Automatic honesty forgoing reward acquisition and punishment avoidance: a functional MRI investigation

Mei Yoneda[a], Ryuhei Ueda[a,c], Hiroshi Ashida[a] and Nobuhito Abe[b]

Recent neuroimaging investigations into human honesty suggest that honest moral decisions in individuals who consistently behave honestly occur automatically, without the need for active self-control. However, it remains unclear whether this observation can be applied to two different types of honesty: honesty forgoing dishonest reward acquisition and honesty forgoing dishonest punishment avoidance. To address this issue, a functional MRI study, using an incentivized prediction task in which participants were confronted with real and repeated opportunities for dishonest gain leading to reward acquisition and punishment avoidance, was conducted. Behavioral data revealed that the frequency of dishonesty was equivalent between the opportunities for dishonest reward acquisition and for punishment avoidance. Reaction time data demonstrated that two types of honest decisions in the opportunity for dishonest reward acquisition and punishment avoidance required no additional cognitive control. Neuroimaging data revealed that honest decisions in the opportunity for dishonest reward acquisition and those for punishment avoidance required no additional control-related activity compared with a control condition in which no opportunity for dishonest behavior was given. These results suggest that honesty flows automatically, irrespective of the concomitant motivation for dishonesty leading to reward acquisition and punishment avoidance. *NeuroReport* **28:879–883 Copyright © 2017 The Author(s). Published by Wolters Kluwer Health, Inc.**

NeuroReport 2017, **28**:879–883

Keywords: functional MRI, honesty, prefrontal cortex, punishment, reward

[a]Graduate School of Letters, [b]Kokoro Research Center, Kyoto University, Kyoto and [c]The Japan Society for the Promotion of Science (JSPS), Kojimachi Business Center Building, Tokyo, Japan

Correspondence to Nobuhito Abe, PhD, Kokoro Research Center, Kyoto University, 46 Shimoadachi-cho, Yoshida Sakyo-ku, Kyoto 606-8501, Japan
Tel/fax: +81 75 366 7202; e-mail: abe.nobuhito.7s@kyoto-u.ac.jp

Received 22 June 2017 **accepted** 3 July 2017

Introduction

Recent literature regarding cognitive science provides one key debate on the cognitive nature of honesty: do we behave honestly by force of will, or does honesty flow automatically? According to the 'Will' hypothesis, honest behavior results from the active resistance of temptation, comparable to the controlled cognitive processes that enable the delay of reward [1,2]. According to the 'Grace' hypothesis, honest behavior occurs more automatically, without the need for active self-control [3,4]. A series of functional MRI (fMRI) studies have supported the Grace hypothesis, indicating that consistently honest behavior involves no additional cognitive work associated with a longer reaction time and prefrontal activation [5,6].

Although it appears that there is no distinctive neural signature of honest behavior, previous studies have left one important question unaddressed. It is possible that the cognitive nature of honesty differs depending on the concomitant motivation for dishonesty leading to reward acquisition and punishment avoidance. This line of thinking is based on a well-known framing effect [7], which occurs when transparently and objectively identical situations generate dramatically different decisions depending on whether the situations are presented as potential gains or losses. A critical feature of the framing effect is that individuals are loss averse: they are willing to go to greater lengths to avoid a loss than to obtain a gain of a similar size [8]. Notably, this tendency substantially influences ethical judgments and behaviors [9]. Kern and Chugh [9] reported that participants in the loss-frame condition were more likely to favor gathering 'insider information' and lied more than participants in the gain-frame condition.

The previous findings on the framing effect just described enable us to hypothesize that the frequency of dishonest behavior that leads to punishment avoidance would be higher than the frequency of dishonest behavior that leads to reward acquisition. As a result, the honest behavior that forgoes dishonest punishment avoidance would require greater cognitive control compared with the honest behavior that forgoes dishonest reward acquisition. However, there is another potential competing hypothesis. Given that individuals who consistently choose honest behavior are characterized by a relatively tepid motivation for dishonesty, we can assume that both kinds of honest behaviors are not associated with additional cognitive control irrespective of the concomitant motivation for dishonesty leading to reward acquisition and punishment avoidance. If this is

0959-4965 Copyright © 2017 The Author(s). Published by Wolters Kluwer Health, Inc. DOI: 10.1097/WNR.0000000000000848

the case, the frequency of dishonesty would be equivalent between the opportunities for dishonest reward acquisition and for punishment avoidance. Thus, to test these competing hypotheses, we used fMRI combined with an incentivized prediction task [5,6,10–12] and focused on reaction time and prefrontal activity associated with honest decisions in individuals classified as honest.

Participants and methods

Participants

The present results are based on data from 33 right-handed healthy participants (15 women and 18 men, mean age: 25.6 years, age range: 20–38 years) without a history of neurological or psychiatric disease. Participants were classified as honest, ambiguous, or dishonest based on self-reported accuracy in the Opportunity (Op) condition of the incentivized coin-flip task collapsed across Reward (Rew) and Punishment (Pun) conditions. Consistent with protocols used previously [5,6], nine participants who reported improbably high levels of accuracy at the individual level (binomial test, $P < 0.001$) were classified as dishonest (mean 'accuracy' = 86.5%; SD = 13.6). The 20 lowest-accuracy participants (binomial test, $P > 0.05$ for the entire group of 20) were classified as honest (mean accuracy = 51.7%; SD = 4.0). The remaining four participants were classified as ambiguous (mean accuracy = 62.5%; SD = 2.0). Participants were paid 8000 yen (~$80) for participating, in addition to the bonus pay based on performance during the experimental tasks. All of the participants provided written informed consent in accordance with the Declaration of Helsinki and guidelines approved by the Ethical Committee of Kyoto University.

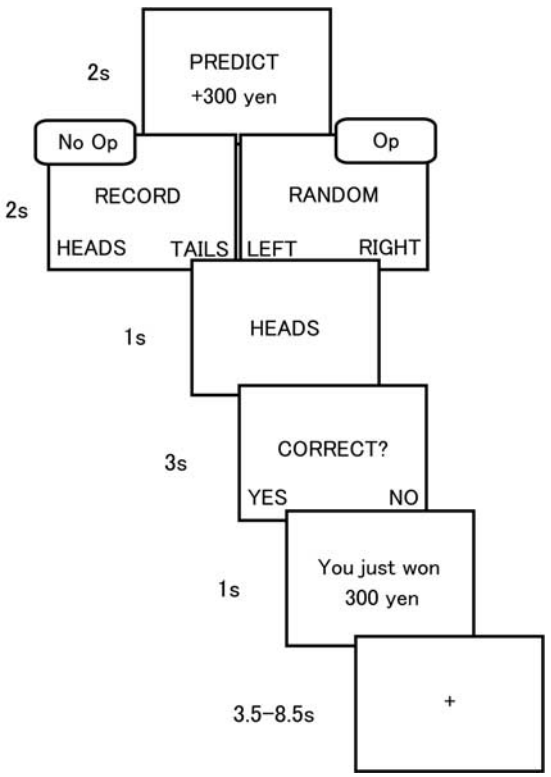
In addition to the data acquired from the 33 participants analyzed, data from a total of 15 participants were discarded for reasons described below. The exclusion criteria used in the present study were identical to those reported previously [5,6]. First, in debriefing, participants were asked what they thought the experiment was about in an open-ended way. At this point in the debriefing, 10 participants classified as dishonest, three participants classified as ambiguous, and six participants classified as honest voiced suspicions that the experiment was about cheating, lying, or dishonesty. Data from the 10 dishonest participants were discarded, but not the others. This was done to exclude data from participants who may have regarded themselves as morally justified in deceiving the experimenters because they believed that the experimenters were attempting to deceive them. Second, participants were eventually informed of the purpose of the experiment and were asked whether they were aware that they could cheat. All but four participants indicated that they were aware of the possibility of cheating. Data from these four participants were excluded because the aim was to investigate honest behavior in the face of opportunity for dishonest gain, and these excluded participants were not aware of the opportunity. Finally, tests to identify and exclude participants who strategically under-reported their accuracy were conducted. In the present paradigm, it is possible to gain money dishonestly while maintaining a chance level of accuracy by cheating in the Op trials that are worth the most (i.e. 300 yen) and deliberately under-reporting accuracy for the Op trials that are worth the least (i.e. 100 yen). To identify such participants, the winnings of each honest participant were compared with those of simulated honest participants (10 000 permutations), with win/loss percentages individually matched to the participant being tested. Based on these findings, the data from one participant classified as honest, whose winnings were improbably large given the participant's win/loss percentage ($P < 0.05$), were discarded.

Cognitive task

To measure dishonesty, an incentivized coin-flip prediction task was used (Fig. 1) [5,6]. This task reliably recruited control-related prefrontal activations in dishonest – but not honest – individuals in previous studies. A 'cover story' was used to justify giving participants obvious opportunities for dishonest gain. This study was presented as an investigation of abilities to predict the future, and aimed at testing the hypothesis that individuals are better able to predict the future when their predictions are (a) private and (b) financially incentivized. Thus, participants were implicitly led to believe that the opportunity for dishonest gain was a known but an unintended by-product of the experiment's design, and that they were expected to behave honestly. It is important to note that, in using this cover story, participants were deceived about the experimenters' interests, but not about the economic structure of the task. Participants were not presented with the cover story until after they had been recruited, thus avoiding self-selection for participants with interests in abilities to predict the future.

The task consisted of 2 (Rew, Pun) × 2 (Op, No-Op) factorial design. In the Rew conditions, participants could earn a monetary reward for the accurate prediction of coin flips, but there was no penalty for failing to predict the outcome of coin flips. In the Pun conditions, participants could avoid monetary punishment for the accurate prediction of coin flips; however, inaccurate predictions caused the participant to lose money. In the Op conditions, participants made their predictions privately and were rewarded on the basis of their self-reported accuracy, affording them the opportunity to cheat. In the No-Op condition, participants recorded their predictions in advance, denying them the opportunity to cheat by lying about their accuracy. Participants completed a total of 240 trials. Within the 60 Rew/Op trials, the values of 100, 200, or 300 yen

Fig. 1



Task sequence of the coin-flip task. The task consisted of a 2 (Rew, Pun) ×2 (Op, No-Op) factorial design. The participant observes the monetary value of the trial and privately predicts the outcome of the upcoming coin flip. Here, positive monetary value is presented in the Rew condition, whereas negative monetary value is presented in the Pun condition. The participant records this prediction by pressing one of two buttons (No-Op condition) or presses one of these buttons randomly (Op condition). The participant then observes the outcome of the coin flip. The participant then indicates whether the prediction was accurate and observes the amount of money won/lost based on the recorded prediction (No-Op) or the self-reported accuracy (Op). Accurate predictions of coin flip allow the participants to get a monetary reward (Rew condition) or to avoid monetary punishment (Pun condition). This is followed by a fixation interval. Op, Opportunity; Pun, Punishment; Rew, Reward.

appeared 20 times each, as was the case in the 60 Rew/No-Op trials. Within the 60 Pun/Op trials, the values 100, 200, or 300 yen appeared 20 times each, as was the case in the 60 Pun/No-Op trials. Trials appeared in random order in a series of four blocks of 60 trials each. Participants' understanding of the experiment was assessed in the debriefing (see above). They were asked about their thoughts and experiences during the experiment in an open-ended way. Subsequently, participants were informed of the true nature of the experiment and were asked whether they were aware of the possibility of cheating. Participants were also asked to roll a tetrahedron dice and were paid the cumulative value of their winnings/losses within one of four blocks according to the side of the dice. This procedure was used to retain participants' motivation in the entire task while avoiding excessive payments. Net losses were capped at 0 yen, and net winnings were capped at 3000 yen.

Image acquisition and analysis

Participants were scanned using a 3.0 T device (Magnetom Verio MRI; Siemens, Erlangen, Germany) equipped with a 12-channel head coil. Data preprocessing and statistical analyses were performed using SPM8 (Wellcome Department of Imaging Neuroscience, London, UK). The details of fMRI parameters and preprocessing steps (i.e. slice-timing correction, realignment, spatial normalization, and smoothing) were exactly the same as those reported in the authors' previous study [13].

The fMRI data were analyzed using an event-related model. All events of interest were modeled through convolution with a canonical hemodynamic response function temporally indexed by participants' responses. The parameter estimates (β) for each condition were calculated for all brain voxels, and the following three contrasts of parameter estimates were computed: Op-Loss versus No-Op-Loss (collapsing across Rew and Pun conditions); Rew-Op-Loss versus Rew-No-Op-Loss; and Pun-Op-Loss versus Pun-No-Op-Loss. These contrasts identify the signal associated with honest behavior in the presence of an opportunity for dishonest reward acquisition and/or punishment avoidance. Here, particular attention was given to the data from honest participants because the Grace hypothesis to be tested in the present study applies, in a strict sense, only to honest decisions in individuals who consistently behave honestly. The contrast images for the 20 honest participants were entered into a series of one-sample *t*-tests. The threshold of significance was set at P value less than 0.05 at the voxel level [corrected for multiple comparisons using family-wise error ([FWE) correction], with the cluster size of five or more voxels. However, given that the conclusions are based on a lack of prefrontal activation, this combination of thresholds is relatively stringent. Therefore, to avoid false-negative errors, a small-volume correction procedure ($P < 0.05$ FWE-corrected, $k > 5$) to reduce the number of comparisons being performed, and to increase the chance of identifying significant results in a particular region of interest [i.e. dorsolateral prefrontal cortex (DLPFC)], was also used. Specifically, small-volume correction was based on a sphere (4 mm radius) centered on coordinates derived from a previous study [5], demonstrating honesty-related DLPFC activity (MNI x, y, z: 34, 14, 54 for right DLPFC and −38, 30, 42 for left DLPFC).

Results

Behavioral data

The mean self-reported %Wins of Rew/Op-Win trials and Pun/Op-Win trials were 0.62 (SD = 0.18) and 0.63 (SD = 0.17), respectively. No significant difference was

Table 1 The mean reaction times of honest participants' responses

	Reaction time (ms)	
Condition	Mean	SD
Rew/Op-Win	694	158
Rew/Op-Loss	753	166
Rew/No-Op-Win	707	145
Rew/No-Op-Loss	765	179
Pun/Op-Win	729	163
Pun/Op-Loss	782	164
Pun/No-Op-Win	722	151
Pun/No-Op-Loss	781	161

Op, Opportunity; Pun, Punishment; Rew, Reward.

found between these two conditions ($t=-0.09$, $P=0.93$, $d=0.008$). A correlation analysis revealed a significant positive correlation between these two conditions ($r=0.87$, $P<0.001$).

Also analyzed were the reaction time data of honest participants, which are summarized in Table 1. Normality of the data was confirmed for all parametric tests (Kolmogorov–Smirnov normality tests, all $P>0.05$). A 2 (motivation: Rew, Pun)$\times$2 (opportunity: Op, No-Op)$\times$2 (outcome: Win, Loss) analysis of variance (ANOVA) revealed significant main effects of motivation [$F(1, 19)=9.32$, $P=0.01$, partial $\eta^2=0.33$] and outcome [$F(1, 19)=8.41$, $P=0.01$, partial $\eta^2=0.31$]. A two-way interaction between motivation and opportunity was significant [$F(1, 19)=7.44$, $P=0.01$, partial $\eta^2=0.28$], whereas the other effects were not significant (all $P>0.1$).

Following up on this three-way ANOVA, a 2 (opportunity: Op, No-Op)$\times$2 (outcome: Win, Loss) ANOVA was performed for the data from the Rew condition. Although significant main effect of outcome was found [$F(1, 19)=6.65$, $P=0.02$, partial $\eta^2=0.26$], there was no significant main effect of opportunity [$F(1, 19)=1.12$, $P=0.30$, partial $\eta^2=0.06$]. Here, the critical test is to determine whether an interaction (i.e. specific increase in Op-Loss trials), which suggests additional cognitive control for honest moral decisions, is significant. However, an interaction between the two factors was not significant [$F(1, 19)<0.001$, $P=1.00$, partial $\eta^2<0.001$]. An ANOVA for the data from the Pun condition yielded similar results; a main effect of outcome was significant [$F(1, 19)=7.33$, $P=0.01$, partial $\eta^2=0.28$], whereas a main effect of opportunity [$F(1, 19)=0.10$, $P=0.75$, partial $\eta^2=0.01$] and an interaction were not significant [$F(1, 19)=0.10$, $P=0.75$, partial $\eta^2=0.01$]. The null results of interaction effect indicate that honest moral decisions forgoing dishonest reward acquisition and punishment avoidance required no additional cognitive control processes.

Imaging data

To identify brain activations associated with honest moral decisions forgoing dishonest reward acquisition and punishment avoidance, Op-Loss trials were compared with No-Op-Loss trials (collapsing across Rew and Pun conditions). This analysis revealed no significant activation in any brain region. Even if a small-volume correction procedure for the DLPFC was used, no significant activation was found. We then compared Rew/Op-Loss trials with Rew/No-Op-Loss trials and found no significant activation in any brain region. We also compared Pun/Op-Loss trials with Pun/No-Op-Loss trials and again found no significant activation in any brain region.

Discussion

We used fMRI and an incentivized prediction task, in which participants were confronted with real and repeated opportunities for dishonest gain leading to reward acquisition and punishment avoidance, to determine whether different types of honest moral decisions that forgo reward acquisition and punishment avoidance are supported by active self-control. The results of the present study are contrary to the prediction inspired by the framing effect [7–9]. The frequency of dishonesty was equivalent between the opportunities for dishonest reward acquisition and for punishment avoidance. Reaction time data demonstrated that, consistent with the Grace hypothesis [5,6], two types of honest decisions in the opportunity for dishonest reward acquisition and punishment avoidance required no additional cognitive control. In line with these behavioral effects, no prefrontal activations were identified during honest decisions in both Rew and Pun conditions, even in the analyses without whole-brain FWE correction. Therefore, the present results take the Grace hypothesis a step further, indicating that both types of honest behaviors that forgo dishonest reward acquisition and punishment avoidance require no additional cognitive work. We speculate that honest individuals who exhibit consistent honest behavior are not tempted by dishonest reward acquisition or by dishonest punishment avoidance. Therefore, they require no cognitive control for either type of honest behavior that forgoes reward acquisition and punishment avoidance. Thus, 'Graceful' honesty is realized automatically, irrespective of the presented frames for gains and losses.

One major limitation of the present study was that our conclusions are based on 'null findings', raising the concern that our study design or the fMRI methodology lacked statistical power to detect real differences between dishonest reward acquisition and dishonest punishment avoidance. Although we believe that this possibility is unlikely given the high consistency between the present findings and previous studies, we leave this question as a topic for future research. Another limitation was the small number of participants classified as dishonest, which prevented us from analyzing honesty-related activity in individuals classified as dishonest. Despite these limitations, our results build on previous studies [5,6], and provide further evidence that supports the Grace hypothesis of honesty,

indicating the automatic honesty forgoing dishonest reward acquisition and punishment avoidance.

Acknowledgements

The authors are grateful to Maki Terao for her assistance in data collection.

This study was conducted using the MRI scanner and related facilities of Kokoro Research Center, Kyoto University. This work was supported by a Grant-in-Aid for Young Scientists (B) (25870337 to N.A.) from the Japan Society for the Promotion of Science (JSPS) and ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan). Nobuhito Abe was supported by the Uehiro Foundation on Ethics and Education.

M.Y. and N.A. developed the study concept and research plan; M.Y., R.U., H.A., and N.A. conducted the research and the data analysis; M.Y. and N.A. drafted the manuscript; and R.U. and H.A. critically appraised and revised the manuscript.

Conflicts of interest

There are no conflicts of interest.

References

1 McClure SM, Laibson DI, Loewenstein G, Cohen JD. Separate neural systems value immediate and delayed monetary rewards. *Science* 2004; **306**:503–507.
2 Metcalfe J, Mischel W. A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychol Rev* 1999; **106**:3–19.
3 Bargh JA, Chartrand TL. The unbearable automaticity of being. *Am Pshychol* 1999; **54**:462–479.
4 Haidt J. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol Rev* 2001; **108**:814–834.
5 Abe N, Greene JD. Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *J Neurosci* 2014; **34**:10564–10572.
6 Greene JD, Paxton JM. Patterns of neural activity associated with honest and dishonest moral decisions. *Proc Natl Acad Sci USA* 2009; **106**:12506–12511.
7 Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science* 1981; **211**:453–458.
8 Tversky A, Kahneman D. Loss aversion in riskless choice: a reference-dependent model. *Q J Econ* 1991; **106**:1039–1061.
9 Kern MC, Chugh D. Bounded ethicality: the perils of loss framing. *Psychol Sci* 2009; **20**:378–384.
10 Ding XP, Gao X, Fu G, Lee K. Neural correlates of spontaneous deception: a functional near-infrared spectroscopy (fNIRS) study. *Neuropsychologia* 2013; **51**:704–712.
11 Hu X, Pornpattananangkul N, Nusslock R. Executive control- and reward-related neural processes associated with the opportunity to engage in voluntary dishonest moral decision making. *Cogn Affect Behav Neurosci* 2015; **15**:475–491.
12 Yin L, Reuter M, Weber B. Let the man choose what to do: neural correlates of spontaneous lying and truth-telling. *Brain Cogn* 2016; **102**:13–25.
13 Yanagisawa K, Abe N, Kashima ES, Nomura M. Self-esteem modulates amygdala-VLPFC connectivity in response to mortality threats. *J Exp Psychol Gen* 2016; **145**:273–283.