Behavioral/Cognitive

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

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

[1]Department of Psychology, New York University, New York, New York 10003, [2]Department of Economics, Rutgers University, New Brunswick, New Jersey 08901, [3]Institute of Neuroscience, National Yang-Ming University, Taipei, 11221, Taiwan, [4]Department of Economics, National Taiwan University, Taipei, 10617, Taiwan, [5]Department of Electrical Engineering, National Taiwan University, Taipei, 10617, Taiwan, [6]Department of Economics, Washington University in St. Louis, St. Louis, Missouri 63130, and [7]Institute of Cognitive Neuroscience, National Central University, Taoyuan, 32001, Taiwan

Both humans and animals are known to exhibit a violation of rationality known as "decoy effect": introducing an irrelevant option (a decoy) can influence choices among other (relevant) options. 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, but may lead to biased choices in certain markets or experiments. We used fMRI and transcranial magnetic stimulation to investigate the neural underpinning of the decoy effect of both sexes. The left ventral striatum was more active when the chosen option dominated the decoy. This is consistent with the hypothesis that the presence of a decoy option influences the valuation of other options, making valuation context-dependent even when choices appear fully rational. Consistent with the idea that control is recruited to prevent heuristics from producing biased choices, the right inferior frontal gyrus, often implicated in inhibiting prepotent responses, connected more strongly with the striatum when subjects successfully overrode the decoy effect and made unbiased choices. This is further supported by our transcranial magnetic stimulation experiment: subjects whose right inferior frontal gyrus was temporarily disrupted made biased choices more often than a control group. Our results suggest that the neural basis of the decoy effect could be the context-dependent activation of the valuation area. But the differential connectivity from the frontal area may indicate how deliberate control monitors and corrects errors and biases in decision-making.

Key words: context dependence; decoy effect; inhibitory control

---

Significance Statement

Standard theories of rational decision-making assume context-independent valuations of available options. Motivated by the importance of this basic assumption, we used fMRI to study how the human brain assigns values to available options. We found activity in the valuation area to be consistent with the hypothesis that values depend on irrelevant aspects of the environment, even for subjects whose choices appear fully rational. Such context-dependent valuations may lead to biased decision-making. We further found differential connectivity from the frontal area to the valuation area depending on whether biases were successfully overcome. This suggests a mechanism for making rational choices despite the potential bias. Further support was obtained by a transcranial magnetic stimulation experiment, where subjects whose frontal control was temporarily disrupted made biased choices more often than a control group.

Introduction

In standard rational choice theory, decision-making is represented as assigning value to each feasible option and choosing the option with the highest value. Moreover, the value of an option should depend only on its intrinsic properties, not on the context in which it appears. Neuroscientific studies suggest that values of available options are indeed encoded (Padoa-

Schioppa and Assad, 2006), but there is mixed evidence on the extent to which these values depend on the context (Nieuwen-huis et al., 2005; Padoa-Schioppa and Assad, 2008). In this article, we use the well-known decoy effect to study context dependence in the human brain.

A decoy is an option that is unlikely to be chosen but may influence choice among other options. To illustrate, consider *The Economist* magazine's advertisement of three options: print-only and print-and-web subscriptions, each at 125 dollars per year, and web-only subscription at 59 dollars per year (Ariely, 2008). The print-only subscription was a decoy because it was dominated by print-and-web: it is clearly worse to get only print access than to get both print and web accesses at the same price. The domination was asymmetric in the sense that the decoy (print-only) was dominated only by one option, called the target (print-and-web), and not by the other, called the competitor (web-only). In the marketing literature, the target is a promoted product, the competitor a similar product not currently being promoted, and the decoy introduced to boost the demand for the target. Although consumers are not expected to choose the decoy, they might be impelled by the asymmetric domination to choose the target that could be favored because of the decoy instead of the competitor that might be penalized.

The marketing strategy contradicts standard rational choice theory, according to which adding an "irrelevant alternative" (the decoy) to the menu should not influence the values assigned to the target and the competitor. However, across taxa (such as humans, gray jays, rufous hummingbirds, ants, and even an amoeboid organism), the existence of a decoy does increase the frequency of choice of the target: there is a decoy effect on choice behavior (Huber et al., 1982; Shafir et al., 2002; Bateson et al., 2003; Latty and Beekman, 2011; Sasaki and Pratt, 2011). Despite its importance and prevalence in choice, the underlying neural basis for the decoy effect is poorly understood. Maintaining the assumption that decision-makers choose the option with the highest value, we will investigate the context dependence of values. Specifically, we hypothesize that, if the context (the menu from which options are chosen) includes a decoy option, then the valuation of the target is increased relative to the competitor. Such context-dependent valuations will necessarily deviate from the intrinsic values of the options, which raises the issue of how the brain can avoid biased choices.

We report two sets of experiments in which subjects made choices from menus that included asymmetrically dominated decoys. In the fMRI experiment, the first question we address is whether the value of the chosen option is influenced by the decoy. Such context-dependent valuation could potentially bias choices toward the target, causing a decoy effect on choice behavior. This raises the next question of how the brain can correct the potential bias. In addressing this question, we constructed a trial-by-trial measure of the strength of the decoy effect and explored potential control modulation from the frontal area for reducing the bias. Because the fMRI data only establish correlations, in the transcranial magnetic stimulation (TMS) experiment, we disrupted the frontal area found in the fMRI experiment to examine whether the decoy effect on choice behavior becomes more pronounced when control is weakened. Our results suggest that observable context-dependent choices or decoy effects may be shaped by the interaction of two processes: an evaluation process that is inherently context-dependent and a control system that can possibly override context dependence.

Materials and Methods

There were two experiments: the fMRI experiment and the TMS experiment. We describe them in turn.

The fMRI experiment

Subjects

Thirty-two right-handed healthy individuals (16 females; mean age 23.94 years) participated in the fMRI experiment approved by the Institutional Review Board of National Yang-Ming University. Each participated in two pretests, the scanning experiment, and a post-test. One trial from the scanning experiment was randomly drawn when a subject left the scanner, and she received a voucher and cash payment corresponding to her choice, in total worth of 1500 New Taiwan dollars (~50 U.S. dollars). Each subject was also paid 300 dollars for each of the pretests and the post-test, totaling 900 dollars.

Task

Subjects made choices from a series of two- and three-option menus. Each option was a restaurant meal with a specified price and a quality rating. The price was the cost of a typical meal at the restaurant. The price and the rating (with a higher percentage implying a better quality) were based on information from the Taiwanese gourmet website www.enjoygourmet.com. In each trial, the subject was endowed with 1500 dollars, from which the price of the meal she selects would be deducted. Four trials are illustrated in Figure 1A. For instance, the menu in the top left panel has two options. The left option is a meal at a restaurant with a 72% rating, at a price of 680 dollars. If she chooses this option, she will be given a voucher for a meal at a restaurant with the specified quality rating. She will receive what remains of her endowment (1500 − 680 = 820 dollars) in cash. The right option is a meal at a better restaurant; it has a 95% rating and costs 1010 dollars. Thus, the trade-off facing the subject is that she can spend more of her endowment to get a better meal. In some trials, one of the options on the menu was an asymmetrically dominated decoy. By definition, such a decoy is dominated in every dimension, in the sense that it is both more expensive and has a lower quality than the target. For instance, in the bottom left panel, the middle option is the decoy and the left-most option is the target. The right-most option is the competitor, which does not dominate the decoy. In the marketing literature, there are many types of decoys. Compromise decoys are also a common type. The idea is to include an undominated yet very unattractive decoy to make the target a compromise and hence chosen. However, because compromise decoys are not dominated and can be chosen with substantial probability, the irrelevance of them is more questionable.

Experimental design and statistical analysis

The experiment consists of three stages: the pretest, the post-test, and the scanning experiment. Each pretest and post-test consisted of 192 trials with two available options. They served to identify how each subject traded off price versus quality without the influence of decoys. The measured trade-off from the pretests was used to tailor-make each subject's decision problems in the scanning experiment. The post-test was administered right after the scanning experiment. The measured trade-off from the post-test was used to estimate each subject's utility function as explained in *Context-dependent values* below.

In the scanning experiment, the main focus was on decoy trials with three available options: two Options A and B, neither dominating the other, and a third decoy option. The decoy trials were paired so that each undominated option would be the target in one trial and the competitor in the other. Such a matching pair of decoy trials has the form {A, B, A'} and {A, B, B'}. In {A, B, A'}, the decoy A' is dominated by A (the target) but not by B (the competitor), whereas in {A, B, B'}, the decoy B' is dominated by B (the target) but not by A (the competitor). The choice is said to be along (Along) the decoy if the target is chosen (e.g., A is chosen from {A, B, A'}) and against (Against) the decoy if the competitor is chosen (e.g., A is chosen from {A, B, B'}).

There were 192 trials in the scanning experiment. Each pair of Options A and B appeared six times: once in a two-option choice problem {A, B}; twice in a three-option choice problem {A, B, A'}; twice in a three-option choice problem {A, B, B'}; once in a three-option choice problem {A, B,

A
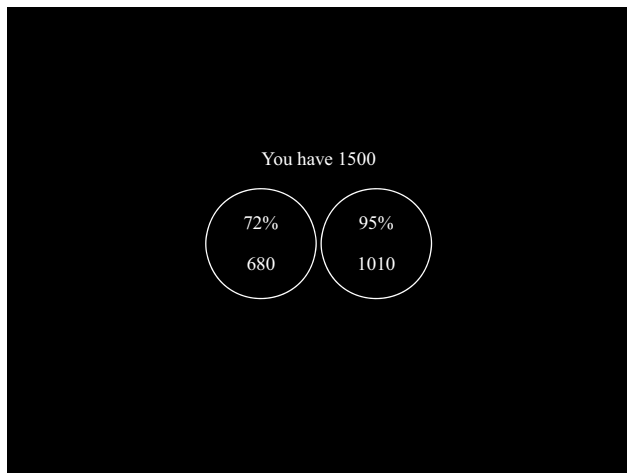
an {A, B} trial
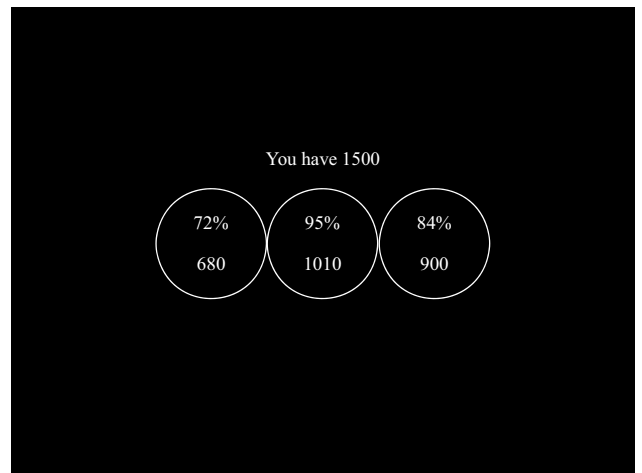
an {A, B, E} trial
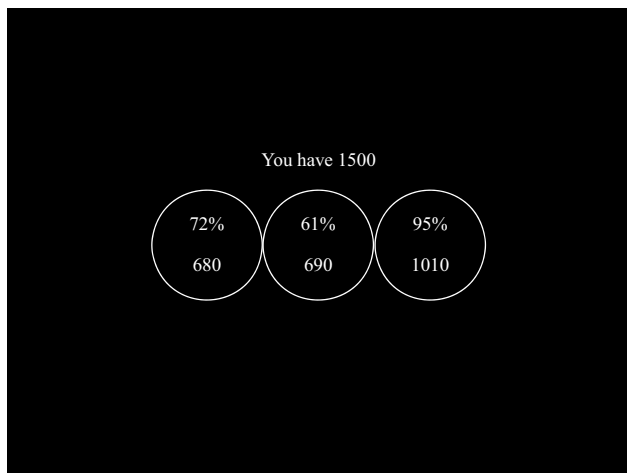
an {A, B, A'} trial in the order of AA'B or TDC
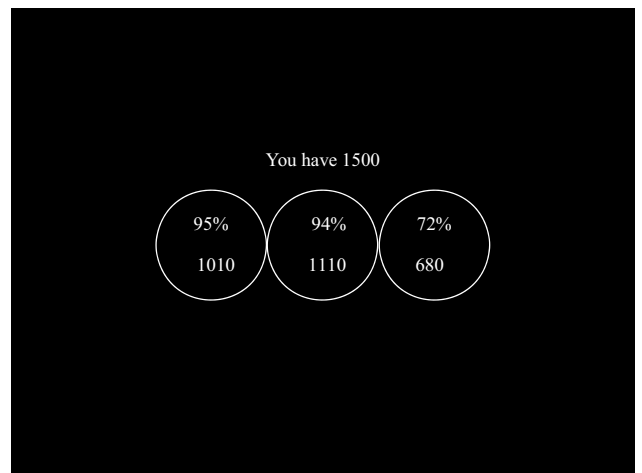
an {A, B, B'} trial in the order of BB'A or TDC

B
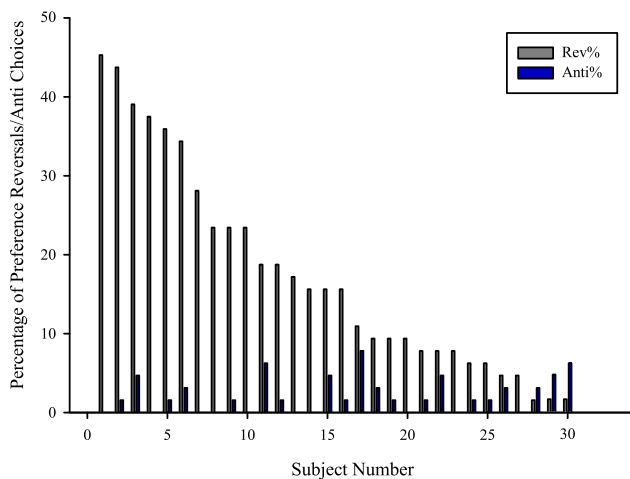
C
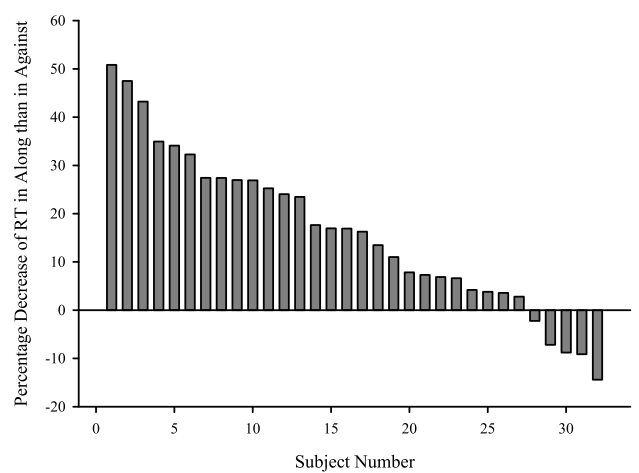
Figure 1. *A*, Sample screens from the experiment. Subjects were endowed with 1500 New Taiwan Dollars to spend in each trial and made choices from a series of two- and three-option menus. The options on the menus were restaurant meals with specified prices and quality levels in which higher percentages imply better qualities. The top left is a two-option trial and the top right is a three-option trial. The bottom two are a matching pair of decoy trials. In this example, Option A has 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 option is dominated by another option (*Figure legend continues.*)

E} where E was an undominated third option. By including {A, B, E} trials with three undominated options, we avoided a situation where all three-option decision problems had an asymmetrically dominated option, which would have been too repetitive and predictable. The order of the trials and the ordering of options in a trial were randomized, with one exception: for every {A, B, A′} trial, there was a matching {A, B, B′} trial in which the ordering of the options on the screen (in terms of target, competitor, and decoy) was controlled to be the same, so that the only difference between these two trials was that the decoy changed and the roles of A and B (in terms of being target or competitor) were reversed (as illustrated in Fig. 1A, two bottom panels). The similarity of the two matching trials could potentially reduce the decoy effect.

The scanning experiment was self-paced. In each trial, pictures of foods (to remind subjects the experiment involved restaurant meals) were presented for 0.5 s, after which a fixation cross (randomized between 0 and 2 s) appeared. The choice screen followed and was on until the subject made her choice. Finally, the fixation cross appeared again during the jittered intertrial interval (randomized between 2 and 6 s) before the next trial began.

For each subject, we obtained contrasts in which we are interested. Parameter estimates from contrasts in single-subject models were entered into the second-level random-effect analysis to perform a one-sample t test on whether the activation of a contrast was significantly different from zero after whole-brain cluster correction. The criterion we used in all analyses is cluster-defining threshold uncorrected p value of 0.001 and familywise error rate of 0.05. We also verified our key results are robust using nonparametric cluster-level inference (Eklund et al., 2016) (data not shown). In all GLMs and psychophysiological interaction (PPI) analysis 1 indicated below, the degree of freedom is 29. In PPI 2, it is 28.

fMRI data acquisition and preprocessing

Imaging was performed using a 3T Siemens Trio MR scanner. An EPI sequence was used to acquire the functional data (TR = 2000 ms, TE = 30 ms, flip angle = 90°, voxel size = 3.4 mm × 3.4 mm × 3.4 mm). The functional (T2*-weighted, BOLD) images were collected in 33 transverse planes parallel to the anterior commissure-posterior commissure line to cover the whole brain. There were four EPI sessions so that each session had 192/4 = 48 trials. A high-resolution T1-weighted MP-RAGE anatomical MRI volume scan of the entire head was run once on each subject (TR = 2530 ms, TE = 3.03 ms, flip angle = 7°, voxel size = 1 mm × 1 mm × 1 mm).

Data analysis was performed using SPM8. The functional images were corrected for differences in slice-acquisition time, realigned, coregistered, normalized to the MNI average template, and smoothed. The first four images were discarded from the analysis for magnetization to reach steady state. Two subjects were removed due to excessive head motions.

fMRI data analysis

The fMRI analysis has two: decoy-induced changes in valuations, and how these changes might be overcome to restore rational (unbiased) choice.

←

(*Figure legend continued.*) 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 option 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) is the decoy and A (72%, $680) is the competitor. *B*, Percentage of preference reversals and anti-choices. The figure is a subject-by-subject breakdown of the percentage of preference reversals and anti-choices of decoy trials. We number subjects 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′}. Two subjects had neither preference reversals nor anti-choices. *C*, Shorter RTs in Along than in Against behaviorally rational pairs. The figure is a subject-by-subject breakdown of the percentage decrease of RT in Along than in Against behaviorally rational pairs. We number subjects in descending order. The percentage decrease of RT is the difference of RT in Against trials and Along trials divided by the average of them in percentage.

Context-dependent values. We first addressed the hypothesis that the valuation of the chosen option is influenced by the decoy. We estimated a linear utility function from choices in the post-test with only two available options. In the absence of decoys, this utility function reflects the intrinsic values of the options. We then constructed a comprehensive model where various value signals and an indicator of Along versus Against trials are included. Our aim was to find regions correlating with valuations and to see how the value signals differentiate between Along and Against trials.

To estimate a linear utility function, consider a trial {A, B} in the post-test where the price of A (resp. B) is P_A (resp. P_B) and the quality is Q_A (resp. Q_B). The intrinsic utility of choosing A is $u(A) = b_P(1500 - P_A) + b_Q Q_A$, where b_P and b_Q are weights put on the cash payment $1500 - P_A$ and the quality Q_A. This is called intrinsic utility because it only depends on the price and quality of the option. The intrinsic utility of B is $u(B) = b_P(1500 - P_B) + b_Q Q_B$. Using the logit function, the probability of choosing A is as follows:

$$\frac{1}{1 + \exp^{-\lambda(u(A) - u(B))}}$$

where λ is the precision parameter. We normalized so that $b_P + b_Q = 1$, and estimated b_P and λ by maximum likelihood. We then used the estimated b_P to calculate the intrinsic utility of each option for each trial of the scanning experiment.

We ran GLM 1 with three regressors of interest. They are convolved with the canonical form of the hemodynamic function and will not be mentioned later on. The first regressor is to model the half-second when subjects saw pictures of foods (called picture). The second is to model the motor response when subjects made choices (called motor). The third is to model the entire duration of the choice. To account for the values of each option and whether choice is Along or Against in every trial, we included six parametric modulations for the third regressor.

We attempted to control for those value signals that would most likely be behaviorally relevant. The value of the chosen option is certainly important and is defined for all trials. Undominated unchosen options may also influence the decision process. For two-option trials and decoy trials, we control for the value of the unique undominated unchosen option. In {A, B, E} trials, there are actually two undominated unchosen options: the one with the highest value is the one most likely to influence the decision process by competing with the chosen option, but we cannot rule out that the worst unchosen option also plays a role. Therefore, for {A, B, E} trials, we separately include the values of both the best and the worst unchosen options. Finally, the value of the decoy is also controlled for.

Specifically, the six parametric modulations are as follows. The first is on the utility of the chosen option (defined for all trials). The second is on the utility of the undominated unchosen option in two-option trials and decoy trials (taking the value of 0 for {A, B, E} trials). The third is on the utility of the highest valued unchosen option in {A, B, E} trials (taking the value of 0 for two-option and decoy trials). The fourth is on the utility of the decoy in decoy trials (taking the value of 0 for two-option and {A, B, E} trials). The fifth is on the utility of the lowest valued unchosen option in {A, B, E} trials (taking the value of 0 for two-option and decoy trials). The sixth parametric modulation is an indicator of Along versus Against (taking the value of 1 when choice is along the decoy, −1 when choice is against the decoy, and 0 for two-option and {A, B, E} trials). We shut down the orthogonalization procedure in SPM. We thus made it harder to obtain significance because the unorthogonalized value modulations are correlated.

Last, the six head motion parameters are included as the regressors of no interest. They are always included in all GLMs and will not be mentioned later on. The contrast of the first parametric modulation allows us to find brain correlates of the utility of the chosen option, other utility-related measures being controlled for. Similarly, the sixth modulation allows us to test which brain region is more active when choice is along (rather than against) the decoy, again controlling for the possible influence of the utility-related measures. This amounts to a simple additive effect of the context. For between-subject correlation of striatal Along-Against activity and two behavioral measures of how "decoyable" a subject is,

we extract the coefficients of the Along-Against contrast within a 4 mm sphere at three peak voxels of putamen and one peak voxel of caudate.

Overcoming the decoy effect. The results in *Context-dependent values* are consistent with the hypothesis that context-dependent values deviate from intrinsic utilities. This may bias choices in favor of the target, which raises the question of whether the brain has some mechanism for correcting the potential bias. We hypothesized that regions related to this mechanism would, on the one hand, correlate negatively with the strength of the decoy effect and, on the other hand, would differentially connect with the valuation area depending on whether or not the decoy effect was successfully overridden. We constructed a trial-by-trial measure of the strength of the decoy effect $d(t)$ (see equation (1) below) and identified brain regions correlating negatively with it. If these regions are part of a mechanism for mitigating a decoy-induced bias, then because context-dependent values suggested that decoy-induced changes in valuations may be reflected in differential striatal activity, the regions identified should differentially connect with the striatum depending on whether the decoy effect is successfully overridden.

To investigate which brain areas correlate with $d(t)$, we ran GLMs 2 and 3. In GLM 2, for each subject, we classified the decoy trials into two halves according to the size of $d(t)$. The half with a stronger decoy effect (a larger $d(t)$) was classified as Strong. The half with a weaker decoy effect (a smaller $d(t)$) was classified as Weak. We include six regressors of interest. The binary classification gives Strong and Weak as two regressors. The next two are on trials with only two options to choose from (classified as two-option trials) and on trials with three undominated options (classified as three-undominated-option trials). These four regressors model the entire duration of the choice. Last, picture and motor regressors are added. The final four regressors (the first modeling two-option trials, the second modeling trials with three undominated options, the picture and the motor regressors) will be called the four standard control regressors.

To find regions coding $d(t)$ parametrically, we tried GLM 3. There are six regressors of interest. The first is a regressor modeling the decoy trials. For this regressor, we included a parametric modulation $d(t)$. These two regressors model the entire duration of the choice. We also included the four standard control regressors.

The binary classification (Weak vs Strong) in GLM 2 is coarser and depends less on the functional form of (1) than the parametric modulation in GLM 3. We identified the inferior parietal lobule (IPL) in both. We also performed several robustness checks on IPL's negative correlation with $d(t)$, including allowing for the possibility that the decision bound T in (1) might have changed during the scanning experiment and measuring the strength of the decoy effect in two model-free ways. We also explored a model of random utility and found its neural correlates.

The fact that IPL negatively correlates with $d(t)$ suggests that it may play a role in a control mechanism for reducing the decoy effect. Because context-dependent values indicated that the decoy effect is reflected in differential striatal activity, we explored how IPL might differentially connect with the striatum depending on whether the decoy effect is successfully reduced. We did not find a direct modulation from IPL in our data. We did find a plausible indirect modulation through the right inferior frontal gyrus (rIFG), which we will further verify by a TMS experiment. The indirect modulation is based on two PPIs.

PPI 1 between IPL and rIFG
A decoy trial {A, B, A′} is a Conflict trial if $u(A) - u(B) < 0$, so the competitor is intrinsically better. It is a NoConflict trial if $u(A) - u(B) > 0$, so the target is intrinsically better. Exerting control to reduce the decoy effect is important in Conflict trials, to prevent the intrinsically inferior target from being chosen. In NoConflict trials, control is less important because the option favored by the decoy effect, the target, is anyway the intrinsically better option.

We hypothesized that, when IPL is active and control is more important, the connectivity between IPL and a control area might be stronger. We took the average activity of a 4 mm sphere surrounding the IPL peak voxel $(-40, -40, 46)$ identified in GLMs 2 and 3 as the seed. First, we estimated a GLM, which was built solely for PPI 1. There are six regres-

sors of interest. We have Conflict trials and NoConflict trials as two regressors. They model the entire duration of the choice. We also included the four standard control regressors. Second, the time series of the average activity of the 4 mm IPL sphere was extracted, removing the variances associated with the head motion parameters. Third, we ran PPI 1 with the psychophysiological interaction of the seed activity and the contrast of Conflict-NoConflict. The psychological contrast Conflict-NoConflict and the bold eigenvariate were also included as regressors.

PPI 2 between rIFG and striatum
We hypothesized that, in trials where control is important, the connectivity between control and valuation areas may be stronger when control is successful compared with when control is unsuccessful. As argued in PPI 1, control is important in Conflict trials, to make sure that the intrinsically best option (the competitor) is chosen against the decoy. Hence, in Conflict trials, control is successful if the competitor is chosen (i.e., in Against trials). It is unsuccessful if the target is chosen (i.e., in Along trials).

To explore the hypothesis, we took the average activity of a 4 mm sphere surrounding the peak voxel $(60, 16, 22)$ of rIFG identified in PPI 1 as the seed. First, we estimated a GLM that was built solely for PPI 2. There are seven regressors of interest. There are mainly two regressors for Conflict trials: one for Against trials (SuccessfulControl) and one for Along trials (UnsuccessfulControl). In addition, a regressor for NoConflict trials is added. These three regressors model the entire duration of the choice. We also included the four standard control regressors. Second, the time series of the average activity of the 4 mm rIFG sphere was extracted, removing the variances associated with the head motion parameters. Third, we ran PPI 2 with the psychophysiological interaction of the seed activity and the contrast of SuccessfulControl-UnsuccessfulControl. The psychological contrast SuccessfulControl-UnsuccessfulControl and the bold eigenvariate were also included as regressors. The results were based on 29 subjects because one subject managed to always choose the intrinsically better option. Eight subjects always chose the intrinsically better options in some EPI session, so the contrast was scaled by the number of sessions in which it could be built. The scaled contrasts were then entered into the second level random-effect analysis.

The TMS experiment
Subjects
Sixty-four healthy individuals (25 females; mean age 22.70 years) participated in the TMS experiment approved by the Institutional Review Board of Taipei Veterans General Hospital. The TMS experiment closely resembled the fMRI experiment. Each subject participated in a pretest, the TMS session, and a post-test. One trial from the TMS session was randomly drawn, and the subject received a voucher and cash payment corresponding to her choice, in total worth of 1500 New Taiwan dollars. Each subject was also paid 300 dollars for the pretest and the post-test, totaling 600 dollars.

Task
The task is the same as that in the fMRI experiment. Subjects made choices from a series of two- and three-option menus.

Experimental design and statistical analysis
The pretests and post-tests are the same as those for the fMRI experiment. The post-tests were administered 1.5 h after TMS was delivered, when the effect should have expired (Huang et al., 2005). The TMS session closely resembles the fMRI experiment, except that all trials have three options. There were 216 trials. Among them, 192 trials are decoy trials in matching pairs. The remaining 24 trials are {A, B, E} trials. The main focus of the analysis is on the matching pairs. Because the effect of TMS is estimated to be 1 h at most, we did not include two-option trials to have as many decoy trials as possible. A few {A, B, E} trials were included to avoid the task being too repetitive and predictable.

The TMS session was self-paced. In each trial, pictures of foods were presented for 0.5 s, after which a fixation cross of 0.25 s appeared. The choice screen followed and was on until the subject made her choice.

Finally, the fixation cross appeared again during the jittered intertrial interval (randomized between 2 and 3 s) before the next trial began.

The degree of freedom for between-group *t* test is 1. That for within-group *t* test is 31.

TMS parameters and sites

We applied theta burst TMS before subjects started the choice task. The TMS pulses were applied in a continuous theta burst TMS pattern as specified by a previous study (Huang et al., 2005). Because this theta burst TMS protocol can induce longer effects on neural activities, it is especially suitable for our current experimental design due to long response time (RT). Subjects received 3 pulses of stimulation at 50 Hz, which were repeated every 200 ms for 40 s at 40% of maximum output of 2.53 tesla. This approximately corresponds to 47%–73% of typical Taiwanese subject motor thresholds. We used a customized figure-of-8 flat coil of 55 mm and a Rapid2 Stimulator (Magstim) to deliver the pulses. A fixed stimulation level was used because it has been proven successful and replicable in many studies and over a wide range of tasks (Kim et al., 2011). Moreover, motor cortex excitability may not provide a good guide to TMS thresholds in other cortical areas (Stewart et al., 2001).

According to the fMRI result, rIFG could be playing a role in implementing control to reduce decoy-induced decision-making bias. Yet because IPL functionally connects more strongly with rIFG in Conflict than in NoConflict trials, potentially, both rIFG and IPL could be our region of interest for the TMS experiment. In the literature, rIFG is often implicated in overriding irrelevant responses. Moreover, IFG and its vicinity have stronger anatomical and functional connections with the basal ganglion than IPL (Manes et al., 2014), and we argue that the striatum might be related to the decoy effect due to its differential Along versus Against activity; we therefore chose rIFG as the region of interest. If rIFG is temporarily disrupted, then the inhibitory control is expected to be weaker and the decoy effect is expected to be stronger.

The design is between-group. Subjects were randomly assigned to two groups so that each group has 32 subjects, the same number as we had for the fMRI experiment. In one group, the site of stimulation was at the peak voxel of rIFG (60, 16, 22) identified from the fMRI experiment. The site was determined by each subject's T1-weighted anatomical image with Brainsight software (Rogue Research). To determine the correspondence between the location on the skull and the site of stimulation on the T1 image, subjects wore goggles with a mounted Polaris infrared tracker (Northern Digital) attached to coregister them with their structural image. This helps pinpoint the skull point overlying the stimulated region to be identified. In the other group, the site of stimulation was the vertex. The vertex is a point midway between the nasion and the inion and of equal distance to the right and left intertragal notches (Muggleton et al., 2008). It serves as a control site for the experiment. To make sure the procedures experienced by vertex subjects were exactly the same as for IFG subjects, we also used the Brainsight pen tool to point out (and localize) the vertex for stimulation. During stimulation over rIFG, the coil handle was oriented toward the superior portion of the head, whereas the coil handle was oriented toward the right side of the head during stimulation over the vertex.

Results

Behavior: choices and RT

In terms of choices, 81.20% of the matching pairs of trials were behaviorally rational: the subject chose the same option from {A, B, A′} and {A, B, B′}; 16.41% exhibited preference reversals: the subject chose the target in both trials (i.e., she switched from A to B when the decoy changed from A′ to B′). These numbers are in line with previous studies, which typically have preference reversals occurring ~20% of the time. Only 2.05% of the matching pairs exhibited anti-choices: the subject chose the competitor in both trials (i.e., she switched from B to A when the decoy changed from A′ to B′). Thus, when choices were not behaviorally rational, there was a strong tendency to favor the target: A in {A, B, A′} and B in {A, B, B′}. The subject-by-subject breakdown

of the percentage of preference reversals and anti-choices is shown in Figure 1*B*.

Because the great majority of matching pairs were behaviorally rational, choices in this experiment were largely context-independent, but this does not necessarily imply context-independent valuations. A decoy-induced increase in the value of the target would not influence choice if the target has a higher intrinsic utility than the competitor. Even if the intrinsic utility of the target is lower, the decoy effect will not influence choice unless it is large enough to overcome the intrinsic utility difference that favors the competitor. However, neural activity can indicate changes in values that do not manifest themselves in choice behavior. The RT difference in behaviorally rational pairs provides the first evidence that values may depend on the context even when choices do not. Because a behaviorally rational pair consists of an Along trial (where the choice is along the decoy) and an Against trial (where the choice is against the decoy), if the same chosen option is valued more when it is the target than when it is the competitor, choosing this option will be easier in the Along trial than in the Against trial. Across all subjects, the average RT was 8.36 s in Along and 9.94 s in Against behaviorally rational trials, indicating that it was indeed easier to make a decision in the former situation. The average percentage decrease of RT in Along trials is 16.20%. The average percentage decrease of RT in Along than in Against behaviorally rational trials is strongly significantly different from zero over all 32 subjects, having the *p* value (two-tailed) of 0.0000052 (Fig. 1*C*).

Neural signature of a decoy effect in the valuation area

We aim to find a neural manifestation of the decoy effect. If valuations are context-independent, then after controlling for the possible influence of various value signals, the valuation area should not differentiate between Along and Against trials. On the other hand, if valuations are context-dependent, then options might be more highly valued when they dominate the decoy. Hence, reward-sensitive areas are predicted to be more active in Along than in Against trials. This prediction was supported by our fMRI data. In GLM 1, we included various value signals and an indicator of Along or Against trials. We found that the striatum activity correlates parametrically with the estimated intrinsic utility of the chosen option positively (Fig. 2*A*; Table 1). Also, the left ventral striatum was more active in Along than in Against trials (Fig. 2*B*; Table 2). The striatum region correlating with the value of the chosen option overlapped with that more active in Along than in Against trials (Fig. 2*C*). That the same neural structure correlating with the value of the chosen option is more active when the decoy favors the chosen option is consistent with context-dependent valuations.

If the striatal differential activity Along versus Against indicates that valuations are biased in favor of the target, then subjects whose striatal activity differentiates more strongly between Along and Against trials should exhibit a stronger decoy effect on choice behavior. Across subjects, we indeed find that the striatal Along versus Against activity is positively correlated with two behavioral measures of how "decoyable" a subject is (Table 3). The first such measure is simply the number of preference reversals (Fig. 2*D*). The second measure μ is the average strength of the decoy effect as explained below (Fig. 2*E*).

For completeness, we also report regions correlating with the utility difference of the chosen option and the undominated unchosen option; the latter competes with the chosen option and presumably influences decision-making (Table 4). Further, we
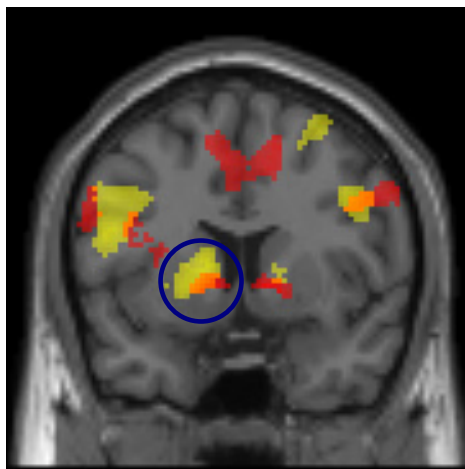
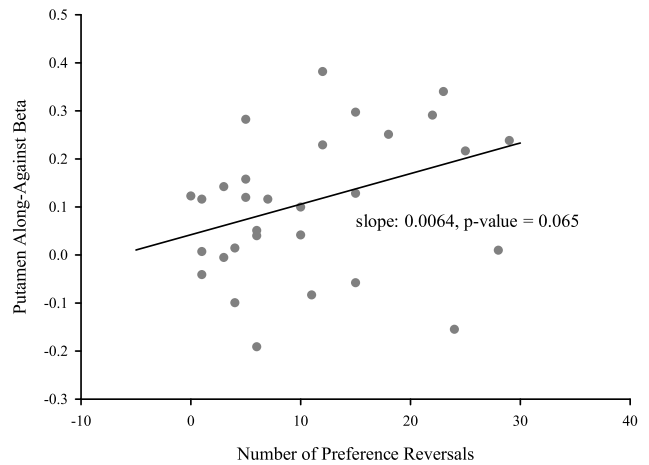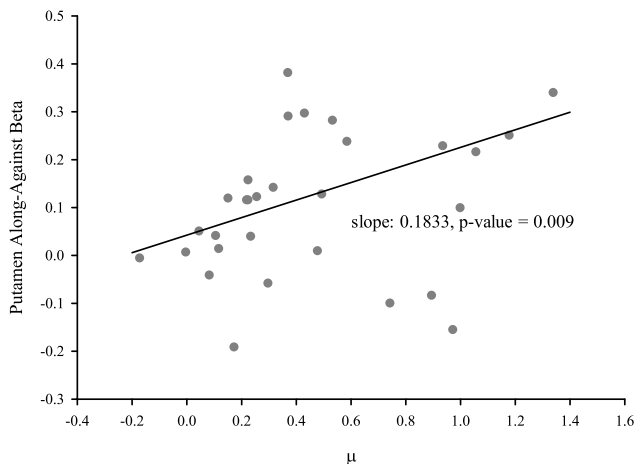**Figure 2.** **A**, The left striatum (shown at $y = 8$) correlated parametrically with the intrinsic utility of the chosen option positively ($p < 0.05$ at whole-brain cluster correction). **B**, The left striatum was more active in Along than in Against trials (shown at $y = 8$) ($p < 0.05$ at whole-brain cluster correction). **C**, Overlap. The left striatum was more active in Along than in Against trials (shown at $y = 8$) (shown in yellow at $p < 0.001$, uncorrected) and correlated parametrically with the intrinsic utility of the chosen option positively (shown in red at $p < 0.001$, uncorrected). The cluster of the left striatum is circled. Orange represents the overlapping area. **D**, Positive correlation between the number of preference reversals and the differential striatal Along versus Against activity. The robust regression slope is 0.0064; $p = 0.065$ (two-tailed). The putamen activity is based on a 4 mm sphere centered at ($-20, 12, 2$). **E**, Positive correlation between the average strength of the decoy effect μ and the differential striatal Along versus Against activity. The robust regression slope is 0.1833; $p = 0.009$ (two-tailed). The putamen activity is based on a 4 mm sphere centered at ($-20, 12, 2$).

Table 1. Regions parametrically correlating with the utility of the chosen option positively (GLM 1)[a]

Cluster size	Brain region	MNI coordinate			z value
		x	y	z	
7324	Left premotor and striatum				
	BA 6	−58	2	38	5.18
	BA 9	−58	4	26	5.06
	BA 4	−36	−30	58	4.80
	BA 24	−2	0	46	4.40
	SMA	14	0	64	4.29
	Caudate	−12	10	0	4.79
	Putamen	−16	8	0	3.74
347	Medial frontal gyrus	−8	50	−14	4.03
985	Middle frontal gyrus				
	BA 9	48	2	34	4.35
	BA 6	58	2	42	4.27
337	Precentral gyrus				
	BA 6	30	−14	56	3.95
323	Postcentral gyrus				
	BA 2	50	−18	26	3.72
277	Middle occipital gyrus				
	BA 19	52	−70	4	4.04
4157	Occipital cortex				
	BA 19	−24	−78	24	4.32
	Cuneus	20	−84	24	4.30
	BA 18	−4	−68	2	4.09

[a]Local maxima of clusters with $p < 0.05$ at whole-brain cluster correction with a t threshold of 3.40 and an extent of 277 voxels. There is no region parametrically correlating with the utility of the chosen option negatively.

Table 2. Regions more active in Along than in Against trials (GLM 1)[a]

Cluster size	Brain region	MNI coordinate			z value
		x	y	z	
1587	Left IFG				
	BA 46	−40	42	10	4.9
	BA 9	−50	6	22	4.4
	BA 10	−36	34	8	4.27
	BA 6	−30	−6	56	3.71
910	Right IFG				
	BA 46	40	32	20	4.17
	BA 10	38	40	12	4.16
	BA 9	42	4	34	4.11
549	Striatum/insula				
	Insula	−28	18	4	4.79
	Caudate	−12	8	8	4.22
	Putamen	−20	8	0	3.86
530	Premotor cortex				
	BA 6	22	−6	56	3.96
	BA 6	32	4	62	3.81
	BA 6	30	18	58	3.77
1356	Right IPL				
	BA 40	34	−60	42	5.42
	BA 40	40	−56	50	4.59
	Precuneus	12	−70	50	4.39
917	Left IPL				
	BA 40	−30	−64	46	4.84
	BA 40	−38	−58	48	4.42
	BA 40	−44	−44	46	4.09

[a]Local maxima of clusters with $p < 0.05$ at whole-brain cluster correction with a t threshold of 3.40 and an extent of 530 voxels. There is no region more active in Against than in Along trials.

Table 3. Positive correlation between the number of preference reversals and the differential striatal activity Along versus Against and positive correlation between μ and the differential striatal activity Along versus Against[a]

	(−20, 12, 2)	(−22, 10, −2)	(−24, 14, −2)	(−12, 14, 4)
Reversals				
Slope	0.0064	0.0076	0.0051	0.0045
p value	0.065*	0.025**	0.063*	0.141
μ				
Slope	0.1833	0.1319	0.1047	0.2010
p value	0.009***	0.099*	0.093*	0.003***

[a]The striatum activity is based on a 4 mm sphere centered at putamen (−20, 12, 2), putamen (−22, 10, −2), putamen (−24, 14, −2), and caudate (−12, 4, 4) (from left to right). The slopes of the robust regressions and their associated p-values (two-tailed) are provided.
*Significance at 10% level; **Significance at 5% level; ***Significance at 1% level.

Table 4. Regions parametrically correlating with the intrinsic utility difference of the chosen option versus the undominated unchosen option positively (GLM 1)[a]

Cluster size	Brain region	MNI coordinate			z value
		x	y	z	
481	Premotor cortex				
	BA 6	−58	0	38	4.7
2039	Sensorimotor cortex				
	BA 4	−28	−16	60	4.1
	BA 3	−46	−20	62	4.0
	BA 6 (medial)	−4	2	42	3.9
888	Visual cortex				
	Cuneus	14	−84	18	3.7
	Cuneus	−12	−68	4	3.5
511	Inferior frontal gyrus				
	BA 9	48	2	34	4.2
	BA 6	58	2	44	4.0

[a]Local maxima of clusters with $p < 0.05$ at whole-brain cluster correction with a t threshold of 3.40 and an extent of 481 voxels. There is no region parametrically correlating with the intrinsic utility difference of the chosen option versus the undominated unchosen option negatively.

find no brain region correlating with the utility sum of the chosen option and the undominated unchosen option.

Trial-by-trial strength of the decoy effect

How well an option 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, as suggested by the higher striatal activity in Along than in Against

trials, choices may become biased toward the target and an intrinsically worse option may be chosen. Unbiased choice may therefore require some cognitive control. To find out how cognitive control could be recruited and its potential effect on valuation, we first constructed a measure of the trial-by-trial strength of the decoy effect using a simplified linear ballistic accumulator (LBA) model (Brown and Heathcote, 2008). The number of preference reversals of a subject can be used to compare between-subject who is more decoyable but cannot be used to determine within-subject when she experiences a stronger decoy effect. For this, we need a trial-by-trial measure. We hypothesized that regions related to a putative control mechanism will correlate negatively with the trial-by-trial strength of the decoy effect. Moreover, if the decoy-induced change in valuation manifest itself in differential striatal activity, regions related to control may exhibit differential connectivity with the striatum depending on whether the bias induced by decoys is successfully overridden. We will explore these hypotheses after discussing the LBA model and the trial-by-trial measure.

The idea behind the LBA model is that the time to make a decision is inversely proportional to the "evidence" in favor of the decision. In a decoy trial, because the decoy is dominated and may no longer be considered relevant as decision goes, the relevant objects of choice are the chosen option and the undominated unchosen option. The intrinsic utility difference between them is $u_{chosen} - u_{unchosen}$, where the intrinsic utilities u_{chosen} and $u_{unchosen}$ were estimated in *Context-dependent values*. We hypothesized that the decoy effect in trial t causes the relative value assigned to the target to be shifted by some amount $d(t)$. We do

Table 5. Strong positive correlation between the number of preference reversals and μ, strong positive correlation between the percentage decrease of RT in Along than in Against and μ, and strong positive correlation between the average RT in decoy trials and T (from left to right)[a]

	Reversals versus μ	% versus μ	RT versus T
Pearson correlation			
Coefficient	0.610	0.517	0.589
p value	0.000***	0.002***	0.000***
Robust regression			
Slope	0.027	0.012	1.310
p value	0.001***	0.002***	0.001***

[a]% stands for the percentage decrease of RT in Along than in Against and RT the average RT in decoy trials.
***Significance at 1% level.

not impose a sign on $d(t)$, although in general we would expect $d(t)$ to be positive. The strength of relative evidence favoring the chosen option will be $u_{chosen} - u_{unchosen} + d(t)$ if the chosen option is the target (an Along trial) and $u_{chosen} - u_{unchosen} - d(t)$ if the chosen option is the competitor (an Against trial). Hence,

$$RT = \frac{T}{u_{chosen} - u_{unchosen} + (I)d(t)}, \qquad (1)$$

where RT is the RT of trial t, $T > 0$ is the decision bound, $u_{chosen} - u_{unchosen}$ is the intrinsic utility difference between the chosen option and the undominated unchosen option of this trial, and $d(t)$ is the strength of the decoy effect of this trial. The indicator function is $I = 1$ for an Along trial and $I = -1$ for an Against trial. The sign in front of $d(t)$ is different in Along and Against trials because the decoy effect favors a different option in the two cases. Assuming $d(t)$ is drawn from a normal distribution with subject-specific mean μ and standard deviation σ, we used maximum likelihood to estimate T, μ, and σ for each subject. The trial-by-trial strength of the decoy effect $d(t)$ was backed out from (1). The LBA model uses the intrinsic utility difference and whether the decoy is helping the chosen option to estimate the RT and hence to back out $d(t)$.

The parameters estimated using (1) have the expected properties, suggesting the model is appropriate (Table 5). In particular, a subject's estimated average strength of the decoy effect, μ, is a measure of how "decoyable" she is. If the decoy makes the target more valuable, μ is expected to be positive. For all but 2 subjects, it is indeed positive. Furthermore, μ is strongly positively correlated with the number of preference reversals, the natural measure of how strongly decoys influence choices, suggesting that μ effectively captures how decoyable each subject is. As mentioned above, μ is also strongly positively correlated with the differential striatum activity in Along versus Against trials. This is evidence in favor of a link between striatal activity and the behavioral decoy effect. Because 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.

Neural correlate of the trial-by-trial strength of the decoy effect
For each subject, 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 IPL was the only region more active contrasting Weak to Strong (Fig. 3A; Table 6). Indeed, its activity parametrically tracks $d(t)$ negatively (Fig. 3B;

Table 7). The overlapping area is shown in Figure 3C. IPL has previously been implicated in goal-directed preparation of attention (Hopfinger et al., 2000; Raz and Buhle, 2006; Shomstein, 2012). In particular, even though IPL is a large region, the area 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 (Sylvester et al., 2003; Liston et al., 2006a; McNab et al., 2008; Ye and Zhou, 2009). The common denominator in these experiments is that IPL may support the allocation of attention to facilitate task-relevant representations.

The fact that the target dominates the decoy is a salient aspect of the choice menu, which may attract attention. This attention to a salient but irrelevant aspect may underlie the decoy effect (Krajbich et al., 2010). Switching attention to the relevant aspects of the menu, such as the intrinsic properties of the undominated options, would reduce the decoy effect. As IPL correlates negatively with $d(t)$, it is possible that attending less to decoys (as reflected by the negative correlation), subjects attend more to relevant information. This may be part of a control mechanism that corrects behavioral biases induced by decoys. However, unlike conventional control circuits, this mechanism would not necessarily slow down decision-making. When the competitor is intrinsically better than the target, the decoy effect works against choosing the competitor. If heightened attention means focusing on the intrinsic utilities of the undominated options, then choosing the competitor becomes easier and the decision may be faster. Behavior will become more rational and the RT will more closely reflect how long it would take to make a decision based only on the relevant aspects of the task.

Because the results of the Along-Against contrast suggest that context-dependent values could be manifested in differential striatal activity, we explored how IPL might differentially connect with the striatum depending on whether the decoy effect is successfully overridden. We did not find a direct modulation from IPL in our data. We did find a plausible indirect modulation through rIFG, which we further verified by the TMS experiment. The indirect modulation is based on the following two PPI analyses. Because IPL plays a key role, we ran several robustness checks on its negative modulation of $d(t)$ as explained in *Overcoming the decoy effect*. To summarize, in either a minor modification of (1) or in two attempts to measure the trial-by-trial strength of the decoy effect in a model-free way, IPL is significant at $p < 0.001$ uncorrected and has cluster size >40 voxels (data not shown).

How control might be recruited: two connectivity analyses
If the decoy effect increases the valuation of the target, it could potentially lead to the wrong decision if the target is not the intrinsically best option. Presumably, it is impossible to inhibit processing of the decoy at the perceptual level. On a higher level, domination may be a salient property that attracts attention, but control may mitigate the interference from the decoy and refocus attention on the intrinsic properties of the undominated options.

First connectivity analysis: between IPL and IFG
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 the highest intrinsic utility, is categorized as NoConflict. In a Conflict trial, control may be required to override the decoy effect and prevent
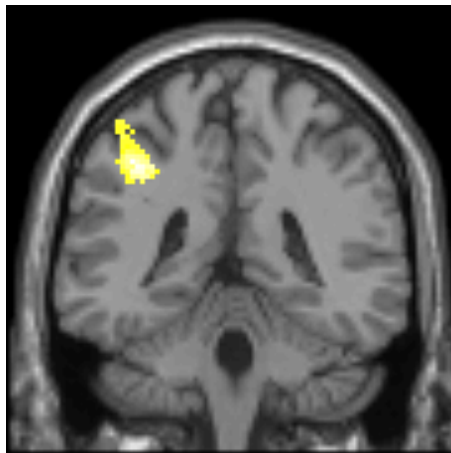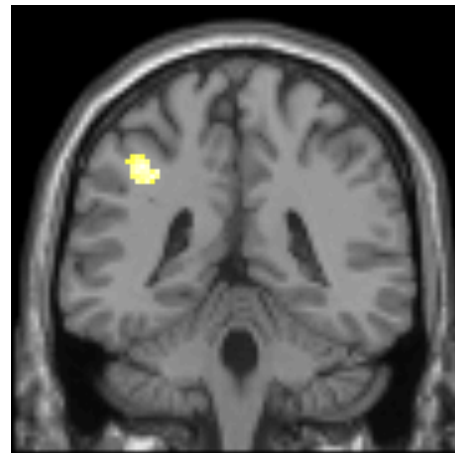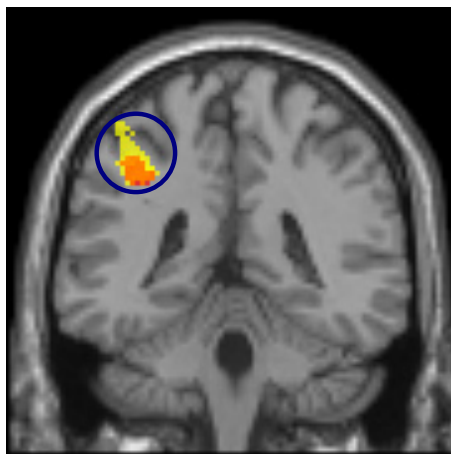
A Weak versus Strong

B Parametric

C Overlap

Figure 3. ***A***, Binary Weak versus Strong decoy effect. The left IPL was more active in Weak than in Strong trials (shown at $y = -40$) ($p < 0.05$ at whole-brain cluster correction). ***B***, Parametric modulation from the LBA model. The left IPL parametrically correlated with $d(t)$ negatively (shown at $y = -40$) ($p < 0.05$ at whole-brain cluster correction). ***C***, Overlap. The left IPL was more active in Weak trials than in Strong trials (shown at $y = -40$) (shown in yellow at $p < 0.001$, uncorrected) and parametrically correlated with $d(t)$ negatively (shown in red at $p < 0.001$, uncorrected). The cluster of the left IPL is circled. Orange represents the overlapping area.

Table 6. Regions more active in Weak than in Strong trials (GLM 2)[a]

Cluster size	Brain region	MNI coordinate			z value
		x	y	z	
461	Inferior parietal lobule				
	BA 40	−36	−42	44	4.1
	BA 40	−40	−40	46	4.1
	BA 40	−38	−38	42	4.1

[a]Local maxima of clusters with $p < 0.05$ at whole-brain cluster correction with a t threshold of 3.40 and an extent of 461 voxels. There is no region more active in Strong than in Weak trials.

Table 7. Regions parametrically correlating with the trial-by-trial decoy effect $d(t)$ negatively (GLM 3)[a]

| Cluster size	Brain region	MNI coordinate			z value
		x	y	z	
234	Medial frontal gyrus				
	BA 32	−16	46	−4	4.5
	BA 10	−12	54	4	4.0
185	Inferior parietal lobule				
	BA 40	−36	−42	44	3.9
	BA 40	−40	−40	46	3.4

[a]Local maxima of clusters with $p < 0.05$ at whole-brain cluster correction with a t threshold of 3.40 and an extent of 185 voxels. There is no region parametrically correlating with the trial-by-trial decoy effect $d(t)$ positively.

the intrinsically inferior target from being chosen. In a NoConflict trial, because the target is the intrinsically best option, the decoy effect is less problematic and control presumably less critical. We hypothesized a stronger connectivity between the area supporting attention and that implementing control when attention is heightened and control is more likely to make a difference.

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 trials as explained in PPI 1. The rIFG is the only region that has this stronger task-related connectivity

(Fig. 4*A*; Table 8). The rIFG is postulated to be a site where goal-directed and stimulus-driven attention converge (Asplund et al., 2010; Shomstein, 2012). Neuroimaging studies implicate this area in supporting inhibition to implement control (Dillon and Pizzagalli, 2007; Levy and Wagner, 2011; Nubert et al., 2012; Aron et al., 2014). 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,

A Conflict versus NoConflict

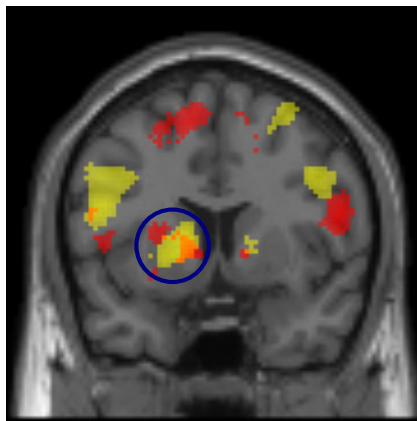**B** SuccessfulControl versus UnsuccessfulControl
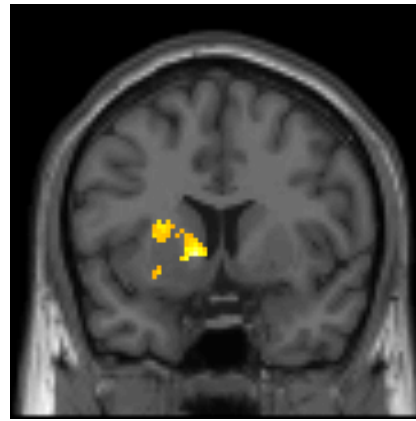
C Overlap

Figure 4. ***A***, The rIFG exhibited stronger functional connectivity with the IPL in Conflict than in NoConflict trials (shown at $y = 16$) ($p < 0.05$ at whole-brain cluster correction). The psychophysiological interaction is based on a 4 mm sphere of IPL centered at $(-40, -40, 46)$. ***B***, The left striatum exhibited stronger functional connectivity with the rIFG in SuccessfulControl than in UnsuccessfulControl trials (shown at $y = 8$) ($p < 0.05$ at whole-brain cluster correction). The psychophysiological interaction is based on a 4 mm sphere of rIFG centered at $(60, 16, 22)$. ***C***, Overlap. The left striatum was more active in Along than in Against trials ($y = 8$) (shown in yellow at $p < 0.001$, uncorrected) and exhibited stronger functional connectivity with the rIFG in SuccessfulControl than in UnsuccessfulControl trials (shown in red at $p < 0.001$, uncorrected). The cluster of the left striatum is circled. Orange represents the overlapping area.

Table 8. Regions exhibiting stronger functional connectivity with IPL in Conflict than in NoConflict trials (PPI 1)[a]

		MNI coordinate			
Cluster size	Brain region	x	y	z	z value
342	Inferior frontal gyrus				
	BA9	60	16	22	4.6

[a]Local maxima of clusters with $p < 0.05$ at whole-brain cluster correction with a t threshold of 3.40 and an extent of 342 voxels. The psychophysiological interaction is based on a 4 mm sphere of IPL centered at $(-40, -40, 46)$. There is no region exhibiting stronger functional connectivity with IPL in NoConflict than in Conflict trials.

all possibly reflecting processes related to overriding prepotent responses (Bunge et al., 2001; Levy and Wagner, 2011). 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, focusing on the facts rather than the framing of a question, or the decision to quit smoking (Goel and Dolan, 2003; De Martino et al., 2006; Berkman et al., 2011). Inhibiting the decoy effect likely is more critical to decision-making in a Conflict trial than in a NoConflict trial. 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 (Sylvester et al., 2003). Anatomical connections between

IPL and IFG have been demonstrated using diffusion-weighted imaging (Rushworth et al., 2006; Caspers et al., 2011). As previous studies indicate that rIFG may play a role in inhibiting irrelevant responses, our connectivity result prompted us to explore further the role of rIFG in our data.

Second connectivity analysis: between IFG and striatum
That the striatum is more active in Along than in Against trials is consistent with the hypothesis that the value of the same chosen option is higher when it is the target. This may lead to biased choices in Conflict trials, where by definition the target is intrinsically worse than the competitor. In Conflict trials, we say that control is unsuccessful if choice is along the decoy and successful if choice is against the decoy. Thus, by definition, if the intrinsically worse target was chosen, then control was unsuccessful; if the intrinsically better competitor was chosen, then control was successful. We hypothesized that, when control is heightened, the connectivity between control and valuation areas will be stronger when control is successful than when 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

Table 9. Regions exhibiting stronger functional connectivity with rIFG in SuccessfulControl than in UnsuccessfulControl trials (PPI 2)[a]

Cluster size	Brain region	MNI coordinate			z value
		x	y	z	
2085	Striatum				
	Medial globus pallidus	−10	2	0	5.0
	Caudate	−8	4	0	4.6
	Putamen	−22	0	4	4.0
660	Inferior frontal gyrus				
	BA 9	60	16	22	4.8
	BA 45	42	24	6	4.3
616	Precentral gyrus				
	BA 6	−10	2	64	4.4
	BA 6	−36	−10	40	4.4
221	Medial frontal gyrus				
	BA 24	16	2	40	4.4
	BA 6	12	2	58	3.6
286	Postcentral gyrus				
	BA 3	46	−22	46	4.0
	BA 3	38	−24	44	4.0
356	Inferior frontal gyrus				
	BA 44	−54	14	14	4.0
	BA 13	−36	22	2	3.7
188	Cerebellum				
		12	−60	−22	3.9
		−6	−52	−20	3.8

[a]Local maxima of clusters with $p < 0.05$ at whole-brain cluster correction with a t threshold of 3.41 and an extent of 188 voxels. The psychophysiological interaction is based on a 4 mm sphere of IFG centered at (60, 16, 22). There is no region exhibiting stronger functional connectivity with IFG in UnsuccessfulControl than in SuccessfulControl trials.

when control was successful than when control was unsuccessful as explained in PPI 2. In a whole-brain search, the left striatum shows a stronger task-related functional connectivity with rIFG (Fig. 4B; Table 9). The active cluster overlaps with what we discovered when contrasting Along to Against trials (Fig. 4C). This is consistent with a possible role of rIFG in reducing the decoy-induced valuation bias. 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 (Liston et al., 2006b).

The TMS experiment: temporarily disrupting the rIFG
The two connectivity results led us to hypothesize about the role of the rIFG in our experiment. Its differential functional connectivity with the striatum hints 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. We applied theta burst TMS to temporarily interfere with the region of interest before subjects started the choice task (Huang et al., 2005; Chao et al., 2011). In one group of subjects, 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 subjects, 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 (Jacobson et al., 2011). 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 (Fig. 5A). In behaviorally rational pairs, for the IFG group, on average, the RT was 7.23 s in Along trials and 9.08 s 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 s in Along trials and 7.47 s in Against behaviorally rational 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 (Fig. 5B).

Discussion
We have investigated the neural underpinning of the decoy effect. The fMRI study revealed differential striatal activity in Along versus Against trials. This is consistent with context-dependent valuations: the same chosen option seems more valuable when it is the target (even when there is no preference reversal). This differential activity correlates with two behavioral measures of the decoy effect, rendering it plausible that context-dependent valuation underlies the behavioral decoy effect. Our results do not contradict the most basic assumption of the standard rational choice model: subjects assign values to the feasible options and choose the option with the highest value. However, if these values are context-dependent, cognitive control may be required to avoid choosing the target option when it is intrinsically worse. We found that IPL, often implicated in subserving voluntary attention, correlates negatively with the trial-by-trial strength of the decoy effect. IPL's functional connectivity with rIFG, a frontal area supporting inhibition to implement control, is stronger when control is more important. rIFG in turn exhibits stronger functional connectivity with striatum when control is successful. These results are summarized in Figure 6. Using TMS to temporarily disrupt rIFG to weaken control, subjects were found to exhibit a stronger decoy effect on behavior. Our results suggest a possible mechanism for how control is recruited and how it may influence valuations. Previous literature supports the role of frontal regions in control. Specifically, rIFG is often implicated in psychological tasks requiring the inhibition of prepotent irrelevant responses or implementation of self-control. The adjacent anterior prefrontal cortex was also found to be a key region in reducing a behavioral bias induced by an irrelevant rewarding feature (Scholl et al., 2015). Our results suggest that rIFG may be involved in inhibiting context-induced preference reversals in simple choice tasks.

A possible source of context dependence is divisive normalization, where the intrinsic utility of each option is divided by the total intrinsic utility of all feasible options (Louie et al., 2013). Thus, if A is more highly valued than B, then the positive normalized difference between A and B is smaller in {A, B, B′} than in {A, B}, resulting in noisier choices and making B more likely to be chosen from the former menu. However, divisive normalization would also predict that B is more likely to be chosen from {A, B, A′} than from {A, B}, contradicting what decoy experiments typically find. Thus, divisive normalization cannot explain the decoy effect. We
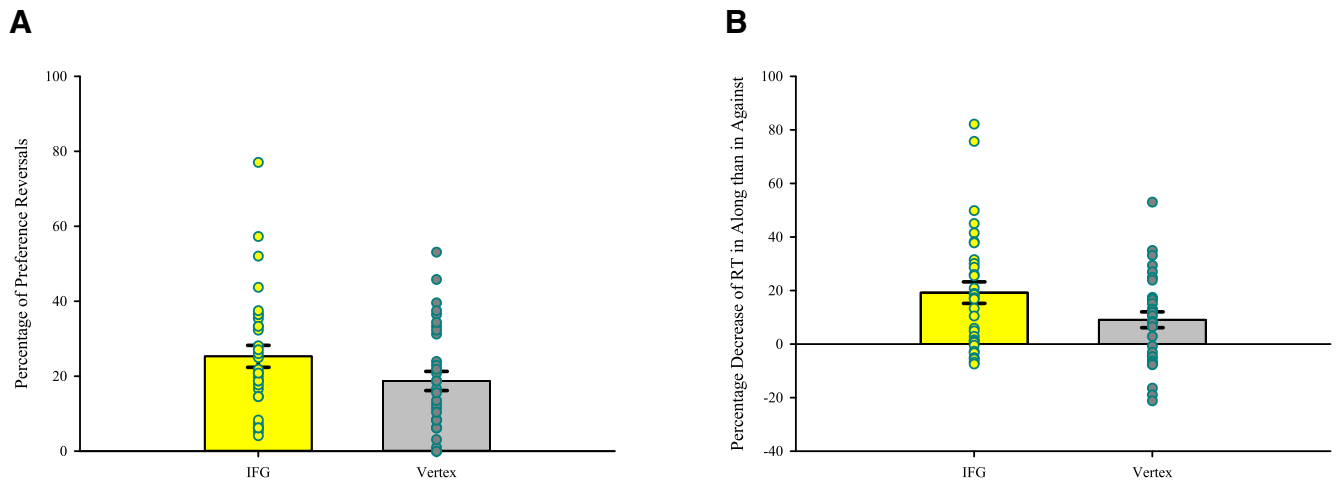
A

B

Figure 5. *A*, Larger percentage of preference reversals in the IFG group than in the vertex group. Overlaid on the mean and the SEM are each subject's data point. *B*, Larger percentage decrease of RT in the IFG group than in the vertex group. The RTs were shorter in Along than in Against behaviorally rational pairs, and this was even more so in the IFG group than in the vertex group. Overlaid on the mean and the SEM are each subject's data point.
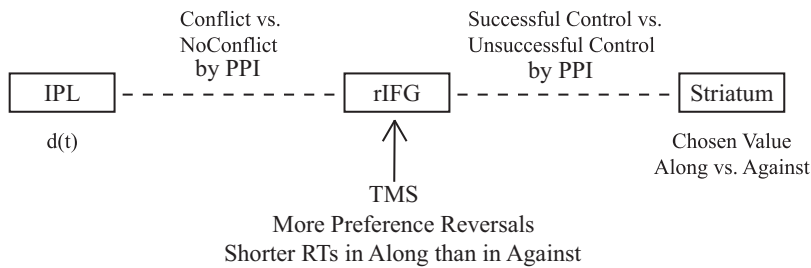

Figure 6. An illustration of the main results.

compare choices along the decoy with choices against the decoy, thereby avoiding the possible confound of comparing three-option menus with two-option menus (Hedgcock and Rao, 2009). Suppose that the better Option A is chosen in a behaviorally rational pair. The normalized difference between A and B is not systematically higher in {A, B, A′} than in {A, B, B′} because, by design, the total intrinsic utility of all options is not systematically lower in the former menu. Therefore, divisive normalization provides no explanation for the systematic favoring of A in {A, B, A′} and B in {A, B, B′}. Similarly, divisive normalization cannot explain the higher striatal activity in Along than in Against trials.

Range normalization might explain the systematic favoring of the target. Consider an {A, B, A′} trial where A is more expensive than B. Then because A′ is dominated by A, A′ is even more expensive. Representing the very expensive decoy A′ neurally may make the price differential between A and B less significant, hence increasing the attractiveness of A. The decoy effect may then be stronger, the more expensive is A′ compared with A (Soltani et al., 2012). More generally, the distance between the target and the decoy might matter. Our experiment was not designed to investigate this, but our data does provide some support for the conjecture. We calculated the trial-by-trial intrinsic utility distance between the target and the decoy, $u_{target} - u_{decoy}$, using the logit estimates from *Context-dependent values*. The average Pearson correlation coefficient between $u_{target} - u_{decoy}$ and $d(t)$ is 0.42. Thus, within the range of variation of our experiment, the decoy effect seems stronger when the intrinsic utility distance is larger. However, a large part of the trial-by-trial variation in the decoy effect is not explained by the intrinsic utility distance between the target and the decoy.

Unavailable distractors have been found to affect value difference signals in ventral medial prefrontal cortex (Chau et al., 2014). These distractors are close to the idea of decoys but did not have specificity, in the sense that the effect in principle was not specific to either the high value or the low value option. In contrast, the decoy is dominated by one option (the target) only, and thus has a strong specificity allowing us to explore how the value of an option depends on whether or not it is the target. Specifically, we found context-dependent activity in the striatum, suggesting a context-dependent valuation of the chosen option. The different results of the two studies are not surprising in view of the different designs. Unlike distractors, decoys are possible objects of choice and may be considered as such at least initially, despite being dominated. Closer to decision, they may no longer be considered relevant options, but the effect of helping the target may linger. We find preliminary evidence (data not shown) supporting this conjecture. Modeling only the first 2 s of a trial, we find that the striatum likely codes the value sum of all options (including the decoy) but at that time, it is not more active in Along than in Against trials. In the middle 2 s, however, the striatum likely codes the value of the chosen option and no longer correlates with the value sum. Moreover, there is a striatal cluster of 173 voxels more active in Along than in Against trials. Even though fMRI does not have good time resolution to explore the dynamics, these observations are consistent with the hypothesis that decoys might initially be considered possible objects of choice, and later their effect is manifested as helping the targets. Among areas often correlated with values, such as ventral medial prefrontal cortex and ACC, although we did not observe their activity in the first 2 s, we found it possibly correlating with the value of the chosen option in the middle 2 s. Hence, the lack of their overall involvement in our experiment could be due to the long RTs. However, unlike striatum, they do not have differential activity in Along than in Against trials, which is the main focus of our experiment. Various value-related signals seem distributed over the brain at different points of time. Biophysically realistic networks have been proposed (Wang, 2002; Hunt et al., 2014), which may produce various value-related signals. How these sig-

nals may be related to each other dynamically remains an open question.

The striatum is often implicated in conflict-eliciting choice situations, and the rIFG is involved in inhibiting a prepotent response. However, interpretations along this line do not easily apply to our experiment, where control may play a subtle role that does not simply show up in elevated mean activity. If striatal activity is linked to conflict-eliciting choices, we would expect it to be more active in Against trials than in Along trials (because choosing against the decoy would presumably involve more of a conflict). In fact we observe the opposite. Moreover, our findings on IPL and rIFG suggest modulation of values rather than simple inhibition. In our data, IPL correlates with the $d(t)$ signal, which is closely linked to values. It is difficult to interpret IPL activity as simple inhibition because, as can be seen from (1), $d(t)$ depends on the value difference and on which option is chosen. As for rIFG activity, it is often observed in No-Go or Stop-Signal tasks that manipulate control explicitly. Our experiment does not explicitly manipulate control: in each trial, subjects just choose an option. We find that the connectivity of rIFG with IPL is stronger in Conflict trials than in NoConflict trials, and its connectivity with striatum is stronger when control is successful than when control is unsuccessful. Thus, intrinsic utility implicitly plays a role in our findings on rIFG as well. IPL is not known as a control region in the same way as rIFG is, but it often activates when attention is needed and in decision tasks (Nieuwenhuis et al., 2005). In view of this, a possible interpretation of our results is in terms of attention being important for the successful implementation of control.

In measuring the trial-by-trial strength of the decoy effect $d(t)$, we attributed variations in the strength of evidence favoring the chosen option to variations in $d(t)$. An alternative assumption would attribute them to random evaluation of intrinsic utility. However, when using (1) to estimate variations in $d(t)$, we take into account which option is target and which is decoy. This information would not be used to estimate a random intrinsic utility noise favoring one option over another. Hence, the brain correlate of $d(t)$ is different from that of random intrinsic utility evaluations. Preliminary evidence (data not shown) shows that the IPL cluster, which correlates negatively with $d(t)$, does not correlate with random utility noises. Among areas correlating with random utility noises, most noticeable is the ventral medial prefrontal cortex, which often codes values.

In our experiment, the decoy is irrelevant to the choice problem; it would be rational to simply disregard the decoy and let the brain save its limited firing range for the viable (undominated) options. However, a comparative valuation heuristic could be adaptive in naturally occurring sequential-choice situations, where whether an organism should choose an option right away depends on the background (Houston, 1997). This heuristic requires the use of memories and contextual information. In an experiment with European starlings, the birds rely on information from the context to make the optimal choice in sequential encounters (Freidin and Kacelnik, 2011). A comparative valuation heuristic may be evolutionarily conserved because sequential decision problems have been prevalent. However, there are several reasons why the heuristic may lead to suboptimal decisions in modern-day choice situations (Slovic et al., 2007). 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 options, marketers distort the decision utility (i.e., the consumer's assessment of the intrinsic utility). *The Economist*'s advertisement mentioned above may be an example of this. Second, the assumptions underlying the normal operation of the system may be violated. For example, the heuristics may be well adapted to naturally occurring sequential decision problems, but not to the static choice problems 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. Rather than being a "design flaw," the decoy effect may be an evolved adaptation that illuminates the online comparative judgments of the decision-making system. It may be analogous to visual illusions (e.g., the Ebbinghaus illusion where the perceived size of an object depends comparatively on the sizes of neighboring objects) (Goodale and Haffenden, 1998; Eagleman, 2001), which is adaptive in ecologically relevant scenarios. Similarly, the heuristic paying attention to dominance may have been adaptive in the past, but modern-day shopping situations may require control, which could have its function originating elsewhere, to prevent biased decisions.

References

Ariely D (2008) Predictably irrational: the hidden forces that shape our decisions, Ed 1. New York: Harper.

Aron AR, Robbins TW, Poldrack RA (2014) Inhibition and the right inferior frontal cortex: one decade on. Trends Cogn Sci 18:177–185. CrossRef Medline

Asplund CL, Todd JJ, Snyder AP, Marois R (2010) A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. Nat Neurosci 13:507–512. CrossRef Medline

Bateson M, Healy SD, Hurly TA (2003) Context-dependent foraging decisions in rufous hummingbirds. Proc Biol Sci 270:1271–1276. CrossRef Medline

Berkman ET, Falk EB, Lieberman MD (2011) In the trenches of real-world self-control: neural correlates of breaking the link between craving and smoking. Psychol Sci 22:498–506. CrossRef Medline

Brown SD, Heathcote A (2008) The simplest complete model of choice response time: linear ballistic accumulation. Cogn Psychol 57:153–178. CrossRef Medline

Bunge SA, Ochsner KN, Desmond JE, Glover GH, Gabrieli JD (2001) Prefrontal regions involved in keeping information in and out of mind. Brain 124:2074–2086. CrossRef Medline

Caspers S, Eickhoff SB, Rick T, von Kapri A, Kuhlen T, Huang R, Shah NJ, Zilles K (2011) Probabilistic fibre tract analysis of cytoarchitectonically defined human inferior parietal lobule areas reveals similarities to macaques. Neuroimage 58:362–380. CrossRef Medline

Chao CM, Tseng P, Hsu TY, Su JH, Tzeng OJ, Hung DL, Muggleton NG, Juan CH (2011) Predictability of saccadic behaviors is modified by transcranial magnetic stimulation over human posterior parietal cortex. Hum Brain Mapp 32:1961–1972. CrossRef Medline

Chau BK, Kolling N, Hunt LT, Walton ME, Rushworth MF (2014) A neural mechanism underlying failure of optimal choice with multiple options. Nat Neurosci 17:463–470. CrossRef Medline

De Martino B, Kumaran D, Seymour B, Dolan RJ (2006) Frames, biases, and rational decision-making in the human brain. Science 313:684–687. CrossRef Medline

Dillon DG, Pizzagalli DA (2007) Inhibition of action, thought, and emotion: a selective neurobiological review. Appl Prevent Psychol 12:99–114. CrossRef Medline

Eagleman DM (2001) Visual illusions and neurobiology. Nat Rev Neurosci 2:920–926. CrossRef Medline

Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A 113:7900–7905. CrossRef Medline

Freidin E, Kacelnik A (2011) Rational choice, context dependence, and the value of information in European starlings (*Sturnus vulgaris*). Science 334:1000–1002. CrossRef Medline

Goel V, Dolan RJ (2003) Explaining modulation of reasoning by belief. Cognition 87:B11–B22. CrossRef Medline

Goodale MA, Haffenden A (1998) Frames of reference for perception and action in the human visual system. Neurosci Biobehav Rev 22:161–172. CrossRef Medline

Hedgcock W, Rao AR (2009) Trade-off aversion as an explanation for the attraction effect: a functional magnetic resonance imaging study. J Marketing Res 46:1–13. CrossRef

Hopfinger JB, Buonocore MH, Mangun GR (2000) The neural mechanisms of top-down attentional control. Nat Neurosci 3:284–291. CrossRef Medline

Houston AI (1997) Natural selection and context-dependent values. Proc Biol Sci 264:1539–1541. CrossRef

Huang YZ, Edwards MJ, Rounis E, Bhatia KP, Rothwell JC (2005) Theta burst stimulation of the human motor cortex. Neuron 45:201–206. CrossRef Medline

Huber J, Payne JW, Puto C (1982) Adding asymmetrically dominated options: violations of regularity and the similarity hypothesis. J Consum Res 9:90–98. CrossRef

Hunt LT, Dolan RJ, Behrens TE (2014) Hierarchical competitions subserving multi-attribute choice. Nat Neurosci 17:1613–1622. CrossRef Medline

Jacobson L, Javitt DC, Lavidor M (2011) Activation of inhibition: diminishing impulsive behavior by direct current stimulation over the inferior frontal gyrus. J Cogn Neurosci 23:3380–3387. CrossRef Medline

Kim JG, Biederman I, Juan CH (2011) The benefit of object interactions arises in the lateral occipital cortex independent of attentional modulation from the intraparietal sulcus: a transcranial magnetic stimulation study. J Neurosci 31:8320–8324. CrossRef Medline

Krajbich I, Armel C, Rangel A (2010) Visual fixations and the computation and comparison of value in simple choice. Nat Neurosci 13:1292–1298. CrossRef Medline

Latty T, Beekman M (2011) Irrational decision-making in an amoeboid organism: transitivity and context-dependent preferences. Proc Biol Sci 278:307–312. CrossRef Medline

Levy BJ, Wagner AD (2011) Cognitive control and right ventrolateral prefrontal cortex: reflexive reorienting, motor inhibition, and action updating. Ann N Y Acad Sci 1224:40–62. CrossRef Medline

Liston C, Matalon S, Hare TA, Davidson MC, Casey BJ (2006a) Anterior cingulate and posterior parietal cortices are sensitive to dissociable forms of conflict in a task-switching paradigm. Neuron 50:643–653. CrossRef Medline

Liston C, Watts R, Tottenham N, Davidson MC, Niogi S, Ulug AM, Casey BJ (2006b) Frontostriatal microstructure modulates efficient recruitment of cognitive control. Cereb Cortex 16:553–560. CrossRef Medline

Louie K, Khaw MW, Glimcher PW (2013) Normalization is a general neural mechanism for context-dependent decision making. Proc Natl Acad Sci U S A 110:6139–6144. CrossRef Medline

Manes JL, Parkinson AL, Larson CR, Greenlee JD, Eickhoff SB, Corcos DM, Robin DA (2014) Connectivity of the subthalamic nucleus and globus pallidus pars interna to regions within the speech network: a meta-analytic connectivity study. Hum Brain Mapp 35:3499–3516. CrossRef Medline

McNab F, Leroux G, Strand F, Thorell L, Bergman S, Klingberg T (2008) Common and unique components of inhibition and working memory: an fMRI, within-subjects investigation. Neuropsychologia 46:2668–2682. CrossRef Medline

Muggleton NG, Cowey A, Walsh V (2008) The role of the angular gyrus in visual conjunction search investigated using signal detection analysis and transcranial magnetic stimulation. Neuropsychologia 46:2198–2202. CrossRef Medline

Nieuwenhuis S, Heslenfeld DJ, von Geusau NJ, Mars RB, Holroyd CB, Yeung N (2005) Activity in human reward-sensitive brain areas is strongly context dependent. Neuroimage 25:1302–1309. CrossRef Medline

Nubert FX, Mars RB, Rushworth MFS (2012) Is there an inferior frontal cortical network for cognitive control and inhibition? In: Principles of frontal lobe function (Stuss DT, Knight RT, eds), pp 332–352. New York: Oxford UP.

Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. Nature 441:223–226. CrossRef Medline

Padoa-Schioppa C, Assad JA (2008) The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat Neurosci 11:95–102. CrossRef Medline

Raz A, Buhle J (2006) Typologies of attentional networks. Nat Rev Neurosci 7:367–379. CrossRef Medline

Rushworth MF, Behrens TE, Johansen-Berg H (2006) Connection patterns distinguish 3 regions of human parietal cortex. Cereb Cortex 16:1418–1430. CrossRef Medline

Sasaki T, Pratt SC (2011) Emergence of group rationality from irrational individuals. Behav Ecol 22:276–281. CrossRef

Scholl J, Kolling N, Nelissen N, Wittmann MK, Harmer CJ, Rushworth MF (2015) The good, the bad, and the irrelevant: neural mechanisms of learning real and hypothetical rewards and effort. J Neurosci 35:11233–11251. CrossRef Medline

Shafir S, Waite TA, Smith BH (2002) Context-dependent violations of rational choice in honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*). Behav Ecol Sociobiol 51:180–187. CrossRef

Shomstein S (2012) Cognitive functions of the posterior parietal cortex: top-down and bottom-up attentional control. Front Integr Neurosci 6:38. CrossRef Medline

Slovic P, Finucane ML, Peters E, MacGregor DG (2007) The affect heuristic. Eur J Oper Res 177:1333–1352. CrossRef

Soltani A, De Martino B, Camerer C (2012) A range-normalization model of context-dependent choice: a new model and evidence. PLoS Comput Biol 8:e1002607. CrossRef Medline

Stewart LM, Walsh V, Rothwell JC (2001) Motor and phosphene thresholds: a transcranial magnetic stimulation correlation study. Neuropsychologia 39:415–419. CrossRef Medline

Sylvester CY, Wager TD, Lacey SC, Hernandez L, Nichols TE, Smith EE, Jonides J (2003) Switching attention and resolving interference: fMRI measures of executive functions. Neuropsychologia 41:357–370. CrossRef Medline

Wang XJ (2002) Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36:955–968. CrossRef Medline

Ye Z, Zhou X (2009) Conflict control during sentence comprehension: fMRI evidence. Neuroimage 48:280–290. CrossRef Medline