Title: The Brain Adapts to Dishonesty

Authors: Neil Garrett[1], Stephanie C. Lazzaro[1], Dan Ariely[2], Tali Sharot[1*]

Affiliations:

[1]Affective Brain Lab, Department of Experimental Psychology, University College London, United Kingdom

[2]Fuqua School of Business, Duke University, North Carolina, United States of America

*Correspondence to: t.sharot@ucl.ac.uk

Abstract:

Dishonesty is an integral part of our social world, influencing domains ranging from finance and politics to personal relationships. Anecdotally, digressions from a moral code are often described as a series of small breaches that grow over time. Here, we provide empirical evidence for a gradual escalation of self-serving dishonesty and reveal a neural mechanism supporting it. Behaviorally, we show that the extent to which participants engage in self-serving dishonesty increases with repetition. Using fMRI we show that signal reduction in the amygdala is sensitive to the history of dishonest behavior, consistent with adaptation. Critically, the extent of amygdala BOLD reduction to dishonesty on a present decision relative to the last, predicts the magnitude of escalation of self-serving dishonesty on the next decision. The findings uncover a biological mechanism that supports a “slippery slope”: what begins as small acts of dishonesty can escalate into larger instances.

Many dishonest acts are speculatively traced back to a sequence of smaller transgressions that gradually escalated. From financial fraud, to plagiarism, online scams and scientific misconduct, deceivers retrospectively describe how minor dishonest decisions snowballed into significant ones over time[1–4]. Despite the dramatic impact of these acts on economics[5,6], policy[7] and education[8], we do not have a clear understanding of how and why small transgressions may gradually lead to larger ones. Here, we set out to empirically demonstrate dishonesty escalation in a controlled laboratory setting and examine the underlying mechanism.

People often perceive self-serving dishonesty as morally wrong[9] and report uneasiness when engaging in such behavior[10]. Consistent with these reports, physiological[11] and neurological[12] measures of emotional arousal are observed when people deceive. Blocking such signals pharmacologically results in significant increases in dishonesty. For example, in one study students who had taken and responded to a mild sympatholytic agent were twice as likely to cheat on an exam than those who took a placebo[13]. Thus, in the absence of an affective signal that can help curb dishonesty, people may engage in more frequent and severe acts.

A large body of research demonstrates that the response to an emotion evoking stimulus weakens with repeated exposure[14,15]. For example, both affective ratings of negative images and amygdala activation have been shown to decrease with each subsequent presentation of those images[16]. It is thus possible that the affective signal that accompanies self-serving dishonesty also diminishes with repetition. If indeed signals that may help curb dishonesty are diminished over time, dishonest acts could increase. Thus, what begins as small deviations from a moral code could escalate to large deviations with potentially harmful consequences.

To test for dishonesty escalation and its underlying neurological mechanism, we combined brain imaging with a behavioral task in which individuals were given repeated opportunities to act dishonestly. We used Neurosynth, a platform for large-scale automated synthesis of thousands of functional magnetic resonance imaging (fMRI) studies[17], to identify voxels within the brain that have been previously associated with emotion. Unsurprisingly, this map predominantly, but not exclusively, identified the amygdala. The amygdala responds to various emotion-eliciting situations, with stronger activity related to a more potent experience (e.g.[18–22]) and research in non-human animals also demonstrates that the amygdala is critical for emotion (e.g.[23]). Our goal was to examine whether dishonesty escalates over time, whether response to dishonesty in the region identified above decreases over time, and whether the extent of decrease in the latter predicts the extent of escalation of the former.

Testing realistic deception in a brain-imaging scanner is notoriously difficult, as participants need to repeatedly, deliberately and voluntarily act dishonestly in a social context without being required to admit to their dishonesty. Our paradigm enables participants to do just that, whilst also allowing us to quantify deception on a trial-by-trial basis. Specifically, we adapted a two party task used previously to elicit and measure dishonesty[24]. Participants advised a second participant played by a confederate, about the amount of money in a glass jar filled with pennies (**Figure 1a**). We changed the incentive structure over the course of two experiments such that dishonesty about the amount of money in the jar would either benefit the participant at the expense of their partner ("Self-Serving-Other-Harming" condition, Experiment 1), benefit both ("Self-Serving-Other-Serving" condition, Experiment 1), benefit the partner at the expense of the participant ("Self-Harming-Other-Serving" condition, Experiment 1), benefit the participant only without affecting the partner (Self-Serving condition, Experiment 2), or benefit the partner only without affecting the participant (Other-Serving condition, Experiment 2). Importantly, the participants believed that their partner was not aware of this incentive structure, but thought that they were working together at all times to provide the most accurate estimate, which would benefit them both equally. A "Baseline condition" enabled us to infer the amount of dishonesty on each trial without the participant being instructed to, or admitting to, dishonesty. Specifically, in this condition, neither the participant nor the partner would benefit from the participant's dishonesty. Here, honest, accurate, advice was the best policy to maximize rewards for all (see Online Methods for details). Because the different jars were repeated under all conditions, the Baseline condition enabled us to estimate how much money the participant thought there was in a specific jar and thus the approximate amount by which a participant was dishonest in the other conditions.

We designed an incentive structure such that adaptation to dishonesty could not be explained by adaptation to reward. Specifically, participants were told that the available reward would vary on each trial. Thus, while on a specific trial a larger magnitude of dishonesty was more likely to result in greater gains, greater dishonesty on that trial relative to the previous one would not necessarily correspond to greater returns (see **Online Methods**). Moreover, neither the amount available, nor the outcome of the trial, would be disclosed at any point during the experiment. In other words, the participants were informed in advance that no feedback would be provided. In addition, participants were told that rewards would not accumulate, but rather one trial would be selected at random at the end of the experiment and the participant would be paid according to that trial. We went to great length to explain the reward structure to the participants and to ensure their comprehension by using a quiz with example trials prior to the beginning of each block of trials.

Results

Self-Serving Dishonesty Escalates. We observed clear evidence of escalation in self-serving dishonesty, such that the magnitude of dishonesty got larger and larger over the course of a block. For each participant (N=55) we regressed dishonesty on trial number and examined standardized regression coefficients at the group level. Entering these scores into a repeated measures ANCOVA (controlling for initial levels of dishonesty and whether the participant was completing the study inside or outside the scanner) revealed a significant effect of condition ($F_{2,49} = 4.65$, $p = 0.014$, **Figure 1b**). This was due to dishonesty increasing over time if it was self-serving (Self-Serving-Other-Harming: $t_{52} = 3.38$, $p = 0.001$; Self-Serving-Other-Serving: $t_{52} = 2.25$, $p = 0.03$), but not otherwise (Self-Harming-Other-Serving: $t_{52} = -0.27$, $p=0.79$). Escalation was significantly greater when dishonesty was self-serving than self-harming (Self-Serving-Other-Harming compared to Self-Harming-Other-Serving: $F_{1,51} = 8.80$, $p = 0.005$; Self-Serving-Other-Serving compared to Self-Harming-Other-Serving: $F_{1,51} = 4.61$, $p = 0.037$).

We further demonstrate this escalation by averaging the amount of dishonesty on each trial (N=60) across participants (N=55) and correlating with trial number. This correlated positively when dishonesty was self-serving (Self-Serving-Other-Harming condition: $r_{58} = 0.66$, $p<0.001$, **Figure 1c**; Self-Serving-Other-Serving condition: $r_{58} = 0.83$, $p<0.001$, **Figure 1d**). In contrast, when dishonesty was self-harming and accuracy was in fact the self-serving choice, a trend in the opposite direction was observed over the course of trials (Self-Harming-Other-Serving: $r_{58} = -0.23$, $p=0.08$, **Figure 1e**), which was significantly smaller than the escalations of self-serving dishonesty (compared to Self-Serving-Other-Harming: $z = 4.83$, $p<0.001$; compared to Self-Serving-Other-Serving: $z = 7.13$, $p<0.001$).

Dishonesty Magnitude. While the focus of this investigation is *escalation of dishonesty* over time rather than its *average magnitude*, for completeness we report the latter here. Entering participants' dishonesty magnitude into a one way repeated measures ANOVA (controlling for study) revealed a significant effect of condition ($F_{2,52} = 24.17$, $p=0.0001$, **Supplementary Figure 1a**). Participants' dishonesty was greater when dishonesty was self-serving (mean Self-Serving-Other-Harming = 7.11, s.d. = 9.61; mean Self-Serving-Other-Serving condition = 12.84, s.d. = 15.46) than when it was self-harming (mean Self-Harming-Other-Serving condition = -0.16, s.d. = 2.99, respective comparisons: $F_{1,53} = 26.44$, $p<0.001$; $F_{1,53} = 39.67$ $p<0.0001$). Dishonesty magnitude was also greater when dishonesty served the self and the other relative to when it only served the self ($F_{1,53} = 7.72$, $p=0.008$). Entering participants' starting dishonesty into a one way repeated measures ANOVA (controlling for study)

revealed a significant effect of condition ($F_{2,52}$ = 12.11, p<0.001, **Supplementary Figure 1b**). Participants' starting dishonesty was greater when dishonesty was self-serving than when it was self-harming (Self-Serving-Other-Harming vs Self-Harming-Other-Serving condition: $F_{1,53}$ = 7.74, p=0.007; Self-Serving-Other-Serving vs Self-Harming-Other-Serving condition: $F_{1,53}$ = 21.99, p<0.001). Starting dishonesty was also greater when dishonesty served the self and the other compared to when it only served the self ($F_{1,53}$ = 8.84, p=0.004). In all conditions, participants' average estimates (Self-Serving-Other-Harming: t_{54} = -43.91, p<0.001; Self-Serving-Other-Serving: t_{54} = -24.54, p<0.001; Self-Harming-Other-Serving t_{54} = -158.97, p<0.001, one sample ttests) and last estimates (Self-Serving-Other-Harming: t_{54} = -37.58, p<0.001; Self-Serving-Other-Serving: t_{54} = -21.37, p<0.001; Self-Harming-Other-Serving t_{54} = -120.34, p<0.001, one sample ttests) were significantly lower than ceiling. Thus, participants always had an opportunity to lie more than they actually did if they wanted to.

Dishonesty, but not its escalation, is evident when it benefits the partner without affecting the self. In the Self-Harming-Other-Serving condition, participants on average did not act dishonestly, presumably because dishonesty in this condition would hurt their pay. It is thus feasible that escalation is not evident in this condition simply because dishonesty is nonexistent. To examine this possibility and tease apart the contribution of self-interest and other-interest in escalating dishonesty, we conducted a follow up study (N=25) in which we had two new conditions in addition to the baseline condition. In the condition of interest, dishonesty only benefitted the self without affecting the other participant (Self-Serving); in the comparison condition dishonesty would only benefit the partner without benefiting or hurting the self (Other-Serving). Each of these two conditions were run twice in two separate blocks, thus creating four counterbalanced blocks in total: in one block over-estimating the amount of money in the jar would benefit the participant only, in another it would benefit the partner only, in a third block underestimating the amount of money would benefit the participant only, and in a fourth the partner only. We added the "underestimation" blocks in this follow up study so that the results could be generalized beyond one specific set of instructions.

While initial levels of dishonesty did not differ between conditions ($F_{1,24}$ = 1.24, p = 0.28, **Figure 2a**), escalation of dishonesty did (**Figure 2c**). Specifically, entering escalation betas into a 2 (condition: Self-Serving/Other-Serving) by 2 (underestimation/overestimation) ANCOVA (controlling for starting dishonesty) revealed a main effect of condition ($F_{1,20}$ = 7.55, p=0.01) which was characterized by the presence of escalation in the Self-Serving condition (overestimate: t_{23} = 2.66, p=0.01; underestimate: t_{23} = 2.66, p=0.01) but not the Other-Serving condition (overestimate: t_{23} =0.85, p=0.40; underestimate: t_{23} = 0.56, p=0.58). There was no main effect of overestimating/underestimating ($F_{1,20}$ = 0.08, p=0.78),

nor an interaction ($F_{1,20}$ = 0.26, p=0.62). Mean dishonesty tended to be larger in the self-serving condition than the other-serving condition ($F_{1,24}$ = 3.50, p=0.07, **Figure 2b**). Escalation of self-serving dishonesty was also significantly greater than of other-serving dishonesty when mean dishonesty was added as a covariate to the ANCOVA described above (main effect of condition; $F_{1,16}$ = 5.47, p=0.01).

These results suggest that the rate at which dishonesty escalates is best explained by self-interest. Furthermore, escalation of dishonesty cannot be attributed to factors such as reduced/enhanced attention or tiredness over the course of the block, nor to a simple tendency to give larger estimates (i.e. to overestimate) over time, or smaller estimates (i.e. to underestimate) over time, as we do not see an escalation of dishonesty when dishonesty benefits only the partner. While 45% of participants reported in the debriefing questionnaire that they believed the aim of the study was to examine dishonesty/trust (**Supplementary Table 1**), there were no differences in behavior between these participants and the rest in mean dishonesty (Experiment 1: Self-Serving-Other-Harming: t_{53} = 0.41, p=0.69; Self-Serving-Other-Serving: t_{53} = 1.02, p=0.31; Self-Harming-Other-Serving: t_{53} = -1.07, p=0.29, independent sample ttest; Experiment 2: Self-Serving: t_{23} = 0.76, p=0.46; Other-Serving: t_{23} = 1.71, p=0.10, independent sample ttest) or dishonesty escalation (Experiment 1:Self-Serving-Other-Harming: t_{53} = -0.65, p=0.52; Self-Serving-Other-Serving: t_{53} = 0.61, p=0.54; Self-Harming-Other-Serving: t_{53} = 0.91, p=0.37, independent sample ttest; Experiment 2: Self Serving: t_{23} = 1.68, p=0.11; t_{23} = 0.93, p=0.36, independent sample ttest). No participant reported they believed the aim was to examine escalation of dishonesty.

Amygdala Sensitivity and Escalation of Self-Serving Dishonesty. Our behavioral findings reveal that self-serving dishonesty escalates over the course of the block. Next, we turned to our fMRI data to ask (1) whether the amygdala's response to dishonesty diminishes over time and (2) whether the extent of this reduction predicts, on a trial by trial basis, subsequent escalation of dishonesty.

1. Amygdala response to dishonesty is reduced over time. Past studies of adaptation examined BOLD reduction in response to a repeated *constant* variable such as an angry face[14,15]. In contrast, here we examine adaptation in response to an escalating noisy variable, analogous to a face that tends to get angrier over time. Thus a signature of adaptation would not simply index reduced brain activity over time. Rather, it should scale simultaneously with trial number and dishonesty magnitude. The intuition here is that the same amount of dishonesty should generate a larger BOLD response if it occurs early in the block than when it occurs late (i.e. after adaptation has taken place). Thus, we constructed a time weighted dishonesty regressor that modulated BOLD signal when participants were presented

with the jar. Note, that we use the word “time” to refer to number of past trials in a block, not absolute time elapsed. This regressor scaled dishonesty by how far through the block participants were such that dishonesty early on in the block received a larger weight relative to if it occurred later on. We controlled for the independent effects of time and dishonesty by entering them as two separate parametric regressors in the model (see **Online Methods**).

As the behavioral results demonstrate that dishonesty is driven both by considerations for self and other, but its escalation is driven only by whether dishonesty benefits or hurts the self, we focused our fMRI analysis on the two cases where the benefit of dishonesty was confined to either the self or the other (i.e. Self-Serving-Other-Harming condition and the Self-Harming-Other-Serving condition). Our ROI was defined by generating a map in Neurosynth, a meta-analysis based on 11,406 studies, reflecting P(Emotion|Activation) thresholded at z = 5 (note that the results are robust and observed for maps generated by using different thresholds). This map corresponds to the likelihood that the term “emotion” is used in a study given the presence of reported activation in a particular voxel. The map reflects the relative selectivity with which voxels activate in relation to the term “emotion” using a comparison between all the studies in the database that contain the term “emotion” and all those that do not (for method details see[17]). The ROI generated was predominately, but not exclusively, confined to bilateral amygdala (**Figure 3a**). Parameter estimates (N=25) of time-weighted-dishonesty were then averaged across all voxels in this ROI. This revealed a significant positive effect when dishonesty was self-serving (t_{24} = 2.36, p = 0.027, one sample ttest vs 0), but not when it was self-harming (trend in opposite direction: t_{24} = -1.93, p=0.066, one sample ttest vs 0) with the former parameter betas significantly larger than the latter (t_{24} = 2.96, p=0.007, paired sample ttest). (**Figure 3b).** Repeating this analysis only on those ROI voxels that were anatomically within the amygdala also generated significant positive results (self-serving: t_{24} = 2.68, p=0.01; self-harming: t_{24} = -0.9, p=0.38; self-serving vs self-harming: t_{24} = 2.60, p=0.02, **Supplementary Figure 2a**). Note, that simply looking at average BOLD signal within the ROI over all trials (without taking into account the effect of time weighted dishonesty) did not reveal positive effects in either of the conditions anywhere in the ROI.

The effects were not due to BOLD signal simply decreasing over time since the model controls for the independent effect of time by including this as the first regressor in the model. Furthermore, examining the effects of the time regressor (rather than time-weighted-dishonesty) did not reveal significant effects in the ROI (self-serving: t_{24} = 0.08, p=0.94, one sample ttest vs 0; self-harming: t_{24} = 1.31, p = 0.20, one sample ttest vs 0; self-serving vs self-harming: t_{24} = -0.85, p=0.40, paired sample ttest). The fact that BOLD signal in the ROI did not simply decrease over time, as well as the fact that our effects remains

significant after controlling for time, suggests that the findings are unlikely to be explained by a reduction of attention/engagement over time. Together, these results show BOLD response consistent with adaptation to self-serving dishonesty in regions that have been previously associated with emotion.

2. Amygdala BOLD Reduction to Self-Serving Dishonesty Predicts its Escalation. Next, we ask whether reduction in the neural response to dishonesty in our ROI predicts the escalation of dishonesty. More precisely, whether the reduction of response in the ROI to one "unit" of dishonesty on a present trial relative to the last predicts escalation of dishonest behavior on the following trial (**Figure 4a**).

To that end, we first computed BOLD signal per £1 of dishonesty (BOLD/dishonesty) for each subject and trial. We then calculated the reduction of this term on trial (t) relative to the previous trial (t-1), (i.e., $BOLD_{t-1}/dishonesty_{t-1}$ - $BOLD_t/dishonesty_t$). We then separately calculated the increase in behavioral dishonesty on the subsequent trial (t+1) relative to trial t. For each subject we then used BOLD signal reduction per one unit of dishonesty on trial t as a predictor of escalation of dishonesty on trial t+1 in a general linear regression (**Figure 4b** shows an example subject).

Across participants (N=25), these betas revealed a significant positive effect when dishonesty was self-serving ($t_{24} = 2.48$, p=0.021, one sample ttest vs 0), but not when it was self-harming ($t_{24} = -1.53$, p=0.14, one sample ttest vs 0) with the former betas larger than the latter ($t_{24} = 2.82$, p=0.01, paired sample ttest), **Figure 4c**. In other words, the greater the reduction in BOLD response to one unit of self-serving dishonesty on a current trial relative to the last, the greater a participant's next self-serving lie. When repeating this analysis only on those voxels that are also in the anatomically defined amygdala, the results are very similar (self-serving: $t_{24} = 2.30$, p=0.03; self-harming: trend in opposite direction $t_{24} = -1.77$, p=0.09; self-serving vs self-harming: $t_{24} = 3.05$, p=0.01, **Supplementary Figure 2b**).

Permutation Test. As a check for the robustness of this result we reran this analysis 1000 times, each time mismatching one participant's fMRI data with another participant's behavioral data. Over 95% of these iterations failed to generate a main effect of self-serving dishonesty, a main effect of other-serving dishonesty or a significant difference between self-serving and self-harming dishonesty. This suggests that BOLD reduction to dishonesty in a particular subject selectively predicts the escalation of dishonesty in that subject and not another subject.

Other Brain Regions. Finally, we tested whether the pattern of results observed above generalize to additional brain regions that play a role in other aspects of deception. This was done post hoc. First we

looked at the Nucleus Accumbens (NA), which is the region most strongly implicated in response to positive stimuli[17,25,26]. As detailed in **Supplementary Table 2** there was no evidence for NA BOLD adaptation to self-serving dishonesty in this task. Next, we examined the Dorsolateral Prefrontal Cortex (DLPFC), which is thought to be associated with cognitive control[27,28] and has previously been observed in dishonesty tasks[29,30]. Again, no evidence for BOLD adaptation to self-serving dishonesty was observed (see **Supplementary Table 2**). Finally, we examined the Anterior Insula (AI), which is suggested to play a role in dishonesty[31,32] and negative affect[33], though less consistently and robustly than the amygdala[17]. As detailed in **Supplementary Table 2** there was indication of BOLD signal adaption to dishonesty in the AI, although not as robust and consistent as that observed in the amygdala. Finally, exploratory analysis in regions outside the ROIs (threshold at FWE<0.05, K>5) did not reveal additional effects of time weighted dishonesty.

Discussion

Dishonesty significantly impacts our personal lives[9] and public institutions[34]. Here, we provide empirical evidence that dishonesty gradually increases with repetition when all else is held constant (see[35] for dishonesty escalation in response to escalating rewards). This experimental result is consistent with anecdotal observations of small digressions gradually snowballing into larger ones[1–4,36]. Our results also offer a mechanistic account of how dishonesty escalates, showing that it is supported by BOLD signal reduction in brain regions previously associated with emotion, predominantly the amygdala. Across individuals the extent of BOLD signal reduction per "one unit" of dishonesty on a present decision relative to the last, predicted escalation of dishonesty on the next decision. This result ties together diminished amygdala sensitivity to dishonest choices and dishonesty escalation, with the former preceding the latter.

We focused our investigation a-priori on brain regions previously associated with emotion by utilizing Neurosynth[17]. The map generated predominantly identified the amygdala, consistent with studies in non-human animals, as well as human patients with damage to the amygdala, which have shown this region to play a key role in signaling, processing and assessing arousal and emotion[17,37,38]. With deference to the caution necessary when making reverse inference[39], we speculate that the blunted response to repeated acts of dishonesty may reflect a reduction in the emotional response to these decisions or their affective assessment and saliency. While the amygdala can also signal positive emotion, this interpretation is less likely as the region most associated with positive affect, the nucleus accumbens, did not reveal a similar pattern of results. The current interpretation is in accord with a previous suggestion that the amygdala signals the rate of averseness to immoral acts[40].

Our results also suggest that dishonesty escalation is contingent on the motivation for the dishonest act. Specifically, while the magnitude of dishonesty was driven both by considerations of benefit to the self and benefit to the other, the escalation of dishonesty, as well as the amygdala's response to it over time, was best accounted for by whether dishonesty was self-serving. When participants were dishonest for the benefit of someone else, dishonesty at a constant rate was observed. This is consistent with the suggestion that the motivation for acting dishonestly contributes to its affective assessment, such that when a person engages in dishonesty purely for the benefit of another it may be perceived as morally acceptable[41]. Thus, the simple act of repeated dishonesty is not enough for escalation to take place, but a self-interest motive need be present.

While here we observe escalation of decisions that involve self-serving dishonesty, we hypothesize that diminished amygdala response may have important behavioral consequences extending to other domains of decision-making. These may include, for example, escalations in risk taking or violent behavior. Our results highlight the importance of considering the temporal evolution of both behavior and brain response in studies that involve repeated decisions.

As we were specifically interested in the effect of repeated dishonesty on its magnitude, we varied whether dishonesty was beneficial/detrimental to the participant and to the target of their deceit (the partner), but held everything else constant. For example, in our design no feedback, such as external punishment for dishonesty or praise for honesty, was provided or expected. In the real world, however, other factors will likely facilitate dishonesty escalation or halt it, including feedback, opportunity and altering rewards and punishment [42,35,43].

Taken together, we reveal a biological mechanism that underlies the escalation of dishonesty, providing new insight into this integral part of human behavior. The results show the possible dangers of regular engagement in small acts of dishonesty, perils that are frequently observed in domains ranging from business to politics and law enforcement[3,34]. These insights can have implication for policy makers in designing deterrents to halt deceit. Despite being small at the outset, engagement in dishonest acts may trigger a process that leads to larger acts of dishonesty further down the line.

Data Availability: The data that support the findings of this study are available from the corresponding author upon request.

Acknowledgments: We thank D.Prelec, B. Baharami, U.Hertz, J.Navajas, D.Bang for helpful discussion; T.Yarkoni, C.Frith and W.Penny for helpful pointers; R.Rutledge, C.Summerfield, M.Cikara, M.Edelson, R.Köster, A.Kappes, C.Charpentier, S.Suarez, L.Coutrot, L.Wittkuhn and P.Czech for helpful comments on previous versions of this manuscript; T.Srirangarajan, R.Anjum, S.Hadden, G.Montinola and M.Wilner for assistance with data collection and scanning; T.Sharot is supported by a Wellcome Trust Career Development Fellowship, N.Garrett by a UCL Impact Award, the research was also supported by funding from the Institute for Advance Hindsight.

Author Contributions: T.S. conceived the study. N.G., .S.C.L., D.A. & T.S. designed the study. N.G. collected behavioral and fMRI data. N.G. & T.S. analyzed the data. N.G. & T.S. wrote the manuscript with edits from S.C.L.

Competing Financial Interests: The authors declare no financial competing interests.

1. Maremont, M. Anatomy of a Fraud. *Bus. Week* **16,** 90–94 (1996).
2. Kirchner, B. *The Bernard Madoff Investment Scam*. (FT Press, 2010).
3. McLean, B. & Elkind, P. *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron*. (Penguin, 2013).
4. Stapel, D. *Ontsporing*. (Prometheus Amsterdam, 2012).
5. Graham, C., Litan, R. E. & Sukhtankar, S. The Bigger They Are, The Harder They Fall: An Estimate of the Costs of the Crisis in Corporate Governance. *The Brookings Institution* (2002). Available at: http://www.brookings.edu/research/papers/2002/07/22corporategovernance-graham.
6. Mauro, P. Corruption and growth. *Q. J. Econ.* 681–712 (1995).
7. Tanzi, V. & Davoodi, H. *Corruption, public investment, and growth*. (Springer, 1998).
8. Heyneman, S. P., Anderson, K. H. & Nuraliyeva, N. The cost of corruption in higher education. *Comp. Educ. Rev.* **52,** 1–25 (2008).
9. Peterson, C. Deception in Intimate Relationships. *Int. J. Psychol.* **31,** 279–288 (1996).
10. DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M. & Epstein, J. A. Lying in everyday life. *J. Pers. Soc. Psychol.* **70,** 979–995 (1996).
11. Gamer, M., Rill, H.-G., Vossel, G. & Gödert, H. W. Psychophysiological and vocal measures in the detection of guilty knowledge. *Int. J. Psychophysiol.* **60,** 76–87 (2006).

12. Abe, N., Suzuki, M., Mori, E., Itoh, M. & Fujii, T. Deceiving Others: Distinct Neural Responses of the Prefrontal Cortex and Amygdala in Simple Fabrication and Deception with Social Interactions. *J. Cogn. Neurosci.* **19,** 287–295 (2007).

13. Schachter, S. & Latané, B. Crime, cognition, and the autonomic nervous system. In D. Levine (Ed.),. *Nebr. Symp. Motiv.* **12,** 221–275 (1964).

14. Breiter, H. C. *et al.* Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17,** 875–887 (1996).

15. Ishai, A., Pessoa, L., Bikle, P. C. & Ungerleider, L. G. Repetition suppression of faces is modulated by emotion. *Proc. Natl. Acad. Sci.* **101,** 9827–9832 (2004).

16. Denny, B. T. *et al.* Insula-amygdala functional connectivity is correlated with habituation to repeated negative images. *Soc. Cogn. Affect. Neurosci.* **9,** 1660–1667 (2014).

17. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8,** 665–670 (2011).

18. Lane, R. D. *et al.* Neuroanatomical correlates of pleasant and unpleasant emotion. *Neuropsychologia* **35,** 1437–1444 (1997).

19. Zald, D. H. & Pardo, J. V. Emotion, olfaction, and the human amygdala: Amygdala activation during aversive olfactory stimulation. *Proc. Natl. Acad. Sci.* **94,** 4119–4124 (1997).

20. Schneider, F. *et al.* Differential amygdala activation in schizophrenia during sadness. *Schizophr. Res.* **34,** 133–142 (1998).

21. Ketter TA, Andreason PJ, George MS & et al. Anterior paralimbic mediation of procaine-induced emotional and psychosensory experiences. *Arch. Gen. Psychiatry* **53,** 59–69 (1996).

22. Irwin, W. *et al.* Human amygdala activation detected with echo-planar functional magnetic resonance imaging. *Neuroreport* **7,** 1765–1769 (1996).

23. Ledoux, J. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. (Simon & Schuster, 1998).

24. Cain, D. M., Loewenstein, G. & Moore, D. A. The dirt on coming clean: perverse effects of disclosing conflicts of interest. *J. Leg. Stud.* **34,** 1–25 (2005).

25. Costa, V. D., Lang, P. J., Sabatinelli, D., Versace, F. & Bradley, M. M. Emotional imagery: assessing pleasure and arousal in the brain's reward circuitry. *Hum. Brain Mapp.* **31,** 1446–1457 (2010).

26. Knutson, B., Adams, C. M., Fong, G. W. & Hommer, D. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci* **21,** RC159 (2001).

27. 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* **288,** 1835–1838 (2000).

28. Miller, E. K. & Cohen, J. D. An Integrative Theory of Prefrontal Cortex Function. *Annu. Rev. Neurosci.* **24,** 167–202 (2001).

29. Greene, J. D. & Paxton, J. M. Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci.* **106,** 12506–12511 (2009).

30. Zhu, L. *et al.* Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nat Neurosci* **17,** 1319–1321 (2014).

31. Langleben, D. D. *et al.* Telling truth from lie in individual subjects with fast event-related fMRI. *Hum. Brain Mapp.* **26,** 262–272 (2005).

32. Mohamed, F. B. *et al.* Brain Mapping of Deception and Truth Telling about an Ecologically Valid Situation: Functional MR Imaging and Polygraph Investigation—Initial Experience 1. *Radiology* **238,** 679–688 (2006).

33. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLoS Biol.* **13,** (2015).

34. Winnett, R. & Rayner, G. *No Expenses Spared*. (Bantam Press, 2009).

35. Welsh, D. T., Ordóñez, L. D., Snyder, D. G. & Christian, M. S. The slippery slope: How small ethical transgressions pave the way for larger future transgressions. *J. Appl. Psychol.* **100,** 114–127 (2015).

36. Schrand, C. M. & Zechman, S. L. C. Executive overconfidence and the slippery slope to financial misreporting. *J. Account. Econ.* **53,** 311–329 (2012).

37. LeDoux, J. E. Emotion circuits in the brain. *Annu. Rev. Neurosci.* **23,** 155–184 (2000).

38. Phelps, E. A. Emotion and Cognition: Insights from Studies of the Human Amygdala. *Annu. Rev. Psychol.* **57,** 27–53 (2006).

39. Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10,** 59–63 (2006).

40. Shenhav, A. & Greene, J. D. Integrative Moral Judgment: Dissociating the Roles of the Amygdala and Ventromedial Prefrontal Cortex. *J. Neurosci.* **34,** 4741–4749 (2014).

41. Wu, D., Loke, I. C., Xu, F. & Lee, K. Neural correlates of evaluations of lying and truth-telling in different social contexts. *Brain Res.* **1389,** 115–124 (2011).

42. Mazar, N., Amir, O. & Ariely, D. The dishonesty of honest people: A theory of self-concept maintenance. *J. Mark. Res.* **45,** 633–644 (2008).

43. Effron, D. A., Bryan, C. J. & Murnighan, J. K. Cheating at the end to avoid regret. *J. Pers. Soc. Psychol.* **109,** 395–414 (2015).

Figure 1. Procedure and dishonesty escalation

(a) Participants were presented with different pictures of glass jars containing one penny coins and entered their advice regarding how much money is in the jar. They were led to believe this advice was relayed to the Estimator via connected computers. **(b)** The magnitude of dishonesty increased over the course of the block when it benefited the self but not otherwise. This was evident when regressing dishonesty on trial number for each participant (N = 55) for each condition and then entering these escalation betas in a linear model across participants controlling for starting dishonesty and study (Self-Serving-Other-Harming: t_{52} = 3.38, p = 0.001; Self-Serving-Other-Serving: t_{52} = 2.25, p = 0.03; Self-Harming-Other-Serving: t_{52} = -0.27, p=0.79). Dishonesty escalation was greater when it benefited the self than not (Self-Serving-Other-Harming vs Self-Harming-Other-Serving: $F_{1,51}$ = 8.80, p = 0.005; Self-Serving-Other-Serving vs Self-Harming-Other-Serving: $F_{1,51}$ = 4.61, p = 0.037; Self-Serving-Other-Harming vs Self-Serving-Other-Serving: $F_{1,51}$ = 0.183, p = 0.670 - statistics reported for ANCOVAs on regression coefficients with condition as a 2 level repeated factor, controlling for initial levels of dishonesty and study). **(c-e)** Averaging mean dishonesty across participants on every trial and correlating with trial number (60 trials) in each condition also revealed significant escalation when dishonesty was self-serving but not otherwise (Self-Serving-Other-Harming: r_{58} = 0.66, p<0.001; Self-Serving-Other-Serving: r_{58} = 0.83, p<0.001; Self-Harming-Other-Serving: r_{58} = -0.23, p=0.08).

Error bars represent standard error of the mean.

n.s. = non-significant. *p<0.05

Figure 2. Replication and extension study

(a) Initial level of dishonesty across participants (N=25) was greater than zero when Other Serving (t_{24} = 3.11, p=0.01, one sample ttest vs 0) but not when Self-Serving (t_{24} = 0.92, p=0.37, one sample ttest vs 0) and did not differ between conditions ($F_{1,24}$ = 1.24, p = 0.28, repeated measures ANOVA). **(b) Mean dishonesty** over the course of the block was greater than zero when Other Serving (t_{24} =

2.11, p=0.046, one sample ttest vs 0) and when Self-Serving (t_{24} = 3.78, p=0.001, one sample ttest vs 0) and did not differ across conditions ($F_{1,24}$ = 3.50, p=0.07, repeated measures ANCOVA). **(c) Escalation of dishonesty** was significant when Self Serving (analysis conducted as in Figure 1b, t_{23} = 4.53, p<0.001), but not when Other-Serving (t_{23} = 1.62, p=0.12) and greater in the former condition than the latter ($F_{1,20}$ = 7.55, p=0.01, repeated measures ANCOVA controlling for starting dishonesty).
Error bars represent standard error of the mean.
*p<0.05, n.s. = non-significant, n.s.t = non-significant trend

Figure 3. Reduction of BOLD sensitivity to dishonesty over time
Parameter estimates of time-weighted-dishonesty were averaged across all voxels in (**a**) an ROI generated from Neurosynth based on a meta-analysis of 11,406 studies, reflecting P(Emotion|Activation). Higher Z scores (lighter colors) indicate higher likelihood that the term "emotion" is used in a study given that a voxel is activated (see[17] for method details), suggesting stronger selective association between that region and emotion [17]. As seen, the ROI predominantly consists of bilateral amygdala. **(b)** A significant positive effect of time-weighted-dishonesty was revealed when dishonesty was Self-Serving-Other-Harming (t_{24} = 2.36, p = 0.027, one sample ttest vs 0) but not when it was Self-Harming-Other-Serving (trend in opposite direction: t_{24} = -1.93, p=0.066, one sample ttest vs 0), with the former significantly larger than the latter (t_{24} = 2.96, p=0.007, paired sample ttest). Repeating this analysis restricting the ROI to voxels in the anatomically defined amygdala revealed the same pattern of results (**Supplementary Figure 2a**). N=25.
Error bars represent standard error of the mean.
*p<0.05, n.s. = non-significant

Figure 4. Reduction in Bold Response to Dishonesty Predicts its Escalation
(a) For each participant, in each condition, reduction in BOLD response to one unit of dishonesty on a current trial relative to the last (extracted across the ROI displayed in **Figure 3a**) is related to escalation in dishonesty on the next trial relative to the current trial. **(b)** Example participant shown for dishonesty in the Self-Serving-Other-Harming condition. **(c)** Across participants these betas revealed a significant positive effect when dishonesty was self-serving-other-harming (t_{24} = 2.48, p=0.021, one sample ttest vs 0), but not when it was self-harming-other-serving (t_{24} = -1.53, p=0.14, one sample ttest vs 0) with the former betas larger than the latter (t_{24} = 2.82, p=0.01, paired sample ttest). Repeating this analysis restricting this ROI to voxels in the anatomically defined amygdala revealed the same pattern of results (**Supplementary Figure 2b**). N=25.
Error bars represent standard error of the mean.

*p<0.05, n.s. = non-significant

Online Methods

Participants (Experiment 1). 58 individuals aged 18 to 65 (mean age (s.d.) = 22.95 (3.55), n female = 36), participated in the study, recruited from University College London (UCL) psychology subject pool. A subset of these (n=28) undertook the experiment in an MRI scanner and the rest undertook the experiment in testing cubicles. 3 MRI participants were subsequently excluded from analysis; 1 because their structural scan showed a suspected brain abnormality; 1 because they repeatedly (≥ 4 occasions) failed a test to check they understood the instructions; 1 because comments to the experimenter revealed they suspected the Estimator was a confederate (final n = 55, n female = 34, mean age (s.d.) = 23.02 (3.59)). The study was approved by the UCL Psychology Ethics Committee. Written informed consent was obtained from all participants and they were paid for participation.

Stimuli. Stimuli consisted of 30 pictures of transparent glass jars containing differing quantities of United Kingdom one penny coins (range: 1,500 to 3,500). Each picture was presented twice in each block to participants in a randomized order, resulting in 60 trials per block.

Procedure and Paradigm. The paradigm was adapted from a previous study[24] that examined self-serving dishonesty. At the start of the study, each participant met the experimenter and a confederate, who they were led to believe was a second participant. The participant was then assigned the role of an "Advisor". In this role they would view large, high resolution pictures of each jar for 3 seconds. They would then send advice (via what they were led to believe were connected computers) to the confederate, indicating how much money they thought was in each jar. The confederate was designated the role of the "Estimator". The participant was led to believe the following; the Estimator would view a smaller picture of the same jar for a shorter period of time (1 second) and would receive advice from the Advisor regarding how much they thought was in the jar. After receipt of this advice, the Estimator would then submit estimates of how much money was in each jar on behalf of both participants. At the end of the experiment, one trial would be selected at random and both parties would be paid according to how accurate the Estimator had been on that trial.

After these instructions, the Advisor and the Estimator were invited to undertake training in separate cubicles. Following training, the Advisor (participant) was privately informed that the range in the jar would always be between £15 and £35 and that the Estimator was not aware of this. They were also told that the Advisor and the Estimator would not always be paid according to how accurate the Estimator was (as they had earlier been told). However, the Estimator would always believe that both parties would be paid according to their accuracy.

On each trial a photo of a glass jar containing one penny coins was presented on screen for 3s. Then participants were instructed to enter their advice of how much money was in the jar to the nearest pound. Participants had up to 4s to do so using a button box with five digits in each hand. If the participant failed to respond, then that trial was excluded from analysis (mean trials with no response over the 4 blocks = 2.45, s.d. = 3.74). Next, the words “Estimator Submitting Estimate” appeared on screen for 2-4s. Finally, a fixation cross appeared for 3.5–7.5s (jittered). See **Figure 1a**.

Participants who completed the study outside the scanner followed the same task except: (1) the jar screen and response screen were combined; participants had up to 10 seconds to view the jar and enter their advice; (2) A subset of these participants (n=15) were asked immediately after the Self-Serving-Other-Harming condition to estimate by how much they lied on average during the block and on the last trial (**Supplementary Figure 3**). Note that an additional 15 participants were also asked by how much they lied during the session. However, we subsequently realized the question was underspecified, with some subjects entering the sum of their lies throughout the session, while others the average amount or the amount on the last trial. Thus we revised the question asking two specific questions instead, and only analyzed answers of the participants who received the revised version of the question. There was no differences in behavior between participants who completed the study outside the scanner (N=30) and inside the scanner (N=25) on either dishonesty escalation (Self-Serving-Other-Harming: t_{53} = 0.79, p=0.44; Self-Serving-Other-Serving: t_{53} = 0.10, p=0.93; Self-Harming-Other-Serving: t_{53} = -0.71, p=0.48, independent sample ttests) nor dishonesty magnitude (Self-Serving-Other-Harming: t_{53} = -1.11, p=0.27; Self-Serving-Other-Serving: t_{53} = -0.58, p=0.56; Self-Harming-Other-Serving: t_{53} = -1.40, p=0.17, independent sample ttests).

The experiment was divided into 4 blocks, consisting of 60 trials each. Before commencing each block, the new reward structure was explained (over an intercom system if the participant was in an MRI scanner). To ensure that the participant fully understood the new reward structure, participants were given different scenarios and required to calculate possible rewards given those scenarios. If participants had difficulty answering these questions, further instruction was provided. They were reminded that the Estimator always believes rewards are contingent on accuracy.

Incentive Structure. Participants were told that on each trial the available reward would be a randomly generated amount. Neither the range of available rewards nor the reward available on each specific trial was disclosed. Making available reward an unknown quantity that fluctuated on each trial meant

that the same amount of dishonesty, in the same block, could lead to differing rewards depending on the reward available on that trial. Thus, increasing dishonesty from one trial to the next would not necessarily increase returns on the subsequent trial. This was done so that escalation of dishonesty could not be accounted for by adaptation to reward. Furthermore, participants were told that rewards would not be accumulating, but rather they will be paid according to the outcome on one trial that was selected randomly at the end of the experiment. Participants were led to believe that the reward gained depended on the incentive structure and the Estimator's estimate. However, as there was no estimator in reality, they were all paid a fixed amount at the end of the study.

The reward structure varied in each block (blocks counterbalanced across participants) as follows:

1. **Baseline** - the participant was instructed that if a trial was selected from this block for payment at the end of the experiment, both s/he (the Advisor) and the Estimator would each be rewarded according to how accurate the Estimator was (i.e. how close to the actual amount in the jar they were). The more accurate the estimate, the greater the reward, with the maximum available reward determined randomly on each trial. Thus, the participant is incentivized to send accurate advice about the amount of money in the jar.
2. **Self-Serving-Other-Harming** - the participant was instructed that s/he (the Advisor) would be rewarded according to how much the Estimator overestimated the amount in the jar, while the Estimator would be rewarded according to how accurate his/her estimate was. Thus, for the participant, the greater the overestimation, the greater the reward up to a certain amount that will be determined randomly on each trial. In this condition, dishonesty (sending high estimates about money in the jar) benefits the participant (Advisor) only and comes at a detrimental cost to the Estimator.
3. **Self-Serving-Other-Serving** - the participant was instructed that if a trial was selected from this block for payment, both s/he (the Advisor) and the Estimator would each be rewarded according to how much the Estimator overestimated the amount in the jar. The greater the overestimation, the greater the reward up to a certain amount that will be determined randomly on each trial. In this condition, dishonesty (sending high estimates regarding money in the jar) benefits both parties.
4. **Self-Harming-Other-Serving** - the participant was instructed that if a trial is selected from this block for payment, s/he (the Advisor) would be rewarded according to how accurate the Estimator's estimate was. The Estimator however would be rewarded according to how much the Estimator overestimated the amount in the jar. Thus for the participant, the more accurate the estimate, the greater the reward, with the maximum available reward determined randomly on each trial. Hence in this condition, falsely sending high estimates regarding money in the jar benefits the Estimator only and comes at a cost to the participant (Advisor).

Experiment 2: We ran a follow up study to extend and replicate our findings to a situation where dishonesty only serves the self vs only serves another. In addition to the baseline condition, where dishonesty does not serve anyone, we also had a condition where dishonesty would only benefit the self without affecting the other participant (Self-Serving) and a comparison condition where dishonesty benefitted the other participant without benefiting or hurting the self (Other-Serving). There were 2 separate blocks for each condition – one in which dishonesty constituted overestimating the amount of money in the jar and another in which dishonesty constituted underestimating the amount of money in the jar. This was done so that the results could generalize to different instructions. Each of these conditions was run twice in two separate blocks (once with overestimation and another with underestimation benefiting the participant/partner).

Self-Serving - the participant was instructed that s/he (the Advisor) would be rewarded according to how much the Estimator overestimated the amount in the jar (in the underestimated condition we simply exchanged the word "over" for "under"), while the Estimator would be rewarded the same amount regardless of their estimate. In this condition, dishonesty benefits the participant (Advisor) only and does not impact the Estimator.

Other-Serving - the participant was instructed that if a trial is selected from this block for payment, s/he (the Advisor) would be rewarded the same amount regardless of the Estimator's estimate. The Estimator however would be rewarded according to how much the Estimator overestimated the amount in the jar (in the underestimated condition we simply exchanged the word "over" for "under"). Thus for the participant, dishonesty benefits the other party with no cost/benefit imposed on themselves.

Participants (Experiment 2): 35 participants were tested at UCL on the task using the same behavioral procedure described above. In this study we employed a strict exclusion procedure; in addition to asking participants if they had any reservations regarding the estimator (the procedure used for exclusion in Experiment I), to which only 3 replied yes (and of these 3, only 2 indicated they suspected that the partner was not a real participant in a follow up question). We also explicitly told participants at the end of the study that the estimator was a confederate and asked them if they suspected as much before. Ten participants replied yes and thus were excluded from analysis (final n = 25, n female = 18, mean age (s.d.) = 20.76 (1.90)).

Behavioral Analysis:

Dishonesty Scores. In each condition dishonesty was estimated on a trial by trial basis by subtracting baseline condition advice for each specific jar from advice sent to the Estimator upon presentation of the same jar (since each jar was presented twice the mean advice for each jar in the baseline condition was used). Thus for each trial:

Dishonesty = Advice (for jar j) – Advice (for jar j in Baseline condition)

For blocks in Experiment II in which dishonesty involved sending low advice, dishonesty was calculated as the reverse, i.e.:

Dishonesty = Advice (for jar j in Baseline condition) - Advice (for jar j)

This meant that high numbers always indicate greater dishonesty in each block in each study.

Dishonesty Escalation. We were interested in the temporal development of dishonesty. To investigate this, for each participant dishonesty was regressed on trial number in each condition and the resulting standardized regression coefficients were entered into an ANCOVA with condition (Self-Serving-Other-Harming, Self-Serving-Other-Serving, Self-Harming-Other-Serving), controlling for starting dishonesty and study (in scanner/outside scanner). To test whether escalation between 2 specific conditions was significantly different from each other, the same ANCOVA was carried out as above on 2 levels for the repeat factor. Simple effects were examined vs 0 in a linear model, controlling for starting dishonesty and whether the study took place inside/outside the scanner. In addition, we calculated mean dishonesty per trial number over all participants, resulting in 60 values per block and correlated these values with their trial number (1-60). Correlations were compared against each other using Steigers Z.

Dishonesty Magnitude. To examine dishonesty magnitude, dishonesty scores were averaged for each participant in each condition and entered into a 3 way repeat measures ANOVA controlling for study. To test whether dishonesty magnitude between 2 specific conditions was significantly different from each other, the same ANOVA as above was conducted with those specific factors. This was followed by one sample t-tests. Starting dishonesty was examined in the same way (with starting dishonesty entered in place of dishonesty magnitude). To test for ceiling effects, one sample ttests were conducted for each condition for average dishonesty and dishonesty on the last trial against a test value of 64 (maximum advice participants could enter [£99] minus maximum amount of pennies that a jar could contain [£35]).

Accuracy in Baseline Condition. To examine how accurate participants were in the baseline condition we calculated the absolute difference between participants' advice on each trial and the actual amount of money in the jar displayed (i.e. error). We also regressed error rate over the 60 trials on trial number for each participant and compared coefficients at the group level (see **Supplementary Figure 4**).

Reaction Time Analysis. Mean reaction time was calculated for each participant for each condition and compared between conditions. In addition we correlated reaction time with trial number for each participant for each condition and compared standardized coefficients at the group level to examine temporal changes in RT. **Supplementary Figure 5**.

Image acquisition.

Scanning was performed at Birkbeck-UCL Centre Neuroimaging, London using a 1.5T Siemens Avanto scanner with a 32-channel Siemens head coil. To correct for inhomogeneity of the static magnetic field we acquired field maps to be used in the unwarping stage of data preprocessing. Anatomical images were acquired using MPRage, which comprised 1mm thick axial slices parallel to the AC-PC plane. Functional images were acquired as echo-planar (EPI) T2*-weighted images. Time of repetition (TR) = 3.40 sec, time of echo (TE) = 30 ms, flip angle (FA) = 90, matrix = 64 X 64, field of view (FOV) = 192 mm, slice thickness = 2 mm. A total of 40 axial slices (-45° tilt) were sampled, in-plane resolution = 3mm x 3mm. Statistical Parametric Mapping (SPM8, Wellcome Trust Centre for Neuroimaging) was used for image analysis. After discarding the first five dummy volumes, images were realigned to the sixth volume and unwrapped using 7^{th} degree B-spline interpolation. Movement plots were studied to ensure scan-to-scan translations greater than one-half of a voxel (1.5mm) or rotations greater than 1 degree did not cause artifacts in the corresponding scans. Structural images were reregistered to mean EPI images and segmented into gray and white matter. These segmentation parameters were then used to normalize and bias correct the functional images. Normalization was to a standard EPI template based on the Montreal Neurological Institute (MNI) reference brain using a non linear (7^{th} degree B-spline) interpolation. Normalized images were smoothed using a Gaussian kernel of 8 mm full-width at half-maximum. Low frequency artifacts were removed using a 1/128 Hz high-pass filter and temporal autocorrelation intrinsic to the fMRI time series was corrected using an AR(1) process.

fMRI analysis:

ROI: Our ROI was defined by generating a map in Neurosynth, a meta-analysis based on 11,406 studies, reflecting P(Emotion|Activation) thresholded at z = 5 using xjView (note that the results are robust and observed for maps generated using different thresholds). This map corresponds to the likelihood that the term "emotion" is used in a study given the presence of reported activation in a particular voxel[17]. The map reflects the relative selectivity with which voxels activate in relation to the term "emotion" using a comparison between all the studies in the database that contain the term "emotion" and all those that do not (for method details see[17]). This ROI predominantly, but not exclusively, occupied the amygdala (**Figure 3a**). We also reran every analysis using a 2nd ROI which was a conjunction of the ROI above and the anatomically defined amygdala. The amygdala was defined using the SPM Anatomy Toolbox, which provides probabilistic cytoarchitectonic maps[44,45].

Time Weighted Dishonesty. For each participant, a design-matrix was created with event onsets time-locked to the temporal positions of jar presentation and presentation of estimator screen. Events were modeled as delta functions and convolved with a canonical hemodynamic response function to create regressors of interest. 6 motion correction regressors estimated from the realignment procedure were entered as covariates of no interest. The onset regressor for jar presentation was modulated by the three parametric regressors: (i) Time, was entered to indicate the number of trials left in the block, so 60 for the first trial, 59 for second trial, and so on up to 1 for the last trial. These numbers were all divided by 100 for normalization purposes (ii) Dishonesty on each trial (calculated as described in behavioral analysis section); (iii) Time weighted dishonesty, calculated as the product of the two regressors above. Time weighted dishonesty weights dishonesty according to how far in the block participants are such that a small amount of dishonesty early on in the block is equivalent to a greater amount later on in the block (and vice versa), all else held constant. In this model, time weighted dishonesty is orthoganalized with respect to dishonesty and time; and dishonesty is orthogonalized with respect to time by the SPM8 software.

Average parameter estimates for time weighted dishonesty for each participant were extracted for Self-Serving-Other-Harming and Self-Harming-Other-Serving conditions using MarsBar. These conditions were compared against each other using paired sample ttests. One sample ttests against a test value of 0 were used to test the significance of specific conditions.

Whole Brain Analyses: Whole brain exploratory analysis was conducted for regions outside the ROIs (threshold $p<0.001$, FWE $P < 0.05$, cluster size (K) > 5).

Prediction Analysis. For each participant we created a design-matrix in which each jar presentation (60 per condition) was modeled as a separate event (without parametric regressors attached to any of these events). Such a procedure has been used many times in the past (e.g.[46,47]). Estimator onset was entered as an additional event. Events were modeled as delta functions and convolved with a canonical hemodynamic response function to create regressors of interest. 6 motion correction regressors estimated from the realignment procedure were entered as covariates of no interest. We then used this model to extract the average BOLD signal on each trial in our ROI using the "spm_summarise.m" function. BOLD activity on each trial generated by this model was then divided by dishonesty on each trial to compute BOLD signal per unit of dishonesty for each subject and trial:

$BOLD\ per\ unit\ dishonesty_t = Bold_t/dishonesty_t$

Trials which were missed or in which dishonesty was 0 were excluded as this division leads to infinity (mean number excluded trials Self-Serving-Other-Harming: 3.04; Self-Serving-Other-Serving: 1.24; Self-Harming-Other-Serving: 6.12).

We then calculated the reduction of this term on trial (t) relative to the previous trial (t-1):

$BOLD\ per\ unit\ dishonesty_{t-1} - BOLD\ per\ unit\ dishonesty_t$

Hence positive numbers indicate a decrease in BOLD per unit dishonesty. Separately we calculated dishonesty escalation observed in participants' *behavior*, as the increase in dishonesty on the subsequent trial (t+1) relative to trial t:

$Dishonesty\ Escalation_{t+1} = Dishonesty_{t+1} - Dishonesty_t$

Hence positive numbers indicate an increase in dishonesty (i.e. dishonesty escalation). For each subject we then used BOLD signal reduction per unit dishonesty on trial t as a predictor of escalation of behavioral dishonesty on trial t+1 in a general linear regression:

$Dishonesty\ Escalation_{t+1} = b0 + b1*BOLD\ signal\ reduction\ per\ unit\ dishonesty_t$

This analysis is therefore independent from our parametric model above as it explores whether change in BOLD predicts dishonesty on the subsequent trial (see **Figure 4a**) whereas the parametric model is used

to identify correlates of time weighted dishonesty on the current trial. We ran the regression for each participant for the Self-Serving-Other-Harming condition and separately ran it for the condition in which dishonesty was Self-Harming-Other-Serving. The resulting standardized regression coefficients were compared against each other using paired sample ttests and followed up with one sample ttests against a test value of 0.

To check the robustness of the results, we repeated this analysis 1000 times. On each occasion fMRI data of each subject was used to predict dishonesty escalation of an alternate subject (randomly selected without replacement) for each of the 25 subjects in the sample. We then examined the number of occasions (in the 1000 iterations) a set of betas was significant for either condition and conditions were significantly different from one another.

Post hoc ROIs. To examine if the results observed could generalize to other regions that are involved in different aspects of deception we run the analysis described above in the NA, AI and DLPFC (see **Supplementary Table 2**). The NA was anatomically defined by creating two 8mm spheres around MNI coordinates +/- 10, 8, -5 (as done previously[48]) using Marsbar region of interest toolbox for SPM (http://marsbar.sourceforge.net). These 2 spheres were then combined to create one ROI of bilateral NA. The DLPFC was anatomically defined using WFU-Pickatlas[49] (Broadman Areas 9 and 46). AI was anatomically defined according to the WFU-Pickatlas tool[49] by dividing the insula at its midpoint (y=0) which approximately demarcates dysgranualar and granular sectors (as done previously[50]).As for NA above, we created a bilateral ROI of DLPFC and a bilateral ROI for the AI..

Debriefing. After the experiment participants were given a debriefing questionnaire. First, they were asked what they thought the goal of the experiment was (see **Supplementary Table 1** for results) Second, their thoughts regarding the estimator were examined (see **Supplementary Table 3** for results). Finally, participants were asked to report the strategy they used in the conditions when dishonesty was self-serving. In Experiment 1, 0% of participants indicated that they attempted to gradually increase overestimations (see **Supplementary Table 4** for results). In Experiment 2, 12% of participants indicated gradual change in dishonesty as a strategy in the Self-Serving condition and 4% in Other-Serving condition (all numbers were averaged over the “overestimate” and “underestimate” conditions). We then gave participants a closed question where subjects picked one strategy from 6 different options which included escalation as one of the options (options were: (1) Underestimate/Overestimate by a little, (2) Underestimate/overestimate by a lot, (3) Gradually increase underestimation /overestimation (4) Gradually decrease underestimation/overestimation, (5) Overestimate/Underestimate, (6) Other (none of the above).

8% of subjects picked escalation as a strategy in the self-serving condition and 4% in the other serving condition. For Self-Serving condition 20% picked the option of over/underestimate by a little and 64% picked over/underestimate by a lot. For Other-Serving the numbers were 12% and 56%, respectively. Finally, we gave subjects the opportunity to pick as many strategies as they wanted from the 6 available; 44% picked the escalation option as one of their choices in the self-serving condition and 28% of participants did so in the other-serving condition.

Statistics. A within-subject design was used. Thus, experimental group randomization or blinding is not applicable. We followed standard procedures in the field and statistical procedures in line with previously published studies. All Student's *t* tests are two-tailed; *F* tests in ANOVAs were one-tailed, as is standard for such comparisons. Data distribution was assumed to be normal, but not formally tested. Sample sizes are based on a behavioral pilot study we administered using this paradigm.

Code Availability: Custom MATLAB scripts were used to relate neural adaptation to dishonesty escalation. These are available upon request.

44. Eickhoff, S. B. *et al.* A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* **25,** 1325–1335 (2005).

45. Eickhoff, S. B., Heim, S., Zilles, K. & Amunts, K. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *Neuroimage* **32,** 570–582 (2006).

46. Charpentier, C. J., Moutsiana, C., Garrett, N. & Sharot, T. The Brain's Temporal Dynamics from a Collective Decision to Individual Action. *J. Neurosci.* **34,** 5816–5823 (2014).

47. Edelson, M. G., Dudai, Y., Dolan, R. J. & Sharot, T. Brain Substrates of Recovery from Misleading Influence. *J. Neurosci.* **34,** 7744–7753 (2014).

48. Drueke, B. *et al.* Neural correlates of positive and negative performance feedback in younger and older adults. *Behav. Brain Funct.* **11,** 1 (2015).

49. Maldjian, J. A., Laurienti, P. J., Kraft, R. A. & Burdette, J. H. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage* **19,** 1233–1239 (2003).

50. Slavich, G. M., Way, B. M., Eisenberger, N. I. & Taylor, S. E. Neural sensitivity to social rejection is associated with inflammatory responses to social stress. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14817–14822 (2010).