Why Do Irrelevant Alternatives Matter? An fMRI-TMS Study of Context-Dependent Preferences

Chen-Ying Huang[1,*,**], Hui-Kuan Chung[2,*], Hsin-Ju Lee[3], Yi-Ta Lu[1], Fu-Yun Tsuo[4], Tzai-Shuen Chen[5], Chi-Fu Chang[6], Chi-Hung Juan[6], Wen-Jui Kuo[3,**], Tomas Sjöström[7,**]

1 Department of Economics, National Taiwan University, Taipei, Taiwan

2 Department of Psychology, New York University, New York, NY 10003, U.S.A.

3 Institute of Neuroscience, National Yang-Ming University, Taipei, Taiwan

4 Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

5 Department of Economics, Washington University in St. Louis, St. Louis, MO 63130, U.S.A.

6 Institute of Cognitive Neuroscience, National Central University, Taoyuan, Taiwan

7 Department of Economics, Rutgers University, New Brunswick, NJ 08901, U.S.A.

* These authors contributed equally to this work.

** To whom correspondence should be addressed:

chenying@ntu.edu.tw, wjkuo@ym.edu.tw or tsjostrom@economics.rutgers.edu

August 2, 2015

Abstract

Both humans and animals are known to exhibit a violation of rationality known as "decoy effect": introducing an irrelevant alternative (a decoy) can influence choices among other (relevant) alternatives[1,2,3]. Exactly how and why decoys trigger this effect is not known. It may be an example of fast heuristic decision-making which is adaptive in natural environments[4,5], but may lead to biased choices in certain markets or experiments. We used functional magnetic resonance imaging (fMRI) and transcranial magnetic stimulation (TMS) to investigate the neural underpinning of the decoy effect. The left ventral striatum was more active when the chosen alternative dominated the decoy (compared with when the same chosen alternative did not). This suggests that the decoy may influence the valuation of other alternatives, making valuation context-dependent. Consistent with the idea that control is recruited to prevent heuristics from producing biased choices[6], the right inferior frontal gyrus (rIFG), often implicated in inhibiting prepotent responses[7,8], connected more strongly with the striatum when participants successfully overrode the decoy effect and made unbiased choices (compared with when they were not successful). This is further supported by our TMS experiment: participants whose rIFG was temporarily disrupted made biased choices more often than a control group. Our results suggest that the decoy effect is due to context-dependent activation of the reward area. But the differential connectivity from the frontal area may indicate how deliberate control monitors and corrects errors in heuristic decision making.

Summary We used functional magnetic resonance imaging and transcranial magnetic stimulation to study the neural context effect caused by a decoy, its trial-by-trial strength, and the possible control mechanism recruited to overcome a potential decision-making bias.

In standard economic theory, decision-making is represented as assigning value to each feasible alternative and choosing the alternative with the highest value. In this theory, the value of an alternative depends only on its intrinsic properties, not on the context in which it appears. Neuroscientific studies suggest that the brain does encode values of available alternatives[9], but there is mixed evidence on the extent to which these values depend on the context[10,11]. Counterfactual outcomes of a lottery (i.e., outcomes that could have, but did not occur) have been shown to be a source of context-dependence for human subjects[12,13]. Intuitively, winning five dollars in a lottery where this is the smallest possible prize is not as rewarding as winning five dollars in a lottery where this is the largest possible prize. Adjusting the responses of neurons in the reward area to the range of possible outcomes in this way may be an efficient way for the brain to utilize the neurons' limited firing range.

Since the counterfactual outcomes of a lottery had a chance of occurring, they may have influenced anticipations, which may affect the evaluation of the final outcome, perhaps by triggering disappointment or elation. In a sense, any outcome with a positive probability of occurring is a relevant part of the lottery. Suppose, however, that a participant chooses from a set of options, one of which should be completely irrelevant to rational decision-making (in a sense to be made precise below). Will the brain manage to disregard the irrelevant alternative, and save its limited firing range for the viable options? Or will the valuations of the viable (relevant) options be influenced by the irrelevant option? We study this question in the context of the decoy problem, well known from the literature on animal behavior[1] and consumer marketing[2].

The decoy problem can be illustrated by the pricing of *The Economist* magazine[3]. A print-only subscription, and a print-and-digital subscription which includes both print and digital access, are priced the same. The print-only subscription is a decoy option which nobody is expected to choose, because it is clearly worse to get only print access than to get *both* print and digital accesses at the same price. However, controlled laboratory experiments[2] have shown a

"decoy effect": the existence of the decoy tends to increase the demand for the print-and-digital subscription. But does the presence of a seemingly "irrelevant" decoy actually influence the valuation of other options ? Our results indicate that it does, even when choices are not influenced by the decoy (so that there is no decoy effect in the usual, choice behavior, sense). Thus, valuation does seem to depend on "irrelevant" aspects of the context. Our results hint at ways the brain may try to prevent this from leading to biased decisions.

Using functional magnetic resonance imaging (fMRI), we scanned 32 participants when they made choices from a series of two- and three-item menus. The items on the menus were restaurant meals with specified prices and quality levels. Formally, alternative X *dominates* alternative X' if and only if X is both cheaper and has a higher quality than X'. (In our notation, an alternative that is dominated by X will always be denoted X'.) Some of the three-item menus contained a decoy item that was dominated by another item on the menu and hence, according to economic theory, should be irrelevant to decision-making. Domination, when it occurred, was always *asymmetric* in the sense that the decoy was dominated only by one item, called the *target*, and not by the other, called the *competitor* (supplementary information (SI), S1.4).

The trials that included a decoy (called the decoy trials) were paired, so that each undominated alternative would be the target in one trial and the competitor in the other. Such a pair of decoy trials has the form $\{A, B, A'\}$ and $\{A, B, B'\}$, and this will be referred to as a *matching pair* (of decoy trials). The undominated alternatives A and B are the same in both trials of the matching pair. The only difference is that, in $\{A, B, A'\}$ the decoy A' is dominated by A but not by B, while in $\{A, B, B'\}$ the decoy B' is dominated by B but not by A[14]. Matching pairs of decoy trials are the main focus of our analysis. We do not compare three-item decoy trials with two-item trials without decoys, because choosing from a larger menu may be different from choosing from a smaller menu, and this may confound the analysis. Four trials are illustrated in Figure 1a.

Behaviorally, 81.20% of the matching pairs were *consistent*, in the sense that the choice was independent of the decoy. That is, the participant chose the same alternative from $\{A, B, A'\}$ and $\{A, B, B'\}$. 16.41% exhibited *preference reversals*, in the sense that the participant switched her preference from A to B when the decoy changed from A' to B'(Figure 1b)[15]. That is, the participant chose A from $\{A, B, A'\}$ and B from $\{A, B, B'\}$. In terms of choice behavior, context-independence seems to be the norm (accounting for 81.20% of the matching pairs). But valuations may be context-dependent even when choice is not, because a change in values does not necessarily lead to a change in the chosen alternative.

In a decoy trial, the choice is said to be *Along* the decoy if the target is chosen and *Against* the decoy if the competitor is chosen. In a consistent matching pair, the same alternative, say A, is chosen both from $\{A, B, A'\}$ and $\{A, B, B'\}$. The choice is hence Along the decoy when A is chosen from $\{A, B, A'\}$ and Against the decoy when A is chosen from $\{A, B, B'\}$. The response time (RT) suggests that it was easier to make a choice along the decoy than against the decoy in consistent trials. Across all participants, the average RT was 8.36 seconds in Along trials and 9.94 seconds in Against trials. The average percentage decrease of RT in Along is 16.20% which is strongly significantly different from zero (Figure 1c). Thus, in consistent trials, where by definition the chosen alternative does not depend on the decoy, choosing along the decoy still seems easier than choosing against it.

We will look for a neural manifestation of a decoy effect, in the sense of context-dependent valuations, in consistent trials. If valuations are context-independent, the valuation area should not differentiate between the Along and Against trials of a consistent matching pair. But if valuations are context-dependent, an alternative will tend to be more highly valued when it dominates the decoy, so the Along trial will be more rewarding than the Against trial even though the chosen alternative is the same. Context-dependence therefore leads to the prediction that reward-sensitive areas will be more activated in

Along than in Against trials of consistent matching pairs. This prediction was supported by our fMRI data. The left ventral striatum was significantly more active in the Along than in the Against consistent trials (Figure 2a). The striatum activity was observed at putamen and extended medially to caudate and anteriorly to insula[16]. Notice that the Along-Against contrast involves a pair of trials with three-item menus, the same undominated items appear on both menus, and the same item is chosen from both menus. The only difference to which differential activity can be attributed is that the chosen item dominates the decoy in the Along but not in the Against trials[17].

Two additional pieces of evidence support the interpretation that choosing along the decoy is more rewarding than choosing against it. First, although previous research has indicated that the striatum does code reward[18,19], we provide direct evidence for the connection between the striatum and reward by estimating the utility of each alternative A, denoted $u(A)$, using data from post-test choices in two-item menus with no decoys (SI, S4.2). The utility $u(A)$ depends on item A's price as well as its quality. We refer to $u(A)$ as the intrinsic utility of A, because it corresponds to the utility from A's intrinsic properties in the absence of a decoy. We found that the striatum activity correlates parametrically with the estimated intrinsic utility of the chosen item positively (Figure 2b). Hence in our data, there is evidence that the striatum may be coding reward. The active cluster overlaps with what we discovered when contrasting Along to Against (SI, Figure S1), rendering our interpretation that it was more rewarding choosing along the decoy than choosing against it plausible. Second, across participants, the differential striatum activity in Along versus in Against positively correlated with two behavioral measures of how "decoyable" a participant is. The first such measure is simply the number of preference reversals (Figure 2c). The second measure, denoted μ, will be introduced below (Figure 2d). Thus, more decoyable participants experienced a larger increase in reward when the chosen alternative dominated the decoy. This match between the behavioral measures and the neural effect is consistent with the hypothesis that

context-dependent valuation causes the decoy effect.

How well an alternative will satisfy a person's wants and needs depends on the intrinsic properties of price and quality. If valuations are influenced by irrelevant aspects of the context, choices may become biased (i.e., the intrinsically less valuable option may be chosen). Unbiased choice may require some cognitive control. To look into how control could be recruited, we first construct a measure of the trial-by-trial strength of the decoy effect, which will guide us to the potential effect of cognitive control on valuation.

The number of preference reversals can be used to compare, between-participant, who is more decoyable. But it cannot be used to determine within-participant when she feels the decoy effect more strongly. Moreover, as argued above, in consistent trials the decoy may have an effect on valuations which is not reflected in choices. To construct a trial-by-trial estimate of the decoy effect, we modify a simplified linear ballistic accumulator (LBA) model. The idea behind LBA is that choice is made when the accumulated evidence in favor of the chosen alternative has reached a threshold. The accumulation of evidence is faster if the evidence is stronger, so on average there is an inverse relationship between the strength of evidence and RT[20]. Translating these ideas into our experiment, if the menu in trial t is $\{A, B, A'\}$, because A' is dominated the race will be between A and B. Intrinsic utilities $u(A)$ and $u(B)$ were estimated in the post-test as mentioned above. Thus, the intrinsic utility difference between A and B is $u(A) - u(B)$. We hypothesized that the decoy effect in trial t causes the relative value assigned to the target to be shifted by some amount $d(t)$. Since A is the target in $\{A, B, A'\}$, due to the presence of the decoy the *decision utility difference* will be $u(A) - u(B) + d(t)$. Here $d(t)$ represents the amount by which the decoy effect favors the target, A, relative to the competitor, B, in trial t. While we might expect $d(t)$ to be positive on average, we make no assumption regarding the sign of $d(t)$.

In terms of the LBA model, the strength of the "evidence" favoring A is

$u(A) - u(B) + d(t)$, the sum of the intrinsic utility difference and the decoy effect of this trial $d(t)$. Thus, if A is chosen from $\{A, B, A'\}$ in trial t, we assume RT is inversely proportional to $u(A) - u(B) + d(t)$,

$$RT = \frac{T}{u(A) - u(B) + d(t)} \tag{1}$$

for some threshold constant $T > 0$. If instead the menu in trial t is $\{A, B, B'\}$, then since B is the target, the decoy effect favors B. Therefore, if $d(t)$ is the strength of the decoy effect in trial t, the decision utility difference, i.e., the strength of the "evidence" favoring A, is $u(A) - u(B) - d(t)$. Thus, if A is chosen from $\{A, B, B'\}$ then RT is inversely proportional to $u(A) - u(B) - d(t)$,

$$RT = \frac{T}{u(A) - u(B) - d(t)}. \tag{2}$$

For any decoy trial, $d(t)$ satisfies equation (1) in Along trials (where A is chosen from $\{A, B, A'\}$) and (2) in Against trials (where A is chosen from $\{A, B, B'\}$). We assume that $d(t)$ is drawn from a normal distribution with participant-specific mean μ and standard deviation σ. Note that μ can be interpreted as a participant's average decoy effect. Even though we expect it to be positive, whether it is positive, zero or negative is left open. We use maximum likelihood to estimate T, μ and σ for each participant. The trial-by-trial decoy effect $d(t)$ is then backed out from (1) in Along trials and from (2) in Against trials. Notice that the sign in front of $d(t)$ is different in (1) and in (2), because the decoy effect favors a different alternative in the two cases. A random component of the intrinsic utility difference would not flip sign in this way, so $d(t)$ cannot be interpreted as an estimate of random utility[21]. Similarly, $d(t)$ is not a monotonic function of RT. In Along trials, the shorter RT is the *stronger* the decoy effect is (because the decoy is "helping" the target to be chosen quickly), but in Against trials, the shorter RT is the *weaker* the decoy effect is (because the decoy is working against the choice of the competitor).

The results of the estimation suggest that the model is appropriate. A participant's estimated average decoy effect, μ, is a measure of how "decoyable"

she is. If the decoy indeed makes the target more valuable, μ is expected to be positive. For all but two participants, it is indeed positive. Furthermore, μ strongly positively correlated with the number of preference reversals (Figure 3a) suggesting it effectively captures how decoyable each participant is. As mentioned above, μ also strongly positively correlated with the differential striatum activity in Along versus in Against. This strengthens the case for the higher activity in Along than in Against being evidence that participants find choosing along the decoy more rewarding than choosing against it. Since T can be interpreted as the threshold for making a decision, a larger T should imply a longer RT, and indeed the positive correlation between T and the average RT of decoy trials is very strong (SI, S4.5).

For each participant, we divided the decoy trials into two halves depending on the size of $d(t)$. The half with large $d(t)$ was classified as Strong trials, suggesting a strong decoy effect. The half with small $d(t)$ was classified as Weak trials. Neurally, no region was found contrasting Strong to Weak. The left inferior parietal lobule (IPL) was the only region more active contrasting Weak to Strong (Figure 3b). Its activity in fact parametrically tracks $d(t)$ negatively (Figure 3c)[22]. IPL has previously been implicated in goal-directed preparation of attention[23,24,25]. In particular, IPL close to our activation cluster is activated more when there is a switch than when a task is repeated, when trials are incongruent than when they are congruent in the Stroop or flanker tasks, when stimulus and response are incompatible than when they are compatible, and when working memory is more heavily taxed[26,27,28,29]. The common denominator in these experiments is that IPL may support the allocation of attention to facilitate task-relevant representations. As part of a general network subserving voluntary attention, this may explain why there is higher IPL activity, suggesting heightened attention, when the decoy effect $d(t)$ is smaller.

If the decoy effect increases the valuation of the target, it could potentially lead to the wrong decision in trials where the target is intrinsically worse than the competitor. Presumably, it is impossible to inhibit processing of the decoy

at the perceptual level. But at a higher level, control may prevent interference from the decoy and help refocus on the relevant aspects of the choice situation (the intrinsic properties of the undominated alternatives). We do not find a direct modulation from the IPL in our data. We do find a plausible indirect modulation which we will further verify by a transcranial magnetic stimulation (TMS) experiment.

We looked into how the putative control was recruited by the following two psychophysiological interaction analyses. We divided the decoy trials into two halves. The half where the target has a lower intrinsic utility than the competitor is categorized as Conflict. The other half, where the target has higher intrinsic utility, is categorized as NoConflict[30]. In a Conflict trial, the target has lower intrinsic utility, and thus control may be required to override the decoy effect and prevent the target from being chosen. In a NoConflict trial, the target has higher intrinsic utility, so the decoy effect is less problematic and control presumably less critical. When attention is heightened and control is more likely to make a difference, we might expect a stronger connectivity between the area supporting attention and that implementing control.

We took the average activity of a 4-mm sphere surrounding a peak voxel of IPL as the seed to examine whole-brain whether any area exhibits stronger functional connectivity with IPL in Conflict than in NoConflict. The right inferior frontal gyrus (rIFG) is the only region that has this stronger task-related connectivity (Figure 4a). The rIFG is postulated to be a site where goal-directed and stimulus-driven attention converge[31,23]. Neuroimaging studies implicate this area in supporting inhibition to implement control[32,33,34]. It is more active in No-Go/Stop-Signal trials than in Go trials, in invalid Posner cueing trials than in valid ones, in trials where recent history may interfere than when it may not – all possibly reflecting processes related to overriding prepotent responses[7,8]. The rIFG has been implicated in a number of other exertions requiring self-control, such as inhibiting an incorrect answer to a problem of logic[35] and focusing on the facts rather than the framing of a question[36]. It has even been implicated

in the decision to quit smoking[37]. The presence of the irrelevant decoy is more critical to decision-making in a Conflict trial than in a NoConflict trial[38]. The stronger functional connectivity between IPL and rIFG could possibly be due to the posterior area signaling the conflicting representations to the frontal cortex for control[28]. Anatomical connections between IPL and IFG have been demonstrated using diffusion-weighted imaging[39,40]. As previous studies indicate that rIFG may play a role in inhibiting irrelevant responses, our connectivity result caused us to explore further the role of rIFG in our data.

We have argued that the striatum seems to code the decision utility of the chosen item. The observation that the striatum is more active in Along than in Against suggests that the decision utility of a chosen item tends to increase when it is the target, which may lead to biased choices in Conflict trials. If the role of control is to reduce this bias, it would tend to offset this differential activity in Along versus in Against by either increasing the decision utility of the chosen option in Against trials, or reducing it in Along trials, or both. This suggests that the control area may correlate more strongly with the reward area in Against than in Along Conflict trials. In Conflict trials the target is intrinsically worse than the competitor, so if the choice is along the decoy then control was not successfully applied, as the intrinsically best alternative is not chosen; if the choice is against the decoy then the intrinsically best alternative is chosen[41]. Hence, by considering the differential connectivity with a control area in Against versus in Along Conflict trials, we are contrasting trials where control is successful with trials where control is unsuccessful.

We took the average activity of a 4-mm sphere surrounding a peak voxel of rIFG identified above as the seed to examine whether the striatum has a differential functional connectivity with rIFG in Against than in Along of Conflict trials. In a whole-brain search, the left striatum shows a stronger task-related functional connectivity with rIFG (Figure 4b). The active cluster overlaps with what we discovered when contrasting Along to Against of consistent trials (SI, Figure S5). This is consistent with a possible role of rIFG in reducing the decoy-

induced bias in the decision utility. Anatomical connections between IFG and striatum have been demonstrated. Restricted frontostriatal diffusion seems to correlate with greater control, hinting at its contribution to the recruitment of control[42].

The two connectivity results provide evidence on the role of rIFG. rIFG exhibits differential functional connectivity with the striatum, hinting at its possible influence on choices. As it is often implicated in overriding irrelevant responses, rIFG may play a role in implementing control to overcome a potential decoy-induced decision-bias. We used TMS to investigate this further (SI, S5). We applied theta burst TMS to temporarily interfere with the region of interest before participants started the choice task[43,44]. In one group of participants, the IFG group, the site of stimulation was at the peak voxel of rIFG identified from the fMRI experiment. In the other group, the vertex group, the site of stimulation was the vertex. Each group had 32 participants, the same number as in the fMRI experiment. If rIFG plays a role in inhibiting the decoy-induced bias, then when it is temporarily disrupted, the inhibitory control is expected to be weaker, and the decoy effect is expected to be stronger[45]. In terms of choices, we thus expected the IFG group to exhibit more preference reversals than the vertex group. In terms of RT, because a strong decoy effect is expected to make it much easier to choose along the decoy than against it, we expected the IFG group to have a larger percentage decrease of RT in Along trials, compared with Against trials, than the vertex group.

These expectations were born out in the TMS experiment. The IFG group on average had a preference reversal rate of 25.30% whereas the vertex group had 18.72%, an increase of 6.58 percentage points. The t-test for comparing the preference reversal rate of the IFG group with that of the vertex group has a p-value (one-tailed) of 0.048 (Figure 4c). For the IFG group, on average, the RT was 7.23 seconds in Along trials and 9.08 seconds in Against trials. The average percentage decrease of RT in Along trials, compared with Against trials, was 19.22%. For the vertex group, on average, the RT was 6.68 seconds in Along

trials and 7.47 seconds in Against trials. The average percentage decrease of RT in Along trials, compared with Against trials, was 9.05%. The t-test for comparing the average percentage decrease of RT in Along trials, compared with Against trials, of the IFG group with that of the vertex group has a p-value (one-tailed) of 0.023 (Figure 4d).

In some naturally occurring choice situations, context-dependent valuation may be adaptive[4]. In sequential decision problems, optimal decisions depend on a comparison of the current option with its background[5]. Discovering an option that dominates other alternatives in the background can make it optimal to choose it right away. Thus, a context-dependent valuation that gives a positive connotation to dominance may be adaptive. There are, however, several reasons why such heuristics may sometimes lead to suboptimal decisions[46]. First, someone may be trying to manipulate the decision maker. For example, a product's intrinsic properties of price and quality will determine how well it will satisfy the consumer's wants and needs, i.e., will determine the experienced (or intrinsic) utility – but by introducing artificial decoy alternatives, marketers make the decision utility exceed the experienced utility. Second, the assumptions underlying the normal operation of the system may be violated. For example, the heuristics may be well adapted to sequential tasks, but not to the rather artificial static tasks in our experiment. Third, some decision problems are too complex for purely affective judgments, and require more deliberative and "rational" cognitive processes. All three reasons are relevant to our experiment. The decoy effect may be an evolved adaptation that illuminates the on-line judgments made by the decision-making system, rather than a "design flaw." It may be analogous to visual illusions, e.g., the Ebbinghaus illusion where the perceived size of an object depends on the sizes of neighboring objects[47,48], which is adaptive in ecologically relevant scenarios.

References and Notes

1. Sasaki, T. & Pratt, S. C. Emergence of group rationality from irrational individuals. *Behav Ecol* **22**, 276-281, doi:DOI 10.1093/beheco/arq198 (2011).

2. Huber, J., Payne, J. W. & Puto, C. Adding Asymmetrically Dominated Alternatives - Violations of Regularity and the Similarity Hypothesis. *J Consum Res* **9**, 90-98, doi:Doi 10.1086/208899 (1982).

3. Ariely, D. *Predictably irrational : the hidden forces that shape our decisions.* 1st edn, (Harper, 2008), chap. 1.

4. Houston, A. I. Natural selection and context-dependent values. *P Roy Soc B-Biol Sci* **264**, 1539-1541, doi:DOI 10.1098/rspb.1997.0213 (1997).

5. Freidin, E. & Kacelnik, A. Rational choice, context dependence, and the value of information in European starlings (Sturnus vulgaris). *Science* **334**, 1000-1002, doi:10.1126/science.1209626 (2011).

6. Kahneman, D. & Frederick, S. Frames and brains: elicitation and control of response tendencies. *Trends in cognitive sciences* **11**, 45-46, doi:10.1016/j.tics.2006.11.007 (2007).

7. Levy, B. J. & Wagner, A. D. Cognitive control and right ventrolateral prefrontal cortex: reflexive reorienting, motor inhibition, and action updating. *Annals of the New York Academy of Sciences* **1224**, 40-62, doi:10.1111/j.1749-6632.2011.05958.x (2011).

8. Bunge, S. A., Ochsner, K. N., Desmond, J. E., Glover, G. H. & Gabrieli, J. D. Prefrontal regions involved in keeping information in and out of mind. *Brain : a journal of neurology* **124**, 2074-2086 (2001).

9. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223-226, doi:10.1038/nature04676 (2006).

10. Padoa-Schioppa, C. & Assad, J. A. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nature neuroscience* **11**, 95-102, doi:10.1038/nn2020 (2008).

11. Padoa-Schioppa, C. Range-adapting representation of economic value in the orbitofrontal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29**, 14004-14014, doi:10.1523/JNEUROSCI.3751-09.2009 (2009).

12. Nieuwenhuis, S. *et al.* Activity in human reward-sensitive brain areas is strongly context dependent. *NeuroImage* **25**, 1302-1309 (2005).

13. Breiter, H. C., Aharon, I., Kahneman, D., Dale, A. & Shizgal, P. Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* **30**, 619-639 (2001).

14. The order of the trials and the ordering of the items in a trial were randomized, except that in the matching pairs $\{A, B, A'\}$ and $\{A, B, B'\}$, the ordering of target, competitor, and decoy was controlled to be the same to make the two trials in a matching pair as similar as possible.

15. We denote a participant by her.

16. Two participants were removed from the fMRI analysis because of excessive head motions.

17. A previous fMRI study[49] contrasted three-item trials of the form $\{A, B, A'\}$ with two-item trials of the form $\{A, B\}$. In the three-item trials, there was decreased activation in the amygdala, MPFC, and right IPL, and increased activation in DLPFC and ACC. Based on this, it was argued that the decoy effect may be caused by a shift toward more reason-based (heuristic) choice processes. Contrasting menus of different sizes introduces a possible confound, as it may be more cognitively demanding to choose from a larger menu. Our fMRI analysis was designed to contrast very similar three-item trials.

18. Delgado, M. R. Reward-related responses in the human striatum. *Annals of the New York Academy of Sciences* **1104**, 70-88, doi:10.1196/annals.1390.002 (2007).

19. Kable, J. W. & Glimcher, P. W. The neural correlates of subjective value during intertemporal choice. *Nature neuroscience* **10**, 1625-1633, doi:10.1038/nn2007 (2007).

20. Brown, S. D. & Heathcote, A. The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive psychology* **57**, 153-178, doi:10.1016/j.cogpsych.2007.12.002 (2008).

21. If we define $d(t)$ by the *same* formula, say (1), in *both* Along and Against trials, then $d(t)$ would correspond to a random component of utility, but it would not be a measure of the decoy effect. In SI, S4.11 and S4.12, we address the issue of random utility and show that if $d(t)$ were defined by the same formula Along and Against then it would not be significantly correlated with the inferior parietal lobule, whereas – as reported below – there is significant correlation when we use (1) in Along trials and (2) in Against trials.

22. We perform several robustness checks on IPL's negative correlation with $d(t)$. See SI, S4.7.

23. Shomstein, S. Cognitive functions of the posterior parietal cortex: top-down and bottom-up attentional control. *Frontiers in integrative neuroscience* **6**, 38, doi:10.3389/fnint.2012.00038 (2012).

24. Hopfinger, J. B., Buonocore, M. H. & Mangun, G. R. The neural mechanisms of top-down attentional control. *Nature neuroscience* **3**, 284-291, doi:10.1038/72999 (2000).

25. Raz, A. & Buhle, J. Typologies of attentional networks. *Nature reviews. Neuroscience* **7**, 367-379, doi:10.1038/nrn1903 (2006).

26. Liston, C., Matalon, S., Hare, T. A., Davidson, M. C. & Casey, B. J. Anterior cingulate and posterior parietal cortices are sensitive to dissociable forms of conflict in a task-switching paradigm. *Neuron* **50**, 643-653, doi:10.1016/j.neuron.2006.04.015 (2006).

27. Ye, Z. & Zhou, X. Conflict control during sentence comprehension: fMRI evidence. *NeuroImage* **48**, 280-290, doi:10.1016/j.neuroimage.2009.06.032 (2009).

28. Sylvester, C. Y. et al. Switching attention and resolving interference: fMRI measures of executive functions. *Neuropsychologia* **41**, 357-370 (2003).

29. McNab, F. et al. Common and unique components of inhibition and working memory: an fMRI, within-subjects investigation. *Neuropsychologia* **46**, 2668-2682, doi:10.1016/j.neuropsychologia.2008.04.023 (2008).

30. The decoy trial $\{A, B, A'\}$ is defined as a Conflict trial if $u(A) - u(B) < 0$, and as a NoConflict trial if $u(A) - u(B) > 0$. As the decoy trials are in matching pairs, exactly half of them are Conflict whereas the other half are NoConflict. For instance, if $u(A) - u(B) < 0$, then $\{A, B, A'\}$ is Conflict and $\{A, B, B'\}$ is NoConflict. See SI, S4.8.

31. Asplund, C. L., Todd, J. J., Snyder, A. P. & Marois, R. A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nature neuroscience* **13**, 507-512, doi:10.1038/nn.2509 (2010).

32. Aron, A. R., Robbins, T. W. & Poldrack, R. A. Inhibition and the right inferior frontal cortex: one decade on. *Trends in cognitive sciences* **18**, 177-185, doi:10.1016/j.tics.2013.12.003 (2014).

33. Nubert, F.-X, Mars, R. B. & Rushworth, M. F. S. Is there an inferior frontal cortical network for cognitive control and inhibition? In *Principles of Frontal Lobe Function*, Stuss, D. T., Knight, R. T. Eds. (Oxford University Press, New York, 2012), pp. 332–352.

34. Dillon, D. G. & Pizzagalli, D. A. Inhibition of Action, Thought, and Emotion: A Selective Neurobiological Review. *Applied & preventive psychology : journal of the American Association of Applied and Preventive Psychology* **12**, 99-114, doi:10.1016/j.appsy.2007.09.004 (2007).

35. Goel, V. & Dolan, R. J. Explaining modulation of reasoning by belief. *Cognition* **87**, B11-22 (2003).

36. De Martino, B., Kumaran, D., Seymour, B. & Dolan, R. J. Frames, biases, and rational decision-making in the human brain. *Science* **313**, 684-687, doi:10.1126/science.1128356 (2006).

37. Berkman, E. T., Falk, E. B. & Lieberman, M. D. In the trenches of real-world self-control: neural correlates of breaking the link between craving and smoking. *Psychological science* **22**, 498-506, doi:10.1177/0956797611400918 (2011).

38. Unlike the previous experiments that manipulate control explicitly, our design is not aimed at studying control. The classification of Conflict and NoConflict is based on the estimation of intrinsic utilities and so the difference between them could be quite subtle.

39. Caspers, S. et al. Probabilistic fibre tract analysis of cytoarchitectonically defined human inferior parietal lobule areas reveals similarities to macaques. *NeuroImage* **58**, 362-380, doi:10.1016/j.neuroimage.2011.06.027 (2011).

40. Rushworth, M. F., Behrens, T. E. & Johansen-Berg, H. Connection patterns distinguish 3 regions of human parietal cortex. *Cerebral cortex* **16**, 1418-1430, doi:10.1093/cercor/bhj079 (2006).

41. In Conflict trials, 66.30% of Along trials (together with their matching trials) exhibit preference reversal (signifying unsuccessful control), while 96.96% of Against trials (together with their matching trials) are consistent. In other words, the differential connectivity in Against than in Along

Conflict trials could also be phrased as the differential connectivity when choices are consistent than when choices largely exhibit reversal.

42. Liston, C. et al. Frontostriatal microstructure modulates efficient recruitment of cognitive control. *Cerebral cortex* **16**, 553-560, doi:10.1093/cercor/bhj003 (2006).

43. Huang, Y. Z., Edwards, M. J., Rounis, E., Bhatia, K. P. & Rothwell, J. C. Theta burst stimulation of the human motor cortex. *Neuron* **45**, 201-206, doi:10.1016/j.neuron.2004.12.033 (2005).

44. Chao, C. M. et al. Predictability of saccadic behaviors is modified by transcranial magnetic stimulation over human posterior parietal cortex. *Human brain mapping* **32**, 1961-1972, doi:10.1002/hbm.21162 (2011).

45. Jacobson, L., Javitt, D. C. & Lavidor, M. Activation of inhibition: diminishing impulsive behavior by direct current stimulation over the inferior frontal gyrus. *Journal of cognitive neuroscience* **23**, 3380-3387, doi:10.1162/jocn_a_00020 (2011).

46. Slovic, P., Finucane, M. L., Peters, E. & MacGregor, D. G. The affect heuristic. *Eur J Oper Res* **177**, 1333-1352, doi:DOI 10.1016/j.ejor.2005.04.006 (2007).

47. Eagleman, D. M. Visual illusions and neurobiology. *Nature reviews. Neuroscience* **2**, 920-926, doi:10.1038/35104092 (2001).

48. Goodale, M. A. & Haffenden, A. Frames of reference for perception and action in the human visual system. *Neuroscience and biobehavioral reviews* **22**, 161-172 (1998).

49. Hedgcock, W. & Rao, A. R. Trade-Off Aversion as an Explanation for the Attraction Effect: A Functional Magnetic Resonance Imaging Study. *J Marketing Res* **46**, 1-13 (2009).

Figure 1

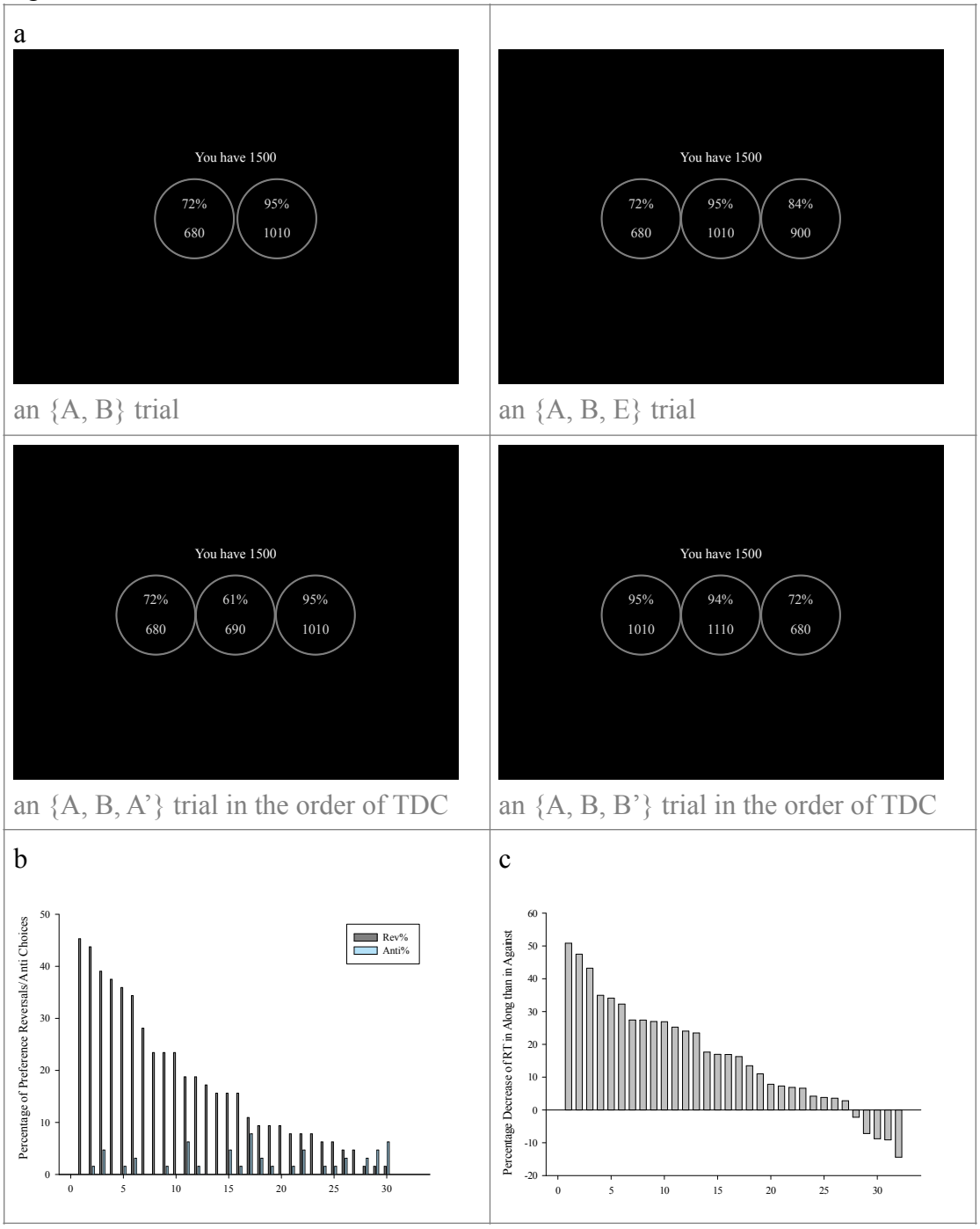

Figure 1a Sample screens from the experiment. Participants were endowed with 1500 New Taiwan Dollars (roughly 50 US dollars) to spend in each trial and made choices from a series of two- and three-item menus. The items on the menus were restaurant meals with specified prices and quality levels in which higher percentages imply better qualities (SI, S1.2 and S1.4). The trial in the top left corner is a two-item trial whereas that in the top right corner is a three-item trial. The two trials in the bottom panel are a matching pair of decoy trials. In this example, option A is the item with quality 72% and price $680; option B has quality 95% and price $1010; option E has

quality 84% and price $900; option A' has quality 61% and price $690; option B' has quality 94% and price $1110. An item is dominated by another item if it has both lower quality and higher price. Thus, A' is dominated by A but not by B, whereas B' is dominated by B but not by A. The target, denoted T, is the item that dominates the decoy, denoted D; there is no domination relationship between the competitor, C, and the decoy. Hence in the bottom left trial, the target is A (quality 72% and price $680), the decoy is A' (61%, $690) and the competitor is B (95%, $1010). In the bottom right trial, B (95%, $1010) is the target, B' (94%, $1110) the decoy and A (72%, $680) the competitor.

Figure 1b Percentage of preference reversals and anti choices. The figure is a participant-by-participant breakdown of the percentage of preference reversals and anti choices of decoy trials (in descending order of the former). Preference reversals occur when A is chosen from {A, B, A'} and B from {A, B, B'} whereas anti choices occur when B is chosen from {A, B, A'} and A from {A, B, B'}. The average percentage of preference reversals was 16.41 and that of anti choices was 2.05 over all 32 participants. Two participants had neither preference reversals nor anti choices.

Figure 1c Response times were shorter in Along trials than in Against trials. In consistent trials, the response times were significantly shorter when choices are along the decoy than when they are against the decoy. The figure is a participant-by-participant breakdown of the percentage decrease of RT in Along trials compared with Against trials (in descending order). The average percentage decrease of RT over all 32 participants was 16.20. The percentage decrease of RT is the difference of RT in Against trials and Along trials divided by the average of them in percentage.

Figure 2

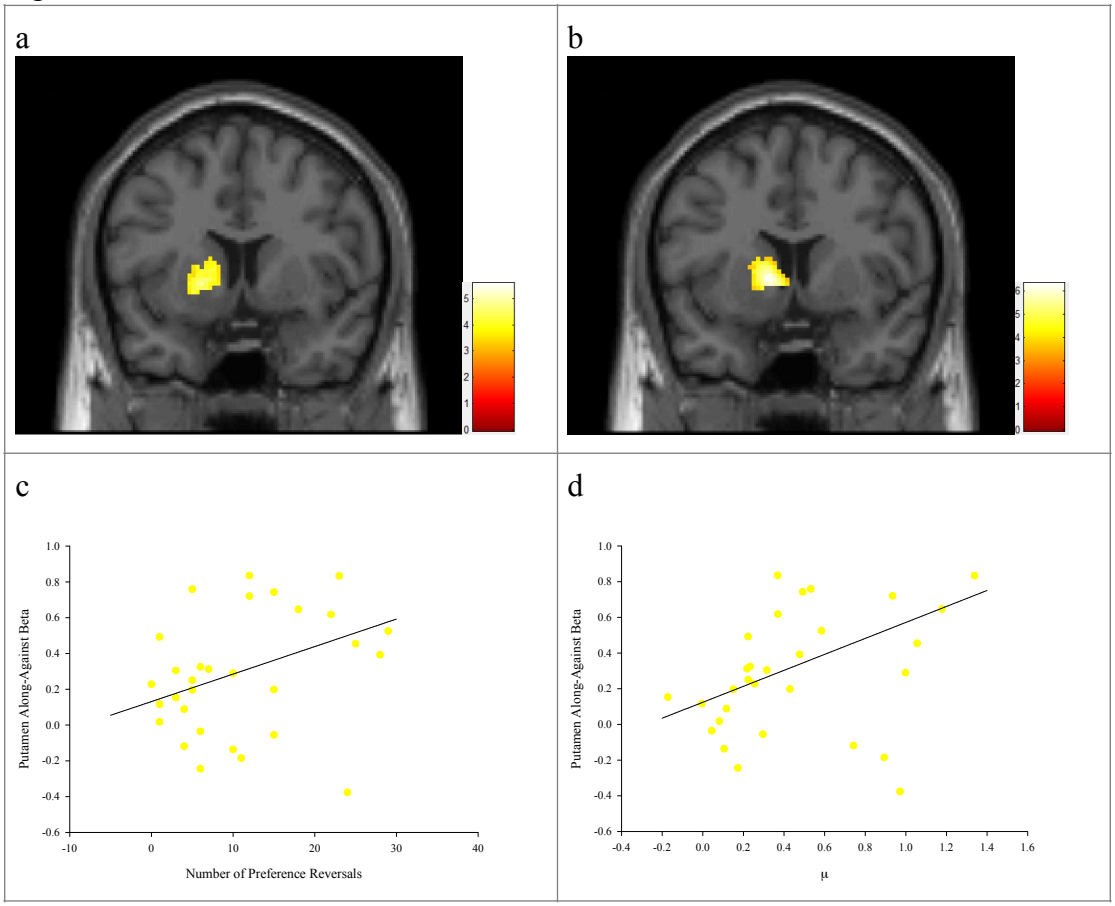

Figure 2a The left striatum was more active Along than Against in consistent trials (y=8) [P<0.05 at whole-brain cluster correction with a t threshold of 3.40 and an extent of 398 voxels] (SI, S4.1 and table S1).

Figure 2b The left striatum (y=8) correlated parametrically with the estimated intrinsic utility of the chosen option positively [P<0.05 at whole-brain cluster correction with a t threshold of 3.40 and an extent of 248 voxels] (SI, S4.2 and table S2).

Figure 2c Positive correlation between the number of preference reversals and the differential striatum activity Along versus Against. The robust regression slope is 0.015 and the p-value (two-tailed) is 0.057. The putamen activity is based on a 4-mm sphere centered at (-18, 8, 0). For other peak voxels of striatum, see SI, S4.3, table S5 and figure S2.

Figure 2d Positive correlation between the estimated μ and the differential striatum activity Along versus Against. The robust regression slope is 0.447 and the p-value (two-tailed) is 0.006. The putamen activity is based on a 4-mm sphere centered at (-18, 8, 0). For other peak voxels of striatum, see SI, S4.3, table S5 and figure S2.

Figure 3

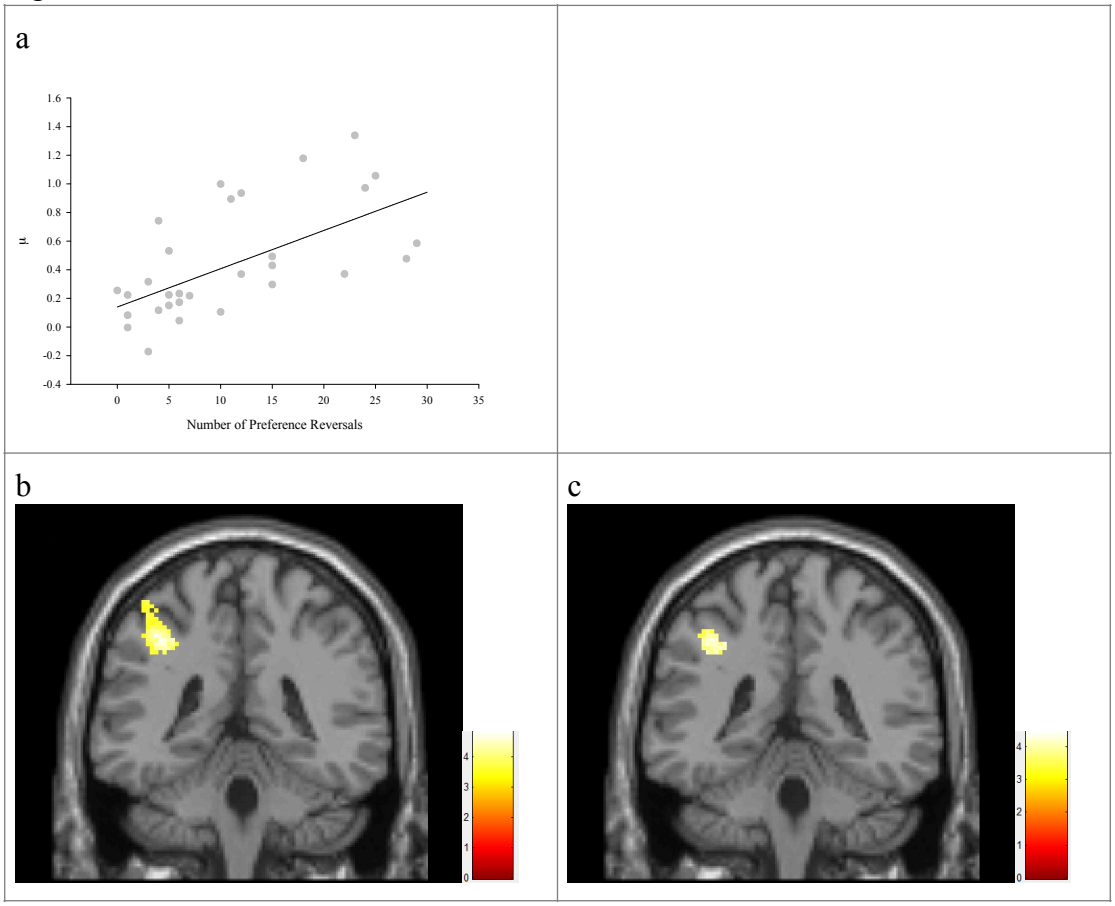

Figure 3a Strong positive correlation between the number of preference reversals and the estimated μ. The Pearson correlation coefficient is 0.610. The robust regression slope is 0.027 and the p-value (two-tailed) is 0.001. For other correlations, see SI, S4.5, table S6 and figure S3.

Figure 3b The left inferior parietal lobule (IPL) was more active in Weak trials than in Strong trials (y=-40) [P<0.05 at whole-brain cluster correction with a t threshold of 3.40 and an extent of 461 voxels] (SI, S4.6 and table S7).

Figure 3c The left IPL activity parametrically tracked the trial-by-trial decoy effect d(t) negatively (y=-40) [P<0.05 at whole-brain cluster correction with a t threshold of 3.40 and an extent of 185 voxels] (SI, S4.6 and table S8).

Figure 4

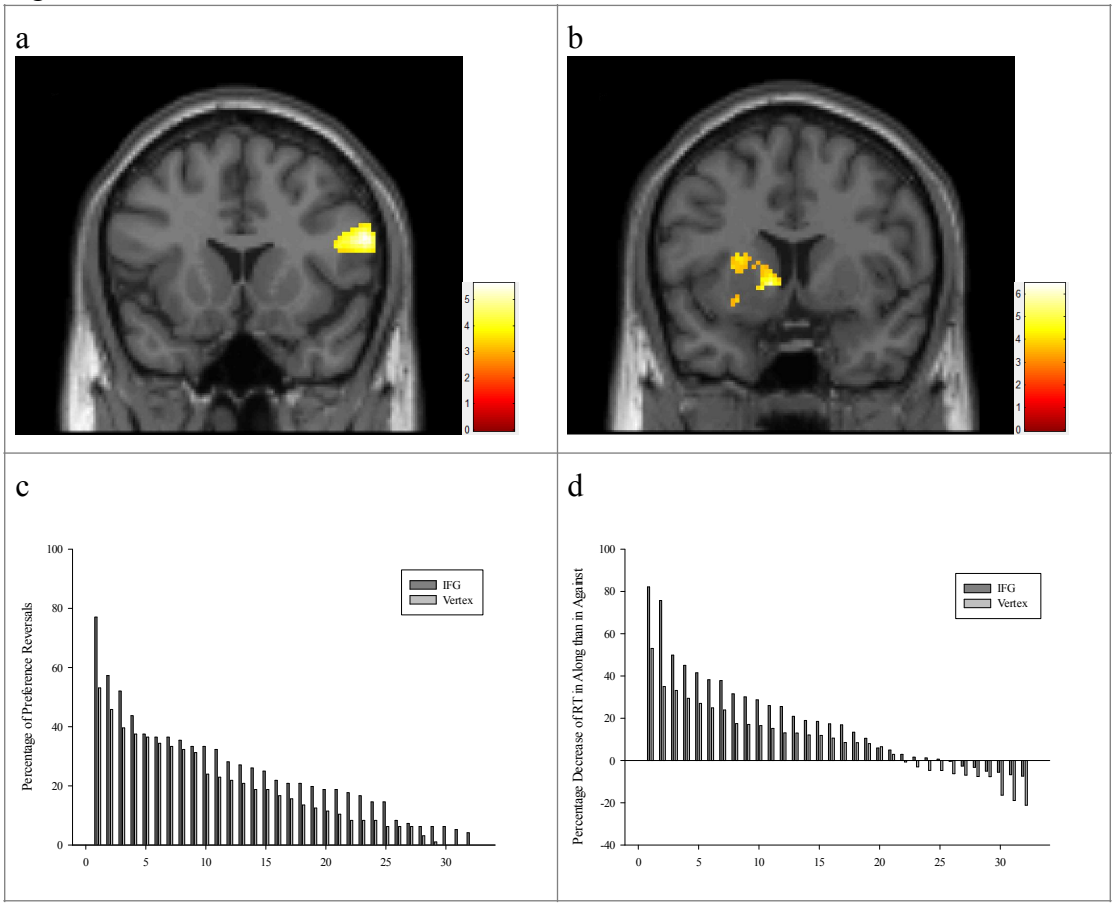

Figure 4a The right inferior frontal gyrus (rIFG) exhibited stronger functional connectivity with the IPL in Conflict trials than in NoConflict trials (y=16) [P<0.05 at whole-brain cluster correction with a t threshold of 3.40 and an extent of 342 voxels] (SI, S4.8 and table S13). The psychophysiological interaction is based on a 4-mm sphere of IPL centered at (-40, -40, 46).

Figure 4b The left striatum exhibited stronger functional connectivity with the rIFG in Against than in Along of Conflict trials (y=8) [P<0.05 at whole-brain cluster correction with a t threshold of 3.41 and an extent of 188 voxels] (SI, S4.9 and table S14). The psychophysiological interaction is based on a 4-mm sphere of rIFG centered at (60, 16, 22).

Figure 4c Percentage of preference reversals of the rIFG TMS group was larger than that of the vertex TMS group. The figure is a participant-by-participant breakdown of the percentage of preference reversals of decoy trials (in descending order), in the IFG group and in the vertex group, respectively. The average percentage of preference reversals was 25.30 for the IFG group and 18.72 for the vertex group. Three vertex participants had no preference reversals (SI, S5.7).

Figure 4d Percentage decrease of response time was larger in the rIFG TMS group than in the vertex TMS group. In consistent trials, the response times were significantly shorter when choices are along the decoy than when they are against the decoy, and this was even more so in the IFG group than in the vertex group. The figure is a participant-by-participant breakdown of the percentage decrease of RT in Along trials compared with Against trials (in descending order), of the IFG group and the vertex group, respectively. The average percentage decrease of RT over all IFG participants was 19.22 and that over all vertex participants was 9.05. The percentage decrease of RT is the difference of RT in Against trials and Along trials divided by the average of them in percentage (SI, S5.8).