HHS Public Access

Author manuscript

Neuron. Author manuscript; available in PMC 2016 July 16.

Published in final edited form as:

Neuron. 2015 July 15; 87(2): 451–462. doi:10.1016/j.neuron.2015.06.031.

A neurocomputational model of altruistic choice and its implications

Cendri A. Hutcherson[1,4], **Benjamin Bushong**[1,3], and **Antonio Rangel**[1,2]

[1]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[2]Computational and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA

[3]Department of Economics, Harvard University, Cambridge, MA 02138, USA

[4]Department of Psychology, University of Toronto, Toronto, ON M1C 1A4, Canada

Summary

We propose a neurocomputational model of altruistic choice and test it using behavioral and fMRI data from a task in which subjects make choices between real monetary prizes for themselves and another. We show that a multi-attribute drift-diffusion model, in which choice results from accumulation of a relative value signal that linearly weights payoffs for self and other, captures key patterns of choice, reaction time, and neural response in ventral striatum, temporoparietal junction, and ventromedial prefrontal cortex. The model generates several novel insights into the nature of altruism. It explains when and why generous choices are slower or faster than selfish choices, and why they produce greater response in TPJ and vmPFC, without invoking competition between automatic and deliberative processes or reward value for generosity. It also predicts that when one's own payoffs are valued more than others', some generous acts may reflect mistakes rather than genuinely pro-social preferences.

Altruism involves helping others at a cost to the self, not only when such behavior is supported by strategic considerations like reciprocity or cooperation (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Nowak and Sigmund, 1998), but even in the absence of expectation for future benefit (e.g. fully anonymous, one-time generosity: Batson, 2011; Fehr and Fischbacher, 2003). A major goal of neuroeconomics is to develop neurocomputational models of altruistic choice, specifying which variables are computed, how they interact to make a decision, and how are they implemented by different brain circuits. Such models have proven useful in domains such as perceptual decision-making (Gold and Shadlen, 2007; Heekeren et al., 2008), simple economic choice (Basten et al., 2010; Hunt et al., 2012; Rangel and Clithero, 2013), self-control (Hare et al., 2009; Kable and Glimcher, 2007; Peters and Büchel, 2011; van den Bos and McClure, 2013), and social learning (Behrens et al., 2008; Boorman et al., 2013). We propose a neurocomputational

Correspondence: chutcher@hss.caltech.edu.

Author Contributions: C.A.H., B.B., and A.R. designed the experiment. C.A.H. and B.B. collected the data. C.A.H. developed the model and its predictions. C.A.H. analyzed the data. C.A.H., B.B., and A.R. wrote the paper.

model of simple altruistic choice and test it using behavioral and fMRI data from a modified Dictator Game in which subjects make choices between pairs of real monetary prizes for themselves ($*Self*) and another ($*Other*). These choices involve a trade-off between what is best for the self and what is best for the other, and thus require people to choose to act selfishly or generously.

Our model assumes that choices are made by assigning an overall value to each option, computed as the weighted linear sum of two specific attributes: monetary prizes for self and other. This type of simple value calculation captures a wide range of behavioral patterns in altruistic choice (Charness and Rabin, 2002; Eckel and Grossman, 1996; Engel, 2011; Fehr and Fischbacher, 2002; Fehr and Fischbacher, 2003). Our model also assumes that the overall value signal is computed with noise and that choices are made using a multi-attribute version of the Drift-Diffusion Model (DDM: Ratcliff and McKoon, 2008; Smith and Ratcliff, 2004). In this algorithm, a noisy relative value signal is integrated at each moment in time and a choice is made when sufficient evidence has accumulated in favor of one of the options. This type of algorithm has been shown to provide accurate descriptions of both choice and reaction time (RT) data (Busemeyer and Townsend, 1993; Hunt et al., 2012; Krajbich et al., 2010; Milosavljevic et al., 2010; Rodriguez et al., 2014; Smith and Ratcliff, 2004), as well as neural response patterns associated with computing and comparing values (Basten et al., 2010; Hare et al., 2011; Hunt et al., 2012) in many non-social domains.

The model suggests neural implementation of two specific quantities. First, values for the attributes $*Self* and $*Other* must be computed independently. Second, an overall value signal must be constructed from the independent attributes. We hypothesized that areas like the temporoparietal junction, precuneus, or medial prefrontal cortex may compute quantities related to the value of these attributes. Prior research strongly implicates these regions in social behavior (Bruneau et al., 2012; Carter and Huettel, 2013; De Vignemont and Singer, 2006; Decety and Jackson, 2006; Hare et al., 2010; Jackson et al., 2005; Moll et al., 2006; Saxe and Powell, 2006; Singer, 2006; Waytz et al., 2012; Zaki and Mitchell, 2011), although their precise computational roles remain poorly understood. Inspired by a large body of work on the neuroeconomics of non-social choice (Basten et al., 2010; Hare et al., 2009; Kable and Glimcher, 2007; Lim et al., 2013; McClure et al., 2004; Tom et al., 2007), we additionally hypothesized that the integration of specific attribute signals would occur in ventromedial prefrontal cortex (vmPFC). We explore these hypotheses with our fMRI dataset.

We also highlight three ways in which the development of a computational model of altruistic choice can be used to generate novel insights into the nature of altruistic choice. First, we compare the model's predictions about RT and neural response for generous versus selfish choices. We find that, for the best-fitting parameters, the model predicts longer RT and higher BOLD response in decision-related regions for generous choices, and that the predicted effect sizes match the observed data. Second, we use simulations to identify how model parameters influence altruistic behavior, and find that several of these variables (including the relative importance of benefits to self and other and the decision boundaries of the DDM) predict observed individual differences in generosity. Third, we show that the model predicts that generous decisions are sometimes unintended mistakes resulting from

the noisy choice process, and exploit an aspect of our experimental design to test this using fMRI data.

Results

We collected whole-brain BOLD responses in male subjects while they made 180 real decisions about different allocations of money between themselves and a real-but-anonymous partner. Each trial consisted of a choice phase and an outcome phase (Figure 1A). During the choice phase, the subject saw a proposal consisting of monetary prizes for himself ($*Self*) and for another person ($*Other*), and had to decide whether to accept or reject it in favor of a constant default prize of $50 for each. On each trial the subject saw one of the nine proposal types depicted in Figures 1B and 2C–D, with ±$1–$4 random jitter added to avoid habituation. All proposals included one payment below and one payment above the default, creating a choice between generous behavior (benefitting the other at a cost to oneself) and selfish behavior (benefitting oneself at a cost to another). Subjects indicated their decision using a four-point scale (1=Strong No, 2=No, 3=Yes, 4=Strong Yes), allowing us to measure both the choice and the value assigned to the proposal. Right-left orientation of the scale varied randomly from scan to scan to reduce motor-related confounds in neural response. Every decision was followed by an outcome phase, during which the decision made by the subject was implemented with 60% probability and reversed with 40% probability. Subjects were told about the 40% probability of choice reversal, and that their partner knew their choices might be reversed, but were encouraged to simply choose the option they most preferred, since their choice made it more likely to occur (see Supplementary Materials for instructions). At the end of the experiment one trial was randomly selected and its outcome implemented. As shown below, the reversal mechanism allows us to test the extent to which different choices may be decision mistakes, while not changing incentives to pick the best option.

Average choices are relatively selfish

Subjects made generous choices—maximizing their partner's payoff ($*Other*) at a cost to their own ($*Self*)—in 21% ±18 (mean ± SD) of trials, sacrificing $3.73 ± 4.64 per trial and giving $8.31 ± 6.86. This level of giving is comparable to other studies of anonymous altruism (Engel, 2011), but also suggests that subjects in general behaved relatively selfishly. There was considerable individual variation in generosity, ranging from 0%–61% generous choices and $0-$22.37 given to the partner. This variation is useful for exploring individual differences, as we do below.

Computational model

The model is a multi-attribute extension of the standard DDM (Ratcliff and McKoon, 2008; Smith and Ratcliff, 2004). On every trial the choice is based on a dynamically evolving stochastic relative decision value (RDV) signal that provides an estimate of the desirability of the proposed prize ($*Self*, $*Other*) relative to the default prize ($50, $50). The signal starts at zero, remaining there for an amount of non-decision time capturing processing and motor delays, given by the parameter *NDT*. Afterwards, it accumulates stochastically at time *t* according to the difference equation

$$RDV_t = RDV_{t-1} + w_{self}(\$Self - \$50) + w_{other}(\$Other - \$50) + \varepsilon_t,$$

where $\$Self$ and $\$Other$ are the proposed prizes for self and other, w_{self} and w_{other} are constant weights, and ε_t denotes white Gaussian noise that is identically and independently distributed with standard deviation σ. A choice is made the first time the RDV crosses one of two pre-specified barriers. The proposal is accepted if the positive barrier is crossed first and rejected if the negative barrier is crossed first. RT equals the sum of the NDT and crossing time t. Building on previous work with time-limited decisions (Churchland et al., 2008; Cisek et al., 2009; Milosavljevic et al., 2010), we allow for the possibility of collapsing barriers, although the model includes fixed barriers as a special case. The upper barrier is described by the equation

$$\overline{B}_t = be^{-td},$$

where $b > 0$ is a parameter denoting the initial height of the barrier, $d \geq 0$ is a parameter denoting its exponential rate of decay, and t is measured from the end of the non-decision period. The lower barrier is symmetric, so that $\underline{B}_t = -\overline{B}_t$. Without loss of generality, we assume that $\sigma = 0.1$, since the model is invariant to affine transformations of the parameters (Ratcliff and McKoon, 2008). The model has five free parameters: NDT, w_{self}, w_{other}, b and d.

Figure 1C illustrates the model. The relative value of the proposal is given by $V = w_{self}(\$self - \$50) + w_{other}(\$Other - \$50)$. When V is positive the optimal choice is to accept the proposal and otherwise to reject it. The decision problem is complicated if V is measured with noise at every instant during the decision phase, especially if the amount of noise is high. The DDM algorithm provides an elegant solution to this problem: by dynamically integrating the instantaneous noisy value measures, RDV generates a posterior estimate of the log-likelihood ratio that the optimal choice is to accept (Bogacz et al., 2006; Gold and Shadlen, 2002). The barriers define a rule for how large this posterior estimate has to become to make a decision. The barriers collapse over time to allow choices to occur in a reasonable timeframe even for trials with low RDV where the signal moves away from zero very slowly.

Several aspects of the model are worth highlighting. First, although choices and RTs are inherently stochastic, the model makes quantitative predictions about how different parameters and proposal amounts affect their distribution. Second, the size of the weights w_{self} and w_{other}, as well as the barrier location, affects the quality and speed of choices, a phenomenon known as the speed-accuracy tradeoff (Bogacz et al., 2010). This implies that mistakes are possible: individuals sometimes act selfishly or altruistically despite their underlying preferences. When the barriers are initially high and decay slowly (b large and d small), decisions are slower but made more accurately. Finally, the relationship of w_{self} to w_{other} plays a critical role. If $w_{self} = w_{other}$, the model predicts that $\$Self$ and $\$Other$ influence choices, RTs, and errors symmetrically. In contrast, if $w_{self} > w_{other}$, changes in

$*Self* have a stronger impact on choices and RTs, and errors are distributed asymmetrically, such that they more frequently involve excessive generosity (more on this below).

The model accurately predicts out-of-sample choice and RT

We used a maximum likelihood method based on simulated likelihood functions to estimate the best-fitting parameters of the model in a randomly-selected half of the data. We used these parameters to test the fit between model predictions and observed data on the other half, separately for each subject (see Methods for details). Model predictions capture inter-individual differences in mean donations (mean Pearson's r_{49}= .94, $P < .0001$) and RTs (r_{49} = .96, $P < .0001$) quite well (Figure 2A, B). The model also captures intra-individual differences in acceptance rates (mean r = .88, one-sample t_{50} = 45.05, $P < .0001$; Figure 2C) and in RTs (mean r = .53, one-sample t_{50} = 11.79, $P < .0001$; Figure 2D) across different trial types.

Although this suggests that the model described above fits well, Figure 2C indicates that the fit for choice behavior (though not RT) was poorer when the proposal involved a sacrifice for the subject (i.e., $*Self* amounts below the default). To investigate this issue, we fit a variant of the model that allows the parameters to depend on whether $*Self* is more or less than $*Other*. This alternative model is motivated by previous behavioral work showing that the value placed on $*Self* and $*Other* can depend on whether the self is coming out ahead or behind (Charness and Rabin, 2002; Engelmann and Strobel, 2004; Fehr and Schmidt, 1999). As detailed in the Supplemental Materials, this analysis improves the fit to observed choice behavior when $*Self* < $*Other*, an effect that derives from a higher weight w_{self}, a lower weight w_{other}, and a higher threshold parameter. However, because there are no qualitative differences in the analyses reported below when using the more complex model, for simplicity the rest of the analyses utilize the simpler version.

Estimated model parameters in the full dataset

Having demonstrated that the model accurately predicts out-of-sample choices and RTs, we next examine the best-fitting parameter values using the full dataset (see Table 1). Several results are worth highlighting. First, the average NDT (868ms) is larger than that usually found for DDMs (Milosavljevic et al., 2010; Ratcliff and McKoon, 2008). We attribute this to the additional time subjects may have needed to determine the payoffs on each trial and translate that into a graded response. Second, both w_{self} and w_{other} are significantly larger than zero on average (both $P < .003$) and w_{self} is considerably larger than w_{other} (paired-t_{50} = 10.83, $P < .001$) Third, we find substantial individual variation in the best-fitting values for all five parameters, which is useful for the individual difference analyses described below.

vmPFC responses encode an integrated value signal at decision

The model suggests that an integrated value signal is used to make choices. We provide neural evidence of such a signal by estimating a general linear model to identify regions in which BOLD responses correlate positively with the value assigned to proposals at the time of decision, measured by the four-point response scale (1 = Strong-No to 4 = Strong-Yes). Several regions satisfy this property, including a region of vmPFC ($P < .05$, whole-brain

corrected [WBC]; Figure 3A; Table S3) that encodes stimulus values at the time of decision in a wide range of tasks (Clithero and Rangel, 2013; Kable and Glimcher, 2007).

Neural representations of $Self and $Other

Our model assumes that the overall value assigned to the proposal (and used by the DDM comparator algorithm to generate a choice) is constructed from information about the independent attributes $*Self* and $*Other*. We show that there are neural signals consistent with representation of these two quantities, using a second model to look for areas in which BOLD responses correlate positively with either $*Self* or $*Other* separately (i.e., inputs to the integrated value signal), as well as regions that reflect both (i.e., overall values).

$*Self* correlates with BOLD responses in a distributed set of regions (Figure 3B, Table S4), including vmPFC ($P < .05$, WBC) and the ventral striatum ($P < .05$, WBC). $*Other* correlates with BOLD responses in a distinct and more circumscribed set of regions (Figure 3C, Table S4), including right temporoparietal junction (rTPJ), precuneus (both $P < .05$, WBC), and vmPFC ($P = .004$ SVC). To determine the specificity of these responses, we looked for regions in which the effect for $*Self* is stronger than for $*Other*, and vice versa. Regions responding more strongly to $*Self* include the ventral striatum ($P < .05$, SVC), vmPFC, and areas of visual and somatosensory cortex ($P < .05$, WBC). No regions respond more strongly to $*Other* at our omnibus threshold, although we observe such specificity in the right TPJ at a more liberal threshold ($P < .005$, uncorrected). A conjunction analysis (Table S4) shows a region of vmPFC (Figure 3D) responding significantly to both $*Self* and $*Other* ($P < .05$, SVC). This area also overlaps fully the vmPFC area correlating with overall preference, supporting the idea that it may represent an area where separate attributes are combined into an integrated value signal.

Together, the behavioral and neural results are consistent with the hypothesis that both $*Self* and $*Other*, quantities required by the computational model, are independently represented in the brain. The results also support the idea that the vmPFC combines information about $*Self* and $*Other* into an overall value, and that choices are made by integrating the proposal values using an algorithm that is well captured by the DDM. These results motivate the second part of the paper, in which we use the best-fitting computational model to derive and test several implications of the theory.

Implication 1: RTs are longer for generous choices, particularly for more selfish individuals

Our computational model has the advantage that it provides a theory of the relationship between choices and RTs. This is of particular interest because differences in RT when choosing to act selfishly or generously have been used in several studies to make inferences about the relative automaticity of pro-social behavior (Piovesan and Wengstrom, 2009; Rand et al., 2012). Simulations from the individual models reveal two interesting predictions. First, in the domain of best-fitting parameters for our subjects, the model predicts that on average RTs are longer for trials that result in a generous (G) choice compared to trials resulting in a selfish (S) choice (predicted $RT_G = 2{,}269$, predicted $RT_S = 2{,}074$, paired-$t_{50} = 9.37$, $P < .0001$; Figure 4A). Second, it predicts that this RT difference is bigger for more

selfish subjects (correlation between predicted generosity and difference in G vs. S RTs r_{49} = −.89, P < .0001; Figure 4B). The observed data displayed both patterns. On average, G choices were significantly slower than S choices (RT_G = 2,300 ms ± 310, RT_S = 2,131 ms ± 280, paired t_{43} = 4.97, P < 0.0001; Figure 4C), and the more generous the individual, the smaller this difference (r_{42} = −.60, P < .001, Fig. 4D).

Implication 2: Neural response in valuation and comparison regions is higher for generous choices

Our computational model suggests a neural corollary of differences in RT: regions whose activity scales with computation in the comparison process should have higher responses during G compared to S choices (predicted comparator response, arbitrary units: $Comp_G$ = 69.68 ± 27.97, S choice = 65.76 ± 27.62, paired-t_{50} = 6.43, P < .0001, see Methods for details). To see why, note that the predicted area-under-the-curve of the accumulator process is larger on longer trials, and that inputs into this process must also be sustained until the process terminates at a decision barrier. This prediction is important, since many studies of altruism have observed differential response during pro-social choices in regions like the vmPFC and TPJ and interpreted it as evidence that such choices are rewarding (Zaki and Mitchell, 2011), or that they require the inhibition of selfish impulses by the TPJ (Strombach et al., 2015). In contrast, our model suggests that such differences could be a straightforward by-product of the integration and comparison process.

To test this prediction, we first defined two independent ROIs shown in previous research to have differential response during G choices: 1) a value-modulated vmPFC region (Figure 5A) based on the set of voxels that correlated significantly with stated preference at the time of choice (P < .0001, uncorrected); and 2) a generosity-related TPJ region (Figure 5B) based on an 8-mm sphere around the peak coordinates of a recent study reporting greater activation in the TPJ when subjects chose generously (Strombach et al., 2015). In both regions, we replicate the pattern of higher response on G vs. S choices (both P < .02). Critically, however, we also find that differential BOLD on G vs. S choices correlates positively with predicted differences in accumulator response in both regions (both r_{42} > .46, P < .001). Moreover, accounting for predicted accumulator differences reduces to non-significance the differential generosity-related response in both vmPFC (P = .92) and TPJ (P = .91). In contrast, response in occipital and motor cortices (which show value modulation but are unlikely to perform value integration and comparison) bear little resemblance to predictions of the model (Figure S2).

Implication 3: Relationships between model parameters and generosity

In order to understand the impact on generosity of variation in the different parameters, we simulate model predictions for our task for a wide range of parameter combinations (see Methods for details). For each parameter combination, we calculate the average generosity (i.e., the average amount of money given to the other over all trials by choosing generously). Then, we use multiple regression to measure the independent influence of the five model parameters on variation in average simulated generosity. In the regression, all parameters are normalized by their mean and standard deviation in order to assess their influence on a common scale. As illustrated in Figure 6A–B, we find the expected association between

average generosity and both w_{Self} (β = −4.86, P < .0001) and w_{Other} (β = 9.26, P < .0001). Intriguingly, the simulations also reveal that a lower starting threshold (β = −.141, P = .0001), and a faster collapse rate (β = .21, P < .0001) increase generosity. That is, individuals with less stringent barriers tend, under the model, to make more G choices, holding w_{Self} and w_{Other} constant. Based on the results below, we attribute this shift to an increase in choice errors for individuals with less stringent decision-criteria, leading to less accurate (and in this case, generosity prone) behavior. The model predicts no relationship with NDTs (P = .79).

We use a similar regression to see if the same relation is evident in the observed data. As with the simulated data, the fitted parameters were z-scored to assess their influence on a common scale. Consistent with model predictions, we find that average observed generosity correlates negatively with w_{Self} (β = −3.11 ± .28, P < .0001) and positively with w_{Other} (β = 6.47 ± .29, P < .0001). Also as predicted, observed generosity correlates negatively with the height of the decision threshold (β = −1.51 ± .41, P = .0006) and positively with the rate at which the threshold collapses toward zero (β = 1.08 ± .37, P = .006). The NDT parameter is non-significant, as expected.

Implication 4: Errors are more likely to involve generous choices

Because choices are stochastic, the model suggests that some may be errors (i.e., a decision in which the option with the higher relative value is not chosen: Bernheim and Rangel, 2005). We use the simulated data to investigate how decision mistakes change with variation in the model parameters, and how this affects generosity. Multivariate regression analyses on the theoretical data, where mistakes can be identified precisely on every trial, show that error rates decrease with w_{Self} (β = −.05, P < .0001) and w_{Other} (β = −.014, P < .0001), and increase with more liberal barrier parameters for b (β = −.04, P < .0001) and d (β =.05, P < .0001). We also assess the relationship between model parameters and the relative percentage of trials that result in *generous* errors (mistakenly choosing to give to the other) vs. *selfish* errors (mistakenly choosing to keep more money). This assesses whether different parameters increase generosity by increasing errors. As shown in Figure 6C, w_{Self} (β = .026, P < .0001) and w_{Other} (β = −.047, P < .0001) influence the relative balance toward generous errors in opposite ways. Increasing the height of the barrier decreases the bias toward generous errors (*b:* β = −.008, P < .0001; *d:* β = +.01, P < .0001; Figure 6D), while NDT has no effect (β = −.0001, P = .89).

We next use the individually-fitted weights w_{Self} and w_{Other} to define the “true” relative value of each proposal, which allows us to estimate the proportion of observed G and S choices for each subject that might reasonably be assumed to be errors. This analysis suggests that G choices were significantly more likely to be errors (M = 49% ± 38%) than S choices (M = 10% ± 21%, paired-t_{50} = 5.45, P < .0001).

We carry out a further test of this prediction, using outcome period BOLD responses, based on the following logic. The model suggests that a proposal’s true value should become increasingly clear to the decision circuitry as the amount of accumulated evidence increases over time, because random fluctuations in the signal will tend to cancel out. If subjects continue to accumulate evidence about the proposal even after making a choice (i.e.,

"double-checking" whether they have made an error), then these signals should be quite clean by the time the subject sees the outcome of his choice. If this increased clarity leads a subject to realize at some point after making his choice that it was a mistake, having that mistake overturned during the outcome period (yielding the unchosen but ultimately preferred option) should be perceived as "good news" (i.e. relief), whereas having it implemented should be experienced as "bad news" (i.e. disappointment). If the original choice is actually correct, then reversal of this choice should be perceived negatively. The model thus predicts that reversal of S choices (which simulations suggest are generally likely to be correct) should be associated with negative affect and lower response in brain regions coding for the utility of an outcome, relative to non-reversal. In contrast, because G choices more likely reflect choice errors, reversing them should be more likely to evoke positive affect and greater neural response compared to implementation. Finally, the response in utility-coding areas to reversing a G choice should increase, across subjects, with the model-estimated likelihood that G choices are mistakes.

We tested these predictions by computing the difference between response in the vmPFC to reversal vs. implementation of G or S choices, controlling both for the strength of preference at the time of decision, and for actual outcomes received (i.e. the amounts $*Self* and $*Other* resulting from choice combined with the random implementation, see GLM 1 in Methods for details). Consistent with predictions, vmPFC response to reversal vs. implementation was significantly higher after G compared to S choices ($P = .02$, SVC, Figure 7A). Also as predicted by the model, the difference in response in this region correlated positively with the estimated excess rate of mistakes for G over S choices ($r_{39} = .43$, $P = .004$, Figure 7B).

Discussion

We have proposed a neurocomputational model of altruistic choice that builds on behavioral, neural and computational work in non-social domains (Basten et al., 2010; Bogacz et al., 2010; Hare et al., 2011; Heekeren et al., 2008; Ratcliff and McKoon, 2008). In the model, decisions result from the stochastic accumulation of a relative value signal that linearly weights information about payoffs for self and other. Despite its simplicity, the model has considerable explanatory power. It accounts for differences in average levels of altruism and RTs within and across subjects, as well as for neural signals encoded in vmPFC, TPJ and striatum at the time of choice. Our results provide insight into the common processes at work in altruistic choice and simple non-social decisions, shed light on some of the neural mechanisms specifically involved in the computation of social value, and provide novel insights into the nature of altruistic behavior.

Simple vs. social decision-making

A growing body of work suggests that in simple non-social choices the vmPFC receives information from regions computing information about different stimulus attributes (Basten et al., 2010; Hare et al., 2009; Kable and Glimcher, 2007; Lim et al., 2013), and combines it into a relative value signal (Hare et al., 2009; Kable and Glimcher, 2007). This signal is then dynamically integrated in comparator regions using algorithms with properties similar to the DDM (Basten et al., 2010; Hare et al., 2011; Hunt et al., 2012). BOLD responses in our

study suggest a similar neural architecture for social choice: we observed attribute-coding regions like the striatum and TPJ (which correlated with $*Self* and $*Other*, respectively), as well as a vmPFC region that represented $*Self* and $*Other* simultaneously and encoded the overall value of a choice.

Neural and psychological bases of pro-social decision-making

Although several studies have shown that the TPJ plays a role in empathic and altruistic decision-making (Decety and Jackson, 2006; Hare et al., 2010; Morishima et al., 2012; Saxe and Powell, 2006) its precise computational nature remains poorly understood. Our findings show that, at least during altruistic choice, signals in this area may reflect computations related to others' interests. This signal differs from two popular alternative accounts of TPJ function: that it represents the beliefs of others (Saxe and Powell, 2006) or that it allows attention to be shifted away from the self (Scholz et al., 2009). Neither theory appears to fully explain the pattern of TPJ response in our task. Although computations about belief might be used to determine how another person would value the proposal, it is not clear why this representation would appear selectively for higher rather than lower amounts of $*Other*. Similarly, if attentional shifts help to incorporate others' rewards into the value signal, they should be equally important for both gains and losses (relative to the default). This predicts a correlation with the absolute value of $*Other* vs. the default, rather than the positive linear response observed here. Tasks observing belief representation or attentional reorienting in TPJ typically examine these processes in an evaluation-free context. We speculate that the explicit use of attention or belief representations to construct *value representations* may produce the pattern of results here. Future work will be needed to determine whether such differences can help to integrate the current findings with previous literature.

The specific processes tapped by our task may also explain why we do not find other areas often implicated in prosociality, such as the anterior insula. While this region contributes to a variety of cognitive and affective functions (Kurth et al., 2010), studies implicating insula in social decision-making typically involve a strong component of negative affect, such as the pain and suffering of a victim (Singer, 2006; Singer et al., 2004). These considerations may play a more limited role in our task, which likely involves more abstract representation of costs and benefits. Exploring how different task features influence the specific neural and psychological processes deployed during altruistic decisions represents an important avenue for research.

Model implications

Several implications of the model showcase the value of computational approaches, and provide novel insights into the nature of pro-social behavior. First, consistent with the data, model simulations predict that generous decisions are made more slowly, but that this slowdown is less pronounced in more generous subjects. This observation has direct relevance for a literature that has made the case for dual-process models of social decision-making based on RT differences between generous and selfish choices (Piovesan and Wengstrom, 2009; Rand et al., 2012; Tinghog et al., 2013). Our results suggest caution in interpreting these RT differences, by showing how they can arise without requiring competition between "fast and automatic" and "slow and deliberative" systems. In our

model, generous choices are made more slowly if the relative weight placed on the self is higher, but more *quickly* if weights on others' payoffs are higher.

This could help to reconcile some of the apparently contradictory results in this literature. Different contexts can evoke dramatically different levels of altruistic or pro-social behavior (Engel, 2011). Studies observing faster RTs for more generous or cooperative choices (Rand et al., 2012) may establish contexts in which, for a variety of reasons, the needs of others are weighted more highly, while studies observing slower RTs (Piovesan and Wengstrom, 2009) may prime subjects toward reduced consideration of others. Note, however, that our results do not undermine the general validity of dual-process frameworks. Indeed, in some respects, our model can be interpreted as involving dual processes with respect to valuing self- and other-interests, but suggests that RT data should be used carefully and *in conjunction with* more formal computational models to derive and test predictions.

Second, the model has similar implications for the interpretation of neural response. It predicts that for subjects with a bias toward the self (i.e., almost everyone), brain areas whose activity scales with computations in the accumulation and comparison process will have greater response on trials resulting in generous choice. We find this pattern in both the TPJ and vmPFC, and show that it can be accounted for by the neurocomputational model. These results urge caution in interpreting generosity-specific activation in TPJ as inhibition of selfish impulses (Strombach et al., 2015), or in concluding from activation differences in vmPFC that choosing generously is rewarding (Strombach et al., 2015; Zaki and Mitchell, 2011). A simple neurocomputational model with identical parameters on every trial reproduces these differences without requiring that choosing generously specifically involve either self-control or a special reward value.

A third implication of the model concerns the relationship between individual differences in generosity and specific model parameters. Not surprisingly, generosity increases with the weight to other and decreases with the weight to self. More surprisingly, generosity also increases with less stringent barriers (i.e., lower starting threshold and a faster collapse rate). Thus, systematic differences in altruistic behavior may not reflect different underlying preferences (i.e. weights on self and other), but simply alterations in the amount of noise in the decision process. This observation has important implications for the large body of social decision-making literature that has used manipulations that might influence barrier height and response caution, such as time pressure (Rand et al., 2012), cognitive load (Cornelissen et al., 2011), or even electrical brain stimulation (Ruff et al., 2013). The results of these studies are often assumed to support a role of self-control in increasing (or decreasing) consideration of others' welfare. Our results point to an alternative interpretation, and suggest that greater attention should be paid to the precise mechanism of action through which different manipulations influence generosity.

The observation that noise can induce systematic shifts in choice without systematic shifts in preferences leads to the final implication of our model: that a significant fraction of generous choices may be decision mistakes. Results from the outcome period in our study suggest that people track these errors, and may feel relieved when the consequences of such errors are avoided due to external contingencies. This insight has profound implications for our

understanding of both basic decision-making and pro-sociality. It adds to other work on impure altruism (Andreoni, 1990; Andreoni and Bernheim, 2009), suggesting that any single generous act can result from many processes that have little to do with the true value we assign to others' welfare.

Experimental Procedures

Participants

Male volunteers ($N = 122$) were recruited in pairs from the Caltech community. Half were active participants who completed the scanning task. The other subjects participated passively as described below. All were right-handed, healthy, had normal/corrected-to-normal vision, were free of psychiatric/neurological conditions, and did not report taking any medications that might interfere with fMRI. All participants received a show-up fee of $30 as well as $0-$100 in additional earnings, depending on the outcome of a randomly chosen experimental trial. We excluded data from ten scanning subjects due to excessive head motion or technical difficulties during scanning (remaining 51 subjects: 18–35 years of age, mean 22.3). Caltech's Internal Review Board approved all procedures. Subjects provided informed consent prior to participation.

Task

Each participant in a pair arrived separately to the lab and waited in a private area where he received instructions. We randomly designated one participant as the active participant (AP), who completed the tasks described below. We designated the other as the passive partner (PP) who, after receiving instructions, waited in a separate room for the study duration. The PP's presence created a real and non-deceptive social context for the AP.

The AP made 180 real decisions in a modified Dictator Game. On each trial he chose between a proposed pair of monetary prizes to himself and his partner and a constant default prize-pair of $50 to both (Figure 1A). Proposed prizes varied from $10 to $100 and were drawn from one of the nine pairs shown in Figure 1B. Each pair appeared 20 times, randomly intermixed across trials, and divided evenly across four scanner runs (5 instances/run). To minimize habituation and repetition effects, proposal amounts were randomly jittered by $1–$4, with the exception that amounts above $100 were always jittered downwards. The side of the screen on which $*Self* appeared was counterbalanced across subjects but was constant throughout the task.

All prize-pairs included one payment below and one payment above the default, and thus involved a choice between generous behavior (benefitting the other at a cost to oneself) and selfish behavior (benefitting oneself at a cost to the other). After presentation of the proposal, subjects had up to four seconds to indicate their choice using a 4-point scale (Strong No, No, Yes, Strong Yes). This allowed us to simultaneously measure both their decision and the relative value of the proposed payment at the time of choice. The direction of increasing preference (right-to-left or left-to-right) varied on each scan. If the subject failed to respond within 4 s, both individuals received $0 for that trial. Although this time limit could be considered a form of time pressure, pilot testing with free response times

suggested that a relative minority of choices (14%) took longer than four seconds and that other basic properties of choice and RT were similar to the current study.

The task also included a second component designed both to increase the anonymity of choices and allow us to test a prediction made by the DDM about the possibility of decision mistakes. After a random delay of 2–4 s following response, the subject's choice was implemented probabilistically: in 60% of trials he received his chosen option, while in 40% his choice was reversed and he received the alternative non-chosen option. This reversal meant that while it was always in his best interest to choose according to his true preferences, his partner could never be sure about the actual choice made. APs were informed that the PPs were aware of the probabilistic implementation. The 40% reversal rate was necessary to test key predictions of the model, but raises the concern that it alters decision computations. Pilot testing with only 10% choice reversals yielded nearly identical behavioral results, suggesting this is not likely an issue.

Behavioral definition of generosity

We label specific decisions as Generous (G) if the AP gave up money to help the PP (i.e., accepting $\$Self < \$50$ or rejecting $\$Self > \$50$), and as Selfish (S) otherwise. Subject-level generosity was measured by the average amount of money per trial that a subject gave to the PP by choosing generously. Alternative measures of generosity (such as money sacrificed) led to similar results.

Model estimation

We use maximum likelihood to estimate the value of the free parameters that provide the best fit to the observed choice and RT data, separately for each AP. For assessing the goodness-of-fit of the model, we estimate these parameters separately for half of the trials and test the accuracy of predictions in the other half of the data. For testing model implications, we use the full set of trials for each AP to fit the parameters. Fitting was done in several steps.

First, we ran 1000 simulations of the DDM to compute the likelihood function over observed choices (Yes/No) and RT bins separately for each proposal-pair used in the experiment and each possible combination of parameters. RT bins were specified in 250-ms increments from 0 to 4 s, and included one additional bin for non-responses (simulations in which the RDV failed to cross a barrier within the 4-sec time limit). The combination of parameters used covers a grid determined by the cross-product of the following sets: $\boldsymbol{w_{Self}} = \boldsymbol{w_{Other}}$ = [−.045, −.003, −.0015, 0, .0015, .0003, .0045, .006, .0075, .009, .0105, .012, .0135], $\boldsymbol{NDT}$ = [.3, .5, .7, .9, 1.1, 1.3], $\boldsymbol{B}$ = [.04, .06, .08, .1, .12, .14, .16, .18, .2, .22, .24, .26, .28, .3, .32] and $\boldsymbol{b}$ = [0, .00005, .0001, .00025, .0005, .00075, .001, .005]. The range of the grid was chosen by trial-and-error so that no more than 10% of subjects fell on a boundary edge for any parameter, while keeping the total number of parameter combinations low to minimize exploding computational costs.

Second, for each subject we identified the parameter combination that minimized the negative log-likelihood (NLL) of the selected trials observed for that subject, based on

likelihoods generated from the simulated data. If more than one parameter combination resulted in the same minimal NLL, one was randomly selected as the solution.

Model simulations—To assess model fits to behavior, we used half of each subject's responses (randomly selected) to find the best-fitting parameters, and simulated 1000 runs of the other half of trials seen by that individual. We then compared observed and simulated values for the average amount given to the partner, average RT on G and S choice trials, and average choices and RTs for particular proposals.

We also simulated data from the best-fitting parameters for *all* trials in each subject, to explore other model implications. First, we predicted overall response in the accumulator (Implication 2 and Figure 6). We speculated based on prior research (Basten et al., 2010) that several brain regions may contribute to this computation, and explored the implications of this architecture for understanding behavioral and neural correlates of generosity. We follow Basten et al. (2010) in defining accumulator response for each trial as $\Sigma_t |\mathrm{RDV}_t|$, and estimate it separately for G and S choices (see GLM 3 below). Second, we explored how individual variation in model parameters affects generosity (Implication 3). Finally, we used simulations to understand the role of choice errors in producing altruistic behavior (Implication 4 and Figure 7).

fMRI

fMRI data was acquired and preprocessed using standard procedures (see Supplemental Experimental Procedures for details). Using this data we estimated three different general linear models (GLMs) of BOLD response.

GLM 1—The first GLM served two purposes: 1) to identify regions associated with the overall decision value of the proposal behaviorally expressed at the time of choice; and 2) to test the hypothesis that many generous choices should be considered errors, and thus be perceived as good news if they are reversed.

For each subject we estimated a GLM with AR(1) and the following regressors of interest: R1) A boxcar function for the choice period on all trials; R2) R1 modulated by the behaviorally expressed preference, ranging from 1 = Strong No to 4 = Strong Yes; R3) a boxcar function of 3 s duration for the outcome period; R4) R3 modulated by the outcome for self on each trial; R5) R3 modulated by the outcome for other on each trial; R6) R3 modulated by a variable consisting of a 1 for every trial in which the subject chose generously but the choice was vetoed, a −1 for every trial in which the subject chose generously and the choice was implemented, and 0 otherwise (i.e. after a selfish choice); R7) R3 modulated by a variable similar to R6, but which was 1 for veto of selfish choices, −1 for implementation of selfish choices, and 0 otherwise. No orthogonalization was used, allowing regressors to compete fully for explained variance. All regressors of interest were convolved with the canonical form of the hemodynamic response. The model also included motion parameters and session constants as regressors of no interest. Missed response trials were excluded from analysis.

We then computed second-level random effects contrasts with one-sample *t*-tests, using the single-subject parameter estimates to construct several contrasts. We used R2 to determine areas correlated with behaviorally expressed preference at the time of choice. We used R6 and R7, and their difference, to explore activation related to choice reversal. Because outcomes for self and other are entered as modulators, R6 and R7 reflect differences in response *over and above* those associated purely with the amounts received when the outcome is revealed.

For inference purposes, we imposed a family-wise error cluster-corrected threshold of $P < .05$ (based on Gaussian random field theory as implemented in SPM5). We also report results surviving small-volume correction within regions for which we had strong *a priori* hypotheses (see *ROI definition* below), including vmPFC and TPJ.

GLM 2—This GLM identified regions in which activity correlates with proposed payments at the time of choice. It included the following regressors: R1) A boxcar function for the choice period on all trials; R2) R1 modulated by *$Self* on each trial; R3) R1 modulated by *$Other* on each trial; R4) a 3 s boxcar function for the outcome period; R5) R4 modulated by the outcome for self on each trial; R6) R4 modulated by the outcome for other on each trial. All other details are as in GLM 1. Using GLM 2, we calculated three single-subject parametric contrasts: R2 vs. zero, R3 vs zero, and R2 vs R3.

GLM 3—We used GLM 3 to test predictions about comparator differences on G vs. S choice trials. It included the following regressors: R1) A boxcar function for the choice period on trials when the subject chose selfishly; R2) R1 modulated by the behaviorally expressed preference at the time of choice; R3) A boxcar function for the choice period on trials when the subject chose generously; R4) R3 modulated by behavioral preference. R5-9 were identical to R3-7 from GLM 1. The contrast R3 vs. R1 identified regions with differential response for G vs. to S choices.

Within two independently-defined ROIs in vmPFC and TPJ (see Figure 5), we calculated the average value of the contrast R3 – R1 for each subject, and regressed it on the model-predicted difference in accumulator activity ($BOLD_{G-S} = \beta_0 + \beta_1 * DDM_{G-S}$). This allowed us to determine if predicted accumulator differences were associated with BOLD differences on G vs. S choice trials, and whether differential generosity-related BOLD response (i.e., β_0) remained significant after controlling for predicted accumulator differences.

ROI Definition—For use in small-volume corrections, as implemented in SPM5, we defined three *a priori* regions-of-interest: a vmPFC region associated with decision value, a vmPFC region associated outcome value, and bilateral TPJ. We defined decision-value related vmPFC using the conjunction of two recent meta-analyses on decision-related reward valuation (Bartra et al., 2013; Clithero and Rangel, 2013). Outcome-value related vmPFC was defined in a similar way, but based on meta-analysis results for value representations at outcome, which may preferentially activate more anterior vmPFC regions (Clithero and Rangel, 2013). The TPJ mask was defined anatomically using the WFU PickAtlas (http://fmri.wfubmc.edu/software/PickAtlas), with a dilation of 3mm to ensure full coverage of the area. It included bilateral angular and superior temporal gyrus, posterior to y = −40 (1975

voxels), a region encompassing activation peaks from several studies of Theory-of-Mind (Decety and Jackson, 2006; Saxe and Powell, 2006). All masks can be obtained from http://www.rnl.caltech.edu/resources/index.html.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Matthew Rabin was an earlier collaborator on this project, which benefited greatly from his insight. This research was supported by NSF-IGERT (B.B.), NSF-Economics (A.R.), NSF-DRMS (A.R.), the Gordon and Betty Moore Foundation (A.R., C.A.H.), and the Lipper Foundation (A.R.). We thank John Clithero and Anita Tusche for helpful comments.

References

Andreoni J. Impure altruism and donations to public-goods - a theory of warm-glow giving. Econ J. 1990; 100:464–477.

Andreoni J, Bernheim B. Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. Econometrica. 2009; 77:1607–1636.

Bartra O, McGuire JT, Kable JW. The valuation system: A coordinate-based meta-analysis of bold fmri experiments examining neural correlates of subjective value. NeuroImage. 2013; 76:412–427. [PubMed: 23507394]

Basten U, Biele G, Heekeren HR, Fiebach CJ. How the brain integrates costs and benefits during decision making. Proc Natl Acad Sci USA. 2010; 107:21767–21772. [PubMed: 21118983]

Batson, CD. Altruism in humans. Oxford University Press; 2011.

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF. Associative learning of social value. Nature. 2008; 456:245–249. [PubMed: 19005555]

Bernheim, BD.; Rangel, A. Behavioral public economics: Welfare and policy analysis with non-standard decision-makers. National Bureau of Economic Research; 2005.

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev. 2006; 113:700. [PubMed: 17014301]

Bogacz R, Wagenmakers EJ, Forstmann BU, Nieuwenhuis S. The neural basis of the speed-ìaccuracy tradeoff. Trends Neurosci. 2010; 33:10–16. [PubMed: 19819033]

Boorman ED, O'Doherty JP, Adolphs R, Rangel A. The behavioral and neural mechanisms underlying the tracking of expertise. Neuron. 2013; 80:1558–1571. [PubMed: 24360551]

Bruneau EG, Pluta A, Saxe R. Distinct roles of the 'shared pain' and 'theory of mind' networks in processing others' emotional suffering. Neuropsychologia. 2012; 50:219–231. [PubMed: 22154962]

Busemeyer JR, Townsend JT. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. Psychol Rev. 1993; 100:432. [PubMed: 8356185]

Carter RM, Huettel SA. A nexus model of the temporal–parietal junction. Trends Cogn Sci. 2013; 17:328–336. [PubMed: 23790322]

Charness G, Rabin M. Understanding social preferences with simple tests. Q J Econ. 2002; 117:817–869.

Churchland AK, Kiani R, Shadlen MN. Decision-making with multiple alternatives. Nat Neurosci. 2008; 11:693–702. [PubMed: 18488024]

Cisek P, Puskas GA, El-Murr S. Decisions in changing conditions: The urgency-gating model. J Neurosci. 2009; 29:11560–11571. [PubMed: 19759303]

Clithero JA, Rangel A. Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neurosci. 2013; 9:1289–1302. [PubMed: 23887811]

Cornelissen G, Dewitte S, Warlop L. Are social value orientations expressed automatically? Decision making in the dictator game. Pers Soc Psychol B. 2011; 37:1080–1090.

De Vignemont F, Singer T. The empathic brain: How, when and why? Trends Cogn Sci. 2006; 10:435–441. [PubMed: 16949331]

Decety J, Jackson PL. A social-neuroscience perspective on empathy. Curr Dir Psychol Sci. 2006; 15:54–58.

Dufwenberg M, Kirchsteiger G. A theory of sequential reciprocity. Games Econ Behav. 2004; 47:268–298.

Eckel CC, Grossman PJ. Altruism in anonymous dictator games. Games Econ Behav. 1996; 16:181–191.

Engel C. Dictator games: A meta study. Exp Econ. 2011; 14:583–610.

Engelmann D, Strobel M. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. Am Econ Rev. 2004; 94:857–869.

Falk A, Fischbacher U. A theory of reciprocity. Games Econ Behav. 2006; 54:293–315.

Fehr E, Fischbacher U. Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. Econ J. 2002; 112:C1–C33.

Fehr E, Fischbacher U. The nature of human altruism. Nature. 2003; 425:785–791. [PubMed: 14574401]

Fehr E, Schmidt K. A theory of fairness, competition, and cooperation. Q J Econ. 1999; 114:817–868.

Gold JI, Shadlen MN. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. Neuron. 2002; 36:299–308. [PubMed: 12383783]

Gold JI, Shadlen MN. The neural basis of decision making. Annu Rev Neurosci. 2007; 30:535–574. [PubMed: 17600525]

Hare T, Camerer C, Knoepfle D, O'Doherty J, Rangel A. Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition: Soms. J Neurosci. 2010; 30:583. [PubMed: 20071521]

Hare TA, Camerer CF, Rangel A. Self-control in decision-making involves modulation of the vmpfc valuation system. Science. 2009; 324:646–648. [PubMed: 19407204]

Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A. Transformation of stimulus value signals into motor commands during simple choice. Proc Natl Acad Sci USA. 2011; 108:18120–18125. [PubMed: 22006321]

Heekeren HR, Marrett S, Ungerleider LG. The neural systems that mediate human perceptual decision making. Nat Rev Neurosci. 2008; 9:467–479. [PubMed: 18464792]

Hunt LT, Kolling N, Soltani A, Woolrich MW, Rushworth MF, Behrens TE. Mechanisms underlying cortical activity during value-guided choice. Nat Neurosci. 2012; 15:470–476. S471–473. [PubMed: 22231429]

Jackson PL, Meltzoff AN, Decety J. How do we perceive the pain of others? A window into the neural processes involved in empathy. NeuroImage. 2005; 24:771–779. [PubMed: 15652312]

Kable JW, Glimcher PW. The neural correlates of subjective value during intertemporal choice. Nat Neurosci. 2007; 10:1625–1633. [PubMed: 17982449]

Krajbich I, Armel C, Rangel A. Visual fixations and the computation and comparison of value in simple choice. Nat Neurosci. 2010; 13:1292–1298. [PubMed: 20835253]

Kurth F, Zilles K, Fox PT, Laird AR, Eickhoff SB. A link between the systems: Functional differentiation and integration within the human insula revealed by meta-analysis. Brain Struct Func. 2010; 214:519–534.

Lim SL, O'Doherty JP, Rangel A. Stimulus value signals in ventromedial PFC reflect the integration of attribute value signals computed in fusiform gyrus and posterior superior temporal gyrus. J Neurosci. 2013; 33:8729–8741. [PubMed: 23678116]

McClure SM, Laibson DI, Loewenstein G, Cohen JD. Separate neural systems value immediate and delayed monetary rewards. Science. 2004; 306:503–507. [PubMed: 15486304]

Milosavljevic M, Malmaud J, Huth A, Koch C, Rangel A. The drift diffusion model can account for value-based choice response times under high and low time pressure. Judgm Decis Mak. 2010; 5:437–449.

Moll J, Krueger F, Zahn R, Pardini M, de Oliveira-Souza R, Grafman J. Human fronto–mesolimbic networks guide decisions about charitable donation. Proc Natl Acad Sci USA. 2006; 103:15623–15628. [PubMed: 17030808]

Morishima Y, Schunk D, Bruhin A, Ruff CC, Fehr E. Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. Neuron. 2012; 75:73–79. [PubMed: 22794262]

Nowak MA, Sigmund K. Evolution of indirect reciprocity by image scoring. Nature. 1998; 393:573–577. [PubMed: 9634232]

Peters J, Büchel C. The neural mechanisms of inter-temporal decision-making: Understanding variability. Trends Cogn Sci. 2011; 15:227–239. [PubMed: 21497544]

Piovesan M, Wengstrom E. Fast or fair? A study of response times. Econ Lett. 2009; 105:193–196.

Rand DG, Greene JD, Nowak MA. Spontaneous giving and calculated greed. Nature. 2012; 489:427–430. [PubMed: 22996558]

Rangel, A.; Clithero, J. The computation of stimulus values in simple choice. In: Glimcher, PW., editor. Neuroeconomics: Decision making and the brain. Academic Press; 2013. p. 125-147.

Ratcliff R, McKoon G. The diffusion decision model: Theory and data for two-choice decision tasks. Neural Comput. 2008; 20:873–922. [PubMed: 18085991]

Rodriguez CA, Turner BM, McClure SM. Intertemporal choice as discounted value accumulation. Plos One. 2014; 9:e90138. [PubMed: 24587243]

Ruff CC, Ugazio G, Fehr E. Changing social norm compliance with noninvasive brain stimulation. Science. 2013; 342:482–484. [PubMed: 24091703]

Saxe R, Powell LJ. It's the thought that counts: Specific brain regions for one component of theory of mind. Psychol Sci. 2006; 17:692–699. [PubMed: 16913952]

Scholz J, Triantafyllou C, Whitfield-Gabrieli S, Brown E, Saxe R. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. Plos One. 2009; 4:4869.

Singer T. The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. Neurosci Biobehav R. 2006; 30:855–863.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, Frith CD. Empathy for pain involves the affective but not sensory components of pain. Science. 2004; 303:1157–1162. [PubMed: 14976305]

Smith PL, Ratcliff R. Psychology and neurobiology of simple decisions. Trends Neurosci. 2004; 27:161–168. [PubMed: 15036882]

Strombach T, Weber B, Hangebrauk Z, Kenning P, Karipidis II, Tobler PN, Kalenscher T. Social discounting involves modulation of neural value signals by temporoparietal junction. Proc Natl Acad Sci USA. 2015; 112:1619–1624. [PubMed: 25605887]

Tinghog G, Andersson D, Bonn C, Bottiger H, Josephson C, Lundgren G, Vastfjall D, Kirchler M, Johannesson M. Intuition and cooperation reconsidered. Nature. 2013; 498:E1–2. [PubMed: 23739429]

Tom SM, Fox CR, Trepel C, Poldrack RA. The neural basis of loss aversion in decision-making under risk. Science. 2007; 315:515–518. [PubMed: 17255512]

van den Bos W, McClure SM. Towards a general model of temporal discounting. J Exp Anal Behav. 2013; 99:58–73. [PubMed: 23344988]

Waytz A, Zaki J, Mitchell JP. Response of dorsomedial prefrontal cortex predicts altruistic behavior. J Neurosci. 2012; 32:7646–7650. [PubMed: 22649243]

Zaki J, Mitchell JP. Equitable decision making is associated with neural markers of intrinsic value. Proc Natl Acad Sci USA. 2011; 108:19761–19766. [PubMed: 22106300]

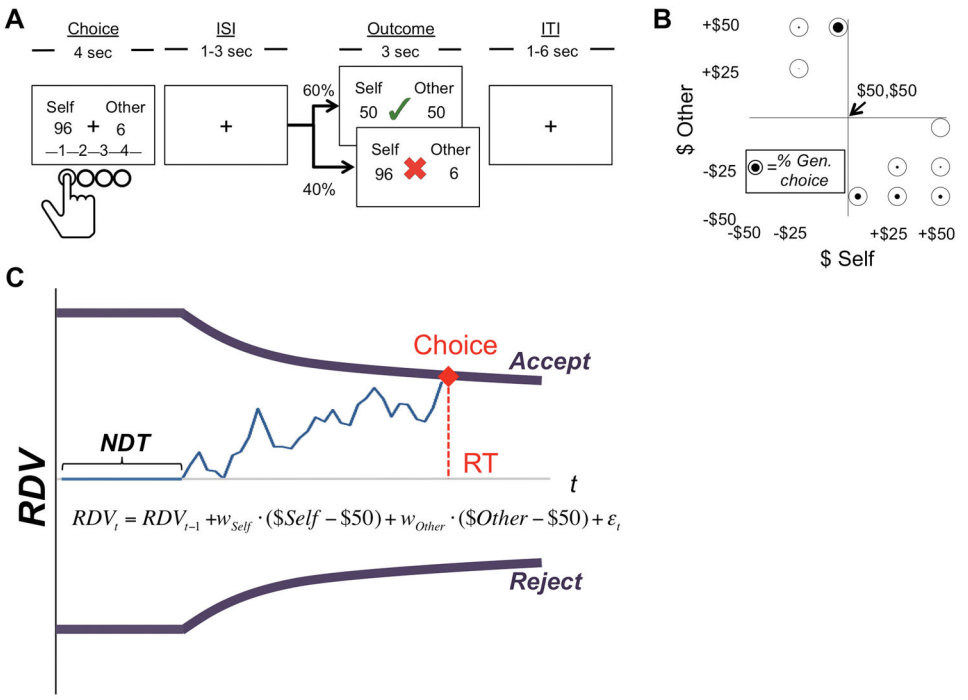

Figure 1.

Task design and model. A) The task consisted of a decision phase, in which subjects chose whether to accept the proposed payment-pair or a default of $50 to both individuals, and a subsequent outcome phase, in which subjects discovered if their choices had been implemented (60% of trials), or reversed, resulting in the non-chosen option (40% of trials). B) Proposed transfers used in the experiment describing $*Self* and $*Other*. The alternative was always a transfer of $50 to both subjects. X- and Y-axes represent distance from default offer. The filled area in each transfer is proportional to the percentage of pro-social choices across all subjects. C) In the DDM model, choices are made through the noisy accumulation of a relative value signal (RDV), based on a weighted sum of the amounts $*Self* and $*Other* available on each trial. A response occurs when this accumulated value signal crosses a threshold, with an RT equal to the total accumulated time + a non-decision time (NDT) to account for sensory and motor-related processes unrelated to the comparison process itself.

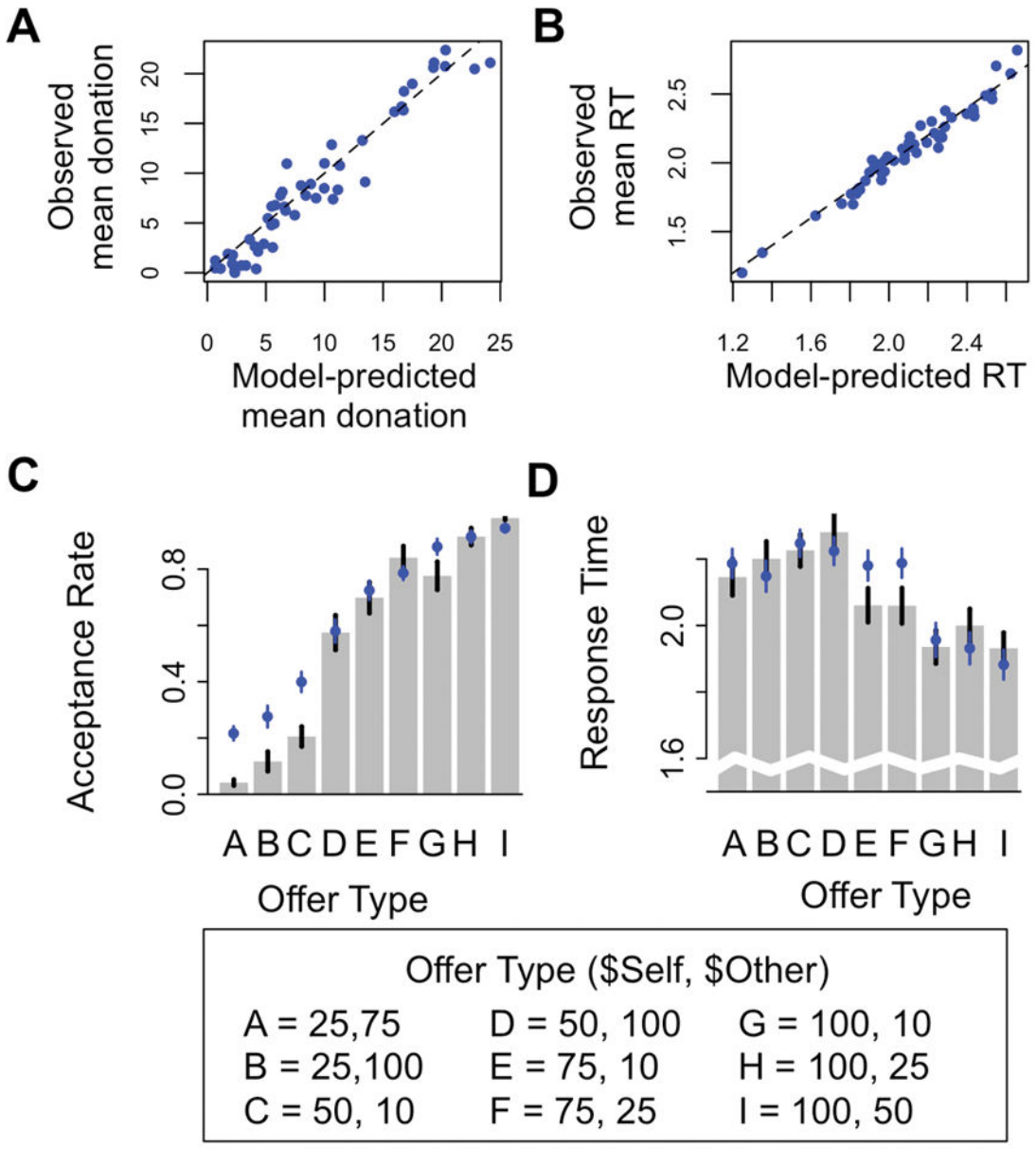

Figure 2.

Model fits to behavior. A) Model-predicted vs. observed average generosity across subjects. Dashed 45 degree line represents a perfect match. B) Model-predicted vs. observed overall response time (RT). C) Within-subject acceptance likelihood (mean ± SEM) and D) RT (mean ± SEM) for each of the 9 proposal-types. Observed behavior: grey bars. Predict behavior: blue circles.

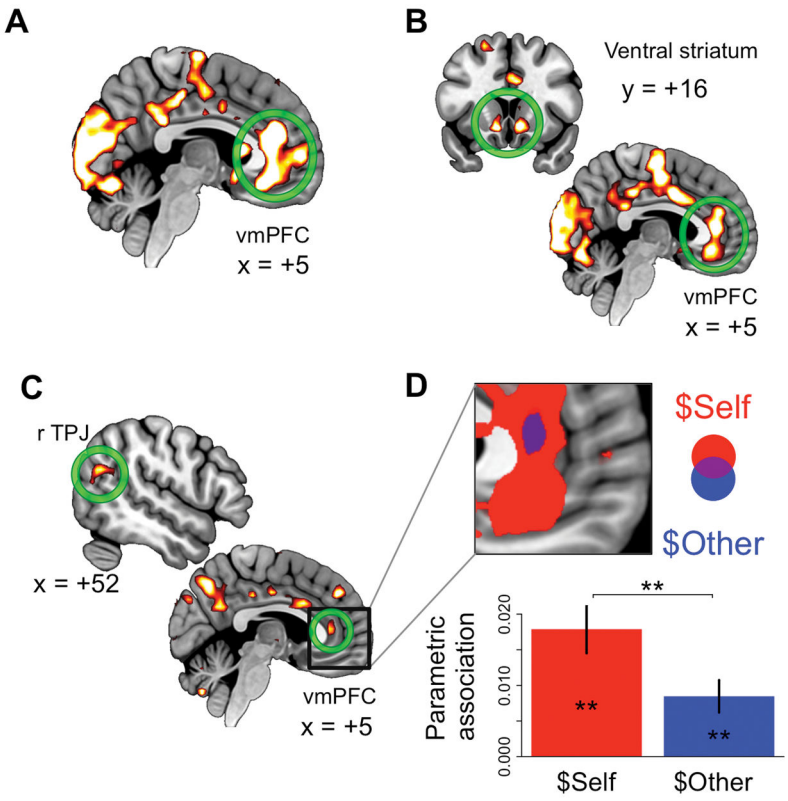

Figure 3.

Neural responses vary parametrically with A) behavioral preference at the time of choice; B) $*Self* on each trial; and C) $*Other* on each trial. D) Conjunction of $*Self* and $*Other* in vmPFC. Bar plot (mean ± SEM) shown for illustrative purposes only. Activations displayed at P < .001, uncorrected.

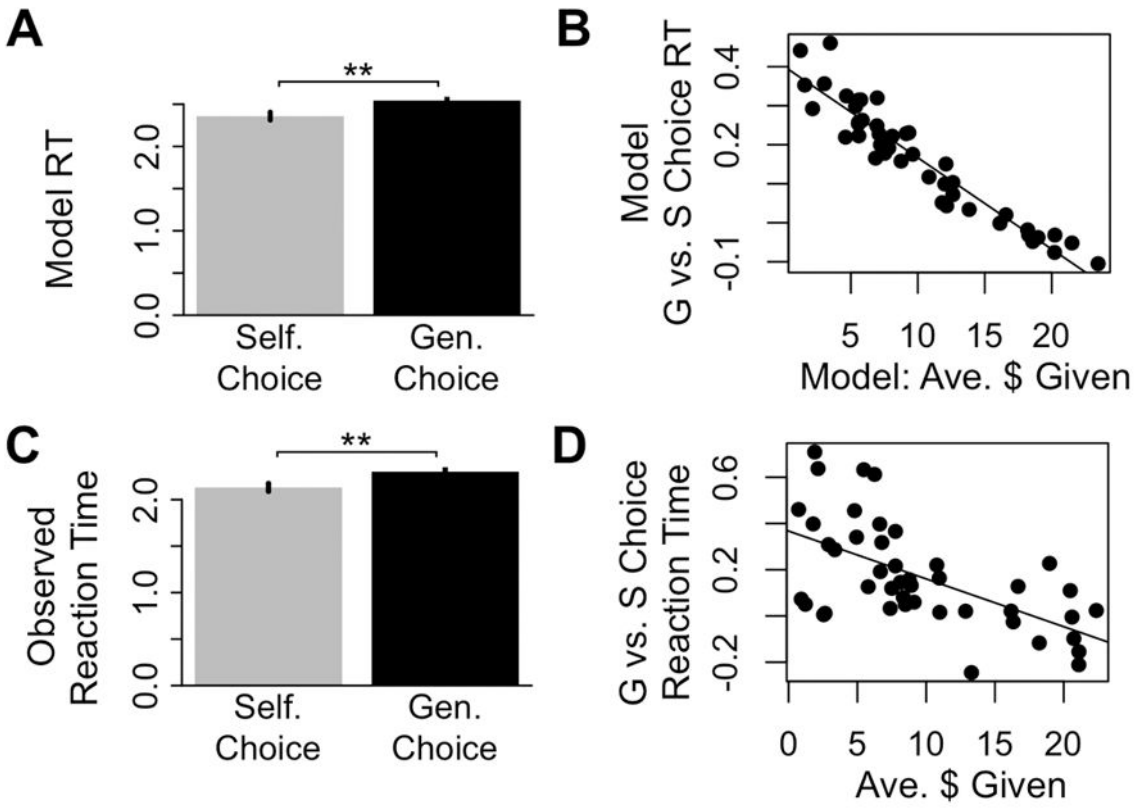

Figure 4.

Model implications for RT differences. A) At the fitted parameters, the model predicts generous (G) choices should take longer than selfish (S) choices. B) The model predicts that overall generosity correlates negatively with RT differences on G vs. S choices. C) Observed RTs for G and S choices. D) Observed relationship between average generosity and G vs. S choice RT differences. Bars show mean RT ± SEM. ** P < .001

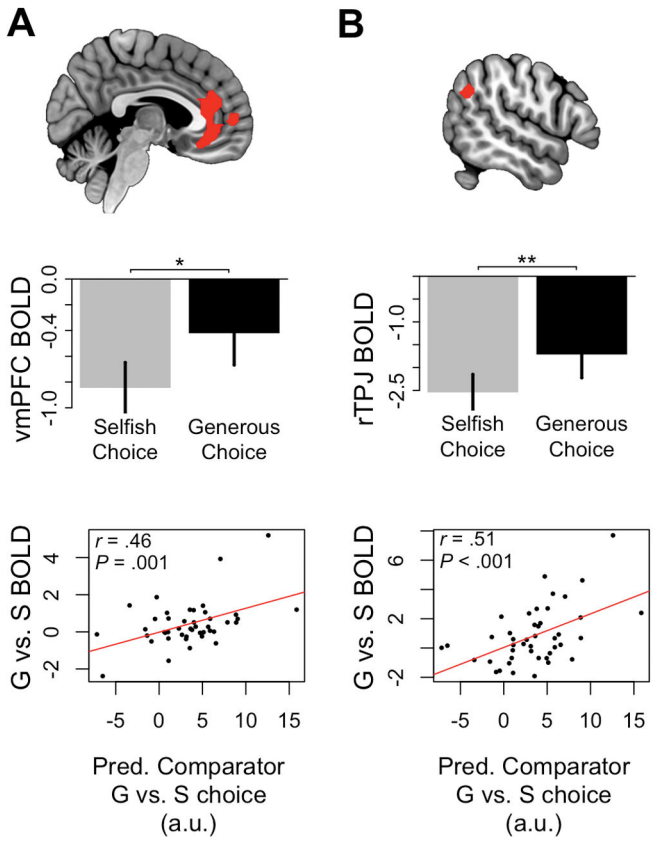

Figure 5.

Model implications for BOLD response (mean ± SEM) during generous (G) vs. selfish (S) choices. Independently defined regions in value-related vmPFC (A) and generosity-related TPJ (B) both show higher response on G choice trials. B) Individual variation in G vs. S choice BOLD response is accounted for by model-predicted comparator differences in both regions. **P*=.02; ***P*=.008.

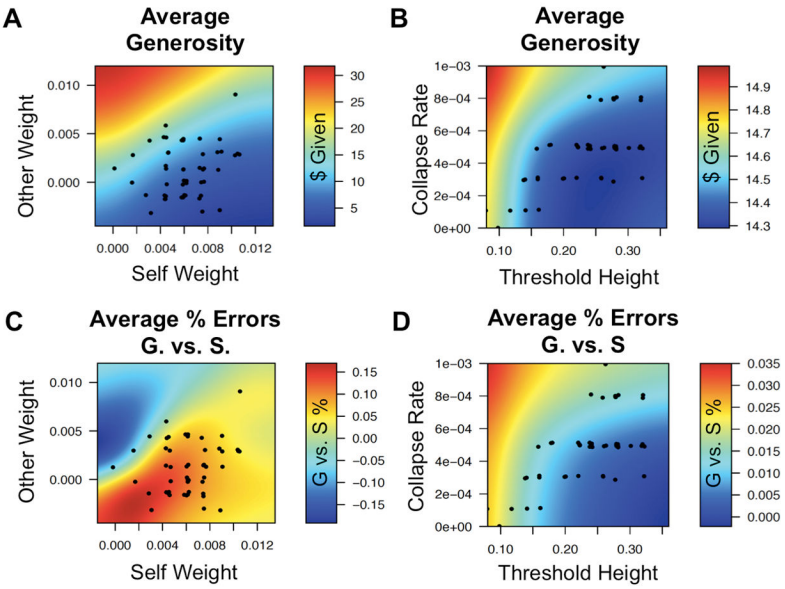

Figure 6.

Model implications for relation between different parameters of the model and behavior. A) Variation in generosity as weights for self and other vary, and B) as threshold starting height and collapse rate vary. C) Variation in likelihood that a generous choice is a mistake as a function of weights for self and other and D) threshold parameters. Dots represent the estimated parameter values for the 51 subjects who completed the fMRI study, jittered randomly by a small amount to allow visualization of subjects with overlapping values.

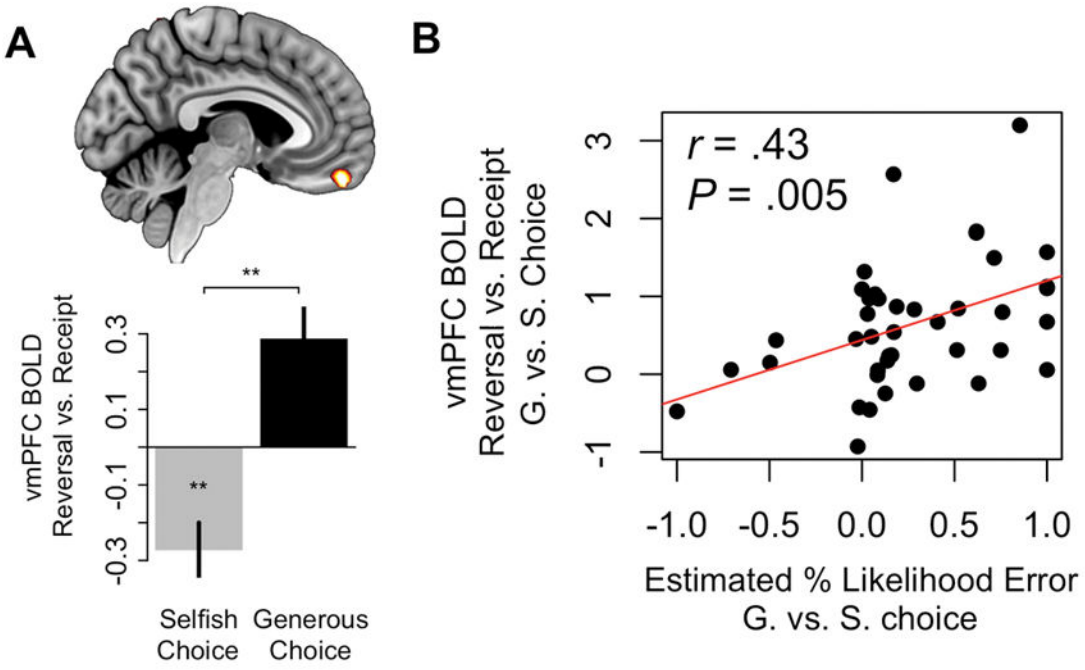

Figure 7.

Model implications for the likelihood that selfish (S) or generous (G) choices are errors. A) A vmPFC region implicated in coding outcome value responded more positively to reversal vs. receipt of G choices compared to S choices ($P < .05$, SVC). Differential BOLD response in this region (mean ± SEM) is shown for illustrative purposes only. B) vmPFC response to reversal vs. receipt of G vs. S choices correlated with the DDM-predicted likelihood that a subject's G choices were more likely to be errors than S choices (i.e. indexing the relief they should feel if those choices are overturned).

Table 1

Parameters of the best-fitting DDM for each subject. w_{Self} and w_{Other} represent weights applied to the value of *$Self* and *$Other* on each trial compared to the default. *NDT*: non-decision time. *b* and *d*: starting value and collapse rate of the decision threshold.

Parameter	Mean	SD	Min	Max
w_{Self}	.006	.002	0	.0105
w_{Other}	.001	.0026	−.003	.009
NDT	868ms	241ms	300ms	1,300ms
b	.23	.065	.08	.32
d	.00046	.00022	0	.001