Neural basis of the undermining effect of monetary reward on intrinsic motivation

Kou Murayama[a,1], Madoka Matsumoto[b,c,d], Keise Izuma[b,c], and Kenji Matsumoto[b,d,1]

[a]Department of Psychology, University of Munich, 80802 Munich, Germany; [b]Brain Science Institute, Tamagawa University, Tokyo 194-8610, Japan; [c]Japan Society for the Promotion of Science, Tokyo 102-8471, Japan; and [d]Cognitive Brain Mapping Laboratory, RIKEN Brain Science Institute, Saitama 351-0198, Japan

Edited by Edward E. Smith, Columbia University, New York, NY, and approved October 20, 2010 (received for review September 8, 2010)

Contrary to the widespread belief that people are positively motivated by reward incentives, some studies have shown that performance-based extrinsic reward can actually undermine a person's intrinsic motivation to engage in a task. This "undermining effect" has timely practical implications, given the burgeoning of performance-based incentive systems in contemporary society. It also presents a theoretical challenge for economic and reinforcement learning theories, which tend to assume that monetary incentives monotonically increase motivation. Despite the practical and theoretical importance of this provocative phenomenon, however, little is known about its neural basis. Herein we induced the behavioral undermining effect using a newly developed task, and we tracked its neural correlates using functional MRI. Our results show that performance-based monetary reward indeed undermines intrinsic motivation, as assessed by the number of voluntary engagements in the task. We found that activity in the anterior striatum and the prefrontal areas decreased along with this behavioral undermining effect. These findings suggest that the corticobasal ganglia valuation system underlies the undermining effect through the integration of extrinsic reward value and intrinsic task value.

crowding-out effect | dopamine | midbrain | neuroeconomics

Performance-based incentive systems have long been part of the currency of schools and workplaces. This predominance of incentive systems may reflect a widespread cultural belief that performance-based reward is a reliable and effective way to enhance motivation in students and workers. However, classic psychological experiments have repeatedly revealed that performance-based reward can also undermine people's intrinsic motivation (1–6), that is, motivation to voluntarily engage in a task for the inherent pleasure and satisfaction derived from the task itself (3–5). In a typical experiment of this "undermining effect" [also called the "motivation crowding-out effect" (7–9) or "overjustification effect" (2)], participants are randomly divided into a performance-based reward group and a control group, and both groups work on an interesting task. Participants in the performance-based reward group obtain (or expect) reward contingent on their performance, whereas participants in the control group do not. After the session, participants are left to engage in any activity, including more of the target task if they wish, for a brief period when they believe they are no longer being observed (i.e., "free-choice period"). A number of studies (4–6) found that the performance-based reward group spends significantly less time than the control group engaging in the target activity during the free-choice period, providing evidence that the performance-based reward undermines voluntary engagement in the task (i.e., intrinsic motivation for the task).

The undermining effect challenges normative economic theories, which assume that raising monetary incentives monotonically increases motivation and, more importantly, that increasing and then removing monetary incentives does not disturb underlying intrinsic motivation (7–9). It also challenges traditional operant learning theory and reinforcement learning theory, which currently constitutes the fundamental theoretical framework for human decision making (10–12). These theories basically predict that performance-based rewards increase the likelihood that the behavior will be voluntarily performed again. It should be noted that, as thoroughly discussed in the literature (4, 5, 13), the undermining effect cannot be explained by the operant conditioning concept of extinction as a result of the withdrawal of reward (i.e., because the reward is no longer promised in the free-choice period, the reinforced response is extinguished and this produces the undermining effect) in several respects. Most importantly, the extinction account predicts that, when rewards are no longer in effect, behavior should revert to its original rate (the baseline) and never decrease below that level (14, 15). Studies on the undermining effect, however, showed less voluntary engagement in the target task in the performance-based reward group than in the control group, which serves as the strict baseline in randomized control design (5, 16). Given that normative economic theories and standard reinforcement learning theory have difficulty explaining the undermining effect, a better understanding of this effect has the potential to enrich and give new insight to these broad research fields (17). However, the neural basis of this provocative and important phenomenon remains unknown.

A source of intrinsic motivation is the intrinsic value of achieving success on a given task (3, 5). As such, the undermining effect may involve the interaction of two different types of subjective values when one succeeds at a task: the extrinsic value of obtaining a reward and the intrinsic value of achieving success. Many neuroscience studies have revealed that a dopaminergic reward network plays a pivotal role in representing and updating various types of subjective valuation (10–12, 18–24). In particular, recent studies have suggested that activation in response to feedback in the anterior part of the striatum (caudate head) is modulated by one's subjective belief in determining the outcome (23, 25), which is considered a key psychological factor in the undermining effect (3–5). Previous studies have also suggested that the midbrain, which has a strong anatomical connection with the anterior striatum (19), is responsive to both monetary reward feedback and cognitive feedback (feedback without monetary reward)—feedbacks that are related to the undermining effect (26, 27). Therefore, we expected that the undermining effect may manifest as brain activity changes in the reward network, especially in the anterior striatum and midbrain, in response to task feedback.

We expected the lateral prefrontal cortex (LPFC) to be another key structure mediating the undermining effect. When confronting an upcoming task, people tend to be more engaged in mental preparation for tasks with higher value (28, 29). As the LPFC is the center for the preparatory cognitive control to achieve goals

Author contributions: K. Murayama, M.M., and K. Matsumoto designed research; K. Murayama, M.M., K.I., and K. Matsumoto performed research; M.M. and K.I. contributed new reagents/analytic tools; K. Murayama analyzed data; and K. Murayama wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: matsumot@lab.tamagawa.ac.jp and murakou@orion.ocn.ne.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1013305107/-/DCSupplemental.

(30–34), and this function has been shown to be modulated by task value (28, 34), the undermining effect may be accompanied by a decrease in LPFC activity upon presentation of the task cue. Here, by using functional MRI (fMRI), we report evidence that these areas are involved in the undermining effect of monetary reward on intrinsic motivation.

The undermining effect is applicable only to interesting tasks that have an intrinsic value of achieving success (3–6). We developed a stopwatch (SW) task in which participants were presented with an SW that starts automatically, and the goal was to press a button with the right thumb so that the button press fell within 50 ms of the 5-s time point (Fig. 1*A*). A point was added to their score when they succeeded. A series of pilot studies showed that the SW task is moderately challenging and inherently interesting to Japanese university students (details provided in *Materials and Methods*). The control task was a watch-stop (WS) task, in which participants passively viewed a SW and were asked to simply press a button when it automatically stopped (Fig. 1*A*). Success and failure were not defined in this task; therefore the WS task was less interesting than the SW task. Both tasks were pseudorandomly intermixed and preceded by a cue that indicated which task to perform.

Twenty-eight participants were randomly assigned to a control group or a reward group. Participants were scanned in two separate sessions (Fig. 1*B*). Before the first session, participants in the reward group were told that they would obtain 200 Japanese yen (approximately $2.20) for each successful trial of the SW task, and indeed they received the performance-based reward after the session. Participants in the control group were told nothing about the performance-based reward and received money just for task participation after the first session. For each control group participant, the amount of monetary reward for the task was matched to that received by another participant of the same sex in the reward group; thus, it was unrelated to the control participant's own task performance. This allowed us to examine the effect of performance-based reward apart from the amount of monetary reward offered. After being released from the scanner and receiving the monetary reward, participants were left alone in a quiet room for 3 min, where they could freely spend time playing the SW or WS task on a computer, read several booklets, or anything else (i.e., free-choice period). The number of times participants played the SW task during this free-choice period was used as the index of intrinsic motivation toward the task (1–6).

To track the brain activity associated with the undermining effect, we asked participants in both groups to perform the SW and WS tasks again after the free-choice period and without performance-based reward inside the scanner (second session; Fig. 1*B*). Both groups of participants were explicitly told in advance that no performance-based rewards would be provided. After being released from the scanner, the second free-choice period followed to confirm that the undermining effect persisted through the second scanning session.

Results

Behavioral Results. We conducted a 2 × 2 mixed ANOVA on the number of times the voluntary SW task was played, with period (first or second free-choice period; within-subject) and group (control or reward group; between-subject) as factors. As predicted, the main effect of group was significant, ($F_{1,26}$ = 6.59, P = 0.016). This result indicates the presence of the undermining effect: participants in the reward group played the SW task during the free-choice period significantly fewer times than did those in the control group (Fig. 1*C*). A significant group difference was observed in both the first and the second free-choice periods (P < 0.05; *SI Results*). To the contrary, neither the main effect of period nor the session-by-group interaction was significant (F < 1, P > 0.32). This suggests that no overall increase or decrease in the voluntary SW task play was observed and that the pattern of change in the voluntary SW task play did not differ across the groups. In fact, we observed no significant increase or decrease of the voluntary SW task play from the first to the second free-choice period in either group (P > 0.19).

We also conducted the same 2 × 2 mixed ANOVA on the number of times participants played the WS task during the free-choice period. Neither the main effects nor the interaction was significant (F < 1, P > 0.34; Fig. 1*C*). The numbers were quite small, suggesting that the WS task was not interesting to the participants.

fMRI Results. In the fMRI analysis, we were interested in finding significant session-by-group interactions, which means that changes in activation across sessions showed different patterns between the two groups. Thus, we applied a 2 × 2 mixed ANOVA

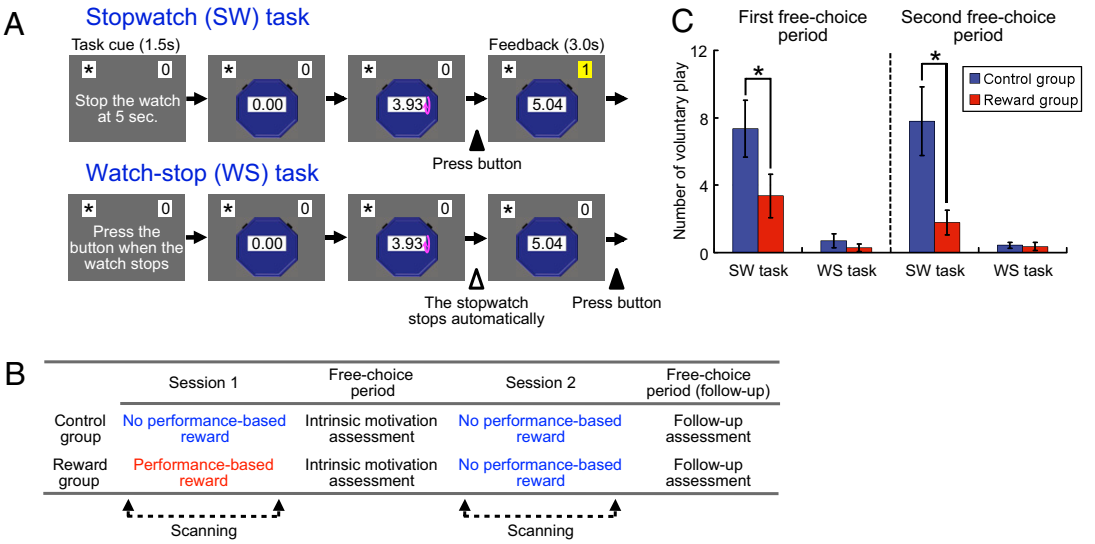

Fig. 1. Experimental protocol and behavioral results. (*A*) Illustration of SW and WS tasks. (*B*) Depiction of the experimental procedure. (*C*) Means and SEs of the number of times participants voluntarily played the SW and WS tasks during the first and the second free-choice periods. Performance-based reward undermined the intrinsic motivation for the SW task for both free-choice periods (Mann–Whitney U = 52.5 and 54.5; P < 0.05).

with session (first or second session) and group (control or reward group) as factors. The significant interactions reported here were based on the regions that survived both a whole-brain analysis ($P < 0.001$, uncorrected) and small-volume correction analysis ($P < 0.05$; details in *Materials and Methods*).

We first focused on a feedback period to examine the neural responses to the success feedback versus the failure feedback. A one-sample t test in the first session showed that the bilateral anterior striatum (caudate head) and midbrain were significantly activated, regardless of the group (P values < 0.05, small-volume corrected). This result indicates that the success feedback in the experimental task we developed involves reward network activation, regardless of whether the feedback was accompanied with monetary reward. This is consistent with previous work (21, 23, 25, 26) and supports the validity of our experimental task for examining brain activation in response to task feedback.

In the 2 × 2 ANOVA, as expected, the bilateral striatum activation showed a significant interaction between session (first or second session) and group (control or reward group) that is a striking parallel with the behavioral undermining effect ($P < 0.05$, small-volume-corrected; Fig. 2 and Fig. S1). During the first session, significant anterior striatum activation was observed in both groups: one-sample t_{13} values of 6.61 (control) and 8.43 (reward); $P < 0.01$ for both. However, the activation was significantly greater in the reward group than in the control group: two-sample $t_{26} = 3.30$ ($P < 0.01$). Previous studies have implied that the striatum functions as a hub of the human valuation process, by converting and integrating different types of reward values onto a common scale (11). Our result can be interpreted by this view such that the significant positive activation in the control group reflects the intrinsic value of achieving success (23, 25) and this activation was elevated by the additional performance-based monetary reward in the reward group. Importantly, whereas this activity during the second session was sustained in the control group (one-sample $t_{13} = 7.33$, $P < 0.01$; no between-session change was observed, $P = 0.41$), there was a dramatic decrease in activation of the bilateral anterior striatum in the reward group, and the activation was no longer significant (one-sample $t_{13} = 0.41$, $P = 0.69$; decrease in the activity from the first to the second session was significant, paired $t_{13} = 7.35$, $P < 0.01$). As a result, the between-group difference in the anterior striatal activation was reversed from the first session to the second session and became significantly smaller in the reward group compared with the control group during the second session (two-sample $t_{26} = 3.75$, $P < 0.01$). Also as predicted, the midbrain showed a similar pattern of interaction ($P < 0.05$, small-volume-corrected; Fig. 3), consistent with the strong anatomical connection between the midbrain and anterior striatum (19, 26).

We next focused on a task cue period to investigate the brain activity associated with preparatory cognitive control in the SW task relative to the WS task. A one-sample t test in the first session revealed that the right LPFC was significantly activated regardless of the group ($P < 0.05$, small-volume corrected). This result indicates that the participants were cognitively engaged in the SW task relative to the WS task when the task cue was presented. This is consistent both with the observation that participants were more willing to engage in the SW task in the free-choice periods and with previous findings that the LPFC is activated in response to a task cue with high value (28, 34). In other words, this finding supports the validity of our experimental task for examining LPFC activation in response to task cue.

In the 2 × 2 ANOVA, as expected, the right LPFC showed a significant session-by-group interaction that is also a parallel with the behavioral undermining effect ($P < 0.05$, small-volume-corrected; Fig. 4). During the first session, the LPFC in the reward group showed significantly larger activation than that in the control group (two-sample $t_{26} = 2.62$, $P < 0.05$), suggesting that participants in the reward group prepared for the SW task more actively than those in the control group when they saw a task cue. However, during the second session, although significant activity in the control condition was sustained in the second session (one-sample $t_{13} = 2.53$, $P < 0.05$), the activation became significantly smaller in the reward group than in the control group (two-sample $t_{26} = 2.27$, $P < 0.05$), and the activation was no longer significant ($P = 0.43$). This result may indicate that the participants in the reward group were not motivated to prepare for the SW task during the second session in comparison with the control group participants. The bilateral striatum also showed a significant interaction for the task cue period, but unlike the activation pattern in the feedback period, no significant between-group difference in activation was detected during the second session (Fig. S2).

Table S1 (for the feedback period) and Table S2 (for the task cue period) list all regions displaying a significant session-by-group interaction in a whole-brain analysis ($P < 0.001$, uncorrected, $k > 5$). The tables also describe the results of simple main effect analyses and one-sample t tests that quantify the pattern of interaction as we conducted for the striatum, midbrain,

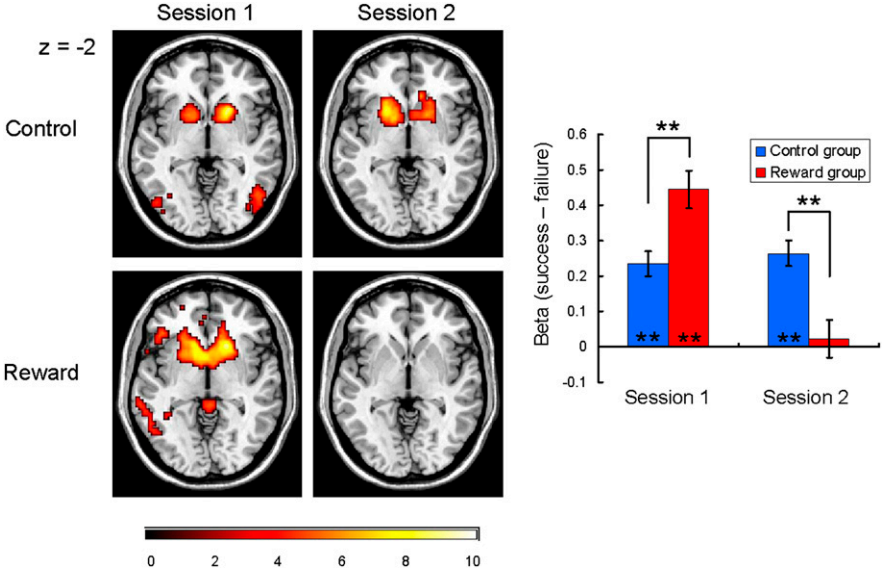

Fig. 2. Bilateral striatum responses elicited by success trials relative to failure trials plotted for each session/group. *Left*: Activations superimposed on transaxial sections ($P < 0.001$, one-sample t test for display). *Right*: Mean contrast values and SEs of the bilateral striatum (averaged) activation are plotted. During the first session, significant bilateral striatum activation was observed in both groups, although the activation was significantly greater in the reward group than in the control group (two-sample $t_{26} = 3.30$, $P < 0.01$). In contrast, during the second session, whereas the control group sustained significant activity, the activation of the bilateral striatum in the reward group decreased significantly below that of the control group (two-sample $t_{26} = 3.75$, $P < 0.01$) and the activation was no longer significant. This striatal response pattern is in parallel with the behavioral undermining effect. Asterisks represent the statistical significance of one-sample/two-sample t tests (**$P < 0.01$).

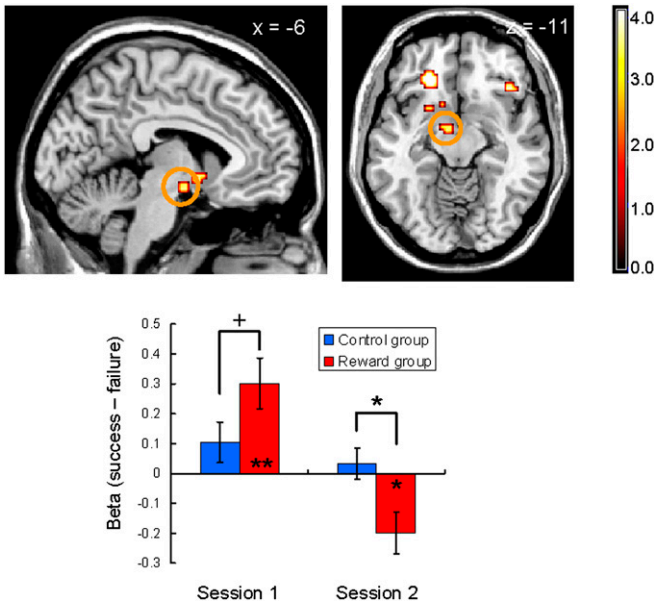

Fig. 3. Midbrain activation (peak at −9, −7, −11) detected in the session-by-group interaction during the feedback period (success trials minus failure trials; $P < 0.05$, small-volume-corrected; the image is shown at $P < 0.001$, uncorrected). Neural responses are displayed in sagittal and transaxial formats. The midbrain was activated when performance-based monetary reward was expected (during the first session; two-sample $t_{26} = 1.80$, $P < 0.10$), but the activation decreased significantly below the control group in the second session (two-sample $t_{26} = 2.63$, $P < 0.05$). Asterisks represent the statistical significance of one-sample/two-sample t tests ($^{+}P < 0.10$, $^{*}P < 0.05$, $^{**}P < 0.01$).

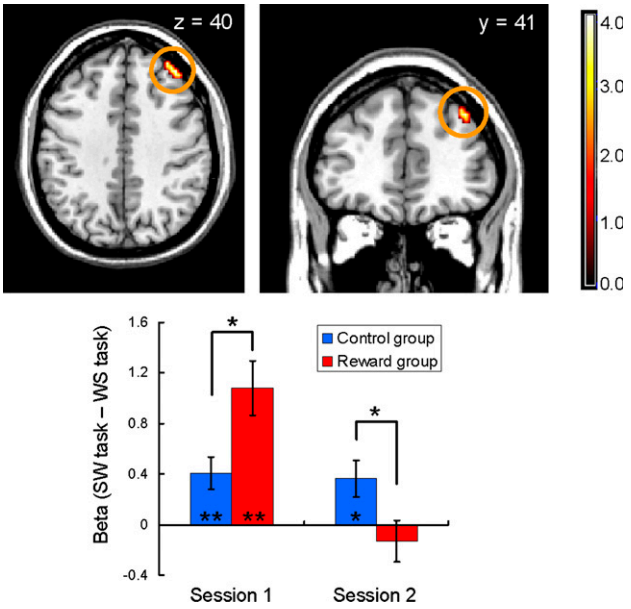

Fig. 4. Right LPFC activation (peak at 39, 41, 40) detected in the session-by-group interaction during the task cue period ($P < 0.05$, small-volume-corrected; image is shown at $P < 0.001$, uncorrected for display). Neural responses are displayed in transaxial and coronal formats. The bar plot represents mean contrast values and SEs for each session/group. During the first session, the LPFC in the reward group showed significantly larger activation than that in the control group (two-sample $t_{26} = 2.62$, $P < 0.05$). However, the activation became significantly smaller in the reward group than in the control group during the second session (two-sample $t_{26} = 2.27$, $P < 0.05$).

and LPFC (*SI Results* includes additional analyses focusing on a possible sex effect).

Brain–Behavior Relation. We conjecture that the observed decreases in activity of the anterior striatum, midbrain, and LPFC are collectively related to the undermining effect. In fact, the magnitudes of the decreases in activation in these regions were highly correlated (mean $r = 0.65$). Accordingly, we calculated a "neural undermining index"—a composite score representing the between-session decreases in activity for these regions, and regressed it on the voluntary SW task play during the free-choice period. Specifically, we computed the magnitude of a decrease in activation by subtracting the contrast value in the second session from the contrast value in the first session for each region of interest (i.e., the striatum and the midbrain for feedback period and the LPFC for cue period) and submitted these values to principal component analysis. Principal component analysis is a statistical technique to compute optimal and reliable composite scores of a set of variables that are less susceptible to random noise by taking into account their variance and covariance (35). The first principal component explained a substantial portion (74%) of the total variance, and we used this component score as the neural undermining index.

As expected, regression analysis revealed a significant negative relationship between the amount of voluntary SW play and the neural undermining index in the reward group (standardized $\beta = -0.49$, $P = 0.037$, one-tailed), indicating that those who did not voluntarily try the SW task during the free choice period showed a larger decrease in activation of the corticobasal ganglia network (Fig. 5). The regression coefficient remained significant even when the confidence interval was based on a (bias-corrected) bootstrapping method to correct for the potential statistical biases resulting from nonnormality and outliers (36). The magnitude of relationship is large according to the Cohen established effect size criterion (37). In contrast, the relationship in the control group was not significant (standardized $\beta = 0.09$, $P = 0.75$). An additional regression analysis including group, the number of voluntary SW play trials, and their interaction as independent variables showed significant interaction ($P = 0.045$), indicating that the aforementioned regression coefficients in the reward and control groups were significantly different.

Discussion

Our study provides evidence that the corticobasal ganglia valuation system plays a central role in the undermining effect. Specifically, our neuroimaging results suggest that, when performance-based reward is no longer promised, (*i*) people do not feel subjective value in succeeding in the task, as indicated by the dramatic decreases in the activation of the striatum and midbrain in response to the success feedback; and (*ii*) they are not motivated to show cognitive engagement in facing the task, as indicated by the decrease in the LPFC activation in response to the task cue. A number of theories have been proposed to explain the undermining effect from value-based and cognitive perspectives (5, 6). Our findings clearly indicate that value-driven and cognitive processes are involved in the undermining effect, and they are linked. Notably, activation in the anterior part of the striatum, which has been implicated in subjective belief in determining outcomes (23, 25), corresponds particularly well to the pattern observed in the behavioral undermining effect. This lends support for the recent psychological theory that the undermining effect is closely linked to a decreased sense of self-determination (3–5).

The precise neural and computational mechanism that accounts for the striatal signal decrease in the reward group merits future inquiry. One explanation is that the striatum, in which incommensurable subjective values are aligned on a unidimensional common scale, integrates the intrinsic value of task success and monetary reward value through relative comparison and rescaling processes (38). Given the relatively stronger salience of monetary reward, the

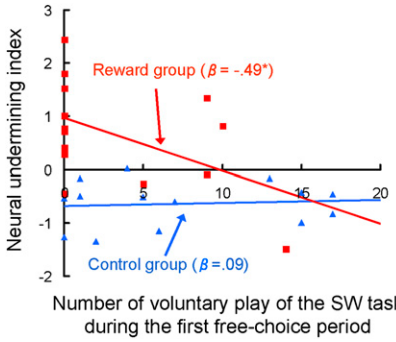

Fig. 5. Relationship between behavioral choice during the first free-choice period and the neural undermining index. Significant negative relationship was observed in the reward group (β = –0.49, $P = 0.037$, one-tailed), indicating that those who did not voluntarily try the SW task during the free-choice period showed a larger decrease in activation of the corticobasal ganglia network. The relationship in the control group was not significant (β = 0.09, $P = 0.75$). Blue triangles represent participants in the control group; red squares represent participants in the reward group.

rescaled value of task success could become smaller than the original magnitude. In other words, the strong incentive value of monetary reward pushed down the intrinsic value of task success. As a result, when the monetary reward was no longer promised, the intrinsic task value was underestimated, resulting in decreased motivation relative to the control group (i.e., less frequent play of the SW task in the free-choice period). This interpretation underscores the importance for future empirical and theoretical work addressing the human value integration process (38, 39).

Neuroscience research has made considerable progress by incorporating concepts of motivation (40), yet most research to date has been confined to extrinsic rewards such as food or money. In comparison with our knowledge of extrinsic rewards, little light has been shed on intrinsic sources of motivation, and much less on the integration of the two. However, given the burgeoning of performance-based incentive systems in contemporary society, the interaction of these motivations is gaining practical importance in guiding human behavior. We believe an expanded understanding of this integration process is a key piece in aligning the undermining effect with leading economic and learning theories, and in reaching a deeper understanding of human behavior in general.

Materials and Methods

Participants. Twenty-eight right-handed healthy participants [mean age, 20.6 ± 1.1 y (SD); 10 male and 18 female] recruited from a pool of Tamagawa University (Tokyo, Japan) students took part in the experiment. Participants were randomly assigned to a control group ($n = 14$) or a reward group ($n = 14$). All participants gave informed consent for the study and the protocol was approved by the Ethics Committee of Tamagawa University.

Experimental Tasks. The undermining effect can be observed only when a task is interesting and has intrinsic value of achieving success; with boring tasks, there is little or no intrinsic motivation to undermine (5). Accordingly, an SW task was developed (Fig. 1*A*) to meet this criterion. A series of pilot studies were conducted to determine the time window for success so that participants can succeed on approximately half the trials on average. Previous literature indicated that people obtain the greatest sense of achievement for the tasks of intermediate difficulty (41, 42). In addition, this rate of success allows a sufficient number of success or failure trials to be obtained for proper fMRI statistical analysis. The participant's total score was displayed in the upper right corner of the display area, and when the participant succeeded in stopping the SW display between 4.95 s and 5.05 s, a point was added to their score (1,500 ms after the button press) and the updated score panel flashed for 1,500 ms. Another pilot study using an independent university student sample ($n = 37$) revealed that this task is sufficiently interesting without any extrinsic incentives (mean enjoyment rating, 4.14; SD, 0.82, on a five-point Likert scale). We also developed a WS task as the control task (Fig. 1*A*). Because success and failure were not defined in this task, no point was added on their response.

The experiment was composed of two separate scanning sessions (approximately 18 min each) and each session consists of 30 SW and 30 WS trials, which were pseudorandomly intermixed with the interstimulus interval alternated between 1,000 and 5,000 ms. Both tasks were preceded by a cue that indicates which of two tasks is to be performed in the next trial. The cue was presented for 1,500 ms and the SW starts 3,000 ms after the cue onset. The timing of the stop for a WS trial is approximately matched (alternating based on random noise) to the time achieved in a prespecified SW trial.

Experimental Procedure. The experimental sessions were conducted individually. Upon arriving at the experiment room, participants were greeted by two experimenters. The experiment consisted of two separate sessions, each of which was conducted by one of the experimenters. We decided to use a different experimenter for each session to prevent the participants from being aware of the relationship between these two sessions. In the postexperimental interviews, no participant reported noticing the fact that the two experiments were conducted under a common purpose. Regardless of the experimental conditions, the payment for the task participation in the second session was provided to the participants before the experiment (2,000 Japanese yen, approximately equal to $22). This procedure allowed us to avoid any possible monetary incentive effects in the second session, the critical session to capture the brain activity of the undermining effect.

Before the first scanning session, participants in the reward group were informed that they would receive 200 Japanese yen (approximately equal to $2.20) for each point they obtained during the session. In contrast, no mention was made about the performance-based monetary reward in the control group. The instruction was provided through a computer program to prevent possible experimenter bias.

On completing the first scanning session, participants were released from the scanner and led to a small waiting room, where participants in the reward group were provided with the performance-based monetary reward. Participants in the control group received payment for task participation. For each participant in the control group, the amount of monetary reward for the task was matched to that of a participant in the reward group (i.e., both groups received the same amount of reward).

After the participants confirmed the amount of money they received, the experimenter asked the participants to wait for a couple of minutes because the other experimenter needed a few more minutes to prepare for the next experiment. There was a computer in the room and participants could freely choose to play the SW or WS task as many times as they wanted with that computer. There were also a few booklets on a desk and participants could freely read them. During this free-choice period, participants believed they were no longer observed by the experimenters, but the computer program confidentially recorded the number of SW and WS task trials they voluntarily played. The number of trials they played on the SW task was used as the index of intrinsic motivation. This is a standard way of assessing intrinsic motivation and has been used in many previous experiments on the undermining effect (1–5). After exactly 3 min, the door of the waiting room was opened and the participants were led to the scanning room again.

The second session was ostensibly conducted by the second experimenter. Before the task, participants were instructed that they would do the SW and WS tasks used in the previous experiment, but it was emphasized that the purpose of the experiment was completely independent. Both groups of participants were also explicitly told that no performance-based rewards would be provided in this experiment. As such, participants in both groups did not expect any performance-based rewards in the second session. After the scanning session, participants were released from the scanner and exposed to a 3-min free-choice period again.

fMRI Data Acquisition. The functional imaging was conducted using a 3-T Trio A Tim MRI scanner (Siemens) to acquire gradient echo T2*-weighted echo-planar images (EPI) with blood oxygenation level-dependent contrast. Forty-two contiguous interleaved transversal slices of EPI images were acquired in each volume, with a slice thickness of 3 mm and no gap (repetition time, 2,500 ms; echo time, 25 ms; flip angle, 90°; field of view, 192 mm^2; matrix, 64 × 64). Slice orientation was tilted –30° from the AC–PC line. We discarded the first three images before data processing and used statistical analysis to compensate for the T1 saturation effects.

fMRI Data Analysis. Image analysis was performed by using Statistical Parametric Mapping software (version 8; http://www.fil.ion.ucl.ac.uk). Images were corrected for slice acquisition time within each volume, motion-corrected with realignment to the first volume, spatially normalized to the standard Montreal

Neurological Institute EPI template, and spatially smoothed using a Gaussian kernel with a full width at half maximum of 8 mm.

For each participant, the blood oxygen-level dependent responses across the scanning run (including both sessions) were modeled with a general linear model. The model included the following regressors of interest: presentation of success feedback in the SW task, presentation of failure feedback in the SW task, presentation of SW task cue, and presentation of WS task cue. The motion parameters, error trials, and session effects were also included as regressors of no interest. The regressors (except for the motion parameters and the session effects) were calculated using a boxcar function convolved with a hemodynamic-response function. The estimates were corrected for temporal autocorrelation by using a first-order autoregressive model. To investigate the feedback effects and cue effects, our primary focus of interest, two contrast values were calculated: (*i*) contrast between success feedback and failure feedback effects (i.e., success minus failure), and (*ii*) contrast between SW task cue and WS task cue effects (i.e., SW minus WS).

We conducted a second-level, whole-brain 2×2 mixed ANOVA with session (first or second session; within-subject) and group (control or reward group; between-subject) as factors, once on the success/failure contrasts and once on the SW/WS contrasts. A number of regions showed a significant session-by-group interaction ($P < 0.001$, uncorrected, $k > 5$ voxels), including anterior striatum, midbrain (for the feedback period), and LPFC (for the task cue period), our primary region of interest (Tables S1 and S2). To confirm the reliability of the significant interaction effects obtained in the regions for which we had an a priori hypothesis (the anterior striatum and midbrain for the feedback period and the LPFC for the task cue period), we also performed a small-volume correction analysis with a corrected significance threshold of $P < 0.05$ within a 12-mm sphere centered on the coordinates identified in the previous empirical studies or meta-analyses (25, 26, 33, 43). All regions of interest survived this analysis.

To quantify the pattern of interaction, we further conducted a series of post-hoc analyses. Specifically, we extracted the contrast values from a 3-mm sphere centered on the peak voxel of each region using rfxplot (44), and subjected these extracted values to a series of simple main-effect analyses (i.e., test for the between-group difference within each session) and one-sample *t* tests (i.e., test for the significance of the absolute contrast values for each session/group).

ACKNOWLEDGMENTS. We thank J. Tanji, C. Camerer, J. O'Doherty, K. Samejima, H. Kim, A. Przybylski, and R. Pekrun for helpful comments; K. D'Ardenne and M. J. Tyszka for providing technical information; T. Haji, R. Iseki, S. Bray, and J. Gläscher for technical advice; J. Helen for English proofing; and Y. Otake for research assistance. This study was supported by Grant-in-Aid for Scientific Research C#21530773 (to K. Matsumoto), Grand-in-Aid for Scientific Research on Innovative Areas 22120515 (to K. Matsumoto); a Tamagawa University Global Center of Excellence grant from the Ministry of Education, Culture, Sports, Science and Technology, Japan; and an Alexander von Humboldt Foundation fellowship (to K. Murayama).

1. Deci EL (1971) Effects of externally mediated rewards on intrinsic motivation. *J Pers Soc Psychol* 18:105–115.
2. Lepper MR, Greene D, Nisbett RE (1973) Undermining children's intrinsic interest with extrinsic rewards: A test of the "overjustification" hypothesis. *J Pers Soc Psychol* 28:129–137.
3. Ryan RM, Mims V, Koestner R (1983) Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *J Pers Soc Psychol* 45:736–750.
4. Deci EL, Koestner R, Ryan RM (1999) A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychol Bull* 125:627–668, discussion 692–700.
5. Deci EL, Ryan RM (1985) *Intrinsic Motivation and Self-Determination in Human Behavior* (Plenum, New York).
6. Morgan M (1984) Reward-induced decrements and increments in intrinsic motivation. *Rev Educ Res* 54:5–30.
7. Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *J Risk Uncertain* 19:7–42.
8. Frey BS, Jegen R (2001) Motivation crowding theory. *J Econ Surv* 15:589–611.
9. Kreps D (1997) Intrinsic motivation and extrinsic incentives. *Am Econ Rev* 87:359–364.
10. Rushworth MFS, Mars RB, Summerfield C (2009) General mechanisms for making decisions? *Curr Opin Neurobiol* 19:75–83.
11. Montague PR, Berns GS (2002) Neural economics and the biological substrates of valuation. *Neuron* 36:265–284.
12. Dayan P, Niv Y (2008) Reinforcement learning: the good, the bad and the ugly. *Curr Opin Neurobiol* 18:185–196.
13. Grolnick WS (2003) *The Psychology of Parental Control: How Well Mean Parenting Backfires* (Lawrence Erlbaum Associates, Hillsdale, NJ).
14. Skinner BF (1938) *The Behavior of Organisms: An Experimental Analysis* (Appleton-Century-Crofts, New York).
15. Hull C (1943) *Principles of Behavior* (Appleton-Century-Crofts, New York).
16. Kirk RE (1995) *Experimental Design: Procedures for the Behavioral Sciences* (Brooks/Cole, Pacific Grove, CA), 3rd Ed.
17. Camerer CF, Loewenstein G, Rabin M (2004) *Advances in Behavioral Economics* (Princeton Univ Press, Princeton, NJ).
18. Seymour B, et al. (2004) Temporal difference models describe higher-order learning in humans. *Nature* 429:664–667.
19. Haber SN, Knutson B (2010) The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology* 35:4–26.
20. Kringelbach ML, Berridge KC (2010) *Pleasures of the Brain* (Oxford Univ Press, London).
21. Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human striatum. *Neuron* 58:284–294.
22. Tobler PN, Fletcher PC, Bullmore ET, Schultz W (2007) Learning-related human brain activations reflecting individual finances. *Neuron* 54:167–175.
23. Tricomi EM, Delgado MR, Fiez JA (2004) Modulation of caudate activity by action contingency. *Neuron* 41:281–292.
24. Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annu Rev Psychol* 57:87–115.
25. Tricomi E, Delgado MR, McCandliss BD, McClelland JL, Fiez JA (2006) Performance feedback drives caudate activation in a phonological learning task. *J Cogn Neurosci* 18:1029–1043.
26. D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* 319:1264–1267.
27. Aron AR, et al. (2004) Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *J Neurophysiol* 92:1144–1152.
28. Jimura K, Locke HS, Braver TS (2010) Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts. *Proc Natl Acad Sci USA* 107:8871–8876.
29. Castel AD, Balota DA, McCabe DP (2009) Memory efficiency and the strategic control of attention at encoding: impairments of value-directed remembering in Alzheimer's disease. *Neuropsychology* 23:297–306.
30. Matsumoto K, Suzuki W, Tanaka K (2003) Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science* 301:229–232.
31. Bunge SA (2004) How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cogn Affect Behav Neurosci* 4:564–579.
32. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.
33. Wager TD, Smith EE (2003) Neuroimaging studies of working memory: A meta-analysis. *Cogn Affect Behav Neurosci* 3:255–274.
34. Leon MI, Shadlen MN (1999) Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron* 24:415–425.
35. Dunteman GH (1989) *Principal Component Analysis* (Sage, Thousand Oaks, CA).
36. Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap* (Chapman & Hall/CRC, Boca Raton, FL).
37. Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Hillsdale, NJ), 2nd Ed.
38. Seymour B, McClure SM (2008) Anchors, scales and the relative coding of value in the brain. *Curr Opin Neurobiol* 18:173–178.
39. FitzGerald TH, Seymour B, Dolan RJ (2009) The role of human orbitofrontal cortex in value comparison for incommensurable objects. *J Neurosci* 29:8388–8395.
40. Berridge KC (2004) Motivation concepts in behavioral neuroscience. *Physiol Behav* 81:179–209.
41. Atkinson JW (1957) Motivational determinants of risk-taking behavior. *Psychol Rev* 64:359–372.
42. Csikszentmihalyi M (1990) *Flow: The Psychology of Optimal Experience* (Harper and Row, New York).
43. Tricomi E, Fiez JA (2008) Feedback signals in the caudate reflect goal achievement on a declarative memory task. *Neuroimage* 41:1154–1167.
44. Gläscher J (2009) Visualization of group inference data in functional neuroimaging. *Neuroinformatics* 7:73–82.

Supporting Information

Murayama et al. 10.1073/pnas.1013305107

SI Results

Randomization Test. To confirm that our behavioral results on the undermining effect were not obtained by chance (e.g., preexisting heterogeneity of the groups), we conducted a randomization test (1) as an additional analysis (using a Monte Carlo simulation; N = 10,000). This test allows us to examine the probability of obtaining the observed between-group difference in the free-choice behavior under the null hypothesis when we use the random assignment procedure. The obtained P values were less than 0.05 ($P = 0.041$ for the first free-choice period and $P = 0.042$ for the second free-choice period), indicating that our behavioral results cannot be attributable to the accidental heterogeneity of the groups.

SW Task Performance. During the first session, SW task performance was significantly better in the reward group than in the control group ($t_{26} = 2.35$, $P < 0.05$; M = 13.07 and 17.79, SD = 5.28 and 5.34). In the second session, the difference became weaker and was no longer significant ($t_{26} = 1.72$, $P = 0.10$; M = 13.07 and 16.79, SD = 6.53 and 4.74), but there is still a trend that participants in the reward condition showed better performance in the SW task. A previous meta-analysis has shown that intrinsic motivation conferred an advantage only for complex, cognitive tasks, but not for simple, noncognitive tasks (2). Therefore, given that the SW task is a noncognitive, motor-response task, the results in the SW task performance are consistent with the previous observations. Indeed, the correlation between the SW task performance in the second session and the number of voluntary plays of the SW task (after the first session) was not significant in either of the groups ($r = -0.01$, $P = 0.98$ for the control group; $r = -0.09$, $P = 0.77$ for the reward group), suggesting that the SW task performance does not reflect participants' intrinsic motivation for the task.

It should also be noted that the skill acquired in motor-response tasks like the SW task is likely to be resistant to the loss (3, 4). This could explain why the participants in the reward group continued to show superior performance (although nonsignificant) in the SW task in the second session. In fact, the correlation between the SW task performance in the first and second sessions is very high ($r = 0.78$, $P < 0.0001$), suggesting the high stability of the SW task performance.

Sex Difference. In the behavioral analysis, we conducted a 2 (free-choice period: first or second time) × 2 (group: control or reward group) × 2 (sex: male or female) mixed ANOVA to investigate a possible sex difference in the behavioral undermining effect. None of the interactions involving sex was significant (P values > 0.20). In the fMRI analysis, we conducted a 2 (session: first or second session) × 2 (group: control or reward group) × 2 (sex: male or female) ANOVA to examine whether the session-by-group interaction (our primary effect of interest) was affected by sex ($P < 0.001$, uncorrected, $k > 5$ voxels). No significant three-way interaction (i.e., session × group × sex) was detected in the striatum, midbrain, or LPFC for the task cue or feedback period. These results, taken together, suggest that our behavioral and fMRI findings are not dependent on participants' sex, although our analyses may be underpowered as a result of the small sample size.

1. Edgington ES, Onghena P (2004) *Randomization Tests* (Chapman & Hall/CRC, Boca Raton, FL), 4th Ed.
2. Utman CH (1997) Performance effects of motivational state: a meta-analysis. *Pers Soc Psychol Rev* 1:170–182.
3. Eslinger PJ, Damasio AR (1986) Preserved motor learning in Alzheimer's disease: implications for anatomy and behavior. *J Neurosci* 6:3006–3009.
4. Doyon J, Ungerleider LG (2002) Functional anatomy of motor skill learning. *Neuropsychology of Memory*, eds Squire LR, Schacter DL (Guilford Press, New York), pp 225–238.

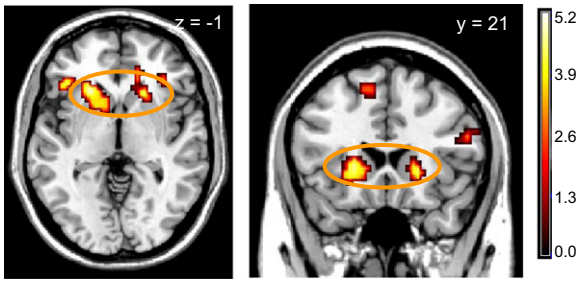

Fig. S1. Bilateral striatum activation (peaks at 21, 20, −2 and −21, 23, 1) detected in the session-by-group interaction during the feedback period (i.e., success trials minus failure trials; $P < 0.05$, small-volume-corrected; image is shown at $P < 0.001$, uncorrected). Neural responses are displayed in transaxial and coronal formats. Plot for the individual session/group is depicted in Fig. 2.

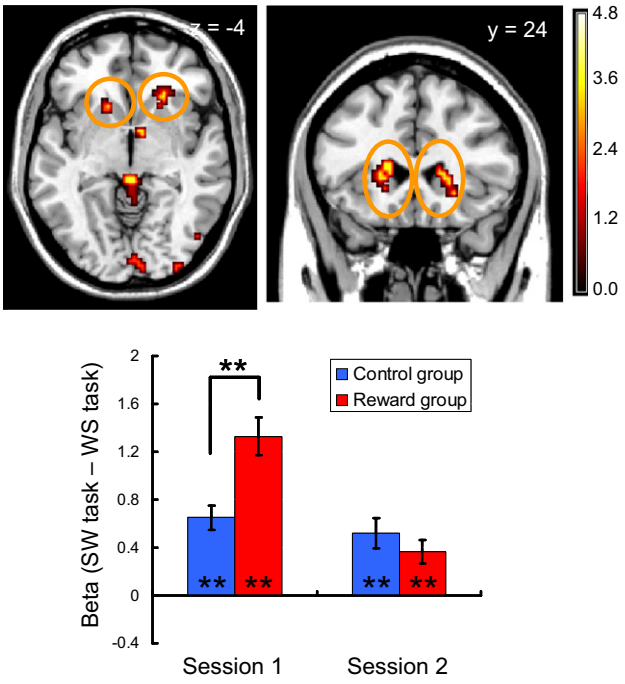

Fig. S2. Bilateral striatum activation (peaks at 15, 8, −11 and −18, 26, 10) showing a significant session-by-group interaction in response to the SW cues relative to the WS cues (image is shown at $P < 0.001$, uncorrected). Neural responses are displayed in transaxial and coronal formats. The pattern of striatal activation was different from that during the feedback period. The graph represents the averaged activation across both the right and left striatum. Asterisks represent the statistical significance of one-sample/two-sample *t* tests (**$P < 0.05$, *$P < 0.01$).

Table S1. Patterns of the session-by-group interaction during the feedback period in response to success (relative to failure) trials

Region	Peak MNI coordinates (x, y, z)			*z* value
$\underline{C_1} < \underline{R_1}$ and $\underline{C_2} > R_2$				
Right anterior striatum	21	20	−2	4.04
Left anterior striatum	−21	23	1	4.75
$\underline{C_1} < \underline{R_1}$ and $C_2 \approx R_2$				
Left inferior frontal gyrus	−42	29	−2	4.08
$C_1 < \underline{R_1}$ and $C_2 > \bar{R}_2$				
Right LPFC	57	32	19	3.79
Right inferior frontal gyrus	42	26	−11	3.52
Midbrain	−9	−7	−11	3.40
Left central OFC*	−24	35	−11	4.40
Presupplementary motor area	−12	20	52	3.63
Right inferior frontal gyrus	33	32	−2	3.38

All regions that showed a significant interaction effect ($P < 0.001$, uncorrected, $k > 5$ voxels) are categorized based on simple main effect analyses within each session. C_1, control group in the first session; R_1, reward group in the first session; C_2, control group in the second session; R_2, reward group in the second session; MNI, Montreal Neurological Institute; ≈, nonsignificant difference. Conditions that showed a significant positive activation (activation is significantly higher in responses to success trials than to failure trials) are underlined (e.g., $\underline{C_2}$). Conditions that showed a significant negative activation (activation is significantly smaller in responses to success trials than to failure trials) have a bar above them (e.g., $\bar{R}_2$).

*In the second session, a marginally significant positive activation was observed in the control group.

Table S2. Patterns of the session-by-group interaction during the task cue period in response to SW (relative to WS) trials

Region	Peak MNI coordinates (x, y, z)			*z* value
$\underline{C_1} < \underline{R_1}$ and $\underline{C_2} > R_2$				
Right LPFC	39	41	40	3.77
Cerebellum	−30	−79	−17	4.34
$\underline{C_1} < \underline{R_1}$ and $\underline{C_2} \approx \underline{R_2}$				
Right striatum	15	8	−11	3.55
Left anterior striatum	−18	26	10	3.75
Right globus pallidus	9	−1	−5	3.42
Presupplementary motor area	3	11	73	4.23
Supplementary motor area	6	−7	79	3.30
Anterior thalamus	−3	−4	7	3.25
$\underline{C_1} < \underline{R_1}$ and $C_2 \approx R_2$				
Right premotor cortex	33	−4	70	3.53
Left premotor cortex*	−30	−4	70	3.87
Right frontal pole	36	62	16	3.88
Parietal lobe†	12	−73	46	3.52
Right temporal lobe†	45	−31	−11	3.92
Left primary motor cortex	−30	−40	61	3.36
$C_1 < \underline{R_1}$ and $C_2 \approx R_2$				
Left primary motor cortex	−42	−34	61	3.52
Cerebellum	−51	−58	−20	3.47

All regions that showed a significant interaction effect ($P < 0.001$, uncorrected, $k > 5$ voxels) are categorized based on simple main effect analyses within each session. C_1, control group in the first session; R_1, reward group in the first session; C_2, control group in the second session; R_2, reward group in the second session; MNI, Montreal Neurological Institute; $\approx$, nonsignificant difference. Conditions that showed a significant positive activation (activation is significantly higher in responses to SW trials than to WS trials) are underlined (e.g., $\underline{C_2}$).

*In the second session, a significant positive activation was observed in the reward group.

†In the second session, a significant (or a marginally significant) positive effect was observed in the control group.