# Neuro-computational mechanisms of action-outcome learning under moral conflict

Alessandra D. Nostro[1*], Kalliopi Ioumpa[1*], Riccardo Paracampo[1], Selene Gallo[1], Laura Fornari[1], Lorenzo De Angelis[1], Alessandro Gentile[5], Michael Spezio[3,#], Christian Keysers[1,2,#] & Valeria Gazzola[1,2,#]

[1] *Netherlands Institute for Neuroscience, KNAW, Amsterdam, The Netherlands*
[2]*Dept of Psychology, University of Amsterdam, Amsterdam, The Netherlands*
[3] *Psychology, Neuroscience, & Data Science, Scripps College, Claremont, CA USA*
[5]*GoDataDriven, Amsterdam, The Netherlands*

[*,#] Equal contribution

## Abstract

Learning to predict how our actions result in conflicting outcomes for self and others is essential for social functioning, but remains poorly understood. We test whether Reinforcement Learning Theory captures how participants learn to choose between two symbols that define a moral conflict between financial gain to self and pain for others. Computational modelling and fMRI imaging show that participants have dissociable representations for self-gain and pain to others. Signals in dorsal rostral cingulate and insulae track more closely with outcomes than prediction errors, while the opposite is true for the ventral rostral cingulate. Cognitive computational models estimated a valuational preference parameter that captured individual variability of choice in this moral conflict task. Participants' valuational preferences predicted how much they chose to spend to reduce another person's pain in an independent task. Learning separate representations for self and others allows participants to rapidly adapt to changes in contingencies during conflicts.

## Keywords:

reinforcement learning theory, prosocial behavior, empathy, pain, reward, mirror neurons, simulation theory

## Introduction

We often have to learn that certain actions lead to favorable outcomes for us, but harm others, while alternative actions are less favorable for us but avoid or mitigate harms to others[1]. Much is already known about the brain structures involved in making moral choices when the relevant action-outcome contingencies are well known[2–9], but how we *learn* these contingencies remains poorly understood, especially in situations pitting gains to self against losses for others.

Reinforcement learning theory (RLT) has successfully described how individuals learn how to benefit themselves[10,11] and most recently, how they learn to benefit others[12,13]. At the core of reinforcement learning is the notion that we do not learn directly by memorizing the outcomes of our actions, but rather that we update expected values (EV) of actions via prediction errors (PE) – the differences between actual outcomes and expected values represented in mind.

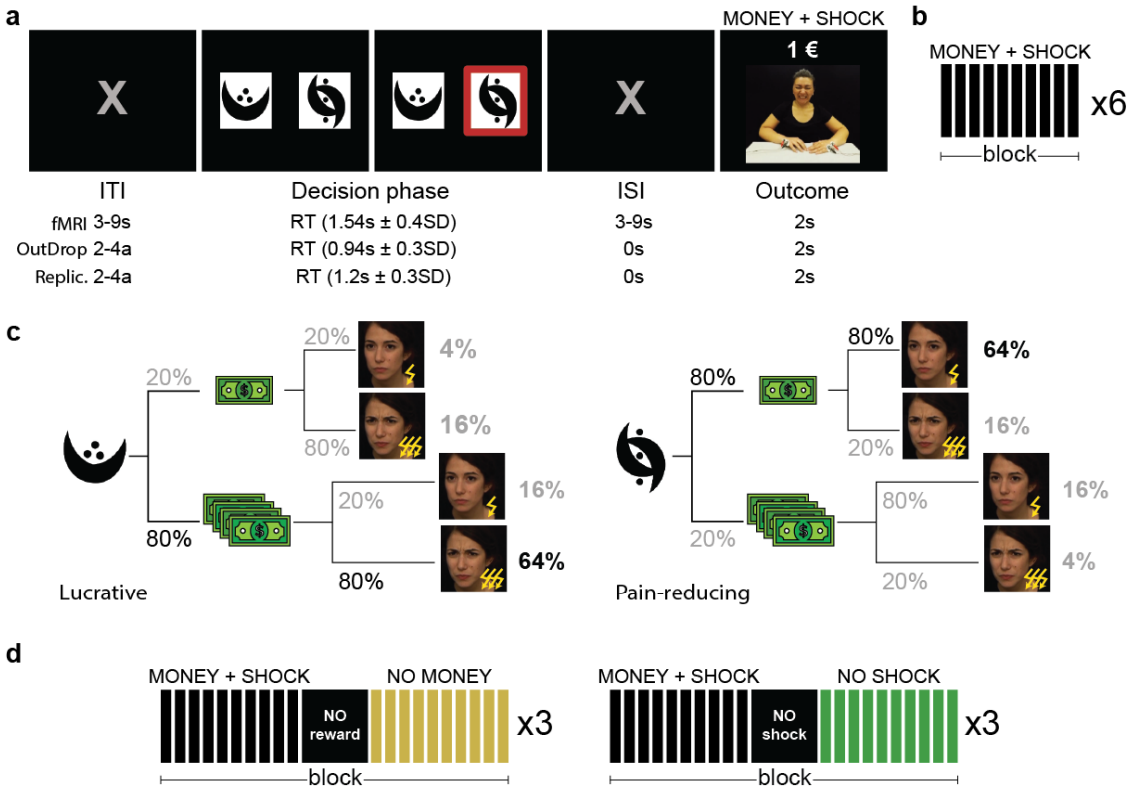
Ambiguity in morally relevant action-outcome associations raises specific questions with regard to RLT, especially if outcomes for self and others conflict. If actions benefit the self and harm others, are these conflicting outcomes typically combined into a common valuational representation, giving rise to a single prediction error that is then used to update the expected value of action alternatives? Or do we typically track separate expectations for benefits to the self and harms to others? In addition, people differ in how they represent benefits and harms to self[14], and in whether they prefer to maximize benefits for the self vs. minimizing harms to others[3,4,6]. How can such differences be computationally represented using RLT? Would people emphasizing one aspect, such as the harm to others, already show increased prediction errors and expected value signals for harms to others, or are expectations tracked independently of one's preferences, such that preferences only play out when decisions are being taken?

To address these questions we designed a learning task in which in which 63 participants (43 women; with 25 of them doing so while undergoing fMRI scanning) learn morally relevant action-outcome associations over 6 blocks of trials (Figure 1a,b). In each block, they see two new symbols. One symbol leads to high monetary gains for the self 80% of the time, and to a painful but tolerable shock to the hand of a confederate with the same probability. We call this symbol 'lucrative' to refer to the associated higher monetary outcomes. The other symbol leads to low monetary gains for the self 80% of the time, and to no painful shock to the confederate with the same probability, and we call this symbol 'pain-reducing' (Figure 1c). Importantly, to partially decorrelate representations of shock and money, the probabilities of high and low monetary reward and pain and no-pain to others are drawn independently. In the beginning of each block, participants do not know which symbol is associated with which outcomes. Choosing which symbol best satisfies the moral values that participants act upon in the task thus involves learning to predict the outcomes associated with each symbol. Shock events are shown to participants as facial expressions from the confederate via recorded video, but participants believe these recordings to be part of a real-time video-feed. We used this video-feedback, instead of the symbolic feedback more often used in neuroeconomic paradigms, to explore the neural systems activated in situations where the consequences of our actions are only available from the facial expressions of the people around us, and to remain closer in design to the paradigms used in nonhuman animal models[15]. Across 10 trials, participants learn the contingencies and express their preferences, prior to resetting the task with new symbols that require new learning.

We used Bayesian model comparison to compare computational models of how people combine the two outcomes in their morally relevant learning. In all of the models we compared, shock and money are additively combined using an individual weighting factor (wf, 1-wf) ranging from 0 to 1. This individually varying weighting factor captures the value of the monetary outcome for self relative to the value of shock to the other and is not unlike the salience alpha in the Rescorla-Wagner Learning Rule[16,17]. We then compare whether choices are better predicted by an RLT model that combines money and shock as soon as outcomes are revealed (M1) or by models that keep separate representations for the two quantities (M2). For M2, we further compared a variant that scales outcomes based on personal preferences for money vs. shock (M2Out) vs a variant that tracks expectations independently of personal preferences regarding outcomes, but that introduces weights at the decision phase of the task (M2Dec).

To more precisely determine how participants represent the outcome (money or shock) that was less influential in guiding their decisions an additional group of 20 participants (12 women) performed an Outcome Dropout task outside of the scanner (Figure 1d). In that novel task within the RLT framework, after an initial 10 trials of learning under conflict, we removed one of the outcome types (money or shock) and examined the resulting pattern of decisions over an additional 10 trials utilizing only a single outcome type.

Finally, we used the neuroimaging data to further inform our understanding of how our participants update values in our tasks. An influential theory posits that regions involved in first person pain experience undergo vicarious activation while witnessing the pain of others (particularly the dorsal rostral cingulate, the anterior insulae and somatosensory cortices), and that these signals contribute to actions that prevent or mitigate pain to others[2,4,18]. In addition to a body of neuroimaging evidence showing that regarding the pain of others recruits regions involved in nociception[19,20], studies with rodent models provide further evidence for a role of the dorsal rostral cingulate in affective responses to observed harm. The existence of pain mirror neurons in the dorsal rostral cingulate cortex[21] in rats shows that in rats the pain of conspecifics recruits a fraction of the same neurons that are active during first hand pain experience. Deactivating this region prevents rodents from showing affective responses to the harm of others[22,23] and prevents them from learning to avoid actions that harm others[15]. Indeed, the dorsal rostral cingulate together with the mediodorsal thalamus (the so-called “prefrontal thalamus” because of its connections to the prefrontal cortex in rodents) are also necessary in mice to learn that an environment is dangerous by witnessing shocks to conspecifics[22]. Unfortunately, the role of the insula and somatosensory cortices have so far not been systematically explored in rodents in the context of witnessing the harm of other animals. We therefore hypothesize that in humans, the dorsal rostral cingulate, and perhaps also the anterior insula, somatosensory cortex and the prefrontal thalamus could contribute to encode the harm of others to influence moral learning. In the context of RLT, in particular, we ask whether activity in these nodes would be better described by prediction errors between expected and observed harm or by how much harm is observed (i.e. the raw outcome). So far, responses in these nodes of the pain matrix have been shown to correlate with witnessed pain intensity[4,19,21,24,25]. However, in the designs used in these studies, observed pain intensities are randomized to avoid predictions, so intensity cannot be distinguished from prediction error. We finally explored whether prediction error signals for shocks scale with the weight that a participant places on these shocks (as predicted by M2Out) or not (as predicted by M2Dec).




**e**

| | Choice in Decision phase | Outcome phase | |
|---|---|---|---|
| **M0** | categorical_logit(tau*[0,0]) | | |
| **M1** | cat._logit(tau*[EV$_{old}$,EV$_{old}$]) | PE = (Out$_M$*wf +Out$_S$*(1-wf)) - EV$_{old}$ | EV$_{new}$ = EV$_{old}$ + LR*PE<br>EV$_{new}$ = EV$_{old}$ |
| **M2 Out** | cat._logit(tau*[EV$_{Mold}$ + EV$_{Sold}$,<br>EV$_{Mold}$ + EV$_{Sold}$]) | PE$_M$= (Out$_M$*wf ) - EV$_{Mold}$<br>PE$_S$ = (Out$_S$*(1-wf)) - EV$_{Sold}$ | EV$_{Mnew}$ = EV$_{Mold}$ + LR*PE$_M$<br>EV$_{Snew}$ = EV$_{Sold}$ + LR*PE$_S$<br>EV$_{Mnew}$ = EV$_{Mold}$<br>EV$_{Snew}$ = EV$_{Sold}$ |
| **M2 Dec** | cat._logit(tau*[EV$_{Mold}$*wf + EV$_{Sold}$*(1-wf),<br>EV$_{Mold}$*wf + EV$_{Sold}$*(1-wf)]) | PE$_M$= Out$_M$ - EV$_{Mold}$<br>PE$_S$= Out$_S$ - EV$_{Sold}$ | |

**Figure 1. Learning task.** (A) Trial structure, with the duration of each event indicated below for each experiment. The red contour indicates an example of a participant's choice. (B) Task structure of the fMRI experiment. Each of the 6 presented blocks included 10 trials (black vertical bars) shown in (A). The outcome phase always showed an outcome for the money and one for the shock. (C) Probabilities associated with each symbol. Note that the probabilities that a symbol is paired with the monetary reward to the participant and the shock for the confederate are computed independently such that there are four possible scenarios and probabilities (numbers in bold) for each symbol. Based on the final probability, choosing a symbol resulted in a more lucrative outcome for the self or a less painful outcome for the other. (D) Task structure of the Outcome Dropout experiment. The first 10 trials (black vertical bars) followed the same contingencies as in (A). After the 10th trial a screen indicated whether money or shock was removed and then 10 additional trials (yellow and green vertical bars) only presented the remaining outcome. Three blocks in which monetary reward (yellow) and three in which shock (green) was removed were randomly presented to participants during the task. (E) Schematic summary of the structure and parameters for the four models, separately for the Decision and Outcome phases of the task. The formulae in the outcome phase illustrate the update if the participant chose the symbol indicated in panel A.

## Results

### Model free description of choices

Participants showed individual variation in their choices (Figure 2a). Averaging participants who preferred the lucrative symbol with participants who preferred the pain-reducing symbol would obscure the learning pattern within each group. Accordingly, we first classified participants into three groups based on their choices in trials 7 to 10, after the action-outcome contingencies had been learned sufficiently for choices to reveal task preferences. We estimated the probability of choosing the pain-reducing option under conditions of no-preference by using a binomial distribution with 24 trials (4 trials x 6 blocks), a choice probability of 0.5 and $p<0.05$. Using this no-preference estimation, we grouped participants as selfish or as considerate depending on whether they showed fewer than 30% or more than 70% pain-reducing, considerate choices, respectively. Over the three experiments, we found that about half of the participants were considerate, one quarter were selfish, and one quarter were neutral (i.e. they do not fall in either category). Over all experiments, considerate and selfish participants learned in the first 6 trials, and choices became stable over the last 4 trials (Figure 2b). Neutral participants presented a pattern of choice that could reflect a failure to learn the action-outcome contingencies or a lack of preference for one expected outcome over another. As we show in Figure 2c, neutral participants in the Outcome Dropout task showed a strong preference for the lucrative option as soon as shock dropped out, and for the pain-reducing option as soon as money dropped out. This set of findings suggests that neutral participants had learned to predict outcomes but had no strong preference during conflictual choices. To understand whether these seemingly indifferent participants alternated their strategy across blocks (i.e. choosing the pain-reducing option in some, and the lucrative option in others), or whether their choices were at chance level also within each block, we examined their choices separately for each block. These plots show that most participants who did not show a particular preference across blocks also made indifferent choices within each block (Figure 1d-f).

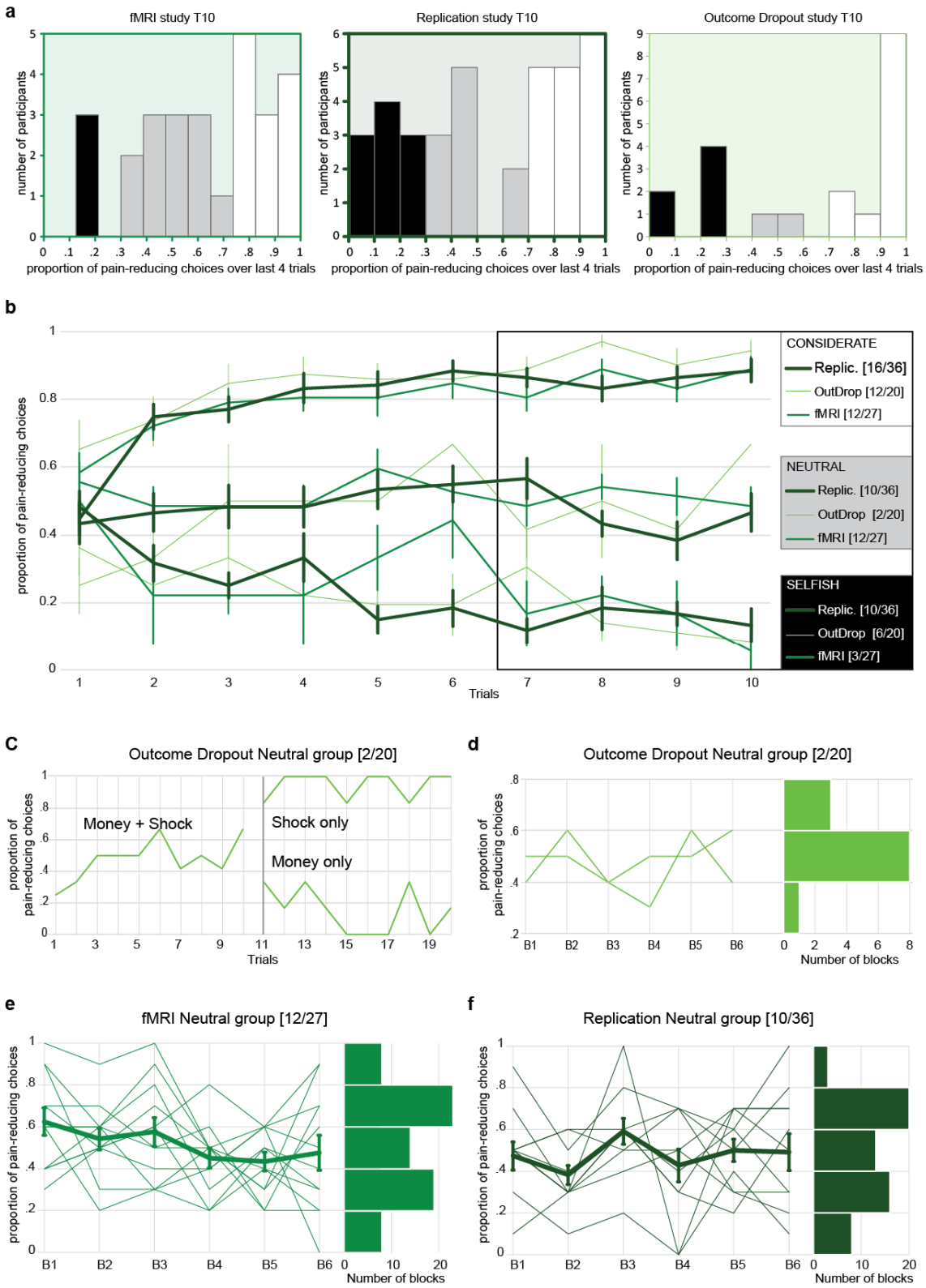



**Figure 2. Participant choices overview. (a)** Histograms of the proportion of pain-reducing choices over trials 7 to 10 for the fMRI, Replication and Outcome Dropout experiments. Participants are labeled as considerate (white) or selfish (black), if they chose the pain-reducing or lucrative symbol above chance, respectively. The remaining participants are labeled as 'neutral' (grey). **(b)** Mean (with s.e.m.) proportion of pain-reducing choices for each trial and group. The number of participants in each group

within each experiment are specified in brackets. The black box around trials 7 to 10 indicates that group classification depended only on the choices in these four trials, after learning had stabilized. Separate lines illustrated the proportion of choices for each experiment. For the Outcome Dropout experiment, in which 20 trials were collected per block, we only show the first 10 trials here, which included both shocks and money. T10 next to all experiments indicates that the first 10 trials are considered in all three experiments. **(c)** Mean proportion of pain-reducing choices for each trial for the two individuals of the Outcome Dropout experiment who did not show a clear pain-reducing or lucrative preference over the first 10 trials (Money + Shock). Choices from the 11$^{th}$ trial onward reveal a clear preference for the remaining outcome. **(d-f)** Thin lines: mean proportion of choices over trials computed for each experiment, block (B1-B6) and participants with a neutral preference separately. Thick lines: average and standard error of the mean of the participants with a neutral preference. Histograms on the right summarize the proportion of pain-reducing choices across participants.

## Computational Model Comparison

To examine the computational processes underlying learning, we estimated parameters for the four models shown in Figure 1e using the choices from the participants in the fMRI study. Model comparison used both the leave one out information criterion (LOOIC, lower values represent better predictive accuracy) and the area under the receiver operating characteristic curve (ROC curve), a metric of accuracy with 0.5, or 50%, indicating chance performance; values approaching 1 indicate better predictive performance). All three learning models (M1, M2Out and M2Dec) outperformed M0, in which no learning occurs (Figure 1e), by a substantial margin (Table 1, and Figure 3a). For the LOOIC, the standard error (se) provides a metric for the likely error in estimating the LOOIC, and the LOOIC difference between M0 and the other models is large relative to that standard error, providing evidence that all learning models outperformed M0. However, we found no robust difference across M1, M2Out and M2Dec: all three predicted participants' choices with similar accuracies, close to 80% (AUC ROC in Table 1), and all follow the average learning curve of the considerate and selfish participants reasonably well (Figure 3b). Comparing the models on the choice data from our Replication study confirms the robustness of this pattern (Table 1, and Figure 3a-b).

To generate data under conditions in which M1, M2Out and M2Dec make dissociable predictions, we reasoned that if one had a preference for pain reduction and found out that shocks were no longer being given, one's choices would differ under M1, M2Out and M2Dec. We predicted that if shock dropped out and if considerate participants maintain differentiable expected values for money and shock, they should switch to the lucrative symbol rapidly, as they would have separable expected values for both symbols. Knowing that shock outcomes will now be zero, one knows which option now has the higher value: the one with the higher expected value in the remaining (monetary) valuational modality. Under M1, one does not have access to monetary expectations stripped of the value of shocks, and thus needs to learn over multiple trials that the lucrative option is now better. We therefore predicted that M1 would take longer to adapt to the new condition, and M2 models would outperform M1. We further hypothesize that M2Dec would predict a faster switch to the lucrative option than M2Out because under M2Out, a considerate person with preferences for reducing pain would downscale expected value for money ($PE_M$=wf*$Out_M$-$EV_M$, Figure 1e). So expected value for money on the 11th trial would differ less across the two symbols under M2Out than under M2Dec, in which there is no downscaling of the expected value for money.

We thus created the Outcome Dropout task in which we informed participants at the 11th trial of each block of 20 trials, that one of the outcomes (i.e. money in 3

blocks and shock in 3 blocks, Figure 1d) drops out for the remaining trials (i.e. trials 11 to 20) in the block. We found that considerate participants continued to favour the pain-reducing symbol if money dropped out, but switched swiftly to the lucrative option when shocks dropped out (green lines, Figure 3d). Conversely, selfish participants continued to favour the lucrative option when shocks were removed, but swiftly switched to the pain-reducing option when money dropped out. Importantly, this shows that selfish participants still reacted in a considerate manner once their own monetary gains were no longer at issue. As hypothesized, this rapid switching was poorly predicted by M1 (magenta lines, Figure 3d), better by M2Out (navy) and best by M2Dec (cyan). Quantitative model comparisons confirmed that over the first 10 trials, M1, M2Out and M2Dec performed similarly well (and outperformed M0), but on the 11th trial, and over all 20 trials, M2Dec and M2Out outperformed both M1 and M0 on all metrics (Table 1, Figure 3b). The difference on the 11th trial between M2Out and M2Dec however does not exceed the margin of error.

| | | **M0** | **M1** | **M2 Out** | **M2 Dec** |
|---|---|---|---|---|---|
| **fMRI (10T)** | LOOIC | 2245.8 (25.0) | 1854.5 (64.9) | 1857.2 (64.7) | 1859.2 (64.3) |
| | WAIC | 2245.0 (22.3) | 1840.2 (65.4) | 1843.3 (64.8) | 1850.0 (35.2) |
| | AUC ROC | 0.50 | 0.77 | 0.78 | 0.77 |
| **Replication (10T)** | LOOIC | 2994.0 (60.0) | 2295.1 (102.0) | 2284.4 (108.0) | 2295.5 (106.5) |
| | WAIC | 2991.0 (10.3) | 2275.9 (106.8) | 2271.6 (106.4) | 2261.8 (45.5) |
| | AUC ROC | 0.50 | 0.82 | 0.82 | 0.82 |
| **Outcome Dropout (10T)** | LOOIC | 1854.5 (30.0) | 1253.4 (62.4) | 1264.5 (62.0) | 1274.7 (64.4) |
| | WAIC | 1663.5 (16.1) | 1236.3 (62.2) | 1242.2 (62.0) | 1244.9 (32.0) |
| | AUC ROC | 0.50 | 0.82 | 0.82 | 0.82 |
| **Outcome Dropout (20 T)** | LOOIC | 3327.1 (50.0) | 2555.3 (102.6) | 2062.3 (97.2) | 1957.0 (108) |
| | WAIC | 3327.1 (40.0) | 2532.5 (46.7) | 2031.2 (47.4) | 1935.3 (50.1) |
| | AUC ROC | 0.51 | 0.82 | 0.89 | 0.90 |
| **Outcome Dropout ($11^{th}$ T)** | LOOIC | 28.0 (2.0) | 35.8 (3.2) | 26.3 (3.1) | 30.2 (4.0) |
| | WAIC | 28.1 (1.8) | 38.6 (3.9) | 28.1 (3.5) | 33.4 (5.2) |
| | AUC ROC | 0.50 | 0.60 | 0.82 | 0.87 |

**Table 1. Model comparison.** For each experiment (most left column) the table reports the predictive performance, expressed in terms of LOOIC, WAIC and AUC ROC, of the computational models. Numbers in brackets indicate the standard error of the metric. For the Outcome Dropout experiment these values were calculated on the first 10 trails (10T), all 20 trials (20T), and also for the 11th trial alone (11th T).

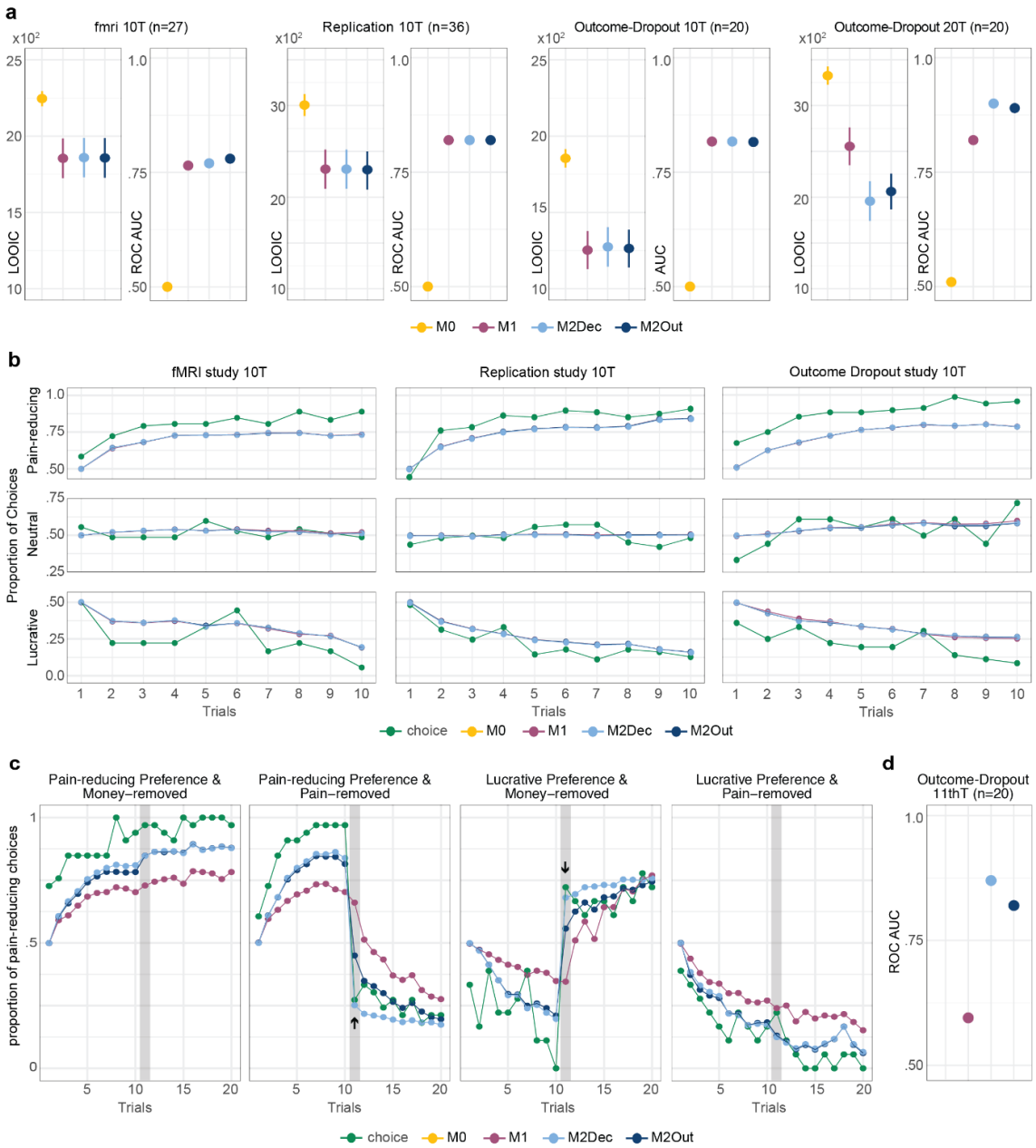



**Figure 3. Model Comparison.** (a) For each experiment, the graphs show the trial by trial proportions of pain-reducing and lucrative choices for the first 10 trials as predicted by the different models. Green lines indicate the actual participant choices and serve as observed data. (b) Predictive performance of the M0 (yellow), M1 (magenta), M2Out (navy), and M2Dec (cyan), expressed in terms of LOOIC (mean ± se) and AUC, for the first 10 trials of each experiment. For the Outcome Dropout study these parameters have also been estimated for all 20 trials. (c) Proportion of pain-reducing and lucrative preference for the Outcome Dropout experiment across the 20 trials. Grey shading highlights the 11$^{th}$ trial, at which one of the outcomes drops out. The dropped out outcome is indicated on top of each panel. (d) Predictive performance of the three models tested (M1, M2Dec and M2Out) for the 11$^{th}$ trial of the Outcome Dropout experiment.

Before exploring fMRI BOLD signal results relating to the computational models, we explored the posterior distributions of the estimated model parameters. Figure 4a shows the posterior distributions of the hyperparameters, and Figure 4b the spread of the individual point estimates. It should be noted that parameters estimated

using M2Out and M2Dec were virtually identical (Kendall T>0.9 for $LR_S$, $LR_M$, Tau and wf). Estimates for the learning rate have a relatively narrow distribution, in particular for $LR_S$. It is not surprising that $LR_S$ have narrower distributions than those for $LR_M$ given that most of our participants weighted shocks more heavily. In the fMRI experiment, the $LR_S$ was higher than the $LR_M$, but this was not consistent across experiments. Some differences related to task design were present across experiments, such as longer intertrial intervals in the fMRI experiment compared to the other studies. Learning rates across experiments were in the moderate range that would be expected for this kind of learning task. $LR_S$ is in a range that is close to optimal in our task (Supplementary Figure 1). Of particular interest in interpreting individual differences is the parameter wf, which showed a wide distribution. As would be expected, wf has a tight relationship with the preference shown in the last 4 trials (Figure 4c): considerate participants had a low wf (i.e., they placed less importance on money). To test whether the wf has external validity, we tested wf values against the average amount that participants donated to reduce the shock intensity to another individual in a different task, namely the Helping Task (see Methods). We found evidence for an association between wf and donation (Figure 4d, Kendall $\tau$=-0.47, $BF_{10}$=76, p<0.001). Bayesian multiple linear regression provided strong evidence that wf explained donation in the Helping Task ($BF_{incl}$=11.46) even in the context of the four subscales of the IRI [26] and MAS [27]questionnaires (Table 2). This analysis also provides moderate evidence that none of these latter questionnaires explained the variance on the Helping Task (all $BF_{incl}$<⅓).

| | Coefficient | Mean | SD | P(incl) | P(incl\|data) | BF $_{inclusion}$ |
|---|---|---|---|---|---|---|
| | Intercept | 2.7080 | 0.183 | 1.000 | 1.000 | 1.000 |
| | wf_10T_M2Dec | -1.6500 | 0.713 | 0.500 | 0.920 | 11.463 |
| IRI | Fantasizing | -0.0050 | 0.023 | 0.500 | 0.201 | 0.251 |
| IRI | Perspective taking | -0.0020 | 0.021 | 0.500 | 0.187 | 0.230 |
| IRI | Empathic Concern | 0.0060 | 0.023 | 0.500 | 0.206 | 0.259 |
| IRI | Personal Distress | 0.0010 | 0.018 | 0.500 | 0.185 | 0.227 |
| MAS | Power-Prestige | -0.0010 | 0.014 | 0.500 | 0.185 | 0.227 |
| MAS | Retention-time | -0.0010 | 0.008 | 0.500 | 0.189 | 0.233 |
| MAS | Distrust | 0.0002 | 0.010 | 0.500 | 0.186 | 0.228 |
| MAS | Anxiety | -0.0005 | 0.015 | 0.500 | 0.189 | 0.232 |

**Table 2. Bayesian linear regression posterior summary of coefficients.** The table summarizes the Bayesian linear regression model comparison for models explaining the donation in the Helping task using the wf of the Learning task, the subscales of the IRI[28] (FS, PT, EC and PD) and MAS [27] (Power-Prestige, Retention-time, Distrust and Anxiety). The first column indicates the variable under consideration, followed by the mean and sd of the regression parameter estimate, followed by the prior probability of including each variable (p(incl)) and the posterior probability of including each variable given the data (p(incl|data)). $BF_{incl}$ indicates how much more likely models including a variable are compared to the average of those not including this variable. $BF_{incl}$>3 is considered moderate, and $BF_{incl}$>10 strong, evidence that a variable explains donation. All other variables have $BF_{incl}$<0.33, showing evidence against a contribution of these variables: models without these variables are more likely given the data than those including the variable [29].

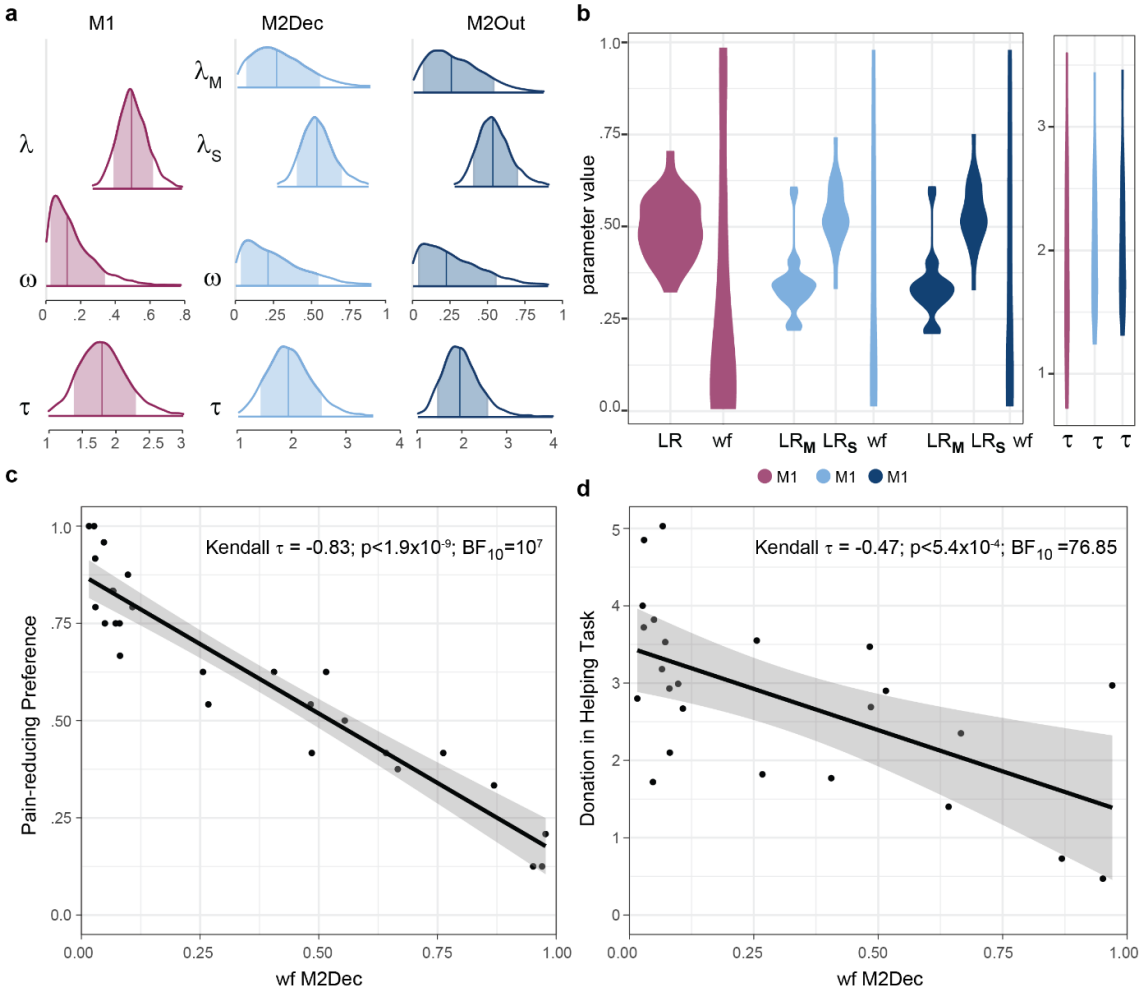



**Figure 4. Parameter estimates for the learning models of the choice data in the fMRI experiment. (a)** Posterior distribution of the hyperparameters of our hierarchical models of the choice data from the fMRI experiment. Greek letters stand for group hyperparameters. λ: learning rate, ω: weighting factor and τ: inverse temperature tau. Vertical lines indicate the median, and the colored range shows the 95% credible interval. **(b)** Violin plots of the individual point estimates of the same parameters, with Latin letters to indicate that these are the individual parameter estimates. **(c)** Scatterplots with regression lines illustrating the relationship between weighting factor and proportion of pain-reducing choices in trials 7 to 10. **(d)** Scatterplots with regression lines illustrating the relationship between weighting factor and the average donation in the Helping Task. Kendall's τ is used instead of r to quantify the correlation because wf does not follow a normal distribution. $BF_{10}$ refers to the Bayes Factor in favour of $H_1$ that Kendall's τ is different from 0. Shaded areas around the regression line indicate the 95% confidence interval on the regression line.

## Neuroimaging results

Participants showed robust brain activity across a wide network of brain regions when the outcomes of their decisions were revealed, including nodes often associated with the observation of facial expressions and monetary reward networks as identified by meta-analyses computed in Neurosynth (Neurosynth.org; Supplementary Figure 2a-b, Supplementary Table 1). As expected, given that in all trials participants viewed someone else receiving shocks (be they low shocks or high shocks), viewing outcomes also triggered activations that overlapped with voxels associated with first-

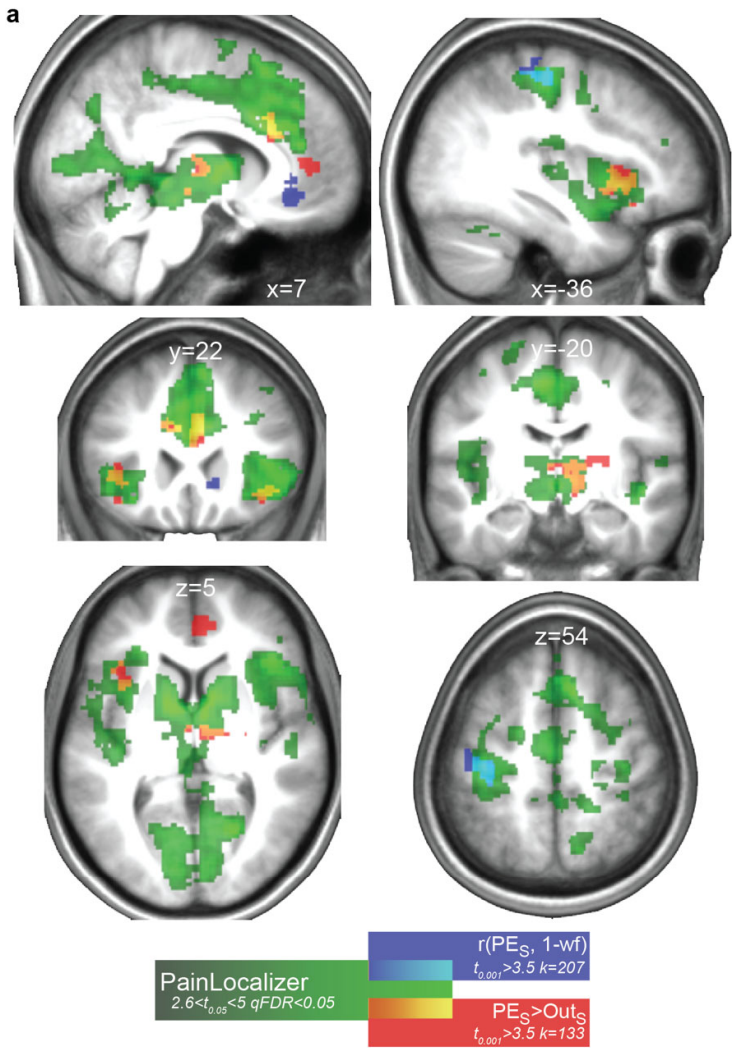
person pain unpleasantness in our Pain-Localizer (Figure 2c; Supplementary Table 3).

The focus of our interest, however, was on the variability of this activity across trials and participants. In particular, we asked: (i) do regions that activate during first person pain and vicarious pain (the rostral cingulate, insula, somatosensory cortices and thalamus) show trial-dependent BOLD signal differences that are better described by $PE_S$ or simple outcomes?; (ii) Is the magnitude of PE signals in the brain dependent on wf (as predicted by M2Out) or not (as predicted by M2Dec)?

To address the first question, we calculated the contrast $PE_S$-$Out_S$ using a matched pair t-test at the second level. To calculate this contrast, we used the trial by trial point estimates for $PE_S$ values from the M2Dec, which do not depend on wf, to simplify the interpretation of the results. We found that voxels with $PE_S$>$Out_S$ fell within the rostral cingulate, ventromedial prefrontal cortex, anterior insulae, and thalamus (Figure 5 red, Table 3; the reverse contrast did not yield results surviving our 5% FWE correction for cluster size).

Overlaying these clusters with voxels that correlate with perceived intensity during first hand pain experience in the Pain-Localizer experiment (Figure 5a green, Supplementary Table 3) from an independent sample of participants revealed that of the 1218 voxels of the $PE_S$-$Out_S$ contrast, 708 (i.e. 58%) fell within the 23154 voxels that were significant in the Pain-Localizer. Given that the search volume was 164375 voxels, the likelihood to fall within the Pain-Localizer by chance is 14%. Because voxels are not independent observations, to compare these proportions, we converted voxel counts to resel counts using the ratio of 1resel=138voxels as determined by SPM[30]. This conservative correction confirmed that more resels of the $PE_S$-$Out_S$ contrast fell within the Pain-Localizer than expected by the proportion of the search volume falling within that localizer ($\chi^2$=9.3 after Yates correction, $p<0.0023$). This overlap between first hand pain unpleasantness and $PE_S$>$Out_S$ representations during learning occurred in the dorsal rostral cingulate, anterior insulae and thalamus (Figure 5a yellow). Only the ventromedial prefrontal (vmPFC) cluster of the $PE_S$-$Out_S$ contrast did not overlap with the Pain-Localizer. Because we used a cluster extent correction for multiple comparison, our confidence that each voxel within a cluster is individually significant is limited. We therefore extracted the average signal in each of the $PE_S$-$Out_S$ clusters from the Pain-Localizer, and examined whether that activity parametrically followed the pain ratings of the participants (Table 3 first column). We found that it did in all but the vmPFC cluster. To shed further light onto the contrast, we extracted the parameter estimates for $PE_S$ and $Out_S$ from each of the 5 ROIs (Figure 5b). We found that in the vmPFC, the contrast was due to a positive parameter estimate for $PE_S$, and a near zero estimate for $Out_S$. This node thus appears to encode positive $PE_S$. In contrast, for the thalamus and left insula that overlap with the Pain-Localizer, the contrast was due to a negative parameter estimate for $Out_S$ and a near zero contrast for $PE_S$ suggesting these nodes encode negative outcomes to the confederate (high shocks were encoded as OutS=-1, low shocks as OutS=+1). Finally, for the dorsal ACC and right insula, the contrast was sue to a slightly positive parameter estimate for positive prediction errors and a slightly negative parameter estimate for OutS. Rather than simply revealing regions better explained by $PE_S$ than $Out_S$, this contrast thus revealed a gradient in the relative magnitude of $Out_S$ and $PE_S$ parameter estimates (Figure 5b; Note: formally testing $PE_S$ or $Out_S$ parameter estimates against zero would be circular, because the ROIs were selected using the contrast $PE_S$-$Out_S$ privileging voxels with positive $PE_S$ and/or negative $Out_S$ parameter

estimates. Interpreting the relative magnitude across these two parameter estimates, however, is less circular[31]).


a
x=7
x=-36
y=22
y=-20
z=5
z=54
PainLocalizer
2.6<t0.05<5 qFDR<0.05
r(PES, 1-wf)
t0.001>3.5 k=207
PES>OutS
t0.001>3.5 k=133

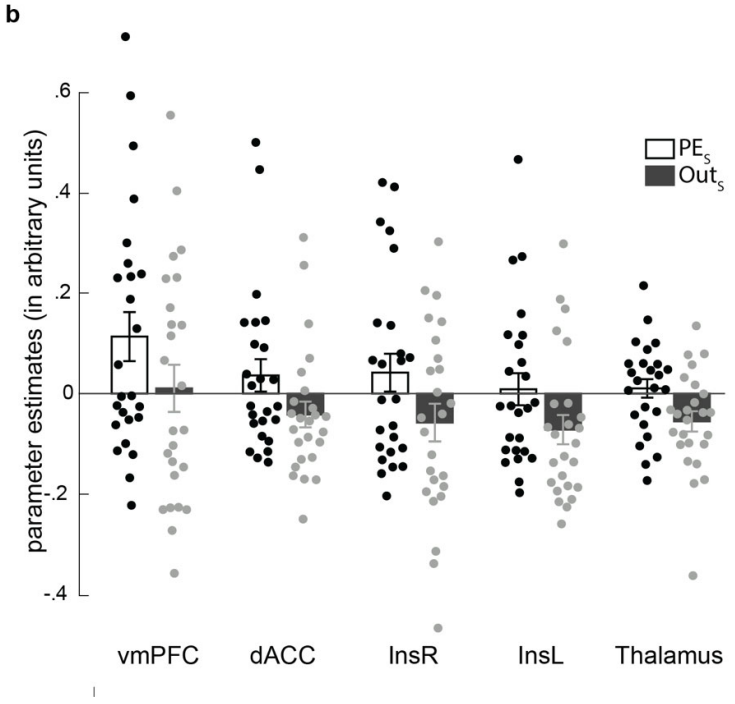

b
parameter estimates (in arbitrary units)
.6
.4
.2
0
-.2
-.4
PES
OutS
vmPFC
dACC
InsR
InsL
Thalamus

**Figure 5. Key fMRI results. (a)** Red: results from the paired-sample-t-tests between $PE_S$ and $Out_S$. t>3.47, p < .001, 5% FWE corrected at cluster level (pFWE=0.05 using k=133 minimum cluster size with 2x2x2mm voxels). The reverse contrast ($Out_S$-$PE_S$) did not survive 5%FWE correction. Peak coordinates can be found in Table 3. Green: areas significantly correlated with the perceived intensity of an electrical stimulation on the hand, derived from an independent sample of participants (n=23, t=2.7, qFDR<0.05 voxelwise). Blue: regression analysis of $PE_S$ using (1-wf) as predictor. t=3.47, p < .001, 5% FWE cluster level correction (pFWE=0.05 using k=204 voxels minimum cluster size with 2x2x2mm voxels). The reverse contrast did not survive 5% FEW correction. All activations shown on the average normalized T1 image of our participants. **(b)** $PE_S$ and $Out_S$ parameters estimated from the mean BOLD signal in the 5 clusters of Table 3 and in red in (a). Bars represent the mean, error bars, the standard error of the mean, and dots, individual participants. This figure is for informational clarity only and is not used for statistical inference.

| Cluster size (Kendall's tau, $BF_{10}$) | Voxels in cyto | % Cluster | Hem | Cyto or Anatomical description | % Area | peak t-value | MNI coordinates (X,Y,Z) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Paired t-test $PE_S$ > $Out_S$** | | | | | | | | | |
| 450 (Pain-Localizer: V=239,p<0.02; wf: $\tau$=0.125, $BF_{10}$=0.37) | 118.0 | 26.2 | R | Prefrontal Thalamus | 21.1 | 5.09 | 6 | -14 | 4 |
| | | | | | | 4.25 | 12 | -14 | -4 |
| | | | | | | 4.02 | 16 | -12 | 6 |
| | 34.4 | 7.6 | R | Premotor Thalamus | 25.8 | 3.92 | 20 | -14 | 10 |
| | 30.0 | 6.7 | R | Parietal Thalamus | 9.0 | 4.51 | 26 | -20 | 0 |
| | 14.8 | 3.3 | R | Temporal Thalamus | 2.7 | | | | |
| | 13.9 | 3.1 | R | Somatosensory Thalamus | 17.1 | | | | |
| | 9.6 | 2.1 | L | Temporal Thalamus | 1.8 | | | | |
| | 9.6 | 2.1 | R | Motor Thalamus | 21.8 | 3.95 | 20 | -18 | 12 |
| 268 (Pain-Localizer:: V=324, p<0.001; wf: $\tau$ =-0.09, $BF_{10}$=0.31) | | | L | Insula Lobe | | 4.61 | -40 | 20 | 2 |
| | | | | | | 4.33 | -36 | 14 | 0 |
| | | | | | | 4.16 | -38 | 12 | -4 |
| | | | | | | 3.86 | -36 | 20 | -4 |
| | | | L | IFG (p. Orbitalis) | | 4.51 | -38 | 22 | -14 |
| | | | | | | 4.42 | -38 | 18 | -14 |
| 241 (Pain-Localizer:: V=320, p<0.001; wf: $\tau$ =0.01, $BF_{10}$=0.26) | 12 | 5.0 | R | dACC Area 33 | 5.5 | 5.41 | 4 | 24 | 22 |
| | 8.9 | 3.7 | L | dACC Area 33 | 4.2 | 4.53 | -4 | 32 | 18 |
| | | | | | | 3.93 | -8 | 34 | 26 |
| 134 (Pain-Localizer:: V=323, p<0.001; wf: $\tau$ =-0.1, $BF_{10}$=0.321) | | | R | Insula Lobe | | 4.27 | 40 | 20 | -8 |
| | | | | | | 3.74 | 48 | 10 | -6 |
| | | | R | IFG (p. Orbitalis) | | 4.23 | 38 | 22 | -10 |
| | | | | | | 3.91 | 36 | 30 | -6 |
| | | | R | Putamen | | 3.56 | 30 | 10 | 0 |
| 133 (Pain-Localizer:: V=190.5, p=0.23, $BF_{10}$=0.022; wf: $\tau$ =-0.152, $BF_{10}$=0.441) | | | R | vmPFC | | 5.13 | 8 | 46 | 4 |
| | | | | | | 4.32 | 14 | 40 | 8 |
| | | | | | | 4.04 | 2 | 38 | 8 |
| **Regression between $PE_S$ and 1-wf** | | | | | | | | | |
| 232 | 54.5 | 23.4 | R | Mid Orbital Gyrus Area s32 | 37.5 | 4.48 | 6 | 36 | -12 |
| | 4.8 | 2.0 | R | sgACC | 2.2 | 4.48 | 8 | 34 | -4 |
| | 4.5 | 1.9 | R | Area s24 | 3.1 | | | | |
| 207 | 62.6 | 30.6 | L | Postcentral Gyrus Area 4p | 19.3 | 3.91 | -34 | -32 | 56 |
| | 52.3 | 25.8 | L | Precentral Gyrus Area 4a | 5.6 | 4.80 | -34 | -28 | 66 |
| | | | | | | 4.75 | -36 | -24 | 54 |
| | | | | | | 4.65 | -40 | -20 | 56 |
| | 34 | 16.7 | L | Postcentral Gyrus Area 3b | 6 | 4.62 | -44 | -18 | 54 |

**Table 3. Results of paired-sample-t-tests between $PE_S$ and $Out_S$**
Only clusters surviving a 5% FWE correction at the cluster size are reported (t=3.47, p < .001, cluster size 133 for the t-test and t=3.48, p < .001, cluster size 207 for the regression). Columns from left to right: cluster size indicated in number of voxels together with the results of a one-tailed Wilcoxon t-test on the parametric modulator for pain-unpleasantness in the Pain-Localizer and a Bayesian analysis of the correlation between the parameter estimate for $PE_S$ and (1-wf); number of voxel of the cluster falling in a cyto-architectonic area (cyto); percentage of the cluster falling in the cyto area; the hemisphere (Hem); the cyto-architectonic description when available or the anatomical description in the other

cases; the percentage of the cyto area activated by the cluster; the t-value of activation peaks identified in the cluster followed by their MNI coordinates. $BF_{10}<1$ is evidence for $H_0$:tau=0, and $BF_{10}<⅓$ is considered moderate evidence for $H_0$[29]; smaller values are stronger evidence for the absence of a correlation. Cytoarchitectural attribution is done using the Anatomy Toolbox for SPM[32,33]. L=left; R=right; IFG=Inferior Frontal Gyrus; ACC=Anterior Cingulate Cortex; p.=pars. The resampled voxel size for all analyses was 2x2x2mm.

To address the second question, we conducted a regression analysis at the second level with 1-wf as predictor for the parameter estimate of $PE_S$. This analysis identifies voxels, in which $PE_S$ signals are stronger in participants that weigh shocks more in their decision-making. We found that the subgenual cingulate (extending into the caudate), and a cluster encompassing the right-hand region of SI and MI indeed showed larger $PE_S$ signals in participants who valued shocks more strongly (Figure 5a blue, Table 3). The clusters more activated by $PE_S$ than $Out_S$ in Figure 5a (red) do not overlap with those where $PE_S$ is associated with 1-wf (blue), suggesting that participants have two separate representations of $PE_S$: one that does and one that does not depend on wf. To establish directly that clusters identified in $PE_S$-$Out_S$ are independent of wf (i.e. evidence of absence) we conducted a Bayesian correlation analysis between wf and the parameter estimate for $PE_S$ estimated using the average signal of all voxels in each of the 5 clusters resulting from the $PE_S$-$Out_S$ contrast[29] (Table 3 first column). In all clusters $BF_{10}$ is close to ⅓, providing evidence that this network indeed processes $PE_S$ independently of wf.

Beyond these planned analyses, we performed a number of additional exploratory analyses. At 5% FWE correction, we found the bilateral middle and inferior frontal gyri to correlate positively with $PE_S$ ($p<.001$, $t=3.47$, $k=136$; red in Supplementary Figure 3, Supplementary Table 4). These clusters therefore showed higher activity on trials in which shocks were less intense than expected. No clusters showed a negative correlation with $PE_S$. We also found a number of regions with signals associated with the update of expectations about shocks (green in Supplementary Figure 3 and 4 Supplementary Table 4). The update of expectations depends on the product of $PE_S$ x $LR_S$, and voxels associated with this product can be identified by performing a second level regression that expresses the magnitude of the $PE_S$ parameter estimate as a function of the covariate $LR_S$. Under M2Out, we would expect the update to be scaled by (1-wf), and such voxels would then be identified using a second level regression expressing the magnitude of the parameter estimate for $PE_S$ as a function of $LR_S$(1-wf). Supplementary Figure 4a shows that, r($PE_S$,$LR_S$) and r($PE_S$,$LR_S$(1-wf)) identified nodes including the insula, subgenual ACC, mid-cingulate and precentral gyrus. In addition, because money was not the dominant motivation for most participants, we did not expect to find prediction error signals for money ($PE_M$) in the brain to be as reliable as those for shocks. However, at a lower threshold ($p_{unc}<0.005$, $t>2.8$), we found $PE_M$ signals in somatomotor, posterior cingulate, temporal, striatal (caudate and putamen) and cerebellum regions in participants with higher wf (Supplementary Figure 4b, Supplementary Table 5). During the outcome phase, signal was higher in high shock compared to low shock outcome trials (i.e. were negatively correlated with $Out_S$ because high shock were coded as -1 and low shock as +1 in the parametric modulator) in clusters along the mid and superior right temporal gyrus (5% FWE correction on cluster size, $p< .001$, $t=3.47$, $k=133$; Supplementary Figure 2e, Supplementary Table 2). No signal was higher in low-shock compared to high-shock trials at $p<0.001$ (but see Supplementary Figure 2e for $p<.005$). We additionally observed signal positively correlated with $Out_M$ in the thalamus and inferior frontal gyrus (5% FWE correction at the cluster size; $p<.001$,

t=3.47, k=137; Supplementary Figure 2d, Supplementary Table 2), while no voxels surviving our 5% FWE correction were negatively correlated with $Out_M$ (even at $p_{unc}$>0.01).

## Discussion

Here we investigated how participants learn action-outcome associations during moral conflict (so called morally salient features[34]) . Participants were faced with two action alternatives with initially unknown consequences. The alternatives created a conflict between selfish and considerate preferences, as in 80% of cases one either chose higher monetary rewards at the expense of increased pain to another person, or less money while reducing the pain of another person. While the action alternatives and the monetary rewards to the self were represented by abstract symbols, the negative consequences of shock to the people that participants saw were represented by embodied facial expressions. Behavioral responses and brain activity were assessed to explore how people learn to value critically important features in applying moral preferences.

About half the participants showed significant preferences for the pain-reducing and a quarter for the lucrative choice. One quarter showed no clear preference. To understand the valuational representation influencing choice, and in particular to find whether participants track separate values for money to self and pain to others, or whether they track a single, combined value for self and other, we used computational modeling and Bayesian model comparison. We estimated: i) a random choice model, ii) an additive value combination model and iii) a separate value model. Among these models, both additive value and separate value models tracked decisions similarly well under conflict. To differentiate their performance, we introduced an Outcome Dropout task in which one of the outcomes, money or shock, suddenly dropped out, thereby removing the conflict. Participants switched their preferences almost instantaneously, which agreed best with the predictions of the separate value models.

Ultimately, even while computing separate values for the money and the shock, deciding which symbol to choose requires combining the two to yield an overall value of each option for the participant. Given the behavioral evidence that participants computationally represent separate expectations for shocks to others and money to self, this raises the question of how and when the two expectations are combined to result in a valuation guiding choice. In our models, we use an individual parameter, called the weighting factor (wf, conceptually similar to the salience factor alpha in the Rescorla-Wagner Learning Model[16]) to capture the relative weight placed on the money vs shocks, with the overall value = $wf*EV_M+(1-wf)*EV_S$. We found this weighting factor to be an effective way to model the individual variability in choices (Figure 4c). Importantly, we found wf to have external validity, in that it predicted how much money the same participants gave to reduce shocks to the same confederate in a different task, the Helping Task, which does not require learning (Figure 4d). We also found that wf was better than our trait measures of empathy and money attitude scales at predicting donation in the Helping Task (Table 2). Recent work has supported the view that state empathy is regulated by motives and context (see[35,36] for reviews). Our moral conflict task creates financial incentives known to downregulate empathy[35]. It is thus perhaps unsurprising that the IRI, which measures self-reported trait empathy and which does not probe empathy under such conflictual situations, fails to predict decisions under conflict, while wf, which is estimated for learning during moral conflict, is better at predicting decisions during moral conflict in

the Helping task. This speaks to the need to develop situated state empathy measures that probe the propensity of participants to deploy their empathy in specific situations, to complement measures of empathy as a context free trait[35,36]. Our paradigm might be particularly well suited in phenotyping antisocial populations and could provide insights into the atypical tuning of their morally relevant learning and decision-making in the spirit of computational psychiatry[37]. This is because psychopathic offenders show atypical instrumental learning[38] and reduced activity in empathy related networks while witnessing the pain of others[39] despite typical levels of trait empathy in self-report questionnaires[40].

The precise time at which participants applied the weighting on the two outcomes is less clear-cut from the data in our tasks. We compared two variants of our separate valuation model, one in which the weighting factor is applied directly to the outcome (M2Out), and one in which it is applied at the point of decision (M2Dec). In M2Out, the prediction errors, and hence the expected values are scaled by the relative importance of each outcome. In this model, for a participant motivated by avoiding shocks to others, the monetary outcomes are scaled down by the weighting factor as soon as they are revealed, and $PE_M$ and $EV_M$ have smaller magnitudes than $PE_S$ and $EV_S$. Decisions are then taken on the simple sum of the expected values for money and shock, as these EVs already reflect personal preferences. In M2Dec, the weighting factor is applied only at the very last, decision stage. In that variant, outcomes, prediction errors and expected values are unscaled, so that a participant motivated by avoiding shocks and one motivated by maximizing money have similar EV magnitudes, but decisions are taken on the weighted sum of these expectations. Computationally, the two models performed similarly well at predicting choices under conditions of stable conflict, and lead to virtually indistinguishable learning rates and weighting factor estimates. However, M2Dec fared slightly better in trials in which we removed one of the consequences and saw that participants swiftly displayed clear-cut preferences based on the outcome that previously carried very little weight (Figure 2c and 3c). This is because M2Out scales down the EV of the less important outcomes, and the difference in EV across the two symbols at the end of the 10th trial for the less important outcome is thus small. Accordingly, decisions on the 11th trials under M2Out were less clear-cut than under M2Dec, and the latter predicted the actual behavior better than the former. This evidence suggests that the participants maintained unscaled representations of the EVs of both options, perhaps in order to rapidly adapt to changing contingencies. Our neuroimaging results confirmed that some voxels, particularly in the ventral ACC have such unscaled $PE_S$ (Figure 5 red): a Bayesian correlation analysis revealed evidence against a correlation between average response in the vmPFC ROI of Figure 5 (red) and wf. However, our neuroimaging results provide evidence that other voxels had signals correlating with $PE_S$*(1-wf) (Figure 5 blue), including in particular the subgenual ACC. This means that prediction errors for shocks in these voxels are larger in participants who place more weight on shocks. By combining computational cognitive modeling and neuroimaging findings, we therefore observe that participants entertain a hybrid neural representation of the values of their actions under conflict: some nodes represent outcomes and their expected values independently of individual moral preferences while other nodes represent prediction errors multiplied with a weight that presumably reflects aspects of the individual's moral preferences during the task.

Most participants showed a preference for the considerate, pain-reducing option. This finding is consistent with work on moral decision-making in which the outcomes of the alternative actions are explicit and require no learning[3,41].

Interestingly, this pattern was true even in the Outcome Dropout and Replication experiments in which we asked participants to choose an amount of money as the high money outcome that they report would be similar in value to the cost of the other receiving the shock-intensity we used (Indifference Point). This bias towards pain-reducing options is consistent with other observations that humans avoid harm to others to an unusual degree compared to other animals[42], although other animals, including rats, also act to prevent harm to others under certain conditions (e.g.[15,43]).

A critical question in moral conflicts is how the perceived pain of another person enters decision-making. The human neuroscience of empathy and emotional contagion has suggested that brain regions involved in first person pain experience are reactivated while witnessing the pain of others, particularly the anterior insula, dorsal rostral cingulate at the edge between the mid- and anterior cingulate, and sometimes the somatosensory cortices[20,44]. Philosophers have proposed that feeling the pain of others by vicariously making it our own is the motive that makes us averse to harming others[18,45], and the reactivation of regions of the pain matrix provides a neural underpinning for this idea[2,4,6]. Indeed, altering activity in the somatosensory cortex alters how much participants are willing to pay to reduce the pain of another[4]. Also, in rats, the dorsal anterior cingulate contains individual neurons whose normalized spike rates when witnessing the pain of other rats are similar to normalized spike rates when experiencing pain[21]. Inhibiting processing in the cingulate prevents rats from avoiding actions that harm others[15]. It was previously unknown whether these brain regions only code the pain states of observed others or whether they also code for prediction errors. Here we show that during action-outcome learning in a moral conflict task, brain activity in some nodes associated with first hand pain unpleasantness in our Pain-Localizer, including the prefrontal thalamus and left anterior insula, preferentially represent negative outcomes. The right insula and dorsal anterior cingulate showed a combination of negative outcomes and positive PES. Prediction errors, considered to be central to learning associations between actions and vicarious shock in reinforcement learning theory, were coded most saliently outside of the pain localizer regions in the vmPFC. Whether the brain transforms sensory evidence about shock outcomes into prediction errors along this gradient of representation from the thalamus to the vmPFC via the insulae and dorsal anterior cingulate remains to be explored.

These findings bring together two of the most influential concepts of social and decision neuroscience - simulation theory and reinforcement learning theory - to provide a computational framework for learning in moral conflict. Our results suggest that people having no special training in moral deliberation or action keep track of rewards for the self and pain for others through separate expectations. Our results also suggest that the pain of others may at least partially enter valuational learning circuits via recruiting the pain matrix that contains signals tracking negative outcomes for others at least when the pain of the other is perceived by directly observing the other person's facial expression. Interestingly, the vmPFC, which has been shown to process outcomes specifically for other people[46,47], processed PE for shocks to others in our task, including more ventral clusters that scales and less ventral clusters that did not scale the $PE_S$ according to the weight individuals placed on the shocks in their decision-making.

In the pain literature, a medial pain system associated with the affective component of pain that includes the medial thalamus, insula and cingulate is often distinguished from a lateral pain system that includes the lateral thalamic nuclei and the somatosensory cortices[20,48]. In the empathy and emotional contagion literature,

the medial pain system has consistently been associated with witnessing the pain of others, while the lateral pain system is only thought to be recruited in situations in which the somatic origin of the pain is salient[20,44]. Mirror touch synesthetes who share the sensations of others particularly strongly and are unusually empathic[49] show stronger activity in the primary somatosensory cortex while witnessing others receive somatosensory stimulation[50]. Our data maps this idea to the realm of action-outcome contingency learning during moral conflict: the lateral pain system, SI in particular, showed $PE_S$ signals that varied across participants, being strongest in the most considerate individuals.

Our study has several limitations. First, we limited our model comparison to a number of hypotheses driven by RLT models. We did not test ratio or logarithmic ratio models of valuational representation in this study. These valuational structures are known to occur but are less often indicated in modeling gains and losses to the self[14]. Future experiments could be optimized to explore whether such alternative ways to combine these values may be more appropriate under certain moral conflicts. Second, our model comparison shows that among the additive models that we compared, M2Dec and M2Out perform best, and that these RLT models perform relatively well at predicting decisions, with AUC around 0.8, suggesting good accuracy in our predictions. However, comparing participants' choices and the prediction of our models shows that our models systematically underestimate the extreme choices of our participants (Figure 3b). This is because RLT models operating in an 80%/20% reward schedule do not predict choice proportions above 80%, while some of our participants chose their preferred action 100% of the time in the last 4 trials (Figure 2a). This discrepancy suggests additional cognitive mechanisms such as a shift to pure exploitation in at least some participants, in addition to the model free RLT models that we show to perform reasonably well. Third, we used the wf as a means of addressing quantitative individual differences. We recognize that wf only indicates preferences during our tasks and that it is risky to interpret the wf as suggesting stable moral values, as is the case for all behavioral studies absent strong evidence regarding stable moral commitments. Future studies may wish to explore whether different individuals may be best captured using qualitatively different models and in conjunction with validated evidence of long-term moral commitments and values. This might be particularly relevant when including participants with independently demonstrated morally considerate commitments on the one hand, or psychiatric disorders affecting social functioning on the other.

Our fMRI results should be interpreted with care. The moral conflict we create generates a complex situation that generates significant individual differences and requires participants to track multiple correlated signals simultaneously, including outcomes for self and other and expected values and prediction errors for both these outcomes and both action alternatives. Here, we focused on specific questions regarding the pain matrix and the difference between M2Dec and M2Out with our limited sample size, and only present the results of the other contrasts for illustrative purposes. A mechanistic understanding of the neuro-computational principles transforming outcomes into decisions via valuation processes will undoubtedly require a number of follow-up experiments. These experiments will require larger samples sufficiently powered to carefully compare selfish and considerate participants. They will also need to include conditions in which participants only need to track one outcome at a time (i.e. benefits for the self, shocks to the self or shocks to others) in addition to their conflictual combinations to trace how the elements of the conflict come together to lead to decisions. Future experiments could also be designed to isolate

signal related to the decision process, which in the current design could occur at any time between revealing the outcome and declaring one's choice. Finally, they will require a combination of animal studies and human studies leveraging the homologies across the ACC[15,22] to understand the causal relationships between activity in individual nodes, action-outcome learning and decisions.

## Materials and Methods

Three independent experiments were performed in the following order: a neuroimaging (fMRI) study, a replication of the behavior observed in the fMRI (Replication), and a behavioral experiment testing whether EV for shocks and money were learned separately (Outcome-Dropout). Table 4 gives an overview of the number of participants and experimental conditions included in each study.

| | **fMRI Experiment** | | **Outcome-Dropout Exp.** | **Replication Experiment** |
|---|---|---|---|---|
| | fMRI Data (T10) | Behav. Data (T10) | Behav. Data (T10+10) | Behav. Data (T10) |
| Helping task | 25 (35y±15SD; 25f) | 27 (37y±17SD; 27f) | | |
| Learning task | 25 (35y±15SD; 25f) | 27 (37y±17SD; 27f) | 20 (24y±4SD; 12f) | 36 (25±6SD; 16f) |
| Indifferent task | | | 20 (24y±4SD; 12f) | 36 (25±6SD; 16f) |
| Other tasks | | 27 (37y±17SD; 27f) | 20 (24y±4SD; 12f) | 36 (25±6SD; 16f) |
| IRI | | 27 (37y±17SD; 27f) | 20 (24y±4SD; 12f) | 36 (25±6SD; 16f) |
| MAS | | 27 (37y±17SD; 27f) | 20 (24y±4SD; 12f) | 36 (25±6SD; 16f) |
| SD3 | | | 20 (24y±4SD; 12f) | 36 (25±6SD; 16f) |

**Table 4. Experiment overview.** The table specifies the number of participants included in the final data analysis each task, with the average age ± standard deviation (SD) and number of females (f) indicated in brackets. In the fMRI experiment, we separated fMRI data (second) and behavioral data (third) columns. This was done, because behavioral data were analyzed from all 27 participants (including two left handed participants that were excluded from the fMRI analysis) and also include additional tasks and questionnaires. The number of trials included in the analyses is indicated in brackets (T10 or T10+10). The gray shade indicates the data used in the current manuscript. IRI: Interpersonal Reactivity Inventory[26], MAS: Money Attitude Scale[27], SD3[51].

### Participants

In total, 88 healthy volunteers with normal or corrected-to-normal vision, and no history of neurological, psychiatric, or other medical problems or any contraindication to fMRI were recruited for our experiments (Table 4). Two of the 27 participants in the fMRI were left handed, and only included in the behavioral part of the experiment, because the stimuli presented in the study showed movements of the right hand of an actor. Three participants in the Outcome-Dropout and two in the Replication studies were excluded from the analyses because did not believe the cover story. The data of 83

participants were therefore included in the final sample. The studies were approved by the Ethics Committee of the University of Amsterdam, The Netherlands (2017-EXT-8201, 2019-EXT-10607 and 2018-EXT-8864). Consent authorization for the publication of images has been obtained.

All participants performed the Learning task. FMRI participants additionally performed a Helping Task, and the remaining 58 participants performed additional tasks (Table 4).

**Learning Task**

Participants performed a probabilistic reinforcement learning task which faces them with a conflictual moral decision. Participants had to choose between symbols 1 and 2. Symbol 1 would most often result in a higher monetary outcome for the participant, and a noxious electrical stimulation to the dorsum of the hand of another person, the 'confederate'. Symbol 2 would most often result in a lower monetary outcome for the participant and a non-noxious electrical stimulation to the confederate (Figure 1a). The outcome of each choice was revealed by showing at the same time the amount of money received (+0.5€ or +1.5€) above a 2s video of the facial expression of the confederate in response to the stimulation. Participants knew that the computer would randomly select 10% of the trials, and pay out the actual amount of the monetary outcome of those trials as an extra bonus. Participants did not initially know what outcomes were associated with the two symbols and needed to learn the symbol-outcome associations over trials. The probabilities governing the money-related outcome (high vs. low amount of money for self) and the pain-related outcome (painful vs painless stimulation to the other) were computed independently (Figure 1c).

In order to i) maximize embodied empathy and a realistic conflictual situation, ii) limit the total number of shocks delivered to the confederate, and iii) avoid uncontrollable variance in the reactions of the victim, we used the cover story used and validated in Gallo et al (2018)[4]. Each participant was paired with what they believed to be another (the confederate), with whom they drew lots to decide who plays the role of the learner and who that of the pain-taker. The lots were rigged so that the participant was always learner and the confederate pain-taker. The participant was then taken to the scanning room (for the fMRI study) or a normal room with a computer (for the Outcome-Dropout and Replication studies) while the confederate was brought to an adjacent room, connected through a video camera. Participants were misled to think that electrical stimulations were delivered to the confederate in real-time, and that what the participants saw on the monitor was a live feed from the pain-taker's room. In reality, we presented pre-recorded videos of the confederate's facial reactions.

In the fMRI experiment only, participants practiced around 10 trials of the task before the actual experiment. The learning task in the fMRI experiment consisted of six blocks of 10 trials each (Figure 1b). At every block, a new pair of symbols was presented and participants had to learn the new symbol-outcome associations. At the end of the paradigm in the fMRI study, participants answered the question 'Do you think the experimental setup was realistic enough to believe it' on a scale from 1 (strongly disagree) to 7 (strongly agree). Five was used as a cut off to discriminate participants who believed in the cover story from those who did not, and participants who reported four or less were excluded from the analyses. At the end of the Outcome-Dropout and Replication studies, participants were instead asked to express the degree of agreement to the statement: "During the study you kept believing the confederate was another participant and received real electrical stimulation", again on

the same scale from 1 to 7. Participants were excluded if they strongly disagreed with the statement. This resulted in 3 participants excluded from the Outcome-Dropout study and 2 from the Replication study. In the Replication study participants performed six blocks of 15 trials, but only the first 10 trials were included in the analyses to keep the results comparable with the fMRI dataset. In the Outcome-Dropout study, participants also performed six blocks but each block consisted of 20 trials (Figure 1d). On the 11$^{th}$ trial, either the money or the electrical stimulation was removed. Participants were informed about which was removed after the 10$^{th}$ trial, via instructions displayed on the screen, but the other outcome associated with the symbols still followed the same contingencies as during the first 10 trials. No outcome was visually given for the removed quantity. The overall task therefore resulted in three randomized blocks without the monetary reward for the self, and three randomized blocks without the electrical stimulation to the other.

All tasks were programmed in Presentation (www.neurobs.com). FMRI tasks were presented under Windows 10 on a 32inch BOLD screen from Cambridge Research Systems visible to participants through a mirror (distance eye to mirror: ~10cm; from mirror to the screen: ~148cm). The behavioral tasks were presented under Window 7 from a 23-inch screen (distance eye to screen ~45 cm).

**Helping Task**

In the fMRI experiment, participants additionally performed the Helping Task presented in Gallo et al (2018) [4]. Only the behavioral results of the Helping Task will be included in the current publication. Briefly, participants performed 60 trials in which they watched a first (pre-recorded) video of the same confederate as in the Learning Task receive a painful stimulation. The intensity of the stimulation could vary between 1 and 6 on a 10 point pain scale, and was chosen on each trial by the computer program. In each trial participants also receive 6 credits, and could decide to donate some of these credits to reduce the intensity of the second stimulation to the confederate. Each credit donated back to the experimenter reduced the next stimulation by 1 point on the 10 point pain scale. Participants then watched a second video showing the confederate's response to the second stimulation. At the end of the task, participants were paid the sum of the amount of money that they had kept for themselves from all the trials divided by 10. We capture individual differences as the average number of credits given up per trials ("donation").

In the Replication and Outcome-Dropout studies, prior to the beginning of the experiment, participants underwent two relevant additional short tasks. **Pain-Threshold:** Before the assignment of the roles, participants underwent a pain-threshold procedure meant to determine the current intensity which would cause a painful but tolerable stimulation. **Indifference Point:** After role attribution but before the learning tasks, we determined the amount of monetary reward that would have a subjective value equivalent to the painful shock received by the confederate. This task enabled us to personalize the amount of money participants were later offered as high reward in the learning tasks to create a meaningful conflict. Participants always had to choose between a pain-reducing option combining 0.5€ for them with a low shock to the confederate, and a lucrative option combining a higher amount of money for themselves with a high shock to the confederate. The amount of money offered in the lucrative option varied in steps of 0.25€ across the 5 types of choices (Table 5). Each of the 5 types of choices were presented 4 times, for a total of 20 decisions. A sigmoid was then fitted to the choice data, and the indifference point was selected based on where the sigmoid crosses the 0.5 pain-reducing proportion. This value was then used

as the high reward in the learning task for the Replication and Dropout experiment. If a participant had always chosen the pain-reducing option, we picked the highest value (2€) as the high reward for the learning experiment. In the Outcome Dropout experiment 1/20 participants and in the Replication study 2/36, were given 2€ based on this rule. These 3 participants had very low wf in the experiment (0.01±0.005SD; N=3). Conversely, if a participant had always chosen the lucrative option, we picked the lowest value (1€) as the high reward for the learning experiment (2/20 Dropout, 1/36 Replication). These 3 participants had very high wf in the experiment (mean wf=0.99±0.005). Over the two experiments, on average the indifferent point was 1.46±0.37SD, which matched the 1.5 chosen in the fMRI experiment.

| | **Pain-reducing option** | **Lucrative option** |
|---|---|---|
| 1 | 0.5€ + Low Shock | 1€ + High Shock |
| 2 | 0.5€ + Low Shock | 1.25€ + High Shock |
| 3 | 0.5€ + Low Shock | 1.5€ + High Shock |
| 4 | 0.5€ + Low Shock | 1.75€ + High Shock |
| 5 | 0.5€ + Low Shock | 2€ + High Shock |

**Table 5. Participant's choices offered during the Indifference Point task.**

## Pain-Localizer Experiment

In an independent sample of 25 participants (age 25 ± 5.6 SD, 15 females, all right handed), we conducted an experiment to identify regions involved in the painfulness of first-hand experience of pain as triggered by an electrical stimulation of the dorsum of the hand. For each participant, ten noxious and ten innocuous electroshocks were applied, in a pseudo-randomized order (i.e. no more than two shocks of the same intensity consecutively). Stimulation consisted of a 100 Hz train of electrical pulses (2 ms each) applied for 0.5 s using an MRI-compatible electrical stimulator attached on the back of the right hand on the 4th *musculus interossei* (stimulation area: 16 mm2) through two bipolar surface electrodes. Before the scanning we measured the pain threshold from each participant. We started from a 0.2 mA current that was then increased until maximally 8.0 mA in 0.2 mA steps. Participants were instructed to evaluate how painful the stimulation was on a 10-point scale (1: not painful at all; 10: most intense imaginable pain). We then chose the current corresponding to a rating of 6 for the noxious condition and of 2 for the innocuous condition. Following each of the 20 stimulation in the scanner, after a random interval ranging from 2 to 5 s, the participant was asked to evaluate how painful the received electroshock was by button press. The participant was instructed to use four buttons of a MRI compatible button-box placed next to their left hand. Each button corresponded to a double step on a scale from 1 to 10. The pain intensity scale was the same 10 point scale used to determine shock intensity (1: not painful at all; 10: most intense imaginable pain), with the starting point set randomly for each trial to disentangle the number of button presses from the rating. A random interval ranging from 8 to 12 s separated the response from the next stimulation.

## Stimuli creation and validation

*Videos*

Videos for the Outcome-Dropout and Replication studies were generated following the procedure in (Gallo et al. 2018)[4]: a Caucasian actress, the same for all videos, showed an initial neutral facial expression followed by the facial expression in response to the electrical stimulation delivered to the right-hand dorsum. The upper part of her body was clearly visible on a black background. We initially recorded 150 videos showing a painful stimulation and 150 videos showing an innocuous stimulation. The actor was encouraged to produce a realistic and clear facial response in response to the temporally unpredictable stimulation. The final pool of stimuli was selected based on an online validation in which 200 participants (aged 18 – 35, 100 females) were recruited through Prolific (https://prolific.ac/). Participants were asked, by a survey created in Gorilla (https://gorilla.sc/), to rate the videos on a scale from 1 to 10, with '1' being 'just a simple touch sensation' and '10' being 'most intense imaginable pain'. High intensity pain stimuli in the final pool had an average rating of 5.30 ± 0.61SD, and low intensity stimuli of 1.42 ± 0.23SD.

A full description of how the videos used in the fMRI study were created can be found in (Gallo et al. 2018)[4]. A different female actor than in the Outcome-Dropout and Replication studies played the confederate in the fMRI experiment. The high intensity pain stimuli for the fMRI experiment had an average rating of 5.10 ± 0.97SD and low intensity stimuli of 2.18 ± 0.39SD ($t_{(52)}$=15.35, $p<0.001$, $BF_{10}=7.7*10^{17}$).

*Symbols*

All the symbols used in the learning task were created using Adobe Illustrator. For the creation of symbols we used simple geometrical shapes without obvious meaning. Each symbol was then paired with a second one created using the same shape elements organized in a different position. All the symbols covered the same area on the screen.

**Computational Modeling of Behavioral Data**

Our experiment represents a variation of a classical two armed-bandit task and was modeled using a reinforcement learning (RL) algorithm with a Rescorla-Wagner updating rule [16]. We compared 4 models explained in Figure 1e. Models were fitted in RStan (version 2.18.2, http://mc-stan.org/rstan/) using a hierarchical Bayesian approach, i.e. by estimating the actual posterior distribution through Bayes rule. Our models were adapted from the R package *hBayesDM* (for "hierarchical Bayesian modeling of Decision-Making tasks") described in detail in Ahn et al (2017) [52]. Model comparison was also performed in a fully Bayesian way, using the leave-one-out informative criterion (LOOIC [53]), which computes a pointwise log-likelihood of the posterior distribution to calculate the model evidence, rather than using only point estimates as with the other methods, e.g. of Akaike information criterion (AIC [54]) and the deviance information criterion (DIC [55]). The LOOIC is on an information criterion scale: lower values indicate better out-of-sample prediction accuracy. In Table 1, we additionally provide the WAIC values, which penalize models based on their complexity [53].

As priors on the hyperparameters LR and wf, we use the recommended Stan method of using a hidden variable distributed along N(0,1), that is then transformed using the cumulative normal distribution to map onto a space from 0 to 1. For Tau, we use the same method but then multiply the result by 5 to have the function map onto the interval [0,5].

For the Outcome-Dropout experiment, in which one of the two outcomes was removed after the 10th trials (Figure 1d), the three models M1, M2Out, M2Dec were modified from the 11th trial to account for the fact that participants were told which quantity would be removed. For the M2 models, this was implemented by setting EV=0 for the removed quantity before decision-making on the 11th trial. In addition, wf is modified only to value the remaining EV (i.e. wf=1 if shocks are removed and wf=0 if money is removed). For M1, we cannot reset expectations for a specific quantity (shock or money) and EV has to remain unchanged. However, the wf is adapted (i.e. wf=1 if shocks are removed and wf=0 if money is removed), which maximizes what can be learned from the following trials. In all cases, the outcomes for the missing quantity is always set to zero.
The stan models can be found in the Github repository https://github.com/anostro88/nin-moral-learning.

**Analysis of Behavioral Data**

Analyses of behavioral data were performed in RStudio (Version 1.1.453) and Matlab (https://nl.mathworks.com/). To illustrate the trend of participants' choices across experiments we split the participants in three groups (considerate, selfish and neutral). We observed that learning typically occurred in the first 6 trials, and therefore based our classification on the choices in trials 7 to 10. Using a binomial distribution we calculated the number of trials a participant needs to select the pain-reducing or lucrative symbol to reveal a preference against the null hypothesis that p(pain-reducing)=p(lucrative)=0.5 using binocdf(x,24,0.5), where x=number of pain-reducing choices across the 4 trials x 6 blocks = 24 choices, and 0.5 the chance decision. The binomial cumulative distribution shows participants need to choose 17 or more pain-reducing options (i.e. ≥70 % of the times) to be significantly classified as considerate, or 7 or fewer (i.e. ≤30%) to be significantly classified as selfish. Group percentages were then calculated as the number of considerate or selfish participants divided by the total number of participants.

Statistical analyses were performed using JASP (https://jasp-stats.org, version 0.11.1). Normality was tested using Shapiro-Wilk's. Since wf violated normality, statistical analyses were performed via non-parametric tests where available. To investigate the relationship between wf and the average donation we used Bayesian Kandel's tau non-parametric association test. The same test was used in Table 3 to verify that parameter estimates for PES in the five ROIs of Figure 5 (red) were not associated with wf.

To investigate whether donation is better explained by the wf, IRI, MAS or their combination, no non-parametric test was available, and considering that the Q-Q plot of the residual looked reasonable, we used Bayesian multiple linear regression test in JASP. In all cases we used the default Cauchy prior.

**MRI Data acquisition**

MRI images were acquired with a 3-Tesla Philips Ingenia CX system using a 32-channel head coil. One T1-weighted structural image (matrix = 240x222; 170 slices; voxel size = 1x1x1mm) was collected per participant together with an average of 775.83 EPI volumes ± 23.11 SD (matrix M x P: 80 x 78; 32 transversal slices acquired in ascending order; TR = 1.7 seconds; TE = 27.6ms; flip angle: 72.90°; voxel size = 3x3x3mm, including a .349mm slice gap).

**fMRI Data processing**

MRI data were processed in SPM12 [56]. EPI images were slice-time corrected to the middle slice and realigned to the mean EPI. High quality T1 images were coregistered to the mean EPI image and segmented. The normalization parameters computed during the segmentation were used to normalize the gray matter segment (1mmx1mmx1mm) and the EPIs (2mmx2mmx2mm) to the MNI templates. Finally, EPIs images were smoothed with a 6mm kernel.

**fMRI Data analysis**

For the Learning Task, our aim was to identify brain activity that scales with the outcome or the PE for money and shock when the outcome is revealed, and to compare this activity across participants with different weighting factors or learning rates. Analyses therefore focused on the outcome phase. In line with other studies (e.g.[57]), activity during the decision phase will not be analysed here for two reasons. First, to isolate activity when outcomes are revealed, we randomized the interval between outcomes and the response screen. As a result, decisions can have occurred at any time between the last outcome and the next button press, making it difficult to capture the activity linked to that decision. Second, the button press required for the response triggers significant brain activity in frontal regions that are hard to dissociate from the valuation processes we are interested in.

To analyse activity during the outcome phase, we ideally would have included predictors for the monetary outcome ($Out_M$) and the shock ($Out_S$) as well as their prediction errors ($PE_M$ and $PE_S$). Unfortunately, it lies in the nature of a moral dilemma, that $Out_M$ and $Out_S$ are negatively correlated, with high $Out_M$ associated with a high shock (i.e. low $Out_S$). This precludes the inclusion of both into a single GLM, as correlated predictors lead to unstable estimates. Similarly, PE and Out are also partially correlated, in that during learning, options with higher outcome will generate more positive PE, precluding the inclusion of outcome and PE in the same model. To obtain stable estimates of each quantity, we therefore only included one outcome or one PE per model, and thus generate four different GLMs (Table 6). In models looking at shocks (be it $Out_S$ or $PE_S$), we additionally modeled the EV in terms of shocks during the decision phase and the choice (i.e. pain-reducing vs. lucrative symbol). In models looking at money, we modeled the EV in terms of money and the choice. However, as mentioned above, we will not report results of this decision phase.

For each GLM, the decision regressor started with the appearance of the two symbols and ended with the button press of the participant. The outcome regressor was aligned with the presentation of the video and had a fixed duration of 2 seconds, corresponding to the duration of the stimulus. Each decision regressor had 3 parametric modulators, and each outcome regressor 1 parametric modulator (Table 6). The modulators were derived from the winning M2Dec, in which PE and EV are expressed in unweighted units. Our parameter estimates of interest were $PE_S$, $Out_S$, $PE_M$ and $Out_M$. For outcomes, the coding was +1 for good outcomes (i.e. high money for $Out_M$ and low shock for $Out_S$) and -1 for bad outcomes (i.e. low money or high shock). EV and PE follow that polarity. The choice regressor had values of ‘1’ or ’2’ corresponding to the lucrative and pan-reducing choice respectively.

| GLM 1 | GLM 2 | GLM 3 | GLM 4 |
|---|---|---|---|
| **Decision phase** | | | |

| | | | |
|---|---|---|---|
| *EVS1* | *EVS1* | *EVM1* | *EVM1* |
| *EVS2* | *EVS2* | *EVM2* | *EVM2* |
| *choice* | *choice* | *choice* | *choice* |
| **Outcome phase** | | | |
| *PES* | *OutS* | *PEM* | *OutM* |

**Table6.** GLMs structure. For each of the four estimated GLM the table indicates the included regressors of interest. In bold the main regressors and italics the parametric modulators, mostly derived from the winning M2Dec.

Because regions of interest are always subjective, we report results at whole brain level. Results were thresholded at $p < .001$ at the second level, with a Family-Wise error corrected at the cluster level over the entire brain.

To explore the relationship between subject specific variables (wf and LR) and the magnitude of PE, we performed regression analyses at the second level. We used the trial-by-trial PE estimates from the M2Dec model in all analysis, which are independent of wf or LR. If in a given voxel, a participant with a lower wf has signals that scale with $PE_S$ twice as strongly as another participant, the parameter estimate for $PE_S$ will be twice as high in the former than in the latter. We then capture the relationship with wf at the second level, by performing a regression analysis in SPM with wf as a predictor. The same approach was used for regressions between $PE_S$ and $LR_S$, $PE_S$ and $LR_S$(1-wf), and between $PE_M$ and wf.

For the Pain Localizer task, at the first, subject level, the data was modelled using the following predictors. One predictor captured the time of the electrical stimulation (duration=500ms) and additionally included the rating given by the participant as a linear parametric modulator. A second predictor contained the rating period from the appearance of the question mark until the participant's button press. Because the task started with a screen of instruction, one additional predictor collected this initial visual stimulus. The predictor was aligned with the presentation of the initial screen and lasted for 5s. Two more predictors were included to isolate errors if they occurred. If any button was pressed outside the rating period, a predictor of duration zero modelled these superfluous button presses. If a participant pressed two or more different buttons after the stimulation, we excluded that trial from the main shock predictor, and modeled it in a further error trial predictor. Six additional motion regressors of no interest were included to account for translations and rotations of the head as quantified during the realignment procedure. None of the participants had head motion parameters exceeding the acquired voxel-size. For each subject, we then brought our main parameter of interest – the parametric modulator for pain unpleasantness on the shock - to a second level t-test against zero either voxel-wise using SPM or within ROIs, using Marsbar and a Wilcoxon t-test against zero. Voxelwise results were thresholded at qFDR=0.05 (false discovery rate) at the voxel level to ensure confidence in localization.

## Data Availability

All data and analysis code will be available in OSF.io after publication.

## Acknowledgements

The work was funded by European Union's Horizon 2020 research and innovation programme under ERC-StG 'HelpUS' 758703 to VG, Dutch Research Council (NWO) VIDI grant (452-14-015) to V.G and VICI grant (453-15-009) to C.K. M.S. thanks the John Templeton Foundation (Grant 21338) for support in developing the ideas regarding self/other valuational representation.
We thank C. Gavanozi, I. Gembutaite and B. Hoekzema for helping with data acquisition in the Replication and Outcome-Dropout study. We thank A. Veggerby Lind for helping recording the stimuli used in the Replication and Outcome-Dropout studies. VG, CK, AN, LdeA, KI, SG, LF and RP thank P. Lockwood as her work inspired the development of the learning tasks.

## Author contribution

The experiments were conceived by VG and CK with input from all authors. In particular, SG and MS helped develop the Helping Task; AN, RP and MS helped develop the tasks of the Replication; AN and RP helped develop the Outcome Dropout experiments. The fMRI data were collected by KI and SG; RP and AN supervised and helped the collection of the Replication study; RP, LF helped collect the data of the Outcome Dropout study; SG collected the data of the Pain-Localizer task. All fMRI data were analyzed by KI with guidance from VG; the RLT models were developed and programmed by AN with the input of AG, LdeA, MS, CK and VG; RP and LF helped with further analyses of the behavioral data. AN, KI, CK, MS and VG wrote the manuscript with edits and comments from all other authors.

## Competing interests

The authors declare no competing interests.

## References

1. Cowell, J. M. *et al.* The development of generosity and moral cognition across five cultures. *Dev. Sci.* (2017). doi:10.1111/desc.12403
2. Hein, G., Silani, G., Preuschoff, K., Batson, C. D. & Singer, T. Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron* **68**, 149–160 (2010).
3. Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P. & Dolan, R. J. Moral transgressions corrupt neural representations of value. *Nat. Neurosci.* **20**, 879–885 (2017).
4. Gallo, S. *et al.* The causal role of the somatosensory cortex in prosocial behaviour. *Elife* **7**, 1–31 (2018).
5. Greene, J. D. The Cognitive Neuroscience of Moral Judgment. in *The cognitive neurosciences* (ed. Mangun, M. S. G. & G. R.) 1013–1023 (MIT Press, 2014).
6. FeldmanHall, O., Dalgleish, T., Evans, D. & Mobbs, D. Empathic concern drives costly altruism. *Neuroimage* **105**, 347–356 (2015).
7. Hutcherson, C. A., Bushong, B. & Rangel, A. A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron* **87**, 451–462 (2015).
8. Rilling, J. K. & Sanfey, A. G. The Neuroscience of Social Decision-Making. *Annu. Rev. Psychol.*

**62**, 23–48 (2011).
9. Hein, G., Morishima, Y., Leiberg, S., Sul, S. & Fehr, E. The brain's functional network architecture reveals human motives. *Science (80-. ).* (2016). doi:10.1126/science.aac7992
10. Sutton, R. S. & Barto, A. G. Reinforcement Learning: An Introduction. **9**, 9227 (1998).
11. Schultz, W. Updating dopamine reward signals. *Current Opinion in Neurobiology* **23**, 229–238 (2013).
12. Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E. & Roiser, J. P. Neurocomputational mechanisms of prosocial learning and links to Empathy. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9763–9768 (2016).
13. Apps, M. A. J., Rushworth, M. F. S. & Chang, S. W. C. The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron* **90**, 692–707 (2016).
14. Talmi, D. & Pine, A. How costs influence decision values for mixed outcomes. *Frontiers in Neuroscience* (2012). doi:10.3389/fnins.2012.00146
15. Hernandez-Lallement, J. *et al.* Harm to Others Acts as a Negative Reinforcer in Rats. *Curr. Biol.* **30**, 949-961.e7 (2020).
16. Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. in *Classical conditioning: current research and theory* (Appleton-Century-Crofts, 1972).
17. Miller, R. R., Barnet, R. C. & Grahame, N. J. Assessment of the Rescorla-Wagner model. *Psychol. Bull.* (1995). doi:10.1037/0033-2909.117.3.363
18. Smith, A. & Adam, S. Theory of Moral Sentiments. *Cambridge Texts Hist. Philos.* (1759).
19. Lamm, C., Decety, J. & Singer, T. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage* **54**, 2492–2502 (2011).
20. Keysers, C., Kaas, J. H. & Gazzola, V. Somatosensation in social perception. *Nat. Rev. Neurosci.* **11**, 417–428 (2010).
21. Carrillo, M. *et al.* Emotional Mirror Neurons in the Rat's Anterior Cingulate Cortex. *Curr. Biol.* (2019). doi:10.1016/j.cub.2019.03.024
22. Jeon, D. *et al.* Observational fear learning involves affective pain system and Ca v 1.2 Ca 2+ channels in ACC. *Nat. Neurosci.* (2010). doi:10.1038/nn.2504
23. Han, Y. *et al.* Bidirectional cingulate-dependent danger information transfer across rats. *PLoS Biol.* **17**, (2019).
24. Ashar, Y. K., Andrews-Hanna, J. R., Dimidjian, S. & Wager, T. D. Empathic Care and Distress: Predictive Brain Markers and Dissociable Brain Systems. *Neuron* **94**, 1263-1273.e4 (2017).
25. Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P. & Singer, T. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nat. Commun.* (2016). doi:10.1038/ncomms10904
26. Davis, M. H. & Association, A. P. A multidimensional approach to individual differences in empathy. *JSAS Cat. Sel. Doc. Psychol.* (1980).
27. Yamauchi, K. T. & Templer, D. I. The Development of a Money Attitude Scale. *J. Pers. Assess.* (1982). doi:10.1207/s15327752jpa4605_14
28. Davis, M. H. Measuring individual differences in empathy: Evidence for a multidimensional approach. *J. Pers. Soc. Psychol.* **44**, 113–126 (1983).
29. Keysers, C., Gazzola, V. & Wagenmakers, E.-J. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absenceitle. *Nat. Neurosci*
30. K.J. Worsley, A.C. Evans, S. M. and P. N. A three-dimensional statistical analysis for rCBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12:900-918**, (1992).
31. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. & Baker, C. I. Circular analysis in systems neuroscience: The dangers of double dipping. *Nat. Neurosci.* (2009). doi:10.1038/nn.2303
32. Eickhoff, S. B. *et al.* A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* **25**, 1325–1335 (2005).
33. Eickhoff, S. B. *et al.* Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *Neuroimage* **36**, 511–21 (2007).
34. Reeder, G. D., Monroe, A. E. & Pryor, J. B. Impressions of Milgram's Obedient Teachers: Situational Cues Inform Inferences About Motives and Traits. *J. Pers. Soc. Psychol.* (2008). doi:10.1037/0022-3514.95.1.1
35. Jamil, Z. Empathy: A motivated account. *Psychological Bulletin* (2014). doi:10.1037/a0037679
36. Keysers, C. & Gazzola, V. Dissociating the ability and propensity for empathy. *Trends in Cognitive Sciences* (2014). doi:10.1016/j.tics.2013.12.011
37. Montague, P. R., Dolan, R. J., Friston, K. J. & Dayan, P. Computational psychiatry. *Trends in*

*Cognitive Sciences* (2012). doi:10.1016/j.tics.2011.11.018
38. Blair, R. J. R. The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends Cogn. Sci.* (2007). doi:10.1016/j.tics.2007.07.003
39. Meffert, H., Gazzola, V., Den Boer, J. A., Bartels, A. A. J. & Keysers, C. Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain* **136**, 2550–2562 (2013).
40. Meffert, H., Gazzola, V., Den Boer, J. A., Bartels, A. A. J. & Keysers, C. Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain* **136**, 2550–2562 (2013).
41. Crockett, M. J. *et al.* Harm to others outweighs harm to self in moral decision making. **112**, (2014).
42. Boehm, C. Costs and benefits in hunter-gatherer punishment. *Behavioral and Brain Sciences* (2012). doi:10.1017/S0140525X11001403
43. Greene, J. T. Altruistic behavior in the albino rat. *Psychon. Sci.* (1969). doi:10.3758/BF03336420
44. Lamm, C., Decety, J. & Singer, T. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage* (2011). doi:10.1016/j.neuroimage.2010.10.014
45. Goldman, A. I. Simulation and Interpersonal Utility. *Ethics* (1995). doi:10.1086/293749
46. Lockwood, P. L. The anatomy of empathy: Vicarious experience and disorders of social cognition. *Behav. Brain Res.* **311**, 255–266 (2016).
47. Chang, S. W. C., Gariépy, J. F. & Platt, M. L. Neuronal reference frames for social decisions in primate frontal cortex. *Nat. Neurosci.* (2013). doi:10.1038/nn.3287
48. Lee, M. C. & Tracey, I. Imaging pain: A potent means for investigating pain mechanisms in patients. *Br. J. Anaesth.* (2013). doi:10.1093/bja/aet174
49. Banissy, M. J. & Ward, J. Mirror-touch synesthesia is linked with empathy. *Nat. Neurosci.* (2007). doi:10.1038/nn1926
50. Blakemore, S. J., Bristow, D., Bird, G., Frith, C. & Ward, J. Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. *Brain* (2005). doi:10.1093/brain/awh500
51. Jones, D. N. & Paulhus, D. L. Introducing the Short Dark Triad (SD3): A Brief Measure of Dark Personality Traits. *Assessment* **21**, 28–41 (2014).
52. Ahn, W.-Y., Haines, N. & Zhang, L. Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Comput. Psychiatry* **1**, 24–57 (2017).
53. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).
54. Akaike, H. Factor analysis and AIC. *Psychometrika* (1987). doi:10.1007/BF02294359
55. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* (2002). doi:10.1111/1467-9868.00353
56. Ashburner, J. *et al.* SPM12 Manual The FIL Methods Group (and honorary members) (2014). (2014).
57. Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E. & Roiser, J. P. Neurocomputational mechanisms of prosocial learning and links to empathy. *Proc. Natl. Acad. Sci.* (2016). doi:10.1073/pnas.1603198113