promote helping of others in dangerous situations. More, the extent to which key defensive regions (amygdala and insula) represent the threat to the self (and not others' distress) predicts more frequent helping decisions.

Reactive fear circuits promote defensive helping

Previous work has shown that increased imminence of an other-directed threat facilitates helping behaviour, suggesting that the same way threat imminence triggers active avoidance (*i.e.*, fight or flight) from self-directed threats, it may promote defensive helping when others are under threat (Vieira et al., 2020). Here, we examined the neural basis of this effect. We found that multivoxel activation within the defensive circuitry (*i.e.*, amygdala, ACC, insula, hippocampus and OFC) tracked with the level of threat. Also, we found dissociable local patterns of activation prior to helping as a function of threat imminence within the insula, hippocampus and dorsal ACC. These findings are

consistent with previous work highlighting the role of the hippocampus and dACC as key regions optimizing escape decisions within cognitive and reactive fear circuits, respectively (Qi et al., 2018). Importantly, we found that average activation within several regions of the defensive circuitry, namely the in the PAG, insula, hippocampus, dACC, OFC and vmPFC, differed between trials leading up to helping or not helping decisions. Specifically, consistent with our predictions, activation in the insula, ACC and PAG was higher before decisions to help, whereas activation in the hippocampus, vmPFC, and OFC was higher prior to decisions not to help. Therefore, although overall we did not find higher frequency of helping decisions under imminent than distal threat, our results suggest that greater engagement of reactive fear circuits facilitated helping behaviour. These results are in line with prior findings suggesting that acute defensive states coordinated by reactive fear circuits promote helping under threat (Vieira et al., 2020). Of note, the differential activation of cognitive and reactive fear circuits based on the subsequent decision (help versus not help) was only found when decisions were made under distal threat (that is, when responses were prompted in the beginning of the trial). Also, contrary to our predictions, when comparing the average activation during distal and imminent threats independently of the decision, we found that the bilateral insula and ACC were more active during distal relative to imminent threats, whereas the bilateral IFG and OFC were more active during imminent threats. These results are opposite to those from prior studies that manipulated the imminence of self-directed threats (e.g., Meyer et al., 2019; Mobbs et al., 2007, 2009; Wendt et al., 2017). It is possible, however, that the disparity between our and previous findings is due to methodological reasons and does not reflect true differences in the processing of self- versus other-directed threats. In our paradigm, every trial started with a distal threat (the circle was static on the left) that always evolved into an imminent threat (the circle would move to the right, and the shock would be administered at the end of the trial). This may have made the threat overly predictable (especially given the high number of trials required for fMRI), engaging reactive fear circuits to a greater extent during the distal phase of the trial.

Nonetheless, our results obtained in a paradigm wherein the threat was directed at another person (the co-participant) were overall consistent with previous fMRI research investigating neural responses to self-directed threats as a function of imminence, suggesting a parallel between responses to self- and other-directed threats. This self-other parallel has been demonstrated in other threat-related processes, such as learning (Olsson et al., 2020). Importantly, in line with previous demonstrations that acute defensive states may promote prosocial outcomes (Vieira et al., 2020), we found that greater engagement of reactive fear circuits may facilitate helping of others in a threatening situation.

The neural representation of threat is associated with helping

To decide whether to help another person in a dangerous situation, one must consider not only their distress, but also the threat in the environment. Here, we determined how the brain, in particular the defensive circuitry, represents another's distress and threat when they are simultaneously present, and how those neural representations guide behaviour. To do so, we collected ratings of both the co-participant's distress, and of how threatened the participant felt during the scan. Crucially, these ratings were obtained after the scan, allowing us to obtain a behavioural metric of how participants independently represented another person's distress and the threat value of the situation. Performing these ratings after the scan also enabled us to avoid explicitly priming participants to consider those distress and threat cues during the scan, which could have influenced their behaviour and neural responses. Using behavioural representations of distress and threat, we assessed the extent to which each brain region in the defensive circuitry represented those cues, and its association with helping decisions. Our results showed that, regardless of threat imminence, the more the left amygdala and left insula represented the threat to oneself, the more participants decided to help.

The association between representation of threat and helping was particularly strong in the left amygdala. The amygdala has for long been known to have a pivotal role in the acquisition and expression of defensive responses in mammals (Tovote et al., 2015). For instance, in both humans and rodents, it has been shown to coordinate switches between defensive states across through the communication between its basolateral nucleus (BLA) and oxytocin(OT)-sensitive neurons in the central amygdala (CeA; Terburg et al., 2018). Our present results suggest that the amygdala's role in defensive responding may also be relevant for helping behaviour. This is consistent with previous work in animals. In rodents, CeA activation by OT not only enables the transition from freezing to fight-or-flight, but has also been shown to trigger offspring care behaviours in females (Rickenbacher et al., 2017), and to enhance maternal aggression (Bosch, 2013). It has additionally been demonstrated that the administration of benzodiazepines, drugs with a known effect on CeA (Griessner et al., 2021), impairs helping behaviour in rats (Ben-Ami Bartal et al., 2016). Taken together with these reports, our findings in humans suggest that amygdala-mediated defensive processes may also enable the provision of care to others, here in the form of defensive helping. Of note, our effects were restricted to the left amygdala, which is consistent with several others demonstrations of hemispheric specialization in amygdala function in emotional and pain processing (e.g., Allen et al., 2021; Baas et al., 2004; Blair et al., 2005; Butler et al., 2018; Young & Williams, 2010).

We found no evidence that the representation of other's distress in any ROI was associated with helping, including in brain regions that have previously linked with empathy for pain and distress states (ACC, insula; Bernhardt & Singer, 2012; Lockwood, 2016). Relatedly, our behavioural analysis revealed

411 that empathic concern, the motivation for care for others, was negatively associated with the
412 percentage of helping decisions. These findings are at odds with prior accounts of altruistic responding,
413 which suggest empathy for distress is the key proximal mechanism driving helping behaviour (e.g.,
414 Batson et al., 1987; de Waal & Preston, 2017). Several explanations may account for lack of evidence
415 of an association between neural representation of distress and helping in our data. One is that the
416 behavioural ratings were not sensitive enough to reflect variations in perceived distress between clips
417 of the confederate. However, this seems unlikely, given that we found significant effects of imminence
418 and threat level on ratings of distress. Another possibility is that distress is mainly represented in brain
419 regions not included in our analysis. The amygdala, insula and ACC would be prime candidate regions
420 to represent another individual's distress, in light of previous research on empathy (e.g., Fallon et al.,
421 2020; Lamm et al., 2011; Marsh et al., 2014; O'Connell et al., 2019), but we did not detect an
422 association between the degree to which these regions represented distress and helping behaviour.
423 Other potential regions would be those previously implicated in mentalizing, such as the temporo-
424 parietal junction (TPJ; Lamm et al., 2011; Patel et al., 2019; Schurz et al., 2014). At an exploratory level,
425 we repeated the RSA analysis within anatomical masks of the left and right TPJ, but did not find
426 evidence that the representation of distress in these regions were related to helping behaviour (see
427 Supplementary material). Further research, namely at the whole brain level, is necessary to further
428 assess the representation of distress in the brain, and its impact on helping behaviour under threat. In
429 any case, our present results suggest that, even if perceiving distress/need in others to some extent
430 triggers altruistic motivation (Preston, 2013), the ability to provide help may ultimately rely on the
431 activation of circuitry implicated in self-defence.

432 *Conclusions*

433 This study is the first to investigate neural responses to other-directed threats varying in imminence.
434 Our results point to a parallel between responses to self- and other-directed threats, and suggest that
435 the engagement of reactive fear circuits facilitates helping of others. Importantly, we showed that the
436 extent to which the amygdala represents the threat to self (and not other's distress) predicts helping
437 decisions. These results challenge the idea that empathy for distress is the only proximal mechanism
438 motivating helping decisions, and that overriding self-defensive responses is necessary to help others
439 under threat. Rather, in dangerous situations, one's own response to the threatening event may
440 enable defensive helping of others, possibly through the activation of neural mechanisms subserve
441 both individual defense and offspring care in mammals.

Materials and methods

This study has been pre-registered (https://osf.io/yvufn) and any deviations from the pre-registration are justified in the Supplementary material. Data and code will be made available at the OSF project page (osf.io/9cuva).

Participants

Forty-nine healthy volunteers (*M*=24.29, *SD*=4.78) participated in the experiment. Participants were recruited via flyers posted on- and off-campus, and local online recruitment systems. All participants were right-handed, had normal or corrected-to-normal vision, and were screened for history of psychiatric or neurological diagnoses, current medication, brain injuries and substance abuse. Participants provided informed consent prior to the experimental, and were compensated for their participation. The study was approved by the regional ethics board in Stockholm, Sweden.

fMRI helping under threat task

In each testing session, a participant and a confederate (henceforth, co-participant) were informed the experiment comprised two parts (only one of those parts involved an MRI scan), which would be randomly assigned to each one by flipping a coin. Participant and co-participant were then accompanied to separate testing rooms (the actual participant was taken to the MRI area) and did not interact again during the experiment (details about testing procedures, post-task questionnaires and debriefing are available in the Supplemental material).

In the MRI, participants performed a task modified from previous work (Vieira et al., 2020) wherein they made trial-by-trial decisions about whether or not to help the co-participant avoid aversive electrical shocks to the wrist, at the risk of also being shocked (Figure 1A). Threat imminence was manipulated by varying the spatial position and movement of a visual cue signaling varying levels of threat on a computer screen. Respectively, a green circle signaled no threat (no shocks), a yellow circle signaled moderate threat (1 upcoming shock) and the red circle signaled high threat (2 upcoming shocks). In addition, a webcam feed of the co-participant was presented on the screen throughout the task. Unbeknownst to the participant, the video feed was in fact pre-recorded, and edited to select unique clips for each trial of the task (details in Supplemental material).

Participants were informed that, throughout the experiment, they and the co-participant would see the same screen. Each trial started with a static cue on the left side of the screen (4 s), which then moved to the right (4 s). In shock trials, the co-participant would be administered an aversive shock to the wrist when the cue reached the right end of the screen, unless participants decided to help him. To decide whether they wanted to help the co-participant avoid the upcoming shock, participants made forced-choice responses by pressing 1 (Help) or 2 (Don’t help) on an MRI-compatible button box

as soon as the response slide was displayed (1.25 – 1.75 s). Responses were prompted sometimes in the beginning of the trial, when the visual cue was static on the left side (distal threat), and other times at the end of the trial, after the visual cue had moved to an endpoint on the right, and thus immediately before shock delivery (imminent threat). Outcomes of participants' decisions were as follows: if they chose not to help, the co-participant would always receive a shock; if they chose to help, there would be around 70% probability of both participant and co-participant receiving 1 shock (in moderate threat trials) or 2 shocks (high threat trials). Shocks were administered on the left ankle. Participants were instructed they should respond as quickly as possible. Also, to discourage missed responses, they were informed that a shock would be delivered to both participants (with 100 % chance) whenever a response was not detected. Finally, to balance the number of helping and non-helping trials during the scanning session, participants were informed that they would have a pre-set number of times they could help on each run, and thus they should try to balance, per run, the number of times they helped and not helped. In reality, participants could help on as many trials as they wished. Shock administration always happened at the end of the trial, and participants were able to see the outcome of their decisions on the screen (*i.e.*, the co-participant receiving or not receiving a shock; 4 s).

Safe trials followed an identical structure, with response slides presented at distal or imminent stages in relation to the end of the trial. However, participants were instructed that no shocks would be given and they should arbitrarily choose to press 1 or 2 when the response slide was displayed. It was made clear to them that their choice would have no consequences for them or the co-participant.

The task included 144 trials split into 8 functional runs (approx. 8 minutes). Each run comprised 18 trials, 9 distal and 9 imminent, and 6 of each threat level (resulting in 24 trials per condition, in total). Distal and imminent trials were presented in blocks, and the order of blocks was in each run. Within each block, safe, moderate threat and high threat trials were randomized. The order of functional runs was randomized across participants. The task was programmed and delivered using E-prime 3.0 (Psychology Software Tools, Inc., www.pstnet.com).

Ratings

After the scan, participants were taken to a different testing room and asked to complete a follow-up task. Here, all video clips showed during the scanning task were presented to participants, in random order. Participants were informed that these had been recorded during the scan, and that their task now was to, for each clip, rate the level of distress, anxiety or concern they perceived in the co-participant, on a 9-point scale. Participants were also presented images of the visual cues at distal (left side of the screen) and imminent positions (right side of the screen), and asked to rate on a 9-point

507 scale how threatened they felt during the scan, whenever they saw those images (Figure 1B). Ratings
508 of distress and threat were presented in separate blocks, and the order was randomized.

509 *fMRI acquisition and preprocessing*

510 Participants were scanned in a single session at the Stockholm University Brain Imaging Center (SUBIC),
511 using a 3T Siemens scanner with a 64-channel head coil. First, a high resolution t1-weighted anatomical
512 scan was obtained (TR = 2300 ms, TE = 2.98ms; FoV = 256mm, flip angle = 9°, 192 axial slices of 1 mm
513 isovoxels), followed by 8 functional runs, of about 8 min each. Functional images were acquired with
514 an echo-planar T2*-weighted imaging sequence with whole-brain coverage while participants
515 performed the fMRI task (TR = 1920ms, TE = 30ms, FoV = 192 mm, flip angle = 70°, 62 interleaved slices
516 of 2mm isovoxels, acceleration factor of 2).

517 Preprocessing of fMRI data was done using SPM12 (Wellcome Trust Centre for Neuroimaging,
518 www.fil.ion.ucl.ac.uk), and included slice timing correction, realignment to the volume acquired
519 immediately before the anatomical scan (i.e., the first image of the first functional sequence) using six
520 parameter rigid-body transformations, coregistration with the structural data, normalization to
521 standard space using the Montreal Neurological Institute (MNI) template with a voxel size of 2 x 2 × 2
522 mm, and smoothing using a Gaussian kernel with an isotropic full-width-half-maximum of 4 mm
523 (Gardumi et al., 2016; Hendriks et al., 2017). Finally, a high-pass filter with a cutoff of 128 s was applied
524 to remove slow signal drifts.

525 *Statistical analysis*

526 Behavioural data

527 Behavioural data was in general analyzed using Generalized Linear Mixed Models (GLMMs), an
528 approach that accounts for variation in the dependent variable that is explained by random sampling
529 of, for instance, participant or trial number (random effects), in addition to the independent variables
530 (fixed effects). Mixed-effects approaches have further been proposed to increase the generalizability
531 of research findings to other individuals and stimuli (Yarkoni, 2020).

532 Our main behavioural variable was helping behaviour, which was operationalized as the percentage of
533 helping responses throughout the task. We modelled helping percentage using GLMMs as a function
534 of threat imminence, threat level, and imminence by threat interaction (fixed effects). The subject was
535 added as a random effect, with random intercept and slope per threat imminence and level. In a
536 separate model, we also added a threat imminence*level*empathic concern interaction, following
537 previous indications that threat imminence may affect helping behaviour more strongly in individuals
538 with higher caregiving tendencies. To account for the possibility that behaviour varied throughout the
539 experiment, we also performed a mixed effects logistic regression on single trial dichotomous

responses (help or no help), including the trial number as a fixed effect in addition to the other fixed effects. Finally, following recent recommendations to consider within-individual effect sizes (Grice et al., 2020), we also calculated the difference between number of helping decisions under imminence and distal threat, per individual.

Reaction times were averaged per condition, and analysed in a GLMM with threat imminence, threat level, and imminence by level interaction as fixed effects, and the subject as a random effect (intercept and slope). Similarly, post-task ratings of other's distress and threat to self were analysed in a GLMM with threat imminence, threat level, and imminence by level interaction as fixed effects, and the subject as a random effect (intercept and slope).

Imaging data

a. First-level analysis

First-level analysis was performed in SPM12 and was based on the general linear model. Time-series of each voxel were normalized by dividing the signal intensity of a given voxel at each point by the mean signal intensity of that voxel for each run and multiplying it by 100. Resulting regression coefficients thus represent a percent signal change from the mean. Regressors were created by convolving the train of stimulus events with a canonical hemodynamic response function. Three different GLMs were estimated based on the goal of the analysis. For assessing differences based on threat imminence and level, 6 regressors of interest were modelled corresponding to the time window of the visual threat cue (distal safe, distal 1 shock, distal 2 shocks, imminent safe, imminent 1 shock and imminent 2 shocks). These regressors were defined based on the position of the threat cue on the screen (static = distal; approaching = imminent), and not relative to when the participant made a decision. In addition, 8 regressors of no interest were added in the model, corresponding to the time window of the response and the outcome, plus the 6 motion parameters estimated during realignment.

To assess differences in neural response based on the type of decision, another model was estimated with 6 regressors of interest: help distal 1 shock, help imminent 1 shock, help distal 2 shocks, help imminent 2 shocks, no help distal, no help imminent. Here, distal and imminent refer specifically to when the decision was prompted in the trial. Of note, due to the reduced number of no helping trials for some participants, the "no help" regressor included both not help decisions, and decisions made during safe trials, wherein no shocks were given. In addition, 8 regressors of no interest (response, outcome and 6 motion parameters) were added.

571 A third model was created to enable subsequent trial-by-trial Representational Similarity Analysis
572 (RSA), wherein one regressor was estimated for each individual trial, modelling the time window of
573 the threat cue.

574 b. Regions of interest (ROIs)

575 Given our focus on defensive brain circuitry, our analysis targeted pre-specified ROIs that were
576 anatomically defined, including the left and right amygdala, left and right hippocampus, left and right
577 insula, midbrain, left and right ACC, left and right vmPFC, left and right vlPFC (Figure 3). ROIs were
578 defined on the Wake Forest University (WFU) Pickatlas toolbox
579 (http://www.fmri.wfubmc.edu/cms/software) (Maldjian et al., 2003).

580 c. MVPA and SVR

581 Beta values derived from first-level analyses were used in multivariate analyses, including multivoxel
582 pattern analysis (MVPA) searchlight, support vector regression (SVR), and representational similarity
583 analysis (RSA). Spatially distributed patterns of activation across voxels can reveal distinguishable
584 neural responses between experimental conditions even in the absence of significant average
585 activation differences in single voxels, making multivariate approaches more sensitive than
586 conventional univariate analysis (Formisano et al., 2008).

587 To identify local activation patterns that distinguish between distal and imminent threats, we used an
588 MVPA Searchlight implemented in The Decoding Toolbox (TDT; Hebart et al., 2015). A spherical
589 searchlight (radius 12 mm) was moved throughout each participant's data and, at each searchlight
590 center, a support vector machine algorithm was trained to *discriminate* activation patterns in response
591 to distal and imminent threats. Training was done iteratively on each 7 functional runs and tested on
592 the 8th (leave-one-out cross-validation). Resulting percentage score at each voxel for participant was
593 calculated and display in individual accuracy maps. Accuracy maps were then analysed at the group
594 level in a one-sample T test implemented in SPM12. A similar approach was taken to the identification
595 of local activation patterns that discriminated helping decisions during distal and imminent threats.
596 Additionally, a SVR was used to identify local patterns that showed a *continuous* linear association with
597 threat level (safe, 1 shock, 2 shocks). As for the classification searchlight, individual accuracy maps
598 were analysed at the group level in a one sample t test. Group results were thresholded at voxelwise
599 FWE < .05 and only clusters with more than 10 voxels were further considered.

600 d. Univariate analysis

601 As a complement to the MVPA, we also performed GLM-based univariate analyses. We analysed BOLD
602 signal to the threat cues, regardless of decision, in a threat imminence (distal, imminent) by threat
603 level (safe, 1 shock, 2 shocks) repeated-measures ANOVA. We also analysed BOLD signal to threat cues

prior to the decision in a decision type (help, not help) by threat imminence (distal, imminent) ANOVA. Full results for these analyses are available in supplementary material. Univariate analysis results were first thresholded at $p<.001$ uncorrected. With this threshold, clusters with more than 10 voxels were significant with FWE-corrected $p<.05$.

e. RSA

One of the goals of the study was to characterize representations of other's distress and of threat to oneself by defensive regions, and determine its relation to helping behaviour. RSA was done separately for imminent and distal trials, and comprised 3 steps (Figure 5). On the first step, we modelled each trial in the first-level analysis (in SPM12), in order to estimate one beta coefficient per trial. Then, for each ROI, we extracted betas from each voxel in each trial to estimate the correlation of beta-values (expressed in r values) between all voxels per trial. Resulting r values were used to construct a representational similarity matrix across all trials, that reflects the correlation between all voxels in each trial. This matrix was then transformed (1-r) to reflect dissimilarity instead of similarity (RDM). On the second step, distress ratings provided after the scan on the unique video clips shown in each trial were used to construct a dissimilarity matrix that reflects the difference in perceived distress of the co-participant between trials (expressed in Euclidean distances). Post-scan threat level ratings were used in an identical manner to construct a dissimilarity matrix that reflects between-trial differences in how threatened the participant felt during the scan. Finally, on the third step, we estimated the second-order similarity (kendall's τ) between neural and behavioural RDMs. In a nutshell, this similarity metric allowed us to assess, for each ROI, whether trial-by-trial multivoxel patterns during the scan primarily represented the co-participant's distress, or the threat to oneself. Importantly, it allowed us to determine whether neural representations of other's distress and of threat to oneself were associated with helping behaviour. To do so, second-order similarity values were entered in a linear model predicting average helping percentage during the scan. Predictors in this model were the similarity between neural and threat RDM, the similarity between neural and distress RDM, as well as threat imminence. Fourteen linear models were estimated, one for each ROI (i.e., left and right amygdala, left and right hippocampus, left and right insula, midbrain, left and right ACC, left and right vmPFC, left and right vlPFC). False discovery rate (FDR) correction was applied to adjust the p value of all coefficient estimates, across all 14 models. FDR-corrected p values below $\alpha=.05$ were considered significant. Beta value extraction was performed in Matlab, and all remaining steps and analyses of the RSA were performed through custom-made scripts in R.

References

Allen, H. N., Bobnar, H. J., & Kolber, B. J. (2021). Left and right hemispheric lateralization of the amygdala in pain. *Progress in Neurobiology*, *196*, 101891. https://doi.org/10.1016/j.pneurobio.2020.101891

Baas, D., Aleman, A., & Kahn, R. S. (2004). Lateralization of amygdala activation: A systematic review of functional neuroimaging studies. *Brain Research Reviews*, *45*(2), 96–103. https://doi.org/10.1016/j.brainresrev.2004.02.004

Batson, C. D., Fultz, J., & Schoenrade, P. A. (1987). Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of Personality*, *55*(1), 19–39. https://doi.org/10.1111/j.1467-6494.1987.tb00426.x

Ben-Ami Bartal, I., Shan, H., Molasky, N. M. R., Murray, T. M., Williams, J. Z., Decety, J., & Mason, P. (2016). Anxiolytic Treatment Impairs Helping Behavior in Rats. *Frontiers in Psychology*, *7*, 850. https://doi.org/10.3389/fpsyg.2016.00850

Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, *35*, 1–23. https://doi.org/10.1146/annurev-neuro-062111-150536

Blair, H. T., Huynh, V. K., Vaz, V. T., Van, J., Patel, R. R., Hiteshi, A. K., Lee, J. E., & Tarpley, J. W. (2005). Unilateral Storage of Fear Memories by the Amygdala. *Journal of Neuroscience*, *25*(16), 4198–4205. https://doi.org/10.1523/JNEUROSCI.0674-05.2005

Bosch, O. J. (2011). Maternal nurturing is dependent on her innate anxiety: The behavioral roles of brain oxytocin and vasopressin. *Hormones and Behavior*, *59*(2), 202–212. https://doi.org/10.1016/j.yhbeh.2010.11.012

Bosch, O. J. (2013). Maternal aggression in rodents: Brain oxytocin and vasopressin mediate pup defence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1631), 20130085. https://doi.org/10.1098/rstb.2013.0085

Bosch, O. J., Meddle, S. L., Beiderbeck, D. I., Douglas, A. J., & Neumann, I. D. (2005). Brain Oxytocin Correlates with Maternal Aggression: Link to Anxiety. *Journal of Neuroscience*, *25*(29), 6807–6815. https://doi.org/10.1523/JNEUROSCI.1342-05.2005

Butler, R. K., Oliver, E. M., Fadel, J. R., & Wilson, M. A. (2018). Hemispheric differences in the number of parvalbumin-positive neurons in subdivisions of the rat basolateral amygdala complex. *Brain Research*, *1678*, 214–219. https://doi.org/10.1016/j.brainres.2017.10.028

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

de Waal, F. B. M., & Preston, S. D. (2017). Mammalian empathy: Behavioural manifestations and neural basis. *Nature Reviews Neuroscience*, *18*(8), 498–509. https://doi.org/10.1038/nrn.2017.72

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, *13*(4), e1005508. https://doi.org/10.1371/journal.pcbi.1005508

Fallon, N., Roberts, C., & Stancak, A. (2020). Shared and distinct functional networks for empathy and pain processing: A systematic review and meta-analysis of fMRI studies. *Social Cognitive and Affective Neuroscience*, *15*(7), 709–723. https://doi.org/10.1093/scan/nsaa090

Fanselow, M. S., & Lester, L. (1988). A functional behavioristic approach to aversively motivated behavior: Predatory imminence as a determinant of the topography of defensive behavior,. *Evolution and Learning, Evol.*, 185–212.

Formisano, E., De Martino, F., & Valente, G. (2008). Multivariate analysis of fMRI time series: Classification and regression of brain responses using machine learning. *Magnetic Resonance Imaging*, *26*(7), 921–934. https://doi.org/10.1016/j.mri.2008.01.052

Gardumi, A., Ivanov, D., Hausfeld, L., Valente, G., Formisano, E., & Uludağ, K. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *NeuroImage*, *132*, 32–42. https://doi.org/10.1016/j.neuroimage.2016.02.033

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as Effect Sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455. https://doi.org/10.1177/2515245920922982

Griessner, J., Pasieka, M., Böhm, V., Grössl, F., Kaczanowska, J., Pliota, P., Kargl, D., Werner, B., Kaouane, N., Strobelt, S., Kreitz, S., Hess, A., & Haubensak, W. (2021). Central amygdala circuit dynamics underlying the benzodiazepine anxiolytic effect. *Molecular Psychiatry*, *26*(2), 534–544. https://doi.org/10.1038/s41380-018-0310-3

Hashemi, M. M., Gladwin, T. E., de Valk, N. M., Zhang, W., Kaldewaij, R., van Ast, V., Koch, S. B. J., Klumpers, F., & Roelofs, K. (2019). Neural Dynamics of Shooting Decisions and the Switch from Freeze to Fight. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-40917-8

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. https://doi.org/10.3389/fninf.2014.00088

Hendriks, M. H. A., Daniels, N., Pegado, F., & Op de Beeck, H. P. (2017). The Effect of Spatial Smoothing on Representational Similarity in a Simple Motor Paradigm. *Frontiers in Neurology*, *8*, 222. https://doi.org/10.3389/fneur.2017.00222

Hernandez-Lallement, J., Attah, A. T., Soyman, E., Pinhal, C. M., Gazzola, V., & Keysers, C. (2020). Harm to Others Acts as a Negative Reinforcer in Rats. *Current Biology*, *30*(6), 949-961.e7. https://doi.org/10.1016/j.cub.2020.01.017

Hortensius, R., & de Gelder, B. (2018). From Empathy to Apathy: The Bystander Effect Revisited. *Current Directions in Psychological Science*, *27*(4), 249–256. https://doi.org/10.1177/0963721417749653

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lindegaard, M. R., Liebst, L. S., Philpot, R., Levine, M., & Bernasco, W. (2021). Does Danger Level Affect Bystander Intervention in Real-Life Conflicts? Evidence From CCTV Footage. *Social Psychological and Personality Science*, 19485506211042684. https://doi.org/10.1177/19485506211042683

Lockwood, P. L. (2016). The anatomy of empathy: Vicarious experience and disorders of social cognition. *Behavioural Brain Research*, *311*, 255–266. https://doi.org/10.1016/j.bbr.2016.05.048

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*(3), 1233–1239. https://doi.org/10.1016/s1053-8119(03)00169-1

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(42), 15036–15041. https://doi.org/10.1073/pnas.1408440111

Meyer, C., Padmala, S., & Pessoa, L. (2019). Dynamic Threat Processing. *Journal of Cognitive Neuroscience*, *31*(4), 522–542. https://doi.org/10.1162/jocn_a_01363

Mobbs, D., Headley, D. B., Ding, W., & Dayan, P. (2020). Space, Time, and Fear: Survival Computations along Defensive Circuits. *Trends in Cognitive Sciences*, *24*(3), 228–241. https://doi.org/10.1016/j.tics.2019.12.016

Mobbs, D., Marchant, J. L., Hassabis, D., Seymour, B., Tan, G., Gray, M., Petrovic, P., Dolan, R. J., & Frith, C. D. (2009). From threat to fear: The neural organization of defensive fear systems in humans. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(39), 12236–12243. https://doi.org/10.1523/JNEUROSCI.2378-09.2009

Mobbs, D., Petrovic, P., Marchant, J. L., Hassabis, D., Weiskopf, N., Seymour, B., Dolan, R. J., & Frith, C. D. (2007). When fear is near: Threat imminence elicits prefrontal-periaqueductal gray shifts in humans. *Science (New York, N.Y.)*, *317*(5841), 1079–1083. https://doi.org/10.1126/science.1144298

Mobbs, D., Yu, R., Rowe, J. B., Eich, H., FeldmanHall, O., & Dalgleish, T. (2010). Neural activity associated with monitoring the oscillating threat value of a tarantula. *Proceedings of the National Academy of Sciences*, *107*(47), 20582–20586. https://doi.org/10.1073/pnas.1009076107

O'Connell, K., Brethel-Haurwitz, K. M., Rhoads, S. A., Cardinale, E. M., Vekaria, K. M., Robertson, E. L., Walitt, B., VanMeter, J. W., & Marsh, A. A. (2019). Increased similarity of neural responses to experienced and empathic distress in costly altruism. *Scientific Reports*, *9*(1), 10774. https://doi.org/10.1038/s41598-019-47196-3

Oliner, S. P. (2004). *Do Unto Others: Extraordinary Acts Of Ordinary People*. Basic Books.

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews. Neuroscience*, *21*(4), 197–212. https://doi.org/10.1038/s41583-020-0276-4

Patel, G. H., Sestieri, C., & Corbetta, M. (2019). The evolution of the temporoparietal junction and posterior superior temporal sulcus. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *118*, 38–50. https://doi.org/10.1016/j.cortex.2019.01.026

Preston, S. D. (2013). The origins of altruism in offspring care. *Psychological Bulletin*, *139*(6), 1305–1341. https://doi.org/10.1037/a0031755

Qi, S., Hassabis, D., Sun, J., Guo, F., Daw, N., & Mobbs, D. (2018). How cognitive and reactive fear circuits optimize escape decisions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(12), 3186–3191. https://doi.org/10.1073/pnas.1712314115

Rand, D. G., & Epstein, Z. G. (2014). Risking Your Life without a Second Thought: Intuitive Decision-Making and Extreme Altruism. *PLOS ONE*, *9*(10), e109687. https://doi.org/10.1371/journal.pone.0109687

Rickenbacher, E., Perry, R. E., Sullivan, R. M., & Moita, M. A. (2017). Freezing suppression by oxytocin in central amygdala allows alternate defensive behaviours and mother-pup interactions. *ELife*, *6*, e24080. https://doi.org/10.7554/eLife.24080

Roelofs, K. (2017). Freeze for action: Neurobiological mechanisms in animal and human freezing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1718), 20160206. https://doi.org/10.1098/rstb.2016.0206

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

Sheleg, M., Yu, Q., Go, C., Wagner, G. C., Kusnecov, A. W., & Zhou, R. (2017). Decreased maternal behavior and anxiety in ephrin-A5-/- mice. *Genes, Brain, and Behavior*, *16*(2), 271–284. https://doi.org/10.1111/gbb.12319

Singer, N., Sommer, M., Döhnel, K., Zänkert, S., Wüst, S., & Kudielka, B. M. (2017). Acute psychosocial stress and everyday moral decision-making in young healthy men: The impact of cortisol. *Hormones and Behavior*, *93*, 72–81. https://doi.org/10.1016/j.yhbeh.2017.05.002

Taylor, S. E. (2006). Tend and Befriend: Biobehavioral Bases of Affiliation Under Stress. *Current Directions in Psychological Science*, *15*(6), 273–277. https://doi.org/10.1111/j.1467-8721.2006.00451.x

Terburg, D., Scheggia, D., Triana Del Rio, R., Klumpers, F., Ciobanu, A. C., Morgan, B., Montoya, E. R., Bos, P. A., Giobellina, G., van den Burg, E. H., de Gelder, B., Stein, D. J., Stoop, R., & van Honk, J. (2018). The Basolateral Amygdala Is Essential for Rapid Escape: A Human and Rodent Study. *Cell*, *175*(3), 723-735.e16. https://doi.org/10.1016/j.cell.2018.09.028

Tomova, L., Majdandžic, J., Hummer, A., Windischberger, C., Heinrichs, M., & Lamm, C. (2017). Increased neural responses to empathy for pain might explain how acute stress increases prosociality. *Social Cognitive and Affective Neuroscience*, *12*(3), 401–408. https://doi.org/10.1093/scan/nsw146

Tovote, P., Esposito, M. S., Botta, P., Chaudun, F., Fadok, J. P., Markovic, M., Wolff, S. B. E., Ramakrishnan, C., Fenno, L., Deisseroth, K., Herry, C., Arber, S., & Lüthi, A. (2016). Midbrain circuits for defensive behaviour. *Nature*, *534*(7606), 206–212. https://doi.org/10.1038/nature17996

Tovote, P., Fadok, J. P., & Lüthi, A. (2015). Neuronal circuits for fear and anxiety. *Nature Reviews. Neuroscience*, *16*(6), 317–331. https://doi.org/10.1038/nrn3945

Vieira, J. B., Schellhaas, S., Enström, E., & Olsson, A. (2020). Help or flight? Increased threat imminence promotes defensive helping in humans. *Proceedings. Biological Sciences*, *287*(1933), 20201473. https://doi.org/10.1098/rspb.2020.1473

von Dawans, B., Ditzen, B., Trueg, A., Fischbacher, U., & Heinrichs, M. (2019). Effects of acute stress on social behavior in women. *Psychoneuroendocrinology*, *99*, 137–144. https://doi.org/10.1016/j.psyneuen.2018.08.031

von Dawans, B., Fischbacher, U., Kirschbaum, C., Fehr, E., & Heinrichs, M. (2012). The social dimension of stress reactivity: Acute stress increases prosocial behavior in humans. *Psychological Science*, *23*(6), 651–660. https://doi.org/10.1177/0956797611431576

Waal, F. B. M. de, & Lanting, F. (1997). *Bonobo: The Forgotten Ape*. https://archive.nytimes.com/www.nytimes.com/books/first/d/dewaal-bonobo.html

Wendt, J., Löw, A., Weymar, M., Lotze, M., & Hamm, A. O. (2017). Active avoidance and attentive freezing in the face of approaching threat. *NeuroImage*, *158*, 196–204. https://doi.org/10.1016/j.neuroimage.2017.06.054

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685

Young, E. J., & Williams, C. L. (2010). Valence dependent asymmetric release of norepinephrine in the basolateral amygdala. *Behavioral Neuroscience*, *124*(5), 633–644. https://doi.org/10.1037/a0020885