

Executive control- and reward-related neural processes associated with the opportunity to engage in voluntary dishonest moral decision making

Xiaoqing Hu · Narun Pornpattananangkul · Robin Nusslock

© Psychonomic Society, Inc. 2015

Abstract Research has begun to examine the neurocognitive processes underlying voluntary moral decision making, which involves engaging in honest or dishonest behavior in a setting in which the individual is free to make his or her own moral decisions. Employing event-related potentials, we measured executive control-related and reward-related neural processes during an incentivized coin-guessing task in which participants had the opportunity to voluntarily engage in dishonest behavior, by overreporting their wins to maximize earnings. We report four primary findings: First, the opportunity to deceive recruited executive control processes involving conflict monitoring and conflict resolution, as evidenced by a higher N2 and a smaller P3. Second, processing the outcome of the coin flips engaged reward-related processes, as evidenced by a larger medial feedback negativity (MFN) for incorrect (loss) than for correct (win) guesses, reflecting a reward prediction error signal. Third, elevated executive control-related neural activity reflecting conflict resolution (i.e., an attenuated executive control P3) predicted a greater likelihood of engaging in overall deceptive behavior. Finally, whereas elevated reward-related neural activity (the reward P3) was associated with a greater likelihood of engaging in overall deceptive behavior, an elevated reward prediction error signal (MFN difference score) predicted increased trial-by-trial moral behavioral adjustment (i.e., a greater likelihood to overreport wins following a previous honest loss than following a previous honest win trial). Collectively, these findings suggest that both

executive control- and reward-related neural processes are implicated in moral decision making.

Keywords Executive control · Medial frontal negativity · Moral decision making · Reward process · Reward prediction error

Real-world deception or cheating behavior, such as a Ponzi scheme, can cause significant harm to society, organizations, and individuals. Examining the neural processes underlying honest and dishonest decision making has important implications for understanding the foundations of both moral and immoral behavior, as well as for the study of ethics, psychology, neuroscience, and law. Previous research has suggested that individuals recruit distinct neural circuits when they are instructed by an experimenter to deceive others (i.e., instructed deception), relative to when they engage in truthful behavior (Abe, 2011; Abe, Suzuki, Mori, Itoh, & Fujii, 2007; Abe et al., 2006; Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2003; Langleben et al., 2002; Lee et al., 2002; Priori et al., 2008; Spence et al., 2001; for a meta-analysis, see Christ, Van Essen, Watson, Brubaker, & McDermott, 2009). Although these studies have contributed to our understanding of the neurocognitive processes associated with deception, real-life deception often involves a voluntary intention to deceive rather than explicit instructions to do so. Thus, examining the neural processes associated with voluntary moral decision making has important implications for understanding immoral behavior as it frequently occurs in real-life settings (Sip, Roepstorff, McGregor, & Frith, 2008).

Researchers have recently begun to investigate the neural processes underlying voluntary deception, which involves

Xiaoqing Hu and Narun Pornpattananangkul contributed equally to this work.

X. Hu (✉) · N. Pornpattananangkul · R. Nusslock
Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208, USA
e-mail: xhu@u.northwestern.edu

engaging in dishonest behavior in a setting in which the individual is free to make his or her own honest or dishonest choices (Abe & Greene, 2014; Baumgartner, Fischbacher, Feierabend, Lutz, & Fehr, 2009; Ding, Gao, Fu, & Lee, 2013; Greene & Paxton, 2009; Sip et al., 2010; Sip et al., 2012). This initial work has suggested both commonalities and distinctions in the neural mechanisms underlying voluntary versus instructed deception. Specifically, both instructed and voluntary deception recruit neural networks involved in executive control, including the dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC). Voluntary deception, however, additionally recruits neural regions implicated in reward-related processing, specifically the ventral striatum (e.g., Abe & Greene, 2014; Baumgartner et al., 2009). Thus, voluntary deception appears to involve both executive control- and reward-related neural processes.

Understanding the neurocognitive processes associated with the opportunity to engage in voluntary dishonest choices not only involves identifying the implicated neural regions, but also when, and in what sequence, these neurocognitive processes occur along the temporal scale. In the present study, we employed scalp-recorded event-related potentials (ERPs), which provide millisecond temporal resolution, to examine the neural temporal dynamics underlying the opportunity to engage in voluntary deception. Specifically, we recorded ERPs during a coin-guessing task in which participants could win rewards if they correctly predicted the outcome of a coin flip (adapted from Greene & Paxton, 2009). Importantly, during certain trials, participants could freely engage in voluntary deception by overreporting the accuracy of their prediction in order to maximize their monetary winnings. ERPs were stimulus locked to the outcome phase of the trial in which participants first learn and evaluate the outcome of the coin flip (heads vs. tails). This outcome serves as a predictive cue for the participant as to whether or not they will win on the basis of their prior prediction. The outcome phase is also the moment participants evaluate their outcomes (e.g., loss or win), and make a decision whether or not to engage in voluntary deception in order to increase their earnings by claiming a correct prediction for trials in which they actually made an incorrect prediction. The present study focuses on three questions. First, we examine the neural temporal dynamics of executive control processes associated with the opportunity to engage in honest or dishonest behavior. Second, we examine the extent to which reward-related neural activity, as indexed by the reward prediction error signal to the outcome cue, is implicated in voluntary dishonest behavior. Finally, we examine whether executive control- and/or reward-related neural activity modulate (a) the overall likelihood of engaging in voluntary deceptive behavior and, (b) moral behavioral adjustment on a trial-by-trial basis (i.e., a greater likelihood of engaging in voluntary deception on a subsequent trial following a loss on a previous trial).

We predict that compared with no opportunity to deceive, having an opportunity to deceive will recruit executive control processes involving elevated conflict monitoring and subsequent conflict resolution. Here, conflict monitoring involves a rapid evaluation about whether response conflict is involved. The detection of conflict then initiates conflict resolution, which is exerted by a resource-limited control system to resolve conflict (Botvinick, Braver, Barch, Carter, & Cohen, 2001). For the present study, analyses of conflict monitoring focused on the fronto-central N2 and analyses of conflict resolution focused on the parietal P3 (we define the parietal P3 that occurs in response to an opportunity to deceive as the “executive control P3”). Elevated N2 has consistently been observed during tasks requiring conflict monitoring, such as the Flanker task and the go/no-go task. The augmented N2 likely reflects the activity of the dorsal ACC associated with the detection of conflict and response monitoring processes (Nieuwenhuis, Yeung, van den Wildenberg, & Ridderinkhof, 2003; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004; van Veen & Carter, 2002; Yeung & Cohen, 2006). Most importantly, enhanced N2 has been found in both instructed deception and information concealment (Carrion, Keenan, & Sebanz, 2010; Gamer & Berti, 2010; Hu, Pornpattananangkul, & Rosenfeld, 2013). These findings suggest that either being deceptive or concealing information requires participants to actively monitor response conflict because two competing response tendencies are activated (honest vs. dishonest response). On the basis of these findings, we predicted that when compared with having no opportunity to deceive, having the opportunity to deceive (i.e., overreport one’s performance) will trigger two competing response tendencies: to honestly report one’s actual performance versus to dishonestly overreport one’s actual performance in an attempt to maximize earnings. Regardless of whether the ultimate behavior is honest or dishonest, we predict these competing response tendencies between making an honest versus dishonest choice will elicit an elevated N2.

The amplitude of the parietal executive control P3 can be modulated by a variety of factors, such as a stimulus’s subjective probability, participants’ devoted attentional resources and stimuli categorization uncertainty, and so forth (Donchin & Coles, 1988; Johnson, 1986, 1993). In particular, it has been found that the parietal executive control P3 is attenuated when experimental manipulations increase the executive control demands involved in the task. These manipulations include perceptual/memory load, dual-task, categorization ambiguity, and stimulus–response incompatibility, among others (Chen et al., 2008; Garcia-Larrea & Cezanne-Bert, 1998; Hu, Hegeman, Landry, & Rosenfeld, 2012; Kok, 2001; Lorist, Snel, Kok, & Mulder, 1996; Wickens, Kramer, Vanasse, & Donchin, 1983). Moreover, research has consistently demonstrated that instructed deception is associated with attenuated parietal executive control P3, which was taken as evidence

that deception involves executive control processes (e.g., Johnson, Barnhardt, & Zhu, 2003). More specifically, it has been argued that deception forces people to manage two competing response tendencies (lie vs. truth) in working memory and that engaging in deception necessitates suppressing the truth in order to give the deceptive responses (Hu, Wu, & Fu, 2011; Johnson et al., 2003; Johnson, Henkell, Simon, & Zhu, 2008). We predicted that having the opportunity to deceive would activate conflict resolution processes to resolve the conflict between competing response tendencies and that these processes would require cognitive effort. We further predicted that this increase in executive control processes would attenuate the executive control P3 when people had the opportunity to deceive relative to when they did not have the opportunity to deceive. Examining both the fronto-central conflict-sensitive N2 and the parietal executive control P3 would allow us to assess multiple psychological processes underlying the opportunity to engage in dishonest moral choices along the temporal scale. Specifically, we proposed that the N2 would inform our understanding of the initial conflict associated with the opportunity to engage in voluntary deception, and that the subsequent executive control P3 would inform our understanding of the higher-level regulatory processes to resolve this conflict. Lastly, a goal of the present study was to examine the extent to which individual differences in executive control-related neural processes modulate one's likelihood of engaging in voluntary deceptive behavior. We predicted that individuals who exhibited increased executive control-related neural activity (i.e., a larger N2 and/or a smaller executive control P3) during trials in which they had an opportunity to deceive would be more likely to engage in voluntary deceptive behavior. This prediction was based on previous studies showing that being dishonest engages executive control-related neural networks (Baumgartner et al., 2009; Greene & Paxton, 2009).

Relative to executive control processes such as conflict monitoring and conflict resolution, less is known about the role that reward-related processes play in voluntary deception. Given deception can serve the goal of maximizing gains, we predict that reward-related neural activity will play an important role in voluntary deception. To investigate this, we examined the medial frontal negativity (MFN; Gehring & Willoughby, 2002) locked to the outcome phase of the coin task, where participants learned whether they had accurately predicted the coin flip (e.g., heads vs. tails). The MFN (also known as the *feedback negativity*, FN, or *feedback error-related negativity*, fERN; Gehring & Willoughby, 2002; Miltner, Braun, & Coles, 1997) is a negative-going, fronto-central-distributed waveform that peaks approximately 200–400 ms after the presentation of negative outcome/feedback, relative to positive outcome/feedback. Research on reward processing has demonstrated that midbrain dopamine neurons encode prediction errors. When a reward prediction is

violated, such that the outcome is worse than desired or does not meet one's goal, the firing rate of midbrain dopamine neurons will temporarily drop, generating a negative reward prediction error (Schultz, 2002). Such reward prediction error signals from the midbrain dopamine system subsequently modulate the activity of the anterior cingulate cortex (ACC), reflected by a scalp-recorded MFN (Holroyd & Coles, 2002; Holroyd & Yeung, 2012). Indeed, the MFN has been proposed as a neural proxy of prediction error and has been studied intensively in the context of decision making or reinforcement-learning tasks that involve gains/losses and trial-by-error learning (for reviews, see Holroyd & Coles, 2002; Nieuwenhuis, Yeung, Holroyd, Schurger, & Cohen, 2004; Walsh & Anderson, 2012). Source localization analyses have suggested that the MFN is likely produced in the ACC (Gehring & Willoughby, 2002; Miltner et al., 1997; but see Carlson, Foti, Mujica-Parodi, Harmon-Jones, & Hajcak, 2011), consistent with the computation model of the MFN (Holroyd & Coles, 2002). The MFN has been demonstrated to update the contingency between behavioral choice and reward outcome, and to guide one's subsequent behavior to pursue wins and/or avoid losses on a trial-by-trial basis (Cohen & Ranganath, 2007; van der Helden, Boksem, & Blom, 2010; Walsh & Anderson, 2011; for a review, see Walsh & Anderson, 2012).

These characteristics of the MFN allowed us to examine the role that reward-related neural activity, and specifically the reward prediction error signal, plays in voluntary deception choices involving possible gains. We hypothesized that an incorrect prediction of the outcome of the coin flip (e.g., the outcome was heads when the prediction was tails) would generate an elevated reward prediction error signal reflected in a larger MFN to the incorrect outcome cue relative to the correct outcome cue. On the basis of the logic that individuals with elevated reward-related neural activity (i.e., an elevated reward prediction error signal) might be more willing to deceive for personal gain, we further predicted that participants with a particularly large MFN (i.e., a larger reward prediction error) to the incorrect outcome cue would be more likely to engage in voluntary deception to maximize gains. Finally, given the role that reward prediction error plays in adjusting one's behavior to maximize performance-based gains, we predicted that a larger MFN during trials in which participants had no opportunity to engage in voluntary deception would be associated with a greater likelihood of engaging in voluntary deception on *subsequent* trials in which there was an opportunity to deceive (i.e., behavioral adjustment).

In addition to the MFN, recent work has indicated that the P3 in response to gain/loss cues (what we refer to as the *reward P3*) is also implicated in reward processing. For example, the reward P3, which follows the MFN, is larger to both unexpected outcomes (Hajcak, Holroyd, Moser, & Simons, 2005; von Borries, Verkes, Bulten, Cools, & de Bruijn,

2013) and rewards of greater magnitudes (Yeung & Sanfey, 2004). Moreover, the reward P3 is larger during tasks that involve active choices rather than mere observations, which may reflect a higher level of attentional engagement associated with making active choices (Yeung, Holroyd, & Cohen, 2005). Because the reward P3 may reflect one's attentional engagement to outcome cues, we predicted that a larger reward P3 response to gain than to loss cues (i.e., higher reward processing) would predict a greater likelihood of engaging in voluntary deception. Moreover, recent studies have also linked the reward P3 with behavioral adjustment strategies in reinforcement-learning tasks (Chase, Swainson, Durham, Benham, & Cools, 2011; Martin, Appelbaum, Pearson, Huettel, & Woldorff, 2013; von Borries et al., 2013). On the basis of these findings, we also tested in the present study whether the MFN or the P3 can predict moral behavioral adjustment strategies to maximize one's gains. Collectively, these analyses have important implications for understanding the extent to which reward-related neural activity influences one's likelihood of engaging in both overall voluntary deception and moral behavioral adjustment on a trial-by-trial basis (i.e., switching from honest to dishonest behavior to maximize earnings).

Materials and methods

Participants

Twenty-six right-handed participants (17 females, nine males; age: 18–22 years) at Northwestern University received partial course credit for their participation. This sample size is consistent with previous studies that employed a similar voluntary dis/honest choice paradigm (e.g., Ding et al., 2013, $N = 18$; Abe & Greene, 2014, $N = 28$; Shalvi & De Dreu, 2014, $N = 30$). Six additional participants were excluded due to excessive blinks and artifacts. Participants were screened for neurological history and had normal or corrected-to-normal vision. The study was approved by the Northwestern Institutional Review Board and participants provided written consents prior to the experiment.

Procedure

Participants were seated in an electromagnetically shielded, sound-attenuated booth in front of a computer monitor. Following electrode application, participants completed an incentive-based coin-guess task (adapted from Greene & Paxton, 2009) to examine their neurocognitive profiles of voluntary moral decision makings. Following the coin-guess task, participants were debriefed and dismissed.

Coin-guess task Participants were instructed that they were to predict the outcome of each coin flip and that they would win raffle tickets for correct guesses (Fig. 1). Specifically, participants were informed they could earn either three or five raffle tickets for each correct guess on a given trial, and that they would not win any raffle tickets for a given trial if their guess was incorrect. The cumulative tickets the participants gained over the course of the task were then placed into a lottery from which they could win one of three \$25 Amazon gift cards (i.e., the better their performance, the higher the probability they would win one of the gift cards). The task consisted of 200 trials and lasted approximately 40 min. Half of the trials (i.e., 100) were no-opportunity-to-deceive trials (NoOp). In these NoOp trials, the word "RECORD" was presented for 3,000 ms at the beginning of the trial. For NoOp trials, participants were instructed to record their prediction about the upcoming coin flip when the word "Press" appeared on the screen by pressing a button labeled "H" if they predicted "heads" and a button labeled "T" if they predicted "tails" for that particular trial. Requiring participants to record their prediction on NoOp trials prevented them from engaging in voluntary deception on these trials. The remaining trials (i.e., 100 trials) were opportunity-to-deceive trials (Op). In these Op trials, the word "RANDOM" was presented for 3,000 ms at the beginning of the trial. When the word "RANDOM" appeared on the screen, participants were instructed to make a prediction in their mind about the upcoming coin flip, but they did not have to record their prediction via external buttonpress. To justify this manipulation, participants were informed that previous research had suggested that people's ability to predict the future (i.e., a coin flip) might be better if they made the predictions privately to themselves (see also Greene & Paxton, 2009). During this coin-guessing task, Op trials and NoOp trials were intermixed and were presented to participants in a randomized order. To balance the motor output across Op and NoOp trials, participants were instructed to randomly press one of two buttons on the button box labeled "R" (random) for the Op trials when the word "Press" appeared on the screen. For all trials, the outcome of the coin flip (visual depiction of a head vs. a tail) was next presented for 2,000 ms. Epochs of electroencephalographic (EEG) data were stimulus-locked to the outcome of the coin flip for all analyses. The question "Correct?" next appeared on the monitor, prompting participants to indicate whether or not their prediction had been accurate. For NoOp (i.e., RECORD) trials, participants were instructed to press either a "YES" button (i.e., correct prediction) or a "NO" button (i.e., incorrect prediction) on the basis of their previously recorded responses. For Op (i.e., RANDOM) trials, participants were instructed to press either the "YES" button (i.e., correct prediction) or the "NO" button (i.e., incorrect prediction) on the basis of their prior, nonrecorded predictions. The fact that participants did not record their predictions during Op trials afforded them the

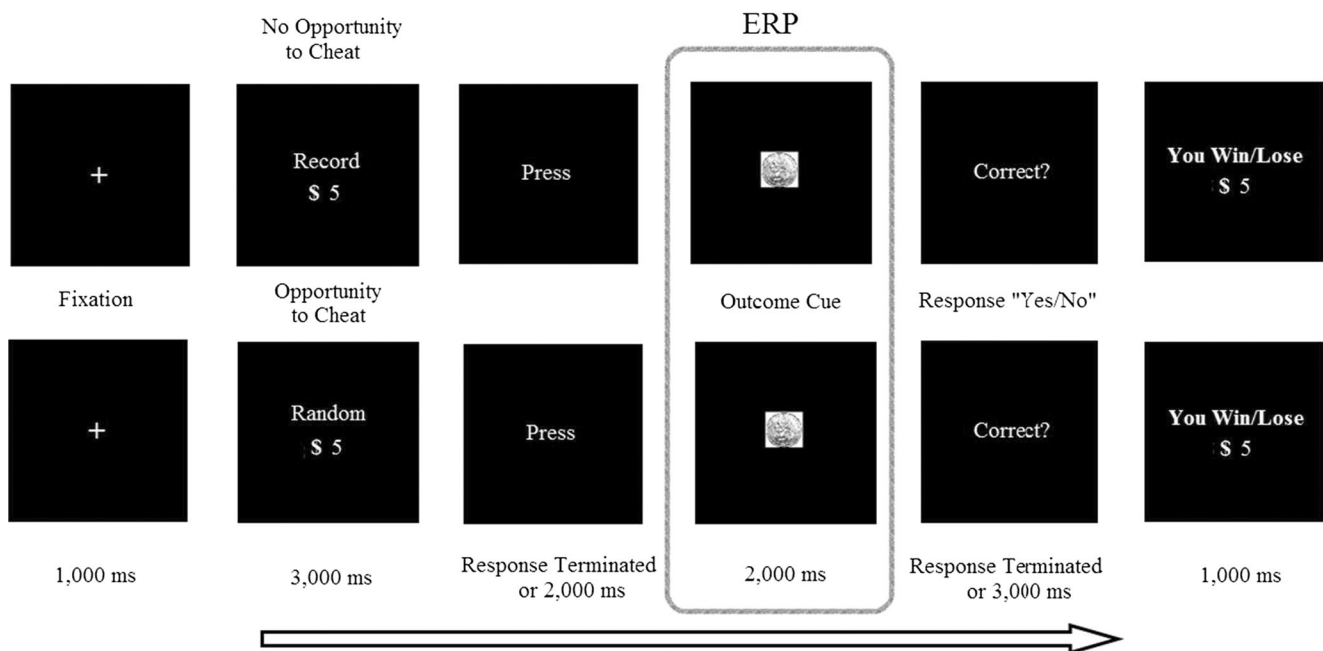


Fig. 1 Task structure of the coin-guess task. Participants made predictions for either three or five raffle tickets (only the \$5 trial is shown here as an exemplar). Note for the “Press” screen participants were instructed during NoOp trials to enter their prediction as either “Heads” or “Tails” by pressing either the “H” or the “T” key. During Op trials, participants

were instructed to randomly press one of the two “R” keys to control for motor activity. ERPs were locked to the outcome cue screen for both Op/NoOp trials (for executive control analyses) and Win/Loss trials, which were coded on the basis of the later report during the response screen (correct/win vs. incorrect/loss, for the reward-processing analyses)

opportunity to overreport their accuracy rate in order to increase their possible winnings (i.e., voluntary deception). During debriefing, all participants reported that they were aware that they could cheat during the Op trials.

Operationalizing overall deception Because participants were instructed to predict the outcome of a coin flip in the coin-guessing task, the expected reported accuracy for honest participants should be comparable across Op and NoOp trials (i.e., ~50 %). Accordingly, a higher reported accuracy for Op than for NoOp trials would suggest a higher likelihood of dishonesty. Thus, by comparing claimed wins between the Op and NoOp trials (reported wins in the Op trials minus reported wins in the NoOp trials), we could infer the likelihood that a participant had engaged in overall voluntary deception during the coin-guess task. A higher difference score

would suggest a higher likelihood of voluntary deception. One could also use 50 % rather than the actual accuracy in the NoOp trials as the baseline, because people’s predictions in the NoOp trials should be 50 %. However, examination of the data suggested that there was a wide range of variance across individuals for their prediction accuracy in the NoOp trials (38 %–60 %; see Fig. 2). We therefore calculated the Op-minus-NoOp performance difference to account for individual differences in prediction accuracy during the NoOp trials.

Operationalizing trial-by-trial moral behavioral adjustment To examine participants’ behavioral adjustment on a trial-by-trial basis, we measured the probability of reported wins during Op trials that were preceded by a NoOp_Loss trial. To ensure that this behavioral adjustment was due to a previous loss rather than to a general craving for wins, we also

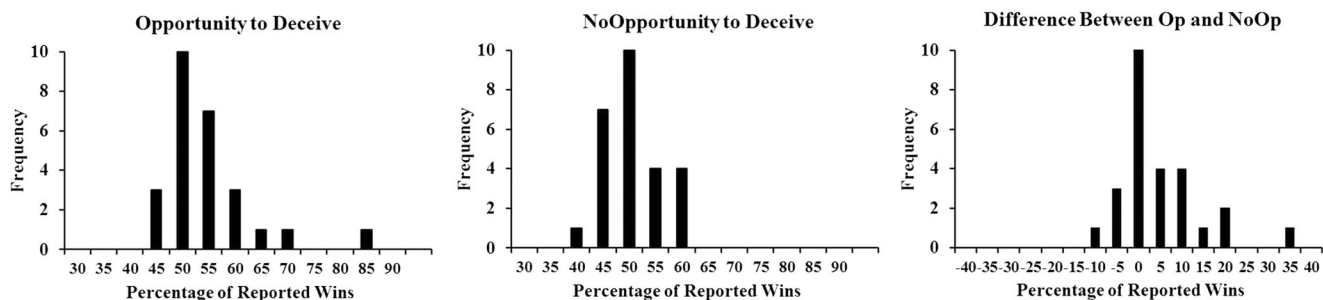


Fig. 2 Distribution of reported wins as percentages, for (a) opportunity-to-deceive (Op) trials, (b) no-opportunity-to-deceive (NoOp) trials, and (c) differences between Op and NoOp trials

calculated the probability of reported wins during Op trials that were preceded by NoOp_Win trials. The behavioral adjustment score was calculated as the difference between the proportion of reported wins during Op trials following NoOp_Loss trials and the proportion of reported wins during Op trials following NoOp_Win trials ($\text{NoOp_LossOpWin} - \text{NoOp_WinOpWin}$). A higher behavioral adjustment score would indicate that a participant was more likely to make a dishonest decision following an honest loss than following an honest win. That is, participants with a high behavioral adjustment score were more likely to engage in voluntary deception on a *subsequent* trial, given a previous honest loss.

EEG recordings and analyses

EEG data were collected from 19 electrodes (FP1/2, Fz, F3/4, F7/8, FCz, Cz, C3/4, C7/8, CPz, Pz, P4/5, and T6/8) grounded at AFz. The online reference was the left mastoid and data were recorded from the right mastoid, enabling computation of an offline linked mastoid reference (impedances < 5 k Ω). The data were filtered (DC–100 Hz), amplified, and digitized (500 Hz).

During offline analyses, eye blinks were first corrected with principal components analysis algorithms implemented in NeuroScan EDIT software (Neuroscan, Inc.). Saccades and movement-related artifacts were removed manually. The EEG data were then high-pass filtered (0.1 Hz, 24 dB). For all analyses, epoched EEG data were stimulus-locked to the onset of the outcome of the coin flip (visual depiction of head vs. tail). A linear detrend algorithm on a large epoch (from –1,100 to 2,000 ms) was used to remove leftover drifts in the data. The ERP epochs were then trimmed (from –200 to 1,000 ms) and the prestimulus baseline (–200–0 ms) corrected. Epochs containing artifacts ($\pm 75 \mu\text{V}$) were rejected, and the remaining clean trials were low-pass filtered (30 Hz, 12 dB).

We stimulus-locked the EEG data to the onset of the outcome of the coin flips for two reasons. First, the outcome of the coin flip occurred right before the question “Correct?” appeared on the monitor, which was the moment that participants would need to decide whether to be honest or dishonest in reporting the accuracy of their prediction about the outcome of the coin toss (i.e., whether or not to engage in voluntary deception). Thus, this time window was relevant to the N2 and P3 executive control analyses. Second, this was the point at which participants first evaluated the outcome of the coin flip [incorrect prediction (loss) vs. correct prediction (win)], and thus was relevant to the MFN and P3 reward-processing analyses.

ERP measurements

All artifact-free EEG epochs were averaged into four categories: Op_Loss, Op_Win, NoOp_Loss, and NoOp_Win, given

our within-subjects 2 (Opportunity: Op vs. NoOp) \times 2 (Outcome: Win vs. Loss) design. Again, whereas the distribution of Op versus NoOp trials was established a priori by the experimenters, whether a trial was a Win or Loss trial was based on participants’ “YES” or “NO” responses to the “Correct?” screen in the coin-guessing task. Given that participants had the opportunity to cheat during the Op trials, an individual’s tendency to engage in dishonest behavior would influence the numbers of Op_Win and Op_Loss trials available for the analyses. However, because the participants in the present study were largely honest (see below for the behavioral results), the numbers of trials for Op_Loss and Op_Win were comparable: On average, 39, 40, 38, and 42 artifact-free trials were available for averaging for the NoOp_Loss, NoOp_Win, Op_Loss, and Op_Win trials, respectively.¹

On the basis of visual inspection of the grand averaged ERPs, we measured the mean ERP amplitude between 200 and 450 ms as the N2/MFN, and the mean ERP amplitude between 450 and 650 ms as the P3. We termed the ERPs that were responsive to Op versus NoOp trials as *executive control-related* ERPs, including the executive control N2 and executive control P3. We termed the ERPs that were responsive to Loss versus Win trials as *reward-related* ERPs, including the reward MFN and reward P3. These labels are based on recent theory and empirical evidence suggesting that, although the N2 and MFN may temporally overlap with each other, the N2 is more sensitive to conflict monitoring, whereas the MFN is more sensitive to reward processing (Baker & Holroyd, 2011; Warren & Holroyd, 2012).

Results

All within-subjects analysis of variance (ANOVA) results are reported with Greenhouse–Geisser-corrected *p* values when the assumption of sphericity was violated. Partial eta-squared values (η_p^2) are used to estimate the effect size in repeated measures ANOVAs, with 0.01, 0.06, and 0.14 considered small, medium, and large effect sizes, respectively. For individual-difference correlational analyses, correlation coefficients of 0.1, 0.3, and 0.5 are considered small, medium, and large effect sizes, respectively. All tests were two-tailed.

¹ Although participants were largely honest in the present study, six “dishonest” participants, who reported significantly more wins than losses (55 vs. 31), $t(5) = 3.996$, $p < .02$, were identified during the Op trials using binominal tests. Although the unequal numbers of Op_Win and Op_Loss trials from these six “dishonest” individuals could introduce confounds into quantifying the ERP amplitudes, this concern was mitigated by the fact that we used the mean amplitude (as opposed to the peak amplitude) to quantify ERP amplitudes. The mean amplitude is less susceptible to bias than is the peak amplitude for studies/conditions with unequal numbers of trials (Luck, 2014).

Behavioral assessment of overall dis/honest behavior

Figure 2 depicts the distributions of reported wins as percentages for the Op and NoOp trials, as well as the difference between the Op and NoOp trials. A paired-sample *t* test was conducted to compare the percentages of reported wins in the Op and NoOp trials. The results indicated that participants did report winning slightly more often in the Op than in the NoOp trials, although the difference was not significant (Op vs. NoOp, mean \pm SE: 52.61 \pm 1.62 % vs. 49.50 \pm 0.97 %), $t(25) = 1.69$, $p > .10$.

We further analyzed participants' accuracy and RTs to the "Correct?" screens in the coin task. For accuracy, because participants made a Yes (Win) or No (Loss) decision based on predictions made in their mind in the Op trials, we could only analyze accuracy in the NoOp trials. The results showed that participants were highly accurate in the NoOp trials (mean \pm SE: 98 \pm 0.54 %), suggesting that they were following the instructions. A 2 (Opportunity Op vs. NoOp) \times 2 (Outcome: Loss vs. Win) within-subjects ANOVA was conducted on the RTs. This ANOVA yielded marginally significant main effects for both opportunity and outcome. With respect to the main effect of opportunity, participants had faster RTs during Op than during NoOp trials (Op vs. NoOp, mean \pm SE: 334.13 \pm 17.18 vs. 346.07 \pm 19.21 ms), $F(1, 25) = 3.95$, $p = .058$, $\eta_p^2 = 0.14$. With respect to the main effect of outcome, participants were slower in reporting a Loss outcome than a Win outcome (Loss vs. Win, mean \pm SE: 350.04 \pm 21.96 vs. 330.16 \pm 14.82 ms), $F(1, 25) = 3.54$, $p = .072$, $\eta_p^2 = 0.12$. Importantly, these marginally significant main effects were qualified by a significant Opportunity \times Outcome interaction: $F(1, 25) = 6.14$, $p < .02$, $\eta_p^2 = 0.20$. Follow-up paired *t* tests showed that during NoOp trials, it took participants longer to report a loss than to report a win (mean \pm SE: 363.35 \pm 22.90 vs. 328.79 \pm 16.54 ms), $t(25) = 3.15$, $p < .01$. In contrast, we found no RT differences between Loss and Win for Op trials (336.73 \pm 21.76 vs. 331.54 \pm 14.26 ms), $t(25) = 0.40$, $p > .70$; see Fig. 3. This suggests that when participants did not have an opportunity to deceive, they found it more difficult to report

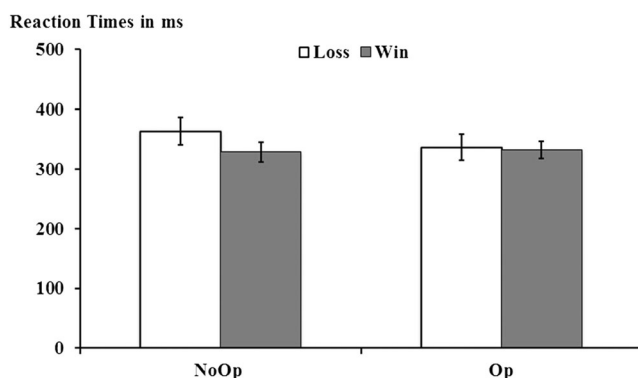


Fig. 3 Reaction times to the "Correct?" screen during the coin-guessing task. Error bars indicate one standard error

a loss outcome than a win outcome, which may have been due to loss aversion. However, when participants had an opportunity to deceive and could voluntarily avoid a loss by cheating, this difficulty in reporting losses disappeared.

To better understand the rates of dishonest behavior in the present study, we conducted binominal tests on each participant's accuracy for Op and NoOp trials. A significantly higher accuracy rate in the Op than in the NoOp trials indicated that a particular participant had likely engaged in dishonest behavior. The results showed that six of the 26 participants (23 %) in the present study could be classified as having engaged in extreme dishonest behavior, $ps < .01$. This rate of dishonest behavior was comparable with those from previous studies that had used either the same or a similar voluntary honest/dishonest moral choice paradigm (Abe & Greene, 2014; Gino & Ariely, 2012; Shalvi & De Dreu, 2014). This rate of dishonest behavior is smaller than that observed in Greene and Paxton (2009), given that the authors of that study intentionally selected participants who were highly likely to cheat, on the basis of a pilot behavioral test.

Behavioral assessment of trial-by-trial moral behavioral adjustment

Across all participants, 663 Op trials were preceded by NoOp_Loss, and 671 Op trials were preceded by NoOp_Win. To estimate the trial-by-trial moral behavioral adjustment, we calculated (1) the percentage of reported wins during Op trials following a previous NoOp Loss trial (NoOp_LossOpWin, mean \pm SE: 53.69 % \pm 0.02) and (2) the percentage of reported wins during Op trials following a previous NoOp Win trial (NoOp_WinOpWin, mean \pm SE: 52.59 % \pm 0.02). A paired *t* test showed no significant difference between these two measures, $t(25) = 0.39$, $p > .70$, suggesting that whether a previous NoOp trial had been a Loss or a Win trial did not modulate participants' decision making on subsequent Op trials. We then calculated each individual's moral behavioral adjustment score, defined as NoOp_LossOpWin – NoOp_WinOpWin, in terms of the percentage of reported wins. No relationship emerged between the moral behavioral adjustment score and the participant's overall deception (as estimated by the difference between claimed wins for Op and NoOp trials), $r(26) = -0.064$, $p > .76$. This suggests that distinct processes may underlie the likelihoods to engage in trial-by-trial moral behavioral adjustment versus overall voluntary deceptive behavior.

ERP results: Within-subjects ANOVAs on the N2/MFN (200–450 ms) and P3 (450–650 ms)

200–450 ms: N2/MFN A 2 (Opportunity: Op vs. NoOp) \times 2 (Outcome: Loss vs. Win) \times 5 (Electrode: Fz, FCz, Cz, CPz, Pz) within-subjects ANOVA on N2/MFN amplitude yielded a

significant main effect of opportunity, $F(1, 25) = 15.10$, $p < .001$, $\eta_p^2 = 0.38$, such that the N2 for the Op trials was significantly greater (i.e., more negative, for this negative wave) than that for the NoOp trials (mean \pm SE: 0.04 ± 1.02 vs. $2.16 \pm 1.23 \mu\text{V}$). We also observed a significant main effect of outcome, $F(1, 25) = 18.98$, $p < .001$, $\eta_p^2 = 0.43$, such that Loss cues elicited a larger MFN than did Win cues (mean \pm SE: 0.32 ± 1.00 vs. $1.88 \pm 1.10 \mu\text{V}$). Finally, there was a significant main effect of electrode, $F(2.14, 53.61) = 96.85$, $p < .001$, $\eta_p^2 = 0.80$. Pairwise comparisons showed that the N2/MFNs at Fz ($-2.71 \pm 0.99 \mu\text{V}$) and FCz ($-1.28 \pm 1.18 \mu\text{V}$) were more negative than the N2/MFNs at Cz ($1.29 \pm 1.07 \mu\text{V}$), CPz ($3.12 \pm 1.08 \mu\text{V}$), and Pz ($5.08 \pm 1.05 \mu\text{V}$, all $ps < .001$). This indicates a fronto-central distribution to the N2/MFN.

None of the two-way or three-way interactions were significant for the N2/MFN [Opportunity \times Outcome, $F(1, 25) = 0.66$, $p > .42$, $\eta_p^2 = 0.03$; Opportunity \times Electrode, $F(1.53, 38.31) = 0.57$, $p > .53$, $\eta_p^2 = 0.02$; Outcome \times Electrodes, $F(2.02, 50.43) = 0.33$, $p > .72$, $\eta_p^2 = 0.01$; Opportunity \times Outcome \times Electrode, $F(1.66, 41.44) = 1.78$, $p > .18$, $\eta_p^2 = 0.07$].

450–650 ms: P3 A 2 (Opportunity: Op vs. NoOp) \times 2 (Outcome: Loss vs. Win) \times 5 (Electrode: Fz, FCz, Cz, CPz, Pz) within-subjects ANOVA on P3 amplitudes yielded a significant main effect of opportunity, $F(1, 25) = 11.71$, $p < .005$, $\eta_p^2 = 0.32$, such that the P3 for Op trials was significantly smaller (i.e., less positive) than that for NoOp trials (mean \pm SE: 3.95 ± 0.84 vs. $5.75 \pm 0.97 \mu\text{V}$). There was also a significant main effect of electrode, $F(2.11, 52.71) = 93.79$, $p < .001$, $\eta_p^2 = 0.79$. Pairwise comparisons showed that the P3 amplitude was greater at Pz ($8.67 \pm 0.96 \mu\text{V}$) than at anterior electrodes: CPz ($7.55 \pm 0.97 \mu\text{V}$), Cz ($5.65 \pm 0.92 \mu\text{V}$), FCz ($2.39 \pm 1.01 \mu\text{V}$), and Fz ($-0.02 \pm 0.78 \mu\text{V}$), all $ps < .001$. We observed no main effect of outcome, $F(1, 25) = 1.98$, $p > .17$, $\eta_p^2 = 0.07$, such that Loss and Win cues generated comparable P3 amplitudes (mean \pm SE: Loss, 5.18 ± 0.89 vs. Win, $4.52 \pm 0.91 \mu\text{V}$).

A significant Outcome \times Electrode interaction emerged for the P3, $F(1.77, 44.27) = 3.78$, $p < .05$, $\eta_p^2 = 0.13$. Follow-up tests indicated that Loss cues elicited a significantly larger P3 than did Win cues at Fz (0.59 ± 0.84 vs. $-0.62 \pm 0.79 \mu\text{V}$), $t(25) = 2.65$, $p < .02$. However, the Loss P3 and Win P3 were comparable at FCz [2.87 ± 1.02 vs. $1.92 \pm 1.07 \mu\text{V}$, $t(25) = 1.88$, $p > .07$], Cz [5.95 ± 0.96 vs. $5.34 \pm 0.96 \mu\text{V}$, $t(25) = 1.23$, $p > .20$], CPz [7.61 ± 0.97 vs. $7.49 \pm 1.03 \mu\text{V}$, $t(25) = 0.22$, $p > .80$], and Pz [8.88 ± 0.93 vs. $8.46 \pm 1.07 \mu\text{V}$, $t(25) = 0.75$, $p > .46$]. None of the other two-way or three-way interactions were significant for the P3 [Opportunity \times Outcome, $F(1, 25) = 0.18$, $p > .60$, $\eta_p^2 = 0.01$; Opportunity \times Electrode, $F(2.53, 63.26) = 1.18$, $p > .32$, $\eta_p^2 = 0.05$; Opportunity \times Outcome \times Electrode, $F(1.72, 43.30) = 0.44$, $p > .60$, $\eta_p^2 = 0.02$].

Individual-differences analyses: Executive control- and reward-related neural activity as predictors of voluntary deception and moral behavioral adjustment

The lack of a significant Opportunity \times Outcome interaction for the N2, MFN, and P3 suggests that having or not having the opportunity to cheat (i.e., opportunity) and processing wins and losses (i.e., the outcome) influenced ERPs independently. This finding is in line with our a priori theorization that the opportunity condition (Op vs. NoOp) would primarily engage executive control processes and that the outcome condition (Loss vs. Win) would primarily engage reward processes. Accordingly, the subsequent individual-differences analyses focused on the two significant main effects [opportunity (Op vs. NoOp) and outcome (Win vs. Loss)]. To isolate the ERPs of interest, we computed two separate difference waves: one for Op-minus-NoOp, to reflect executive control processes, and one for Loss-minus-Win, to reflect reward processes (for similar difference-wave approaches, see Bress & Hajcak, 2013; Cohen & Ranganath, 2007; Yeung et al., 2005).

To quantify the executive control N2, we calculated the mean amplitude in the 200- to 350-ms time window for the Op-minus-NoOp difference wave (collapsing across Win and Loss trials) at FCz (Hu et al., 2011; Johnson et al., 2008). An elevated (i.e., more negative) N2 during Op than during NoOp trials (reflected by a larger N2 Op-minus-NoOp difference wave) suggests that a participant experienced greater conflict between honest and dishonest response tendencies during Op trials.

The executive control-related P3 was calculated as the mean amplitude for the 450- to 650-ms time window for the Op-minus-NoOp difference wave (collapsing across Win and Loss trials) at Pz (Hu et al., 2011). An attenuated executive control P3 during Op relative to NoOp trials (as reflected by a smaller executive control P3 Op-minus-NoOp difference wave) suggests that a participant devoted more effort to resolving conflict between honest and dishonest responses tendencies during Op trials (for the executive control-related ERPs, see Fig. 4a).

We calculated the MFN as the mean amplitude for the 300- to 450-ms time window for the Loss-minus-Win difference wave at FCz (collapsing across Op and NoOp trials). We chose FCz to be consistent with previous MFN literature (e.g., Cohen & Ranganath, 2007). The Loss-minus-Win MFN difference score served as an indicator of reward prediction error, with a larger MFN difference score (i.e., more negative) reflecting a larger reward prediction error.

We calculated the Loss-minus-Win P3 as the mean amplitude for the 450- to 650-ms time window for the Loss-minus-Win difference wave (collapsing across Op and NoOp trials) at Pz (Yeung & Sanfey, 2004). A larger value for this difference score indicated a larger P3 (i.e., a higher level of attention engagement) for Loss than for Win cues (for the reward-related ERPs, see Fig. 5a).

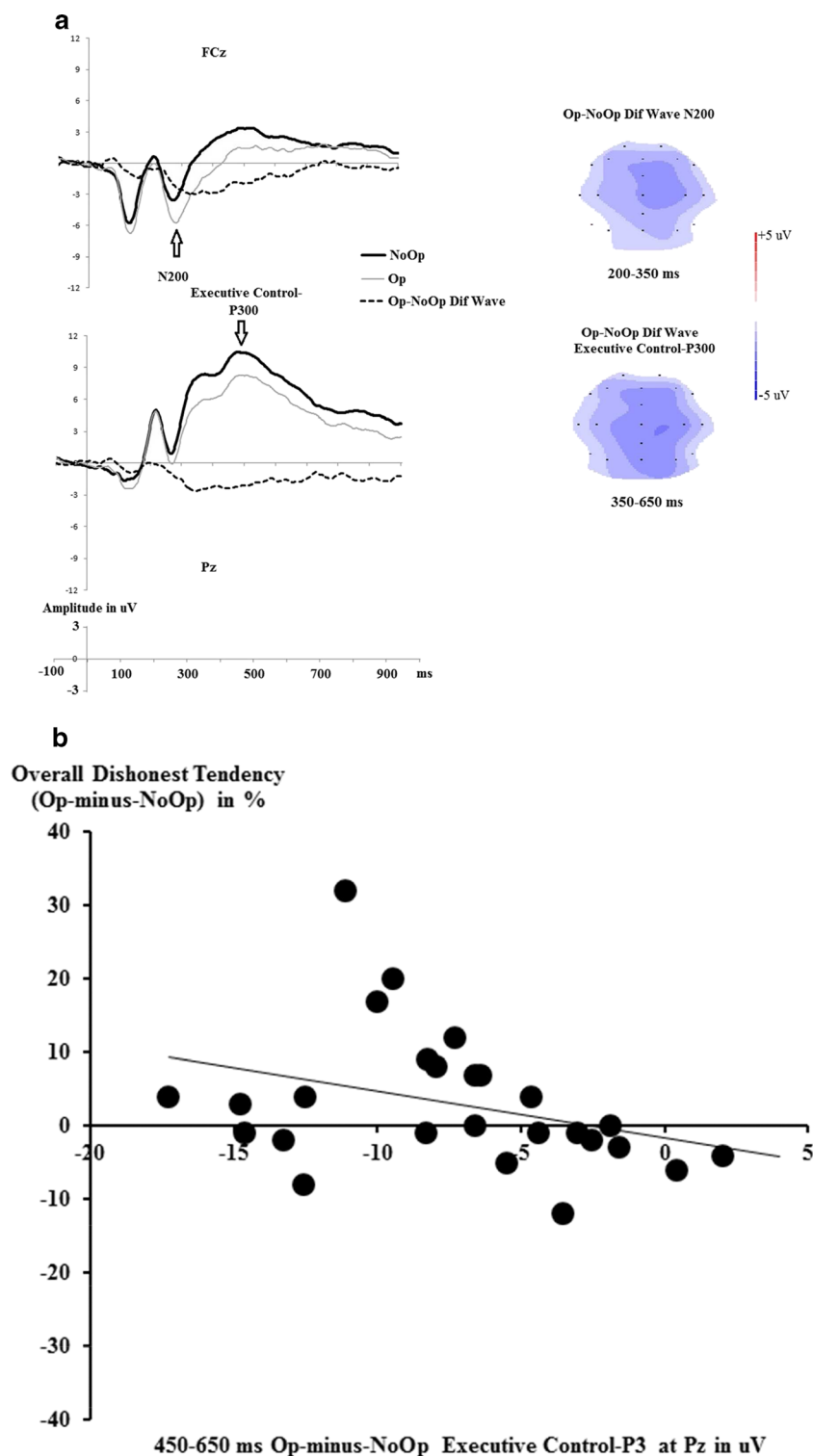


Fig. 4 (a) No-opportunity-to-deceive trials (NoOp) versus opportunity-to-deceive trials (Op) and the Op-minus-NoOp difference wave event-related potentials at fronto-central (FCz) and parietal (Pz) electrodes. The topographical maps depict the mean amplitudes of the N2 and the executive control P3 during each time window, based on the grand

average waveforms. (b) Relationship between the Op-minus-NoOp executive control P3 and participants' overall voluntary dishonest tendencies (percentage differences in reported wins between Op and NoOp trials)

Analyses of overall voluntary honest and dishonest decision making Prior to proceeding with the individual-differences

analyses, we first examined whether the distribution of participants' overall dishonest behavioral tendencies was normal,

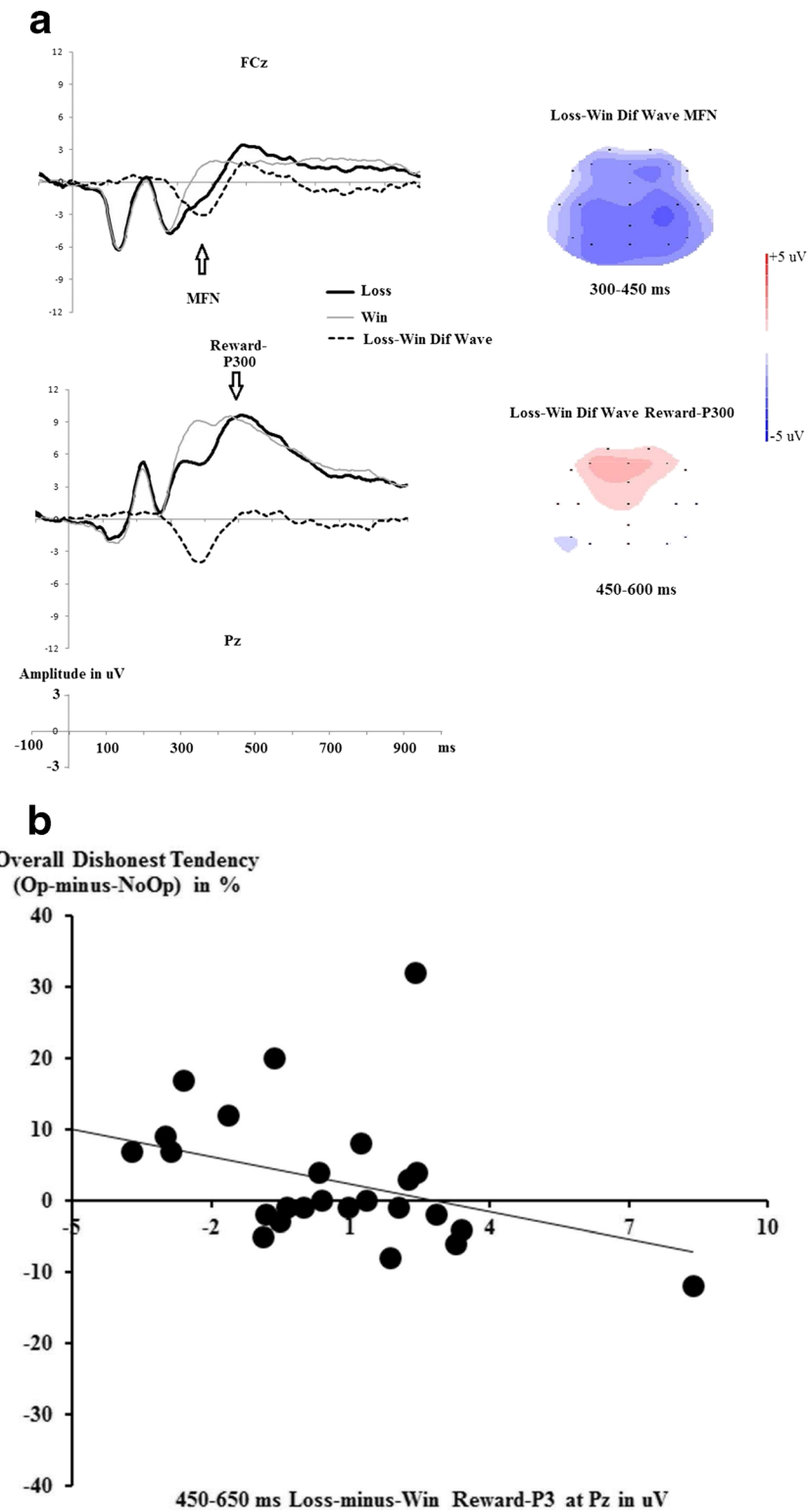


Fig. 5 (a) Loss trials (incorrect predictions of the coin toss) versus Win trials (correct predictions of the coin toss) and the Loss-minus-Win difference wave event-related potentials at fronto-central (FCz) and parietal (Pz) electrodes. The topographical maps depict the mean amplitudes of

the MFN and reward P3 during each time window, based on the grand average waveforms. (b) Relationship between the Loss-minus-Win reward P3 and participants' overall voluntary dishonest tendencies (percentage differences in reported wins between Op and NoOp trials)

using the Shapiro–Wilk test. A normal distribution of dishonest tendencies would justify the use of parametric tests—that is, the Pearson correlation. By contrast, a nonnormal distribution of participants' dishonest behavioral tendencies would indicate that nonparametric tests would be more appropriate—that is, the Spearman rank tests. The results indicated that participants' overall dishonest behavioral tendency was not normally distributed ($W = 0.91$, $df = 26$, $p = .022$). Therefore, we report results from Spearman rank-order tests for correlational analyses involving participants' overall dishonest behavioral tendencies.

In line with prediction, the Op-minus-NoOp executive control P3 at Pz was negatively correlated with participants' overall dishonest behavioral tendencies [$r_s(26) = 0.404$, $p < .045$]. This indicates that individuals with a smaller P3 amplitude during Op trials relative to NoOp trials were more likely to overreport their wins (i.e., to engage in overall deception; see Fig. 4b). We found no relationship between individual differences in the Op-minus-NoOp executive control N2 at FCz and participants' likelihood to engage in overall deception [$r_s(26) = 0.059$, $p > .70$].

Regarding reward-related neural activity (i.e., the Loss-minus-Win ERPs), and in line with prediction, the Loss-minus-Win reward P3 at Pz was negatively associated with participants' overall tendencies to engage in voluntary deception [$r_s(26) = -0.47$, $p < .02$]. Thus, a larger P3 to win than to loss was associated with a greater likelihood of engaging in voluntary deception (see Fig. 5b). There was no significant relationship between the Loss-minus-Win MFN difference score at FCz and participants' overall tendencies to engage in voluntary deception [$r_s(26) = 0.14$, $p > .50$].

Analyses of moral behavioral adjustment We next examined the relationship between trial-by-trial moral behavioral adjustment and reward-related neural activity (Loss-minus-Win ERPs) exclusively during the NoOp trials. This analysis was restricted to NoOp trials because, by definition, moral behavioral adjustment involved being dishonest following a NoOp loss trial. We did not examine the relationship between executive control-related neural activity and behavioral adjustment because this analysis would require the use of both Op and NoOp trials.

The Shapiro–Wilk normality test indicated that the distribution of moral adjustment was normal ($W = 0.97$, $df = 26$, $p > .60$), justifying the employment of parametric tests. Pearson correlational analyses suggested that neither the Loss-minus-Win MFN at FCz [$r(26) = -0.195$, $p > .30$] nor the Loss-minus-Win P3 at Pz [$r(26) = -0.017$, $p > .90$] predicted behavioral adjustment. We next conducted exploratory correlational analyses using all five midline electrodes. To reduce the chance of a false positive, we interpolated Fz/FCz/Cz as a fronto-central cluster and CPz/Pz as a centro-parietal cluster. Analyses indicated that a larger centro-parietal Loss-minus-

Win MFN during NoOp trials uniquely predicted a greater likelihood of trial-by-trial moral behavioral adjustment in order to maximize gains [$r(26) = -0.390$, $p < .05$; see Fig. 6a for the grand average ERPs and Fig. 6b for the scatterplot]. This correlation was not significant when the fronto-central MFN was used: $r(26) = -0.229$, $p > .25$. Finally, we found no relationship between the reward P3 and trial-by-trial moral behavioral adjustment [for the fronto-central P3, $r(26) = 0.002$, $p > .99$; for the centro-parietal P3, $r(26) = -0.062$, $p > .76$].

Discussion

In the present study, we investigated the neurocognitive processes associated with having the opportunity to engage in voluntary deception. In this context, an individual was free to make his or her own honest or dishonest moral choices. We reported four primary findings. First, when individuals had the opportunity to deceive, they experienced elevated conflict monitoring (a more-negative N2) and devoted increased cognitive resources to resolve this conflict (an attenuated executive control P3), as compared to when they did not have the option to deceive. Second, evaluating the outcome cue of the coin toss (win vs. loss) elicited reward-related neural activity. Specifically, an incorrect prediction of the outcome of the coin toss (e.g., the outcome was heads when the prediction was tails) generated an elevated reward prediction error signal, reflected in a larger MFN to the incorrect outcome cue than to the correct outcome cue. Third, elevated executive control-related neural activity reflecting conflict resolution (attenuated executive control P3) was associated with a greater likelihood of engaging in overall deceptive behavior. Finally, whereas elevated reward-related neural activity (reward P3) was associated with a greater likelihood of engaging in overall deceptive behavior, an elevated reward prediction error signal (MFN difference score) predicted increased trial-by-trial moral behavioral adjustment.

Opportunity to engage in voluntary deception recruits executive control-related neural activity

The opportunity-to-deceive (Op) versus no-opportunity-to-deceive (NoOp) contrast in the present study mimicked real-life scenarios in which individuals navigate the temptation to deceive or engage in dishonest behavior for personal gains. We reported that having the opportunity to engage in voluntary deception was associated with a more-negative N2 than was not having the opportunity to engage in such deception. The N2 has been observed during tasks involving conflict monitoring and response uncertainties, such as the go/no-go or flanker tasks, in which participants need to override one response tendency to execute an alternative, goal-directed response (for a review, see Folstein & Van Petten, 2008). The

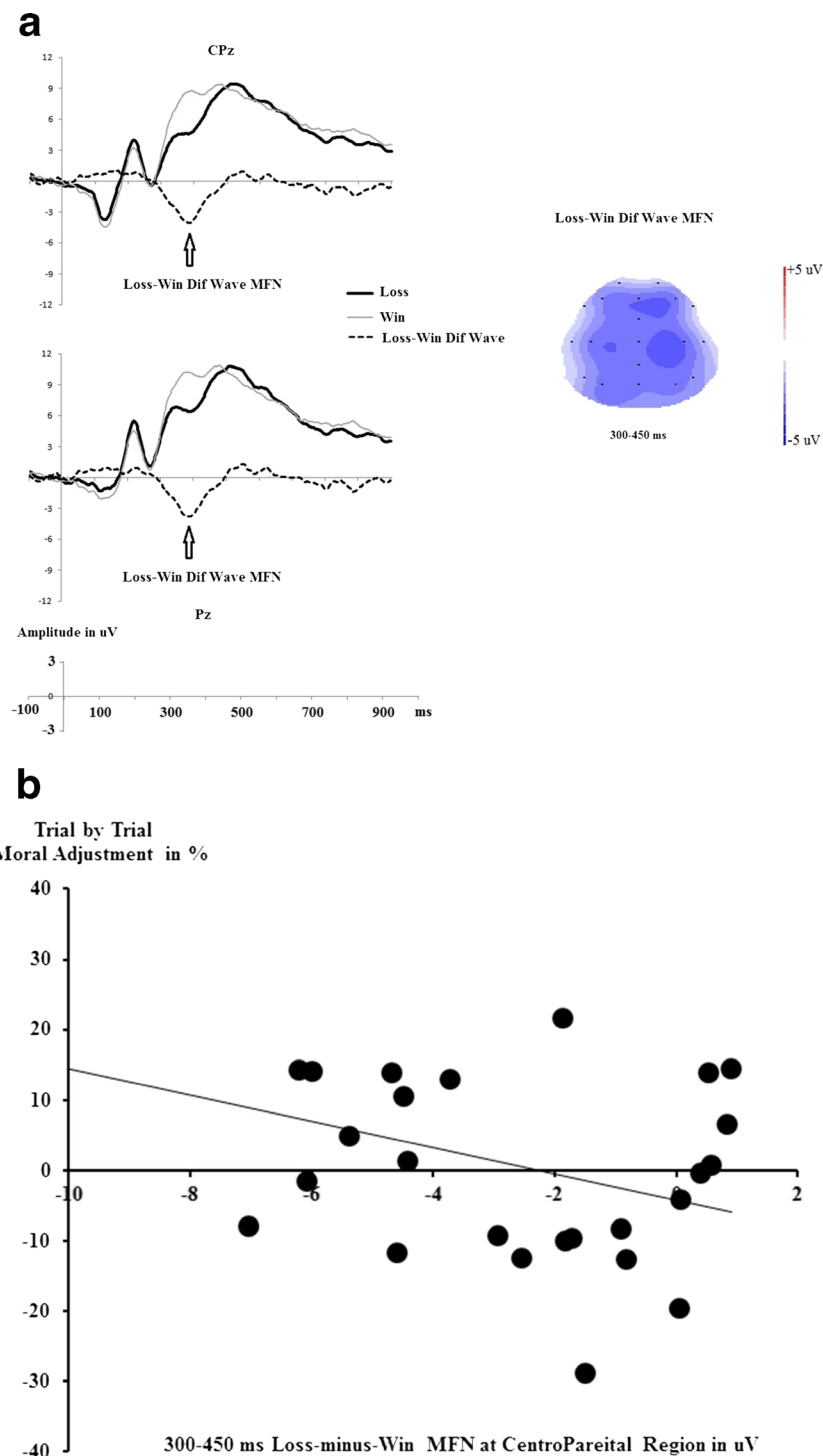


Fig. 6 Moral behavioral adjustment analyses: (a) Within NoOp trials, Loss trials (incorrect predictions of the coin toss) versus Win trials (correct predictions of the coin toss) and the Loss-minus-Win difference wave event-related potentials at centro-parietal (CPz and Pz) electrodes. The topographical map depicts the mean amplitude of the MFN during the

300- to 450-ms time window, based on the grand average waveforms. (b) Relationship between the Loss-minus-Win MFN (within NoOp trials) and participants' moral behavioral adjustments (in percentages). A larger Loss-minus-Win MFN reflects a larger reward prediction error

increased N2s in the present study suggest that having the opportunity to engage in voluntary deception triggers two competing response tendencies: to honestly report one's actual

performance versus to dishonestly overreport one's actual performance in order to maximize earnings. A larger (i.e., more-negative) N2 has also been documented in previous studies of

instructed deception, reflecting a conflict between participants' automatic tendency to engage in honest behavior and the instructions from an experimenter in order to engage in deceptive behavior (Hu et al., 2011). Considering that the N2 typically reflects activity in the ACC (van Veen & Carter, 2002), the results from the present study concur with fMRI studies that have shown elevated ACC activation during both instructed and voluntary deception (Abe & Greene, 2014; Baumgartner et al., 2009; Christ et al., 2009; Greene & Paxton, 2009). Collectively, these data suggest that both voluntary and instructed deception are associated with elevated conflict monitoring-related neural activity.

An advantage of the temporal resolution afforded by ERPs is that it allowed us to assess multiple executive control processes associated with having the opportunity to deceive along the temporal scale. In line with prediction, having the opportunity to engage in voluntary deception was associated with a more attenuated executive control-related P3 than was not having the opportunity to deceive. A similar attenuation of executive control P3s has been observed during studies of instructed deception (Hu et al., 2011; Johnson et al., 2003; Johnson et al., 2008). Drawing on this previous research on instructed deception, we argue that an attenuated executive control-related P3 reflects the engagement of regulatory processes to resolve the detected conflict (i.e., N2) between the tendencies to honestly report one's actual performance versus to engage in deceptive behavior to maximize earnings. In line with this argument, theories of cognitive control have stated that when conflict is detected within the ACC, it alerts higher cognitive control systems (such as the DLPFC) to expend resources in an attempt to resolve such conflict (Botvinick et al., 2001). The reduced executive control-related P3 in this context likely reflects the engagement of these higher-level executive control processes to resolve the detected conflict and implement the behavioral response. The present study is the first to report an attenuated executive control-related P3 when people have the opportunity to engage in voluntarily deception, suggesting that both instructed and voluntary deception engage conflict-resolution-related neural activity aimed at resolving the tension between honest and deceptive behavior.

Opportunity to engage in voluntary deception recruits reward-related neural activity

Previous research has highlighted the involvement of motivational/reward-related processes in voluntary moral decision making (Abe & Greene, 2014; Baumgartner et al., 2009; Ding et al., 2013). By investigating ERPs locked to the outcome of the coin flip, we have provided the first electrophysiological evidence regarding how reward processing, and in particular the reward prediction error signal, contributes to voluntary moral decision making.

In line with prediction, cues that were indicative of loss (incorrect prediction of the outcome of the coin flip) elicited a larger MFN (i.e., more negative) than did cues indicative of win (correct prediction of the outcome of the coin flip). The reinforcement-learning model of the MFN proposes that the MFN reflects the impact of reward prediction error signals from the midbrain dopamine system on the ACC. The ACC then uses this signal to improve one's behavior to obtain a desired goal and maximize one's personal gains (Holroyd & Coles, 2002; Walsh & Anderson, 2012). The MFN has been observed in numerous previous studies of reward processing and reinforcement learning in nonmoral domains (Bress & Hajcak, 2013; Foti, Weinberg, Dien, & Hajcak, 2011; Gehring & Willoughby, 2002; Hajcak et al., 2005; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; Miltner et al., 1997; Nieuwenhuis, Holroyd, Mol, & Coles, 2004; Yeung et al., 2005). The present study is the first to report a reward prediction error signal in the context of moral decision making when such (im)moral behavior was associated with gains/losses. Collectively, these findings suggest that subsequent research on voluntary moral decision making should assess both executive control- and reward-related neural activity, since reward processes play an important role in modulating voluntary deceptive behavior.

It should be noted that recent empirical evidence puts in question the assumption that the MFN encodes reward prediction errors. In particular, the MFN was enhanced when an anticipated pain was omitted (i.e., a rewarding event; see Talmi, Atkinson, & El-Dereby, 2013). The MFN thus appears to encode salience prediction errors (Talmi et al., 2013). The results in our study, however, are consistent with both the reward and salience perspectives of the MFN: That is, an experienced loss can be perceived as more salient than an experienced win, which is reflected by an enhanced MFN.

Examining the P3 that follows the MFN to win versus loss cues (what we refer to as the *reward P3*) provides a window into subsequent stages of reward processing in the context of voluntary moral choices. Consistent with previous research (von Borries et al., 2013; Yeung et al., 2005; Yeung & Sanfey, 2004), we did not find that the outcome valence modulated the reward P3. Despite the lack of a main effect of valence, individual differences in the reward P3 may reflect participants' different levels of attentional engagement while processing gain and loss outcomes (Yeung et al., 2005). In line with this perspective, the reward P3 predicted one's overall dishonest tendency, as we report below.

Individual differences in overall dishonest tendencies and moral behavioral adjustment

Analyses of individual differences allow us to examine the extent to which executive control-related and/or reward-related processes modulate one's propensity to engage in

voluntary deception. In addition to mechanistic implications, examining this topic has important practical implications, because it can provide insight into which specific processes should be targeted in the promotion of honest and ethical behavior.

Here we operationalized voluntary deception at two different levels of analysis. First, we obtained an overall measure of deception for each individual, defined as the likelihood of overclaiming one's gains throughout the task. Second, we obtained a trial-by-trial moral behavioral adjustment estimate for each individual. This was defined as the likelihood of one making a dishonest or deceptive decision on the present trial if that trial followed an honest loss trial. Thus, participants with a high behavioral adjustment score were more likely to engage in voluntary deception on a *subsequent* trial, given a previous honest loss.

With respect to overall dishonest tendencies, we reported here for the first time that both executive control-related and reward-related P3s modulate one's likelihood of engaging in voluntary deceptive behavior. A reduced executive control-related P3 in this context likely reflects the engagement of conflict resolution during Op trials to resolve the conflict between honest and dishonest response tendencies. The present study reports that individuals with a smaller P3 amplitude during Op than during NoOp trials were more likely to over-report their wins (i.e., to engage in overall deception). This finding suggests that the more that people deliberate and resolve conflict between honest and dishonest responses, the higher the likelihood that they will violate moral norms and exhibit morally questionable behavior. This result is consistent with previous neuroimaging research showing that activity in the executive control neural network (e.g., DLPFC, parietal lobe) is positively correlated with actual cheating behavior (Baumgartner et al., 2009; Greene & Paxton, 2009). Collectively, these findings support the hypothesis that (1) being honest or adhering to moral norms is a default behavioral tendency, and (2) cheating or violating moral norms to pursue self-interests actually requires active cognitive control (Baumgartner et al., 2009; Greene & Paxton, 2009; Rand, Greene, & Nowak, 2012; Sip et al., 2010).

Another neural indicator that predicted one's overall dishonest tendency is the reward P3. Here, a larger reward P3 to Win than to Loss cues was associated with a greater likelihood of engaging in overall voluntary deception. Given the P3's close relationship with attention allocation (Donchin & Coles, 1988; Johnson, 1986) and the fact that the P3 is associated with reward magnitude in gambling tasks (Yeung & Sanfey, 2004), this finding suggests that individuals who were more sensitive to Win cues were more likely to engage in voluntary deception to maximize personal gain.

Regarding one's trial-by-trial moral behavioral adjustment, we reported that the larger the magnitude of the Loss-minus-Win MFN during previous NoOp trials (i.e., larger reward

prediction errors), the more likely participants were to report a win on the subsequent Op trial. Because participants' responses on the NoOp trials were always honest, the higher reported wins on the subsequent Op trials suggest that participants switched their response tendencies to be more dishonest once they were given the opportunity. As is argued in the reinforcement-learning theory of the MFN, the reward prediction error signal from the midbrain dopamine system alerts the ACC and the PFC to update the stimulus/response–reinforcement contingency, and to adjust behavior for optimal outcomes such as reward (Holroyd & Coles, 2002). In accordance with this perspective, the MFN predicts a range of feedback-based choices and learning efficiencies on a trial-by-trial basis in nonmoral tasks (Cohen & Ranganath, 2007; van der Helden et al., 2010). The present study has been the first to illustrate that the reinforcement-learning signal MFN is involved in trial-by-trial moral behavioral adjustment to maximize personal gains.

Collectively, these results reveal an interesting dissociation. Whereas the early (300–450 ms) Loss-minus-Win MFN during NoOp trials predicts one's moral behavioral adjustment on a trial-by-trial basis, the later (450–650 ms) executive control P3 and reward P3 are associated with one's overall dishonest tendency. It is possible that because the MFN provides rapid and initial evaluations of ongoing performance (Hajcak, Moser, Holroyd, & Simons, 2006), it is more likely to drive *trial-by-trial* deceptive behavior instead of modulating *overall* dishonest response tendencies. In contrast, given that maintaining a general dishonest behavioral tendency requires increased cognitive control, it seems appropriate that this overall dishonest tendency would be uniquely predicted by late P3 activity and not by earlier MFN or N2 activity. Future research is needed to more fully examine these hypotheses and the possible dissociation in neural processes underlying *trial-by-trial* versus *overall* deceptive behavior.

Given the novelty and complexity of the coin-guessing task, it is not surprising that the Loss-minus-Win MFN observed in the present study shows some differences from the MFNs reported in previous literature (e.g., Gehring & Willoughby, 2002). First, the MFN in the present study occurred relatively late (~300 ms) as compared to previous studies, in which the MFN typically peaked between 200 and 300 ms (Gehring & Willoughby, 2002). The delayed MFN observed in the present study may reflect the nature of the coin-guess task. Upon the receipt of the outcome cue of the coin toss, participants need to maintain their predictions in working memory and to choose between a subsequent honest or dishonest response. The prolonged latency of the MFN may reflect the involvement of such complex cognitive processes (Baker & Holroyd, 2011).

Second, whereas we found that the centro-parietal MFN predicted moral behavioral adjustment, most previous studies had found that a fronto-central MFN modulated nonmoral

behavioral adjustment (for a review, see Walsh & Anderson, 2012). This discrepancy may reflect the nature of the coin task used in the present study, which assessed moral behavioral adjustment, as opposed to the tasks used in previous MFN research that had assessed more basic reward-based processes outside the scope of moral behavioral adjustment (Walsh & Anderson, 2012). Specifically, moral decisions likely involve elevated conflict detection and control-related processes, as compared to nonmoral choices (Greene & Paxton, 2009; Sip et al., 2010), and our posterior MFN may reflect such processes that are critical for moral behavioral adjustment (see Folstein & Van Petten, 2008). Future studies will be needed to replicate the present findings and investigate similarities and differences between the reward-related neural processes within and outside the scope of moral behavioral adjustment.

A potential limitation of the present study is that we cannot know for certain at which point during the trial of the coin-guess task a participant decided to make an honest or dishonest decision. For example, participants could have decided to cheat as soon as they learned that the next trial was an Op trial. In addition, a participant may have initially decided to make an honest decision but then changed his or her mind to be dishonest during the “Correct?” screen. Similarly, a participant may even have flipped between honest and dishonest decisions as the trial proceeded. These possibilities would make it precarious to focus analyses on the earlier stages of a trial, given the ambivalence that participants may have experienced at such points during the trial. In contrast, the Outcome stage of the trial clearly involves Loss versus Win outcome evaluations, and these outcome-locked ERPs can serve as ideal neural signals for reward processing, as was shown in previous literature (Walsh & Anderson, 2012). Thus, despite uncertainty as to when a participant precisely decided to make an honest or dishonest decision, we believe that the Outcome stage of the trial was the ideal time period in which to focus analyses in the present study.

Another limitation of the present study was the relatively small sample size. Although our sample size was consistent with previous research that had used a similar paradigm ($Ns = 18$ – 30 ; Abe & Greene, 2014; Ding et al., 2013; Shalvi & De Dreu, 2014), studies with this sample size may have difficulty detecting smaller effect sizes (i.e., Type II errors). This problem is rather common in cognitive, affective, and behavioral neuroscience (Lieberman & Cunningham, 2009). One practical solution to obtain a balance between Type I and Type II errors would be to aggregate research findings by meta-analyses (Lieberman & Cunningham, 2009). Thus, it will be important for future research to investigate whether the pattern of relationships observed in the present study can be replicated across studies and laboratories, and meta-analytic research should be used to provide more precise estimates of effect sizes.

In conclusion, in the present study we examined the neurocognitive processes and neural temporal dynamics involved in voluntary honest and dishonest moral decision making. This work has implications for a broad range of topics, including ethics, philosophy, neuroscience, and forensic science. We found that having the opportunity to cheat recruited both executive control (i.e., N2, executive control P3) and reward (i.e., MFN, reward P3) related neural activity. Moreover, early/late ERPs differentially predicted one’s trial-by-trial moral behavioral adjustment and overall dishonest tendency. This work sheds new light on the neurocognitive processes underlying both voluntary deception and moral decision making.

References

- Abe, N. (2011). How the brain shapes deception: An integrated review of the literature. *The Neuroscientist*, 17, 560–574. doi:10.1177/1073858410393359
- Abe, N., & Greene, J. D. (2014). Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *Journal of Neuroscience*, 34, 10564–10572.
- Abe, N., Suzuki, M., Mori, E., Itoh, M., & Fujii, T. (2007). Deceiving others: Distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *Journal of Cognitive Neuroscience*, 19, 287–295. doi:10.1162/jocn.2007.19.2.287
- Abe, N., Suzuki, M., Tsukiura, T., Mori, E., Yamaguchi, K., Itoh, M., & Fujii, T. (2006). Dissociable roles of prefrontal and anterior cingulate cortices in deception. *Cerebral Cortex*, 16, 192–199.
- Baker, T. E., & Holroyd, C. B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by feedback error-related negativity and N200. *Biological Psychology*, 87, 25–34. doi:10.1016/j.biopsycho.2011.01.010
- Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron*, 64, 756–770.
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624–652. doi:10.1037/0033-295X.108.3.624
- Bress, J. N., & Hajcak, G. (2013). Self-report and behavioral measures of reward sensitivity predict the feedback negativity. *Psychophysiology*, 50, 610–616.
- Carlson, J. M., Foti, D., Mujica-Parodi, L. R., Harmon-Jones, E., & Hajcak, G. (2011). Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: A combined ERP and fMRI study. *NeuroImage*, 57, 1608–1616. doi:10.1016/j.neuroimage.2011.05.037
- Carrion, R. E., Keenan, J. P., & Sebanz, N. (2010). A truth that’s told with bad intent: An ERP study of deception. *Cognition*, 114, 105–110.
- Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, 23, 936–946. doi:10.1162/jocn.2010.21456
- Chen, A., Xu, P., Wang, Q., Luo, Y., Yuan, J., Yao, D., & Li, H. (2008). The timing of cognitive control in partially incongruent categorization. *Human Brain Mapping*, 29, 1028–1039.

- Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta analyses. *Cerebral Cortex*, *19*, 1557–1566. doi:10.1093/cercor/bhn189
- Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience*, *27*, 371–378. doi:10.1523/JNEUROSCI.4421-06.2007
- Ding, X. P., Gao, X., Fu, G., & Lee, K. (2013). Neural correlates of spontaneous deception: A functional near-infrared spectroscopy (fNIRS) study. *Neuropsychologia*, *51*, 704–712.
- Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*, 357–374. doi:10.1017/S0140525X00058027
- Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology*, *45*, 152–170.
- Foti, D., Weinberg, A., Dien, J., & Hajcak, G. (2011). Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Temporospatial principal components analysis and source localization of the feedback negativity. *Human Brain Mapping*, *32*, 2207–2216. doi:10.1002/hbm.21182
- Gamer, M., & Berti, S. (2010). Task relevance and recognition of concealed information have different influences on electrodermal activity and event-related brain potentials. *Psychophysiology*, *47*, 355–364.
- Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: An fMRI investigation. *Cerebral Cortex*, *13*, 830–836.
- Garcia-Larrea, L., & Cezanne-Bert, G. (1998). P3, positive slow wave and working memory load: A study on the functional correlates of slow wave activity. *Electroencephalography and Clinical Neurophysiology*, *108*, 260–273.
- Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*, 2279–2282.
- Gino, F., & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, *102*, 445–459.
- Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, *106*, 12506–12511. doi:10.1073/pnas.0900152106
- Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*, 161–170.
- Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*, 148–154. doi:10.1016/j.biopsycho.2005.04.001
- Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709. doi:10.1037/0033-295X.109.4.679
- Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *NeuroReport*, *14*, 2481–2484.
- Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, *16*, 122–128. doi:10.1016/j.tics.2011.12.008
- Hu, X., Hegeman, D., Landry, E., & Rosenfeld, J. P. (2012). Increasing the number of irrelevant stimuli increases ability to detect countermeasures to the P300-based complex trial protocol for concealed information detection. *Psychophysiology*, *49*, 85–95.
- Hu, X., Pornpattananangkul, N., & Rosenfeld, J. P. (2013). N200 and P300 as orthogonal and integrable indicators of distinct awareness and recognition processes in memory detection. *Psychophysiology*, *50*, 454–464. doi:10.1111/psyp.12018
- Hu, X., Wu, H., & Fu, G. (2011). Temporal course of executive control when lying about self- and other-referential information: An ERP study. *Brain Research*, *1369*, 149–157.
- Johnson, R. (1986). A triarchic model of P300 amplitude. *Psychophysiology*, *23*, 367–384.
- Johnson, R. (1993). On the neural generators of the P300 component of the event-related potential. *Psychophysiology*, *30*, 90–97.
- Johnson, R., Jr., Barnhardt, J., & Zhu, J. (2003). The deceptive response: Effects of response conflict and strategic monitoring on the late positive component and episodic memory-related brain activity. *Biological Psychology*, *64*, 217–253.
- Johnson, R., Jr., Henkell, H., Simon, E., & Zhu, J. (2008). The self in conflict: The role of executive processes during truthful and deceptive responses about attitudes. *NeuroImage*, *39*, 469–482. doi:10.1016/j.neuroimage.2007.08.032
- Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*, 557–577.
- Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., . . . Childress, A. R. (2002). Brain activity during simulated deception: An event-related functional magnetic resonance study. *NeuroImage*, *15*, 727–732.
- Lee, T. M. C., Liu, H.-L., Tan, L.-H., Chan, C. C. H., Mahankali, S., Feng, C.-M., . . . Gao, J.-H. (2002). Lie detection by functional magnetic resonance imaging. *Human Brain Mapping*, *15*, 157–164.
- Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience*, *4*, 423–428. doi:10.1093/scan/nsp052
- Lorist, M. M., Snel, J., Kok, A., & Mulder, G. (1996). Acute effects of caffeine on selective attention and visual search processes. *Psychophysiology*, *33*, 354–361.
- Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
- Martin, R. S., Appelbaum, L. G., Pearson, J. M., Huettel, S. A., & Woldorff, M. G. (2013). Rapid brain responses independently predict gain maximization and loss minimization during economic decision making. *Journal of Neuroscience*, *33*, 7011–7019.
- Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a “generic” neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798. doi:10.1162/jocn.1997.9.6.788
- Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. H. (2004). Reinforcement-related brain potentials from medial frontal cortex: Origins and functional significance. *Neuroscience & Biobehavioral Reviews*, *28*, 441–448. doi:10.1016/j.neubiorev.2004.05.003
- Nieuwenhuis, S., Yeung, N., Holroyd, C. B., Schurger, A., & Cohen, J. D. (2004). Sensitivity of electrophysiological activity from medial frontal cortex to utilitarian and performance feedback. *Cerebral Cortex*, *14*, 741–747.
- Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., & Ridderinkhof, K. R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. *Cognitive, Affective, & Behavioral Neuroscience*, *3*, 17–26. doi:10.3758/CABN.3.1.17
- Priori, A., Mamelì, F., Cogiamanian, F., Marceglia, S., Tiriticco, M., Mrakic-Sposta, S., . . . Sartori, G. (2008). Lie-specific involvement of dorsolateral prefrontal cortex in deception. *Cerebral Cortex*, *18*, 451–455. doi:10.1093/cercor/bhm088
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*, 427–430.

- Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*, 443–447. doi:10.1126/science.1100301
- Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*, 241–263.
- Shalvi, S., & De Dreu, C. K. W. (2014). Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences*, *111*, 5503–5507. doi:10.1073/pnas.1400724111
- Sip, K. E., Lyngé, M., Wallentin, M., McGregor, W. B., Frith, C. D., & Roepstorff, A. (2010). The production and detection of deception in an interactive game. *Neuropsychologia*, *48*, 3619–3626. doi:10.1016/j.neuropsychologia.2010.08.013
- Sip, K. E., Roepstorff, A., McGregor, W., & Frith, C. D. (2008). Detecting deception: The scope and limits. *Trends in Cognitive Sciences*, *12*, 48–53.
- Sip, K. E., Skewes, J. C., Marchant, J. L., McGregor, W. B., Roepstorff, A., & Frith, C. D. (2012). What if I get busted? Deception, choice, and decision-making in social interaction. *Frontiers in Neuroscience*, *6*, 58. doi:10.3389/fnins.2012.00058
- Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *NeuroReport*, *12*, 2849–2853.
- Talmi, D., Atkinson, R., & El-Derey, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction errors. *Journal of Neuroscience*, *33*, 8264–8269. doi:10.1523/JNEUROSCI.5695-12.2013
- van der Helden, J., Boksem, M. A. S., & Blom, J. H. G. (2010). The importance of failure: Feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex*, *20*, 1596–1603. doi:10.1093/cercor/bhp224
- van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*, *14*, 593–602. doi:10.1162/08989290260045837
- von Borries, A. K., Verkes, R. J., Bulten, B. H., Cools, R., & de Bruijn, E. R. (2013). Feedback-related negativity codes outcome valence, but not outcome expectancy, during reversal learning. *Cognitive, Affective, & Behavioral Neuroscience*, *13*, 737–746. doi:10.3758/s13415-013-0150-1
- Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, *108*, 19048–19053.
- Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*, *36*, 1870–1884. doi:10.1016/j.neubiorev.2012.05.008
- Warren, C. M., & Holroyd, C. B. (2012). The impact of deliberative strategy dissociates ERP components related to speeded responding vs. reinforcement learning. *Frontiers in Decision Neuroscience*, *6*, 43. doi:10.3389/fnins.2012.00043
- Wickens, C., Kramer, A., Vanasse, L., & Donchin, E. (1983). Performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information-processing resources. *Science*, *221*, 1080–1082.
- Yeung, N., & Cohen, J. D. (2006). The impact of cognitive deficits on conflict monitoring: Predictable dissociations between the error-related negativity and N2. *Psychological Science*, *17*, 164–171. doi:10.1111/j.1467-9280.2006.01680.x
- Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*, 535–544. doi:10.1093/cercor/bhh153
- Yeung, N., & Sanfey, A. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, *24*, 6258–6264. doi:10.1523/JNEUROSCI.4537-03.2004