Supporting Information

Greene and Paxton 10.1073/pnas.0900152106

SI Text

SI Methods and Related Discussion. The present experimental design differs substantially from those used previously in cognitive neuroscience and moral psychology. For this reason, we here attempt to anticipate concerns and misunderstandings that are likely to arise from our methods and interpretation. This section includes supplemental methodological information and addresses related concerns. The *SI Discussion* that follows addresses further concerns related to the interpretation of our data.

Exclusion of Subject for Strategic Underreporting of Accuracy. We classified subjects as "honest" or "dishonest" based on their reported levels of accuracy in the Opportunity condition. However, it is possible to gain money dishonestly while maintaining a chance level of accuracy by cheating in relatively high-value Opportunity trials and deliberately underreporting accuracy for relatively low-value Opportunity trials. Subjects who use this strategy should exhibit improbably high levels of cumulative reward given their win/loss percentages. To identify such subjects we compared the winnings of each honest subject to those of simulated honest subjects (10,000 permutations) with win/loss percentages individually matched to the subject being tested. Based on these findings, we discarded the data of one subject initially classified as honest whose winnings were improbably large given that subject's win/loss percentage ($P = 0.005$). The winnings of all other honest subjects were consistent with their respective win/loss percentages ($P > 0.05$), making the excluded subject an extreme outlier. This subject was excluded because s/he could not be classified as "honest" (for obvious reasons) and did not meet our established, and rather conservative, criteria for inclusion in the "dishonest" group, which is based on self-reported accuracy in the Opportunity condition. Likewise, it did not make sense to include this subject in the "ambiguous" group because his/her self-reported accuracy appears to be distorted, and it is this accuracy report that is used in the individual differences analysis that includes the "ambiguous" subjects.

Exclusion of Subjects Based on Suspicion or Ignorance. In debriefing, subjects were first asked, in an open-ended way, what they thought the experiment was about. At this point in debriefing, 4 subjects initially classified as dishonest, 1 subject classified as ambiguous, and 4 subjects classified as honest voiced suspicions that the experiment was about cheating/lying/dishonesty. We discarded the data from the 4 dishonest subjects, but not the others. Our aim in doing this was to exclude data from subjects who may be seen as morally justified in deceiving the experimenters because they believed that the experimenters were attempting to deceive them. We adopted this policy as a conservative measure, anticipating that some may hesitate to call such deception dishonest. (See the following discussion concerning our operational definitions of honesty and dishonesty.) We included the remaining subjects because it is not essential to our design that honest behavior be motivated by purely moral (rather than prudential) considerations. (See the following discussion.) Additional analyses verified that our key findings held when the 4 suspicious honest subjects were excluded.

Subjects were eventually informed of the purpose of the experiment and were asked whether they were aware that they could cheat. All but one subject indicated that they were aware of this. Data from this subject were excluded because our aim is to investigate honest behavior in the face of opportunity for dishonest gain, and this subject was not aware of the opportunity.

Inclusion of Subjects with Prior Participation. To ensure an adequate supply of dishonest behavior for our fMRI experiment, we recruited subjects who, based on their performances in pilot testing, were likely to exhibit high levels of dishonest behavior in a second testing session, and while undergoing brain scanning. These subjects were not debriefed before their participation in the fMRI experiment. Two consequences of this procedure deserve attention. First, the distribution of honest/dishonest performances observed in the fMRI study (Fig. 2) is not necessarily representative of our subject pool. (The proportions of subjects reaching dishonesty threshold in pilot testing and in the present experiment were comparable, both at ≈40%, depending on exclusions. However, only 26% of first-time subjects reached dishonesty threshold in the present experiment, suggesting that the brain scanning environment may have reduced the level of dishonesty.) Second, the proportion of first-time and repeat subjects differs between the honest and dishonest groups, raising the possibility that our findings could be accounted for by differences in task experience rather than differences in honest/dishonest behavior (11 of 14 honest subjects were first-time subjects; 5 of 14 dishonest subjects were first-time subjects). This alternative hypothesis could possibly explain why we observed differences in control network activity between groups. However, it cannot explain within-group (first-time group or repeat-group) correlations between levels of control network activity and frequency of dishonest behavior.

Thus, to test this alternative hypothesis, we reexamined the results of our regression analysis correlating individual differences in control network activity with individual levels of dishonesty (Fig. 4 and Table S2). To determine whether the success of the regression model depends on a confound based on first-time ($n = 19$) vs. repeat ($n = 16$) subjects, we separately assessed the accuracy of the model predictions for both groups. The correlations between model predictions and actual values were very high for both groups: $r = 0.89$ ($P < 0.0001$) for first-time subjects and $r = 0.95$ ($P < 0001$) for repeat subjects. Because the model accounts for most of the variance *within* the first-time subjects and *within* the repeat subjects, the success of the model cannot be explained in terms of confounding differences between these 2 groups. We note that this regression analysis is based on percent signal changes in ROIs identified by our 2 critical within-subject contrasts: Opportunity Wins > No-Opportunity Wins and Opportunity Losses > No-Opportunity Losses.

Probabilistic Classification of Subjects as Honest, Dishonest, or Ambiguous. One might object to our use of statistical methods to classify subjects as honest and dishonest. More specifically, one might claim that it is illegitimate to label behavior as dishonest simply because the evidence indicates that the subject in question *probably* cheated. We note, however, that most scientific conclusions are supported by statistical analyses culminating in probability estimates (P values). Thus, this objection, if taken seriously, would discredit not only our classification system, but the conclusions of most scientific papers. We emphasize further that our threshold for classifying an individual subject as dishonest is very conservative ($P < 0.001$). It is true that our method does not allow us to identify individual responses as dishonest, but this does not prevent us from identifying individual subjects

as dishonest. (See discussion of implications for brain-based lie detection in following text.) Finally, we emphasize again that in labeling subjects as dishonest, we are describing their present behavior only and not ascribing to them stable personality traits.

Characteristics of Honest vs. Dishonest Subjects. There were no significant differences in age (t test, $P = 0.16$), gender (χ^2, $P = 0.7$), or paranormal belief (t test, $P = 0.83$) between honest and dishonest subjects.

Procedural Deviations. For 13 subjects, a stimulus programming error caused the properly randomized sequence of Opportunity and No Opportunity trials used in the first run to be repeated for subsequent runs. This error, although regrettable, does not compromise the findings presented here. Subjects were given no additional information that would allow them to make more accurate predictions, and the resulting changes in trial sequence did not confound the comparisons made in our analyses. The primary consequence of this error is that subjects did not necessarily respond to equal numbers of each trial type, thus reducing statistical power. Subjects may also have been able to anticipate upcoming trial types, but, once again, the repetition in sequencing provided subjects with no strategically useful information.

Subject Instructions. The following instructions were presented to subjects on a computer:

Thank you for participating. In this study your job is to predict the outcomes of computerized random coin flips. You may not think that you have the ability to do this, and that's okay. Just do your best. You may be surprised at what you can do! Press any key to continue.

It has been suggested that people make more accurate predictions when they are motivated to predict accurately. To test this idea, we will be providing you with varying levels of financial incentive. Before each coin flip happens, an amount of money will appear on the screen (e.g., $0.25 or $5.00). This is the amount of money that you will win or lose depending on whether you accurately predict the outcome of the coin flip. If your prediction is correct, then you win the amount of money shown. If your prediction is incorrect, you lose the amount of money shown. The computer will keep track of all of your wins and losses. If, at the end of the experiment, your money total is positive, you will be paid that amount. If your total is negative or zero, you will not win any additional money. This is not pretend money. This is real money that you will be paid based on your performance in the experiment. However, your winnings cannot exceed $75. Press any key to continue.

It has been suggested that people's ability to predict the future is disrupted if they have to record their predictions externally (i.e., outside of their minds). To test this idea, we will sometimes ask you to report your prediction in advance. In other cases, you will simply tell us after the fact whether or not your prediction was correct. Press any key to continue.

Before each coin flip you will see the dollar amount that the trial is worth and, below it, the word "PREDICT" on the screen. At that point you should make your prediction in your mind. Next you will either see the word "RECORD" or the word "RANDOM." If you see the word "RECORD" you should press the button on the LEFT to indicate that you are predicting HEADS or the button on the RIGHT to indicate that you are predicting TAILS. If you see the word "RANDOM" then you should randomly press either the LEFT button or the RIGHT button. When you make random responses, you should not follow any fixed pattern. Press any key to continue.

Next you will see the word "HEADS" or "TAILS" appear on the screen. This is the outcome of the computerized coin flip. After that you will see a screen that says "CORRECT?" At that point you must indicate whether or not your prediction was correct. If you were told to hit a random button, it does not matter which button you hit. Whether or not your prediction was correct depends only on the prediction you made in your mind and the outcome of the coin flip. Press the LEFT key to indicate (YES) that your prediction was correct. Press the RIGHT key to indicate (NO) that your prediction was incorrect. If you failed to form a prediction before the outcome of the coin flip was revealed, then you should indicate that your prediction was incorrect. After you have said whether your prediction was correct, the computer will tell you how much money you won or lost on that coin flip. Press any key to continue.

You will make a total of 210 predictions. You will do these in 7 groups of 30 trials. There will be about 10 seconds between the end of one trial and the beginning of the next one. After each group of trials you will have a chance to rest. The whole task will take a little less than 90 min. Press any key to continue.

You are now ready to practice. Remember, first comes the dollar amount telling you what the coin flip is worth and the word "PREDICT." At that point you will make your prediction privately to yourself. (Note that the dollar amounts presented here will not count toward your final total.) Then you will see either "RECORD" or "RANDOM." If you see "RECORD" enter your prediction (LEFT key HEADS, RIGHT key for TAILS). If you see "RANDOM" press either the LEFT key or the RIGHT key randomly. Then you will see the outcome of the coin-flip (HEADS or TAILS). Then you will see the word "CORRECT?" on the screen. At that point you indicate whether the prediction you made in your mind was correct. Press the LEFT key (YES) if your prediction was correct or the RIGHT key (NO) if your prediction was incorrect. Then the computer will tell you how much money you won or lost on that coin flip. Then you wait for the next coin flip, which will begin with a dollar amount, as before. Press any key to begin practicing.

SI Discussion

Defining Honesty and Dishonesty. In attempting to study honesty and dishonesty scientifically, one cannot avoid making assumptions about what it means to be honest or dishonest, despite that these terms are not precisely defined in ordinary discourse. For present purposes we have defined honesty and dishonesty in minimal behavioral terms, i.e., as behaviors that a reasonable person would regard as honest or dishonest given the circumstances. Were the honest people really honest? In refraining from lying, they knowingly "left money on the table." This behavior must have some motivational basis, which we here refer to as honesty. This minimal, behavioral conception of honesty does not involve ascribing noble motivations to these individuals. In calling them honest, we are claiming only that they chose not to behave dishonestly. [It is a controversial philosophical question whether, and to what extent, more noble forms of honesty and other virtues exist (1).] Were the "dishonest" people really dishonest? These individuals violated the rules of the game, to which they had agreed, and gained money as a result. What's more, most of the individuals we tested either did not violate these rules or did so less than they could have. This suggests a prevailing norm against the behavior we have called dishonest. We are agnostic as to whether this dishonest behavior is conscious or unconscious. In our opinion, the observed association between control network activity and dishonest behavior is no less significant, and is perhaps more significant, if it turns out that the dishonest behavior in question is largely unconscious.

Interpretation of Control Network Activity and Reverse Inference. Because our conclusions do not depend on any specific interpretation of the observed control network activity, or even on the appropriateness of the "control network" label, our conclusions do not depend on any kind of problematic reverse inference (2). With respect to the honest subjects, our key finding is that no brain regions, whether in the control network or elsewhere,

exhibited significant increases in activity when honest subjects chose to forgo opportunities for dishonest gain (as compared with matched trials with no opportunity). Here there is no reverse inference because there are no regional brain activations to interpret. To the extent that we may accept the "control network" label as valid, we may infer that an analogue of the Grace hypothesis applied to dishonesty is probably false: Dishonest behavior appears to involve the engagement of additional controlled cognitive processes.

Attribution of fMRI BOLD Effects to Accuracy Reports. As noted in the main text, it is unlikely that the fMRI BOLD effects attributed to dishonest decisions (Fig. 3 *A* and *B*) are related to the preceding behavioral responses whereby subjects recorded their predictions (No Opportunity) or pressed random buttons (Opportunity). Once again, this is because the honest subjects (who also recorded their predictions/pressed random buttons) did not exhibit such effects and because the fMRI data are correlated with the frequency of dishonest behavior (Fig. 4). We also noted that the timing of the BOLD signal is more consistent with its being related to the accuracy reports than to the prediction/random responses. This is illustrated in Fig. S2, which depicts the mean time course of fMRI BOLD activity in the regions depicted in Fig. 3 *A* and *B* for the conditions that exhibited greater activity in the relevant contrasts. As Fig. S2 illustrates, the signal tends to peak ≈5 sec following the accuracy report, consistent with the typical 4- to 6-sec lag in peak BOLD response following a neural event (3). If the signal were primarily related to the earlier behavioral responses, one would expect the signal to peak ≈3 sec earlier.

The RT data also speak against this alternative interpretation. As noted in the main text, accuracy reports took longer for Opportunity Loss trials than for No-Opportunity Loss trials ($P < 0.0001$) and for Opportunity Win trials ($P < 0.0001$), but only within the dishonest group. We performed parallel analyses on the RTs for the earlier behavioral responses. For the first contrast (dishonest: Opportunity Loss vs. No-Opportunity Loss) we found a marginally significant effect ($P = 0.04$) in the direction *opposite* that predicted by the alternative hypothesis. That is, the dishonest subjects took slightly longer to record their predictions (No Opportunity) than to make their random button presses (Opportunity). This is consistent with their putting more effort into prediction in the No Opportunity condition (when they have to make a prediction), but this result cannot explain why Opportunity trials are associated with more control network activity. The second contrast (dishonest: Opportunity Loss vs. Opportunity Win) did not reveal any significant difference in the random button-press RTs ($P = 0.29$). Thus, the RT data for the moral decisions converge with the fMRI data, but the RT data for the earlier behavioral responses do not.

Is It Self-Evident That the Grace Hypothesis Is Correct? A common criticism of social-psychological research is that the conclusions reached are self-evident. Here, one might suppose that it is self-evident that the Grace hypothesis is correct. Indeed, the Grace hypothesis may be self-evidently correct with respect to some situations. For example, it seems highly unlikely (although not impossible) that ordinary law-abiding citizens actively resist the temptation to shoplift whenever they walk through a store with minimal security. Thus, one might wonder whether the situation examined here is also one in which it is self-evidently the case that honest behavior involves little active self-control.

To assess commonsense expectations concerning the psychology of honest behavior in our coin-flip prediction experiment, we conducted an additional survey. We emphasize, however, that this survey was *not* conducted to assess the validity of the conclusions drawn from our main experiment. Rather, we conducted this survey to empirically assess the extent to which our main conclusion is self-evident. [Other researchers have used similar techniques to assess the self-evidence of their conclusions, most famously Milgram (4).]

Fifty subjects (27 females, mean age 27.5) completed a 1-page survey in Harvard Square and were compensated $2. The survey described the behavioral aspect of the coin-flip prediction experiment in detail and asked people to respond to the following 2 questions:

Question 1: Please circle the answer below that best describes how things would go if you were to participate in this experiment:

A. I would not be tempted to cheat, at least not for most of the experiment.

B. I would be tempted to cheat during much of the experiment, but I would resist that temptation and not cheat.

C. I would cheat.

Question 2: Which of the following statements do you think best describes people who choose NOT to cheat in this experiment?

A. These people are not tempted to cheat, at least not for most of the experiment.

B. These people are tempted to cheat during much of the experiment, but they resist that temptation and don't cheat.

The results were as follows:

Question 1: A. 38% (19/50), B. 46% (23/50), C. 16% (8/50)

Question 2: A. 32% (16/50), B. 68% (34/50).

Thus, a majority of survey subjects who thought that they themselves would behave honestly in this experiment thought that they would do so through substantial resistance of temptation (Will). Here, respondents did not significantly favor one hypothesis over the other (binomial test, $P > 0.05$), despite the fact that a majority favored the Will hypothesis. In response to question 2, the tendency to favor the Will hypothesis (answer B) was significant (binomial test, $P < 0.02$). Thus, it is by no means self-evident that the findings of our experiment would end up supporting the Grace hypothesis, and, if anything, common sense appears to favor the Will hypothesis.

1. Kavka, G (1986) *Hobbesian Moral and Political Theory* (Princeton Univ Press, Princeton, NJ).
2. Poldrack RA (2006) Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci* 10:59–63.
3. Huettel S, Song A, McCarthy G (2004) *Functional Magnetic Resonance Imaging* (Sinauer, Sunderland, MA).
4. Milgram, S (1974) *Obedience to Authority* (Harper and Row, New York).

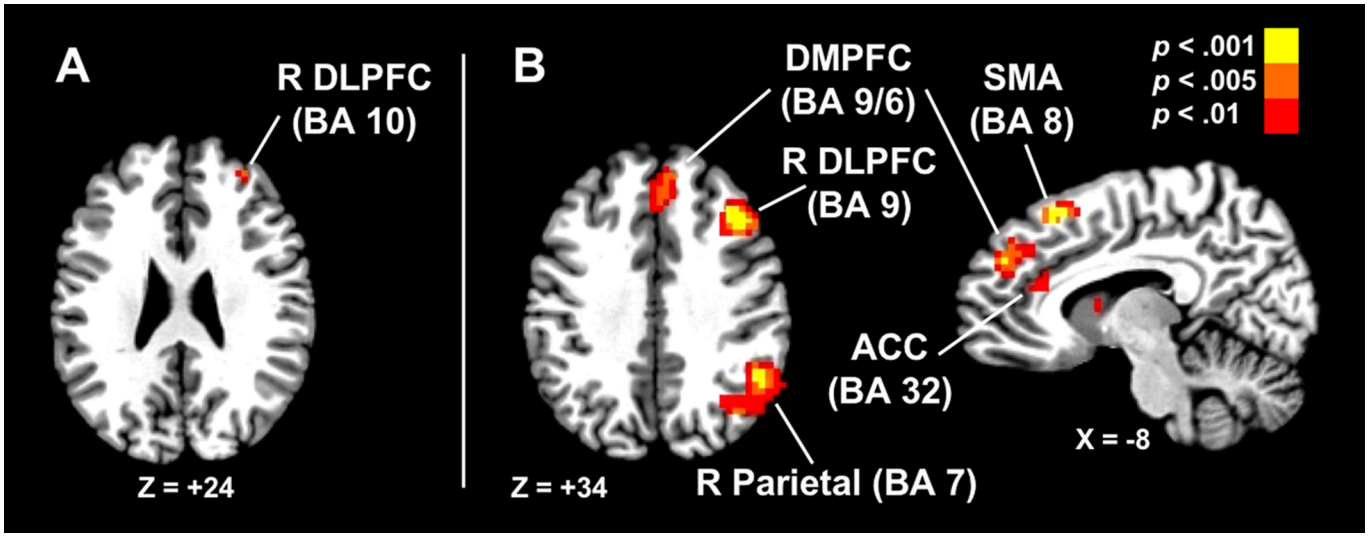

Fig. S1. Selected brain regions exhibiting interactions between group (honest vs. dishonest) and condition (Opportunity vs. No Opportunity) within Win trials (*A*) and Loss trials (*B*). fMRI data are projected onto a reference anatomical image. See Table S2 for further details. BA, Brodmann area.

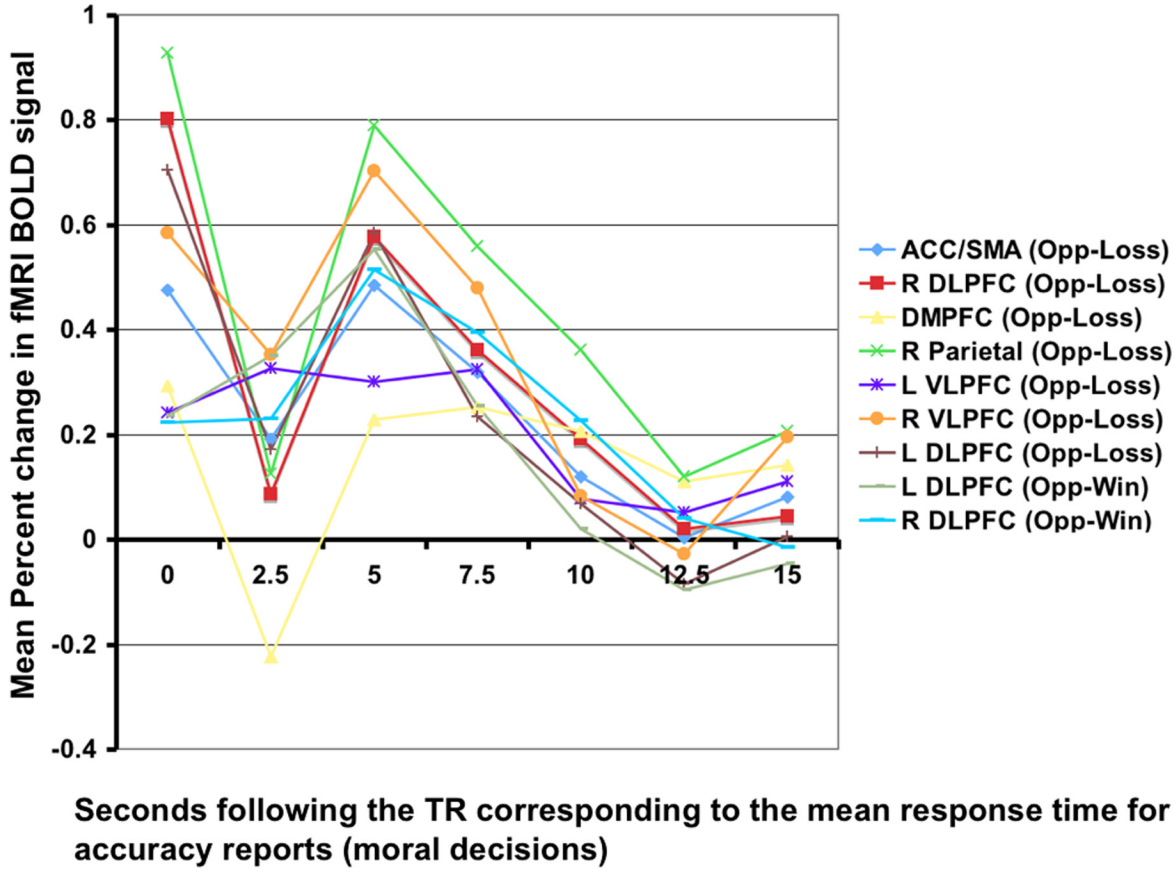

Fig. S2. Time course of activity in brain regions exhibiting increased activity in the Opportunity condition (vs. No Opportunity) within dishonest subjects (see Fig. 3 *A* and *B*). Data are shown for the Opportunity condition only. Bold responses tend to peak ≈5 sec following the accuracy report (moral decision). This is consistent with BOLD effects in these regions being related to accuracy reports, rather than prior behavioral responses, which occurred ≈8 sec before the peak responses in most regions.

Table S1. Results of planned fMRI contrasts

Group/contrast/region	R/L/M	BA	Max *t* (df = 13)	k	Talairach coordinates	Group × Contrast *F*(1, 26)	*P*
Dishonest							
Op Wins > No-Op Wins							
Superior frontal gyrus (DLPFC)	L	9/10	5.72	11	35, −47, 30	1.20	0.29
	R	9/10	5.53	9	−29, −50, 27	4.46	0.04
Op Losses > No-Op Losses							
Anterior cingulate (ACC)/	M	32	9.13	201	−8, −23, 50	11.02	0.003
Superior frontal gyrus (SMA)		8/6					
Middle frontal gyrus (DLPFC)	R	9/10	7.36	133	−35, −32, 32	13.42	0.001
	L	10/46	5.10	9	46, −38, 23	3.12	0.09
Medial frontal gyrus (DMPFC)	M	9	4.84	17	7, −47, 29	9.03	0.006
Inferior/superior parietal lobe	R	39/7	5.07	16	−38, 67, 44	12.86	0.001
Inferior frontal gyrus (VLPFC)	L	47	5.06	15	40, −23, −10	4.71	0.04
	R	47	5.06	11	−44, −20, −1	3.59	0.07
Op Wins > Op Losses							
Postcentral gyrus	R	2	5.04	15	−32, 38, 69	0.02	0.89
Postcentral gyrus	R	2	5.59	10	−44, 35, 63	0.48	0.49
Honest							
Op Wins > No-Op Wins							
Inferior frontal gyrus (VLPFC)	L	47/13	6.01	36	31, −20, −13	2.58	0.12
	R	47/13	5.12	9	−29, −14, −13	5.76	0.02
Op Losses > No-Op Losses (no significant effects)							
Op Wins > Op Losses							
Postcentral gyrus	R	3	6.88	264	−41, 29, 54	2.04	0.17
Superior frontal gyrus	R	8	6.39	21	−14, −41, 54	2.63	0.12
Middle frontral gyrus	L	6	4.56	13	23, −20, 54	4.12	0.05

Note: No brain regions exhibited increased activity for the contrasts opposite those above. Voxelwise threshold is $P < 0.001$, uncorrected; cluster threshold = 8 voxels; df = 13. To test for Group × Contrast interactions, we computed for each subject the mean percent signal change from baseline in each of the above ROIs. We then computed difference scores for each ROI for each subject, subtracting the percent signal change scores for the 2 cells that generated the ROI. We then made a between-group comparison of these difference scores for each ROI (2 rightmost columns). BA, Brodmann area; k, cluster size, Op, opportunity.

Table S2. Regions exhibiting Group (Honest vs. Dishonest) × Condition (Op vs. No Op) interactions

Trial type/region	R/L/M	BA	k	Talairach Coordinates	Max $F(1, 26)$	Uncorrected threshold, $P <$
Within Win trials						
Superior frontal gyrus (DLPFC)	R	10	8	−32 −47 27	10.33	0.01*
Superior frontal gyrus (DLPFC)	L	10	10	35 −56 18	6.56	0.05
Superior frontal gyrus (DLPFC)	L	8	13	20 −38 54	7	0.05
Within Loss trials						
Middle frontal gyrus (DLPFC)	R	9	33	−41 −26 36	16.34	0.001*
Superior frontal gyrus (DLPFC)	R	10	19	−26 −53 18	9.66	0.005*
Middle frontal gyrus/superior frontal gyrus (DLPFC)	L	6/8	18	38 −11 54	7.17	0.05
Anterior cingulate (ACC)	R	32	11	−5 −38 18	7.71	0.01*
Anterior cingulate (ACC)	L	24/32	10	8 −32 21	8.41	0.01*
Superior frontal gyrus (SMA)	R	8	18	−5 −17 51	14.56	0.001*
Inferior parietal lobe/supramarginal gyrus	R	40	15	−50 53 36	14.41	0.001
Superior parietal lobe	R	7	24	−38 65 54	10.38	0.005
Medial frontal gyrus (DMPFC)	M	6/9	44	−2 −41 36	9.45	0.005*
Inferior frontal gyrus (VLPFC)	R	47	8	−47 −23 −1	4.71	0.05
	L	47	11	38 −20 −4	5.42	0.05

*Survives partial-volume correction ($P < 0.05$) performed over prefrontal cortex. Results are from whole-brain voxelwise analyses with a cluster threshold of 8 voxels.

Only effects consistent with a priori regions of interest are listed. For all effects, (Dishonest Op − Dishonest No Op) > (Honest Op − Honest No Op). See Table S1 for functional ROI-based interaction analyses. BA, Brodmann area; k, cluster size; Op, opportunity.

Table S3. Reduced regression model predicting individual subjects' percent Wins in the Op condition

Predictor	Condition	Estimate	SE	*t*	*P*
Intercept		65.95	2.19	30.16	<0.0001
L superior frontal gyrus (DLPFC)	OpWin	42.46	9.73	4.36	0.0002
medial frontal gyrus (DMPFC)	OpLoss	49.56	11.53	4.3	0.0002
medial frontal gyrus (DMPFC)	OpWin	−55.57	14.1	−3.94	0.0005
L inferior frontal gyrus (VLPFC)	OpWin	−60.3	16.39	−3.68	0.001
R inferior/superior parietal lobe	OpWin	−24.7	8.68	−2.84	0.008
R inferior/superior parietal lobe	OpLoss	14.7	7.57	1.94	0.06
R inferior frontal gyrus (VLPFC)	OpWin	21.72	11.97	1.81	0.08

Probability to leave = 0.1. Op, opportunity. $R^2 = 0.79$, adjusted $R^2 = 0.74$, $r = 0.89$, $N = 35$, model df = 7, $P < 0.0001$.