nature neuroscience

Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest

Lusha Zhu[1], Adrianna C Jenkins[2], Eric Set[2,3], Donatella Scabini[4,5], Robert T Knight[4,5], Pearl H Chiu[1,6,7], Brooks King-Casas[1,6–8] & Ming Hsu[2,5]

Substantial correlational evidence suggests that prefrontal regions are critical to honest and dishonest behavior, but causal evidence specifying the nature of this involvement remains absent. We found that lesions of the human dorsolateral prefrontal cortex (DLPFC) decreased the effect of honesty concerns on behavior in economic games that pit honesty motives against self-interest, but did not affect decisions when honesty concerns were absent. These results point to a causal role for DLPFC in honest behavior.

A wealth of field and laboratory studies have shown that humans are often willing to sacrifice their own economic payoffs in the interest of being honest, even in the absence of punishment or reputational factors[1,2]. At the neural level, there is substantial evidence from both neuroimaging[3–6] and developmental[7,8] literatures that the prefrontal cortices, in particular dorsolateral prefrontal (DLPFC) and orbitofrontal (OFC) cortices, are critical to decisions involving honesty. Owing to the inherently correlational nature of such data, however, the specific roles of these regions in honesty and dishonesty remains unclear. Here we sought to characterize the causal contribution of these regions by comparing the behavior of patients with focal lesions to either the DLPFC or OFC to that of healthy comparison participants in a battery of signaling games extensively studied in behavioral economics and evolutionary biology[9,10] (**Fig. 1**, **Supplementary Figs. 1** and **2**, **Supplementary Table 1** and Online Methods). These games capture a core dilemma involved in honest behavior where interests of the signaler conflict with those of the signal receiver, such as that of a seller (signaler) choosing to either truthfully disclose or misrepresent information about a product's quality, which has direct monetary consequences for the buyer (signal receiver).

First, in the 'message' condition, the participant in the role of the signaler can send one of two messages to an anonymous counterpart in the role of the signal recipient, on the basis of which the recipient chooses one of two monetary allocations associated with the messages (**Fig. 2a** and Online Methods)[2,10]. Importantly, both players were instructed that only the signaler would be informed about the monetary consequences associated with each option, and that recipients would never know whether a message they received was true (Online Methods). This highlights the fact that the signal recipient is entirely reliant upon the signaler for potential information about the options and prevents the recipient from using payoff information to make inferences about signaler behavior[2,10].

Second, to account for possible baseline differences in altruistic tendencies, we included a 'choice' condition that contained matching monetary consequences to those in the message condition (Online Methods). The only difference between the conditions was that, in the choice condition, participants directly chose between option A and option B. An individual who is completely insensitive to honesty concerns will behave identically in the two conditions, whereas those sensitive to honesty concerns are predicted to behave more generously in the message condition. All choices were conducted using hypothetical payoffs and no feedback, with order of message and choice blocks counterbalanced across participants within each cohort (Online Methods and **Supplementary Table 2**).

We first investigated how introduction of honesty concerns affected choice behavior in healthy participants by comparing altruistic giving in the message and choice conditions, defined as the amount received by the recipient following implementation of the participant's decision,

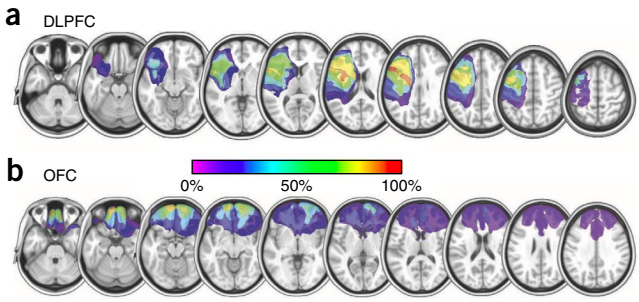

Figure 1 Lesion reconstruction. Structural magnetic resonance imaging slices illustrate the lesion overlap across the two patient groups. (**a**) For the DLPFC group ($n = 6$), mean lesion volume was 125.76 cm^3 and maximal cortical lesion overlap (>50%) was in the Brodmann areas 6, 8, 9 and 46, encompassing portions of the middle and superior frontal gyri in all patients. All dorsolateral prefrontal cortex lesions (5 left, 1 right) were shown overlaid on the left hemisphere for comparison purposes. For lateralized and individual reconstruction, see **Supplementary Figures 1** and **2** and **Supplementary Table 1**. (**b**) For the orbitofrontal cortex group ($n = 7$), mean lesion volume was 72.29 cm^3 and maximal cortical lesion overlap (>50%) was in Brodmann areas 10, 11 and 47, centered in the OFC and including portions of inferior and superior frontal gyri in some patients. See Online Methods for details.

[1]Virginia Tech Carilion Research Institute, Roanoke, Virginia, USA. [2]Haas School of Business, University of California, Berkeley, Berkeley, California, USA. [3]Department of Economics, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. [4]Department of Psychology, University of California, Berkeley, Berkeley, California, USA. [5]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, California, USA. [6]Department of Psychology, Virginia Tech, Blacksburg, Virginia, USA. [7]Department of Psychiatry, Virginia Tech Carilion School of Medicine, Roanoke, Virginia, USA. [8]Virginia Tech-Wake Forest School of Biomedical Engineering and Sciences, Blacksburg, Virginia, USA. Correspondence should be addressed to M.H. (mhsu@haas.berkeley.edu).

Received 3 June; accepted 14 July; published online 31 August 2014; doi:10.1038/nn.3798

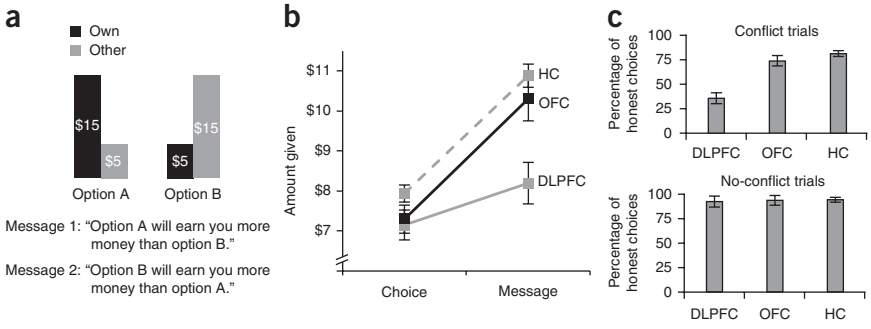

Figure 2 Experimental procedure and behavioral results. (**a**) Experimental procedure. In the message condition, the participant in the role of the signaler is presented with two options, A and B, associated with different monetary consequences. For example, option A corresponds to $15 to the participant and $5 to an anonymous signal recipient—i.e., ($15, $5)—and option B corresponds to ($5, $15). There are two actions available to the participant in the form of two statements describing the monetary consequences of the options to the recipient: the participants must choose between sending a truthful message (message 2) that sacrifices economic self-interest in favor of honesty or a false message (message 1) that satisfies self-interest at the expense of being honest. See Online Methods for details. (**b**) Amount given. In the choice condition, all cohorts gave similar amounts to the recipient (healthy comparison (HC), $7.44 ± 0.22; DLPFC, $6.65 ± 0.38; OFC, $6.79 ± 0.35; Kruskal-Wallis test, $P = 0.15$, two-tailed). In the message condition with identical monetary consequences but with the addition of honesty concerns, the HC cohort increased giving by $2.94 ± 0.44. In contrast, the DLPFC cohort's giving increased by less than half this amount ($1.05 ± 0.43), significantly lower than the increase by the HC cohort (Wilcoxon rank-sum test, $P < 0.001$, two-tailed). OFC participants were nearly identical to HC participants ($3.01 ± 0.55; Wilcoxon rank-sum test, $P = 0.65$, two-tailed), and significantly different from DLPFC participants (Wilcoxon rank-sum test, $P < 0.001$, two-tailed). (**c**) Conflict and no conflict trials. On trials in the message condition where honesty motives conflicted with those of self-interest (top), DLPFC patients made a significantly lower proportion of honest choices (36.7% ± 5.75%) compared to OFC and HC cohorts (OFC patients, 75.7% ± 5.44%; HC, 83.3% ± 3.00%; Fisher's exact test, $P < 0.01$ for both, two-tailed). In contrast, on trials where conflict was absent (bottom), there were no significant differences between cohorts (Fisher's exact test, $P = 0.25$, two-tailed). All error bars indicate s.e.m.

which for simplicity we refer to as "amount given" (Online Methods). Using paired comparisons on decisions with identical monetary consequences, we found that, as consistent with previous studies in healthy participants[1,2,10], inclusion of honesty concerns in the message condition substantially increased altruistic giving compared to the choice condition (Wilcoxon signed-rank test, $P < 0.001$, two-tailed; **Fig. 2b**).

To test the extent to which prefrontal regions are causally involved in trade-offs between honesty concerns and economic self-interest, we next compared amount given between the message and choice conditions in patients with lesions to either DLPFC or OFC versus healthy participants. We found significant main effects of both condition (Wilcoxon signed-rank test, $P < 0.001$, two-tailed), such that participants on average gave more in the message condition ($7.90 ± 0.20) than in the choice condition ($4.14 ± 0.24), and cohort (Kruskal-Wallis test, $P < 0.001$, two-tailed), such that DLPFC patients ($7.17 ± 0.33) on average gave less than healthy participants ($8.91 ± 0.19) and OFC patients ($8.30 ± 0.35). Critically, we observed a significant interaction between condition and cohort (Kruskal-Wallis test on paired difference in amount given across three cohorts, $P < 0.001$, two-tailed), such that damage to DLPFC was associated with significantly lower giving amounts than other cohorts in the message condition but not in the choice condition, suggesting a reduction in the sensitivity to honesty concerns without changes in baseline altruistic tendencies on the part of DLPFC patients (**Fig. 2b** and **Supplementary Figs. 3–5**). All results were robust to using parametric statistical tests. For additional details on the relationship between behavior and demographic variables and lesion laterality, see **Supplementary Figure 3** and **Supplementary Table 3**.

To assess the possibility that deficits in cognitive processes unrelated to honesty may have produced the observed behavioral differences, we first separated decisions in the message condition where honesty and self-interest were in conflict from decisions where the two were aligned (Online Methods). If behavioral patterns observed in DLPFC cohort reflected general impairments such as misunderstanding of payoffs or different beliefs about the behavior of the signal recipients, we would expect DLPFC patients to be affected on both types of decisions. In contrast, we found that DLPFC patients were selectively affected in conflict trials (**Fig. 2c**, top) and were indistinguishable from healthy or OFC cohorts in no-conflict trials (**Fig. 2c**, bottom; **Supplementary Fig. 3b**). In addition, we did not find support for the hypothesis that DLPFC patients exhibited more random choice behavior in the message condition, therefore exerting downward bias on the effect of honesty (**Supplementary Fig. 6**). For additional behavioral results validating task design, see **Supplementary Figures 7** and **8**.

The above results are thus consistent with previous suggestions that DLPFC influences value computations by diminishing subjective value associated with the pursuit of immediate self-interest[11,12]. To formally test this mechanistic hypothesis, we used a computational approach to characterize how parametric variation in costs and benefits associated with honesty influenced choice behavior in our different cohorts. Specifically, we assumed that the subjective value of an option is influenced not only by monetary consequences to self and other but also the means (honest or dishonest) by which these outcomes are obtained (Online Methods and **Supplementary Table 4**)[10]. We found that the weight placed on participants' own payoff decreased in the message condition (α_M) for the OFC and healthy comparison cohorts by approximately 50% relative to the choice condition (α_C; **Fig. 3a**). Strikingly, DLPFC patients' choices did not exhibit a significant discrepancy in the weight across two conditions (**Fig. 3b**), and were significantly different from those of both healthy comparison and OFC cohorts (**Fig. 3b**).

Together, our findings suggest that DLPFC is necessary for promoting honesty concerns over self-interested motives and argue against the widely proposed view that the involvement of prefrontal regions in honesty reflects the need to engage regulatory processes to over-ride truthful responses and implement self-interest[3,13]. Under the latter hypothesis, damage to prefrontal regions should have been associated with an increased sensitivity to honesty concerns, resulting in greater altruistic tendencies when honesty came into conflict with self-interest. Instead, the current results are consistent with the idea that control is necessary to curb self-interest motives in order to communicate the truth, and further suggest that previous neuroimaging findings of DLPFC engagement during dishonest behavior reflect active, but ultimately unsuccessful, engagement of control processes, consistent with observations that individuals with control deficits often engage DLPFC more[14,15].

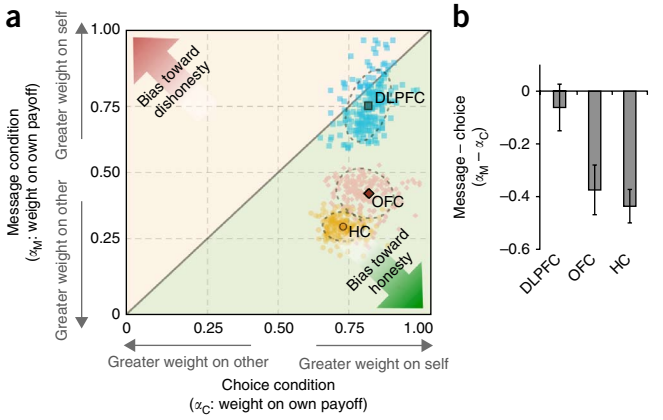

Figure 3 Computational modeling. (**a**) Green shaded region captures willingness to sacrifice one's own payoffs to send the true message—that is, bias toward honesty, where weight on self-interest in the message condition (α_M) is reduced relative to the choice condition (α_C). Conversely, red shaded region captures willingness to sacrifice one's own payoffs to send the false message—that is, bias toward dishonesty, where α_M is greater than α_C. All cohorts placed similar weights on one's own payoff in the choice condition (DLPFC, 0.82 ± 0.05; OFC, 0.79 ± 0.07; healthy comparison (HC), 0.73 ± 0.05). In the message condition, OFC and HC participants showed a significant reduction in weight on own payoff, whereas DLPFC participants did not differ significantly between the two conditions (DLPFC, 0.75 ± 0.09; OFC, 0.43 ± 0.06; HC, 0.29 ± 0.04). Dark points represent parameter estimates and smaller points represent bootstrap pseudo-sample estimates. Dashed ellipses correspond to bootstrapped s.e.m. (**b**) Taking paired-wise differences in pseudo-sample estimates of α_M and α_C, OFC and HC participants showed significantly lower weights on own payoff in the message condition as compared to the choice condition ($P < 0.01$, two-tailed), whereas the DLPFC cohort did not exhibit a significant difference ($P = 0.49$, two-tailed; all error bars indicate bootstrap s.e.m.).

In contrast to the DLPFC, we did not observe an effect of OFC damage on behavior, which might reflect a number of features of our task, including the reduction of anticipated guilt and lack of strong affective components (**Supplementary Fig. 9**)[16,17]. At the same time, we cannot completely rule out possible contributions from non-PFC-based processes to honesty owing to the presence of damage to white matter and in some cases the extension of damage into adjacent regions in our lesion sample (**Fig. 1** and **Supplementary Figs. 1** and **2**). Future studies combining larger lesion cohorts with functional connectivity measures will be needed to address these questions[18]. More broadly, by connecting tools and ideas from behavioral economics and theoretical biology with those of cognitive neuroscience, our study raises exciting questions regarding to what degree the neurocomputational substrates of honesty are shared with other types of norm-guided and moral behavior[19,20], as well as what neural mechanisms arbitrate between such norms in cases of conflict.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Auerbach, Z. Robertson and C. Clayworth for assistance with data collection, analyses and lesion reconstruction. This research was supported by the US National Institutes of Health (R01 MH098023 to M.H., R01 MH087692 to P.H.C., R01 DA036017 to B.K.-C. and R01 NS21135 to R.T.K.), Hellman Family Faculty Fund (M.H.) and the Nielsen Corporation (R.T.K.).

AUTHOR CONTRIBUTIONS

L.Z., E.S., P.H.C., B.K.-C. and M.H. designed the experiments; E.S. and D.S. carried out the experiments; L.Z., A.C.J., E.S., R.T.K. and M.H. carried out statistical analyses; and all authors wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Sally, D. *Rationality Soc.* **7**, 58–92 (1995).
2. Gneezy, U. *Am. Econ. Rev.* **95**, 384–394 (2005).
3. Greene, J.D. & Paxton, J.M. *Proc. Natl. Acad. Sci. USA* **106**, 12506–12511 (2009).
4. Nuñez, J.M., Casey, B., Egner, T., Hare, T. & Hirsch, J. *Neuroimage* **25**, 267–277 (2005).
5. Christ, S.E., Van Essen, D.C., Watson, J.M., Brubaker, L.E. & McDermott, K.B. *Cereb. Cortex* **19**, 1557–1566 (2009).
6. Spence, S.A. *et al. Phil. Trans. R. Soc. Lond. B* **359**, 1755–1762 (2004).
7. Somerville, L.H. & Casey, B. *Curr. Opin. Neurobiol.* **20**, 236–241 (2010).
8. Sodian, B. & Frith, U. *J. Child Psychol. Psychiatry* **33**, 591–605 (1992).
9. Searcy, W.A. & Nowicki, S. *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems* Princenton University Press (2010).
10. Camerer, C. *Behavioral Game Theory: Experiments in Strategic Interaction* Princenton University Press (2003).
11. Figner, B. *et al. Nat. Neurosci.* **13**, 538–539 (2010).
12. Hare, T.A., Camerer, C.F. & Rangel, A. *Science* **324**, 646–648 (2009).
13. Sip, K.E., Roepstorff, A., McGregor, W. & Frith, C.D. *Trends Cogn. Sci.* **12**, 48–53 (2008).
14. Rosano, C. *et al. Biol. Psychiatry* **57**, 761–767 (2005).
15. Tapert, S.F. *et al. Psychopharmacology (Berl.)* **194**, 173–183 (2007).
16. Koenigs, M. *et al. Nature* **446**, 908–911 (2007).
17. Krajbich, I., Adolphs, R., Tranel, D., Denburg, N. & Camerer, C.F. *J. Neurosci.* **29**, 2188–2192 (2009).
18. He, B.J. *et al. Neuron* **53**, 905–918 (2007).
19. Greene, J.D., Sommerville, R., Nystrom, L.E., Darley, J.M. & Cohen, J. *Science* **293**, 2105–2108 (2001).
20. Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. *Science* **314**, 829–832 (2006).

ONLINE METHODS

Subjects. Patients with focal brain lesions to the dorsolateral prefrontal cortex ($n = 7$) and orbitofrontal cortex ($n = 7$) were included in the experiment (see **Supplementary Table 1** for details). Healthy comparison participants ($n = 27$) were recruited from the San Francisco Bay area, California. All subjects provided informed consent approved by the University of California, Berkeley, California. One DLPFC lesion patient answered incorrectly on more than 50% of post-instruction questionnaires and was excluded from the study. In comparison, all other subjects answered at least 90% of the questions correctly. All statistical results reported in the study were robust to inclusion of this participant.

Lesion reconstruction. Software reconstructions were performed using MRIcron[21]. For both patient groups, testing took place at least 6 months after the date of the stroke or accident. A neurologist (R.T.K.) inspected patient MRIs to ensure that no white matter hyperintensities outside the lesioned area were observed in either patient group. All traumatic brain injury patients had low-impact force injuries with no clinical or MRI evidence of axonal shear.

Signaling games. We used a battery of signaling games extensively studied in behavioral economics and evolutionary biology[9,10]. These games capture a core dilemma involved in honest behavior where interests of the signaler conflict with those of the signal receiver, such as that of a seller (signaler) choosing to either truthfully disclose or misrepresent information about a product's quality, which has direct monetary consequences for the buyer (signal receiver).

These games have three important advantages as an assay of decisions involving tradeoffs between honesty and self-interest. First, to isolate the effects of honesty, we included a set of message and choice conditions. Because the latter condition does not include honesty concerns, we remove the tension between honesty and other social preferences and are able to control for participants' concerns for equity and efficiency. As a result, systematic deviations in behavior between the two sets of games can be interpreted as being affected by honesty concerns. Specifically, an individual who is completely insensitive to honesty concerns will behave identically in the two conditions, whereas those sensitive to honesty concerns are predicted to behave more generously in the message condition. In previous experiments using these games, introduction of honesty concerns in the message condition has been found to increase cooperation rates and altruistic giving by approximately 50% (refs. 1,2,10,22).

Second, the clearly delineated cost-benefit relationship associated with self-interest and honesty facilitates a computational account of honesty, which allows us to better connect the potential behavioral differences to their computational substrates. Finally, and importantly in the context of lesion studies, by explicitly presenting honest and self-interested actions to subjects, the message condition allows us to hold constant the available action set across cohorts and verify understanding. This included both comprehension tests and control trials with no conflict between honesty and self-interest.

Message and choice conditions. In the message condition, the participant in the role of signaler was presented with two options, A and B, which yielded different monetary outcomes. For example, in **Supplementary Figure 10**, option A corresponded to \$6 to the signaler and \$5 to an anonymous random signal recipient—that is, (\$6, \$5)—and option B corresponded to (\$5, \$10). Only the signaler knew the payoffs associated with the options, and the signaler had to send either an honest or dishonest message to an anonymous recipient. The recipient did not know the associated payoffs but had to choose one of the two options. That is, the signaler could either choose to convey the truth, "Option B will earn you more money than option A," or a falsehood, "Option A will earn you more money than option B". Importantly, all signalers were informed that recipients would never know the payment information associated with each option and therefore whether senders' messages were true or not.

The monetary outcomes varied across trials. In particular, in some trials we pitted self-interest against honesty. That is, honest choices were associated with allocations that yielded less payment to the participant and more to the recipient (for example, \$5 for self, \$15 for other in option A; versus \$6 for self, \$5 for other in option B). We refer to these trials as "conflict trials." In "no-conflict trials," honest choices were associated with allocations that yielded more payment to both participant and recipient (for example, \$8 for self, \$10 for other in option A; versus \$10 for self, \$12 for other in option B). A full list of trial options is presented in **Supplementary Table 2**.

As a control condition, we also included the choice condition associated with the same set of payoff allocations. In particular, participants were asked to directly choose either option A or option B. Following the procedure of previous experiments using the message and choice condition[2], participants were informed that in the choice condition (i) their decisions would be implemented 80% of the time, while the other 20% of the time the alternative option would be implemented; and (ii) receivers would not know the monetary payoff associated with each option and would just receive money passively.

Procedure. Following task instructions and a comprehension quiz, participants were administered two blocks of message and choice condition trials, each containing 12 trials. All choices were conducted using hypothetical payoffs and no feedback, with order of message and choice blocks counterbalanced across participants within each cohort. Within each block, questions were presented in a random order. For complete experimental instructions, see http://neuroecon.berkeley.edu/papers.html.

Behavioral analysis. In both conditions, the behavioral measure of altruistic giving was defined as the amount that would be received by the recipient if the participant's decision was implemented, which for simplicity we refer to as "amount given." Using payoffs given in **Supplementary Figure 10** as an example, the amount given in the message condition by a participant choosing the truthful (false) message 2 (1) would be defined as \$10 (\$5). Similarly, in the choice condition, the amount given by a participant choosing option A (B) would be defined as \$5 (\$10).

Computational modeling. To characterize the relative contributions of economic self-interest, distributional preference and honesty consideration to allocation decisions, we adapted an economic model that was previously applied to study social preferences[23] to our tasks.

First, denote M_s and M_o as monetary payoffs for self and other respectively. The indicator function I is equal to 1 when the monetary payoff is achieved through dishonesty and 0 otherwise. That is, I indicates whether honesty concerns are over-ridden. We propose that the decision-maker's utility is modulated by honesty in addition to monetary allocations to self and other

$$U(M_s, M_o) = \left[(\alpha - I \cdot \delta) M_s^{\rho} + (1 - \alpha + I \cdot \delta) M_o^{\rho}\right]^{\frac{1}{\rho}}$$

Here α and ρ are parameters capturing distributional preferences that solely depend upon the monetary allocation between self and other, whereas δ quantifies the biasing effects of honesty concerns. The functional form follows the well-established Constant Elasticity of Substitution utility function[24]. Specifically, the parameter α quantifies the relative weight between monetary payoffs for self and other. A large α indicates a larger weight on one's own economic gain. The parameter ρ reflects the elasticity of substitution between M_s and M_o. For example, if ρ approaches 1, the utility function will reduce to a linear function representing the preference of welfare maximizing. If ρ approaches negative infinity, the utility function will reduce to $U(M_s, M_o) = \min(M_s, M_o)$, which corresponds to the preference of maximal inequity aversion.

In the context of our game, we refer to α as the weight placed on own payoff in the choice condition, as there is no tradeoff between self-interest and honesty. That is, $\alpha_C = \alpha$. In contrast, the weight placed on one's own payoff in the message condition is defined by $\alpha_C = \alpha - \delta$. Critically, the parameter δ can be interpreted as the degree to which honesty reduces self-interested motives. If $\delta > 0$, the signaler suffers from a disutility of deception and is more likely to sacrifice self-interest in favor of honesty concerns. In contrast, if $\delta < 0$, the signaler receives an additional utility from dishonesty, and thus is more likely to choose dishonest options. Finally, if $\delta = 0$, the signaler is indifferent between honest or dishonest actions and will behave as if the tradeoff between honesty and dishonesty does not exist. The combination of these parameters thus nests a wide range of social preferences proposed by existing theory and allows for rich interactions among economic self-interest, distributional preferences and honesty considerations.

To calibrate the model given the binary choice behavior of each cohort in the game, we adopted the standard logit assumption, aggregated observations

doi:10.1038/nn.3798

conditional on lesion cohorts and experimental conditions and conducted maximal likelihood estimation, specifically maximizing the log likelihood function $\sum_i \sum_t \log(P_{i,t}(y_{it};\alpha_M,\alpha_C,\rho))$. The standard errors of estimated parameters were obtained through the bootstrap procedure with 200 iterations for each cohort.

A **Supplementary Methods Checklist** is available. See full version of experimental instructions at http://neuroecon.berkeley.edu/papers.html.

21. Rorden, C. & Brett, M. *Behav. Neurol.* **12**, 191–200 (2000).
22. Crawford, V. *J. Econ. Theory* **78**, 286–298 (1998).
23. Charness, G. & Rabin, M. *Q. J. Econ.* **117**, 817–869 (2002).
24. Andreoni, J. & Miller, J. *Econometrica* **70**, 737–753 (2002).