ESSAYS IN BEHAVIORAL AND NEURO-ECONOMICS

Thesis by

Benjamin Bushong

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

2013

California Institute of Technology

Pasadena, California

(Defended April 22, 2013)

Acknowledgement:

The author wishes to thank his family for their support through the development of this thesis. He is particularly grateful to the members of his committee: Antonio Rangel for his guidance, support, and his lab's research vision throughout the development of this thesis; Matthew Rabin for expanding his horizons into neuroscience and, along the way, teaching the author an immense amount of psychology and economics; and Colin Camerer, whose fierce passion for neuroeconomics convinced the author to begin in the field five years ago.

This thesis would not have been possible without the generous financial support of the NSF IGERT training program and the Moore Foundation.

Abstract:

This thesis examines foundational questions in behavioral economics—also called psychology and economics—and the neural foundations of varied sources of utility. We have three primary aims: First, to provide the field of behavioral economics with psychological theories of behavior that are derived from neuroscience and to use those theories to identify novel evidence for behavioral biases. Second, we provide neural and micro foundations of behavioral preferences that give rise to well-documented empirical phenomena in behavioral economics. Finally, we show how a deep understanding of the neural foundations of these behavioral preferences can feed back into our theories of social preferences and reference-dependent utility.

The first chapter focuses on classical conditioning and its application in identifying the psychological underpinnings of a pricing phenomenon. We return to classical conditioning again in the third chapter where we use fMRI to identify varied sources of utility—here, reference dependent versus direct utility—and cross-validate our interpretation with a conditioning experiment. The second chapter engages social preferences and, more broadly, causative utility (wherein the decision-maker derives utility from making or avoiding particular choices).

Table of Contents:

List of Illustrations and Tables:

PAVLOVIAN PROCESSES IN CONSUMER CHOICE: THE PHYSICAL PRESENCE OF A GOOD INCREASES WILLINGNESS-TO-PAY[1]

Benjamin Bushong[2]

Lindsay M. King[2]

Colin F. Camerer[3]

Antonio Rangel[3]

[1] The authors would like to thank Yuval Rottenstreich for his extremely useful feedback on various aspects of this chapter.

[2] Benjamin Bushong and Lindsay M. King: California Institute of Technology, HSS, 1200 E. California Blvd, Pasadena, CA 91125.

[3] Colin Camerer & Antonio Rangel: California Institute of Technology, HSS and CNS, 1200 E. California Blvd, Pasadena, CA 91103, camerer@caltech.edu and rangel@caltech.edu.

Abstract. Consider choosing a meal in a restaurant by reading a text-based menu, looking at a picture-based menu, or being exposed to a buffet table. Does this affect the choices that you would make? This paper describes a series of laboratory experiments studying whether the form in which items are displayed at the time of choice affects the dollar value that subjects place on them. We sell foods and trinkets to subjects using a Becker-DeGroot auction under three different conditions: (1) text displays, (2) image displays, and (3) displays of the actual items. We find that subjects' willingness-to-pay is between 40 and 61% higher when they were presented with the real items, as compared to the text or image displays. Furthermore, a series of follow-up experiments suggest the presence of the real item matters because it triggers pre-programmed consummatory Pavlovian processes that promote behaviors that lead to contact with appetitive items whenever they are present and accessible.

A basic principle in many social science models is that choices among objects should not vary with innocuous changes in the procedure by which they are made, or with their description. In this "consequentialist" view, choices should only depend on their likely consequences. Ordering groceries online and shopping in a store should lead to identical contents in the refrigerator at home (holding information about quality and other relevant variables constant).

However, many experiments have shown that the cognitive processes that guide choice appear to violate description-invariance and consequentialism. A large research effort in psychology is devoted to studying the ways in which preferences are "constructed" (see S. Lichtenstein and P. Slovic, 2006, for a recent compendium of articles). Economists are increasingly interested in what policies make sense if preferences are constructed (e.g., D. McFadden, 2006). And more recent efforts in neuroeconomics focus on the computational and neurobiological mechanisms used in the computation and comparison of values (for a recent review see A. Rangel et al., 2008).

Most of the constructed-preferences studies describe goods or services abstractly and alter descriptions without changing consequences. For example, in one classic study medical students were asked to choose between hypothetical surgery and radiation treatment options. Outcomes were described in terms of mortality statistics in one "frame", and in terms of survival statistics in another frame, while maintaining statistical equivalence (B. J. McNeil et al., 1982). Medical students and physicians made different choices when faced with the different frames. In practice, differences in physical displays of goods also appear to be important. For example, marketing firms spend enormous resources creating the packaging of a product (which is typically later discarded), the lighting and location of sales displays, the selection and training of salespeople, and so forth.

This paper extends the literature on constructed-preferences by studying whether the form in which items are displayed affects the value that subjects place on them. We compare three conditions: a text display, a picture display, and putting the actual items in front of subjects. We investigate these three conditions because of their theoretical interest and because they approximate some archetypes of situations in which consumers often find themselves. Consider, for example, choosing a meal in a restaurant by reading a text-based menu, looking at a picture-based menu (as is common in some countries), or being exposed to a buffet table.

We describe the results of two separate experiments suggesting that, in comparison to a text or high-resolution displays, the physical presentation of a food item or a trinket has a sizable effect on its value (as measured by incentive-compatible monetary bids). This presents a puzzle for the behavioral sciences, and especially for the emerging field of neuroeconomics: Why do the brain's valuation systems treat these three types of displays so differently?

We propose and test three different explanations of the real-exposure effect based on recent research in psychology and neuroscience. Our results suggest that Pavlovian consummatory mechanisms, which are unfamiliar to economists but have been well established in behavioral neuroscience, might be at work (B.W. Balleine et al., 2008, A. Rangel et al., 2008, B. Seymour et al., 2007). The function of these mechanisms is to deploy behaviors that lead to the consumption of appetitive items when those items are physically exposed to them. Furthermore, these types of processes are thought to influence behavior by changing the value that the brain assigns to particular items.

I. EXPERIMENT 1: BASIC FOOD EXPERIMENT

A. METHODS

Fifty-seven Caltech undergraduate and graduate students participated in the experiment. Individuals were excluded if they had a history of eating disorders, had dieted in the past year, were vegetarian, disliked junk food, or were pregnant. The selection criteria were designed to recruit individuals who liked junk food and were not trying to control their diet. Individuals received $20 for their participation and provided informed consent. Participants were asked to eat and then fast for three hours prior to the experiment. All testing took place in mid-afternoon.

The experiment took approximately 30 minutes. Subjects were told that they would receive $20 for their participation and may receive additional money and food prizes depending on their decisions in the experiment. During this initial instruction, we emphasized that no deception was used in the experiment. Subjects received their instructions through a computer monitor.

At the beginning of the instruction period participants were informed that they would have to stay in the lab for an additional 30 minutes at the end of the experiment (regardless of its outcome). During this time they were allowed to eat as much as they wanted of the single food item they purchased from us during a bidding task, but no other foods or drinks were allowed. If they did not purchase an item they still had to stay in the lab for 30 minutes at the end of experiment. The foods that they could purchase were 80 different popular snacks such as candy bars (e.g., Snickers Bars) and potato chips (e.g., Lay's), which are available at local convenience stores.

Every participant performed three tasks: (1) a liking-rating task, (2) a familiarity-rating task, and (3) a bidding task.

During the liking-rating task subjects had to answer the question "How much would you like to eat this item at the end of the experiment?" on a scale of -7 ("not at all") to 7 ("very much"), with 0 denoting indifference. The timeline of the liking-rating trials started with a 1s central fixation cross, followed by a 3 s presentation of a high-resolution picture of the item to be rated. Pictures were 400x300 pixels in size and showed both the package and the food; the name of the food was also displayed above the picture. Afterwards subjects entered their liking rating at their own pace using the keyboard. The items were shown in random order. There was a 1 s inter-trial interval with an empty screen.

Familiarity-rating trials were similar except that subjects answered the question "How familiar are you with this item?" on a scale of 1 (not much) to 3 (very much). The purpose of these two tasks was two-fold: First, the liking and familiarity ratings were used in analyses reported below. Second, both tasks increased the familiarity of the subjects with the foods and their names.

The bidding trials were the core of the experiment. In addition to the participation fee, each subjects received an endowment of $3 that they could use to purchase food from us. At the end of the experiment one of the bidding trials was selected by drawing a ball from an urn. The subject's bid on this selected trial determined whether he got the item and the price that he had to pay for it.

Items were sold to the subjects on the selected trial by applying the rules of a Becker-DeGroot auction. A random number between $0 and $3 dollars (in $0.25 increments) was selected from an urn. Let n denote the random number that was selected and let b denote the

subject's bid. If $b \geq n$, the subject got the food item and paid \$$n$. If $b < n$, the subject did not get the item but kept the \$3 of bidding money. Note that since the subjects kept whatever funds they did not use, they were *de facto* spending their own money to purchase the food.

A key feature of the auction procedure is that it satisfies "incentive compatibility": if a subject's true value for an item was v, her best response was to bid exactly v. Any deviation from this strategy resulted in a lower expected payoff. Bidding below v does not save money on the price (which is determined by n), it only increases the chance that an item which is liked will not be bought.

Since the rules of the auction are somewhat complicated, we spent significant time training the subjects. In particular, we emphasized that their best strategy was to "go with their gut feeling" about how much each item was worth to them, and then to bid that amount. Debriefing during a pilot experiment confirmed that subjects complied with these instructions. Furthermore, even if there is some bias in bidding relative to true underlying valuation, it should not vary systematically with the display treatments.

The bidding trials were structured as follows: A representation of an item was shown for a certain amount of time, and immediately afterward the subjects entered a bid between \$0 and \$3 by clicking with a mouse on an analog bid bar. There were three between-subjects experimental conditions that differed on how the stimuli were presented in the bidding trials: (1) a text condition (N=20), in which only the text descriptor (the product name) was shown, (2) an image condition (N=17), in which the high-resolution image of the food was shown, and (3) a real condition (N=20), in which an open package of the food item was displayed on a tray. In the text and image conditions the item was presented for 3 s and there was a 3 s inter-trial interval. In the real condition, the item was also presented for 3 s (although the time was not controlled as

precisely) but the inter-trial interval varied since it was determined by the amount of time that it took the experimenter to locate the next food and present it to the subject. The real items were displayed in a way that resembled the presentation in the images (including the use of a black cloth on a tray to resemble the black background of the computer screen). In the text and picture conditions data from 6 to 12 subjects was collected in parallel. Each subject received instructions and performed the task through his own computer terminal. In the real condition only one subject was run at a time.

B. RESULTS

Figure 1 provides a succinct description of the results. There are two main results. First, as can be seen in the top panel, the average bid in the text condition (68 cents, S.D.=0.52) is approximately equal to the average bid in the picture condition (71 cents, S.D.=0.53, two-sided t-test $p=0.88$), and both of them are significantly smaller than the average bid in the real condition (113 cents, S.D.=0.61 two-sided t-test $p<0.004$). Note that the average liking ratings were marginally higher in the text condition (mean = 1.43), than in the real (mean = 1.16) or picture conditions (mean = 0.58), which implies that the effect cannot be attributed to differences in the underlying value of the food items. As the bottom panel illustrates, a random effects linear model with random intercepts and slopes showed no significant differences between the slopes of the bidding curves (i.e., bids as a linear function of liking-rating) in any of the three conditions.

In order to investigate the possibility that the effect might only work with unfamiliar items, we compared average bids for familiar and unfamiliar items. The main effect of displaying the

real item was similar across the two groups: for highly familiar items (familiarity rating=3) the average bid in the real condition was 50 cents higher than in the two other conditions ($p<0.006$); and for less familiar items (rating <3) the bid difference was 41 cents higher ($p<0.001$).

C. DISCUSSION

These results suggest that the form in which an item is displayed can have a sizable impact on real choices: subjects' willingness-to-pay for snacks increased by 61% when they were presented with the real items as opposed to text or image displays.

Several aspects of the results are worth highlighting. First, contrary to our prior expectations, there was no difference between the text and image displays. This is particularly puzzling since the text and image displays contain different amounts and types of information. Second, the display mode had no effect on the relationship between the liking-ratings, which are an independent measure of the consumption value of the items, and the value that is computed at the time of bidding. Instead, the real display basically added a constant markup to all of the items. Third, subjects bid positive amounts even for some items that they had earlier rated as aversive (i.e., negative liking-rating). There are two potential explanations for this. One is that the constant exposure to food during the experiment increased their hunger, and thus made some of the aversive items desirable. The other is that the liking-rating scale did not do a good job picking up the valence of the foods (the mean bid for neutrally rated items was 63 cents, S.D.= 53).

The results raise two important questions. First is the question of robustness: Does the effect occur only with foods and hungry subjects? Or, does a similar phenomenon occur in other subjective states and for other types of items? The second question has to do with the underlying mechanisms generating the effect: What can explain the difference between the text, picture, and real conditions?

A natural hypothesis for economists is that the real condition increases the amount of information that subjects have about the goods and that, by decreasing uncertainty, it increases their willingness-to-pay for them. This is hard to reconcile with all the evidence. Most of these items were highly familiar to our subjects, and the magnitude of the real-condition effect was similar for familiar and unfamiliar items. Furthermore, the largest increase in information takes place between the text and display conditions, instead of between the display and real ones. Finally, additional evidence against the information hypothesis is provided in Experiments 3 and 4 below.

Note also that the experiment cannot be explained in terms of changes in transaction costs: since only one food item is chosen for consumption at the end of the experiment, the cost of actually getting the item is the same across display conditions.

Instead we developed three alternative hypotheses based on previous findings from psychology and neuroscience. The first one focused on the role of odors, which are potentially unconscious. Previous research (G. Loewenstein, 1996) has argued that real items, especially highly appetitive ones, can trigger visceral urges that affect valuation in a more potent way than images or words. Thus, one potential explanation for our findings is that real displays involve the sense of smell, and that adding it into the sensory representation of the item might trigger emotional responses that affect valuation in a way that images alone do not. In addition, the

activation of multiple sensory representation of the choice item might have a super-additive effect on the valuation systems (N. P. Holmes and C. Spence, 2005).

The second hypothesis is based on the idea that activating the experienced reward circuitry at the time of choice might increase the decision value that is assigned to it. In an extreme example of this phenomenon, consider the impact that taking a puff could have on a smokers' desire for a cigarette. Based on this, we hypothesized that exposure to the real items at the time of choice might induce an especially strong activation of the experience reward circuitry (in comparison to the picture and text representations) and that this might lead to an increase on subjects' willingness-to-pay.

The third hypothesis is taken from the animal learning and behavioral neuroscience literatures. A sizable and growing body of evidence suggests that environmental cues can have an effect on the value assigned to items at the time of choice (B.W. Balleine, N. Daw and J. O'Doherty, 2008, A. Rangel, C. Camerer and P. R. Montague, 2008, B. Seymour, T. Singer and R. Dolan, 2007). In particular, B.W. Balleine (2005) and B.W. Balleine et al. (2008) have argued that the physical presence of an appetitive item can trigger Pavlovian consummatory processes that lead animals to make contact with the reward. In the language of animal learning theory, the physical presence of the appetitive stimulus (e.g., food) serves as an unconditioned stimulus (US) triggering the consummatory response. As with every Pavlovian process, it is possible for organisms to learn to associate other cues (called conditioned stimuli, CS) with the presence of the US (given by the actual presence of the appetitive item). When the pairing is sufficiently strong, the mere presence of the CS can trigger the approach/consummatory responses. This type of learning explains, for example, why highly trained pigeons peck at a light that predicts the delivery of actual food.

According to this hypothesis, the Pavlovian consummatory response is triggered in Experiment 1 for the real condition, since the presence of the food serves as a US, but not in the text or picture conditions, because these stimuli are not CSs which are as strongly associated with the US. Although a priori there is nothing precluding the text or pictures from serving as a CS capable of triggering the approach response, the data suggests that they have not acquired the required association with the US and thus the pairing is weaker than that in the real treatment. One potential reason why this might be the case is that our subjects are unlikely to have been trained repeatedly to pair the text and pictorial stimuli with the US. That is, in contexts outside the experiment, the names and pictures of foods (e.g., in advertisements) are not frequently associated with the presence of the foods, so there is no associative link that can trigger the approach response.

In order to test this third hypothesis, we conjectured that the Pavlovian consummatory processes might not be activated in situations in which the items cannot be accessed. This could happen because the response is not activated in the first place, or because it is overridden by competing behavioral responses that take into account the fact that the stimulus is not accessible.

The rest of the paper describes the results of three additional experiments designed to address the issue of robustness and to investigate the relative contribution of the three proposed mechanisms.

II. EXPERIMENT 2: ROBUSTNESS AND THE ROLE OF SMELL

In order to address the issue of robustness, and to investigate the role that smell plays in the previous results, we repeated Experiment 1 using trinkets instead of foods.

A. METHODS

Sixty Caltech undergraduate and graduate students participated in this experiment. Since the design is extremely similar to Experiment 1, here we only describe the differences between them. First, instead of snack foods, subjects bid on 20 different small-value trinkets such as Caltech mugs and various DVDs. All of the items were sold at the Caltech bookstore at the time of the experiment and had a maximum in-store price of $20. Second, subjects were not required to stay for 30 minutes at the end of the experiment. Instead, any trinkets purchased during the experiment were mailed to them at the end of the day. Third, prospective subjects faced no exclusion criteria. Fourth, the trinkets were displayed without any packaging in all of the pictures and in the real condition. Twenty subjects participated in each of the conditions.

B. RESULTS

Figure 2 summarizes the results. As can be seen in the top panel, the average bid in the text condition (1.02 cents, S.D.=0.54) is approximately equal to the average bid in the picture condition (1.01 cents, S.D.=0.53, two-sided t-test $p=0.9806$), and both of them are significantly smaller than the average bid in the real condition (142 cents, S.D.=0.61 two-sided t-test $p<0.008$). This represents a 41% increase in the subjects' willingness-to-pay for the items, which is commensurate with the effect size that we found in Experiment 1. Note that the average liking

ratings were not significantly different in the three conditions (minimum p-value in a two-sided test 0.45), which implies that the effect cannot be attributed to differences in the underlying value of the trinkets. As the bottom panel illustrates, a random effects linear model with random intercepts and slopes showed no significant differences between the slopes of the bidding curves in any of the three conditions.

C. DISCUSSION

A comparison of Figures 1 and 2 shows that the results for the food and trinket experiments are remarkably similar. It follows that the real-exposure effect is not limited to the case of snack foods. Furthermore, since smells are unlikely to play a role in the case of the trinkets, we can conclude that they are not the mechanism behind the real-exposure effect in both experiments.

III. EXPERIMENT 3: THE ROLE OF EXPERIENCED REWARD PROCESSESS

The next experiment addressed the role of experienced reward on the valuation processes. To do this we repeated the picture food condition with a twist: subjects had to eat a small sample of each food while deciding how much to bid. The idea behind the experiment is that if experiencing the rewards generated by an item has a positive effect on valuations, then a taste of an appetitive food should have a positive effect on the bids even when it is not physically present.

A. METHODS

Seventeen Caltech undergraduate and graduate students participated in this experiment. Since the design is extremely similar to the picture condition of Experiment 1, here we only describe the differences between them. First, instead of 80 snack foods, subjects only placed bids on 20 of them. The 20 foods were chosen at random from those used in Experiment 1. We reduced the number of foods to facilitate the process of data collection given the additional difficulties described below.

Second, after seeing the picture of the food item, subjects were asked to taste and swallow a small amount (about 10 grams) of it prior to entering their bids. This was done as follows. The experimenter sat next to the subject and had access to 20 small paper cups, each containing a sample of one of the foods. After the image of a food was presented in the screen for 3 seconds the experimenter handled a sample of that item to the subject who had to eat it before entering a bid. The picture of the food stayed on the screen until a bid was entered. In order to facilitate the process of running the experiment, the order of food presentation was randomized but kept constant across subjects.

Note that, in contrast to the real condition in Experiment 1, the subjects were not exposed to packages or full samples of the foods. In fact, most of the time subjects did not even take a look at the contents of the paper cups since they knew that it was just a sample of the food displayed in the screen.

B. RESULTS

Figure 3 summarizes the results. For comparison purposes, the figure compares the results of this experiment with those of the picture and real conditions in Experiment 1. As can be seen in the top panel, the average bid in the taste condition (74 cents, S.D.=0.51) is approximately equal to the average bid in the picture condition (70 cents, S.D.=0.53, two-sided t-test p=0.85), but substantially smaller than the average bid in the real condition (114 cents, S.D.-0.53, two-sided t-test $p<0.029$). Note that the average liking ratings in the taste condition (mean = 2.01) were marginally higher than those in the real condition (mean = 1.16), which implies that the effect cannot be attributed to differences in the underlying value of the food items. As the bottom panel illustrates, a random effects linear model with random intercepts and slopes showed no significant differences between the slopes of the bidding curves in any of the three conditions.

C. DISCUSSION

The results of this experiment show that giving subjects a taste of the item has no effect on their willingness-to-pay when the item is not physically present. In fact, as can be seen in Figure 3, the bidding curves for the taste and picture and picture only conditions are nearly identical. These results are valuable for two reasons. First, they provide evidence against the hypothesis that experienced reward processes are responsible for the real-exposure effect. Second, since getting a taste of the item should increase the amount of information that subjects have about the foods, it provides further evidence against informational explanations for the effect.

IV. EXPERIMENT 4: THE ROLE OF PAVLOVIAN CUES

The final experiment investigated the Pavlovian consummatory mechanisms explanation of the real-exposure effect. The experiment is almost identical to the real condition of experiment except that we placed a fully transparent Plexiglas wall between the subject and the food, while keeping the physical distanced between subject and food constant. Our hypothesis was that if consummatory cues are at work, then the presence of a physical barrier would decrease the likelihood that the processes would be deployed (because the subjects knew that the barrier made the items unavailable), thus reducing the impact of real exposure on the subject's willingness-to-pay. Keep in mind that the use of clear Plexiglas means that all sensory cues are still present, so the information hypothesis predicts that the results of this experimental treatment should be much like the original finding of a real-exposure effect.

A. METHODS

Thirty Caltech undergraduate and graduate students participated in this experiment. Since the design is almost identical to the real condition of Experiment 1, here we only describe the differences between them. First, instead of 80 snack foods, subjects only placed bids on 20 of them. These were the same 20 foods used in Experiment 3 and were chosen at random from those used in Experiment 1. Second, although the physical set-up of the experiment was unchanged (including the distance of the experimenter to the subject), a fully transparent

Plexiglas wall (dimensions 8 ft by 8 ft by 1/4 inches) was placed midway between the subject and the experimenter. The barrier was large enough so that the foods shown by the experiment were out of the subject's reach.

B. RESULTS

Figure 4 summarizes the results. As before, the figure compares the results of this experiment with those of the picture and real conditions in Experiment 1. As can be seen in the top panel, the average bid in the Plexiglas condition (81 cents, S.D.=0.53) is approximately equal to the average bid in the picture condition (70 cents, S.D.=0.53, two-sided t-test $p=0.51$), but substantially smaller than the average bid in the real condition (114 cents, S.D.-0.53, two-sided t-test $p<0.042$). Note that the average liking ratings were marginally higher in the Plexiglas condition (mean = 1.62), than in the real condition (mean = 1.16), which implies that the effect cannot be attributed to differences in the underlying value of the food items. As the bottom panel illustrates, a random effects linear model with random intercepts and slopes showed no significant differences between the slopes of the bidding curves (i.e., bids as a liner function of liking-rating) in any of the three conditions.

C. DISCUSSION

The introduction in the real condition of a transparent Plexiglas barrier between the subject and the food, which has no impact on the sensory information available to the subject, reduces the willingness-to-pay almost to the level of the picture condition, thus eliminating the real-exposure effect. Given that this was a surprising and somewhat farfetched prediction of the Pavlovian account, that it is quite hard to explain the effect of the Plexiglas barrier using an alternative theory, and that there exists a considerable amount of neural evidence for the presence of these types of mechanisms (B.W. Balleine, N. Daw and J. O'Doherty, 2008, A. Rangel, C. Camerer and P. R. Montague, 2008, B. Seymour, T. Singer and R. Dolan, 2007), the experiment provides significant support in favor of this theory of the real-exposure effect. Note, in addition, that the amount of information provided in the real and Plexiglas conditions is identical, and therefore the experiment provides further evidence against an informational explanation of the phenomenon.

V. GENERAL DISCUSSION

The experiments in this study suggest the following three main results. First, the physical presence of an accessible appetitive (i.e., desirable) item at the time of choice leads to a sizable increase in subject's willingness-to-pay for it, a phenomenon that we have labeled the real-exposure effect. Second, the effect is at work in the evaluation of basic rewards such as high-caloric items for hungry subjects and non-basic rewards such as low value consumer products.

Third, Pavlovian consummatory processes triggered by the item's presence might be responsible for the real-exposure effect.

We emphasize three aspects of the Pavlovian consummatory processes theory that we posit as a potential explanation for our findings. First, the text and picture of the stimuli do not seem to be able to serve as CSs capable of triggering the consummatory response through their association with the US given by the actual presence of the food. One potential explanation for this finding is that subjects are unlikely to have received extensive training in pairing these stimuli with the actual physical presence of the foods, which is the US triggering the Pavlovian approach response. Second, the Pavlovian consummatory are not deployed when the stimuli cannot be acquired because, for example, it is placed behind a large Plexiglas wall. This could happen because the Pavlovian processes are sophisticated and take into account that there is no point on deploying a Pavlovian response that cannot succeed, or because in these circumstances the response is inhibited by alternative competing processes. Third, the Pavlovian consummatory processes are triggered by the presence of very different appetitive items, which is necessary to explain why we get similar results for foods and trinkets.

An important open question for future research is to explore further what makes a stimulus a predictive CS capable of triggering a Pavlovian consummatory response, and the extent to which appropriate CS are domain specific. Our limited understanding of the nature of these cues is highlighted by the fact that a taste of the food, which one might have speculated should be a powerful CS associated with the presence of food, did not activate the approach response. This is puzzling since the taste of a food is typically associated with having more of the same food available.

The results have practical implications in a number of domains. First, consider again the problem of restaurateur that has to decide whether to provide its customers with a written menu, a picture-based menu, or a dessert tray. The results in this paper suggest that dessert sales should go up significantly if the restaurant uses the dessert tray as opposed to the other two options. Furthermore, the results of the Plexiglas experiment suggest that a transparent glass dome should not cover the desert tray, as is the practice in some establishments. Second, the results also help to explain companies' efforts to find the right packaging and display for their products. In particular, they suggest that stores might want to display real products to consumers and allow more sensory interaction (e.g., test-driving cars which have the "new car smell"). Producing these effects is especially challenging for Internet commerce since, by necessity, Internet sellers are restricted to image, text, and sound displays. Third, the results described above suggest a scope for government regulation of packaging and displays of items that are associated with unhealthy consumption, such as addictive substances and junk foods, to help consumers self-regulate (K.J. Wenterbroch, 1998). Finally, our findings might also extend to social bargaining situations. A common legal practice is to present a plaintiff with a signed check when making an offer for a settlement. Our results suggest that this practice might increase the likelihood that the settlement offer is accepted.[1]

Our results also provide insight into the findings of two recent studies on the valuation of economics goods. The first one shows that subject's valuations for small toys (e.g., a slinky)

[1] In a sequential trust game, S. Solnick (2007) found that subjects in the second-mover trustee role returned only half as much actual cash as other subjects who were asked to return play money or make a numerical statement of the intended cash return. Since money is a highly conditioned stimulus, the results of this experiment can also be explained through our mechanism. Under this explanation, the physical presence of money triggers approach responses that makes it hard to transfer it to the other player.

increase when they are allowed to touch them (J. Peck and S.B. Shu, forthcoming). This can be explained within our framework by the fact that the touch manipulation involves direct and unencumbered proximity to the items, which can trigger Pavlovian approach mechanisms towards the desirable items. The second study shows that the occurrence of the classic endowment effect (D. Kahneman et al., 1990, J.L. Knetsch and J.A. Sinden, 1984), in which subjects valuations for items depend on whether or not they own them, depends on the actual items being physically present at the time of the experiment. In fact, a recent study (J.L. Knetsch and Wei-Kang Wong, 2009) found no endowment effect when two goods were simply passed around and inspected by subjects (but not physically proximate at the time of decision), and a strong effect when an endowed good was in front of a subject. Again, this observed difference is explained by the Pavlovian consummatory mechanisms described here.

Our results are related to several other findings about the effects of displays and environmental cues on decision-making. Here we describe these findings briefly and discuss their similarities and differences with the real-exposure effect and the Pavlovian consummatory mechanisms that we think are at work.

First, a series of experiments have studied the impact of display mode on self-control. (W. Mischel and B. Moore, 1973, W. Mischel and B. Underwood, 1974, B. Shiv and A. Fedorikhin, 1999, K.J. Wenterbroch, 1998) These studies show that subjects are less likely to choose a tempting option when it is represented symbolically (e.g., in a picture) than when it is put in front of the subjects. Previous interpretations of the experiments have emphasized the tempting nature of the goods, but a mechanistic explanation has not been provided. The results in this paper suggest that Pavlovian consummatory processes could be at work in these studies, and that this might contribute to self-control problems when the tempting good is present. In fact, choosing

between immediate and delayed rewards sometimes confounds an actual physical display of the immediate reward with a symbolic or imagined delayed reward. It follows that some aspects of preference for immediacy may be intimately related with the real-exposure effect we document.

Second, several studies have found that cues associated with being watched by others seem to increase pro-sociality in simple economic games. For example, K. Haley and D.M.T. Fessler (2005) demonstrated the effect of subtle social cues on the dictator game by using a pair of eyes (to cue a sense of being watched) and noise-muffling headphones (to cue a sense of being alone). They found that the eyes cue increased giving, but the headphones did not decrease it. M. Bateson et al. (2006) found that eye pictures increased voluntary payments for coffee in an office. M. Rigdon et al. (2008) demonstrated that three small dots at the top of a piece of paper, when oriented in a way that mimics a face, increases giving in a dictator game for males (but not for females). T.C. Burnham and B. Hare (2007) found that people give more in a public good game in the presence of a robot that was built to appear lifelike. All of these are examples of how social cues at the time of decision-making can affect behavior. Although the exact mechanisms at work in these results are not known, it might also be the case that they activate highly evolved behavioral programs in response to the presence of others (which the brain might detect through the perception of real or artificial faces). Note, however, that the types of cues and mechanisms at work are different from the real-exposure effect. In our case, the triggering cue is the presence of the item itself and the mechanisms at work are Pavlovian consummatory processes that activate behaviors that lead to making contact with appetitive items. In contrast, the cue here are real or abstract faces and the underlying psychological processes are unknown.

Third, cues have also been shown to have strong effects in drug cravings and consumption. Addicts often experience a craving, and more likely to consume, when cues associated with

previous drug use are present. This often leads to relapse even after years of abstinence (see B.D. Bernheim and A. Rangel (2004) for a review of the evidence). Although direct exposure to a drug of choice is thought to trigger the type of Pavlovian consummatory mechanisms discussed in this paper, other drug cues can trigger cravings and recidivism even if the actual drug is not present. Such cues include seeing a place or friend associated with drug use, or watching films showing the use of drug paraphernalia. This is thought to operate through at least two separate mechanisms. First, drug cues trigger physiological "opponent process" which causes unpleasant withdrawal like symptoms (S. Siegel, 1975; D. Laibson, 2001). This is thought to increase the marginal utility of consuming the substance. Second, cues are also thought to trigger habitual behavioral responses that promote drug seeking behaviors even if utility maximization calculations suggest that this is not the optimal course of behavior (B.D. Bernheim and A. Rangel, 2004; D. Reddish, 2004; A. Rangel et. al., 2008).

Finally, cues can also affect behavior through a mechanism known in the behavioral economics literature as projection bias (G. Loewenstein et al., 2003). Here, environmental cues that change the current experienced utility of consuming an item (e.g., the current level of hunger or weather) can affect choices, even if cues (which can also be thought of as states) are not predictive of the actual state of the world at the time of consumption. Because such states or cues do not affect eventual consequences, their effects on choice violate the "consequentialist view" of idealized choice described in the introduction. For example, Gilbert, Gill and Wilson (2002) showed that shoppers who were given a muffin to eat before entering a supermarket were more likely to restrict their purchases to the items in their shopping list, rather than adding unplanned impulse purchases. This finding shows that the value assigned to foods that will not be eaten until much later depended on the level of hunger at the time of decision, which is presumably

uncorrelated with the hunger state at the time of consumption (see also D. Read and B. van Leeuwen, 1998, for a closely related result). M. Conlin et al. (2007) report field evidence for a similar effect of weather: unusually cold weather at the time of ordering cold-weather clothes from a catalog predicts whether goods are later returned. Note that projection bias is quite distinct from the real-exposure effect that we have identified in this paper. In projection bias, cues affect behavior because subjects overestimate the extent to which the future experience utility of consuming an item will be equal to the experienced utility of consuming it now. Thus, it is due to a cognitive bias. In addition, the cues at work have nothing to do with the physical presence of the good itself.

Figure 1. Results for Experiment 1: Consumer's willingness-to-pay for a food item is larger when it is physically present. A) Average bids and standard error bars in the three treatments: text, image, and real presentation. There was no significant difference between the text and picture conditions, but both were significantly lower than bids in the real condition ($p<0.004$). B) Bids as a function of self-reported liking ratings for each of the treatments. There was no statistically significant change in the linear slope of these curves across conditions.

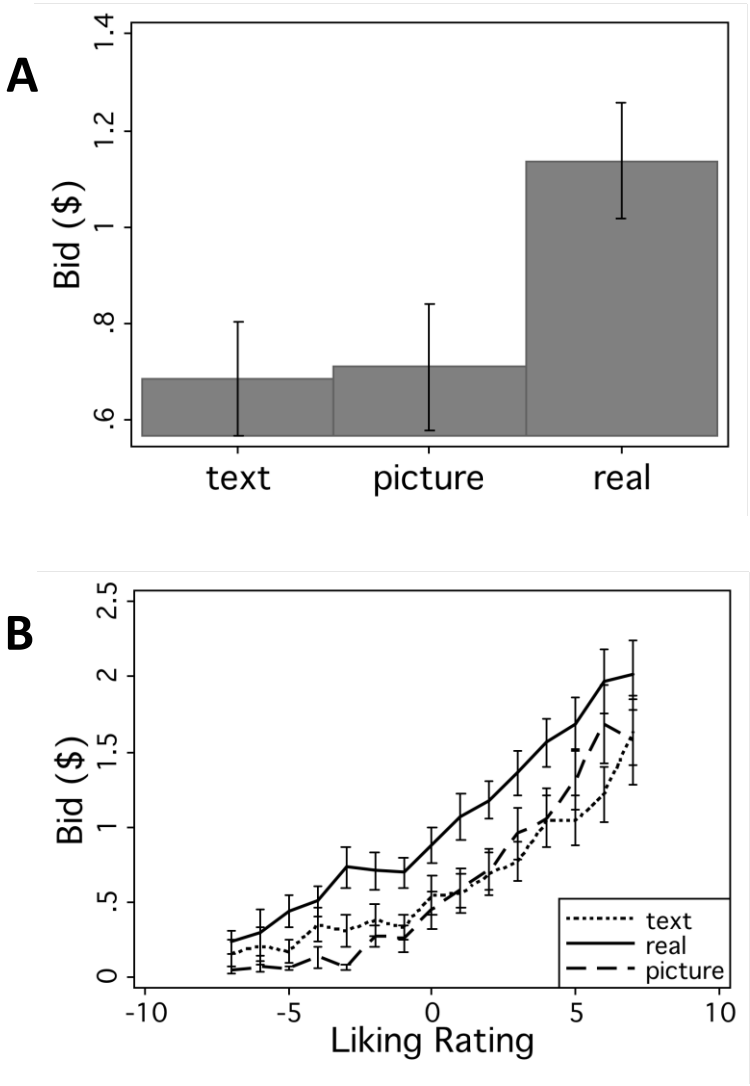


Figure 2. Results for Experiment 2: Consumer's willingness-to-pay for a trinket is larger when it is physically present. A) Average bids and standard error bars in the three treatments: text, image, and real presentation. There was no significant difference between the text and picture conditions, but both were significantly lower than bids in the real condition ($p<0.008$). B) Bids as a function of self-reported liking ratings for each of the treatments. There was no statistically significant change in the linear slope of these curves across conditions.

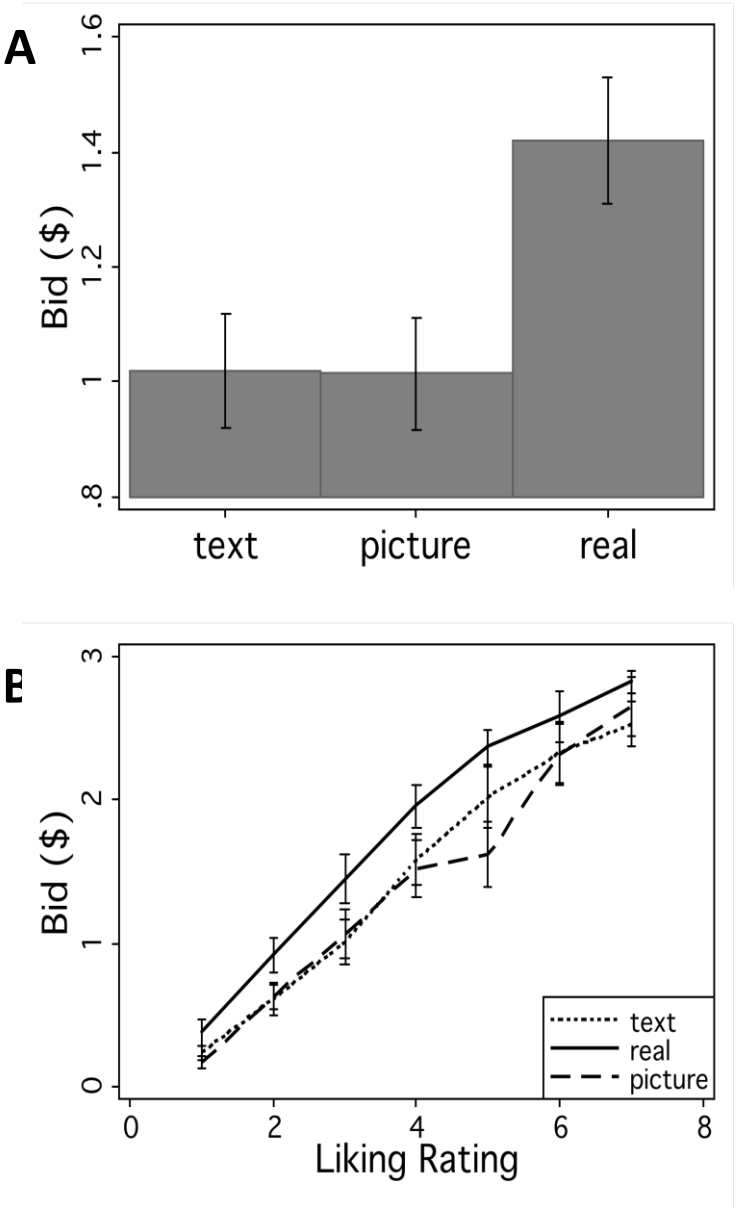


Figure 3. Results of Experiment 3: A small taste of the food item in the picture condition has no impact on subject's willingness-to-pay. A) A comparison of average bids and standard error bars in the picture, taste, and real presentation treatments. There was no significant difference between the picture and taste conditions, but the bids in the taste case were lower than in the real condition ($p<0.029$). B) Bids as a function of self-reported liking ratings for each of the treatments. There was no statistically significant change in the linear slope of these curves across conditions.

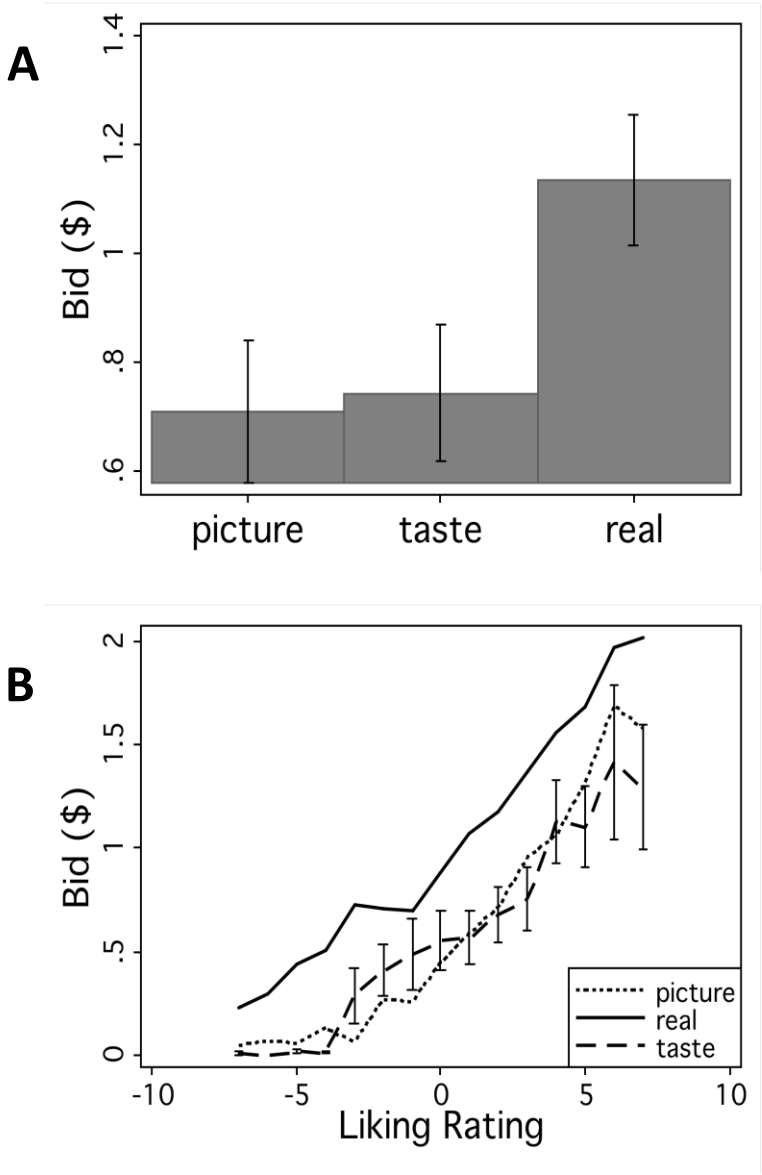
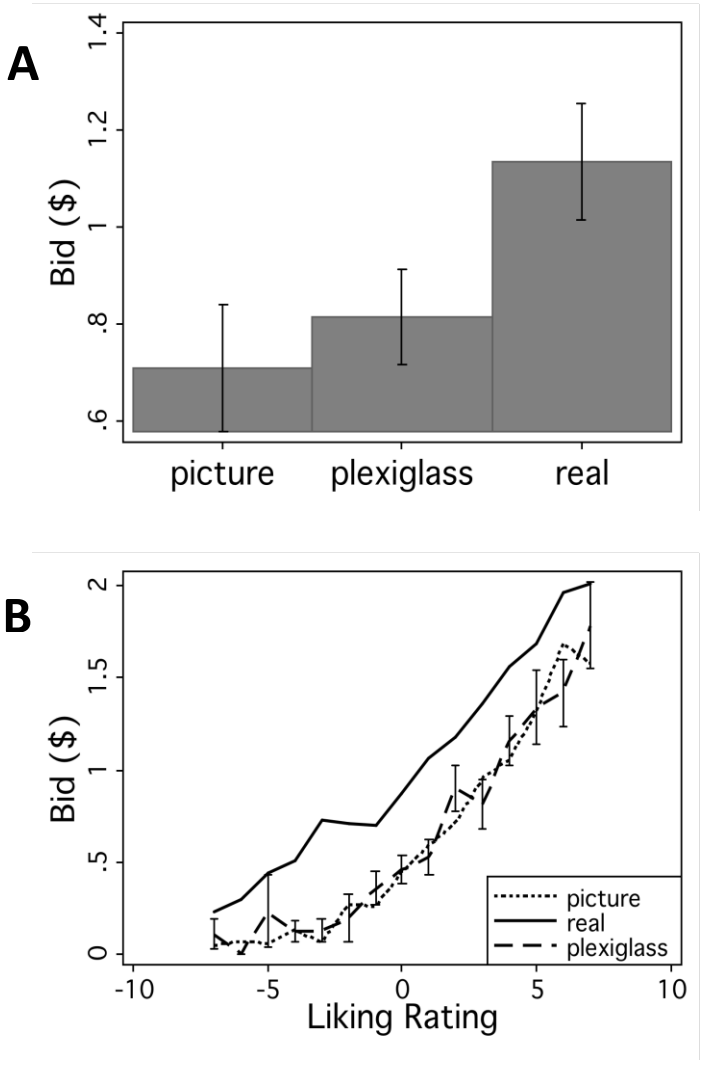


Figure 4. Results of Experiment 4: The introduction of a transparent Plexiglas barrier between the subjects and the foods eliminates the difference between the real and picture conditions. A) A comparison of average bids and standard error bars in the picture, real with Plexiglas, and real without Plexiglas conditions. There was no significant difference between the picture and Plexiglas conditions, but the bids in the Plexiglas case were lower than in the real condition ($p<0.042$). B) Bids as a function of self-reported liking ratings for each of the treatments. There was no statistically significant change in the linear slope of these curves across conditions.



References (alphabetical)

Balleine, Bernard W; Daw, N. and O'Doherty, J. "Multiple Forms of Value Learning and the Function of Dopamine," P. W. Glimcher, E. Fehr, C. Camerer and R. A. Poldrack, *Neuroeconomics: Decision-Making and the Brain*. New York: Elsevier, 2008.

Balleine, B. W. "Neural Bases of Food-Seeking: Affect, Arousal and Reward in Corticostriatolimbic Circuits." *Physiol Behav*, 2005, 86(5), 717–30.

Bateson, M.; Nettle, D. and Roberts, G. "Cues of Being Watched Enhance Cooperation in a Real-World Setting." *Biol Lett*, 2006, 2(3), 412–4.

Bernheim, B.D. and Rangel, A. "Addiction and Cue-triggered Processes." *American Economic Review*, 2004, 94(5), 1558–1590.

Burnham, T.C. and Hare, B. "Engineering Human Cooperation: Does Involuntary Neural Activation Increase Public Good Contributions?" *Human Nature*, 2007, 18(2), 88–108.

Conlin, Michael; O'Donoghue, Ted; and Rabin, Matthew. "Projection bias in catalog orders." *American Economic Review*, 2007, 97(4), 1217–1249.

Gilbert, Daniel T., Michael J. Gill, and Timothy D. Wilson, "The Future Is Now: Temporal Correction in Affective Forecasting," *Organizational Behavior and Human Decision Processes*, 2002, 430–444.

Haley, Kevin, and Fesseler, Daniel M.T. "Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game." *Evolution and Human Behavior*, 2005, 26, 245–256.

Holmes, N. P. and Spence, C. "Multisensory Integration: Space, Time and Superadditivity." *Curr Biol*, 2005, 15(18), 762–4.

Kahneman, Daniel; Knetsch, Jack L and Thaler, Richard. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy*, 1990, 99, 1325–48.

Knetsch, Jack L and Sinden, Jack A. "Willingness to Pay and Compensation Demanded: Experimental Evidence for an Unexpected Disparity in Measures of Value." *Quarterly Journal of Economics*, 1984, 99, 507–21.

Knetsch, Jack L and Wong, Wei-Kang. "To Trade or Not to Trade: The Endowment Effect and Manipulaitons of the Reference State," *Simon Fraser University Working Paper*. 2009.

Laibson, David. "A Cue-theory of consumption." *Quarterly Journal of Economics*, 2001, 116(1), 81–119.

Lichtenstein, S. and Slovic, P. The Construction of Preference. Cambridge University Press. Cambridge, 2006.

Loewenstein, G.; O'Donohue, T.; and Rabin, M. "Projection Bias in Predicting Future Utility", 2003, 118(4), 1209–1248.

Loewenstein, G. "Out of Control: Visceral Influences on Behavior." *Organizational Behavior and Human Decision Processes*, 1996, 65, 272–92.

McFadden, D. "Free Markets and Fettered Consumers." *American Economic Review*, 96(1), 5–29.

McNeil, B. J.; Pauker, S. G.; Sox, H. C., Jr. and Tversky, A. "On the Elicitation of Preferences for Alternative Therapies." *N Engl J Med*, 1982, 306(21), 1259–62.

Mischel, W. and Moore, B. "Effects of Attention to Symbolically Presented Rewards on Self-Control." *Journal of Personality and Social Psychology*, 1973, 28, 172–79.

Mischel, W. and Underwood, B. "Instrumental Ideation in Delay of Gratification." *Child Dev*, 1974, 45(4), 1083–8.

Peck, Joann and Shu, Suzanne B. "The Effect of Mere Touch on Perceived Ownership." *Journal of Consumer Research*, forthcoming, 36.

Rangel, A.; Camerer, C. and Montague, P. R. "A Framework for Studying the Neurobiology of Value-Based Decision Making." *Nat Rev Neurosci*, 2008, 9(7), 545–56.

Read, Daniel and van Leeuwen, Barbara. "Predicting Hunger: The Effects of Appetite and Delay on Choice." *Organizational Behavior and Human Decision Processes*, 1998, LXXVI, 189–205.

Reddish, David. "Addiction as a Computational Process Gone Awry." *Science,* 2004, 306(5703), 1944–7.

Rigdon, Mary; Ishii, Keiko,; Watabe, Motoki; and Kitayama, Shinobu. "Minimal Social Cues in the Dictator Game." University of Michigan, manuscript, 2008.

Seymour, B.; Singer, T. and Dolan, R. "The Neurobiology of Punishment." *Nat Rev Neurosci*, 2007, 8(4), 300–11.

Siegel, Shepard. "Evidence from Rats that Morphine Tolerance is a Learned Response." *Journal of Comparative and Physiological Psychology*, 1975, 89, 498–506.

Shiv, Baba and Fedorikhin, Alexander. "Heart and Mind in Conflict: The Interplay of Affect and Cognition in Consumer Decision Making." *Journal of Consumer Research*, 1999, 26(3), 278–92.

Solnick, S. "Cash and Alternate Methods of Accounting in an Experimental Game." *Journal of Economic Behavior and Organization*, 2007, 62, 216–31.

Wenterbroch, K.J. "Consumption Self-Control Via Purchase Quantity Rationing of Virtue and Vice." *Marketing Science*, 1998, 17, 317–37.

ORBITOFRONTAL CORTEX RESPONSES REFLECT CONSUMPTION AND NEWS-BASED EXPERIENCED UTILITY SIGNALS

Benjamin Bushong[1]

Matthew Rabin[2]

Colin F. Camerer[1,3]

Antonio Rangel[1,3]

[1] Benjamin Bushong, Antonio Rangel, and Colin Camerer: California Institute of Technology, HSS, 1200 E. California Blvd., Pasadena, CA, 91125.
[2] Matthew Rabin: University of California, Berkeley, Department of Economics, 549 Evans Hall, #3880, Berkeley, CA 94720.
[3] Antonio Rangel and Colin Camerer: California Institute of Technology, CNS, 1200 E. California Blvd., Pasadena, CA, 91125, camerer@caltech.edu and rangel@caltech.edu.

Abstract. People's pleasant and aversive experiences are accompanied by hedonic responses, also called experienced utility (EU). A basic and open question is the extent to which the EU signals encoded by the brain depend only on what is consumed, or also on the extent to which the experienced level of pleasure was better or worse than anticipated. In particular, economic and psychological theories suggest that good news increases hedonic responses and bad news decreases them. We investigated this question using a human fMRI experiment in which thirsty subjects consumed liquid rewards that were anticipated and liked to various degrees. We found that areas of the orbitofrontal cortex that are known to correlate with subjective pleasure were modulated both by the direct pleasure from consumption and by the extent to which that outcome was a surprise. We also found that subjects were more likely to later choose stimuli (fractal images) that were previously associated with positive surprises, and avoid those associated with negative surprises, which provides further support that good news is pleasurable and bad news is aversive.

Classical models in economics and psychology assume that a person's hedonic response depends only on the identity of what is consumed (e.g., water versus juice), and the individual's physiological state at the time of consumption (e.g., thirsty versus satiated) (*7*). In contrast, other models assume that the hedonic response also depends on the degree to which the consumption event was anticipated: consumption of a liked item is more pleasurable when unanticipated, and consumption of a disliked item is more aversive when unanticipated (*8–11*). By the same token, a given level of consumption is more pleasurable when unexpected. This surprise component of subjective well-being can be stated simply: good news is pleasurable and bad news is unpleasant. There is a growing consensus that the surprise component of experienced utility (EU) is an important force in people's choices and is necessary to understand risk preferences and other important economic phenomena.

Yet the focus in neuroscience has been on consumption as the sole carrier of value. In this paper, we present a pair of experiments that together suggest that both consumption and surprise about consumption are hedonic components. Experiment 1 shows that activity in the OFC is determined simultaneously by both the amount of juice subjects consumed and their surprise in the amount they consumed. With treatments where subjects received different levels of juice but fully expected what they received, we show that the OFC signal encodes consumption independent from the surprise component. With treatments where subjects received the same juice amount but did so with varying degrees of surprise, we show that the OFC signal also encodes the surprise component independent from the consumption component. Experiment 2 suggests that these OFC signals are encoding hedonic value. We associated fractal images with surprise and non-surprise juice rewards and found that the value induced in a later choice task over those fractals reflected both the pure consumption value of the juice and the surprise value.

Our results are intrinsically inconclusive and tentative in ways we return to in the discussion and the end of the paper. But taken together these experiments strongly suggest a re-orientation in focus and interpretation in neural studies and an avenue for future research to put greater emphasis on the comparable "value" nature of consumption and news. Within an experimental session, consumption of juice and the surprise over that consumption independently induced the same sort of activity in OFC commonly associated with value, and do so in a way that matched the two components emphasized in behaviorally and psychologically motivated models of utility from economic and psychological research, and induced the same sort of value in an associated stimulus.

I. EXPERIMENT 1: FMRI TASK

The first experiment, summarized in Fig. 1A (see SOM for details), used human fMRI to study the EU responses in areas of the orbitofrontal cortex (OFC) that have been previously shown to correlate with subjectively experienced pleasure *(12–19)*. Thirsty participants were given small amounts of two types of liquid rewards inside the scanner: a highly liked juice (mean pleasantness rating=3.50, SD=0.508, scale -4 to 4) and a neutral control solution (mean rating=-0.406, SD=0.712; p<.0001 one-sided paired t-test). We varied parametrically the extent to which the two possible outcomes were expected: a cue at the beginning of every trial indicated the probability p of obtaining the juice (p = 1, 2/3, 1/3, 0). The neutral solution was delivered with probability $1-p$. Fig. 1B-C describe the basic idea of the experiment. If the EU signals encoded in OFC depend only on what is consumed, then they should respond to what is consumed, but not

to the extent to which it is a surprise (i.e., they should be independent of *p*; Fig. 1B). In contrast, if the EU signals encoded in OFC depend also on expectations, then their response should depend both on what is consumed, and on the extent to which that consumption event was anticipated (i.e., they should be dependent on both outcome and *p*, as shown in Fig. 1C).

In order to evaluate these predictions we analyzed the BOLD data in several steps. First, we estimated a simple general linear model of BOLD activity at the time of consumption to identify regions of the orbitofrontal cortex (OFC) that were more responsive to the more liked juice than to the less liked neutral liquid (see SOM for details). Since both models predict that activity in an area responsive to EU should be stronger for the juice than for the neutral liquid, this provides a localizer for areas that respond in a manner consistent with EU coding, without biasing responses in favor of either model. We found an area of medial OFC in which responses were stronger during the consumption of the juice (Fig. 2A, Table S1). This result is consistent with a growing number of previous studies that have shown that activity in this area of OFC correlates positively with measures of EU for a wide class of stimuli using fMRI (*12–19*) and PET (*20, 21*). Furthermore, devaluation paradigms show that the OFC response when consuming a stimulus is lower after subjects are fed to satiation (*12, 13, 22–24*).

Second, we carried out a region-of-interest analysis of the individual responses in this area of OFC to determine if they responded only to the type of liquid received, or also to the degree of surprise. We did this by estimating a mixed effects linear regression of how the individual responses in OFC for each of the six possible experimental outcomes were modulated by two variables: an indicator variable for receiving the more liked juice, and a variable measure the extent to which the experienced pleasure exceeded or fell below expectations for the trial (see SOM for exact definitions). We found that the response in OFC was significantly affected by

both ($p<.023$ for outcome regressor; $p<.056$ for surprise regressor), which is consistent with the existence of a surprise component of EU. As described in the SOMs, this regression model also allowed us to estimate the relative weight given to the direct and surprise components of the EU signal in the OFC responses. We also found that the effect of the surprise component on the OFC responses was four times stronger than the direct component (coefficients: 0.795 vs. 0.215, $p<0.003$).

Third, we carried out a Bayesian model comparison (*25*) of the extent to which the following two models accounted for responses in the area of OFC identified in Fig. 2A. The first model assumed that activity at consumption was modulated only by the direct component of EU, while the second model assumed that it was modulated by both (with relative weights equal to those estimated in the previous ROI). We found that the model with both components had an exceedance probability of $>$ 99.2% over the model that only included direct components.

Fourth, we carried out an additional region-of-interest analysis of the responses in OFC to investigate if positive and negative surprises had symmetric effects. We did this by estimating a new mixed effects linear model of the individual OFC responses with three independent variables: 1) the direct components of EU for each trial, 2) the surprise component of EU interacted with a dummy variable for positive surprise trials, and 3) the surprise component of EU interacted with a dummy variable for negative surprise trials. We found that the effect of the negative surprise measure was not significantly different from the effect of the positive surprise measure.

II. EXPERIMENT 2: CONDITIONING TASK

Although the results from the first experiment are consistent with the surprise model of EU, they are not sufficient to establish that the effect of surprises have an effect on experienced hedonics. Armed with only the neural data we cannot rule out the possibility that only the non-surprise component of the EU actually affects the hedonic experience (perhaps because this area contains a mixture of neurons encoding separately for hedonics and degree of surprise). The surprise component of the EU signal is mathematically identical with a (cognitive) prediction-error signal that has been found to play a critical role in reinforcement learning, and which is distinct from a hedonic response (*2, 3, 26–29*). As a result, without further evidence we cannot rule out that the OFC might contain a combination of surprise-independent EU and non-hedonic prediction-error signals.

We addressed both concerns with a second behavioral conditioning task that is closely related to the first experiment. As before, subjects received either the juice or a neutral solution after being exposed to a cue indicating the probability of receiving each of the two liquids. There were three key differences with the first experiment. First, there were only three lottery conditions ($p=1,1/2,0$), which leads to the four outcomes depicted in Fig. 3A. Second, at the time the liquid was revealed, subjects were exposed to one of four fractal stimuli. Importantly, for each subject the same fractal was always paired with the same outcome condition. Third, after 60 rounds of this conditioning task, subjects were asked to make two choices. Both decisions involved a choice between two pairs of equal probability lotteries between two fractals, which are depicted in Fig. 3B. Afterwards, we randomly selected one of the fractals from the chosen

pairs, and participants received the outcome (juice or neutral liquid) that was associated with those fractals.

The first choice allowed us to examine if the hedonic response was modulated by the type of outcome: the first lottery gave 50% probability to the fractals associated with the juice outcome, whereas the second lottery gave 50% probability to the fractals associated with the neutral liquid outcome. Note that in As a result, if the EU signal increases with the type of outcome, individuals should choose the first lottery over the second one. This is what we found: 27 out of 30 participants ($p<.00001$, binomial test) chose the lotteries associated with the juice outcome.

The second choice allowed us to examine if the hedonic response was modulated by surprise: the first lottery gave 50% probability to the fractal associated with the surprise appetitive outcome and 50% he fractal associated with the fully expected neutral liquid; the second lottery gave 50% to the fractals associated the fully expected appetitive outcome and 50% to the fractal associated with surprise neutral liquid. Note that since the value acquired by the fractals is equal to the hedonic response of the outcome to which they are paired, if there is a surprise signal the individuals should choose the first lottery over the second one. This is what we found: 23 out of 30 ($p<.003$, binomial test) chose the first lottery over the second one.

The specific design we employed allowed us to conclude that second experimental results of the conditioning experiment can only be explained if there are surprised-related hedonics. To support this claim, we ran a standard reinforcement learning (RL) analysis of our experiment under the assumption that there are no surprise-related hedonics, and found that this generated predictions inconsistent with the choice experiment (see Methods). We then employed a standard RL analysis of our experiment under the assumption that there are surprise-related hedonics, and show that this generates predictions consistent with the choice experiment.

Together, the results of the two experiments allow us to draw several conclusions. First, there is strong evidence in favor of the hypothesis that the EU signal encoded in OFC is modulated both by the direct pleasure derived from consumption, and by the extent to which the level of pleasure obtained exceeds or falls short of expectations. Second, a neurometric estimate suggests that the surprise component might be stronger than the direct component of EU. Third, we found that negative and positive surprises have a similar impact in OFC. Finally, we found that the surprise component of EU affects the value that is learned for stimuli through affective conditioning.

III. DISCUSSION

Several previous studies have addressed related questions, but their results could not establish if the EU signals in OFC are responsive to the extent to which the pleasure at consumption exceeds or fall behind expectations. Some studies have tried to measure the impact of expectations on EU directly by manipulating the expectation of obtaining a monetary reward and then measuring EU using subjective reports (*11, 30*). Their results have been consistent with the surprise model, but did not address how EU modulated brain activity. Second, several neurophysiology (*31, 32*) and fMRI (*33, 34*) studies have found responses in orbitofrontal cortex at the time of consumption that were also consistent with the surprise model of EU. Unfortunately, their results were inconclusive because they could not rule out that this area was encoding prediction errors instead of reference-dependent EU (and in fact, they have generally been interpreted as prediction errors). This study suggests that such signals are inconsistent with non-hedonic prediction-error signals and are better explained as surprise-related hedonics.

The fact that the surprise component of EU was passed to the fractals in the conditioning task is important because it demonstrates that the value that subjects learn to assign to a stimulus depend on the predictability of their outcomes during learning. Some psychology studies have argued that individuals who make choices assuming that their preferences have a surprise component are making a mistake (*35*). The results presented here argue against that interpretation of the data—if, in fact, good news leads to a pleasurable experience, then it is perfectly reasonable for individuals to seek out such pleasure.

The consumption and surprise components of utility might be computed in distinct areas before they are integrated into a net hedonic signal. A natural hypothesis is that the surprise component is based on a prediction-error signal computed by dopamine release into the striatum. This signal might then be passed to many other areas, including the OFC areas involved in encoding total hedonics. However, it must also be the case that these components are integrated somewhere into a hedonic signal so that it can perform all of its roles. The only such signal that we found (at our omnibus threshold) is in the OFC.

These results have significant implications in various domains. First, economic and psychological theories based on the surprise models of EU have distinctive normative and policy implications (*9*, *36*, *37*). The results of this paper provide support for these types of models and their implications. Second, our results suggest that the simple type of value learning models that have been used in neuroscience and psychology, such as reinforcement learning (*1*, *38*), which ignore the surprise components of EU, are likely to mispredict the values learned by subjects in a large class of situations. Third, the results provide neural foundations for the fact that expectations can affect subjective well-being so that, for example, a fully anticipated increase in consumption could have a smaller impact on EU than one might otherwise expect (*39*).

Figure 1. Basic experiment and model predictions. A) Timeline for a typical trial of the fMRI task. Participants were shown a cue describing a lottery between receiving an appetitive juice (with probability *p*) and a neutral solution (with probability *1-p*). at the end of the trial. After a short 1–6 s random delay, one of the liquids was delivered and participants held the liquid in their mouth for 6s until they were instructed to swallow it. B) Predicted *consEU* signals for each condition. C) Predicted *totalEU* for each condition, for the case in which *totalEU = 0.5 consEU + 0.5 newsEU*.

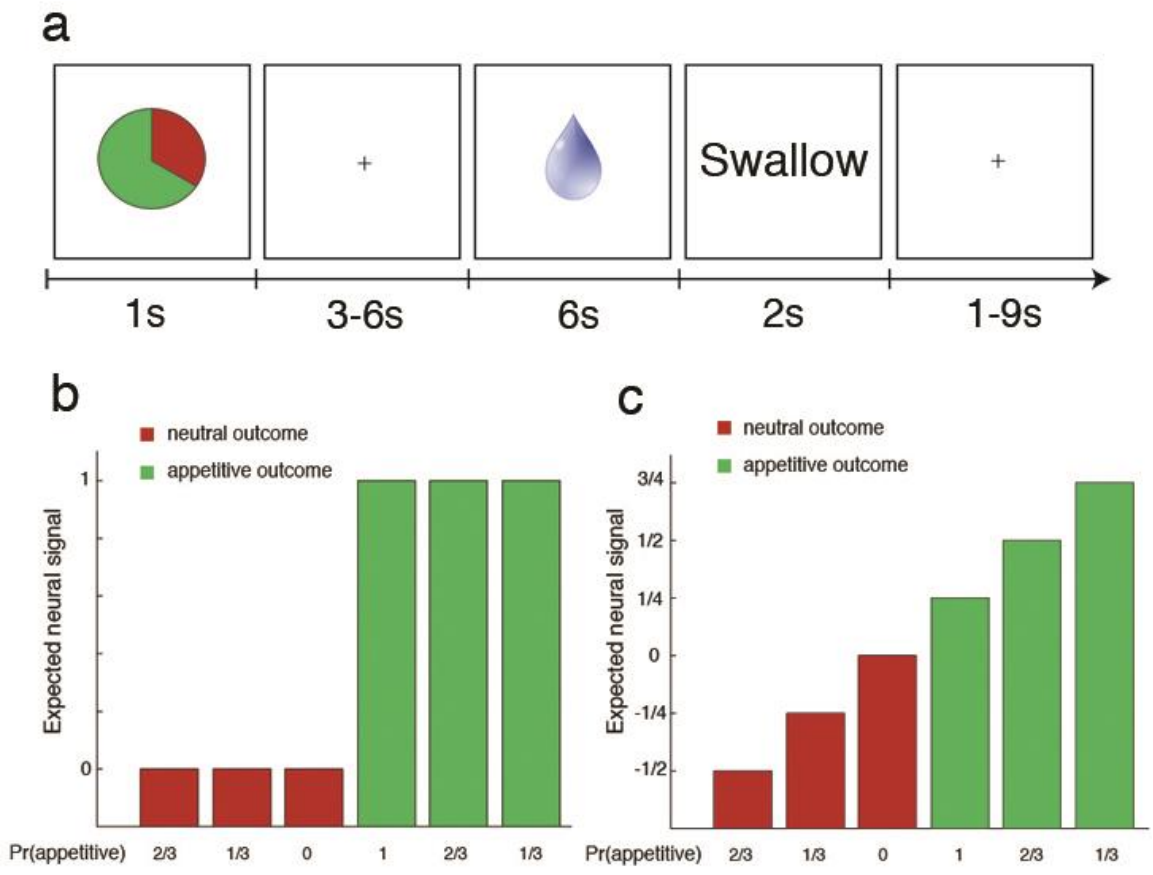


Figure 2. Experiment 1 results. A) Area of mOFC that responded more strongly to the highly liked juice than to the neutral liquid at the time of consumption (shown at p<.001 uncorrected with a 10 voxel extent threshold; overlaid on average anatomical image). B) Mean brain response in the mOFC region of interest as a function of the experimental condition.

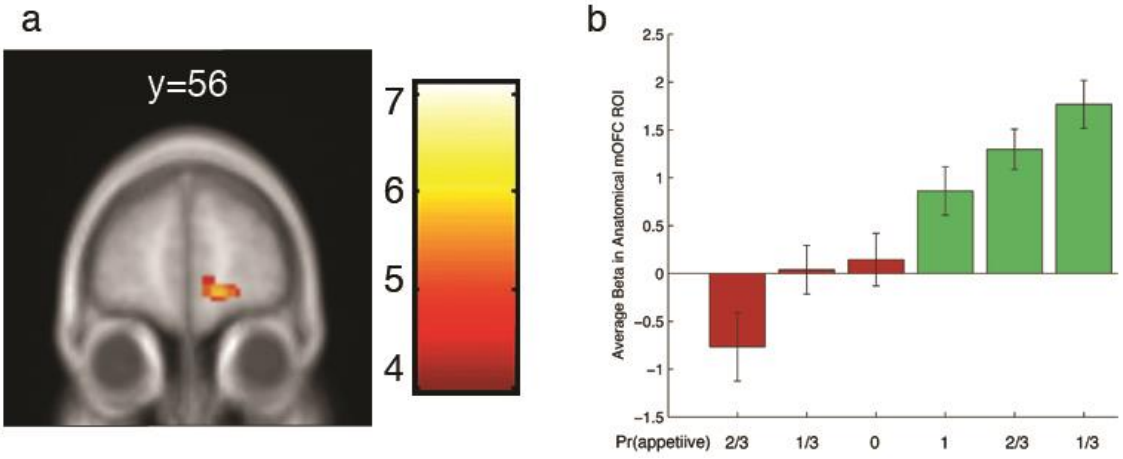


Figure 3. Experiment 2 design and results. A) Training phase of the conditioning experiment. Four fractals were used, and each fractal was always paired with the same experimental condition. The assignment of fractals to conditions was randomized across subjects. B) Description of choices presented to the subjects after the training phase.

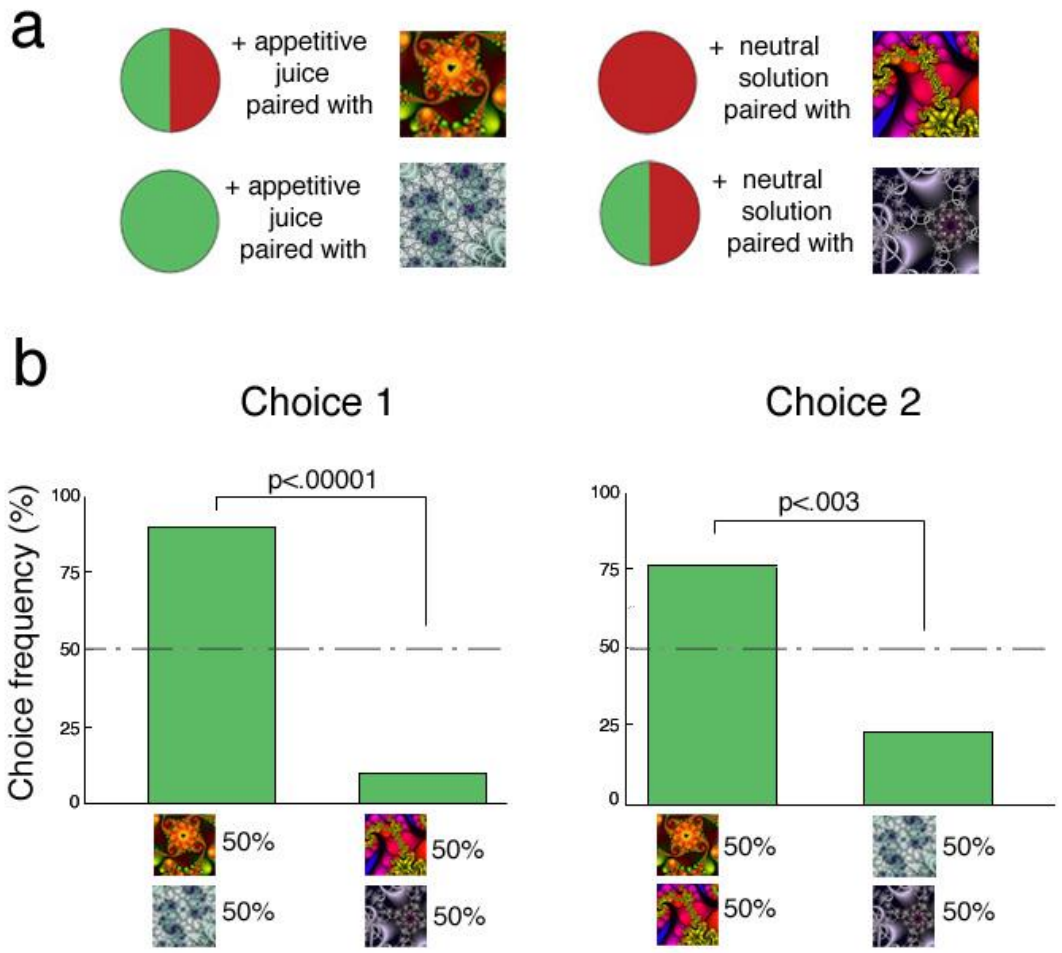


Table 1. GLM 1. Areas in which BOLD responses are higher for the juice than the neutral liquid at the time of swallowing.

Region	Side	BA	MNI Coordinates (peak voxels)	Z
Pre/Postcentral Gyrus	L R	2,4,6	-57, -15, 39 66, -18, 36	4.96 4.30
Parietal	L R	7	-18, -60, 66 21, -66, 69	4.85 4.63
Middle temporal	L R	19,39	-56, -66, 6 60, -63, 9	4.50 4.54
Orbitofrontal Cortex	R	10	15, 57, -9	4.32
Results thresholded at p<.05, whole-brain cluster correction (t = 4.2155)				

Appendix 1.

METHODS

I. THEORY

In this study we compare two popular models of experienced utility (EU). In order to describe the models we need some notation. Let U denote the EU signal at the time of consumption; x=J,N denote the liquid consumed (juice or neutral liquid); and p=1,2/3,1/3,0 denote the probability of receiving the juice for that consumption event.

Model 1: Standard experienced-utility (SEU) model.

This model assumes that the EU signal depends only on the amount of reward generated by the liquid consumed. Thus, it predicts that $U(J|p) = U(J|p') = u(J) > U(N|p) = U(N|p') = u(N)$ for all p and p', where $u(x)$ is a function measuring the direct pleasure derived directly from consuming x, independent of expectations. Fig. 1B describes the predictions of this model for our experiment for the case in which $u(J) = 1$ and $u(N) = 0$. In order to facilitate the discussion below, we refer to $u(x)$ as the direct component of EU.

Model 2: Reference-dependent experienced-utility (RDEU) models.

This class of models, which have been proposed by Koszegi and Rabin (*9*,*36–37*), Loomes and Sugden (*10*), and Bell (*8*) and others, assumes that the EU signal is given by

$$U(x|p) = (1-a)*u(x)+ a* (u(x)-E[u(x)|p]),$$

where a is a weighting factor representing the relative contribution of the two components, E is the expectation operator, and a is restricted to be between 0 and 1.

The predictions of this model are depicted in Fig. 1C for the case of $a=3/4$, $u(J) = 1$, and $u(N) = 0$. We refer to the term "$u(x)$- E[$u(x)|p$]" as the surprise component of the model, which measures how much of the basic reward received was unexpected. Note that this model assumes that EU is a linear combination of the direct and surprise components of consumption, and that the surprise term can be either positive or negative, depending on whether the actual level of direct consumption is above or below the amount that was expected.

This simple version of the model assumes that EU is a linear combination of the direct and surprise components. More general versions of the theory relax this assumption.

The regressors used in the parametric fMRI analyses are derived from the two models described here. The direct component of EU is given by $u(J)=1$ or $u(N)=0$, depending on which item was consumed. The surprise component of EU is given by $u(x)$-E[$u(x)|p$], and it also assumes that $u(J)=1$ and $u(N)=0$. The normalization of the direct component of EU is without loss of generality because it does not affect any of the statistical contrasts used in the analysis.

II. METHODS FOR EXPERIMENT 1

Subjects. Thirty-four subjects recruited from Caltech's student body participated in the experiment (38% female; mean age=21.7; age range 18–27). However, we dropped data from two subjects, as they did not meet an *a priori* excessive movement exclusion criterion during the scanning sessions. All subjects were right-handed, healthy, had normal to corrected-normal vision, and were not taking any medications that would interfere with the performance of fMRI. No subjects reporting having any food or beverage allergies. The review board of Caltech approved the study and subjects provided informed consent prior to their participation.

Task. Subjects abstained from drinking any liquids for 4 hours prior to the experiment. Upon their arrival, they tasted a very small amount (< 5 ml) of 14 liquids and rated the subjective pleasantness of each. The scale for the ratings task began with -4 = strong dislike, through 0 = neutral, to +4 = strong like. All liquids were sweet, zero pulp, and served at room temperature. For all subjects, the highest-rated juice was used in the scanning portion as the liked juice. The neutral liquid was always a neutral control solution made with water and the ionic components of saliva.

Once inside the scanner, participants received juice directly into their mouths from tubes. The timeline of events can be seen in Fig. 1A. Subjects were shown four different lotteries which differed on the probability p=1,2/3,1/3/,0 with which the juice was delivered. The neutral solution was delivered with probability $1\text{-}p$. Each lottery was depicted by a pie chart indicating the probability of the two outcomes. In order to insure a sufficient number of trials for each type of outcome the frequency of trials was given by the following:

Table 2. Frequency of delivery by experiment condition, Experiment 1.

P	**Frequency**
1	1/6
2/3	1/3
1/3	1/3
0	1/6

Subjects were explained the meaning of the lottery cues prior to scanning and were given a short quiz to insure that they understood. Subjects were asked to hold the liquid in their mouth for 6 s until they were cued to swallow it. Inter-trial intervals were pseudo-randomized between 1 and 9 seconds.

There were three runs of the task, each lasting 12 minutes 28 seconds. Each run consisted of 40 trials of the task. In order to minimize satiation effects, we limited the juice intake to ~ 60 milliliters over the course of the experiment (.5 ml per trial).

fMRI Acquisition. The fMRI data was acquired in a Siemens Trio 3.0 Tesla machine with an eight-channel phased array coil at the Caltech Brain Imaging Center. We acquired gradient echo T2*-weighted echoplanar (EPI) images with BOLD contrast. We also acquired T1-weighted structural images (1 mm cubic volumes) for anatomical localization. We acquired the BOLD data using an orientation of 30 degrees with respect to the line of anterior commissure-posterior commissure (ACPC), which has been shown to improve signal-to-noise ratio in OFC (*6*). Functional scanning parameters were as follows: echo time, 30 ms; field of view, 192 mm; in-plane resolution and slice thickness, 3 mm; repetition time, 2.75 s; flip angle, 80 degrees.

fMRI Preprocessing. We performed all image analyses using SPM5 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). We corrected images for slice

acquisition time within each volume, motion corrected the images with realignment based on the first volume, and spatially normalized the final product to the Montreal Neurological Institute EPI template. Finally, we spatially smoothed images using a Gaussian kernel with width 8 mm. We also applied intensity normalization and high-pass filtering with filter width of 128 s. We excluded data from two subjects who consistently moved in excess of 3 mm.

General Linear Model 1. The first step in the analysis of fMRI data involved estimating a model designed to identify brain regions that responded more strongly to the consumption of the juice than to the consumption of the neutral liquid. This first analysis is important because both theories predict that an area that encodes EU should exhibit a stronger hedonic response to the juice, which means that it provides a localizer for regions involved in computing the EU signal without biasing it towards either of the two models. This GLM was estimated in three steps.

First, for each individual we estimated a GLM with first order autoregression and the following three regressors: (R1) at lottery cue, and (R2) at swallowing juice, and (R3) at swallowing neutral liquid. R1 was parametrically modulated by p. R2 and R3 were unmodulated. R1 was modeled as an event with a 0 s duration aligned to the appearance of the lottery cue. R2 and R3 were modeled as an event with a 2 s duration aligned to the appearance of the swallowing instruction. All of these regressors were convolved with a canonical hemodynamic response function. The model also included session constants and motion parameters as regressors of no interest.

Second, we calculated the contrast R2 > R3 for every subject using a one-tailed t-test.

Third, we estimated a second-level mixed effects analysis over all of the subjects by computing one-sample t-tests on the single-subject contrast coefficients. The results are shown in

Table S1 and Fig. 2A. For visualization purposes only, all of the images shown in the paper and supplementary materials are thresholded at $p < .001$ uncorrected with an extent threshold of five voxels. For inference purposes, we use whole-brain corrections at the cluster level based on the algorithm implemented in the CorrClusTh program by Thomas Nichols (http://www.sph.umich.edu/~nichols/JG5/CorrClusTh.m). All anatomical localizations were performed by overlapping the t-maps on a normalized structural image averaged across subjects, and with reference to an anatomical atlas (*7*).

About GLM 1. GLM 1 assumes that the EU response takes place at the time of swallowing, but not during the 6 seconds tasting period that takes place between the delivery of the liquids and the swallowing instruction. Since previous studies have found hedonic responses at the time of tasting (*8–10*), it is important to justify this modeling assumption.

Note that, depending on how exactly the liquids are tasted and swallowed, a priori they can generate hedonic responses at tasting, at swallowing, or both. We did not have any a priori hypothesis on when within these two intervals will we observe the strongest EU like responses. As a result, we carried out the following model comparison to determine that GLM 1 best describe the hedonic responses. This was done in two steps.

First, in addition to GLM 1, we estimated an additional model GLM 1' in which hedonic responses were only modeled at the time of tasting (with a duration of 6 s). We then carried out a contrast analogous to the one for GLM 1 to identify areas that responded more strongly to the juice than to the liquid neutral at the time of tasting at our omnibus threshold of $p<0.05$ whole-brain corrected. We did not find any such areas.

Second, we estimated two additional models: a GLM 1'', which was identical to GLM 1

except that it restricted the response to both liquids at swallowing to be identical, and a GLM 1''' that included regressors at tasting (with 6 second duration) and at swallowing (with 2 second duration), separately for the juice and neutral liquid outcomes. Note that the three models are nested, with GLM 1'' being a special case of GLM 1, and GLM 1 being a special case of GLM 1'''. As a result, it is possible to compare the models using standard F-tests , which weigh the relative increase in fit to the increase in degrees of freedom.

First, we compared models GLM 1'' and GLM 1 using a simple F-test. For each individual, we averaged the F-statistics over the area of orbitofrontal cortex (OFC) identified in GLM 1 (Fig. 2A, Table S1). We found that in 31 out of 32 subjects we could reject the hypothesis at $p<0.05$ that the responses at swallow for juice and the neutral liquid were identical when comparing GLM 1'' and GLM 1. We also found that in 31 out of 32 subjects we could not reject the hypothesis that the coefficients for the regressors at tasting were equal to zero within our ROI at $p<0.05$. This is the key test when comparing GLM 1 and GLM 1'''.

Together, these results provide supporting evidence for the fact that in our experiment the responses of the OFC were consistent with the encoding of an EU signal at swallowing, but not at taste.

It is also important to emphasize that all of the analyses described so far are used to identify the area of OFC that responds in a manner consistent with EU encoding in our sample, but that these analyses are independent of the next set of tests designed to test between the SEU and RDEU models.

ROI definition and signal extraction. Based on the results of GLM 1 we defined a mask of OFC given by the intersection of the group contrast described above (thresholded at $p<.001$ unc., five

voxel extent threshold) and an anatomical mask of the ventromedial prefrontal cortex (vmPFC). The resulting mask is shown in Fig. 2.

We then constructed individual measures of the response in this area of OFC for each of the six possible experimental outcomes: (*x=N*, *p=2/3*), (*x=N*, *p=1/3*), (*x=N*, *p=0*), (*x=J*, *p=1*), (*x=J*, *p=1/3*), (*x=N*, *p=2/3*). To do this we had to estimate a new GLM 2, which is similar to GLM 1 except that now each of the six outcomes is modeled as a separate event.

In order to insure the independence of the individual response measures from the ROI analyses, the response was constructed as follows for each target subject. First, we identified the region of ROI that responds more strongly for juice than neutral liquid at consumption by running the second leaving contrast excluding the target subject. Second, we computed the average response (beta value) for each subject and condition in this mask.

ROI analysis 1. We investigated the shape of the EU signals in OFC by estimating a random effects linear regression of its responses. The dependent variable was the extracted response in OFC activity for each subject and outcome condition. The independent variables for were the direct and the surprise components of EU for each trial, which were constructed as described in the theory section above. Let *bDEU* and *bRDEU* denote the estimated coefficients of both variables. Note that *bRDEU* / (*bDEU* + *bRDEU*) provides an unbiased estimate of a. For later references, the estimated value was a= 0.795.

ROI analysis 2. We estimated a second ROI model to investigate if the correlation between the surprise component of EU and OFC activity depended on the sign of the surprise. The model was almost identical to ROI 1 except that now there were three independent variables: 1) the direct

components of EU for each trial, 2) the surprise component of EU interacted with a dummy variable for positive surprise trials, and 3) the surprise component of EU interacted with a dummy variable for negative surprise trials.

***Bayesian model comparison*.** Finally, we used a Bayesian model comparison procedure (*12*) to compare which of the two candidate models fit the BOLD responses in OFC better. The first model was GLM 1, which is the case of SEU. The second model, GLM 2, was identical except that activity at swallowing was modulated by the predictions of the RDEU model, using a value $a = 0.8$, $u(J)=1$, and $u(N)=0$ (which was taken from the results of ROI 1).

Briefly, we used the Bayesian comparison methods to test which of the two models fit best the log evidences for BOLD responses within a 12 mm sphere (1.5x smoothing kernel size) centered on the OFC group peak for GLM 1. Note that this procedure treats the model as a random variable and estimating the parameters of a Dirichlet distribution, which describes the probabilities for all models considered. These probabilities then define a multinominal distribution over model space, allowing one to compute how likely it is that a model generated the subjects' data. To decide which model is more likely, we use the conditional model probabilities to quantify an exceedance probability, i.e., a belief that a particular model is more likely than the other model, given the group data.

III. METHODS FOR EXPERIMENT 2

***Task*.** As in Experiment 1, subjects abstained from drinking any liquids for 4 hours prior to the experiment. Upon their arrival, they tasted a very small amount (< 5 ml) of 14 liquids and rated the subjective pleasantness of each. The scale for the ratings task began with -4 = strong dislike,

through 0 = neutral, to +4 = strong like. All liquids were sweet, zero pulp, and served at room temperature. For all subjects, the highest-rated juice was used in the conditioning portion as the liked juice. The neutral liquid was always a neutral control solution made with water and the ionic components of saliva.

Next, participants received juice directly into their mouths from tubes while seated at a table with a keyboard. Subjects were shown four different lotteries which differed on the probability p=1,1/2,1 which indicated the probability that the juice would be delivered. The neutral solution was delivered with probability $1-p$. Each lottery was depicted by a pie chart indicating the probability of the two outcomes.

Each lottery was paired with a fractal which was shown at the delivery stage (the "conditioned fractal"; see Table 3 below). Fractals were generated from the Mandelbrot set from an online archive—the exact fractals used for particular subjects was randomized. During the conditioning stage, the subject passively viewed the lotteries and fractals and drank the juice. There was no choice task. Conditioning lasted for ~ 60 rounds.

Table 3. Frequency of delivery by experiment condition, Experiment 2.

P	**Outcome**	**Fractal shown**
1	Juice	1
½	Juice	2
½	Neutral	3
0	Neutral	4

After the conditioning task, subjects were asked to choose between two lotteries over fractals—the outcome of the lottery (determined by a coin flip) would determine a final sip of either juice or neutral solution. Subjects received the outcome depending on their choice, were paid and left.

References

1. Sutton, R.S. & Barto, A.G. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 1998.

2. Schultz, W., Dayan, P. & Montague, P.R. "A neural substrate of prediction and reward." *Science*, 1997, 275, 1593–9.

3. Niv, Y. & Montague, P.R. Theoretical and empirical studies of learning. in Neuroeconomics: Decision-Making and the Brain (ed., P.W. Glimcher, E. Fehr, C. Camerer & R.A. Poldrack). New York: Elsevier, 2008.

4. Poldrack, R.A., et al. "Interactive memory systems in the human brain." *Nature*, 2001, 414, 546–50.

5. Adcock, R.A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B. & Gabrieli, J.D. "Reward-motivated learning: mesolimbic activation precedes memory formation." *Neuron*, 2006, 50, 507–17.

6. Kahneman, D., Wakker, P.P. & Sarin, R. "Back to Bentham? Explorations of Experienced Utility." Quarterly Journal of Economics, 1997, 112, 375–405.

7. Mas-Colell, A., Whinston, M. & Green, J. Microeconomic Theory. Cambridge: Cambridge University Press, 1995.

8. Bell, D.E. "Disappointment in decision making under uncertainty." *Operations Research*, 1985, 33, 1–27.

9. Koszegi, B. & Rabin, M. "A model of reference-dependent preferences." *Quarterly Journal of Economics*, 2006, 1221, 1133–65.

10. Loomes, G. & Sugden, R. "Disappointment and dynamic consistency in choice under uncertainty." *Review of Economic Studies*, 1986, 53.

11. Mellers, B.A., Schwartz, A., Ho, K. & Ritov, I. "Decision affect theory: Emotional reactions to the outcomes of risky options." *Psychological Science*, 1997, 8, 423–9.

12. Kringelbach, M.L., O'Doherty, J., Rolls, E.T. & Andrews, C. "Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness." *Cereb Cortex*, 2003, 13, 1064–71.

13. de Araujo, I.E., Kringelbach, M.L., Rolls, E.T. & McGlone, F. "Human cortical responses to water in the mouth, and the effects of thirst." *J Neurophysiol*, 2003, 90, 1865–76.

14. de Araujo, I.E., Rolls, E.T., Kringelbach, M.L., McGlone, F. & Phillips, N. "Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain." *Eur J Neurosci*, 2003, 2059–68.

15. Grabenhorst, F., Rolls, E.T., Parris, B.A. & d'Souza, A.A. "How the brain represents the reward value of fat in the mouth." *Cereb Cortex*, 2010, 20, 1082–91.

16. McClure, S.M., et al. "Neural correlates of behavioral preference for culturally familiar drinks." *Neuron*, 2004, 44, 379–87.

17. Plassmann, H., O'Doherty, J., Shiv, B. & Rangel, A. "Marketing actions can modulate neural representations of experienced pleasantness." *Proc Natl Acad Sci*, 2008, 105, 1050–4.

18. Anderson, A.K., et al. "Dissociated neural representations of intensity and valence in human olfaction." *Nat Neurosci*, 2003, 6, 196–202.

19. Vollm, B.A., et al. "Methamphetamine activates reward circuitry in drug naive human subjects." *Neuropsychopharmacology*, 2004, 29, 1715–22.

20. Blood, A.J. & Zatorre, R.J. "Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion." *Proc Natl Acad Sci*, 2001, 98, 11818–23.

21. Blood, A.J., Zatorre, R.J., Bermudez, P. & Evans, A.C. "Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions." *Nat Neurosci*, 1999, 2, 382–7.

22. O'Doherty, J., et al. "Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex." *Neuroreport*, 2000, 11, 893–7.

23. Critchley, H.D. & Rolls, E.T. Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *J Neurophysiol*, 1996, 75, 1673–1686.

24. Small, D.M., Zatorre, R.J., Dagher, A., Evans, A.C. & Jones-Gotman, M. "Changes in brain activity related to eating chocolate: from pleasure to aversion." *Brain*, 2001, 124, 1720–33.

25. Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J. & Friston, K.J. "Bayesian model selection for group studies." *NeuroImage*, 2009, 46, 1004–17.

26. McClure, S.M., Berns, G.S. & Montague, P.R. "Temporal prediction errors in a passive learning task activate human striatum." *Neuron*, 2003, 38, 339–46.

27. O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. "Temporal difference models and reward-related learning in the human brain." *Neuron*, 2003, 38, 329–37.

28. Abler, B., Walter, H., Erk, S., Kammerer, H. & Spitzer, M. "Prediction error as a linear function of reward probability is coded in human nucleus accumbens." *NeuroImage*, 2006, 31, 790–5.

29. Aharon, I., et al. "Beautiful faces have variable reward value: fMRI and behavioral evidence." *Neuron*, 2001, 32, 537–51.

30. Larsen, J.T., McGraw, A.P., Mellers, B.A. & Cacioppo, J.T. "The agony of victory and thrill of defeat: Mixed emotional reactions to disappointing wins and relieving losses." *Psychological Science*, 2004, 15, 325–30.

31. Tremblay, L. & Schultz, W. "Relative reward preference in primate orbitofrontal cortex." *Nature*, 1999, 398, 704–8.

32. Hosokawa, T., Kato, K., Inoue, M. & Mikami, A. "Neurons in the macaque orbitofrontal cortex code relative preference of both rewarding and aversive outcomes." *Neurosci Res*, 2007, 57, 434–45.

33. Elliott, R., Agnew, Z. & Deakin, J.F. "Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans." *Eur J Neurosci*, 2008, 27, 2213–8.

34. Knutson, B., Fong, G.W., Bennett, S.M., Adams, C.M. & Hommer, D. A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *NeuroImage*, 2003, 18, 263–272.

35. Kermer, D., Driver-Linn, E., Wilson, T.D. & Gilbert, D.G. "Loss aversion is an affective forecasting error." *Psychological Science*, 2006, 17, 649–53.

36. Koszegi, B. & Rabin, M. "Reference-dependent risk attitudes." *American Economic Review*, 2007, 97, 1047–73.

37. Koszegi, B. & Rabin, M. "Reference-dependent consumption plans." *American Economic Review*, 2009, 99, 909–36.

38. Rescola, R.A. & Wagner, A.R. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In Classical Conditioning II: Current

Research and Theory (ed., A.H. Black & W.F. Prokasy) 406–412. New York: Appleton Century Crofts, 1972.

39. Keely, L.C. "Why isn't growth making us happier? Utility on the hedonic treadmill." *Journal of Economic Behavior and Organization*, 2004, 5, 333–55.

COMPUTATIONS IN DORSOLATERAL PREFRONTAL CORTEX AND TEMPOROPARIETAL JUNCTION SUPPORT DISSOCIABLE MOTIVES FOR ALTRUISTIC BEHAVIOR

Cendri A. Hutcherson[1]

Benjamin Bushong[1]

Matthew Rabin[2]

Antonio Rangel[1,3]

[1] Cendri Hutcherson, Benjamin Bushong, and Antonio Rangel: California Institute of Technology, HSS, 1200 E. California Blvd., Pasadena, CA, 91125.
[2] Matthew Rabin: University of California, Berkeley, Department of Economics, 549 Evans Hall, #3880, Berkeley, CA 94720.
[3] Antonio Rangel: California Institute of Technology, CNS, 1200 E. California Blvd., Pasadena, CA, 91125, rangel@caltech.edu.

Abstract. What are the computational and neurobiological mechanisms that support altruistic behavior? We used fMRI to investigate the neural mechanisms underlying two distinct motivations for altruism: 1) normative preferences, in which individuals care about appearing ethical to themselves or others, and 2) direct preferences, where individuals care about their own and others' outcomes. We found that the motives are supported by dissociable neural processes. The dlPFC supports normative preferences both by suppressing the influence of valuation regions like the vmPFC and by directly influencing response selection. In contrast, the temporoparietal junction and anterior cingulate cortex support direct preferences by computing a value that includes others' outcomes. In addition, we found that the strength with which the two processes are deployed is negatively correlated across individuals (with normative preferences driving generosity only when direct preferences are relatively weak) and are associated with different levels of generosity.

Altruism, as defined here, involves helping others at a material cost to the self, and is distinct from other types of pro-social behavior supported by strategic considerations[1–3]. Altruism is an apparently universal feature of human societies[4], and understanding the motives and processes that support it has been a major goal in economics, psychology, philosophy, and neuroscience. A large number of models have been proposed for why individuals engage in altruistic behavior, ranging from inequality aversion[5] and social efficiency concerns[6,7] to social signaling[1] or internal guilt[8]. Although different in their precise formulations, these theories can be classified into two broad categories: people behave generously because 1) they have *normative preferences* to comply with their beliefs about the social and ethical norms that dictate appropriate actions[9–12], or 2) they have *direct preferences* over the distributions of outcomes between themselves and others[5,6,13,14]. An individual motivated by normative preferences behaves altruistically because he derives value from the appearance of his *actions*, either to himself or observers, independent of the ultimate effects of those actions. In contrast, an individual with direct preferences (sometimes referred to as distributive preferences) behaves altruistically because he derives value from others' *outcomes* as well as his own. To illustrate the distinction, consider someone confronted by a beggar. Normative preferences might dictate giving money to the beggar when looking the other way seems too morally callous, regardless of the impact of his choice on the beggar. In contrast, direct preferences might dictate giving regardless of appearances, because of value derived directly from improving the beggar's well-being.

Determining the relative role of these two motives in different domains may be important for the design of institutions and policies that promote social welfare and pro-social behavior. Yet this goal has proven difficult to achieve using purely behavioral methods, because both mechanisms may be at work within or across individuals and often produce similar behavior[6].

For example, as predicted by normative preference mechanisms, decreasing the saliency of social norms (by priming anonymity, obscuring the effect of one's choices, or changing implicit experimental demands[9,12,15]) decreases—but does not entirely eradicate—generosity. Yet it is difficult to determine whether the remaining altruism results from direct preferences, or from incomplete crowding out of normative preferences.

To supplement continued efforts with behavioral and other approaches, some researchers have turned to neuroscience to gain better insight into underlying neural mechanisms of altruism. This research shows that regions related to thinking about others, like the temporoparietal junction (TPJ)[16], as well as reward-related regions like the ventral striatum (vStr) and medial prefrontal cortex (mPFC)[17,18], are active during decisions that reflect an altruistic concern for either charities[17–19] or other individuals[20,21]. These results have generally been interpreted as support for the existence of direct preferences in altruistic decision-making. But the value of the choice to help others also correlates with the *normative* value of behaving generously. Unfortunately, none of these studies was designed to test this, which means that these reward signals could be associated with either process.

Other work has examined the neural bases of normative mechanisms during reciprocal interactions involving cooperation and punishment. These studies suggest that the dorsolateral prefrontal cortex (dlPFC) is involved in suppressing self-interested behavior[8,22–24] in this class of social decisions. It is tempting to interpret this as evidence that dlPFC supports the implementation of normative preferences, since there is a well-established link between dlPFC and *behavioral inhibition* (preventing prepotent or automatic processes from controlling actions[25,26]). Under this interpretation, the dlPFC supports behavior consistent with normative preferences through behavioral control that prevents the expression of selfish impulses[22].

However, recent research on self-control in non-social domains suggests an alternative computational role for the dlPFC in these tasks that is also consistent with implementing direct preferences. More concretely, the dlPFC appears to also affect behavior by modulating *value* signals in vmPFC [27,28]. This modulation could be consistent with implementing either normative preferences (if it increases the sensitivity of the vmPFC to normative concerns) or direct preferences (if it increases the sensitivity of the vmPFC to others' outcomes). Since no studies have examined the precise mechanism via which dlPFC promotes altruism, it is unclear whether its activation reflects normative preferences, direct preferences, or both.

Thus, while the existing literature provides important hints as to some of the regions involved in supporting altruistic behavior, it leaves unanswered several critical questions. 1) Are computations in areas such as mPFC, vStr, and dlPFC associated with implementing normative preferences, direct preferences, or both? 2) How much do individuals use these two neural mechanisms, and how do they interact within individuals? 3) What is the association between the two mechanisms and the tendency to make generous choices?

To answer these questions, we used human functional magnetic resonance imaging (fMRI). Several key experimental features allowed us to independently measure value computations related to generous *actions* and generous *outcomes*, and thus to dissociate the neural mechanisms that support normative and direct preferences. To measure value computations related to generous actions, fifty-one participants made 180 real decisions between a proposed pair of payments to themselves and an anonymous partner or a constant default payment-pair of $50 to both subjects (Fig. 1A, see Online Methods for details). The proposed payments to each person in the payment-pair varied from $10 to $100 (Fig. 1B). All payment-pairs included one payment below and one payment above the default, and thus always involved a choice between generous

behavior (benefitting one's partner at a cost to oneself) and selfish behavior (benefitting oneself at a cost to one's partner). Subjects indicated their choice using a 4-point scale (Strong No, No, Yes, Strong Yes), which allowed us to measure both their decision and the value they assigned to the proposed payment at the time of choice. To measure computations related to generous *outcomes*, each trial also included a probabilistic outcome period: in 60% of trials the participant received his chosen option, while in 40% of trials his choice was vetoed and he received the alternative, non-chosen option. This probabilistic outcome appeared at the end of each trial; one trial was picked randomly at the end of scanning to determine payoffs for both participants. The partner, who knew that 40% of choices were vetoed, saw only the final outcome.

Critically, this experimental design allowed us to distinguish neural mechanisms implementing normative preferences from those implementing direct preferences, as well as to measure the extent to which each individual relies on the two motives. Our results suggest that the dlPFC supports normative preferences, both by suppressing the influence of valuation regions like the vmPFC and by directly influencing response selection; and that the temporoparietal junction and anterior cingulate cortex support direct preferences, by integrating the value of others' outcomes into the decision process. Moreover, the deployment of these two mechanisms is negatively correlated across individuals (i.e., normative mechanisms drive generosity only when direct mechanisms are relatively weak), and is associated with different levels of anonymous giving.

I. EXPERIMENT 1 RESULTS

A. INDIVIDUAL DIFFERENCES IN GENEROSITY

On average, subjects made generous choices—maximizing their partner's payoff ($Partner) at a cost to their own ($Self)—in 21% of trials, sacrificing $3.73 from a mean possible payoff of $27.75 per trial in order to give $8.31 to their partner. While this amount is fairly low, there was considerable individual variation: 0%-61% generous choices, $0-$17.08 sacrificed per trial, and $0-$22.37 donated to the partner.

The distribution of money sacrificed (Fig. 1C) suggested that our subjects may have been drawn from two distinct distributions: a large group who acted comparatively selfishly, and a second, smaller group who acted more generously. We carried out a formal test of this possibility (see Online Methods for details) by using Markov-chain Monte Carlo methods to estimate the mean and standard deviation of the two separate distributions as well as the probability that subjects were drawn from each distribution. We then compared the likelihood of this model to one in which generosity levels were drawn from a single population. A likelihood ratio test indicated that the mixture model fit the data considerably better than the single distribution model (P = .001). Based on the Bayesian posterior probability that each participant belonged to the more selfish group (mean generosity = $0.95, s.d. = $0.71) or the more generous group (mean generosity = $8.53, s.d. = $4.82), we classified 34 participants into the Selfish group (< $2.50 sacrificed) and 17 participants into the Generous group. We used this behavioral classification in several of the analyses reported below.

B. DIFFERENCES IN GENEROSITY CORRELATE WITH DIFFERENCES IN DECISION TIME.

Previous studies of reciprocal interactions suggest that overriding self-interest recruits cognitive control regions[8,22,23]. We analyzed reaction times (RTs) to look for behavioral evidence of similar patterns in our experiment. Across the entire group (excluding 7 participants who made fewer than four generous choices), generous choices took significantly longer than selfish choices (RT_G = 2,131 ms, s.d. = 280 ms, RT_S = 2,300 ms, s.d. = 310 ms, paired $t_{43} = 4.97$, $P < 0.0001$). However, we found considerable differences between the two behaviorally defined groups (Fig. 1D). Although the groups did not differ in overall RT ($t_{49} = .36$, $P = 0.72$), Selfish participants took considerably longer to make generous compared to selfish choices (RT_G = 2,358 ms, s.d. = 313 ms, RT_S = 2,110 ms, s.d. = 242 ms, paired $t_{26} = 6.29$, $P < 0.0001$), while Generous participants showed no evidence of this effect (RT_G = 2,207 ms, s.d. = 291 ms, RT_S = 2,164 ms, s.d. = 338 ms, paired $t_{16} = .87$, $P = 0.39$; two-sample $t_{42} = 3.25$, $P = 0.002$). The pattern of reaction times shows that computations associated with more altruism take longer, but they are silent about the extent to which direct or normative processes are at work. For this question we turn to the neural data.

C. DLPFC SUPPORTS NORMATIVE PREFERENCES

One possibility (supported by our analysis) is that normative preferences promote behavior that is consistent with a shared social or ethical belief (e.g., give to another whenever the cost is sufficiently low). A neural mechanism supporting this preference could affect behavior either by

changing the values assigned to the options in areas such as vmPFC, or by directly influencing the action selection process. Critically, to the extent that the value of complying with the shared belief is reflected in value signals in vmPFC, this should be the case at the time of decision, but need not be so when the probabilistic outcome is announced. For an individual driven solely by normative preferences, the ideal situation is one in which he behaves altruistically (thus reaping the benefits from complying with the shared belief), but his choice is reversed exogenously (so that he ends up not having to give up his own money).

Based on this, the following three markers should characterize a region involved in implementing normative preferences: 1) it should be more active when a person behaves altruistically (since the more active the mechanism, the more likely it is to influence behavior); 2) its connectivity profile should be consistent with either the modulation of value signals at the time of choice in areas like vmPFC, or with the inhibition of selfish impulses by directly influencing the action selection process; and 3) reliance on these processes implies that an exogenous, random reversal of those generous choices should be experienced as subjectively rewarding, because the person reaps the reward related to acting in compliance with social norms and the reward associated with the outcome he would otherwise prefer in the absence of social norm considerations. Based on previous work in self-control [26,29] and altruistic choice[8,22], we hypothesized that the dlPFC might satisfy these criteria.

To test the first hypothesis, we looked for regions displaying greater activity during generous compared to selfish choices during the decision period (see Online Methods, GLM 1). As expected, generous choices were characterized by greater activation in the right dlPFC (Fig. 2A, Supplementary Table 1, $P < 0.001$ uncorrected, $P = .04$ small-volume corrected, SVC), as well as a region of the dorsal anterior cingulate cortex (dACC) previously associated with conflict

monitoring and cognitive control[30,31] ($P < 0.05$, whole-brain corrected). ROI analysis further indicated that increased activity in the right dlPFC on generous compared to selfish choice trials correlated positively with longer RTs to choose generously (robust reg. coef. = .27, p = .01), supporting a relation between this region and use of cognitive control. Activity in the dACC did not correlate with RT differences (robust reg. coef. = .21, p = .17).

To test the second hypothesis—that normative preferences imply inhibition of selfish impulses and the selection of norm-consistent behavioral responses—we estimated a psychophysiological interaction (PPI) model to compare the connectivity profile of the dlPFC during generous and selfish choices. We focused on this region rather than the dACC both because of the considerable literature linking it to self-control[27] and behavioral inhibition[25,26], and because only the dlPFC correlated with differences in RT. Consistent with the characteristics of a normative mechanism for generous choice, we found that that during generous compared to selfish choice the dlPFC exhibited stronger negative connectivity with an area of the vmPFC associated with the computation of values at the time of choice[27,32] as well as stronger positive connectivity with areas of supplementary motor (SMA) and inferior parietal (IPL) cortex previously associated with response selection[33–35] (Fig. 2B, Supplementary Table 2, p < .05, corrected).

Finally, to test the hypothesis that implementation of normative mechanisms via behavioral suppression results in a distinction between preferences over actions and preferences over outcomes, we exploited a unique feature of our experimental design: the 40% of trials on which a subject's choice is vetoed. We hypothesized that reward-related response to veto vs. receipt of generous choices in an independently defined area of vmPFC shown to track the subjective pleasantness of consuming a stimulus[36] should correlate with activation in the dlPFC, if the

dlPFC implements normative mechanisms during choice. Consistent with this hypothesis, response in the vmPFC when generous choices were vetoed (resulting in the non-chosen option) was *higher* than when generous choices were received, to the extent that participants activated the dlPFC during generous compared to selfish choices (Fig. 2C, robust reg. coef. = .44, $p < .001$).

Taken together, these three results are consistent with the hypothesis that the dlPFC implements normative mechanisms for norm-consistent behavior, both by suppressing the influence of value-related regions like the vmPFC, and by directly modulating behavioral response selection.

D. TEMPOROPARIETAL JUNCTION AND ROSTRAL ANTERIOR CINGULATE CORTEX SUPPORT DIRECT PREFERENCES.

We used a similar logic and set of steps to identify regions associated with direct preferences. Direct preferences imply that a person represents not only his own welfare ($Self) but also his partner's ($Partner). Importantly, these computations apply not just during choice, but also at outcome. An individual driven solely by direct preferences cares about others' *outcomes*, and not about his actions per se. This individual would be indifferent between having made the decision to act generously and having that decision determined for him, as long as the distribution of outcomes remains the same.

Based on this, the following three markers should characterize a region involved in implementing direct preferences: 1) its activity should correlate with the value of $Partner on each trial; 2) its activity should correlate with stated preferences for different options; 3) its

connectivity profile should be consistent with feeding of information about $Partner to regions involved in valuation and/or response selection; 4) use of this neural mechanism to choose generously implies a genuine preference over outcomes, making exogenously imposed, random reversal of those generous choices (resulting in unchosen outcomes) subjectively unpleasant. Based on research in the social cognition[16] and value-based decision making[18,32,34,37] literature, we hypothesized that regions related to social cognition, such as the TPJ, and/or regions related to valuation at the time of choice, such as the vmPFC or ACC, might satisfy these minimally necessary criteria.

To test the first hypothesis, we looked for brain regions whose activity was consistent with representing the amount $Partner at the time of choice. This analysis identified several regions in which responses during the choice period correlated with $Partner (Fig. 3A, Supplementary Table 3), including right temporoparietal junction (TPJ), left TPJ/angular gyrus, and precuneus ($p < .05$, corrected) as well as the rostral ACC (rACC; $p < .001$ uncorrected, $p = .004$ SVC). Several regions also represented $Self (Fig. 3B, Supplementary Table 4), including ventral striatum and ACC ($p < .05$, corrected). A direct contrast of response to $Partner and $Self confirmed that activity in both right ($p < .05$, corrected) and left ($p < .001$, uncorrected) TPJ was specific to $Partner while other regions, including both the amygdala ($p < .05$, corrected) and the ventral striatum, marginally ($p < .08$, SVC), responded uniquely to $Self (Supplementary Table 5) but not to $Partner. In contrast, rACC responded equally to both, consistent with its hypothesized role in computating integrated value signals[27]. Importantly, the representation of $Partner did not differ as a function of whether participants chose generously or selfishly on a given trial for either the left TPJ or the rACC (all P = n.s.), suggesting that these regions

represented $Partner to a similar degree on all choices regardless of whether that representation resulted in generosity or not, and thus could serve as precursors to choice.

To test the second hypothesis that areas representing $Partner also contribute to valuation during decision-making we looked for overlap between the neural correlates of stated preferences (given by the participant's ratings of the proposal), and regions representing $Partner. Note that these two analyses are independent of each other: identifying regions correlated with stated preferences does not assume any particular relationship to $Self or $Partner, although it does not preclude such a relationship. As expected, we identified several regions where subjective preferences (Fig. 3C) overlapped with the representation of $Partner, including the left TPJ/angular gyrus and rACC (Fig. 3D, Supplementary Table 6,7).

To test the third hypothesis, we examined the connectivity profile of the left TPJ region correlating with both $Partner and stated preferences. Although the TPJ has not often been implicated in decision-value computation[37], we hypothesized that it may influence behavior during social decision-making by providing inputs to the rostral ACC region associated with valuation and/or to response selection regions. Consistent with this hypothesis, a psychophysiological interaction (PPI) analysis (see SOM for details) indicated that, during choice, functional connectivity increased between left TPJ and the area of ACC associated with $P as well as with areas of IPL and SMA involved in response selection[33-35] (Fig. 4A, $P < 0.05$ corrected, Supplementary Table 8), suggesting that social value computations supported by the TPJ may influence choices not only indirectly through modulation of value computations in the rACC, but also directly by modulating the selection of behavioral responses.

Finally, we tested the prediction that the use of direct mechanisms implies similar preferences at both decision and outcome. In contrast to normative mechanisms, where randomly

vetoing generous choices results is experienced as rewarding, direct preferences imply that veto of generous choices should be associated with less reward. To test this, we correlated representation of $Partner in the TPJ and rACC (where representation and valuation of $Partner overlapped) with response to generous choice veto vs. receipt at outcome in the region of vmPFC associated with subjective utility. Consistent with the hypothesis that direct preferences imply *consistency* between preferences at choice and outcome, response in the vmPFC when generous choices were vetoed was lower to the extent that rACC demonstrated sensitivity to $Partner (Fig. 4B, robust reg. coef. = -.32, $P = 0.009$) and non-significantly for the lTPJ (robust reg. coef. = -.2, $P = 0.13$).

Taken together, these results are consistent with the hypothesis that direct preferences are implemented through the representation and valuation of $Partner in lTPJ and rACC.

E. DEPLOYMENT OF NORMATIVE AND DIRECT MECHANISMS IS NEGATIVELY CORRELATED ACROSS PARTICIPANTS.

The previous results establish the existence of two computationally and neurobiologically dissociable mechanisms that promote altruistic behavior. This leads to a natural question: Is the extent to which these mechanisms are deployed independent, positively correlated, or negatively correlated? The answer is not obvious a priori, and it is hard to speculate based on previous data. For example, given their distinct neural bases, the two mechanisms might be deployed independently. This would imply that more generous individuals would be more likely to engage in both types of processes. Perhaps more interestingly, subjects might deploy the normative preference mechanism only when their direct preferences are not sufficiently strong to lead them

to comply with the shared social norms, in which case they would be negatively correlated in the population.

Our design allows us to address this question, by using dlPFC responses as a measure of the extent to which the normative mechanism is deployed, and rACC and lTPJ responses as a measure of the extent to which the direct mechanism is deployed. We found a negative correlation across participants between recruitment of dlPFC on generous choices and representation of $Partner in both the lTPJ (robust reg. coef. = -0.33, $p < .001$) and rACC (robust reg. coef. = -.4, $p < .005$). This observation suggests that, to the extent that a person uses direct mechanisms (which may more naturally lead to generous choices), he tends not to use normative self-control mechanisms, perhaps because using self-control is effortful and thus deployed only when necessary.

F. NORMATIVE AND DIRECT MECHANISMS SHOW DIFFERENT PATTERNS OF CORRELATION WITH LEVELS OF GENEROSITY ACROSS INDIVIDUALS.

Finally, we examined in the data the extent to which individual generosity was correlated with the relative reliance on the two processes. Note that the answer to this question is also not obvious a priori. On the one hand, since individuals activate the normative mechanism only when their direct preferences are weak, one might expect a negative correlation between generosity and reliance on normative preferences. On the other hand, there are no a priori constraints in our experiment on the strength of the shared social belief, and thus the activation of normative motives could be enough to compensate for weak direct preferences.

We found that stronger dlPFC activation was associated with *less* generosity overall (robust reg. coef. = -.61, $p < .001$). Indeed, examining Selfish and Generous participants separately, we observed that the Selfish group showed a robust increase in dlPFC, while Generous participants showed none (Fig. 5A). In contrast, we observed a positive correlation between behavioral generosity and neural representation of $Partner in the left TPJ (robust reg. coef. = .6, $p < .001$), although this relation failed to reach significance in the rACC (robust reg. coef. = .14, $p = .19$). Indeed, whereas the TPJ showed a strong and specific sensitivity to $Partner in the Generous group, it showed little differential response in the Selfish group (Fig. 5B). Because these two mechanisms were negatively correlated with each other, we also used multiple regression to test the degree to which each mechanism correlated with generosity when controlling for the other. This analysis indicated that lTPJ response to $Partner independently correlated with generosity ($b = 53.46$, s.e. = 17.39, $p = .004$), while dlPFC response marginally negatively correlated with generosity ($b = -1.15$, s.e. = .59, $p = .06$).

II. DISCUSSION

This study examined the computational and neurobiological basis of simple, anonymous altruistic decision-making. The study allowed us to address the following three open questions: 1) Are computations in areas such as mPFC, vStr, and dlPFC associated with implementing normative preferences, direct preferences, or both? 2) How much do individuals use these two neural mechanisms, and how do they interact within individuals? 3) What is the association

between the two mechanisms and the tendency to make generous choices? Our results advance our understanding of each of these questions.

First, our results support the existence of dissociable neurobiological mechanisms for each type of preference, and provide insights about the computations carried out by areas like vmPFC, vSt, dlPFC, and TPJ in altruistic choice. Activity in a region of dlPFC associated with cognitive control[27,29] was consistent with the deployment of normative preferences: 1) when participants chose generously, this region showed greater activity and correlated with reaction times; 2) connectivity analysis suggested that it down-modulated a region of vmPFC involved in valuation and up-modulated motor control areas, and 3) activity in this region during decision-making predicted positive responses at outcome to generous choice veto in a region of vmPFC associated with subjective utility. We interpret this as evidence that the dlPFC mechanism enables pro-social choices even when participants did not care directly about another's well-being (or did not care *enough*). By contrast, responses in the TPJ and rACC, regions associated with social cognition and valuation[16,32] were consistent with the deployment of direct preferences: 1) these regions displayed activity that correlated both with the magnitude of the partner's payoff and with participants' stated preferences; 2) these regions displayed functional connectivity with each other and with regions involved in motor planning at choice; and 3) weighting of $Partner in rACC during decision-making predicted negative responses at outcome to generous choice veto in vmPFC.

Second, our results suggest that different individuals deploy these two processes to varying degrees. Behaviorally, levels of generosity suggested the existence of two groups: one relatively selfish group and a second considerably more generous group. These behavioral distinctions were associated with qualitatively different patterns of neural activation. The majority of

participants fell into the first group and for the range of choices we studied displayed evidence of relying when giving on normative preferences implemented by the right dlPFC. A smaller group displayed evidence of a reliance on direct preferences implemented in left TPJ and rACC. Importantly, the more participants represented their partner's outcomes—both behaviorally and neurally—the less they recruited the dlPFC mechanism in order to make a generous choice. This suggests that while both motives may be active in some individuals, they tend to be deployed in complementary fashion.

Third, we found that when our subjects relied on the dlPFC-based mechanism for normative preferences they tended to give less, while when they used the TPJ-based mechanism for direct preferences they tended to give more. This finding provides the seed for further investigations of the relative role of the two mechanisms across settings and across individuals. The particular correlations we observe in our data might reflect more on our subjects' shared view of the social norms in this particular context than about underlying differences in the effectiveness of the two mechanisms. Indeed, we suspect that in situations where social norms dictate higher levels of giving (either in other cultures or other contexts), and in contexts where social norms have more bite (e.g., non-anonymous giving), use of normative mechanisms may be associated with *greater* generosity. Future work will be needed to determine whether such contexts induce greater giving in the subset of people relying on self-control and normative motives to give.

Our results have potential implications for several domains. First, social neuroscience studies have found that the TPJ, particularly in the right hemisphere, plays a critical role in the ability to characterize the mental states of others[16,38] (e.g., by taking their perspective), and our results corroborate other accounts that implicate the TPJ in charitable giving[18,39]. However, our results differ in important ways from prior work. While both right and left TPJ correlated with $Partner

in our study, the process of translating this into a reward-related value seemed to be more strongly related to a region of the left TPJ extending posteriorly into the angular gyrus. This suggests that there may be some important structural distinctions between simply representing and actually valuing another's welfare, and future work will need to address the key computations required to internalize the well-being of another person when computing the values of social decisions. In particular, it will be important to understand whether differences in TPJ-related generosity have purely social-cognitive roots, or whether they arise out of more basic, non-social functions, such as attentional orienting[40]. These questions have practical implications for understanding diseases like autism, which are characterized by structural differences in regions like the TPJ and ACC[41] as well as differences in social behavior, including charitable giving[42].

Our work also has implications for understanding the role that basic individual traits like the ability to exercise self-control have in altruistic choice. In our study, the deployment of self-control via the dlPFC was associated with an apparent disconnect between what people chose to do and what they wanted to occur. Studies of non-social decision-making have found that a similar area of dlPFC to the one identified here plays a critical role in the ability to choose options that promote long-term goals[27,43] and may induce a similar disconnect between stated preferences and underlying motivation in non-social contexts as well[44]. In these contexts, difficulty making adaptive, far-sighted choices is often seen as a *failure* of self-control. By analogy, selfish choices should not automatically be seen as evidence that people do not care about others, or do not care about social norms. Rather, such behavior may simply reflect an inability to properly exert self-control. This problem may be particularly acute for children, because dlPFC areas mediating the ability to control selfish impulses develop only slowly[24].

Educational programs that improve both direct mechanisms (e.g., perspective-taking) and normative mechanisms (e.g., self-control) may thus yield more consistent pro-social behavior.

Finally, the use of neuroeconomic insights to make inferences about motives for giving at the individual level may be a valuable tool with diverse applications. In scientific contexts, it can be used to contribute insight into long-standing scientific debates about the prevalence of different motives in different types of altruistic behavior. Much more speculatively, it may influence the design of institutions concerned with promoting public welfare and pro-social behavior. For example, our results suggest that the answer to economic policy debates on the relative benefits of tax-provided public goods vs. voluntary giving to charity may depend largely on the individual who is taxed. Mandatory taxation reduces the welfare of those motivated primarily by normative preferences (perhaps the majority of individuals) by denying them the "warm glow" of acting charitably[45], while perhaps increasing the welfare of those motivated by direct preferences by improving the outcomes of others. The ability to identify these motives within individuals may help to determine the aggregate (group-level) benefit resulting from different interventions.

Figure 1. Design and behavioral results. A) Trial structure. B) Proposed transfers used in the experiment describing $ Self and $ Partner. The alternative was always a transfer of $50 to both subjects. X- and Y-axes represent distance from default offer. The filled area in each transfer is proportional to the percentage of pro-social choices across all participants. C) Distribution of generosity across participants, given by the cumulative amount of money he sacrificed over the course of the experiment. A model in which observations came from two distributions best accounts for the observed levels of generosity (maximum likelihood fit; Selfish group—green, Generous group—blue). D) Differences in average reaction time for Selfish (dark bars) and Generous (light bars) choices, separately for Selfish and Generous groups. Error bars indicate standard error of the mean. ** $P < .01$

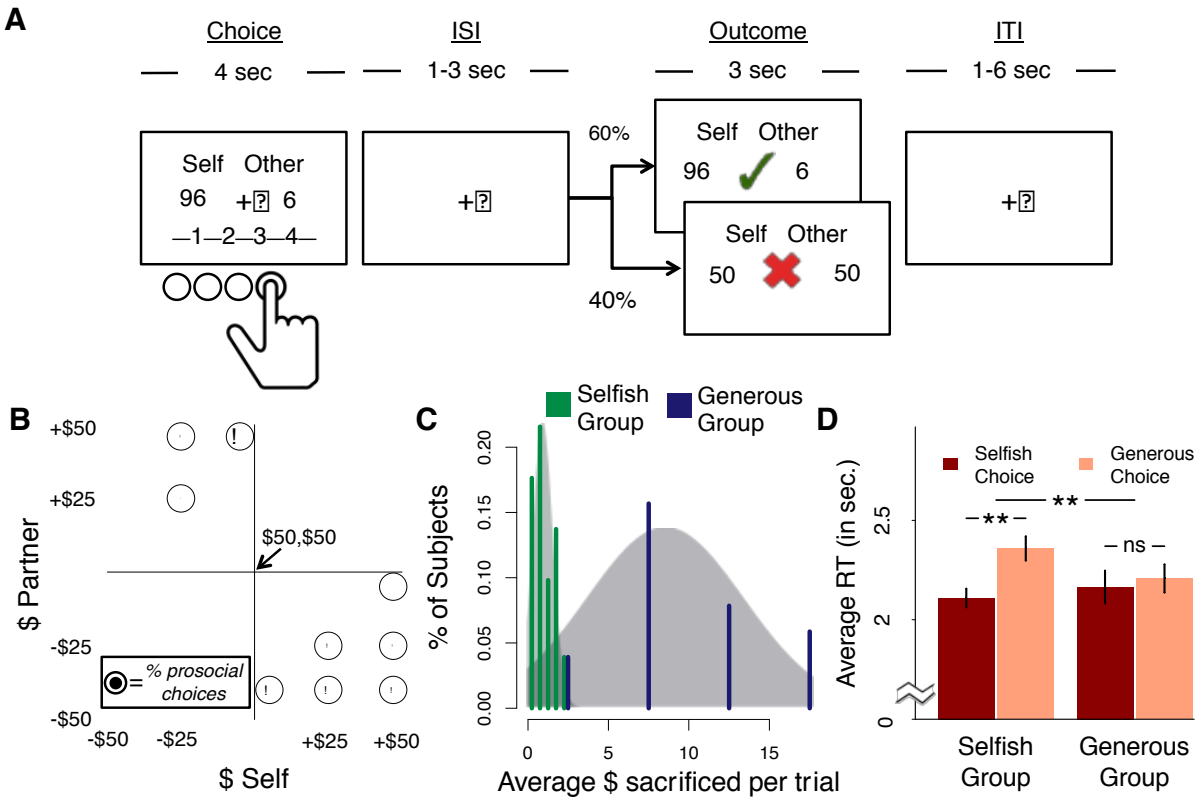


Figure 2. Right dlPFC and normative mechanisms of generous choice. A) Right dlPFC responded more strongly on generous vs. selfish choice trials. B) vmPFC showed decreased connectivity with the dlPFC on generous choice trials, while SMA and inferior parietal cortex showed increased connectivity. C) Right dlPFC increases during generous choice negatively predict vmPFC response to vetoed vs. received generous choices at outcome. To make the robust regression coefficient equivalent to a standard correlation statistic, both variables were normalized prior to regression.

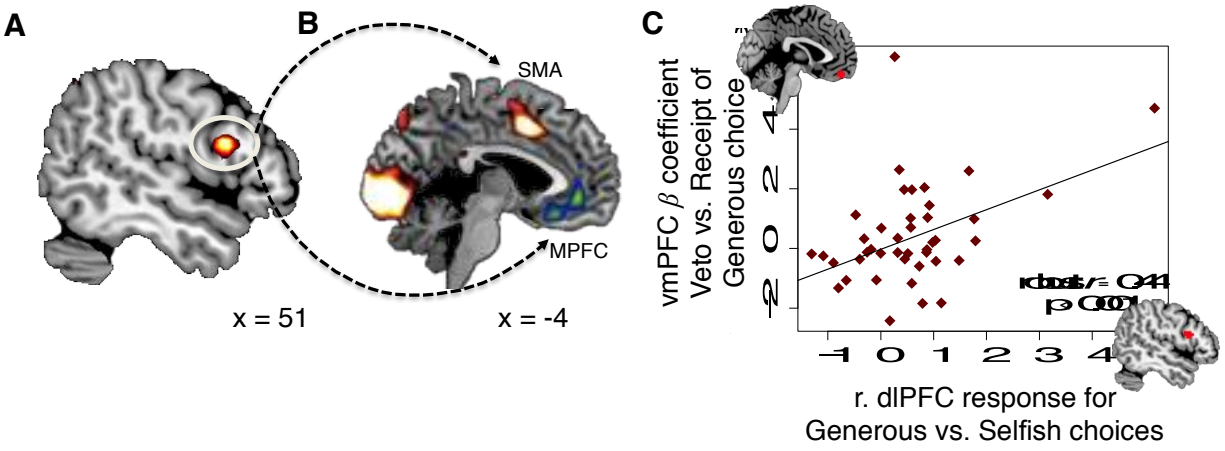


Figure 3. Neural responsiveness to $S, $P, and subjective preference. A) Regions in which activity increased with $Partner; B) Regions in which activity increased with $Self. C) Regions demonstrating sensitivity to an independent estimate of stated preference for the proposal. D) Overlap. Red = $Partner, Blue = $Self, Green = Stated Preference. Images thresholded at $P <$ .0005, uncorrected. Red circles - left and right TPJ; black circles - ventral striatum; green circles - mPFC/rACC.

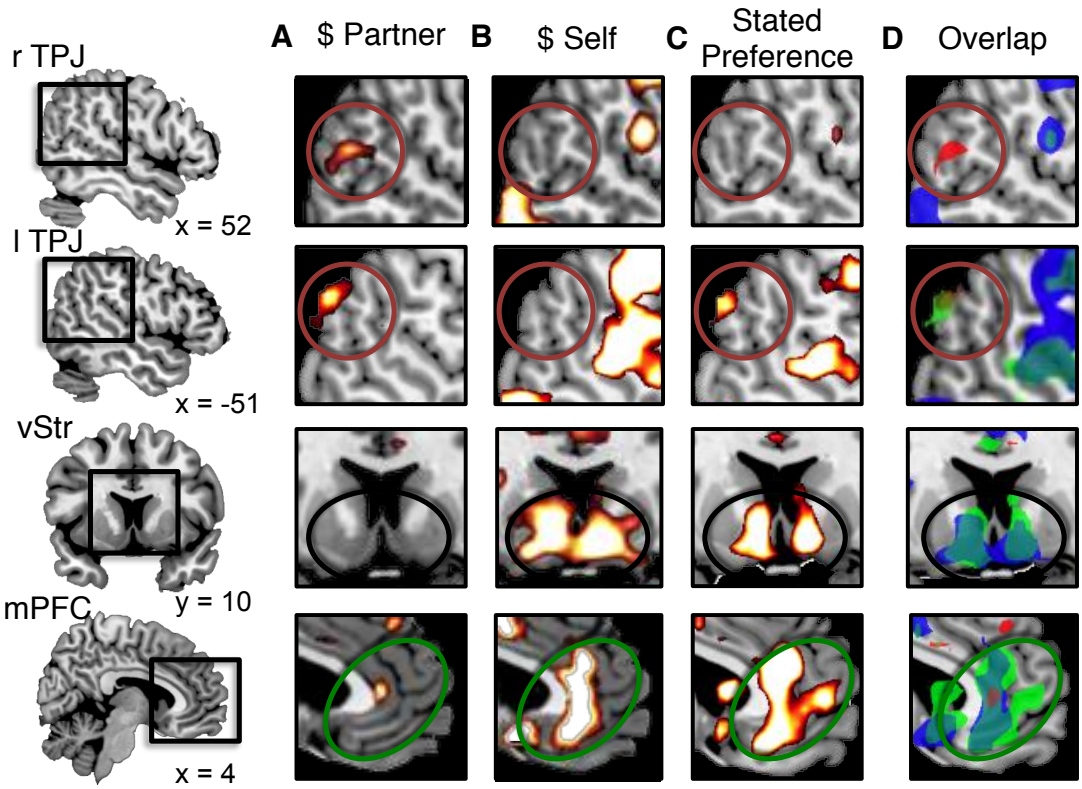


Figure 4. Left TPJ, rACC and direct mechanisms of generous choice. A) Left TPJ demonstrated increased functional connectivity with the rACC, SMA and inferior parietal cortex during the choice period. B) rACC relative weighting of $Partner positively predicts vmPFC response to vetoed vs. received generous choices at outcome. To make the robust regression coefficient equivalent to a standard correlation statistic, both variables were normalized prior to regression.

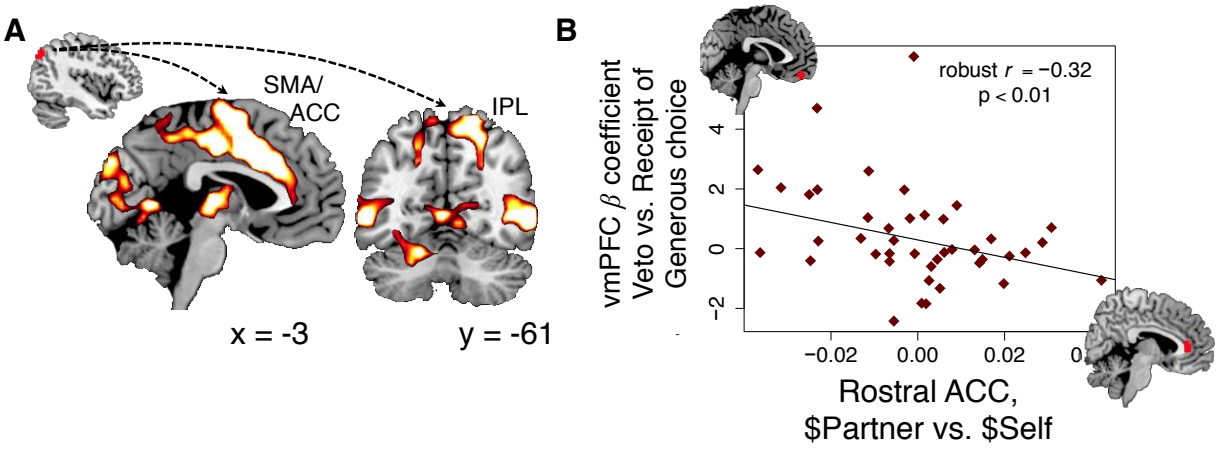
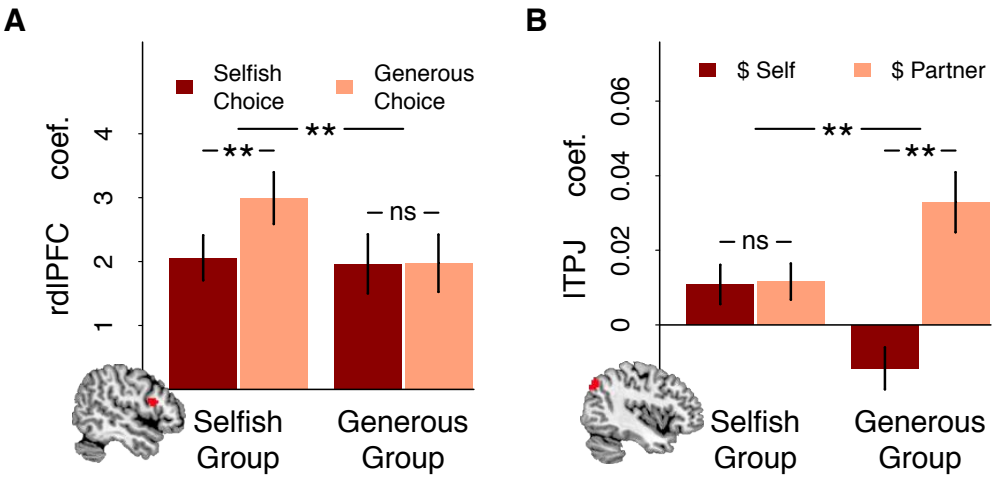


Figure 5. Extracted betas from ROIs. A) dlPFC was more active on generous (light bars) vs. selfish choice trials (dark bars) only in the Selfish group of subjects. B) lTPJ represents $Partner (light bars) rather than $Self (dark bars) only in the Generous group of subjects. Error bars indicate standard error of the mean. ** $P < .01$



Supplementary Table 1. Regions displaying differential activity on generous vs. selfish choice trials at the time of choice (GLM 1)

Region		BA	Cluster Size	Z score	x	y	z
Unselfish > Selfish							
L	Anterior cingulate cortex	32	86	4.03	-3	33	36
R	DLPFC	45/46	24	4.26*	54	12	21
Unselfish < Selfish							
R	Occipital Cortex	18	577	4.81	27	-96	12

Note:
Regions are reported if they passed two thresholds: $p < .001$ uncorrected and $p < .05$ cluster corrected, unless otherwise noted.
* $p = .04$, SVC within a bilateral anatomical mask of lateral PFC (see Online Methods for details).

Supplementary Table 2. Regions exhibiting differential connectivity with dlPFC on generous vs. selfish trials at the time of choice (PPI 1)

	Region	BA	Cluster Size	Z score	x	y	z
Generous > Selfish							
R	Occipital cortex	17/18/19	7365	6.52	24	-78	-12
L	Occipital cortex	17/18/19	a	6.51	-24	-96	9
L	Inferior parietal lobule	40	a	6.3	-33	-51	48
R	Inferior parietal lobule	40	a	6.1	27	-63	39
L	Supplementary motor cortex		349	5.75	-6	9	42
L	Thalamus		264	5.28	-12	-15	6
R	Anterior insula	13	126	5.04	36	21	3
R	Middle frontal gyrus	6/9	229	4.93	39	3	36
R	Middle frontal gyrus	6	48	3.94	24	-6	51
R	Posterior cingulate gyrus	23	52	3.83	6	-30	30
R	Thalamus		53	3.75	12	0	0
L	Dorsolateral prefrontal cortex	10/46	57	3.74	-33	42	18
Generous < Selfish							
L	Angular gyrus	40	287	5.85	-36	-81	42
L	Middle frontal gyrus	8	49	4.74	-21	27	57
L	Anterior cingulate/ventromedial prefrontal cortex	10/32	165	4.49	-6	42	-6
R	Angular gyrus	40	131	4.4	45	-75	42
L	Superior frontal gyrus	9/10	36	3.99	-15	51	27

Note:
Regions are reported if they passed two thresholds: $p < .001$ uncorrected and $p < .05$ cluster corrected, unless otherwise noted.
a. Distinct peak in larger cluster of activation, reported separately for completeness.

Supplementary Table 3. Regions correlating with $Partner (GLM 2)

Region		BA	Cluster Size	Z score	x	y	z
R	Anterior cingulate cortex	24	45	3.82*	9	36	3
L	Inferior parietal lobule/ temporoparietal junction	7/39	217	4.71	-24	-48	24
L	Precuneus	7/31	494	4.89	-9	-60	45
R	Temporoparietal junction	39	63	3.92	39	-63	21
L	Occipital cortex	30	61	4.51	-30	-63	12
R	Occipital cortex	18	76	4.22	21	-90	-9
L	Cerebellum		152	4.24	-24	-93	-27
	Occipital cortex	18	a	4.17	-18	-96	-12

Note:
Regions are reported if they passed two thresholds: $p < .001$ uncorrected and $p < .05$ cluster corrected, unless otherwise noted.
* $p = .003$, small-volume corrected within an anatomically defined mask of the mPFC (see Online Methods for details).
a. Distinct peak in larger cluster of activation, reported separately for completeness.

Supplementary Table 4. Regions correlating with $Self (GLM 2)

Region		BA	Cluster Size	Z score	x	y	z
L	Anterior cingulate/	24/32	180	4.8	-3	39	6
	Ventromedial prefrontal cortex	11/32	a	4.49	6	33	-12
R	Inferior frontal gyrus	44	24	4.62	48	3	15
L	Precentral gyrus	6	19	4.59	-48	0	45
R	Supplementary motor area	6	441	5.62	3	-9	54
R	Precentral gyrus	3/4	68	4.15	42	-18	57
R	Postcentral gyrus	3	55	5.06	57	-21	30
L	Postcentral gyrus	4	1234	6.43	-33	-27	51
	Posterior cingulate cortex	23/31	a	5.12	-3	-36	39
	Occipital cortex (L)	18/19	a	6.29	-36	-81	-12
R	Occipital cortex	18/19	4882	6.36	21	-99	-3
R	Ventral striatum		84	5.33	9	12	-9
L	Ventral striatum		52	5.1	-9	12	-3

Note:
Regions are reported if they passed two thresholds: $p < .0001$ uncorrected and $p < .05$ cluster corrected, unless otherwise noted. This more stringent threshold was used to better separate clusters of activation.
a. Distinct peak in larger cluster of activation, reported separately for completeness.

Supplementary Table 5. Differences in correlation with $Partner and $Self (GLM 2)

	Region	BA	Cluster Size	Z score	x	y	z
Correlation with Other Amount > Self Amount							
L	Middle frontal gyrus	9	24	3.75	-39	15	36
R	Temporoparietal junction	39/40	29	3.93	54	-54	27
L	Angular gyrus/temporoparietal junction	39	10	3.29*	-48	-69	39
Correlation with Self Amount > Other Amount							
L	Postcentral gyrus	3/4	240	4.17	-36	-27	72
R	Fusiform gyrus	19	17	3.52	27	-66	-12
L	Inferior occipital cortex	19	45	3.94	-45	-72	-6
R	Inferior occipital cortex	19	30	3.69	36	-72	-12
R	Occipital cortex	18/19	347	4.82	30	-75	21
R	Lingual gyrus	18	20	3.62	12	-78	-12
L	Middle occipital gyrus	18/19	332	4.15	-27	-96	9
R	Cerebellum		74	4.61	21	-60	-27
L	Ventral striatum		16	3.27†	-9	9	-9
R	Amygdala		21	3.94	30	0	-18

Note:
Regions are reported if they passed two thresholds: $p < .001$ uncorrected and $p < .05$ cluster corrected, unless otherwise noted.
* $p < .001$, uncorrected; $p = .07$, small-volume corrected within an anatomically defined mask of the bilateral temporoparietal junction
† $p < .005$, uncorrected; $p = .08$ small-volume corrected within an anatomically defined mask of the ventral striatum (see Online Methods for details)

Supplementary Table 6. Regions correlating with stated preferences at the time of choice (GLM 3).

Region	BA	Cluster Size	Z score	x	y	z
L Superior frontal gyrus	10	75	4.9	-12	60	27
B Anterior cingulate cortex	24/32	680	5.37	3	39	18
L Ventromedial prefrontal cortex	11/32	a	5.01	-6	33	-12
R Ventral striatum		a	5.29	9	12	-6
L Ventral striatum		a	5.13	-9	12	-6
L Middle frontal gyrus	6/8	49	4.36	-21	24	54
R Precentral gyrus	6	32	4.33	63	3	24
L Mid-cingulate cortex	24	53	4.27	-3	-6	39
R Supplementary Motor Area	6	16	4.28	6	-12	72
L Precentral gyrus	4	90	4.86	-39	-15	57
L Postcentral gyrus	4	216	5.06	-21	-27	72
R Superior temporal gyrus	21/22	38	4.63	60	-30	6
L Superior temporal gyrus	22/41	179	4.91	-63	-36	9
L Posterior cingulate cortex	31	186	5.64	-6	-42	42
R Inferior temporal gyrus	37	35	5.42	54	-42	-21
L Inferior parietal cortex	7	72	4.48	-36	-75	42
B Occipital cortex	18/19	3430	6.23	-6	-102	0
Occipital cortex	18/19	a	5.3	18	-96	15

Note:
Regions are reported if they passed two thresholds: $p < .0001$ uncorrected and $p < .05$ cluster corrected, unless otherwise noted.
a. Distinct peak in larger cluster of activation, reported separately for completeness.

Supplementary Table 7. Regions exhibiting overlap between $Partner or $Self and stated preferences (GLM 2 & GLM 3).

Region		BA	Cluster Size	x	y	z
Overlap between $Partner and Stated Preference						
L	Precuneus	7	90	-9	-54	42
R	Occipital cortex	17/18	62	21	-90	-9
L	Occipital cortex	17/18	43	-18	-96	-12
L	Angular gyrus/temporoparietal junction	39	31	-48	-69	39
L	Cerebellum		31	-24	-93	-27
R	Precuneus	7	28	3	-78	33
R	Anterior cingulate cortex	24/32	21	9	36	3
Overlap between $Self and Stated Preference						
R	Occipital cortex	17/18	4559	21	-99	-3
L	Occipital cortex	17/18	a	-36	-81	-12
L	Postcentral gyrus	4	604	-33	-27	51
L	Anterior cingulate cortex	32	350	-3	39	6
L	Mid-cingulate cortex	31	204	-3	-36	39
R	Supplementary Motor Area	6	192	3	-9	54
R	Ventral Striatum		175	9	12	-9
L	Superior temporal gyrus	41	175	-51	-33	18
R	Inferior frontal gyrus	6/44	36	48	3	15
R	Superior parietal lobule	7	31	30	-57	54
L	Insula	13	29	-45	-9	21
R	Supramarginal gyrus	41	24	57	-21	30

Note:
Regions are reported if they were jointly significant for both contrasts at $p < .0005$, uncorrected, with at least 20 voxel overlap.
a. Distinct peak in larger cluster of activation, reported separately for completeness.

Supplementary Table 8. Regions exhibiting functional connectivity with left TPJ at choice (PPI 2).

Region		BA	Cluster Size	Z score	x	y	z
L	Supplementary motor area/anterior cingulate cortex	8/32	3649	6.09	-3	9	39
R	Middle frontal gyrus	4	a	4.71	45	3	39
L	Inferior parietal cortex	7	a	4.9	-24	-48	51
R	Inferior parietal cortex	7	a	4.91	24	-54	54
R	Precuneus	7	a	7.12	9	-48	54
R	Anterior cingulate cortex	32	5	3.91*	0	36	12
R	Middle frontal gyrus	9	178	5.89	33	39	39
L	Insula	13	317	5.8	-36	-9	-6
L	Supramarginal gyrus	40	312	5.66	-60	-33	24
R	Posterior middle temporal gyrus	37	221	5.07	54	-57	0
L	Middle frontal gyrus	9	59	4.93	-30	48	39
L	Thalamus		189	4.92	-12	-9	15
L	Postcentral gyrus	4	25	4.7	-42	-15	51
L	Amygdala	34	40	4.68	-21	3	-18
R	Thalamus		32	4.38	18	-15	15
R	Occipital cortex	19	31	4.36	33	-90	21
L	Cerebellum		21	4.19	-36	-51	-33
L	Cerebellum		16	4.19	-18	-63	-24
R	Subcallosal gyrus	34	16	4.1	18	9	-15

Note:
Regions are reported if they passed two thresholds: $p < .0001$ uncorrected and $p < .05$ cluster corrected, unless otherwise noted.
a. Distinct peak in larger cluster of activation, reported separately for completeness.
* $p < .05$, SVC within a mask of the ACC region associated with $P at $p < .001$, uncorrected, reported for completeness.

Appendix 1.

METHODS

Subjects. Male volunteers ($N = 122$) were recruited in pairs from the Caltech community. Subjects were right-handed, healthy, with normal or corrected-to-normal vision, no history of psychiatric or neurological conditions, and free of medications that might interfere with fMRI. Data from ten pairs was excluded from the analyses due to excessive head motion (more than 3mm total translation or 3 degrees total rotation in any single volume) or technical difficulties during the scanning session (mean age 22.3, range 18-35, for the remaining subjects). All subjects received a show-up fee of $30, as well as $0-$100 more depending on the outcome of a randomly chosen experimental trial. All procedures were approved by the internal review board of the California Institute of Technology, and subjects provided informed consent prior to their participation in the experiment.

Task. Each subject in a pair arrived separately to the lab, and was escorted to separate waiting areas where he received task instructions. In each pair, one subject was randomly designated as the (active) participant, and completed the tasks described below. The other was designated as the (passive) partner, and after receiving instructions, waited in a separate room for the duration of the study. Although participants never met, and the partner made no decisions, his presence was important to provide a real social context for the participants.

The active participant read the rules of the decision task, and completed three practice trials. Then, on each of 180 trials in the scanner, he saw a proposed transfer consisting of an amount of money for himself ($Self) and an amount for his partner ($Partner). The side of the screen on

which $Self appeared was counterbalanced across subjects. Participants had up to four seconds to choose between the proposed transfer, and a default transfer of $50 to both players. If the participant failed to respond within four seconds, the payoff for both individuals was $0 for that trial. Participants indicated their decisions using a 4-point scale: Strong No, No, Yes, Strong Yes. This allowed us to measure simultaneously their choice, and the strength of their preferences at the time of decision. To decouple motor activity from preference signals, the direction indicating increasing preference (right-to-left or left-to-right) varied randomly each trial.

The proposal shown in each trial was drawn from one of the nine pairs shown in Fig. 1B. Each pair appeared 20 times, randomly intermixed across subjects, and divided evenly across four scanner runs (5 instances per run). To minimize habituation and repetition effects, proposals (which ranged from $10 to $100) were randomly jittered by $1–$4 in each dimension, with the exception that amounts above $100 were always jittered downwards. Due to an error in coding, both self and other proposals were jittered by the same amount.

After the participant's response, a fixation cross appeared for a random delay of 2–4 seconds (average = 3 sec), followed by the trial's outcome displayed for 3 seconds (Fig. 1A). The outcome was stochastic: 60% of trials resulted in the chosen option (green check), while 40% resulted in the non-chosen option (red cross).

At the end of the experiment the outcome of one trial was selected randomly to determine the pay for both participants. To minimize the extent to which reciprocity considerations play a role in the experiment, participants never met their partner, and were assured that their identity would never be revealed. Moreover, participants knew that their partner would only know the final outcome for the randomly selected trial, and not their actual choice, and that during the

instruction period the partner was informed about the outcome randomization procedure described above. Participants were also assured that their partner was a real person, that no deception was being used in the experiment, and were asked to affirm that they believed this. All participants indicated a belief that their decisions affected a real partner, both before and after the scanner task.

Behavioral analyses. Subject-level generosity was measured by the average amount of money per trial that a subject sacrificed to increase their partner's payoffs. For example, compared to the default ($S = $50, $P = $50), accepting ($S = $25, $P = $100) entails a sacrifice of $25 on that trial. Fig. 1C depicts the distribution of generosity across subjects. We considered several alternative natural measures of subject-level generosity (such as money given to partner), which led to very similar results. In addition, we label a choice as Generous if the participant sacrificed to help the partner (i.e., accepting $Self < $50, rejecting $Self > $50), and Selfish otherwise.

The histogram of subject-level generosity suggested that the population might consist of a mixture of types with different levels of generosity. We tested this hypothesis formally by estimating the parameters of two models using Bayesian methods, using the package rjags[46] written for R[47]), and then carrying out a model comparison between them. The first model assumes that subjects' mean generosity is drawn from a mixture of two normally distributed populations. The model is characterized by five parameters: means (μ_1, μ_2) and standard deviations (σ_1, σ_2) for the two groups' distributions, and a mixing parameter λ indicating the probability that a subject's generosity level is drawn from the first group. We assumed largely uninformative priors for all parameters: a uniform distribution on [0,5] for μ_1, a uniform distribution on [2.5,20] on μ_2, uniform distributions on [0,100] for σ_1 and σ_2, and a Dirichlet

distribution with 1 observation for λ (this last simplifies the computations while maintaining an approximately uniform prior). Results of this analysis were robust to a wide range of prior specifications. The second model assumes that subject's generosity levels are drawn from a single normally-distributed population, and thus is characterized by two parameters: its mean μ and its standard deviation σ. The priors for this model were: uniform on [0,20] for μ and on [0,100] for σ. We compared the two models by computing the difference in the log-likelihood for each. To compute the significance value for this difference, we ran the same procedure 1000 times using simulated data based on the null hypothesis (normally distributed population with mean and standard deviation of the observed data) and computing the probability of a difference in the simulated log-likelihoods equal or larger in magnitude to the observed statistic[48].

fMRI data acquisition. BOLD responses were acquired using a Siemens 3.0 Tesla MRI scanner (Erlangen, Germany) to acquire gradient echo T2*-weighted echoplanar images (EPI). To optimize functional sensitivity in the orbitofrontal cortex (OFC), a key region of interest, we used a tilted acquisition in an oblique orientation of 30° to the anterior commissure–posterior commissure line[49]. In addition, we used a standard eight-channel phased array coil. Each volume comprised 45 axial slices. A total of 960 volumes were collected over four sessions during the experiment in an interleaved ascending manner. The first two volumes of each session were discarded to allow for scanner equilibration. The imaging parameters were as follows: echo time, 30 ms; field of view, 192 mm; in-plane resolution and slice thickness, 3 mm; repetition time, 2.75 s. Whole-brain high-resolution T1-weighted structural scans (1 x 1 x 1 mm) were acquired from the 51 subjects and coregistered with their mean EPI images and averaged together to permit anatomical localization of the functional activations at the group level.

fMRI data pre-processing. Image analysis was performed using SPM5 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). Images were corrected for slice acquisition time within each volume, motion corrected with realignment to the last volume, spatially normalized to the standard Montreal Neurological Institute EPI template using affine transformation, and spatially smoothed using a Gaussian kernel with a full-width at half-maximum of 8 mm. Intensity normalization and high-pass temporal filtering (using a filter width of 128 s) were also applied to the data.

GLM 1. We estimated several general linear models (GLMs) of the BOLD data. The first model was designed to compare patterns of activity between trials in which subjects acted generously and trials in which they acted selfishly, and to compare the response at outcome as a function of whether a pro-social choice was vetoed or received. The model was analyzed in three steps.

First, for each subject we estimated a GLM with AR(1) and the following regressors of interest: R1) A boxcar function for the choice period, which extends from proposal onset to the subject's response for the trial, for generous trials only; R2) R1 modulated by the stated preference on that trial; R3) A boxcar function for the choice period during selfish trials only; R4) R3 modulated by the stated preference on that trial (coded as -1.5 = Strong No, -.5 = No, .5 = Yes, 1.5 = Strong Yes); R5) an indicator function of the outcome period onset (i.e., a stick function of 0s duration) on trials in which a generous choice occurred and was implemented; R6) an indicator function of the outcome period onset on trials in which a generous choice occurred and was reversed; R7) an indicator function of the outcome period onset on trials in which a selfish choice was made and was implemented; R8) an indicator function of the outcome period

onset on trials in which a selfish choice was made and reversed.

Second, we calculated the following single-subject contrasts: 1) Generous vs. Selfish choice (R1–R3). 2) Preference on generous vs. selfish choice (R2–R4). 3) Reversed vs. implemented, generous choice (R6–R5). 4) Reversed vs. implemented selfish choice (R8–R7). All regressors were convolved with the canonical form of the hemodynamic response. Missed response trials were excluded from the above analysis. The model also included motion parameters and session constants as regressors of no interest. Due to estimation constraints, eight participants were excluded from contrasts 1 and 2 because they chose generously fewer than five times over the course of the experiment.

Third, we computed second-level random effects analyses for each group of interest (All subjects, Selfish group, Generous group) by computing one-sample and two-sample *t*-tests on the single-subject contrast coefficients.

For inference purposes, we imposed a cluster-corrected threshold of $p < .05$ (based on Gaussian random field theory as implemented in SPM5). We also report results in this and all analyses below that survived small-volume correction within regions for which we had strong *a priori* hypotheses (see *ROI definition* below), including medial prefrontal cortex (mPFC), ventral striatum (vStr), dorsolateral prefrontal cortex (dlPFC), and bilateral temporoparietal junction (TPJ).

GLM 2. The second model was designed to identify regions in which activity responded to $Self or $Partner at the time of choice. It consisted of the following regressors: R1) A boxcar function for the choice period on all trials; R2) R1 modulated by the value of $Self on each trial; R3) R1

modulated by the value of $Partner on each trial; R4) an indicator function of the outcome period onset (i.e., a stick function of 0 s duration); R5) R4 modulated by the monetary outcome for self on each trial; R6) R4 modulated by the monetary outcome for partner on each trial. Parametric modulators were orthogonalized as follows: $Partner was orthogonalized with respect to $Self, and outcome for partner was orthogonalized with respect to outcome for self. We then calculated the following single-subject contrasts: $Self vs. baseline, $Partner vs. baseline, and $Self - $Partner. All omitted details (e.g., contrast estimates, group-level analyses, etc.) are as in GLM 1.

***GLM* 3.** The third model was designed to identify regions in which activity responded monotonically to the stated preference expressed on each trial. It included the following regressors of interest: R1) An boxcar function for the choice period, which extends from proposal onset to the subject's response for the trial; R2) R1 modulated by the stated preference on each trial; R3) R1 modulated by the value of $Self on each trial; R3) R1 modulated by the value of $Partner on each trial; R5) an indicator function of the outcome period onset (i.e., a stick function of 0s duration); R6) R5 modulated by the monetary outcome for self on each trial; R7) R5 modulated by the monetary outcome for partner on each trial. All omitted details are as in GLM 1.

PPI 1. We estimated a psychophysiological interaction (PPI) analysis to investigate the functional connectivity of the area of right dlPFC associated with generous vs. selfish choices

(see ROI Definition below). The analysis proceeded in three steps:

1. Individual BOLD time-series were computed using the first eigenvariate of the time series within a sphere (4 mm radius) centered on individual subject peak sensitivity generous vs. selfish choices (GLM1) within a functional mask of the right dlPFC region that that exhibited stronger response in GLM at the time of choice during generous compared to selfish choice trials (see ROI definition below).
2. Seed time-courses were deconvolved based on the formula for the canonical hemodynamic response, in order to construct a time series of neural activity in rdlPFC. This was done following the procedures described in [50].
3. A GLM was estimated with the following regressors: R1) an interaction between neural time series of the region and contrast function for the psychological regressor (generous choice period = +1, selfish choice period = -1) on all trials. R2) the psychological variable (R1–R3 in GLM1), and R3) the first BOLD eigenvariate time series from the 4 mm sphere.

The first two regressors were convolved with a canonical form of the hemodynamic response. The model also included motion parameters as regressors of no interest. Single subject contrasts for the first regressor were calculated and submitted to one-sample t-test to determine group activations.

PPI 2. We estimated a second PPI to investigate functional connectivity of the left TPJ region associated with $Partner in GLM 2 and stated preferences in GLM 3 (see ROI Definition and Analyses below). All details are identical to PPI 1, with the following differences:

1. Individual BOLD time series were computed of average activation within a sphere (4 mm radius) centered on individual subject peak sensitivity to $Partner (see GLM1) within a functional mask of the left TPJ.
2. Individual peaks within this mask were identified using this same contrast.
3. The psychological variable used to determine R1 and R2 of the GLM consisted of all choice periods, regardless of whether the subject chose generously (R1 from GLM 2).

ROI Definition and Analyses. For use in small-volume correction, we defined four regions anatomically using WFU PickAtlas (http://fmri.wfubmc.edu/software/PickAtlas), with a dilation of 3 mm to ensure full coverage of an area. The mPFC mask included bilateral anterior cingulate cortex, rectus, and medial orbitofrontal gyrus from the AAL atlas (2,733 voxels total). This region encompasses the peak voxels related to value computation in several independent studies[27,29,32,37]. The striatum encompassed the head of the caudate bilaterally from the Talairach Daemon atlas (438 voxels). The dlPFC was defined using a combined bilateral mask of the middle frontal and inferior frontal gyrus (6,085 voxels total). The TPJ region included bilateral angular and superior temporal gyrus, posterior to y = -40 (1975 voxels), a region which encompassed peaks of activation from several studies of theory-of-mind[16,18,40].

We also defined four regions based on functional criteria and extracted specific contrasts of interest from each: 1) The right dlPFC defined by the set of voxels where group activity was significantly greater on generous compared to selfish trials (GLM 1), at $p < .001$, uncorrected. Parameter estimates of the contrast between generous vs. selfish choice (R1–R3 in GLM 1) for all voxels included in the mask were averaged to extract a single value for this region. This estimate was correlated with differences in reaction time, differences in vmPFC correlation with

preference on generous vs. selfish choices, response in the vmPFC to generous choice veto, and average generosity levels. 2) A region of the vmPFC showing a significant negative PPI with the right dlPFC (PPI 1, R1, thresholded at $p < .001$, uncorrected). The average parameter estimate of the difference in stated preference during generous vs. selfish choice (R2–R4 in GLM 1) was extracted from this region and correlated with dlPFC response during generous vs. selfish choices. 3) The left angular gyrus/TPJ and 4) rostral ACC regions were defined by the set of overlapping voxels that correlated both with $Partner in GLM 2 and stated preference in GLM 3, both thresholded at $p < .0005$, uncorrected. Parameter estimates for the difference of $Partner vs. $Self (R3–R2 in GLM 2) were extracted from each of these regions and correlated with response to generous choice veto in the vmPFC (see below), average generosity, and dlPFC response during generous choice.

To examine responses to generous choice veto vs. receipt, we defined a fifth vmPFC region (independent of functional activations in the current study) related to the subjective pleasantness of outcomes, using a sphere (6 mm radius) centered on the coordinates of activation described by a previous study of hedonic encoding (see Figures 2 and 4)[36]. This measure was correlated with activity from the function ROIs defined above.

To minimize the influence of outliers on conclusions drawn from these ROI analyses, all correlations used robust regression. To make the robust regression coefficient equivalent to a standard correlation statistic, the dependent and independent variables were normalized prior to regression.

References

1. **Nowak, M.A. & Sigmund, K.** "Evolution of indirect reciprocity by image scoring." *Nature*, 1998, 393, 573–7.

2. **Dufwenberg, M. & Kirchsteiger, G.** "A theory of sequential reciprocity." *Games Econ Behav*, 2004, 47, 268–98.

3. **Falk, A. & Fischbacher, U.** "A theory of reciprocity." *Games Econ Behav*, 2006, 54, 293–315.

4. **Henrich, J., et al.** "In search of Homo economicus: Behavioral experiments in 15 small-scale societies." *American Economic Review,* 2001, 91, 73–8.

5. **Fehr, E. & Schmidt, K.** "A theory of fairness, competition, and cooperation." *Q J Econ*, 1999, 114, 817–68.

6. **Charness, G. & Rabin, M.** "Understanding Social Preferences with Simple Tests." *Q J Econ*, 2002, 117, 817–869.

7. **Engelmann, D. & Strobel, M.** "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments." *American Economic Review*, 2004, 94, 857–69.

8. **Chang, L.J., Smith, A., Dufwenberg, M. & Sanfey, A.G.** "Triangulating the neural, psychological, and economic bases of guilt aversion." *Neuron*, 2011, 70, 560–72.

9. **Dana, J., Weber, R.A. & Kuang, J.X.** "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Econ Theor*, 2007, 33, 67–80.

10. **Andreoni, J.** "Impure Altruism and Donations to Public-Goods—a Theory of Warm-Glow Giving." *Econ J*, 1990, 100, 464–77.

11. Andreoni, J. & Bernheim, B. "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects." *Econometrica*, 2009, 77, 1607–36.

12. List, J. "On the interpretation of giving in dictator games." *J Polit Econ*, 2007, 115, 482–93.

13. Belk, R. & Coon, G. "Gift giving as agapic love: An alternative to the exchange paradigm based on dating experiences." *J Consum Res*, 1993, 393–417.

14. Loewenstein, G., Thompson, L. & Bazerman, M. "Social utility and decision making in interpersonal contexts." *J Pers Soc Psych*, 1989, 57, 426–41.

15. Haley, K.J. & Fessler, D.M.T. "Nobody's watching? Subtle cues affect generosity in an anonymous economic game." *Evol Hum Behav*, 2005, 26, 245–56.

16. Saxe, R. & Powell, L.J. "It's the thought that counts: Specific brain regions for one component of theory of mind." *Psychol Sci*, 2006, 17, 692–9.

17. Moll, J., et al. "Human fronto–mesolimbic networks guide decisions about charitable donation." *Proc Natl Acad Sci USA*, 2006, 103, 15623.

18. Hare, T.A., Camerer, C.F., Knoepfle, D.T. & Rangel, A. "Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition." *J Neurosci*, 2010, 30, 583–90.

19. Harbaugh, W., Mayr, U. & Burghart, D. "Neural responses to taxation and voluntary giving reveal motives for charitable donations." *Science*, 2007, 316, 1622.

20. Tricomi, E., Rangel, A., Camerer, C.F. & O'doherty, J.P. "Neural evidence for inequality-averse social preferences." *Nature*, 2010, 463, 1089–91.

21. Zaki, J. & Mitchell, J.P. "Equitable decision making is associated with neural markers of intrinsic value." *Proc Natl Acad Sci USA*, 2011, 108, 19761–6.

22. Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C. & Fehr, E. "Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice." *Nat Neurosci*, 2011, 14, 1468–74.

23. Sanfey, A., Rilling, J., Aronson, J., Nystrom, L. & Cohen, J. "The neural basis of economic decision-making in the ultimatum game. *Science*, 2003, 300, 1755.

24. Steinbeis, N., Bernhardt, B.C. & Singer, T. "Impulse Control and Underlying Functions of the Left DLPFC Mediate Age-Related and Age-Independent Individual Differences in Strategic Social Behavior." *Neuron*, 2012, 73, 1040–51.

25. Aron, A.R., Robbins, T.W. & Poldrack, R.A. "Inhibition and the right inferior frontal cortex." *Trends Cogn Sci*, 2004, 8, 170–7.

26. Konishi, S., et al. "Common inhibitory mechanism in human inferior prefrontal cortex revealed by event-related functional MRI." *Brain*, 1999, 122 pt. 5, 981–91.

27. Hare, T.A., Camerer, C.F. & Rangel, A. "Self-control in decision-making involves modulation of the vmPFC valuation system." *Science*, 2009, 324, 646–8.

28. Hare, T.A., Malmaud, J. & Rangel, A. "Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice." *J Neurosci*, 2011, 31, 11077–87.

29. McClure, S.M., Laibson, D.I., Loewenstein, G. & Cohen, J.D. "Separate neural systems value immediate and delayed monetary rewards." *Science*, 2004, 306, 503–7.

30. MacDonald, A.W., Cohen, J.D., Stenger, V.A. & Carter, C.S. "Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control." *Science*, 2000, 288, 1835.

31. Kerns, J.G., et al. "Anterior cingulate conflict monitoring and adjustments in control." *Science*, 2004, 303, 1023–6.

32. Plassmann, H., O'Doherty, J. & Rangel, A. "Orbitofrontal cortex encodes willingness to pay in everyday economic transactions." *J Neurosci*, 2007, 27, 9984–8.

33. Wunderlich, K., Rangel, A. & O'Doherty, J.P. "Neural computations underlying action-based decision making in the human brain." *Proc Natl Acad Sci USA*, 2009, 106, 17199–204.

34. Basten, U., Biele, G., Heekeren, H.R. & Fiebach, C.J. "How the brain integrates costs and benefits during decision making." *Proc Natl Acad Sci USA*, 2010, 107, 21767–21772.

35. Hare, T.A., Schultz, W., Camerer, C.F., O'Doherty, J.P. & Rangel, A. "Transformation of stimulus value signals into motor commands during simple choice." *Proc Natl Acad Sci USA*, 2011, 108, 18120–5.

36. Plassmann, H., O'Doherty, J., Shiv, B. & Rangel, A. "Marketing actions can modulate neural representations of experienced pleasantness." *Proc Natl Acad Sci USA*, 2008, 105, 1050–4.

37. Kable, J.W. & Glimcher, P.W. "The neural correlates of subjective value during intertemporal choice." *Nat Neurosci*, 2007, 10, 1625–33.

38. Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A. & Saxe, R. "Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments." *Proc Natl Acad Sci USA*, 2010, 107, 6753–8.

39. Tankersley, D., Stowe, C. & Huettel, S. "Altruism is associated with an increased neural response to agency." *Nat Neurosci*, 2007, 10, 150–1.

40. Decety, J. & Lamm, C. "The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition." *Neuroscientist*, 2007, 13, 580.

41. Barnea-Goraly, N., et al. "White matter structure in autism: preliminary evidence from diffusion tensor imaging." *Biol Psychiatry*, 2004, 55, 323–6.

42. Izuma, K., Matsumoto, K., Camerer, C.F. & Adolphs, R. "Insensitivity to social reputation in autism." *Proc Natl Acad Sci USA*, 2011, 108, 17302–7.

43. Figner, B., et al. "Lateral prefrontal cortex and self-control in intertemporal choice." *Nature*, 2010, 13, 538–9.

44. Luo, S., Ainslie, G., Giragosian, L. & Monterosso, J.R. "Behavioral and neural evidence of incentive bias for immediate rewards relative to preference-matched delayed rewards." *J Neurosci*, 2009, 29, 14820–7.

45. Harbaugh, W.T. "The prestige motive for making charitable transfers." *Am Econ Rev*, 1998, 88, 277-82.

46. Plummer, M. rjags: Bayesian graphical models using MCMC. R package version 2.2.0-4. http://CRAN.R-project.org/package=rjags. 2011.

47. Team, R.D.C. R: A language and environent for statistical computing. in *R Foundation for Statistical Computing.* Vienna, Austria: 2008.

48. McLachlan, G.J. "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture." *J R Stat Soc Ser C Appl Stat*, 1987, 36, 318–24.

49. Deichmann, R., Gottfried, J.A., Hutton, C. & Turner, R. "Optimized EPI for fMRI studies of the orbitofrontal cortex." *NeuroImage*, 2003, 19, 430–41.

50. Gitelman, D.R., Penny, W.D., Ashburner, J. & Friston, K.J. "Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution." *Neuroimage*, 2003, 19, 200–7.