

**A network for computing value equilibrium  
in the human medial prefrontal cortex**

Keno Juechems<sup>1\*</sup>, Jan Balaguer<sup>1</sup>, Santiago Herce Castañón<sup>1,2</sup>, María Ruz<sup>3</sup>, Jill X. O'Reilly<sup>1</sup>,  
and Christopher Summerfield<sup>1</sup>

<sup>1</sup>Dept. Experimental Psychology  
University of Oxford  
Walton Street, Oxford OX2 6AE, UK

<sup>2</sup>Dept. Psychology and Educational Sciences  
University of Geneva  
1202 Geneva, Switzerland

<sup>3</sup>Dept. Experimental Psychology  
Mind, Brain and Behavior Research Center  
University of Granada, Granada, Spain

\* Lead Contact and Corresponding author: keno.juechems@psy.ox.ac.uk

## Summary

Humans and other animals make decisions in order to satisfy their goals. However, it remains unknown how neural circuits compute which of multiple possible goals should be pursued (e.g. when balancing hunger and thirst) and how to combine these signals with estimates of available reward alternatives. Here, humans undergoing functional magnetic resonance imaging (fMRI) accumulated two distinct assets over a sequence of trials. Financial outcomes depended on the minimum cumulate of either asset, creating a need to maintain “value equilibrium” by redressing any imbalance among the assets. BOLD signals in the rostral anterior cingulate cortex (rACC) tracked the level of imbalance among goals, whereas the ventromedial prefrontal cortex (vmPFC) signalled the level of redress incurred by a choice, rather than the overall amount received. These results suggest that a network of medial frontal brain regions compute a value signal that maintains value equilibrium among internal goals.

## Introduction

Canonical models in psychology, economics, and machine learning assume that decisions are made in order to maximise expected reward. One popular view posits that the brain evolved dedicated structures that represent the value of stimuli or actions. For example, in non-human primates and rodents, neuronal firing rates in the orbitofrontal cortex scale with the value of primary reinforcers (Padoa-Schioppa and Assad, 2008; Rich and Wallis, 2016; Strait et al., 2014). In humans, BOLD signals in the orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (vmPFC) code for the value of diverse assets including food items (Hare et al., 2011, 2009), money (De Martino et al., 2006; Frydman et al., 2014) and social cues (Lin et al., 2012). This has been interpreted as revealing a domain-general value code or “common neural currency” in this region. Representing diverse prospects on a common value function may allow organisms to estimate the value of multidimensional prospects and to make decisions among otherwise incommensurable goods (Chib et al., 2009; Kable and Glimcher, 2009; Kim et al., 2011).

However, in natural environments, survival depends on the minimum level of any one of a number of competing internal needs. For example, a hungry animal needs food more than water, whereas the reverse is true for a thirsty animal. As the world changes dynamically, distinct internal assets (e.g. satiety or hydration) are continuously being depleted and replenished, so that equilibrium among internal resource levels needs to be monitored and maintained in neural circuits for valuation and choice (Cannon, 1929; Keramati and Gutkin, 2014; Korn and Bach, 2015). A related framework for considering reward-guided choices, thus is that animals strive to satisfy multiple needs in parallel, maintaining a balance among relative asset levels, such as satiety, warmth, or reputation (O’Reilly et al., 2014). This class of model requires that value is coded on multiple distinct axes, each pertaining to a currently relevant goal (rather than being represented on a single monolithic value function) that mapping from

these axes to choices occurs rapidly and flexibly (Valentin et al., 2007), and that axes interact and compete for action selection (Hunt and Hayden, 2017). In the current work we asked how the brain evaluates which of multiple possible goals should be pursued at any one time, and how it dynamically updates goal values according to time-varying resource levels.

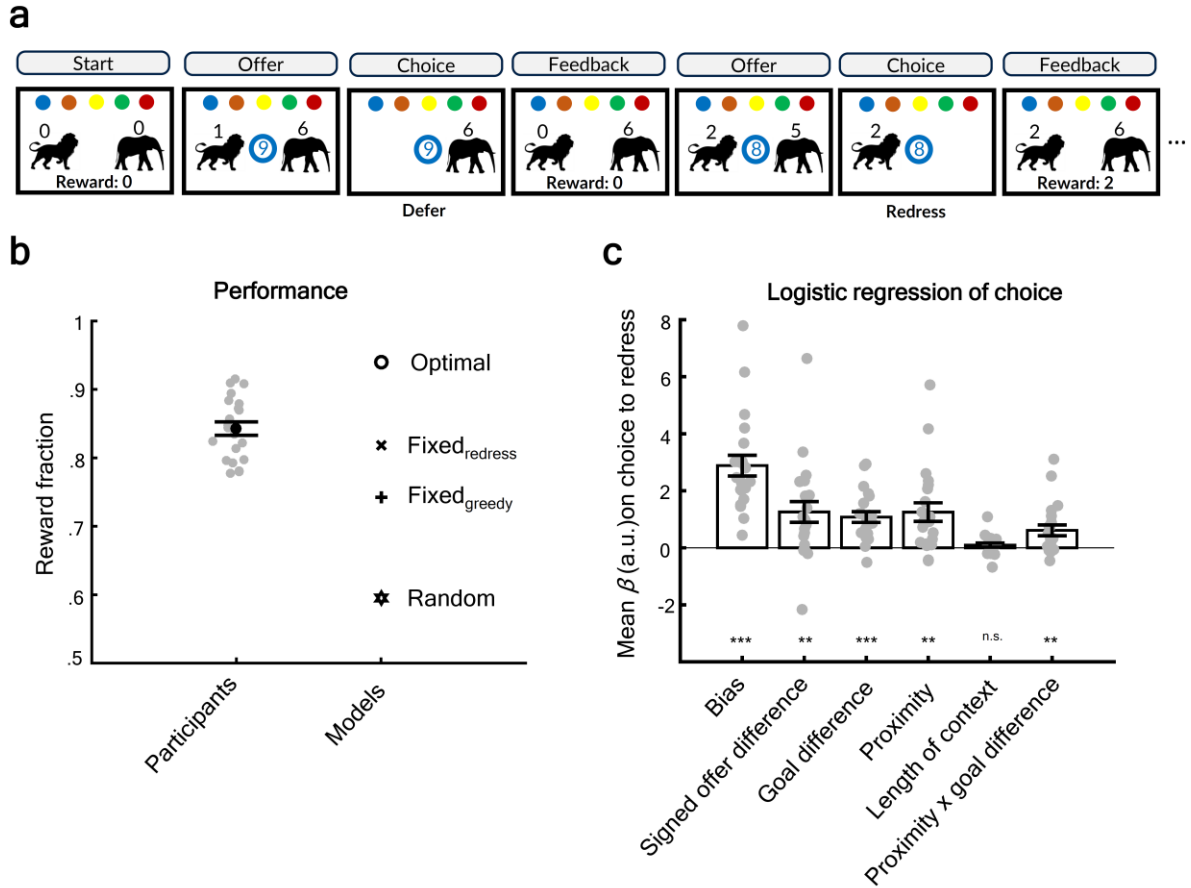
This question has been hitherto unaddressed because most previous studies have used paradigms (e.g. bandit or food choice tasks) that involve maximisation of a single asset, rather than tasks that require the balancing of multiple assets. Nevertheless, there is reason to predict that the OFC/vmPFC may play a role in evaluating stimuli in the context of internal goal states (e.g. satisfy hunger or quench thirst). Firstly, animals with OFC lesions continue to choose actions that lead to devalued outcomes, as if they were failing to maintain or follow currently active goals (Bouret and Richmond, 2010; Burke et al., 2008; Izquierdo et al., 2004). Secondly, humans and primates with OFC lesions fail to integrate multiple attributes of a stimulus when making decisions, implying that this structure plays an active role in constructing composite value estimates from available stimulus dimensions (Fellows and Farah, 2007; Glascher et al., 2012; Izquierdo et al., 2004; Wallis, 2011). Finally, vmPFC tracks the internal state given by cumulative satiety or wealth over a sequence of trials, a critical requirement for computing which assets may be most needed (Juechems et al., 2017; Tsetsos et al., 2014).

We reasoned that the medial OFC/vmPFC and interconnected areas of the medial prefrontal cortex might allow animals to compute ongoing levels of competing assets, and dynamically adjust choices to both replenish immediate needs and guard against future scarcity. We tested this using a task in which participants were encouraged to maximise the minimum of two assets (animals in a virtual zoo). We report evidence for a new frame of reference for human neural value signals, with rACC encoding the imbalance among assets (or goals), and the vmPFC coding how this imbalance is redressed by a choice.

## Results

Participants ( $n = 21$ ) performed a “virtual zookeeper” task in the fMRI scanner. The task involved managing a zoo during contexts of variable length (10-20 trials) that housed two assets (lions and elephants). On each trial, participants were offered the opportunity to expand their zoo by a variable number of animals (e.g. 4 lions or 2 elephants). Subsequently, trialwise reward was offered in proportion to the minimum number of animals in the zoo (e.g. if they had a total of 12 elephants and 8 lions in the zoo, they received 8 points) creating a pressure to redress any imbalance in the assets. To mimic the autocorrelation in natural environments, we created “streaks” in which numerically more of one species (4-6 uniform) were offered than the other (1-3 uniform); these numbers reversed with probability 0.3 on each trial. This manipulation ensured that the optimal policy was neither to always choose the more plentiful animal, nor to always satisfy immediate needs (redress the imbalance), but to vary these strategies according to the quality of the offer, the need to redress, and the time remaining in the block (see **Fig. 1a**).

**Human choices are driven by both offer values and internal state.** Participants received 84.3% of the maximum attainable cumulative reward as calculated by numerical simulation (see Methods; **Fig. 1b**). This was significantly better than chance (mean: 59.6%, Wilcoxon sign-rank test,  $p < 0.001$ ), better than a fixed strategy of always choosing the higher offer (“greedy”; mean: 74.3%, sign-rank test,  $p < 0.001$ ), and better than a model that used a fixed strategy of always satisfying immediate needs, assuming perfect memory for the tallies of animals in the zoo (“redress”; mean: 81.8%, sign-rank test,  $p < 0.03$ ).



**Figure 1.** Task and behavior. (A) Task with 2 example trials from the first block in a scanner run. The number of animals on offer (“offer” screen), the number of animals chosen (“choice” screen) and the number of animals in the zoo (“feedback”) are shown numerically above a picture of a lion and elephant. Coloured circles (top) indicate the zoos (blocks) for the current run, with the central circle coloured according to the current block. The central number indicates the number of trials (including the current trial) remaining before the end of the block, i.e. the “countdown”. Trialwise reward was indicated on the feedback screen as the minimum of the two assets. Background screens were grey during testing, but are shown in white here for illustration. (B) Black circle indicates average earnings as fraction of maximum possible; black error bars are standard error of the mean (SEM) and grey circles show the data from individual participants. Other markers indicate averaged reward fractions for various models. (C) Beta coefficients from a logistic regression of various predictors on choice to redress. “Proximity” was coded as 1 over the number of trials left in a context, “length of context” reflected the total number of trials in a context (10-20). Bar height indicates mean beta, error bars are SEM, grey circles represent individual participant data. Significance was assessed using a two-tailed t-test against zero. \*\*\* is  $p < 0.001$ , \*\* is  $p < 0.01$ , \* is  $p < 0.05$ .

These data suggest that participants were not simply maximising their receipt of animals, or myopically seeking to balance their assets, but rather that choices were jointly driven by offer quality and relative needs. To test this contention more formally, we defined choices as either

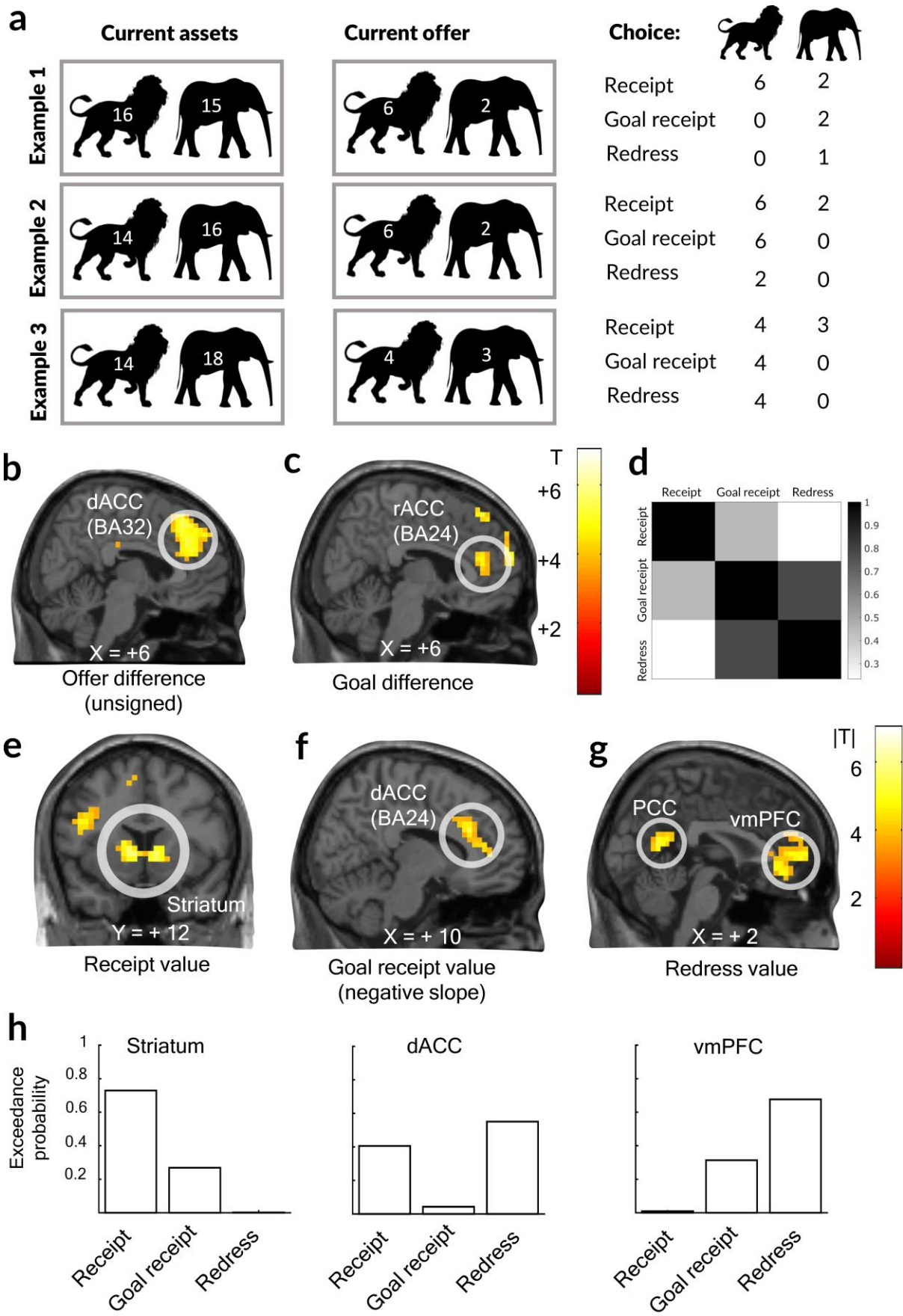
“defer” or “redress”, where the latter trials were those on which participants chose to acquire more animals of the species with the lower current tally (mean = 80.85%, SD = 10.78%). We then used a logistic regression model to identify the factors that predicted the decision to redress or defer. Although participants were strongly biased to choose to redress ( $t_{20} = 7.74$ ,  $p < 1 \times 10^{-6}$ ), this tendency was stronger when the (signed) offer difference was higher, i.e. when more of the currently minimal asset was offered ( $t_{20} = 3.36$ ,  $p < 0.005$ ), and also stronger when there was greater need to redress, i.e. when the level of imbalance among the two assets (or “goal difference”) was higher ( $t_{20} = 5.61$ ,  $p < 1 \times 10^{-4}$ ). In other words, participants were neither greedily choosing the highest offer, nor slavishly satisfying immediate needs, but rather trading off the relative offer values and their internal needs or goals when making decisions. This was also evident in a regression of reaction time (**Fig. S4b**). We also verified that a simple bias to redress was not the optimal policy in our task; indeed, those participants with an overall stronger tendency to redress tended on average to reap lower rewards ( $r_{21} = -0.36$ ,  $p > 0.94$ , one-sided test).

**The dACC encodes offer difference and the rACC encodes goal difference.** Behavioural analyses indicated that decisions to redress depended on both the offer difference (which asset was offered most plentifully?) and the goal difference (how much pressure was there to restore asset imbalance?). Hence, in our first neural analysis (GLM1) we asked whether these quantities are encoded in blood-oxygenation-level dependent (BOLD) signals at the time of choice, i.e. time-locked to offer onset. We observed a dissociation whereby more dorsal regions of the ACC responded to offer difference and more rostral regions of the ACC responded to goal difference. Specifically, a correlate of goal difference (asset imbalance) was observed in rACC (BA24; peak: 10, 40, 14 [x,y,z], cluster-false-discovery-rate-q (FDRq) < 0.01; **Fig. 2c and Table S2**), accompanied by a cluster in the ventral occipital lobe (peak: -34, -48, -18, cluster FDRq < 0.005) and in some surrounding regions of dorsomedial PFC. By contrast, a

correlate of offer difference was observed in dACC (BA32), extending into the overlying dorsomedial prefrontal cortex (dmPFC) (see **Fig. 2b**; peak: -6, 28, 38, global peak: -14, 28, 50, cluster -  $FDRq < 1 \times 10^{-11}$ ). Additional clusters were found in the bilateral insula (left peak: -34, 20, -2; right peak: 46, 20, 2; cluster –  $FDRqs < 0.02$ ), bilateral inferior parietal lobule (iPL), and angular gyri (left peak: -46, -56, 30; right peak: 50, -56, 30, cluster –  $FDRqs < 0.005$ ). We additionally included the signed offer difference in the regression, as in the behavioural analysis above (i.e. most– least needed asset) but did not observe significant clusters that encoded this quantity signed by needs.

One potential concern is that these effects are secondary to time on task. However, neither unsigned offer difference nor goal difference were strongly correlated with reaction time (RT;  $r = 0.046$ ,  $r = 0.006$ , respectively), and we found the same neural effects when controlling for RT in GLM1 (**Fig. S1**), so we think it unlikely that this finding reflects the confounding influence of time-on-task. Rather, these data suggest that distinct regions of the ACC encode the two factors that determine whether a need should be addressed: the relative quality of the rewards on offer (dACC), and the relative balance of internal needs (rACC).





**Figure 2.** Representation of choice-relevant task features and value codes. (A) Illustration of the different value codes described in the paper for three example trials. The “current assets” panel depicts the state of assets before participants made a choice, whereas the “current offer” depicts the available choices. The panel to the right shows the values of the three value codes used for analyses depending on choice: Receipt value reflected the increase in the overall number of animals achieved by choice. Goal receipt value codes the receipt value in the frame of reference of the goal. Finally, redress value encodes the realized redress between the goals or, equivalently, the increase in trial reward. It is computed as  $\min(\text{receipt value}, \text{goal difference})$  if participants chose according to their needs, or zero otherwise. (B) Positive correlation between offer difference (high minus low) and BOLD signals in the medial PFC, sagittal slice at  $x = 6$ . (C) Correlation matrix for the three value codes. (D) Positive correlation of goal difference with voxels on the same slice as in B). (E) Encoding of receipt value in voxels rendered on a coronal slice at  $y = 12$ . (F) Encoding of goal receipt value on a sagittal slice at  $x = 10$  (G) Encoding of redress value on a sagittal slice at  $x = 2$ . Voxels were thresholded at  $p < 0.001$ , uncorrected. (H) Results of Bayesian model selection. Bar height corresponds to exceedance probability when testing which of the three value models fit best in a random effects analysis across subjects within each ROI (striatum, dACC, vmPFC). No error bars can be computed for exceedance probabilities.

**vmPFC encodes a value redress signal.** Previous research has reliably demonstrated that vmPFC encodes decision outcome but left open the question of how current needs may impact reward encoding. Our task allowed us to dissociate three possible frames of reference for the outcome signal that might be computed on each trial: the total number of animals received independent of need (receipt value), the number of animals received in the frame of reference of need (goal receipt value), and the level of redress among asset imbalance that was incurred by a choice (redress value). To illustrate, consider a participant with 10 lions and 12 elephants in their zoo who receives 3 more lions. The receipt value (and goal receipt value) would be 3, whereas the redress value would be 2. However, if the participant chose 2 more elephants, the receipt value would be 2, and the goal receipt value and redress value would be zero (see **Fig. 2a** for an illustration). By design, receipt value and redress value were only weakly correlated ( $r = 0.23$ ,  $R^2 < 0.06$ ; correlation matrix **Fig. 2d**) and thus dissociable; however, goal receipt value and redress value were inevitably more correlated, because they were identical whenever the choice did not fully restore the asset imbalance ( $r = 0.78$ ,  $R^2 < 0.61$ ).

We first used a standard fMRI analysis approach, allowing these three value regressors compete for variance in the BOLD signal at the time of choice (GLM2). Receipt value positively activated clusters in the bilateral dorsal striatum (caudate nucleus), extending into ventral striatum (peak: -6, 12, 2; cluster – FDRq <  $1 \times 10^{-6}$ ; **Fig. 2e**). Goal receipt value was negatively correlated with an area in the dACC (BA 24; peak: 14, 28, 26; cluster – FDRq < 0.05 **Fig. 2f**). This cluster lay between the dACC cluster that responded to offer difference and the rACC cluster that encoded goal difference, as if the ACC variously encoded the offer, the need, and the receipt in the frame of needs –three quantities required to compute the level of balance or imbalance among assets. Finally, redress value activated a cluster in the vmPFC (peak: 6, 52, 2, cluster – FDRq < 0.001; **Fig. 2g**) and the posterior cingulate cortex (peak: -6, -60, 14, cluster-FDRq < 0.005), two regions that are often found to coactivate in value-based decision tasks (Bartra et al., 2013) (see **Tables S3-5** for full results). This association between redress value and vmPFC signals was also observed when we coded redress value relative to unchosen redress, in line with the previous observation that the vmPFC codes for the value of a chosen relative to an unchosen alternative (Boorman et al., 2009; Rushworth et al., 2012).

We were initially concerned about the reliability of this effect given the non-negligible correlation between receipt value and redress value. However, when we adopted the conservative approach of first orthogonalising this relative redress value with respect to the other two predictors, ensuring that any effect of redress value must be over and above the other two value codes, we observed the same pattern of activations in dorsal striatum, dACC and vmPFC (**Fig. S2, Tables S6-8**). Thus, whilst mindful of the challenge of dissociating correlated regressors in BOLD analysis, we think the most probable explanation is that vmPFC encodes the redress among asset imbalance in our task.

One reasonable alternative hypothesis is that the vmPFC is simply encoding the monetary receipt on each trial. We note that redress value does not equal the overall monetary reward

received on a given trial, but rather encodes the expected *increase* in reward conditional on the choice. Nevertheless, we used a non-circular ROI approach to test whether the vmPFC simply coded for the monetary reward observed on each trial by testing whether it encoded the level of the lower (trial reward-determining) asset at the time of choice, but found that this was not the case ( $t_{20} = 0.02$ ,  $p > 0.98$ ; GLM3). This leads us to conclude that vmPFC encodes the level of redress incurred by a choice, and not simply hedonic reward incurred on any given trial.

One interesting feature of these data is that they seem to suggest that vmPFC, dACC and caudate code for the value signal in distinct frames of reference: the caudate codes the level of receipt independent of needs; the dACC codes the level of receipt in the frame of reference of needs; and the vmPFC signals the degree to which a choice restores value equilibrium. To test this claim, we used a more stringent model fitting approach in which each predictor (parametric value signal) was entered alone into the GLM to estimate which provided the best fit to BOLD responses in each ROI. We compared models using a Bayesian model selection method that relaxes the assumption that all participants are explained by a single model, allowing inference at the random effects level (Stephan et al., 2009). We compared models based on the exceedance probability, which estimates the likelihood that one model outperformed all other models in the set. It is thus a continuous measure indicating the confidence one may place in a model being better than the other ones that were tested.

For each model we computed the log evidence (as approximated by the model's log likelihood fit) in each ROI (caudate, dACC, vmPFC) selected from the contrasts previously described for GLM2 using a non-circular, leave-one-subject-out approach. The pattern we observed corroborates our previous finding that the caudate, dACC, and vmPFC exhibit different responses to these value codes, but added further nuance: Caudate was best fit by model 1 (containing receipt value; exceedance probability ( $ep$ )  $> 0.73$ ), whereas vmPFC was best fit by model 3 (redress value;  $ep > 0.67$ ). Within the dACC, however, no model showed a clear

advantage over the others, although model 3 was numerically best ( $ep > 0.55$ ; **Fig. 2h**). This analysis thus confirmed our contention from GLM2 that striatum and vmPFC were best described by two different value codes. The pattern of results in dACC, however, was less clear, indicating that it may encode a wide range of different values relevant for a decision (Kolling et al., 2016).

Together, thus, we find that the dorsal striatum and vmPFC fulfilled dissociable roles in value encoding in our task. The dorsal striatum encoded the number of animals received irrespective of needs (receipt value). In contrast, the vmPFC encoded the redress in imbalance incurred by a given choice, i.e. the extent to which internal needs were returned to a balanced state by a given choice. The ACC, by contrast, seems to form a hub that brings together information about the offer, the needs and the change in asset levels incurred by a choice.

**Computational model and optimal policy.** The analyses above leave open the question of how an agent should behave in order to maximise reward in our task. One way to approach this question is to use dynamic programming (DP) to compute the policy that will maximise expected future return over each zoo. DP searches through possible future states and chooses a reward-maximising action according to the expected future return, assuming optimal subsequent choices. Because it accounts for possible future offers up to the horizon defined by the end of the zoo, the DP algorithm will sometimes choose the immediately less needed option in order to maximize overall return. For instance, where an offer of the asset not immediately required is generous, it may be sensible to nevertheless choose that asset, to guard against future scarcity when the offer probabilities reverse. This policy will be less useful at the end of the block, because such future benefits are curtailed when the zoo ends. Our use of a DP model does not imply a commitment to dynamic programming as a plausible algorithmic account of the computations that humans performed during the task. Rather, we adopted this approach to explore the normative policy for maximising reward.

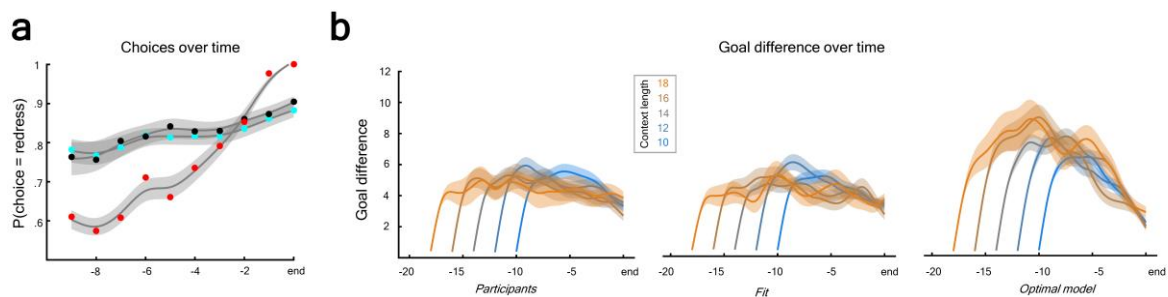
Armed with this computational tool, we first computed the upper bound on reward shown in **Fig. 1b**, as reported above. Next, we asked whether humans, like the model, considered the potential time horizon when choosing to defer an immediate redress in asset imbalance. We found that they did: as for the DP model, proximity to the end of a zoo (defined as  $1/\text{countdown}$ ) positively predicted redress responses ( $t_{20} = 3.74$ ,  $p < 0.005$ ) and that this tendency was stronger when the goal difference was greater ( $t_{20} = 3.13$ ,  $p < 0.01$ ). However, this was not driven by the overall length of the zoo alone, which failed to predict redress probability ( $t_{20} = 1.26$ ,  $p > 0.22$ ). Next, we evaluated choices made by the DP model using the same logistic regression approach. Despite some similarities in the betas obtained between humans and the DP model, the latter exhibited a stronger impact on  $p(\text{redress})$  of both offer value and proximity to the end of the zoo (**Fig. S3a**), but predicted the general trend in participants' choices (**Fig. S3b**).

These findings might be explained if humans adopt a strategy that is approximately optimal but myopic i.e. has insufficient search depth when computing expected future return. We thus reimplemented the DP model, but allowed the planning horizon (in steps) to vary as a free parameter, as well as the subjective reversal probability, and policy terms that allow for bias and choice variability. When we fit the output of this “myopic” DP models to participants choices using maximum likelihood estimation with five-fold cross-validation (one fold per scanner run), we found that, on average, participants estimated the reversal probability to be approximately 0.44 (SD = 0.33) and planned only 7.5 trials ahead (SD = 6.04), significantly lower than the theoretical value of 20. The overall model log-likelihood was -2047.2, which was lower than both the logistic regression model with proximity (7 parameters; LL = -2111.1) and without proximity (4 parameters; LL = -2189.2), and the DP model was strongly favoured by Bayesian model selection (exceedance probability = 0.99).

This best-fitting (but myopic) DP model was also able to reproduce several qualitative features of the data. First, it captured how the average probability of redress varied with proximity to

the end of the block (**Fig. 3a**). Second, it captured how the average need to redress (i.e. tolerance for disparity among the two assets in the zoo) varied as a function of time for blocks of different length, **Fig. 3b**. By contrast, a fully optimal DP model is more tolerant of larger goal difference values in the middle portion of the block, indicative of a longer time horizon for planning.

In conclusion, the DP analyses revealed that although humans planned for future needs, they were suboptimal in two ways: firstly, they did not plan ahead sufficiently, and secondly, they over-estimated the reversal probability.



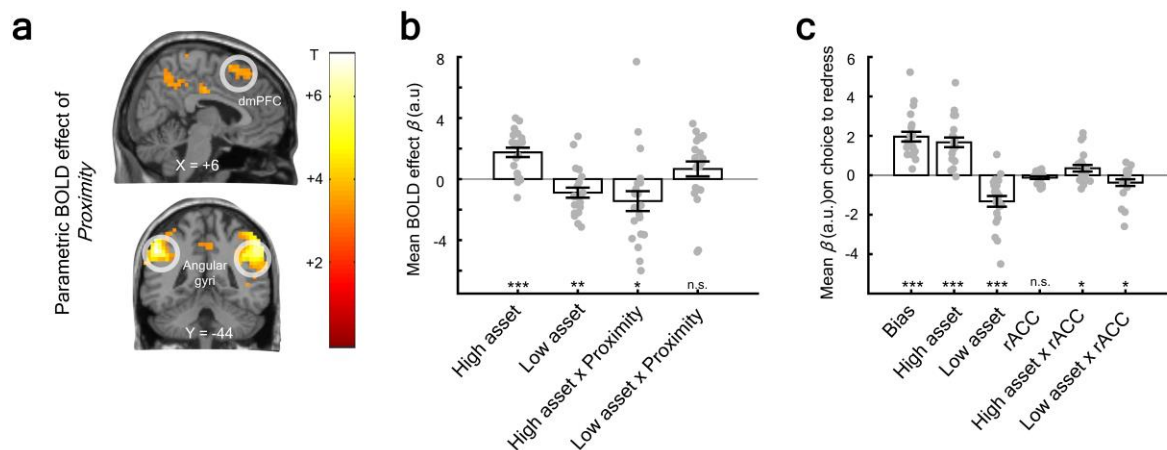
**Figure 3.** Comparison of optimal policy model and participants' behaviour. (A) Plot of redress choices against countdown toward the end of a block for optimal DP model (red dots), human participants (black dots), and best fitting "myopic" DP model (turquoise dots). Circles are mean and shaded areas are SEM around the spline-smoothed means (for display purposes only). (B) Plot of goal difference against time for blocks (zoos) of different length. The three subpanels show data averaged across participants (left) the simulated fit of the myopic model (middle), and the optimal model (right panel). Solid lines are mean across participants/models, shaded areas are SEM. Curves are spline-smoothed for clarity of viewing. Blocks with 20 trials were excluded from this analysis, as these represented less than 1% of all blocks.

**Neural coding of time-varying pressure to redress.** One algorithmic account that would be consistent with both this description of human performance and the neural data above is that the brain keeps track of the goal difference (pressure to redress, as a function of the two asset tallies) and codes pressure to redress in a way that varies with proximity to the end of the block.

To test this view, we returned our attention to the neuroimaging data, and asked whether BOLD signals covaried with the interactions of the two asset tallies and proximity, i.e. whether cortical regions reflected a time-varying pressure to redress. In GLM3 we found that the rACC (BA 24) region covarying with goal difference (i.e. the quantity that participants need to keep in balance) also encoded the interaction of the higher asset and proximity (negatively;  $t_{20} = -2.16$ ,  $p < 0.05$ ), but not of the lower asset and proximity ( $t_{20} = 1.33$ ,  $p > 0.19$ ). Thus, encoding of the higher asset (and by extension goal difference) decreased over time in this region (**Fig. 4b**). We also observed that a more posterior region in the dmPFC (GLM1; peak: -36, 16, 50, cluster – FDR<sub>q</sub> < 0.001), as well as bilateral angular gyri (left peak: -50, -48, 34; right peak: 54, -44, 38; cluster – FDR<sub>q</sub>s < 0.001), and dorsolateral PFC (left peak: -34, 20, 46; right peak: 30, 4, 62; cluster – FDR<sub>q</sub>s < 0.001, **Table S9**) encoded the main effect of proximity to the block end, independent of asset difference, as if they were tracking the available time remaining to redress (**Fig. 4a**).

Finally, we asked whether single-trial activity (from GLM4) in the same rACC ROI (BA 24) had an impact on behaviour. We constructed a logistic regression model in which choices to redress were predicted by the two goal tallies, the rACC activity, and the interactions of rACC with the goal tallies. Choices to redress were negatively predicted by the lower tally ( $t_{20} = -4.75$ ,  $p < 0.0002$ ), and this effect was accentuated when rACC signals were stronger ( $t_{20} = -2.11$ ,  $p < 0.05$ ), whereas choices to redress were predicted positively by the higher tally ( $t_{20} = 6.69$ ,  $p < 1 \times 10^{-5}$ ), and this effect was also stronger when rACC was more active ( $t_{20} = 2.09$ ,  $p < 0.05$ ; **Fig. 4c**). No main effect of rACC activity was observed ( $t_{20} = -1.72$ ,  $p > 0.10$ ). Thus, high rACC activity guided choices based on the current need to redress.





**Figure 4.** Neural effects of trial sequence. (A) Main effect of proximity on voxels rendered on a sagittal slice at  $x = 6$  and a coronal slice at  $y = -44$  from GLM1. Voxels were thresholded at  $p < 0.001$ , uncorrected. (B) Higher and lower asset encoding in leave-one-out ROI in rACC (BA 24; ROI defined from goal difference contrast in **Fig. 2**). Bar height indicates mean neural beta, error bars are SEM, grey circles represent individual betas. (C) rACC activity predicted choice to redress on a trial-by-trial basis. Bars are mean parameter estimates from logistic regression, error bars are SEM. Significance was assessed using a two-tailed t-test against zero. \*\*\* is  $p < 0.001$ , \*\* is  $p < 0.01$ , \* is  $p < 0.05$ .

## Discussion

We asked humans to perform a value-guided decision task requiring them to balance levels of two possible assets (“value equilibrium”). We found that they did so strategically, basing their choices about whether or not to redress or defer on the value of current offers, knowledge of value imbalance, and the time horizon available for future redress. In other words, humans make choices that actively arbitrate among current goals, and plan for possible future scarcity. We argue that considering this alternative frame of reference for the neural coding of value sheds new light on the function of key brain areas involved in reward-guided decision-making. Previous studies have identified human BOLD responses that code for the economic value of offers and outcomes in various frames of reference, observing for example that the vmPFC codes positively for the value of a chosen option and negatively for an unchosen option, with the reverse pattern observed in the dACC (Bartra et al., 2013; Boorman et al., 2009). Other,

less well-established frames of reference for value signals have been reported, for example that value is coded relative to an average or default stimulus in either the vmPFC (Cox and Kable, 2014; Lopez-Persem et al., 2016) or dACC (Boorman et al., 2013), or that these regions respectively signal the value of a proximal choice, and the need to switch to a new context of known average value (Hayden et al., 2011; Kolling et al., 2014, 2012). However, the foundational premise that has underpinned these (and related) past studies is that an outcome can be indexed on a single value function that denotes the change in overall reward or wellbeing that is incurred by its receipt.

Our approach is rather different. We begin with the assumption that agents strive to maintain equilibrium among multiple different classes of asset, because depletion of any one asset can be detrimental to survival even if levels of other assets are high (O'Reilly et al., 2014). Thus, a sated and hydrated animal can still die of cold, or a human who has accumulated strong social capital can still fall into ill health through personal neglect. This intuition suggests a novel frame of reference for the neural coding of value, in which the brain keeps track of the relative disequilibrium among goals (which indicates how pressing it is to maximise the minimum asset) and the degree to which any one choice redresses this imbalance (which is a proxy for how that choice promotes survival, at least over the short term). In support of this view, we find that BOLD signals in distinct brain regions vary with these quantities. For example, the rACC encodes the need to redress imbalance among assets, and the vmPFC signals the extent to which choices restore value equilibrium. In other words, in our task, values are coded with reference to internal needs: the relative levels of depletion or replenishment of a multiplicity of different assets.

Turning to the detail of our neural findings, the current work offers new insights into the function of three regions previously implicated in value-guided choice: the striatum, vmPFC and ACC. Previous studies have variously implicated these regions in computing the value of

stimuli, actions and goals (Dolan and Dayan, 2013; Doll et al., 2012), with vmPFC and ACC typically coding inversely for value in an at least partially redundant fashion and the vmPFC and striatum seemingly coding for the value of proximal rewards and goals in a perplexingly overlapping fashion (Bartra et al., 2013). By contrast, our work suggests some unusually clear dissociations among the quantities coded by these regions. Firstly, the results we report dissociate value encoding in the striatum and vmPFC, with the former responding to outcomes independent of goals and the latter responding to the level of redress among goals incurred by a choice. Secondly, we observed a dissociation among ACC signals, with more dorsal regions signalling the quality of an offer and more rostral regions indicating the need to restore imbalance among goals, with the latter signal modulated by the available time remaining to do so.

**Dissociable value codes in vmPFC and striatum.** On each choice, the striatum responded to the magnitude of the chosen offer (i.e. the number of animals received), but unlike cortical regions, it did so in a fashion that was insensitive to internal needs – the signal went unmodulated by estimated cumulative assets. Strikingly, these outcomes do not readily relate to reward received on a given trial, but simply indicate a change in one’s overall internal state. Previous studies have suggested that where a single asset (e.g. money) is to be maximised, both vmPFC and striatum seem to code the long-run average value of an action, and the value computed via tree search through future states or outcomes (Bartra et al., 2013; Doll et al., 2015, 2012). These common value signals in cortical and subcortical regions may reflect the fact that animals use a mixture of model-free and model-based information when making decisions (Daw et al., 2011; Dolan and Dayan, 2013; Doll et al., 2012). However, our work instead suggests that cortical systems compute values in the frame of reference of ongoing internal needs, whereas the striatum codes for overall receipt independent of those needs. It should be noted, however, that the peak of the striatal cluster identified here was slightly more

dorsal than commonly identified for hedonic value (Bartra et al., 2013), although it also included voxels in the ventral striatum.

Importantly, however, our task allowed us for the first time to dissociate outcomes *per se* from their impact on current goals. We found that unlike in previous tasks, vmPFC did not simply mirror the code employed by the striatum, but rather coded for the redress in goal difference – that is, the extent to which any given choice restores the imbalance among internal needs. This signal remained the best explanation of vmPFC signals even when we partialled out the actual outcome received in number of animals, or the trialwise monetary reward that motivated participants to perform the task. We note that in our task, as in a natural environment where wellbeing is determined by the minimum of multiple possible assets, redress value is inevitably identical with the change in reward or hedonic value. This presumably explains the ubiquitous observation that vmPFC correlates positively with reward outcome when a single asset is being optimised, and why this signal is often observed to be modulated by the local context provided by average value (Cox and Kable, 2014; Padoa-Schioppa, 2009; Yamada et al., 2018). Further research is needed to more fully distinguish our account of the vmPFC from other models that emphasise encoding of trial-wise change in reward (measured on a single or multiple attributes). However, our task may provide new insights into *how* the vmPFC may compute hedonic value. Choice outcomes (i.e. the number of animals received) had to be related to the goal difference to compute value. Thus, our findings differ on two accounts: First, vmPFC did not simply reflect the value of a trial, but rather the *between-trial* increase in value. Second, it could only do so by utilizing the goal difference encoded in the rACC. We note, however, that our design did not allow us to differentiate vmPFC activity immediately preceding a choice from activity post-choice as participants were allowed to respond at any time after stimulus onset.

These findings build on our previous work implicating the human vmPFC in coding an internal, unsignalled representation of cumulative assets in a way that maximises a specific goal (maintaining net positive aggregate reward), over and above any momentary hedonic value that arises from choices (Juechems et al., 2017). However, the current work additionally implies *why* the brain keeps track of cumulative assets – because this allows any imbalance among internal needs to be redressed by future choices. This helps understand a perplexing paradox in the decision literature – why is it that neural signals in the vmPFC code ubiquitously for rewards, but lesions to the vmPFC incur only subtle deficits where decisions involve maximisation of a single asset (Noonan et al., 2010)? The vmPFC outcome signals ubiquitously observed in fMRI studies may reflect the tracking of redress value to one of multiple potential goals – obtaining food, or money, or maximising accuracy – rather than computing the relative strength of one offer over another. Indeed, in studies of navigation the vmPFC signal also ramps up over multiple steps that are made towards a destination, presumably because each step redresses the distance between current and goal state (Balaguer et al., 2016; Howard et al., 2014). Thus, we would expect patients with vmPFC lesions to arbitrate effectively between offers of differing value but to fail to allocate decisions strategically over the long term, precisely the pattern that is observed in both humans and monkeys (Glascher et al., 2012; Noonan et al., 2010). Furthermore, patients with lesions in the medial PFC tend to search for information within individual options, failing to integrate over options and attributes (Fellows, 2006).

**ACC employs multiple frames of reference to arbitrate between goals.** We also observed that two distinct areas of the ACC encoded the current offer (dACC) and internal state (rACC). Firstly, a cluster in the rACC (BA24) responded to the current imbalance among assets, i.e. the extent to which one asset needed to be replenished in order to match another. This cluster did not simply reflect how this imbalance evolved up to the current trial, but exerted influence on

participants' choices indicating that participants actively sought to maintain equilibrium between assets. Previous studies of reward-guided decisions have emphasised that the dACC codes positively for the relative value of an "unchosen" option— i.e. that which was foregone when a choice was made (Boorman et al., 2009; Daw et al., 2006) – or that it signals the need to switch away from a current context towards a new, potentially richer, source of rewards (Kolling et al., 2016, 2012). Secondly, a cluster in the dACC reflected a generic disequilibrium in the offers irrespective of needs. Interestingly, a cluster that lay in the intermediate region between the dACC and rACC seemingly employed multiple frames of reference for value encoding. It has been recognised previously that the ACC employs multiple frames of reference and encodes variables on several timescales (Kolling et al., 2018, 2016; Meder et al., 2017). Our framework provides an intriguing explanation about *why* this may be the case: The ACC integrates offers with goals to inform the best course of action in a given context. In order to do so, it must reflect i) the overall magnitude of offers, ii) how a choice will change the cumulative assets, and iii) how several assets relate to one another. Its role in updating all relevant cumulative assets explains why dACC is often found to encode an unchosen (i.e. a counterfactual) value (Boorman et al., 2013), while its role in relating assets to one another fits with the finding that dACC encodes the value of switching away from a currently active context (Kolling et al., 2012). Together, our findings argue for a critical role for the ACC in computing which of multiple possible goals should be pursued at any one time.

The choice of whether to satisfy an immediate need, or to build up less pressing resources to offset future scarcity, is a ubiquitous problem for humans and other animals (Hurly, 1992; Mullainathan and Shafir, 2014). Critically, this choice depends on the time horizon over which choices can be made: an optimal agent with knowledge of the time horizon will tolerate an asset imbalance early in the block, where it can be later redressed if opportunities change, but respond to pressure to redress later in the block. Human participants were less tolerant of

imbalance than the optimal model, which can be explained if they are somewhat myopic in their planning (Frydman et al., 2014; Kolling et al., 2018). Note, however, that this is akin to, and indistinguishable from a parameter discounting future rewards in favour of short-term gain. Nevertheless, reversal probability, planning horizon and an alternative discount parameter would all predict that participants chose to redress more than was optimal, especially at the early stages of the block. Consistent with this view, the rACC region that coded for goal difference did so in a fashion that varied with proximity to the end of the block. However, the direction of this effect was somewhat surprising to us, in that goal difference encoding decreased with time. It is possible that participants had a bias to redress towards the end irrespective of the size of the goal difference – thus, goal difference encoding may not be needed to drive choice towards the final few trials in a block. Nevertheless, this finding requires further corroboration in future studies.

We acknowledge one important limitation of our approach: working in the restricted environment of the laboratory, where varying the true internal needs of the participant is technically challenging if not impossible, we instead adopted a task in which participant payment depended overtly on asset balancing. Thus, whilst we would like to argue that animals intrinsically seek to maintain value equilibrium in the natural world, in our task they were required to do so in order to maximise a single asset (money). Given that we were unable to vary the need for primary reinforcement, our task is thus perhaps a closer simulacrum of the uniquely human task of maintaining equilibrium among more complex needs, such as when a busy academic balances the goals of conducting impactful research and providing effective teaching. We also acknowledge that an alternative computational framework exists for understanding how animals might solve a task such as ours, namely that the brain searches forward through a tree of possible future states in order to calculate which course of action will reap the greatest long-term reward – indeed, this is the normative framework that we used as a

benchmark for comparing human performance. Although our findings do not explicitly rule out this possibility, it is not clear to us why, if humans were not tracking equilibrium among goals, the brain would code for goal disequilibrium (rACC) and redress (vmPFC).

Our findings may more generally have implications for understanding health and wellbeing in humans. We assume that wellbeing is related to the minimum among current needs, and find evidence that neural circuits for valuation and choice have evolved to maintain a balance among needs via goal-directed choice. Disorders of valuation and choice, such as depression, might be associated with failures of a system that attempts to maintain value equilibrium such as an exaggerated representation of goal difference, or a failure to update internal asset estimates when they are redressed.

## **Acknowledgments**

This work was supported by European Research Council to CS; and the Economic and Social Research Council studentship to KJ (ES/J500112/1); and a Wellcome Trust 4-year-PhD grant to SHC (0099741/Z/12/Z). We would like to thank Nathaniel Daw and his lab for useful suggestions, and all members of the Summerfield lab, especially Andreas Jarvstad, for helpful discussions from initial idea to publication, and Hannah Tickle, Hildward Vandormael, and Vickie Li for assistance collecting the data.

## **Author contributions**

KJ, CS, and JXO designed the study and interpreted the results, KJ, JB, SHC, MR, and CS collected the data, JB provided code for fMRI analyses, KJ analysed the data and provided code, KJ and CS wrote the paper.



534    **Declaration of Interests**

535    The authors declare no competing interests.

536

537

## References

- Balaguer, J., Spiers, H., Hassabis, D., Summerfield, C., 2016. Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. *Neuron* 90, 893–903. doi:10.1016/j.neuron.2016.03.037
- 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. doi:10.1016/j.neuroimage.2013.02.063
- Boorman, E.D., Behrens, T.E., Woolrich, M.W., Rushworth, M.F., 2009. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* 62, 733–743. doi:S0896-6273(09)00389-4 [pii]10.1016/j.neuron.2009.05.014
- Boorman, E.D., Rushworth, M.F., Behrens, T.E., 2013. Ventromedial Prefrontal and Anterior Cingulate Cortex Adopt Choice and Default Reference Frames during Sequential Multi-Alternative Choice. *J. Neurosci.* 33, 2242–2253. doi:10.1523/JNEUROSCI.3022-12.2013
- Bouret, S., Richmond, B.J., 2010. Ventromedial and Orbital Prefrontal Neurons Differentially Encode Internally and Externally Driven Motivational Values in Monkeys 30, 8591–8601. doi:10.1523/JNEUROSCI.0049-10.2010
- Burke, K.A., Franz, T.M., Miller, D.N., Schoenbaum, G., 2008. The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. *Nature* 454, 340–344. doi:10.1038/nature06993
- Cannon, W.B., 1929. Organization for Physiological Homeostasis. *Physiol. Rev.* 9, 399–431. doi:10.1152/physrev.1929.9.3.399
- Chib, V.S., Rangel, A., Shimojo, S., O’Doherty, J.P., 2009. Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *J. Neurosci.* 29, 12315–12320. doi:10.1523/JNEUROSCI.2575-09.2009
- Chumbley, J.R., Friston, K.J., 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage* 44, 62–70. doi:10.1016/j.neuroimage.2008.05.021
- Cox, X.K.M., Kable, J.W., 2014. BOLD Subjective Value Signals Exhibit Robust Range Adaptation. *J. Neurosci.* 34, 16533–16543. doi:10.1523/JNEUROSCI.3927-14.2014
- Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans’ choices and striatal prediction errors. *Neuron* 69, 1204–1215. doi:10.1016/j.neuron.2011.02.027
- Daw, N.D., O’Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879. doi:10.1038/nature04766
- De Martino, B., Kumaran, D., Seymour, B., Dolan, R.J., 2006. Frames, biases, and rational decision-making in the human brain. *Science* (80-. ). 313, 684–687. doi:10.1126/science.1128356
- Dolan, R.J., Dayan, P., 2013. Goals and Habits in the Brain. *Neuron* 80, 312–325. doi:10.1016/j.neuron.2013.09.007
- Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., Daw, N.D., 2015. Model-based choices involve prospective neural activity. *Nat. Neurosci.* 18, 767–772. doi:10.1038/nn.3981
- Doll, B.B., Simon, D.A., Daw, N.D., 2012. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* 22, 1075–1081. doi:10.1016/j.conb.2012.08.003
- Eklund, A., Knutsson, H., Nichols, T.E., 2018. Cluster Failure Revisited: Impact of First Level Design and Data Quality on Cluster False Positive Rates. *bioRxiv*. doi:10.1101/296798
- Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci.* 113, 7900–7905. doi:10.1073/pnas.1602413113
- Fellows, L.K., 2006. Deciding how to decide : ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. *Brain* 129, 944–952. doi:10.1093/brain/awl017
- Fellows, L.K., Farah, M.J., 2007. The role of ventromedial prefrontal cortex in decision making: Judgment under uncertainty or judgment per se? *Cereb. Cortex* 17, 2669–2674. doi:10.1093/cercor/bhl176
- Frydman, C., Barberis, N., Camerer, C., Bossaerts, P., Rangel, A., 2014. Using neural data to test a theory of investor behavior: An application to realization utility. *J. Finance* 69, 907–946. doi:10.1111/jofi.12126
- Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15, 870–878. doi:10.1006/nimg.2001.1037
- Glascher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., Paul, L.K., Tranel, D., 2012. Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proc. Natl. Acad. Sci.* 109, 14681–14686. doi:10.1073/pnas.1206608109
- Hare, T.A., Camerer, C.F., Rangel, A., 2009. Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science* (80-. ). 324, 646–648.
- Hare, T.A., Malmaud, J., Rangel, A., 2011. Focusing Attention on the Health Aspects of Foods Changes Value Signals in vmPFC and Improves Dietary Choice. *J. Neurosci.* 31, 11077–11087.

doi:10.1523/JNEUROSCI.6383-10.2011

Hayden, B.Y., Pearson, J.M., Platt, M.L., 2011. Neuronal basis of sequential foraging decisions in a patchy environment. *Nat. Neurosci.* 14, 933–939. doi:10.1038/nn.2856

Howard, L.R., Javadi, A.H., Yu, Y., Mill, R.D., Morrison, L.C., Knight, R., Loftus, M.M., Staskute, L., Spiers, H.J., 2014. The Hippocampus and Entorhinal Cortex Encode the Path and Euclidean Distances to Goals during Navigation. *Curr. Biol.* 24, 1331–1340. doi:10.1016/j.cub.2014.05.001

Hunt, L.T., Hayden, B.Y., 2017. A distributed, hierarchical and recurrent framework for reward-based choice. *Nat. Rev. Neurosci.* 18, 172–182. doi:10.1038/nrn.2017.7

Hurly, T.A., 1992. Energetic reserves of marsh tits (*Parus palustris*) : food and fat storage in response to variable food supply. *Behav. Ecol.* 3, 181–188.

Izquierdo, A., Suda, R.K., Murray, E.A., 2004. Bilateral Orbital Prefrontal Cortex Lesions in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward Contingency. *J. Neurosci.* 24, 7540–7548. doi:10.1523/JNEUROSCI.1921-04.2004

Juechems, K., Balaguer, J., Ruz, M., Summerfield, C., 2017. Ventromedial Prefrontal Cortex Encodes a Latent Estimate of Cumulative Reward. *Neuron* 93, 705–714.e4. doi:10.1016/j.neuron.2016.12.038

Kable, J.W., Glimcher, P.W., 2009. The Neurobiology of Decision: Consensus and Controversy. *Neuron* 63, 733–745. doi:10.1016/j.neuron.2009.09.003

Keramati, M., Gutkin, B., 2014. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *Elife* 3, 1–26. doi:10.7554/eLife.04811

Kim, H., Shimojo, S., O'Doherty, J.P., 2011. Overlapping Responses for the Expectation of Juice and Money Rewards in Human Ventromedial Prefrontal Cortex. *Cereb. Cortex* 21, 769–776. doi:10.1093/cercor/bhq145

Kolling, N., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., 2012. Neural Mechanisms of Foraging. *Science* (80-. ). 336, 95–98. doi:10.1126/science.1216930

Kolling, N., Scholl, J., Chekroud, A., Trier, H.A., Rushworth Correspondence, M.F.S., 2018. Prospection, Perseverance, and Insight in Sequential Behavior. *Neuron* 99, 1069–1082.e7. doi:10.1016/j.neuron.2018.08.018

Kolling, N., Wittmann, M., Rushworth, M.F.S., 2014. Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron* 81, 1190–1202. doi:10.1016/j.neuron.2014.01.033

Kolling, N., Wittmann, M.K., Behrens, T.E.J., Boorman, E.D., Mars, R.B., Rushworth, M.F.S., 2016. Value, search, persistence and model updating in anterior cingulate cortex. *Nat. Neurosci.* 19, 1280–1285. doi:10.1038/nn.4382

Korn, C.W., Bach, D.R., 2015. Maintaining Homeostasis by Decision-Making. *PLoS Comput. Biol.* 11, 1–19. doi:10.1371/journal.pcbi.1004301

Lin, A., Adolphs, R., Rangel, A., 2012. Social and monetary reward learning engage overlapping neural substrates. *Soc. Cogn. Affect. Neurosci.* 7, 274–281. doi:10.1093/scan/nsr006

Lopez-Persem, A., Domenech, P., Pessiglione, M., 2016. How prior preferences determine decision-making frames and biases in the human brain. *Elife* 5, 1–20. doi:10.7554/eLife.20317

Meder, D., Kolling, N., Verhagen, L., Wittmann, M.K., Scholl, J., Madsen, K.H., Hulme, O.J., Behrens, T.E., Rushworth, M.F., 2017. Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nat. Commun.* 1–39. doi:10.1101/195842

Mullainathan, S., Shafir, E., 2014. *Scarcity: The True Cost of Not Having Enough*. Penguin.

Noonan, M.P., Walton, M.E., Behrens, T.E.J., Sallet, J., Buckley, M.J., Rushworth, M.F.S., 2010. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex 107, 20547–20552. doi:10.1073/pnas.1012246107

O'Reilly, R.C., Hazy, T.E., Mollick, J., Mackie, P., Herd, S., 2014. Goal-Driven Cognition in the Brain: A Computational Framework. *arXiv*. doi:10.1097/CCM.0000000000001615

Padoa-Schioppa, C., 2009. Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. *J. Neurosci.* 29, 14004–14014. doi:10.1523/JNEUROSCI.3751-09.2009

Padoa-Schioppa, C., Assad, J.A., 2008. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat. Neurosci.* 11, 95–102. doi:10.1038/nn2020

Rich, E.L., Wallis, J.D., 2016. Decoding subjective decisions from orbitofrontal cortex. *Nat. Neurosci.* 19, 1–10. doi:10.1038/nn.4320

Rushworth, M.F.S., Kolling, N., Mars, R.B., 2012. Valuation and decision-making in frontal cortex : one or many serial or parallel systems ? ' ro 946–955. doi:10.1016/j.conb.2012.04.011

Sladky, R., Friston, K.J., Tröstl, J., Cunningham, R., Moser, E., Windischberger, C., 2011. Slice-timing effects and their correction in functional MRI. *Neuroimage* 58, 588–594. doi:10.1016/j.neuroimage.2011.06.078

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. *Neuroimage* 46, 1004–1017. doi:10.1016/j.neuroimage.2009.03.025

Strait, C.E., Blanchard, T.C., Hayden, B.Y., 2014. Reward value comparison via mutual inhibition in

657 ventromedial prefrontal cortex. *Neuron* 82, 1357–1366. doi:10.1016/j.neuron.2014.04.032  
 658 Tsetsos, K., Wyart, V., Shorkey, S.P., Summerfield, C., 2014. Neural mechanisms of economic commitment in  
 659 the human medial prefrontal cortex. *Elife* 3, e03701. doi:10.7554/eLife.03701  
 660 Valentin, V. V., Dickinson, A., O'Doherty, J.P., 2007. Determining the Neural Substrates of Goal-Directed  
 661 Learning in the Human Brain. *J. Neurosci.* 27, 4019–4026. doi:10.1523/JNEUROSCI.0564-07.2007  
 662 Wallis, J.D., 2011. Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nat.*  
 663 *Neurosci.* 15, 13–19. doi:10.1038/nn.2956  
 664 Yamada, H., Louie, K., Tymula, A., Glimcher, P.W., 2018. Free choice shapes normalized value signals in  
 665 medial orbitofrontal cortex. *Nat. Commun.* 9, 1–11. doi:10.1038/s41467-017-02614-w  
 666

## **METHODS**

### **CONTACT FOR REAGENT AND RESOURCE SHARING**

All requests for further information, code, or fMRI data should be addressed to and will be fulfilled by the Lead Contact: Mr. Keno Juechems, Department of Experimental Psychology, University of Oxford, keno.juechems@psy.ox.ac.uk

### **EXPERIMENTAL MODEL AND SUBJECT DETAILS**

Twenty-two healthy volunteers with no history of psychiatric or neurological disorder participated in the study. One participant was excluded from analyses due to scanning equipment failure, leaving  $N = 21$  (5 male, mean age = 23.6,  $SD = 2.77$ ). All participants gave written informed consent and the study received approval from the ethics board of the University of Granada. Participants were compensated for their time with €25 plus a performance-based bonus (mean = €7.15,  $SD = €1.42$ , range: €4.52 to €11).

### **METHOD DETAILS**

#### **Task details**

Participants performed a task that involved managing a virtual zoo which housed two types of animals: lions and elephants. Each zoo began empty and animals were offered over a block lasting a variable number of trials (10-20), before a new zoo began. Participants encountered 5 zoos on each of 5 scanner runs, with each run lasting a total of 65 trials. On each trial, participants were first shown a fixation cross (150 ms), followed by a “choice” screen (2.5 seconds) in which they were offered a variable number of lions or elephants. This was indicated by an arabic number above a drawing of the animal on the left and right of the screen. There were always a high (4-6) offer of one animal and a low (1-3) offer of the other. The assignment

of high or low offers to each animal was autocorrelated over trials, reversing with a probability of 0.3, which led to “streaks” in which one animal could be harvested in more abundance. Participants indicated their choice by pressing a button with their left or right index finger (for options on the left and right, respectively). The side on which each animal was offered varied randomly across trials. Once a choice was made, the unchosen option disappeared from the screen, leaving only the chosen option (“response” screen), and this lasted the remainder of the trial (2.5s minus reaction time), followed by a blank screen for a uniformly jittered interval (1.5-4.5 seconds). Subsequently, a feedback screen (1 second) indicated the number of animals of each species in the zoo (the “assets”) as a number above the respective drawing. Trial-wise reward was indicated below the drawings and was directly proportional to the lower asset (e.g. it was 5 if there were 10 elephants and 5 lions in the zoo). Inter-trial intervals were drawn uniformly between 2s and 6s (1-3 time of repetition; TR).

Throughout the run, five coloured discs just below the top of the screen were shown, representing in the 5 zoos ordered from left to right in time. Each zoo had a colour, and the colour of the current zoo was indicated by a central circle that appeared on the “choice” and “response” screens. Colours were drawn randomly on each run and were only included to aide participants in grouping trials into discrete blocks. During the “choice” and “response” phases of the trial, a central number was displayed within the circle, also in a colour matching the colour of the current zoo. This number counted down the trials still remaining in the current zoo (including the current trial). At the conclusion of a block, its cumulative earnings were displayed on a “bonus” screen and this number was shown underneath the corresponding coloured disc on all subsequent trials. A lottery at the end of each run displayed to participants which of these bonuses was chosen as extra payment. Each block had equal probability of being selected in the lottery.

*Training and instruction.* Participants completed one run immediately before scanning to familiarize themselves with the task. They were told that if lions were currently the high offer, they were more likely to be the high offer on the following trial, but that this was not certain and this contingency could reverse unpredictably. However, they were not told the exact reversal probability (0.3). They were told that on some occasions it was more beneficial to defer choosing according to their current needs as this could yield better long-term payoff. Importantly, participants were not given specific instructions to maintain the goal difference at a certain level (e.g. as close to zero as possible). They were also told that they would receive the sum of the lottery values across runs after the end of the experiment.

#### **FMRI Data Acquisition and pre-processing**

MRI data were acquired on a 3T Siemens scanner. T1 weighted structural images were recorded directly prior to the task using an MPRAGE sequence:  $1 \times 1 \times 1 \text{ mm}^3$  voxel resolution,  $176 \times 256 \times 256$  grid, TR = 1900ms, TE = 2.52ms, TI = 900ms. Each fMRI image contained 32 axial echo-planar images (EPI) acquired in descending sequence with a voxel resolution of 3.5 mm isotropic, slice spacing of 4.2mm, TR = 2000ms, flip angle = 80, and TE of 30ms. 1850 EPI images were recorded per participant, resulting in a scanning time of around 62 minutes. Data were pre-processed using SPM12 (Wellcome Trust, London). As EPI acquisition used a descending sequence, images were corrected for slice time acquisition with the middle slice (at  $\text{TR}/2 = 1 \text{ s}$ ) as reference to minimize interpolation errors (Sladky et al., 2011). Scans were realigned to the first scan within each session. The anatomical scan was co-registered to the mean functional image. Anatomical scans were normalized to the standard MNI152 template brain. The functional EPI images were then normalized, resliced to  $4 \times 4 \times 4 \text{ mm}$  resolution, and smoothed with a full width half maximum Gaussian kernel of 8mm.

738

## 739 QUANTIFICATION AND STATISTICAL ANALYSIS

740 Throughout the paper, a statistical test was deemed significant when the resulting p-value for  
741 a two-tailed test was lower than 0.05 unless otherwise noted. Details about the specific test  
742 used, its N, measures of center and dispersion, and degrees of freedom can be found in the  
743 results section and in the corresponding figure captions.

### 744 **Analysis of behavior**

745 All analyses were carried out using MATLAB 8.6 (R2015b) and custom in-house code.

746 Choices were coded as “redress” when participants chose the offer corresponding to the  
747 currently lower asset in their zoo and “defer” if they chose the other option. All behavioural  
748 results discussed in the main text are based on this dichotomy (but see **SI** for behavioural  
749 analyses that classed choices as “greedy” with respect to the offers on the screen; **Fig. S4a**).  
750 Proximity was coded as 1/countdown.

### 751 **Logistic regression**

752 The dependent measure for the logistic regression was the binary variable 1=redress, 0=defer.  
753 We fit the models to individual subjects using z-scored regressors in five-fold cross-validation  
754 (one for each scanner run). We then averaged the betas across runs for each subject and used a  
755 two-tailed one-sample t-test to determine whether the group mean of each parameter differed  
756 significantly from zero. We excluded trials where the goal difference was zero from analysis  
757 ( $n = 1071$  out of  $N = 7150$  trials, of which almost half were first trials in a new block), because  
758 participants’ choice would be trivial (choose the higher offer) and the signed offer difference  
759 would be undefined in these cases.



The first logistic regression model contained the following regressors: a constant term, signed offer difference (need option minus not-needed option), absolute asset difference, and overall length of block. The second logistic regression model added proximity and the interaction of goal difference and proximity.

An additional model testing for the influence of rACC activity on behaviour differed from these models by splitting the asset difference into high and low asset which allowed us to test interactions of these terms with rACC activity.

Model fits were compared using five-fold cross-fitting via Matlab's fitglm function, summing log-likelihoods across participants and trials. In addition, we performed Bayesian model selection to compute the exceedance probabilities for each model, which measures whether a given model was more likely to be the true model than all competitor models (Stephan et al., 2009).

### **Dynamic Programming model**

We developed a Markov model to derive the optimal solution for the task. The model's state space was given by the goal difference, the number of trials left in the block, the two offers in the frame of reference of the lower asset, and the reversal probability (constant for all models, except when fitting to participants). The model was optimized using the following dynamic programming formula

$$Q(s, a) = R(s, a) + \sum_{s'} P(s'|s, a) \times \max_{a'} Q(s', a')$$

Where  $s$  and  $s'$  are the current and successor state,  $a$  and  $a'$  are the actions taken in each state,  $Q(s, a)$  is the value of taking action  $a$  in state  $s$ , and  $P(s'|s, a)$  is the probability of transitioning from state  $s$  to state  $s'$  by choosing action  $a$ , and  $R(s, a)$  is the reward obtained in the current state. Thus, the model chose the action which maximized the sum of immediate reward ( $R(s, a)$ )

and the expected value of the best action in all subsequent states. As the task had a finite horizon, we do not include a discount factor. We fit the model with different reversal probabilities (from 0 in increments of 0.05 to 1, where 0.3 was the true value) in order to compare it to participants' behaviour. As the model computed the expected value for both choosing to redress and choosing to defer, we computed the difference between the two and passed it through a sigmoid to allow for noisy decisions. The output of this sigmoid –  $p(\text{redress})$  – was then fit to participant's behaviour with a weighting parameter and intercept using maximum likelihood estimation. Maximum possible reward was calculated by permuting across all possible combinations of choices in each sequence and finding the highest reward score.

#### **FMRI data analysis**

Data were analysed using SPM12 and custom scripts. Data from the five scanning sessions were concatenated and constants identifying each run were added manually to the GLM in order to account for potential differences in mean activation and scanner drift. Stimulus onsets were incremented by 1s to account for slice time correction to the middle slice, which occurred at  $TR/2 = 1\text{s}$  after stimulus onset (Sladky et al., 2011). All results reported came from GLMs in which the canonical haemodynamic response function was convolved with a delta function (i.e. with zero seconds duration) time-locked to event onset (e.g. offer or feedback screen). The six vectors of motion parameters derived from pre-processing were included as nuisance regressors. Automatic orthogonalization was disabled. Group-level contrasts were constructed as simple t-contrasts using subject-level contrast images as input. Throughout this paper, we report only clusters that survived false-discovery-rate-correction (FDR) for multiple comparisons (Chumbley and Friston, 2009; Genovese et al., 2002) unless otherwise noted. This

avoids the potentially inflated false positive rates at the cluster level (Eklund et al., 2016). The FDR was calculated using a minimum cluster extent of 10 voxels and cluster-defining threshold of  $p < 0.001$ , uncorrected. These settings have been found to perform well in most cases (Eklund et al., 2018). Data were visualized using the xjview toolbox for SPM (<http://www.alivelearn.net/xjview>).

Our analyses are based on 4 GLM models. All 4 GLMs included two predictors of interest (onset of the choice screen, onset of the feedback screen) and two nuisance predictors (trials on which the goal difference was zero, and all no-response trials). GLM1 additionally included the following parametric regressors time-locked to the offer (i.e. at the time of choice): congruency (whether the high offer was of the currently needed type), response hand (left vs. right), signed and unsigned offer difference, sum of the offers, proximity, and goal difference. For completeness, we also included the following parametric modulators on trials with zero goal difference: response hand, unsigned offer difference, offer sum, proximity. Finally, goal difference (post-choice) was included as parametric modulator time-locked to the feedback screen. GLM2 included parametric modulators for the receipt value, goal receipt value, and redress value, as well as proximity, all time-locked to the onset of the offer. For completeness, we also included chosen value and proximity on trials with zero goal difference, and goal difference and cumulative value at feedback. GLM3 included parametric modulators corresponding to the two zoo tallies (up to that trial) their interactions with proximity, and their increases, as well as proximity. These modulators were time-locked to the onset of the choice screen. These regressors were z-scored to allow comparisons between their beta values. For completeness, we also included chosen and foregone receipt value on trials with zero goal difference, as well as goal difference and total sum of assets at feedback. GLM4 modelled each trial and feedback screen as a single event without any parametric modulation.

*Bayesian model selection.* We tested three variations of GLM2 which only included one of the three value codes – receipt value, goal receipt value, redress value. All other settings were identical to GLM2. We estimated these models in the striatum, dACC, and vmPFC ROIs based on the contrasts reported for GLM2 extracted using a leave-one-subject-out approach and Matlab’s fitglm function to obtain the log likelihood of the model. We then summed the log likelihoods across voxels as an approximation to the model’s log evidence. We used a standard Bayesian Model Selection approach in SPM described in Stephan et al. (2009). This approach uses a random effects Variational Bayes algorithm which iteratively estimates the probability of a model given the data. We can then estimate the exceedance probability – a measure of the probability that a given model outperformed all others in the comparison. Unlike other model selection approaches, this does not assume that all subjects were generated (best fit) by the same model. We repeated this procedure for all three regions.

*Region of interest extraction.* ROI were extracted using a leave-one-subject-out procedure to avoid potentially circular inference. A mask was created for each participant using the first-level contrast images from all other participants (i.e. leaving out the current participant). ROI were then constructed by taking all significant voxels ( $p < 0.001$ , uncorrected) within the relevant region (approximately given by xjview’s database, e.g. within medial PFC). We extracted one rACC ROI from the “goal difference” contrast from GLM1. In addition, we extracted one vmPFC ROI from the “redress value” contrast, a striatal ROI from “receipt value”, and a dACC ROI from “goal receipt value”, all from GLM2.

## DATA AND SOFTWARE AVAILABILITY

856 Behavioral data and custom code to reproduce figures can be accessed on our GitHub profile  
857 at [https://github.com/summerfieldlab/juechems\\_etal\\_zookeeper](https://github.com/summerfieldlab/juechems_etal_zookeeper) .

858

859