**Response to Reviewer 1:**

We would like to thank the reviewer for his/her detailed comments, which we have found extremely useful in revising and improving the quality of the paper.

For the reviewer’s convenience, we have copied his/her original comments in bold-type face.

**Comment R1-1: This is a thought-provoking paper that attempts to connect altruistic behavior with drift diffusion models that have been widely applied to perceptual decision-making (and increasingly value-based decision-making as well). I enjoyed reading this paper, although I felt it has a review paper quality to it in that its main contribution was an integration and reframing of previous results, rather than new results per se. My main comments are below:**  
Response**:** We thank the reviewer for the helpful and detailed comments.

We have implemented multiple improvements to address the reviewers’ comment, including the collection of new behavioral data, additional new analyses, and the introduction of additional results that we had left out from the first version. We have also rewritten extensive sections of the paper to improve its clarity on a number of areas.

**Comment R1-2: 1. Neuro- vs. behavioral focus. Mostly the paper is about the model and its behavioral implications; the actual neuro part is fairly minimal (and not particularly novel, as it mostly involves confirming what past research has suggested). The fMRI analysis primarily serves as a validation of their behavioral model; they use it to confirm that the model is capturing relevant components of decision-making, based on the implication of similar regions in similar computations in past work. (The reversal analysis, providing insight into erroneous generosity, is the one part of the fMRI portion that is more novel.) This raises some question as to whether the most appropriate home for the paper is a neuroscience journal.**

Response**:** We hope to convince the reviewer that the link of the computational model to the neural data is more novel and interesting than suggested in the comment.

There are several reasons for this.

First, despite some previous evidence that TPJ responses are important for altruistic behavior, the precise computational role of the TPJ in altruistic choice, and how it interacts with the basic decision circuitry, remains unclear. This is the first paper to show that a neurocomputational model of altruistic choice successfully captures several key patterns of response, not only in choice behavior, but also in neural response.

Second, we have added a new set of neural results that had been omitted in the previous draft. Specifically, we test the model prediction that neural regions whose activity scales with the comparison and accumulation process should show greater activation on trials that resulted in a generous compared to a selfish choice. We replicate previous results showing that both the vmPFC and TPJ display this pattern. However, our model suggests a novel and fundamentally different neural interpretation from what has been previously argued: activation in these regions does NOT indicate that choosing generously requires inhibition of selfish tendencies (as proposed by Strombach et al., 2015), or that it is inherently rewarding (as proposed by Zaki et al., 2011, PNAS and Strombach et al., 2015, PNAS). Rather, the model implies that greater generosity-related response may be a by-product of how the decision-circuitry makes choices. We provide additional support for this alternative interpretation, showing that differences in BOLD response in vmPFC and TPJ when choosing generously correlate with and are fully accounted for by model predictions (see new Implication 2 in the revised paper, pg. 12-13). We believe that this is an important point, and that it has ramifications for the interpretation of neural response not just during social decision making, but in many areas of cognitive neuroscience. The new draft of the paper includes new sections to discuss this result and its implications.

To provide further evidence for this important point, we have also conducted several other supplementary analyses. We describe these analyses here, but in the interest of keeping the paper at manageable length, we do not currently report these analyses in the paper. We could of course include them if the editor or reviewers deemed it appropriate.

For example, a hypothesis that has been advanced in the literature is that TPJ response represents a special modulatory influence on vmPFC calculations, and that this modulatory influence is highest on trials when the TPJ is highly active (Morishima et al., 2014; Strombach et al., 2015). In contrast, our neurocomputational model suggests that TPJ computes a quantity related to another’s welfare, and that this represents an input into the integration and comparison process that is weighted equally and consistently across all trials, regardless of the final choice. Our model does suggest that TPJ representations could be more extensively recruited by the integration and comparison process on trials that result in a generous choice (because computations on these choices take longer to cross the choice-determining threshold). However, this recruitment should not specifically change the weight attached to welfare considerations. To test these competing hypotheses, we have estimated DDM parameters separately for the half of trials with highest TPJ activation compared to the half of trials with lowest TPJ activation. The “modulatory hypothesis” predicts that parameters of the DDM related to weights on self and other should change as a function of higher TPJ response (i.e., wself should be lower, or wother should be higher). In contrast, our neurocomputational model predicts that TPJ responses represent an input to vmPFC that is integrated in a consistent way across all trials, and thus that there should be no difference in any parameters as a function of high vs. low TPJ response. Results strongly support the latter hypothesis. Estimating DDM parameters separately based on a median split in right TPJ response (either in the area sensitive to $Other at P < .001, uncorrected, or in an area that is specifically more active during generous vs. selfish trials at P < .005, uncorrected) yields nearly identical estimates for all parameters of the model, regardless of whether TPJ response is high or low (all Ps > .2). Note also that this is not a function of our inability to detect differences in parameters as a function of activation within a region: we observed significant differences in the estimated threshold as a function of high vs. low dorsal ACC response (an area that is more active on generous choice trials). Such differences could reflect the use of conflict-related computations to dynamically adjust response caution. Collectively, these results suggest that information in the TPJ may represent a direct input into the vmPFC value integration processes rather than a modulatory influence.

Third, as highlighted by the reviewer, this is the first paper to combine modeling and neural data to test the hypothesis that a significant fraction of generous choices are mistakes, and that their prevalence is negatively correlated with average levels of observed generosity.

Last, but not least, the connection between computational modeling and neural data has proven critical to make progress in many areas of neuroscience, from vision to motor control and simple decision-making. In fact, these types of models are considered so relevant for neuroscience journals that they are frequently published in the field’s best journals. This study contributes to this tradition, by providing the first such model of altruistic choice and linking it to neural data.

**Comment R1-3: 2. Correlation between generosity and RT: Perhaps the most novel finding in the study is the positive correlation between generosity and RT. This, as the authors note, goes against a growing body of research observing the opposite relationship. Again, it's unclear what the role of neuroimaging here beyond response time data.  
  
Another possible issue is that subjects in the current study were obliged to respond within 4 s. Given the role of RT in the model and many of the results, it would be helpful if the authors discussed the extent to which this deadline may have imposed time pressure on participants and the extent to which their effects should be expected to generalize to unconstrained response windows.**

Response**:** Several comments.

First, as we alluded to above, we have now included a section in which we identify neural concomitants of the model-predicted differences in activation on generous vs. selfish trials. We believe that this neural data plays a complementary and valuable role over and above the RT data, because it provides an alternative explanation to some commonly observed patterns of neural response in the literature.

Second, with respect to the reviewer’s second point in this section, we selected the 4 second response window based on a pilot study (N = 22) in which participants were allowed to respond freely, and which also included a wider range of proposals. In that study, we observed that the average response time was 2.55 seconds, with only 13.8% of responses taking longer than 4 seconds. Notably, in that study, average generosity (11.5% generous choices) was also slightly lower than in the observed study, exactly as would be predicted by the DDM.

However, the same basic patterns predicted by the DDM held even under conditions of free response. Generous choices took longer to make than selfish choices (P = .06, two-tailed), and there was a negative (though slightly weaker) correlation across individuals between reaction time differences and generosity (r = -.23). We thus would expect the implications of the DDM to extend to free response times, but with the assumption that this should result in a higher threshold for responding, longer average response times, and less error-prone choices.

As suggested by the reviewer, we have added a short discussion of this issue in the current paper, in the Methods section on pg. 22 “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.”

**Comment R1-4: 3. Role of vmPFC. The authors report that "vmPFC responses encode an integrated value signal at decision", but as the table shows, there are actually ~14 regions (not counting subregions) that emerge from this analysis. Is there a reason to think that vmPFC is doing something different from these other regions? Additionally, the values for $Self and $Other are both associated with value in (an overlapping region of) vmPFC. The authors suggest that this supports the idea that vmPFC "may represent an area where separate attributes are combined into an integrated value signal for the proposal" (pg. 11), but the data show they are each represented independently AND represented overall there. Then, they also report that vmPFC response is associated with choice reversals (vs. implementations), especially for generous choices. Is there an account of vmPFC that can encompass its involvement in all of these?**

Response**:** Several comments with respect to this important comment.

First, several studies and meta-analyses have shown that, although the vmPFC as a whole plays an important role in the valuation of stimuli, there are very different types of value signal encoded there, and that different sub-regions might represent particular subsets of these signals (e.g., see Clithero & Rangel, 2014; Bartra et al., 2013; McNamee et al. 2013)

More importantly, as has been demonstrated in multiple studies in the literature, in many paradigms there are many signals that are not directly related to decision values computed at the time of choice, but still highly correlated with them. A typical example includes responses in visual processing regions in right and left occipital cortex, which are likely to reflect modulations in visual attention. In this paper, we focus on vmPFC because a wealth of data points to the special role of the vmPFC in the computation of decision values at the time of choice across a wide class of paradigms (Rangel, 2008; Clithero & Rangel, 2014; Bartra et al., 2013).

Second, a key hypothesis of the model is that there should be an overall value signal that reflects the integration of the $Self and $Other attributes. Note that, although we find many regions that show an association with decision value/preference, there are only a few that show a triple conjunction between decision value, $Self, and $Other. These include visual processing regions in right and left occipital cortex, cuneus, and cerebellum, as well as the vmPFC.

Third, the area of the vmPFC where we observe an overlap in decision value as well as individual attribute signals is somewhat different from the area of the vmPFC that we observe correlating with outcome value. This outcome-value region is more anterior. A recent meta-analysis supports this posterior-anterior distinction, suggesting that the more anterior region of the vmPFC that we find correlating with reversal signals is also the region that is more clearly linked to processing of reward outcome compared to decision value (Clithero & Rangel, 2014). We therefore think it premature to try to generate a computational account of vmPFC that encompasses ALL of these different functions within a single region. We make this point clearer on page 27 of the current manuscript. We also now use an outcome-value meta-analysis mask for small-volume correction of outcome-value signals, in the same way as we used the decision-value meta-analysis mask. This analysis confirms that there is an area in the anterior vmPFC that is significantly more active on generous choice reversals compared to selfish choice reversals at P < .02, small-volume corrected. We think that this new method is an improvement to the paper because it represents a more consistent approach to statistical thresholds and a more robust approach to identifying outcome-value related signals, while making it clearer to the reader why we think that decision-value and outcome-value signals may be somewhat different quantities psychologically and neurally.

**Comment R1-5: 4. TPJ interpretation. The authors suggest that the two main theories of TPJ function (belief representation and attention) cannot account for the observation in the current study that TPJ response correlates with $Other. Can the authors offer a way to reconcile their findings with the (abundant) findings that support these theories? Currently they simply suggest that the processes involved in their task may "represent an additional computation supported by this area", which is acceptable but not especially satisfying.**

Response**:** This is a very fair and important point. To be clear, we did not intend to argue (nor do we believe) that the extensive literature on TPJ involvement in belief representation and attention are incorrect.

To address the reviewer’s suggestion, we now offer some speculation about a computational account that might reconcile these different findings on pg. 17 “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.”

**Response to Reviewer #2:**

We would like to thank the reviewer for his/her detailed comments, which we have found extremely useful in revising and improving the quality of the paper.

For the reviewer’s convenience, we have copied his/her original comments in bold-type face.

**Comment R2-1: I have enjoyed reading the manuscript from Hutcherson and colleagues entitled "A neurocomputational model of altruistic choice and its implications". I found the approach to the problem original and the conclusions drawn from the model and data analysis interesting and in some aspects unexpected. I think this manuscript gives fresh insights into computational mechanisms that generate altruistic behaviors. However, I do have a few issues that I would like to be clarified by the authors:**  
Response**:** We thank the reviewer for his supportive comments.

We have implemented multiple improvements to address the reviewers’ comment, including the collection of new behavioral data, additional analyses, the introduction of additional results that we had left out from the first version. We have also rewritten extensive sections of the paper to improve its clarity in a number of areas.

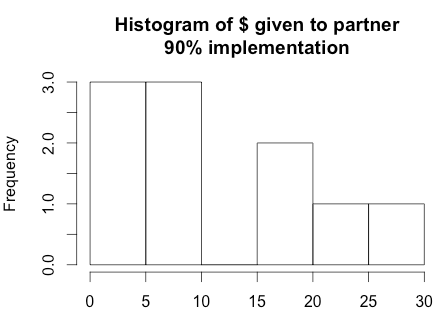
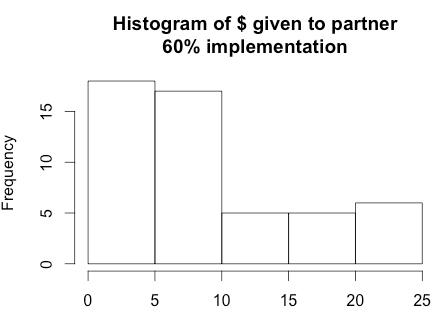
**Comment R2-2: Weak link between choices and outcomes: I am a bit worried that the probabilistic link devised by the authors (60% of trials implemented - 40% reversal) had an impact on the subjects' behaviors, significantly reducing their altruistic behavior. In fact, an important drive for altruistic behavior might be social norms (acting selfishly is usually not socially accepted). However if the causal link between choice and outcome is so weak, it might be that participants are encouraged to act more selfishly, and the altruistic acts are driven by stochastic behavior where the role of error might be more relevant (as shown later in the manuscript). It will be important to check that this is not the case, and that these findings are generalizable to real world situations in which choices are often more strongly linked to outcomes. A small behavioral study in which choice produce outcomes in a deterministic fashion might be used to address this issue.**

Response**:** Although we explicitly instructed participants not to choose something they did not want as a way to game the reversal process, we agree with the reviewer that the weaker link between choices and outcomes is a potential worry.

To address this point, we ran a small additional study of 10 subjects, in which choices were implemented with 90% probability. We chose the 90% parameter based on two principles: 1) we wished to assess whether this might have changed participants’ responses, and 2) we also wished to retain the “plausible choice anonymity” of the current design to test internal motivations for altruism, rather than external motivations related to reputation management which would appear if there were no reversals. The 90% probability of implementation creates a tighter link between choice and outcome, but retains plausible choice anonymity.

Despite the difference in implementation probability, choice behavior looked almost identical to that in the current study.

1. Average percentages of generous behavior were roughly comparable (27% in the pilot study, 21% in the study reported in the paper).
2. Distributions of average generosity in dollar amounts were similar:

1. Response time differences between generous and selfish behavior were nearly identical (mean difference in 90% implementation = 198ms, mean difference in 60% implementation = 135ms).
2. The negative correlation between RTs and generosity emerged just as strongly (r = -.87 in 90% implementation, r = -.64 in 60% implementation)

For these reasons, we think that the behavior we observed is unlikely to be due simply to people worrying that their choices were rarely implemented.

We have included a short statement on pg. 22 of the current manuscript to indicate that behavior is not substantially affected by the high reversal rate: “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.”

However, we also agree wholeheartedly with the reviewer that social reputation is likely an important part of altruism, and hope in future studies to address exactly this issue. We think that the computational models we have begun to develop will be uniquely suited to study which aspects of computation change as a function of manipulations that influence observed levels of giving.

**Comment R2-3: A related point - was the other participant (the receiver) informed of the actual choice or were they just shown the outcome? And if they were informed of the choice, did the actual participant know this? I would recommend appending the instructions given to participants.**

Response**:** As we stated in the original manuscript, passive participants were aware of the probabilistic reversal, and active participants were explicitly informed about their partner’s knowledge: “APs were informed that the PPs were aware of the probabilistic implementation” (pg. 22 of revised manuscript).

Also, as requested, we have appended the full set of instructions given to participants in the revised supplementary materials.

**Comment R2-4: Asymmetric design: From figure 1b and the panel in fig c, it seems that the matrix of gambles offered were not symmetric, with significantly more 'selfish' offers. Could the authors explain the rationale of this asymmetric design? Could this imbalance have had an effect on the parameters estimation? Or even affect the subjects' behavior?(If generous choices are rarer it might be more acceptable to act selfishly)**

Response**:** We selected the matrix of payoffs based on a pilot study in which it was observed that these payoffs evoked reasonably high levels of generosity. However, we have data from a follow-up study where we included a much wider range of payoffs, symmetrically distributed around the 50/50 point. In this follow-up study we also more frequently presented a subset of offers adaptively selected for each subject to encourage a minimum 30% of generous choices. We find very little evidence that the asymmetric distribution of choices, or the raw frequency of generosity, substantially influences behavior. For example, model fitting using the same parameters as those used in the current paper is capable of capturing behavior under more balanced distributions of proposals just as closely as in the current study, and with similar average parameters. We also observe similar effects in RT. We therefore think it unlikely that the specific proposals used in the current study strongly affected the observed behavior or range of applicable parameters observed in the study.

In the interest of keeping the paper length as close as possible to Neuron guidelines, we have opted not to include a discussion of this concern in the current manuscript. However, if the editor prefers, we could add in a short description of the evidence suggesting that this is not a concern in the supplementary materials.

**Comment R2-5: Orthogonalisation of the regressors in the GLMs: When the authors say that regressors were orthogonalised, are they referring to the default option in SPM5 that uses serial (Gram-Schmidt) orthogonalisation of successive parametric modulators? It is important to highlight that this default has been now changed in SPM12. The problem is that serial orthogonalization assumes a natural hierarchy to which one can arbitrarily assign shared variance. By orthogonalising the second regressor with respect to the first, one essentially forces the first regressor to explain the shared variance. This way of dealing with shared variance makes sense only if there is good reason to believe that the first regressor accounts for the shared variance. This is a rare situation that needs to be justified very carefully and this does not seem the case in this study (e.g. why does the $others regressor need to be orthogonalised with respect to $self and not the other way around?). This might have affected the result of the study in particular in vmPFC. A more conservative approach would be to omit serial orthogonalisation, making the two regressors compete for shared variance (an alternative and more robust approach to dealing with shared variance is reducing it by design). If this is not possible, it would be better to acknowledge how much variance is shared - can the authors clarify how much the $self and $others regressors are correlated?**

Response**:** To address the reviewer’s comment, we have replaced the results originally reported in the manuscript with a set of results corresponding to an unorthogonalized model.

As can be seen, for most regions, this does not affect the basic conclusions, although it changes significance values slightly. In addition, the ventral striatum, which before correlated with $Self at sub-threshold levels, now correlates significantly and specifically for $Self at a whole-brain level. There are two regions where our original conclusions are tempered somewhat. One is in the TPJ. Although this region continues to correlate with $Other at whole-brain-corrected thresholds, the specificity of this response when compared to $Self becomes somewhat weaker (P < .005, uncorrected). However, TPJ does *NOT* correlate with $Self significantly on its own. Given these results, we think it most conservative to temper our conclusions about representations in TPJ: while it clearly codes for $Other, there may be small but intermingled populations of neurons coding for $Self or related quantities as well. These populations, while failing to produce a significant correlation on their own, may subtly reduce the specificity of $Other coding in TPJ. Note that this observation does nothing to change the main point of the paper, which is that the computational model requires the quantities $Self and $Other to construct an integrated value, and that both of these quantities appear to be represented in the brain in multiple regions. Finally, the only region where our original conclusion changes substantially is in the amygdala, which no longer correlates significantly with $Self at corrected levels. We have modified the paper and figures to reflect these results on pg. 10-11.

**Comment R2-6: Errors: The authors are looking for confirmation of their new hypothesis (i.e. generous acts are driven by errors) using some simulations of parameters in their models. Can the authors test this more directly in their RT data? According to a speed-accuracy tradeoff should we not expect that accepting a generous choice would then be faster? Is this the case? Also errors might produce some degree of auto-correlation in the data. For example one can imagine that after an 'error' (i.e. accepting a generous offer) the subject would try to correct his or her error in the next generous offer, therefore the correlation in a generous choice at time 1 (t) should be inversely correlated with the behaviour at the next generous choice (t+1). The best way to test this would be by designing a study in which time for response is varied by the experimenter.**  
Response**:** The analysis suggested by the reviewer is interesting, but unfortunately underpowered in the current study, which was not optimally designed to assess serial effects of choice (particularly since generous choices are relatively spaced out over the task). A study that varies time pressure would also be illuminating.

Fortunately, we have data from a completed behavioral study that we believe *more directly* addresses the issues/questions raised by the reviewer: 1) how do we know that people sometimes make errors they would have wished to correct, 2) that generous choices are more likely to be errors than selfish choices, and 3) that having an error reversed might be pleasant?

The basic design of the study is exactly like the one used in the scanner: participants see proposal payment pairs that vary in $Self and $Other and have to choose between the proposed and default payment options, with 40% of their choices reversed. Critically however, in this behavioral study participants made choices by moving a computer mouse around on a screen and clicking on their final choice. They were also subsequently explicitly asked to evaluate whether the outcome of their choice (reversed vs. not) made them happy or sad. “Mouse-tracking” techniques have been used in the past to show that people sometimes make errors in perceptual choice (i.e., start moving toward the incorrect option), and that they can use incoming perceptual evidence to subsequently correct those errors before committing to a final choice (Resulaj et al., 2009, *Science*). We used mouse-tracking here for a similar purpose.

On every trial, we asked whether participants initially selected either the proposal or the default option, by assessing the initial direction of movement of the mouse. We then asked whether at any point during the trial, participants showed evidence of a “change of mind”, assessed by asking whether the direction of the mouse reversed course and headed toward the initially non-chosen option. Finally, we developed a computational model (adapted for use with mouse-tracking using techniques described in Resulaj et al., 2009) to estimate the same parameters of the model that we estimated in the fMRI study.

Using this design and these analyses, we can ask whether initial or final choices are likely to be errors, and test the questions raised by the reviewer.

1. Question 1 - How do we know that people sometimes make errors they would have wished to correct?: We can show using analyses of the mouse-tracking data that if the initial movement toward an option (i.e., the preliminary choice) was identified by the computational model as an error, participants are significantly more likely to change their mind later in the trial, that is, to correct their original error (P < .0001).
2. Question 2 - How do we know that generous choices are more likely to be errors?: We can show that people are more likely to change their mind on trials where their first movement was initially generous (P < .0001), and that the extent to which a given person is more likely to change their mind after initially making a generous choice is correlated with the model-predicted likelihood that generous choices are more likely to be errors than selfish choices (r = .78, P < .0001).
3. Question 3 – How do we know that people subjectively experience reversal of an error as pleasant?: Examining participants’ outcome evaluations reveals that, controlling for amounts received by self and other, in general people like having their choices implemented (estimated positive effect of choice implementation on outcome evaluation = .53, P < .0001). However, when that choice was estimated by the model to be an error, people report being significantly less happy if the choice is implemented than if it is reversed (B = -.23, P < .0001).

Taken together, we think these results provide *powerful* support for our hypothesis. Given the space constraints of the journal, we have not included a description of the results in the revision. However, we are working on a follow-up paper describing these results. We believe that mouse-tracking has considerable potential to yield novel insights into the processes described here, and that the mousetracking results described above should alleviate the concerns of both Reviewer 2 in this and the following comment, as well as reviewer 3’s question about the utility of the reversal period (see comment 3-2).

**Comment R2-7: Finally, I am not so convinced about the rationale of the analysis presented in figure 7. The reasoning presented by the authors relies on 2 layers of reverse inference: lower response in a region is a marker of a negative affect and (since errors might produce negative affect) this should be taken as support of the error hypothesis. The same problem arises in defining vmPFC (a region involved in many computational processes including decision-making, memory, emotional regulation, extinction, simply resting state to mention a few) a 'utility coding area' - I am sure the authors are aware of the fallacy of this chain of reasoning, which should be avoided.**

Response**:** Several points in relation to this very important comment.

First, we agree completely that traditional reverse inference arguments have a lot of problems. But this is not what is being done here.

In a traditional reverse inference analysis, the analyst carries out the contrasts of interest between two conditions of interest, and then uses the resulting activations to infer mental function by appealing to some published studies. We agree that this is problematic since areas like vmPFC are likely to be modulated by many different computations.

But the logic of the exercise carried here is different. We identified ex-ante, at the time of experimental design, an area of vmPFC that has been reliably shown to be correlated with the computation of experienced utility signals in multiple studies and in meta-analyses. We then designed a task that, if our hypothesis is correct, should change the amount of experienced utility in predictable ways. We used these pre-defined markers of experienced utility to test the model prediction that reversing mistake should lead to a higher experienced utility than reversing correct choices. Note that we don’t look for this test everywhere, but only in the ex-ante predefined area of interest.

It is important to emphasize that, unlike traditional applications of reverse inference, the logic of the test is valid even if there are other signals encoded in this area.

Second, as detailed in our response to the previous comment, we also have novel behavioral data that *strongly* supports the interpretation we have given here regarding vmPFC responses. As we note above, if the editor or reviewer believes that the addition of this behavioral data is feasible given space constraints, and contributes to the strength of the paper, we would be more than happy to include it.

**Response to Reviewer #3**

We would like to thank the reviewer for his/her detailed comments, which we have found extremely useful in revising and improving the quality of the paper.

For the reviewer’s convenience, we have copied his/her original comments in bold-type face.

**Comment R3-1: The authors use a drift diffusion model (DDM) and fMRI to analyze participants' behavior and neural responses while allocating (pre-defined) financial prizes to themselves or an anonymous other person. The main neural results show A) that activation in the ventromedial prefrontal cortex (vmPFC) correlates with value assigned to the proposal as indicated on a rating scale, B) that the proposed prize for self (i.e., the participant) correlates with activation in a number of brain regions, including vmPFC and amygdala, and C) that the proposed prize for the "other" correlates with activations in the temporoparietal junction and the precunius. Moreover, there was an overlap of activations for self and other in the vmPFC. One main result of the DDM analysis showed longer reaction times for generous as compared to selfish choices.   
  
The manuscript is well written, and covers a timely topic. However, I have some important reservations that considerably reduce my enthusiasm for this paper.**

Response**:** We thank the reviewer for their comments here and below.

We have implemented multiple improvements to address the reviewers’ comment, including the collection of new behavioral data, additional new analyses, the introduction of additional results that we had left out from the first version, and rewritten extensive sections of the paper to improve its clarity on a number of areas.

**Comment R3-2: “Major points.**

**The authors develop a drift diffusion model that includes the monetary payoff for "other". The development of such a model is, in principle, welcome and could bring more rigor to the analysis. However, the authors greatly oversell the value of their model. What I find most problematic is that they create the impression that the model has been verified by the data. But this is not the case. To give an example of what I mean. They estimate 5 parameters based on the choice data in the behavioral task. Then they use the model with the estimated parameters and "predict" subjects behavior based on these parameters. Not surprisingly, this model provides a reasonable fit for subjects' behavior because, after all, the parameters of the model have been estimated on the basis of subjects' observed behavior. It is therefore totally unsurprising that the model is consistent with the data. The statistical estimator does exactly that. It chooses those parameter values for the five parameters that fit the behavioral data best. But this has nothing to do with "predicting" data. The paper is full of problematic claims (e.g. final paragraph on p.4 which extends to p. 5) that seemingly validate the model predictions by the observed behaviors although - as explained above, this argumentation is entirely circular. Nothing is predicted, and the behavioral data cannot validate the model. The whole exercise is all about "fitting" and not "predicting" or "testing" the model.”**

Response**:** Several points with respect to this comment.

First, and perhaps more importantly, we agree with the reviewer that we were too liberal in some cases with our use of the word “prediction.” We had done out-of-sample prediction analyses, but originally opted to not include them in the paper for the sake of brevity. However, we now include them in the first part of the paper where we discuss fits between model predictions and observed data (see the section titled “The model accurately predicts out-of-sample choice and RT” and the updated Figure 2). As can be seen, fitting the data from a randomly-selected half of the data and testing model predictions using the other half shows no difference in the main conclusion, and almost identical fits to the data, both within and across participants. We hope that this alleviates part of the reviewer’s concern about the ability of the model to make predictions about new data, and avoids the circularity of argument to which the reviewer objected.

Second, we also would like to point out that there are several other points in the paper where we engage in genuine prediction exercises. For example, although we use only the behavior to fit the model, the computations involved make several important predictions about neural patterns of response. We test 3 of those predictions here: 1) that different regions of the brain should code for attributes deemed to be important in the model (e.g. $Self and $Other), and that these attributes should be integrated and compared somewhere (e.g. vmPFC); 2) in a newly added section, we show that the model makes the prediction that accumulator computation should be more extensive on trials resulting in generous vs. selfish choices. We find that there are several regions of the brain whose activation conform to the predictions of the model. 3) Finally, we make predictions about the outcome period based entirely on data from the decision period, and show that activations in the vmPFC reflect exactly the predictions of the model.

We believe that the addition of true out-of-sample prediction with respect to model-fitting, as well as the link between model predictions and neural response, constitute novel predictive exercises that should be of considerable interest to the wide readership of Neuron.

Finally, it might be useful to emphasize that fitting the many patterns explained by the model using only 5 parameters, and connecting them to the neural data, is a non-trivial exercise. Plainly, most five model parameters will not be able to provide a full account of all of these variables.

**Comment R3-3: 2) The design of the behavioral task is quite strange. If a subject makes an altruistic choice there is a 40% probability that this choice is reversed by the experimenter. Likewise, if subjects make a selfish choice there is a 40% chance that the experimenter invalidates the choice and implements an outcome that gives both players the same payoff. This gives subjects an excuse for making selfish choices because perhaps the computer (experimenter) is reversing their choice, and it provides very weak incentives for altruistic choices because there is a high chance that the experimenter will undo an altruistic choice. I could not find a good justification for this very unusual method.**

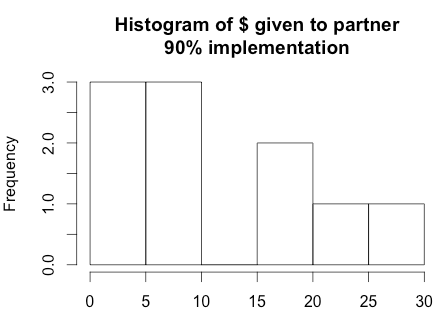
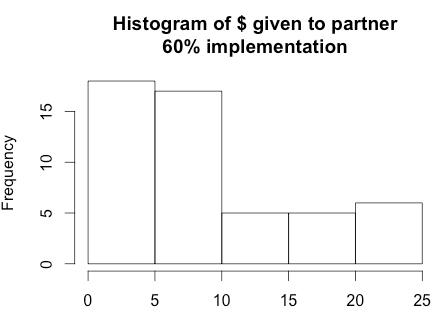
Response**:** Several points with respect to this comment.

First, we fear that there is some confusion regarding the logic of this aspect of the experiment, most likely due to imperfect exposition in our part. As described in more detail in the new version of the paper, this feature of the experiment was chosen because it allowed us to test the hypothesis about mistakes using neural data, but does not change the incentives to choose the favorite option in every trial.

Second, to address this point (and related comment R2-2), we ran a small additional study of 10 subjects, in which choices were implemented with 90% probability. We chose the 90% parameter based on two principles: 1) we wished to assess whether this might have changed participants responses, and 2) we also wished to retain the “plausible choice anonymity” of the current design to test internal motivations for altruism, rather than external motivations related to reputation management which would appear if there were no reversals. The 90% probability of implementation creates a tighter link between choice and outcome, but retains plausible choice anonymity.

Choice behavior looked almost identical to that in the current study.

1. Average percentages of generous behavior were roughly comparable (27% in the pilot study, 21% in the study reported in the paper).
2. Distributions of average generosity in dollar amounts were similar:

1. Response time differences between generous and selfish behavior were nearly identical (mean difference in 90% implementation = 198ms, mean difference in 60% implementation = 135ms).
2. The negative correlation between RTs and generosity emerged just as strongly (r = -.87 in 90% implementation, r = -.64 in 60% implementation)

For these reasons, we think that the behavior here is unlikely to be due simply to people worrying that their choices were rarely implemented.

We have included a short statement on pg. 22 of the current manuscript to indicate that behavior is not substantially affected by the high reversal rate: “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 that this is not likely an issue.”

**Comment R3-4: 3) Because of this method it is not surprising that only 21% of the choices were altruistic. The manuscript does not provide details about the distributions of generous choices across participants, but states a maximum of generous choices of 61%. Together with the small average, this implies that many participants made exclusively selfish, or only very few generous choices. I wonder about the validity of a neuro-computational model of altruistic choices that is derived from a data set with mainly selfish decisions. For example, the slowdown of reaction times for generous choices was less pronounced in more generous participants (Results, p. 12, first paragraph). Given that these "more generous participants" were the minority in the sample, findings like this raise questions about the interpretation of the majority of the data where generous decisions were very rare events or did not occur at al. What do we learn from a slowdown in reaction time of a decision, which was hardly made?**

Response**:** The reviewer is correct that we observe somewhat low levels of generosity in the current paradigm. However, we would like to note that these levels are consistent with levels of giving that have been observed in similar contexts across a wide variety of studies (e.g., Engel, 2011, Experimental Economics).

We also have data from follow-up behavioral studies in which we attempt to explicitly control levels of generosity, both by adaptively presenting offers that a subject is likely to respond generously or selfishly to, and by explicitly manipulating subjects’ attention to produce higher rates of giving (e.g. by asking them focus on empathy for the other person). In these studies, we are able to achieve much higher rates of generosity (30-50%), but still observe patterns of response that are entirely consistent with the predictions of the computational model. Despite greater frequency, generous choices take longer to make. We also observe the negative correlation between average generosity and reaction time differences to choose generously. Finally, we are also able to confirm a prediction of our model, which is that manipulations that increase the weight placed on the other (e.g. through manipulation of empathic concern) *decrease* reaction times for generous trials. We observe these changes in reaction time *within the same subject*, in a manner that conforms precisely to the predictions of the model. For this reason, we think that the relative rarity of generous behavior observed in the study described here is not likely to mean that the model or the conclusions we draw from it cannot be extended to other conditions. In fact, we have explicitly shown that it can be applied to a wide variety of contexts that vary substantially in their baseline levels of generosity. Although we cannot include the results of these other studies here, we are working on a follow-up manuscript describing their implications.

More importantly, however, we believe that the reviewer’s comments, both here and below (see response to the next comment), might reflect a misunderstanding of the point of our computational model. We would argue that the model should apply not only for subjects who are quite generous, but even for subjects who never make a single generous choice. A selfish person who never makes a generous choice should have a high weight on her own payoffs, and a weight of near zero (or even negative) on the other person’s. A generous person might have a higher weight on the other person, but the actual implementation of choices (i.e. integration and comparison) will be identical for both people. This can be seen by the fact that the accuracy of our model in predicting both choices and reaction times is high for both selfish and generous subjects. The mean within-subject correlation between model-predicted and actual choices is .95 (P < .001) in the most selfish half of subjects, and .81 (P < .001) in the most generous subjects. The mean correlation between model-predicted RT and observed RT for the most selfish subjects is r = .61 (P < .001), and r = .46 for the most generous subjects. Although we do see slightly less accurate model predictions for generous compared to selfish subjects, we believe that this might result from the greater variability in their responses (since they are incorporating $Other and performing a more complex tradeoff). However, our point remains: the model accurately captures both choice and RT, regardless of whether one is a generous or selfish individual, and regardless of whether one made many generous choices or only a few.

In fact, we believe this is a strength of our model over many others: based on the RT data, our model can infer that someone who never makes a single generous choice in this experiment might still care about others’ welfare at a minimal level (i.e. with low but positive weight). Over the range of offers presented in the current study, this low weight would not be enough to outweigh their own benefits, and therefore would not result in a generous choice in the current study. However, we would expect our model to more accurately predict the likelihood of a generous choice for these subjects in cases where one had to sacrifice just $1 in order for someone else to receive $1000 or $1,000,000. Here, the low but positive weight on another person would result in a calculus favoring generosity. In contrast, other models would have inferred a weight of 0 on the other, and thus would continue to predict selfishness even at these extremes of minimal self-sacrifice and maximal benefit to others. Although a test of this hypothesis with extreme benefits to others is beyond our ability, we hope that it helps to illustrate why we think our model is a better and more accurate one to infer behavior about a rare phenomenon.

**Comment R3-5: 4) The neural results are determined based on parametric analyses across ALL decisions (selfish and generous). Given that the majority of decisions was selfish, the reported neural activation should rather be associated with the selfish domain, or with allocation decisions in general. To establish a neural model of altruistic choices, it would be necessary to conduct these analyses based on altruistic decisions only, or on the contrast between altruistic and selfish decisions. Given the small number of altruistic choices in the data set, I suspect that these analyses might not be feasible here.**

**Taken Point 1 and 2 together, I do not think that these results "provide insight into the common processes at work in altruistic choice […], and provide novel insights into the nature of altruistic behavior" as stated in the Discussion (p. 16, end of first paragraph).**

Response**:** As we have noted above, we believe that this comment represents a misunderstanding of the nature of the computational model, and therefore misses the nature of our study. We argue that the model should apply EQUALLY to trials on which a person chooses generously and trials on which a person chooses selfishly, and EQUALLY for generous or selfish individuals. The point of our model is that several complex patterns of choices, reaction times, and neural response can emerge from a model in which the only thing that varies across trials are the INPUTS to the system, not the way those inputs are transformed into outputs. The computational process can be EXACTLY THE SAME at every instance, and yet, given the computational architecture, different inputs will produce different outputs. People will make generous choices if the weighted sum of the inputs produces an integrated value that favors the generous choice, and will make selfish choices otherwise. Thus, it would be inappropriate, computationally speaking, to estimate the model using only trials that resulted in a generous or a selfish response. This would in fact produce a form of over-fitting (which the reviewer’s earlier comments rightfully point out that we should avoid), since one would be trying to predict responses using a training dataset that had already been sorted as a function of the desired criterion to be matched (i.e., the response). We can also show that parametric neural representations of $Self and $Other do not differ substantially as a function of generous vs. selfish trials, even when we restrict our analysis to subjects with a large fraction of generous decisions.

The point of our paper is precisely to show that the kind of thinking that tries to divide choices into generous or selfish responses, and then examine these choices as if different processes were active during the two cases, is problematic. A simple model in which the SAME computational algorithms and neural mechanisms are in force equally on every trial can produce patterns of choice, RT, and neural response that LOOK like different mechanisms are active when people choose generously instead of selfishly. Our model suggests that people will take longer to choose on trials in which the inputs ultimately lead to a generous choice, and will be more likely to show greater response in several brain areas, to the extent that they have low weights on the other person’s welfare. This is exactly what we observe in the data, and is a pattern observed in other people’s work (i.e. Piovesan and Wengstrom, 2009). Critically, we can show that these patterns are entirely accounted for by the predictions of the model.

We hope to convince the reviewer that this misunderstanding actually highlights how fundamentally important and novel our model is: much of the literature to date makes inferences about the need for self-control, or the rewarding nature of choosing generously, by trying to compare patterns on generous vs. selfish choices. Our work suggests that such inferences need to take into account the predictions of a simpler model before making conclusions about more complex processes, and that this point applies equally well to choice and RT data as it does to neural response. For this and many other reasons, we believe that the paper represents a much more important and novel advance than the reviewer concludes in his/her comments.

**Comment R-6: Minor points.  
5) One region of interest (ROI) in the vmPFC was defined based on a conjunction of two masks generated from a meta-analysis on reward processing, another ROI in vmPFC based on a conjunction of activation related to the prize of self and other in the current study. What is the reason for defining two different ROIs within the same region? It is preferable to use the same independent ROI for all analyses.**

Response**:** The reviewer here is suggesting that a mask used for small-volume correction should also be the mask used for extracting parameters of interest for use in further analyses.

We respectfully disagree that this is the most appropriate method. For example, the meta-analyses that we use to define a vmPFC region for small-volume correction are based on a large set of studies that include both social and non-social values. Because of this, and because individual brains will differ somewhat, we favor a two-step approach: first use a general, non-study specific mask to define a priori regions of potential interest, and then define within this larger area a region that shows activation for a particular quantity of interest specific to the individuals in the current study. We have thus opted for now to retain the original approach used in the manuscript.

That said, the mask we used in the paper and the conjunction mask based on the meta-analysis produce nearly identical results: $Self and $Other are both significantly represented in the meta-analysis based mask (*P* < .001 and *P* = .04, respectively), with $Self being significantly more represented than $Other. Although we think that the mask defined by functional activations in the current study is more likely to represent an accurate picture of important voxels for the subjects used in this study, we would be happy to replace the current results with the meta-analysis mask results if the editor deemed it appropriate.

**Comment R-7: 6) Is there a measure which assesses whether the participants really believed that they are playing with real people? Maybe the majority behaved so selfishly because they did not believe that a human partner would actually receive the financial prize.**

Response**:** Although we do not have data from this particular study, we have data from several other studies run on similar populations, suggesting that only a minority disbelieve the presence of the partner. While this disbelief does reduce giving somewhat, we see that it does not reduce it entirely. In study 1: 17% disbelieved (Ave generosity if a Pp believed: $5.25, didn’t believe: $3.70, P = n.s.), in study 2, 18% disbelieved (believed: $1.65, didn’t believe: $0.43, P = .03), in study 3: 13% disbelieved (believed: $4.15, didn’t believe: $3.76, P = .67). Note also that there are several people who reported they believed the other person was there and STILL behaved entirely selfishly in these studies. So it is not the case that one behaves selfishly ONLY if you don’t believe someone is there. Given the rates of disbelief in these other studies, we would expect only 7-9 participants in the current study to have had doubts about the presence of the other person. While this might account for a small portion of the low donation rates, we think it unlikely to be the dominant force in shaping responses here.

**Comment R-8: 7) eToc statement: "The model explains neural responses and suggests that some generous choices are mistakes that are later regretted." This interpretation seems bold, given that regret was not assessed.**

Response**:** We have revised the eToc statement to read: “Hutcherson et al. show that a computational model of altruism accounts for behavioral and neural effects attributed to self-control and/or the value of generosity, without requiring either. The model suggests that some generosity represents choice errors, not genuine social preferences.”