

Research Articles | Behavioral/Cognitive

Neural reward representations enable utilitarian welfare maximization

<https://doi.org/10.1523/JNEUROSCI.2376-23.2024>

Received: 19 December 2023

Revised: 4 March 2024

Accepted: 24 March 2024

Copyright © 2024 the authors

This Early Release article has been peer reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Neural reward representations enable utilitarian welfare maximization

Alexander Soutschek^{1*}, Christopher J. Burke^{2*}, Pyungwon Kang², Nuri Wieland³,
Nick Netzer², and Philippe N. Tobler²

¹ Department of Psychology, Ludwig Maximilian University Munich, Munich, Germany

² Department of Economics, University of Zurich, Zurich, Switzerland

³ Catholic University of Applied Sciences North Rhine-Westphalia, Cologne, Germany

* These authors contributed equally

Author contributions: CJB, NN, and PNT designed research; CJB, NW, and PK performed research; AS and CJB analyzed data; AS wrote manuscript, with contributions from CJB, NN, and PNT; all authors approved final manuscript version

Contact author:

Alexander Soutschek
Ludwig Maximilian University Munich
Department of Psychology
Leopoldstr. 13
80802 Munich, Germany
Email: alexander.soutschek@psy.lmu.de

Abstract

From deciding which meal to prepare for our guests to trading-off the pro-environmental effects of climate protection measures against their economic costs, we often must consider the consequences of our actions for the well-being of others (welfare). Vexingly, the tastes and views of others can vary widely. To maximize welfare according to the utilitarian philosophical tradition, decision makers facing conflicting preferences of others should choose the option that maximizes the sum of subjective value (utility) of the entire group. This notion requires comparing intensities of preferences across individuals. However, it remains unclear whether such comparisons are possible at all, and (if they are possible) how they might be implemented in the brain. Here, we show that female and male participants can both learn the preferences of others by observing their choices, and represent these preferences on a common scale to make utilitarian welfare decisions. On the neural level, multivariate support vector regressions revealed that a distributed activity pattern in the ventromedial prefrontal cortex (VMPFC), a brain region previously associated with reward processing, represented preference strength of others. Strikingly, also the utilitarian welfare of others was represented in the VMPFC and relied on the same neural code as the estimated preferences of others. Together, our findings reveal that humans can behave as if they maximized utilitarian welfare using a specific utility representation and that the brain enables such choices by repurposing neural machinery processing the reward others receive.

Key words: utilitarianism, VMPFC, neural reward system, social decision neuroscience, multivariate decoding

Significance statement

In many situations politicians and civilians strive to maximize the welfare of social groups. If the preferences of group members are in conflict, identifying the utilitarian welfare-maximizing option requires that decision makers can compare the strengths of conflicting preferences on a common scale. Yet, there is a fundamental lack of understanding which brain mechanisms enable such comparisons of conflicting utilities. Here, we show that brain regions involved in reward processing compute welfare comparisons by representing the preferences of others with a common neural code. This provides a neurobiological mechanism to compute utilitarian welfare maximization as desired by moral philosophy in the Humean tradition.

Introduction

Imagine you have invited two guests for dinner, one preferring pasta over sushi, the other preferring sushi over pasta. Which meal should you cook? Or, how should a politician decide if one part of the population values environmental protection over economic interests, while the other part prefers avoiding potential job losses over the benefits of pro-environmental actions? According to moral philosophers in the utilitarian tradition (Bentham, 1935; Hume, 2003), decision makers facing conflicting preferences of others should choose the option that maximizes the utilitarian welfare of the entire group. For example, if one guest strongly prefers pasta over sushi, whereas the other has only a slight preference for sushi over pasta, serving pasta to both will maximize the sum of their utilities compared to serving sushi to both.

Identifying the option that maximizes utilitarian welfare requires comparing intensities of preferences across individuals. Put differently, it is necessary to represent preferences of different individuals on a common (cardinal) scale rather than to just rank-order their preferences. Yet, there is a fundamental lack of conceptual understanding of how the required cardinal utility properties can be measured and compared interpersonally, or whether this is possible at all (Arrow, 1951; Binmore, 2007; Hammond, 1993; Harsanyi, 1955). For example, simply observing that one guest would choose pasta over sushi and the other would choose sushi over pasta does not reveal any information about the intensity of the preference. Proponents of utilitarian welfare have attempted to solve the problem by positing that humans possess perspective-taking skills that enable them to represent the intensities of others' preferences in a similar fashion as their own (Harsanyi, 1955). This notion requires that decision makers can represent and compare the preferences of others on a common scale. However, there is a fundamental lack of evidence whether decision makers actually can compare the preference strengths of others as well as which neurobiological substrate enables such comparisons.

The neuroscientific literature suggests that the subjective value underlying the decision maker's own preferences is encoded in a network comprising the ventromedial prefrontal cortex (VMPFC) and the striatum (Bartra et al., 2013; Clithero & Rangel, 2014). Crucially, these regions appear sensitive not only to selfish interests but also to social motives (Ruff & Fehr, 2014; Soutschek et al., 2017) as well as to vicarious rewards (Morelli et al., 2015). In particular, a subregion in VMPFC is active both when agents receive rewards and when they observe others receiving rewards, suggesting that the VMPFC might compute a receiver-invariant value signal. Moreover, VMPFC is involved in learning to predict others' behavior (Apps & Sallet, 2017), computes the subjective values of rewards when deciding on others' behalf (Nicolle et al., 2012), and processes the difference between others' and one's own preferences (Garvert et al., 2015). This makes the VMPFC and striatum plausible candidates for a neural substrate encoding the preferences of others on a common scale. In fact, the striatum was reported to encode efficiency (Hsu et al., 2008), though only under the unrealistic assumption that all recipients possess the same preference. Under more realistic assumptions (e.g., when inviting guests for dinner), others' preferences can differ and be in conflict, such that decision makers need to learn and integrate the potentially conflicting preferences. Thus, there is a knowledge gap regarding whether (and if so, how) regions involved in reward processing can represent and integrate conflicting preferences of others.

In-keeping with the positive approach to welfare analysis (Ambuehl & Bernheim, 2021), we studied whether and how the brain implements welfare decisions when others' preferences are in conflict. We first tested whether VMPFC and striatum encode the observed preferences of others like our own (hypothesis 1). The neural representations of others' preferences might then enable humans to compare the welfare-consequences of options for other agents on a common scale, using the representations of others' utility to maximize welfare (hypothesis 2). This further predicts that the change in utilitarian welfare associated with different options can

be decoded from activity in reward-encoding regions, such that representations of welfare should rely on the same neural code as the estimated preferences of others (hypothesis 3). Testing these hypotheses together reveals the neural mechanisms that enable the brain to compare and integrate the estimated, conflicting preferences of others when deciding about welfare distributions.

Materials and Methods

Participants

We invited 46 participants to take part in the experiment (mean age 23.3 years; standard deviation 3.3 years; 22 females, all right-handed). Data were collected in two cohorts: The first cohort of 23 participants was scanned in 2013. While such a sample size was common for imaging studies at that time, we decided to collect data from a further cohort of 23 participants at the beginning of 2023 to fulfill the higher standards with regard to statistical power and replicability in current fMRI research. We note that we statistically control for potential differences between the cohorts (and state of the scanner) in all analyses. Three out of the 23 participants of the first cohort were excluded from data analysis (one participant's data was lost due to a software glitch, two other participants did not complete all three tasks). All participants had normal or corrected-to-normal vision, were screened to exclude those with a previous history of neurological or psychiatric disease and gave informed consent. The study was approved by the Research Ethics Committee of the Canton of Zurich. Participants were asked not to eat anything for 4 hours prior to scanning.

Experimental procedures and task design

Upon arrival at the laboratory, participants were presented with food items (peanuts, almonds, raisins, crackers, gummy bears, and smarties) in sealed transparent plastic bags. Each

bag contained approximately 10g of a particular food item. Participants provided complete demographic information and indicated how much they liked each food item (11 point Likert scales from -5 = “strongly disliked” to +5 = “strongly liked”) and how many hours had passed since they had last eaten anything.

All computer tasks were scripted in the Cogent toolbox (Wellcome Laboratory of Neurobiology, University College, London, UK) for Matlab (Mathworks). For each participant, we chose a pair of food items that had been rated positively and differed by a rating of between 1 and 2. The more preferred food was labelled ‘A’ and the less preferred food was labelled ‘B’ for analysis purposes. For each participant, the two food items were randomly assigned a color (blue or red), which stayed constant across all tasks. Participants then engaged in a short learning task (outside the scanner) to associate the number of squares of a certain color on the screen with a quantity of food (1 square = 1 bag of food). In this task, red or blue colored squares were randomly presented on one side of the screen, and participants were asked to indicate this side using the left and right keys on the keyboard. Following a correct response, photographs of the associated food item were displayed under the squares, with the number of images corresponding to the number of bags of food represented by the squares. After completing 50 trials and verbally confirming which color was associated with each food item, participants entered the scanner and progressed to the individual decision task. All tasks after this learning task (individual decision task, preference learning task, and welfare maximization task) were performed inside the MRI scanner.

Individual decision task. In this task, participants were asked to choose between different quantities of A and B, ranging between 1 bag (10g) and 5 bags (50g). The quantities available in a given choice option were depicted by the number of colored (blue or red) squares in a 3-by-3 isoluminant matrix, to reduce the confounding factor of visual salience with food quantity (Figure 1A). Nine possible choice scenarios were displayed (5A:1B, 4A:1B, 3A:1B,

2A:1B, 1A:1B, 1A:2B, 1A:3B, 1A:4B, 1A:5B). Participants completed ten trials per choice scenario, giving 90 choices in total per participant. The side on which item A or B was presented was randomly determined on each trial. After a variable intertrial interval (ITI, 2-14s), the choice situation was presented for up to 4 seconds. Participants made their choices by pressing a response box with the index and middle fingers of the right hand. Registration of the response was then indicated to the participant by fading the chosen option for 2 seconds. We informed participants that at the end of the experiment we would randomly select one of their decisions and give them the chosen snack in the quantity of the selected decision.

Preference learning task. In each trial of this task, participants predicted the choices of two agents performing the individual decision task participants had performed before. Participants again viewed two choice options but also a photograph of one of the agents above the fixation cross (Figure 2A). The photos were from pilot participants who had displayed preferences and were gender-matched to participants. In a pilot study, we had measured the snack preferences of a large sample of participants and from these we selected two agents who displayed exchange rates of $1A = 2.5B$ and $2.5A = 1B$ for the main experiment. Note that we selected two agents with opposing snack preferences for the main experiment, but the identification of an agent having similar versus dissimilar preferences could be done only after the determination of a participant's exchange rate in the main experiment. Thus, A and B here do not refer to the snacks preferred and non-preferred, respectively, by the observer but to the two snacks (e.g., peanuts and raisins) used in an experimental session. In the main experiment, participants were required to predict which option the agent would choose within a maximum of 4 seconds. The side of the screen that A or B was displayed on was randomly determined for each trial. Agent 1 and agent 2 trials were randomly interspersed. Participants made their prediction by pressing a response box with the index and middle fingers of the right hand. The prediction was displayed to the participant by an arrow for 2 seconds. The actual choice of the

agent was then illustrated by fading the chosen option for 2 seconds. If the prediction of the participant was correct, ‘+10’ appeared below the fixation cross for 2 seconds, indicating the winning of 10 points which were later converted to Swiss Francs (CHF) at a rate of 1 point = 0.3 CHF. Conversely, wrong predictions were followed by a loss ‘-10’. Trials were separated by a variable ITI (2-14s). The agents faced the same choice scenarios as those used in the individual decision task. Crucially, the A:B exchange rate was different for agent 1 and agent 2, allowing for one of the agents and the participant to share a similar relative preference between the two foods (preferring A to B), while the other agent had opposing preferences (preferring B to A). The script generated agents’ choices using a softmax action selection rule,

$$p(A) = \frac{e^{\beta \times V(A)}}{e^{\beta \times V(A)} + e^{\beta \times V(B)'}}$$

where $V(xA) = x$ and $V(yB) = y/e$ (e = agent’s exchange rate). $P(A)$ is the probability of choosing option A, and β is an inverse temperature parameter (set at 3.5 for both agents to ensure that they differed only regarding their exchange rates, not their decision noise). Note again that during the generation of the observed choices in an experimental session the preferences of the observing participant were not known yet, such that for the choice generation one snack was arbitrarily defined as A and the other as B. Participants performed 12-20 trials per choice situation (more trials around the exchange rate/indifference point for each agent), giving 312 trials in total, split equally across two sessions.

Welfare maximization task. In each trial of the welfare maximization task, the participant was asked to make choices that assigned different quantities of A and B to agent 1 and agent 2. There was no incentive for participants in this task, and they received no specific instructions for how they should make choices. Options were of the form

- “xA for agent 1, yB for agent 2 versus yB for agent 1, xA for agent 2” (Figure 3A),
- “xA for agent 1, yA for agent 2 versus yB for agent 1, xB for agent 2”, or
- “xA for agent 1, yA for agent 2 versus xB for agent 1, yB for agent 2”,

with x and y ranging from 1 to 5. This format of the choice situations meant that there would often be a conflict of interest between the agents based on the estimated preferences of the participant. To maximize utilitarian welfare, the participant should choose distributions that were strongly preferred by one agent and only mildly disliked by the other, over distributions that were strongly disliked by one agent and mildly liked by the other, which requires an interpersonal comparison of preference intensities.

Participants were presented with options on the left and right side of the screen for up to 4 seconds. Option sides were randomly determined on each trial. In the example shown in Figure 3A, the participant chose between (left option) one blue item for agent 1, three red items for agent 2 versus (right option) three red items for agent 1, one blue item for agent 2. Participants made their choices using a response box with the index and middle finger of the right hand. After making their choice, the chosen option was faded for up to 6 seconds. Trials were separated by a variable ITI of 2-14 s. Participants made a total of 204 decisions, split across four sessions.

Data Acquisition

Images were acquired using a Philips Achieva 3 T whole-body scanner with an eight-channel sensitivity-encoding head coil (Philips Medical Systems) at the Laboratory for Social and Neural Systems Research, University Hospital Zurich. The task was projected on a display, which participants viewed through a mirror fitted on top of the head coil. We acquired gradient echo T2*-weighted echo-planar images (EPIs) with blood-oxygen-level-dependent (BOLD) contrast (slices = 37; repetition time = 2 s). We collected 380 – 550 volumes in each session of the experiment, together with five “dummy” volumes at the start and end of each scanning session. Participants each completed seven experimental sessions in the scanner, with short breaks between each session. Scan onset times varied randomly relative to stimulus onset times.

A T1-weighted 3D turbo field echo structural image was also acquired for each participant. Volumes were acquired at a -30° tilt to the anterior commissure–posterior commissure line, rostral more than caudal. Imaging parameters were the following: echo time = 30 ms; field of view = 220 mm. The in-plane resolution was 2.75×2.75 mm, with a slice thickness of 3 mm and an interslice gap of 0.5 mm. High-resolution T1-weighted structural scans were coregistered to their mean EPIs and averaged together to permit anatomical localization of the functional activations at the group level.

Data analysis

Behavioral analysis. Behavioral analyses were conducted with the lme4 package in R (Bates et al., 2014). In the individual decision task, we determined individual subjective value functions by regressing binary choices on scenario type (1 = 5A/1B, 2 = 4A/1B, ...), modelled both as fixed-effect predictor and participant-specific random slope in addition to random intercepts. As in all other analyses, we also added a predictor for Cohort to control for potential differences between the two samples. The participant-specific coefficients for slope and intercept allowed us to determine the exchange rate between the goods for each individual, which corresponds to the indifference point where an individual chooses both options with a probability of 0.5 ($exchange\ rate = -\frac{intercept}{slope}$). The same approach was used to determine the estimated exchange rates of the observed agents in the preference learning task, separately for the similar and the dissimilar agent.

In the welfare maximization task, we conducted a generalized linear mixed model that regressed binary choices (0 = right option chosen, 1 = left) on the welfare difference between the presented options (welfare left option minus welfare right option) and Cohort as fixed-effect predictors. In the baseline model, the welfare of an option was defined as the sum of the utilities of both agents, which were derived from their exchange rates as estimated by the participant in

the social learning task, using a linear utility representation. In a further regression model, we assessed whether participants made decisions according to the Rawlsian maximin principle that defines welfare as the utility of the worse-off agent. The utility of the worse-off agent for each option was defined as the lowest value of the agents' utilities of an option. As for utilitarian welfare, we regressed binary choices on the difference in the utilities for the worse-off agent between the left and the right option. Lastly, we also considered a model where welfare comparisons were based on the product rather than the sum of the agents' utilities.

Imaging analysis. Analysis of neuroimaging data was performed with SPM12 in Matlab (www.fil.ion.ucl.ac.uk/spm). The functional images of each participant were motion corrected, unwarped, slice-timing corrected (temporally aligned to the middle image), and co-registered to the anatomical image. Following segmentation, we spatially normalized the data into standard MNI space. Finally, data were smoothed with an 8 mm FWHM Gaussian kernel and high-pass filtered (filter cutoff = 128 seconds).

We performed univariate and multivariate analyses of the fMRI data. *Univariate analyses* were conducted separately for each task. In the individual decision task, we computed a general linear model (GLM-1) which included an onset regressor for the decision (modelled at the time of the response) and was modulated by a parametric modulator for the z-transformed absolute value difference between the options in the given trial. The value difference between the options was computed with the exchange rate between goods A and B as determined in the behavioral data analysis.

For the analysis of the preference learning task, GLM-2 included separate onset regressors for the prediction of the choice of the similar and dissimilar agent (modelled at the decision time), again modulated by the estimated value difference on the given trial. As second parametric modulator, we modelled the feedback participants received at the end of each trial

(+10 or -10) in order to disentangle the neural correlates of value representations from reward learning. We note that the mean correlation between these two parametric modulators was small-to-moderate, $r = 0.22$. As we orthogonalized feedback with respect to value difference, controlling for feedback did reduce the variation of the value difference regressor but allowed us to explain additional variance related to feedback processing. Lastly, GLM-3 modelled the welfare maximization task and included onset regressors for choices, which were parametrically modulated by the welfare difference between the options. In all models, we added 6 movement (3 translation and 3 rotation) parameters as covariates of no interest and convolved regressors with the canonical hemodynamic response function in SPM.

For statistical assessment, we first computed participant-specific contrasts for value difference (GLM-1, GLM-2, and GLM-3). For the second-level analysis, we entered the resulting contrast images from all participants in a between-participant, random effects analysis and conducted whole-brain second-level analyses using one-sample t -tests. In all second level analyses, we added a covariate for Cohort as control variable. For these analyses, we report results that survive whole-brain family-wise error (FWE) corrections at the peak or cluster level. To test the directed hypothesis that interpersonal utility comparisons are encoded in the VMPFC or striatum, we used a mask with pre-defined region of interest (ROIs) based on a meta-analysis of the neural basis of value coding (Bartra et al., 2013). This mask can be downloaded from <http://www.kablelab.com/resources.html>. We used it to account for multiple comparisons using small-volume FWE correction (note that this single ROI included both VMPFC and striatum regions, avoiding the problem of multiple hypothesis testing with multiple ROIs). As robustness check, we defined a further, anatomical rather than functional ROI for the VMPFC based the automated anatomical labelling atlas in SPM 12 (Rolls et al., 2020). All effects in VMPFC reported in this manuscript were robust to employing this anatomical mask for small-volume

correction, all $p < 0.04$. For the figures, we set the individual voxel threshold to $p < 0.001$. We report results in the MNI coordinate system.

As *multivariate analyses*, we performed support vector regressions (SVRs) with the decoding toolbox (Hebart et al., 2014). Following previous procedures (Kahnt et al., 2014), we performed the SVR on condition-specific estimated regression weights rather than on raw data. To obtain regression weights for neural activity related to the choice options in the three tasks, we performed additional analyses on the unsmoothed data with separate onset regressors for each combination of choice options (e.g., an onset regressor modelling trials with “5A/1B” offer for the similar agent). These participant- and run-specific regression weights were submitted to a first-level searchlight analysis (Haynes et al., 2007; Kriegeskorte et al., 2006), which for each participant examined the information in local fMRI patterns surrounding each voxel. For each voxel, we defined a spherical cluster of radius = 2 voxels (corresponding to a volume of 63 mm³) centered around that voxel. We conducted four different SVR analyses of interest: In SVR-1, we trained a linear SVR to predict the value difference between the choice options for one (e.g., the similar) agent from these local response patterns (training data). Using the SVR model, we then tested how well the value-related information for that agent predicted the value information for the other (e.g., dissimilar) agent (test data) within the cluster. In SVR-2, we assessed whether value differences in the preference learning task can be decoded from value differences in the individual decision task. Specifically, we defined the neural activation pattern in each task either as training or test data (i.e., linear SVR was trained on the data from the individual decision task and tested on the data from the preference learning task, and vice versa). SVR-3 used the data from the preference learning task or the welfare maximization task as training data and the data from the other task as test data. Lastly, SVR-4 matched SVR-3 but replaced the preference learning task with the individual decision task. In all analyses, we analyzed prediction accuracy via Fisher's Z-transformed correlation coefficient between the

predicted value/welfare differences and the actual value/welfare differences in the test data. We then entered the accuracy maps to second-level analyses and performed statistical tests in the same way as for the contrast maps in the univariate analyses.

Neural Model Comparisons. In our main analyses described above, we computed the estimated subjective values of rewards via the individually determined exchange rates of the two goods assuming a linear relationship between the quantities of a good and its subjective value. One can alternatively determine subjective values using non-linear relationships between good quantity and subjective value (e.g., squared, square rooted, logarithmic, exponential) or from the estimated logistic curves of each participant (individual decision task) or observed agent (preference learning and welfare maximization tasks). To interrogate these alternatives, we assessed whether activation in the neural reward system was better explained by value differences determined via these alternatives. For this purpose, we re-computed the GLMs for all tasks with value differences derived from the described alternative non-linear utility functions (e.g., value difference = $| \text{value}_{\text{left}}^2 - \text{value}_{\text{right}}^2 |$). We then compared how well the different GLMs explained neural activity in the value coding ROI by computing the Akaike information criterion (AIC) for each participant with the MACS toolbox in SPM (Soch & Allefeld, 2018).

Results

VMPFC encodes the estimated preferences of others irrespective of the preferences of the observer

In the individual decision task, participants chose between two types of food items, such as raisins and peanuts, offered in different quantities. Food types were represented by color and quantities by the number of squares, with one colored square representing approximately 10 g

of the item (Figure 1A). Fitting sigmoidal curves to participants' stochastic choice data allowed us to quantify the individual trade-off between food type and food quantity. Equating a choice probability of 0.50 with indifference, we can identify an "exchange rate" between A and B for each participant. On average, the exchange rate between A and B, where A always refers to the participant's more preferred food item, was $1A=2.41B$ (Figure 1B). Thus, on average 1 unit of food item A was roughly equivalent to 2.4 units of food item B for the participants.

We first assume a linear value function (Levy & Glimcher, 2011; Padoa-Schioppa & Assad, 2006) so that the unsigned subjective value difference between x_A and y_B is $|x - y/e|$, where e is a participant's exchange rate described above. As to be expected, decision times were longer for choice problems with subjective value differences closer to zero, $\beta = -0.10$, $t(41) = 2.53$, $p = 0.01$, suggesting that choices closer to the indifference point were more difficult for the participants (Figure 1C). On the neural level, we found that differences in subjective value between the presented options correlated with activation in the VMPFC (peak coordinates: $x = -6$, $y = 44$, $z = -7$, $t(41) = 3.84$, $p = 0.02$). This activation was small-volume family-wise error (FWE) corrected within an a priori-defined region-of-interest (ROI) in bilateral striatum and VMPFC based on meta-analysis of neural value coding (Bartra et al., 2013) (Figure 1D and Table 1). Consistent with previous findings (Bartra et al., 2013; Clithero & Rangel, 2014), the VMPFC thus encoded own preferences during individual decision making.

Next, we used the second task to test whether the neural reward system also encoded the preferences of others. In this task, participants observed the (previously recorded) choices of two other agents (Fig. 2A). This task resembled the individual decision task in the use of visual symbols and primary rewards, but the objective of participants was to predict the choices of both agents. The preferences of one agent were similar to those of the participant (i.e., preferring food item A to B with an exchange rate of $1A=2.5B$), whereas the preferences of the

other agent were dissimilar to those of the participant (i.e., preferring food item B to A with an exchange rate of 2.5A=1B). Correct predictions won the participant 10 points, which were converted to money at the end of the experiment, while incorrect predictions yielded -10 points. Participants were able to successfully predict the choices of the two agents (91.2% correct predictions) (Figure 2B), and exchange rates estimated by fitting sigmoid curves to participants' predictions were close to the true exchange rates of both agents (1A:2.40B for the similar agent; 2.42A:1B for the dissimilar agent). A generalized linear mixed model regressing correct predictions on the observed agents' true absolute value differences between the options (assuming a linear value function) suggested that prediction accuracy decreased around the true indifference points of both agents, $z > 11.38$, $p < 0.001$, with no significant differences in prediction accuracy between similar and dissimilar agents, $z = 1.24$, $p = 0.21$ (Figure 2C). Thus, participants learned the differing preferences of the two agents similarly well.

We then tested our first hypothesis that the neural reward system encodes not only one's own preferences but also the estimated preferences of others. We computed subjective (linear) value differences between options based on participants' estimates of the two agents' exchange rates. We then examined the neural encoding of these value differences at the time of choice prediction, controlling for feedback on each trial and collapsing across the similar and dissimilar agent. Univariate analyses revealed that the value differences between the presented options correlated with activation in VMPFC (Figure 2D; peak coordinates: $x = 3$, $y = 44$, $z = -10$, $t(41) = 3.60$, $p = 0.04$, small-volume FWE-corrected with the meta-analysis-based VMPFC/striatum ROI). Moreover, hippocampal areas and the temporoparietal junction showed whole-brain corrected correlations with estimated value difference (Table 2). No brain regions showed significant differences in value coding between similar and dissimilar agents (all $p > 0.88$ for whole-brain FWE-correction at cluster or peak level); note that a Bayes factor of 2.8 computed with the BayesFactor package on parameters extracted from the value-coding ROI provided

only weak evidence in favor of the null (similar = dissimilar) relative to the alternative (similar $\neq$ dissimilar) hypothesis. A conjunction analysis testing the overlap between individual (individual decision task) and social value representations (preference learning task) revealed no significant activation for whole-brain FWE correction or small-volume correction with the VMPFC/striatum ROI, all $p > 0.23$. Taken together, in our task the VMPFC encoded the estimated preferences of others (supporting hypothesis 1), irrespective of the similarity of others' preferences with one's own.

As we had found no significant difference between the neural representations of the preferences of others with either similar or dissimilar preferences, we next assessed whether the preferences of similar versus dissimilar others were represented in the same neural code. In principle, two different but intermingled populations of neurons could separately represent value for similar and dissimilar agents. To investigate this possibility, we performed a multivariate support vector regression (SVR) with a searchlight procedure testing whether the neural representations of others with dissimilar preferences can be decoded from neural preference representations for similar others, and vice versa (Hebart et al., 2014). We found that the predicted value differences of others can be decoded in the VMPFC and striatum (peak coordinates VMPFC: $x = 6$, $y = 41$, $z = -10$, $t(41) = 3.67$, $p = 0.04$, small-volume FWE-corrected; peak coordinates striatum: $x = -6$, $y = 5$, $z = -4$, $t(41) = 4.16$, $p = 0.005$, small-volume FWE-corrected, Table 3). As robustness check, we tested whether the predicted value differences of similar and dissimilar agents can be cross-decoded directly within the VMPFC and striatum ROIs (rather than using a whole-brain searchlight). Prediction accuracy significantly differed from zero both in the VMPFC, $t(42) = 2.29$, $p = 0.03$, and the striatum ROI, $t(42) = 3.32$, $p = 0.002$. Thus, value regions appear to represent the observed preferences of similar and dissimilar others with the same neural code.

To test whether the same neural code represents individual preferences and the estimated preferences of others, we performed a between-task decoding analysis, training classifiers on neural correlates of value differences in the individual decision task and testing the trained classifiers on value differences in the preference learning task, and vice versa. However, there was no evidence for successful between-task decoding in the VMPFC or any other brain region, all $p > 0.36$, FWE-corrected (Figure 2F), and this null result also held when we decoded directly from the VMPFC and striatum ROIs instead of using a searchlight procedure, both $t < 1$, both $p > 0.5$. Moreover, a Bayes factor of 5.6 (computed based on parameters extracted from the value-coding ROI) provided positive evidence for the null (no between-task decodability) relative to the alternative (between-task decodability) hypothesis. Thus, VMPFC appears to use a different neural code to represent one's own and others' preferences. The common coding of subjective value across different (similar and dissimilar) agents fulfills a precondition for interpersonal value comparisons performed by the brain.

By design, the behavior of the two observed agents in the preference learning task is partly related to the number of food items presented on the screen (despite their opposing tastes, both agents preferred e.g. 5 portions of peanuts over one portion of raisins). We therefore asked whether the findings above were driven by visual properties of presented stimuli and conducted a control analysis where we decoded the number of displayed items for one agent from neural activation related to the number of displayed items for the other agent. This SVR revealed no significant activation in VMPFC or striatum, neither with a whole-brain searchlight (although there was a non-significant trend-level effect in the striatum: $p = 0.07$, small-volume corrected) nor with an ROI decoding approach (both $p > 0.31$ for VMPFC and striatum ROI). Decoding was possible only in primary visual and motor cortex (both $p < 0.003$, whole-brain FWE-corrected at cluster level). Prediction accuracy was marginally higher in the preference-decoding than in the item number-decoding SVR in the VMPFC ROI, $t(42) = 2.00$, $p = 0.05$,

but not in the striatum ROI, $t(42) = 1.51$, $p = 0.14$. These findings suggest that the number of displayed food items cannot explain the role of the VMPFC for representing others' preferences.

VMPFC encodes Utilitarian welfare

After observing the choices of the similar and dissimilar agent, participants progressed to the welfare maximization task. In this task, participants chose between two different distributions of goods A and B for agent 1 and agent 2 (for example, “1A for agent 1 and 3B for agent 2” versus “3B for agent 1 and 1A for agent 2”; Figure 3A). In most of these decisions, there was a conflict of interest between the two agents based on the preferences estimated by the participant who made the allocation decision, with one agent inferred to favor the first option and the other agent favoring the second option. Since there were only two agents and two options, it was not possible in those cases to make a justified decision based only on the estimated ordinal preferences of the agents. However, given that participants seem to encode the preference strengths of others in overlapping neural codes (as suggested by the preference learning task), they might be able to compare these conflicting utilities to identify the welfare-maximizing option.

In line with this prediction, participants chose the utilitarian welfare-maximizing option in 71% of all trials, and the (linear) utilitarian welfare difference between the two options impacted decisions strongly, $\beta = 1.29$, $z = 10.61$, $p < 0.001$ (Figure 3B). In contrast, prominent alternative models explained choices less well. For example, a Rawlsian decision strategy (Rawls, 2020) where participants are meant to maximize the utility of the worse-off agent could explain only 49% of choices. A model according to which welfare comparisons are based on the product (rather than the sum) of utilities (Nash, 1953) explained 68% of choices.

Thus, participants often appeared to maximize welfare in the standard utilitarian sense and were sensitive to the utilitarian welfare consequences of the two options.

Having established that individuals often made choices in the welfare maximization task as if they were maximizing utilitarian welfare using linear utility representations of estimated exchange rates, we tested the second hypothesis that these choices rely on brain regions involved in reward processing. Univariate imaging analyses (GLM-3) revealed that the utilitarian welfare difference between the two options correlated with activation in the VMPFC (Figure 3C, peak coordinates: $x = -3$, $y = 56$, $z = -7$, $z = 3.55$, $t(41) = 3.87$, $p = 0.02$, small-volume FWE-corrected within value-coding ROI). Thus, decision makers not only behave as if they represent and compare subjective values of different agents on a common scale, but also process on the neural level the resulting welfare differences parametrically and categorically. Moreover, GLMs assuming welfare comparisons based on the Rawlsian maximin rule (AIC = 2215) or the product rule (AIC = 2215) provided no better fit than the utilitarian welfare maximization rule (AIC = 2214).

Given that the VMPFC represents both the estimated preferences of others and seems to perform utility comparisons between two agents, one might ask whether these distinct but related value signals rely on the same neural code. If this is the case, it should be possible to decode utilitarian welfare differences in the welfare maximization task by training an SVR analysis on the estimated value differences in the preference learning task, and vice versa (between-task cross-decoding). We found this to be the case: Activity patterns in the VMPFC allowed decoding utilitarian welfare differences with classifiers trained on the estimated value differences of others, and vice versa (peak coordinates: $x = 0$, $y = 50$, $z = -10$, $t(41) = 4.07$, $p = 0.03$, small-volume FWE-corrected within value-coding ROI). This result was robust to directly decoding from the VMPFC ROI ($t(42) = 2.65$, $p = 0.01$; Figure 3D) but not from the striatum ROI ($t(42) = 1.68$, $p = 0.10$). We also assessed whether utilitarian welfare differences could be

decoded from individual utility representations in the individual decision task. We found no significant VMPFC activity either with a whole-brain searchlight approach ($t(41) = 3.55$, $p = 0.15$, small-volume corrected) or with an ROI-based approach (both $t < 1$, both $p > 0.66$), providing no evidence for a common coding of individual values and interpersonal welfare comparisons in VMPFC (Figure 3E). Prediction accuracy in the VMPFC ROI also tended to be higher for decoding utilitarian welfare differences from the predicted value differences of others in the preference learning task than from participants' own values in the individual decision task, $t(42) = 1.83$, $p = 0.07$. However, we note that this non-significant trend should be interpreted with caution. Taken together, the data suggest interpersonal utility comparisons and estimated subjective values of others to be implemented by related neural codes.

Neural model comparisons

While the results for the neural analyses (GLM-1 – GLM-3) assumed linear value functions where the subjective value differences between x_A and y_B are given by $|x - (y/e)|$ (with e representing the individual exchange rate), we also tested alternative non-linear value functions that would generate the same behavior using squared (so that the value difference becomes $|x^2 - (y/e)^2|$), square-rooted ($|x^{0.5} - (y/e)^{0.5}|$), logarithmic ($|\log(x) - \log(y/e)|$), and exponential transformations ($|\exp(x) - \exp(y/e)|$). In the individual decision task (GLM-1), direct comparisons at the neural level suggested that no alternative specification (all $AIC \geq 718$) explained activity in the unbiased value coding ROI better than the linear function ($AIC = 718$). The same held when we derived value differences directly from logistic curves fitted to choice data, which corresponds to the estimation of a logistic random utility model ($AIC = 719$). This suggests that the linear value function is a reasonable approximation to how the brain encodes individual preferences, at least with regard to the choice set of our task.

Also in the preference learning task (GLM-2), we explored whether non-linear value functions (squared, square-rooted, logarithmic, or exponential transformations, as well as values directly derived from the fitted logistic curves) explained value-related VMPFC activity better than the assumed linear model. Again, the linear specification ($AIC = 1464$) was not outperformed by any tested alternative function (all $AIC \geq 1464$), compatible with the linear model representing a reasonable assumption for the current data.

Finally, neural model comparisons for the welfare maximization task (GLM-3) provided no evidence that activation in the value-coding ROI was better explained by utilitarian welfare differences based on non-linear (all $AIC \geq 2216$) rather than the assumed linear utility functions ($AIC = 2214$). When we compared different utility functions on the behavioral level, the linear model explained 71% of all choices and was not outperformed by the squared (71%), square-rooted (71%), exponential (71%), or logarithmic model (68%). However, when we used the AIC to assess how well continuous welfare differences explained welfare choices, the linear utility model ($AIC = 10,419$) explained the choice data better than the squared ($AIC = 11,292$) and the exponential model ($AIC = 12,235$) but was outperformed by the square-rooted ($AIC = 9,523$) and the log-transformed models ($AIC = 9,235$). Thus, concave value transformations might provide the best fit for the behavioral (though not the neural) data, consistent with previous findings (Ambuehl & Bernheim, 2021). Importantly, the imaging results for the linear utility function are robust to applying square-rooted and logarithmic utility transformations: In GLM-3, welfare differences based on such transformations correlated with activation in the VMPFC, both $p < 0.001$, small-volume FWE-corrected within value-coding ROI. Moreover, in the SVR analyses welfare differences could be decoded from the estimated value differences of others in the VMPFC when applying concave transformations to both the welfare and estimated value differences, both $p < 0.01$. We note that these comparisons of utility functions should be considered as exploratory given that the experiment was not designed to determine the precise

shape of the utility function underlying utilitarian welfare comparisons on the behavioral and neural level. In any case, we emphasize that our conclusions regarding the VMPFC's role for welfare comparisons are robust to assuming alternative, non-linear shapes for the utility function.

Discussion

Utilitarian welfare comparisons are impossible based on ordinal preferences alone, as they require comparing the strength of others' preferences (Arrow, 1951; Hammond, 1993). It is therefore an important question whether and how humans and their brains can learn and represent others' preferences with the required cardinal properties (Alós-Ferrer et al., 2021; Harsanyi, 1955). Here, we show that, through observation, humans can learn to predict the choices of others and use that knowledge as if they represented and compared the strength of the inferred preferences in the brain, which allows selection of the option that maximizes the utilitarian welfare of the group. Our data suggest that humans can decide as if they maximized utilitarian welfare using a utility representation of the observed behavioral preferences of others.

The estimated preferences of others were represented in the observer's VMPFC, a central region processing value (Bartra et al., 2013; Clithero & Rangel, 2014). The VMPFC fulfilled three criteria (reflecting our three hypotheses) that enabled individuals to decide as if they were making utilitarian welfare comparisons. First, VMPFC activity encoded the estimated preferences of other agents irrespective of individual preferences. Second, VMPFC activity integrated the estimated preferences of other agents, enabling humans to compare the welfare consequences of different options and to choose the option that maximizes utilitarian welfare based on the encoded utility representation. Third, this neural comparison process was possible

because others' preferences, as well as the computed welfare difference signal, were represented in similar neural codes in VMPFC, as revealed by the multivariate analyses.

Our findings thus suggest novel important properties of regions that have been linked to the representation of rewards: While past studies ascribed VMPFC and striatum a role for social decision making (Cooper et al., 2012; Ruff & Fehr, 2014; Soutschek et al., 2017; Strombach et al., 2015), they focused on how these regions encode prosocial outcomes in comparison to selfish rewards. Our finding that VMPFC can represent others' preferences independently of one's own is consistent with previous findings linking the VMPFC to vicarious rewards or representing subjective values for others (Morelli et al., 2015; Nicolle et al., 2012). While some studies reported that rewards for others are represented in more dorsal parts of VMPFC than rewards for the self (Hill et al., 2016; Sul et al., 2015), Nicolle et al. (2012) suggest that the VMPFC encodes the value of self- or other-regarding rewards depending on the current task demands. The current findings go beyond these existing studies by showing that VMPFC can not only represent others' preferences but also compare them in order to compute a utilitarian net utility signal. Thus, the generality of the VMPFC for value coding in social interactions enables humans to represent and compare the revealed preferences of others even if those are irrelevant for their selfish interests and different from their own preferences. Similar and dissimilar preferences were represented in overlapping neural codes (as suggested by our multivariate analysis). We note though that cross-agent decoding was not perfect, suggesting that activation patterns were not identical. This difference may allow the neural system to both parsimoniously distinguish between different others and to integrate their preferences. In contrast, neural representations of others' preferences could not be decoded from activity patterns related to the preferences of the participants. Thus, the VMPFC encodes both the value of selfish and other-regarding rewards with different multivariate codes. This coding scheme may complement that of distinct univariate representations of one's own and others' preferences

in different subsystems during social learning (Sul et al., 2015). In line with the idea that different subregions within VMPFC correlate with self- and other-related rewards, a conjunction analysis revealed no overlap between self-related and other-related value representations in VMPFC. In any case, our findings advance the neuroscience of value-based decision making by revealing that within one given value region (VMPFC), individual values and social values are represented with distinct neural patterns.

Representations of others' preferences with the required cardinal properties (Alós-Ferrer et al., 2021; Harsanyi, 1955) enable decision makers to distribute resources between recipients by directly comparing the utilitarian welfare consequences of different options. Our findings demonstrate such representations in VMPFC. Specifically, the difference in welfare between different options was encoded in VMPFC, suggesting that the VMPFC encodes not only the single preferences of other agents but also the combined net utilitarian welfare. Strikingly, we could decode differences in welfare from the neural representations of others' preferences (and vice versa), such that interpersonal value comparisons and the estimated preferences of others appear to rely on a shared neural code. Thus, by integrating the welfare consequences for multiple agents, we identify the VMPFC as the neurobiological substrate enabling utilitarian welfare maximization.

These neuroscientific findings directly speak to theoretical accounts in moral philosophy and economics of how to aggregate the conflicting preferences of different agents. The classical social choice literature studies the problem under the assumption that only ordinal preferences of agents, which can be deduced from observed behavior, are used in the aggregation (Ambuehl & Bernheim, 2021; Arrow, 1951). However, accounts based on rank-ordered preferences have little to say about many real-world choice problems and the choice problem we studied here where ordinal preferences are not sufficient to make a decision on the optimal outcome distribution. Utilitarian welfare maximization (Harsanyi, 1977) provides a

possible solution to these problems. This approach is in line with philosophical utilitarianism (Bentham, 1935; Hume, 2003) but requires cardinal utility information and interpersonal comparisons of utility. Crucially, our data show that utilitarian welfare maximization is enabled by the neural reward system's ability to compare the preferences of others on a cardinal scale. While our empirical data cannot contribute to the normative philosophical debate on how welfare should be defined in societies, they suggest that interpersonal comparisons of imputed utilities take place in the brain, and that people and their brains tend to favor a utilitarian over a Rawlsian approach that puts particular weight on the agent who is worst off (Rawls, 2020). However, our findings should not be misinterpreted as implying that people generally prefer utilitarian over Rawlsian welfare maximization, as various aspects of the choice menu like a difference in endowments could affect preferences for Rawlsian versus utilitarian outcomes.

It is worth mentioning some limitations of our study. First, participants were aware that their choices in the welfare maximization task had no real consequences for the others' payoff. While decisions affecting one's own payoff can (Camerer & Mobbs, 2017; Vlaev, 2012), but often do not (Brañas-Garza et al., 2021; Taylor, 2013), differ between real and hypothetical rewards, such differences remain largely unexplored for choices that affect only people other than the decision maker. Moreover, while in our study participants knew only the food preferences of the other agents, future studies may investigate how additional information about the receivers (e.g., political affiliation, demographics, group membership, hobbies, profession) affects welfare distributions. Lastly, participants' decisions in the welfare maximization task had no consequences for their own payoff, which may rarely be the case in real world decisions (when inviting friends for dinner, we may consider not only their but also our own food preferences). In the extreme case where the welfare is to be distributed only between the decision maker and one other person, the choice situation reflects the canonical dictator game. In team dictator games, decision makers decide how to distribute endowments between groups

of dictators and receivers, but contrary to our design decisions require a consensus between dictators (Kocher et al., 2020). It yet needs to be determined how strongly decision makers consider their own preferences when distributing welfare among different people and themselves.

In conclusion, we show that the preferences of others can be represented and compared on a common scale by the VMPFC, which we identify as a neural substrate of welfare representations, irrespective of similarity to the preferences of the observer. This may enable humans in principle to impartially maximize the well-being of other people when their decisions affect others, although in the real world, such decisions may often be influenced by partisanship.

Acknowledgments

We thank Karl Treiber for expert assistance with data acquisition. This study was funded by the Swiss National Science Foundation (PP00P1_128574; 100014_165884; IZKSZ3_162109; 100019_176016; 10001C_188878 to PNT). AS received an Emmy Noether fellowship (SO 1636/2-1) from the German Research Foundation.

Data availability statement

The behavioral data supporting the findings of this study are publically available on the Open Science framework (OSF; https://osf.io/6n4af/?view_only=70200271b8b840f2a34e5c69525927ea). Neuroimaging data are available on NeuroVault (<https://neurovault.org/collections/ZFRLNBMB/>).

References

- Alós-Ferrer, C., Fehr, E., & Netzer, N. (2021). Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy*, 129(6), 1828-1877.
- Ambuehl, S., & Bernheim, B. D. (2021). *Interpreting the will of the people: a positive analysis of ordinal preference aggregation*.
- Apps, M. A. J., & Sallet, J. (2017). Social Learning in the Medial Prefrontal Cortex. *Trends Cogn Sci*, 21(3), 151-152. <https://doi.org/10.1016/j.tics.2017.01.008>
- Arrow, K. J. (1951). An extension of the basic theorems of classical welfare economics. Proceedings of the second Berkeley symposium on mathematical statistics and probability,
- Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, 76, 412-427. <https://doi.org/10.1016/j.neuroimage.2013.02.063>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Bentham, J. (1935). *An Introduction to the Principles of Morals and Legislation: James Mills. Essays on Government, Jurisprudence, Liberty of the Press, and Law of Nations*. Doubleday, Doran, Incorporated.
- Binmore, K. (2007). Interpersonal comparison of utility.
- Brañas-Garza, P., Estepa-Mohedano, L., Jorrat, D., Orozco, V., & Rascón-Ramírez, E. (2021). To pay or not to pay: Measuring risk preferences in lab and field. *Judgm Decis Mak*, 16(5), 1290-1313.
- Camerer, C., & Mobbs, D. (2017). Differences in Behavior and Brain Activity during Hypothetical and Real Choices. *Trends Cogn Sci*, 21(1), 46-56. <https://doi.org/10.1016/j.tics.2016.11.001>
- Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Soc Cogn Affect Neurosci*, 9(9), 1289-1302. <https://doi.org/10.1093/scan/nst106>
- Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *J Cogn Neurosci*, 24(1), 106-118. <https://doi.org/10.1162/jocn.a.00114>
- Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E., & Dolan, R. J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron*, 85(2), 418-428. <https://doi.org/10.1016/j.neuron.2014.12.033>
- Hammond, P. J. (1993). 7. Interpersonal comparisons of utility: Why and how they are and should. *Interpersonal Comparisons of Well-Being*, 200.
- Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4), 309-321.
- Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social research*, 623-656.
- Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr Biol*, 17(4), 323-328. <https://doi.org/10.1016/j.cub.2006.11.072>
- Hebart, M. N., Gorgen, K., & Haynes, J. D. (2014). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Front Neuroinform*, 8, 88. <https://doi.org/10.3389/fninf.2014.00088>
- Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nat Commun*, 7(1), 12722. <https://doi.org/10.1038/ncomms12722>
- Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: distributive justice and neural encoding of equity and efficiency. *Science*, 320(5879), 1092-1095. <https://doi.org/10.1126/science.1153651>
- Hume, D. (2003). *A treatise of human nature*. Courier Corporation.
- Kahnt, T., Park, S. Q., Haynes, J. D., & Tobler, P. N. (2014). Disentangling neural representations of value and salience in the human brain. *Proc Natl Acad Sci U S A*, 111(13), 5000-5005. <https://doi.org/10.1073/pnas.1320189111>

- Kocher, M. G., Praxmarer, M., & Sutter, M. (2020). Team decision-making. *Handbook of labor, human resources and population economics*, 1-25.
- Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proc Natl Acad Sci U S A*, 103(10), 3863-3868. <https://doi.org/10.1073/pnas.0600244103>
- Levy, D. J., & Glimcher, P. W. (2011). Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain. *Journal of Neuroscience*, 31(41), 14693-14707.
- Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *Neuroimage*, 112, 244-253. <https://doi.org/10.1016/j.neuroimage.2014.12.056>
- Nash, J., F. (1953). Two-person cooperative games. *Econometrica*, 21(1), 128-140.
- Nicolle, A., Klein-Flugge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron*, 75(6), 1114-1121. <https://doi.org/10.1016/j.neuron.2012.07.023>
- Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090), 223-226. <https://doi.org/10.1038/nature04676>
- Rawls, J. (2020). *A theory of justice: Revised edition*. Harvard university press.
- Rolls, E. T., Huang, C. C., Lin, C. P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *Neuroimage*, 206, 116189. <https://doi.org/10.1016/j.neuroimage.2019.116189>
- Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nat Rev Neurosci*, 15(8), 549-562. <https://doi.org/10.1038/nrn3776>
- Soch, J., & Allefeld, C. (2018). MACS - a new SPM toolbox for model assessment, comparison and selection. *Journal of neuroscience methods*, 306, 19-31. <https://doi.org/10.1016/j.jneumeth.2018.05.017>
- Soutschek, A., Burke, C. J., Beharelle, A. R., Schreiber, R., Weber, S. C., Karipidis, I. I., . . . Tobler, P. N. (2017). The dopaminergic reward system underpins gender differences in social preferences. *Nature Human Behaviour*, 1(11), 819-827. <https://doi.org/10.1038/s41562-017-0226-y>
- Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, II, Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proc Natl Acad Sci U S A*, 112(5), 1619-1624. <https://doi.org/10.1073/pnas.1414715112>
- Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proc Natl Acad Sci U S A*, 112(25), 7851-7856. <https://doi.org/10.1073/pnas.1423895112>
- Taylor, M. P. (2013). Bias and brains: Risk aversion and cognitive ability across real and hypothetical settings. *Journal of Risk and Uncertainty*, 46, 299-320.
- Vlaev, I. (2012). How different are real and hypothetical decisions? Overestimation, contrast and assimilation in social interaction. *Journal of Economic Psychology*, 33(5), 963-972.

Table 1. Characteristics of peak activations correlating with the value difference between the two choice options in the individual decision task. We report significant effects that survived whole-brain family-wise error correction at the peak or cluster level.

Hem = Hemisphere (L = left, R = right); BA = Brodmann area

Region	Hem	BA	MNI Coordinates			k	t
			X	Y	Z		
Supramarginal gyrus	R	40	42	-25	26	129	4.67

Table 2. Characteristics of the peak activations correlating with the observed value differences in the preference learning task. We report significant effects that survived whole-brain family-wise error correction at the peak or cluster level.

Hem = Hemisphere (L = left, R = right); BA = Brodmann area

Region	Hem	BA	MNI Coordinates			k	t
			X	Y	Z		
Hippocampus	R		39	-43	-1	38	5.00
Sensory association cortex	R	5	15	-43	68	378	4.72
Supramarginal gyrus/TPJ	R	40	51	-25	20	133	4.56

Table 3. Anatomical locations and MNI coordinates of the peak activations where value differences in the preference learning task could be decoded from neural pattern in a searchlight support vector regression. We report significant effects that survived whole-brain family-wise error correction at the peak or cluster level.

Hem = Hemisphere (L = left, R = right); BA = Brodmann area

Region	Hem	BA	MNI Coordinates			k	T
			X	Y	Z		
Supramarginal gyrus	R	40	39	-49	53	1105	8.22
Premotor cortex	L	6	-27	8	62	2204	6.73
Orbitofrontal cortex	R	47	36	23	-22	223	6.35
Superior parietal lobule	L	7	-24	-67	53	1266	6.14
Cerebellum	R		15	-85	-34	22	5.87
Fusiform gyrus	R	37	57	-49	-13	82	5.36
Hippocampus	R		15	-10	-10	148	5.24
Operculum	L	44	-48	14	11	47	5.18
Middle temporal gyrus	R	21	66	-10	-16	29	5.16
Dorsolateral prefrontal cortex	L	9	-48	32	32	38	5.13
Occipital cortex	L	19	-42	-70	2	214	5.12
Ventromedial prefrontal cortex	R	11	6	47	-16	52	5.10
Orbitofrontal cortex	L	47	-33	23	-16	261	5.01

Figure captions

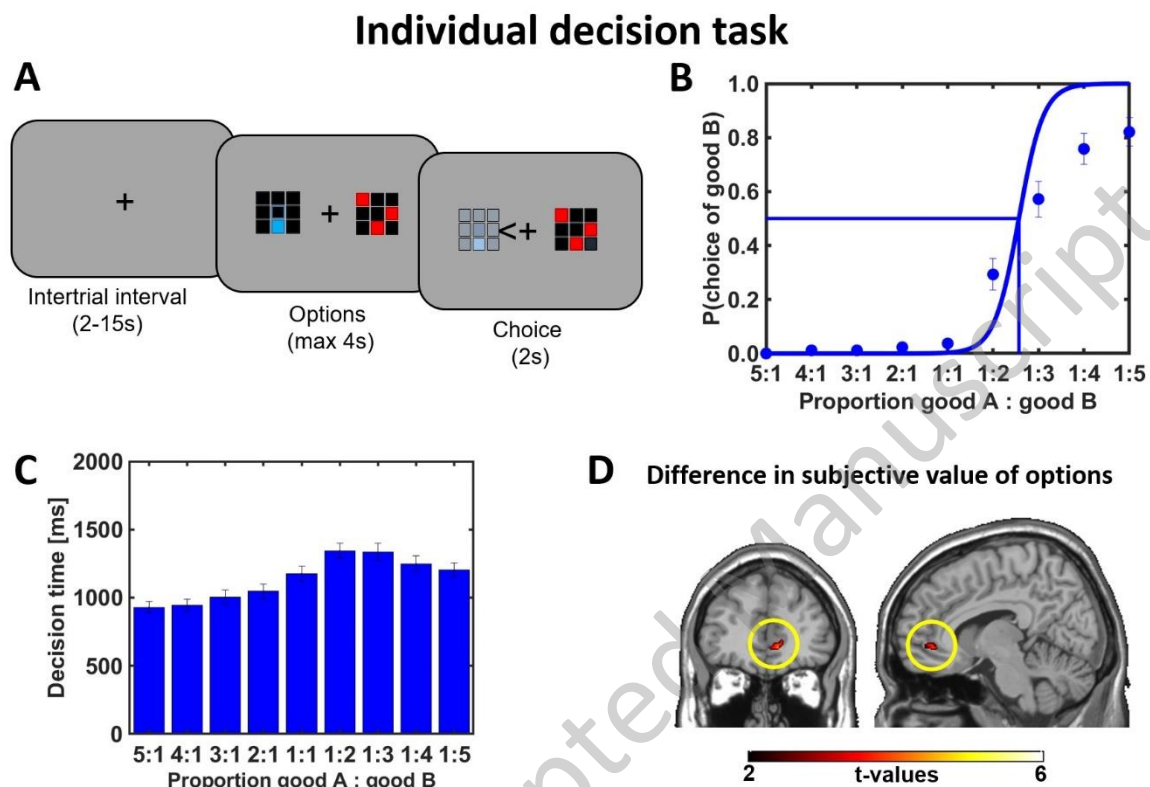
Figure 1. Individual decision task: design, behavior, and individual value-related activity. (A) In each trial of the individual decision task, participants chose between different quantities of goods A and B (e.g., 1 portion of peanuts versus 3 portions of raisins) for themselves. (B) Fitting logistic curves to participants' choice data allowed determining the indifference point where a participant was equally likely to choose A and B. The indifference point (i.e., the intersection of the vertical and horizontal blue lines) can be interpreted as an exchange rate expressing the value of one good in quantities of the other (e.g., $1A = 2.4B$). In the figure, A and B refer to the more preferred and to the less preferred good for each participant. (C) Decision times were slowest around participants' indifference points, compatible with higher choice difficulty. (D) In the brain, absolute value difference between the two choice options correlated with activation in the VMPFC.

Figure 2. Preference learning task: design, behavior, and social value-related activity. (A) In each trial of the preference learning task, participants predicted the choices of two agents whose preferences were either similar or dissimilar to their own. After each prediction, participants received feedback about whether the prediction was correct (“+10”) or incorrect (“-10”). Instead of drawings, photos of other agents were shown. (B, C) Participants could reliably infer the other agents' preferences from their choices, with expected drops in accuracy around the other agents' indifference points (similar agent: $1A=2.5B$; dissimilar agent: $2.5A=1B$; illustrated by the green and red crossing lines in B). (D) Univariate analyses revealed that VMPFC activity encoded the observed agents' estimated value difference between goods A and B, and (E) a multivariate support vector regression indicated that the estimated value differences of the dissimilar agent could be decoded from those of the similar agent, and vice versa. (F) In

contrast, the estimated value differences between the agents observed in the preference learning task could not be decoded from VMPFC activity related to value differences in the individual decision task. Thus, the preferences of others are represented in a common neural code in VMPFC independently of their similarity with our own preferences.

Figure 3. Welfare maximization task: design, behavior, and welfare-related activity. (A) Each trial of the welfare maximization task required participants to choose between two different distributions of goods for the two agents (e.g., “1 raisin for agent 1 and 3 peanuts for agent 2” versus “3 peanuts for agent 1 and 1 raisin for agent 2”) whose preferences they had learned in the preference learning task. Instead of drawings, photos of other agents were shown. (B) Participants strongly preferred the option that maximized the utilitarian welfare of the two agents (based on their estimated preferences represented by linear utility functions). Blue dots indicate mean choices for binned levels of utilitarian welfare differences (quartiles); error bars depict standard errors of the mean. (C) A univariate analysis revealed that larger differences in utilitarian welfare between the options were associated with stronger activation in VMPFC. (D) Activity patterns in the VMPFC allowed cross-decoding utilitarian welfare differences in the welfare maximizing task from estimated value differences of the observed agents in the preference learning task (and vice versa) using support vector regression analyses. By contrast, (E) VMPFC activity related to welfare differences could not be decoded from participants’ value differences in the individual decision task. Thus, neural representations of utilitarian welfare and the preferences of others appear to rely on related neural codes.

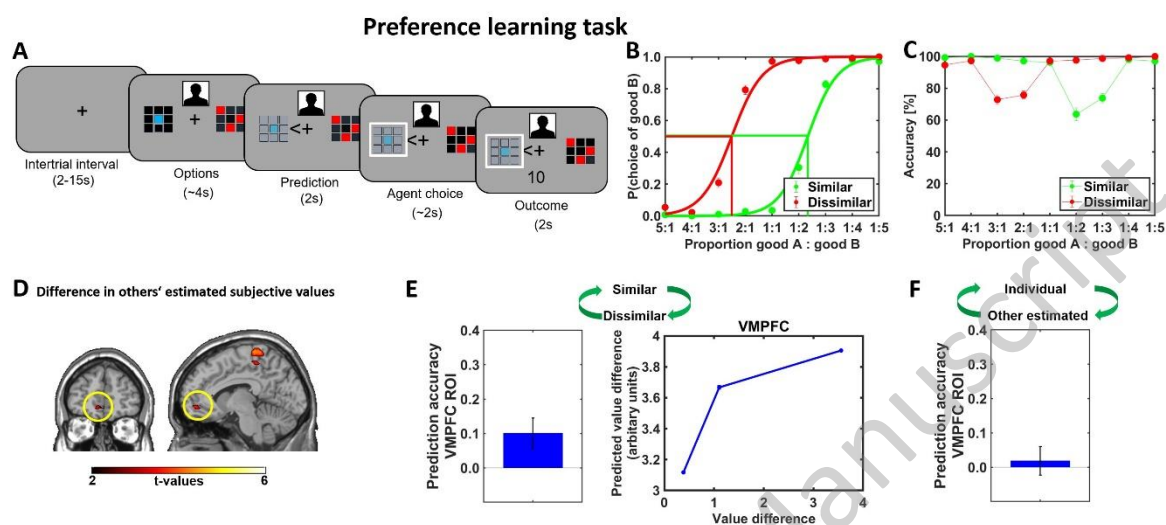
845 Figure 1



846

847

848 Figure 2



851 Figure 3

